Do we always need a zero inflated model to capture an apparent excess of zeros?

Alexandra M. Schmidt[1]* João Batista M. Pereira[1]
and Pedro Paulo Vieira[2]

[1]IM - UFRJ [2]FMTAM - Amazonas, Brazil

April 2008

Abstract

Count data are usually modeled as following a Poisson distribution. Implicit in the use of a Poisson distribution is the assumption that the observations are generated from a distribution with the same mean and variance. However, this is rarely true in practice. Frequently, data present an observed variance greater than the observed mean, i.e., the data are overdispersed. In Epidemiology, when modelling the number of cases of a disease, overdispersion might be caused by a great amount of zeros in the data. However, we do not know if this is a *true* zero or not (the disease might be *present* but was *not observed*). Here we discuss, from a Bayesian point of view, how to deal with this problem on a time series of registered number of cases of malaria for a municipality located in the Amazon region of Brazil. We have available the monthly number of cases of malaria and 71% of the observations are zeros. We investigate if this is the source of overdispersion. Generalized dynamic models are used to capture the temporal structure of the data, and different families of distributions, like the Poisson, Poisson-Gamma, Poisson-Log-normal and the zero inflated versions of these, are used to fit the data. Markov chain Monte Carlo methods are used to obtain a sample from the posterior distribution and efficient sampling schemes, which take into account the correlation structure of the parameters, are used to build the sampling algorithm. Analysis of the posterior predictive distribution and two other model comparison criteria, indicate the negative binomial distribution as the best among those fitted. However, the zero inflated negative binomial version provides an estimate of the probability that the ”observed zero” was expected from the negative binomial part of the model.

Key Words: Bayesian paradigm; Generalized Dynamic models; Latent variables; Mixture models; Zero inflation.

*[*Address for correspondence*: Alexandra M. Schmidt, Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Caixa Postal 68530, Rio de Janeiro, RJ, Brazil. CEP 21945-970. *Tel.:* `0055 21 2562-7505 x 204`. *Fax:* `0055 21 2562-7374`. *E-mail*: `alex@im.ufrj.br`. *Homepage*: www.dme.ufrj.br/∼alex

1 Introduction

In Epidemiology, we are usually interested in modelling the number of cases of a certain disease over a period of time. It is quite common to assume that the number of cases at time t is a realization from a discrete random variable. The most used distribution is the Poisson. However, the Poisson assumes that mean and variance are equal, which is hardly true in practice. Usually, the variance is much greater than the mean. When this happens it is said that data are overdispersed. Another point of concern is that if the disease is rare, or if we span a short period of time, we might get a great amount of observed zeros. Also, the presence of too many zeros might be the source of overdispersion. Typically, however, in such situations, we do not know if the observed zero is a *true* one or not (the disease might be *present* but was *not observed*). Our contribution here lies on the investigation, on a time series of malaria counts, of the source of overdispersion, and the probability, at time t, that the observed zero is a *true* one.

There has been, in the literature, a lot of effort in tackling these problems. To account for overdispersion it is quite common to use a mixture between the Poisson and Gamma distributions, which marginally results on a negative binomial distribution. Another mixture which is quite used is between a Poisson and Log-Normal distributions. From a classical point of view the latter is more complicated to fit as the marginal distribution is unknown. If there are too many zeros, considering only these mixtures might not be enough to capture a possible excess of zeros. Lambert (1992) proposed a zero inflated Poisson model that captures the excess of zeros through a mixture between a Bernoulli and Poisson. Essentially, this mixture is defined by the introduction of an indicator variable of the presence/absence of the disease. Usually, the likelihood of such models is based on the marginalized distribution with respect to this indicator variable. In other words, this indicator variable is assumed to be known. Here, however, we will assume it to be known only when the number of counts is greater than zero. Otherwise, these become parameters of the model and have also to be estimated.

To account for overdispersion, Scollnik (1995) presents a Bayesian analysis of ordinary Poisson and generalized Poisson distribution models. On a spatio-temporal setting, Kim et al. (2002) propose a quasi-multiplicative spatio-temporal model with gamma extra variation effects, and compare the performance of the proposed model to a loglinear model with lognormal extra variation effects. They conclude that for their dataset, the models are interchangeable when gamma and lognormal distributions have similar location and scale parameters. Agarwal et al. (2002) propose a zero inflated Poisson model for spatial count data which present an excess number of zeros. In their case, they estimate the probability of the observed zero coming from the Poisson part of the mixture. Therefore, when a zero is observed, it is not considered as a "true" one. Yau et al. (2004) propose a zero inflated negative binomial model to analyze a set of pancreas disorder length of stay data. They make use of random effects to account for inter-hospital variations. And the likelihood is based on the negative binomial distribution and not on the mixture between the Poisson and Gamma distributions. Inference procedure is performed via the maximization of an appropriate log-likelihood function through the use of an EM algorithm. Dagne (2004) also proposes a fully Bayesian hierarchical model which incorporates both overdispersion and zero inflation. More specifically, he considers only the case of a zero inflated Poisson model, and includes independent random effects to capture possible heterogeneity present in the data. Warton (2005) investigates the need of a zero inflated model to describe abundance data that have many zeros. He interchanges among the fitting of the Poisson and the negative binomial models, and the zero inflated versions of these. The estimation procedure is done via maximum likelihood and the method of moments. An important issue is that when he fits the zero inflated models, he assumes the zeros as being all observed, when in practice these are unknown, as all is known is that a particular species was not observed.

This paper is organized as follows. Next section describes the data which motivated this study. Then, Section 3 introduces the general structure of the model which will be fitted to the data, and describes all the particular cases. Inference is performed following the Bayesian paradigm, so therein the associated prior distributions as well as the MCMC procedure are also discussed. In Section 4 we present the results of the analysis of the proposed models. We discuss the significance of overdispersion for our dataset, as well as the advantage (or not) of fitting a zero inflated model. Therein the benefits of a Bayesian analysis are clear, as all the uncertainty about the parameter's estimates is naturally described. Finally, Section 5 concludes this study.

2 Motivation

Malaria is a public health problem in more than 90 countries inhabited by approximately 2400 million people, representing 40% of the population of the world. Best estimates currently describe the annual burden of malaria as 1,12 million deaths and 300-500 million clinical cases. More than 90% of the burden of the disease falls in Sub-Saharan Africa where almost all deaths are attributable to *P.falciparum* infections. Most of the remaining is distributed between the Indian sub-continent, South-east Asia, Oceania and the Americas. After *Plasmodium falciparum* the largest burden of the disease is caused by *Plasmodium vivax* (WHO, 1998).

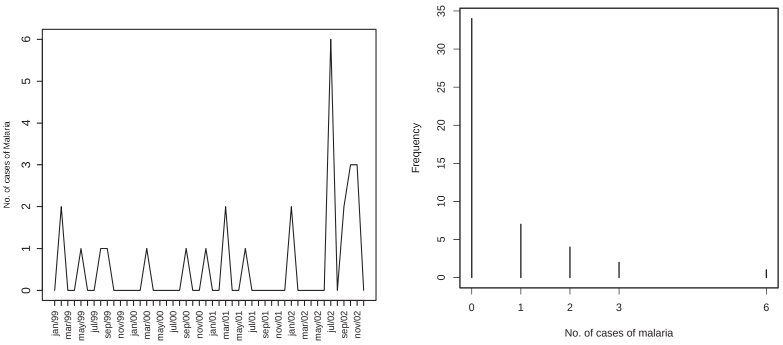
Across the Brazilian Amazon basin, more than 500 thousand people are infected every year, and the disease is a major threat to human health, despite considerable national and international control efforts. Continued progress in prevention, treatment and the development of innovative tools for the control of malaria is required[1].

In this study we are particularly interested on a hipoendemic malarial area called Barreirinha, a municipality located at the northeastern part of the Amazonas state (57o 13' 43,31" W / 2o 45' 47, 73" S), 372 kilometers far from the capital Manaus, and inhabited by approximately $25,000$ people. This region, located by the Amazonas river, has registered a particularly low number of positive malaria cases

[1] From the report *Situação Epidemiológica da Malária no Brasil* published by the Health Ministery of Brazil, 2006.

(< 10/year) from 1999 to 2002 as provided by the Brazilian Epidemiological Surveillance System (SIVEP)[2].

We have available monthly time series of registered number of cases of malaria for Barreirinha. Observations correspond to January 1999 until December 2001. Panel (a) of Figure 1 shows the observed time series whereas panel (b), of this same figure, presents the observed counts of malaria cases in Barreirinha during this period. Note that 71% of the observations are zeros.



(a) Times Series

(b) Observed Counts

Figure 1: Panel (a) Time series of the registered cases of malaria from January, 1999 until December, 2001, for the municipality of Barreirinha, located in the Amazon state of Brazil. Panel (b) observed counts of the number of cases of malaria registered between Jan 1999-Dec 2001.

We aim to investigate if such a big proportion of zeros was expected or if they are a sign of under reported cases. Moreover, we want to estimate the probability of presence of the disease at time t, given it was not observed.

3 Proposed Models

Assume that Y_t represents the number of cases of malaria over a specific region for time $t \in \{1, 2 \cdots\}$. Usually, it is assumed that Y_t follows a Poisson distribution with mean λ_t. In this case, it is well known that the variance of Y_t is also equals λ_t. However, in practice, it is quite unusual to observe data in which mean and variance are similar. Usually, there is the presence of some extra variability, that is $V(Y_t) \gg E(Y_t)$. When this is the case there are some alternatives to the usual Poisson model. One possibility is to assume a continuous mixture such that

$$Y_t \mid \lambda_t, \delta_t \sim Poi(\lambda_t \, \delta_t), \tag{3.1}$$

and $p(\delta_t)$ follows a continuous distribution assuming positive values. It is well known that if δ_t follows a Gamma distribution with mean α/β and variance α/β^2, then the

[2] `www.saude.gov.br/sivep_malaria` .

distribution of $Y_t \mid \lambda_t$, obtained through,

$$\int p(y_t \mid \lambda_t, \delta_t) p(\delta_t \mid \alpha, \beta)\, d\delta_t = p(y_t \mid \lambda_t, \alpha, \beta), \tag{3.2}$$

follows a negative binomial distribution with mean and variance given, respectively, by

$$\begin{aligned} E(Y_t \mid \lambda_t, \alpha, \beta) &= \lambda_t \frac{\alpha}{\beta}, \\ V(Y_t \mid \lambda_t, \alpha, \beta) &= \lambda_t \frac{\alpha}{\beta} + \lambda_t^2 \frac{\alpha}{\beta^2}. \end{aligned} \tag{3.3}$$

Another possibility is to consider a lognormal mixture, that is, to assume $\delta_t \sim LN(\mu, V)$, where $LN(a, b)$ stands for the lognormal distribution whose associated normal has mean a and variance b. From a frequentist view point, this model is not used because the integral in (3.2) does not have an analytical solution. But using the properties of conditional expectation, it can be shown that

$$\begin{aligned} E(y_t \mid \lambda_t) &= \lambda_t \exp\left(\mu + \frac{V}{2}\right), \\ V(y_t \mid \lambda_t) &= \lambda_t \exp\left(\mu + \frac{V}{2}\right) + \lambda_t^2 \exp(2\mu + V)(\exp(V) - 1). \end{aligned} \tag{3.4}$$

Notice that both mixtures capture overdispersion as the variance in both cases is the mean plus a positive quantity.

When modelling the number of cases of a certain disease, usually, there is a great amount of zeros in the data. This might be the source of overdispersion. In this case, one might use a zero inflated version of the Poisson, Poisson-Gamma or Poisson-Lognormal model.

Let X_t be a random variable representing the presence ($X_t = 1$) or absence ($X_t = 0$) of the process being observed. Assume that $X_t \mid \theta$ follows a Bernoulli distribution with probability of success θ. Let $p(y_t \mid \lambda_t, \delta_t)$ be a model for the process being observed *given* it is *present* ($X_t = 1$). We use this general notation, conditioned on δ_t, but in the Poisson case, δ_t is known and equals 1. By definition, $P(Y_t = y_t \mid \lambda_t, \delta_t, X_t = 1) = p(y_t \mid \lambda_t, \delta_t)$ and $P(Y_t = 0 \mid \lambda_t, \delta_t, X_t = 0) = 1$. The joint density function of X_t and Y_t is given by

$$p(y_t, x_t \mid \lambda_t, \delta_t, \theta) = [\theta\, p(y_t \mid \lambda_t, \delta_t)]^{x_t} (1 - \theta)^{1 - x_t}. \tag{3.5}$$

Therefore, the marginal distribution of Y_t, with respect to X_t, is given by

$$p(y_t \mid \lambda_t, \delta_t, \theta) = \theta\, p(y_t \mid \lambda_t, \delta_t) + (1 - \theta) p_0(y_t \mid \lambda_t, \delta_t), \tag{3.6}$$

where $p_0(y_t \mid \lambda_t, \delta_t)$ represents the probability of observing $[y_t = 0]$.

Notice that X_t is a latent variable. Following Agarwal et al. (2002), its inclusion in the model results that $P(Y_t = 0 \mid X_t = 1, \theta, \lambda_t) = p_0(y_t \mid \lambda_t, \delta_t)$, $P(Y_t = y_t \mid X_t = 0, \theta, \lambda_t, \delta_t) = 0$, $P(X_t = 1 \mid Y_t = y_t > 0, \theta, \lambda_t, \delta_t) = 1$ and, more interestingly,

$$P(X_t = 1 \mid Y_t = 0, \theta, \lambda_t, \delta_t) = \frac{\theta\, p_0(y_t \mid \lambda_t, \delta_t)}{(1 - \theta) + \theta\, p_0(y_t \mid \lambda_t, \delta_t)}, \tag{3.7}$$

and marginally, as previously defined, X_t follows a Bernoulli distribution with probability of success θ, that is $X_t \mid \theta \sim ber(\theta)$. Notice that Equation (3.7) provides an estimate of the probability of presence of the process, at time t, given that it was not observed. That is, this represents the probability that the observed 0 comes from the model $p(y_t \mid \lambda_t, \delta_t)$. The quantity $1 - P(X_t = 1 \mid Y_t = 0, \theta, \lambda_t, \delta_t) = P(X_t = 0 \mid Y_t = 0, \theta, \lambda_t, \delta_t)$ might be used as a guidance to Epidemiologists to point those regions which are "suspect" of having under reported cases.

Usually, the likelihood is obtained through the probability function in Equation (3.6). In our context, X_t is considered unknown when $Y_t = 0$, therefore, these X_t's are parameters of the model which need to be estimated. From a Bayesian point of view this is a not problem, since we can compute the posterior full conditional distribution for these X_t and sample from it in the sampling scheme which will be described in Subsection 3.2.

Modelling the temporal structure

As we do not have many observations, it is not possible to investigate for any trend or seasonal structure. Following the literature on generalized dynamic models (West and Harrison, 1997), we assume that $\log(\lambda_t)$ varies smoothly with time, that is,

$$\log \lambda_t \;=\; \mu_t \quad \text{and} \quad \mu_t = \mu_{t-1} + \omega_t \;\text{ with }\; \omega_t \sim N(0, W), \;\text{ for } t = 1, \cdots, T,$$

and $\mu_0 \sim N(0, C_0)$, C_0 is a known (large) variance, and W is another parameter of the model which needs to be estimated.

When analyzing our data we will entertain among 6 models, the Poisson, Poisson-Gamma, Poisson-Lognormal and the zero inflated versions of these, ZIP, ZIP-Gamma, and ZIP-LN. Our aim is to model the cases of malaria in Barreirinha, such that we check which one, among these, fits the data best.

3.1 Likelihood Function and Prior Distributions

Assume we have observations y_t, for $t = 1, \cdots, T$. We can write the likelihood based on the models described in the previous subsection as

$$p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\lambda}, \boldsymbol{\delta}, \theta) = \prod_{t=1}^{T} \theta^{x_t} (1-\theta)^{1-x_t} \left[\frac{(\lambda_t \delta_t)^{y_t} \exp(\lambda_t \delta_t)}{y_t!} \right]^{x_t}.$$

Let $\boldsymbol{\Theta}$ be the parameter vector comprising the quantities of the model which need to be estimated. Broadly speaking $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \boldsymbol{\delta}, W, \mathbf{X}, \theta)$, where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_T)$, $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_T)$, and $\mathbf{X} = (X_1, \cdots, X_T)$. In the general expression above, if we assume $0^0 = 1$, for each model specification we then have:

1. Poisson (Pois): $\delta_t = 1$, $x_t = 1 \;\forall\, t$, and $\theta = 1$;

2. Negative Binomial (NB): δ_t follows a Gamma distribution and $x_t = 1$, $\forall\, t$, $\theta = 1$;

3. Poisson-Lognormal (Pois-LN): δ_t follows a lognormal distribution and $x_t = 1$, $\forall\, t$, $\theta = 1$;

4. Zero Inflated Poisson (ZIP): $\delta_t = 1$, $\forall t$, and $x_t = 1$ if $y_t > 0$ but x_t is unknown when $y_t = 0$;

5. Zero Inflated Negative Binomial (ZINB): δ_t follows a Gamma distribution and $x_t = 1$ if $y_t > 0$ but x_t is unknown when $y_t = 0$;

6. Zero Inflated Poisson-Lognormal (ZIP-LN): δ_t follows a lognormal distribution and $x_t = 1$ if $y_t > 0$ but x_t is unknown when $y_t = 0$;

These are the 6 models which will be fitted to the malaria dataset discussed in Section 2.

Prior Specification

From a Bayesian point of view we are now left to assign the prior distribution of the parameters involved in the model. Here, in particular, for the Poisson-Gamma model, we assume that $\delta_t \mid \epsilon \sim Ga(\epsilon, \epsilon)$. In this case, (3.3) becomes

$$E(Y_t \mid \lambda_t, \epsilon) = \lambda_t, \qquad V(Y_t \mid \lambda_t, \epsilon) = \lambda_t + \frac{\lambda_t^2}{\epsilon}.$$

For the Poisson-Log-Normal model we assume $\delta_t \mid \epsilon \sim LN(0, \epsilon)$. Therefore, equation (3.4) becomes

$$E(y_t \mid \lambda_t) = \lambda_t \exp\left(\frac{\epsilon}{2}\right), \quad V(y_t \mid \lambda_t) = \lambda_t \exp\left(\frac{\epsilon}{2}\right) + \lambda_t^2 \exp(\epsilon)(\exp(\epsilon) - 1).$$

For both specifications, it is assumed, *a priori*, that $\epsilon \sim Ga(0.1, 0.1)$. For W, the variance of the evolution equation of μ_t, we assume an inverse gamma prior with an infinite variance and mean equals 1. This reflects our prior belief that W should assume very large values with very low probability, *a priori*. When θ is unknown, we assume a uniform prior in the interval $(0, 1)$.

3.2 Inference Procedure

Following the Bayesian paradigm, the posterior distribution is proportional to likelihood times prior. For all 6 models considered for the malaria dataset, a unknown posterior distribution results. We resort to efficient Markov Chain Monte Carlo algorithms to obtain samples from the posterior distribution of interest. In particular, we make use of the Gibbs sampling with some steps of the Metropolis-Hastings (M-H) method. See Gamerman and Lopes (2006) for more details.

Our main concern is when sampling the μ_t's. Conditional on δ_t, we have a generalized dynamic linear model (West and Harrison, 1997). It is well known that these parameters are highly correlated, *a posteriori*. The posterior full conditional distributions have a unknown form. We make use of the conjugate updating backward sampling (CUBS) algorithm proposed by Ravines et al. (2007), which takes this correlation structure into account. More specifically, the proposal for the Metropolis-Hastings step is a normal distribution, whose mean and variance are computed according to a Linear Bayes approximation. See Ravines et al. (2007) for details. For the models where θ is unknown, its posterior full conditional distribution follows a beta distribution, which is easy to sample from. In the zero inflated models, when $y_t = 0$ we have to sample the corresponding indicator variable X_t. Its posterior full conditional follows a Bernoulli distribution with probability of success given by equation (3.7). The δ_t's, when unknown, also result on unknown posterior full conditional distributions. They are sampled one at a time, through a M-H step, based on a lognormal distribution, whose mean of the associated normal is equal to the current value of the chain and the variance of the proposal is tunned according to the algorithm proposed by Roberts and Rosenthal (2001).

4 Results

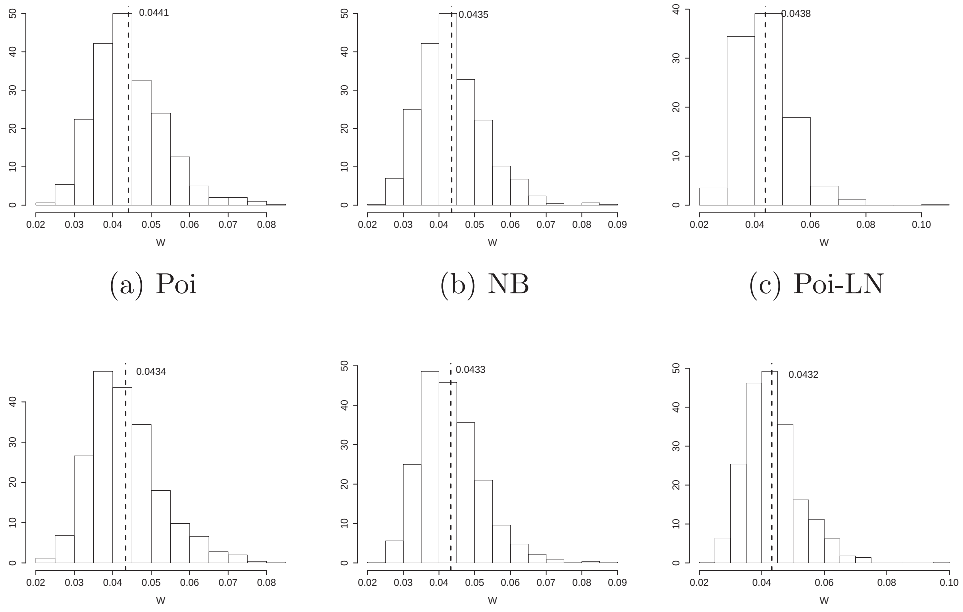
We let the MCMC run for 200,000 iterations, used 40,000 as burn in and stored every 160th iteration, to avoid autocorrelation among the sampled values. Convergence was checked following standard procedures available in CODA (Plummer et al., 2006). We show the results for all fitted models in order to investigate the contributions provided by each of them.

The panels in Figure 2 show the posterior distribution of W under each of the fitted models. We notice that there is not much difference among the estimated values, and on the resultant posterior samples.

For all models, $\lambda_t = \exp(\mu_t)$, as we have a sample from the posterior distribution of μ_t, due to this one-to-one relationship we automatically obtain a sample for the posterior of λ_t. Figure 3 shows the posterior mean (solid lines) and respective 95% posterior credible intervals of λ_t, for each time t. It is very clear, that they all present an increasing pattern and they seem not to differ much from model to model.

On the other hand, for each time t, Figure 4 summarizes the posterior mean with respective 95% credible intervals for δ_t, the parameter which is capturing the overdispersion present in the observed time series. For most of the observed times, $\delta = 1$ is included in the 95% posterior credible interval, giving an indication that there is no need for this parameter. However, for the negative binomial model, in the end of the series this is not true, that is $\delta = 1$ is not included in the posterior credible interval. Apparently, the same characteristic happens under model ZINB.

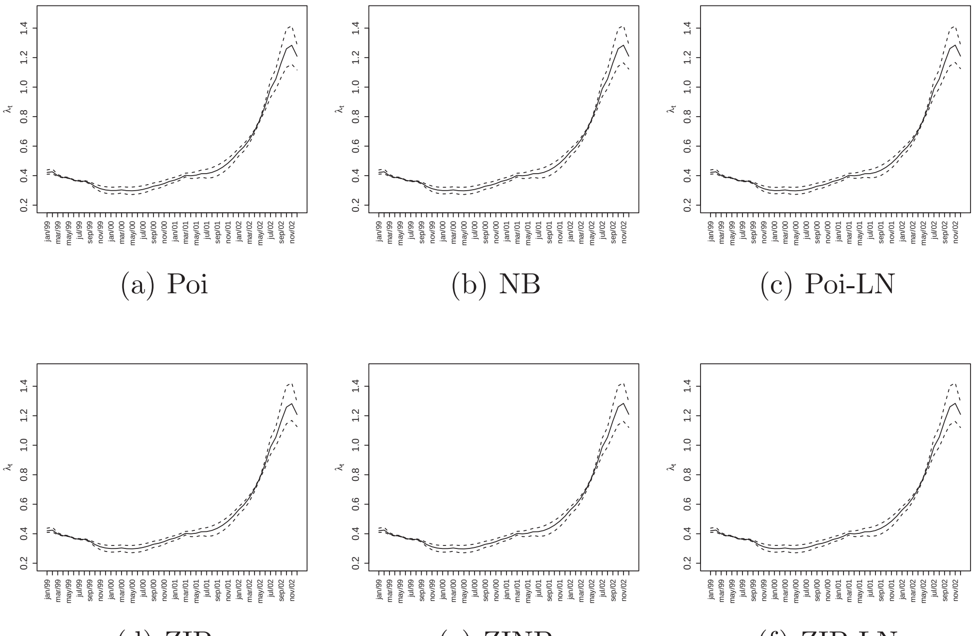
To investigate this better, panels in Figure 5 show the posterior summary of λ_t^2/ϵ and $\lambda_t \exp\left(\frac{\epsilon}{2}\right) + \lambda_t^2 \exp(\epsilon)(\exp(\epsilon) - 1)$, the terms that capture overdispersion, under models (a) NB, (c) ZINB, and (b) Poi-LN, and (d) ZIP-LN, for each time t, respectively. From these panels, it is clear that models NB and ZINB present values which are significantly greater than zero towards the end of the time series. The same is not true for the Poi-LN and ZIP-LN models.



(a) Poi (b) NB (c) Poi-LN

(d) ZIP (e) ZINB (f) ZIP-LN

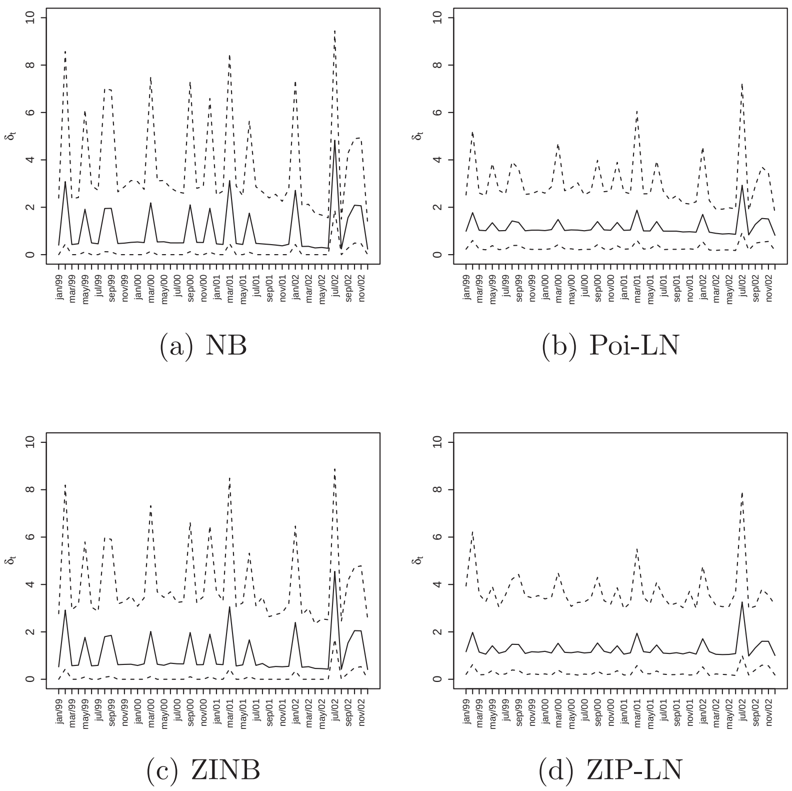
Figure 2: Posterior sample of W, under each model specification. The vertical line represents the estimated posterior mean.



(a) Poi (b) NB (c) Poi-LN

(d) ZIP (e) ZINB (f) ZIP-LN

Figure 3: Posterior mean (solid lines) of λ_t for each time t, and respective 95% posterior credible interval (dotted lines).



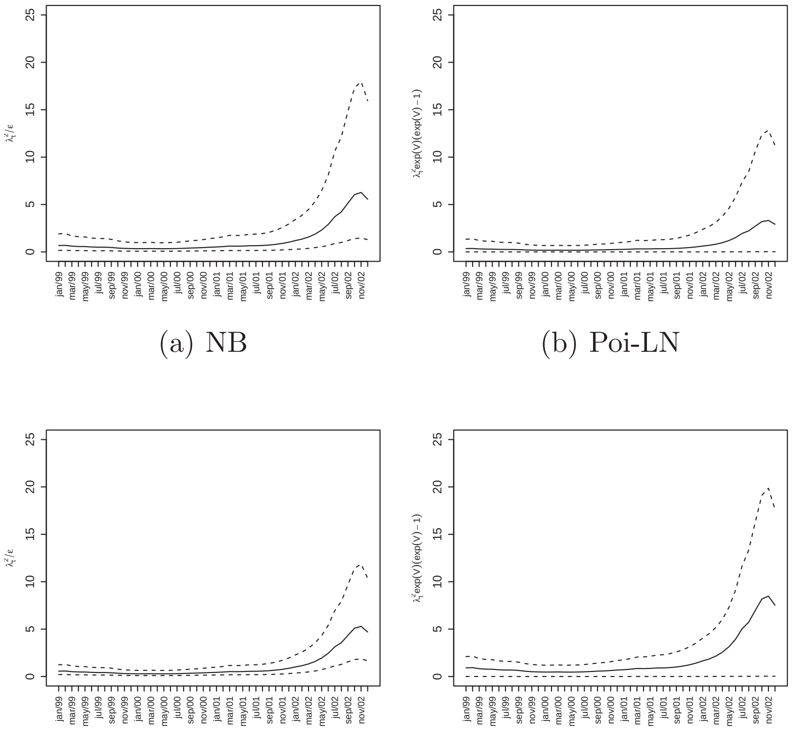
(a) NB

(b) Poi-LN

(c) ZINB

(d) ZIP-LN

Figure 4: Posterior mean (solid lines) of δ_t for each time t and respective 95% posterior credible interval (dashed lines).



(a) NB

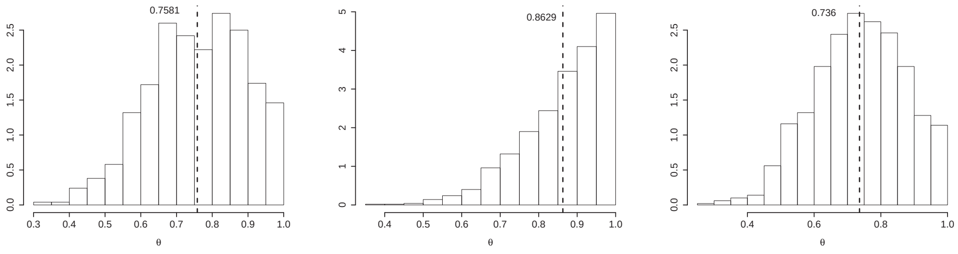
(b) Poi-LN

(c) ZINB

(d) ZIP-LN

Figure 5: Posterior mean (solid lines) and respective 95% credible intervals limits (dashed lines) of λ_t^2/ϵ in panels (a) and (c) and of $\lambda_t^2 \exp(\epsilon)(\exp(\epsilon) - 1)$ in panels (b) and (d), for each time t.

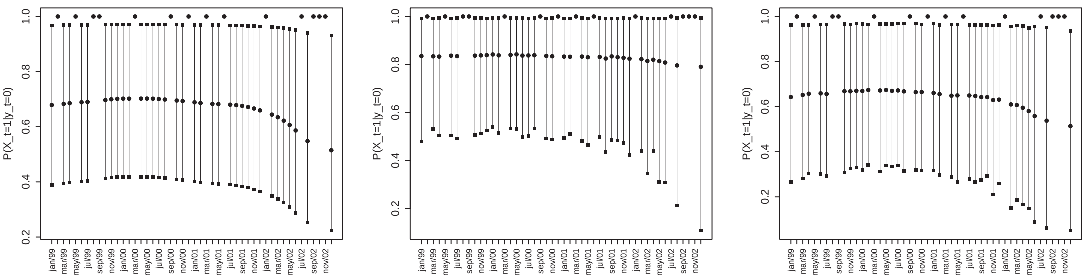
Recall that when we fit the zero inflated models, we have a parameter, θ, indicating the prior probability of presence of the disease. The panels in Figure 6 show the histograms of the posterior sample for θ under the (a) ZIP, (b) ZINB, and (c) ZIP-LN models. We notice that the ZINB model provides a skewed distribution, with probability mass quite concentrated near 1. On the other hand, under the ZIP and ZIP-LN models, apparently we are more uncertain about the estimate of θ and for these models, the estimated value of θ is smaller than under ZINB. Apparently, although the ZIP-LN has a parameter to capture overdispersion, this is not able to explain the zeros with high probability, differently from the ZINB model.



(a) ZIP (b) ZINB (c) ZIP-LN

Figure 6: Posterior sample of θ, under each zero inflated model specification: (a) ZIP, (b) ZINB, and (c) ZIP-LN. The vertical dashed line represents the estimated posterior mean.

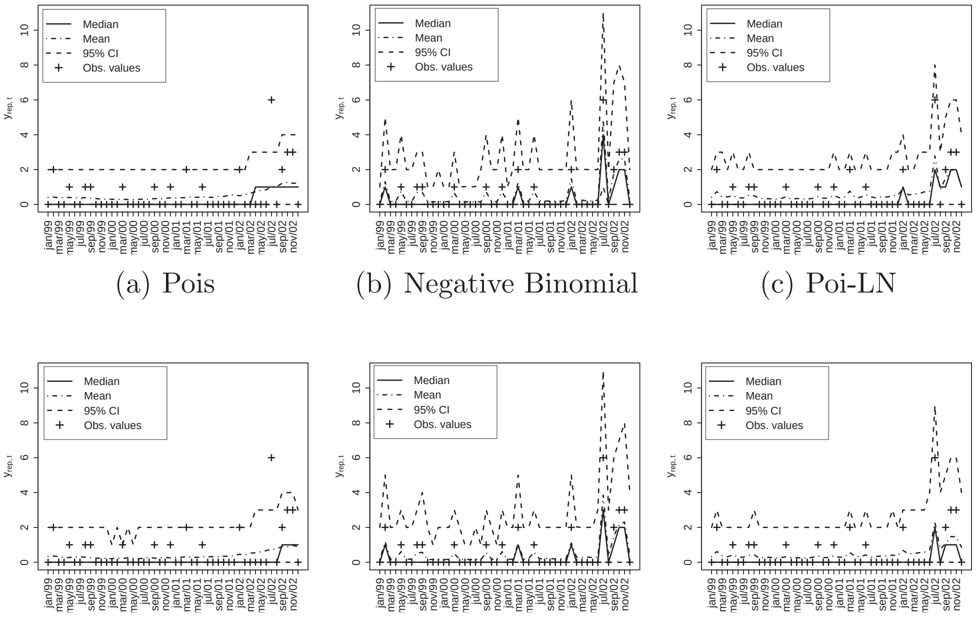
We can go further and compute the posterior probability of a "zero" coming from the distribution $p(y_t \mid .)$ (see eq. (3.7)). Therefore, we can estimate the posterior distribution of the probability of presence of the disease given it was not observed. Figure 7 shows, for each time t, the posterior summary (mean (dot) and 95% credible intervals (extremes of the vertical lines)) of the probability of presence given malaria was not observed at time t for models (a) ZIP, (b) ZINB and (c) ZIP-LN. According to this figure, the ZIP model indicates that the posterior probability of the "observed zero" coming from the Poisson part of the model is around 70%, and for each time, these probabilities vary from around 39% up to approximately 1. A similar behaviour is observed for the ZIP-LN model. On the other hand, the ZINB model provides a stronger result. We notice from panel (b) of this figure that for most of the times, the probability in equation (3.7) is estimated at above 80%, and they vary from around 48% up to 1. What is really interesting from these panels is that all of them indicate that, for the last periods of time, specially for August, 2002 and December 2002, although the estimated probability that the zero was expected is at 80%, the estimated uncertainty is quite high, as it varies from almost 0 up to 1. From the time series we notice that August 2002 is soon after the month that had the highest value of the number of cases. In other words, these zero inflated models are indicating that they are quite uncertain that these months were expected to present counts equal zero. Policy makers should try and understand the source of this variability.



(a) ZIP (b) ZINB (c) ZIP-LN

Figure 7: Posterior summary of the probability of presence of Malaria, for each time t, given it was not observed, under models (a) ZIP; (b) ZINB; (c) ZILN.

Lastly, panels in figure 8 present the mean (dot-dashed), median (solid line) and 95% credible intervals (dashed lines) of the posterior predictive distribution for each of the fitted models. From these panels it is clear that, neither the Poisson, nor the ZIP models fit the data well. They estimate zeros for most instants in time. The Poi-LN and ZIP-LN models are slightly better but do not fit the data as well as the NB and ZINB. Notice that both, NB and ZINB, seem to capture reasonably well the structure of the data.



(a) Pois (b) Negative Binomial (c) Poi-LN

(d) ZIP (e) ZINB (f) ZIP-LN

Figure 8: Replication of the observations under each fitted model.

Model Comparison

In order to compare these fitted models we use two different criteria (i) the *Deviance Information Criterion* (DIC) (Spiegelhalter et al., 2002) and (ii) the posterior predictive loss (EPD) introduced by Gelfand and Ghosh (1998). Here the EPD is based on the deviance computed for the Poisson distribution. Both criteria are based on the sum of two components, one which indicates the goodness of fit (G in EPD and $\overline{D}$ in DIC), the other which penalizes for the number of parameters (P in EPD and p_D in DIC). For both, the smallest value, among the fitted models, indicate the best one. Table 1 presents the results of both criteria.

Table 1: Values of DIC and EPD for each fitted model.

Model	EPD			DIC		
	G	P	D	$\overline{D}$	p_D	DIC
Pois	154.00	49.59	203.59	92.40	7.64	100.04
ZIP	187.64	76.21	263.85	128.14	14.59	142.73
BN	48.85	48.44	**97.29**	70.39	22.64	**93.02**
ZINB	78.42	60.83	139.25	97.46	35.18	132.64
Pois-LN	114.51	54.44	168.94	107.76	12.26	120.02
ZIP LN	160.97	72.93	233.91	134.79	8.50	143.28

It is clear that both agree that the NB model gives the best results. As expected (because they have more parameters), the zero inflated versions of the models present higher values of both EPD and DIC when compared to the simpler versions. Among the zero inflated models, ZINB is the one that result on the smallest values of both EPD and DIC. In other words, both criteria agree with the analysis when we look at the posterior predictive distribution (see Figure 8).

5 Conclusions

This paper discusses the need of fitting a zero inflated model on a time series corresponding to malaria counts in the municipality of Barreirinha, Brazil. We entertain among different distributions, the Poisson, Poisson-Gamma and Poisson-Log-Normal, to check which, among these, fits the data best. We go further and fit the zero inflated versions of these distributions. And in this matter, our main contribution lies on estimating, for every time t, the probability of presence of the disease, together with its associated uncertainty, given it was not observed. This measurement might give policy makers indications of periods of time that underreporting cases might have occurred.

The temporal structure of the data was naturally accounted for through the use of dynamic generalized linear models. These models naturally impose a correlation structure among the parameters and care must be taken when building a

MCMC algorithm to obtain samples from the posterior distribution. We made use of the conjugate updating backward sampling method (CUBS), recently proposed by Ravines et al. (2007), and the chains seemed to converge well.

For the analyzed time series, if we had to choose a model, we would probably choose the NB. Actually, all fitted models give an indication of an increase on the level of cases of malaria in Barreirinha. But the analysis shows that if we start from the simplest model, the Poisson, the fitted values are not good. Then, if we include a component to capture the excess of zeros, we do not get a good fit yet. In other words, this might be an indication that the overdispersion present in the data is not coming solely from the excess of zeros (71% of the observations). The NB model is able to explain the data quite well, even to capture the peak at August 2002. The zero inflated version of the NB model is interesting as it gives an estimate of the probability of the "observed zero" coming from the negative binomial part of the model. From the ZINB model is clear that the observations come from the negative binomial with probability 80%. Only the last observations of the series seem to be suspicious, that is, to indicate that there were some kind of underreport of the number of cases of malaria for these months. So if we were searching for the model that fits the data best, we would not need a zero inflated component to explain this time series. However, the zero inflated version might be interesting to investigate if there are indications of underreports.

Acknowledgements

The authors wish to thank Cynthia de Oliveira Ferreira, from the Malaria Department at FMTAM and Raul Amorim, from the Epidemiology and Public Health Department at FMTAM, for making the dataset available. Also, they thank Marco A. Rodríguez for some suggestion which improved the presentation of this manuscript. Alexandra M. Schmidt is grateful to CNPq and FAPERJ for funding this research. João Batista M. Pereira was supported by a scholarship from PIBIC/CNPq.

References

Agarwal, D. K., Gelfand, A. E. and Citron-Pousty, S. (2002) Zero inflated models with application to spatial count data. *Environmental and Ecological Statistics*, **9**, 341–355.

Dagne, G. A. (2004) Hierarchical Bayesian analysis of correlated zero-inflated count data. *Biometrical Journal*, **46**, 653–663.

Gamerman, D. and Lopes, H. F. (2006) *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference.* 2nd Edition, Chapman & Hall.

Gelfand, A. E. and Ghosh, S. K. (1998) Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

Kim, H., Sun, D. and Tsutakawa, R. K. (2002) Lognormal vs. gamma: Extra variations. *Biometrical Journal*, **3**, 305–323.

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

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11. URL`http://CRAN.R-project.org/doc/Rnews/`.

Ravines, R. R., Migon, H. S. and Schmidt, A. M. (2007) An efficient sampling scheme for dynamic generalized models. *Tech. rep.*, No. 201/2007. Departamento de Métodos Estatísticos, IM-UFRJ, Brazil.

Roberts, G. O. and Rosenthal, J. S. (2001) Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, **16**, 351–367.

Scollnik, D. P. M. (1995) Bayesian analysis of two overdispersed Poisson models. *Biometrics*, **51**, 1117–1126.

Spiegelhalter, D., Best, N., Carlin, B. and Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. B*, **64**, 583–639.

Warton, D. I. (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, **16**, 275–289.

West, M. and Harrison, P. J. (1997) *Bayesian Forecasting and Dynamic Models.* Springer-Verlag New York, Second Edition.

WHO (1998) *Fact Sheet No 94.* World Health Organization Press Office, Geneva, Switzerland.

Yau, K. K. W., Wang, K. and Lee, A. H. (2004) Zero inflated Negative Binomial mixed regression modelling of over-dispersed count data with extra zeros. *Biometrical Journal*, **46**, 653–663.