BIODIVERSITAS
Volume 19, Number 5, September 2018
Pages: 1871-1876

ISSN: 1412-033X
E-ISSN: 2085-4722
DOI: 10.13057/biodiv/d190538

Latent variable models for multi-species counts modeling in ecology

RIKI HERLIANSYAH[1,♥], IRMA FITRIA[2]
[1]Department of Mathematics, Institut Teknologi Kalimantan. Kampus ITK Karang Joang, Balikpapan 76127, East Kalimantan, Indonesia
Tel.: +62-542-8530801, Fax.: +62-542-8530800, ♥email: rherliansyah@itk.ac.id

Manuscript received: 10 July 2018. Revision accepted: 18 September 2018.

Abstract. *Herliansyah R, Fitria I. 2018. Latent variable models for multi-species counts modeling in ecology. Biodiversitas 19: 1871-1876.* High-dimensional multi-species counts are often collected in ecology to understand the spatial distribution over different locations and to study effects of environmental changes. Modeling multivariate abundance is challenging as we need to consider the possibility of interactions across species. Latent variable models are the recent popular approaches in statistical ecology to address such issue that has a similar framework to ordinary regression models. In this paper, we employed the poisson distribution for modeling count responses and a negative binomial distribution for more frequent zeros in observations. The implementation of a latent variable model, Generalized Linear Latent Variable Models (GLLVMs), was demonstrated on multi-species counts of endemic bird species collected in 37 different sites in Central Kalimantan, Indonesia. The main objectives were to study the effect of logging activities on abundance of endemic species and their interactions and to observe the habitat preference of certain species. Our study found that out of four endemic species, Alophoixus bres and Eurylaimus javanicus species were significantly affected by logging activities. The sign of parameters was negative indicating the logging activities in 1989 and 1993 bring significantly negative impacts on those species. The interaction created among species was strongly negative for major endemic species especially Alophoixus bres and Eurylaimus javanicus that prefer living in primary forest than in logging areas.

Keywords: Multi-species counts, latent variable, endemic species

INTRODUCTION

Studying a spatial distribution of a group of species and their interactions with the ecosystem becomes the main objective in species modeling. Various researches regarding species richness and biodiversity, in Indonesia especially, have been widely carried out to understand the behavior of species towards changes in environment. Janiawati et al. (2016), for instances, studied the relationship between environmental characters and 21 species of reptiles in Gianyar Regency, Bali; Pritchett et al. (2016) analyzed 20 species of rattan palms to observe the behavior of rattan species towards edaphic niches; Kaban et al. (2017) recorded 40 bird species Gunung Walat, West java, Indonesia to explain the response of birds to various plantation forests; Kurniawan et al. (2018) studied Arthropod community of Semedi Show Cave in Gunungsewu Karst Area, Pacitan, East Java, Indonesia. These researches successfully collected groups of species (multivariate counts) during the observations but statistical tools used were limited to the analysis in univariate cases or to simple descriptive statistics to explain the data. Working on multivariate cases, however, where large groups of species are collected is still minor and difficult to do. In order to do this, a joint statistical model is necessarily required as we need to induce the correlations across species into the model; interactions among species are not independent.

In statistical ecology, one model that could explain multivariate inference and has been rapidly expanded in a wide range of applications is latent variable models. Recent applications of latent variable models on species modeling can be seen in Warton et al. (2015), Thorson et al. (2016), Ovaskainen et al. (2016) and Caraka et al (2018). Latent variables are used in the model to explain the unobserved quantities in environment and to incorporate the interactions across species. In this paper, we introduce a latent variable model, Generalized Linear Latent Variable Models (GLLVMs), for modeling multivariate count data. GLLVMs is an extension of Generalized Linear Models, ordinary regression models, that similarly aims to study the effect of explanatory variables and can also be used for species ordination (Hui et al. 2015). The distributional choices for multivariate count responses considered in this paper for modeling were Poisson and negative binomial distributions. These distributions have been shown to fit count data types better for GLLVMs (Warton 2005).

The main challenge in most latent variable models that it is complicated to use especially for non-statistical background users since the marginal likelihood function is not straightforward, involving the integration on latent variables. Hence, certain approaches are required to approximate the function that cannot be straightforwardly used in practice. Laplace approximation has been popularly employed to estimate parameters of GLLVMs (Huber et al. 2004; Niku et al. 2017). Hui et al. (2016) proposed a variational approximation that produced similar outcomes in terms of accuracy and computation to Laplace approximation. A recent study by Herliansyah et al. (2017) regarding how to improve computational issues of GLLVMs using Template Model Builder (TMB) leads to unexpected outcomes. This research later was used to

create a new package designed especially for fitting GLLVMs either using Laplace or variational approximations for various choice of distributions and easy to implement.

To demonstrate the application of GLLVMs, we used multivariate endemic bird species collected in three different habitat structures in Central Kalimantan, Indonesia (Cleary et al. 2005). Our objectives were to study the effect of logging activities in 1989 and in 1993 on bird abundance, to explain the interaction among endemic species (whether the relationship is random or associated), to show spatial distributions of endemic species over three habitats (species ordination) and to obtain species clustering. The rest of this paper is structured as follows. Section 2, we provide a brief description of data used for modeling followed the idea of GLLVMs in next section. Section 3 presents the application of GLLVMs under assumptions of poisson and negative distributions on responses with the last section containing discussions and conclusions.

MATERIALS AND METHODS

Data of endemic species

Indonesia is a country that has a higher number of endemic birds than any other country. This condition is supported by the size, tropical climate, and archipelagic of Indonesia. The data of endemic species used in this paper refers to Cleary et al. (2005) that collected data from Indonesia. Based on that paper, the sampling was collected in the area that represents the natural vegetation and the regional topography of the inland, upstream region in Kalimantan. Specifically, the data was obtained in 300,000 ha Kayu Mas logging, near Sangai, Central Kalimantan, from June until October in 1997 and 1998 respectively. Thirty-seven samples were collected from habitat classes, unlogged primary forest, and forest logged in 1997 and 1993-1994. The total area was about 196 km2.

The study found over 170 different bird species observed in three different habitat classes where seven of them were endemic species as presented in Table 1. In this paper, we consider four of seven endemic species used for modeling species with sums of more than 10. Due to a large number of zeros from the remaining three species, we decided to exclude them from the model as it would affect the estimation. Finally, habitat classes, primary and logged forests, are the only explanatory variables used in the model. See Cleary et al. (2005) for more details.

Statistical analysis

Generalized linear latent variable models (GLLVMs) is a statistical model that can be thought as an extended version of Generalized linear model (GLM) with the addition of random effects (Moustaki 1996) and often be used for species ordination. The random effects in GLLVMs are described as hidden variables or unobserved environmental factors and used to induce correlations across species. Let Y_{ij} is count responses with $i = 1,2,\dots,n$ being the site where data are collected and $j = 1,2,\dots,p$ being the number of species. The functional relationship between mean responses and the linear predictor, η_{ij} is defined by

$$\mathrm{E}(Y_{ij}) = \mu_{ij} = g^{-1}(\eta_{ij})$$

With $g(.)$ is a link function. Linear components of the predictor are similar to GLM structures with the inclusion of multivariate random effects as follows:

$$\eta_{ij} = \alpha_i + \boldsymbol{\beta}_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\lambda}_j$$

Where: α_i represents the site variation or row effect treated as fixed parameters, β_j contains a matrix of the regression coefficient to corresponding independent variables, $\boldsymbol{x}_i'$. In many papers, the distributional choice of latent variables, $\boldsymbol{u}_i$, is a normal distribution with mean zero and constant variance and assumed to be independent of each other. These latent variables are often used for species ordination, to show distributions of species across sampling sites. The term $\boldsymbol{u}_i'\boldsymbol{\lambda}_j$ is a random component that has the variance-covariance matrix $\boldsymbol{\Sigma}$ controlling the correlations across species:

$$\boldsymbol{\Sigma} = \boldsymbol{\lambda}_j'\boldsymbol{\lambda}_j$$

Where the number of latent variables is less than the number of species, $q<p$. Loading factors λ_j represents parameters connecting the unobserved environmental variables to responses.

Table 1. Data of endemic species collected at thirty-seven sites from three different habitat classes in Central Kalimantan, Indonesia with references to Cleary et al. (2005)

Species/habitats	Primary forest (P)	Logged forest in 1989 (L89)	Logged forest in 1993 (L93)	Total
Alcedo euryzona	0	0	1	1
Alophoixus bres	85	21	35	142
Chloropsis cochinchinensis	41	46	66	160
Eurylaimus javanicus	17	8	7	32
Hemicircus concretus	2	0	1	3
Meiglyptes tristis	5	7	6	18
Pellorneum capistratum	8	2	0	10

In GLLVMs, we also need to choose the link function and distributions for responses. In this paper, we only consider poisson and negative binomial distributions for modeling multi-species counts. A negative binomial distribution was proven to be more appropriate choices than poisson and zero inflated distributions for more frequent zeros in data (Warton 2005). The choice of link function can be based on selected distributions for responses. See details in Skrondal and Rabe-Hesketh (2004). For poisson and negative binomial distributions, the obvious choice for a link function is the log link. Hence, the relationship between mean responses and a linear predictor can be rewritten as

$$\mu_{ij} = \exp\left(\alpha_i + \boldsymbol{\beta}_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\lambda}_j\right)$$

Where, for the poisson regression the variance is equal to its mean, μ_{ij}. For a negative binomial model, the variance is equal to $\mu_{ij} + \phi_j\mu_{ij}^2$ with ϕ_j being the overdispersion parameters.

To assess the adequacy of models, residuals of the model are often presented. The similarity between GLLVMs and GLMs leads to the same issue in assessing goodness of fit. For most cases in GLMs, the model does not have a constant variance and a zero mean especially when overdispersion occurs. Pearson and deviance residuals are two common used residuals for diagnostic checking. For overdispersion cases, however, both residuals are often not normally distributed (Dunn and Smyth 1996). Hence, in this paper we use Dunn-Smyth or quantile residuals for assessing models. Residuals are defined as follows (Hui et al 2015):

$$r_{ij} = \Phi^{-1}\left(z_{ij}F_{ij}(y_{ij}) + (1 - z_{ij})F_{ij}^{-}(y_{ij})\right)$$

Where: $\Phi(.)$ and $F_{ij}(.)$ are the cumulative density functions of a standard normal distribution and responses, y_{ij} and z_{ij} being generated from a standard uniform distribution.

To fit the model, we use a new created package in R, gllvm(), designed especially for fitting GLLVMs with the help of Template Model Builder (TMB). See details in Kristensen et al. (2015) for more information about TMB. This package can be found at https: //cran.r-project.org/package=gllvm. The choice of distributions for fitting GLLVMs available in this package are a binomial distribution for binary responses, poisson, negative binomial and zero inflated poisson distributions for count data with two optional approaches, Laplace and variational approximations.

RESULTS AND DISCUSSION

In this section, we fitted GLLVMs on endemic bird data to study the species distribution over habitat classes, the environmental effects, species interaction, species prediction and species clustering. To begin with, we present descriptive statistics of data in the following figure. Distributions of abundance of endemic species over three habitat structures showed in Figure 1.A are roughly identical each other with slightly higher mean and more outliers in primary forest. Each species was distributed variously across sampling sites given in Figure 1.B. *Meiglyptes tristis* and *Eurylaimus javanicus* were found to be less abundant than other two species with average numbers of individual close to zero.

To fit GLLVMs, we assumed poisson and negative binomial distributions for responses and a fixed number of latent variables, $q = 2$ for ordination purposes while it can also be selected based on either information criteria, AIC or BIC, or coverage probabilities. We used habitat structures as explanatory variables defined as follow: D_1 (logged forest in 1989 = 1; otherwise = 0) and D_2 (logged forest in 1993 = 1; otherwise = 0) with primary forest as the reference variable (P = 0). Predicted values of abundance across sites, $\hat{y}_i = \sum_{j=1}^{p}\hat{y}_{ij}$, were computed and compared to actual counts given in Figure 2. Both models seem fit data very well where predicted values lie closely to actual counts. There is almost no distinction between poisson and negative binomial models. Either poisson or negative binomial models can be selected in this case for modeling; diagnostic checking on residuals, however, can be used for model selections.

In this paper we use Dunn-Smyth residuals for assessing the adequacy of models as explained in previous section. A good model is supposed to be independently and normally distributed as in GLM. As we can observe, residual models for a negative binomial model are randomly spread over linear predictors, no clear pattern, while a poisson model produce fan-shaped pattern as an indication of overdisversion (Hui et al. 2015). Both residuals, a negative binomial and a poisson regressions lie closer to a normal distribution line with no major outlier is found in the model. In conclusion, a negative binomial model is a more appropriate model to use for further analysis; testing our hypothesis about logging effects on abundance and species ordination as we observe no sign for a lack of fitting.

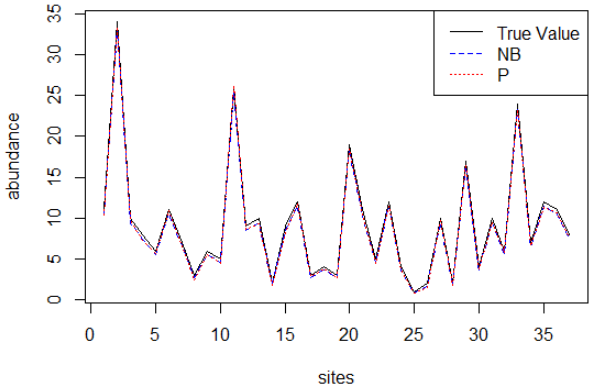
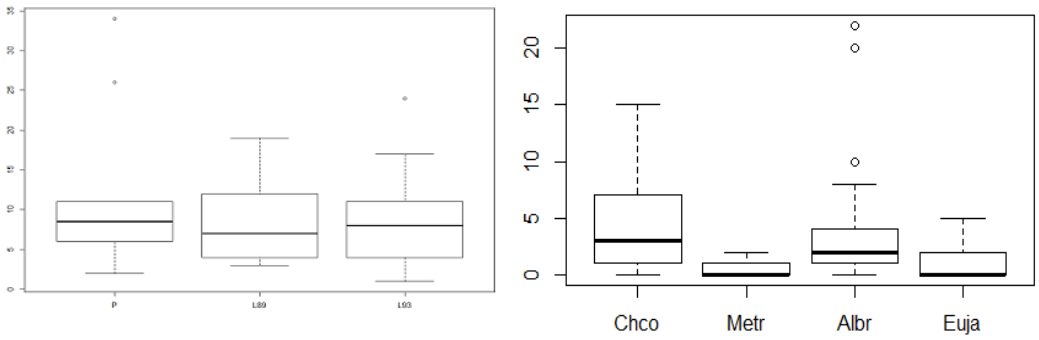


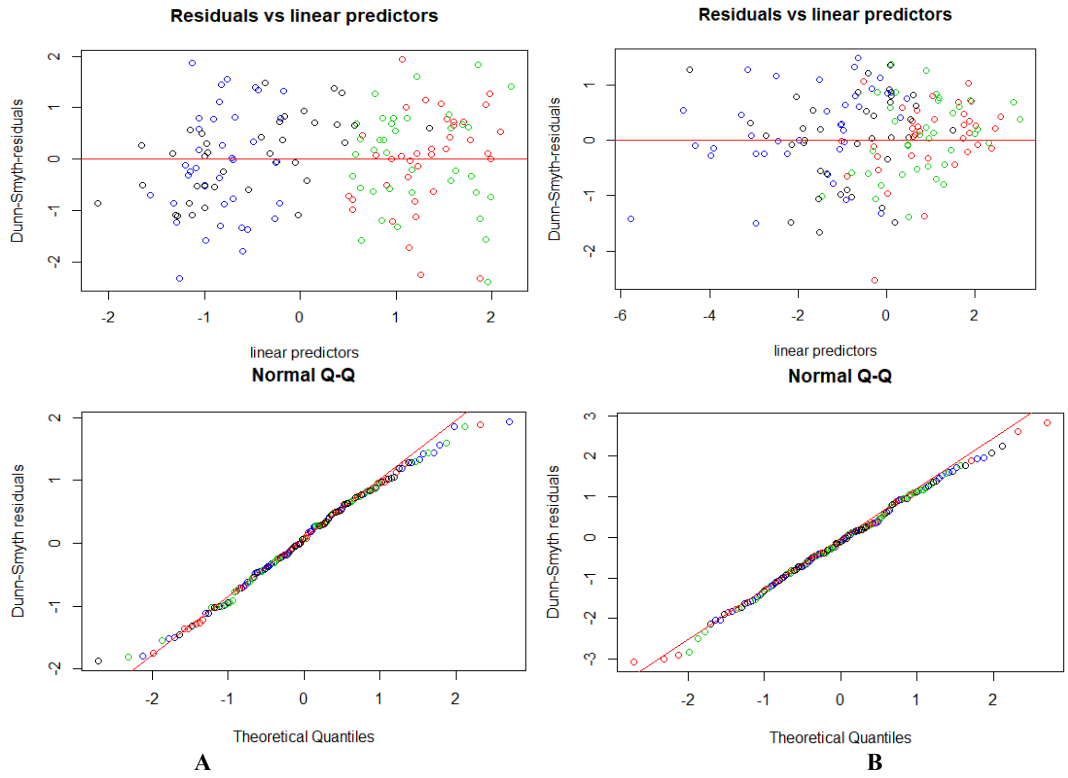
Figure 2. Comparisons of observed values and predicted values of poisson and negative binomial regression models

To answer our questions in introduction regarding the logging effects on abundance of endemic bird in Central Kalimantan, we run a hypothesis testing for two dummy variables defined before. Figure 4 displays parameters of four endemic species with 95% confidence interval for poisson and negative binomial regression models. If 95% confidence intervals contain zero then the corresponding parameter is not statistically significant. As we can observe from the following figure, two models give the same conclusion about species that were significantly affected by logging activities. For a dummy variable, D_1, we found that *Alophoixus bres* was the only species receiving a significant effect from logging activities in 1989. The parameter lies in the negative area indicating that logged forest in 1989 brought negatively a significant impact on abundance of *Alophoixus bres*. Logged forest in 1993 also affected the same species as in 1989, *Alophoixus bres*. This species received a negative effect on its abundance, while remaining species seem not too much different over three habitat structures.



A B

Figure 1. Distributions of endemic species for (A) different habitats: Primary forest, logged forests in 1989 and 1993 respectively (B) each species



A B

Figure 3. Diagnostic checking on residuals model: (A) Dunn-Smyth residuals of a negative binomial model, (B) Dunn-Smyth residuals of a poisson model

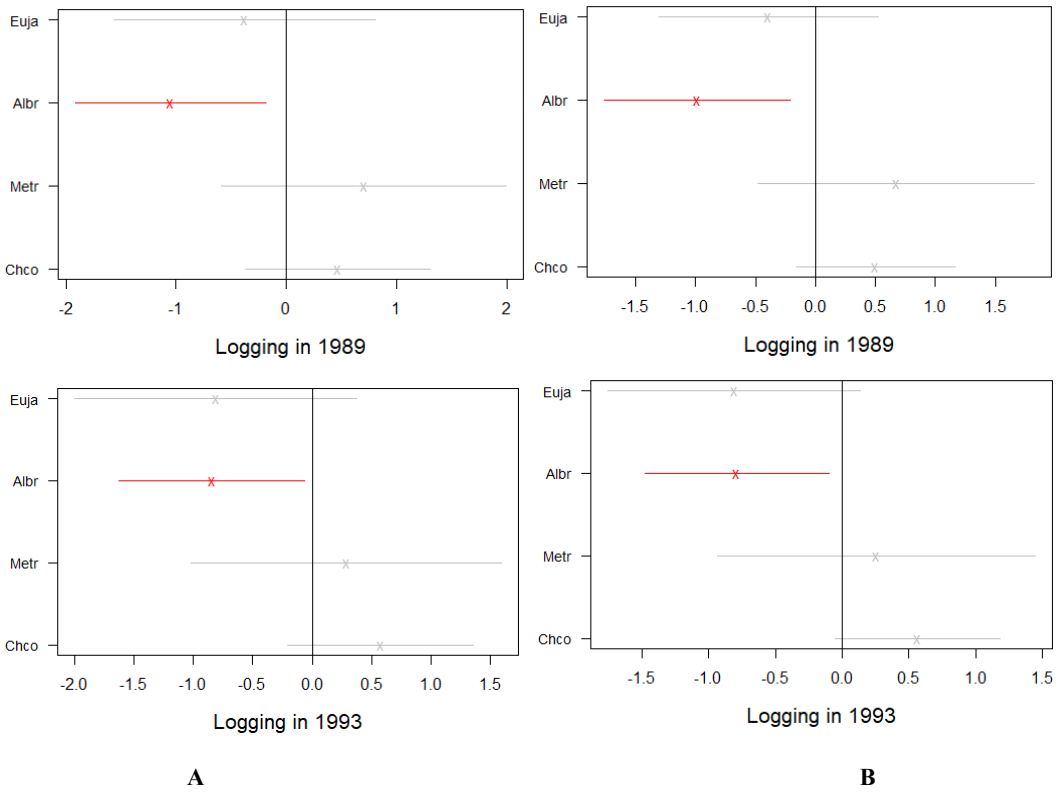


Figure 4. Estimation of parameters and their 95% confidence intervals: (A) a negative binomial regression and (B) a poisson regression.

Finally, GLLVMs also can be used for species ordination like those in multidimensional scaling and principal component analysis and for explaining interactions among species. Estimated latent variables were used to create ordination species while the correlation matrix describing association among species was computed through variance-covariance matrix using loading factors, $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\lambda}}'\hat{\boldsymbol{\lambda}}$. To create ordination plot, we used an unconstrained model, a model without explanatory variables, and loading factors used to create correlation matrix were loading factors after controlling habitat structures effects, a model with D_1 and D_2. Figure 5.A shows the distributions of endemic specie over three different locations. As we can see, distributions of endemic species in logged forests in 1989 and in 1993 were roughly similar over different sites. In primary forest, however, some sites were located further from logged forest indicating a different pattern of species distribution from those two habitats. For instance, *Eurylaimus javanicus* was found to be more abundant in one site of primary forest while *Meiglyptes tristis* was observed more in one logged forest in 1989 sites. Most species seem to be strongly associated with more negative interactions. Positive correlations indicate endemic species tend to be close to each other, their existence is positively relied on each other while negative correlations imply endemic species prefer to stay further from one to another. Figure 5.C presents species clustering of endemic species in Central Kalimantan. This clustering was measured based on the Euclidean distance

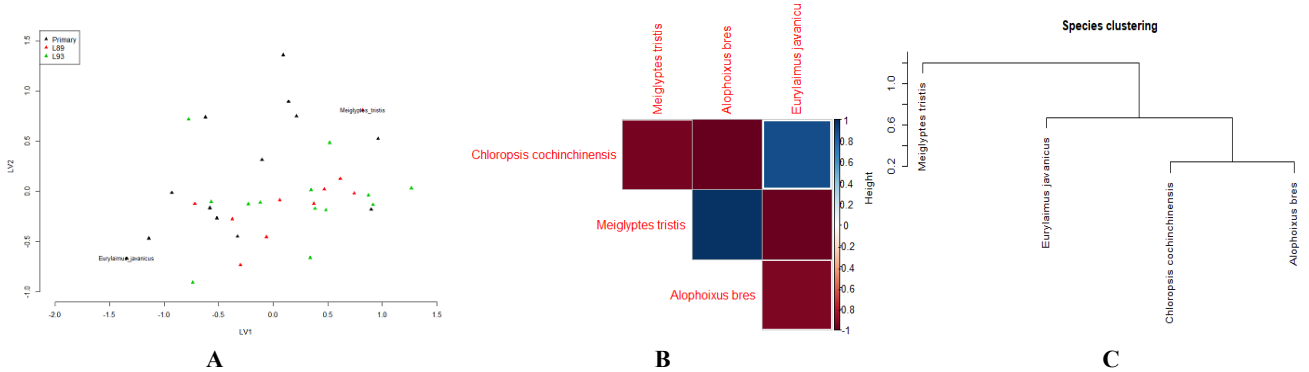


Figure 5. Ordination species using latent variables (A), correlation matrix (B) and species clustering (C) using loading factors for a negative binomial model

More information can be extracted using GLLVMs from data compared to univariate cases e.g., species distributions, effects of environmental changes on abundance, species interactions, the prediction and the species clustering. In this paper, we only use four endemic species as responses for a demonstration while in practice it could handle larger numbers of species. Our results show that a negative binomial seems fit the data better, satisfying normal assumption and more random than a poisson model even though they lead to the same inferences. Logging activities have been shown to negatively affected abundance of some endemic species where most species tended not to interact each other. Remaining endemic species, however, gave a positive response to logging activities regarding their insignificant parameters. We also describe the relationship between mean responses and only one independent variable, habitat structures while more environmental factors can be included into to the model to analyze either qualitative or quantitative types. Further works in species modeling, GLLVMs can be extended to explain the association between species traits and environmental variables also known as a forth corner model where we included species traits into the model and their interactions with environments as explanatory variables (Brown et al. 2014).

ACKNOWLEDGEMENTS

This research was supported and funded by Institute for Research and Community Service (LPPM), Institut Teknologi Kalimantan, Indonesia.

REFERENCES

Brown A, Warton DI, Andrew N, Binns M, Cassis G and Gibb H. 2014. The fourth-corner solution - using predictive models to understand how species traits interact with the environment. Meth Ecol Evol 5: 344-352.

Caraka RE, Shohaimi S, Kurniawan ID, Herliansyah R, Budiarto A, Sari SP and Pardamean B. 2018. Ecological show cave and wild cave: Negative binomial gllvm's arthropod community modelling. Procedia Comput Sci 135, 377-388.

Cleary DFR, Genner MJ, Boyle TJB, Setyawati T, Angraeti CD, Menken SBJ. 2005. Associations of bird speciess richness and community composition with local and landscape-scale environmental factors in Borneo. Landsc Ecol 20: 989-1001.

Dunn K, Smyth GK. 1996. Randomized quantile residuals. J Comput Graph Stat 5: 236-244.

Herliansyah R, Brook W, Warton DI. 2017. Fast estimation of generalized linear latent variable models. [Thesis]. University of New South Wales, Sydney.

Huber P, Ronchetti E, Victoria-Fese MP. 2004. Estimation of generalized linear latent variable models. J R Stat Soc B 66: 893-908.

Hui FKC, Taskinen S, Pledger S, Foster SD, Warton DI. 2015. Model-based approaches to unconstrained ordination. Meth Ecol Evol 6: 399-411.

Hui FKC, Warton DI, Ormerod JT, Haapaniemi V, Taskinen S. 2016 Variational approximations for generalized linear latent variable models. J Comput Graph Stat. DOI: 10.1080/10618600.2016.1164708.

Janiawati IAA, Kusrini MD, Mardiastuti A. 2016. Structure and composition of reptile communities in human modified landscape in Gianyar Regency, Bali. Hayati J Biosci 23 (2): 85-91.

Kaban A, Mardiastuti A, Mulyani YA. 2017. Response of bird community to various plantation forests in Gunung Walat, West Java, Indonesia. Hayati 24 (2): 72-78.

Kristensen K, Nielsen A, Berg CW, Skaug H, Bell B. 2015. TMB: Automatic differentiation and laplace approximation. J Stat Software. arXiv: 1509.00660v1.

Kurniawan ID, Rahmadi C, Caraka RE, Ardi TA. 2018. Short Communication: Cave-dwelling Arthropod community of Semedi Show Cave in Gunungsewu Karst Area, Pacitan, East Java, Indonesia. Biodiversitas 19 (3): 857-66. 3.

Kurniawan ID, Soesilohadi RCH, Rahmadi C, Caraka RE, Pardamean B. 2018. The difference on Arthropod communities' structure within show caves and wild caves in Gunungsewu Karst area, Indonesia. Ecol Environ Conserv 24 (1): 81-90.

Moustaki I. 1996. A latent trait and a latent class model for mixed observed variables. British J Mathematical and statistical Psychology 49: 313-334.

Niku J, Warton, DI, Hui Francis K.C, Taskinen S. 2017. generalized linear latent variable models for multivariate abundance data in ecology. J Agric Biol Environ Stat 22: 498-552.

Ovaskainen O, Abrego N, Halme P, Dunson D. 2016. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. Meth Ecol Evol 7: 549-555.

Pritchett R, Phillips A, Mardiastuti A, Powling A. 2016. Rattan diversity and broad edaphic niches in a tropical rainforest of Buton, Sulawesi, Indonesia. Reiwardtia 15 (2): 99-110.

Skrondal A, Hesketh-Rabe S. 2004. Generalized Latent Variable Modeling. Chapman & Hall/CRC, Boca Raton.

Thorson JT, Ianelli JN, Larsen EA, Ries L, Scheuerell MD, Szuwalski C, Zipkin EF. 2016. Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. Global Ecol Biogeogr 25: 1144-1158.

Warton DI, Blanchet FG, O'Hara RN, Ovaskainen O, Taskinen S, Walker SC, Hui FKC. 2015. So many variables: joint modeling in community ecology. Trends Ecol Evol. DOI: 10.1016/j.tree.2015.09.007.

Warton DI. 2005. Many zeros does not mean zero inflation: comparing the goodness-of-it of parametric models to multivariate abundance data. Environmetrics 16: 275-289.