OPEN ACCESS Freely available online

PLOS ONE

A Generalized Approach to the Modeling of the Species-Area Relationship

Katiane Silva Conceição[1], Werner Ulrich[2], Carlos Alberto Ribeiro Diniz[3], Francisco Aparecido Rodrigues[1], Marinho Gomes de Andrade[1]*

1 Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, São Paulo, Brazil, **2** Department of Animal Ecology, Nicolaus Copernicus University in Toruń, Toruń, Poland, **3** Department of Statistics, Federal University of São Carlos, São Carlos, Brazil

Abstract

This paper proposes a statistical generalized species-area model (GSAM) to represent various patterns of species-area relationship (SAR), which is one of the fundamental patterns in ecology. The approach enables the generalization of many preliminary models, as power-curve model, which is commonly used to mathematically describe the SAR. The GSAM is applied to simulated data set of species diversity in areas of different sizes and a real-world data of insects of *Hymenoptera* order has been modeled. We show that the GSAM enables the identification of the best statistical model and estimates the number of species according to the area.

Citation: Conceição KS, Ulrich W, Diniz CAR, Rodrigues FA, Andrade MGd (2014) A Generalized Approach to the Modeling of the Species-Area Relationship. PLoS ONE 9(8): e105132. doi:10.1371/journal.pone.0105132

Editor: Enrico Scalas, Universita' del Piemonte Orientale, Italy

Received December 3, 2013; **Accepted** July 21, 2014; **Published** August 29, 2014

Funding: K. S. Conceição would like to acknowledge FAPESP (grant # 2008/10613-1). F. A. Rodrigues would like to thank CNPq (grant # 305940/2010-4), Fapesp (grant # 2013/26416-9) and NAP eScience - PRP - USP for financial support. The authors also thank Fapesp (grant #2014/15860-8) for funding the publication of this paper. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: marinho@icmc.usp.br

Introduction

The variation in the number of species with area, known as species-area relationship (SAR), is one of the most important ecological patterns [1]. The models of SAR enable the prediction of the number of species that coexist and share resources, as well as the impact of the extinction of species caused by habitat loss. Sampled data for a single species, or all species of a specific trophic level within a particular site have shown that the SAR has a well-defined shape, most often described by power and exponential curves [2]. The number of species in an area increases with increasing island area, but the rate of increase slows for larger islands. Many hypotheses have been proposed to explain the SAR [3,4,6]. For instance, some are based on the immigration and extinction of species [4], random sampling processes [6] or the Second Law of Thermodynamics [5].

These different hypotheses have generated many mathematical models for the description of the SAR [1–3,7–10]. The early models were based on deterministic modeling, which assumes that every set of variable states is uniquely determined by the parameters in the model. For instance, Arrhenius considered that the number of species (S) is related to area (A) through a power law form [11] (called the power-function), i.e. $S = S_0 A^z$ (or $\log S = \log c + z \log A$), where S_0 represents the number of species in a unit area ($A = 1$) and $0 < z < 1$. Due to the random nature of the sampled data, statistical modeling is more suitable for SAR description than the deterministic approach [6], therefore, many statistical models have been developed (e.g. [2]). Moreover, statistical models can be thought of as general cases of deterministic models, because the mean value of the random variable of interest yields the results of the deterministic model.

Because there exist many models to address the SAR (e.g. [3,9,12]), a natural question is how to select the best model for a given data set. To address this issue, here we integrate different models within a common framework and consider the problem of curve fitting by the transformed generalized linear model (TGLM) [13]. We propose the use of the generalized species-area model (GSAM) to describe the SAR. GSAM includes many models, such as those described in [3,9] as special cases. We also consider a model that simulates the colonization process of a region by different species and show that the GSMA has best fitted the data in comparison with traditional power-curve models. Finally, we use the data on the cumulative species richness of parasitic *Hymenoptera* from 25 nested plots in a beech forest on limestone [14]. Our results show that the GSAM can determine the best model for the data and estimate the number of species accurately.

Methods

Generalized species-area models

In species-area curves, the number of species (S) is the dependent variable and the area (A) is the explanatory variable. Some mathematical models of SAR propose that the number of species is related to the area as

$$S_i = \mu(A_i, \boldsymbol{\beta}), \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}$ is a p-parameter vector [1]. Function μ can be derived from laws governing the physical system that gave rise to the data. As such models are deterministic and the properties related to the random nature of variable S are neglected, the deterministic models are often inadequate due to the stochastic nature of the

data [6]. The statistical modeling usually assumes that Eq. 1 can be written as

$$S_i = \mu(A_i, \boldsymbol{\beta}) + \varepsilon_i, \tag{2}$$

where $\{\varepsilon_i\}$ are independent and identically distributed random noises (*i.i.d*) — usually, $\varepsilon_i \sim N(0,\sigma^2)$. Note that the mode in Eq. 2 is a generalization of the determinic model, i.e. $E(S_i) = \mu(A_i, \boldsymbol{\beta})$. The model given by Eq. 1 can be understood as a particular case of the model $E\{\Lambda(S_i,\lambda)\} = \mu(A_i, \boldsymbol{\beta})$, where $\Lambda(\cdots,\lambda)$ is a monotonic transformation and λ is a scalar parameter defining such a transformation. For instance, in cases whose data suggested by Eq. 2 are unsatisfactory, the experimenter can assume a model with logarithmic transformation, i.e.

$$\log(S_i) = \mu(A_i, \boldsymbol{\beta}) + \varepsilon_i. \tag{3}$$

This paper proposes a new model called generalized species-area model (GSAM), which is based on the TGLM approach proposed in [13]. The GSAM works with a general parametric family of transformations from the dependent variable S to $S^{(\lambda)} = \Lambda(S;\lambda)$ and postulates that the transformed random variable $S^{(\lambda)}$ follows a continuous probability distribution belonging to the exponential family. Furthermore, the GSAM assumes that there exists some λ value such that $S^{(\lambda)}$ satisfies the usual assumptions of the generalized linear models (GLM) [15].

A suitable choice of the family of transformations enables the representation of power-curves, their recent extensions (see [11,16–19]), the models presented in [2,3,9] and the logarithmic model described in [16] as special cases of the GSAM. We have considered the Box-Cox power transformation [20], which is effective at turning skewed unimodal distributions into nearly symmetric normal-like distributions.

Let $\boldsymbol{S} = (s_1, \ldots s_n)^T$ be the vector of observations. By using

$$S^{(\lambda)} = \begin{cases} \dfrac{S^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(S) & \text{if } \lambda = 0, \end{cases} \tag{4}$$

we can obtain the transformed observations $\boldsymbol{S}^{(\lambda)} = (s_1^{(\lambda)}, \ldots s_n^{(\lambda)})^T$. The GSAM assumes that there exists some λ value such that the transformed random variables $\{S_1^{(\lambda)}, \ldots S_n^{(\lambda)}\}$ can be considered independently distributed. Each $S_i^{(\lambda)}$ follows an exponential family distribution with a probability density function of the form

$$\pi(s_i^{(\lambda)}; \theta_i, \phi) = \exp\left[\phi^{-1}\left\{s_i^{(\lambda)}\theta_i - b(\theta_i)\right\} + c(s_i^{(\lambda)}, \phi)\right], \tag{5}$$

where $b(x)$ and $c(x,\phi)$ are appropriate known functions. The dispersion parameter ϕ is assumed to be the same for all observations. The mean and variance of $S_i^{(\lambda)}$ are, respectively, $E\{S_i^{(\lambda)}\} = \mu_i = db(\theta_i)/d\theta_i$ and $Var\{S_i^{(\lambda)}\} = \phi\, V(\mu_i)$, where $V(\mu_i) = d^2b(\theta_i)/d\theta_i^2 = d\mu_i/d\theta_i$ is the variance function. Parameter $\theta_i = \int V^{-1}(\mu_i)d\mu_i = q(\mu_i)$ is a known one-to-one function of μ_i.

The GSAM also considers a systematic component given by

$$g(\mu_i) = \eta_i = \boldsymbol{x_i^T}\boldsymbol{\beta}, \tag{6}$$

where the link function $g(\cdot)$ is a known one-to-one continuously differentiable function and $\boldsymbol{x_i}$ is a specified vector $(p \times 1)$ of the explained variables, which include the area, known functions of the area, and other environmental variables. Matrix $\boldsymbol{X}$ whose rows are vectors $\boldsymbol{x_i^T}$, $i = 1, \ldots, n$, is a specified $n \times p$ model matrix of full rank $p < n$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a set of unknown linear parameters to be estimated. The link function is assumed to be monotonic and differentiable.

The GSAM proposed here considers three components of structural importance: (i) the Box-Cox family of transformations (Eq. 4) in association with a more general form for the distribution of the transformed variable $S^{(\lambda)}$ (Eq. 5); (ii) a linear predictor function and (iii) a possible nonlinear link function for the regression parameters (Eq. 6). Moreover, when the variance function $V(\mu_i)$ is not constant, i.e. when the variance is correlated with mean μ, some distributions of the exponential family enable the handling of data presenting heteroscedasticity. In this context, GSAM is a generalization of the previous mathematical models that describe the SAR.

Species-area relationship models. Many models have been proposed for the description of the SAR and some can be linearized by a logarithmic transformation of the response variable (i.e. diversity of species). Models that are special cases of the GSAM have the following properties: (i) the transformation parameter in Eq. 4 is $\lambda = 0$, i.e. a log-transformation is adopted, $S^{(0)} = \log(S)$ and $\mu = E\{S^{(0)}\} = E\{\log(S)\}$ or no transformation is considered for variable S, i.e. we assume $\lambda = 1$ and $\mu = E\{S\}$; (ii) the distribution in Eq. 5 is the normal distribution; and (iii) the link function is the identity function, $g(\mu) = \mu$ and the systematic component in Eq. 6 is given by $\mu = \boldsymbol{X}\boldsymbol{\beta}$. The elements of matrix $\boldsymbol{X}$ may be area A, $\log(A)$ or additional variables, as in [12]. Very simple forms of the systematic component are given by $\mu = \beta_0 + \beta_1 \log A$ or $\mu = \beta_0 + \beta_1 A$. This special case of the GSAM can be understood as the particular cases proposed in [11,16]. A list of some models of SAR is provided.

1. Considering the stochastic nature of variable S, which represents the number of species, the power model proposed by Arrhenius [11] can be written as,

$$E\{S\} = b_0 A^{b_1}. \tag{7}$$

The logarithm of variable S yields

$$E\{\ln(S)\} = \beta_0 + \beta_1 \ln(A), \tag{8}$$

where the parameters of the power-curve in log-log space are $\beta_0 = \ln(b_0)$ and $\beta_1 = b_1$.

2. The persistence model (P1-full) proposed by Plotkin et al. [17] is given by

$$E\{S\} = b_0 A^{b_1} \exp\left\{\sum_{k=1}^{n} b_{k+1} A^k\right\}, \tag{9}$$

or, considering the logarithm of variable S,

$$E\{\ln(S)\} = \beta_0 + \beta_1 \ln(A) + \sum_{k=1}^{n} \beta_{k+1} A^k, \tag{10}$$

where $\beta_0 = \ln(b_0)$, $\beta_1 = b_1$ and $\beta_{k+1} = b_{k+1}$, $k = 1, \ldots, n$. A special case, when $n = 1$ (P1 model [17]) is given by

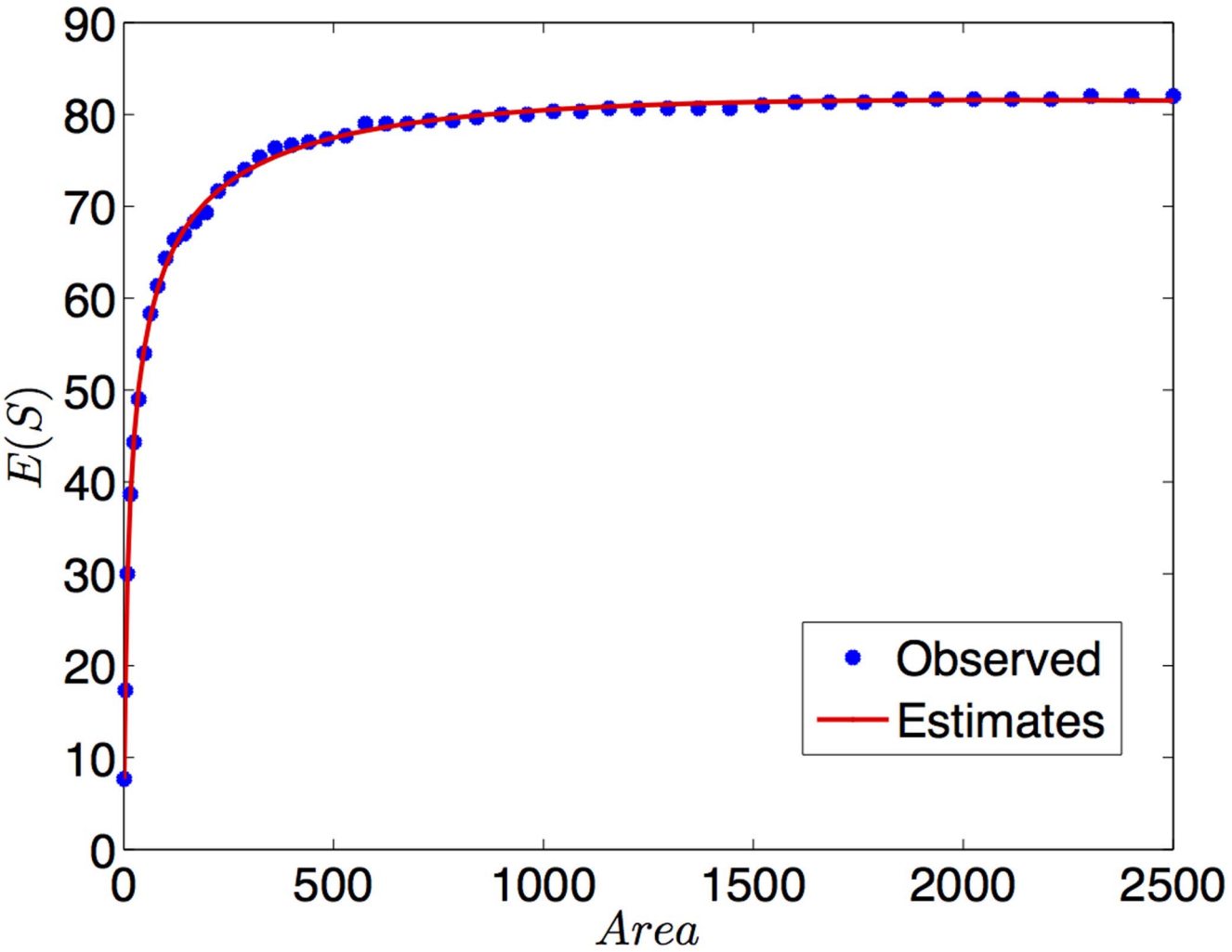

Figure 1. Data and fitted curve obtained by the GSMA model for the simulated species-area relationship.
doi:10.1371/journal.pone.0105132.g001

$$E\{S\} = b_0 A^{b_1} \exp\{b_2 A\}, \tag{11}$$

or, considering the logarithm of variable S,

$$E\{\ln(S)\} = \beta_0 + \beta_1 \ln(A) + \beta_2 A \tag{12}$$

where $\beta_0 = \ln(b_0)$, $\beta_1 = b_1$ and $\beta_2 = b_2$.

3. The persistence model proposed by Ulrich and Buszko [18,19] is given by

$$E\{S\} = b_0 A^{b_1} \exp\left\{\frac{b_2}{A}\right\}, \tag{13}$$

or, considering the logarithm of variable S,

$$E\{\ln(S)\} = \beta_0 + \beta_1 \ln(A) + \frac{\beta_2}{A}, \tag{14}$$

where $\beta_0 = \ln(b_0)$, $\beta_1 = b_1$ and $\beta_2 = b_2$.

4. The polynomial power-function model proposed by Chiarucci et al. [21] is defined as

$$E\{S\} = 10^{\left(\sum_{k=0}^{n} b_k A^k\right)}, \tag{15}$$

or, considering the logarithm of variable S,

$$E\{\ln(S)\} = \sum_{k=0}^{n} \beta_k A^k, \tag{16}$$

where $\beta_k = \ln(10^{b_k})$, $k = 0,1,\ldots,n$. The quadratic power-function model proposed in [21] considers $n = 2$, i.e.

$$E\{S\} = 10^{\{b_0 + b_1 \log(A) + b_2 (\log(A))^2\}}, \tag{17}$$

or, considering the logarithm of variable S,

$$E\{\ln(S)\} = \beta_0 + \beta_1 \log(A) + \beta_2 (\log(A))^2, \tag{18}$$

where $\beta_k = \ln(10^{b_k})$, $k = 0,1,2..$

Some species-area relationships may also be represented by linear functions, specifically:

Table 1. Some of the traditional models adjusted with Gaussian errors.

$\mu = E\{S\}$	**Selection Criteria**				
Models	***AIC***	***BIC***	$-2logL$	***MSE***	***MAPE***
[4]	406.59	424.09	400.59	0.608	0.305
[16]	297.87	315.34	291.87	0.069	0.086
[22]	228.63	251.93	220.63	0.016	0.050
$\mu = E\{\log(S)\}$	**Selection Criteria**				
Models	***AIC***	***BIC***	$-2logL$	***MSE***	***MAPE***
[11]	405.91	423.38	399.91	0.340	0.155
[17])	358.24	381.54	350.25	0.123	0.091
[18]	331.76	355.05	323.76	0.082	0.071
[21]	221.09	244.39	213.09	0.010	0.025

doi:10.1371/journal.pone.0105132.t001

5. Linear model proposed by MacArthur and Wilson [4]

$$E\{S\} = b_0 + b_1 A. \tag{19}$$

6. Logarithmic function proposed by Gleason [16]

$$E\{S\} = \beta_0 + \beta_1 \ln(A). \tag{20}$$

7. Quadratic logarithmic function proposed by Gitay et al. [22]

$$E\{S\} = \{b_0 + b_1 \ln(A)\}^2, \tag{21}$$

or

$$E\{S\} = \beta_0 + \beta_1 \ln(A) + \beta_2 (\ln(A))^2, \tag{22}$$

where $\beta_0 = b_0^2$, $\beta_1 = 2b_0 b_1$ and $\beta_2 = b_1^2$

8. General power-logarithmic function proposed by Gitay et al. [22]

$$E\{S\} = \{b_0 + b_1 \ln(A)\}^{b_2}. \tag{23}$$

If b_2 is any real number and $|b_0| > |b_1 \ln(A)|$, from Newton's generalized binomial expansion we obtain

$$\{b_0 + b_1 \ln(A)\}^{b_2} = (b_0)^{b_2} \sum_{k=0}^{\infty} \binom{b_2}{k} \left(\frac{b_1}{b_0}\right)^k (\ln(A))^k, \tag{24}$$

where the binomial coefficients with an arbitrary upper index can be defined as

$$\binom{b_2}{k} = \frac{b_2(b_2-1)\cdots(b_2-k+1))}{k!} = \frac{(b_2)_k}{k!}. \tag{25}$$

Therefore, the logarithmic function model can be written as

$$E\{S\} = \beta_0 + \sum_{k=1}^{\infty} \beta_k (\ln(A))^k, \tag{26}$$

where $\beta_0 = (b_0)^{b_2}$, $\beta_k = (b_0)^{b_2}\{(b_2)_k/k!\}(b_1/b_0)^k$.

Full-scale generalized species-area relationship model. The right side of all equations presented in the previous section always involves polynomial terms, such as $\ln(A)$, A and/or $1/A$. Here, we propose a generalization of these models by considering the right side of the full-scale model consists of three polynomials, i.e.

$$P_1(A) = \sum_{k=1}^{m} \beta_k (\ln(A))^k, \tag{27}$$

Table 2. The GSAMs fitted with different models according to the likelihood method.

Models	$g(\mu)$	**Systematic component**
normal	μ	$\beta_0 + \beta_1 \ln(A) + \beta_2 (\ln A)^2 + \beta_3/A$
gamma	$\frac{1}{\mu}$	$\beta_0 + \beta_1 \ln(A) + \beta_2 (\ln A)^2 + \beta_3/A + \beta_4/A^2$
I.G.	$\frac{1}{\mu^2}$	$\beta_0 + \beta_1 \ln(A) + \beta_2 (\ln A)^2 + \beta_3/A + \beta_4 A^2$

doi:10.1371/journal.pone.0105132.t002

Table 3. Selection criteria for the GSAMs fitted with $\hat{\lambda}$ adjusted according to the likelihood method.

Model	Parameter	Selection Criteria				
	$\hat{\lambda}$	*AIC*	*BIC*	$-2logL$	*MSE*	*MAPE*
normal	0.822(0.729; 0.922)	79.90	109.02	69.90	0.08	0.61
gamma	0.328(−0.003,0.680)	116.82	151.76	104.82	0.09	0.64
I.G.	−0.089(−0.322,0.141)	142.87	177.82	130.88	0.11	0.68

doi:10.1371/journal.pone.0105132.t003

$$P_2(A)=\sum_{k=m+1}^{n}\beta_k\left(\frac{1}{A}\right)^{k-m}, \tag{28}$$

and

$$P_3(A)=\sum_{k=m+n+1}^{q}\beta_k A^{k-(m+n)}. \tag{29}$$

The left sides of those equations have the number of species (S) or $\ln(S)$. In order to generalize them, we have assumed that S is a random variable and considered the Box-Cox power transformation (Eq. 4).

The curves defined by the GSAM assume a linear predictor function and a nonlinear link function $g(\mu)$ for the systematic component (Eq. 6). The systematic component of the GSAM is given by

$$\begin{aligned} g(\mu) &= P_1(A)+P_2(A)+P_2(A) \\ &= \beta_0+\sum_{k=1}^{m}\beta_k(\ln(A))^k+\sum_{k=1}^{n}\beta_{m+k}\left(\frac{1}{A}\right)^k+\sum_{k=1}^{q}\beta_{m+n+k}A^k. \end{aligned} \tag{30}$$

The GSAM has other models as special cases. For instance, the persistence function, P2-full model, is a special case of GSMA if we consider the identity link function $g(\mu)=\mu$, where $\mu=E\{\log(S)\}$, i.e.

$$\mu=\beta_0+\beta_1\ln(A)+\sum_{k=1}^{n}\beta_{k+1}\left(\frac{1}{A}\right)^k. \tag{31}$$

The persistence model, P2-full, can be written as

$$\mu=\beta_0+\sum_{k=1}^{m}\beta_k(\ln(A))^k+\sum_{k=1}^{n}\beta_{m+k}\left(\frac{1}{A}\right)^k, \tag{32}$$

where $\beta_0=\ln(b_0)$, $\beta_1=b_1$ and $\beta_{k+1}=b_{k+1}$, $k=1,\ldots,n$.

Although the model defined by Eq. 30 has a large number of parameters (theoretically, it can have an infinite number of parameters), in practice the fitted models have no more than six parameters. The advantage of such a model is that it enables the formulation of hypothesis testing for the choice of the parameters to be removed from those that are significant for better describing the SAR.

Model fitting. The parameters to be estimated in the GSAM are λ, $\boldsymbol{\beta}$ and ϕ (Eqs. 4 and 6). In order to obtain maximum likelihood estimates for the vector of parameters $\boldsymbol{\beta}$ and dispersion parameter ϕ, we have defined a profiled likelihood function for λ and used the same algorithm proposed in [13]. By assuming the model given by Eq. 5, the log-likelihood function for the vector of the transformed observations $\boldsymbol{S}^{(\lambda)}=(s_1^{(\lambda)},\ldots,s_n^{(\lambda)})^T$ can be written as

$$\begin{aligned} \mathcal{L}(\beta,\phi,\lambda) &= \frac{1}{\phi}\sum_{i=1}^{n}\left\{s_i^{(\lambda)}\theta_i-b(\theta_i)\right\} \\ &\quad+\sum_{i=1}^{n}\left[c(s_i^{(\lambda)},\phi)+\log\{J(\lambda,s_i)\}\right], \end{aligned} \tag{33}$$

where $\theta_i=\int V^{-1}(\mu_i)d\mu_i=q(\mu_i)$ and $J(\lambda,s_i)=|s_i|^{(\lambda-1)}$, $i=1,\ldots,n$ is the Jacobian of the transformation from S to $S^{(\lambda)}$.

The procedure described in [13] is used for making inferences about parameters $(\boldsymbol{\beta},\phi)$ first assuming that λ is fixed and obtains the log-likelihood equations for estimating $\boldsymbol{\beta}^{(\lambda)}$ and $\phi^{(\lambda)}$. The maximum likelihood estimates (MLE) of $\boldsymbol{\beta}$, η, μ and ϕ for a given λ are denoted by $\hat{\boldsymbol{\beta}}^{(\lambda)}$, $\hat{\boldsymbol{\eta}}^{(\lambda)}=\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda)}$, $\hat{\boldsymbol{\mu}}^{(\lambda)}=g^{-1}(\hat{\boldsymbol{\eta}}^{(\lambda)})$ and $\hat{\boldsymbol{\phi}}^{(\lambda)}$, respectively. $\hat{\boldsymbol{\beta}}^{(\lambda)}$ can be calculated, without knowledge on $\phi^{(\lambda)}$, adjusting the GSAM (Eqs. 5–6) to $\boldsymbol{S}^{(\lambda)}$ by iteration.

The iteration starts with an initial set of values $\hat{\boldsymbol{\beta}}^{(\lambda)(k)}$, $k=1$, used to evaluate $\boldsymbol{W}^{(\lambda)(k)}$ and $\boldsymbol{z}^{(\lambda)(k)}$, where $\boldsymbol{W}^{(\lambda)(k)}=\mathrm{diag}\{w_1^{(\lambda)(k)},\ldots,w_n^{(\lambda)(k)}\}$ is a diagonal matrix with weights

$$w_i^{(\lambda)(k)}=V(\mu_i^{(\lambda)(k)})^{-1}(d\mu_i^{(\lambda)(k)}/d\eta_i^{(\lambda)(k)})^2 \tag{34}$$

Table 4. Normal GSAM model fitted by the systematic component shown in Table 2.

	β_0	β_1	β_2	β_3
Coefficients	−10.0405	14.1548	−0.9251	15.3104
(SD)	(0.4593)	(0.1713)	(0.0154)	(0.5575)

doi:10.1371/journal.pone.0105132.t004

Table 5. Richness of *Hymenoptera* species in different sample areas (m^2).

Area	Species	Area	Species	Area	Species	Area	Species	Area	Species
1	55	6	133	12	246	20	274	36	379
2	89	6	157	12	204	20	310	49	409
4	167	6	147	12	200	24	311	61	454
4	116	6	146	16	258	24	311	73	473
4	148	8	226	16	260	30	344	89	521

doi:10.1371/journal.pone.0105132.t005

and $\boldsymbol{z}^{(\lambda)(k)}=(z_1^{(\lambda)(k)},\ldots,z_n^{(\lambda)(k)})^T$ is a working vector whose components are given by

$$z_i^{(\lambda)(k)}=\eta_i^{(\lambda)(k)}+(s_i^{(\lambda)(k)}-\mu_i^{(\lambda)(k)})\left(\frac{d\eta_i^{(\lambda)(k)}}{d\mu_i^{(\lambda)(k)}}\right). \tag{35}$$

The next estimate $\hat{\boldsymbol{\beta}}^{(\lambda)(k+1)}$ can be obtained by

$$\hat{\boldsymbol{\beta}}^{(\lambda)(k+1)}=(\boldsymbol{X}^T\hat{\boldsymbol{W}}^{(\lambda)(k)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}^{(\lambda)(k)}\hat{\boldsymbol{z}}^{(\lambda)(k)}. \tag{36}$$

This new value is used to update $\boldsymbol{W}^{(\lambda)(k+1)}$ and $\boldsymbol{z}^{(\lambda)(k+1)}$ and the procedures are repeated until convergence has been achieved.

Estimating parameter $\phi^{(\lambda)}$ is more difficult than estimating $\boldsymbol{\beta}^{(\lambda)}$. In principle, $\phi^{(\lambda)}$ could also be estimated by maximum likelihood, although there may be practical difficulties associated with this task for some members of Eq. 5. Details about the technique used for finding the MLE $\hat{\phi}^{(\lambda)}$ for a fixed λ can be found in [13].

In order to obtain the MLE $\hat{\lambda}$, we replace MLE $\hat{\boldsymbol{\beta}}^{(\lambda)}$ and $\hat{\phi}^{(\lambda)}$ in (33), which results in the profile log-likelihood function $l_P(\lambda)=l(\hat{\boldsymbol{\beta}}^{(\lambda)},\hat{\phi}^{(\lambda)},\lambda)$. The plot of the profile likelihood function $l_P(\lambda)$ against λ for a sequence of values of λ numerically determines the MLE for λ. Once the MLE for $\hat{\lambda}$ has been obtained, it can be used to produce the unrestricted estimates $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{(\hat{\lambda})}$ and $\hat{\phi}=\hat{\phi}^{(\hat{\lambda})}$.

Assuming that the estimated $\hat{\lambda}$ is known, the confidence intervals for parameters $\boldsymbol{\beta}^{(\lambda)}$ and $\phi^{(\lambda)}$ can be calculated in the usual context of the GLM and using the adjusted values $\hat{\boldsymbol{\beta}}^{(\lambda)}$ and $\hat{\phi}^{(\lambda)}$. We consider the approximate covariance matrix of $\hat{\boldsymbol{\beta}}^{(\hat{\lambda})}$ and the variance of $\hat{\phi}^{(\hat{\lambda})}$ given in [13] to make inferences about these parameters. Here, we have considered the gamma, Gaussian and inverse Gaussian distributions for the probability density function (Eq. 5)

We also performed likelihood ratio (LR) tests [23] using a statistic $w=2\{l_P(\hat{\lambda})-l_P(\lambda^{(0)})\}$, which has an asymptotic χ_1^2 distribution for testing $\lambda=\lambda^{(0)}$ and constructed a large sample confidence interval for λ by inverting the LR test.

Results and Discussion

Simulation of the colonization process of a region

The parameters of SAR curves are determined from the survey data. As a proof of concept we first used simulated data for 80 species placed in a 50×50 cell lattice according to a neutral model [24] without dispersal limitation, as applied in [25]. This lattice was then resampled so as to establish the shape of the SAR (Figure 1).

The adjusted models are variations of the full-scale model with six parameters, given by

$$\begin{aligned} g(\mu)=\beta_0+\beta_1\ln(A)+\beta_2(\ln A)^2 \\ +\beta_3/A+\beta_4/A^2+\beta_5A+\beta_6A^2. \end{aligned} \tag{37}$$

The following canonical link functions were considered: (i) $g(\mu)=\mu$ for the normal GSAM, (ii) $g(\mu)=1/\mu$ for the gamma GSAM and (iii) $g(\mu)=1/\mu^2$ for the inverse Gaussian GSAM.

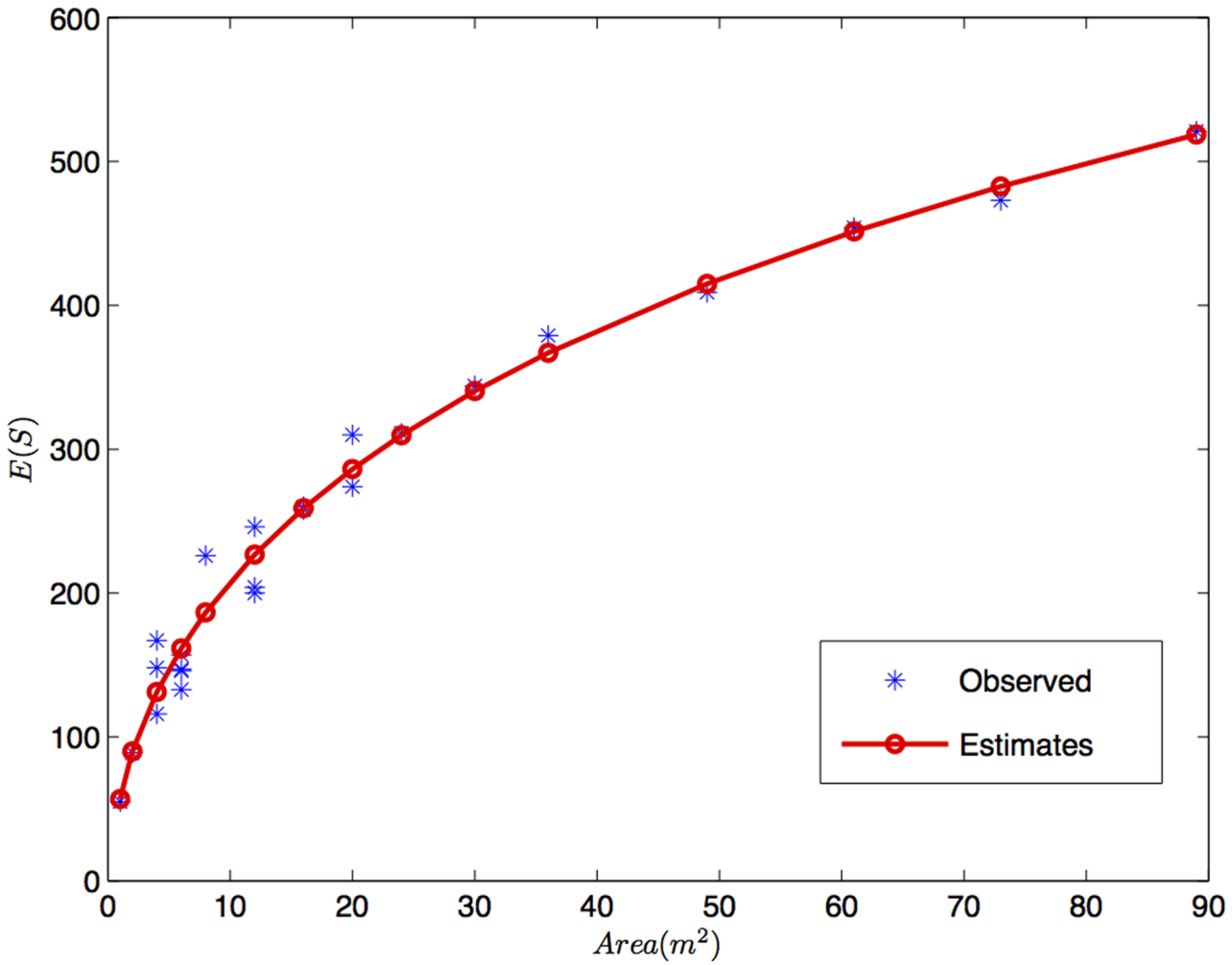

Figure 2. Data and fitted model for the real species-area relationship of *Hymenoptera* in a beech forest on limestone.
doi:10.1371/journal.pone.0105132.g002

Moreover, the traditional models presented in the previous section were considered by assuming that the random variable S is normally distributed. We could estimate the mean of the transformed data $E(S^{(\lambda)})=\mu$, but to predict the expected value of the untransformed dependent variable S, when the GSAM is adjusted to the data, $E(S)$ must be estimated. The dependent variable S can be explained by subtracting μ on both sides of Eq. 4 and solving this equation for S. When $\lambda \neq 0$ we can write

$$S=(\lambda S^{(\lambda)}+1)^{1/\lambda}==(\lambda S^{(\lambda)}-\lambda\mu^{(\lambda)}+1+\lambda\mu^{(\lambda)})^{1/\lambda}=$$
$$=[1+\lambda\mu^{(\lambda)}]^{1/\lambda}\left\{1+\frac{\lambda}{1+\lambda\mu^{(\lambda)}}(S^{(\lambda)}-\mu^{(\lambda)})\right\}^{1/\lambda}. \quad (38)$$

The expected value of the species number S, on the original scale, can be evaluated by a first-order approximation of the binomial expansion (Eq. 38), as given in detail in [13]:

$$E(S)=(1+\lambda\mu)^{1/\lambda}\left\{1+\frac{(1-\lambda)\phi\, V}{2(1+\lambda\mu)^2}\right\}. \quad (39)$$

The best model can be chosen by using the AIC and BIC criteria [26], which are measurements of the relative goodness of fit of a statistical model for a given set of data. The mean square error (MSE) and mean absolute percent error (MAPE) are given, respectively, by

Table 6. Normal GSAM model fitted to SAR of *Hymenoptera*.

	β_0	β_1	β_2
Coefficients	101.9707	20.6625	−46.1669
(SD)	(9.4348)	(0.8211)	(22.3287)

doi:10.1371/journal.pone.0105132.t006

$$MSE = \frac{100\%}{n\hat{\sigma}_s^2}\sum_{i=1}^{n}(s_i - \hat{s}_i)^2 \quad (40)$$

and

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{s_i - \hat{s}_i}{s_i}\right|, \quad (41)$$

where $\hat{\sigma}_s^2$ is the sample variance of S and $\hat{s}_i$ is the estimate of s_i given by Eq. 39.

Table 1 shows some of the traditional models adjusted with normal error. The selected model was the logarithmic quadratic function proposed by [21], with minimum $AIC = 221.09$, $BIC = 244.39$, $MSE = 1.02\%$ and $MAPE = 2.50\%$.

Table 2 shows the selected models adjusted by the GSAM with normal, gamma and inverse Gaussian (I.G.) distributions. Note that the adjusted models are variations of the full-scale model.

Table 3 shows the GSAM models fitted with $\hat{\lambda}$ adjusted by the profile likelihood. The selected normal GSAM has minimum $AIC = 79.90$ and $BIC = 109.02$, which are the lowest values among the adjusted models. MSE and $MAPE$ of the model are also smaller than those of the adjusted model with gamma and inverse Gaussian distributions. The adjusted value of parameter λ is $\hat{\lambda} = 0.822$ with confidence interval $(0.729; 0.922)$. Because λ is different from zero or one, there is a significant difference between the results achieved with this model or by using the traditional models given in Table 1. Therefore, for this data set, the normal GSAM is the model that has best fitted the analyzed data. The MLE estimates of the systematic component and standard-deviation (SD) of the systematic component are shown in Table 4.

Figure 1 shows the systematic observation of SAR on the original scale and the fitted curve with the adjusted GSMA models. $E(S)$ was calculated by Eq. 39 and for the adjusted GSAM we obtained $MSE = 0.08\%$ and $MAPE = 0.61\%$, respectively.

Application to real data

The GSAM model was applied to a data set that consisted of 25 observations of parasitic insects of the *Hymenoptera* order in a beech forest on limestone. *Hymenoptera* is one of the largest orders of insects that comprise sawflies, wasps, bees and ants. The total number of Hymenopteran species in Europe exceeds 20,000. The data considered here contain the summary of a long-term study of the ecology of parasitic Hymenoptera in a German beech forest, i.e. the Göttingen forest, which is approximately 120 years old and has grown over a ground limestone. The climate of the forest is typical of Central Europe and the work area covered approximately four acres. The study was conducted for 8 years (starting in 1980) in 144 square meters of forest soil.

The analysis of the SAR for Hymenoptera is essential, because the insects that belong to this order are the most important environmental agents fundamental for nutrient recycling and control of harmful species. The group is ubiquitous and it is common sense to assume that there is at least one species of parasitic insects for each species of herbivore insects [14]. Many of such species can be considered for the biological control of plague in agriculture. For instance, wasps, from *Symphyta* suborder, are plague conifers in the Northern hemisphere and several species of ants cause losses of millions of dollars for agriculture. Such insects act as special indicators and enable the inference of the diversity of arthropods of a broad spectrum of niches. *Hymenoptera* parasitoids are sensitive to environmental pollution, therefore fluctuations in their population are observed earlier than in their hosts [27]. This sensitivity makes this group an ideal candidate for studies on conservation. Therefore, the knowledge on how the number of species scales with area is fundamental for the prediction of the impact of such insect parasitic on both ecosystems and agriculture.

Table 5 shows the species richness in different sample areas (see also [14]). We modeled the data by taking into account all the models presented in previous sections. The results show that the normal GSAM with $\hat{\lambda} = 1$ was the best fitted model. No transformation of the original data was necessary:

$$E(S) = \beta_0 + \beta_1(\ln A)^2 + \beta_2/A.$$

The parameters of the fitted model are shown in Table 6. The fitted mean together with the data provided in Table 5 are shown in Figure 2. The adjusted model resulted in $AIC = 220$, $BIC = 238$, $MSE = 1.86\%$ and $MAPE = 6.69\%$, therefore, the GSMA model has proved very accurate. Our fit has provided a very good description of the increase observed in species richness and differs from the simple power-function presented in [14]. Interestingly our best fitting model includes features of the modified persistence model [18], but it has not been predicted by any recent macroecological theory, which calls for a fresh look on the patterns and constraints of spatial species distribution.

Conclusions

The generalized species-area model (GSAM) proposed here has provided a generalized model to mathematically describe the SAR. The GSMA can reduce the efforts devoted to finding the best model and can more accurately represent the effect of the area over the diversity of species than the power-curve models commonly used. This fact has been verified in simulated and real-world data.

Acknowledgments

The authors acknowledge Angela C. P. Giampedro, who provided a careful review of the text.

Author Contributions

Conceived and designed the experiments: WU KSC MGA. Performed the experiments: WU KSC MGA. Analyzed the data: KSC CARD FAR MGA. Wrote the paper: KSC WU CARD FAR MGA.

References

1. Rosenzweig ML (1995) Species diversity in space and time. Cambridge, UK.: Cambridge Univ.Press.
2. Tjørve E (2003) Shapes and functions of species-area curves: a review of possible models. J Biogeogr 30: 827–835.
3. Dengler J (2009) Which function describes the species-area relationship best? a review and empirical evaluation. J Biogeogr 36: 728–744.
4. MacArthur RH (1967) The theory of island biogeography, volume 1. Princeton University Press.
5. Würtz P, Annila A (2008) Roots of diversity relations. Journal of Biophysics 2008.
6. Connor EF, McCoy ED (1979) The statistics and biology of the species-area relationship. The American Naturalist 113: 791–833.

7. Scheiner SM (2003) Six types of species-area curves. Global Ecology & Biogeography 12: 441–447.
8. Tjørve E, Tjørve KMC (2008) The species-area relationship, self-similarity, and the true meaning of the z-value. Ecology 89: 3528–3533.
9. Tjørve E (2009) Shapes and functions of species-area curves (ii): a review of new models and parameterizations. J Biogeogr 36: 1435–1445.
10. Guilhaumon F, Mouillot D, Gimenez O (2010) mmsar: an r-package for multimodel species-area relationship inference. Ecography 33: 420–424.
11. Arrhenius O (1921) Species and area. J Ecol 9: 95–99.
12. Kallimanis AS, Mazaris AD, Tzanopoulos J, Halley JM, Pantis JD, et al. (2008) How does habitat diversity affect the species-area relationship? Global Ecol Biogeogr 17: 532–538.
13. Cordeiro GM, Andrade MG (2009) Transformed generalized linear models. J Statist Plann Inference 139: 2970–2987.
14. Ulrich W (2001) Hymenopteren in einem Kalkbuchenwald: Eine Modellgruppe zur Untersuchung von Tiergemeinschaften und ökologischen Raum-Zeit-Mustern. Schriftenr. Forschzentr. Waldökosysteme. Göttingen A 171: 249 S.
15. Dobson AJ (2010) An introduction to generalized linear models. CRC press.
16. Gleason HA (1922) On the relation between species and area. Ecology 3: 158–162.
17. Plotkin JB, Potts MD, Leslie N, Manokaran N, LaFrankie J, et al. (2000) Species-area curves, spatial aggregation, and habitat specialization in tropical forests. J Theor Biol 207: 81–99.
18. Ulrich W, Buszko J (2003) Self-similarity and the species-area relation of polish butterflies. Basic Appl Ecol 4: 263–270.
19. Ulrich W, Buszko J (2004a.) Habitat reduction and patterns of species loss. Basic Appl Ecol 5: 231–240.
20. Box GEP, Cox DR (1964) An analysis of transformation. J R Statist Soc B 26: 211–252.
21. Chiarucci A, Viciani D, Winter C, Diekmann M (2006) Effects of productivity on species-area curves in herbaceous vegetation: evidence from experimental and observational data. Oikos 115: 475–483.
22. Gitay H, Roxburgh SH, Wilson JB (1991) Species-area relations in a new zealand tussock grassland, with implications for nature reserve design and for community structure. Journal of Vegetation Science 2: 113–118.
23. Mood AM (1950) Introduction to the theory of statistics.
24. Hubbell SP (2001) The Unified Neutral Theory of Biodiversity and Biogeography. Princeton, NJ.: Princeton University Press.
25. Zillio T, Condit R (2007) The impact of neutrality, niche differentiation and species input on diversity and abundance distributions. Oikos 116: 931–940.
26. Burnham KP, Anderson DR (1998) Model selection and inference: a practical information-theoretic approach/.
27. LaSalle J, Gauld I (1991) Parasitic hymenoptera and the biodiversity crisis. Redia 74: 315–334.