International Journal of Statistics and Probability; Vol. 3, No. 3; 2014
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Bayesian Approach Using Latent Variable for Zero Truncated Poisson Distribution: Application for Species-Area Relationship

Claude Thiago Arrabal[1], Marinho Gomes de Andrade Filho[2] & Karina Paula dos Santos Silva[3]

[1] Department of Statistic, Federal University of São Carlos, São Paulo, Brazil

[2] Department of Applied Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

[3] Institute of Mathematic and Statistics, University of São Paulo, São Paulo, Brazil

Correspondence: Claude Thiago Arrabal, Department of Statistic, Federal University of São Carlos, Rodovia Washington Luis, km 235 , São Carlos - SP, Brazil. Tel: 55-11-957-477-819. E-mail: claude.arrabal@gmail.com

Received: April 7, 2014 Accepted: May 31, 2014 Online Published: June 11, 2014

doi:10.5539/ijsp.v3n3p18 URL: http://dx.doi.org/10.5539/ijsp.v3n3p18

Abstract

In ecology, understanding the species-area relationship (SARs) is extremely important to determine species diversity. SARs are fundamental to evaluate the impact in this diversity due to destruction of natural habitats, to create biodiversity maps and to determine the minimum area to preserve. In this study, the number of species is observed in different area sizes. These studies are referred in the literature through nonlinear models without assuming any distribution of the data. In this situation, it only makes sense to consider areas in which the number of species is greater than zero. As the dependent variable is a count data, we assume that this variable comes from a known distribution for discrete positive data. In this paper, we used the zero truncated poisson distribution (ZTP) to represent the probability distribution of the random variable "species diversity" and we considered some nonlinear models to describe the relationship between species diversity and habitat area. Among the proposed models in literature, we considered the Arrhenius power function, Persistence function (P1 e P2), Negative Exponential and Chapman-Richards to describe the abundance of species. In this paper, we take a Bayesian approach to fit models. With the purpose of obtaining conditional distributions, we propose the use of latent variables to implement the Gibbs Sampler. In order to progress using the best possible models for data, a comparison of performance between models referred in this paper will be verified through the criteria Extended Akaike Information Criterion (EAIC), Extended Bayesian Information Criterion (EBIC), Deviance Information Criterion (DIC) and Conditional Predictive Ordinate Criterion (CPO). In addition to selecting the best model, it will also assist to define the best selection criterion.

Keywords: zero truncated poisson distribution, species-area relationship, nonlinear models, count data, latent variables, gibbs sampling

1. Introduction

One of the fundamental aspects from ecology is the relationship between habitat area and species diversity (Lomolino, 2000). This relationship is essential to understand the biological distribution of species diversity and it is determined by counting the number of distinct species in different sized areas. This studies, known as species - area relationship (SAR), are one of the most important tools to create biodiversity maps, to establish relationships between extinction and migration rates and to determinate minimum size areas for species preservation (Arrhenius, 1921; Ulrich et al., 2003). Generally, only one type of organism is observed; the fish diversity in a lagoon, for example. In order to describe SAR, several nonlinear models are suggested in literature (Tjorve, 2003). However, few have been published about comparing different adjustment models. Among suggested models, we are going to consider Arrhenius model, the Persistence models (P1 and P2), Negative Exponential and Chapman-Richards to describe the species abundance (Dengler, 2009).

In these models, the parameter β_0 corresponds to the maximum expected value of species in a specific region, called asymptote. The parameter β_1 is related to the average rate of growth in species diversity. The parameter β_2 defines the shape of the curve and dislocates the inflection point of the function, reflecting the conditions of the region favorable to the growth in species diversity.

Table 1. Nonlinear models

Name	Model	Parameters	Shape	Asymptotes
Arrhenius	$\beta_0 x^{\beta_1}$	2	Convex	No
Exponential	$\beta_0 [1 - \exp(-\beta_1 x)]$	2	Convex	Yes
Persistence 1	$\beta_0 x^{\beta_1} \exp(-\beta_2 x)$	3	Sigmoid	No
Persistence 2	$\beta_0 x^{\beta_1} \exp(-\beta_2 / x)$	3	Sigmoid	No
Chapman	$\beta_0 [1 - \exp(-\beta_1 x)]^{\beta_2}$	3	Sigmoid	Yes

Since the dependent variable is a count data, it is plausible to assume that it comes from some positive discrete distribution. The most commonly used distribution in count data is Poisson. However, when we have a structural absence of zeros in the data, the most commonly used form is to truncate at the point zero. The objective of this paper is to consider the Zero Truncated Poisson distribution (ZTP) as an alternative to represent the probability distribution of the number of species in determined area.

In this paper, we propose a bayesian approach using latent variables to obtain conditional distributions and Gibbs algorythm implementation.

2. Method

2.1 Truncated Distributions

We will discuss some definitions of truncated distributions used in this paper. Initially we can redistribute the probability distribution functions in a way that the sum keeps being unitary. This way, we can define another variable $W = Y$ in the values of interest, in a way that $g(y) = kf(y)$, $y = 1, 2, 3, ...$, where f is the probability function of W. As we want Y truncated in zero, we should have $\sum_{y=1}^{\infty} g(y) = 1$, so

$$\sum_{y=1}^{\infty} kf(y) = 1 \Rightarrow k = \frac{1}{\sum_{y=1}^{\infty} f(y)} = \frac{1}{f(0)} = \frac{1}{P(W=0)}. \tag{1}$$

So we have the zero-truncated distribution function given by:

$$g(y) = \begin{cases} \frac{f(y)}{1-P(W=0)}, & y = 1, 2, 3, ... \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

with the main moments

$$E(Y) = \frac{E(W)}{1 - P(W=0)} \tag{3}$$

$$V(Y) = \frac{1}{1 - P(W=0)} \left[V(W) - \frac{E^2(W)}{[P(W=0)]^{-1} - 1} \right]. \tag{4}$$

So it is evident that the expected value of the truncated distribution is higher when compared to the non-truncated distribution. This difference is also evident in the variance, however the variance is lower once we have a truncated distribution.

2.1.1 Zero Truncated Poisson Distribution

Since the dependent variable is a count data, it is plausible to assume that it comes from some known positive discrete distribution, such as Poisson distribution. Considering that the event $y_i = 0$ is not observed, we can obtain the zero truncated distributions conditioning the probability functions at the point zero (Van Der Heijden, 2003). The probability function of the Zero-Truncated Poisson model (Tate, 1958) is defined by:

$$P(Y = y_i | y_i > 0) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i! (1 - e^{-\mu_i})}, \tag{5}$$

and the first and second moments are

$$E(Y_i | y_i > 0) = \frac{\mu_i}{1 - e^{-\mu_i}} \tag{6}$$

$$V\left(Y_i|y_i>0\right)=\frac{\mu_i}{1-e^{-\mu_i}}\left[1-\frac{\mu_i}{e^{\mu_i}-1}\right]. \tag{7}$$

This way, it is noticeable that the conditional variance is inferior to the conditional mean.

The fact that a distribution belongs to the exponential family provides great properties to the estimators and has good emphasis in generalized linear models (McCullagh e Nelder, 1989). The probability density functions of exponential family distributions can be expressed as:

$$p(\boldsymbol{\beta}|\mathbf{y})\propto\exp\left\{y_i\theta_i-b_1(\theta_i)-b_2(\theta_i)\right\}p(\boldsymbol{\beta}), \tag{8}$$

in which $\theta_i=\log(f(x_i,\boldsymbol{\beta}))$, $b_1(\theta_i)=f(x_i,\boldsymbol{\beta})$ and $b_2(\theta_i)=\log(1-exp(-f(x_i,\boldsymbol{\beta})))$ for Truncated Poisson.

We will use the systematic component of the model with identity link function. In this case, μ_i is specified as a non-linear function that represents the expected value of the species diversity due to the area. We consider $\mu_i=f(x_i;\beta)$ and to $f(\cdot)$ the following functions (Table 1).

2.1.2 Inference

Considering an independent sample $(y_1,...,y_n)$, the log-likelihood function of the model is:

$$\log\left(l_P\right)=\sum_{i=1}^{n}\left\{y_i\log\left(\mu_i\right)-\mu_i-\log\left(1-e^{-\mu_i}\right)-\log\left(y_i!\right)\right\}. \tag{9}$$

The estimation of the parameters can be obtained by the maximization of the log-likelihood function. The score function of the log-likelihood (9) is expressed by:

$$\frac{\partial l}{\partial\beta_i}=\sum_{i=1}^{n}\left[\frac{y_i}{\mu_i}-1-\frac{e^{-\mu_i}}{1-e^{-\mu_i}}\right]\left[\frac{\partial\mu_i}{\partial\beta_i}\right]. \tag{10}$$

The log-likelihood function of the model (5) considering the function $\mu_i=f(x_i,\boldsymbol{\beta})$ is given by:

$$\log\left(l_P\right)=\sum_{i=1}^{n}y_i\log\left\{f\left(x_i;\boldsymbol{\beta}\right)\right\}-\sum_{i=1}^{n}f\left(x_i;\boldsymbol{\beta}\right)-\sum_{i=1}^{n}\log\left(1-e^{-f(x_i;\boldsymbol{\beta})}\right)-\sum_{i=1}^{n}\log\left(y_i!\right). \tag{11}$$

2.2 Bayesian Inference

We introduce a Bayesian approach using non-informative prior densities. The non-informative priors are used when we expect that the information of the data is dominant.

For the estimation of $\boldsymbol{\beta}$ through bayesian model, we assume as prior a Multivariated Normal distribution $N(\boldsymbol{\beta}_0,\Sigma_0)$, with $\boldsymbol{\beta}_0$ given by the estimation of maximum likelihood and Σ_0 given by the inverse of the observed hessian matrix.

$$p(\boldsymbol{\beta})\propto\exp\left\{-\frac{\tau}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}. \tag{12}$$

By multiplying the expressions (5) and (12), the joint posterior distribution is given by:

$$p(\boldsymbol{\beta}|\mathbf{y})\propto\exp\left[-\frac{\tau}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right]\left[\prod_{i=1}^{n}\frac{\left[f\left(x_i;\boldsymbol{\beta}\right)\right]^{y_i}e^{-f(x_i;\boldsymbol{\beta})}}{y_i!\left(1-e^{-f(x_i;\boldsymbol{\beta})}\right)}\right]. \tag{13}$$

In this case, we may use as an approximation to $\boldsymbol{\beta}_0$, Σ_0 the estimates based on the likelihood function (11), with Σ_0 approximated by the inverse of the hessian matrix and τ being a known parameter that controls the prior uncertainty about the parameters $\boldsymbol{\beta}$.

Since the conditional distributions are unknown, we use the MCMC method through the Metropolis-Hastings algorithm to obtain the estimates of the parameters. The Bayesian confidence interval (*Highest Posterior Density*) (HPD), fixing the confidence (α) in 95% is given by:

$$(\beta',\beta'')=\exp\left\{\beta:p(\boldsymbol{\beta}|\mathbf{y})\geq k(\alpha)\right\}, \tag{14}$$

in which $k(\alpha)$ is the highest constant that $p(\beta'<\beta<\beta'')=1-\alpha$.

Considering a *flat* Normal distribution as priori, the posterior density to the parameters is given by:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \left[\prod_{i=1}^{n} \frac{[f(x_i;\boldsymbol{\beta})]^{y_i} e^{-f(x_i;\boldsymbol{\beta})}}{y_i!\left(1-e^{-f(x_i;\boldsymbol{\beta})}\right)}\right] p(\boldsymbol{\beta}). \tag{15}$$

We have a situation where the posterior conditional distributions don't have the form of any known distribution and, in this case, the Metropolis-Hastings algorithm is necessary to the posterior calculation.

In order to simplify the conditional distribution for the Gibbs algorithm, we resort to the use of latent variables that allow writing the posterior density as the product of model components.

2.2.1 Latent Variables

The idea of using latent variables is to obtain known forms of complete conditional distributions (Higdon, 1998). Considering the variables $\boldsymbol{u} = (u_1, ..., u_n)$, $\boldsymbol{v} = (v_1, ..., v_n)$ and $\boldsymbol{w} = (w_1, ..., w_n)$, the joint density to $(\boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \mathbf{w})$ is:

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}|\boldsymbol{y}) \propto \prod_{i=1}^{n} \exp\{-(v_i + w_i)\} I\{u_i \leq \exp(y_i\theta_i), v_i > b_1(\theta_i), w_i > b_2(\theta_i)\} p(\boldsymbol{\beta}). \tag{16}$$

We can prove the above equation by verifying that

$$\begin{aligned} p(\boldsymbol{\beta}|y_i) &\propto \int_{u_i}\int_{v_i}\int_{w_i} p(\beta, u_i, v_i, w_i|y) dw_i dv_i du_i. \\ &\propto p(\boldsymbol{\beta}) \int_0^{\exp(y_i\theta_i)} du_i \int_{b_1(\theta_i)}^{\infty} \exp\{-v_i\} dv_i \int_{b_2(\theta_i)}^{\infty} \exp\{-w_i\} dw_i. \\ &\propto p(\boldsymbol{\beta}) \exp(y_i\theta_i) \exp(-b_1(\theta_i)) \exp(-b_2(\theta_i)). \end{aligned}$$

In order to obtain the complete condicional distributions, we can evaluate the situations represented by the variable indicated by the expression (16).

In the case of the latent variable u_i, we have:

$$u_i \leq \exp(y_i\theta_i)$$

with $\theta_i = log(f(x_i, \beta))$, we have

$$\log u_i \leq y_i \log\{f(x_i;\beta)\}$$

i.e.,

$$f(x_i;\boldsymbol{\beta}) \geq \exp\left(\frac{1}{y_i}\log u_i\right).$$

As such, the vector of parameters $\boldsymbol{\beta} = \left(\beta_0, ..., \beta_q\right)$ must satisfy the restrictions above to every x_i. In this case, we define a subset of the parametric space for $\boldsymbol{\beta}$ given by:

$$S_\beta^u = \left\{\beta | \min_i(f(x_i;\boldsymbol{\beta})) \geq \exp\left(\frac{1}{y_i}\log u_i\right)\right\} \quad i = 1, ..., n. \tag{17}$$

In the case of the latent variable v_i, we have:

$$v_i > b_1(\theta_i)$$

$$f(x_i;\boldsymbol{\beta}) < v_i.$$

This way, we have:

$$S_\beta^v = \{\beta | \max_i(f(x_i;\boldsymbol{\beta})) < v_i\} \qquad i = 1, ..., n. \tag{18}$$

Considering the latent variable w_i, we have:

$$w_i > b_2(\theta_i)$$

$$w_i > \log\{1 - \exp(-f(x_i;\boldsymbol{\beta}))\}$$

$$f(x_i;\boldsymbol{\beta}) < \log\{1 - \exp\{w_i\}\}.$$

Hence, we define the set as:

$$S_\beta^w = \{\beta | \min_i(f(x_i;\boldsymbol{\beta})) < \log[1 - \exp(w_i)]\} \qquad i = 1, ..n \tag{19}$$

In this case, the indicative variable in the equation (16) is equal to 1 if, and only if, $\beta \in S_\beta^u \cap S_\beta^v \cap S_\beta^w$. The conditional probability distribution can be expressed by:

To the variable u_i,

$$f(u_i|\boldsymbol{\beta}) \propto I\{u_i \leq \exp(y_i\theta_i)\} \qquad i = 1, ..., n \tag{20}$$

hence $u_i \backsim$ Uniform $(0, \exp\{y_i\theta_i\})$.

To the variable v_i,

$$f(v_i|\boldsymbol{\beta}) \propto \exp(-v_i)\, I\{v_i > b_1(\theta_i)\} \qquad i = 1, ..., n \tag{21}$$

hence $v_i \backsim$ Truncated Exponential $(1, b_1(\theta_i))$.

To the variable w_i,

$$f(w_i|\boldsymbol{\beta}) \propto \exp(-w_i)\, I\{w_i > b_2(\theta_i)\} \qquad i = 1, ..., n \tag{22}$$

hence $w_i \backsim$ Truncated Exponential $(1, b_2(\theta_i))$.

2.2.2 Algorithm to the Generation of Samples

Given $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_q^0)$, we use the estimates calculated by the maximum likelihood method as initial values. The values will be updated following these steps:

1) $u_i|\boldsymbol{\beta}, y_i \backsim$ U $(0, \exp(y_i\theta_i))$

2) $v_i|\boldsymbol{\beta}, y_i \backsim$ Truncated Exp $(1, b_1(\theta_i))$

3) $w_i|\boldsymbol{\beta}, y_i \backsim$ Truncated Exp $(1, b_2(\theta_i))$

4) We generate a new candidate, $\boldsymbol{\beta}^c$, from the multivariate normal distribution, N$(\hat{\boldsymbol{\beta}}, \Sigma)$, corresponding to the estimates of the log-likelihood (9) and the hessian matrix. We submit the candidate to the acceptance according the conditions $(\boldsymbol{\beta} \in S_\beta^u \cap S_\beta^v \cap S_\beta^w)$. Otherwise the vector is discarded and $\boldsymbol{\beta}^1 = \boldsymbol{\beta}^c$ is used.

The process is repeated using $\boldsymbol{\beta}^1 = (\beta_1^1, ..., \beta_q^1)$ as initial values until the necessary sample is reached.

2.3 Model Selection

Several model selection method are proposed in literature. In the paper we will consider the criteria EAIC, EBIC (Brooks, 2002), DIC (Spiegelhalter et al., 2002) and CPO (Pettit & Young, 1990). The first three criteria suggest that the comparison among models are made based on the *deviance* calculation.

$$D(\boldsymbol{\beta}, M_i) = -2\log(L(\mathbf{y}|\boldsymbol{\beta}, M_i)) + C,$$

where $L(\mathbf{y}|\boldsymbol{\beta}, M_i)$ is the likelihood function associated to the model M_i and C is a constant that is canceled out. A Monte Carlo estimate for the standard deviation is:

$$\overline{D}(\boldsymbol{\beta}, M_i) = \frac{1}{m}\sum_{j=1}^{m} -2\log(L(\mathbf{y}|\boldsymbol{\beta}^{(j)}, M_i)),$$

i.e., the posterior average deviation.

2.3.1 Extended Akaike Information Criterion (EAIC)

The criterion proposed by Akaike (1974) is based on the likelihood function, $AIC(\boldsymbol{\beta}, M_i) = -2\log(L(\mathbf{y}|\boldsymbol{\beta}, M_i)) + 2p$, penalized by the numbers of parameters in the model. Whereas the bayesian information criterion (BIC) ponders the sample size using $BIC(\boldsymbol{\beta}, M_i) = -2\log(L(\mathbf{y}|\boldsymbol{\beta}, M_i)) + p\log(n)$. The selection criteria, within the Bayesian context, are obtained through a natural extension considering the posterior density of the parameters of the model.

$$EAIC = -2E[\log(L(\mathbf{y}|\boldsymbol{\beta}, M_i))] + 2p = D(\overline{\boldsymbol{\beta}}, M_i) + 2p,$$

$$EBIC = -2E[\log(L(\mathbf{y}|\boldsymbol{\beta}, M_i))] + p\log(n) = D(\overline{\boldsymbol{\beta}}, M_i) + p\log(n),$$

where p is the number of parameters of the model, n is the sample size and $D(\overline{\boldsymbol{\beta}}, M_i) = -2\log(L(\mathbf{y}|\overline{\boldsymbol{\beta}}, M_i)$ with $\overline{\boldsymbol{\beta}}$ equal to the mean of the posterior density. Both criteria (EAIC e EBIC) indicate the best models the lower the obtained value.

2.3.2 Deviance Information Criterion (DIC)

The DIC is consisted by the posterior mean deviance *(deviance)* penalized by the number of parameters of the model. This criterion is interesting since it can be incorporated during the Monte Carlo simulation. Just like the other criteria, it is an asymptotic approximation to large samples and it is valid when the posterior distribution follows an approximately multivariate normal distribution. Lower values of DIC indicate better adjustment. The DIC is obtained by:

$$DIC = \overline{D}(\boldsymbol{\beta}, M_i) + p_{di},$$

where $p_{di} = \overline{D}(\boldsymbol{\beta}, M_i) - D(\overline{\boldsymbol{\beta}}, M_i)$ measures the complexity of the model i. The criterion suggests a comparison between the mean deviance and the deviance applied in the posterior mean.

2.3.3 Conditional Predictive Ordinate Criterion (CPO)

The Conditional Predictive Ordinate (CPO) (Gelfand et al., 1994) is another criterion widely used in literature to evaluate the model and it is based on cross-validation density. Considering $\boldsymbol{Y} = (y_1, y_2, ..., y_n)$, the cross-validation obtained through density $p(y_k|\boldsymbol{Y}_{-k})$, with $\boldsymbol{Y}_{-k}$ denoting the set of all elements except the k-the observation y_{-k}. This statistic represents the most likely values when the model is adjusted to every observation except y_k. For the k-th observation, the CPO is defined by:

$$CPO_k = p(y_k|\boldsymbol{Y}_{-k}) = \int p(y_k|\boldsymbol{\beta}, \boldsymbol{Y}_{-k}) p(\boldsymbol{\beta}|\boldsymbol{Y}_{-k}) d\boldsymbol{\beta}. \tag{23}$$

In this case, $p(\boldsymbol{\beta}|Y_{-k})$ represents the posterior density of $\boldsymbol{\beta}$ based on data of Y_{-k}. Therefore, from the Equation (23) the CPO_k is defined as the posterior predictive marginal density of y_k given Y_{-k}. Larger values for the CPO_k imply better models and lower values indicate influential observations. For most models, there isn't a closed form for the CPO_k. Thus, using the samples generated by the Monte Carlo method, we can approximate the calculation of the CPO (Li et al., 2006) by:

$$CPO_k \approx \frac{1}{M} \sum_{j=1}^{m} [L(y_k|\boldsymbol{\beta}^{(j)})]^{-1}.$$

Summing the information in a simple measure, we chose the model with larger value applied in the natural logarithm of CPO's, called log pseudo marginal likelihood $LPML = \sum_{i=1}^{n} \log(CPO_i)$.

3. Results and Discussion

In this simulation, we fixed the arbitrary values for β of the Table 1 models, and areas units going from 1 to 40. This areas were divided in sections of size 100, with values generated from the Zero Truncated Poisson distribution, representing the number of different species observed in a specific area.

By calculating the statistics, 10, 40 and 100 replicas were generate *Mean Square Error* (MSE) and *Mean Absolute Percent Error* (MAPE) to evaluate the quality adjustment of the model (Table 2, Table 3 and Table 4). The statistics MSE and MAPE are given by:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{\beta}_i - \beta_i\right)^2$$

$$MAPE = \frac{1}{m} \sum_{i=1}^{m} \left|\frac{\hat{\beta}_i - \beta_i}{\beta_i}\right|.$$

The adjustment of the models were evaluated by four bayesian criteria: DIC, EAIC, EBIC and CPO (Table 5, Table 6 and Table 7) (Spiegelhalter et al., 2002).

Table 2. Bayesian estimation for Poisson distribution n = 10

	Parameters	Reals	Mean	CI (95 %)	MSE	MAPE	Coverage
Arrh	β_0	15	15.141	(10.228;20.059)	0.945	0.051	91.6%
	β_1	0.5	0.502	(0.397;0.606)	0.056	0.032	93.5%
Expo	β_0	100	100.114	(91.132;109.098)	4.949	0.039	92.1%
	β_1	0.15	0.152	(0.103;0.201)	0.027	0.138	92.3%
	β_0	10	10.107	(5.347;14.826)	0.927	0.074	90.8%
Ps 1	β_1	0.9	0.907	(0.664;1.159)	0.137	0.117	92.2%
	β_2	0.02	0.02	(0.007;0.035)	0.007	0.294	94.5%
	β_0	50	62.133	(46.467;85.196)	11.458	0.181	93.0%
Ps 2	β_1	0.22	0.149	(0.036;0.233)	0.077	0.297	90.5%
	β_2	0.1	0.331	(0.021;0.843)	0.256	1.123	94.4%
	β_0	100	109.639	(92.993;126.552)	16.9	0.126	75.6%
Chap	β_1	0.1	0.073	(0.039;0.102)	0.033	0.288	53.0%
	β_2	3	1.978	(0.867;2.592)	1.090	0.341	22.9%

Table 3. Bayesian estimation for Poisson distribution n = 40

	Parameters	Reals	Mean	CI (95 %)	MSE	MAPE	Coverage
Arrh	β_0	15	15.092	(12.429;17.764)	1.395	0.074	94.2%
	β_1	0.5	0.499	(0.442;0.556)	0.03	0.047	94.3%
Expo	β_0	100	100.086	(95.637;104.564)	2.355	0.019	94.2%
	β_1	0.15	0.15	(0.127;0.174)	0.013	0.068	95.6%
	β_0	10	9.964	(7.16;12.78)	1.511	0.119	91.3%
Ps 1	β_1	0.9	0.908	(0.754;1.063)	0.084	0.074	92.5%
	β_2	0.02	0.02	(0.012;0.029)	0.005	0.192	93.2%
	β_0	50	56.091	(45.158;70.241)	6.836	0.104	91.4%
Ps 2	β_1	0.22	0.158	(0.08;0.218)	0.052	0.2	90.5%
	β_2	0.1	0.376	(0.039;0.85)	0.212	1.626	93.5%
	β_0	100	100.029	(89.69;110.373)	5.532	0.044	93.4%
Chap	β_1	0.1	0.102	(0.077;0.128)	0.014	0.11	92.9%
	β_2	3	3.096	(2.269;3.932)	0.468	0.119	94.5%

Table 4. Bayesian estimation for Poisson distribution n = 100

	Parameters	Reals	Mean	CI (95 %)	MSE	MAPE	Coverage
Arrh	β_0	15	15.01	(13.296;16.738)	0.945	0.051	91.6%
	β_1	0.5	0.5	(0.463;0.537)	0.02	0.032	92.5%
Expo	β_0	100	100.087	(97.266;102.888)	1.481	0.012	95.5%
	β_1	0.15	0.15	(0.135;0.165)	0.008	0.042	96.0%
	β_0	10	9.951	(8.104;11.797)	0.927	0.074	95.0%
Ps 1	β_1	0.9	0.906	(0.804;1.008)	0.052	0.046	94.1%
	β_2	0.02	0.02	(0.015;0.026)	0.003	0.12	93.4%
	β_0	50	54.034	(46.467;63.279)	4.388	0.068	93.4%
Ps 2	β_1	0.22	0.173	(0.117;0.215)	0.039	0.155	92.6%
	β_2	0.1	0.221	(0.026;0.496)	0.16	1.211	92.9%
	β_0	100	100.098	(93.556;106.663)	3.341	0.026	94.2%
Chap	β_1	0.1	0.101	(0.085;0.117)	0.008	0.066	93.3%
	β_2	3	3.034	(2.522;3.55)	0.266	0.069	94.3%

Table 5. Criteria of selection of the models n = 10

Criteria		Arrhenius	Exponential	Persistence 1	Persistence 2	Champan
Arrh	DIC	**52.1%**	7.4%	2.1%	0.7%	37.7%
	EAIC	**78.9%**	6.9%	4.4%	0.3%	9.5%
	EBIC	**94.8%**	3.9%	0.2%	0%	1.1%
	CPO	**86.3%**	7.5%	4.3%	1.9%	0%
Expo	DIC	0.6%	**83.0%**	10.9%	0.9%	4.6%
	EAIC	0.3%	**81.1%**	15.3%	0.6%	2.7%
	EBIC	2.1%	**77.2%**	13.6%	1.2%	5.9%
	CPO	0.3%	**82.9%**	15.5%	1.3%	0%
Ps 1	DIC	2.4%	64.1%	**2.1%**	0.2%	31.2%
	EAIC	4.8%	64.4%	**4.0%**	0.2%	26.6%
	EBIC	17.7%	59.1%	**2.1%**	0.6%	20.5%
	CPO	13.5%	77.3%	**6.7%**	2.5%	0%
Ps 2	DIC	64.5%	0.7%	18.2%	**8.9%**	7.7%
	EAIC	73.3%	0.7%	24.6%	**1.2%**	0.2%
	EBIC	93.4%	0.5%	4.8%	**0.7%**	0.6%
	CPO	66.4%	0.5%	20.4%	**12.7%**	0%
Chap	DIC	5.7%	29.5%	0%	10.6%	**54.2%**
	EAIC	8.6%	20.7%	0%	10.6%	**60.1%**
	EBIC	14.9%	13.3%	0%	10.6%	**61.2%**
	CPO	17.5%	71.7%	0%	10.8%	**0%**

Table 6. Criteria of selection of the models n = 40

Criteria		Arrhenius	Exponential	Persistence 1	Persistence 2	Champan
Arrh	DIC	**65.5%**	0.5%	3.1%	3.4%	27.5%
	EAIC	**80.2%**	0.3%	5.4%	1.9%	12.2%
	EBIC	**98.1%**	0.3%	0.8%	0.3%	0.5%
	CPO	**80.3%**	0.3%	10.2%	9.2%	0%
Expo	DIC	0%	**69.8%**	11.4%	3.6%	15.2%
	EAIC	0%	**72.6%**	16.5%	3.5%	7.4%
	EBIC	0%	**60.5%**	13.8%	4.8%	20.9%
	CPO	0%	**77.2%**	17%	5.8%	0%
Ps 1	DIC	0.1%	72.6%	**7.4%**	0.8%	19.1%
	EAIC	0.2%	68.4%	**15.9%**	1.1%	14.4%
	EBIC	0.9%	64.1%	**14%**	1.6%	19.4%
	CPO	0.1%	72.3%	**24.9%**	2.7%	0%
Ps 2	DIC	55%	0%	21.2%	**14.6%**	9.2%
	EAIC	65.8%	0%	28.3%	**5.9%**	0%
	EBIC	95.3%	0%	2.3%	**2.3%**	0.1%
	CPO	58.3%	0%	22.8%	**18.9%**	0%
Chap	DIC	0%	0%	0%	3.3%	**96.7%**
	EAIC	0%	0%	0%	2.8%	**97.2%**
	EBIC	0%	0%	0%	2.8%	**97.2%**
	CPO	31.3%	4.6%	20%	44.1%	**0%**

Table 7. Criteria of selection of the models n = 100

Criteria		Arrhenius	Exponential	Persistence 1	Persistence 2	Champan
Arrh	DIC	**72.2%**	0%	4.1%	7%	16.7%
	EAIC	**77.7%**	0%	15.1%	3%	4.2%
	EBIC	**98.2%**	0%	0.3%	0.3%	1.2%
	CPO	**72.7%**	0%	16%	11.3%	0%
Expo	DIC	0%	**71%**	8.4%	1.6%	19%
	EAIC	0%	**78.5%**	10.5%	1.4%	9.6%
	EBIC	0%	**63.6%**	9.4%	1.7%	25.3%
	CPO	0%	**86.1%**	11.6%	2.3%	0%
Ps 1	DIC	0%	61.7%	**28.5%**	0.5%	9.3%
	EAIC	0%	55.1%	**40.5%**	0.4%	4%
	EBIC	0%	52.2%	**36%**	0.5%	11.3%
	CPO	0%	55.2%	**44.1%**	0.7%	0%
Ps 2	DIC	44.5%	0%	19.3%	**22.7%**	13.5%
	EAIC	58.6%	0%	30.4%	**10.5%**	0.5%
	EBIC	94.4%	0%	2.3%	**3.3%**	0%
	CPO	49%	0%	22.6%	**28.4%**	0%
Chap	DIC	0%	0%	0%	0.1%	**99.9%**
	EAIC	0%	0%	0%	0.1%	**99.9%**
	EBIC	0%	0%	0%	0.1%	**99.9%**
	CPO	0%	0%	0%	0.1%	**99.9%**

Using Table 2, we notice that every parameter of the model indicates good adjustment when we compare the actual values to the mean. In these models, the coverage probabilities are close to the expected value of 95. The Table 5 shows the difference between the selection criteria. The Arrhenius, Exponential and Chapman models were correctly selected.

However, a competition between Persistence models and the others is noticeable. It means that, despite the samples being generated with distinct parameters, the models adjust to the data due to their flexibility (Figure 1), with curves approximately superimposed for data generated from Persistence models, resulting in close values to the selection criteria.

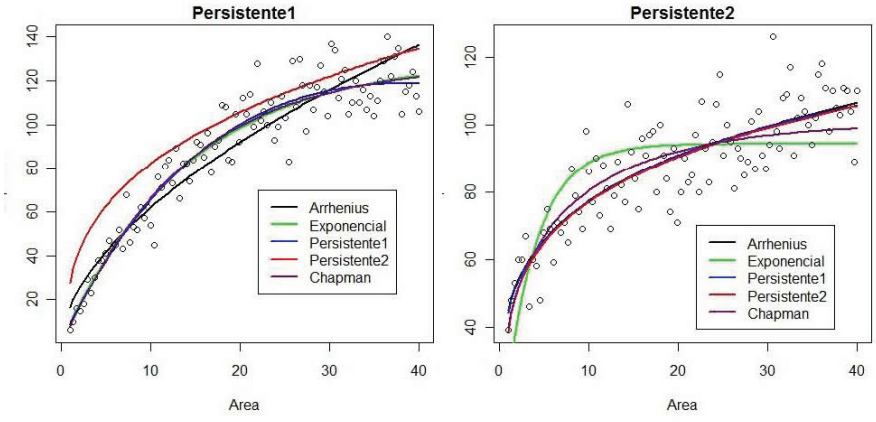

Figure 1. Comparison between models

3.1 Application to Real Data

The data refer to the species diversity from Hymenoptera order (wasps, bees and ants), observed in a beech forest. The objective of the study is to relate the number of collected species with the area and estimate the extinction and immigration rates (Ulrich, 2001). In order to estimate the posterior distribution parameters given by 8, 10,000 values were simulated for each parameter and Bayesian criteria of selections were calculated. The priori distributions are given by the maximum likelihood estimates.

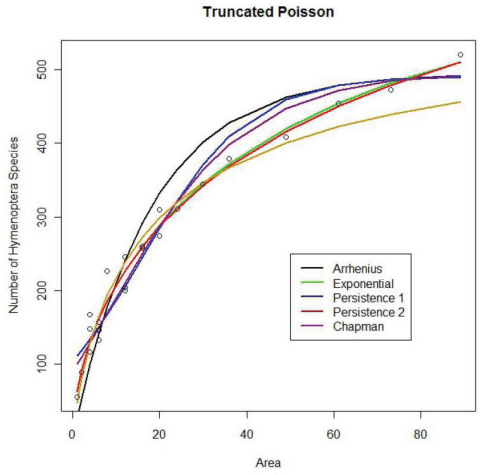

Figure 2. Model Adjustment for Truncated Poisson for Hymenopteras species

Table 8. Bayesian selection for the study of Hymenoptera

Criteria	Arrhenius	Exponential	Persistence 2	Persistence 1	Chapman
DIC	205.610	220.155	239.789	148.905	153.176
EAIC	413.167	383.007	285.438	274.101	237.871
EBIC	422.042	398.320	300.751	287.414	251.185
CPO	-196.838	-181.956	-135.576	-130.122	-111.306

The Chapman and Persistence 1 models using Poisson distribution show better adjustment, due to lower values of criteria EAIC, DIC and EBIC (Table 8). This fact may be verified by observing Figure 2, in which we notice the two models adjustment adequate to the data.

4. Discussion

In the proposed approach, we obtained good estimates to the simulated data. For every parameters, the real values are in between the credible intervals. The coverage percents obtained are close to 95% as the sample size is increased. We can notice a depletion in the credible intervals amplitudes. The only exception was verified in the Chapman model for a sized 10 sample, which presented different coverage probabilities from the other models.

We notice a depletion in the sampling errors, which means that the adjusted values are close to the real values, with slightly increase in MSE and MAPE of Arrhenius and Persistence 1 models once we increase from 10 to the 40th sample.

Through the Gibbs algorithm, we notice that the EAIC, EBIC, DIC and CPO models identify, with high probability, the Exponencial, Chapman and Arrhenius models. However, it wasn't possible to verify one single selection criterion that correctly identifies every model. What we have is a high competitivity between the Persistence models. It means that, despite the samples being generated with distinct parameters, the models adjust to the data due to their flexibility.

For the approach using latent variables, the method showed adequate, viable and easy implementation.

References

Arrhenius, O. (1921). Species and Area. *J. Ecol., 9*, 95-99. http://dx.doi.org/10.2307/2255763

Dengler, J. (2009). Which function describes the species-area relationship best? A review and empirical evaluation. *Journal of Biogeography, 36*, 728-744. http://dx.doi.org/10.1111/j.1365-2699.2008.02038.x

Gelfand, A., & Dey, K. D. (1994). Bayesian Model Choice: Asymptotics and Exacts Calculations *J. R. Stat. Soc, 56*, 501-514.

Higdon, D. M. (1998). Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications. *Journal of the American Statistical Association, 3*, 585-595. http://dx.doi.org/10.1080/01621459.1998.10473712

Lomolino, M. V. (2000). Ecology most general, yet protean pattern: the species-area relationship. Millennium Issue. *Journal of Biogeography, 27*, 17-26. http://dx.doi.org/10.1046/j.1365-2699.2000.00377.x

McCullagh, P. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall. http://dx.doi.org/10.1007/978-1-4899-3242-6

Scheiner, S. M. (2003). Six types of species-area curves. *Global Ecology and Biogeography, 96*, 1141-1151.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde. (2002). The species-area relationship: Separating the effects of species-abundance and spatial distribution. *Journal of the Royal Statistial Society, Series B, 64*, 583-639.

Tate, R. F., & Goen, R. L. (1958). Minimum Variance Unbiased Estimation for the Truncated Poisson Distribution. *The Annals of Mathematical Statistics, 29*, 3, 755-765. http://dx.doi.org/10.1214/aoms/1177706534

Tjorve, E. (2009). Shapes and functions of species-area curves (II): A review of possible models. *Journal of Biogeography, 30*, 827-835. http://dx.doi.org/10.1046/j.1365-2699.2003.00877.x

Tjorve, E., Kunin, W. E., Polce, C., & Tjorve, K. M. C. (2003). The species-area relationship: Separating the effects of species-abundance and spatial distribution. *Journal of Ecology, 12*, 441-447.

Ulrich, W. (2001). Hymenopteren in einem Kalkbuchenwald: Eine Modellgruppe zur Untersuchung von Tiergemeinschaften und okologischen Raum-Zeit-Mustern. *Schriftenr. Forschzentr. Waldokosysteme A 171*, *Gottingen, 249 S.*

Ulrich, W., & Buszko, J. (2003). Self-similarity and the species-area relation of Polish butterflies. *Basic and Applied Ecology, 4*, 263-270. http://dx.doi.org/10.1078/1439-1791-00139

Ulrich, W., & Buszko, J. (2006). Sampling design and the shape os species-area curves on the regional scale. *Acta Oecol, 31*, 54-59. http://dx.doi.org/10.1016/j.actao.2006.03.005

Van Der Heijden, Peter; Bustami, R., Cruyff, M., Engbersen, G., & Van Houwelingen, H. (2003). Point and interval estimation of the population size using the truncated Poisson regression model *Statistical Modelling: An International Journal, 3*(4), 305. http://dx.doi.org/10.1191/1471082X03st057oa