Les Cahiers du GERAD ISSN: 0711–2440

A Nonhomogeneous Poisson process predictive model using maximum entropy prior random effects with application to predict purchases

L. Khribi, M. Fredette,
B. MacGibbon, J.-F. Ouellet

G–2018–75

October 2018

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2018
– Bibliothèque et Archives Canada, 2018

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2018
– Library and Archives Canada, 2018

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

A Nonhomogeneous Poisson process predictive model using maximum entropy prior random effects with application to predict purchases

Lotfi Khribi [b]
Marc Fredette [a,c]
Brenda MacGibbon [a,b]
Jean-François Ouellet [d]

[a] *GERAD, Montréal (Québec), Canada, H3T 2A7*

[b] *Department of Mathematics, Université du Québec à Montréal, (Québec) Canada, H3C 3A7*

[c] *Department of Management Sciences, HEC Montréal, Montréal (Québec), Canada, H3T 2A7*

[d] *Department of Entrepreneurship and Innovation, HEC Montréal, Montréal (Québec), Canada, H3T 2A7*

khribi.lotfi@uqam.ca
marc.fredette@hec.ca
macgibbon.brenda@gmail.com
jean-francois.ouellet@hec.ca

October 2018
Les Cahiers du GERAD
G–2018–75

Abstract: Top-tier customers—that is, those 20% of customers that typically bring in 80% of all profits—are extremely valuable to companies. In the many instances in which organizations attribute top-tier status to customers based on their consumption behaviour within a specific period, such as a year, it becomes very important to determine, during this period, how likely those gold customers are to retain their top-tier status going into the next period. This allows better planning at the corporate level, but can also allow for corrective measures or special retention efforts to be deployed. For this, we develop a model of intra-periodic forecasting of customer behaviour that allows for a continuous re-estimation of customer status or value according to calendar time, based on historical data and year-to-date information rather than existing models that predict customer churn or customer lifetime value either at the beginning of a period or on a continuous basis according to the evolution of inter-purchase time. Our model uses nonhomogeneous Poisson processes with possible heterogeneity amongst the individual units modeled with higher moment maximum entropy prior random effects instead of the gamma prior. We empirically assess the performance of such a model with a real data set from a loyalty program at a major commercial airline and compared its adequacy to the negative binomial model using the conjugate gamma prior.

Keywords: Recurrent events, mixed-Poisson, nonhomogeneous Poisson process, maximum entropy principle, maximum likelihood, forecasting

1 Introduction

It is rather well accepted in the academic and business literature that around 20% of any firm's customers bring in some 80% of corporate revenues. Past research has indeed confirmed this Pareto-like distribution of customer lifetime value across various segments. For instance, in their research conducted in the airline industry, Rust, Lemon, and Zeithaml (2004) show that only 11.6% of customers at a leading U.S. airline have a lifetime value of $500 or more and that this segment makes up approximately 50% of that airline's total customer equity.

Considering the importance of so-called "Gold Customers" for every organization, predicting the activity levels of these customers, and figuring out whether they will remain top-tier customers over the next period should hold considerable appeal. Indeed, such prediction would allow better planning and allocation of resources across segments and customer tiers, and drive the overall strategy as to whether it should focus on retaining top-tier customers or acquiring new ones. In this context, several general approaches have been developed to model customer behaviour, most of which stemming from research pertaining to the evaluation and forecasting of customer lifetime value (e.g., Borle, Singh, and Jain 2008 and Venkatesan, Kumar, and Bohling 2007). Under this paradigm, the focus lies on evaluating the total number of transactions or revenues per customer by the end of a specific period in order to steer investments towards certain specific customers as a way to manage customer equity. Although not specifically developed with the top-tier customer in mind, these models could still be used to determine, at the beginning of a new period, whether each customer is most likely to remain a "Gold customer," or to move down or even defect.

While these models perform generally well and could provide significant benefits in the evaluation and forecasting of top-tier customer "downward-migration," their contribution to other down-to-earth and pragmatic managerial concerns and marketing preoccupations is rather limited. For example, an advertising manager might be interested in knowing which top-tier customers are most at risk of loosing top-tier status, allowing better allocation of resources to better reach these customers. Or maybe, a product manager might be interested in finding out at which period demand is likely to be higher or lower in order to manage pricing differentially across sub-periods to capture as much of the value as possible. Also taking the example of a customer relationship manager in a loyalty program at an airline may want to know in October what is the likelihood of a Gold Customer to have accrued enough air miles to remain a Gold Customer going into the next year in order to help plan end-of-year special offers. Or, a key account executive may want to know whether a client is to remain a "Gold Customer" this year despite a low number of transactions over the first three quarters, and whether he or she should trigger special recovery efforts.

Under the hood, existing models consider the flow of transactions to be homogeneous within the period. However, in many product categories, sales are affected by cyclical events like seasons or holiday periods. Alternatively, these existing models could be used to forecast the level of transaction in a specific period for each customer. Because it relies on past data for comparable periods, this approach yields two major drawbacks: (1) Due to the usually limited amount of historic data (e.g., companies preserve customer data for a limited period of time, oftentimes for 3 years or less), it greatly extends prediction intervals, oftentimes past a reasonable and managerially useful threshold; and (2) it considers these periods to be isolated from one another—that is, it fails to take into account the effects of preceding (or future) periods.

To overcome these problems and address these managerial preoccupations, we develop a predictive model allowing finite-horizon prediction of recurrent events using flexible nonhomogeneous Poisson processes with higher moment maximum entropy prior random effects instead of the gamma prior used in the model proposed by Fredette and Lawless (2007) to forecast automobile warranty claims. We empirically assess the performance of such predictive model with a real data set from a loyalty program at a major commercial airline. Note that the choice of a gamma prior in the model proposed by Fredette and Lawless (2007) was motivated by its nice mathematical properties when used with Poisson processes.

A key objective in our research is to improves the predictive model proposed by Fredette and Lawless (2007) for the problem at hand in three ways.

- The use of the higher moment maximum entropy prior instead of the gamma prior as a random effects when we have a possible heterogeneity amongst the individual units.
- The use of spline function instead of Laguerre polynomials for the non-negative function $f(t;\beta)$ which represents the shape of the rate function. It is important to use a function that is very flexible and that would be decreasing quickly as we approach the end of the year and spline function have nice properties with good ability to fit sharply curving shapes.
- The use of the Weibull survival function to reflect the fact that some customers are likely to leave the program over time.

The remainder of this paper is organized as follows. In Section 2, we outline the assumptions relative to the use of mixed nonhomogeneous Poisson models for predicting recurrent events. Then, we describe the maximum entropy principle before providing details on the development and estimation of our model. In Section 3, the performance of the proposed approach is studied in a particular data setting from a major airline company and its comparison with the negative binomial (NB) model using the conjugate gamma prior. We finally conclude with a discussion of our results, limitations, and avenues for future research in Section 4.

2 Prediction of recurrent events with mixed Poisson models

Here we present the nonhomogeneous Poisson processes, then we recall the maximum entropy principle used to take account the possible heterogeneity amongst the individuals, finally we define the Poisson-maximum entropy model.

2.1 Mixed nonhomogeneous Poisson processes

The motivation for our work lies in the prediction of a real data set from a loyalty program or other events that occur for individual units or subjects in a population. That is, there is a finite population of units $i = 1, \ldots, n$ and we wish to predict the total number of events, for an individual or the whole population over a specified time period $(0, T]$ 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, this interval $(0, T]$ would typically refer to a calendar or fiscal year time period, and the various t_i's would usually take the same value for all units.

Let t represent the number of days elapsed since the beginning of the calendar year, and let $N_i(u, v)$ denote the number of events in the age interval $u < t \leq v$. The objective is then to predict $N_i(0, T)$ the total number of events associated with unit i up until time T, where T could possibly represent the end of the year. Of course, as an added benefit, it may eventually be useful to predict the total number of events during the whole year by predicting

$$N_+ (0, T) = \sum_{i=1}^{n} N_i (0, T) . \tag{1}$$

Of course, $N_i(0, T)$ will ultimately be known for each i once the calendar year is over. However, in several situations, it might be useful to predict this value on the basis of previous experience with this unit but also on the basis of events that have already occurred during that calendar year. Because $N_i(0, t_i)$ is known for each $i = 1, \ldots, n$ where $0 \leq t_i \leq T$, prediction of $N_i(0, T)$ is equivalent to predicting $N_i(t_i, T)$.

For convenience, we consider continuous time processes where two events cannot occur simultaneously. From this point on, we also write $N(t)$ for $N(0, t)$. Different types of recurrent events processes are discussed in the literature on point processes (Grandell, 1997). These are all characterized by an event intensity function

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

where $H(t)$ denotes the history of the process up to time t. Poisson processes are Markovian because (2) depends only on t. The intensity, or rate, function is then simply denoted by $\lambda(t)$, and

$$N(t) \sim PP(\lambda(t))$$

means that $N(t)$ is a nonhomogeneous 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 $N(s_1, t_1)$ and $N(s_2, t_2)$ in two nonoverlapping time intervals $(s_1, t_1]$ and $(s_2, t_2]$ are independent. These two properties make Poisson processes easy to use with prediction problems involving recurrent events. However, in populations with heterogeneous units, it is generally 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 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 $\mathrm{E}[N(0, T)] = \alpha$, 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 $(0, T]$.

2.2 The Maximum entropy principle

The entropy of a probability density $\pi(\alpha)$ is a measure of the amount of information contained in the density which was first defined by Shannon (1948) as

$$H \quad = \quad -\int_{\alpha} \pi(\alpha) \ln(\pi(\alpha))d\alpha.$$

The goal is to maximize H subject to certain side conditions. The usual choice to determine $\pi(\alpha)$ is to use a finite set of expectations $\mu_j = \mathbb{E}[\phi_j(\alpha)]$ of known functions $\phi_j(\alpha), j = 0, ..., k$. This is called the matching moment (MM) estimation method. These known functions $\phi_j(\alpha)$ are often the arithmetic non-central moments of the form $\phi_j(\alpha) = \alpha^j, j = 0, ..., k$. In this simple case using the arithmetic non-central moments maximizing the likelihood yields the same estimates as the matching moment method (Mohammad-Djafari (1992)).

To find the function $\pi(\alpha)$ that maximizes the entropy of this nonlinear problem using matching moments we form the Lagrangian

$$L \quad = \quad \int_{\mathbf{R}^+} \pi(\alpha) \ln(\pi(\alpha))d\alpha + \sum_{j=0}^{k} \gamma_j \Big(\int_{\mathbf{R}^+} \alpha^j \pi(\alpha)d\alpha - \mu_j \Big).$$

where γ is a vector of Lagrange multipliers. Applying the Lagrange's multiplication method (Weinstock, 1952). The following k moment maximum entropy prior distribution is defined by:

$$\pi(\alpha|\gamma) \quad = \quad A \exp\,(-\sum_{j=1}^{k} \gamma_j \alpha^j), \tag{3}$$

where $\gamma = (\gamma_1, \gamma_2, ..., \gamma_k)$ and with normalization constant defined by:

$$A \quad = \quad \frac{1}{\int_{\mathbb{R}^+} \exp\,(-\sum_{j=1}^{k} \gamma_j \alpha^j)d\alpha}.$$

2.3 Model specification of the general Poisson-maximum entropy model

To consider scenarios in which heterogeneity is observed among the processes for different units, we incorporate unobservable *iid* random effects in our model by using the k moment maximum entropy distribution given k non-central moments. The general Poisson-MaxEnt model considered in this article is

$$N_i(t)|\alpha_i \sim PP(\alpha_i f(t; \beta)), \tag{4}$$

$$\pi(\alpha_i;\gamma) = A\exp\,(-\sum_{j=1}^{k}\gamma_j\alpha_i^j), \quad i=1,\ldots,n.$$

We will propose later an efficient criterion which allows us to determine the number of moments necessary for the k-moment priors in the model (4). And for our particular data set studied here it will be seen further that the model (4) performs very well when the number of moments k is equal to 4.

2.4 Prediction

We seek to construct prediction intervals for a future random variable Y, given observed data $X = x$. Such intervals are of the form $(L(x), U(x))$, and we attempt to find intervals where $P[L(X) \leq Y \leq U(X)]$ equals some specified fixed value $1-\zeta$, in which case $(L(x), U(x))$ is called a $1-\zeta$ prediction interval (e.g., Lawless and Fredette, 2005) and $1-\zeta$ is called its coverage probability.

In the context discussed in this article, we wish to use the information regarding the n processes that are available at a certain given time to make predictive statements about the remaining number of events that would be observed. As it is the focus of our article, only the prediction of a single count $N_i(t_i, T)$ is discussed here. It is easy to make the extension to predict the sum of all counts (1).

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, $\tau_i(t_i) = \{\tau_{i1},\ldots,\tau_{iN_i(t_i)}\}$. The conditional distribution $\pi(\alpha|(\mathbf{N}(t),\tau);\gamma,\beta)$ of the random effects is defined by:

$$\begin{aligned}\pi(\alpha|(\mathbf{N}(t),\tau);\gamma,\beta) &= \frac{\mathbf{L}(\alpha,\beta|(\mathbf{N}(t),\tau))\pi(\alpha;\gamma)}{\int_\alpha \mathbf{L}(\alpha,\beta|(\mathbf{N}(t_1),\tau))\pi(\alpha;\gamma)d\alpha} \\ &= \prod_{i=1}^{n}\frac{\alpha_i^{N_i(t_i)}\exp\Big(-\alpha_i(\gamma_1+F(t_i;\beta))-\sum_{j=2}^{k}\gamma_j\alpha_i^j\Big)}{\int_{\alpha_\mathbf{i}}\alpha_i^{N_i(t_i)}\exp\Big(-\alpha_i(\gamma_1+F(t_i;\beta))-\sum_{j=2}^{k}\gamma_j\alpha_i^j\Big)d\alpha_i} \end{aligned} \tag{5}$$

where

$$\mathbf{L}(\alpha,\beta|(\mathbf{N}(t),\tau)) = \prod_{i=1}^{n}\left(\prod_{j=1}^{N_i(t_i)}\alpha_i f(\tau_{ij};\beta)\right)e^{-\alpha_i F(t_i;\beta)}$$

and $\mathbf{N}(t) = (N_1(t),\ldots,N_n(t))$.

Then for each process i the $\alpha_i|(\mathbf{N}(t),\tau)$ has a distribution defined by:

$$\pi(\alpha_i|(\mathbf{N}(t),\tau);\gamma,\beta) = \frac{\alpha_i^{N_i(t_i)}\exp\Big(-\alpha_i(\gamma_1+F(t_i;\beta))-\sum_{j=2}^{k}\gamma_j\alpha_i^j\Big)}{\int_{\alpha_i}\alpha_i^{N_i(t_i)}\exp\Big(-\alpha_i(\gamma_1+F(t_i;\beta))-\sum_{j=2}^{k}\gamma_j\alpha_i^j\Big)d\alpha_i}$$

where $F(t;\beta) = \int_0^t f(u;\beta)du$.

Hence, using this conditional density, the density function for $N_i(t_i,T)|N_i(t_i;\gamma,\beta)$ is given by

$$\begin{aligned}P[N_i(t_i,T) = n|N_i(t_i);\gamma,\beta] &= \frac{(F(T;\beta)-F(t_i;\beta))^n}{n!\int_{\alpha_i}\alpha_i^{N_i(t_i)}\exp\Big(-\alpha_i(\gamma_1+F(t_i;\beta))-\sum_{j=2}^{k}\gamma_j\alpha_i^j\Big)d\alpha_i} \\ &\times\int_{\alpha_i}\alpha_i^{(N_i(t_i)+n)}\exp\Big(-\alpha_i(\gamma_1+F(T;\beta))-\sum_{j=2}^{k}\gamma_j\alpha_i^j\Big)d\alpha_i. \end{aligned} \tag{6}$$

Note that the occurrence times do not appear in this distribution; only knowledge of $N_i(t_i)$ is needed to determine this conditional distribution. However, the occurrence times will enter into the estimation of the model parameters β.

2.5 Discrepancy measure

In order to compare the adequacy of the point prediction method for $\mathbf{N}_i(t_i, T)$ obtained from our model, we calculate the prediction error between the real value of $\mathbf{N}_i(t_i, T)$ and its predictor $\hat{\mathbf{N}}_i(t_i, T)$.

For this, we used a discrepancy measure. Discrepancy is measure equal to the root mean square prediction error between the predicted value obtained using a specific prediction model and the estimator obtained here using our prediction method. It is used in many different types of applications (Cooray, 2006). Our discrepancy measure is defined as follow:

$$D = \sqrt{\frac{\sum_{i=1}^{n}\left(N_i(t_i,T) - \hat{N}_i(t_i,T)\right)^2}{n}}. \tag{7}$$

where the point predictor $\hat{N}_i(t_i, T)$ is defined by $\hat{N}_i(t_i, T) = \mathbb{E}[N_i(t_i, T)|\mathbf{N}(t_1); \gamma, \beta] = (F(T; \beta) - F(t_i; \beta))\mathbb{E}[\alpha_i|\mathbf{N}(t_1); \gamma, \beta]$ with $\mathbb{E}[\alpha_i|\mathbf{N}(t); \gamma, \beta]$ is the posterior mean of $\alpha_\mathbf{i}|(\mathbf{N}(t); \gamma, \beta)$ given by (5) and where all the unknown parameters are replaced by their estimations (see Section 2.6).

The posterior distributions (5) will not have a known closed form, but it is a rather complicated high dimensional density, which makes direct inference almost impossible. For this reason, we can generate from this posterior distribution a large number of samples using Markov chain Monte Carlo (MCMC) implemented in WinBUGS (Spiegelhalter et al., 2003), and from these samples, we can obtain appropriate parameters estimate like the posterior mean of $\alpha|(\mathbf{N}(t_1); \gamma, \beta)$, where γ and β are estimated by the method described in the next section.

2.6 Estimating unknown Poisson-Maximum Entropy parameters

In this section, we will discuss ways to estimate the vector of the parameters γ and β in the general Poisson-MaxEnt model (4). These estimates will then substitute for the real parameters in the point prediction and prediction intervals previously mentioned. In the study (Khribi et al., 2016) where the predictive model for the prediction of recurrent events uses homogeneous Poisson processes, the parameters of the maximum entropy prior distribution were estimated by two methods, the usual maximum entropy estimation method which uses matching moments (MM) and the maximum likelihood method, referred to as the Pseudo-MaxEnt. Because we have seen in Khribi et al., (2017) that the MLE-MaxEnt method is computationally less complex than MM method when $k > 2$, we use in this study the MLE-MaxEnt method to estimate the parameters γ and β.

2.6.1 The MLE-Maximum Entropy method for the Poisson-MaxEnt model

For the empirical Bayes MaxEnt model (4), we introduce the MLE-maximum entropy (MLE-MaxEnt) method using MLE for estimating the vector of the parameters γ and β. We start by construct the marginal likelihood L of the empirical Bayes general Poisson-Maximum Entropy model (4)

$$\begin{aligned}
L(\gamma, \beta|(\mathbf{N}(t), \tau)) &= \int_\alpha L(\alpha, \beta|(\mathbf{N}(t), \tau))\pi(\alpha; \gamma)d\alpha \\
&= \int_\alpha \left[\prod_{i=1}^{n}\left(\prod_{j=1}^{N_i(t_i)} \alpha_i f(\tau_{ij}; \beta)\right)e^{-\alpha_i F(t_i;\beta)}\right]\left(\prod_{i=1}^{n}\frac{e^{(-\sum_{j=1}^{k}\gamma_j\alpha_i^j)}}{\int_{\alpha_i} e^{(-\sum_{j=1}^{k}\gamma_j\alpha_i^j)}d\alpha_i}\right)d\alpha \\
&= \prod_{i=1}^{n}\left[\frac{\left(\prod_{j=1}^{N_i(t_i)} f(\tau_{ij}; \beta)\right)}{\int_{\alpha_i} e^{(-\sum_{j=1}^{k}\gamma_j\alpha_i^j)}d\alpha_i}\int_{\alpha_\mathbf{i}} \alpha_i^{N_i(t_i)} e^{\left(-\alpha_i(\gamma_1 + F(t_i;\beta)) - \sum_{j=2}^{k}\gamma_j\alpha_i^j\right)}d\alpha_i\right] \\
&= \prod_{i=1}^{n}\left[\frac{\left(\prod_{j=1}^{N_i(t_i)} f(\tau_{ij}; \beta)I_i(N_i(t_i))\right)}{\int_{\alpha_i} e^{(-\sum_{j=1}^{k}\gamma_j\alpha_i^j)}d\alpha_i}\right]
\end{aligned} \tag{8}$$

with

$$I_i(N_i(t_i)) = \int_{\alpha_i} \alpha_i^{N_i(t_i)} e^{\left(-\alpha_i(\gamma_1 + F(t_i;\beta)) - \sum_{j=2}^{k} \gamma_j \alpha_i^j\right)} d\alpha_i. \tag{9}$$

The log-likelihood is given by

$$l(\gamma, \beta|(\mathbf{N}(t), \tau)) = \sum_{i=1}^{n} \left[-\log\Big(\int_{\alpha_i} e^{(-\sum_{j=1}^{k} \gamma_j \alpha_i^j)} d\alpha_i \Big) + \sum_{j=1}^{N_i(t_i)} \log\Big(f(\tau_{ij};\beta)\Big) + \log\Big(I_i(N_i(t_i))\Big) \right]. \tag{10}$$

Using Lebesgue's Dominated Convergence Theorem, (Talvila, 2001) gave necessary and sufficient conditions to interchange the order of differentiation and integration for (10) which are verified here. We can find the estimate of the vector of the parameters γ and β by solving the score equations,

$$\frac{\partial l(\gamma, \beta|(\mathbf{N}(t), \tau))}{\partial \gamma_j} = \sum_{i=1}^{n} \left[\frac{\int_{\alpha_i} \alpha_i^j e^{-\sum_{j=1}^{k} \gamma_j \alpha_i^j} d\alpha_i}{\int_{\alpha_i} e^{-\sum_{j=1}^{k} \gamma_j \alpha_i^j} d\alpha_i} - \frac{I_i(N_i(t_i)+j)}{I_i(N_i(t_i))} \right] = 0,$$

$$\frac{\partial l(\gamma, \beta|(\mathbf{N}(t), \tau))}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{\frac{\partial f(\tau_{ij};\beta)}{\partial \beta}}{f(\tau_{ij};\beta)} - \frac{\partial F(t_i;\beta)}{\partial \beta} \frac{I_i(N_i(t_i)+1)}{I_i(N_i(t_i))} \right] = 0,$$

The analytic solutions to these score equations are impossible to obtain; we thus use a numerical method to estimate directly the vector of the parameters γ and β that maximize the log-likelihood (10). We have chosen MATLAB "fminsearchbnd", a nonlinear optimization method which is derivative-free and allows bounds on the variables.

2.7 Plug-in prediction intervals

A prediction interval for $N_i(t_i, T)$ is an interval $[L(N(t), \tau(t)), U(N(t), \tau(t))]$ such that

$$P[L(N(t), \tau(t)) \leq N_i(t_i, T) \leq U(N(t), \tau(t)); \gamma, \beta] = 1 - \zeta.$$

Such an interval is called an exact $1 - \zeta$ prediction interval for $N_i(t_i, T)$. In most settings (including the one considered in this paper), one cannot find exact prediction intervals when the parameters γ, and β 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 $1 - \zeta$. This can be accomplished in one way by finding an interval $[L, U]$ such that

$$P[L \leq N_i(t_i, T) \leq U; \hat{\gamma}, \hat{\beta}] = 1 - \zeta, \tag{11}$$

where only $N_i(t_i, T)$ is treated as a random variable, and where $\hat{\gamma}$ and $\hat{\beta}$ are the MLE estimates obtained from the likelihood function based on the observed data and defined by (8).

The interval (11) is called a "plug-in" $1 - \zeta$ prediction interval. Essentially, our method assumes that (6) is the true distribution and that the true parameter values are in fact $(\hat{\gamma}, \hat{\beta})$ and thus ignores completely the uncertainty in $(\hat{\gamma}, \hat{\beta})$ relative to (γ, β). When the observed data set is very large, so that $(\hat{\gamma}, \hat{\beta})$ can be assumed close to (γ, β), then the coverage probability of this interval will be close to 1-ζ. However, in the case were the observed data set is not very large, our method can be improved by calibrating the plug-in intervals as was done by Fredette and Lawless (2007). We note that the calibration procedure still provides an approximate coverage probability for the prediction interval (11).

Plug-in prediction intervals with an approximate coverage probability of $1 - \zeta$ can easily be obtained from the $\zeta/2$ and the $1 - \zeta/2$ quantiles based on the predictive probability function $P[N_i(t_i, T) = n|N_i(t_i); \hat{\gamma}, \hat{\beta}]$ given by (6).

3 Predicting the number of flights taken by frequent flyers

The context of this research is frequent flyer status within a specific airline loyalty program. Frequent flyer programs involve the systematic collection of detailed information regarding members' flying activities, thus allowing prediction of individual activity level based on the data already observed. The database at hand was obtained from the loyalty program of a major American commercial airline. It includes information on individual top-tier frequent flyers for a period of 3 years starting January 1st, 2004 and provides, for each frequent flyer, a unique identifier along with the various dates that flights have been flown. To qualify for top-tier "Gold" membership, each frequent flyer had to fly at least 20 times over the first calendar year—that is, between January 1st, 2004 and December 31st, 2004 inclusively.

The quantities we wish to predict are $N_i(366, 731)$ for each frequent flyer i—that is, the number of flights taken by each frequent flyer between the first and last day of the second calendar year. The dataset contains such data for 5,000 frequent flyers. In the second year, each of them had actually flown between 0 and 158 flights. Table 1 gives the distribution of total number of flights in year 2 for those who had qualified for top-tier membership at the end of year 1.

Table 1: Distribution of the number of flights taken over year 2 by frequent flyers who had qualified for top-tier status by the end of year 1

Number of flights	Total
0	94
1∼10	721
11∼19	1112
20+	3073

During a given year, the managers want to estimate, for the new year in progress, the eventual number of flights to be flown by each frequent flyer according to past data. Here we show how the methods in Section 2 can be used to predict the number of flights to be flown by each member, or the total flights to be flown for a group of, or all, frequent flyers.

Figure 1 provides a plot of the number of flights each day for the first 3 years considered, for those with top-tier frequent flyer membership after the first year. This graph clearly shows the seasonal (that is, nonhomogeneous) character of the flying habits of top-tier frequent flyers. It also shows that a number of flights flown daily diminishes with time as members from this top-tier cohort leave the program and/or the company, or diminish flying habits.

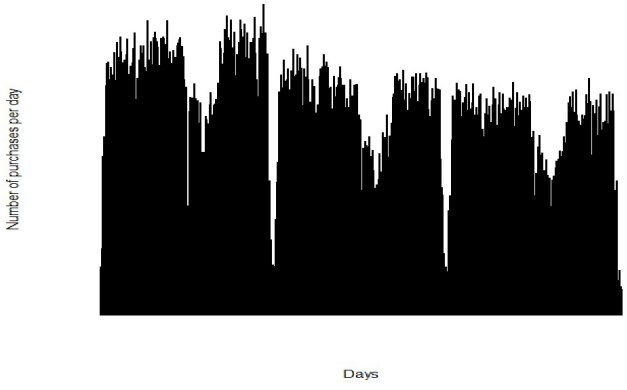

Figure 1: Total number of purchases per day over 3 years

We now propose to use model (4) to predict the total number of flights by a given top-tier frequent flyer over a calendar year. Fredette and Lawless (2007) proposed a similar prediction model for forecasting automobile warranty claims, but instead of using a higher order maximum entropy prior for the random effects he uses the gamma prior and Laguerre polynomials function for $f(t;\beta)$. The choice of a suitable parametric form for $f(t;\beta)$ in (4) is crucial, because our predictions necessarily involve extrapolation into the

future. Ideally, the shape of this function would be the same every year to reflect the periodicity of flying habits of frequent flyers. In addition, we would like to allow for a potential reduction of the amplitude of this function to reflect the fact that the number of flights usually diminishes over time.

We thus consider the function

$$f(t;\beta) = p(t-366;\beta_1,\beta_2) \times \exp\{C(d_t;\beta_3,\ldots,\beta_{K+3})\},$$

where:

- d_t is the number of the day of the year. For example, $d_1 = d_{366+1} = d_{366+365+1} = 1$ (the first year was a leap year). This will allow the function to retain the same shape year after year.
- At the beginning of the second year, we incorporate a decreasing proportion $p(.;\beta_1,\beta_2)$ to reflect the fact that some customers are likely to leave the program over time. Because of the obvious relationship between this phenomenon and a survival problem, we opted for a survival function $p(t-366,\beta_1,\beta_2) = S(t;\beta_1,\beta_2)$ such that $S(0;\beta_1,\beta_2) = 1$ and decreases thereafter. We used the Weibull survival function $S(t;\beta_1,\beta_2) = \exp\{(-t\beta_1)^{\beta_2}\}$ which is probably, along with the log-normal survival function, the most popular distribution for survival problems.
- $C(t;\beta)$ is a cubic spline. Cubic splines are continuous piecewise cubic polynomials used in curve fitting. They have been found to have nice properties with good ability to fit sharply curving shapes (Harrel, 2001). In order to use a cubic spline, we first have to determine an appropriate number of knots. Between each of these knots, the continuous function $C(t;\beta)$ is a cubic polynomial. Based on the data available after the first year, we found out that it was sufficient here to use $K = 4$ knots. In order to have approximately the same number of recurrent events between consecutive knots, the knots are the 20%, 40%, 60%, and 80% quantiles of all the occurrence times observed that 1$^{\text{st}}$ year (i.e., 70, 140, 220, and 300 days). The explicit form of this piecewise cubic polynomial is given by:

$$\begin{aligned} C(t;\beta_3,\ldots,\beta_9) =& \beta_3 t + \beta_4 t^2 + \beta_5 t^3 + \beta_6 (t-70)_+^3 + \\ & \beta_7 (t-140)_+^3 + \beta_8 (t-220)_+^3 + \beta_9 (t-300)_+^3 \end{aligned}$$

  where $(.)_+$ is the positive part of what is inside the parenthesis.

As Figure 2 shows, the use of splines in this case does allow for our model to follow rather well the bimodal distribution of flying behaviour among the top-tier frequent flyers over the first year of data, used to estimate our model.

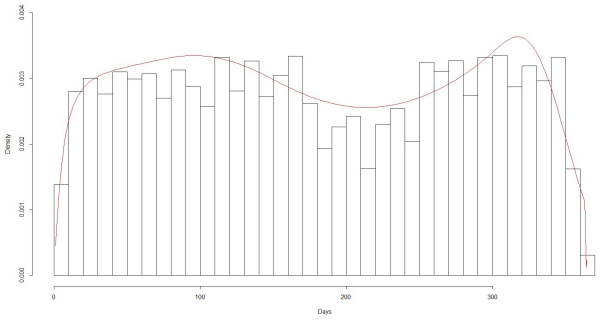

Figure 2: Adequacy of the nonhomogeneous process

3.1 Empirical tests of the prediction model proposed

In this section, we apply the general Poisson-MaxEnt model using higher moment maximum entropy prior and compared its adequacy to the NB model using the gamma prior proposed by Fredette and Lawless (2007). For this, we explore the performance of our approach by predicting the number of flights at the end of each month in year 2 to be flown by 5,000 members of such a cohort of top-tier frequent flyers within this loyalty program.

3.1.1 Likelihood ratio tests

The likelihood ratio test (LRT) is used to determine the value of k in (3). Let say we want to compare the 2-moment maximum entropy and the 4-moment maximum entropy priors on the 4-moment and 6-moment maximum entropy priors, then the test statistic is the ratio between the log-likelihood of the null model to the alternative model:

$$\Gamma \quad = \quad -2\log\Big(\frac{l(\alpha_1|N(t))}{l(\alpha_2|N(t))}\Big) \tag{12}$$

where $l(\alpha_1|N(t))$ and $l(\alpha_2|N(t))$ are the log-likelihood of the null and alternative models respectively. This is a statistical test for nested models which reject the null hypothesis with a given significance level based on the chi-squared distribution. Through successive testing using the likelihood ratio test (Wilks, 1938), we can determine the number of moments necessary for the k-moment prior in the general Poisson-MaxEnt model.

Table 2 present the likelihood ratio test (LRT) results where the last two columns indicate respectively the p-values using 2 and the 4-moment maximum entropy prior model as the null models versus the alternative models with 4 and 6 moments. Note that the last column shows us the number of moments required for our predictive model.

Based on the results in Table 2 with a significance level equal to 5%, we can say that the LRT always rejects the model (4) with 6 moments compared with the one with 4 moments. However, it always supports the model (4) with 4 moments against the model with 2 moments. This means that the LRT always recommends the use of 4 moments at the end of each month in year 2.

Table 2: The likelihood ratio test (p-value=.05) using data from the loyalty program

t_i (in days)	p-value of LRT (MaxEnt_2Moments_vs_4Moments)	p-value of LRT (MaxEnt_4Moments _vs_6Moments)	Number of Moments Suggested
397 (13 months)	**< 0.01%**	75.82%	**4**
425 (14 months)	**< 0.01%**	82.59%	**4**
456 (15 months)	**< 0.01%**	89.55%	**4**
486 (16 months)	**< 0.01%**	92.41%	**4**
517 (17 months)	**< 0.01%**	98.74%	**4**
547 (18 months)	**< 0.01%**	99.23%	**4**
578 (19 months)	**< 0.01%**	100%	**4**
609 (20 months)	**0.14%**	100%	**4**
639 (21 months)	**0.75%**	100%	**4**
670 (22 months)	**0.95%**	100%	**4**
700 (23 months)	**2.24%**	100%	**4**
731 (24 months)	**3.87%**	100%	**4**

Table 3 presents the discrepancy between the real value of $\mathbf{N}_i(t_i, T)$ and its predictor $\hat{N}_i(t_i, T)$ defined by (7), where $T = 731$ (the end of the year 2), using the different values of $t_i = (366 + 31, 366 + 31 + 28, ..., 731)$ where t_i is the number of days at the end of each month i in year 2. For example, a value of **11.03** in the first line of Table 3 means that a prediction for the end of the first month in year 2 based on this model (the general Poisson-MaxEnt model with the 4-moment prior) would be on average **11.03** flights from the real value of $\mathbf{N}_i(t_i, T)$. The likelihood ratio test stopping rule, that is, to stop at 4 moments result is confirmed in Table 3, where the average discrepancy values for the general Poisson-MaxEnt model with the 4-moment maximum entropy prior (values in bold font) are always very close to the smallest absolute error discrepancy given by the model using the 6-moment maximum entropy prior.

As a another example of the usefulness of this approach, let us consider a scenario in which, as she prepares her marketing activities for the fall season, a customer relationship manager of this loyalty program is concerned about deploying extra effort to retain those Gold customers that are in danger of not qualifying for Gold status the next year. In order to target the right customers with a costly special offer, this manager wishes to target those with a moderate chance of actually qualifying for top-tier membership as assessed using data available on August 1st of 2004. For each of these "gold" customers, our model returns the probability

Table 3: Discrepancy of point predictors with different values of t_i using data from the loyalty program

t_i (in days)	Gamma (Fredette (2007)) $(\widehat{a}_{\text{mle}}, \widehat{b}_{\text{mle}})$	*MLE 4Moments*	*MLE 6Moments*
397 (13 months)	15.72	**11.03**	10.99
425 (14 months)	13.93	**9.94**	9.92
456 (15 months)	12.37	**8.78**	8.76
486 (16 months)	11.91	**8.03**	8.02
517 (17 months)	9.28	**7.89**	7.87
547 (18 months)	8.57	**6.77**	6.75
578 (19 months)	7.62	**5.49**	5.49
609 (20 months)	6.05	**4.94**	4.92
639 (21 months)	4.43	**4.05**	4.03
670 (22 months)	3.31	**3.00**	3.00
700 (23 months)	1.82	**1.70**	1.70
731 (24 months)	0.00	**0.00**	0.00

that these customers will remain top-tier members in 2006—that is, the probability that they will fly 20 flights or more during 2005. Of course, those frequent flyers having already flown these 20 flights have a probability of 100%. Table 4 shows these probabilities for 11 segments according to how likely they are to remain "gold" customers for both predictive models: the model (4) using 4-moment maximum entropy prior and the NB model using the gamma prior proposed by Fredette and Lawless (2007). From this table, we notice that the values of the probability of being Gold for customers define by the general Poisson-MaxEnt predictive model with the 4-moment maximum entropy prior (values in bold font) are always closest to the actual proportion of customers who retained top-tier membership by Dec. 31$^{\text{st}}$, 2005. Hence, our predictive model with the 4-moment maximum entropy prior performs better when we compare it to the NB model using the gamma prior where the parameters were estimated using the MLE method.

Table 4: Models Fit According to Likelihood of Retaining Top-Tier Frequent Flyer Status

Probability intervals for customers already qualified	Probability of being Gold for customers with gamma prior	Probability of being Gold for customers with 4-moment prior	Actual proportion of customers who retained top-tier membership by Dec. 31$^{\text{st}}$, 2005
[0-10%[	1.63%	**2.89**%	3.31%
[10-20%[	13.17%	**15.87**%	16.40%
[20-30%]	19.89%	**28.03**%	29.61%
[30-40%[	30.48%	**33.79**%	33.33%
[40-50%[	41.03%	**46.93**%	48.51%
[50-60%[	56.82%	**52.02**%	50.00%
[60-70%[	67.36%	**60.51**%	56.00%
[70-80%[	77.34%	**71.29**%	69.39%
[80-90%[	86.29%	**76.47**%	71.71%
[90-100%[	99.05%	**93.33**%	91.76%
[100%]	100%	**100**%	100%

To further assess the predictive performance of our model compared to the NB model using the gamma prior proposed by Fredette and Lawless (2007), we use the data available up to August 1$^{\text{st}}$, 2005 to extrapolate the rate function of our nonhomogeneous Poisson process between August 1$^{\text{st}}$, 2005 and December 31$^{\text{th}}$, 2005. As Figure 3 shows, our model allows rather precise prediction past August 1$^{\text{st}}$. This analysis demonstrates the high degree of validity of using our nonhomogeneous mixed Poisson model for the purposes of forecasting a customer's future purchasing, conditional on his past buying behaviour and his activity to date.

Finally, we present a last example of the usefulness of this approach for various scenarios. We can imagine our customer retention manager is interested in predicting the likelihood of remaining top-tier customers at the beginning of each month. Let us consider the example of two Gold customers who both flew 26 flights over the first year. They both have an 86.2% likelihood of remaining Gold customers at the beginning of the second year. Ultimately, Customer A will fly 23 qualifying flights this year thus conserving his top-tier status, whereas Customer B will fly only 19, meaning he will loose his top-tier status at the end of the year.

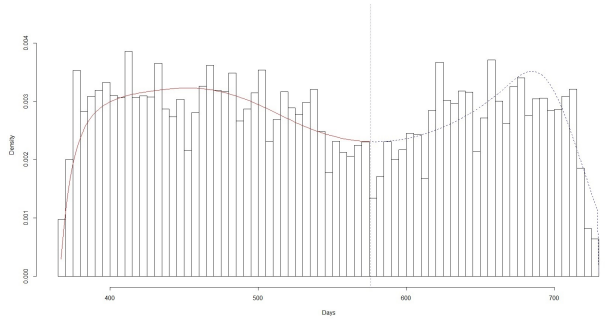

Figure 3: Accuracy of the forecasting based on the data available on August 1^{st} ($t = 578$)

Figures 4 and 5 provide the 12 monthly 95% prediction intervals for Customers A and B. The dotted lines on both graphs indicate the total numbers of flights actually flown by the end of the year while the increasing solid curve represents the total number of flights taken at that point in time. As can be seen, and as an additional demonstration of the predictive ability of our model, the forecasted intervals always contain the actual, final number of flights taken for each of those two customers. Of course, the prediction interval also becomes smaller with time, as data accrue regarding both customers' actual behaviour.

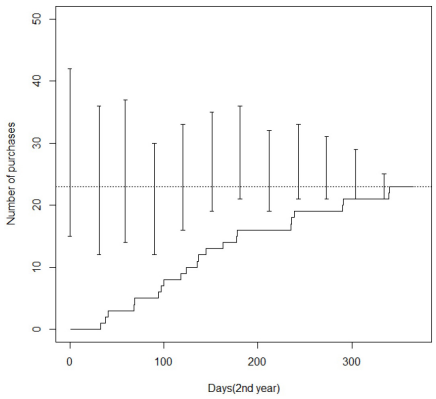

Figure 4: 95% prediction intervals for customers A

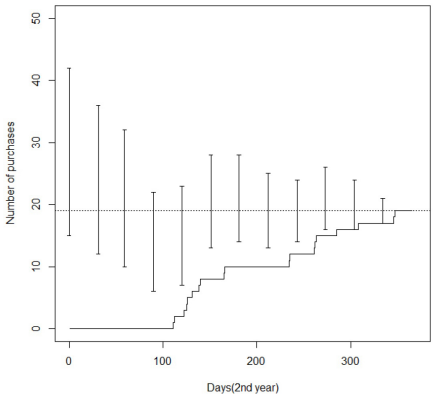

Figure 5: 95% prediction intervals for customer B

On the basis of their respective flying activities, our model allows us to estimate at any point in time the probability that each of these two customers will take at least 20 flights. For instance, a monthly review would provide the probabilities of taking at least 20 flights before the end of the year for each member (see Table 5).

Table 5: Models Fit According to Likelihood of Retaining Top-Tier Frequent Flyer Status.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Customer A (23 flights actually taken at the end of the year)	.862	.698	0.798	0.520	0.819	0.963	0.995	0.975	0.999	0.996	1.000	1.000
Customer B (19 flights actually taken at the end of the year)	.862	.698	.485	.074	.097	.502	.562	.308	.332	.630	.428	.122

As can be seen in Table 5, while the probability of Customer A retaining his top-tier status by the end of the year remains high—above 69.8% except for one month—throughout the year except for one month, Customer B can be identified as potentially losing his top-tier status as early as April. Considering that only 2 flights actually made the difference in the end, the airline company could have used such approaches as

reminding Customer B of the value of his Gold membership as an incentive to fly more in order to retain top-tier benefits into the next year. Adopting such "corrective" actions early on during the year would have likely left enough time for Customer B to better plan his flying activities for the remainder of the year.

According to the results of the different empirical tests applied to our prediction model for the data setting from a major airline company, we can say that the general Poisson-MaxEnt model using 4-moment maximum entropy prior and a spline function for the non-negative function $f(t;\beta)$ with a Weibull survival function $S(t;\beta_1,\beta_2)$ clearly outperformed the NB model proposed by Fredette and Lawless (2007) as a predictive model.

4 Summary and discussion

Retaining top-tier customers holds considerable importance for companies because of the net effect these customers have on any organization's bottom-line. One key assumption in the customer lifetime value associated with top-tier customers is that they are not costly to market to since they do not need marketing—they are already acquired by the firm. However, in real life, even top-tier customers can leave the company for reasons other than service or product failures. For relationship managers, understanding which customers are likely to leave, and identifying which of these the organization still has a chance to retain as top-tier customers can be extremely important. Indeed, managers will want to time their retention efforts, and target them precisely towards these customers they are likely to loose but may still retain provided the right actions are taken.

In this study, we have proposed a predictive model allowing prediction of recurrent events using flexible nonhomogeneous Poisson processes with higher moment maximum entropy priors to model possible heterogeneity amongst the individual units modeled. The motivation for our work lies in the prediction of individual activity level using the data already observed or other events that occur for individual units or subjects in a population. The database at hand was obtained from the loyalty program of a major commercial airline where the behaviour is observed and stored for each customer. The efficiency of our predictive model is compared to the NB model using the conjugate gamma prior proposed by Fredette and Lawless (2007). Also, we have seen throughout this paper that an accurate prediction depends on choosing a satisfactory model for $f(t;\beta))$ in (4) representing the shape of event rate functions for individual units or processes. Using the spline functions for $f(t;\beta))$ also makes our approach especially well suited for situations with irregular purchase behaviour, such as seasonal or cyclical products or services. It has been shown that for the database at hand the use of 4-moment maximum entropy prior provides us a realistic prediction model than the one given by Fredette and Lawless (2007).

Finally, though some detailed development remains to be done, our predictive model considered here can be extended to others situation where there are costs or other values associated with events and we may wish to predict future costs.

References

Borle, Sharad; Singh, Siddharth S.; Jain, Dipak C. (2008), "Customer Lifetime Value Measurement," Management Science, 54(1), 100–112.

Cooray, K. (2006). "Generalization of the Weibull distribution: the odd Weibull family." Statistical Modelling, 6(3), 265–277.

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

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

Harrel, F.E. (2001). Regression Modeling Strategies. New York: Springer-Verlag.

Khribi, L. Fredette, M., and MacGibbon, B. (2016). "The Poisson maximum entropy model for homogeneous Poisson processes." Communications in Statistics - Simulation and Computation, 45(9), 3435–3456.

Khribi, L. Fredette, M., and MacGibbon, B. (2017). "Choosing between higher moment maximum entropy models and its application to homogeneous point processes with random effects." Entropy, 19(12), 687.

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

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

Mohammad-Djafari, A.(1992). "Maximum likelihood estimation of the Lagrange parameters of the maximum entropy distributions." Appeared in Maximum Entropy and Bayesian Methods. Series Fundamental Theories of Physics, 50, 131–139.

Rust, Roland T., Lemon, Katherine N., and Zeithaml, Valarie A. (2004), "Return on Marketing: Using Customer Equity to Focus Marketing Strategy," Journal of Marketing, 68(1), 109–127.

Shannon, C.E.(1948). "The mathematical theory of communication." Bell System Technical Journal, 27, 379–423.

Spiegelhalter, D. Thomas, A. Best, N., and Lunn, D. (2003). WinBUGS User Manual. Version 1.4 (http://www.mrc-bsu.cam.ac.uk/bugs). Technical Report. Medical Research Council Biostatistics Unit. Cambridge.

Talvila, E. (2001). "Necessary and sufficient conditions for differentiating under the integral sign." American Mathematical Monthly, 108, 544–548.

Venkatesan, Rajkumar; Kumar, V; Bohling, Timothy (2007), "Optimal Customer Relationship Management Using Bayesian Decision Theory: An Application for Customer Selection," Journal of Marketing Research, 44(4), 579–594.

Weinstock, R. (1952). Calculus of Variations - With Applications to Physics and Engineering. New York: McGraw-Hill.

Wilks, S. S. (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses." The Annals of Mathematical Statistics, 9(1), 60–62.