SPECIAL PAPER

Shapes and functions of species–area curves (II): a review of new models and parameterizations

Even Tjørve*

Lillehammer University College, Lillehammer, Norway

ABSTRACT

This paper has extended and updated my earlier list and analysis of candidate models used in theoretical modelling and empirical examination of species–area relationships (SARs). I have also reviewed trivariate models that can be applied to include a second independent variable (in addition to area) and discussed extensively the justifications for fitting curves to SARs and the choice of model. There is also a summary of the characteristics of several new candidate models, especially extended power models, logarithmic models and parameterizations of the negative-exponential family and the logistic family. I have, moreover, examined the characteristics and shapes of trivariate linear, logarithmic and power models, including combination variables and interaction terms. The choice of models according to best fit may conflict with problems of non-normality or heteroscedasticity. The need to compare parameter estimates between data sets should also affect model choice. With few data points and large scatter, models with few parameters are often preferable. With narrow-scale windows, even inflexible models such as the power model and the logarithmic model may produce good fits, whereas with wider-scale windows where inflexible models do not fit well, more flexible models such as the second persistence (P2) model and the cumulative Weibull distribution may be preferable. When extrapolations and expected shapes are important, one should consider models with expected shapes, e.g. the power model for sample area curves and the P2 model for isolate curves. The choice of trivariate models poses special challenges, which one can more effectively evaluate by inspecting graphical plots.

Keywords

Extended power models, interaction term, Kobayashi logarithmic, logistic family, model choice, negative-exponential family, P2 model, persistence functions, species–area curves, trivariate models.

*Correspondence: Even Tjørve,
Lillehammer University College,
PO Box 952, 2604 Lillehammer, Norway.
E-mail: even.tjorve@hil.no

INTRODUCTION

The species–area relationship (SAR) is one of the most fundamental patterns in macroecology and an important tool in assessing species diversity, extinction rates and species hotspots. SARs are at the heart of understanding the distribution of diversity in space because they address the manner in which the total number of species (S) increases with area (A). The observation that S almost always increases with A (i.e. grows monotonically) has been dubbed one of the few true laws in ecology (Lawton, 1999).

Preston (1962) had early on distinguished between curves compiled from 'isolates' and those compiled from 'samples', and both types are commonly encountered in the literature, though this distinction is seldom made explicit. Isolates are areas where some less hospitable surrounding matrix reduces biological interactions with other isolates (of the same type) and where biotic and abiotic factors define their boundaries and sizes. Examples of isolates include forest remnants, ponds, mountaintops and islands. Sample areas, on the other hand, are survey areas where sampling design defines boundaries and sizes. Examples of samples include survey units (e.g. quadrats and latitude–longitude blocks) and political units (e.g. counties and states). This distinction is important because processes producing SARs should influence the two in different ways. Sample area curves and isolate curves should, therefore, have different shapes and consequently require different interpretations. Sample area curves are expected to be constantly

decelerating in arithmetic space, whereas isolate curves may become sigmoid and include an inflection point (Tjørve, 2003). Also, sampling design affects curve shape (Hill *et al.*, 1994; Keeley & Fotheringham, 2005; Ulrich & Buszko, 2007).

The purpose of this paper is to review developments since my earlier paper on shapes and functions of species–area curves (Tjørve, 2003) and to discuss the criteria for applying different models. In this short time, several new models have been proposed or applied to species–area data sets, most of which have been extensions (i.e. inclusion of new parameters) or parameterizations of earlier models. In particular, one group of candidate models that has been proposed recently is the group of modified or extended power models (EPMs).

Accordingly, the first aim of this paper is to review these extensions and parameterizations and to discuss their usefulness. Whatever its theoretical and practical limitations may be, the power function remains the most frequently preferred model for fitting curves to species–area data. Yet, the realization of the shortcomings of the power model in describing the growth curve of SARs has inspired a search for extensions and alternative candidate models (Tjørve, 2003). This search has not been unproblematic. This paper thus brings attention to the effects of over-parameterization and the use of location parameters. Also, the use of the same name for alternative parameterizations in the literature further confuses matters. The second aim of the paper is to review and assess models with two predictor variables, i.e. one predictor variable in addition to area. The extra variable is included in the regression model either as a weighting of area or as an additional predictor variable for a multiple regression. By reviewing and assessing models, it should be possible both to identify models that are more likely to be useful candidates and to discuss when different models may be better alternatives than others.

FAMILIES AND PARAMETERIZATIONS

The parameters of a model help to define its properties and behaviour. In a linear regression the parameters are slope and intercept. The power function and the logarithmic function (see Tjørve, 2003, for details) are examples of simple two-parameter models. Some parameterizations alter the shape of the curve. Others only shift the curve horizontally or vertically. The latter is called a location parameter. A model and its parameterizations are often described as a family of models.

The addition of parameters, or 'parameterization', increases the flexibility of the model, improving the fit of the model (i.e. producing higher R^2-values). Here I use the term 'parameterization' to denote the extension of the model with one or several new parameters (Ratkowsky, 1990), even though elsewhere this term is often used to denote the process of defining or deciding the parameters (i.e. their values) in the model. A totally different model may still produce a better fit than the addition of one or more parameters to an initially favoured model, and estimates obtained may also be severely biased and non-normally distributed. The inclusion of too

many parameters, which is often called *overparameterization*, can also result in overfitting to what are just chance variations in data points. It may even cause non-convergence when using an iterative procedure to estimate parameters by the ordinary least squares (OLS) method in nonlinear regression models, and attempts to fit the overly complex model therefore fails, though this is probably more often caused by the optimization algorithm than the model itself. One should, therefore, choose structurally simpler and more parsimonious models out of concern for overparameterization. Regardless of the problem of failure to converge, the application of an F-test can assess the usefulness of extra parameters (Motulsky & Christopoulos, 2004).

Extended power models

Today, most ecologists still model SARs by fitting a linear model in log–log space, i.e. applying the power model as, $\log f(x) = \log a + b \log x$. Unfortunately, as species–area curves are usually assumed to be power functions, studies almost never evaluate the shape of the curve by plotting the curve (Scheiner, 2004). In reality, the power model (or indeed any two-parameter model) is very inflexible (Ratkowsky, 1990). In addition, the power model assumes that the same (scale-invariant) proportion of new species is added every time the area is doubled (Harte *et al.*, 1999; Tjørve & Tjørve, 2008). The inflexibility is one reason why the power model consistently produces poor fits to many empirical data sets. A reasonable attempt to find a better fitting model is to add parameters to the power model. Such parameterizations can be made to correct either parameter a or b (or both) so that either (or both) becomes a function of x, denoted here by $a^\star$ and $b^\star$, respectively. These EPMs typically change shape with different ranges of parameter values.

Let $y = ax^{b^\star}$, and consider then $b^\star(x) = bx^{-c}$ or $b^\star(x) = b-(c/x)$ (see EPM 1 and 2 in Table 1), which are both parameterizations of the Freundlich model (Sibbesen, 1981). I have not yet found these types of 'modified power models', where b is a function of x, fitted to species–area data sets, but they have been applied to other types of biological data (e.g. energy expenditure as a function of age or body mass; Tjørve *et al.*, 2007, 2008b). They were fitted because their data plots were not linear in log–log space. Unfortunately, these Freundlich-type curves have, or may have, maxima or minima, which are sometimes included when fitting the curve to data. Such models may become useless when one aims to extrapolate species richness for scales outside the scale window of the sampled area.

Some studies have fitted species–area data to the second type of EPM, where $a^\star$ is a function of x, $(a^\star(x))$ (Plotkin *et al.*, 2000; Ulrich & Buszko, 2003, 2004). These models are referred to in the literature as *persistence functions*. Plotkin *et al.* (2000) were the first to fit a persistence function, $ax^b \exp(-cx)$, to a SAR, applying the model to data sets from tropical forests. I shall here denote this first persistence model as the P1 model. It has a maximum (for positive values of c). Plotkin *et al.*

Table 1 Features of the power model, candidate extended power models (EPMs) and persistence functions, plus logarithmic models. The EPMs and persistence models are parameterizations of the power model and are more flexible than the power model because of the extra parameter. The new terms added can be seen as extensions to either parameter a or parameter b of the power model. The Kobayashi logarithmic, unlike the logarithmic model, goes through the origin of axes and is therefore a useful alternative to the latter.

Model name	Model	Extension to:	Shape characteristics	Asymptote	Source
Power model	ax^b	–	Upward convex, linear in log–log space	No asymptote	Arrhenius (1921)
Extended power model 1 (EPM1)	$ax^{bx^{-c}}$	b-value	May have inflection point and maximum	a for some values c	Sibbesen (1981)
Extended power model 2 (EPM2)	$ax^{b-(c/x)}$	b-value	Inflection point and minimum	ax^b as x becomes large	Tjørve et al. (2007, 2008b)
Persistence function 1 (P1)	$ax^b \exp(-cx)$	a-value	Inflection point and maximum	ax^b as x becomes small + zero as x large	Plotkin et al. (2000)
Persistence function 2 (P2)	$ax^b \exp(-c/x)$	a-value	Inflection point, convex in log–log space	ax^b as x becomes large	Ulrich & Buszko (2003, 2004)
Logarithmic model	$a + b \log x$	–	Upward convex, linear in log-linear space	No asymptote	Gleason (1922, 1925)
Kobayashi logarithmic	$a \log(1 + x/b)$	–	Upward convex, not linear in log-linear space	No asymptote	Kobayashi (1975)

(2000) introduced this model in order to predict (extrapolate) large-scale diversity from small-scale sample area data because their SARs showed consistent deviations from the power law at all spatial scales. They found that the proportion of species that persisted with bisections (i.e. remained when an area was halved) was not scale constant (i.e. not the result of a power relationship) but increased with scale. They did not find the model parameters of their persistence model by SAR regression but on the basis of a persistence relationship; that is, area vs. the proportion of species remaining with the halved area. Ulrich & Buszko (2003, 2004) modified the P1 function and suggested a second persistence function (P2), also in response to data that did not follow the allometric relationship (power relationship). They did not explicitly discern between the two models and referred to the P2 model as the Plotkin (P1) model, which implies that they did not recognize the difference between the P1 and P2 models at the time. They applied (by regression) the P2 model, $ax^b \exp(-c/x)$, to data on Polish butterflies. For both, the additional term in this model, compared with the power model, can be seen as an extension of a, by $a^{\star}(x) = a \exp(-cx)$ or $a^{\star}(x) = a \exp(-c/x)$. However, these two models behave very differently. The P1 model asymptotically approaches ax^b as x becomes small, whereas the P2 asymptotically approaches ax^b as x becomes large.

Parameterizations of the logarithmic model

Like other two-parameter models, the logarithmic model, $a + b \log x$, is quite inflexible. It assumes the same (scale-invariant) number of new species added every time the area is doubled. Another inherent problem of the logarithmic model is that it has no value for $x = 0$. One remedy for this is to modify the model by replacing $\log x$ with $\log(1 + x)$ (e.g. Tjørve, 2002), which gives $a + b \log(1 + x)$. This model acquires a convex-downward shape (rather than a straight line) in log-linear space but does not go through the origin of axes.

Based on the log-series species abundance distributions (SADs), Kobayashi (1975) proposed an alternative version of the logarithmic model, $a \log(1 + x/b)$ (Table 1) as a model to describe SARs generated from continuous sampling where the species abundances follow a logarithmic-series distribution. Given that spatial distribution of individuals and species abundances affects curve shape (He & Legendre, 2002; Tjørve et al., 2008a, however, even SARs from log-series SADs may not fit logarithmic family models. As for the ordinary logarithmic model, this also results from the same number of species being added when the area is doubled. Even though this model is now rarely applied, it is the more realistic one in that it goes through the origin of axes.

Parameterizations of the negative-exponential family

One can regard the family of 'power models' as parameterizations of the power model and the family of 'logarithmic models' as parameterizations of the basic logarithmic model. Many of the other models that have been fitted to SARs are also related to each other in this way, meaning that the more complex models are parameterizations of simpler models in the same family. We may, for example, regard both the Chapman–Richards model and the cumulative Weibull distribution as parameterizations of the cumulative negative exponential distribution (Table 2) (see Tjørve, 2003, for details). (The Weibull distribution and the negative-exponential distribution discussed here are the cumulative versions of these distributions and, therefore, increase monotonically with inflection points where their non-cumulative counterparts have maxima.)

All three models of the negative-exponential family have an upper asymptote given by parameter a, but only the negative-exponential model and the Weibull distribution go through the origin of axes. Both sigmoid models have floating inflection points determined by the parameter values (contrary

Table 2 Parameterizations of the 'negative-exponential family'. We may view the Chapman–Richards model and the cumulative Weibull distributions as parameterizations of the negative exponential model. They both have an extra parameter and are, in contrast to the negative exponential, both sigmoid in arithmetic space.

Model name	Model	Parameters	Shape characteristics	Asymptote
Negative exponential	$a[1 - \exp(-bx)]$	2	Convex	Parameter a
Chapman–Richards	$a[-\exp(-bx)]^c$	3	Sigmoid with floating inflection point	Parameter a
Cumulative Weibull distribution	$a[1 - \exp(-bx^c)]$	3	Sigmoid with floating inflection point	Parameter a

to the logistic family, the Gompertz model, and the EVF model; see below). Parameter b or c, or both, define their position relative to the asymptote (parameter a). In the Weibull distribution, parameters b and c define the inflection (relative to the asymptote). The inflection property of the Chapman–Richards model is special in that the inflection point falls at $a[1 - (1/d)]^d$. One can thus fix the inflection point relative to the upper asymptote (for example, when $d = 2$ at 25% of a).

Parameterizations of the logistic family

The most commonly applied logistic model is a three-parameter model, $a/[1 + \exp(-bx + c)]$ (Ratkowsky, 1990; Tjørve, 2003), where parameter a sets the upper asymptote. Archibald's model (Archibald, 1949), $a/(b + c^x)$, which was developed from theoretical arguments, is an alternative three-parameter model (see Table 3) for values of parameter $c < 1$. It has a/b (parameter a over parameter b) as an upper asymptote and an inflection point at $x = 0$ that makes it unrealistic at small scales and when an inflection point is expected in an isolate curve. Archibald (1949) argued that this model fitted his data plots (sample area curves derived from plant communities), claiming that they were sigmoid in log-linear space. Given the above assumptions, the model stays sigmoid in log-linear space but with an inflection point at some positive value of x.

The common logistic model has an inflection point fixed at 50% of the upper asymptote (so that each half of the curve becomes a mirror image of the other) in arithmetic (and log-linear) space. This feature is distinct from the Gompertz model and the extreme-value function (EVF) (below) that also have fixed inflection points but at 36.8% and 63.2% of their upper asymptotes, respectively.

In a reparameterization (of the common logistic model), $a/\{1 + \exp[-b(x - c)]\}$ (Farebrother, 1998), the parameter b changes the slope of the curve at the inflection point, that is, the maximal steepness of the slope, without changing the position of the inflection point or the asymptotes (see also below and Table 4). This parameter is also known as the *HillSlope* (or *growth*) *coefficient* and the model accordingly as a *HillSlope equation* or simply *Hill equation* (Motulsky & Christopoulos, 2004). With this reparameterization, parameter c becomes a location parameter that shifts the whole curve horizontally without changing its shape.

The location parameter

We can usually recognize the location parameter as a constant added to the x-value (see Tjørve, 2003). Some sigmoid models, like the logistic (see above), the Gompertz and the EVF, can be reparameterized to make one parameter act as a location parameter (i.e. without adding an extra parameter). In the reparameterized versions, in Table 4, parameter c now acts as a location parameter (without stretching or compressing the curve horizontally).

A location parameter could be useful for fitting regression lines to isolate data if one anticipates a minimum-area requirement for the group of species studied, i.e. that no species (e.g. in a given taxon, guild or pre-defined group) are expected to be present on islands smaller than a given size.

TRIVARIATE MODELS

The list of possible variables that may affect the number of species, in addition to area, include time, energy, climate, habitat diversity, isolation and elevation. We can measure isolation as the distance between areas or as an isolation index (Hanski, 1994; Hanski & Thomas, 1994; Hanski & Kuussaari, 1995), which also takes into account the size of the closest area or island. Available energy or productivity, as measured by, say, climate (temperature and precipitation), photosynthesis or

Table 3 Parameterizations of the 'logistic family'. The Archibald logistic can be seen as a reparameterization of the common logistic model. Both models are sigmoid, but the Archibald model, owing to its inflection point falling at the origin of the axes, represents an alternative to convex models rather than other sigmoid models.

Model name	Model	Parameters	Shape characteristics	Asymptote
Common logistic	$a/[1 + \exp(-bx + c)]$	3	Sigmoid with fixed inflection point	Parameter a
Archibald logistic	$a/(b + c^x)$	3	Sigmoid but only convex (decelerating) part used	Parameter a/b

Table 4 Illustration of how one can reparameterize the logistic model, the Gompertz model and the extreme-value function (EVF). These models share a special characteristic in that they can be written in a form (reparameterized) so that one parameter acts as a location parameter. In other words, one can move the curve horizontally without altering its shape. The table shows reparameterizations of these models, which make parameter c into a location parameter.

Model name	Reparameterized model
Logistic with location parameter	$a/\{1 + \exp[-b(x - c)]\}$
Gompertz with location parameter	$a \exp\{-\exp[-b(x - c)]\}$
EVF with location parameter	$a(1 - \exp\{-\exp[b(x + c)]\})$

normalized difference vegetation index (NDVI) is probably the factor most regularly considered as a second predictor (see Wright, 1983; Dodson, 1992; Storch *et al.*, 2005; Kalmar & Currie, 2006; Evans *et al.*, 2007).

A simple linear trivariate regression model is expressed as

$$f(x, y) = a + bx + cy. \tag{1}$$

The weighting of the dependent or the independent (predictor) variables changes the model and, therefore, its shape. The most common are log-transformations. If both independent variables, x and y in equation 1, are log-transformed, then a logarithmic model results:

$$f(x, y) = a + b \log x + c \log y. \tag{2}$$

If species number (the dependent variable) is also log-transformed, we have a power-law model:

$$f(x, y) = ax^b y^c \tag{3}$$

(see Table 5). Adler & Lauenroth (2003) and Adler *et al.* (2005) have assessed the latter as a possible model for species–time–area relationship (STAR). One often employs a log-transformation to improve the fit. On the other hand, the dependent variable y, is also often transformed to remedy non-normality or heteroscedasticity, or both. Such weightings of y other than log-transformations include power transformations (*weighting by 'model'$^{'exponent'}$*) (e.g. Kalmar & Currie, 2007), relative weighting (*weighting by 1/'model'^{2}*), Poisson weighting (*weighting by 1/'model'*), and weighting by observed variability (see, e.g. Motulsky & Christopoulos, 2004, for a review of weighting methods). Unfortunately, such weightings are often performed without considering whether the resulting shapes in arithmetic space are realistic.

Multiple regressions assessing the effect of two or more independent variables (transformed or not) on species numbers are also very common (e.g. Dodson, 1992; Adler & Lauenroth, 2003; Triantis *et al.*, 2003; Adler *et al.*, 2005; Storch *et al.*, 2005; Kalmar & Currie, 2006, 2007; Ulrich, 2006; Evans *et al.*, 2007).

The above models with two independent (predictor) variables are all linear or can be linearized and fitted by a simple or a multiple linear regression. The model with a power term plus logarithmic term,

$$f(x, y) = ax^b + c \log y, \tag{4}$$

proposed by Ulrich (2006) for STARs, relies, on the other hand, on nonlinear regression. With all the new software available, one might have expected nonlinear multivariate regression to have become more common than it actually has.

Combination variables

One can also include a second predictor variable in the model by simply combining the two predictors into a new one. Triantis *et al.* (2003) have presented such a model for the species–area–habitat relationship (which they termed the Choros model), in which the predictor variable is substituted with a new variable, here termed a combination variable (CV), which is the product of two variables (see Table 5), given as $a(xy)^b$. Here, the new variable is included without adding a new parameter (constant) to the model. In reality, the second predictor, y, is here simply used as a corrective term or, in other words, to weight the first predictor (or vice versa). The model is fitted as a straight line in log–log space. Adler & Lauenroth (2003) and Adler *et al.* (2005) have also tried this model, with a multiplicative combination term, as a candidate STAR model. One may also consider models with additive combination terms, such as $a + b(x + y)$ and $a + b(\log x + \log y)$ or other. The term produced by the combination of two variables may also be seen as an interaction term (see below). Note that such additions of new parameters may also cause overparameterization in trivariate models.

Table 5 Illustration of models with interaction terms, here called INT models, and models with no interaction terms, here called no-INT models. Models where two independent (predictor) variables are combined into one are called CV models. The table shows examples of possible simple trivariate (no-INT) models with two independent (predictor) variables and how they can be reduced or extended to CV models and to INT models, respectively. In models with (additive or multiplicative) combination variables (CVs), the number of parameters is reduced from three to two, whereas in the models with interaction terms (INTs) the number of parameters is increased to four.

Basic model	No-INT model	CV model	INT model
Simple linear regression	$a + bx + cy$	$a + b(x + y)$	$a + bx + cy + d(xy)$
Logarithmic model	$a + b \log x + c \log y$	$a + b(\log x + \log y)$	$a + b \log(x) + c \log y + d(\log x)(\log y)$
Power law model	$ax^b y^c$	$a(xy)^b$	$ax^{[b + d(\log y)/2]} y^{[c + d(\log x)/2]}$

Adding an interaction term

In a linear regression model with two predictor variables, $a + bx + cy$, the effect of one predictor is the same for all values of the other. If there is instead a linear relationship between the two predictors, then $b = b^* + dy$, so that $a + (b^* + dy)x + cy$, which we can write as

$$f(x,y) = a + b^*x + cy + d(xy), \tag{5}$$

where $d(xy)$ is the interaction term. This type of interaction-term model (INT model) is useful if we expect the effect of one variable in the model to vary with the second variable; for example, if we expect the number of species to increase faster with area when species are sampled over a longer period rather than a shorter period. Log transformations of x and y in model (5) give:

$$f(x,y) = a + b \log x + c \log y + d(\log x)(\log y) \tag{6}$$

and power-law models give:

$$f(x,y) = ax^{[b+d(\log y)/2]}y^{[c+d(\log x)/2]} \tag{7}$$

Simple trivariate models may not turn out to be appropriate because of the way the effect of one variable depends on the second. In models (5), (6) and (7), multiplicative interaction terms as $d(xy)$ and $d(\log x)(\log y)$ (where d represents a fourth parameter) investigates this effect. Model (7) is the same as the 'full' species–time–area model used by Adler & Lauenroth (2003) and Adler *et al.* (2005). Storch *et al.* (2005) (using NDVI as a measure of energy) and others have applied the same model to species–area–energy relationships. Figure 1a–c illustrates how the three models behave in space with positive interaction parameters.

It may be argued that a linear unweighted trivariate model with a (linear times linear) interaction term (5) constitutes an unrealistic model when we want to describe, for example, species–time–area relationships, species–area–energy relationships or species–area–habitat relationships. If the parameter for the interaction term, d, is positive, then model (5) becomes unrealistic. Its slope will increase along with an increase in the independent variables, as illustrated in the three-dimensional plot in Fig. 1a. Species numbers (as found in empirical data) cannot, however, be expected to grow (with increasing predictor variables) at an ever increasing rate (producing monotonically accelerating curvatures). If the interaction parameter is negative, then as the two independent variables, x and y, increase, the model will reach a maximum and thereafter decline. Thus, both positive and negative interaction parameters may cause unrealistic shapes, especially for large values of the independent variables. This is only a serious problem when extrapolating the model beyond the range of observed data, otherwise even such interaction terms might be useful.

When the parameter for the interaction term is negative, the two other models, (6) and (7), also describe unrealistic results when x and y become large, as $f(x,y)$ produces a maximum, decreases to become negative (which was also reported by Adler & Lauenroth, 2003, and is visible in Fig. 3 in Storch *et al.*, 2005). Therefore, if extrapolations are attempted with a negative
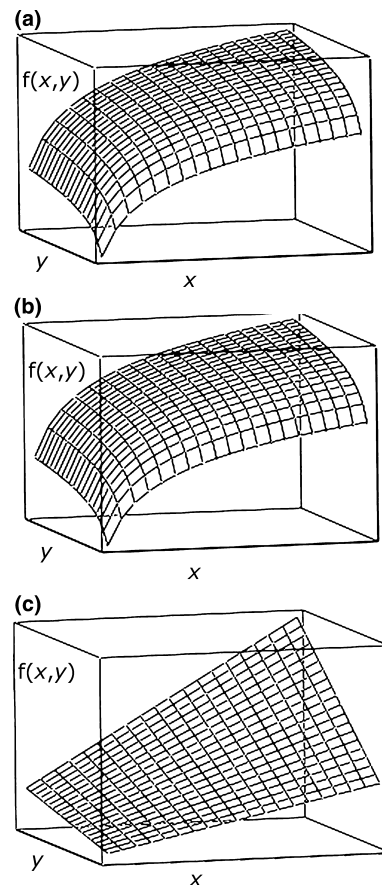

Figure 1 Illustration in 3D of how plots of the three models in Table 5, with two predictor variables and an interaction term (INT models), behave in arithmetic space with positive interaction parameters (d). The first model (a) is an extended version of the simple trivariate (no-INT) model with untransformed variables, where the interaction term $d(xy)$ is added to give $y = a + bx + cy + d(xy)$. The second model (b) is the logarithmic version of an INT model, where both predictor variables are log-transformed, so that $f(x, y) = a + b \log x + c \log y + d(\log x)(\log y)$. The third model (c) is the power version of the INT model, where the dependent variable is also log-transformed, $\log f(x, y) = a + b \log x + c \log y + d(\log x)(\log y)$. In arithmetic space, this is represented by $f(x, y) = ax^{[b + d(\log y)/2]}y^{[c + d(\log x)/2]}$.

parameter d, maxima and 'dips' will occur. Still, we may expect some predictor variables to produce accelerating curvatures over at least part of their total ranges. Therefore, if fitted surfaces are only discussed inside the observed ranges then we may also consider models or parameter values that have the ability to produce unrealistic shapes (or results). Thus, both Adler *et al.* (2005) and Storch *et al.* (2005) have reported negative interaction terms (when applying model (7)).

DISCUSSION OF DIFFERENT MODELS

Curves are fitted to species–area plots for *descriptive*, *explicative* and *predictive* purposes. The identification of curve shape and the comparison of areas and studies constitute descriptive

purposes. Explicative purposes refer to the understanding of how observed or expected patterns in nature affect curve shape and the interpretation of parameters and parameter values. Predictive purposes include extrapolation of total species numbers, prediction of extinction caused by fragmentation and habitat loss, and the identification of species hotspots. We may usually consider those areas displaced above the fitted curve as hotspots.

One can choose a model from these three purposes based on *statistical arguments* and *non-statistical arguments*. The main statistical arguments refer to the compliance with statistical assumptions of the regression method (specifically the assumptions of normality and homoscedasticity) and 'fit' or rather explanatory power (e.g. R^2, residual sum of squares and standard deviation). We can compare candidate models with Akaike's information criterion (AIC) (Akaike, 1974) or with an *F*-test (the latter when models are parameterizations of each other, i.e. nested) (see, e.g. Motulsky & Christopoulos, 2004). The two main non-statistical arguments for choice of model are compliance with expected and observed curve shapes and the purpose of comparing parameter values with those found in the literature. If the latter is an important aim, one may be forced to fit the same model, but it may also be useful to fit an alternative model.

The arguments below demonstrate that there is a plurality of different models from which to choose. In other words, there may neither be one best model for a given data set nor, a fortiori, a best model for all data sets. Absolute rules on applying, evaluating and choosing a model are difficult to establish because there are too many factors and purposes involved in the fitting of a curve. The choice of model will, therefore, depend ultimately on the specific purposes of the exercise. In some cases, both minima and maxima can be real if the SAR is taken as a simple fit of observed data. Nevertheless, because species–area curves are expected to increase monotonically (see above), we may deem models with minima or maxima (displayed within the scale window considered) to be unrealistic.

Given that we should consider expected curve shapes when choosing a model, another important discussion is whether we can expect species–area curves to have upper asymptotes, i.e. a levelling out towards a finite number of species as the area becomes large (e.g. Williamson *et al.*, 2001, 2002; Lomolino, 2002).

The 20 bivariate models reviewed in this paper and in my earlier paper (Tjørve, 2003) are all listed in Appendix 1. One may use this '20-questions game' to help decide which model to apply. I have chosen to discuss mainly the power model and the logarithmic model, the most common bivariate models, and a few others that I regard as the most useful, the P2 model and the negative-exponential family. In addition, I shall discuss trivariate models briefly.

The power, the logarithmic and the P2 models

There are several good reasons to choose the power model. First, it is easy to compare parameter values with data sets from other authors. Secondly, it is easily linearized and consequently easy to fit, though this only holds true when there are no observations with zero species (as one cannot linearize the logarithmic model with zeros in the species count, one would have to use nonlinear optimization). Thirdly, log transformation of the *y*-axis may rectify heteroscedasticity and non-normality of the data set.

Because the power model is linear in log–log space and the logarithmic model assumes a convex-upward shape, they complement each other nicely. The Kobayashi (1975) logarithmic that goes through the origin of axes may be preferable to the commonly applied logarithmic model. It is also tempting to have a second look at the EPMs, especially the P2 model, which can be made to have a depressed curve compared with the power model at small scales, resulting in a sigmoid curve in arithmetic space and a decelerating curve in log–log space. Ulrich & Buszko (2007) presume that the model may not be the ideal candidate for all SARs but expect it to be useful at intermediate scales. Figure 2a,b shows the shapes of P2 for values of $c > 0$ in arithmetic and in log–log space.
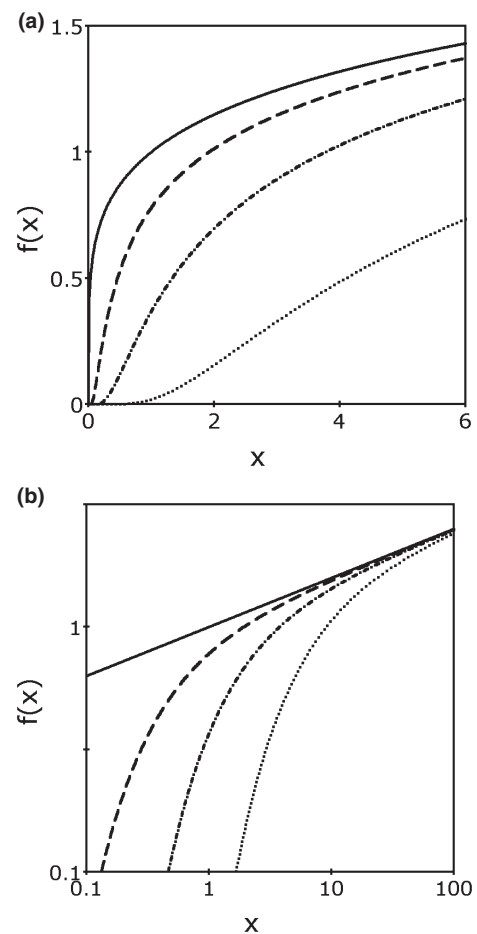

Figure 2 Comparison of curves of the second persistence model (P2), $ax^b \exp(-c/x)$, with the curve of the power model, ax^b (which equals the P2 model when $c = 0$) (unbroken line). The P2 model is calculated for $a = 1$, $b = 0.2$ and $c = 0.25$ (dashed), $c = 1$ (dash-dotted) and $c = 4$ (dotted). The figure shows how the models behave in (a) arithmetic space and (b) log–log space, respectively.

Some of the models reviewed, noticeably the EPM1 and P1 models (Table 1), and the trivariate models with interaction terms (INTs) (Table 4) have, or can have, unrealistic shapes (in which the number of species decreases with an increase in independent variables such as area and time). This distortion occurs because these models sometimes produce maxima or minima that may be included within the scale window studied or will be included if extrapolations are attempted. With a maximum, larger values of the independent variable may cause the model to approach zero or even to become negative. Models with minima may increase with decreasing independent variable. Still, these models can be useful if care is taken and if they are applied only within scale windows defined so that apparent maxima and minima are excluded.

Bearing in mind that one should fit models that have the expected shape of the relationship studied, one should consider the P2 function as an alternative when fitting curves to isolate (island) data sets. Isolates curves are expected to be sigmoid in arithmetic space or to intersect the x-axis above zero, though a possible inflection point may lie outside the scale window sampled. Extended power models may, therefore, also produce a good fit, whereas the power model and the logarithmic model do not, as when the observed curvature falls between these two models.

Other sigmoid models such as the logistic, the Gompertz and the cumulative Weibull distribution all have an upper asymptote. The P2 model may, therefore, prove useful, especially when fitting curves to isolate or island data sets, where one should expect inflection points or depression of the curve at small scales.

Asymptotic models

The cumulative negative exponential distribution, the cumulative Weibull distribution and the Chapman–Richard model are all useful both because they are flexible and because a location parameter can easily be added (Tjørve, 2003). Yet, they all have asymptotes that make them less suitable when extrapolating above the scale window sampled. The negative exponential model is convex-upward (constantly decelerating) whereas the Weibull distribution and the Chapman–Richards model are sigmoid. These are also more flexible than many other sigmoid models, because the inflection point is not fixed proportionally to an upper asymptote as is the case for the logistic model, the Gompertz model and the EVF. The Archibald logistic is decelerating for all positive values of x, and, therefore, is an alternative to the negative exponential model.

Most of the candidate models reviewed here and in Tjørve (2003) have been applied to empirical species–area data sets, though only a handful of studies have applied and compared more than two or three different models. Flather (1996), who applied six different models to species-accumulation curves based on bird occurrences from roadside stops, reported that the Weibull distribution and the beta-P distribution produce the best fit. Fattorini & Fowles (2005) applied 10 different

models to SARs of beetle species on Greek islands and have reported that the power model and the Weibull distribution produced the best fits. Fattorini (2007), who applied five models plus a linearized power model to the distribution of butterflies in Europe, reports the linearized power function and the Archibald logistic (which he calls the 'He–Legendre model') produces the best fit. Stiles & Scheiner (2007) have applied 15 models to SARs of Sonoran Desert plants in remnant habitat patches. They have reported that the logistic model produced the best fit for sample area curves constructed from sample areas within a habitat remnant, whereas the power model produced the best fit for isolate curves made from species numbers of complete patches of different sizes. They have also noted that the cumulative Weibull distribution and the cumulative beta-P distribution did not converge, which may be a result of overparameterization (or be caused by the algorithm applied).

Trivariate models

The choice of trivariate models requires different considerations. The combination of two predictor variables into one is tempting, since bivariate rather than multivariate regression can be applied. Such models are, on the other hand, difficult to justify owing to, for example, the combination of predictors simply by multiplication.

An interaction term is only useful when the effect of one independent variable on the model is affected by a second value. Such cases occur, for example, when time affects species number more at small scales than at large scales or when the effect of a higher number of habitats is stronger at small scales than at large ones.

To assess the usefulness of an interaction term, both a trivariate model with an interaction term and a trivariate model without an interaction term may be fitted by nonlinear regression and compared with the use of an F-test or with the AIC. Alternatively, we may evaluate the value of the interaction term by a (stepwise) multiple regression. Kalmar & Currie (2007), when modelling the relationship between global bird richness and the two predictors area and climate (precipitation and temperature), found that the best model included climate plus the interaction term (i.e. the area term was excluded from the model), thus not excluding the interaction term but instead one of the others. If the multiple regression had thrown out the other two terms, one would have been left with a model consisting only of the interaction term, which is equivalent to the multiplicative term in the above described *combination variable* approach.

RECOMMENDATIONS

It may be hard to predict *a priori* which model is preferable, but I suggest the following recommendations. By grading models according to their fit, one may end up with unlikely curve shapes, especially for scales outside the sampled scale window. A model with a shape similar to the expected is,

therefore, often preferable. On the other hand, if the aim is only a best fit and scales outside the scale window of the data set are not discussed, any model may suffice given that it produces a good fit and produces no maxima or minima inside the scale window studied. One should keep in mind that a model that produces a good fit for one data set may not do so for another. One is often forced to fit the same model, as other researchers have applied to their data, in order to be able to compare parameter values, but one can do this in addition to the application of a better fitting model or models with better expected shapes, or both.

When there are few data points and large scatter, models with fewer parameters are often preferable, e.g. the power model, the logarithmic model (or maybe preferably the Kobayashi logarithmic) and the negative exponential model. With narrow-scale windows, even such inflexible models as the power model and the logarithmic model may produce good fits. With wider-scale windows and where inflexible models do not fit well, more flexible models may give better results, e.g. the P2 model and the Weibull distribution. Any model (whether convex or sigmoid, and whether with an upper asymptote or not) may be applied, as long as scales where the curve is J-shaped are not included. If extrapolations and expected shapes are important, one should consider models that produce curves exhibiting such shapes, e.g. the power model for sample area curves and the P2 model for isolate curves.

ACKNOWLEDGEMENT

REFERENCES

Adler, P.B. & Lauenroth, W.K. (2003) The power of time: spatiotemporal scaling of species diversity. *Ecology Letters*, **6**, 749–756.

Adler, P.B., White, E.P., Lauenroth, W.K., Kaufman, D.M., Rassweiler, A. & Rusak, J.A. (2005) Evidence for a general species–time–area relationship. *Ecology*, **86**, 2032–2039.

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

Archibald, E.E.A. (1949) The specific character of plant communities, I. Herbaceous communities. *Journal of Ecology*, **37**, 260–273.

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

Dodson, S. (1992) Prediction crustacean zooplankton species richness. *Limnology and Oceanography*, **37**, 848–856.

Evans, K.L., Lennon, J.J. & Gaston, K.J. (2007) Slopes of avian species–area relationships, human population density, and environmental factors. *Avian Conservation and Ecology*, **2**, article 7 (http://www.ace-eco.org/vol2/iss2/art7/).

Farebrother, R.W. (1998) Non-linear curve fitting and the true method of least squares. *The Statistician*, **47**, 137–147.

Fattorini, S. (2007) To fit or not to fit? A poorly fitting procedure produces inconsistent results when the species–area relationship is used to locate hotspots. *Biological Conservation*, **16**, 2531–2538.

Fattorini, S. & Fowles, A.P. (2005) A biogeographical analysis of the tenebrionid beetles (Coleoptera, Tenebrionidae) of the island of Thasos in the context of the Aegean islands (Greece). *Journal of Natural History*, **39**, 3919–3949.

Flather, C.H. (1996) Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, **23**, 155–168.

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

Gleason, H.A. (1925) Species and area. *Ecology*, **6**, 66–74.

Hanski, I. (1994) A practical model of metapopulation dynamics. *Journal of Animal Ecology*, **63**, 151–162.

Hanski, I. & Kuussaari, M. (1995) Butterfly metapopulation dynamics. *Population dynamics* (ed. by N. Cappuccino and P.W. Price), pp. 149–171. Academic Press, San Diego, CA.

Hanski, I. & Thomas, C.D. (1994) Metapopulation dynamics and conservation: a spatially explicit model applied to butterflies. *Biological Conservation*, **68**, 167–180.

Harte, J., Kinzig, A. & Green, J. (1999) Self-similarity in the distribution and abundance of species. *Science*, **284**, 334–336.

He, F. & Legendre, P. (2002) Species diversity patterns derived from species–area models. *Ecology*, **83**, 1185–1198.

Hill, J.L., Curran, P.J. & Foody, G.M. (1994) The effect of sampling on the species–area curve. *Global Ecology and Biogeography Letters*, **4**, 97–106.

Kalmar, A. & Currie, D.J. (2006) A global model of island biogeography. *Global Ecology and Biogeography*, **15**, 72–81.

Kalmar, A. & Currie, D.J. (2007) A unified model of avian species richness on islands and continents. *Ecology*, **88**, 1309–1321.

Keeley, J.E. & Fotheringham, C.J. (2005) Plot shape effects on plant species diversity measurements. *Journal of Vegetation Science*, **16**, 249–256.

Kobayashi, S. (1975) The species–area relation II. A second model for continuous sampling. *Researches on Population Ecology*, **16**, 265–280.

Lawton, J.H. (1999) Are there general laws in ecology? *Oikos*, **84**, 177–192.

Lomolino, M.V. (2002) Correspondence: ''... there are areas too small, and areas too large, to show clear diversity patterns...'' R. H. MacArthur (1972: 191). *Journal of Biogeography*, **29**, 555–557.

Motulsky, H.J. & Christopoulos, A. (2004) *Fitting models to biological data using linear and nonlinear regression*. Oxford University Press, Oxford.

Plotkin, J.B., Potts, M.D., Yu, D.W., Bunyavejchewin, S., Condit, R., Foster, R., Hubbell, S., Lafrankie, J., Manokaran, N., Lee, H.S., 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.

Preston, F.W. (1962) The canonical distribution of commonness and rarity: Part II. *Ecology*, **43**, 410–432.

Ratkowsky, D.A. (1990) *Handbook of nonlinear regression models*. Marcel Dekker, New York.

Scheiner, S.M. (2004) A mélange of curves – further dialogue about species–area curves. *Global Ecology and Biogeography*, **13**, 479–484.

Sibbesen, E. (1981) Some new equations to describe phosphate sorption by soils. *Journal of Soil Science*, **32**, 67–74.

Stiles, A. & Scheiner, S.M. (2007) Evaluation of species–area functions using Sonoran Desert plant data: not all species–area curves are power functions. *Oikos*, **116**, 1930–1940.

Storch, D., Evans, K.L. & Gaston, K.J. (2005) The species–area–energy relationship. *Ecology Letters*, **8**, 487–492.

Tjørve, E. (2002) Habitat size and number in multi-habitat landscapes: a model approach based on species–area curves. *Ecography*, **25**, 17–24.

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

Tjørve, E. & Tjørve, K.M.C. (2008) The species–area relationship, self-similarity, and the true meaning of the z-value. *Ecology*, **89**, 3528–3533.

Tjørve, E., Kunin, W.E., Polce, C. & Tjørve, K.M.C. (2008a) The species–area relationship: separating the effects of species-abundance and spatial distribution. *Journal of Ecology*, **96**, 1141–1151.

Tjørve, K.M.C., Schekkerman, H., Tulp, I., Underhill, L.G., De Leeuw, J.J. & Visser, G.H. (2007) Growth and energetics of a small shorebird species in a cold environment: the little stint *Calidris minuta* of the Taimyr Peninsula, Siberia. *Journal of Avian Biology*, **38**, 552–563.

Tjørve, K.M.C., Underhill, L.G. & Visser, G.H. (2008b) The energetic implications of precocial development for three shorebird species breeding in a warm environment. *Ibis*, **150**, 125–138.

Triantis, K.A., Mylonas, M., Lika, K. & Vardinoyannis, K. (2003) A model for the species–area–habitat relationship. *Journal of Biogeography*, **30**, 19–27.

Ulrich, W. (2006) Decomposing the process of species accumulation into area dependent and time dependent parts. *Ecological Research*, **21**, 578–585.

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.

Williamson, M., Gaston, K.J. & Lonsdale, W.M. (2001) The species–area relationship does not have an asymptote! *Journal of Biogeography*, **28**, 827–830.

Williamson, M., Gaston, K.J. & Lonsdale, W.M. (2002) An asymptote is an asymptote and not found in species–area relationships. *Journal of Biogeography*, **29**, 1713.

Wright, D.H. (1983) Species–energy theory: an extension of species–area theory. *Oikos*, **41**, 496–506.

BIOSKETCH

Even Tjørve is an Associate Professor at Lillehammer University College. His main research interests are in biogeography, landscape ecology and macroecology, with emphasis on the effect of spatial patterns and species-abundance distributions on species diversity. He is currently working with species–area curves and species-diversity models based on species–area curves. His other interests include regression analysis, bird migration, wader studies, endocrinology and metabolism, and nature-based travel and ecotourism.

Editor: Walter Jetz

APPENDIX 1

The '20 candidates list' of bivariate models reviewed in this and the previous paper on shapes and functions of species–area curves (Tjørve, 2003). Shape is described for arithmetic space. Number of parameters, whether the model curve has an upper asymptote (plus possible maxima) and whether the model curve goes through the origin of the axes are also given. Models may have other shapes or characteristics for negative parameter values or values outside certain ranges. Various reparameterizations and additions of location parameters are not included.

Model name	Model	Shape (used part)	Parameters	Upper asymptote	Intersects origin
Power model	ax^b	Convex	2	No	Yes
EPM 1 model	$ax^{bx^{-c}}$	Sigm./conv.	3	No/may have max.	Yes
EPM 2 model	$ax^{b-(c/x)}$	Sigmoid	3	No/have min.	No
P1 model	$ax^b \exp(-cx)$	Convex	3	No/have max.	Yes
P2 model	$ax^b \exp(-c/x)$	Sigmoid	3	No	Yes

Appendix 1 Continued

Model name	Model	Shape (used part)	Parameters	Upper asymptote	Intersects origin
Logarithmic model	$a + b (\log x)$	Convex	2	No	No
Kobayashi logarithmic	$a \log (1 + x/b)$	Convex	2	No	Yes
Negative exponential	$a[1 - \exp(-bx)]$	Convex	2	Yes	Yes
Common logistic	$a/[1 + \exp(-bx + c)]$	Sigmoid	3	Yes	No
Archibald logistic	$a/(b + c^x)$	Sigmoid	3	Yes	No
Gompertz	$a \exp[-\exp(-bx + c)]$	Sigmoid	3	Yes	No
Extreme-value function (EVF)	$a\{1 - \exp[-\exp(bx + c)]\}$	Sigmoid	3	Yes	No
Monod	$a[x/(b + x)]$	Convex	2	Yes	Yes
Asymptotic regression	$a - bc^{-x}$	Convex	3	Yes	May
Rational function	$(a + bx)/(1 + cx)$	Convex	3	Yes	May
Chapman–Richards	$a[1 - \exp(-bx)]^c$	Sigmoid	3	Yes	Yes
Cumulative Weibull distribution	$a[1 - \exp(-bx^c)]$	Sigmoid	3	Yes	Yes
Morgan–Mercer–Flodin (Hill function)	$ax^c/(b + x^c)$	Sigmoid	3	Yes	Yes
Lomolino function	$a/[1 + (b^{\log(c/x)})]$	Sigmoid	3	Yes	Yes
Cumulative beta-P distribution	$a\{1 - [1 + (x/c)^d]^{-b}\}$	Sigmoid	4	Yes	Yes