A Generalized Species-Area Relationship for Estimating Species Diversity: The Poisson Distribution Case

K.S. Conceição[a], R.M. Pires[a], F. Louzada-Neto[a], M.G. Andrade[b], C.A.R. Diniz[a]

[a] DEs, Universidade Federal de São Carlos, São Carlos-SP
[b] SME-ICMC, Universidade de São Paulo, São Carlos-SP

1 Introduction

The species diversity is related to several factors, such as the mutation, interaction, competition and the amount of available resources for survival, amongst others. Besides, species diversity is related to the size of the habitat area (species-area relationship). The mutation allows the action of natural selection and contributes to the emergence of new species (speciation), while the resources represent the issue of homogeneity and heterogeneity of the studied habitat. Besides these factors, species diversity is also related to the size of the habitat area, and are usually refereed as species-area relationship.

The relationship between species diversity and area is also influenced by the shape of the habitat sampling, which can modify the pattern (concave or convex) and the properties of the curve describing this relationship.

Williams *et al.* (2009) points out three deficiencies observed in the studies of species-area relationships: (a) too much emphasis on maximizing the goodness of fit between species diversity and area, ignoring the effects of other factors; (b) assumptions about the species diversity distribution are often inadequate and untested; and (c) comparison of models using coefficient of determination (R^2) with different distributions of error and/or number of parameters.

Studies involving the species-area relationships are based on a structure composed of two components: one of them deterministic, represented by the function that describes this species-area relationship, and other random one, represented by the probability distribution of the random variable species diversity number. Here we shall assume that the number of species diversity follows a usual Poisson distribution. In this paper, the main idea is to present a generalization of the species-area relationship that has as particular cases relations proposed earlier, including ones that consider the effects of minimum area and upper asymptote. The proposed model is suitable for areas of different scales (small, intermediate and large) and considers a discrete probability distribution for the species diversity. The advantage of our formulation is to lead to a unique function for species-area relationship, which takes into account both effects of minimum area and asymptotic behavior of the growth curve for large areas, providing a unique algorithm for fitting different dataset, and given the opportunity of choosing the best model in the light of the data.

2 Proposed Model

Let S be a variable that represents the diversity of species and A the size of area, the generalized species-area relationship ($gSAR$) is given by,

$$S = \beta_0 + \beta_1 \cdot [g(A)]^{\beta_2} \cdot e^{-\beta_3 A^{\beta_4}}, \tag{1}$$

where β_i is the parameter of the model, $i = 0, \ldots, 4$, such that $\beta_0,\ \beta_1,\ \beta_3 \geq 0$, $\beta_2 \in (0,1)$ and $\beta_4 \in \Re$; $g(A)$ is a function of the area, given by $g(A) = A$ or $g(A) = \log(A)$, set to $A \geq 1$. The restrictions imposed on the parameters $\beta_0,\ \beta_1,\ \beta_2$ and β_3 are related to their biological interpretation. For $g(A) = \log(A)$, β_0 represents the density of species, in other words, species diversity for a unit area ($A = 1$). However, this interpretation does not apply to $g(A) = A$. The parameter β_1 is the slope of the curve representing the growth of species diversity with increasing area. The exponent β_2 reflects a typical feature of the power law traditionally adjusted for the species-area relationship. The parameter β_3 introduces in the model the flexibility of simultaneously represent of the effects of minimum area and upper asymptote, depending on the sign of the parameter β_4. The term $e^{-\beta_3 A^{\beta_4}}$, with $\beta_3 > 0$, represents the persistence function that assigns a pattern to the curvature of the relationship that reflects the effect of minimum area, if $\beta_4 < 0$, or the asymptotic effect for large areas, if $\beta_4 > 0$.

2.1 Particular Species-Area Relationship Cases

Several usual species-area relationship can be seen as particular cases of $gSAR$ (1). In the section we describe five of them as follows.

2.1.1 Model Proposed by Arrhenius (1921)

A pioneering study (Arrhenius, 1921) expressed the species-area relationship as a power law given by,

$$S = \beta_1 A^{\beta_2}, \tag{2}$$

where the variable S represents the number of species in the sampled area A, β_1 and β_2 correspond, in the $\log - \log$ scale, the number of species expected in a unit of area ($A = 1$) and the linear coefficient, respectively. Variations in values of β_1 and β_2 are of interest because they might indicate that different processes define the species-area relationship at different spatial scales. The author emphasizes that, the larger the area, the greater the number of species. This model is widely used by ecologists (not always for biological causes, usually for convenience) (Tjørve, 2003; Ulrich & Buszko, 2007). This model is a particular case of the $gSAR$ model (1) when the function $g(A) = A$ and parameters $\beta_0 = 0$ and $\beta_3 = 0$.

2.1.2 Model Proposed by Gleason (1922)

Gleason (1922) noticed that the formula of Arrhenius suggests extremely high estimates for the number of species in large areas. To avoid this problem, the author proposed an alternative based on this model, but considering that the diversity of species is a linear function of the logarithm of the area, such as,

$$S = \beta_0 + \beta_1 \log(A), \tag{3}$$

which may be more convenient to describe the relationship over large areas. This model is a particular case of the $gSAR$ model (1) when the function $g(A) = \log(A)$ and parameters $\beta_2 = 1$ and $\beta_3 = 0$.

2.1.3 Model Proposed by Connor & McCoy (1979)

Connor & McCoy (1979) proposed a simple linear regression to express number of species S in function of the area A,

$$S = \beta_0 + \beta_1 A, \tag{4}$$

where β_0 represents the intercept parameter and β_1 the slope of the regression. This model fits well for species-area relationships in small areas. A limitation of this model is the inconsistency in the estimation of species diversity when $A = 0$. This model is a particular case of the $gSAR$ model (1) when the function $g(A) = A$ and parameters $\beta_2 = 1$ and $\beta_3 = 0$.

2.1.4 Model Proposed in Plotkin *et al.* (2000)

The model proposed in Plotkin *et al.* (2000) seeks a better fit of the relationship for large areas. This model can be seen as an extension of the Arrhenius model, with the addition of a persistence function. It is given by,

$$S = \beta_1 A^{\beta_2} e^{-\beta_3 A}, \tag{5}$$

where the variable S represents the number of species, A is the sampled area, β_1 is a constant, β_2 is the $\log - \log$ linear coefficient and β_3 is a parameter that, when greater than zero, reduces the curvature of the power function for large areas. This model is a particular case of the $gSAR$ model (1) when the function $g(A) = A$ and parameters $\beta_0 = 0$ and $\beta_4 = 1$.

2.1.5 Model Proposed by Ulrich & Buszko (2003, 2004)

To model the effect of small areas in the growth of species diversity Ulrich & Buszko (2003, 2004) proposed a modification of the model presented by Plotkin *et al.* (2000), which resulted in the following expression,

$$S = \beta_1 A^{\beta_2} e^{-\beta_3 / A}, \tag{6}$$

where the parameter β_3, if greater than zero, reflects the need for a minimum area that certain species may need to survive and reproduce. This model is a particular case of the $gSAR$ model (1) when the function $g(A) = A$ and parameters $\beta_0 = 0$ and $\beta_4 = -1$.

3 Inference

Consider the random variable S denoting the species diversity number, which has a Poisson distribution with parameter μ, such that $\mu = E(S) = f(A)$. Consider $\boldsymbol{S} = \{S_1, S_2, \ldots, S_n\}$ the vector of observations of the random variable S associated to the vector of observations on the size of area $\boldsymbol{A} = \{A_1, A_2, \ldots, A_n\}$, such that $E(S_i) = \mu_i = f(A_i)$. The log-likelihood function associated with the observation vector $\boldsymbol{S}$ is given by,

$$\ell(\boldsymbol{S}, \boldsymbol{A}; \boldsymbol{\beta}) = \sum_{i=1}^{n} S_i \log f(A_i) - \sum_{i=1}^{n} (f(A_i)) - \sum_{i=1}^{n} \log S_i!. \tag{7}$$

For the proposed model, $f(A)$ is given by the equation (1) and then, the likelihood function (7) can be rewrite as,

$$\ell(\boldsymbol{S}, \boldsymbol{A}; \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) \quad = \quad \sum_{i=1}^{n} S_i \log \left(\beta_0 + \beta_1 (g(A_i))^{\beta_2} e^{-\beta_3 A_i^{\beta_4}} \right) - \sum_{i=1}^{n} \log S_i! -$$

$$\sum_{i=1}^{n} \left(\beta_0 + \beta_1 (g(A_i))^{\beta_2} e^{-\beta_3 A_i^{\beta_4}} \right). \tag{8}$$

The maximum likelihood estimates (MLEs) are obtained by direct maximization of the log-likelihood function (7). The advantage of this procedure is that it runs immediately using existing statistical packages such R (R Development Core Team, 2009). We consider the software R through the BFGS algorithm (Nocedal & Wright, 2006) to compute the MLEs. As usual, large-sample inference for the parameters are based on the MLEs and their estimated standard errors.

There are several criteria for choosing models that fit best in a data set. We consider here the BIC and AIC criteria. Smaller values of BIC and AIC indicate better model but the BIC is more rigorous, as the models with larger number of parameters have greater penalties.

From the practical point of view, it may be interesting to test some specific hypothesis about the parameters, in order to verify if any of the particular cases of the $gSAR$ model (1), that is, to test the following hypothesis: *(i)* $H_0 : \beta_0 = 0$; *(ii)* $H_0 : \beta_2 = 1$; *(iii)* $H_0 : \beta_3 = 0$; *(iv)* $H_0 : \beta_4 = 1$; and *(v)* $H_0 : \beta_4 = -1$.

4 Application

In this section, we illustrate the flexibility of the proposed $gSAR$ model and compare it with its particular cases on one dataset. A real data on fish diversity from 70 lakes in the world analyzed in Stein & Juritz (1988) and Barbour & Brown (1974).

4.1 Real Data

Consider the data set extreacted from Barbour & Brown (1974) on fish diversity in 70 lakes, whose areas were measured in Km^2, belonging to the following regions: Africa, Canada, Great Britain, Guatemala, Italy, Japan, Mexico, Nicaragua, Costa Rica, Peru, Bolivia, Soviet Union, United States, Romania and Yugoslavia. These data were analyzed by Barbour & Brown (1974) and Stein & Juritz (1988) considering a linear relationship between the species diversity and the area logarithm. All models considered in this work were fitted to the data aiming to check which function is the best to describe the relationship between species diversity and area of the lakes. The Table 1 shows the selection criteria values for each model. Our $gSAR$ model, with $g(A) = A$, presents the lowest values for the AIC and BIC, indicating that it is the best fit among the models considered for these data.

Table 1: Selection criteria for the fitted models.

Criteria	$g(A) = A$					$g(A) = \log(A)$	
	Connor & McCoy	Arrhenius	Plotkin *et al.*	Ulrich & Buszko	$gSAR$	Gleason	$gSAR$
AIC	2426.49	1900.36	1910.55	1902.38	**1855.48**	2169.36	2175.36
BIC	2430.98	1904.86	1917.29	1909.12	**1866.73**	2173.86	2186.60

The Figure 1 illustrates the $gSAR$ model fitted for the data of fish diversity in function of the lakes areas in the original and logarithmic scale. The proposed model adequately adjusted the data set, capturing mainly the minimum area effect (see Figure 1 **(B)**). Comparing the points in Figure 1 **(A)**, we notice that, for some areas, there are points with high fish diversity. In contrast, there are overly large areas where the diversity are not as high as those of a few relatively minor areas. For this reason, we believe that these extreme points may have influenced the quality of the adjustments.

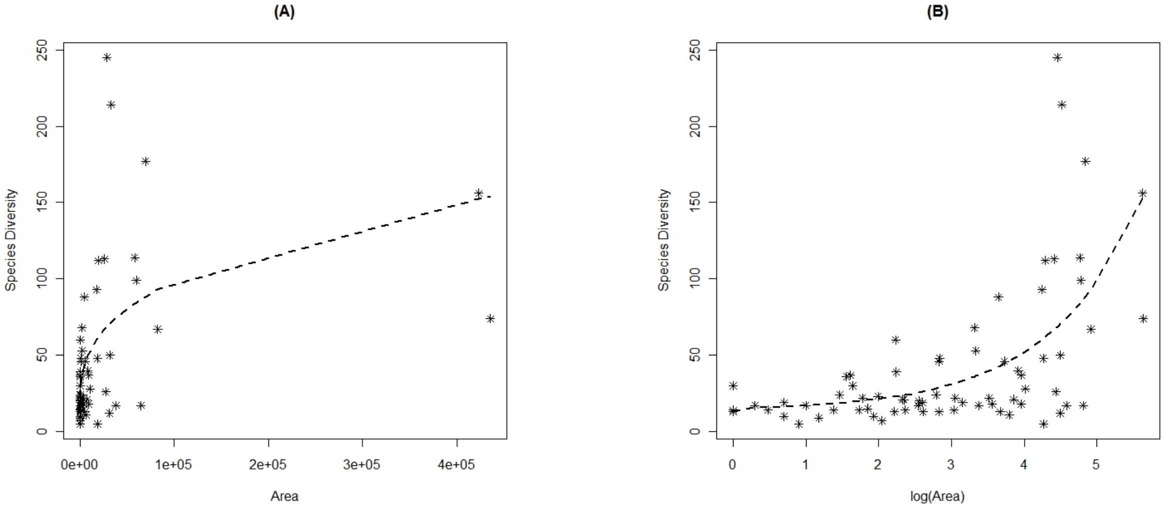

Figure 1: Fitted model for fish diversity data from 70 lakes in the world: **(A)** area in the original scale and **(B)** area in logarithmic scale. * Real data; **- -** Fitted model.

It is worth to notice that, in order to analyze this data set, Barbour & Brown (1974) considered a multiple linear regression model. That is, the expected value of species diversity is a linear function of area and latitude. Therefore, assuming approximately, a Normal distribution for the species diversity number. In Stein & Juritz (1988) was considered a model for the logarithm of the expected value of species diversity number as a linear function of the logarithm of the area, the error distribution being a Poisson-Inverse Gaussian. The fitted model shows large deviations between the estimated and observed value of species diversity numbers for large areas, resulting in a mean square error of 70% of the variance of S. The proposed $gSAR$ model however reduced this error to 65%.

5 Final Comments

Many models have been proposed in the literature in order to explain the relationship between the species diversity number and the habitat area. In this paper we propose the $gSAR$ model (1), which takes into account the effect of minimum area and also the behavior pattern of this relationship for large areas, while presents several usual specie-area relationship as particular cases. The results of a conducted simulation study showed that the selection criteria, such as AIC and BIC, are suitable to decide for the best model to be used for describing the species-area relationship in the light of a particular dataset. The empirical study based on a known real dataset on fish diversity in 70 lakes revels that the proposed extension over performs its particular cases to fit this dataset.

Acknowledgement

This work had financial support from FAPESP, CAPES and CNPq.

References

Arrhenius, O. (1921). Species and area. *Journal of Ecology*, **9**, 95–99.

Barbour, C. D. & Brown, J. H. (1974). Fish Species Diversity in Lakes. *The American Naturalist*, **962**, 473–489.

Connor, E. F. & McCoy, E. D. (1979). The Statistics and Biology of the Species-Area Relationship. *The American Naturalist*, **6**, 791–833.

Gleason, H. A. (1922). On the Relation Between Species and Area. *Ecology*, **3**, 158–162.

Nocedal, J. & Wright, S. J. (2006). *Numerial Optimization*. Springer-Verlag, New York, second edition.

Plotkin, J. B., Potts, M. D., Yu, D. W., Bunyavejchewin, S., Condit, R., Foster, R., Hubbell, S., LaFrankie, J., manokaran, N., Seng, L. H., Sukumar, R., Nowak, M. A. & Ashton, P. S. (2000). Predicting species diversity in tropical forests. *Proceedings of the National Academy of Sciences USA*, **97**, 10850–10854.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Stein, G. Z. & Juritz, J. M. (1988). Linear models with an inverse Gaussian-Poisson error distribution. *Communications in Statistics - Theory and Methods*, **17**, 557–571.

Tjørve, E. (2003). Shapes and functions of species-area curves: a review of possible models. *Journal of Biogeography*, **30**, 827–835.

Ulrich, W. & Buszko, J. (2003). Self-similarity and the species-area relation of Polish butterflies. *Basic and Applied Ecology*, **4**, 263—270.

Ulrich, W. & Buszko, J. (2004). Habitat reduction and patterns of species loss. *Basic and Applied Ecology*, **5**, 231–240.

Ulrich, W. & Buszko, J. (2007). Sampling design and the shape of species-area curves on the regional scale. *Acta Oecologica*, **31**, 54–59.

Williams, M. R., Lamont, B. & Henstridge, J. (2009). Species-area functions revisited. *Journal of Biogeography*, **36**, 1994–2004.