Les Cahiers du GERAD ISSN: 0711–2440

Random forests for non-homogeneous Poisson processes with excess zeros

W. Mathlouthi, D. Larocque, M. Fredette

G–2015–77

August 2015

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

Random forests for non-homogeneous Poisson processes with excess zeros

Walid Mathlouthi
Denis Larocque
Marc Fredette

GERAD & Department of Decision Sciences,
HEC Montréal, Montréal (Québec) Canada,
H3T 2A7

denis.larocque@hec.ca
marc.fredette@hec.ca

August 2015

Les Cahiers du GERAD
G–2015–77

Abstract: We propose a method to build trees and forests when the response is a non-homogeneous Poisson process with excess zeros, based on two forests. The first one is used to estimate the probability of having a zero. The second forest is used to estimate the Poisson parameter using trees built with a splitting criterion derived from the zero truncated non-homogeneous Poisson likelihood. Simulation studies show that the proposed method performs well in hurdle (zero-altered) and zero-inflated settings.

Key Words: Hurdle model, zero-altered Poisson (ZAP), zero-inflated Poisson (ZIP), non-homogeneous Poisson process, tree-based method, random forests.

Acknowledgments: The authors acknowledge the financial support of Natural Sciences and Engineering Research Council of Canada (NSERC) and of Fonds de recherche du Québec – Nature et technologies (FRQNT).

1 Introduction

It is often of interest to model the number of times that an event occurs during a given period of time. Such count data, taking only non-negative integer values are encountered in many situations. Some examples are the number of visits to the emergency unit in a hospital, the number or purchases made by a client, and the number of accidents on a road. Many parametric models are available to study the impact of covariates on the count response, including the Poisson and negative binomial models; see Cook and Lawless (2007) and Hilbe (2011). Sometimes, the number of zeros in the response is large and cannot be adequately accounted by the usual models and the available covariates. Hurdle models, also called zero-altered models (Mullahy, 1986), and zero-inflated models (Lambert, 1992) were designed to handle such situations.

In this paper, we are interested in tree-based methods (Breiman et al., 1984) and more particularly in random forests (Breiman, 2001). The main advantage of these methods is their flexibility, meaning they can adapt to the data at hand without having to specify a parametric form. Within the CART (Classification and Regression Tree) paradigm, Poisson regression trees can be fitted in R (R Core Team, 2015) with the package rpart (Therneau et al., 2014). The GUIDE approach (Loh, 2002) provides another way to build Poisson trees using the splitting rule of Chaudhuri et al. (1995). But these methods are not aimed at handling excess zeros in the response. Lee and Jin (2006) proposed a tree-based method for the excess zero case. They use the zero-inflated Poisson distribution to derive a splitting criterion. However, all the methods discussed above work under the basic assumption that the Poisson process generating these count data is homogeneous with respect to time. That is, the rate function of the response, given the covariates, does not vary with time. Mathlouthi, Fredette and Larocque (2015) developed tree-based methods for a response from a non–homogeneous Poisson process, but their method is not aimed at the excess zero case. In this paper, we present a method that extends both the Lee and Jin (2006) and Mathlouthi, Fredette and Larocque (2015) methods. We propose a method to build trees and forests for a non-homogeneous Poisson response with excess zeros. Hence, the method can handle simultaneously a non-homogeneous rate function and excess zeros in the response.

The paper is organized as follows. Section 2 describes the usual parametric models for a Poisson response with excess zeros. Section 3 describes the basic tree and forest methodology and the method of Lee and Jin (2006). Section 4 presents the proposed methods. The results from a simulation study are presented in Section 5. Concluding remarks and possibilities for future work are given in Section 6.

2 Zero-altered Poisson (ZAP) and zero-inflated Poisson (ZIP) regression model

We are interesting in modeling a count response Y with a set of q covariates $\mathbf{X} = (X_1, \ldots, X_q)'$. In Poisson hurdle, or zero-altered Poisson (ZAP), and zero-inflated Poisson (ZIP) models, it is assumed that Y follows a Poisson distribution modified to account for an excess number of zeros. We will denote by λ, the parameter of the Poisson distribution, by p, the probability of having an excess (with respect to the Poisson distribution) 0, and by θ, the total probability of having a 0. These parameters can depend on the covariates and it will be clear form the context whether we are talking about the generic parameters or the covariate dependent parameters. We note that the Poisson variable used in this section could also represent the total number of events obtained from a Poisson process. The parameter λ then represents the integral of the corresponding rate function over a given time interval.

In the ZAP regression model, it is assumed that $Y = 0$ with probability θ ($0 \leq \theta < 1$), and that Y follows a zero truncated Poisson distribution with parameter λ, ($\lambda > 0$), given that $Y > 0$. Consequently,

$$P(Y = y) = \begin{cases} \theta & y = 0 \\ \frac{(1-\theta)\exp(-\lambda)\lambda^y}{(1-\exp(-\lambda))y!} & y = 1, 2, \ldots \end{cases} \tag{1}$$

Many possibilities are available to link the covariates to the response, through p and θ. The common link functions are given by:

$$\log(\lambda) = \mathbf{X}'\beta \quad \text{and} \quad \log\left(\frac{\theta}{1-\theta}\right) = \mathbf{X}'\gamma, \tag{2}$$

where β, γ are vectors of parameters to be estimated.

In the ZIP regression model, it is assumed that $Y = 0$ with probability p, $(0 \leq p < 1)$ and that Y follows a Poisson distribution with parameter λ, $(\lambda > 0)$, with probability $1 - p$. Hence, there are two sources for the zeros. Consequently,

$$P(Y = y) = \begin{cases} p + (1-p)\exp(-\lambda) & y = 0 \\ (1-p)\exp(-\lambda)\lambda^y/y! & y = 1, 2, \ldots \end{cases} \tag{3}$$

Again, many possibilities are available to link the covariates to the response. The common link functions are the same as for the ZAP model:

$$\log(\lambda) = \mathbf{X}'\beta \quad \text{and} \quad \log\left(\frac{p}{1-p}\right) = \mathbf{X}'\gamma, \tag{4}$$

where β, γ are vectors of parameters to be estimated. Note that different covariates can be used for the (excess) zero part and the Poisson part of these models, but we use this notation for simplicity.

Note that, without covariates, models (1) and (3) are just two parameterizations of the same model. The difference between the two approaches lies in the way the covariates are linked to the parameters. In the ZAP model, the covariates are linked directly to the total probability of having a 0 while, for the ZIP model, they are linked directly to the probability of having an excess 0 with respect to the Poisson distribution.

One key observation is that the ZAP model is formed by two models. Consequently, the covariates can have different effects on the zero and the Poisson parts. The same is true for the ZIP model, except that this time, the covariates can have different effects on the excess zero and the Poisson parts.

Assume that a sample of size n is available. Namely, $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, where $\mathbf{X}_i = (X_{1i}, \ldots, X_{qi})'$. Then the parameters of the ZAP and ZIP regression models can be estimated with maximum likelihood. It is straightforward to do it with the ZAP model because the likelihood for the zero and the zero truncated Poisson parts can be factored. Hence, the two models can be fitted separately. Things are a slightly more complex with the ZIP model but Lambert (1992) proposed an estimation method based on the EM algorithm.

3 Trees and forests

3.1 Basic tree and forest methodology

We assume the reader is familiar with the CART paradigm (Breiman et al., 1984) as only a brief description is given here. Tree-based methods partition the covariates space by splitting it recursively with rules based on covariates. The basic ingredient for building a tree is the splitting criterion, which is problem dependent. For example, the least-squares splitting criterion is the usual one when the response is continuous. Suppose we are at a given node t and we want to split it into two children nodes, t_L (left node) and t_R (right node). The best split is chosen among all possible binary splits obtained from a covariate. If x is continuous (or at least ordinal), the possible splits take the form $I(x \leq c)$. If x is categorical, the possible splits take the form $x \in \{c_1, \ldots, c_s\}$ where $\{c_1, \ldots, c_s\}$ is a subset of the possible values of x. The best split is the one maximizing the splitting criterion. If a single tree is required, than the usual procedure builds a large tree and then uses a pruning algorithm to avoid over-fitting. However, it is now well-established that using an ensemble of trees is generally preferable to a single tree. One the most popular ensemble method is random forests, introduced by Breiman (2001). Here we describe the generic random forest algorithm that will be used in this paper:

1. Draw B bootstrap samples from the original data.
2. For each bootstrap sample, grow a tree with the selected splitting criterion. At each node, randomly select q_0 out of q covariates where $q_0 \leq q$ and is a user-specified parameter. Splitting ends when a stopping criterion is reached; for instance, when a node has less than a predetermined number of observations. No pruning is performed.
3. To obtain an estimation of a parameter (or a prediction) for a new observation, take the average estimates (predictions) from the B trees.

3.2 Maximum likelihood splitting criterion

A simple method for deriving a splitting criterion is to use the log-likelihood of an adequate two-node model; see Su, Wang and Fan (2004) and Bou-Hamad et al. (2009) for some examples. Basically, the best split at a given node is the one that maximizes the observed log-likelihood, i.e. the one evaluated at the maximum likelihood estimates, among all allowable splits. Moreover, if the parameters are estimated separately in the two children nodes, then the best split is the one that maximizes

$$\widehat{LL}(\text{left node}) + \widehat{LL}(\text{right node}), \tag{5}$$

where $\widehat{LL}(\text{left node})$ and $\widehat{LL}(\text{right node})$ are the observed log-likelihood in the left and right nodes, respectively.

3.3 ZIP tree of Lee and Jin (2006)

Lee and Jin (2006) proposed a decision tree method for zero-inflated count data based on the CART paradigm. They call it a ZIP tree. They basically fit the ZIP distribution (3) separately in the two children nodes, and use (5) as the splitting criterion. Let N^+ denote the number of observations such that $Y_i > 0$. The log-likelihood function in that case is

$$\begin{aligned} LL_{ZIP} = (n - N^+)\log(p + (1-p)\exp(-\lambda)) + N^+(\log(1-p) - \lambda) \\ + \sum_{Y_i>0} Y_i \log(\lambda) - \sum_{Y_i>0} \log(Y_i!). \end{aligned} \tag{6}$$

One crucial observation is that the covariates are used to find the tree structure for both the excess zero part and the Poisson part jointly. Hence a single model is used to model both parts and a single tree is built. But as we just saw, the ZIP regression model (3) uses two models, one for the excess zero part and one for the Poisson part. Hence the covariates can have different effects on both parts. This is why we propose in this paper a new approach, more in the spirit of the ZIP and ZAP models, where two models are used.

4 Random forests for Poisson data with excess zeros

In the ZAP model, assume that, instead of a rigid parametric model like (2), we use a general nonparametric model given by

$$\theta = f_\theta(\mathbf{X}) \quad \text{and} \quad \lambda = f_\lambda(\mathbf{X}), \tag{7}$$

where f_θ and f_λ are general unknown link functions. Similarly, assume a same general setup for the ZIP model, given by

$$p = g_p(\mathbf{X}) \quad \text{and} \quad \lambda = g_\lambda(\mathbf{X}), \tag{8}$$

where g_p and g_λ are general unknown link functions. Since $\theta = g_p(\mathbf{X}) + (1 - g_p(\mathbf{X}))\exp(-g_\lambda(\mathbf{X}))$, we see that in this general nonparametric framework, the ZIP and ZAP models are again only different parameterizations of the same model. Indeed, we can just define $f_\theta(\mathbf{X}) = g_p(\mathbf{X}) + (1 - g_p(\mathbf{X}))\exp(-g_\lambda(\mathbf{X}))$ in (7). Hence, it does not matter whether we specify model (7) or (8). Namely, if a general nonparametric and flexible procedure is used to estimate f_θ and f_λ in (7), it can be used to obtain estimates

$$\hat{\theta} = \hat{f}_\theta(\mathbf{X}) \quad \text{and} \quad \hat{\lambda} = \hat{f}_\lambda(\mathbf{X}), \tag{9}$$

for a given value of $\mathbf{X}$. But it can also be used to estimate p through the equation $\hat{\theta} = \hat{p} + (1 - \hat{p})\exp(-\hat{\lambda})$. Solving for $\hat{p}$ gives $\hat{p} = (\hat{\theta} - \exp(-\hat{\lambda}))/(1 - \exp(-\hat{\lambda}))$. However, since this value can be less than 0, we will use $\hat{p} = \max(0, (\hat{\theta} - \exp(-\hat{\lambda}))/(1 - \exp(-\hat{\lambda})))$, in this paper. The key point is that the method proposed in this paper is valid for both the ZAP and the ZIP settings.

4.1 Description of the basic method

The basic idea is to fit two random forests, one for the zero part to estimate f_θ, and one for the observations that are greater than 0 to estimate f_λ. More specifically, for the zero part, the response is $I(Y > 0)$, that is the binary variable taking a value of 1 if $Y > 0$ and the value 0 if $Y = 0$. This is a standard problem and many implementations are available in R, for example through the packages `randomForest` (Liaw and Wiener, 2002) and `randomForestSRC` (Ishwaran, Kogalur, Blackstone and Lauer 2008, Ishwaran and Kogalur, 2015). For the the observations that are greater than 0, we propose a forest of trees built using a splitting criterion derived from the zero truncated Poisson likelihood. Only the observations where $Y > 0$ are used. Assume there are N^+ such observations denoted by $Y_1^+, \ldots, Y_{N^+}^+$. The probability mass function from the truncated Poisson distribution is

$$P(Y^+ = y) = P(Y = y|Y > 0) = \frac{\exp(-\lambda)\lambda^y}{y!(1 - \exp(-\lambda))} \quad y = 1, 2, \ldots \tag{10}$$

Hence, the log-likelihood function for the sample is

$$LL^+ = -N^+ \log(1 - \exp(-\lambda)) + \log(\lambda) \sum_{i=1}^{N^+} Y_i^+ - N^+\lambda - \sum_{i=1}^{N^+} \log(Y_i^+!). \tag{11}$$

The estimated λ is obtained by solving $\partial LL^+/\partial\lambda = 0$ which reduces to

$$\frac{\sum_{i=1}^{N^+} Y_i^+}{N^+} = \frac{\lambda}{1 - \exp(-\lambda)}. \tag{12}$$

For a given candidate split, the zero truncated Poisson model is fitted separately in the two children nodes and the splitting criterion is given by (5) with (11) as the log-likelihood function.

4.2 Extension to the non-homogeneous case

Mathlouthi, Fredette and Larocque (2015) proposed tree and random forest methods for non-homogeneous Poisson processes. It was achieved by considering a model with a piecewise constant rate function. Here we extend this method to the case of a non-homogeneous Poisson process with excess zeros. We are again interested in a count response but this time we want to allow the rate function to vary over time. Assume we have a fixed time period T and assume that it is partitioned into K subperiods, $T_1, \ldots, T_K$, such that

$$\bigcup_{k=1}^{K} T_k = T \quad \text{and} \quad T_i \cap T_j = \emptyset \quad \text{for all} \quad i \neq j.$$

Each subperiod T_k may be an interval or a finite union of disjoint intervals. Denote by N_k the number of events in subperiod T_k. We assume that N_k follows a Poisson distribution with parameter λ_k and that all the N_k's are independent. The K subperiods can be adjacent time intervals $(a_0, a_1], (a_1, a_2], \ldots, (a_{K-1}, a_K]$ covering the whole period $T = (a_0, a_K]$. But the above formulation is more general than that and the subperiods can represent non–adjacent periods. For example, they could represent days (T_1=all the Mondays in a year, T_2=all the Tuesdays in a year and so on). The total number of events observed over T is denoted by Y, that is

$$Y = \sum_{k=1}^{K} N_k.$$

We are interested in allowing Y to have excess zeros with respect to the non-homogeneous Poisson process described above. Once again, the idea is to fit two random forests, one to estimate $P(Y = 0)$, that is the

probability that no events at all occurred, and one for the observations with at least one event. For the zero part, the binary response is $I(Y > 0)$, which can be fitted by standard algorithm as described in the preceding section. For the the observations with at least one event, we propose a forest of trees built using a splitting criterion derived from a zero truncated non-homogeneous Poisson model likelihood, as described next. Let $(n_1, \ldots, n_K)$ be non-negative integers such that at least one of them is greater than 0. The joint probability mass function of $(N_1, \ldots, N_K)$ given that at least one event occurred is

$$P(N_1 = n_1, \ldots, N_K = n_k | Y > 0) = \frac{1}{(1 - \exp(-\lambda))} \prod_{k=1}^{K} \exp(-\lambda_k)\lambda_k^{n_k}/n_k! \tag{13}$$

where $\lambda = \sum_{k=1}^{K} \lambda_k$.

For a sample, $(N_{i1}, \ldots, N_{iK})$, $i = 1, \ldots, N$, assume we have N^+ observations such that $Y_i = \sum_{k=1}^{K} N_{ik} > 0$, then the log-likelihood function for those observations is

$$LL_{NH}^{+} = -N^+ \log(1 - \exp(-\lambda)) + \sum_{k=1}^{K} \log(\lambda_k) \sum_{i=1}^{N^+} N_{ik} - N^+\lambda - \sum_{k=1}^{K}\sum_{i=1}^{N^+} \log(N_{ik}!). \tag{14}$$

Note that if we have a single time period, i.e., $K = 1$, then $Y_i = N_{i1}$ and we fall back to the setting of Section 4.1. The estimated λ_k's are obtained by solving the K equations $\partial LL_{NH}^{+}/\partial\lambda_k = 0$, giving

$$\sum_{i=1}^{N^+} \frac{N_{ik}}{N^+} = \frac{\lambda_k}{1 - \exp\left(-\sum_{j=1}^{K} \lambda_j\right)}, k = 1, \ldots, K.$$

This system can be solved with the Newton-Raphson algorithm. For a given candidate split, the zero truncated non-homogeneous Poisson model is fitted separately in the two children nodes and the splitting criterion is given by (5) with (14) as the log-likelihood function.

5 Simulation study

In this section, we investigate the performance of the proposed method compared to various competitors including parametric models and forests. In the first set of simulations, the data generating process (DGP) is homogeneous. This will allow comparisons with the usual parametric ZIP and ZAP models, with a forest of Poisson trees but also with a forest built with the Lee and Jin (2006) ZIP tree approach. Then, in the second set of simulations, non-homogeneous DGPs will be considered. This will allow a comparison with the approach of Mathlouthi et al. (2015).

5.1 Description of the simulation study

In all cases, nine covariates $X_1, ..., X_9$, are available. They are independent and uniformly distributed over the interval $[0, 10]$. X_1, X_2, and X_3 are related to the Poisson intensity parameter, while X_1, X_4, and X_5 are related to zero part of the model. Hence X_6, X_7, X_8 and X_9 are noise covariates unrelated to the outcome.

DGPs from ZAP and ZIP settings with varying proportions of zeros are considered. Consider first the following logistic regression DGP that will be used to generate either the zeros (ZAP) or excess zeros (ZIP):

DGP zero

$$\log\left(\frac{1-\tau}{\tau}\right) = c - 3\log(X_1 + 0.5) + 0.2(10X_4 - X_4^2) + 0.4X_5.$$

The choices of intercepts $c = 2.6$, $c = 0.55$ and $c = -1.1$ produce approximately 15%, 35% and 55% of zeros for this DGP. The parameter τ represents either θ, the total probability of having a 0, for a ZAP DGP or p, the probability of having an excess 0, for a ZIP DGP.

Consider the following three models for the Poisson part of the outcome, governed by the parameter λ.

DGP A: Poisson model with main effects only

$$\ln(\lambda) = -0.105 + 0.1X_1 - 0.1X_2 + 0.1X_3.$$

DGP B: Poisson model with more complicated effects

$$\ln(\lambda) = -0.7 + 0.05X_1 + 2(X_2 > 5) + 0.05(10X_3 - X_3^2) + 0.04(X_2 > 5)(X_1 - 5)^2.$$

DGP C: Poisson model with a tree structure

Leaf 1. If $X_1 \leq 5$ and $X_2 \leq 5$ then $\lambda = 1.5$.
Leaf 2. If $X_1 \leq 5$ and $X_2 > 5$ then $\lambda = 3.0$.
Leaf 3. If $X_1 > 5$ and $X_3 \leq 5$ then $\lambda = 2.5$.
Leaf 4. If $X_1 > 5$ and $X_3 > 5$ then $\lambda = 2.0$.

Nine scenarios are considered for the ZAP DGPs, by crossing the three Poisson DGPs with the three different probabilities of zero. For these scenarios, a binary outcome is first generated from DGP zero. If a 0 is generated, then this is the value of Y. If not, then Y is generated by using the Poisson model (either A, B or C) but truncated at zero. Indeed, for a ZAP DGP, the total probability of having a 0 is governed only by DGP zero. Hence, for these scenarios, the total proportions of zeros are going to be approximately 15%, 35% or 55%.

Nine scenarios are also considered for the ZIP DGPs, again by crossing the three Poisson DGPs with the three different probabilities of zero. For these scenarios, a binary outcome is first generated from DGP zero. If a 0 is generated, then this is the value of Y. If not then Y is generated by using the Poisson model (either A, B or C). This time, a 0 can also be generated from the Poisson part. The binary outcome model only generates excess zeros with respect to the Poisson model. Hence, for these scenarios, the total proportions of zeros are going to be higher than 15%, 35% or 55%.

Nine other scenarios are considered for the non-homogeneous ZAP case. Using the notation of Section 4.2, we have $K = 12$ subperiods. The probability that no event at all occurred is still given by DGP zero above. The three following non-homogeneous Poisson DPGs are used for the Poisson part. This time, 12 parameters, $\mathbf{\Lambda} = (\lambda_1, \ldots, \lambda_{12})$ are required.

DGP D: Non-Homogeneous Poisson model with main effects only

$$\ln(\lambda_k) = \log(0.1 * k) - 0.3 + 0.1X_1 - 0.1X_2 + 0.1X_3, k = 1, \ldots, 12.$$

DGP E: Non-Homogeneous Poisson model with more complicated effects

$$\ln(\lambda_k) = -0.5 * k + 0.05X_1 + 2(X_2 > 5) + 0.05(10X_3 - X_3^2) + 0.04(X_2 > 5)(X_1 - 5)^2, k = 1, \ldots, 12.$$

DGP F: Non-Homogeneous Poisson model with a tree structure

Leaf 1. If $X_1 \leq 5$ and $X_2 \leq 5$ then $\mathbf{\Lambda} = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2)$.
Leaf 2. If $X_1 \leq 5$ and $X_2 > 5$ then $\mathbf{\Lambda} = (1.2, 1.2, 1.2, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 1.2, 1.2, 1.2)$.
Leaf 3. If $X_1 > 5$ and $X_3 \leq 5$ then $\mathbf{\Lambda} = (0.1, 0.1, 0.1, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 0.1, 0.1, 0.1)$.
Leaf 4. If $X_1 > 5$ and $X_3 > 5$ then $\mathbf{\Lambda} = (1.2, 1.1, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$.

For the homogeneous DGP's, five models are compared. They are

1. Parametric ZAP model described in (1) and (2). The nine covariates are used as main effects only both for the zero and Poisson parts of the model.

2. Parametric ZIP model described in (3) and (4). The nine covariates are used as main effects only both for the excess zero and Poisson parts of the model.
3. Poisson forest. A forest of basic Poisson trees is built.
4. Lee-Jin forest. A forest with the ZIP tree approach of Lee and Jin (2006).
5. Proposed approach, the ZAP forest (Section 4.1).

For the non-homogeneous DGP's, two models are compared. They are the NHPPF method of Mathlouthi et al. (2015), and the proposed non-homogeneous ZAP forest (Section 4.2).

The parametric models were fitted with the R package `pcsl` (Jackman, 2015, and Zeileis, Kleiber and Jackman, 2008). The Poisson parts of the forests were implemented in C. However, the R package `randomForest` was used to build the forest for the zero part of the proposed ZAP forest. All forest are built with 500 trees. Three out of the nine covariates are randomly selected at each node of each tree. This comes from the value $\sqrt{q}$ typically used to build a regression random forest. Thirty observations are needed to attempt splitting and the resulting nodes must have at least ten observations. The models are estimated with a training sample of size 1000. Parameter estimates are then obtained for each observation in a test sample of size 5000. The number of simulation runs is 100.

The mean absolute error (MAE) is used as the performance criterion. It is defined by

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T} |\gamma_t - \hat{\gamma}_t|,$$

where T is the size of the test set (5000 here), γ_t and $\hat{\gamma}_t$ are the true and estimated values of the parameter of interest for the t^{th} test observation. Here γ represents one of the three parameters of interests which are λ, θ and p.

5.2 Results

The results are presented in Tables 1 to 3. Table 1 presents the average MAE, over the 100 simulation runs, of all methods for the three parameters of interests, λ, θ, p, for the nine ZAP DGPs. For each line in the table, the value of the best (smallest) MAE is in bold. The values between parentheses are the standard deviations of the MAE over the simulation runs. It is striking that the proposed method has the smallest MAE in all but three of the 27 cases. For DGP A, the Poisson part is a main effect DGP. Hence, the Poisson part of the parametric ZAP model contains the true effects. It is then not surprising that it has the smallest MAE for estimating λ whatever the proportion of zeros is. However, the zero part of the parametric ZAP model is not well-specified and thus the proposed method is better to estimate both θ and p. The Lee-Jin forest is better than a Poisson forest for estimating λ and p but the opposite is true for estimating θ.

Table 2 presents the MAE for the nine ZIP DGPs. This time, the proposed method has the smallest MAE in 20 cases and is close to the best one in four other cases. Similarly to what we saw with the ZAP DGPs, the parametric model is the best one when the poisson part of the model contains main effects only, that is for DGP A. The parametric ZIP model is then the best one for estimating λ in these three cases (with either 5%, 35% and 55% of excess zeros). There are four instances where the Lee-Jin forest is slightly better than the proposed method. But apart from that, the proposed method is globally the best one for the DGPs considered.

Table 3 presents the results for the non-homogeneous ZAP DGPs. This time only the proposed method and the NHPPF method of Mathlouthi et al. (2015) are compared. The NHPPF fits a forest of non-homogeneous Poisson trees but does not account for excess zeros. The proposed method has the smallest MAE in all 27 cases. This, combined with the fact that it is also always better than a Poisson forest in Tables 1 and 2, clearly shows the importance of modeling the potential excess zeros.

Table 1: Simulation results for the homogeneous ZAP DGPs. The average MAE are reported for the three parameters of interest: λ, the Poisson intensity; θ, the probability of zero; p, the probability of an excess zero. Standard deviations of the MAE are reported between parentheses. The smallest value of the MAE for a given scenario is in bold. In the first column, the percentage corresponds to the total probability of having a 0.

DGP	Parameter	Parametric Zap	Parametric Zip	Poisson forest	Lee-Jin forest	Zap Forest
A(15%)	λ	**0.1304** (0.0295)	0.3984 (0.0365)	0.5222 (0.0261)	0.3989 (0.0315)	0.2703 (0.0244)
	θ	0.1190 (0.0040)	0.1787 (0.0143)	0.1714 (0.0044)	0.2136 (0.0072)	**0.0719** (0.0045)
	p	0.0889 (0.0032)	0.0799 (0.0117)	0.0991 (0.0027)	0.0817 (0.0030)	**0.0496** (0.0039)
A(35%)	λ	**0.1651** (0.0402)	0.3470 (0.0661)	0.7264 (0.0295)	0.4365 (0.0429)	0.3260 (0.0368)
	θ	0.1741 (0.0032)	0.2163 (0.0084)	0.1994 (0.0078)	0.2303 (0.0066)	**0.1016** (0.0055)
	p	0.1575 (0.0034)	0.1672 (0.0040)	0.2738 (0.0048)	0.1820 (0.0076)	**0.0927** (0.0064)
A(55%)	λ	**0.2098** (0.0506)	0.3260 (0.0488)	1.0019 (0.0313)	0.6341 (0.0584)	0.4397 (0.0565)
	θ	0.1656 (0.0028)	0.1916 (0.0036)	0.1863 (0.0099)	0.2489 (0.0097)	**0.1049** (0.0050)
	p	0.1653 (0.0030)	0.1772 (0.0037)	0.4689 (0.0054)	0.3285 (0.0134)	**0.1165** (0.0081)
B(15%)	λ	0.4748 (0.0167)	0.5919 (0.0220)	0.4672 (0.0261)	0.2994 (0.0343)	**0.2857** (0.0249)
	θ	0.1190 (0.0040)	0.1447 (0.0097)	0.1551 (0.0051)	0.1992 (0.0081)	**0.0718** (0.0046)
	p	0.0852 (0.0033)	0.0671 (0.0072)	0.0906 (0.0026)	0.0747 (0.0027)	**0.0490** (0.0038)
B(35%)	λ	0.4868 (0.0232)	0.5639 (0.0491)	0.6225 (0.0284)	0.3615 (0.0348)	**0.3243** (0.0315)
	θ	0.1741 (0.0032)	0.1848 (0.0181)	0.1729 (0.0078)	0.2043 (0.0067)	**0.1015** (0.0055)
	p	0.1678 (0.0039)	0.1716 (0.0063)	0.2649 (0.0047)	0.1807 (0.0074)	**0.0979** (0.0065)
B(55%)	λ	0.5095 (0.0292)	0.5374 (0.0414)	0.8870 (0.0302)	0.5356 (0.0496)	**0.4008** (0.0420)
	θ	0.1656 (0.0028)	0.1754 (0.0038)	0.1659 (0.0094)	0.2254 (0.0089)	**0.1050** (0.0050)
	p	0.1813 (0.0035)	0.1885 (0.0089)	0.4682 (0.0054)	0.3314 (0.0116)	**0.1233** (0.0084)
C(15%)	λ	0.4748 (0.0167)	0.5919 (0.0220)	0.4672 (0.0261)	0.2994 (0.0343)	**0.2857** (0.0249)
	θ	0.1190 (0.0040)	0.1447 (0.0097)	0.1551 (0.0051)	0.1992 (0.0081)	**0.0718** (0.0046)
	p	0.0852 (0.0033)	0.0671 (0.0072)	0.0906 (0.0026)	0.0747 (0.0027)	**0.0490** (0.0038)
C(35%)	λ	0.4868 (0.0232)	0.5639 (0.0491)	0.6225 (0.0284)	0.3615 (0.0348)	**0.3243** (0.0315)
	θ	0.1741 (0.0032)	0.1848 (0.0181)	0.1729 (0.0078)	0.2043 (0.0067)	**0.1015** (0.0055)
	p	0.1678 (0.0039)	0.1716 (0.0063)	0.2649 (0.0047)	0.1807 (0.0074)	**0.0979** (0.0065)
C(55%)	λ	0.5095 (0.0292)	0.5374 (0.0414)	0.8870 (0.0302)	0.5356 (0.0496)	**0.4008** (0.0420)
	θ	0.1656 (0.0028)	0.1754 (0.0038)	0.1659 (0.0094)	0.2254 (0.0089)	**0.1050** (0.0050)
	p	0.1813 (0.0035)	0.1885 (0.0089)	0.4682 (0.0054)	0.3314 (0.0116)	**0.1233** (0.0084)

Table 2: Simulation results for the homogeneous ZIP DGPs. The average MAE are reported for the three parameters of interest: λ, the Poisson intensity; θ, the probability of zero; p, the probability of an excess zero. Standard deviations of the MAE are reported between parentheses. The smallest value of the MAE for a given scenario is in bold. In the first column, the percentage corresponds to the probability of having an excess 0.

DGP	Parameter	Parametric Zap	Parametric Zip	Poisson forest	Lee-Jin forest	Zap Forest
A(15%)	λ	0.1469 (0.0348)	**0.1467** (0.0480)	0.4248 (0.0326)	0.3067 (0.0234)	0.2857 (0.0263)
	θ	0.1237 (0.0047)	0.1022 (0.0075)	0.1044 (0.0057)	**0.1014** (0.0053)	0.1035 (0.0052)
	p	0.1401 (0.0110)	0.1237 (0.0107)	0.1520 (0.0029)	0.1225 (0.0086)	**0.1212** (0.0110)
A(35%)	λ	0.1817 (0.0480)	**0.1783** (0.0460)	0.7471 (0.0330)	0.4152 (0.0433)	0.3474 (0.0419)
	θ	0.1565 (0.0028)	0.1398 (0.0044)	0.1470 (0.0083)	0.1511 (0.0053)	**0.1082** (0.0059)
	p	0.1881 (0.0072)	0.1799 (0.0082)	0.3515 (0.0045)	0.2184 (0.0094)	**0.1437** (0.0101)
A(55%)	λ	0.2385 (0.0594)	**0.2384** (0.0578)	1.0627 (0.0336)	0.6572 (0.0614)	0.4721 (0.0614)
	θ	0.1353 (0.0024)	0.1268 (0.0035)	0.1427 (0.0108)	0.1947 (0.0107)	**0.0992** (0.0056)
	p	0.1729 (0.0054)	0.1710 (0.0064)	0.5493 (0.0048)	0.3715 (0.0141)	**0.1485** (0.0102)
B(15%)	λ	0.4999 (0.0296)	0.5015 (0.0277)	0.3880 (0.0270)	**0.3002** (0.0236)	0.3062 (0.0252)
	θ	0.1259 (0.0040)	0.1184 (0.0057)	0.0992 (0.0056)	0.0982 (0.0060)	**0.0976** (0.0053)
	p	0.1494 (0.0134)	0.1397 (0.0166)	0.1520 (0.0029)	0.1188 (0.0088)	**0.1153** (0.0106)
B(35%)	λ	0.5109 (0.0328)	0.5127 (0.0325)	0.6636 (0.0325)	0.3833 (0.0357)	**0.3446** (0.0314)
	θ	0.1544 (0.0026)	0.1472 (0.0031)	0.1363 (0.0092)	0.1458 (0.0055)	**0.1059** (0.0058)
	p	0.1918 (0.0083)	0.1875 (0.0088)	0.3515 (0.0045)	0.2194 (0.0107)	**0.1406** (0.0093)
B(55%)	λ	0.5310 (0.0332)	0.5368 (0.0352)	0.9655 (0.0334)	0.5822 (0.0514)	**0.4230** (0.0430)
	θ	0.1386 (0.0021)	0.1341 (0.0022)	0.1385 (0.0104)	0.1927 (0.0104)	**0.1011** (0.0055)
	p	0.1750 (0.0051)	0.1731 (0.0052)	0.5493 (0.0048)	0.3741 (0.0145)	**0.1454** (0.0115)
C(15%)	λ	0.4124 (0.0124)	0.4158 (0.0183)	0.4583 (0.0322)	**0.2988** (0.0271)	0.3210 (0.0265)
	θ	0.1303 (0.0040)	0.1221 (0.0063)	0.1049 (0.0054)	**0.0941** (0.0053)	0.0956 (0.0057)
	p	0.1339 (0.0082)	0.1253 (0.0092)	0.1520 (0.0029)	0.1033 (0.0068)	**0.0983** (0.0078)
C(35%)	λ	0.4274 (0.0185)	0.4305 (0.0192)	0.8384 (0.0368)	0.4070 (0.0415)	**0.3664** (0.0303)
	θ	0.1660 (0.0026)	0.1596 (0.0030)	0.1630 (0.0103)	0.1535 (0.0051)	**0.1082** (0.0061)
	p	0.1825 (0.0055)	0.1797 (0.0055)	0.3515 (0.0045)	0.1936 (0.0089)	**0.1208** (0.0078)
C(55%)	λ	0.4519 (0.0295)	0.4580 (0.0363)	1.2576 (0.0378)	0.6704 (0.0591)	**0.4479** (0.0444)
	θ	0.1504 (0.0024)	0.1475 (0.0027)	0.1766 (0.0121)	0.2165 (0.0078)	**0.1041** (0.0054)
	p	0.1695 (0.0040)	0.1692 (0.0040)	0.5493 (0.0048)	0.3363 (0.0136)	**0.1214** (0.0067)

Table 3: Simulation results for the non-homogeneous ZAP DGPs. The average MAE are reported for the parameters of interest: $\mathbf{\Lambda} = (\lambda_1, \ldots, \lambda_{12})$, the Poisson intensities; θ, the probability of zero; p, the probability of an excess zero. Standard deviations of the MAE are reported between parentheses. The smallest value of the MAE for a given scenario is in bold. Note that in the case of $\mathbf{\Lambda}$, the average MAE over the 12 parameters are reported. In the first column, the percentage corresponds to the total probability of having a 0.

DGP	Parameter	NHPPF	Non-homogeneous Zap Forest
D(15%)	$\mathbf{\Lambda}$	0.2216 (0.0118)	**0.1334** (0.0055)
	θ	0.1511 (0.0037)	**0.0721** (0.0040)
	p	0.1515 (0.0038)	**0.0721** (0.0040)
D(35%)	$\mathbf{\Lambda}$	0.3962 (0.0135)	**0.1462** (0.0073)
	θ	0.3334 (0.0055)	**0.1012** (0.0056)
	p	0.3499 (0.0050)	**0.1016** (0.0056)
D(55%)	$\mathbf{\Lambda}$	0.5656 (0.0143)	**0.1724** (0.0105)
	θ	0.4534 (0.0126)	**0.1053** (0.0054)
	p	0.5476 (0.0057)	**0.1059** (0.0054)
E(15%)	$\mathbf{\Lambda}$	0.2274 (0.0132)	**0.1481** (0.0062)
	θ	0.1517 (0.0038)	**0.0721** (0.0041)
	p	0.1524 (0.0038)	**0.0721** (0.0041)
E(35%)	$\mathbf{\Lambda}$	0.4131 (0.0141)	**0.1627** (0.0063)
	θ	0.3363 (0.0055)	**0.1011** (0.0056)
	p	0.3512 (0.0051)	**0.1011** (0.0056)
E(55%)	$\mathbf{\Lambda}$	0.6093 (0.0143)	**0.1885** (0.0096)
	θ	0.4682 (0.0114)	**0.1053** (0.0055)
	p	0.5490 (0.0057)	**0.1054** (0.0055)
F(15%)	$\mathbf{\Lambda}$	0.1414 (0.0072)	**0.0923** (0.0043)
	θ	0.1492 (0.0037)	**0.0721** (0.0040)
	p	0.1522 (0.0038)	**0.0721** (0.0040)
F(35%)	$\mathbf{\Lambda}$	0.2496 (0.0082)	**0.1079** (0.0045)
	θ	0.3182 (0.0066)	**0.1012** (0.0056)
	p	0.3510 (0.0051)	**0.1012** (0.0056)
F(55%)	$\mathbf{\Lambda}$	0.3677 (0.0087)	**0.1376** (0.0079)
	θ	0.4164 (0.0128)	**0.1052** (0.0055)
	p	0.5489 (0.0058)	**0.1053** (0.0055)

6 Concluding remarks

We proposed a method to build trees and forests for a non-homogeneous Poisson response with excess zeros, based on two forests. Unlike the ZIP tree proposed by Lee and Jin (2006), our method has two parts, like the usual parametric ZAP and ZIP models. This allows for different covariates effects for the zero part and the Poisson part. Any flexible method to model the probability for a binary response can be used for the zero part. Here we used the traditional random forest for a binary response. For the Poisson part, we used a forest of trees built with a splitting criterion derived from the zero truncated non-homogeneous Poisson likelihood. Our method extends the work of Lee and Jin (2006) in the sense that it can handle a non-homogeneous rate function. Our method also extends the work of Mathlouthi, Fredette and Larocque (2015) by allowing for excess zeros.

The results from extensive simulation studies clearly show the merits of the proposed method. A mix of homogeneous and non-homogeneous ZAP and ZIP models were used to generate artificial data. The proposed method was compared to parametric ZAP and ZIP models, to a basic Poisson forest and to a forest built with the Lee and Jin (2006) ZIP tree approach. The proposed method had the smallest mean absolute error for 71 out of the 81 estimation problems considered, and it was a close second for four others. The six cases where the proposed method was not the best or a close second was for estimating the Poisson intensity when the parametric model was correctly specified for this parameter.

Our approach is very general and many extensions are possible. Firstly, if the Poisson assumption is not appropriate, other models, like the negative binomial, could be used. This could be useful in cases where extra-Poisson variation is observed, for instance with clustered data. Secondly, we only considered covariates that do not vary with time. It would be interesting to generalize our approach to be able to include time-varying covariates.

References

Bou-Hamad, I., Larocque, D., Ben-Hameur, H., Msse, L. C., Vitaro, F. and Tremblay, R.E. (2009). Discrete-time survival trees. Canadian Journal of Statistics, 37, 17–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth International Group.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

Chaudhuri, P., Lo, W.-D., Loh, W.-Y., Yang, C.-C. (1995). Generalized regression trees. Statistica Sinica, 5, 641–666.

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

Hilbe, J.M. (2011). Negative Binomial Regression, 2nd edition. Cambridge University Press. Cambridge.

Ishwaran H., Kogalur U.B., Blackstone E.H. and Lauer M.S. (2008). Random survival forests. Annals of Applied Statistics, 2, 841–860.

Ishwaran H. and Kogalur U.B. (2015). Random forests for survival, regression and classification (RF-SRC), R package version 1.6.1. http://cran.r-project.org/web/packages/randomForestSRC/index.html.

Jackman, S. (2015). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University. Stanford, California. R package version 1.4.8. http://pscl.stanford.edu/.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34, 1–14.

Lee, S.K. and Jin, S. (2006). Decision tree approaches for zero-inflated count data. Journal of Applied Statistics, 33, 853–865.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. R News, 2, 18–22.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. Statistica Sinica, 12, 361–386.

Mathlouthi, W., Fredette, M. and Larocque, D. (2015). Regression trees and forests for non-homogeneous Poisson processes. Statistics and Probability Letters, 96, 204–211.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics, 33, 341–365.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Su, X., Wang, M. and Fan, J. (2004). Maximum likelihood regression trees. Journal of Computational and Graphical Statistics, 13, 586–598.

Therneau, T.M., Atkinson, B. and Ripley, B. (2014). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-8. http://CRAN.R-project.org/package=rpart.

Zeileis, A., Kleiber, C. and Jackman, S. (2008). Regression models for count data in R. Journal of Statistical Software, 27(8), 1–25.