TOOLBOX

Frontiers of Biogeography
the scientific journal of
the International Biogeography Society

Thresholds and the species–area relationship: a set of functions for fitting, evaluating and plotting a range of commonly used piecewise models in R

Thomas J. Matthews[1, 2] and François Rigal[3,2]

[1]GEES (School of Geography, Earth and Environmental Sciences) and Birmingham Institute of Forest Research, University of Birmingham, Birmingham, B15 2TT; [2]CE3C – Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity Group and Universidade. dos Açores – Depto de Ciências Agráriase Engenharia do Ambiente, PT-9700-042, Angra do Heroísmo, Açores, Portugal; [3]CNRS - Université de Pau et des Pays de l'Adour - E2S UPPA, Institut Des Sciences Analytiques et de Physico Chimie pour L'environnement et les Materiaux UMR5254, 64000 Pau, France.
Corresponding author: Thomas J. Matthews: txm676@gmail.com

Abstract

An increasing number of studies have focused on identifying thresholds in the species–area relationship (SAR). The most common approach in such studies is to use piecewise regression models. While a few software packages are available for fitting piecewise models, these resources are general regression packages (i.e., they are not specifically designed for the analysis of SAR data) and tend to only provide functions for fitting a subset of the piecewise models proposed in the SAR literature. Given the large number of SAR studies now fitting piecewise models, there is a need for a software package that provides functions for fitting a range of piecewise models, including continuous, left-horizontal and discontinuous models in addition to supplementary functions for analysing model fits, in the context of SAR data. To this end, we provide a set of functions for fitting six piecewise regression models to SAR data, calculating confidence intervals around the breakpoint estimates (for certain models), comparing the models using various information criteria, and plotting the resultant model fits. Here, we present these functions and illustrate them using a selection of empirical datasets. These functions are implemented in the freely available and open-source R package 'sars.'

Highlights

- The possibility of thresholds in the species–area relationship (SAR) has long been recognized, but there are few software resources for fitting different kinds of threshold models to SAR data.
- Our functions allow users to fit a range of different one- and two-threshold piecewise regression models to SAR data, as well as providing tools to undertake a range of additional tasks, such as plotting model fits and comparing models using different criteria.
- These functions will allow authors interested in the SAR, and other types of diversity–area relationship, to test for thresholds in their data, ultimately expanding our understanding of how diversity scales with area.

Keywords: species–area relationship, piecewise regression, threshold, breakpoint, diversity–area relationship, islands, small island effect

Introduction

The species–area relationship (SAR) describes the relationship between the number of species found in an area and the size of the area, and is a commonly studied diversity pattern in island biogeography and macroecology (Rosenzweig 1995, Triantis et al. 2012, Matthews et al. 2016, Chase et al. 2019). Various types of SAR have been described (see Scheiner 2003 for a detailed overview of the different types) and here we focus primarily on island species–area relationships (the number of species found on islands of different sizes; ISARs) for ease, but the functions described can be used with any type of SAR. An increasing number of studies have focused on identifying thresholds in the ISAR, including studies i) focused on identifying the small island effect (i.e., a different relationship between island area and species richness on smaller compared to larger islands ; SIE; Lomolino and Weiser 2001, Gao

e-ISSN: 1948-6596 https://escholarship.org/uc/fb doi:10.21425/F5FBG49404

and Perry 2016, Wang et al. 2018), ii) determining whether the ISAR has an upper asymptote (Lomolino 2000), iii) looking for signals of speciation in the ISAR (Losos and Schluter 2000), and iv) identifying thresholds in habitat ISARs, which are often systems of conservation concern (Fahrig 2001, Matthews et al. 2014). The most common approach in such studies is to use piecewise regression (e.g., Lomolino and Weiser 2001, Toms and Lesperance 2003, Gentile and Argano 2005, Matthews et al. 2014, Gao et al. 2019). Piecewise regression (also known as breakpoint regression, segmented regression, and broken-stick regression) models are those where the independent variable is broken up into segments and separate lines are fitted to each segment (Toms and Lesperance 2003). The values of the independent variable where the breaks occur are termed breakpoints and are often interpreted as thresholds in the relationship between the independent and dependent variables (e.g., Toms and Lesperance 2003, Matthews et al. 2014).

A wide range of piecewise models have been fitted to ISAR data, with Gao et al. (2019) providing functions for 14 different models. These can be broadly grouped into continuous models, where the slope of the line changes at a breakpoint (Fig. 1a), and discontinuous models (Fig. 1b), where both the slope and intercept of the line can change at the breakpoint (i.e., the lines are not continuous). Left-horizontal models are a subset of continuous models where the slope of the line to the left of the first breakpoint is zero (Fig. 1c). The slope of the line to the right of a breakpoint (in all piecewise models) can be steeper or shallower than the slope of the line to the left, leading to what have been termed *shallow-steep* and *steep-shallow* relationships, respectively (Matthews et al. 2014, Gao et al. 2019). Most work to date has focused on one-threshold (i.e., single breakpoint) piecewise models, but it has recently been suggested that the number of thresholds in the ISAR often increases with the range in island area analysed (Gao et al. 2019). For example, it has been argued that the ISAR exhibits both a SIE and a change in slope at large island areas where *in situ* speciation becomes a dominant assembly process (Losos and Schluter 2000, Lomolino and Weiser 2001).

Piecewise models can be fitted to untransformed ISAR data (i.e., where both richness and area are untransformed), data where only area is log-transformed, and data where both richness and area are log-transformed. Different transformations can result in better model fits depending on the data (and models) at hand (Dengler 2009). However, the use or not of variable transformation has been found

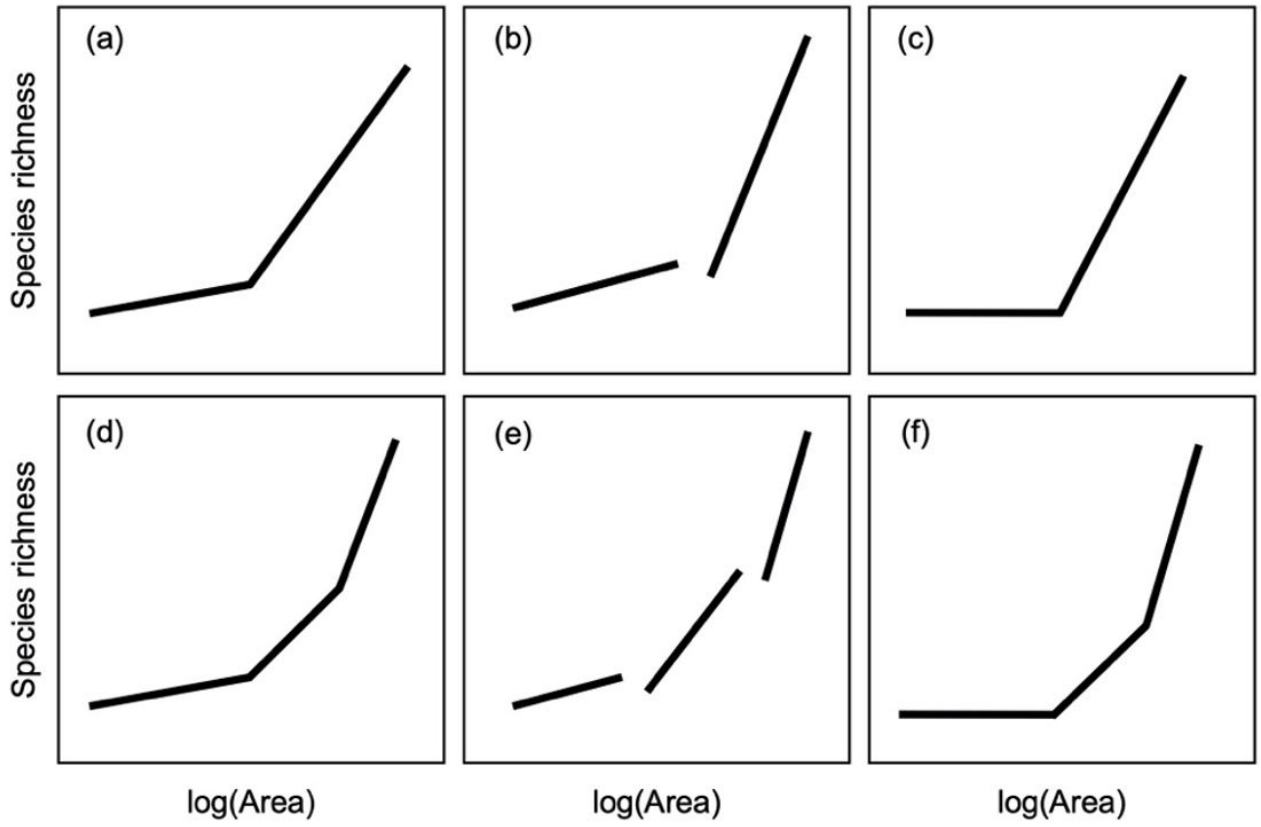

Figure 1. Schematic illustration of the six piecewise models implemented in our study, comprising three one-threshold models (a-c) and three two-threshold models (d-f). The models are the continuous one-threshold (a) and two-threshold (d), discontinuous one-threshold (b) and two-threshold (e), and left-horizontal one-threshold (c) and two-threshold (f) models. The models are presented using a semi-log transformation (area is log-transformed but species richness is not), but the models can also be fitted to untransformed or log–log transformed data.

to affect the conclusions drawn in regard to whether or not a threshold is present in the ISAR (Burns et al. 2009, Matthews et al. 2014), and is thus an important consideration in ISAR threshold studies.

A few software packages are available that enable users to fit piecewise regression models, including the 'SiZer' (Sonderegger 2020) and 'segmented' (Muggeo 2008) R packages, the latter being the most widely used package for fitting these types of models in the R language (R Core Team 2019). However, these resources are general regression packages (i.e., they are not specifically designed for the analysis of [I]SAR data) and tend to only provide functions for fitting a subset of the breakpoint models proposed in the ISAR literature. Given the large number of ISAR studies now fitting piecewise models, there is a need for a software package that provides functions for fitting a range of piecewise models (including continuous, left-horizontal and discontinuous models) in addition to supplementary functions for analysing model fits in the context of ISAR data. To this end, we provide a set of functions for fitting six piecewise regression models to ISAR data, calculating confidence intervals around the breakpoint estimates (for certain models), comparing the models using various information criteria, and plotting the resultant model fits. The framework and functions we present can also be easily adapted to fit additional threshold models (e.g., those with three breakpoints).

Models and model fitting

The six piecewise models included here (Fig. 1) are a selection of 6 models out of the 14 listed by Gao et al. (2019) and comprise the continuous one-threshold (Eq. 1, Fig. 1a) and two-threshold (Eq. 2, Fig. 1d), discontinuous one-threshold (Eq. 3, Fig. 1b) and two-threshold (Eq. 4, Fig. 1e), and left-horizontal one-threshold (Eq. 5, Fig. 1c) and two-threshold (Eq. 6, Fig. 1f) models (see Gao et al. 2019 for further information). The equations of these six models are:

$$S = c_1 + (\log A \le T) z_1 \log A + (\log A > T)\left[z_1 T + z_2 (\log A - T)\right] \quad (1)$$

$$S = c_1 + (\log A \le T_1) z_1 \log A + (\log A > T_1 \ \&\ \log A \le T_2) \left[z_1 T_1 + z_2 (\log A - T_1)\right] + (\log A > T_2)\left[z_2 (T_2 - T_1) + z_3 (\log A - T_2)\right] \quad (2)$$

$$S = (\log A \le T)(c_1 + z_1 \log A) + (\log A > T)(c_2 + z_2 \log A) \quad (3)$$

$$S = (\log A \le T_1)(c_1 + z_1 \log A) + (\log A > T_1 \ \&\ \log A \le T_2) (c_2 + z_2 \log A) + (\log A > T_2)(c_3 + z_3 \log A) \quad (4)$$

$$S = c_1 + (\log A > T) z_2 (\log A - T) \quad (5)$$

$$S = c_1 + (\log A > T_1 \ \&\ \log A \le T_2)\left[z_2 (\log A - T_1)\right] + (\log A > T_2)\left[z_2 (T_2 - T_1) + z_3 (\log A - T_2)\right] \quad (6)$$

where *logA* is the log-transformed island area, and c_i (intercept), z_i (slope), and T_i (thresholds) are fitted parameters. The logical expressions (e.g., *logA* > *T*) return a value of 1 if they are true and a value of 0 if they are false. We have presented the equations using the semi-log transformation approach (area was log-transformed but not species richness) but these models can also be fitted using either the untransformed or the log–log approach (Matthews et al. 2014).

Within the functions, the models are fitted using ordinary least squares regression (OLS) and the 'lm' function in R. The optimum threshold / breakpoint values are chosen by iterating across values and selecting those that result in the minimum residual sum of squares (RSS). For the continuous models (i.e., the continuous and left-horizontal one and two-threshold models), this iteration process works by selecting the smallest island area value as the first fitted threshold, and then adding an increment (set by the user; discussed below) at each iteration up to one increment below the maximum area value to avoid fitting models with no island in the right segment. For the discontinuous piecewise models, it does not make sense to choose values that are not observed area values as thresholds because segments are disconnected and doing so would leave a gap between the last island and the first island of two consecutive segments. Therefore, for these models, we simply used the full set of observed area values as the sequence of thresholds for fitting. As for continuous models, we did not include the maximum area value in the set of potential thresholds to avoid fitting models with no island in the right segment.

For the discontinuous models, the actual identified threshold (i.e., the value returned by the function) between two consecutive segments can be the area of either the last island of the first segment or the first island of the second segment. In our functions, we decided to report the value of the last island of the first segment as it represents the value after which another relationship between species richness and area is estimated. However, due to this issue the threshold values identified in the discontinuous models have to be interpreted with caution and cannot be fully compared to the thresholds of continuous models, the latter of which represent the predicted area where the slope of the relationship changes. For both the continuous and discontinuous models, the threshold values that result in the minimum RSS are chosen.

For the two-threshold models, the first breakpoint (T1) was estimated before the second (T2). With the discontinuous two-threshold model, the first breakpoint (T1) was first assigned to one of the observed area values, and the second breakpoint (T2) was then assigned to one of the observed area values between T1 and the maximum observed area value (not including the max area value). T1 was then shifted to the next observed area value and the process repeated, and so on. With the continuous two-breakpoint models, for each value of T1, T2 was then assigned to all values between T1 and the maximum observed value (again, not including the max area), in units of the increment argument. T1 was then shifted according to the increment argument and the process repeated, and so on. In rare cases, multiple breakpoint values can return the same minimum RSS. In these cases, we just randomly choose and return one breakpoint value

(see Dengler et al. 2020) and also produce a warning. However, if this occurs, it is worth checking the data and model fits carefully as this tends to be a sign that the resultant model fit is poor and/or the dataset is noisy (i.e., there is no discernible ISAR).

Functions

All functions are written in R (R Core Team 2019) and are contained in the most recent version of the 'sars' R package (version 1.3.0; Matthews et al. 2019), available on CRAN and GitHub ("txm676/sars"). In line with the rest of the 'sars' package, the functions have been programmed using standard S3 methods. The main function is 'sar_threshold', which fits the six piecewise models, in addition to a linear model and an intercept only model for comparison (using the 'non_th_models' argument). The 'mod' argument can be used to select individual piecewise models to be fitted, a selection of models, or all six models ('All'). The 'logAxes' argument can be used to fit the models to untransformed data ('logAxes' = "none") or log(Richness) ~ log(Area) data ('logAxes' = "both"); the default (Richness ~ log(Area) data; 'logAxes' = "area") fits models to semi-log transformed data, as this transformation is often used in ISAR threshold studies (e.g., Morrison 2014, Matthews et al. 2020). When richness is log-transformed, a constant (set using the 'con' argument, default = 1) is added to all richness values if any zeros are detected. Other alternatives (e.g., removing all zero species islands or only adding constants to the zero species islands) would need to be undertaken by the user prior to using the function. The log-function used can be selected using the 'logT' argument with logT = "log" for natural logarithms (the default setting), logT = "log2" for log to the base 2, and logT = "log10" for log to the base 10. The 'nisl' argument can be used to constrain the minimum of number of islands that should be included in the first and last segment. This argument was added to the function because we observed that, in a small number of cases, thresholds that result in only one island being present in the first or last segment can be identified. Constraining the number of islands could therefore be useful to avoid such situations and thus reduce the risk of returning model fits with questionable ecological meaning. By default, the argument is set to NULL, and if the user selects a minimum of number of islands that is equal or larger than half of the total number of islands, the functions stop and a warning message is returned.

For the continuous and left-horizontal models, an important part of the model fitting process is selecting a suitable value (using the 'interval' argument) for incrementing the threshold during the iterative fitting procedure. The defaults used in 'sar_threshold' are different depending on whether untransformed or log-transformed area is used as the independent variable. For untransformed area, the default 'interval' is 1, while for log-transformed area it is 0.1. However, depending on the range of island areas in a given dataset, these defaults may not be appropriate; for example, when 'interval' is too large (relative to the size of the smallest island in a dataset), it can result in a single data point being included within one or more of the model segments (discussed above). It can also bias the confidence intervals (discussed below) of the one-threshold continuous models downwards. Thus, users are advised to select their own 'interval' values given their data. As discussed above, the observed area values are used to select the optimal breakpoint in the discontinuous models, and thus the 'interval' argument is not relevant in these cases. If the selected interval is small (relative to the range of observed island area), fitting the continuous and left-horizontal models can be relatively time consuming; however, we would argue that it is better to select a smaller interval argument (and thus take longer in fitting the models) and be more confident that the optimum breakpoint has been found. Fitting the continuous and left-horizontal two-threshold models can be particularly time consuming if the range in area is large and/or the selected interval is small. To deal with this, we have integrated parallel processing into the fitting process for these two-threshold models, which can be set using the 'parallel' and 'cores' arguments. While the default for 'parallel' is set to FALSE (as parallel processing may not be appropriate for all users), we advise its use when fitting the two-threshold models.

More generally, due to the increased number of parameters, fitting piecewise models to datasets with few islands is not recommended. In particular, we would advise against fitting the two-threshold models to small SAR datasets. A rough rule of thumb would be to not fit one-threshold models to datasets with fewer than 10 islands and two-threshold models to datasets with fewer than 20 islands. However, it should be noted that a recent study has argued that a minimum sample size of at least 25 islands is required for fitting even simple linear SAR models (i.e., models with no thresholds) when there is high variance in the data (Jenkins and Quintana-Ascencio 2020). This leads on to a more general point that all threshold model fits (as with any regression model fit) should be checked carefully, the fitted relationship plotted and different parameter settings tested (e.g., the interval arguments).

```
#load an example dataset, and fit the continuous
#two-threshold model
#to the data (with area transformed using log to the
#base 10), using an interval of 0.1 (for
#speed) and parallel processing. Plot the resultant
#model fit (plotting discussed in more detail
#below).

library(sars)
data(aegean2)
fit <- sar_threshold(data = aegean2, mod = c("ContTwo"),
interval = 0.1,
    non_th_models = FALSE, logAxes = "area", con = 1,
    logT = log10, nisl = NULL, parallel = TRUE, cores = 3)
plot(fit, cex = 0.8, cex.main = 1.1, cex.lab = 1.1, pcol
= 'grey') #Figure 1
```

The 'sar_threshold' function returns a list of class 'threshold' and class 'sars' with five elements, containing various details about the model fitting procedure. The individual model fit objects (returned from the 'lm' function) are provided in the first element. These 'lm' fit objects can be used to generate classic diagnostic plots for linear regression using the standard 'plot' function (e.g., QQ-plot, Cook's distance). This is recommended as

there are no automatic model validation tests undertaken within the sar_threshold function. Summary and plot generic functions are available which return user-friendly output. The 'summary.sars' generic function (when applied to an item of class 'threshold') generates a list with three elements, where the second element is a model summary table. The table provides, for each fitted model, a range of useful information, including AIC, BIC and AIC_c values, the model R^2 and adjusted R^2, the threshold / breakpoint values, and the number of data points in each fitted segment. Models are ordered in the table according to a selected information criterion. Table 1 provides an example model summary table. Because we considered the search for the breakpoint to represent a free parameter (cf. Matthews et al. 2014), we added one unit per threshold to the original number of model parameters in the piecewise models, increasing the number of model parameters to k=4 for the left-horizontal one-threshold model, k=5 for the continuous one-threshold, k=6 for the discontinuous one-threshold and left-horizontal two-threshold models, k=7 for the continuous two-threshold model, and k=9 for the discontinuous two-threshold model. These parameter numbers are then used in the calculation of the various information criteria.

```
#fit all six piecewise models to the aegean2 dataset,
#along with the linear and intercept-only
#models, using parallel processing for the two-threshold
#continuous models, and generate a
#summary table (Table 1) where models are ordered
#by BIC. Note that the number of cores to
#select depends on the specific computer being used

fit2 <- sar_threshold(data = aegean2, mod = "All",
interval = 0.1, nisl = NULL, non_th_models = TRUE,
logAxes = "area", logT = log10, parallel = TRUE, cores = 4)
s <- summary(fit2, order = "BIC")
s[[2]] #Table 1
```

As the coefficients in the fitted breakpoint regression models do not all represent the intercepts and slopes of the different segments (for these it is necessary to add different coefficients together), a separate function ('get_coef') can be used to calculate these. Table 2 provides an example of this output.

```
#load a dataset of invertebrates on 90 Aegean islands,
#fit four piecewise models to the
#dataset (and not the intercept-only and linear models)
#using parallel processing (for the two
#threshold models) and log-transformed area and
#richness, and generate the intercept and
#slope values from the model fits (Table 2)

data(aegean)
fit3 <- sar_threshold(data = aegean, mod = c("ContOne",
"ContTwo", "ZslopeOne", "ZslopeTwo"), interval =
0.1, non_th_models = FALSE, logAxes = "both", logT
= log10, parallel = TRUE, cores = 3)
get_coef(fit3) #intercept and slope values (Table 2)
```

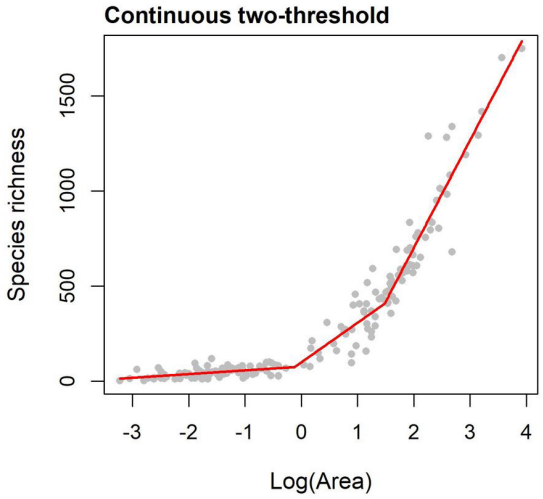

Figure 2. The fit (red lines) of the continuous two-slope piecewise model (Eq. 2, above) to a dataset of plants on 173 islands (black circles) in the Aegean Sea. Island area was log-transformed (using log10) prior to model fitting.

Table 1. Model summary table, generated using the 'summary.sars' generic function. The table provides information about the fit of six piecewise models, in addition to linear and intercept-only models. See the legend of Fig.2 for dataset information. "Zslope" models are the left-horizontal models, "Cont" are the continuous models, and "Disc" are the discontinuous models; the numbers ("One" or "Two") relate to the number of thresholds in the model. LL is the log likelihood of the model and Pars is the number of parameters. The R2 and R2a are the model R^2 and adjusted R^2 values. Th1 and Th2 are the $\log_{10}$ threshold value(s), and seg1, seg2, seg3 provide the number of data points within each segment (for the threshold models). Note that in the function output, the dashes are actually NAs.

Model	LL	Pars	AIC	AICc	BIC	R2	R2a	Th1	Th2	seg1	seg2	seg3
ZslopeTwo	-1004.13	6	2020.25	2020.76	2039.17	0.95	0.95	-0.22	1.48	86	37	50
ContTwo	-1002.74	7	2019.48	2020.16	2041.55	0.95	0.95	-0.12	1.48	86	37	50
DiscTwo	-1002.21	9	2022.42	2023.52	2050.80	0.95	0.95	0.16	1.69	88	46	39
ContOne	-1017.97	5	2045.94	2046.30	2061.71	0.94	0.94	0.98	-	105	68	-
DiscOne	-1017.89	6	2047.78	2048.28	2066.70	0.94	0.94	0.96	-	104	69	-
ZslopeOne	-1026.85	4	2061.70	2061.94	2074.31	0.94	0.94	0.58	-	94	79	-
Linear	-1149.52	3	2305.05	2305.19	2314.51	0.74	0.74	-	-	-	-	-
Intercept	-1265.92	2	2535.84	2535.91	2542.14	0.00	0.00	-	-	-	-	-

A 'plot.threshold' generic function uses the base R plotting framework and allows users to plot individual model fits, or a collection of fits in the same plotting window (e.g., Fig. 3). Standard plotting arguments (e.g., to change point and line size, point and line colours, plotting window dimensions, and axes and plot titles) enable users to edit most aspects of the plots.

```
#Plot the fit object from above (Table 1), and change
#the x-axis name, provide a vector of
#model titles, and set the line and point colour

par(mai = c(0.6, 0.6, 0.3, 0.3)) #change the graph margins
plot(fit2, xlab = "Area (Log10)", ModTitle = c("a) ContOne", "b) ZslopeOne", "c) DiscOne", "d) ContTwo", "e) ZslopeTwo", "f) DiscTwo", "g) Linear", "h) Intercept"), pcol = "grey", lcol = "orange")
```

Table 2. The intercepts (ci) and slopes (zi) of the different segments in four piecewise regression models fitted to a dataset of invertebrates on 90 islands in the Aegean Sea. Both area and richness were log-transformed prior to model fitting. Dashes indicate a parameter is not in a given model. Note that in the function output, the dashes are actually NAs.

Model	c1	z1	c2	z2	c3	z3
ContOne	0.70	0.13	-	0.27	-	-
ZslopeOne	0.53	-	-	0.26	-	-
ContTwo	0.66	0.10	-	0.51	-	0.23
ZslopeTwo	0.37	-	-	0.13	-	0.27

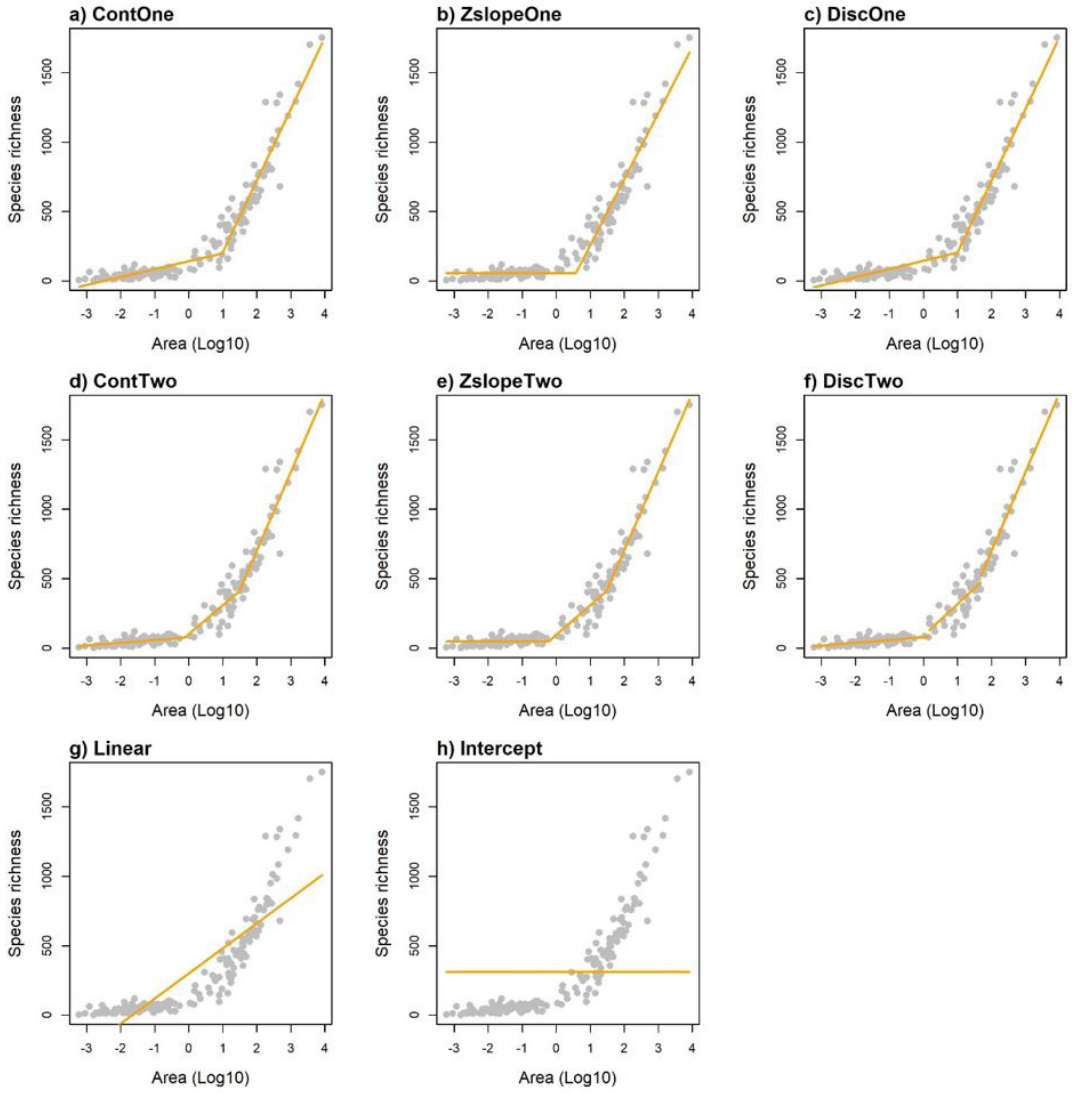

Figure 3. The fits (orange lines) of six piecewise models, a linear model and an intercept-only model to a dataset of plants on 173 islands (grey circles) in the Aegean Sea. Island area was log-transformed (using $\log_{10}$) prior to model fitting. Table 1 provides information about the fit of the six piecewise models (a-f) including information criteria and R^2, in addition to the linear (g) and intercept-only (h) models.

A separate function ('threshold_ci') generates the confidence intervals around the breakpoints of the one-threshold continuous and left-horizontal models. Two types of confidence interval can be implemented (using the 'method' argument): (1) a confidence interval derived from an inverted *F* test ('method' = "F"), and (2) an empirical bootstrap confidence interval ('method' = "boot"). Full details of the two approaches can be found in Toms and Lesperance (2003). The argument 'cl' can be used to select the confidence level, and the level is set to 0.95 (95%) by default. When the bootstrapping method is selected, all bootstrap samples are returned. It should be noted that, depending on the number of bootstrap samples selected and the interval value chosen, the function can take a (very) long time to run. On the other hand, if the selected interval is too large, every estimated bootstrap value will simply be the same as the fitted value and the confidence interval will be zero. Following Toms and Lesperance (2003), we therefore recommend the use of the inverted *F* test confidence interval when sample size is large, and bootstrapped confidence intervals when sample size is smaller. As the breakpoints in discontinuous models can only be observed area values, these two methods are not applicable for these models. Further work is needed to extend these approaches to the two-threshold continuous models.

```
#fit the one-threshold left-horizontal model to the
#aegean2 dataset and generate the 95%
#confidence interval around the breakpoint estimate
#using bootstrapping (100 bootstrap
#samples) and the same interval argument as in the
#model fit

fit4 <- sar_threshold(data = aegean2, mod =
c("ZslopeOne"), interval = 0.1, non_th_models = FALSE,
logAxes = "area", logT = log10)
fit4[[3]] #print the breakpoint (n.b. is on log10 area scale)
[1] 0.5781513

threshold_ci(fit4, method = 'boot ', interval = 0.1,
Nboot = 100)

Threshold confidence interval summary
Model: ZslopeOne
Confidence interval of the breakpoint: 0.48 - 0.68
```

A warning on the use of discontinuous models

It is necessary to mention that several authors have cautioned the use of discontinuous piecewise functions for modelling SARs. There are two points in particular that are worth highlighting here. First, studies have recently questioned the ecological logic of discontinuous relationships (e.g., Yu et al. 2020), arguing that the modelling of macro scale processes in nature using discontinuous models is inappropriate (Dengler 2010). Indeed, discontinuous relationships could reflect the absence of confounding factors (variables) that were not included in the models (e.g., isolation) rather than a true mechanistic link between species richness and area. Second, through testing the models and working with piecewise models in previous work, we have observed potential overfitting issues in relation to discontinuous models, even when more stringent model selection criteria, such as BIC, are used. Future research on this issue is needed. Despite these issues, given that several studies continue to fit discontinuous models, we decided it was preferable to provide the functions to fit these models and let authors decide on their utility.

Conclusions

It is worth re-emphasising that, in certain circumstances (e.g., large datasets), the various functions can take a while to run. Overall, our approach of iterating across different potential threshold values using the 'interval' argument is slower than, for example, simply using non-linear least squares and the 'nls' R package. However, if 'interval' is sufficiently small, the approach has the advantage of undertaking a more comprehensive search of parameter space, providing a greater chance of locating the optimum parameter estimates for any type of dataset. The use of 'nls' in particular is very sensitive to the (required) user-provided starting parameter estimates. If a user wants to reduce computation time when using the functions, we have three recommendations. First, increase the 'interval' argument, although it is important to remember that there is a trade-off between the accuracy of the breakpoint estimation and the computing time. Second, use parallel processing (i.e., set the 'parallel' argument to TRUE and specify the number of cores to use) if fitting the two-threshold continuous and left-horizontal models. Third, if calculating confidence intervals around the threshold estimate, use an inverted *F* test confidence interval when sample size is large.

The functions outlined here provide a set of tools for fitting, evaluating, and plotting a range of commonly used piecewise models in ISAR threshold research. Whilst the six piecewise models included in the package represent some of the most commonly used models (e.g., Lomolino and Weiser 2001, Matthews et al. 2014, 2020, Gao and Perry 2016, Wang et al. 2018), they are by no means the only models with thresholds that have been used in ISAR studies (see Toms and Lesperance 2003, Gao et al. 2019). However, using the source code, it would be relatively straightforward to add additional piecewise models into this fitting framework, and interested users are welcome to contribute code for fitting any models not included here (e.g., through GitHub).

While the focus of this paper has been on ISAR data (i.e., where the response variable represents the number of species), there is no reason these functions cannot be used to fit piecewise models to other diversity–area relationships, including island functional diversity–area relationships (Whittaker et al. 2014) and phylogenetic diversity–area relationships (Helmus and Ives 2012). Indeed, a number of recent studies have explored the small island effect in plants using breakpoint models in combination with island functional and phylogenetic diversity (Schrader et al. 2020, Matthews et al. 2020). Expansion of the approaches in these studies to a wider range of systems and taxonomic groups will likely increase our understanding of the mechanisms

underpinning the small island effect. The functions provided here should aid these future endeavours, and the investigation of thresholds in diversity–area relationships more generally.

A final point on the calculation of information criteria

Piecewise models are not the only approach that have been proposed for testing for the presence of thresholds (e.g., the SIE) in the ISAR and users might want to compare the performance of our 6 models to other functions, such as sigmoidal models (Schrader et al. 2019, Tjørve and Tjørve in press; the 'sars' package provides functions for fitting several sigmoid models to SAR data, see Matthews et al. 2019). In this vein, we would like to highlight a specific issue inherent to model comparison and selection. There are different ways to calculate the various information criteria (IC) used for model comparison (e.g., AIC, BIC). One difference relates to whether the RSS or the log-likelihood (LL) is used in the IC formulas. Under standard assumptions (e.g., independence of data points, homoscedasticity and normality of the residuals), the two approaches produce identical parameter estimates for regression models. However, the formulas are different and thus can produce different absolute IC values for the same model. For example, historically in the 'sars' package we have calculated IC values using formulas based on the RSS (Burnham and Anderson 2002, Guilhaumon et al. 2008). This meant that the IC values generated in 'sars' were not comparable with values generated in packages using different formulas. For example, in the 'nls' (the main function for non-linear regression in R) and 'lm' functions in the stats R package, a LL approach is used, meaning IC values from models fitted using 'nls' could not be compared with IC values from 'sars' models. To re-iterate, the parameter estimates are comparable, and the relative IC values (calculated using the same formula) are the same, it is simply that the absolute IC values will differ. In this new version of the package (version 1.3) we have changed our IC formulas to match those in 'nls' and 'lm'. Thus, if users wish to compare IC values with models fitted in 'sars', this is now straightforward. To re-create IC values from previous studies (i.e., those using a version of sars pre 1.2.2), it will be necessary to download 'sars' Version 1.2.1 or earlier (either from CRAN or GitHub; version 1.1.1 was published as a release on GitHub). It is important to note that these are not the only two ways of calculating ICs for regression models, and other formulas exist. Thus, if building models using other functions and packages (i.e., other than 'nls' or 'lm'), users should make sure to check how these packages calculate IC values before comparing with models fitted in 'sars'. In 'sars', as in 'nls', we include an additional parameter for estimation of the variance. Finally, if users are comparing models fitted in the package with their own models fitted using other packages, it is essential that IC values are all calculated using the same dependent variable (e.g., untransformed richness when using 'sar_average', and either untransformed or log-transformed richness when using 'sar_threshold').

Acknowledgments

Joe Wayman tested the new functions and reviewed an earlier draft of the paper. The information criteria formula changes were made in collaboration with François Guilhaumon, and the code for calculating confidence intervals using an inverted *F* test was written with the help of Christian Paroissin. The Aegean datasets were provided by Kostas Triantis. Our understanding of ISAR thresholds has benefited from discussions with Kostas Kougioumoutzis, Manuel Steinbauer, Kostas Triantis, Panayiotis Trigas and Rob Whittaker.

References

Burnham, K.P. & Anderson, D.R. (2002) Model selection and multi-model inference: a practical information-theoretic approach, 2nd edn. Springer, New-York.

Burns, K.C., Paul McHardy, R. & Pledger, S. (2009) The small-island effect: fact or artefact? Ecography, 32, 269-276.

Chase, J.M., Gooriah, L., May, F., Ryberg, W.A., Schuler, M.S., Craven, D. & Knight, T.M. (2019) A framework for disentangling ecological mechanisms underlying the island species–area relationship. Frontiers of Biogeography, 11, e40844.

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

Dengler, J. (2010) Robust methods for detecting a small island effect. Diversity and Distributions, 16, 256-266.

Dengler, J., Matthews, T.J., Steinbauer, M.J., et al. (2020) Species–area relationships in continuous vegetation: evidence from Palaearctic grasslands. Journal of Biogeography, 47, 72-86.

Fahrig, L. (2001). How much habitat is enough? Biological conservation, 100, 65-74.

Gao, D. & Perry, G. (2016) Detecting the small island effect and nestedness of herpetofauna of the West Indies. Ecology and Evolution, 6, 5390-5403.

Gao, D., Cao, Z., Xu, P. & Perry, G. (2019) On piecewise models and species–area patterns. Ecology and Evolution, 9, 8351-8361.

Gentile, G. & Argano, R. (2005) Island biogeography of the Mediterranean Sea: the species–area relationship for terrestrial isopods. Journal of Biogeography, 32, 1715-1726.

Guilhaumon, F., Gimenez, O., Gaston, K.J. & Mouillot, D. (2008) Taxonomic and regional uncertainty in species-area relationships and the identification of richness hotspots. Proceedings of the National Academy of Sciences USA, 105, 15458-15463.

Helmus, M.R. & Ives, A.R. (2012) Phylogenetic diversity–area curves. Ecology, 93, S31-S43.

Jenkins, D.G. & Quintana-Ascencio, P.F. (2020) A solution to minimum sample size for regressions. PLoS One, 15, e0229345.

Lomolino, M.V. (2000). Ecology's most general, yet protean pattern: the species-area relationship. Journal of Biogeography, 27, 17-26.

Lomolino, M.V. & Weiser, M.D. (2001) Towards a more general species–area relationship: diversity on all islands, great and small. Journal of Biogeography, 28, 431-445.

Losos, J.B. & Schluter, D. (2000) Analysis of an evolutionary species–area relationship. Nature, 408, 847-850.

Matthews, T.J., Steinbauer, M.J., Tzirkalli, E., Triantis, K.A. & Whittaker, R.J. (2014) Thresholds and the species–area relationship: a synthetic analysis of habitat island datasets. Journal of Biogeography, 41, 1018-1028.

Matthews, T.J., Guilhaumon, F., Triantis, K.A., Borregaard, M.K. & Whittaker, R.J. (2016) On the form of species–area relationships in habitat islands and true islands. Global Ecology and Biogeography, 25, 847–858.

Matthews, T.J., Triantis, K., Whittaker, R.J. & Guilhaumon, F. (2019) sars: an R package for fitting, evaluating and comparing species–area relationship models. Ecography, 42, 1446-1455.

Matthews, T. J., Rigal, F., Kougioumoutzis, K., Trigas, P. & Triantis, K. A. (2020) Unravelling the small-island effect through phylogenetic community ecology. Journal of Biogeography, DOI: 10.1111/jbi.13940.

Morrison, L. W. (2014) The small-island effect: empty islands, temporal variability and the importance of species composition. Journal of Biogeography, 41, 1007–1017.

Muggeo, V.M.R. (2008) Segmented: an R package to fit regression models with broken-line relationships. R News, 8, 20-25.

R Core Team (2019) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Rosenzweig, M.L. (1995) Species diversity in space and time. Cambridge University Press, Cambridge.

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

Schrader, J., Moeljono, S., Keppel, G., & Kreft, H. (2019) Plants on small islands revisited: the effects of spatial scale and habitat quality on the species–area relationship. Ecography, 42, 1405-1414.

Schrader, J., König, C., Triantis, K.A., Trigas, P., Kreft, H. & Weigelt, P. (2020) Species–area relationships on small islands differ among plant growth forms. Global Ecology and Biogeography, 29, 814-829.

Sonderegger, D. (2020). SiZer: Significant Zero Crossings. R package version 0.1-7. https://CRAN.R-project.org/package=SiZer

Tjørve, E. & Tjørve, K.M.C. (in press) Mathematical expressions for the species–area relationship and the assumptions behind the models. In: The species–area relationship: theory and application (ed. by T.J. Matthews, K.A. Triantis and R.J. Whittaker). Cambridge University Press, Cambridge.

Toms, J.D. & Lesperance, M.L. (2003) Piecewise regression: a tool for identifying ecological thresholds. Ecology, 84, 2034-2041.

Triantis, K.A., Guilhaumon, F. & Whittaker, R.J. (2012) The island species–area relationship: biology and statistics. Journal of Biogeography, 39, 215-231.

Wang, Y., Chen, C. & Millien, V. (2018) A global synthesis of the small-island effect in habitat islands. Proceedings of the Royal Society B, 285, 20181868.

Whittaker, R.J., Rigal, F., Borges, P.A.V., Cardoso, P., Terzopoulou, S., Casanoves, F., Pla, L., Guilhaumon, F., Ladle, R.J. & Triantis, K. (2014) Functional biogeography of oceanic islands and the scaling of functional diversity in the Azores. Proceedings of the National Academy of Sciences USA, 111, 13709–13714.

Yu, J., Li, D., Zhang, Z. & Guo, S. (2020) Species–area relationship and small-island effect of bryophytes on the Zhoushan Archipelago, China. Journal of Biogeography, 47, 978–992.

Submitted: 3 August 2020
First decision: 21 September 2020
Accepted: 9 October 2020

Edited by Janet Franklin

