00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

Ecography 33: 420–424, 2010
doi: 10.1111/j.1600-0587.2010.06304.x

Subject Editor: Thiago Rangel. Accepted 29 April 2010

mmSAR: an R-package for multimodel species–area relationship inference

François Guilhaumon, David Mouillot and Olivier Gimenez

F. Guilhaumon (Francois.Guilhaumon@univ-montp2.fr) and D. Mouillot, Laboratoire Ecosystèmes Lagunaires, Unité Mixte de Recherche 5119, Centre National de la Recherche Scientifique-IFREMER-UM2, Univ. Montpellier 2, cc 093, Place Eugène Bataillon, FR-34095 Montpellier Cedex 5, France. – O. Gimenez, Centre d'Ecologie Fonctionnelle et Evolutive, Unité Mixte de Recherche 5175, Campus Centre National de la Recherche Scientifique, 1919 Route de Mende, FR-34293 Montpellier Cedex 5, France.

The species–area relationship (SAR) is one of the most fundamental tools in ecology. After almost a century of quantitative ecology, however, the quest for a ''best SAR model'' still remains elusive, with a substantial uncertainty about the best fitting SAR model frequently being observed. Recent research has required that this uncertainty be addressed, and a multimodel SAR framework has been devised. Here we introduce the mmSAR R-package, which is a flexible and scalable implementation of the multimodel SAR framework for species-area datasets, and provide some examples of its use. This R-package provides functions for fitting SAR models, performing model selection, and the build up of multimodel SARs.

One of the most ancient and ubiquitous patterns that has been recognized in ecology is the increase in species richness (S) with increasing sampling area (A): the species–area relationship (SAR). The SAR has been mystifying ecologists for more than 150 years (De Candolle 1855, MacArthur and Wilson 1967, Connor and McCoy 1979, Drakare et al. 2006, Southwood et al. 2006) and its modelling remains a central issue for theoretical ecologists and conservationists (Rosenzweig 1995, Smith 2010). Inference about the SAR is mandatory in the wide range of conservation applications that require the comparison of diversity patterns when regions differ in area, such as global scale conservation priority-setting schemes (Brooks et al. 2006, Lamoreux et al. 2006, Wilson et al. 2007). In theoretical studies, SARs are considered to be fundamental properties of biological systems and are, for example, explained in terms of species abundances and spatial distribution of individuals (He and Legendre 2002, Martin and Goldenfeld 2006) and constitute a cornerstone for macroecological investigations (Šizling and Storch 2004, Drakare et al. 2006). After Arrhenius (1921), the SAR has mainly been modelled using a power law ($S = cA^z$, where c and z are constants to be estimated). Despite this historical hegemony, however, several studies have highlighted other functional forms for SARs (Gleason 1922, Coleman et al. 1982, Lomolino 2000, Tjørve 2003, 2009). Moreover, quantitative studies focusing on comparisons among models have indicated that the power law SAR is not ubiquitous (Connor and McCoy 1979, Flather 1996, Stiles and Scheiner 2007), stressing the importance of testing the relative fit of various different models in SAR analyses (Smith 2010). Furthermore, recent analyses have often demonstrated substantial uncertainty in selecting the best SAR model for a given dataset (Stiles and Scheiner 2007, Guilhaumon et al. 2008). The multimodel selection framework (Burnham and Anderson 2002) is an approach that can account for such uncertainties in inferring the SAR, allowing the investigator to perform inferences while incorporating variability in both model selection and parameter estimation (multimodel SARs; Guilhaumon et al. 2008).

Here, we introduce the mmSAR R-package for the freeware and open-source R software (R Development Core Team 2009). mmSAR is a flexible and scalable implementation of the multimodel SAR framework for species–area datasets and provides several functionalities: fitting several relevant SAR models, performing a selection among this set of models, averaging the prediction of the SAR obtained from different models to establish a consensual inference and to provide robust confidence intervals. The present software note describes the different components of the multimodel SAR framework, as well as their implementation in the mmSAR R-package (Fig. 1). We illustrate the framework with the results of an analysis of a species–area dataset for the plants of the Galapagos Islands (Preston 1962). The users interested in the methodological details of the multimodel SAR framework are referred to Guilhaumon et al. (2008).

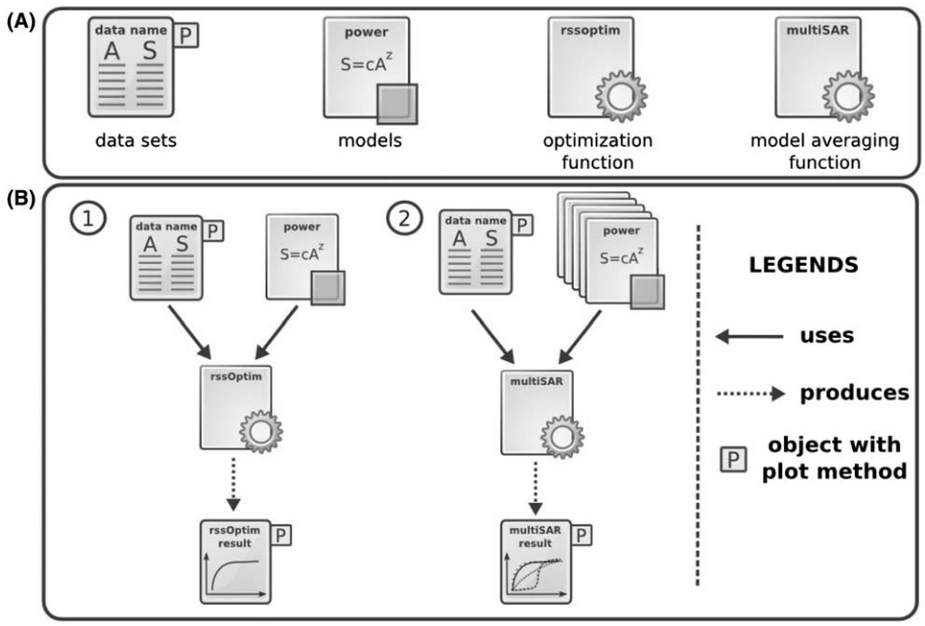

Figure 1. Main components of mmSAR (A) and sample "use cases" (B). B1 simple nonlinear SAR model fitting. B2 multimodel SAR calculation.

The multimodel SAR framework

The components of the mmSAR implementation of the multimodel SAR framework are presented in Fig. 1. Apart from the species–area dataset itself, mmSAR provides R objects to handle SAR models, facilitating the fit of SAR models through non-linear regression and the construction of consensual prediction for the SAR with associated confidence intervals (Fig. 1A). Different applications can be envisaged with the mmSAR components, from simple model fitting to selection and average across sets of models (Fig. 1B).

Models

For a given dataset, a multimodel SAR inference is made simultaneously using the predictions of several non-linear regression models. Obtaining a consistent set of models is one of the most important challenges in information-theoretic analyses (Burnham and Anderson 2002). mmSAR proposes a comprehensive set of SAR models (Table 1), including five convex models (power, exponential, negative exponential, Monod and rational function) and three sigmoid models (logistic, Lomolino, and cumulative Weibull). This includes convex, sigmoid, asymptotic, and non asymptotic functions, thus encompassing the various shapes attributed to SARs in the literature. Note that the linearized forms (via logarithmic transformations) of the power and exponential models, which require using $\log(S)$ in place of S, were not implemented in mmSAR, otherwise precluding comparisons across the entire set of models. In mmSAR, models are implemented as R objects and new non linear SAR models should easily be specified by the user and added to the available collection.

Model fitting

mmSAR performs nonlinear regressions to obtain model parameter estimates by minimizing the residual sum of squares with an unconstrained Nelder–Mead optimization algorithm. Assuming normality of the observations, this approach produces optimal maximum likelihood estimates of model parameters (Burnham and Anderson 2002). To avoid numerical problems, such as local minima, and speed up the convergence process, starting values used to run the optimization algorithm are carefully chosen. For directly interpretable parameters (e.g. an asymptote), corresponding values in the datasets are used (e.g. the observed maximum of species richness in the case of an asymptote), otherwise the standard procedures described by Ratkowsky (1983, 1990) are implemented. Finally, mmSAR gives the option to provide custom starting values, allowing users to implement exhaustive searches for best fits. We provide example fits of the eight SAR models implemented in mmSAR to the Galapagos Islands dataset in Fig. 2A1–A8.

Regression validation

Regressions are usually evaluated by statistical examination of normality and homoscedasticity of residuals. In mmSAR, two tests for the normality of the residuals are available: the Lilliefors extension of the Kolmogorov normality test, which is advocated when sample size is large or when the data show a substantial variability (e.g. continental scale studies) and the Shapiro–Wilk test for normality, which focuses on skewness and kurtosis of the empirical distribution of the residuals and is useful for small sample size or when data results from small scale sampling. mmSAR tests

Table 1. Functional forms for the SAR implemented in mmSAR. In these equations, S and A represent, respectively, species richness and area, while c, z, f and d are fitted parameters. The parameter d is an upper asymptote, except for the rational function for which the upper asymptote is z/d.

Name	Code	Formula	Number of parameters	Shape	Asymptotic nature
Power	Power	$S = cA^z$	2	Convex	No
Exponential	Expo	$S = c + z\log(A)$	2	Convex	No
Negative exponential	Negexpo	$S = d(1 - \exp(-zA))$	2	Convex	Yes
Monod	Monod	$S = d/(1 + cA^{-1})$	2	Convex	Yes
Rational function	Ratio	$S = (c + zA)/(1 + dA)$	3	Convex	Yes
Logistic	Logist	$S = d/(1 + \exp(-zA + f))$	3	Sigmoid	Yes
Lomolino	Lomolino	$S = d/1 + (z^{\log(f/A)})$	3	Sigmoid	Yes
Cumulative Weibull	Weibull	$S = d(1 - \exp(-zA^f))$	3	Sigmoid	Yes

for homoscedasticity by evaluating the correlation between residual magnitude and areas or fitted values (Pearson's product moment correlation coefficient). Generally, a model is considered not to be valid for a given dataset if one of the tests of normality or homoscedasticity is significant at the 5% level.

Model selection

The information-theoretic framework for model-selection is based on the evaluation of multiple working hypotheses (Burnham and Anderson 2002). This evaluation of competing hypotheses, which are each represented by a different model, is achieved through the estimation, for each, of the probability to be the best in explaining the data. In mmSAR, these probabilities are materialized by Akaike weights (Burnham and Anderson 2002) derived from information criteria (IC) such as the Akaike information criterion (AIC) or its correction for small sample bias (AICc) and the Bayesian information criterion (BIC). AIC and other model selection criteria that estimate Kullback–Leibler information are used widely in the ecological literature, but other criteria such as the BIC are also commonly used to carry out model selection (see Burnham and Anderson 2002 for a review of model selection and multimodel inference). AIC and BIC do not share the same conceptual bases and penalize differently for the dimension of the models (BIC tends to select models with fewer parameters than AIC), and although the results of (mm)SAR analyses are generally robust as regards the criterion used for model selection (Guilhaumon et al.

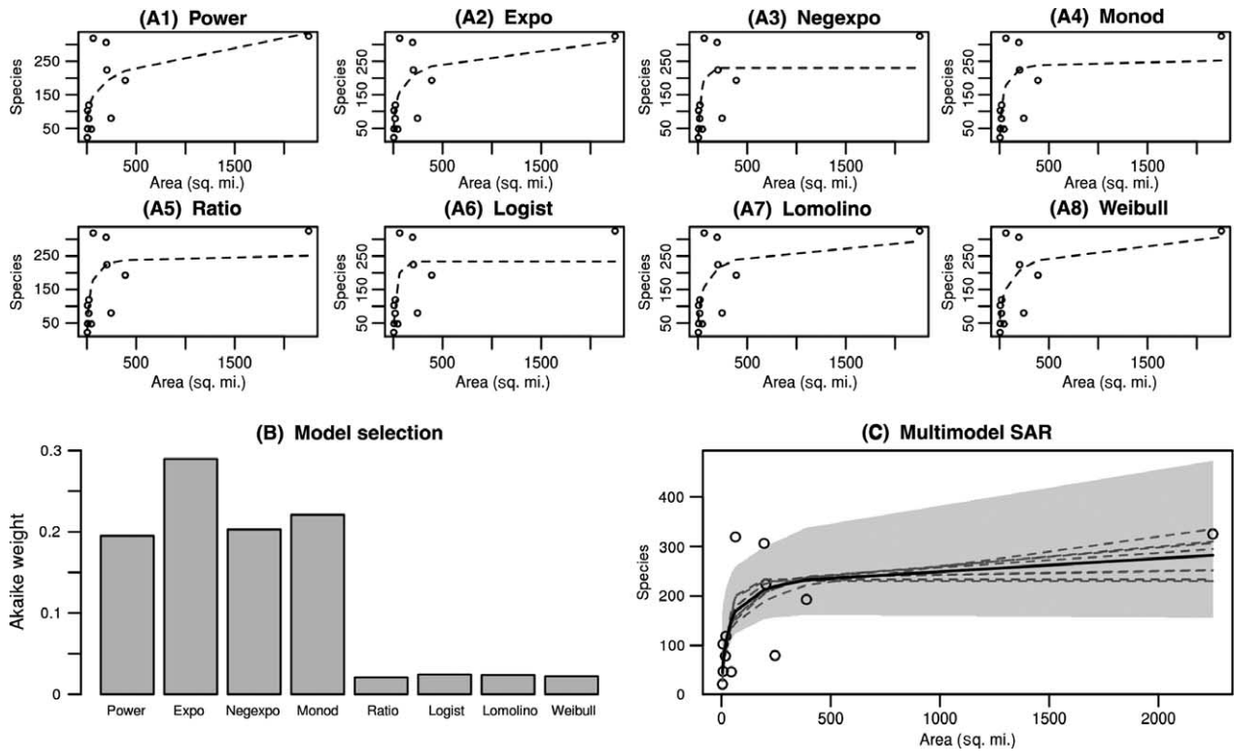

Figure 2. mmSAR results for a species–area dataset describing the plants of the Galapagos Islands (Preston 1962). (A1–A8) Fit of the eight SAR functional forms implemented in mmSAR (see Table 1 for equation descriptions). (B) Results of a model selection procedure (the eight of a bar indicates the probability (i.e. Akaike weights derived from the AICc criterion in this example) of the model being the best in fitting the data). (C) Model fits (dashed lines), multimodel SAR (black line) and associated non parametric confidence interval (grey shading).

2008), mmSAR implements both Kullback–Leibler and Bayesian strategies for model selection. For a fitted model i, its weight w_i is given by:

$$w_i = \frac{e^{-1/2\Delta_i}}{\sum_{r=1}^{M} e^{-1/2\Delta_r}} \quad (1)$$

where M is the number of models in the set and Δi is defined as $\Delta_i = IC_i - IC_{min}$ with $ICmin$ the IC value for the best model.

Akaike weights are a straightforward means of interpreting the IC values of each model, as model likelihood, and provide the basis of multimodel inference. For the Galapagos Islands data set, the best fitting model was exponential but three others models (power, negative exponential, and Monod) had almost equivalent probabilities in explaining the data (AICc Akaike weights in Fig. 2B). The four remaining models (rational function, logistic, Lomolino, and cumulative Weibull) have negligible likelihood and should contribute only marginally to the multimodel SAR (AICc Akaike weights in Fig. 2B).

Model averaging and confidence interval building

In the model selection framework, model selection uncertainty arises when the dataset support several models with a similar strength (i.e. for a given dataset, no w_i is higher than 0.9; Burnham and Anderson 2002), as this is the case with the data from the Galapagos Islands (Fig. 2B). In such cases, it is not adequate to rely exclusively on the best model only; multimodel inference can construct a more robust final inference (Burnham and Anderson 2002). As advocated for differently parameterized models, mmSAR implements model averaging and considers the weighted average of all valid model predictions (see Regression validation), with respect to model weights, to construct multimodel SARs:

$$\bar{\hat{S}} = \sum_{i=1}^{M} \hat{S}_i w_i \quad (2)$$

where $\bar{\hat{S}}$ is the multi-model averaged species richness and $\hat{S}_i$ is the species richness inferred from model i, M is the number of valid models. The multimodel SAR for the Galapagos Islands data set is presented in Fig. 2C.

Finally, in mmSAR, confidence intervals incorporating uncertainty regarding both model selection and parameter estimation can be constructed using the percentile method and a non-parametric bootstrap scheme (Efron 1979, Buckland et al. 1997). For a given species–area dataset, a large number of bootstrap samples are obtained in the following manner: 1) one of the SAR models included in the analysis is selected with a probability equal to its weight as calculated from eq. 1. 2) The selected model is fitted to the observed dataset under study. 3) The vectors of inferred species richness (regression line) and residuals are obtained from the regression and the residuals are standardized. 4) The residuals are sampled with replacement until sample size reaches that of the dataset, to form a vector of modified residuals. 5) The vector of modified residuals is added to the vector of inferred species richness, to form the resample (bootstrap set of pseudo responses).

A collection of multi-model SARs inferred from each of the resamples is gathered by applying the whole procedure of model selection and averaging, while the bootstrap estimates of species richness are sorted in ascending order to provide the percentile confidence intervals (Buckland et al. 1997): the limits of an approximate $(1-\alpha)100\%$ confidence interval are given by picking the rth and sth values in the ordered vector of bootstrap estimates, such that $r=(b+1)\alpha$ and $s=(b+1)(1-\alpha)$.

For the Galapagos Islands dataset, the number of resamples was fixed to 9999, thus the limits of the 95% confidence interval for a point estimate of species richness (Fig. 2C) are given by the 250th and the 9750th values.

The mmSAR R-package may have potential uses in both theoretical and conservation analyses. For example, in theoretical applications such as investigations about how SARs may differ among different systems, model selection patterns (i.e. relative likelihoods of different SAR shapes) can be compared for the different systems. Allowing one, for example, to state about the saturation or non saturation of species richness with increasing area. These kind of analyses may help to extend discussions beyond the comparison of slopes of log-linear power SARs (Guilhaumon et al. 2008). In conservation applications, multimodel non-parametric confidence intervals can inform about the reliability of the multimodel SAR for a given dataset but also have more practical applications. For example, these confidence intervals were used by Guilhaumon et al. 2008 to rank regions of a dataset with respect to their biological richness. By positioning the observed richness of each region in the associated vectors of ordered bootstrap species richness estimates (the higher the position of the observed species richness in the vector of bootstrap estimates the higher the ecoregion in the ranking), these authors were able to devise a hotspot ranking methodology that was robust to the underlying form of SARs.

The mmSAR R-package and a detailed user's guide is available at the R-Forge website <http://mmsar.r-forge.r-project.org>.

To cite mmSAR or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for "Version 0":

Guilhaumon, F., Mouillot, D. and Gimenez, O. 2010. mmSAR: an R-package for multimodel species–area relationship inference. – Ecography 33: 420–424 (Version 0).

Acknowledgements – We thank two anonymous reviewers for comments or helpful discussions and are grateful to David Mckenzie for editing and correcting the language.

References

Arrhennius, O. 1921. Species and area. – J. Ecol. 9: 95–99.
Brooks, T. M. et al. 2006. Global biodiversity conservation priorities. – Science 313: 58–61.
Buckland, S. T. et al. 1997. Model selection: an integral part of inference. – Biometrics 53: 603–618.

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multimodel inference: a practical information-theoretic approach. – Springer.
Coleman, B. D. et al. 1982. Randomness, area and species richness. – Ecology 63: 1121–1133.
Connor, E. F. and McCoy, E. D. 1979. The statistics and biology of the species–area relationship. – Am. Nat. 113: 791–833.
De Candolle, A. 1855. Géographie botanique raisonnée; ou exposition des faits principaux et des lois concernant la distribution géographique des plantes de l'époque actuelle. – Maisson.
Drakare, S. et al. 2006. The imprint of the geographical, evolutionary and ecological context on species–area relationships. – Ecol. Lett. 9: 215–227.
Efron, B. 1979. Bootstrap methods: an other look at the jackknife. – Ann. Stat. 7: 1–26.
Flather, C. H. 1996. Fitting species–accumulation functions and assessing regional land use impacts on avian diversity. – J. Biogeogr. 23: 155–168.
Gleason, H. A. 1922. On the relation between species and area. – Ecology 3: 158–162.
Guilhaumon, F. et al. 2008. Taxonomic and regional uncertainty in species–area relationships and the identification of richness hotspots. – Proc. Nat. Acad. Sci. USA 105: 15458–15463.
He, F. L. and Legendre, P. 2002. Species diversity patterns derived from species–area models. – Ecology 83: 1185–1198.
Lamoreux, J. F. et al. 2006. Global tests of biodiversity concordance and the importance of endemism. – Nature 440: 212–214.
Lomolino, M. V. 2000. Ecology's most general, yet protean pattern: the species–area relationship. – J. Biogeogr. 27: 17–26.
MacArthur, R. H. and Wilson, E. O. 1967. The theory of island biogeography. – Princeton Univ. Press.
Martin, H. G. and Goldenfeld, N. 2006. On the origin and robustness of power-law species–area relationships in ecology. – Proc. Nat. Acad. Sci. USA 103: 10310–10315.
Preston, F. W. 1962. The canonical distribution of commonness and rarity: part I. – Ecology 43: 185–215.
R Development Core Team 2009. R: a language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria, <www.R-project.org>.
Ratkowsky, D. A. 1983. Nonlinear regression modelling. A unified practical approach. – Dekker.
Ratkowsky, D. A. 1990. Handbook of nonlinear regression models. – Dekker.
Rosenzweig, M. L. 1995. Species diversity in space and time. – Cambridge Univ. Press.
Šizling, A. and Storch, D. 2004. Power-law species–area relationships and self-similar species distributions within finite areas. – Ecol. Lett. 7: 60–68.
Smith, A. B. 2010. Caution with curves: caveats for using the species–area relationship in conservation. – Biol. Conserv. 143: 555–564.
Southwood, T. R. E. et al. 2006. Observations on related ecological exponents. – Proc. Nat. Acad. Sci. USA 103: 6931–6933.
Stiles, A. and 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.
Tjørve, E. 2003. Shapes and functions of species–area curves: a review of possible models. – J. Biogeogr. 30: 827–835.
Tjørve, E. 2009. Shapes and functions of species–area curves (II): a review of new models and parameterizations. – J. Biogeogr. 36: 1435–1445.
Wilson, K. A. et al. 2007. Conserving biodiversity efficiently: what to do, where, and when. – PLoS Biol. 5: 1850–1861.