# Observation-Driven Model for Zero-Inflated Daily Counts of Emergency Room Visit Data

Gary Sneddon[a,*], Wasimul Bari[b] and M. Tariqul Hasan[c]

[a]*Department of Mathematics and Computer Sciences, Mount Saint Vincent University, Halifax, Nova Scotia, Canada*

[b]*Department of Statistics, Biostatistics and Informatics, University of Dhaka, Dhaka, Bangladesh*

[c]*Department of Mathematics and Statistics, University of New Brunswick, Fredericton, NB, Canada*

**Abstract:** Time series data with excessive zeros frequently occur in medical and health studies. To analyze time series count data without excessive zeros, observation-driven Poisson regression models are commonly used in the literature. As handling excessive zeros in count data is not straightforward, observation-driven models are rarely used to analyze time series count data with excessive zeros. In this paper an observation-driven zero-inflated Poisson (ZIP) model for time series count data is proposed. This approach can accommodate an autoregressive serial dependence structure which commonly appears in time series. The estimation of the model parameters by using the quasi-likelihood estimating equation approach is discussed. To estimate the correlation parameters of the dependence structure, a moment approach is used. The proposed methodology is illustrated by applying it to a data set of daily emergency room visits due to bronchitis.

**Keywords:** Autocorrelation structure, non-stationary, observation-driven model, quasi-likelihood, zero-inflated Poisson.

## 1. INTRODUCTION

### 1.1. Motivating Example

In the past number of years there have been increasing concerns with how the environment affects our health. In particular, the role that air quality, specifically air pollution levels, plays in affecting the general health of the population has become more important. Given these concerns, it is important to develop statistical models that can be used to investigate the relationship between air quality and measurements of a person's health. This work is motivated by data collected in Prince George, British Columbia on daily emergency room visits, air pollution and meteorological variables [1]. The health measurement of interest is the daily number of emergency room visits that are due to bronchitis. The air pollution variables are sulphur (total reduced sulphur compounds, in parts per billion) and particulates (total suspended particulates, in $\mu g / m^3$). The meteorological measurements used are average daily temperature, maximum daily relative humidity and minimum daily relative humidity. The number of emergency room visits used in the analysis were recorded daily from April 1, 1984 to March 31, 1986.

*Address correspondence to this author at the Department of Mathematics and Computer Sciences, Mount Saint Vincent University, Halifax, Nova Scotia, Canada; Tel: 1-902-457-6261; Fax: 1-902-457-6656;
E-mail: gary.sneddon@msvu.ca

### 1.2. Background

The data on emergency room visits has been studied previously. Jorgensen *et al.* [1] assumed the daily counts of emergency room visits to be a Poisson process driven by a latent Markov process. Knight *et al.* [2] used a log-linear model, assuming the counts were independent over time. Hasan *et al.* [3] developed a multilevel, parameter-driven zero-inflated Poisson (ZIP) mixed model to account for excess zeros in the response.

There are a number of issues that need to be considered with this data. It is a time series of over 700 observations, so the presence of some serial correlation is to be expected. This is because the number of emergency room visits on a particular day is likely to be related to the number of visits on the previous day. However, a standard ARIMA (autoregressive integrated moving average) model is inappropriate because the responses are counts. A Poisson (log-linear) model that can incorporate correlated responses would be more appropriate. The data set also has another issue that needs to be addressed. A large number of responses (about 28%) are zero. In other words, on about 28% of days there were no emergency room visits that were due to bronchitis. This pattern, along with the positive skew in the distribution, is seen in Figure **1**. We need to develop a model that accounts for these excess zeros as well.

E-ISSN: 1929-6029/13 © 2013 Lifescience Global

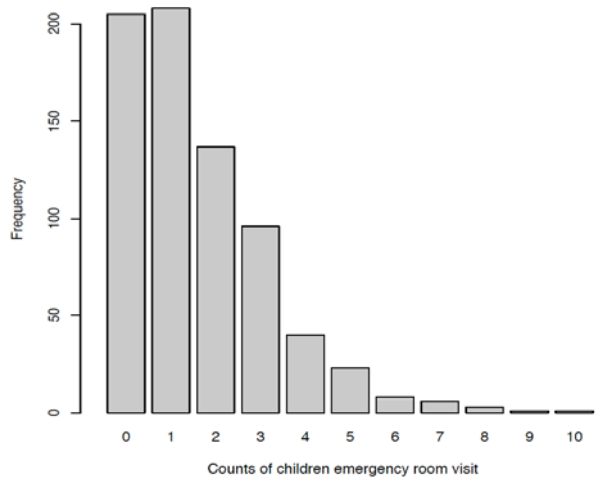


**Figure 1:** Bar plot of counts for children emergency room visit data due to bronchitis.

Lambert [4] developed zero-inflated Poisson (ZIP) models to deal with excess zeros in independent count responses. The ZIP model can be thought of as a mixture of a Poisson and a degenerate component putting all of its mass at zero [4]. Ngatchou-Wandji and Paris [5] discuss a variety of applications of ZIP models. Others have worked on extending the ZIP model to allow for correlated observations [3, 6-8], using random effect-based models. Hasan and Sneddon [9] developed an observation-driven ZIP model for longitudinal count data with excessive zeros.

### 1.3. Proposed Method

In this paper we modify the work in [9] and develop a non-stationary, observation-driven ZIP model for a time series of counts with excessive zeros. The work in [9] addressed the issue of modelling longitudinal count data with excessive zeros, whereas this paper addresses the issue of modelling time series count data with excessive zeros. The proposed approach will be used for the analysis of the daily emergency room visits described previously. This observation-driven model will be based on binomial thinning. The advantage of the obervation-driven model is that the correlation structure will have a very similar form to that of an autoregressive of order 1 (AR(1)) type model in a traditional time series. We introduce the observation-driven model in Section 2. We discuss a quasi-likelihood approach for estimating the regression parameters and a moment estimator of the correlation parameter in Section 3. The analysis of the daily counts of emergency room visits data is presented in Section 4. The performance of the proposed model is examined through a simulation study and the results are presented in Section 5, with a concluding discussion in Section 6.

## 2. OBSERVATION-DRIVEN MODEL FOR COUNT DATA WITH EXCESSIVE ZEROS

### 2.1. The Model

In this section we present a non-stationary observation-driven model for count data with excessive zeros. Let $Y_t$ denote the non-stationary time series count response recorded at the $t$ th ($t=1,2,\ldots,T$) time point. In the proposed observation-driven model, the response at the $t$ th time point depends on the responses of the previous $1,2,\ldots,(t-1)$ time points which can be constructed based on the following four assumptions:

*Assumption 1*: Let $U_t$, for $t=1,2,\ldots,T$, independently follows a Poisson distribution with parameter $(\lambda_t-\rho\alpha\lambda_{(t-1)})$. That is, the mean and variance of $U_t$ can be expressed as

$$E(U_t)=Var(U_t)=(\lambda_t-\rho\alpha\lambda_{(t-1)}). \tag{1}$$

In (1), $\lambda_t=\exp(X_t'\beta)$ with the vector of covariates $X_t'=(1,X_{t1},X_{t2},\ldots,X_{tn})$ and the vector of regression parameters $\beta=(\beta_0,\beta_1,\beta_2,\ldots,\beta_n)'$. Also in (1), $\alpha$ and $\rho$ are the probability and the correlation parameters respectively, which satisfy the range restriction of $0\le(\alpha,\rho)\le 1$. As the mean and variance of the Poisson random variable have to be positive i.e. $(\lambda_t-\rho\alpha\lambda_{(t-1)})\ge 0$, it then implies that $\rho<\dfrac{\lambda_t}{\alpha\lambda_{(t-1)}}$. Thus the correlation parameter $\rho$ must satisfy the range restriction $0<\rho<\min\left(\dfrac{\lambda_1}{\alpha\lambda_0},\dfrac{\lambda_2}{\alpha\lambda_1},\ldots,\dfrac{\lambda_T}{\alpha\lambda_{(T-1)}},1\right)$.

*Assumption 2*: Let $Y_t$ be the count response recorded at the $t$ th time point ($t=1,2,\ldots,T$). The distribution of the zero inflated count response $Y_t$ can be developed through the binomial thinning operation. To be specific, for a given $Y_{(t-1)}$, $\rho\otimes Y_{(t-1)}$ is the sum of $Y_{(t-1)}$ binary observations, where each observation is generated with probability $\rho$. We write this as

$$\rho\otimes Y_{(t-1)}=\sum_{j=1}^{Y_{(t-1)}}b_j(\rho)=Z_{(t-1)},\qquad say \tag{2}$$

with $Pr[b_j(\rho)=1]=\rho$ and $Pr[b_j(\rho)=0]=1-\rho$. This implies that the conditional distribution of $Z_{(t-1)}$ given $Y_{(t-1)}$ follows a binomial distribution with parameters $Y_{(t-1)}$ and $\rho$. Then the conditional expectation and variance of $Z_{(t-1)}$ given $Y_{(t-1)}$ can be expressed as

$E\left(Z_{(t-1)} \mid Y_{(t-1)}\right) = \rho Y_{(t-1)}$ *and*
$Var\left(Z_{(t-1)} \mid Y_{(t-1)}\right) = \rho(1-\rho)Y_{(t-1)}$.

Then the unconditional mean of $Z_{(t-1)}$ can be expressed as

$$E\left(Z_{(t-1)}\right) = E[E\left(Z_{(t-1)} \mid Y_{(t-1)}\right)] = \rho\alpha\lambda_{(t-1)}.$$

Similarly the unconditional variance of $Z_{(t-1)}$ can be expressed as

$$Var\left(Z_{(t-1)}\right) = Var[E\left(Z_{(t-1)} \mid Y_{(t-1)}\right)] + E[Var\left(Z_{(t-1)} \mid Y_{(t-1)}\right)]$$
$$= \rho\alpha\lambda_{(t-1)}.$$

Thus it is easy to show that the unconditional distribution of $Z_{(t-1)}$ is a Poisson random variable with parameter $\rho\alpha\lambda_{(t-1)}$.

*Assumption 3*: The random variables $Z_{(t-1)}$ for the $(t-1)$ th and $U_t$ for the $t$ th time points are independent of one another and follow Poisson distributions with parameters $\rho\alpha\lambda_{(t-1)}$ and $(\lambda_t - \rho\alpha\lambda_{(t-1)})$, respectively. After some algebra it can be shown that the sum of these two random variables i.e. $Y_t^* = Z_{(t-1)} + U_t$ follows a Poisson distribution with parameter $\lambda_t$.

*Assumption 4*: The non-stationary response $Y_t$ recorded at the $t$ th ($t = 1,2,\ldots,T$) time point has the following form:

$$Y_t = \begin{cases} 0 & with\ probability\ 1-\alpha \\ Y_t^* & with\ probability\ \alpha \end{cases}. \quad (3)$$

After some algebra it can be shown that $Y_t$ follows a Poisson distribution with parameter $\alpha\lambda_t$. Thus the unconditional mean and variance can be expressed as $E(Y_t) = \alpha\lambda_t$ and $var(Y_t) = \alpha\lambda_t$, respectively. The mean of $E(Y_t) = \alpha\lambda_t$ is smaller than the mean of $E(Y_t^*) = \lambda_t$ because of the mixture of the excess zeros in the time series count responses. The mixture of zero and Poisson components here is modeled through the observation driven technique, in accordance with ZIP models.

### 2.2. Variance-Covariance Structure

We calculate the variance-covariance structure for the proposed model by induction. As $Y_t$ follows a Poisson distribution with parameter $\alpha\lambda_t$, the variance of $Y_t$ for $t = 1,2,\ldots,T$ can be expressed as

$Var(Y_t) = \alpha\lambda_t$.

To calculate the covariance between lag-$l$ ($l = |t - t'|$) apart observations such as $Y_t$ and $Y_{t'}$, first we calculate the covariance between lag-1 apart observations $Y_t$ and $Y_{(t-1)}$ under the proposed model. To do this we write

$$Cov(Y_t, Y_{(t-1)}) = E(Y_t Y_{(t-1)}) - E(Y_t)E(Y_{(t-1)}). \quad (4)$$

In (4), $E(Y_t Y_{(t-1)})$ can be expressed as $E(Y_t Y_{(t-1)}) = EE\left(Y_t Y_{(t-1)} \mid Y_{(t-1)}\right)$, which after some lengthy algebra can be simplified as

$$E(Y_t Y_{(t-1)}) = \rho\alpha^2\lambda_{(t-1)} + \alpha^2\lambda_t\lambda_{(t-1)}.$$

This implies that $Cov(Y_t, Y_{(t-1)})$ in (4) is

$$Cov(Y_t, Y_{(t-1)}) = \rho\alpha^2\lambda_{(t-1)}. \quad (5)$$

Similarly we can calculate the covariance for lag-2 apart observations $Y_t$ and $Y_{(t-2)}$ as

$$Cov(Y_t, Y_{(t-2)}) = \rho^2\alpha^3\lambda_{(t-2)}. \quad (6)$$

Consequently, following (5) and (6), we can write the lag-$l$ apart covariance as

$$Cov(Y_t, Y_{(t-l)}) = \rho^l\alpha^{l+1}\lambda_{(t-l)}. \quad (7)$$

Let $C(\phi)$ denote the correlation matrix of the time-series response vector $= (Y_1, Y_2, \ldots, Y_T)$, which can be expressed as

$$C(\phi) = \begin{bmatrix} 1 & \phi_{(1,2)} & \phi_{(1,3)} & \cdots & \phi_{(1,T)} \\ \phi_{(2,1)} & 1 & \phi_{(2,3)} & \cdots & \phi_{(2,T)} \\ \phi_{(3,1)} & \phi_{(3,2)} & 1 & \cdots & \phi_{(3,T)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \phi_{(T,1)} & \phi_{(T,2)} & \phi_{(T,3)} & \cdots & 1 \end{bmatrix}. \quad (8)$$

where $\phi_{(t,t')}$ represents the correlation between lag-$l$ ($l = |t - t'|$) apart responses $Y_t$ and $Y_{t'}$. After some simple algebra, $\phi_{(t,t')}$ can be simplified as

$$\phi_{(t,t')} = Corr(Y_t, Y_{t'}) = \rho^{|t-t'|}\alpha^{|t-t'|}\sqrt{\frac{\lambda_{(t-|t-t'|)}}{\lambda_t}}. \quad (9)$$

The correlation structure presented in (9) can be considered as autoregressive of order 1 type

correlation structure for the non-stationary count responses as the correlation decays with the increase of the lag. Note that if $\lambda_1 = \lambda_2 = \ldots = \lambda_T = \lambda$, then the lag-$l$ ($l = |t - t'|$) apart correlation structure in (9) can be simplified as

$$\phi_{(t,t')} = Corr(Y_t, Y_{t'}) = \rho^{|t-t'|}\alpha^{|t-t'|}, \quad (10)$$

which is an autoregressive of order 1 correlation structure for the stationary count responses with excessive zeros.

## 3. PARAMETER ESTIMATION

In this section we discuss the estimation procedures to estimate the regression parameter $\beta$ and the correlation parameter $\rho$. In Section 3.1, following [10], we introduce a quasi-likelihood (QL) approach to estimate $\beta$. Note that the QL approach involves the unknown correlation parameter $\rho$ and probability parameter $\alpha$ which need to be estimated. In Section 3.2 we use a moment approach to estimate $\rho$. The logistic regression based estimation approach of the probability parameter $\alpha$ using the covariate information is discussed in Section 3.3.

### 3.1. Estimation of Regression Parameters

Following [10], the QL estimating equation can be written as

$$g(\beta) = \frac{\partial \mu}{\partial \beta}\Sigma^{-1}(Y - \mu) = 0. \quad (11)$$

In (11), $Y = (Y_1, \ldots, Y_t, \ldots, Y_T)'$ is a $T \times 1$ response vector with mean vector $\mu = (\mu_1, \ldots, \mu_t, \ldots, \mu_T)'$ where $\mu_t = E(Y_t \mid X_t') = \alpha\lambda_t$ and $\frac{\partial \mu'}{\partial \beta} = \left[\frac{\partial \mu_1}{\partial \beta}', \ldots, \frac{\partial \mu_T}{\partial \beta}'\right]$. Also in (11), $\Sigma = A^{1/2}C(\rho)A^{1/2}$ where $C(\rho)$ is the correlation matrix defined in (8) and $A$ is the $T \times T$ diagonal matrix given by $A = diag[\upsilon_1, \ldots, \upsilon_t, \ldots, \upsilon_T]$, where $\upsilon_t = var(Y_t \mid X_t') = \alpha\lambda_t$. For known $\rho$ and $\alpha$, the solution of $\beta$ from (11) is achieved by using the Gauss-Newton iteration procedure, which can be written as

$$\beta(r+1) = \beta(r) + \left[\frac{\partial \mu'}{\partial \beta}\Sigma^{-1}\frac{\partial \mu}{\partial \beta}\right]_{(r)}^{-1}\left[\frac{\partial \mu'}{\partial \beta}\Sigma^{-1}(Y - \mu)\right]_{(r)}, \quad (12)$$

where $[.]_{(r)}$ denotes that the expression within the square brackets is evaluated at $\beta = \hat{\beta}(r)$, the values of $\beta$ at the $rth$ iteration. Let $\hat{\beta}$ be the final estimator of $\beta$. Following [10], it can be shown that, under some mild regularity conditions, $T^{\frac{1}{2}}(\hat{\beta} - \beta)$ is asymptotically normal with mean vector $0$ and covariance matrix $V$ given by

$$V = \lim_{T\to\infty} T\left[\frac{\partial \mu'}{\partial \beta}\Sigma^{-1}\frac{\partial \mu}{\partial \beta}\right]^{-1}. \quad (13)$$

Note that the solution of (11) for $\beta$ computed by (12) is a consistent estimate for $\beta$. This is because for known $\rho$ and $\alpha$, the estimating function $g(\beta)$ is unbiased for zero.

### 3.2. Estimation of Correlation Parameter

As mentioned before, the estimation of $\beta$ requires the correlation matrix $C(\rho)$. Even though the structure of the correlation matrix $C(\rho)$ is known under the proposed model, the correlation parameter $\rho$ involved in $C(\rho)$ is however still unknown. We need to estimate $\rho$ consistently in order to obtain the consistent and efficient estimate for the regression parameter $\beta$. To do that we assume that the probability parameter $\alpha$ is known in this section. Under the proposed model (3), the correlation structure of $Y_t$ and $Y_{(t-l)}$ have the form derived in (9). In order to know all lag correlations under the proposed model, it is sufficient to know the estimate of $\rho$. Consequently, to develop a moment equation for $Y_t$ $(t = 1, 2, \ldots, T)$, we consider a statistic $\Psi_1$, as a function of the sample autocovariances of the response variable given by

$$\Psi_1 = \sum_{t=2}^{T}\left\{\tilde{Y}_t\tilde{Y}_{(t-1)}\right\}/(T-1), \quad (14)$$

where $\tilde{Y}_t = \frac{Y_t - E(Y_t)}{[var(Y_t)]^{1/2}}$. Now to develop the moment equation we solve $\Psi_1$ with $E(\Psi_1)$, where

$$E(\Psi_1) = \varphi_1(\rho) = \sum_{t=2}^{T}\left\{\rho\alpha\frac{\sqrt{\lambda_{(t-1)}}}{\sqrt{\lambda_t}}\right\}/(T-1). \quad (15)$$

For the estimate of $\rho$ *i.e.* $\hat{\rho}$, we consider the moment equation

$$\Psi_1 - E(\Psi_1) = \Psi_1 - \varphi_1(\hat{\rho}) = 0 \quad (16)$$

and solve for $\rho$ by using the iterative equation

$$\rho(r+1) = \rho(r) - \left[\frac{\partial}{\partial\rho}\{\varphi_1(\rho) - \Psi_1\}\right]_{(r)}^{-1}[\varphi_1(\rho) - \Psi_1]_{(r)}, \quad (17)$$

where

$$\frac{\partial}{\partial\rho}\{\varphi_1(\rho)-\Psi_1\}=\sum_{t=2}^{T}\left\{\alpha\frac{\sqrt{\lambda_{(t-1)}}}{\sqrt{\lambda_t}}\right\}/(T-1),$$

and $[.]_{(r)}$ denotes that the expression in [ ] is evaluated at $\rho=\hat{\rho}(r)$. Once the correlation parameter $\rho$ is estimated using the appropriate moment estimating equation for known correlation structures, one can use this estimate of $\rho$ in the QL estimating equation (12), which also involves the estimation of the probability parameter $\alpha$. In Section 3.3 we present the estimation technique for estimating the probability parameter.

### 3.3. Estimation of Probability Parameter

In the proposed observation-driven model, the unknown probability parameter $\alpha$ can be considered as a nuisance parameter. So a consistent estimate of the probability parameter $\alpha$ should be enough to obtain estimates of the regression and correlation parameters. To do that we follow [9] and estimate $\alpha$ from the data using logistic regression. Note that logistic regression is commonly used in the ZIP model to estimate the probability parameters consistently [4-8]. To do this we convert the count data to binary data as

$$Y_t^{**}=\begin{cases}0 & if\ Y_t=0\\ 1 & if\ Y_t>0\end{cases}.$$

Then the probability parameter $\alpha$ can be estimated as

$$\alpha=E(Y_t^{**})=\frac{1}{T}\sum_{t=1}^{T}\frac{\exp\left(X_t^{'}\beta^{**}\right)}{1+\exp\left(X_t^{'}\beta^{**}\right)},$$

using the quasi likelihood estimate of $\beta^{**}$. We then use the maximum likelihood estimate of $\alpha$ to update the estimates of the regression parameters iteratively.

## 4. DATA ANALYSIS

In this section, we illustrate our proposed observation-driven method to the daily counts of emergency room visits due to bronchitis from Prince George, British Columbia over a two year period [1]. As there may be no visit to a single hospital due to bronchitis in many days, we may observe excessive zero counts as compared to the nominal counts of zeros. To check the presence of extra zeros in the data set, we first conduct exploratory data analysis which shows that about 28% of days have zero responses. We then calculate the nominal percentage of zeros from a Poisson distribution using its sample average as the mean parameter which is 19%. Therefore the observed percentage of zeros is nearly 50% higher than the nominal percentage. So we need to accommodate these excessive zeros to analyze this data set appropriately.

In our analysis, the daily counts of emergency room visits by residents to the single hospital in Prince George, British Columbia due to bronchitis have been considered responses. To examine the various covariates' effects on the daily counts of ER visits due to bronchitis, we consider temperature, maximum relative humidity and minimum relative humidity as covariates, which were collected at the Prince George airport. The measurements refer to the daily average readings in degrees Celsius, largest reading and smallest reading of humidity, respectively. We considered two air quality variables: TRS (total reduced sulphur) and TSP (total suspended particulates) as in [1]. These two air quality variables refer to the daily average reading collected from the six stations at Prince George. To consider effects of the air quality of previous days, we considered the lag 0, 1 and 2 of log(TRS) and log of lag TSP [1].

We also considered the effect of different days of the week in the initial analysis and found out that only weekends are significantly different than weekdays. Therefore we incorporated an indicator variable (Day) which is 1 if the response is recorded on the weekend or 0 otherwise. The occurrence of bronchitis can also be affected by periodical variation as more cases may occur during flu season. To capture the periodical variation in the data we considered cosine and sine terms of month and season. Therefore the covariates in our final model are temperature, sum of log humidities (Sum Hum.), difference of log humidities (Diff. Hum.), lag 0 of log TSR (Log TRS.L0), lag 1 of log TSR (Log TRS.L1), lag 2 of log TSR (Log TRS.L2), log of lag TSP (Log TSP.L), day, cosine of month, sine of month, cosine of season and sine of season.

In this paper, our scientific interest is to assess the effects of air pollution on the emergency room visits while accounting for the large number of zero observations as well as serial correlation which commonly occurs in time series data. Let $Y_t$ represent the observed number of daily emergency room visits at the $t$ th time point, which can be analyzed using our model (3) with the Poisson mean parameter being specified as

$$\lambda_t = \exp\left(\begin{array}{l}\beta_0 + \beta_1\ Temperature + \beta_2\ Sum\ Hum. + \\ \beta_3\ Diff.Hum. + \beta_4\ Log\ TRS.L0 \\ +\beta_5\ Log\ TRS.L1 + \beta_6\ Log\ TRS.L2 + \beta_7\ Log\ TRS.L + \\ \beta_8\ Day + \beta_9\ Cosine\ of\ Month \\ +\beta_{10}\ Sine\ of\ Month + \beta_{11}\ Cosine\ of\ Season + \\ \beta_{12}\ Sine\ of\ Season\end{array}\right), \quad (18)$$

where $t = 1,\ldots,728$. Our data analysis results are presented in Table **1**.

Our analysis shows that temperature has a negative significant effect. That is, if temperature increases then number of emergency room visits decreases. The covariates sum of log humidities and difference of log humidities also have positive significant effect, which indicate that the increase of the daily variation in humidities increases the emergency room visits. Jørgensen *et al.* [1] and Hasan *et al.* [3] found that the sum of log humidities and difference of log humidities have insignificant effects. Other covariates such as lag 0, 1 and 2 of log(TRS) appear to be insignificant as in [1]. In their data analysis, Jørgensen *et al.* [1] found that log(TSP.L) has a negative significant effect and the number of emergency room visits during weekends as well as Mondays and Wednesdays are significantly higher as compared to the other days of the week. After accounting for excessive zeros in our analysis, we found Log TSP.L has an insignificant effect and the number of emergency room visits is significantly higher on the weekends as compared to the weekdays, which is similar to the results presented in [3]. One of the reasons for this is probably due to the unavailability of family doctors during weekends.

To capture the seasonal variation of the responses we included the cosine and sine terms of months and seasons. Our results show that cosine and sine term of months and sine of season have negative significant effects. To understand the overall effect of the periodical variation, we plot the overall effect against the day in Figure **2** where the values of $\hat{\beta}_9, \hat{\beta}_{10}, \hat{\beta}_{11}$ and $\hat{\beta}_{12}$ are given in Table **1**:

$$overall\ effects = \hat{\beta}_9\ Cosine\ of\ Month + \hat{\beta}_{10}\ Sine\ of\ Month + \hat{\beta}_{11}\ Cosine\ of\ Season + \hat{\beta}_{12}\ Sine\ of\ Season$$

The overall periodic pattern shown in Figure **2** would repeat annually over these two years, so the plot is drawn for only for one year from January 1 to December 31. There appears to be a quadratic shape in the periodic pattern of daily emergency room visits due to bronchitis over a year with its peak in fall and winter. The highest peak in the months is in winter and the lowest peak is in the summer months. This is probably beacuse fall and winter are considered to be flu season in British Columbia. The dotted line indicates the overall effect and the solid line indicates the nonparametric fit of the overall effect (using the function supsmu in R).

**Table 1: Estimates of Parameters with Standard Errors and *p*-Values for Daily Emergency Room Visit Data**

| Covariates | Estimate | St. Error | *p*-value |
|---|---|---|---|
| Intercept | 0.2512 | 0.6788 | 0.7113 |
| Temperature | -0.0111 | 0.0026 | 0.0000 |
| Sum Hum. | 0.2852 | 0.1025 | 0.0054 |
| Diff. Hum. | 0.1497 | 0.0717 | 0.0367 |
| Log TRS.L0 | 0.0156 | 0.0199 | 0.4340 |
| Log TRS.L1 | -0.0425 | 0.0211 | 0.0538 |
| Log TRS.L2 | 0.0177 | 0.0194 | 0.3609 |
| Log TSP.L | -0.0438 | 0.0473 | 0.3539 |
| Day | 0.6723 | 0.0319 | 0.0000 |
| Cosine of Month | 0.4755 | 0.0763 | 0.0236 |
| Sine of Month | -0.0500 | 0.0641 | 0.0000 |
| Cosine of Season | -0.2442 | 0.0619 | 0.4352 |
| Sine of Season | -0.1394 | 0.0616 | 0.0000 |
| $\alpha$ | 0.7184 | | |
| $\rho$ | 0.5217 | | |

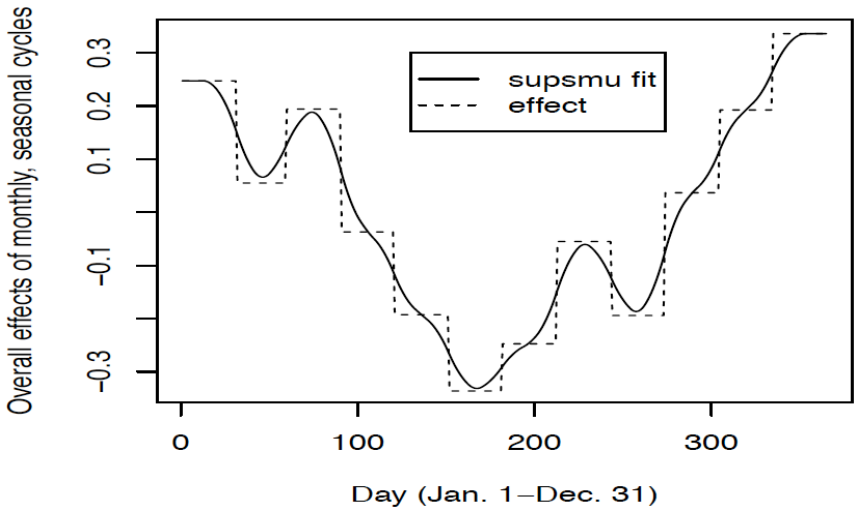


**Figure 2:** Overall periodical effect plot for children emergency room visit data due to bronchitis.

The estimates of $\alpha$ and $\rho$ are $0.7184$ and $0.5217$, respectively. The serial correlation of $0.5217$ indicates that the serial dependence of the daily counts between consecutive days is strong. This pattern is common as the temperature, flu activity and variation in humidity are similar on consecutive days.

## 5. SIMULATION STUDY

To examine the performance of the proposed methodology we conducted a simulation study. In the simulation study, we generate count responses similar to those in the emergency room dataset 500 times *via* simulation by using the proposed observation driven

**Table 2: Simulated Means, Simulated Standard Errors and Estimated Standard Errors Based on 500 Simulation**

| Covariates | True Values | SM[a] | SSE[b] | ESE[c] |
|---|---|---|---|---|
| $\beta_0$ | 0.2512 | 0.1797 | 0.5424 | 0.5615 |
| $\beta_1$ | -0.0111 | -0.0081 | 0.0039 | 0.0027 |
| $\beta_2$ | 0.2852 | 0.1986 | 0.0568 | 0.0792 |
| $\beta_3$ | 0.1497 | 0.1026 | 0.0341 | 0.0595 |
| $\beta_4$ | 0.0156 | 0.0083 | 0.0100 | 0.0170 |
| $\beta_5$ | -0.0425 | -0.0272 | 0.0137 | 0.0176 |
| $\beta_6$ | 0.0177 | 0.0155 | 0.0085 | 0.0157 |
| $\beta_7$ | -0.0438 | -0.0522 | 0.0262 | 0.0493 |
| $\beta_8$ | 0.6723 | 0.5322 | 0.0186 | 0.0261 |
| $\beta_9$ | 0.4755 | 0.4517 | 0.0336 | 0.0673 |
| $\beta_{10}$ | -0.0500 | -0.0438 | 0.0250 | 0.0729 |
| $\beta_{11}$ | -0.2442 | -0.2278 | 0.0319 | 0.0797 |
| $\beta_{12}$ | -0.1394 | -0.1291 | 0.0387 | 0.0793 |
| $\alpha$ | 0.7184 | 0.7166 | 0.0173 | |
| $\rho$ | 0.5217 | 0.5099 | 0.2320 | |

[a]Simulated means.
[b]Simulated standard errors.
[c]Estimated standard errors.

ZIP model specified in our example. We consider the covariates listed in Table **1** from the data example as the covariates for our simulation study. The true values of the regression parameters in the simulation study are taken from the estimates of the corresponding regression parameters listed in Table **1**. The true values of the probability and the correlation parameters also taken from the estimates of the corresponding parameters from our data example. They are listed in Table **2** as 'True value' $\beta_0$, $\beta_1$, …, $\beta_{12}$, $\alpha$ and $\rho$ respectively.

In each simulation, we generated 728 count responses based on the assumptions described in Section 2. We then analyze the simulated data set under the estimation technique presented in Section 3. The simulation results are presented in Table **2**. The simulated means for the regression parameters are appear to be very close to the true values. The average biases (the absolute difference between the true values and the simulated mean) for the regression parameter estimates over 500 simulations are all less than 0.05, with the exception of $\beta_0$ and $\beta_8$, which each have bias around 0.1. The simulated means of the probability and the correlation parameters estimates over 500 simulations are also very close to the corresponding true values, with average biases less than 0.02. The sample standard errors of the estimates over 500 simulations and the averages of the 500 estimated standard errors are termed simulated and estimated standard errors respectively. All estimated standard errors for the regression parameter estimates are similar to the simulated standard errors. This limited simulation study demonstrates that our approach is performing well for the analysis of time series count data with excessive zeros.

## 6. DISCUSSION

In this paper we have proposed an observation-driven zero inflated Poisson model for analyzing time series count data with excessive zeros. In this observation-driven model the regression coefficients can be interpreted as the proportional change in the marginal expectation of the response variable on the logarithm scale given a unit change in the regressor variables. This property makes the interpretation of the regression parameter easy to understand. Our choice of observation-driven model allows us to accommodate the excessive zeros in the time series count responses. Our proposed approach can also capture the serial dependence which is likely to be present in time series data. Although in this paper we assumed an AR(1)-type correlation structure due to its common occurrence in time series analysis, our model is easily modified for other serial dependence structures such as moving average of order 1 or exchangeable correlation structures. A computationally efficient estimation method for the regression parameters has been developed through the quasi-likelihood approach. Results based on our limited simulation study show that the proposed approach performs well. In our example, the large value of the estimate of serial correlation indicates the importance of capturing the serial dependence among the time series responses. Under the longitudinal count data with excessive zeros, it was shown in [9] that ignoring serial dependence among the responses may lead to inefficient estimates of the model parameters. This conclusion can be extended to the context of time series count responses with excessive zeros.

## ACKNOWLEDGMENTS

This research is partially supported by the Natural Sciences and Engineering Research Council of Canada and University Research Fund and Start-up fund, University of New Brunswick and Mount Saint Vincent University. The authors would like to thank the referees for their helpful comments which improved the original manuscript.

## REFERENCES

[1] Jørgensen B, Lundbye-Christensen S, Song XK, Sun L. A longitudinal study of emergency room visits and air pollution for Prince George, British Columbia. Statist Med 1996; 15: 823-36. http://dx.doi.org/10.1002/(SICI)1097-0258(19960415)15:7/9<823::AID-SIM252>3.0.CO;2-A

[2] Knight K, Leroux BG, Millar J, Perkau AJ. Air pollution and human health: a study based on data from Prince George, British Columbia, Technical report 1991.

[3] Hasan MT, Sneddon G, Ma R. Regression analysis of zero-inflated time-series counts: application to air pollution related emergency room visit data. J Appl Statist 2012; 39: 467-76. http://dx.doi.org/10.1080/02664763.2011.595778

[4] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 1992; 34: 1-14. http://dx.doi.org/10.2307/1269547

[5] Ngatchou-Wandji J, Paris C. On the zero-inflated count models with application to modelling annual trends in incidences of some occupational allergic diseases in France. J Data Sci 2011; 9: 639-59.

[6] Yau KKW, Lee AH, Carrivick PJW. Modeling zero-inflated count series with application to occupational health. Comp Methods Progr Biomed 2004; 74: 47-52. http://dx.doi.org/10.1016/S0169-2607(03)00070-1

[7] Hall DB. Zero-inflated Poisson and binomial regression with random effects: A case study. Biometrics 2000; 56: 1030-39. http://dx.doi.org/10.1111/j.0006-341X.2000.01030.x

[8] Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. Statist Model 2005; 5: 1-19. http://dx.doi.org/10.1191/1471082X05st084oa

[9] Hasan MT, Sneddon G. Zero-inflated poisson regression for longitudinal data. Communications in Statistics: Simulation and Computation 2009; 38: 638-53. http://dx.doi.org/10.1080/03610910802601332

[10] Zeger SL. A regression model for time series of counts. Biometrika 1988; 75: 621-29. http://dx.doi.org/10.1093/biomet/75.4.621

Received on 04-07-2013 Accepted on 24-07-2013 Published on 31-07-2013

http://dx.doi.org/10.6000/1929-6029.2013.02.03.7