Finite horizon prediction of recurrent events with application to forecasts of warranty claims *

Marc Fredette
HEC Montréal

J.F. Lawless
University of Waterloo

July 6, 2006

Abstract

In this paper we present prediction methods for recurrent events which occur for individuals or units in some population. The events are modeled using flexible non-homogeneous Poisson processes and possible heterogeneity amongst the individuals is modeled using random effects. We also present effective calibration techniques which provide prediction intervals with coverage probabilities close to a desired nominal level. We apply these methods to a particular warranty data setting from the automobile industry. The number of processes in this context being very large, we emphasize methods providing rapid computation of prediction intervals.

Key Words: calibration, non-homogeneous Poisson processes, prediction intervals, random effects.

1 INTRODUCTION

Recurrent events are ubiquitous in populations and processes; consider, for example, occurrences of disease in a human population, the creation of jobs in an economic sector, or the occurrences of stoppages in production or service processes due to equipment failure. Often

*Corresponding author : Marc Fredette, Department of management sciences; HEC Montreal; 3000, Cote-Sainte-Catherine; Montreal, Canada; H3T 2A7; marc.fredette@hec.ca

there is some sort of adverse connotation to the events, as in the case of disease episodes or equipment failures, and many experimental or observational studies are directed at understanding factors affecting event occurrence, and at reducing their frequency. In many settings it is, on the other hand, important to predict the numbers of events that will occur in future time periods. For example, in the case of insurance or warranty claims in populations of contract holders, such predictions are used for fiscal planning and for taxation purposes. In testing and debugging new software systems (e.g. Dalal & McIntosh 1994), decisions about when to stop testing and to release software are influenced by predictions of the number of new bugs that would be found if testing were to continue further.

This paper deals with such prediction problems. The motivation for our work lies in the prediction of warranty claims (e.g. Robinson & McDonald 1991, Chen, Lynn & Singpurwalla 1996) or other events that occur for individual units or subjects in a population. That is, there is some population of units $i = 1, \ldots, k$ and we wish to predict the number of events for individual units, or the aggregate number across the population or a sub-population of units. Specifically, we consider what we term finite horizon aggregate prediction: the objective is to predict the total number of events in the population over a specified time period, which without loss of generality we denote as $(0, T)$, on the basis of events that have already occurred up to given times $t_i \leq T$ for the units in the population. In practice the interval $(0, T)$ may refer to either a calendar time period or to a time period that is related to the "age" of units in the population.

We now discuss a motivating example to which we will later apply the methodology developed. The context is warranty claims on manufactured products, where each of k product units sold is under warranty for a specified time T from the date of sale. Let t represent the age of a unit, meaning time since it was sold (here expressed for the sake of discussion in days), and let $\mathbf{N}_i(u, v)$ denote the number of claims in the age interval $u < t \leq v$. The objective is then to predict

$$\mathbf{N}_+(0, T) \quad = \quad \sum_{i=1}^{k} \mathbf{N}_i(0, T), \tag{1}$$

the total number of claims that will eventually accrue in the population. Of course, $\mathbf{N}_+(0, T)$ will eventually be known, once every unit is sold and completes its warranty period. However, it is desired to predict $\mathbf{N}_+(0, T)$ at various calendar times on the basis of the claims already observed. Product units are sold at different times so at any given calendar time the age of the ith unit will be some value t_i, where $0 \leq t_i \leq T$. This allows that some units may not have yet been sold ($t_i = 0$) and that some may have already completed their warranty period ($t_i = T$).

Since $N_i(0, t_i)$ is known for each $i = 1, \ldots, k$, the prediction of (1) is equivalent to prediction of $\sum_{i=1}^{k} \mathbf{N}_i(t_i, T)$. In practice, it is often of interest to update predictions of (1) at various calendar times. These problems will be addressed for a real car warranty setting in Section 4, but also apply in a variety of other contexts.

The prediction of recurrent events has been discussed in specific contexts such as warranty claims (e.g. Robinson & McDonald 1991, Kalbfleisch, Lawless & Robinson 1991, Lawless 1998), insurance claims (e.g. England & Verrall 2002) and software reliability (e.g. Singpurwalla & Wilson 1999) but there has been little general discussion and intervals are generally not well calibrated in terms of coverage probability. Novel features of our presentation here include the provision of probabilistically well-calibrated prediction intervals, the ability to handle age or time trends in the event processes, and the ability to handle large populations of heterogeneous units. In Section 2, we present non-homogeneous Poisson processes with random effects, which will be used as the basis for our methodology. Section 3 develops a flexible method of prediction, and Section 4 illustrates the methodology on car warranty data. Section 5 discusses the robustness of the methodology and some extensions.

2 MIXED POISSON MODELS

Let $\mathbf{N}(s, t)$ be the random variable representing the number of events occurring for a subject in the time interval $(s, t]$; we write $\mathbf{N}(t)$ for $\mathbf{N}(0, t)$. We consider for convenience continuous time processes where two events cannot occur simultaneously; settings where events are recorded in discrete time units such as days or weeks can be handled by grouping. Different types of such processes are discussed in the literature on point processes (e.g. Snyder & Miller 1991, Grandell 1997), but Poisson processes and renewal processes are the two most popular types used to model recurrent events. We can distinguish them through the event intensity function

$$\lambda(t|H(t)) \quad = \quad \lim_{\Delta_t \to 0} \frac{P[\mathbf{N}(t, t + \Delta_t) = 1 | H(t)]}{\Delta_t}, \tag{2}$$

where $H(t)$ denotes the history of the process up to time t. Note that conditional on $H(0)$, (2) fully specifies the process $\{N(t), t > 0\}$. Renewal processes make the assumption that (2) depends only on the time elapsed since the last event; these processes are semi-Markovian. On the other hand, the Poisson processes are Markovian because (2) depends only on t. The intensity, or rate, function is then simply denoted by $\lambda(t)$, and

$$\mathbf{N}(t) \sim \mathcal{PP}(\lambda(t))$$

means that $\mathbf{N}(t)$ is a non-homogeneous Poisson process (NHPP) with rate function $\lambda(t)$.

It is well known that in a Poisson process the total number of events over any interval has a Poisson distribution, and that the number of events $\mathbf{N}(s_1, t_1)$ and $\mathbf{N}(s_2, t_2)$ in two non-overlapping time intervals (s_1, t_1) and (s_2, t_2) are independent. These two properties combined with the fact that event times are often interval-censored make Poisson processes easy to use with prediction problems involving recurrent events. On the other hand, the distribution of the future number of events can be difficult to obtain when renewal processes are used. Experience has shown that Poisson processes apply to a wide range of settings (e.g. Ascher & Feingold 1984, Grandell 1997) and in this paper we focus on them. However, in populations with heterogeneous units it is generally found necessary to extend the models by including unit-specific random effects. Such models are termed random effects, or mixed, Poisson processes (e.g. Lawless 1987, Grandell 1997).

We henceforth consider mixed non-homogenous Poisson processes. We will model the rate function for a single process with parametric forms

$$\lambda(t; \alpha, \beta) = \alpha f(t; \beta),$$

where α is a scalar and β is a vector of low dimension. This parameterization is convenient because $f(t; \beta)$ and α measure different aspects of a NHPP: the function $f(t; \beta)$ describes the shape of the rate function and α represents the overall event frequency. In the finite horizon problems it is convenient to choose α so that $\mathbb{E}[\mathbf{N}(0, T)] = \alpha$, in which case $\int_0^T f(t; \beta)dt = 1$. That is, $f(t; \beta)$ has the form of a probability density function over $(0, T)$.

To consider scenarios where heterogeneity is observed amongst the processes for different units, we incorporate unobservable i.i.d. random effects in our model. The model considered in this paper is

$$\begin{aligned} \mathbf{N}_i(t)|\alpha_i &\sim \mathcal{PP}(\alpha_i f(t; \beta)), \\ \alpha_i &\sim \text{Gamma}(a, b), \end{aligned} \tag{3}$$

where $i = 1, \ldots, k$. The parameterization for the gamma distribution is such that $\mathbb{E}[\alpha_i] = a/b$ and $\mathbb{V}ar[\alpha_i] = a/b^2$. Appropriate forms for $f(t; \beta)$ will be discussed in Section 4 for the warranty claims problem; in general, they depend on the setting being considered.

The model (3) may seem like a stringent assumption, but it allows for cases where the rates may be affected by unobserved covariates in a multiplicative fashion, and has been found to work well in many settings. Simulations in Fredette (2004, Chapter 3) also suggest that predictions based on models (3) are robust to some types of misspecification. For example, they provide adequate predictions if the α_i's are not actually random or if they are random but their actual distribution is not gamma. More generally, (3) can be extended by allowing β to contain random components. Some further comments on extensions and robustness are given in Section 5.

3 PREDICTION

Prediction has been discussed in general terms by various authors, for example, Aitchison & Dunsmore (1975), Geisser (1993), Barndorff-Nielsen & Cox (1996), Beran (1990), and Meeker & Escobar (1999). Like those authors, we consider frequentist methods and, in particular, we seek to construct prediction intervals for a future random variable $\mathbf{Y}$, given observed data $\mathbf{X} = x$. Such intervals are of the form $(L(x), U(x))$, and we attempt to find intervals where $P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X})]$ equals some specified fixed value γ, in which case $(L(x), U(x))$ is called a γ prediction interval (e.g. Barndorff-Nielsen & Cox 1996), and γ is called its coverage probability.

In the context discussed in this paper, we wish to use the information regarding the k processes that is available at a certain given time to make predictive statements about the remaining number of events to be observed. Since processes were not necessarily observed for the same amount of time, we let t_i represent the "age" of the ith process at that time and the remaining number of events is then represented by $\sum_{i=1}^{k} \mathbf{N}_i(t_i, T)$. For each process, the information available to make our prediction consists of the total number of events $N_i(t_i)$ and the set of occurrence times $\tau_i(t_i) = \{\tau_{i1}, \ldots, \tau_{i\,N_i(t_i)}\}$. It is shown in Appendix A.1 that conditional on this information each $\mathbf{N}_i(t_i, T)$ has a negative binomial distribution with parameters $a + N_i(t_i)$ and $(b + F(t_i; \beta))/(b + F(T; \beta))$, where $F(t; \beta) = \int_0^t f(u; \beta) du$. From now on, this distribution will be denoted $NB(a + N_i(t_i), (b + F(t_i; \beta))/(b + F(T; \beta)))$, with a probability function given by

$$\begin{aligned} P[\mathbf{N}_i(t_i, T) = n | N_i(t_i); a, b, \beta] &= \frac{\Gamma(a + N_i(t_i) + n)}{\Gamma(a + N_i(t_i)) n!} \left(\frac{F(T; \beta) - F(t_i; \beta)}{b + F(T; \beta)} \right)^n \times \\ & \left(\frac{b + F(t_i; \beta)}{b + F(T; \beta)} \right)^{a + N_i(t_i)} . \end{aligned} \quad (4)$$

Note that the occurrence times do not appear in this distribution; only the knowledge of $N_i(t_i)$ is required to determine this conditional distribution. However, the occurrence times will enter in the estimation of model parameters. Note also that although claim times are recorded in discrete units (days), we utilize the continuous time likelihood, given below in Section 3.1. This is typically done when, as here, failure times or event times are recorded in units that are short relative to the length of the processes in question.

3.1 Plug-in prediction intervals

Let $\underset{\sim}{N}(t) = (N_1(t_1), \ldots, N_k(t_k))$ and $\underset{\sim}{\tau}(t) = \{\tau_{ij}; i = 1, \ldots, k \text{ and } j = 1, \ldots, N_i(t_i)\}$. A prediction interval for $\sum_{i=1}^k \mathbf{N}_i(t_i, T)$ is an interval $[L(\underset{\sim}{N}(t), \underset{\sim}{\tau}(t)), U(\underset{\sim}{N}(t), \underset{\sim}{\tau}(t))]$ such that

$$P[L(\underset{\sim}{\mathbf{N}}(t), \underset{\sim}{\tau}(t)) \leq \sum_{i=1}^{k} \mathbf{N}_i(t_i, T) \leq U(\underset{\sim}{\mathbf{N}}(t), \underset{\sim}{\tau}(t)); a, b, \beta] = \gamma.$$

Such an interval is called an exact γ prediction interval for $\sum_{i=1}^k \mathbf{N}_i(t_i, T)$. Note that this random variable as well as the random variables $\underset{\sim}{N}(t)$ and $\underset{\sim}{\tau}(t)$ enters in the calculation of the probability above.

In most settings (including the one considered in this paper) we cannot find exact prediction intervals when the parameters are unknown. This is analogous to the non-existence of exact confidence intervals for parameters in most statistical models. The alternative is to find an interval with an approximate coverage probability of γ. This can be accomplished in one way by finding an interval $[L, U]$ such that

$$P[L \leq \sum_{i=1}^{k} \mathbf{N}_i(t_i, T) \leq U); \hat{a}, \hat{b}, \hat{\beta}] = \gamma, \tag{5}$$

where only $\sum_{i=1}^k \mathbf{N}_i(t_i, T)$ is treated as a random variable, and where $\hat{a}, \hat{b}$, and $\hat{\beta}$ are the maximum likelihood estimates (mle's) obtained from the likelihood function based on the observed data, which is (e.g. Lawless 1987)

$$\begin{aligned} L(a, b, \beta | (\underset{\sim}{N}(t), \underset{\sim}{\tau}(t))) \;\; = \;\; & \prod_{i=1}^{k} \left[\left(\prod_{j=1}^{N_i(t_i)} f(\tau_{ij}; \beta) \right) \left(\frac{b^a}{(b + F(t_i; \beta))^{a + N_i(t_i)}} \right) \times \right. \\ & \left. \left(\frac{\Gamma(a + N_i(t_i))}{\Gamma(a)} \right) \right]. \end{aligned}$$

The interval (5) is called a "plug-in" γ prediction interval. Essentially, this method assumes that (4) is the true distribution and that the true parameter values are in fact $\hat{a}, \hat{b}$, and $\hat{\beta}$ and thus ignores completely the uncertainty in $(\hat{a}, \hat{b}, \hat{\beta})$ relative to (a, b, β). When the observed data set is very large , so that $(\hat{a}, \hat{b}, \hat{\beta})$ can be assumed close to (a, b, β) (provided the model is correct), then the coverage probability of this interval will be close to γ. However, this method can be improved upon by "calibrating" the plug-in intervals obtained (Beran 1990, Meeker & Escobar 1999, Lawless & Fredette 2005). We will now show how to obtain these plug-in intervals and then we will explain how to calibrate them so that they have close to a stated nominal coverage probability.

For the prediction problem at hand, we can see from (4) that the distribution of $\sum_{i=1}^{k} \mathbf{N}_i(t_i, T)$ given $\underset{\sim}{N}(t)$ is a convolution of k $NB(a + N_i(t_i), (b + F(t_i; \beta))/(b + F(T; \beta)))$ distributions with a probability function given by

$$\begin{aligned} p(n|\underset{\sim}{N}(t); a, b, \beta) &= P[\sum_{i=1}^{k} \mathbf{N}_i(t_i, T) = n|\underset{\sim}{N}(t); a, b, \beta] \\ &= \sum_{\{z_i : \sum_{i=1}^{k} z_i = n\}} \prod_{i \in \mathcal{S}_t} \frac{\Gamma(a + N_i(t_i) + z_i)}{\Gamma(a + N_i(t_i)) z_i!} \left(\frac{F(T; \beta) - F(t_i; \beta)}{b + F(T; \beta)} \right)^{z_i} \times \\ & \quad \left(\frac{b + F(t_i; \beta)}{b + F(T; \beta)} \right)^{a + N_i(t_i)}, \end{aligned}$$

where $\mathcal{S}_t$ is the set of cars already sold at time t.

Now let $Q(\alpha; a, b, \beta)$ be the α quantile of $\sum_{i=1}^{k} \mathbf{N}_i(t_i, T)$ given $\underset{\sim}{N}(t)$, and we see that according to (5) a two-sided plug-in $1 - \gamma$ prediction interval is given by

$$[Q(\gamma/2; \hat{a}, \hat{b}, \hat{\beta}), Q(1 - \gamma/2; \hat{a}, \hat{b}, \hat{\beta})].$$

These quantiles may be hard to compute; we next discuss ways to do this.

3.2 Approximation of $p(n|\underset{\sim}{N}(t); \hat{a}, \hat{b}, \hat{\beta})$

The quantiles mentioned above are retrieved using the probability function $p(n|\underset{\sim}{N}(t); \hat{a}, \hat{b}, \hat{\beta})$. However this function is obtained by summing over $\binom{n+k-1}{n}$ terms, which may not be feasible when k or n is large. Nevertheless, this problem in principle can be solved by using a recursive formula to find $p(n|\underset{\sim}{N}(t); \hat{a}, \hat{b}, \hat{\beta})$. In particular, a recursive formulae can be obtained by a method in Klugman, Panjer & Willmot (2004, Example 4.60); see Fredette (2004, pp. 52-54).

When n is very large, which will be the case for the data studied in Section 4, it is more convenient to approximate $p(n|\underset{\sim}{N}(t); \hat{a}, \hat{b}, \hat{\beta})$ instead of using a recursive formula. A good approximation can be obtained by generating convolutions of gamma random variables. This is due to the fact that the probability function can be written as

$$\begin{aligned} p(n|\underset{\sim}{N}(t);\hat{a},\hat{b},\hat{\beta}) &= \int_{\underset{\sim}{\alpha}} P[\sum_{i=1}^{k}\mathbf{N}_i(t_i,T)=n|\underset{\sim}{\alpha};\hat{\beta}]\pi(\underset{\sim}{\alpha}|\underset{\sim}{N}(t);\hat{a},\hat{b},\hat{\beta})d\underset{\sim}{\alpha} \\ &= \int_{\underset{\sim}{\alpha}} P\left[\text{Poisson}\left(\sum_{i=1}^{k}(F(T;\hat{\beta})-F(t_i;\hat{\beta}))\alpha_i\right)=n\right]\pi(\underset{\sim}{\alpha}|\underset{\sim}{N}(t);\hat{a},\hat{b},\hat{\beta})d\underset{\sim}{\alpha} \\ &= \int_0^\infty \frac{\exp\{-u\}u^n}{n!}\psi(u;\hat{a},\hat{b},\hat{\beta})du, \end{aligned}$$

where $u=\sum_{i=1}^{k}(F(T;\hat{\beta})-F(t_i;\hat{\beta}))\alpha_i$, and since each α_i has a Gamma$(\hat{a}+N_i(t_i),(\hat{b}+F(t_i;\hat{\beta})))$ distribution, $\psi(u;\hat{a},\hat{b},\hat{\beta})$ is the density of the convolution of k Gamma$(\hat{a}+N_i(t_i),(\hat{b}+F(t_i;\hat{\beta}))/(F(T;\hat{\beta})-F(t_i;\hat{\beta})))$ random variables. The density (4) does not have a simple closed form, and we approximate the integral above by simulation. Once we generate B convolutions of gammas, where each simulated convolution is denoted by u_i^* $(i=1,\ldots,B)$, we can approximate the predictive density function with

$$p(n|\underset{\sim}{N}(t);\hat{a},\hat{b},\hat{\beta}) \simeq \sum_{i=1}^{B}\frac{\exp\{-u_i^*\}(u_i^*)^n}{n!}.$$

3.3 Calibration

Letting $\theta=(a,b,\beta)$, we mentioned in Section 3.1 that a γ plug-in prediction interval is likely to have a coverage probability below the desired level since it ignores the uncertainty in $\hat{\theta}$ relative to θ. Therefore, we should try to find a level greater than γ, say γ', such that a γ' plug-in prediction interval would have an actual coverage probability of γ. This procedure is called calibration. Calibration can be done using asymptotic expansions or simulations. We refer the reader to Komaki (1996) and Barndorff-Nielsen & Cox (1996) to learn more about calibration procedures using asymptotic expansions and to Beran (1990), Meeker & Escobar (1999), and Lawless & Fredette (2005) for calibration procedures using simulation. Calibrating an interval by simulation is usually more tractable for finite horizon prediction problems involving recurrent events.

From a theoretical point of view, it is clear that calibrated plug-in prediction intervals are more adequate than simple plug-in prediction intervals (see Theorem 1 in Lawless & Fredette (2005) for an example). However, practical prediction problems are mostly handled using non-calibrated intervals. Foregoing the fact that some scientists could be unaware of this procedure, the main reason why intervals are not calibrated is probably because this procedure often requires a significant amount of computational time.

In the next section, we will apply the methods mentioned above to a particular database. We will calibrate the prediction intervals using an algorithm proposed in Lawless & Fredette (2005), and we will see that the calibrated prediction intervals provide a much more realistic measure of uncertainty than the uncalibrated intervals.

4 PREDICTION OF CAR WARRANTY CLAIMS

With products under warranty, manufacturers usually collect detailed claims data. When this database is maintained properly, it can be used to predict the eventual total number of warranty claims based on the data already observed. In this subsection, we apply our prediction model to a car warranty database. The number of units (cars) in this dataset being very large, we will consider efficient computational methods for providing prediction intervals.

The setting and database that we consider are an update of that presented in Kalbfleisch et al. (1991); this dataset contains warranty information on one subsystem for cars of one production year and model type. The database includes information over a span of 571 days since the first car was produced and the following times were recorded for each car: production time, sale time and the claim time(s). Each car had a one year or 12,000 mile warranty, whichever came first.

We let $N_i(t) = N_i(-\infty, t)$ be the total number of warranty claims for the ith car up to t days after it was sold. Note that $N_i(0) = N_i(-\infty, 0)$ is not necessarily equal to 0 as some claims could occur between the production day and the day of sale. Therefore, $N_i(t)$ now represents the total number of warranty claims while $N_i(0, t)$ represents the number of claims observed after the sale of the ith car.

The quantity we wish to predict is $\sum_{i=1}^{k} \mathbf{N}_i(365)$, the eventual total number of warranty claims for this fleet of cars. The prediction of claim numbers is of interest because they can be compared across production years or other periods, and because they drive costs. The prediction of total warranty cost is also important. A possible approach, that we wish to explore in separate article, is to model the cost process. One way to do this is by assigning a cost distribution to a claim, combining this with the claim occurrence modeling of the present paper. The methods herein could then be extended to provide calibrated prediction intervals for cost.

To be able to assess the performance of our predictions, we will here only consider cars for which $\mathbf{N}_i(365)$ is known, *i.e.* cars sold at least 365 days before the end of the data collection period for this particular database. The resulting dataset contains 15,775 cars, manufactured over a period of 206 days. Each of them had between 0 and 10 claims for a total of 2,620 claims. Table 1 shows the distribution of total claims amongst all the cars. We see that most of the cars never had a warranty claim, and only a few cars had more than 2 claims.

Put Table 1 around here

Early in a production year and before cars have been in service for a substantial period of time, manufacturers estimate the eventual number of warranty claims per unit from past years' data. Once data begin to accrue for the model year in question, however, the methods in Section 3 can be used to predict the average claims per vehicle, or the total claims for a specified number of vehicles sold. We will illustrate this methodology here by beginning prediction based on data accumulated in the first 150 days of production, and then updating the predictions every 50 days thereafter. We will be able to assess how well the prediction methodology has performed, since $T = 365$ and $\sum_{i=1}^{15{,}775} N_i(365) = 2,620$ is eventually known for the vehicles we consider.

Put Figure 1 around here

Before implementing the prediction methodology, we provide some discussion of the full data set, which will give some insight into predictive performance, and ideas for extensions to the models used. Figure 1 shows a plot of the occurrence times (in days before or after sale) for each claim, for each of the 15,775 vehicles produced; there are thus 2,620 points in the plot. The claims for each car are represented along an invisible horizontal line. The first car produced appears at the bottom while the last one, produced 203 days later, appears at the top. Some interesting features are revealed in this figure. First we see that the rate of claims decreases as age increases; one explanation for this is that a significant number of cars are no longer under warranty at higher ages because of the mileage limit. In addition, we can see that the cars produced towards the end appear to have more claims than the first ones produced and that some cars manufactured during a certain early period had fewer claims than the others.

Figure 2 shows another view of the data in Figure 1; Nelson-Aalen estimates of the cumulative mean function for the total number of claims per vehicle are given for cars stratified by manufacturing period. This figure originally appeared in Lawless & Nadeau (1995). The sample of cars we are considering in this paper appeared in the first three manufacturing periods. It is more clear in this figure that cars produced later (period 3) indeed appear to have more claims. Pointwise 95% confidence bands for that period are displayed and we can see other mean functions for the data at hand (periods 1 and 2) are outside these bands.

We must keep in mind that the characteristics mentioned above are not known at first to the analysts facing this prediction problem. Therefore, the features mentioned will not be taken into account in our prediction model. However, it is possible that the analysts could be able to detect such fundamental reliability shifts and adapt the model accordingly during the prediction process. An example of this will be discussed in Section 4.7.

Put Figure 2 around here

4.1 Prediction model proposed

We will now propose a model to predict the total number of warranty claims or, equivalently, the average number per vehicle. This model is a simple extension of model (3) which takes into account that $\mathbf{N}_i(0)$ is not necessarily 0. The choice of a suitable parametric form for $f(t;\beta)$ in (3) is crucial, since our predictions necessarily involve extrapolation into the future. The time scale for prediction is car age t, in days since sale. To deal with the calendar times at which vehicles are sold, we let s denote calendar time, with $s = 0$ the day of manufacture of the first vehicle in the database and $s = 571$ the day the last warranty expired among the set of 15,775 vehicles considered here. We also introduce the following notation:

$$\begin{aligned} \tau_{ij} &= \text{The age of car } i \text{ when the } j\text{th claim after its sale occurred.} \\ s_i &= \text{Sale time of the } i\text{th car.} \\ S_s &= \{i : s_i \leq s\} \;\; \text{(set of cars already sold at calendar time } s\text{).} \\ \mathbf{W}_+(s) &= \sum_{i \in S_s^c} \mathbf{N}_i(365) + \sum_{i \in S_s \setminus S_{s-365}} \mathbf{N}_i(s - s_i, 365) \\ &= \text{The number of future warranty claims at calendar time } s. \end{aligned}$$

Note that $\mathbf{W}_+(0)$ corresponds to $\sum_{i=1}^{15{,}775} \mathbf{N}_i(365)$, the total number of warranty claims for the cars in the database. Our objective is to be able to, at any calendar time s, predict $\mathbf{W}_+(s)$.

Adding this to the number of claims already observed equivalently gives a prediction for $\mathbf{W}_+(0)$. The prediction model proposed is

$$\begin{aligned} \mathbf{N}_i(0)|\alpha_i &\sim \text{Poisson}(c\alpha_i), \\ \mathbf{N}_i(0,t)|\alpha_i &\sim \mathcal{PP}(\alpha_i f(t;\beta)), \\ \alpha_i &\sim \text{Gamma}(a,b), \end{aligned} \tag{6}$$

where $i = 1, \ldots, 15,775$ and $0 < t \leq 365$. In this model, the unknown parameters are a, b, c, and the vector β. We point out that a model using different random effects for $\mathbf{N}_i(0)$ and $\mathbf{N}_i(0,t)$ was also fitted but we are presenting the simpler one since the predictions obtained with the other model were not much different.

The following proposition shows that the distribution of the quantity we want to predict will still be a convolution of negative binomials; the proof of this proposition is given in Appendix A.2:

Proposition 1. *Using model (6), the probability function for $\mathbf{W}_+(s)$, the number of future warranty claims for the $k = 15,775$ cars, given the information available at calendar time s, is a convolution of $k - |S_{s-365}|$ negative binomials. Of these, $k - |S_s|$ have a $NB(a, b/(b+c+F(365;\beta)))$ distribution, while the remaining $|S_s| - |S_{s-365}|$ have a $NB(a + N_i(s - s_i), (b+c+F(s-s_i;\beta))/(b+c+F(365;\beta)))$ distribution where $i \in S_s \backslash S_{s-365}$.*

Like in Section 3, the distribution of the quantity being predicted is a convolution of negative binomials. In fact, $\mathbf{W}_+(s)$ can be written in the form $\sum_{i=1}^{k} \mathbf{N}_i(t_i, T)$, where $T = 365$ and

$$t_i = \begin{cases} \min(s - s_i, 365) \text{ if } s \geq s_i, \\ -\infty \text{ otherwise.} \end{cases}$$

Thus, like in Section 3, plug-in prediction intervals are found by finding the quantiles of this distribution with the unknown parameters (a, b, c, β) replaced with their mle's, based on the data observed up to calendar time s.

We will now consider the estimation of the unknown parameters a, b, c, and β, and the specification of the form of the function $f(t;\beta)$. Since it is important to use a function that is flexible, we consider functions such as

$$f(t;\beta) = \exp\{\beta_1 t + \beta_2 t^2 + \ldots + \beta_q t^q\}. \tag{7}$$

Even for moderate values of q, this function can have different changes of slope, with up to $q-1$ critical points. In the cases like here where t is large, we have found that it is better to use $f(\log(1+t);\beta)$ instead of $f(t;\beta)$. Since polynomials may not extrapolate well too far into the future, we also adopt a strategy whereby a small value of q is used early in the data collection period, with a larger value used later. Note, however, that since $t \leq 365$ we are not forced to extrapolate too far forward with respect to the age of a vehicle. Note also that with these (and most other) functions, $F(t;\beta)$ has to be evaluated numerically when $q > 1$. However, simple numerical integration techniques like the trapezoidal method give very adequate approximations.

When comparing different types of function where β has the same dimension, we have no reason to believe *a priori* that our log-polynomial model would be the most adequate, so we may want to explore polynomials in different functions of t, or other functions altogether.

Now let $H(s)$ be the information available for all the processes at calendar time s; $H(s)$ contains information about the total number of claims for each car and their associated claim times. Given this information, the likelihood function $L(a,b,c,\beta|H(s))$ based on model (6) is derived and stated in Appendix B.

This likelihood function will also be used to find a suitable value for q in (7), the dimension of the vector of unknown parameters β. We will initially fit the model with $q = 1$ and increase q if there is a substantial increase in the likelihood function (13). However, because the model is to be used for prediction, we also pay close attention to the fit of the model at higher ages, and to the shape of $f(t;\hat{\beta})$ there. We now expand on the model fitting process.

4.2 Fitting the model

With $q+3$ parameters and up to 15,775 processes, the computer time required by non-linear maximization routines to obtain the mle's is non-negligible. However, there are ways to substantially reduce this computational time: the use of a conditional likelihood, a reparameterization for the function $f(t;\beta)$ and methods to obtain appropriate starting values. We now present these ideas.

The likelihood function (13) given in Appendix B can be rewritten as

$$\begin{aligned} L\left(a,b,c,\beta|H(s)\right) &= \left(\prod_{i\in S_s} P[\mathbf{N}_i(0)=N_i(0), \mathbf{N}_i(0,s-s_i)=N_i(0,s-s_i);a,b,c,\beta]\right)\times \\ &\quad \left(\prod_{i\in S_s}\prod_{j=1}^{N_i(0,s-s_i)}\frac{f(\tau_{ij};\beta)}{F(s-s_i;\beta)}\right), \qquad (8) \end{aligned}$$

where $P[\mathbf{N}_i(0) = N_i(0), \mathbf{N}_i(0, s-s_i) = N_i(0, s-s_i); a, b, c, \beta]$ is the product of the probability function of a $NB(a, \frac{b}{b+c}))$ and a $NB(a+N_i(0), \frac{b+c}{b+c+F(s-s_i;\beta)})$. Note that to ease the notation, we are using expressions like $N_i(0, s-s_i)$ and $F(s-s_i;\beta)$ instead of $N_i(0, \min(s-s_i, 365))$ and $F(\min(s-s_i, 365); \beta)$: when a car has been sold for more than 365 days at calendar time s, it no longer provides additional information to the likelihood function, since the warranty coverage ceases at 365 days.

Now let $\underset{\sim}{N}(0) = \{N_i(0) : i \in S_s\}$, $\underset{\sim}{N}(0, s) = \{N_i(0, s-s_i) : i \in S_s\}$, and $\tau(s) = \{\tau_{ij} : i \in S_s \text{ and } j = 1, \ldots, N_i(0, s-s_i)\}$, so that the likelihood function is then the product of

$$L(a, b, c, \beta | \underset{\sim}{N}(0), \underset{\sim}{N}(0, s)) = P[(\underset{\sim}{\mathbf{N}}(0), \underset{\sim}{\mathbf{N}}(0, s)) = (\underset{\sim}{N}(0), \underset{\sim}{N}(0, s); a, b, c, \beta] \tag{9}$$

and

$$L_c(\beta|\tau(s)) = \prod_{i \in S_s} \prod_{j=1}^{N_i(0, s-s_i)} \frac{f(\tau_{ij}; \beta)}{F(s-s_i; \beta)}. \tag{10}$$

We can see that $L_c(\beta|\tau(s))$ is the likelihood function for β when we condition on the total number of claims per car. Our empirical studies suggest that (9) has relatively little information on β. This means that the parameter maximizing (10) will be close to the $\hat{\beta}$ obtained by maximizing (8). Thus, a good strategy is to first find the estimate maximizing $L_c(\beta|\tau(s))$ and use this value as a starting value to maximize the original likelihood. These two maximizers being similar, we are close to maximizing q parameters and then 3 instead of $q+3$ parameters all at once, a strategy that often leads to a substantial decrease in computational time.

We can also reparameterize $f(t;\beta)$ to shorten the time required to obtain the mle's. The way this function is defined in (7) leads to high correlations between the components of $\hat{\beta}$. To correct this problem, we consider the following reparameterization:

$$f(t; \beta) = \exp\{\beta_1 L_1(t) + \ldots + \beta_q L_q(t)\}, \tag{11}$$

where

$$L_n(t) = \exp\{t\} \frac{\partial^n}{\partial t^n}(t^n \exp\{-t\}).$$

These polynomials, called Laguerre polynomials, are of interest here because they are orthogonal with respect to a certain inner product. In addition, they are easily obtained via the recursive formula

$$L_{n+2}(t) = [2(n+1) - t + 1] L_{n+1}(t) - (n+1)^2 L_n(t).$$

For example, when the model is fitted using the complete dataset and $q = 4$, Table 2 shows the substantial reduction of the correlation of the $\hat{\beta}_i$'s when (11) is used instead of (7). The correlations in Table 2 are obtained from the asymptotic covariance matrix for the parameters, obtained by inverting the observed information matrix.

Put Table 2 around here

When we first fit model (6) using $q = 1$, $\beta = 0$ is usually an appropriate starting value to find $\hat{\beta}$ using a non-linear maximization routine. When investigation reveals that it may be better to increase q, good starting values to find $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_q)$ are given by $\beta = (\hat{\beta}^{(q-1)}, 0)$ where $\hat{\beta}^{(q-1)}$ is the mle of β obtained when we fitted the model with $q - 1$ terms. This approach works well for both likelihood functions (10) and (13). Appropriate starting values for the other parameters a, b, c can also be found using a technique similar to the moment matching approach. This technique can be derived from Gaver & O'Muircheartaigh (1987) and the actual formulae can be found in Fredette (2004, Section 5.3).

4.3 Assessment of fit

We fitted the model and obtained prediction intervals for the total number of warranty claims using the information available after 150 days and every subsequent 50 days. Table 3 shows the number of warranty claims observed by calendar times $s = 150, 200, \ldots, 550$, as well as by day 571. Based on the information available at each given time, the value selected for $q = \dim(\hat{\beta})$ was 1 after 150 days, 2 after 200 days, and 4 thereafter.

Put Table 3 around here

Before presenting the prediction intervals, we present some methods to assess how well model (6), with $q = 4$ for the function $f(log(1+t); \beta)$ given in (11), fits the complete dataset. We cannot of course assess this at the earlier prediction times, but this provides insight into the performance of the predictors discussed below. First we compare the distribution of the total claims amongst all the cars and their corresponding fitted values. We do not observe any significant difference between the observed and fitted values and the p-value of Pearson's goodness-of-fit test is 0.15.

In addition to the total number of claims for each car, we are also interested to assess the validity of the function $f(t; \beta)$ to model the occurrence times. We can show from Snyder &

Miller (1991, Section 2.3) that under the stated mixed Poisson model the set of all

$$u_{ij}(\beta) = \frac{F(\tau_{ij};\beta)}{F(365;\beta)},$$

where $i = 1, \ldots, 15,775$ and $j = 1, \ldots, N_i(0, 365)$, forms a sample of independent observations uniformly distributed on $[0, 1]$. Therefore, when $\hat{\beta}$ is a precise estimate of β, the sample of all $u_{ij}(\hat{\beta})$'s should behave like a sample of independent uniforms if model (11) is right. Figure 3 shows the empirical quantiles of the $u_{ij}(\hat{\beta})$'s versus the theoretical quantiles of a $U(0, 1)$ distribution using the complete dataset with $q = 4$. This figure provides no evidence that $f(t;\beta)$ is not adequate.

Put Figure 3 around here

Another way to assess the adequacy of $f(t;\beta)$ is by comparing the histogram of the claim times with the estimated density $f(t;\hat{\beta})/F(365;\hat{\beta})$. This is done in Figure 4 where the estimated density is plotted for each q=1, 2, 3, and 4. It is clear from this figure that the estimated density with $q = 4$ (the value chosen at calendar times $s = 250, 300, \ldots$) is preferable for the full set of claims. The model with $q = 2$ fits only moderately well, but it should be noted that because of the narrow bin widths in Figure 4, there is substantial sampling variation in the bin frequencies. The histogram suggests that warranty claim rate is a little higher early after the sale of a vehicle but decreases in the next couple of weeks. After that, it suggests an increase in the rate of claims for approximately 4-5 months and a decrease thereafter, probably because of the mileage drop-out. Note that this feature was also observed on Figure 1 and Figure 2.

Put Figure 4 around here

4.4 Non-calibrated prediction intervals

We now look at the ability of the proposed models to predict $\mathbf{W}_+(s)$. The upper panel of Figure 5 shows 95% non-calibrated plug-in prediction intervals, presented in Section 4.1, obtained using the information available after $150, 200, \ldots, 550$ days. The digit next to each interval is the value of q used, and the increasing solid curve represents the total number of claims known at that point. Early prediction intervals are seen to be substantially lower than the eventual value of $W_+(0) = \sum_{i=1}^{15,775} N_i(365) = 2,620$, which is represented by a line in the plot. However, the models start to give better prediction intervals after about 300 days. Factors other than model imperfection (which is related to extrapolation, and cannot be detected from the data used to fit the models for earlier calendar times) can also affect the performance of the early prediction intervals. For example, some of these intervals were obtained when few claims were observed and therefore the mle's were not very precise. In addition, we saw in Figure 1 that a group of cars manufactured early in the production year had very few claims relative to the others. Since these cars were mostly sold early, this contributes to the early under-prediction. The lower panel of Figure 5 shows predictions obtained by using the mle's with $q = 4$ from the full data set; it suggests that the early under-prediction is more due to poor estimation of the unknown parameters than model imperfection.

Put Figure 5 around here

Even though they usually do not include the real value, the prediction intervals obtained after about 300 days are reasonably adequate, and close to the eventual value of $\mathbf{W}_+(0)$. It should be stressed that the prediction horizon here is 571 days, and we are thus able to obtain reasonably adequate predictions approximately 571-300=271 days before the last warranty claim was observed. However, we should keep in mind that we are here studying only the cars sold at least 365 days before the end of the data collection period (s=571). In the complete database, the warranty of the last car sold expired more than 900 days after the first car was produced. Therefore, the real prediction horizon is more than 571 days. In addition, we are ignoring the sampling variability in the estimation of up to 7 unknown parameters $(a, b, c, \beta_1, \beta_2, \beta_3, \beta_4)$ in getting these plug-in prediction intervals. We will see in the next section that properly calibrated prediction intervals are much more satisfactory.

4.5 Calibrated prediction intervals

Letting $\theta = (a, b, c, \beta)$, we will calibrate our prediction intervals at a given calendar time s using an algorithm derived from Lawless & Fredette (2005):

1 Simulate B complete datasets using $\hat{\theta}(s)$ the mle's obtained from the claims data available at calendar time s. By complete datasets, we mean the full claim histories $\{N_i(t), 0 \leq t \leq 365\}$ for each vehicle $i = 1, \ldots, 15,775$; see Ross (2003, Section 11.5.1) for details on how to simulate a NHPP.

2 For each simulated dataset, use the information available at time s to obtain a set of mle's denoted by $\hat{\theta}_i^*(s)$, $i = 1, \ldots, B$.

3 Let $W_i^*(s)$ be the remaining number of warranty claims at time s for the ith simulated dataset and $\underset{\sim}{N_i}^*(s)$ be a vector giving the total number of claims per car based on the ith simulated dataset. Calculate $u_i^* = P[\mathbf{W}_+(s) \leq W_i^*(s) | \underset{\sim}{N_i}^*(s); \hat{\theta}_i^*(s)]$.

4 Let u_L and u_U be the $\alpha/2$ and the $1 - \alpha/2$ empirical quantiles based on the sample $(u_1^*, \ldots, u_B^*)$. It is shown in Lawless & Fredette (2005) that a calibrated two-sided $1 - \alpha$ prediction interval for $\mathbf{W}_+(s)$ is obtained by finding the u_L and u_U quantile of the distribution of $\mathbf{W}_+(s)$ given in Proposition 1, with the unknown parameters θ replaced with $\hat{\theta}(s)$.

Put Figure 6 around here

Using this method, we simulated $B = 2,000$ datasets at times $t = 150, 200, \ldots, 550$ to provide calibrated 95% prediction intervals. These new intervals are presented in Figure 6. We can see that these intervals predict well by 300 days. Moreover, early predictions have wide limits, indicating a high degree of uncertainty, and the upper limits are within 15-30 percent of the eventual number of claims. Because of the small number of claims observed early on (see Table 3) and patterns in the data mentioned earlier, it is difficult for any early prediction to be accurate. We will discuss this issue further in Section 5.

The importance of calibrating the prediction intervals is also illustrated in Table 4, which shows the approximate coverage probability of nominal two-sided 95% plug-in prediction intervals. These are obtained from the simulation mentioned above by finding the empirical coverage for the plug-in intervals; see Lawless & Fredette (2005) for more detail. We can see that unless s is large, the non-calibrated prediction intervals do not have close to the nominal coverage probability.

Put Table 4 around here

4.6 Approximating the calibration process

It is clear from Table 4 and Figure 6 that plug-in prediction intervals should be calibrated. However, the amount of computational time required to calibrate these intervals is non negligible. In our calibration algorithm, the step where we obtain B sets of mle's is especially long. We can significantly reduce the amount of computational time if we replace the second step of the calibration algorithm by:

2 Simulate B sets of mle's $\hat{\theta}^*_1(s), \ldots, \hat{\theta}^*_B(s)$ using their asymptotic normal distribution $\mathcal{N}(\hat{\theta}(s), I^{-1}(\hat{\theta}(s)))$ where $I(\theta)$ is Fisher's observed information matrix.

With this modification, we are essentially using the asymptotic normal distribution of the mle's instead of approximating their real distribution via simulations.

The calibrated intervals using the "real" and the "normal" mle's are shown in Figure 7. We can see that these prediction intervals are very similar. Since the approximate intervals are obtained rapidly, they could be used on their own, or to determine if intervals should be calibrated: if the approximate interval obtained is substantially different than the non-calibrated interval, one could perform simulations to obtain a calibrated prediction interval.

Put Figure 7 around here

4.7 Monitoring of the data

Although our proposed model (6) provides adequate predictions once a sufficient number of claims have been observed, such models should of course be compared with the observed data. In addition, the data should be monitored to detect any shifts over time. Amongst others, an item that is important to monitor is the stationarity of the claims process for cars produced in different time periods; a potential problem we discussed at the beginning of this section. We now suggest how monitoring can be performed and used in connection with prediction.

Put Figure 8 around here

An simple way to assess stationarity of the claims process is to look at the claims per vehicle curves for different production periods, as illustrated in Figure 2. Figure 8 shows final values for the cars produced during the first 12 weeks; we can see that cars produced in week 3 had substantially fewer claims than those for the other weeks. In fact, they correspond to the group of reliable cars we observed in Figure 1 and that were potential contributors to our early under-prediction problem. Of course, the values in Figure 8 only become apparent over time, as we see more and more of the claims curves analogous to those in Figure 2. In practice, monitoring should be done on a frequent basis so as to pick up features like this as early as possible. Combined use of figures like Figures 1 and 2 are useful for this purpose. For example, Figure 9 resembles Figure 1 except that the x-coordinate is now the calendar time instead of the age of each car. After 250 days we would have observed all the claims on the left of the vertical line in the figure. It is likely that analysts would by then be able to identify the unusual behavior of cars produced during the third week (all the processes between the two horizontal lines) at or before that time and take this into account in subsequent predictions.

Put Figure 9 around here

In order to study their actual impact on the adequacy of our predictions, we removed these 1,100 cars from week 3 and their 60 warranty claims and redid our predictions. Figure 10 gives the 95% calibrated prediction intervals for the total number of claims when these cars are removed. We can see that the intervals obtained after about 250 days are now quite good. Obviously, however, the interest is to predict the claims for all the cars and thus the claims process for cars produced during the third week also have to be taken into account. We discuss how this can be achieved in the next section.

Put Figure 10 around here

5 CONCLUDING REMARKS

We have presented a flexible approach to the prediction of recurrent events, which is easy to implement. Successful prediction of course depends on choosing a satisfactory model for $f(t;\beta)$ in (7), representing the shape of event rate functions for individual units or processes. Model specification should include an assessment of fit, as illustrated in Section 4.3. Among two or more models that fit the data equally well, we recommend choosing the model that best captures the time trend in the event rate at the largest observed values of t, and which provides defensible extrapolations. In practice, there is often historical data on event occurrence that suggests families of models for $f(t;\beta)$. In the warranty claims field, these are sometimes referred to as claim development curves, and are often relatively stable across different production years.

The methodology here uses unit-level random effects. This captures unit-to-unit heterogeneity and provides realistic prediction intervals. Studies with such models suggest that the methodology is robust to the assumption of a gamma distribution for the random effects, but if desired other distributions can be considered (e.g. see Grandell 1997).

In many settings, as for the warranty data process described here, there may be variations in event rate functions according to the time a process begins. An extension of the models used here would be to include terms in the rate function to represent the time a vehicle is manufactured. However, this increases the complexity of the models and the variance of predictions based on them. If there is substantial concern about this issue, a compromise is to stratify the vehicles into a small set of production periods, and then to model separately by (6) and (7) the claims process for each group. The computational and simulation methods in Section 3 and 4 are readily extended to deal with this.

Prediction intervals made when only a little data is available are necessarily very uncertain. In this case, data from previous experience or expert knowledge is valuable, and can be used to modulate predictions based on the current data alone. A Bayesian approach to prediction (e.g. Kuo & Yang 1996, Singpurwalla & Wilson 1999) is useful for this purpose, and is currently being investigated. We also remark that the simulation approaches used here for our frequentist prediction intervals can be related to Bayesian posterior and predictive probability calculations, and computational approaches such as MCMC methods could be considered.

The problem addressed in this paper is one of aggregate prediction: we wish to predict the total number of events across a population of units. We have adopted here what one can term micro modeling, by using models that describe the event processes for individual units. This is appealing in settings where the processes for individual units start at different times. However, it is also possible to develop macro models that consider aggregate event occurrence; in the warranty claims setting, see Chen et al. (1996) for such an approach. We also note that the micro models we use can provide predictors for individual units of empirical Bayes type. These are useful in problems such as maintenance planning for multiple pieces of equipment.

Finally, there are many settings (including the warranty data setting) where there are costs or other values associated with events, and we may wish to predict future costs. The methods considered here can be extended to deal with this through the use of compound mixed Poisson processes (Snyder & Miller 1991, Grandell 1997, Klugman et al. 2004), though detailed development remains to be done.

ACKNOWLEDGMENT

This research has been partially supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors also want to thank the referees for their helpful suggestions.

APPENDIX A

The following Lemma will be used in both Appendix A.1 and A.2:

Lemma 1. *Let* $\mathbf{X}|\lambda$ *and* $\mathbf{Y}|\lambda$ *have a Poisson distribution with mean* $K_1\lambda$ *and* $K_2\lambda$*. If* λ *has a Gamma(*a, b*) distribution, then* $\mathbf{Y}|\mathbf{X} = x$ *has a* $NB(a + x, (b + K_1)/(b + K_1 + K_2))$ *distribution.*

Proof. Let $\pi(\lambda)$ be the density function of λ. The conditional density of $\lambda|\mathbf{X} = x$ is then proportional to

$$\begin{aligned} \pi(\lambda|x) &\propto (\exp\{-\lambda K_1\}\lambda^x) \left(\exp\{-\lambda b\}\lambda^{a-1}\right) \\ &= \exp\{-\lambda(b + K_1\}\lambda^{a+x-1}. \end{aligned}$$

Thus $\lambda|\mathbf{X} = x$ must have a Gamma$(a + x, b + K_1)$ distribution.

Using this conditional density, we now have

$$\begin{aligned} P[\mathbf{Y} = y|\mathbf{X} = x] &= \int_\lambda P[\mathbf{Y} = y|\lambda]\pi(\lambda|x; a, b)d\lambda \\ &= \int_\lambda \left[\frac{\exp\{-\lambda K_2\}(\lambda K_2)^y}{y!}\right] \left[\frac{(b + K_1)^{a+x}\exp\{-\lambda(b + K_1)\}\lambda^{a+x-1}}{\Gamma(a + x)}\right] d\lambda \\ &= \frac{(K_2)^y(b + K_1)^{a+x}}{y!\Gamma(a + x)} \int_\lambda \exp\{-\lambda(b + K_1 + K_2)\}\lambda^{a+x+y-1}d\lambda \end{aligned}$$

$$= \quad \frac{\Gamma(a+x+y)}{\Gamma(a+x)n!}\left(\frac{K_2}{b+K_1+K_2}\right)^y\left(\frac{b+K_1}{b+K_1+K_2}\right)^{a+x},$$

which is the probability function of a $NB(a+x,(b+K_1)/(b+K_1+K_2))$. □

A.1: Conditional distribution of $\mathbf{N}_i(t_i,T)|N_i(t_i)$

By letting $\mathbf{X} = \mathbf{N}_i(t_i)$, $\mathbf{Y} = \mathbf{N}_i(t_i,T)$, $K_1 = F(t_i;\beta)$, and $K_2 = F(T;\beta) - F(t_i;\beta)$, we can apply Lemma 1 to show that the distribution of $\mathbf{N}_i(t_i,T)|\mathbf{N}_i(t_i) = N_i(t_i)$ is $NB(a + N_i(t_i),(b+F(t_i;\beta))/(b+F(T;\beta)))$.

A.2: Proof of Proposition 1

Proof. At calendar time s, only the cars that have not been sold yet and those sold over the last 365 days are at risk of having any additional warranty claims. Therefore, the variable of interest, $\mathbf{W}_+(s)$, is a sum of $k - |S_{t-365}|$ random variables:

$$\mathbf{W}_+(s) \quad = \quad \sum_{i\in S_s^c}\mathbf{N}_i(365) + \sum_{i\in S_s\backslash S_{s-365}}\mathbf{N}_i(s-s_i,365), \tag{12}$$

where S_s^c is the set of cars still unsold at calendar time s and $S_s\backslash S_{s-365}$ is the set of all cars sold during the year before calendar time s.

First, we use Lemma 1 to find the distribution of each random variable in the first summation on the right hand side of (12). By letting $\mathbf{X} = 0$, $\mathbf{Y} = \mathbf{N}_i(365)$, $K_1 = 0$, and $K_2 = c + F(365;\beta)$, we show that each $\mathbf{N}_i(365)$ has a $NB(a, b/(b+c+F(365;\beta)))$ distribution.

At calendar time s, the random variables in the second summation on the right hand side of (12) have been observed for less than 365 days and thus some information is available about them. We can show that conditional on this information, each of these processes has a $NB(a+N_i(s-s_i),(b+c+F(s-s_i;\beta))/(b+c+F(365;\beta)))$ distribution. This is achieved by using Lemma 1 again with $\mathbf{X} = \mathbf{N}_i(s-s_i)$, $\mathbf{Y} = \mathbf{N}_i(s_i,365)$, $K_1 = c+F(s-s_i;\beta)$, and $K_2 = F(365;\beta) - F(s-s_i;\beta)$. □

APPENDIX B

The likelihood function based on model (6) is given by (e.g. Lawless 1987)

$$\begin{aligned}
L(a,b,c,\beta|H(s)) &= \int_{\underset{\sim}{\alpha}} L(\underset{\sim}{\alpha},c,\beta|H(s))\pi(\underset{\sim}{\alpha};a,b)d\underset{\sim}{\alpha} \\
&= \prod_{i\in S_s}\int_{\alpha_i}\left[\frac{\exp\{-\alpha_i c\}(\alpha_i c)^{N_i(0)}}{N_i(0)!}\right]\times \\
&\quad \left(\exp\{-\alpha_i F(s-s_i;\beta)\}\prod_{j=1}^{N_i(0,s-s_i)}\alpha_i f(\tau_{ij};\beta)\right)\left[\frac{b^a\exp\{-\alpha_i b\}\alpha_i^{a-1}}{\Gamma(a)}\right]d\alpha_i \\
&= \prod_{i\in S_s}\frac{c^{N_i(0)}b^a(\prod_{j=1}^{N_i(0,s-s_i)}f(\tau_{ij};\beta))}{N_i(0)!\Gamma(a)}\int_{\alpha_i}\exp\{-\alpha_i(b+c+F(s-s_i;\beta))\}\times \\
&\quad \alpha_i^{a+N_i(0)-1}d\alpha_i \\
&= \prod_{i\in S_s}c^{N_i(0)}\left(\prod_{j=1}^{N_i(0,s-s_i)}f(\tau_{ij};\beta)\right)\left(\frac{\Gamma(a+N_i(s-s_i))}{\Gamma(a)N_i(0)!}\right)\times \\
&\quad \left[\frac{b^a}{(b+c+F(s-s_i;\beta))^{a+N_i(s-s_i)}}\right]. \qquad (13)
\end{aligned}$$

References

Aitchison, J. & Dunsmore, I. (1975), *Statistical Prediction Analysis*, Cambridge University Press.

Ascher, H. & Feingold, H. (1984), *Repairable Systems Reliability*, Marcel Dekker, New York.

Barndorff-Nielsen, O. & Cox, D. (1996), 'Prediction and asymptotics', *Bernoulli* **2**, 319–340.

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

Chen, J., Lynn, N. & Singpurwalla, N. (1996), Forecasting warranty claims, *in* W. Blischke & D. Murthy, eds, 'Product Warranty Handbook', Marcel Dekker, New York, pp. 803–816.

Dalal, S. & McIntosh, A. (1994), 'When to stop testing for large software systems with changing code', *IEEE Transactions on Software Engineering* **20**, 318–323.

England, P. & Verrall, R. (2002), 'Stochastic claims reserving in general insurance', *British Actuarial Journal* **8**, 443–518.

Fredette, M. (2004), Prediction of recurrent events, PhD thesis, University of Waterloo. http://etd.uwaterloo.ca/etd/mfredett2004.pdf.

Gaver, D. P. & O'Muircheartaigh, I. G. (1987), 'Robust empirical bayes analyses of event rates', *Technometrics* **29**, 1–15.

Geisser, S. (1993), *Predictive Inference: An Introduction*, Chapman & Hall, London.

Grandell, J. (1997), *Mixed Poisson Processes*, Chapman & Hall, London.

Kalbfleisch, J. D., Lawless, J. F. & Robinson, J. (1991), 'Methods for the analysis and prediction of warranty claims', *Technometrics* **33**, 273–285.

Klugman, S., Panjer, H. & Willmot, G. (2004), *Loss Models: From Data to Decisions*, 2nd edn, Wiley, Hoboken.

Komaki, F. (1996), 'On asymptotic properties of predictive distributions', *Biometrika* **83**, 299–313.

Kuo, L. & Yang, T. Y. (1996), 'Bayesian computation for nonhomogeneous Poisson processes in software reliability', *Journal of the American Statistical Association* **91**, 763–773.

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

Lawless, J. (1998), 'Statistical analysis of product warranty data', *International Statistical Review* **66**, 41–60.

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

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

Meeker, W. Q. & Escobar, L. A. (1999), 'Statistical prediction based on censored life data', *Technometrics* **41**, 113–124.

Robinson, J. & McDonald, G. (1991), Issues related to field reliability and warranty data, *in* G. E. Liepins & V. Uppuluri, eds, 'Data Quality Control: Theory and Pragmatics', Marcel Dekker, New York, pp. 69–90.

Ross, S. M. (2003), *Introduction to Probability Models*, 8th edn, Academic Press, San Diego.

Singpurwalla, N. & Wilson, S. (1999), *Statistical Methods in Software Engineering*, Springer, New York.

Snyder, D. & Miller, M. (1991), *Random Point Processes in Time and Space*, Springer-Verlag, New York.

Number of claims	Number of cars
0	13,987
1	1,243
2	379
3	103
4	34
5+	29
2,620	15,775

Table 1: Frequency distribution of warranty claims.

With reparameterization				Without reparameterization			
1.00	0.83	-0.01	0.65	1.00	-0.98	0.95	-0.91
	1.00	-0.53	0.96		1.00	-0.99	0.97
		1.00	-0.73			1.00	-0.99
			1.00				1.00

Table 2: Correlation matrix of $\hat{\beta}$.

Time	Number of claims	%	$q = \dim(\hat{\beta})$
150	184	7.0%	1
200	457	17.4%	2
250	874	33.4%	4
300	1,392	53.1%	4
350	1,791	68.4%	4
400	2,160	82.4%	4
450	2,426	92.6%	4
500	2,555	97.5%	4
550	2,615	99.8%	4
571	2,620	100.0%	4

Table 3: Number of claims observed at every given time.

s	150	200	250	300	350	400	450	500	550
Prob.	34.6%	32.6%	25.6%	38.3%	53.6%	64.2%	75.8%	86.2%	96.0%

Table 4: Approximated coverage probability of 95% plug-in intervals.

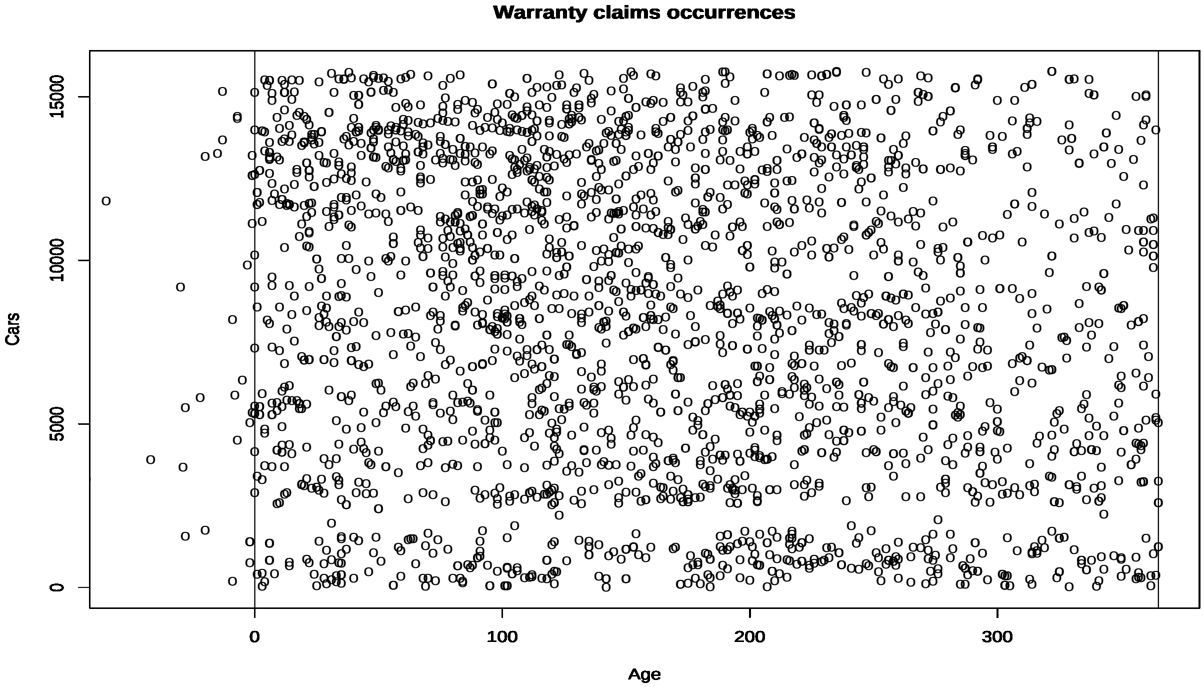

Figure 1: Warranty claims occurrences (time of sale is the origin).

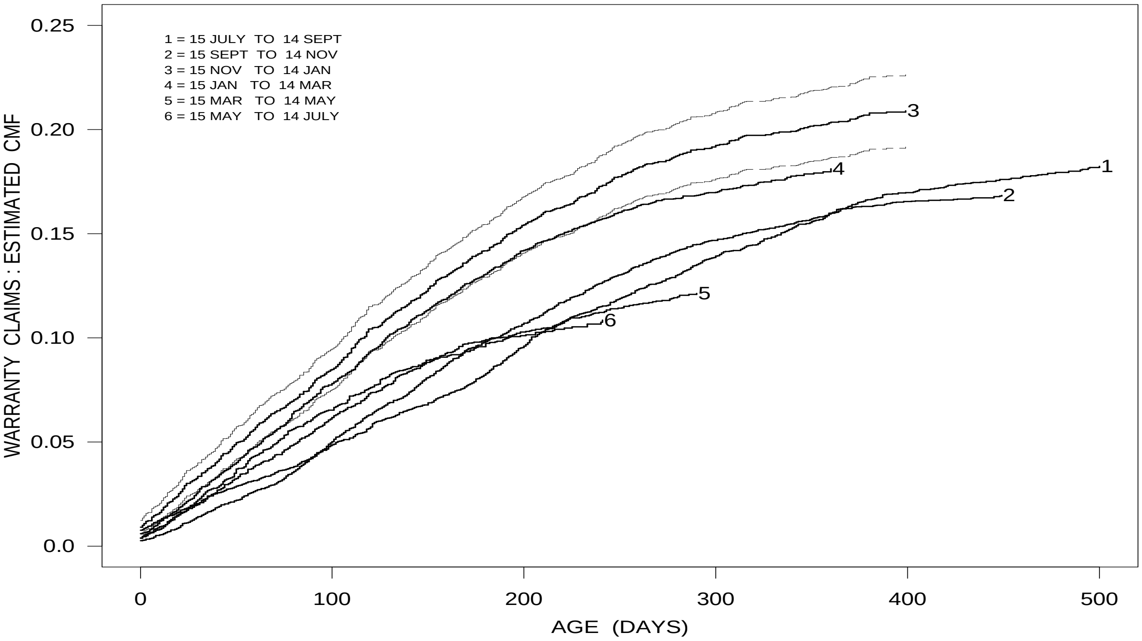

Figure 2: Cumulative mean function estimates for the six production periods; 95% pointwise confidence bands are displayed for the third group.

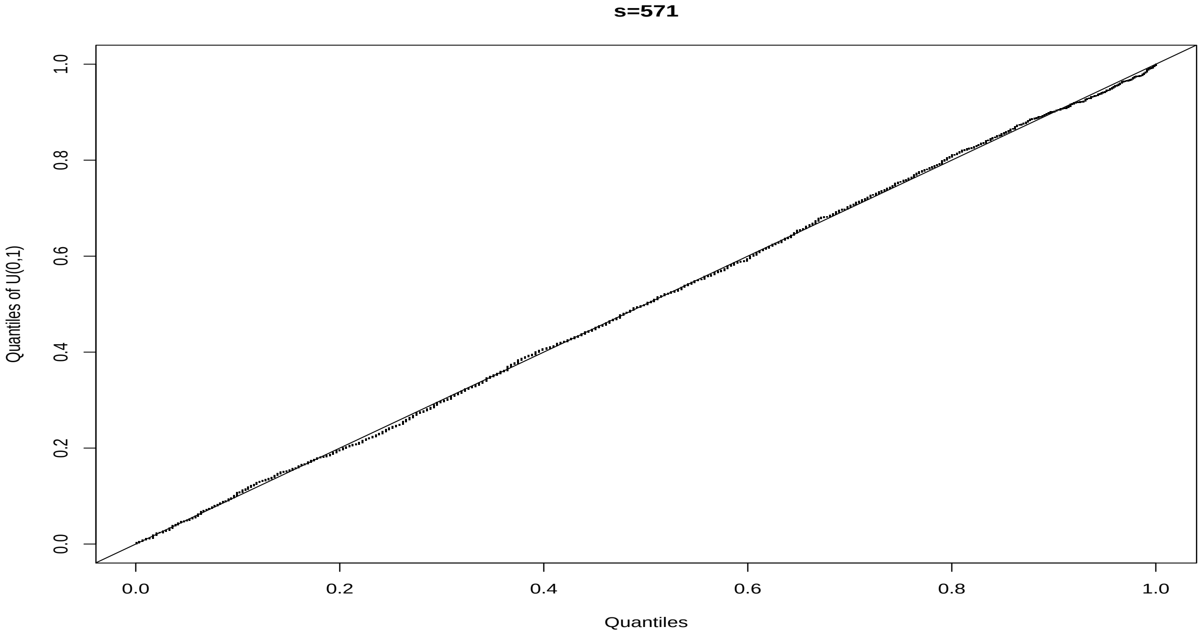

Figure 3: Quantile-quantile plot of the $u_{ij}(\hat{\beta})$'s.

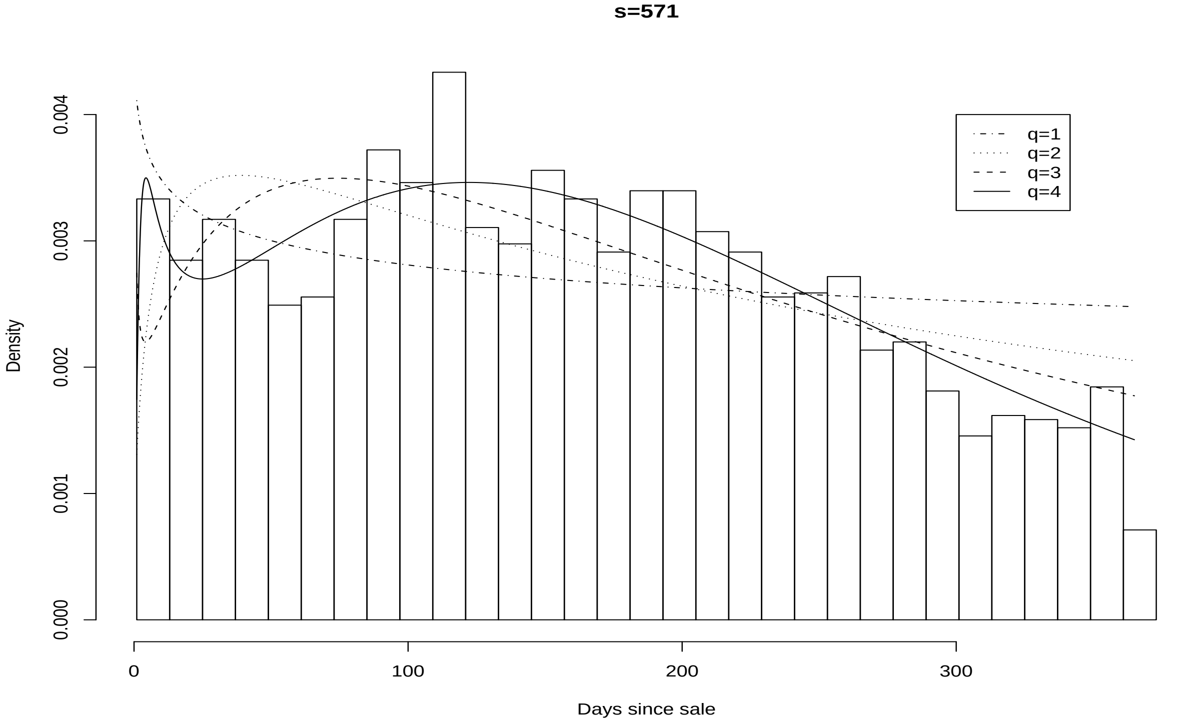

Figure 4: Histogram of the occurrence times with different estimated densities.

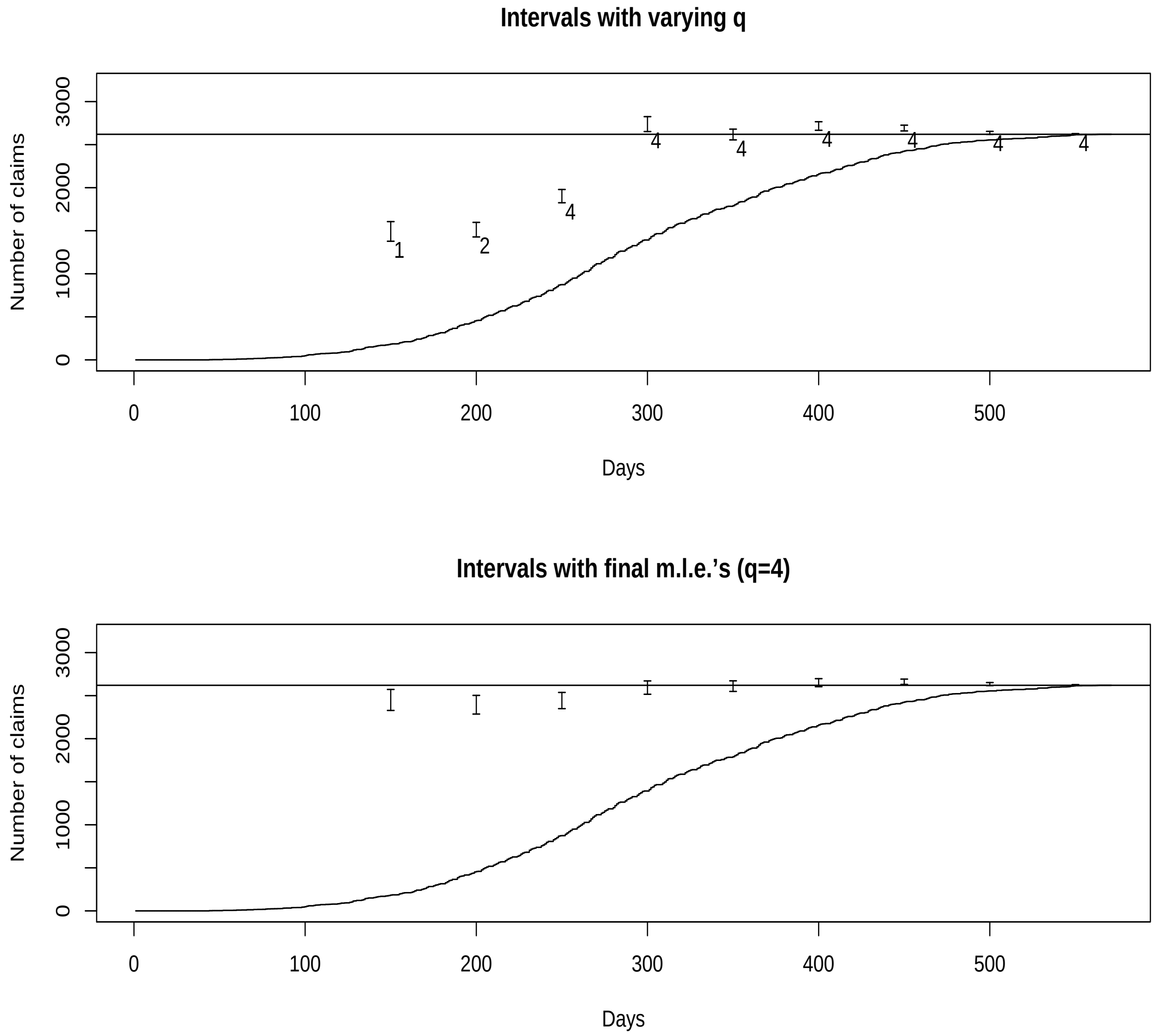

Figure 5: Non-calibrated 95% prediction intervals for the total number of warranty claims. The increasing solid curve represents the total number of claims known at that point and the digit next to each interval in the upper panel is the degree of the polynomial used at that time.

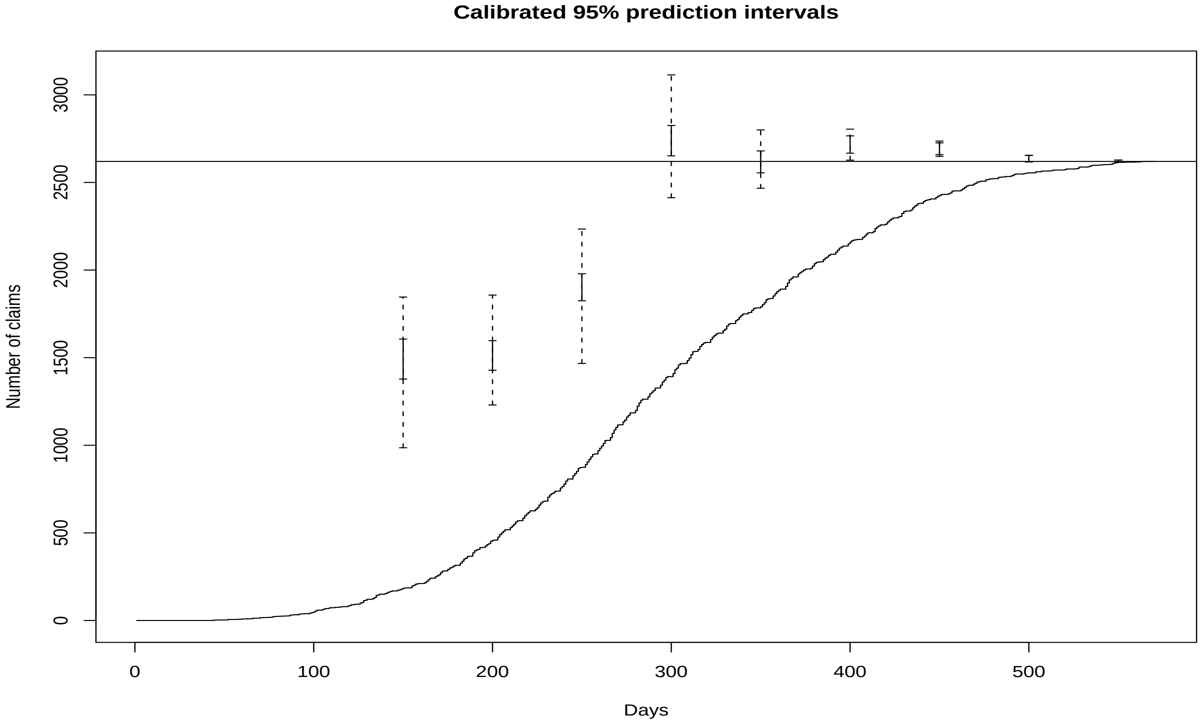

Figure 6: Calibrated 95% prediction intervals for the total number of warranty claims. A dotted line shows the difference between a calibrated and a non-calibrated interval.

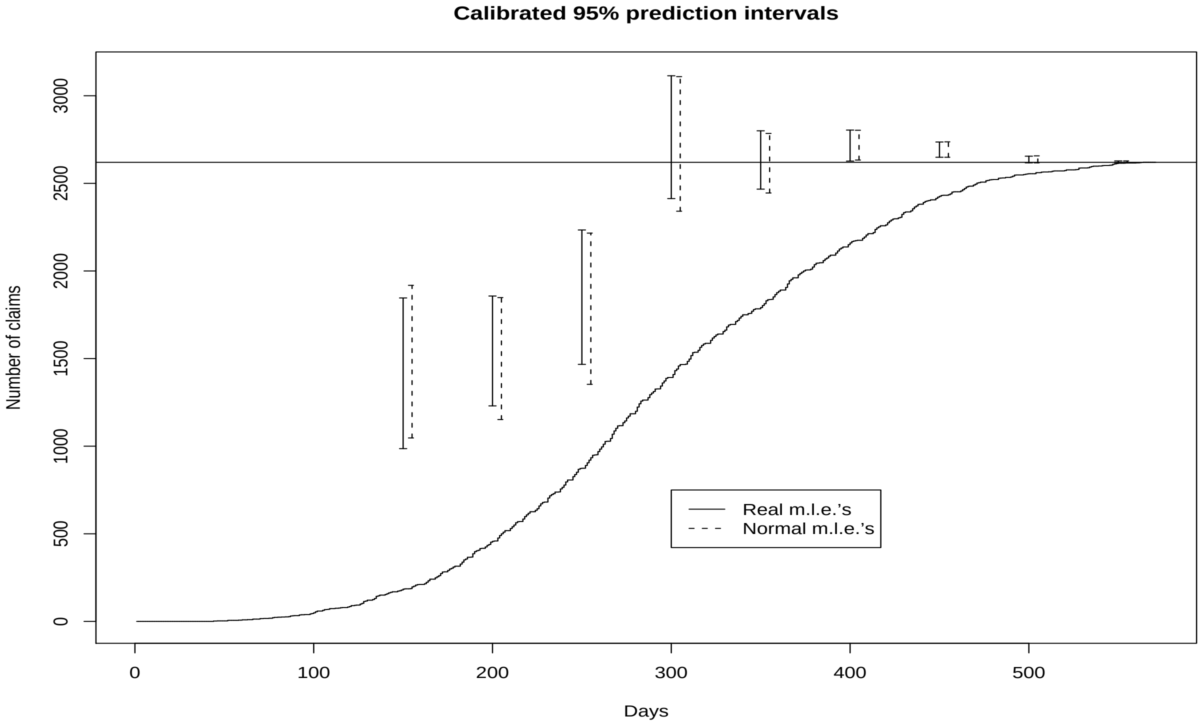

Figure 7: Comparison of calibration methods; we compare 95% prediction intervals for the total number of warranty claims calibrated using the "real" and the "normal" mle's

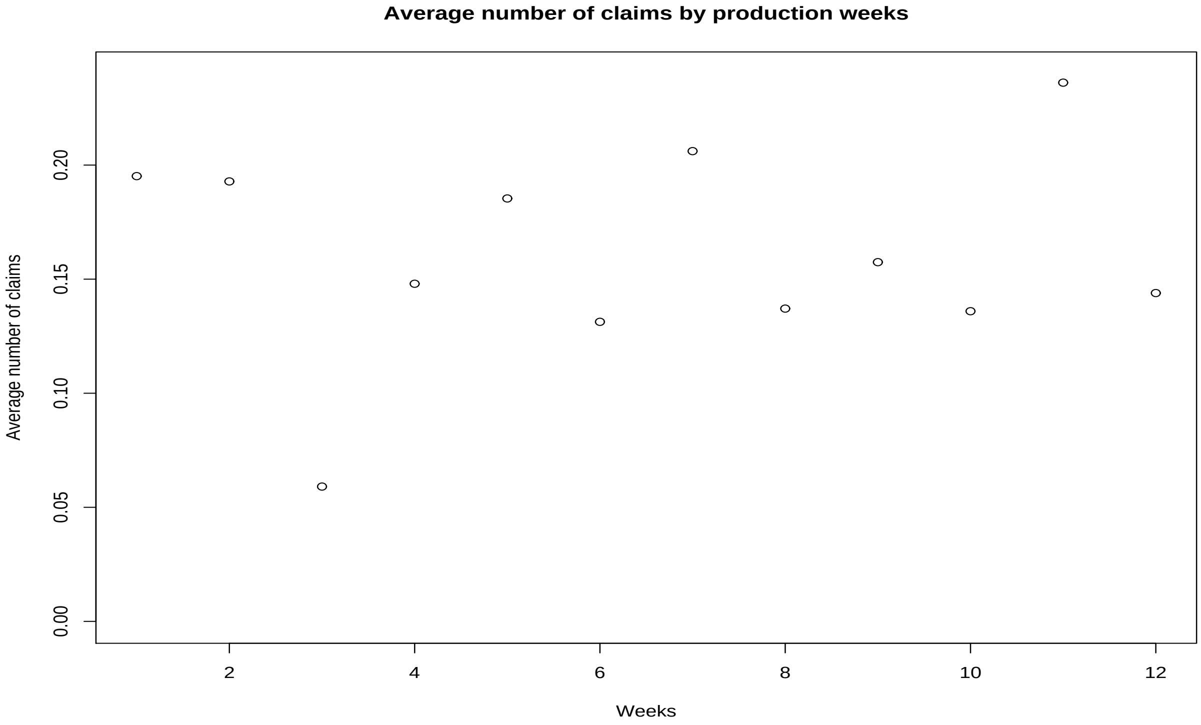

Figure 8: Average number of warranty claims observed for cars produced over different weeks.

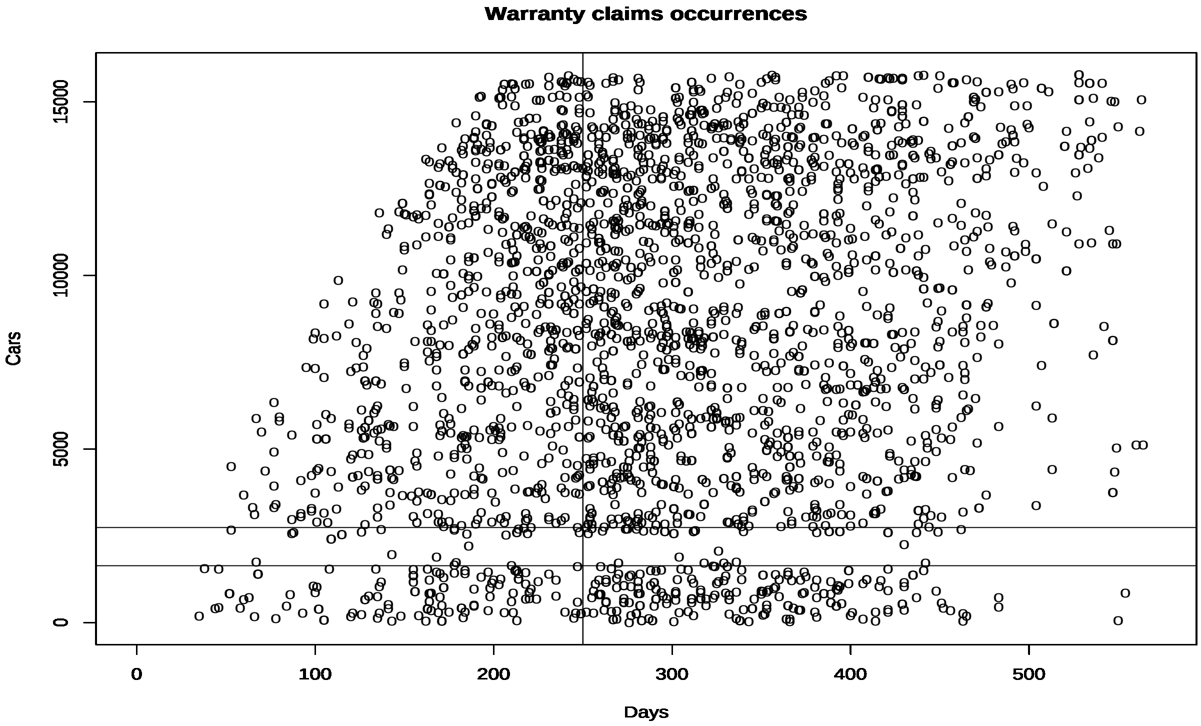

Figure 9: Warranty claims occurrences where the x-coordinate is the calendar time. The processes between the two horizontal lines are the cars produced during the third week.

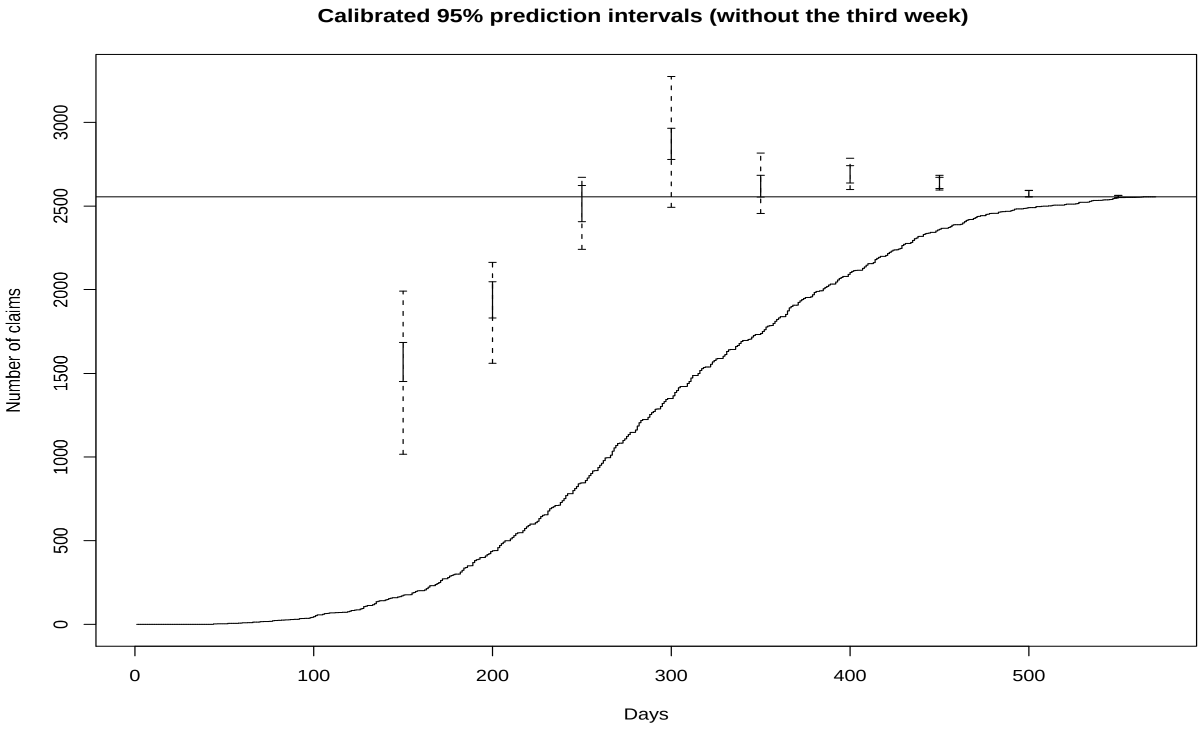

Figure 10: Calibrated 95% prediction intervals for the total number of warranty claims when cars produced during the third week are removed. A dotted line shows the difference between a calibrated and a non-calibrated interval.