Brazilian Journal of Probability and Statistics
2014, Vol. 28, No. 3, 446–460
DOI: 10.1214/13-BJPS219

A generalized species-area relationship: The Poisson distribution case

Katiane S. Conceição[b], Rubiane M. Pires[a], Francisco Louzada[b], Marinho G. Andrade[b] and Carlos A. R. Diniz[a]

[a]*UFSCar*
[b]*ICMC–USP*

Abstract. Species richness is related to various factors, such as mutation, interaction, competition and the amount of available resources for survival, among others. Additionally, species richness is related to the size of the habitat area (species-area relationship). A major problem in ecology is to identify the best function that models this relationship. The main aim of the paper is to propose a generalized fitting formula for the species-area relationship. The advantage of our model is that it leads to a unique formulation for species-area relationships, which takes into account both the effects of the minimal area and asymptotic behavior of the growth curve for large areas. This approach provides a unique algorithm for fitting different datasets, and choosing the best model in light of the data. The applicability of our approach was tested via a simulation study conducted to determine if two usual selection criteria are suitable to choose the best formulation to be considered to describe species-area relationships. Real datasets involving a species richness of fish in 70 lakes was analysed using the proposed methodology.

1 Introduction

The relationship between the number of species (S) and area (A) of a habitat is one of the fundamental aspects of ecology. This relationship is critical to understand the biological distribution of species richness and is determined by counting the number of distinct species in different area sizes. Mathematically, this relationship is represented by a function (species-area curve) expressing the number of species in different area sizes in a given habitat. The species-area (*SAR*) curve is an essential tool for creating map biodiversity, predicting losses of species due to habitat destruction and determining the minimum area requirements for preserving a specie or species.

The species-area curve can be classified according to the types of habitats (Scheiner, 2003), continents (counting in contiguous habitats, with quadrats or areas sampled with gradually increasing sizes) or islands (counting in noncontiguous habitats). Different habitats or characteristics of sampling will result in

Key words and phrases. Species-area relationship, species richness, generalized model, discrete probability distribution, effect of minimum area.
Received September 2012; accepted January 2013.

different shapes of species-area curves. Thus, it is advantageous to find theoretical expressions that describe the various empirical curves.

One of the first proposals of a equation used to describe the species-area relationship was a simple function, known as the Arrhenius power law (Arrhenius, 1921), given by $S = \beta_1 A^{\beta_2}$, where S represents the number of species, A is the area and β_1 and β_2 are the parameters of the model. Plotkin et al. (2000) incorporated an upper asymptote parameter into the Arrhenius power law, indicating that the number of species in a habitat cannot increase infinitely. Ulrich and Buszko (2003) also proposed an Arrhenius power law extension, including the capability to represent the effect of the minimal area.

A number of functions used for *SAR* are presented in Tjørve (2003, 2009), Dengler (2009). Guilhaumon et al. (2010) presented an R package (see R Development Core Team, 2009) used to fit different *SAR*s. Different functions proposed in the literature have distinct properties (upper asymptote or effect of the minimal area). There is no single function that includes both upper asymptote and the effect of the minimal area, which are essential properties for a good fit to most datasets. However, Williams et al. (2009) pointed out three deficiencies observed in the proposition of different *SAR*s: (a) there is too much emphasis on maximizing the goodness of fit between species richness and area, ignoring the effects of other factors; (b) the assumptions about the species richness distribution are often inadequate and have not been tested; and (c) the existing fitting formulas use coefficients of determination (R^2) with different error distributions and/or parameter numbers. An important feature, which is often overlooked in studies that propose some fitted method for *SAR*, is the fact that the number of species S is a random variable that takes only integer and non-negative values. This feature requires special assumptions for the probability distribution of this variable.

The main aim of this paper is to present a general *SAR* model which reduces to previously proposed formulas. The proposed fitting formula is suitable for areas of different scales (small, intermediate and large) and allows the inclusion of other covariates, beyond the area, to explain species-area relationship. Moreover, our model considers the Poisson probability distribution for the number of species richness. The advantage of our formulation is that it leads to a unique formula for representing the *SAR*, which takes into account both the effects of the minimum area and asymptotic behavior of the growth curve for large areas. This approach provides a unique algorithm for fitting different datasets, and gives, by the hypotheses test formulation, the opportunity to choose the *SAR* which best fits the data.

This paper is structured as follows. Section 2 presents the proposal generalized species-area relationships ($gSAR$) and the inferential procedure for calculating the maximum likelihood estimators and some criteria for selecting the *SAR* which best fits the data. Section 3 presents a simulation study, which was conducted to determine if two usual selection criteria are suitable to choose the best *SAR* model, and the analysis of artificial datasets and real datasets in fish species richness in

70 lakes, using the proposed (*gSAR*) model. Final discussions are presented in Section 4.

2 Material and methods

Let S be a random variable that represents the species richness, which has a Poisson distribution with parameter μ, such that $\mu = f(A)$, where A is the size of the area. The *gSAR* is given by

$$\mu = E(S) = \beta_0 + \beta_1 \cdot \left[g(A)\right]^{\beta_2} \cdot e^{-\beta_3 A^{\beta_4}}, \tag{2.1}$$

where β_i is the unknown parameters, $i = 0, \ldots, 4$, such that $\beta_0 \geq 0$, $\beta_1 > 0$, $\beta_2 \in (0, 1]$, $\beta_3 \geq 0$ and $\beta_4 \in \Re$; $g(A)$ is a function of the area, given by $g(A) = A$ or $g(A) = \log(A)$. The restrictions imposed on the parameters $\beta_0, \beta_1, \beta_2$ and β_3 are related to particular *SAR*s which will be considered later in this paper. For $g(A) = \log(A)$, β_0 represents the initial species richness, in other words, the species richness for a unit area ($A = 1$). However, this interpretation does not apply to $g(A) = A$. Parameter β_1 is the slope of the curve representing the growth of species richness with an increasing area. Exponent β_2 reflects a typical feature of the power model traditionally fitted for the species-area relationship. Parameter β_3 introduces the flexibility of simultaneously representing the effects of the minimum area and upper asymptote into the model, depending on the sign of parameter β_4. The term $e^{-\beta_3 A^{\beta_4}}$, with $\beta_3 > 0$, represents the persistence function which assigns a pattern to the curvature of the relationship that reflects the effect of the minimum area, if $\beta_4 < 0$, or the asymptotic effect for large areas, if $\beta_4 > 0$.

The proposed *gSAR* has five parameters to be fitted. This can be a problem if datasets do not have many data points, However, it enables the general formulation including several particular cases together in a unique structure. Besides, with the amount of five parameters, even if all particular *SAR*s are not adequate to accommodate a particular dataset, the general formulation keeps enough flexibility to accommodate it. Taking this into account, we carried out a simulation study in order to verify if it is possible to distinguish the *gSAR* and its particular cases in light of datasets. The performed simulation can be seen as a misspecification study and the main concern was to observe whether it was possible to make a distinction between the various formulations, even in the presence of small and moderate sized samples considering two usual statistical criteria (*AIC* and *BIC*). We discovered that, considering both statistical criteria, most fitted models indicated that the best *SAR* was the one used to generate samples, even with small samples, resulting in quite satisfactory percentages of choosing the correct model.

2.1 Particular *SAR*s

Various usual *SAR*s can be seen as particular cases of *gSAR* (2.1). These *SAR*s are usually fitted by ordinary least squares (OLS) and do not make assumptions about the probability distribution of S. In this section, we describe five of them.

2.1.1 *SAR Proposed by Arrhenius* (1921). In a pioneering study, Arrhenius (1921) expressed the species-area relationship as a power law given by,

$$E(S) = \beta_1 A^{\beta_2}, \tag{2.2}$$

this *SAR* is called the power law model, where the variable S represents the number of species in the sampled area A, β_1 and β_2 correspond, in log–log scale, to the expected number of species 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 emphasized that the larger the area, the greater the number of species. This *SAR* is widely used by ecologists (not always for biological causes, usually for convenience) (Tjørve, 2003, Ulrich and Buszko, 2007). Equation (2.2) is a particular case of the *gSAR* (2.1) when the function $g(A) = A$ and parameters $\beta_0 = 0$ and $\beta_3 = 0$.

The identification of the power law as a particular case of *gSAR* facilitates the interpretation of fitted *gSAR* parameters, allowing for comparison of the fitted *gSAR* function with other commonly used functions. In this case, parameters β_1 and β_2 in the *gSAR* model have identical interpretation to these same parameters in the Arrhenius (1921) model.

2.1.2 *SAR Proposed by Gleason* (1922). Gleason (1922) noticed that the Arrhenius *SAR* produces extremely high estimates for the number of species in large areas. To avoid this problem, the author proposed an alternative *SAR* considering that the richness of species is a linear function of the logarithm of the area, such as,

$$E(S) = \beta_0 + \beta_1 \log(A), \tag{2.3}$$

this formulation is called the logarithmic *SAR*, which may be more appropriate to describe the relationship over large areas. Equation (2.3) is a particular case of the *gSAR* (2.1) when the function $g(A) = \log(A)$ and parameters $\beta_2 = 1$ and $\beta_3 = 0$.

2.1.3 *SAR Proposed by Connor and McCoy* (1979). A simple linear regression was proposed by Connor and McCoy (1979) to express the number of species S in function of area A,

$$E(S) = \beta_0 + \beta_1 A, \tag{2.4}$$

which is called the linear *SAR*, where β_0 represents the intercept parameter and β_1 the slope of the regression. This *SAR* is appropriate to small areas. A limitation of this *SAR* is the inconsistency in the estimation of species richness when $A = 0$. Equation (2.4) is a particular case of the *gSAR* (2.1) when function $g(A) = A$ and parameters $\beta_2 = 1$ and $\beta_3 = 0$.

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

$$E(S) = \beta_1 A^{\beta_2} e^{-\beta_3 A}, \tag{2.5}$$

which is called the persistence (P1) *SAR*, where 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. Equation (2.5) is a particular case of the *gSAR* (2.1) when the function $g(A) = A$ and parameters $\beta_0 = 0$ and $\beta_4 = 1$.

The model proposed by Plotkin et al. (2000) provides, in some situation, the existence of an upper asymptote in the species-area curve which limits the increase in the number of species in very large areas. However, in this model the parameter related to the upper asymptote is fixed and always assumes value one. On the other hand, if the presence of an asymptote is revealed when analyzing a given dataset, then the *gSAR* model is able to identify it by fitting the parameter β_4. Moreover, the introduction of this parameter makes the *gSAR* model more flexible and accurate in the identification of the asymptote. Parameters β_3 and β_4 in the *gSAR* model are interpreted as these same parameters in the model proposed by Plotkin et al. (2000).

2.1.5 *SAR Proposed by Ulrich and Buszko* (2003, 2004). To model the effect of small areas in the growth of species richness, Ulrich and Buszko (2003, 2004) proposed a modification of the *SAR* presented in Plotkin et al. (2000) resulting in the following expression,

$$E(S) = \beta_1 A^{\beta_2} e^{-\beta_3 / A}, \tag{2.6}$$

which is called the persistence (P2) *SAR*, 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. Equation (2.6) is a particular case of the *gSAR* (2.1) when function $g(A) = A$ and parameters $\beta_0 = 0$ and $\beta_4 = -1$.

In the model proposed by Ulrich and Buszko (2003, 2004) the effect of small areas is related to the term A^{-1}, this is equivalent to always assign a fixed value to parameter $\beta_4 = -1$, in *gSAR* model. When parameter β_4 assumes a negative value other than -1, *gSAR* models comprise the effect of small areas in a more flexible way.

2.2 Summary of the *SAR*

Table 1 presents a summary of the five *SAR*s, which are particular cases of the proposed *gSAR* (2.1).

Table 1 *Particular cases of the gSAR (2.1) and their characteristics*

Function	$E(S) = f(A)$	Param.	Asymp.	Author
Linear	$\beta_0 + \beta_1 A$	2	No	Connor and McCoy (1979)
Power law	$\beta_1 A^{\beta_2}$	2	No	Arrhenius (1921)
Persistence (P1)	$\beta_1 A^{\beta_2} e^{-\beta_3 A}$	3	Yes	Plotkin et al. (2000)
Persistence (P2)	$\beta_1 A^{\beta_2} e^{-\beta_3 / A}$	3	No	Ulrich and Buszko (2003, 2004)
Logarithmic	$\beta_0 + \beta_1 \log(A)$	2	No	Gleason (1922)

Interpretation of the *gSAR* model parameters is similar to the usual models, which are particular cases of the *gSAR* model. One advantage of using *gSAR* model is that there is no need to fit various models to completely analyze a dataset. Therefore, a dataset analysis is carried out by fitting the *gSAR* model and checking the significance of its parameters, which enables the identification of small area effect, or the asymptote with parameter β_4. The analysis in then completed by checking whether the value of the slope parameter β_2 is within the normal range found in different ecological habitats. Section 3.2.2 presents more details on the advantages of using *gSAR* model to analyze real datasets.

2.3 Inference

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

$$\ell(\mathbf{S}, \mathbf{A}; \beta) = \sum_{i=1}^{n} S_i \log f(A_i) - \sum_{i=1}^{n} \big(f(A_i)\big) - \sum_{i=1}^{n} \log S_i!. \tag{2.7}$$

For the proposed *SAR*, $f(A)$ is given by equation (2.1) and the likelihood function (2.7) can be rewritten as,

$$\begin{aligned}\ell(\mathbf{S}, \mathbf{A}; \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = &\sum_{i=1}^{n} S_i \log\big(\beta_0 + \beta_1 \big(g(A_i)\big)^{\beta_2} e^{-\beta_3 A_i^{\beta_4}}\big) \\ &- \sum_{i=1}^{n} \log S_i! - \sum_{i=1}^{n} \big(\beta_0 + \beta_1 \big(g(A_i)\big)^{\beta_2} e^{-\beta_3 A_i^{\beta_4}}\big).\end{aligned}$$

The maximum likelihood estimates (MLEs) are obtained by direct maximization of the log-likelihood function (2.7). The advantage of this procedure is that it runs immediately using existing statistical packages such as R (R Development Core Team, 2009). We consider software R using the BFGS algorithm (Nocedal

Table 2 *Combinations of H_0 hypotheses for the gSAR*

H_0	$g(A)$	Particular *SAR*
(ii) & (iii)	A	Linear
(i) & (iii)	A	Power law
(i) & (iv)	A	Persistence (P1)
(i) & (v)	A	Persistence (P2)
(ii) & (iii)	$\log(A)$	Logarithimic

and Wright, 2006) to compute the MLEs. As usual, large-sample inferences for the parameters are based on the MLEs and their estimated standard errors.

There are several criteria for choosing *SAR*s that best fit datasets. We consider here the *BIC* (Bayesian information criteria) proposed in Schwarz (1978) and the *AIC* (Akaike information criteria) proposed in Akaike (1974). These criteria are composed of a balance between the maximum log-likelihood function and the number of parameters of the formulation. The *BIC* is given by,

$$BIC = -2\ell(\mathbf{S}, \mathbf{A}; \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) + k\ln(n), \tag{2.8}$$

where $\ell(\mathbf{S}, \mathbf{A}; \beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ is the log-likelihood function value, k is the number of parameters in the formulation and n is the sample size. The *AIC* is a measure of the goodness of the formulation fit given by

$$AIC = -2\ell(\mathbf{S}, \mathbf{A}; \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) + 2k. \tag{2.9}$$

Smaller values of *BIC* and *AIC* indicate a better *SAR* fit, but the *BIC* is more rigorous, as the formulations with a larger number of parameters have greater penalties.

From a practical standpoint, it may be interesting to test some specific hypothesis about the parameters in order to verify the particular cases of the *gSAR* (2.1), that is, to test the following hypotheses: (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) $H_0 : \beta_4 = -1$. The different combinations of these situations result in specific cases of the *gSAR*, as presented in Table 2.

3 Results

3.1 Simulation study

In this section, we present a study which was carried out to observe if it is possible to distinguish the *gSAR* and its particular cases, based on datasets and the criterion described in Section 2.3. We generated samples of sizes $n = 20, 50, 100, 200, 500$ and 1000 from the species richness (S) for area sizes n between 1 and 100, considering the *gSAR*—with $g(A) = A$ and $g(A) = \log(A)$—and each of the *SAR*s described in Table 1. The parameter values considered to generate the data are presented in Table 3. These values were chosen so that all *SAR*s show a growth

Table 3 *Values of the parameters for each model*

Model	β_0	β_1	β_2	β_3	β_4	$g(A)$
Linear	2	2	1	0	0	A
Power law	0	8	0.75	0	0	A
Persistence (P1)	0	8	0.75	0.005	1	A
Persistence (P2)	0	8	0.75	50	-1	A
Logarithimic	2	100	1	0	0	$\log(A)$
gSAR with $g(A) = A$	2	20	0.75	50	-0.8	A
gSAR with $g(A) = \log(A)$	8	100	0.75	0.005	0.5	$\log(A)$

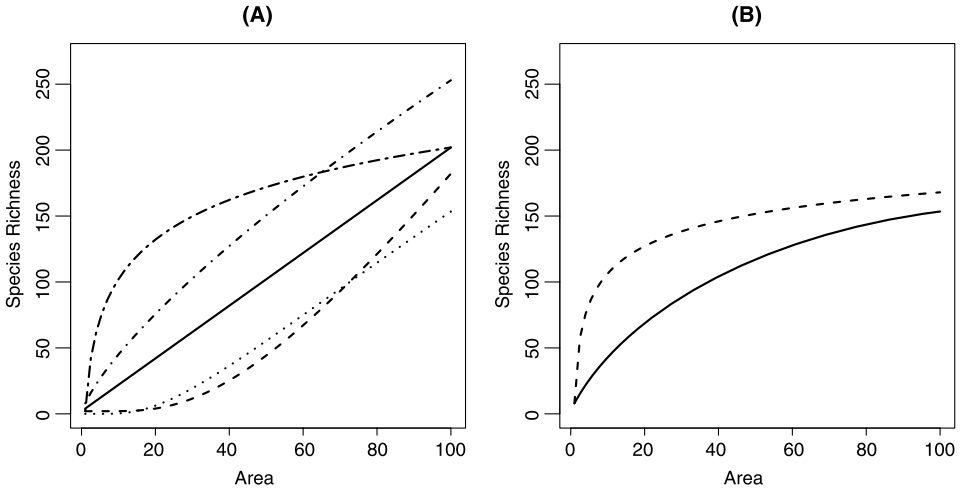

Figure 1 *SAR scatterplots*: (A) *without asymptote*; - *Linear*; -·- *Power law*; ··· *Persistence* (*P*2); – - – *Logarithmic*; - - - *gSAR with* $g(A) = A$. (B) *with asymptote*; - *Persistence* (*P*1); - - - *gSAR with* $g(A) = \log(A)$.

pattern on the same scale. Figure 1 shows the behavior of the *SAR*s, considering the parameters in Table 3.

We considered two groups of *SAR*s: the first, in which $g(A) = A$, and the second, where $g(A) = \log(A)$. For both groups, we generated a sample of each *SAR* and fitted all seven *SAR*s to each sample. This procedure was repeated 1000 times to verify if the *SAR*s selected by the adopted criteria coincide with the true *SAR*s used to generate the data. Table 4 shows the percentages of choosing the correct model of each criterion for different sample sizes In general, it can be stated that the *SAR* from which the sample was generated is indicated as the best model. We observed that most fitted models indicated that the best *SAR* is the one used to generate the samples. A relevant point is that, regardless of the sample size, the use of the *AIC* and *BIC* criteria for selecting the true *SAR* resulted in quite satis-

Table 4 *Percentages of choosing the correct model of each criterion for selecting the true model*

		$g(A) = A$					$g(A) = \log(A)$	
n	Criteria	Linear	Power law	Persistence (P1)	Persistence (P2)	*gSAR*	Logarithmic	*gSAR*
20	*AIC*	74.5	83.6	69.9	94.3	99.1	90.6	99.9
	BIC	75.8	89.5	64.7	98.0	98.9	98.2	99.9
50	*AIC*	91.2	90.5	85.2	94.4	99.7	85.5	99.9
	BIC	92.9	97.4	82.2	98.0	99.7	98.9	99.9
100	*AIC*	88.5	85.3	93.9	93.2	99.9	82.4	99.9
	BIC	90.9	90.0	94.9	99.2	99.9	99.6	99.9
200	*AIC*	94.5	97.4	63.5	93.2	99.9	75.5	99.9
	BIC	95.1	99.4	66.4	97.8	99.9	99.4	99.9
500	*AIC*	99.4	99.2	99.8	84.6	99.9	82.5	99.9
	BIC	99.6	99.7	99.9	99.6	99.9	99.9	99.9
1000	*AIC*	99.6	99.9	78.3	88.9	99.9	75.3	99.9
	BIC	99.6	99.9	80.5	95.6	99.9	99.9	99.9

factory percentages of choosing the correct model. The proposed *SAR*, with both structures $g(A) = A$ and $g(A) = \log(A)$, obtained the highest percentages.

Table 4 shows the percentages of choosing the correct model of each criterion for different sample sizes. In general, it can be stated that the *SAR* from which the sample was generated is indicated as the best model.

3.2 Application

In this section, we illustrate the flexibility of the proposed *gSAR* and compare it with its particular cases in artificial datasets and real datasets of fish species richness from 70 lakes in the world previously analyzed in Barbour and Brown (1974) and Stein and Juritz (1988).

3.2.1 *Artificial data.* In the study with artificial data, a sample size $n = 100$ was generated from the *gSAR* (2.1) with $g(A) = \log(A)$, considering areas with sizes between 1 and 100. Subsequently, all models were fitted in order to find the relationships that best explains the data behavior. Table 5 shows the selection criteria values considered for all models. We observed an agreement in the choice of the best *SAR* by considering both the *AIC* and *BIC* criteria. They indicate that the best fit is obtained from the *SAR* in which the data were generated, that is, with the *gSAR* one.

The panel (A) of Figure 2 presents the scatterplot of the artificial dataset, the function curve used to generate the data and the fitted one. It can be seen that the proposed *SAR* represents the data behavior well, and it also captures the smoothness of the growth curve for large areas and the minimum area effect. Another

Table 5 *Selection criteria of the SARs fitted for the artificial data*

	$g(A) = A$					$g(A) = \log(A)$	
Criteria	Linear	Power law	Persistence (P1)	Persistence (P2)	*gSAR*	Logarithmic	*gSAR*
AIC	923.52	789.40	731.38	707.19	707.00	1261.27	**696.18**
BIC	928.73	794.61	739.19	715.01	720.02	1266.48	**709.20**

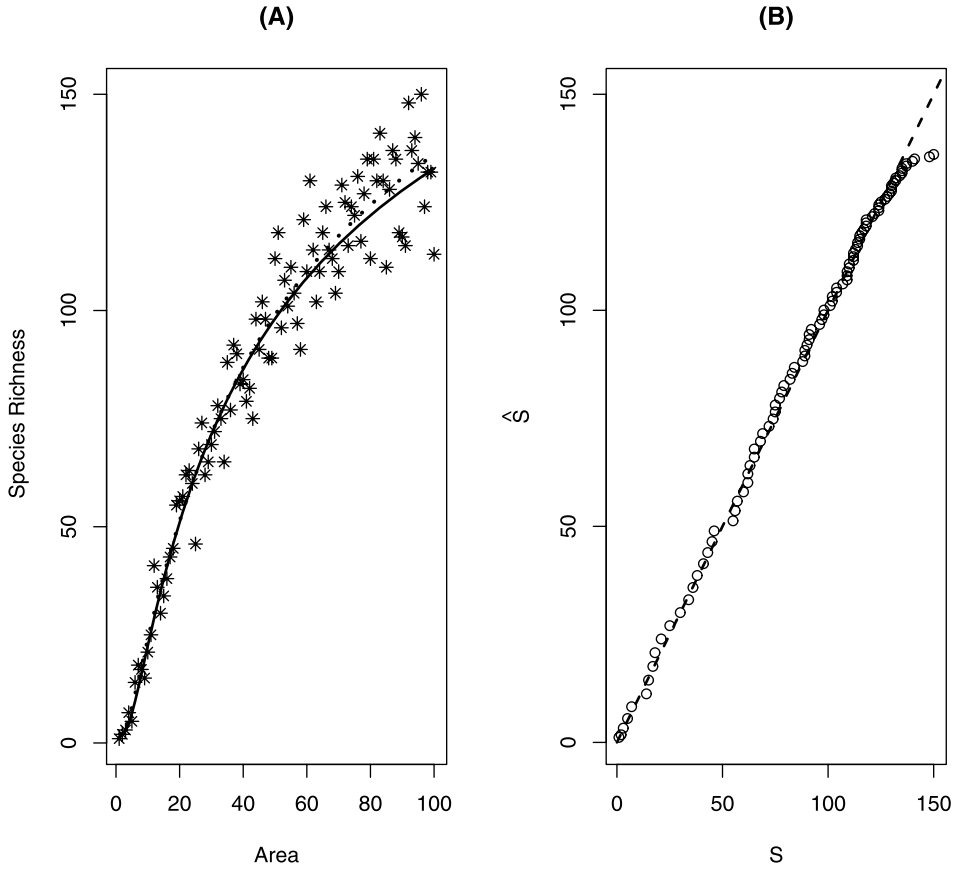

Figure 2 *Artificial model and fitted model to the simulated data of gSAR*, $g(A) = \log(A)$, *with sample size* $n = 100$: (A) *area in the original scale* (* *artificial data*; — *artificial model*; - - - *fitted model*) *and* (B) *QQ-plot*.

way to check the goodness of fit is to construct the QQ-plot of the estimated species richness with each fitted *SAR*. Indicative of goodness fit is the proximity of points to the 45° line. The panel (B) of Figure 2 shows that the *gSAR*, with $g(A) = \log(A)$, has a higher adhesiveness to the points.

Table 6 *Estimates of parameters of the fitted SARs*

Model	Parameters					Criteria	
	β_0	β_1	β_2	β_3	β_4	*AIC*	*BIC*
Linear	31.17	4×10^{-4}	–	–	–	2426.49	2430.98
Power law	–	8.73	0.20	–	–	1900.36	1904.86
Persistence (P1)	–	8.00	0.22	5×10^{-7}	–	1910.55	1917.29
Persistence (P2)	–	8.55	0.20	2×10^{-3}	–	1902.38	1909.12
gSAR with $g(A) = A$	13.50	1.65	0.34	71.88	−18.85	**1855.48**	**1866.73**
Logarithmic	6.99	11.91	–	–	–	2169.36	2173.86
gSAR with $g(A) = \log(A)$	6.98	11.91	1.00	0.01	−15.87	2175.36	2186.60

3.2.2 *Real data*. We consider the fish species richness dataset presented in Barbour and Brown (1974). The dataset consists on fish species richness 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, the Soviet Union, the United States, Romania and Yugoslavia. These data were analyzed by Barbour and Brown (1974), considering the power law of Arrhenius (1921), and by Stein and Juritz (1988), considering a linear relationship between the expected value of the number of species and the logarithm of the area.

We fitted the *gSAR* and all its particular cases to the data to check which function best describes the relationship between species richness and area of the lakes. Table 6 shows the parameter estimates and the selection criteria for each *SAR*. The *gSAR*, with $g(A) = A$, presents the lowest values for the *AIC* and *BIC*, indicating that it has the best fit among the *SAR*s considered for these data.

Figure 3(A) illustrates the *gSAR* fitted for the data in fish species richness according to the lake areas in logarithmic scale. The proposed *SAR* adequately fits the dataset. Comparing the points in this figure, we notice that, for some areas, there are points with high fish species richness. In contrast, there are overly large areas where the diversities are not as high as those of a few relatively small areas. For this reason, we believe these extreme points, hereafter outliers, may have influenced the quality of the adjustments (see Cook, 1986, Paula and Cysneiros, 2010). These outliers represent the fish species richness in the lakes Malawi, Tanganyika and Victoria (all in Africa).

For the sake of illustration, we dropped the outliers and fitted the model again. Table 7 presents the estimates of each parameter and the selection criteria, which even after dropping the outliers continue indicating the *gSAR* with $g(A) = A$ as having the best fit. Figure 3(B) illustrates the *gSAR* fitted to the fish species richness data, without outlier points, according to the lake areas on a logarithmic scale.

Barbour and Brown (1974) considered the power law to fit the species-area curve to this fish richness dataset. The analysis resulted in a low slope to species-area curve ($\beta_1 = 0.16$, in Table 7) falling below the usual range reported for islands

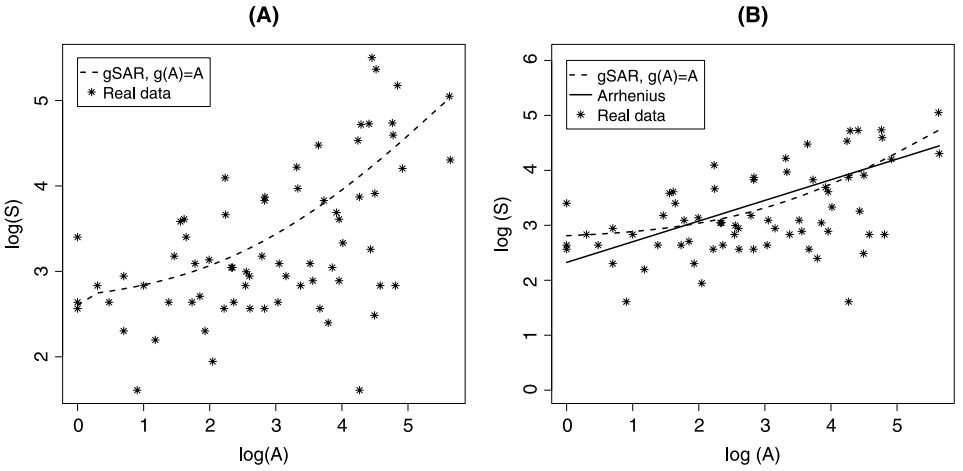

Figure 3 *Fitted model for fish species richness data from lakes in the world*: (A) *complete dataset and* (B) *dataset without outlier points*.

Table 7 *Estimates of parameters of the fitted SARs, without outlier points*

	Parameters					Criteria	
Model	β_0	β_1	β_2	β_3	β_4	*AIC*	*BIC*
Linear	27.65	3×10^{-4}	–	–	–	1413.15	1417.56
Power law	–	10.22	0.16	–	–	1252.27	1256.68
Persistence (P1)	–	10.23	0.16	4×10^{-9}	–	1254.48	1261.09
Persistence (P2)	–	10.24	0.16	2×10^{-7}	–	1254.27	1260.88
gSAR with $g(A) = A$	15.68	2.63	0.31	1.03	−0.07	**1208.93**	**1219.95**
Logarithmic	8.96	8.85	–	–	–	1371.21	1375.62
gSAR with $g(A) = \log(A)$	18.89	16.42	0.97	14.57	−0.30	1238.49	1249.51

(0.2–0.35), Bell et al. (2005). On the other hand, the slope value estimated for this same dataset using the *gSAR* model was $\beta_1 = 0.31$ (see Table 7 with $g(A) = A$). We note that models of persistence (P1) and (P2) fitted to this dataset led to the same power law slope. Moreover, the low value estimated for parameter β_3 in both (P1) and (P2) persistence models indicate no upper asymptote or small area effect. However, fitting the *gSAR* model to the fish richness dataset revealed a small area effect. We also can note that for the same area variation, the fitted *gSAR* model shows an increase in the number of species faster than the Barbour and Brown model. The power law model fitted by Barbour and Brown (1974) and the fitted *gSAR* model are shown in Figure 3(B).

Barbour and Brown (1974) also considered the power law model to separately analyze two datasets consisting of the number of fish species in 14 lakes in North America and 14 lakes in Africa. We have also fitted the *gSAR* model to these two

Table 8 *Estimates of parameters of the fitted Species-Areas curves, for North American and African lakes*

Lakes	Model	Parameters					Criteria	
		β_0	β_1	β_2	β_3	β_4	*AIC*	*BIC*
North American	Power law	–	10.38	0.19	–	–	336.71	337.99
	gSAR with $g(A) = A$	26.15	2.14	0.37	4.23	−0.16	334.44	337.63
African	Power law	–	0.89	0.47	–	–	360.77	362.05
	gSAR with $g(A) = A$	14.59	0.13	0.67	4.27	−2.53	330.87	334.06

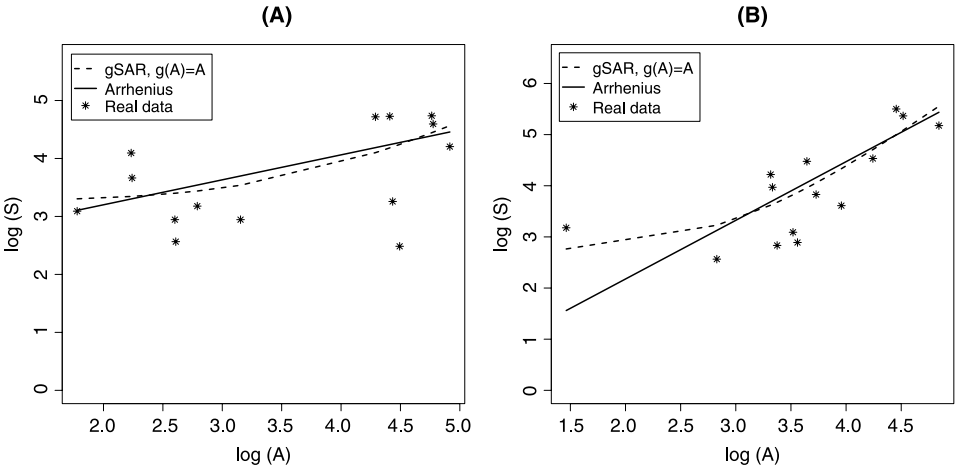

Figure 4 *Fitted model for fish species richness data from lakes in*: (A) 14 *North American lakes*; (B) 14 *African lakes*.

datasets and parameter estimates are given in Table 8. We note that the *gSAR* model provided a slope value of $\beta_2 = 0.37$ (for the lakes in North America) and $\beta_2 = 0.67$ (for the lakes in Africa), both being greater than the values obtained using the power law (Barbour and Brown, 1974), which were $\beta_2 = 0.19$ (for the lakes in North America) and $\beta_2 = 0.47$ (for the lakes in Africa). We can also note that, for both datasets, the parameter β_4 assumes a negative value, meaning that there is effect of small areas. A comparison of the fitted models is shown in Figure 4.

4 Discussion

Many *SAR*s have been proposed in the literature to explain the relationship between the species richness number and the habitat area. In this paper, we propose the *gSAR* (2.1), which takes into account the effect of the small area and also the behavior pattern of this relationship for large areas. The proposed *SAR* has vari-

ous usual formulations used to explained the species-area relationship as particular cases, which can be properly selected via two different criteria (*AIC* and *BIC*). The results of a simulation study showed that both *AIC* and *BIC* selection criteria are suitable to decide on the best *SAR* to describe the relationship between the species richness number and the habitat area in light of a particular data set. The empirical study based on a real dataset of 70 lakes in the world and two datasets consisting of 14 lakes in North America and 14 in reveals that the proposed extension outperforms its particular cases to fit these datasets.

It is worth noting that, to analyze these datasets on fish species richness, Barbour and Brown (1974) considered a multiple linear regression model. That is, the expected value of species richness number is a linear function of logarithm of the area and latitude, and a normal distribution is assumed for the species richness number. However, we believe that this procedure is questionable from a statistical point of view since the response variable (species richness number) is discrete, and by using this procedure it is assumed to be continuous. Stein and Juritz (1988) considered the logarithm of the expected value of the species richness number as a linear function of the logarithm of the area, assuming a Poisson–Inverse Gaussian error distribution. The fitted model showed large deviations between the estimated and observed value of species richness numbers for large areas, resulting in a mean square error of 70% of the variance of S. As a crude comparison, the proposed *gSAR* reduced this error to 58%, a gain of about 17% in a mean square error.

A natural extension of this work may be obtained by considering different probability distributions for the species richness number. In particular, distributions involving overdispersion and underdispersion. In this context, del Castillo and Pérez-Casany (2005) may be investigated further.

Acknowledgments

K. S. Conceição is supported by the Brazilian organization FAPESP (2008/10613-1). M. G. Andrade is supported by the Brazilian organizations FAPESP (2007/07612-0) and CAPES (PROCAD 177/2007). F. Louzada is supported by the Brazilian organization CNPq (301912/2009-2). We are indebted to the Editorial Boarding and Referees for their valuable comments, criticisms and suggestions which have substantially improved the text of the manuscript.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716–723. MR0423716

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

Barbour, C. D. and Brown, J. H. (1974). Fish species diversity in lakes. *The American Naturalist* **962**, 473–489.

Bell, T., Ager, D., Song, J.-I., Newman, J. A., Thompson, I. P., Lilley, A. K. and Gast, C. J. (2005). Larger islands house more bacterial taxa. *Science* **308**, 1884.

Connor, E. F. and McCoy, E. D. (1979). The statistics and biology of the species-area relationship. *The American Naturalist* **113**, 791–833. MR0597208

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society (Series B)* **48**, 133–169. MR0867994

del Castillo, J. and Pérez-Casany, M. (2005). Overdispersed and underdispersed Poisson generalizations. *Journal of Statistical Planning and Inference* **134**, 486–500. MR2200069

Dengler, J. (2009). Which function describes the species-area relationship best? A review and empirical evaluation. *Journal of Biogeography* **36**, 728–744.

Gleason, H. A. (1922). On the relation between species and area. *Ecology* **3**, 158–162.

Guilhaumon, F., Mouillot, D. and Gimenez, O. (2010). mmSAR: An R-package for multimodel species-area relationship inference. *Ecography* **33**, 420–424.

Nocedal, J. and Wright, S. J. (2006). *Numerial Optimization*, 2nd ed. New York: Springer. MR2244940

Paula, G. A. and Cysneiros, F. J. A. (2010). Local influence under parameter constraints. *Communications in Statistics—Theory and Methods* **39**, 1212–1228. MR2745375

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. and Ashton, P. S. (2000). Predicting species diversity in tropical forests. *Proceedings of the National Academy of Sciences of the USA* **97**, 10850–10854.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org, ISBN 3-900051-07-0.

Scheiner, S. (2003). Six types of species-area curves. *Global Ecology and Biogeography* **12**, 441–447.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464. MR0468014

Stein, G. Z. and Juritz, J. M. (1988). Linear models with an inverse Gaussian–Poisson error distribution. *Communications in Statistics—Theory and Methods* **17**, 557–571. MR0939667

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

Tjørve, E. (2009). Shapes and functions of species-area curves (II): A review of new models and parameterizations. *Journal of Biogeography* **36**, 1435–1445.

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

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

Ulrich, W. and 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. and Henstridge, J. (2009). Species-area functions revisited. *Journal of Biogeography* **36**, 1994–2004.

K. S. Conceição
R. M. Pires
C. A. R. Diniz
Departamento de Estatística
Universidade Federal de São Carlos
C.P. 676, 13565-905
São Carlos-SP
Brasil
E-mail: katianeconceicao@gmail.com
rubianemariapires@gmail.com
dcad@ufscar.br

F. Louzada
M. G. Andrade
Departamento de Matemática Aplicada e Estatística
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
C.P. 668, 13560-970
São Carlos-SP
Brasil
E-mail: louzada@icmc.usp.br
marinho@icmc.usp.br