Earthworm Abundance and Habitat Characteristics – A Quasi-Poisson Model

Ksh. Anand Singh, M. Arun Singh

ABSTRACT

Poisson regression models are the most reasonable regression models while the observation is a count data. In many cases where ecological data are analyzed, 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 methods 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 investigated through various soil and environmental characteristics in three subtropical forest ecosystems of Manipur, India. Counts of different earthworm species are being observed in three subtropical forest ecosystems of Manipur; viz., Mixed Reserved forest, Disturbed forest and Plantation forest during the 12 months of the year. The count of species of a particular type of earthworm observed during the 12 months of the year is regressed on 9 soil characteristics. Three random effects Quasi-Poisson models consisting of continuous predictor variables are fitted separately for the three different sites. The main effects models and interaction effects model are separately interpreted for all the three sites. Adequacy of fitted models are checked by using diagnostic plots. Soil temperature and soil moisture are two dominant characteristics which significantly influence the abundance of earthworm in all three sites. Carbon, nitrogen and phosphorus are also significant predictors of earthworm abundance. Some interaction effects also contribute to species abundance.

Keywords: Quasi-Poisson model, Main effects, Interaction effects, Forest ecosystems.

Dr.Ksh. Anand Singh*
Assistant Professor,
Department of Statistics, Manipur University,
Imphal, India 795003
Email: kshanand_1@yahoo.com

M. Arun Singh
Research Scholar
Manipur University
*Corresponding author

INTRODUCTION

Analysis of ecological data is generally complex because of the complexity of data present in the data itself. 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. 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.

Any statistical model typically consists of two parts, a deterministic component and a stochastic component, the later usually designated as error component. Most of the differences among methods are due to differences in the assumptions about either the response variable, the deterministic component, or the error component and most are extensions or modification to the basic general linear model that expresses the response Y as linear function of predictor X, where all observed values are assumed to be independent and normally distributed with a constant variance. In the present study, we consider count of species of earthworm as response Y variable and a set of soil characteristics as vector of predictor variables X and then develop a multivariable model which takes care of all possible modelling steps.

Soil animals are of immediate concern in activities related to agriculture, forestry and environment monitoring (Sileshi 2008). However, the complexity and diversity of soil animals and habitats in which they live pose challenges to those who seeks to quantify the effects of land use and management practices on the abundance soil animals (Lavelle *et al*. 2003). Earthworms are widely distributed in most ecosystems 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. Distribution and abundance of earthworms are governed by several ecological factors viz., soil temperature, soil moisture, soil pH value, available organic matters. 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). The diversity of the earthworm community at a given locality is influenced by the characteristics of the soil, climate and 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 matter that helps in increasing soil nutrients, increase air water infiltration, soil aggregation, increase the availability of plant nutrients, worm cast as bio-fertilizer.

Source of data

In this paper, data on count of species of 12 earthworm species collected from three subtropical forests ecosystems of Manipur, India are used. The three forest ecosystem are (1) Mix reserve sub-tropical forest ecosystem located at Koirengei (24° 52'51.36''North latitude and 93°54' 49.75 East longitude and altitude 800 – 917 m above MSL); (2) Oak dominated Langol Hills (24°52 '51.6 N and 93°55'26.59 E and altitude of 797- 848 m above MSL) (3) Managed oak plantation Forest (valley area) at Mantripukhri (24°52' 52.9'' N, 93°056' 016'' E and altitude of 786 m above MSL). Data were collected in the 12 months of the year. We designated the three forest ecosystems as Site 1, Site 2 and Site 3 respectively. All three sites have different types of biotic interference. Site 1 is located 11 km from Imphal city having an area of 25 hectares and the collection site is protected from various biotic interference and Site 2 is located at Langol Hills 8 km. from Imphal city where frequent biotic interference takes place. Site 3 is managed oak plantation forest ecosystem at the valley area of Mantripukhri 4 km North of Imphal city. Number of different earthworm species are collected from six different locations (replicates) at each of the study sites during January to December. A maximum of 12 different species are found in Site 1 whereas only 6 and 4 different species are found in Site 2 and Site 3 respectively. Each replicate has a depth of 10 cm inside the soil from the surface. Some of the locations shows 0 counts of a particular species in a particular month. Data on the number of counts of species are thus obtained for the 12 months of the year during 2012 and 2013 (Haokip 2015). Altogether, there are 216 (12×6×3) sampling points. Measurements on 9 soil and environmental characteristics are recorded for each sampling point at the time of species collection. The nine soil characteristics are Soil Temperature (Temp), Soil Moisture (Moist), Soil bulk density (bdensity), Soil porosity (porosity), Soil pH (pH), Soil carbon, Soil nitrogen, Phosphorus (p) and Potassium (k).

MATERIALS AND METHODS

The primary objective of this study to investigate the abundance of earthworm species, its variability in these three different forest ecosystems of Manipur. Regression models are attempted in order to fit the count of species on soil and environmental variables. Thus, the objective is to develop a suitable multivariate model for count of earthworm species on various soil characteristics in a parsimonious environment. Different sites have different levels of soil characteristics which affect the earthworm abundance, thereby three different models in three different sites are attempted. This will include various steps demonstrated in the model building strategies.

Modelling strategies

Regression modelling of the relationship between an outcome variable and independent predictor variables is commonly employed in virtually all fields of research. The popularity of this approach is due to the fact that biologically plausible models may be easily fitted, evaluated and interpreted. Every model building strategy would include a through understanding of the data structure to choose a suitable model from a variety of models available in the literature. Statistically, the specification of a model requires choosing both systematic and error components. The choice of the systematic component involves an assessment of the relationship between an average of the outcome variable and the independent variables. The choice of an error component involves specifying the distribution of what remains to be explained after the model is fitted (the residuals).

A good place to start is to use a model with linear systematic component and normally distributed errors, the normal linear regression model. However, when the outcome variable is a count variable the assumption of normal error is often misinterpreted because the count dataset shows a skewed distribution. Thus, the use of normal linear regression for the count of species is not a suitable model.

The Poisson regression

An alternative and more appropriate model when the response variable is count data would be the Poisson regression which is a generalization of the general linear model. Poisson regression is similar to regular multiple regression except that the dependent (Y) variable is an observed count that follows the Poisson distribution. Thus, the possible values of Y are the non-negative integers: 0, 1, 2, 3 and so on. It is assumed that large counts are rare. Hence, Poisson regression is similar to logistic regression, which also has a discrete response variable. However, the response is not limited to specific values as it is in logistic regression. The approach for developing the final model will be the same except for the error component which will be assumed to be distributed as Poisson with some non-negative parameter (McCullagh and Nelder 1989).

The Poisson distribution models the probability of y incidences with the formula :

$$Pr(Y=y/\mu) = \frac{e^{-\mu}\mu^{y}}{y'} \quad (y = 0, 1, 2, 3\ldots) \quad (1)$$

The Poisson distribution is specified with a single parameter μ. This is the mean incidence rate of a rare incidence per unit of exposure. Because exposure is often a period of time, we use the symbol t to represent the exposure. The parameter μ may be interpreted as the risk of a new occurrence of the event during a specified exposure period, t. The probability of y incidences is then given by

$$Pr(Y=y/\mu,t) = \frac{e^{-\mu}(\mu t)^{y}}{y!} \quad (y = 0, 1, 2, 3\ldots) \quad (2)$$

In Poisson regression, we suppose that the Poisson incidence rate μ is determined by a set of k regressor variables (the X's). The expression relating these quantities is :

$$\mu = t \exp(\beta_0+\beta_1 x_1+\beta_2 x_2++\beta_k x_k) \quad (3)$$

The regression coefficients β_0, β_1, $\beta_2\ldots$, β_k are unknown parameters that are estimated from a set of data. Using this notation, the fundamental Poisson regression model for an observation i is written as :

$$P(Y_i=y_i/\mu_i t_i) = \frac{e^{-\mu_i t_i}(\mu t)^{y_i}}{y_i!} \quad (4)$$

Where μ_i=tiexp $(\beta_0+\beta_1\ x_{1i}+\beta_2\ x_{2i}+ ... +\beta_k\ x_{ki})$, xji denotes the jth variate of the ith observation.

That is, for a given set of values of the regressor variables, the outcome follows the Poisson distribution.

The regression coefficients are estimated using the method of maximum likelihood.

The multivariable model

We start with a bivariate analysis in the sense that the response variate i.e. the count of species is fitted in a Poisson regression with all the predictor variables (9 soil characteristics) separately i.e. one predictor variable at one time.

Where Y = Count of species, X_i = i^{th} predictor variable, i = 1,… , 9

The observed data on count of species is first checked to ascertain whether normality assumption can be valid. Using the Shapiro-wilks test for normality all three data sets corresponding to the three sites show significant p-values which indicates violation of normal distribution. Graphical verification reassures that the assumption of normality is far beyond hope. The histogram is skewed to the right which is typical of count data. Possible transformations on the count of species such as logarithmic or square root do not show any improvement to validate normality assumption. Thus, we have to take recourse to Poisson regression model in order to suitably fit the count data on the soil and environment variables. However, another difficulty arises in the present data while fitting a Poisson model. In the general Poisson distribution, the assumption is that mean and variance are equal, if not and if variance is larger than the mean, the fitting is not appropriate due to problem of overdispersion. Overdispersion is a problem which arises when the conditional variance(residual variance) is larger than the conditional mean (Faraway 2006). The present data requires to address the problem of overdispersion. Two alternative solutions are available in the literature: (a) Fit it by using the method of quasi likelihood, call it Quasi-Poisson, (b) Change the model to negative binomial distribution. We choose

the first one and fit the model using Quasi-Poisson.

The influence of a particular site on earthworm abundance is of particular interest in this study, so that effect of site is considered fixed. All other soil characteristics which are thought to be potential regressors of earthworm abundance are continuous random effects. Three separate models one for each site are attempted separately for each site in order to identify those soil characteristics which could significantly influence the abundance of earthworm species of type 1(DrawidaJapanica). All other species types are not considered in this study. In each model, analysis is based on 72 sample points which are collected from 6 different locations/replication in 12 months of the year.

First, we begin with a bivariate analysis i.e. species count is regressed on each of the soil variables taken one at a time. Those variables which are significant at 5-10% level is considered as a potential candidate in the multivariable model. We use R software package to fit the above model. The result of bivariate analysis shows that all of the soil variables are significant at 5% level, thereby indicating that all are potential candidatesto include while developing the multivariable model in all the three sites. The result of fitting bivariate analysis is shown in Table 1 for Site 1. The same analysis is also carried out for Site 2 and Site 3, but for the sake of space the results are not presented.

Next to the bivariate analysis is to attempt a full main effects model which seeks to retain all influential variable as significant predictors. We employ

Table 1 (A). Bivariate Analysis -Site 1. *All variables are significant at 5%.

Variables (Xi)	Estimate of Coeff	SE	p-value
Temp	0.103	0.01	<0.001
Moisture	0.054	0.01	<0.001
Bulk density	-5.79	2.59	0.03
Porosity	0.15	0.07	0.03
pH	-1.36	0.01	<0.001
Carbon	0.918	0.11	<0.001
Nitrogen	10.10	0.86	<0.001
P	102.88	10.8	<0.001
K	-19.74	8.70	0.03

a forward addition technique to expand the model starting from a simple one which includes only one variable. In this, we fit a Quasi-Poisson model containing only the variable "Temp" call it fit1. Next, we add one more variable "Moist" to fit1 and call it fit^2. The two models are then compared to see whether there is any improvement in the 2nd fit over the 1st fit. A significant F-value in the "anova(fit^1,fit^2, test='F')" (R command) indicates that the later fit is well improved over the earlier fit. Thus, in fit^2 with two variables shows significant improvement over fit^1. We retain both the variables. However, in fit^3 where one more variable "bdensity" is added to fit^2, we cannot see any improvement, thereby supporting to remove "bdensity" and retain the earlier fit^2. The process of adding a new variable in the existing fit and then checking if there is any improvement in the new fit over the former fit continues till all variables are exhausted. Finally, in Site 1, the main effects model consists of five factors viz., Temp, Moist, pH, carbon and nitrogen. Thus, the adjusted main effects model in Site 1 is

$$\mu_1= t.e~(\beta_0+\beta_1~Temp+\beta_2~Moist+\beta_3~pH+\beta_4~carbon+\beta_5~nitrogen)~(5)$$

The same technique is applied to Site 2 and Site 3 where we get two different main effects model as shown in equations (6) and (7) :

$$\mu_2= t.e~(\beta_0+\beta_1~Temp+\beta_2~Moist+\beta_3~nitrogen)\quad(6)$$

$$\mu_3= t.e~(\beta_0+\beta_1~Temp+\beta_2~Moist+\beta_3~porosity+\beta_4~pH+\beta_5~carbon+\beta_6P)\qquad(7)$$

Where, μ_1, μ_2 and μ_3 are the respective means in the three sites, β's are the regression coefficients.

The results of fitting the three models are presented in Tables 2(A), 2(B) and 2(C). It consists of the estimated coefficients, std. error and t- values with

Table 2(A). Site1-Main effects models.

Variable	Coeff	SE	T	p-value
(Intercept)	2.243	1.253	1.790	0.078
Temp	0.034	0.013	2.650	0.010*
Moist	0.009	0.006	1.505	0.137
pH	-0.337	0.171	-1.96	0.054
Carbon	0.327	0.103	3.161	0.002**
Nitrogen	3.249	1.306	2.48	0.015 *

associated p-values. For Site 1 three variables viz., Temp, carbon and nitrogen are significant at 5% level. The variable pH is nearly significant (at 6% level) and so we retain it. Even though the variable Moist is not significant we keep it because of its importance in the interaction effects. In Site 2, all three variables Temp, Moist and nitrogen are significant and in Site 3, Temp, carbon and P are significant. Also, the variable Moist is nearly significant (at 9% level) and so we retain it.

Before further expanding the model by including interaction effects we perform the diagnostics of the fitted models which checks for adequacy of the models. At this stage, we can well assume that the variables in the three respective models are good predictors of the count of species in the respective sites (Diagnostic plots are not shown for saving space).

The next step is to see whether there is any interaction effect of each of the predictors in the main effects models which could significantly influence the earthworm counts so that we can arrive at a final multivariate interaction model. Here, we will consider only two factor interaction. Only those variables which are significant at 5% are considered for inclusion in the interaction. The reason is that an interaction term cannot be significant if the any of its factors is not significant.

Final Interaction models

The method of forward addition is again employed

Table 2(B). Site 2-Main effects model.

Estimate	Estimate	SE	T	p-value
(Intercept)	-0.91	0.320	-2.861	0.005 **
Temp	0.043	0.011	3.881	<0.001 ***
Moist	0.017	0.008	2.015	0.047 *
Nitrogen	8.490	1.463	5.801	<0.001 ***

Table 2(C). Site3-Main effects models.

Variables	Coeff	SE	T	p-value
Intercept	-6.20	0.91	-6.83	<0.001**
Temp	0.04	0.014	2.93	0.004 **
Moist	0.018	0.01	1.16	0.085
Carbon	1.63	0.38	4.20	<0.001**
p	57.09	20.43	2.79	0.006**

here in order to arrive at a final interaction model. In Site1, only one interaction is considered important i.e. carbon : nitrogen. Other interactions do not show any significant improvement in the final model, so we remove them (Table 3(A)). In Site 2, also one interaction is included i.e. Temp:Moist (Table 3(B)) and in Site 3, none of the interactions are significant. In other words, by adding any of the possible interactions the fit does not improve the model. Hence, we retain only the main effects model, (Table 3(C)).

Adequacy of fitted models

While interpreting a fitted model it is first checked for model adequacy in order to verify whether the data fits well into the proposed model. The steps for checking model adequacy is similar to the one already discussed in the main effects models. A graphical display of diagnostic plot is shown in Figure1 for Site

1, where we can, at this stage assume that there is no evidence of lack of an adequate fit. The residuals vs fitted plot shows the fit is good as there is no obvious pattern on the line and the residuals are more or less equally spread around the horizontal line. The normal Q-Q plot supports that the residuals are normally distributed. Further there are no influential observations in the fit (Cook's distance). The diagnostic plots for Site 2 (eqn. (9)) and Site 3 (eqn. (10)) also support reasonably good fits which for the sake of space, are not shown.

RESULTS AND DISCUSSION

Model summary

Site 1: Mix reserve sub-tropical forest ecosystem located at Koirengei protected from frequent biotic interference.

$$E(y) = \mu_1 = e^{(-2.73+0.0267Temp+0.012Moist-0.248pH+1.86carbon+18.33\,nitrogen-4.96carbon*nitrogen)} \tag{8}$$

Site 2: Disturbed forest atLangol Hills- disturbed forest ecosystem where frequent biotic interference takes place.

$$E(y) = \mu_2 = e^{(-2.745+0.125Temp+0.069Moist+8.922nitrogen-0.002Temp*Moist)} \tag{9}$$

Site 3: Managed Oak Plantation Forest ecosystem at Mantripukhri.

$$E(y) = \mu_3 = e^{(-6.2+0.04Temp+0.02Moist+1.63carbon+57.1P)} \tag{10}$$

The results of fitting the final interaction models in equations (8), (9) and (10) are presented in Tables 3(A) through Table 3(C). In Site 3 no interaction term is significant. The tables consist of the estimated

Table 3 (A). Final Interaction Model for Site-1.

Variable	Coeff	SE	T	p-value
(Intercept)	-2.7314	2.408	-1.134	0.261
Temp	0.0267	0.012	2.059	0.043 *
Moist	0.0122	0.006	1.876	0.065
pH	-0.2486	0.170	-1.456	0.150
Carbon	1.8654	0.647	2.879	0.005**
Nitrogen	18.329	6.411	2.859	0.005**
Carbon: Nitrogen	-4.964	2.071	-2.397	0.019 *

(Dispersion parameter for Quasi- Poisson family taken to be 2.136168).
Null deviance: 739.74 on 71 degrees of freedom,
Residual deviance: 141.85 on 65 degrees of freedom,
AIC: NA.

Table 3 (B). Final Interaction Model for Site-2.

Variable	Coeff	SE	t	p-value
(Intercept)	-2.745	0.7116	-3.859	<0.001***
Temp	0.1258	0.0304	4.131	<0.001 ***
Moist	0.0699	0.0203	3.440	0.0010 **
Nitrogen	8.9223	1.3874	6.431	<0.001 ***
Temp : Moist	-0.002	0.0008	-2.861	0.005 **

(Dispersion parameter for Quasi- Poisson family taken to be 1.366089).
Null deviance: 405.139 on 71 degrees of freedom,
Residual deviance: 88.914 on 67 degrees of freedom,
AIC: NA.

Table 3 (C). Final main effects models for Site-3.

Variables	Coeff	SE	T	p-value
Intercept	-6.20	0.91	-6.83	<0.001**
Temp	0.04	0.014	2.93	0.004 **
Moist	0.018	0.01	1.16	0.085*
Carbon	1.63	0.38	4.20	<0.001**
P	57.09	20.43	2.79	0.006**

(Dispersion parameter for Quasi- Poisson family taken to be 2.912755).
Null deviance: 610.08 on 71 degrees of freedom,
AIC : NA.

coefficients (β's), the standard error (s.e.), t values and associated p-values. In Table 3 (A), the variables Temp, carbon and nitrogen are significant at 5% level along with the interaction carbon : nitrogen, but Moist, and pH are not significant. Moist and pH are still retained in the model due to their importance as a predictor. In Table 3(B), all variables, Temp, Moist and nitrogen are significant along with the interaction Temp:Moist. In Table 3 (C), all variables Temp, Moist carbon and P are significant.

All predictor variables in these models are all continuous variables. The estimated coefficient for a predictor represents the change in the link function for each unit change in the predictor variable while the other predictors in the model are held constant. Generally, positive coefficient makes the response more likely (to increase) while negative coefficients make it less likely. An estimated coefficient near 0 indicates that the effect of the predictor is small. In the models in equations (9), (10) and (11), the link function is the natural logarithm there by the relationship between the response variable and the predictors can be done by exponentiation. As for example in Site 1 the coefficient for Temp is 0.026 which shows that a one unit change in the soil temperature results in the mean number of earthworms increased by

exp (0.026) = 1.03 times. However, the coefficients, for soil pH is -0.24, which indicates that a one unit increase in soil pH value will result in the decrease of exp (-0.24)= 0.79 times in the mean no. of earthworms. The interaction terms cannot be interpreted in this way as the relationship of one predictor to the response variable depends on the other term of the interaction. We can do similar interpretation of the coefficients in the other two sites.

The effect of soil temperature plays an important role in the earthworm species count of type 1 in all the three sites. However, in disturbed forest ecosystem soil temperature affects the species count more dominantly than the other two forest ecosystems. Soil moisture is not significant in Site 1 and Site 3 but it is significant in Site 2. Soil carbon content and nitrogen are highly significant in Site 1. The positive coefficients indicate that these variables contribute to the increase in species counts as their values increase. The interaction effect of carbon and nitrogen is also significant in Site 1. Soil pH value does not show significant contribution to the species count in Site 1. Soil nitrogen has significant contribution in Site 2 whereas soil carbon has significant contribution to species count in Site 3. In Site 2 the interaction effect
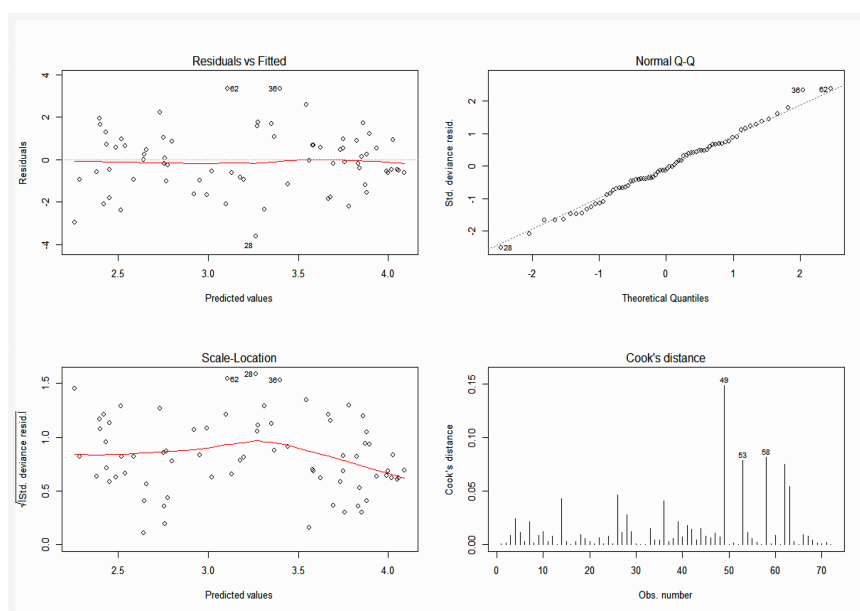

Fig. 1. Diagnostic plots in Site- 1 (Final Interaction Model).

of Temperature and moisture is significant whereas none of the interactions is significant in Site 3. In Site 3 P (phosphorus) also contributes significantly

While ecologists are often concerned with the study of species abundance and diversity, the variations in abundance and diversity of species across different habitats is measured by the species count data. However, 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 data set. 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 data set is referred to as zero inflated. The present data set exhibits over dispersion due large number of zero counts which poses crucial issues while fitting the ordinary Poisson model. A solution to the problem is to take care of the excess variance. Thus, a Quasi-Poisson model is being fitted here. According to the different soil habitat characteristics in the three different forest ecosystems which we have mentioned above three different multivariate models are obtained. While alternative methods to tackle the excess zero problem are available in the literature improved models will be attempted in our next paper.

REFERENCES

Edwards CA, Bohlen PJ (1996) Biology and ecology of earthworm.Book. 3rd edn. Chapman and Hall, London.

Faraway Julian J (2006) Extending the Linear Model with R, Generalized Linear, Mixed Effects and Non-parametric Regression Models. Taylor and Francis.

Haokip S (2015) Ecological study of earthworm in a subtropical forest ecosystem, Manipur. Thesis submitted to Manipur University, CanchipurImphal, India.

Lavelle P, Senapati B, Barros E (2003) Soil macrofauna. In: Schroth G Sinclair FL (eds). Trees,Crops and Soil Fertility : Concepts and Research Methods. CAB International, Wallingford 303—323.

McCullagh P, Nelder J (1989) Generalized Linear Models. 2nd edn. London : Chapman and Hall.

Sileshi G (2008) The excess zero problem in soil animal count data and choice of appropriate models for statistical inference. Pedobiologica 52 : 1 ——17.