Journal of Environmental Statistics

September 2015, Volume 7, Issue 4. http://www.jenvstat.org

Parametric Approaches for Estimating the Number of Species of Water Birds of Bangladesh

Syed S. Hossain
ISRT, University of Dhaka

Marzana Chowdhury
ISRT, University of Dhaka

Farhana Sadia
ISRT, University of Dhaka

Abstract

Water birds are important indicators of ecological change in wetland ecosystems. By estimating the species richness of waterbirds, the changes in the environmental behaviour of Bangladesh may be assumed substantially. The leading idea of this paper is to estimate the number of species of water birds of Bangladesh. Many parametric models are used for the estimation of species richness now-a-days. We fit four parametric models to the data of water birds and compared them on the basis of AIC values and standard errors. Among these models, two-mixed exponential mixed Poisson model has proven to be the best fit model.

Keywords: Species richness, parametric model, two-mixed exponential mixed Poisson model.

1. Introduction

Species richness is used to describe the number of species which reside in a certain biosphere or belong to a particular population. Knowing the species richness of a population helps researchers understand bio-diversity. One of the most remarkable components of global bio-diversity is the water bird. Their long migrations and tendency to concentrate in large numbers on particular wetlands make them both visible and charismatic. They are important indicators of the ecological condition and productivity of wetland ecosystems, and their presence is widely valued by numerous stake holders including local human populations, tourists, associated enterprisers, hunters (both sport and subsistence) and research biologists throughout the world. The presence of water birds also offers many opportunities for using wetlands sustainably, particularly through eco-tourism. This is particularly important for developing countries.

Bangladesh is a country with lots of wetlands including rivers and streams, freshwater lakes

and marshes, haors, baors, beels, water storage reservoirs, fish ponds, flooded cultivated fields and estuarine systems with extensive mangrove swamps. The haors, baors, beels and jheels are of fluvial origin and are commonly identified as freshwater wetlands. These freshwater wetlands occupy four landscape units - flood plains, freshwater marshes, lakes and swamp forests. For many reasons bio-diversity in these areas is reducing, many species of flora and fauna are threatened, wetland-based ecosystem is degenerating, and the living conditions of local people are changing as livelihoods, socioeconomic institutions, and cultural values are affected (Islam 2010; Shafiqul-Islam 2012). So, it is important and interesting for many researchers to know how many unseen species of water birds are there. This, practically, can be treated as an indicator of the environmental change. Application of proper statistical tools is a requirement for prediction of the total number of species of water birds.

Predicting the number of microbial species in biology is recently a very popular issue (Hong, Bunge, Jeon, and Epstein 2006). Bunge and Barger (2008) compared among seven parametric models to get the best fit for the microbial data and found that two-mixed exponential model best fits the data. They also showed the connections between parametric models for abundance and incidence data. Interest in the number of species is not restricted to specialized fields in biology. Bulmer (1974) estimated the number of species of butterflies and the light trapped moth using Poisson lognormal model where a method of fitting the compound Poisson lognormal distribution by maximum likelihood is described. In Keith and Whitemore (1986) the Poisson- inverse gaussian has been used as a model for species abundance.

The objective of this paper is to estimate the number of classes of the water birds by fitting parametric models such as Poisson model, exponential mixed Poisson model, two-mixed exponential mixed Poisson model, lognormal mixed Poisson model and present a comparison among the models to find the best fitted model.

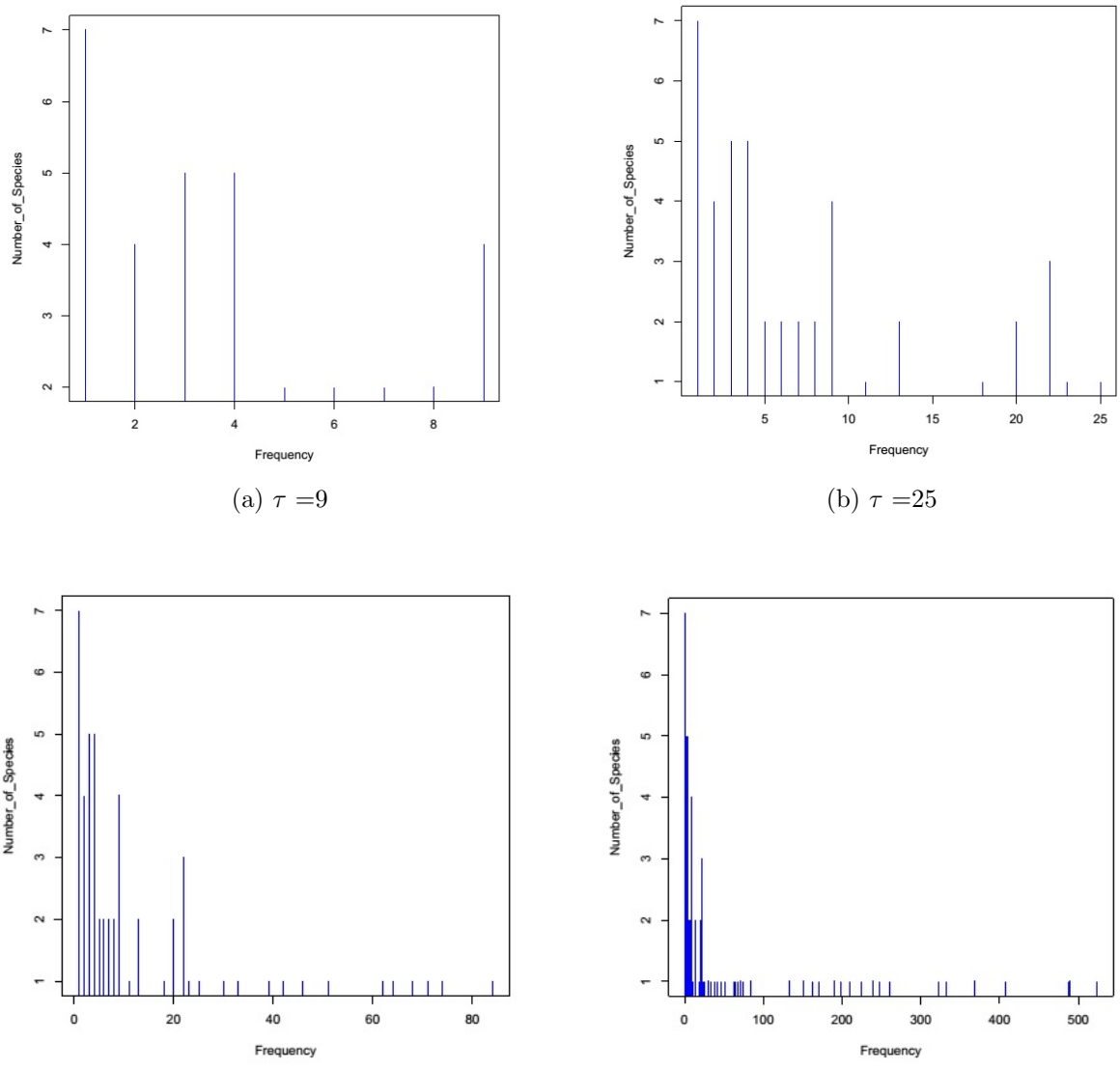
2. Data at a glance

The data analyzed in this paper is the collection of the water birds of Bangladesh which is the result of the the Asia Water bird Census: 2002-2004 by Wetlands International. During the study period, a total of 355,754 birds were sighted and classified into 99 species ranging from 7 species observed once to a maximum frequency of 85,109. A part of the data is presented in Table 1. It shows that 7 species are observed only once, 4 species are observed twice, 5 species are observed thrice and so on.

Table 1: Frequency Distribution of Observed Number of Species of Water Bird of Bangladesh

Frequency	Number of Species	Frequency	Number of Species
1	7	9	4
2	4	11	1
3	5	13	2
4	5	18	1
5	2	20	2
6	2	22	3
7	2	23	1
8	2	> 23	56

The *cut-off point* is a value that separates frequency counts into abundant and rare species. The number of species with a very high frequency is usually low, that is why, a cut-off point for the frequency counts is set so that the parametric model be fit for the species with maximum frequency counts equal to the cut-off point. Often, the choice of the cut-off point is made by trial and error basis from an arbitrarily chosen set. The cut-off points are also termed as the *tuning parameter*. The data at different cut-off points (τ) are analyzed in this paper. Figure 1 shows the number of species against frequencies for cut-offs 9, 25, 84 and 523 for the water birds data. This illustrates how the distributional behaviour of observed data changes as the cut-off value increases.



(a) τ =9

(b) τ =25

(c) τ =84

(d) τ =523

Figure 1: Water Bird of Bangladesh

3. Estimating Species Richness Via Parametric Models

In estimating species richness parametric models are historically being used [see R. A. Fisher and Williams (1943)] and Bunge and Barger (2008) justified the use of parametric models for such scenario but argued that no consensus has yet emerged regarding the choice of parametric distribution for real applications. In a comprehensive study, they considered all the important models (seven in total) proposed in the literature and discussed the comparative issues of goodness of fit, model selection, maximal use of the data etc. In this paper, the four parametric models are chosen in line of the merits and demerits stated by them and suitability for the water birds data of Bangladesh. Of the seven models they compared, the ones that were not considered in this paper are given in the following along with the reasons for not considering them:

1. G-mixed Poisson or negative binomial: Although popular, this model has almost never been found to fit Bunge and Barger (2008) data sets.

2. Inverse Gaussian: This distribution was found to be little used in practice;

3. Pareto: No previous application of this distribution in the species problem were found by Bunge and Barger (2008).

Consider a population divided into C classes, where $C < \infty$ is known. Let, the number of individuals observed in the given time interval be denoted by n and we consider these n observed individuals as the *sample* data. Note that the sample size referred to as n is not pre-determined or fixed. Assuming each individual can be identified by only one class, all of the individuals observed can be sorted into their respective classes. Let c be the number of unique classes that have appeared in the sample. Define Y_j to be the number of observed individuals associated with class j where $j = 1, 2, \cdots, C$. We can describe Y_j as a Poisson random variable with mean $j > 0$. The probability density function for Y_j is

$$p(y_j) = P[Y_j = y_j] = \frac{exp(-\lambda_j)\lambda_j^{y_j}}{y_j!} \ . \tag{1}$$

We assume that the Y_j are independent for all j. We will call $Y = (Y_1, Y_1, Y_2, \cdots, Y_C)$ the non-truncated data vector. If $Y_j = 0$, then no individuals from class j were observed in the sample.

3.1. Ordinary or Unmixed Poisson

Ordinary or unmixed Poisson model, which assumes equal abundances, rarely fits data except in the case of very coarsely defined taxa, but serves as a useful lower-bound benchmark. We will now assume that C is unknown. This means when class j is not found in the sample, we do not know the class exists. The truncated data for each class is X_i, where $i = 1, 2, \cdots, c$ and c is the number of unique classes observed in the sample. We observe a subset, $X_i \subseteq Y_j$ such that $Y_j > 0$. In other words, $X_i = Y_j$ if $Y_j > 0$. Denote the truncated (or more specifically, zero-truncated) data vector as $X = (X_1, X_2, \cdots, X_c)$. So, X_i for $i = 1, \cdots, c$ will represent the observed data. By properties of conditional probability, we have

$$P[X_i = x_i] = P[Y_j = x_i | Y_j > 0] = \frac{P[Y_j = x_i, Y_j > 0]}{P[Y_j > 0]} = \frac{P[Y_j = x_i]}{1 - P[Y_j > 0]} \quad ,$$

where $x_i = 1, 2, \cdots$, and we call the distribution of X_i a zero-truncated Poisson distribution.

3.2. Exponential Mixed Poisson

Consider Y_j, the number of individuals observed from species j, where $j = 1, \cdots, C$ and C is the total number of species. Let $Y_j|\lambda_j$ follows Poisson(λ_j), independently for each j. Now let λ_j follows Exponential(θ), where θ is the mean of the distribution, i.e., if $p(\lambda_j|\theta)$ represents the probability density for λ_j, then

$$p(\lambda_j|\theta) = \frac{1}{\theta} exp\left(-\frac{\lambda_j}{\theta}\right) \ .$$

We assume that the λ_j are independent for all j. The marginal distribution of Y_j is given by a geometric $(\frac{1}{1+\theta})$, that is

$$p(y_j) = \int p(y_j|\lambda_j)p(\lambda_j)d\lambda_j = \int \frac{exp(-\lambda_j)\lambda_j^{y_j}}{y_j!}\frac{1}{\theta}exp\left(-\frac{\lambda_j}{\theta}\right) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{y_j}, y_j = 0, 1, \cdots \ .$$

Now for the observed data the zero-truncated geometric distribution is

$$p[X_i = x_i] = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{x_i - 1} \ .$$

Thus X_i is also a Geometric distribution with probability of success $\frac{1}{1+\theta}$.

Using the zero-truncated Geometric distribution we find the maximum likelihood estimator of θ (Bunge and Barger 2008) as $\hat{\theta} = \bar{x} - 1$, where $\bar{x} = \frac{1}{c}\sum_{i=1}^{c} x_i$.

Let $C_0 = C - c$ denotes the number of unobserved classes. We estimate C_0 as

$$C_0 = Cp_0(\hat{\theta}) = (C_0 + c)p_0(\hat{\theta}) \ . \tag{2}$$

From equation 2, we have

$$\hat{C_0} = \frac{cp_0(\hat{\theta})}{1 - p_0(\hat{\theta})}.$$

For the single Exponential Mixed Poisson Model,

$$\hat{C_0} = \frac{c}{\bar{x} - 1}.$$

The asymptotic standard error of $\hat{C_0}$ is (Sanathanan 1972)

$$SE(\hat{C}_0) = \left[\frac{1}{\hat{C}}\left(\frac{1-p_0(\theta)}{p_0(\theta)} - \frac{1}{p_0^2(\theta)}\left(\frac{dp(0)}{d\theta}\right)^T I(\theta)^{-1}\left(\frac{dp(0)}{d\theta}\right)\right)^{-1}\right]^{-\frac{1}{2}} ,$$

where

$$\hat{C} = \frac{c}{1-p_0(\hat{\theta})}.$$

3.3. Two-mixed Exponential Mixed Poisson Model

Mixture of two exponential distribution as a species abundance model was proposed and applied by Barger (2006). It may also be named as mixture of two geometric. Consider a three parameter probability density for λ_j as

$$p(\lambda_j|\theta_1,\theta_2,\theta_3) = \theta_3 f_1 + (1-\theta_3) f_2 \ ,$$

where $f_1 = \frac{1}{\theta_1}e^{-\frac{\lambda_j}{\theta_1}}$ and $f_2 = \frac{1}{\theta_2}e^{-\frac{\lambda_j}{\theta_2}}$ are two exponential distributions with parameter θ_1 and θ_2 respectively. Thus we get the two-mixed exponential mixed Poisson model

$$p(\lambda_j|\theta_1\theta_2\theta_2) = \theta_3\frac{1}{\theta_1}e^{-\frac{\lambda_j}{\theta_1}} + (1-\theta_3)\frac{1}{\theta_2}e^{-\frac{\lambda_j}{\theta_2}}.$$

The marginal distribution of Y_j is, hence given by the weighted sum of the two geometric probabilities as

$$\begin{aligned} P[Y_j = y_j] &= \int p(y_j|\lambda_j)p(\lambda_j|\theta_1,\theta_2,\theta_3) \\ &= \theta_3\int f_1 p(y_j|\lambda_j)d\lambda_j + (1-\theta_3)\int f_2 p(y_j|\lambda_j)d\lambda_j \\ &= \frac{\theta_3}{1+\theta_1}\left(\frac{\theta_1}{1+\theta_1}\right)^{y_j} + \frac{1-\theta_3}{1+\theta_2}\left(\frac{\theta_2}{1+\theta_2}\right)y_j. \end{aligned} \tag{3}$$

The zero truncated term of equation (3) is given by

$$P[X_i = x_i] = \frac{P[X_i = x_i]}{1-P[X_i=0]} = \frac{\frac{\theta_3}{1+\theta_1}\left(\frac{\theta_1}{1+\theta_1}\right)^{x_i} + \frac{1-\theta_3}{1+\theta_2}\left(\frac{\theta_2}{1+\theta_2}\right)x_i}{1-\left(\frac{\theta_3}{1+\theta_1}+\frac{1-\theta_3}{1+\theta_2}\right)}, \qquad x_i = 1,2,\cdots.$$

3.4. Lognormal Mixed Poisson Model

The Poisson lognormal model was proposed by Bulmer (1974) as a model for estimating species abundance. Like section 3.2, let $Y_j|\lambda_j$ follows Poisson (λ_j), independently for each j but λ_j follows Lognormal (M,V), where M is the mean and V is the variance of the distribution, i.e. if $p(\lambda_j|M,V)$ represents the probability density for λ_j, then

$$p(\lambda_j|M,V) = \frac{1}{\lambda_j(\sqrt{2\pi V})^{-\frac{1}{2}}} exp\left(-\frac{(log\lambda_j - M)^2}{2V}\right).$$

The marginal distribution of Y_j is

$$\begin{aligned} P(y_j) &= \int p(y_j|\lambda_j)p(\lambda_j|M,V) \\ &= \frac{(2\pi V)^{-\frac{1}{2}}}{y_j!}\int \lambda_j^{y_j-1} exp(-\lambda_j) exp\left(-\frac{(log\lambda_j - M)^2}{2V}\right) d\lambda_j,\ y_j = 0,1,2,\cdots. \end{aligned} \quad (4)$$

The probabilities in equation 4 are cumbersome to evaluate even with computers. An approximation for large values of y_j was used by Bulmer (1974) and it is

$$P(Y_j) = \frac{(2\pi V)^{-\frac{1}{2}}}{y_j!} exp\left(-\frac{(log\lambda_j - M)^2}{2V}\right)\left[1 + \frac{1}{2Vy_j}\left(\frac{(lny_j - M)^2}{V} + lny_j - M - 1\right)\right].$$

Comparisons with results obtained by numerical integration shows that this approximation has a relative error less that 10^{-3} when $y_j \geq 10$ for values of M and V likely to be encountered in practice. The distribution can be fitted to observed data by estimating the parameters M and V, by the method of maximum likelihood. The observed data is considered as a sample from a zero truncated Poisson lognormal model. In the case of a truncated distribution the total number of species, including the missing ones can be estimated as

$$\hat{C} = \frac{c}{1 - \hat{P}_0}. \quad (5)$$

The total number of species is $\hat{C} + \sum I[x_i > \tau]$, where τ represents the cut off points. The approximate sampling variance of the estimate in equation 5 is

$$\text{Var}(\hat{C}) = \frac{\text{var}(c)}{(1-\hat{P}_0)^2} + \frac{c^2}{(1-P_0)^4}\text{var}(\hat{P}_0) \quad ,$$

where, $\text{var}(c) = CP_0(1-P_0) \approx c\hat{P}_0$ and
$\text{var}(\hat{P}_0) = \hat{P}_1^2\text{var}(\hat{M}) + \hat{P}_1(\hat{P}_1 - 2\hat{P}_2)\text{cov}(\hat{M},\hat{V}) + \frac{1}{4}(\hat{P}_1 - 2\hat{P}_2)^2\text{var}(\hat{V})$.

4. Analysis

For different value of the tuning parameter (cut-off point), τ, the observed and estimated number of water bird species along with the standard error and associated AIC values for ordinary Poisson model (section 3.1) are given in Table 2.

Table 2: Observed and fitted values of number of water bird species under Poisson model alongside standard error and AIC

Cutoff	Observed	Estimated	SE	AIC
9	33	99.55	0.7737	46.06
11	34	99.46	0.7006	50.74
13	36	99.29	0.5537	67.25
20	39	99.10	0.3207	125.79
25	44	99.02	0.1290	229.98
30	45	99.01	0.1019	262.21
40	47	99.003	0.0578	350.25
50	50	99.0004	0.0198	539.36
100	56	99.00000081	0.0008	1180.94

The ordinary Poisson model assumes equal abundance, so it is not considered as a realistic model. We can see in Table 2 that the total number of unseen species at cutoff ($\tau = 9$) is 99.55 with a large AIC values whereas there are already 99 species in the data. So there is barely a chance that any species is unseen, which is almost impossible in reality. As the cut-off increases the estimate tends to 99, that is the model fails to predict any unseen species from the data.

The fitted curve of Poisson model at $\tau = 6$ is given in Figure 2 which shows the apparent lack of fit by this model.

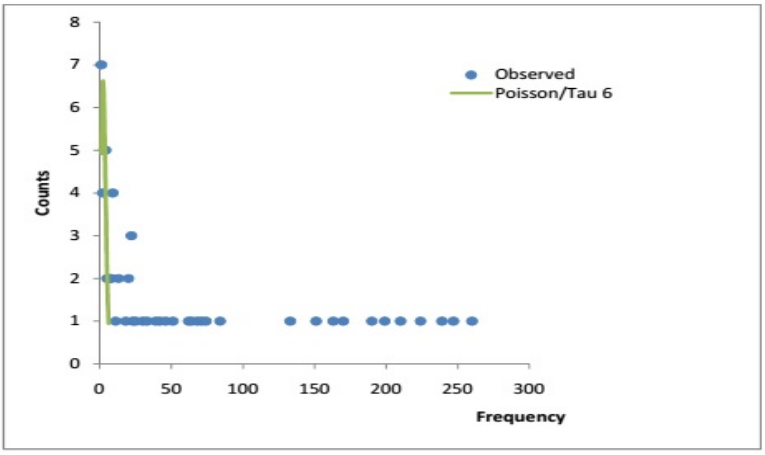


Figure 2: *Observed and fitted values of number of water bird species under Poisson model ($\tau = 6$)*

Table 3 presents the estimated value of number species of water birds for each cut-off point using exponential mixed Poisson model along with standard error and AIC. Table 3 exhibits that the highest estimated total is 109.37 at cut-off point 9 and as the cut-off point increases the the value decreases to 99.90 for cut-off value 500, That is if the cut-off value is increased the value will tend to 99, the number of species in the sample,that is there are no unseen

species. The standard error of the estimates are also showing a decreasing pattern with the increasing values of τ . As we look at the AIC values, they are much lower than the AIC values of ordinary Poison, which means Exponential models is giving a better fit comparing to the Poisson model. Figure 3 exhibits the fitted curve on the observed data for the Exponential mixed Poisson model at $\tau = 9$.

Table 3: Observed and fitted values of number of water bird species under exponential mixed Poisson model alongside standard error and AIC

Cutoff	Observed	Estimated	SE	AIC
9	33	109.37	4.23	38.17
11	34	109.05	4.11	39.35
13	37	107.78	3.66	47.82
20	39	106.84	3.36	56.46
25	44	105.39	2.89	76.08
30	45	105.10	2.80	80.27
40	47	104.50	2.62	90.25
50	50	103.65	2.36	107.95
100	56	102.28	1.92	151.002
500	73	99.96	0.99	355.07

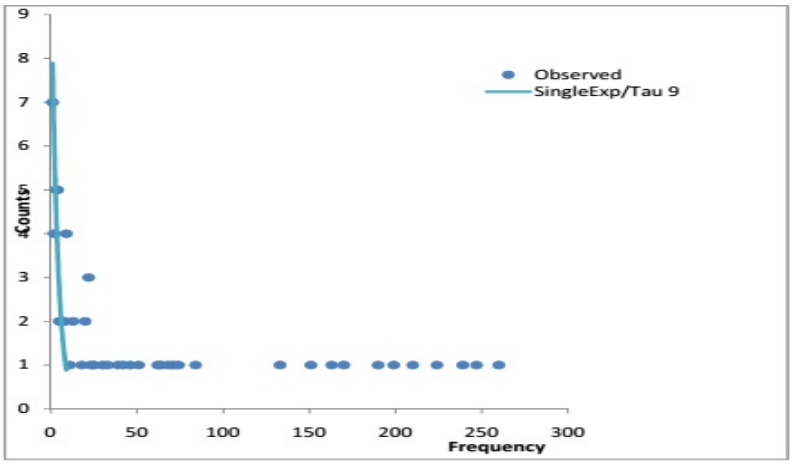


Figure 3: *Observed and fitted values of number of water bird species under single exponential mixed Poisson model ($\tau = 9$)*

The observed and predicted number of water bird species as well as the standard error and associated AIC values after fitting the two-mixed exponential mixed model on the water bird data is showed in Table 4.

Table 4: Observed and fitted values of number of water bird species under two mixed exponential mixed Poisson model alongside standard error and AIC

Cutoff	Observed Sp	Estimated Total Sp	SE	AIC
9	33	109.37	4.13	42.17
11	34	109.05	4.15	43.35
13	36	108.32	4.17	48.28
20	39	106.84	3.68	60.46
25	44	105.39	3.10	80.08
30	45	107.83	6.80	83.50
40	47	107.92	5.80	92.46
50	50	107.96	5.06	107.55
100	56	107.79	4.40	141.99
200	62	106.73	3.62	193.05
500	74	105.38	3.01	297.55

Table 4 shows a similar pattern as in single exponential mixed Poisson model. The estimated value of total number of water bird species is 109.37 at $\tau = 9$ and decreases to 105.38 for $\tau = 500$. But as we may notice that a sudden decrease at cut-off points 20 to 25. The change is also noticeable in AIC values. But this change of data pattern is ignored by the Poisson, exponential mixed Poisson. From the AIC values we can see that, this model has lower AIC values comparing to the previous two models (Poisson and Exponential mixed Poisson).

The fitted curve of two mixed exponential mixed Poisson at $\tau = 260$ is showed in Figure 4 which shows a good fit.

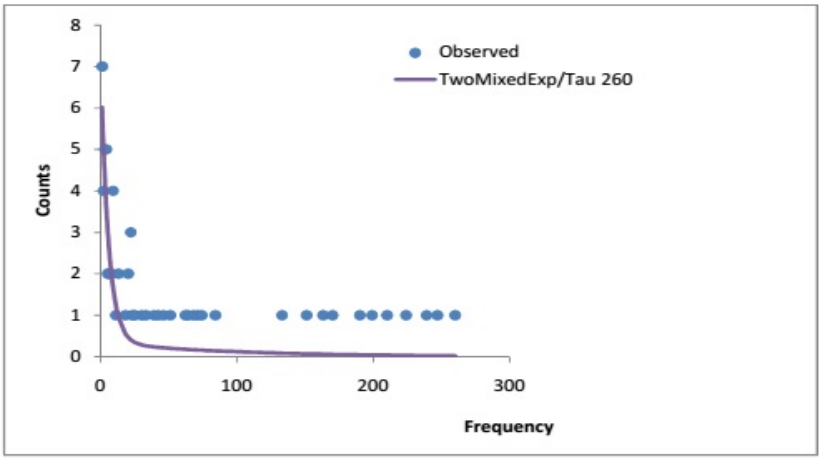


Figure 4: *Observed and fitted values of number of water bird species under two-mixed exponential mixed Poisson model ($\tau = 260$)*

Poisson lognormal model is fitted in the data using the Bulmer‘s approximation **?**. Table 5 shows The observed and predicted number of water bird species under lognormal Poisson

model together with the standard error and associated AIC values for each value of the tuning parameter value, τ .

Table 5: Observed and fitted values of number of water bird species under lognormal mixed Poisson model alongside standard error and AIC

Cutoff	Observed Sp	Estimated Total Sp	SE	AIC
9	33	101.44	2.12	154.75
11	34	101.52	2.16	163.39
13	36	101.76	2.02	182.20
20	39	102.58	2.57	215.79
25	44	102.97	2.67	269.99
30	45	103.08	2.71	281.68
40	47	103.37	2.81	306.25
50	50	103.84	2.98	345.18
100	56	104.76	3.28	429.08
200	62	106.50	3.91	528.16
210	63	106.70	3.98	544.91
500	74	108.32	4.44	739.52

Table 5 shows that the estimated values increase with the value of cut-offs but the estimated values increase slowly from 101.44 ($\tau = 9$) to 108.32 ($\tau = 500$). This model has small standard errors comparing to the previous models (exponential mixed Poisson model and two mixed exponential mixed Poisson model). But AIC values are highest among the all other models, proving a possible worse fit for the data.

5. Discussion and Conclusion

Our basic objective is to estimate the number of unseen species of the water birds in Bangladesh with the best model among the used tools. Our sample contains 99 species of water birds. Four candidate parametric models have been fitted. Among them the ordinary Poisson model is not a realistic one since it assumes equal abundance. This model has failed to predict any unseen species because estimated number of total species ranges from 99.54 to 99. On the basis of AIC values of the remaining models we may see that the Lognormal mixed Poisson has the highest AIC value. Now the remaining two models are, single exponential mixed Poisson model and two mixed exponential mixed Poisson model. The summary of the fittings with these two models are given in Table 6.

Table 6: Summary findings of the selected best models

Model Name	Cutoff	Estimated Total Sp	Number of unseen species	SE	AIC
TwoMixedExp	100	107.79	8	4.40	141.99
SingleExp	9	109.4	9	4.2	38.17

For each of the two models, the model with minimum AIC (Akaike Information Criterion) is

selected. Models combinations for which SE > estimate/2 are eliminated. Another criteria for the selection of the model is to choose which has utilized the maximum information with a small AIC. After analyzing the selection characteristics we may say two mixed exponential mixed Poisson is the best model at $\tau = 100$, because of the maximum usage of the information and covering the right tail of the data. The estimated number of species is 107.79. That is there may be eight more species of water birds that are still unseen.

References

Barger K (2006). *Mixtures of exponential distributions to describe the distribution of Poisson means in estimating the number of unobserved classes.* Master's thesis, Cornell University.

Bulmer MG (1974). "On fitting the Poisson lognormal distribution to species-abundance data." *Biometrics*, **30**, 101–110.

Bunge J, Barger K (2008). "Parametric models for estimating the number of classes." *Biometrical Journal*, **50**(6), 971–982.

Hong SH, Bunge J, Jeon SO, Epstein SS (2006). "Predicting microbial species richness." *Proceedings of the National Academy of Sciences of the United States of America*, **103**(1), 117–122.

Islam SN (2010). "Threatened wetlands and ecologically sensitive ecosystems management in Bangladesh." *Frontiers of Earth Science in China*, **4**(4), 438–448.

Keith OJ, Whitemore A (1986). "The Poisson-inverse gaussian distribution as a model for species abundance." *Communications in Statistics-Theory and Methods*, **15**, 853–871.

R A Fisher ASC, Williams CB (1943). "The relation between the number of species and the number of individuals in a random sample of an animal population." *Journal of Animal Ecology*, **12**, 42âĂŞ58.

Sanathanan L (1972). "Estimation The Size of a Multinomial Population." *The Annals of Mathematical Statistics*, **43**, 142–152.

Shafiqul-Islam M (2012). "Present Status of Wetland Biodiversity-A Study in Sujanagar Upazila, Pabna, Bangladesh." *IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)*, **3**(1), 06–13.

Affiliation:

Syed S. Hossain
Professor
Institute of Statistical Research and Training (ISRT), University of Dhaka
Dhaka, Bangladesh
E-mail: shahadat@isrt.ac.bd
URL: http://www.isrt.ac.bd/faculty/shahadat

Journal of Environmental Statistics
Volume 7, Issue 4
September 2015

http://www.jenvstat.org
Submitted: 2015-01-26
Accepted: 2015-09-01