MODELLING SOIL EARTHWORM COUNT DATA WITH EXCESS ZEROES PROBLEM AND CHOICE OF APPROPRIATE MODELS IN DIFFERENT FOREST ECOSYSTEMS OF MANIPUR.

Ksh. Anand Singh[1], M. Arun Singh[2],

Assistant Professor, Research Scholar

Department of Statistics, Manipur University, Imphal, India.

Abstract:

Analysis of ecological data is complex because the data structure in any ecological set up is in itself complex. In many cases, the normality assumption is often violated and as such fitting the normal linear models to ecological data is not at all the usual way. There are variety of other models which are conflicting in themselves and so choosing an appropriate one is another point of discussion. In this paper, the abundance of earthworm species is modelled through various soil and environmental characteristics in three subtropical forest ecosystems of Manipur, India. Earthworm count is observed in Mixed reserved forest, Disturbed forest and Plantation forest ecosystems in six different locations of each observation site at a soil depth of ten cm. The count of species of a particular type of earthworm observed during the twelve months of the year is regressed on 9 soil characteristics. One of the crucial challenges which ecologists often encounter in dealing with species count data is its inherent complexity arising out of sampling procedure which is further complicated by the presence of excess number of zeroes in the dataset. When the frequency of zeroes is very large and do not readily fit into any of standard distributions mainly because of skewness and over-dispersion, the dataset is referred to as zero inflated. Starting with the Standard Poisson model, we fitted four other different models viz. quasi-Poisson, Zero-inflated Poisson, Negative Binomial and Zero-inflated Negative binomial models. The appropriateness of the models are checked using AIC, BIC and Vuong test. The appropriateness of the models depends on the particular species type and type of forest.

Key words: Quasi-Poisson, zero-inflated, AIC, BIC, Vuong test.

1. Introduction:

The field of ecological data modelling has grown amazingly complex over the years. One of the greatest challenges in modelling ecological data by way of learning statistics is to figure out how the various methods relate to each other and determining which method is most appropriate for any particular problem. There are a number of statistical methods available to ecologists which are derived and available in the literature as a consequence of the fact that ecological data is complex. However, no single method can accommodate the myriad problems we encounter with ecological data. Thus, we have to look for various methods available and derive a meaningful model to choose, while seeking for an appropriate analysis.

In most ecosystems, both in natural and plantation forest grasslands and agro-ecosystem, earthworms represent a major portion (>80%) of the soil invertebrate biomass and involve in the process of soil formation and maintenance of soil fertility. Soil earthworm abundance is a concern for the ecologists in activities such as agriculture, forestry and environmental monitoring. However, the complexity and diversity of soil animals and the habitats in which they live pose unique challenges. Distribution and abundance of earthworm species like other soil animals are governed by several environmental factors such as temperature, moisture, soil pH, soil porosity, soil bulk density and available organic matter etc. The number of species in a given earthworm community, which is the simplest measure of species diversity range from 1 to 15 species (Edwards and Bohlen, 1996). Apart from the various soil and environmental factors, the diversity of soil animals depends on the organic resources of the locality as well as its history of land use and soil disturbance. Earthworms perform several beneficial functions which include decomposition of organic matters that helps in increasing soil nutrients, increase air water filtration, soil aggregation, increase the availability of plant nutrients, worn cast as biofertilizers etc.

Recent studies have revealed that soil animal counts exhibit two features: a substantial proportion of the values are zero's and the remainder have a skewed distribution (Sileshi and Mafongoya, 2006a; Sileshi and Mafongoya, 2007). When the frequency of zeroes is very large and do not readily fit to any of standard distributions, the dataset is referred to as zero inflated (Lambert, 1992; Martin et al.,2005). In earthworm count data, some species which are rare in different soil types shows a large number of zero counts during the unfavourable seasons. These zeros are referred to as structural 0's which are true zeroes. Sampling zeroes, often referred to as false zeroes (Mackenzie et al. 2002) occur when the species count present at the time of sampling is not detected by the observer. Another issue in count of species data occur when some sampling points show very large counts. This happens when a nest or foraging party is encountered (Jones et al. 2005). Density estimates can therefore have high variance, making it difficult to compare statistically significant differences among sites, seasons, species types etc. Zero inflated datasets are often accompanied with the problem of overdispersion, a case where the variance is larger than the mean beyond the expected limit. Overdispersion creates problem with ordinary statistical inference by violating basic assumption implicit in standard distribution (Martin et, al., 2005). Overdispersion leads to underestimation of std. error of regression parameters, Confidence Intervals and p-values.

The most common analysis used for soil animal counts are non-parametric tests or log-normal least-square regression, e.g. ANOVA. Both the methods do not take care of the problem of overdispersion arising out due to excess zeros in the dataset. Thus, the analysis of ecological data are in general of a complex nature as the data is very complex in itself and often data do not support only one model as clearly best (Dayton, 2003; Johnson and Omland, 2004). When testing differences among sites, or other treatment effects, the assumptions made on the response variable can lead to biased inferences. This arises the issue of comparing models to assess which ones are adequate for the data and which ones could be chosen for interpretation and prediction or subsequent use. The present study aims at developing models for soil earthworm counts and examine their adequacy by comparing among themselves and suggest appropriate methods for abundance estimation.

2. Methods and Materials

The data used in this study were collected from three sub-tropical forest ecosystems of Manipur, India (Sharon Haokip, 2014), to study abundance of different earthworm species and their diversity. We designate site 1: as Mixed reserved protected forest ecosystem which are protected from various biotic interference; site 2: Disturbed forest ecosystem dominated by oak and sustained to frequent biotic interference, and site 3: Oak dominated plantation forest ecosystems developed by men. The three forest ecosystem are (1) Mix reserve sub-tropical forest ecosystem located at Koirengei ($24^0\ 52'51.36''$ North latitude and $93^054'\ 49\cdot 75''$ East longitude and altitude 800 – 917 m above MSL); (2) Oak dominated Langol Hills ($24^052^151.6N\ and\ 93^055'26.59''E\ \ and\ altitude\ of\ 797-848\,m\ \ above\ MSL$ (3) Managed oak plantation Forest (valley area) at Mantripukhri ($24^052'\ 52\cdot 9''N, 93^056'0\cdot 16''E$ and altitude of 786m above MSL). In all the three sites, 6 locations were identified at least 6 feet apart between the locations for sampling counts. At each location, 3 different soil depths 10, 20 and 30 cm are observed for earthworm counts. Sampling was done during January to December in 2013-14; once in every month. The sampling period constitutes three seasons viz. Rainy season from May to August, Dry Season from September to October and March to April whereas winter season covers from November to February. Altogether, there are 12×6×3×3=648 sampling points. In site 1, 12 different species were detected, whereas in site 2 only 5 different species are detected and in site 3, only 4 different species are detected. During sampling, data on soil characteristics such as soil temperature, Soil Moisture, soil pH, Porosity, bulk density, Carbon(C), N (Nitrogen) , P(Phosphorous) are recorded at each sampling points.

2.1 Modelling Strategies

To begin with we generally think of a normal linear regression model to any given dataset. The ordinary least square regression (OLS) assumes that the probabilistic model of the original data, suitably transformed could well be assumed to be normal. However, soil animal count data generally do not follow a normal distribution thereby the use of OLS regression is prohibited. We have tested our data on different species type using the Shapiro-Wiik's normality test as well as graphical test and no evidence of normality is observed. The log transformed data also do not improve to qualify for normality assumption.

An alternative and more appropriate model when the response variable is a count data would be Poisson regression model which is generalisation of general linear model. The response variable Y in a Poisson regression takes the non- negative integer values 0,1,2,3,….. thereby it resembles the logistic regression except for the error component (Mc Culdagh P and J. Nelder,1989).

Theoretically, the parameter μ of Poisson distribution which is often referred to as incidence rate is expressed as

$$\mu = t\exp(\beta_0+\beta_1x_1+\ldots+\beta_kx_k) = t\exp\sum\beta_{ij}x_{ij} \quad (1)$$

Where $X_1,X_2,..X_k$ are the regressors and regression coefficient $\beta_1\ldots\ \beta_k$ are to be estimated using the data and t is specified constant indicating the time of exposure, and t is time.

Thus, in eqn. (1) incidence rate μ is related to regressors by a link function which is the log.

Using the notation in eqn. (1), the Poisson regression model for an observation i is written as

$$P(Y_i=y_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!} \quad (2)$$

$$\text{Where } \mu_i=t_i\exp(x'\beta) \quad (3)$$

The coefficient β's are estimated using the method of maximum likelihood.

The Poisson model involves explicitly modelling the distribution of the count assuming that the variance σ^2 is proportional to the mean (μ) say $\sigma^2=\Phi\mu$, where Φ is a dispersion parameter (Cameron and Trivedi,1998). Parameter of standard Poisson regression model for count of different soil earthworm species are estimated using the *glm* function in R by setting *family=Poisson*. The *glm* function with *family=Poisson* allows the mean of the population (μ) to depend on a linear predictor through a nonlinear function and permit the probability distribution to be a member of exponential family. In the present study the count of the earthworm species (Y_i) is assumed to be distributed as Poisson whose parameter depends on a vector of linear predictors (X_i) (such as soil temperature, soil moisture, densities etc.) through a log-linear link function.

The earthworm counts (Y_i) in the 3 sites are fitted separately with the soil variables. Only the type 1 species (Drawida Sp) is being considered in the beginning standard Poisson model. We use the untransformed original observed data. Although Poisson regression is recommended for fitting count data, it often does not fit overdispersed data very well. We examine the goodness of fit using graphical displays which implies that we should not accept the above fit in a reasonable way. The reason for lack of a well fitted model of a Poisson regression arise with overdispersed data.

The present data contains a good no. of zero counts which may arise from seasonal variation or otherwise sampling zeroes. The presence of large proportion of 0's in the data results in overdispersion. In modelling overdispersed count data, quasi-likelihood adjustment are normally done when a reasonable lack of fit to the standard Poisson is found (McCullagh and Neider,1989). In the quasi-likelihood method, we introduced a variance inflation factor estimated by Maximum likelihood. However, the introduction of the inflation factor does not arise a new probability distribution. It adjusts the standard error and provide wider confidence intervals and P- values larger than what is obtained under the standard Poisson model. In the present study we call the quasi-Poisson model as Poisson with corrected for overdispersion (PCO). The fitted model for species type 1 shows reasonably acceptable criteria.

The PCO produces an appropriate inference only if the overdispersion is moderate. The count data for species type1 show a reasonably good fit for PCO as the magnitude of overdispersion is moderate. However, for the other types of species viz type 2, 3 and 4 the no. of zero counts are very high resulting in heavily overdispersed data. The frequencies of zeroes are shown for four types of species in Table 1. Thus to fit data with highly overdispersed data we have to look for other options such as zero-inflated Poisson, Negative Binomial and Zero inflated Negative Binomial. We have fitted for the present data for all possible models mentioned above and are compared using the Akaike Information criteria(AIK) and Bayesian information criteria(BIC).

Comparison of Generalized Poisson, Poisson corrected for overdispersion (PCO), Zero inflated Poisson (ZIP), Negative Binomial distribution (NBD) and Zero inflated negative binomial (ZINB) null model (without covariates) and full models (with covariate) for first four earthworm species types in three forest ecosystems using AIC (Akaike Information Criteria) and BIC(Bayesian Information Criteria) are shown in Tables 2A – 2C.

3. Results and Discussion

In Table 2A the AIC for null model and full model (With covariates) are compared in Reserved forest. In the null model the ZINB provides a better description of the earthworm count data with Species type 1 and type 4 whereas for species type2 and type 3 ZIP and NBD respectively give better description of the data. In the full model ZINB is better for species type 1 and type 4 whereas ZIP is better for Species type2 and type 3. In both the cases the standard Poisson cannot give a better model for all the species type.

In Table 2B the AIC for null model and full model (With covariates) are compared in Disturbed forest. ZINB is better in the null model for Species type 1 and type 4 whereas NBD and ZIP respectively give better fit for species type 2 and type 3. As expected the standard Poisson GLP model is no good here also.

In Table 2C the AIC for null model and full model (With covariates) are compared in Plantation forest. In this forest type, ZINB is better than all other models for both null and full models in all the species types except species type 1 in null model.

While comparing the GLP and ZIP the later is always better in all the sites for all species types in both the null and full models. But for NBD and ZINB both the models are competing closely in site 1 and site 2. In the site 1 and 2 ZINB is better for species types 1 and 4 and NBD is better for species types 2 and 3 in the null model. In the full model in site 1, ZINB is better than NBD for species types 1, 2 and 4 and in site 2, ZINB is always better than NBD for all species types. In site 3 ZINB is better in almost all cases except the one in species type 1 null model. The AICs for PCO models are not obtained as the output of the fitted models in R and thus are not shown. The BIC data are not shown as the results are almost the same as AIC results.

In our earlier paper we have attempted to model the species count through nine soil characteristics in each of the sites regardless of the species type. The soil characteristics that could influence the count of species include soil temperature, moisture, porosity, bulk density, pH, carbon, nitrogen, potassium and phosphorous. We pick up those soil variables in each site which are significant in the quasi-Poisson model here to accommodate them into different models. AIC and BIC criteria is again used to examine the appropriateness of the different models in the three sites for the first four earthworm species. Table 4A, 4B and 4C presents the AIC values for four different models viz. GLP, ZIP, NBD and ZINB for comparison among the models. The AIC for quasi-Poisson model or PCO cannot be computed in R. The BIC values are not shown.

The Vuong non-nested test is based on a comparison of the predicted probabilities of two models that do not nest. Examples include comparisons of zero-inflated count models with their non-zero-inflated analogs (e.g., zero-inflated Poisson versus ordinary Poisson, or zero-inflated negative-binomial versus ordinary negative-binomial). A large, positive test statistic provides evidence of the superiority of model 1 over model 2, while a large, negative test statistic is evidence of the superiority of model 2 over model 1.

4. Conclusion

The present dataset on count of species of earthworm significantly deviated from normal linear regression assumption and the logarithmic transformation of the data did not achieve the desired result. Researchers often transform the data or use non-parametric tests to analyse count data. However, these procedures have their own limitations. Alternative models are being employed here and comparisons are made among themselves in order to find appropriate models in specific sites and species. In most of the cases, the NBD and ZINB models perform better than the standard Poisson and ZIP models. The PCO model here cannot accommodate the excess zeroes in all types of earthworm species as number of zeroes is large. We do not include the PCO model here in the comparison.

We apply the vuong test for comparing goodness fit models for non-nested models (Table 4A, 4B, 4C). We compare GLP with ZIP and NBD with ZINB. In that the ZIP is significantly better than the standard Poisson GLP and ZINB is significantly better than NBD in both the null and full models in site 1. However, in site 2, though the ZIP model is still better than the GLP model in both the models, the NBD and ZINB are equally good in the null model as indicated by insignificant p-values for raw and AIC. In Site 3, in both the models ZIP is better than GLP and ZINB is better than NBD.

TABLES

Table 1: Frequency of zeroes in counts of earthworm species

Forest Type	Drawida sp (Type-1),	Drawida Japouica (Type-2),	Drawida nepalensis (Type-3),	Eutyphoeus sp. (type 4)
Site1 Reserved Forest	39%	66%	56%	47%
Site 2: Disturbed Forest	53%	68%	69%	69%
Site 3 Plantation Forest	50%	51%	56%	58%

Table 2A: Comparing Models using AIC in Site 1(Reserved Forest)

Forest type	Animal Species	Null Model					Full Model (Species count ~ Season + depth)			
Site-1 Reserved Forest		GLP	ZIP	NBD	ZINB		GLP	ZIP	NBD	ZINB
	Type-1	5118	2850	1363	**1338**		2115	1564	1245	**1164**
	Type-2	455	**437**	438	439		396	**393**	399	395
	Type-3	602	565	**556**	557		488	**483**	490	492
	Type-4	4408	2206	1203	**1174**		1712	1174	1077	**963**

GLP: Generalized Poisson, PCO: Poisson corrected for overdispersion, ZIP: Zero inflated Poisson, NBD: Negative binomial, ZINB: Zero inflated negative binomial

Table 2B: Comparing Models using AIC in Site 2(Disturbed Forest)

Forest type	Animal Species	Null Model					Full Model (Species count ~ Season + depth)			
Site-2 Disturbed Forest		GLP	ZIP	NBD	ZINB		GLP	ZIP	NBD	ZINB
	Type-1	2834	1493	987	**974**		829	**742**	771	748
	Type-2	2225	1196	**683**	685		1351	666	639	**624**
	Type-3	409	**395**	396	397		378	**366**	384	368
	Type-4	3352	1156	3353	**771**		1296	801	703	**636**

Table 2C: Comparing Models using AIC in Site 3(Plantation Forest)

	Animal Specie	Null Model					Full Model (Species count ~ Season + depth)			
Site-3 Plantation forest		GLP	ZIP	NBD	ZINB		GLP	ZIP	NBD	ZINB
	Type-1	1992	1368	**885**	887		985	878	752	**742**
	Type-2	6415	3344	1164	**1154**		2181	1815	1023	**974**
	Type-3	2423	1226	941	**908**		1014	837	790	**726**
	Type-4	4033	1479	4034	**979**		1887	1078	948	**837**

Table 3: Vuong test for non nested models to compare goodness fit among different models for count of earthworm species

Table 3A

Site 1: Reserved Forest	Null Model p-value			Full Model p-value		
	Raw	AIC Corrected	BIC Corrected	Raw	AIC Corrected	BIC Corrected
GLP Vs ZIP	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
NBD Vs ZINB	0.008	<0.001	<0.001	<0.001	0.001	0.04

Table 3B

Site 2: Disturbed Forest	Null Model p-value			Full Model p-value		
	Raw	AIC Corrected	BIC Corrected	Raw	AIC Corrected	BIC Corrected
GLP Vs ZIP	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
NBD Vs ZINB	0.38	0.06	<0.001	<0.001	<0.001	0.003

Table 3C

Site 3: Plantation Forest	Null Model p-value			Full Model p-value		
	Raw	AIC Corrected	BIC Corrected	Raw	AIC Corrected	BIC Corrected
GLP Vs ZIP	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
NBD Vs ZINB	0.0003	0.0007	0.003	<0.001	<0.001	<0.001

Table 4A: Comparison of different models in Site1 using AIC

Forest type	Animal Species	Full Model (Species count ~ Temp + moist+ pH + carbon + nitrogen)			
Site-1 Reserved forest		GLP	ZIP	NBD	ZINB
	Type-1	3309	1709	1337	**1214**
	Type-2	409	**400**	407	402
	Type-3	525	**497**	522	499
	Type-4	2818	1283	1167	**1032**

Table 4B: Comparison of different models in Site2 using AIC

Forest type	Animal Species	Full Model (Species count ~ Temp + moist+ + nitrogen)			
Site-2 Disturbed forest		GLP	ZIP	NBD	ZINB
	Type-1	1885	925	946	**854**
	Type-2	1430	745	649	**622**
	Type-3	403	**389**	395	**389**
	Type-4	1815	726	757	**656**

Table 4C: Comparison of different models in Site3 using AIC

Forest type	Animal Species	Full Model (Species count ~ Temp + moist+ Soil porosity+ pH+ carbon + p)			
Site-3 Plantation forest		GLP	ZIP	NBD	ZINB
	Type-1	1018	841	811	**788**
	Type-2	3177	1745	1095	**980**
	Type-3	1283	879	872	**776**
	Type-4	2612	1156	1045	**895**

References

[1] Dayton, C.M., 2003. *Model comparison using information measures*. J. Mod. Appl. Stat. Method 2, 281–292.

[2] Edwards, C. A. and Bohlen, P. J., 1996. *Biology and ecology of earthworm. (book 3rd edn.)*, Chapman and Hall, London(1996).

[3] James, F.C. & McCulloch, C.E.. 1990. *Multivariate Analysis in Ecology and Systematics: Panacea or Pandora's Box*? Annual Review of Ecology and Systematics, **Vol. 21**(1990),129−166.

[4] Jones, D.T., Verkerk, R.H.J., Eggleton, P., 2005. *Methods for sampling termites. In: Leather, S. (Ed.), Insect Sampling in Forest Ecosystems*. Blackwell Publishing, Oxford, UK, pp. 221–253.

[5] Johnson, J.B., Omland, K.S., 2004. *Model selection in ecology and evolution.* Trends Ecol. Evol. 19, 101–108.

[6] Lambert, D., 1992. *Zero-inflated Poisson regression, with an application to random defects in manufacturing.* Technometrics 34, 1–14.

[7] Mackenzie, D.I., Nichols, J.D., Lachman, G.B., Droege,S., Royle, J.A., Langtimm, C.A., 2002. *Estimating site occupancy rates when detection probabilities are less than one*. Ecology 83, 2248–2255.

[8] Martin, T.G., Wintel, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A., Possingham, H.P., 2005. *Zero tolerance ecology: improving ecological inference by modelling the source of zero observations.* Ecol. Lett. 8, 1235–1246.

[9] McCullagh P. and Nelder, J. A. (1989) *Generalized Linear Models.* London: Chapman and Hall.

[10] Sharon Haokip. 2014. *Ecological study of Earthworm in a subtropical forest ecosystem, Manipur*. Thesis submitted to Manipur University, (2014).

11] Sileshi, G., Mafongoya, P.L., 2006a. *The short-term impact of forest fire on soil invertebrates in the miombo.* Bidiver. Conserv. 15, 3153–3160.

[12] Sileshi, G., Mafongoya, P.L., 2007. *Quantity and quality of organic inputs from coppicing leguminous trees influence abundance of soil macrofauna in maize crops in eastern Zambia*. Biol. Fertil. Soils 43,

.