involve
a journal of mathematics

Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities

Sadie Beckett, Joshua Jee, Thapelo Ncube, Sophia Pompilus, Quintel Washington, Anshuman Singh and Nabendu Pal

msp

2014 vol. 7, no. 6

INVOLVE 7:6 (2014)
dx.doi.org/10.2140/involve.2014.7.751

Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities

Sadie Beckett, Joshua Jee, Thapelo Ncube, Sophia Pompilus, Quintel Washington, Anshuman Singh and Nabendu Pal

(Communicated by Sat N. Gupta)

This work deals with estimation of parameters of a zero-inflated Poisson (ZIP) distribution as well as using it to model some natural calamities' data. First, we compare the maximum likelihood estimators (MLEs) and the method of moments estimators (MMEs) in terms of standardized bias (SBias) and standardized mean squared error (SMSE). We then proceed to show how datasets from some recent natural disasters can be modeled by the ZIP distribution.

1. Introduction

A random variable X following the usual Poisson distribution with parameter λ, Poi(λ), with the probability mass function

$$P(X=k)=\exp(-\lambda)\frac{\lambda^k}{k!},\quad k=0,1,2,3,\ldots \tag{1-1}$$

is widely used to model many naturally occurring events where X represents the "number of events per unit of time or space". Note that X takes only nonnegative integer values. However, the Poi(λ) distribution may not be useful (or it gives a bad fit) when X takes the value 0 with a high probability. In such a case a modified version of a regular Poi(λ) distribution known as the zero-inflated Poisson (ZIP) distribution becomes useful. The ZIP distribution with parameters π and λ, denoted by ZIP(π,λ), has the following probability mass function:

$$P(X=k)=\begin{cases}\pi+(1-\pi)\exp(-\lambda) & \text{if } k=0\\ (1-\pi)\exp(-\lambda)\lambda^k/k! & \text{if } k\in\{1,2,\ldots\},\end{cases} \tag{1-2}$$

where $0\le\pi\le 1$ and $\lambda\ge 0$.

MSC2010: primary 62F10; secondary 62F86, 62P12.
Keywords: method of moments estimation, maximum likelihood estimation, bias, mean squared error.

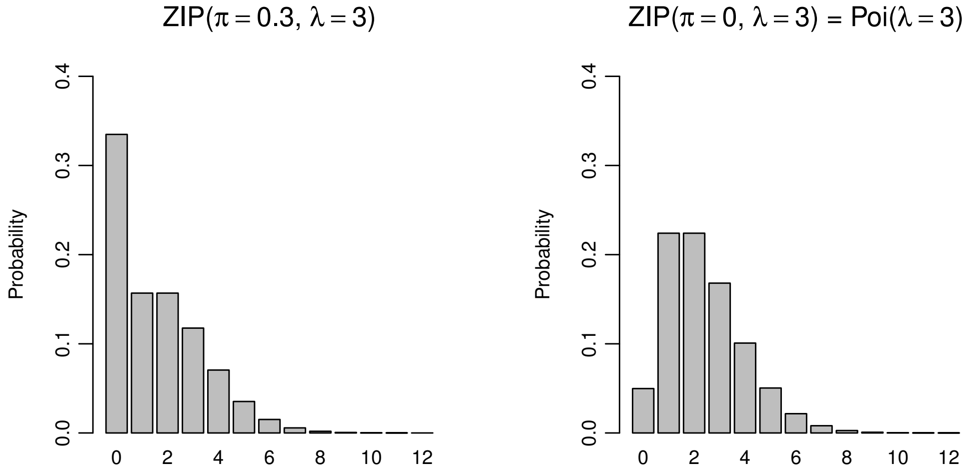

Figure 1. Probability distributions of ZIP and regular Poisson.

The parameter π gives the extra probability thrust at the value 0; when it vanishes, ZIP(π, λ) reduces to Poi(λ). Figure 1 shows visually the difference between these two distributions for selected values of π and λ.

The mean and variance of ZIP(π, λ) are

$$\begin{aligned} E(X) &= \lambda(1-\pi), \\ V(X) &= \lambda(1-\pi)(1+\lambda\pi). \end{aligned} \tag{1-3}$$

For example, for ZIP(0.3, 3), these characteristics are $E(X) = 3(1-0.3) = 2.1$ and $V(X) = 3(1-0.3)(1+3(0.3)) = 3.99$.

In the following, we provide a brief but comprehensive literature review to show how other researchers have used the ZIP distribution to model real-life data. Other important references can be found in these papers as well.

Lambert [1992] shows how a ZIP regression is better than a Poisson regression in fitting a data set with many zeros. The dataset she uses to compare these models is the number of manufacturing defects on wiring boards. Lambert concludes that ZIP regression is a straightforward model to interpret, and is convenient to use.

The decayed, missing and filled teeth (DMFT) index is used in dental epidemiology research to measure the dental health of individuals. The study [Böhning et al. 1999] used data from Brazilian school children to determine which processes were the most beneficial in preventing dental cavities. Böhning et. al state that the Poisson model often underestimates the dispersion of the data, which is why the ZIP is used instead. The ZIP model was used in this study to account for the number of children who had a DMFT of 0 (which represents good dental health). Researchers graphed the distribution of the DMFT values of the children before and after the preventive measures were implemented in their respective schools, in order to compare the results. Besides preventive measures, intervention effects based on the ZIP model were also discussed.

Böhning [1998] asserts that the simple Poisson distribution is oftentimes inappropriate for datasets due to the numerous zeros in the data. As an example, Böhning refers to a study done with 98 HIV-positive men that provides the number of urinary tract infections experienced by these men. When the data is seen graphically, we see a huge spike at zero. We also see that there is a lack of a good fit with the Poisson model, but a good one with the ZIP model. Thus, Böhning maintains that the ZIP is a better application when there is an inflation of zeros in the count data.

Ridout, Demetrio and Hinde [1998] argue that the Poisson model does not account for high occurrences of zeros in the dataset, and therefore a better model is needed, namely the ZIP. The ZIP distribution is a slight generalization of the Poisson model, but it gives a better fit for the extra zeros.

The research described in [Davidson 2012] relates to the recurrent colorectal adenomas and the usage of the ZIP distribution. Davidson mentions that though the Poisson distribution may be used for estimated recurrences of polyp prevention trials, the ZIP is the adequate model for dealing with an inflation of zeros. This inflation was due to the fact that a large number of patients did not have recurring adenomas after being observed and treated.

The rest of the paper is organized as follows. In Section 2 we discuss the two estimation techniques and the challenges we face in using them. Section 3 covers our comprehensive simulation study to compare the two estimation techniques in terms of standardized bias (SBias) and standardized mean squared error (SMSE). In Section 4 we present some data from natural calamities where the ZIP distribution appears to provide a better fit than the usual Poisson model.

2. Estimation of ZIP parameters

Assume that we have independent and identically distributed (iid) observations $X_1, X_2, \ldots, X_n$ from ZIP(π, λ). Our first objective is to estimate the model parameters π and λ. We are going to follow two estimation techniques, namely the Method of Moments Estimation (MME) and the Maximum Likelihood Estimation (MLE).

2.1. ***The MME estimators.*** Here we obtain the estimators by equating the first two sample moments with their corresponding theoretical expressions:

$$E(X) = (1-\pi)\lambda \approx \bar{X}, \tag{2-1}$$

$$V(X) = (1-\pi)\lambda(1+\pi\lambda) \approx s^2, \tag{2-2}$$

$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ being the sample average and $s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$ the sample variance.

By solving (2-1) and (2-2), the MMEs are found as

$$\hat{\lambda}_{MM} = \overline{X} + \frac{s^2}{\overline{X}} - 1 \tag{2-3}$$

and

$$\hat{\pi}_{MM} = \frac{s^2}{\overline{X}} - \frac{1}{\hat{\lambda}_{MM}} = \frac{s^2 - \overline{X}}{\overline{X}^2 + (s^2 - \overline{X})}. \tag{2-4}$$

However, it must be noted that the above estimators may have the undesirable property of being negative though the parameters are nonnegative. When $\overline{X} > s^2$, then $\hat{\pi}_{MM}$ can become negative, whereas the actual parameter π is always between 0 and 1. Therefore we are going to modify the MME by truncating $\hat{\pi}_{MM}$ at zero and $\hat{\lambda}_{MM}$ at $\overline{X}$ when $\overline{X} \geq s^2$. The resultant estimators are called *corrected MMEs* (CMMEs) and denoted by

$$\hat{\pi}^{c}_{MM} = \begin{cases} 0 & \text{if } \overline{X} \geq s^2, \\ \hat{\pi}_{MM} & \text{otherwise,} \end{cases}$$

and

$$\hat{\lambda}^{c}_{MM} = \begin{cases} \overline{X} & \text{if } \overline{X} \geq s^2, \\ \hat{\lambda}_{MM} & \text{otherwise.} \end{cases} \tag{2-5}$$

The above CMMEs make sense because under the ZIP model $V(X) > E(X)$ always (see (1-3)). Therefore, it is expected to have s^2 to be greater than $\overline{X}$. Hence, a corrective measure is taken when $\overline{X} \geq s^2$.

2.2. *The MLE estimators.* For *iid* observations $\widetilde{X} = (X_1, \ldots, X_n)$ from ZIP(π, λ), the likelihood function $L(\pi, \lambda | \widetilde{X})$ is defined as

$$L(\pi, \lambda | \widetilde{X}) = \prod_{i=1}^{n} P(X = X_i). \tag{2-6}$$

Define Y to be the number of X_i's taking the value 0. Then

$$L(\pi, \lambda | \widetilde{X}) = (\pi + (1-\pi)e^{-\lambda})^{Y} \prod_{\substack{i=1 \\ X_i \neq 0}}^{n} (1-\pi)e^{-\lambda} \frac{\lambda^{X_i}}{X_i!}, \tag{2-7}$$

so our log likelihood function, denoted by L_*, is

$$L_* = Y \ln(\pi + (1-\pi)e^{-\lambda}) + (n-Y)\ln(1-\pi) - (n-Y)\lambda + n\overline{X} \ln\lambda - \ln \prod_{i=1}^{n} X_i! \tag{2-8}$$

By taking the derivatives of L_* with respect to π and λ, and setting them equal to

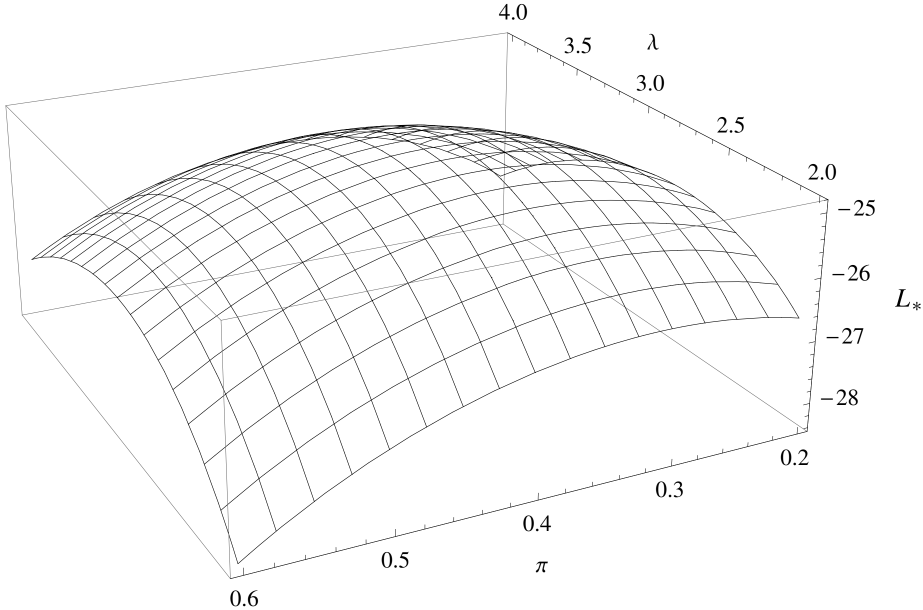

Figure 2. 3D diagram of L_* plotted against π and λ.

zero, we get the following system of equations:

$$\frac{n\overline{X}}{\lambda} = \frac{Y(1-\pi)e^{-\lambda}}{\pi + (1-\pi)e^{-\lambda}} + n - Y, \tag{2-9}$$

$$\frac{Y(1-\pi)(1-e^{-\lambda})}{\pi + (1-\pi)e^{-\lambda}} = n - Y. \tag{2-10}$$

The MLEs of π and λ, henceforth denoted by $\hat{\pi}_{ML}$ and $\hat{\lambda}_{ML}$ respectively, are the solutions of (2-9) and (2-10). Unlike the MMEs (or the CMMEs) we do not have explicit expressions for $\hat{\pi}_{ML}$ and $\hat{\lambda}_{ML}$.

As a demonstration, we draw a random sample of size $n = 15$ from ZIP(0.3, 3), giving us the following dataset: 0, 3, 3, 4, 0, 2, 0, 5, 0, 0, 0, 1, 3, 4, 3. The resultant log-likelihood function L_* is plotted against π and λ in Figure 2. The plot appears to have only one maximum and this has been our experience with all the replications of our simulation.

All of our computations are done using R. Widely used by statisticians, R is a free programming software for statistical computations and graphing purposes. It provides a plethora of both graphing and computational techniques, and is especially helpful for data analysis.

3. Comparison of two estimation techniques

In this section, we compare $\hat{\pi}^c_{MM}$ against $\hat{\pi}_{ML}$ and $\hat{\lambda}^c_{MM}$ against $\hat{\lambda}_{ML}$ in terms of *standardized bias* (SBias) and *standardized MSE* (SMSE), which are defined as

$$\text{SBias}(\hat{\theta}) = \frac{1}{\theta}\text{Bias}(\hat{\theta}) = \frac{1}{\theta}E(\hat{\theta}-\theta),$$

$$\text{SMSE}(\hat{\theta}) = \frac{1}{\theta^2}\text{MSE}(\hat{\theta}) = \frac{1}{\theta^2}E(\hat{\theta}-\theta)^2,$$

where $\hat{\theta}$ is a generic estimator for the parameter θ. (Note that θ can be either π or λ, and $\hat{\theta}$ can be the corresponding CMME or MLE.)

The usual Bias and MSE of an estimator $\hat{\theta}$ of θ are defined as $\text{Bias}(\hat{\theta}) = E(\hat{\theta}-\theta)$ and $\text{MSE}(\hat{\theta}) = E(\hat{\theta}-\theta)^2$. However, a true picture of the performance of $\hat{\theta}$ can be judged only through SBias and/or SMSE.

In the following, we provide the SBias and SMSE of each estimator for various values of n as well as (π, λ). For a fixed n and (π, λ), we generate $X_1, \ldots, X_n$ from the specified ZIP(π, λ) 10^5 times and for each replication we compute the parameter estimates. Then an expectation is approximated by taking the average of N replicated expectants. In other words, if in the j-th replication $(1 \le j \le N = 10^5)$ we estimate a parameter θ by $\hat{\theta}^{(j)}$, based on $\widetilde{X}^{(j)} = \left(X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)}\right)$, then the SBias and SMSE are obtained by the approximations

$$\text{SBias}(\hat{\theta}) = \frac{1}{N}\sum_{j=1}^{N}\frac{\hat{\theta}^{(j)}-\theta}{\theta}, \quad \text{SMSE}(\hat{\theta}) = \frac{1}{N}\sum_{j=1}^{N}\frac{(\hat{\theta}^{(j)}-\theta)^2}{\theta^2}.$$

Figures 3–10 provide the plots of SBias and SMSE of estimators of π and λ. These are some of the results of our comprehensive simulation study. Every third plot across the row in each figure shows the difference between the SBias (SMSE) of CMME and SBias (SMSE) of MLE. The simulated results have been presented for small ($n = 15$), moderate ($n = 30$) and large ($n = 50$) sample sizes. The findings of our simulation study have been summarized in the following remark.

Remark 3.1. For λ estimation, it has been observed that:

(i) $\hat{\lambda}_{ML}$ has mostly smaller |SBias| than that of $\hat{\lambda}^c_{MM}$ when plotted against λ.

(ii) When plotted against π, |SBias| of $\hat{\lambda}_{ML}$ is smaller than that of $\hat{\lambda}^c_{MM}$ for moderate to large sample sizes. For small n, $\hat{\lambda}_{ML}$ has worse |SBias| than that of $\hat{\lambda}^c_{MM}$ over a small region of π.

(iii) In terms of SMSE, $\hat{\lambda}_{ML}$ is much superior to $\hat{\lambda}^c_{MM}$ for all values of λ, except for small n when it is the other way around for small λ.

(iv) The MSE of λ estimators, when plotted against π, shows superiority of $\hat{\lambda}_{ML}$ over $\hat{\lambda}^c_{MM}$.

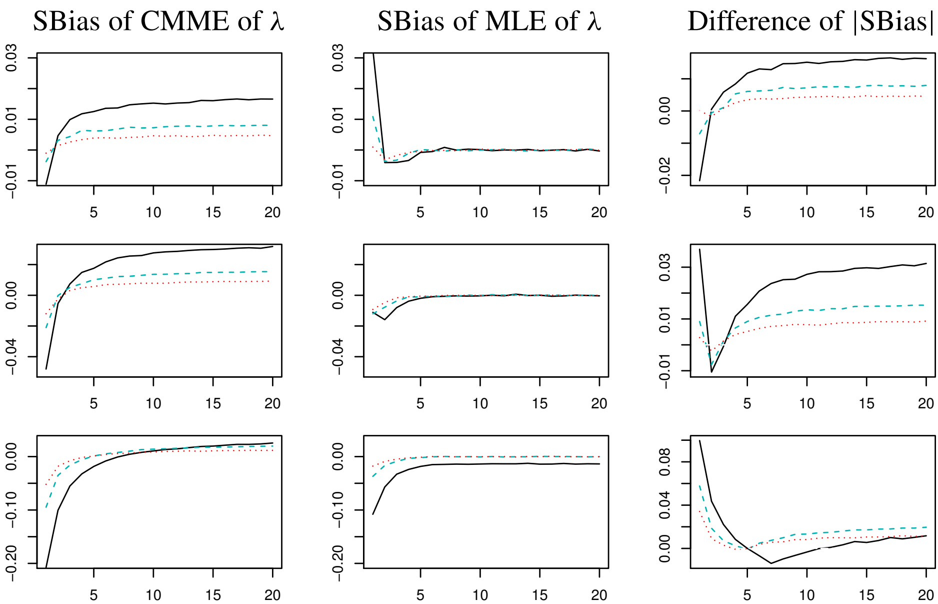

Figure 3. SBias study of λ estimators plotted against λ, for $\pi = 0.25$ (top row), $\pi = 0.5$ (middle), $\pi = 0.75$ (bottom)and for $n = 15$ (solid black line), 30 (dashed blue), 50 (dotted red).

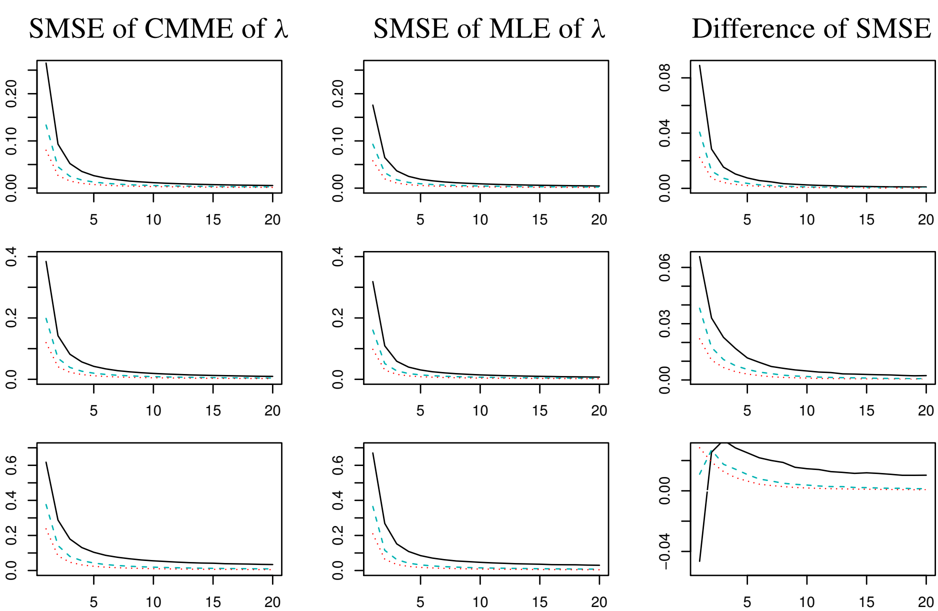

Figure 4. SMSE study of λ estimators plotted against λ, for $\pi = 0.25$ (top row), $\pi = 0.5$ (middle), $\pi = 0.75$ (bottom)and for $n = 15$ (solid black line), 30 (dashed blue), 50 (dotted red).

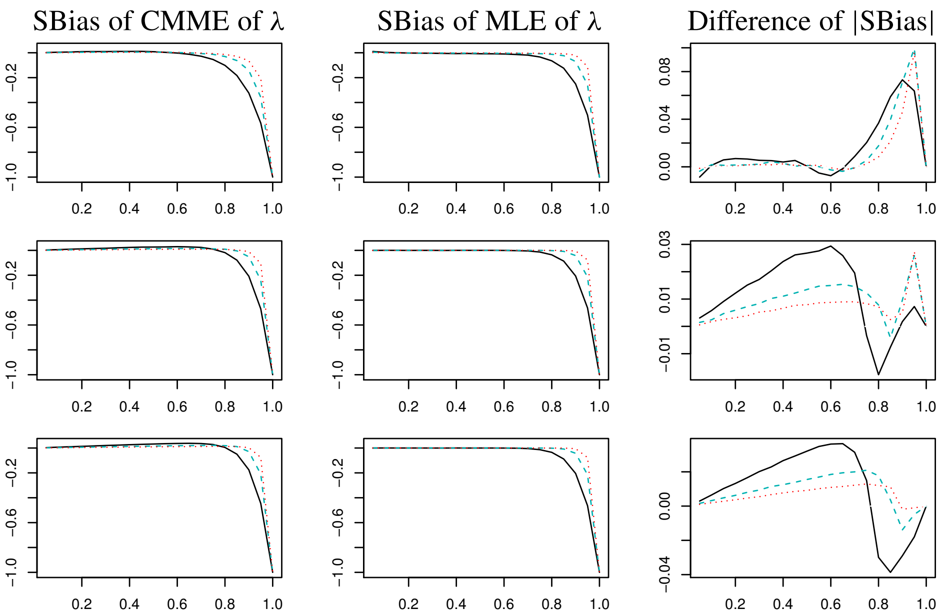

Figure 5. SBias study of λ estimators plotted against π, for $\lambda = 3$ (top row), $\lambda = 10$ (middle), $\lambda = 50$ (bottom)and for $n = 15$ (solid black line), 30 (dashed blue), 50 (dotted red).

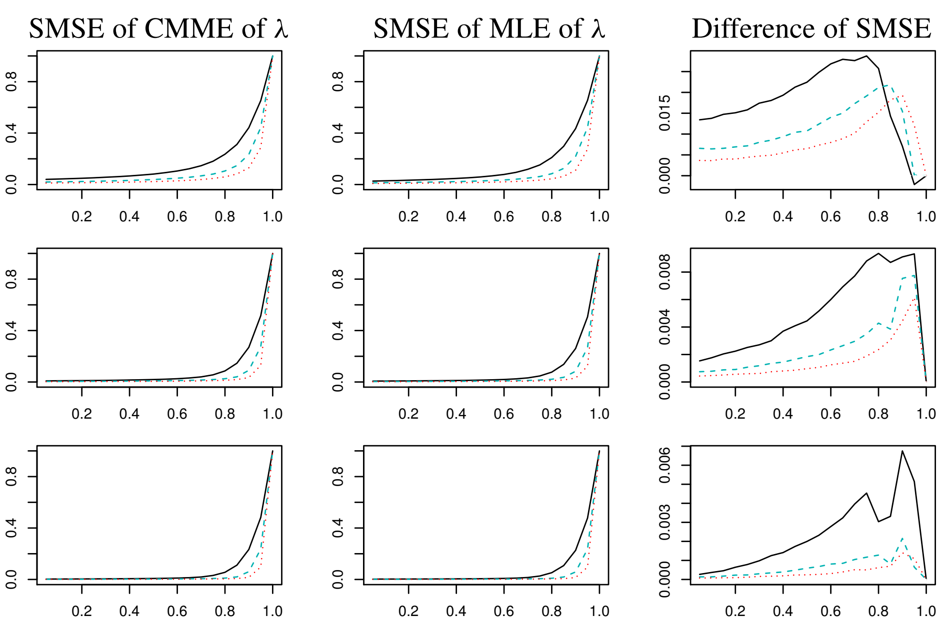

Figure 6. SMSE study of λ estimators plotted against π, for $\lambda = 3$ (top row), $\lambda = 10$ (middle), $\lambda = 50$ (bottom)and for $n = 15$ (solid black line), 30 (dashed blue), 50 (dotted red).

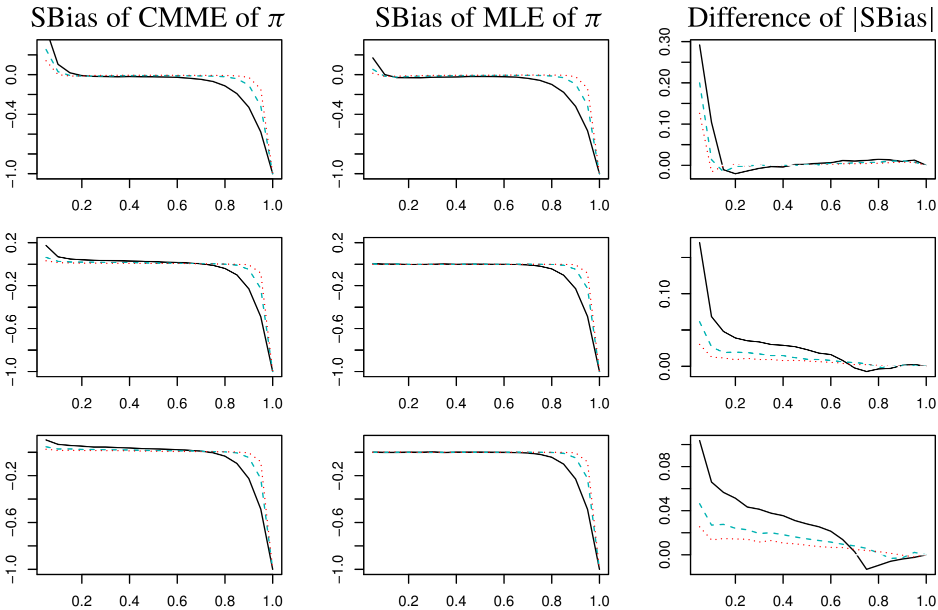

Figure 7. SBias study of π estimators plotted against π, for $\lambda = 3$ (top row), $\lambda = 10$ (middle), $\lambda = 50$ (bottom)and for $n = 15$ (solid black line), 30 (dashed blue), 50 (dotted red).

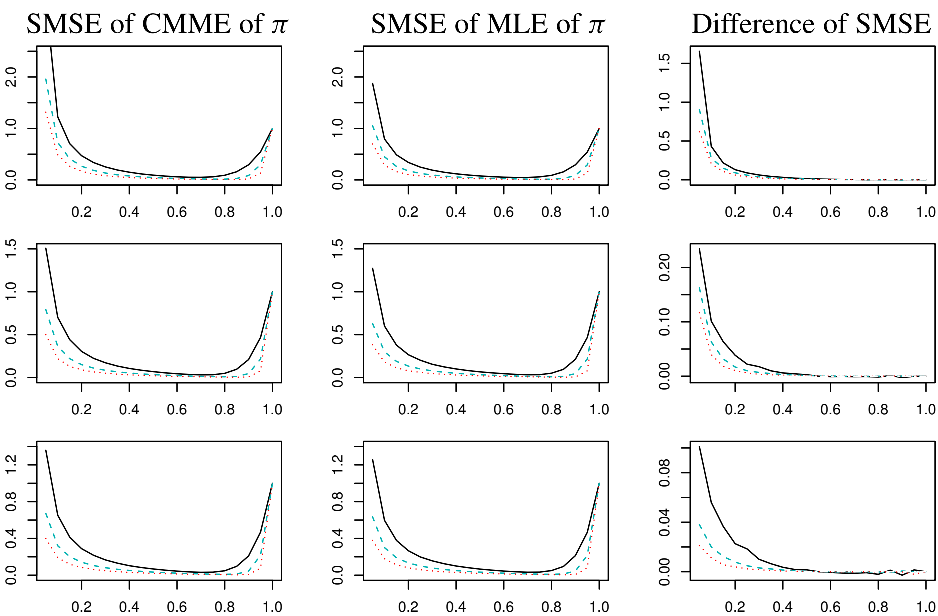

Figure 8. SMSE study of π estimators plotted against π, for $\lambda = 3$ (top row), $\lambda = 10$ (middle), $\lambda = 50$ (bottom)and for $n = 15$ (solid black line), 30 (dashed blue), 50 (dotted red).

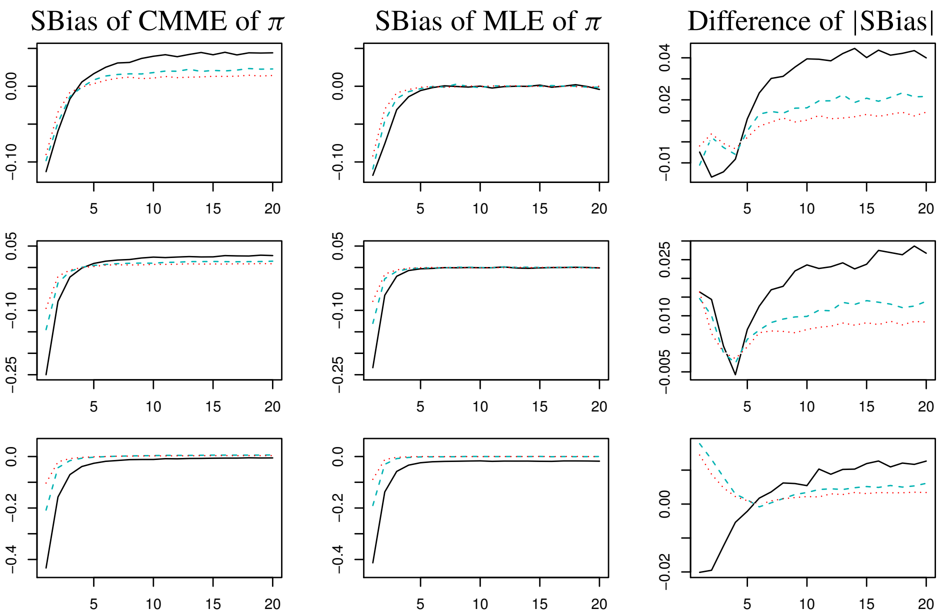

Figure 9. SBias study of π estimators plotted against λ, for $\pi = 0.25$ (top row), $\pi = 0.5$ (middle), $\pi = 0.75$ (bottom)and for $n = 15$ (solid black line), 30 (dashed blue), 50 (dotted red).

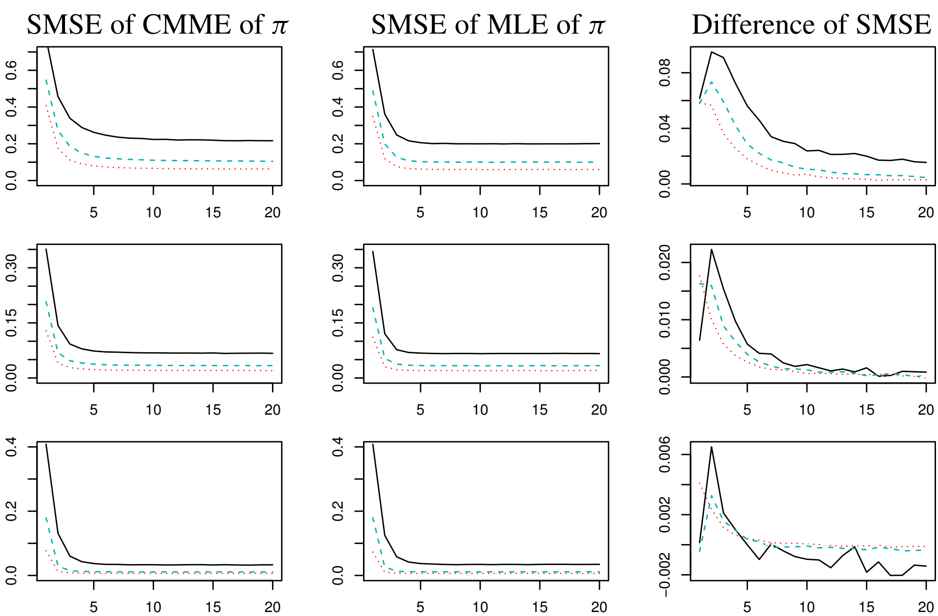

Figure 10. SMSE study of π estimators plotted against λ, for $\pi = 0.25$ (top row), $\pi = 0.5$ (middle), $\pi = 0.75$ (bottom)and for $n = 15$ (solid black line), 30 (dashed blue), 50 (dotted red).

Similar trends hold for π estimators as well:

(i) In terms of |SBias|, both $\hat{\pi}^c_{MM}$ and $\hat{\pi}_{ML}$ are very close for small λ ($\lambda = 3$) when plotted against π but, as λ increases, $\hat{\pi}_{ML}$ tends to perform better than $\hat{\pi}^c_{MM}$ for most of π values.

(ii) When |SBias| is plotted against λ, again $\hat{\pi}_{ML}$ tends to perform better than $\hat{\pi}^c_{MM}$ for most values of λ, especially for moderate to large sample sizes.

(iii) In terms of SMSE, except for small n, $\hat{\pi}_{ML}$ performs better than $\hat{\pi}^c_{MM}$ for most of the λ values.

(iv) When SMSE is plotted against π, $\hat{\pi}_{ML}$ appears to be superior to $\hat{\pi}^c_{MM}$ uniformly.

Based on our simulation study, the MLEs of π and λ appear to be superior estimators over their CMME counterparts, and therefore they are recommended for usage as done in Section 4 where ZIP is used to model data from natural calamities. The superiority of the MLEs has also been partially corroborated by Schwartz and Giles [2013], who observed that the MLEs exhibit very little bias even for small samples.

4. Applications with real-life data

In this section we are going to present a few datasets from natural calamities. In each case we show the empirical probability distribution, the fitted ZIP probability distribution as well as the fitted regular Poisson probability distribution. In each of the following figures the estimated λ parameter under ZIP and Poisson models are denoted by $\hat{\lambda}^z$ and $\hat{\lambda}^p$, respectively.

Earthquake dataset. Table 1 shows the number of major US earthquakes (those of magnitude at least 7.0) per year from 1950 through 2012. Figure 11 shows the plots using $\hat{\pi}_{ML} = 0.17$ and $\hat{\lambda}^z_{ML} = 0.69$ for ZIP and using $\hat{\lambda}^p = 0.57$ for Poisson.

Decade	Count of yearly events									
1950–1959	0	0	1	1	1	0	0	5	2	1
1960–1969	0	0	0	0	1	2	1	0	0	0
1970–1979	0	0	1	0	0	2	0	0	0	1
1980–1989	1	0	0	0	0	0	1	1	1	0
1990–1999	0	1	2	1	1	0	1	0	0	1
2000–2009	0	0	2	2	0	1	0	1	0	0
2010–2019	0	0	0	-	-	-	-	-	-	-

Table 1. Number of major US earthquakes per year from 1950 through 2012 [USGS 2012].

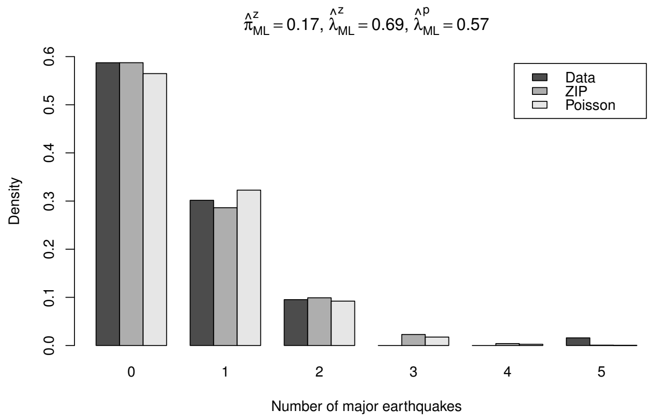

Figure 11. Empirical, fitted ZIP, and fitted Poisson models of the number of major earthquakes per year.

Wildfire dataset. Table 2 shows the number of major US wildfires (covering 400,000 acres or more) per year from 1997 through 2012. Figure 12 shows the plots using $\hat{\pi}_{ML} = 0.29$ and $\hat{\lambda}^z_{ML} = 1.6$ for ZIP and using $\hat{\lambda}^p = 1.25$ for Poisson.

Decade	Count of yearly events									
1990–1999	-	-	-	-	-	-	-	1	0	0
2000–2009	0	0	3	0	5	1	1	1	0	3
2010–2019	0	1	2	-	-	-	-	-	-	-

Table 2. Number of major US wildfires per year from 1997 through 2012 [NIFC 2012].

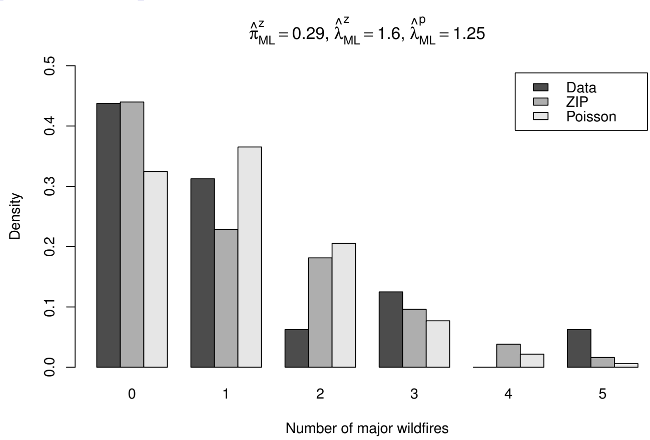

Figure 12. Empirical, fitted ZIP, and fitted Poisson models of the number of major US wildfires per year.

Decade	Count of yearly events									
1980–1989	-	-	-	-	-	-	-	0	0	1
1990–1999	0	0	1	0	0	1	0	0	2	2
2000–2009	0	0	1	1	3	4	0	0	2	0
2010–2019	0	0	0	-	-	-	-	-	-	-

Table 3. Number of major Atlantic hurricanes per year having landfall in the US from 1987 through 2012 [UNISYS 2012].

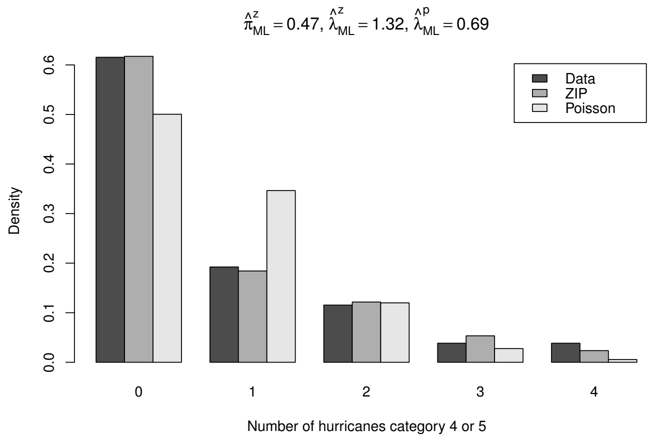

Figure 13. Empirical, fitted ZIP, and fitted Poisson models of the number of major Atlantic hurricanes per year to have landfall in the US.

Hurricane dataset. Table 3 shows the number of major Atlantic hurricanes (category 4 or 5) per year to have made landfall in the US from 1987 through 2012. Figure 13 shows the plots using $\hat{\pi}_{ML} = 0.47$ and $\hat{\lambda}^z_{ML} = 1.32$ for ZIP and using $\hat{\lambda}^p = 0.69$ for Poisson.

Tornado dataset. Table 4 shows the number of tornado occurrences in Lafayette Parish, Louisiana, US per year from 1950 through 2012. Figure 14 shows the plots using $\hat{\pi}_{ML} = 0.27$ and $\hat{\lambda}^z_{ML} = 0.93$ for ZIP and using $\hat{\lambda}^p = 0.63$ for Poisson.

Lightning dataset. Table 5 shows the number of lightning fatalities in Louisiana caused by a tree, out in the open, on golf courses, and on boats, per year from 1995 through 2012. Figure 15 shows the plots as well as estimated parameters.

Remark 4.1. Table 6 provides the *goodness of fit* (GOF) test results. For each dataset, k represents the number of categories (i.e., the values of X) which is determined so that each category has at least one frequency, and the last category has been taken as $X \geq k$. The GOF test statistic is $\Delta_{\text{GOF}} = \sum_{i=0}^{k} (O_i - E_i)^2 / E_i$,

Decade	Count of yearly events									
1950–1959	0	0	0	1	0	0	0	1	0	0
1960–1969	1	0	0	0	1	1	0	0	0	2
1970–1979	0	0	0	0	1	3	0	2	1	0
1980–1989	1	0	0	1	0	1	0	0	2	1
1990–1999	0	1	2	0	0	1	0	1	2	0
2000–2009	0	0	3	0	2	0	1	1	3	0
2010–2019	1	1	1	-	-	-	-	-	-	-

Table 4. Number of tornado occurrences in Lafayette Parish, Louisiana per year from 1950 through 2012 [NOAA 2012b].

Decade	Fatalities by a tree									
1990–1999	-	-	-	-	-	0	1	0	0	0
2000–2009	0	0	1	0	0	0	0	0	0	0
2010–2019	0	0	2	-	-	-	-	-	-	-
	Fatalities in the open									
1990–1999	-	-	-	-	-	1	0	0	2	1
2000–2009	0	1	1	0	0	1	0	0	0	0
2010–2019	1	0	0	-	-	-	-	-	-	-
	Fatalities on golf courses									
1990–1999	-	-	-	-	-	0	0	0	0	2
2000–2009	0	0	0	0	0	0	0	0	0	0
2010–2019	0	1	0	-	-	-	-	-	-	-
	Fatalities on boats									
1990–1999	-	-	-	-	-	0	0	0	2	1
2000–2009	2	0	0	0	1	0	0	0	0	1
2010–2019	0	0	0	-	-	-	-	-	-	-

Table 5. Number of lightning fatalities by category per year 1995 through 2012 [NOAA 2012a].

where O_i is the observed frequency of the event $X = i$ for $i \leq k-1$, and the event $X \geq k$ when $i = k$, and E_i is the expected frequency obtained by multiplying the corresponding fitted probability by the sample size n. Note that the p-values are all more than 75%, and mostly 90% or higher. This clearly shows that the ZIP model gives a very good fit to model the severe natural calamities which occur rarely.

Remark 4.2. In some of these plots it is seen that the fitted Poisson model comes very close to the fitted ZIP model, namely when the estimated π is very close to

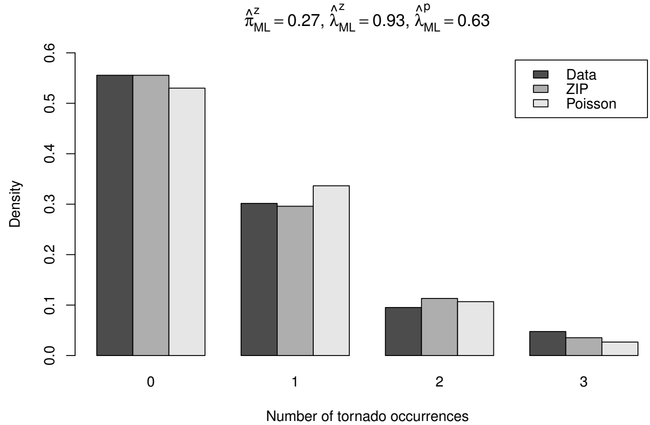

Figure 14. Empirical, fitted ZIP, and fitted Poisson models of the number of tornado occurrences per year in Lafayette, Louisiana.

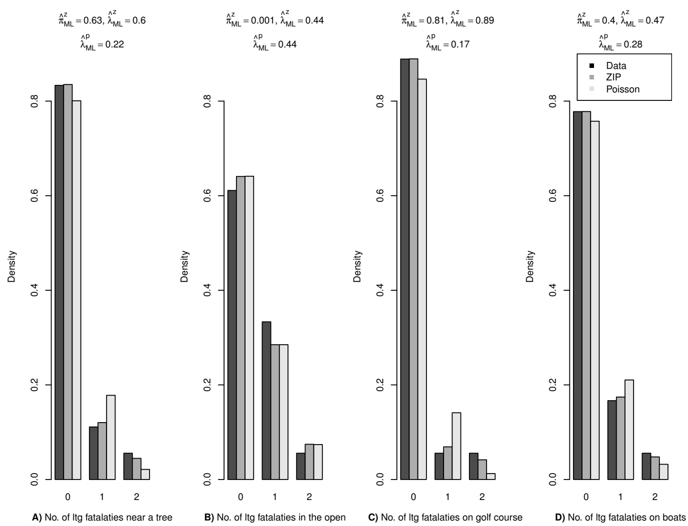

Figure 15. Empirical, fitted ZIP, and fitted Poisson models of dhe number of lightning fatalities in Louisiana for specified situations.

zero. It is reasonable to expect that $\hat{\pi}_{ML}$ being close to 0 implies that $\pi = 0$, i.e., that the ZIP model reduces to the regular Poisson model. Currently, hypothesis testing on the ZIP parameters is under consideration, and will be reported in near future.

Concluding remark

Using the fitted ZIP model, one can estimate that in any year, the probability of having at least one major earthquake in the US is 0.4136 or approximately 41%.

	Observed Δ_{GOF} value δ	p-value = $P(\chi^2_{k-2} > \delta)$	Conclusion
Earthquake data	$\delta = .3575$, $k = 4$	.9489	good fit
Wildfire data	$\delta = 1.892$, $k = 5$	.7556	good fit
Hurricane data	$\delta = .3680$, $k = 5$	.9850	good fit
Tornado data	$\delta = .3637$, $k = 5$	.9853	good fit
Lightning data			
outside	$\delta = .2628$, $k = 3$	.8769	good fit
near tree	$\delta = .0606$, $k = 3$	.9702	good fit
on a golf course	$\delta = .1218$, $k = 3$	.9409	good fit
on a boat	$\delta = .4158$, $k = 3$	.8123	good fit

Table 6. Goodness of fit results.

Similarly, the probability of Lafayette Parish getting hit by a tornado in any year is 0.4420 or approximately 44%. Hopefully these probabilities may find applications in the insurance industry, and this study will stimulate further research in this direction.

Acknowledgements

This work is a part of the undergraduate summer research and professional development experience, organized by the University of Louisiana at Lafayette, called "Smooth Transition for Advancement to Graduate Education" (STAGE) for Underrepresented Minorities (URM) in Mathematical Sciences. STAGE is a Pilot Project supported by the National Science Foundation under the Grant DMS-1043223. Five URM Students — S. Beckett (University of Louisiana at Lafayette, LA), J. Jee (Louisiana College, Pineville, LA), T. Ncube (University of West Florida, Pensacola, FL), S. Pompilus (Georgetown University, Washington, D.C.), and Q. Washington (Southern University, New Orleans, LA) — constituted the statistics group of the STAGE 2013 program. With help from A. Singh (graduate assistant) on numerical computations and statistical simulations, the statistics group worked under the supervision of Professor N. Pal, Principal Investigator of STAGE.

References

[Böhning 1998] D. Böhning, "Zero-inflated Poisson models and C.A.MAN: a tutorial collection of evidence", *Biometric Model* **40**:7 (1998), 833–843. Zbl 0914.62091

[Böhning et al. 1999] D. Böhning, E. Dietz, P. Schlattmann, L. Mendonça, and U. Kirchner, "The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology", *J. R. Statist. Soc. A* **162**:2 (1999), 195–209.

[Davidson 2012] C. L. Davidson, *A model selection paradigm for modeling recurrent adenoma data in polyp prevention trials*, Master's thesis, University of Arizona, 2012, http://arizona.openrepository.com/arizona/handle/10150/228465.

[Lambert 1992] D. Lambert, "Zero-inflated Poisson regression with an application to defects in manufacturing", *Technometrics* **34**:1 (1992), 1–14. Zbl 0850.62756

[NIFC 2012] NIFC, "1997–2012 large fires (100,000+ acres)", 2012, http://www.nifc.gov/fireInfo/fireInfo_stats_lgFires.html. Table.

[NOAA 2012a] NOAA, "Natural hazard statistics: lightning", 2012, http://www.nws.noaa.gov/om/hazstats.shtml.

[NOAA 2012b] NOAA, "Storm prediction center warning coordination meteorologist page", statistical report, 2012, http://www.spc.noaa.gov/wcm/#data.

[Ridout et al. 1998] M. Ridout, C. G. B. Demetrio, and J. Hinde, "Models for count data with many zeros", pp. 179–192 in *Proceedings of the XIXth International Biometric Conference* (Cape Town, 1998), International Biometric Society, Washington, DC, 1998.

[Schwartz and Giles 2013] J. Schwartz and D. E. Giles, "Biased-reduced maximum likelihood estimation for the zero-inflated Poisson distribution", preprint, 2013, http://econpapers.repec.org/RePEc:vic:vicewp:1102. To appear in *Commun. Stat. Theory Methods*.

[UNISYS 2012] UNISYS, "Atlantic tropical storm tracking by year", statistical report, 2012, http://weather.unisys.com/hurricane/atlantic/index.php.

[USGS 2012] USGS, "Historic earthquakes in the United States and its territories", statistical report, 2012, http://earthquake.usgs.gov/earthquakes/states/historical.php.

Received: 2013-07-24 Revised: 2013-12-06 Accepted: 2014-01-03

smbeck17@aol.com — *University of Louisiana at Lafayette, 104 E University Avenue, Lafayette, LA 70504, United States*

joshua92891@yahoo.com — *Louisiana College, Pineville, LA 71360, United States*

tmn11@students.uwf.edu — *University of West Florida, 11000 University Parkway, Pensacola, FL 32514, United States*

sp485@georgetown.edu — *Georgetown University, 3700 O St NW, Washington, DC 20057, United States*

washingtell@gmail.com — *Southern University New Orleans, 6400 Press Drive, New Orleans, LA 70126, United States*

anshumandotsingh@gmail.com — *University of Louisiana at Lafayette, 104 E University Avenue, Lafayette, LA 70504, United States*

nxp3695@louisiana.edu — *Department of Mathematics, University of Louisiana at Lafayette, PO Box 41010, Lafayette, LA 70504, United States*

involve

msp.org/involve

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2014 is US $120/year for the electronic version, and $165/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFlow® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

involve

2014 vol. 7 no. 6