Journal of Vegetation Science **25** (2014) 1148–1153

SPECIAL FEATURE: VEGETATION PATTERNS AND THEIR UNDERLYING PROCESSES

Mathematical contributions to link biota with environment

J.O. Cerdeira, T. Monteiro-Henriques, M.J. Martins, P.C. Silva, D. Alagador & A.M.A. Franco

Keywords
BIOCLIM; Combinatorics; Convex hull; Data depth; HABITAT; Niche; Overlap; Suitability modelling

Received 30 September 2013
Accepted 12 March 2014
Co-ordinating Editor: Martin Zobel

Cerdeira, J.O. (corresponding author, jo.cerdeira@fct.unl.pt): Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516 Caparica, Portugal and Centro de Estudos Florestais, Universidade de Lisboa, Tapada da Ajuda, 1349-017, Lisboa, Portugal
Monteiro-Henriques, T. (tmh@isa.ulisboa.pt): Centro de Estudos Florestais, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal
Martins, M.J. (mjmartins@isa.ulisboa.pt) & **Silva, P.C.** (pcsilva@isa.ulisboa.pt): Centro de Estudos Florestais, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal
Alagador, D. (alagador@uevora.pt): CIBIO/InBio-UE, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade de Évora, 7000-890 Évora, Portugal
Franco, A.M.A. (A.Franco@uea.ac.uk): School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK

Abstract

Hutchinson's pioneering work on the niche concept, dating from 1957, inspired the development of many ecological models. The first proposals, BIOCLIM and HABITAT, were simple geometric approximations to the shape of the niche. Despite their simplicity, they combine two features that make them adequate for the purpose of exploring the niche: they fit a predefined shape to the empirical data; and produce binary or ordinal predictions rather than continuous predictions. Thus, both explicitly delineate a precise boundary for the niche. However, the two methods present some limitations: BIOCLIM assumes that the variables are independent in their action on the species; and HABITAT, although not having that limitation, only delineates the boundaries of the niches without distinguishing levels of suitability for the species. We propose, discuss and illustrate: (1) the use of depth functions to identify regions with distinct suitability inside the niche; and (2) a general framework to assess overlap of the niches of two species, which can be applied to predictions from models that decompose the niche into a finite number of measurable regions.

Introduction

In 1957 Hutchinson formalized the niche of species as 'an *n*-dimensional hypervolume where every point in it corresponds to a state of the environment which would permit the species to exist indefinitely' (Hutchinson 1957, p. 416). Hutchinson called this hypervolume in the space defined by *n* environmental variables, the *fundamental niche*, which bounds the species physiological limits on each variable. Nowadays there is a proliferation of methods that assess the adequacy of ecological conditions to support species, using the Hutchinson's environmental hyperspace.

Some methods, such as BIOCLIM (Nix 1986; Busby 1991) and HABITAT (Walker & Cocks 1991), have the main purpose of delineating the niche boundary, fitting a predefined geometric shape to the sampling set of species occurrences, in the environmental hyperspace. In this article, we focus on these kinds of delineation-oriented models, which produce binary or ordinal predictions, and therefore a finite number of iso-suitability regions.

Doi: 10.1111/jvs.12188

We start by (1) pointing out some mathematical aspects underlying Hutchinson's niche concept; then we (2) analyse the extent by which the geometric approaches BIOCLIM and HABITAT capture some properties behind Hutchinson's niche, and indicate their limitations; (3) present a mathematical concept, *data depth*, and show how this concept can be used (4) to resolve limitations of the geometric models; and (5) to assess niche overlap.

Some considerations on Hutchinson's niche concept

Hutchinson (1957) clearly states that the fundamental niche is a region bounded by 'the limiting values (of each variable) permitting a species to survive and reproduce'. When represented in a space of variables, this implies that the fundamental niche is contained in the hyperrectangle defined by the limiting values (minimum and maximum) of each of the n variables. Hutchinson also states that, 'if the variables are independent in their action on the species', the fundamental niche is the above hyperrectangle, 'each point of which corresponds to a possible environmental state permitting the species to exist indefinitely'. In particular, this implies, in the case of independence, that the fundamental niche defines a *convex region*. A set of points is convex if every point on the straight line segment joining two points in the set is also in that set. In this context, convexity reads as: if the species exists indefinitely in two environmental states, it also exists in any state that is a weighted average (where the non-negative weights sum to 1) of the two states.

It is unrealistic to assume that all variables act independently on a species. For instance, many species only exist at their maximum tolerated temperatures if humidity is higher and far from the humidity lower toleration point. Indeed, it can be inferred from Hutchinson's words that the effects of variable dependencies over the niche's shape are not straightforward to find, although 'the area (niche) will exist whatever the shape of its sides'. Therefore, it is reasonable to assume that, in the case of dependencies, peripheral regions of the hyperrectangle could be excluded, as representing combinations of variables not tolerated by the species.

Hutchinson was also concerned about how species' interactions and intersection of their niches relate. Besides the quantification of this intersection, also its location within each niche seems to be relevant for species competition, co-occurrence and other ecological phenomena. This implicitly assumes that there are distinct parts of the fundamental niche. Indeed, Hutchinson's original niche concept assumes 'that all points in each fundamental niche imply equal probability of persistence of the species', but he discusses this property as a strong limitation given that 'ordinarily there will however be an optimal part of the niche with markedly suboptimal conditions near the boundaries'. This has since been reinforced by Pianka (2000) when referring to fitness gradients inside the niche space. That is, species find distinct levels of suitable conditions inside their tolerance limits, and optimal conditions are deemed to occur in the most interior regions.

Geometric approaches to Hutchinson's niche

Geometric procedures (i.e. approaches that only rely on topological relationships, not depending on metric properties), such as BIOCLIM and HABITAT, enclose some features that make them particularly adequate for the purpose of exploring the niche: (1) they fit a predefined shape, explicitly delineating a precise boundary for the niche (a hyperrectangle in the case of BIOCLIM and a convex hull in the HABITAT procedure); (2) they produce binary or ordinal predictions; and (3) in the case of BIOCLIM, it decomposes the niche into a finite number of regions, with non-null volumes, of environments (i.e. non-*negligible* sets of environments) with similar effects on species. Other methods, which are more prediction-oriented (such as GLM, GAM, Maxent, Mahalanobis distance, DOMAIN, ENFA, among others) produce continuous predictions and, therefore, depend on thresholds to delineate a niche boundary (see Elith et al. 2006; Tsoar et al. 2007). Some methods, such as GLM, GAM, Maxent, Genetic Algorithm for Rule-set Production (GARP) and artificial neural networks, even after applying user-defined suitability/probability thresholds, may produce unlimited hypervolumes. These are quite distinct features, compared to the delineation-oriented models referred above, where unbounded regions do not occur.

BIOCLIM estimates the niche as the bounding hyperrectangle enclosing all records of a species in the n-dimensional environmental space, creating a rectilinear envelope defined by the most extreme values of each variable in the set of the occurrences. It thus assumes independence of the variables in their effect on species (see Fig. 1a).

Different environments presumably affect differently the species performance (i.e. fitness, growth rate, intake capacity, etc.) and, as mentioned above, species are expected to perform optimally in environmental conditions closer to the innermost parts of the niche. BIOCLIM discriminates the niche accordingly, identifying nested hyperrectangles of environments with increasing levels of suitability. This is geometrically achieved using the percentiles of the environmental values recorded on each occurrence point. More specifically, a point in the n-dimensional environmental space has predicted suitability $2a$ if the coordinate with the minimum percentile is a $100a$ or $100(1-a)$ percentile, considering the values of the corresponding variable for all the occurrence points. A

Doi: 10.1111/jvs.12188

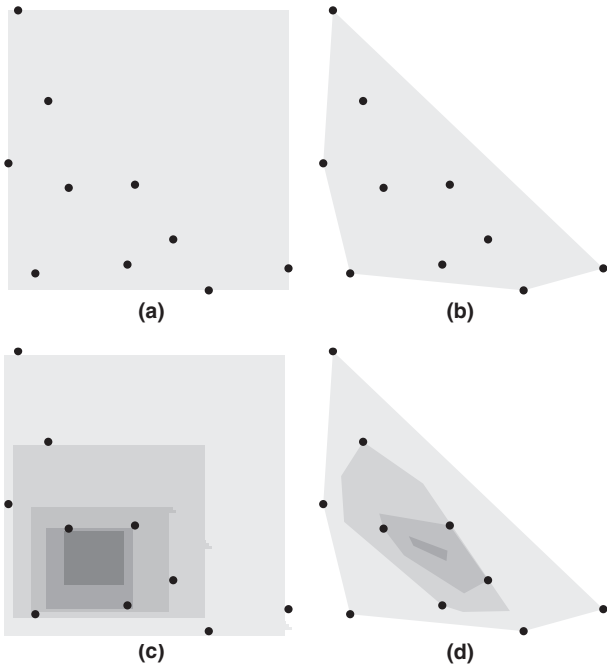

Fig. 1. Geometric representations of species niche. Points are presences in a two-dimensional environmental space. Grey areas define the niche. (**a**) outer rectangle obtained by BIOCLIM; (**b**) convex hull used by HABITAT; (**c**) nested regions of increasing levels of suitability obtained by BIOCLIM with percentile; (**d**) nested regions of increasing levels of suitability obtained by the convex hull with Tukey depth.

point in the region with higher suitability (the median region) has all coordinates in the 50-percentile. Figure 1c depicts the six different regions of iso-suitability obtained by BIOCLIM with percentiles procedure, in respect to the set of ten points (occurrences) represented in the four panels of Fig. 1.

If R and S are consecutive nested hyperrectangles, with $R \supset S$, the points in $R \setminus S$ are environmental states considered similar with respect to species performance. Inner hyperrectangles are expected to include more favourable environmental states for the species.

It is noteworthy that the process used by BIOCLIM to discriminate suitability regions within the niche reinforces the convexity assumption, leading to the following strongest assumption: if two environments e and e' have quantified levels of suitability $s(e)$ and $s(e')$, respectively, and if e'' is any environment on the straight line segment joining e to e', then the suitability is $s(e'') \geq min\{s(e), s(e')\}$. We call this assumption '*reinforced convexity*' on the niche. In other words, it states that any environment that is a weighted average (where the weights are non-negative and sum to 1) of two environmental states, is at least as suitable as the less suitable of the two. Reinforced convexity is not exclusive from this particular approach. Actually, any given collection of nested convex regions of the environmental space, combined with any function that assigns the same values to points in $R \setminus S$, where $R \supset S$ are two consecutive nested regions, with values increasing when moving to innermost sets, satisfy the reinforced convexity assumption.

The HABITAT procedure outlines the niche as the convex hull of all the records of the species in the n-dimensional environmental space. Even if HABITAT is a multistep procedure, hereafter we consider HABITAT as the step that outlines the niche as a convex hull. The convex hull of a set of points is the smallest convex region containing all points (see Fig. 1b). For HABITAT, the niche is tighter to the occurrence points when compared with the hyperrectangle. HABITAT does not assume independence of variables for their action on the species.

Although the convex hull seems to be more realistic (assuming a representative sampling) than the hyperrectangle to delineate niche boundaries, no consistent geometric procedure, such as the BIOCLIM with percentile approach, has been proposed to distinguish suitable regions inside it. However, this can be achieved using a mathematical concept called *data depth* that is a multivariate analogue of univariate order statistics.

Data depth

In statistics, several measures have been introduced as generalizations of the median and percentiles to dimensions >1. The motivation for these generalizations came from the need for robust measures of central location in multivariate data, given that the mean is highly sensitive to extreme observations. Such measures are generally called *data depth*. Fukuda & Rosta (2005) provide a unified framework for the main data depth measures.

A depth function is a process to measure the centrality of a point within a data cloud on a multi-dimensional space. Each function determines a particular centre-outward ranking of points within a given multivariate data set. A depth function is any function that satisfies certain postulates, including affine invariance (i.e. depth does not change under linear transformations of data, thus, in particular, it is invariant on scale of variables), and monotone on rays (i.e. depth monotonically decreases when moving from a point of maximum depth along a straight line). Interestingly, it is worth noting that BIOCLIM with percentiles verifies all postulates that define a depth function.

Tukey depth (Tukey 1975) is a prominent example of depth function, particularly suitable to the purpose of discriminating regions within the convex hull.

The *Tukey depth* of a point x with respect to a set X of k points in R^n is m/M, where m is the minimum number of points that have to be removed from X such that x is no

Doi: 10.1111/jvs.12188

longer in the convex hull of the remaining points of X, and M is the maximum value of m, which is the largest integer not greater than $(k-n+1)/2$. Tukey depth ranges from 0 to 1. In particular, points outside the convex hull of X have 0 Tukey depth, whereas points located in the innermost part of the convex hull of X, called *Tukey median* region, score 1. Figure 1d represents five nested regions of increasing Tukey depths, with respect to the set of ten points.

Convex hull with Tukey depth

Tukey depth is the natural depth function to incorporate in HABITAT in order to discriminate inside the niche, and we propose to score suitability using depth values with respect to the set X of occurrences. Thus, the set of points with positive suitability is the convex hull of X, which is the niche defined by HABITAT.

The use of Tukey depth within the convex hull is an analogue for the percentile procedure used in BIOCLIM. Indeed, similar to what happens in BIOCLIM, Tukey depth enables us to discriminate geometrically, decomposing the environmental space into a finite number of regions with similar suitability values (with non-zero volumes), and satisfies the reinforced convexity property.

The main difference between the two procedures is that, while in BIOCLIM with percentiles the region R_s of suitability greater than or equal to s is a rectangle with the sides parallel to the axis, assuming independence of the effects of each variable on the species; in convex hull with Tukey depth in this region is tighter. Thus, R_s contains the corresponding region obtained by convex hull with Tukey depth.

These approaches based on the median and percentiles, or their multivariate generalizations, have several interesting properties: (1) only use presences and are invariant to background; (2) are invariant to scale, as a result of being obtained from depth functions; (3) are robust to outliers exactly in the same way as univariate percentiles are (while the outer regions are very sensitive to outliers, their influence vanishes for the interior); and (4) keep a tight relation with the niche concept assuring high interpretability of results. However, it should be noticed that these approaches assume that maximum suitability occurs in a unique central region of the hypervolume, which may not always be the case. The approach of Silva et al. (2014) can be used to identify configurations for which the methods should not be used.

Niche overlap

Several indices have been proposed to assess the niche overlap between two species using raw presence/absence data or using predictions from an environmental niche model (see Warren et al. 2008 for a survey). Applications of the same indices can be extended for comparisons of the niches of two populations of a same species (native vs invasive or geographically distant populations) or for assessments of niche evolution, by evaluating the similarity of species niche at different time periods. For a recently proposed statistical framework, see Broennimann et al. (2012).

Most of these indices are different ways of comparing vectors, where each vector refers to the suitability of the ecological conditions of each cell for a particular species, assuming a discretization of the geographical space in a set of cells. An important drawback in some of these approaches is that the suitability of the environments (corresponding to these cells) is compared disregarding the position that the environments occupy within the niches. For instance, Sørensen similarity index (Sørensen 1948) and Schoener's statistic for index overlap (Schoener 1968) only account for the absolute values of differences d between the suitability of two species at every cell, e.g. the differences of suitability $d = 1.0–0.7$ and $d = 0.4–0.1$, are equally accounted as 0.3. Some proposals distinguish these situations using algebraic manipulations such that differences between the higher suitability values account for less than differences between the low suitability values.

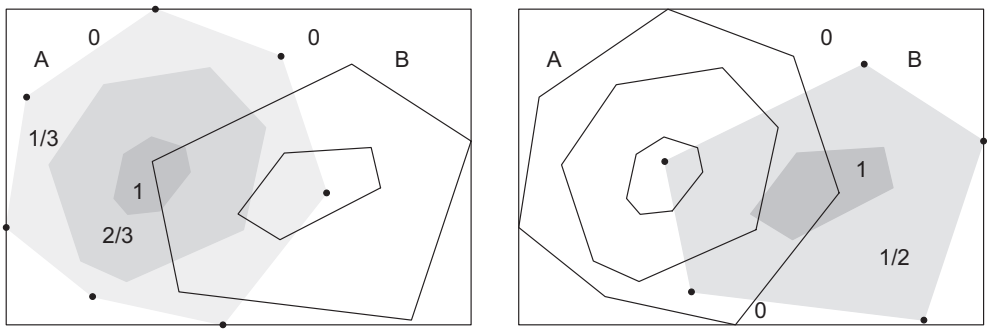

Fig. 2. Overlap of niches, for species A and B, estimated by convex hull with Tukey depth. Left (right) panel highlights the regions of iso-suitability based on seven (five) occurrences for species A (B).

Doi: 10.1111/jvs.12188

For instance, the similarity statistic index introduced by Warren et al. (2008) uses a square root function to achieve this goal (e.g. the 'modified difference' between suitability 1.0 and 0.7 is $(\sqrt{1.0}-\sqrt{0.7})^2 = 0.027$, while the 'modified difference' between 0.4 and 0.1 is $(\sqrt{0.4}-\sqrt{0.1})^2 = 0.1$). These approaches are formulae refinements with no clear ecological explanation. Moreover, indices that are based on comparisons of vectors of suitability only account for a finite number of points on the environmental space, instead of considering the predictions for the whole environmental space.

We next describe an approach for niche overlap estimation that takes into account suitability of every point of the environmental space, and that has a straightforward ecological interpretation. The procedure applies, but is not limited, to the outputs of BIOCLIM or HABITAT incorporating the depth functions described in the previous sections. It can be applied whenever there is a finite number of measurable iso-suitability regions, such as the regions arising from binary or ordinal predictions.

Consider two species A and B and the partitions of the environmental space into k_A and k_B finite sets of predicted iso-suitability for each species. Compute, for every pair (i,j) of predicted suitability i for species A and j for species B, the volume $V_{i,j}$ of the (possibly disconnected) region composed by all the environments having suitability i for species A and j for species B. Let $M_V = [V_{i,j}]$ be the $k_A \times k_B$ matrix with the volumes for all these regions. Figure 2 illustrates the regions of iso-suitability of the two species, obtained from the convex hull with Tukey depth based on seven occurrences for species A and five occurrences for species B. The corresponding matrix of volumes M_V is presented in Table 1.

Matrix M_V can be viewed as an encoding of the intersection pattern of A and B niches. Together with its row and column indices, which are in fact the distinct levels of suitability for each of the two species, enclose all the information given by the predictive model to assess niche overlap.

From M_V we define the matrix M_A (M_B) of the asymmetric *overlap of species* B (A) on A (B), multiplying the rows (columns) of M_V by its row (column) index vector (see Table 1). Matrix M_A (M_B) contains, for every region with iso-suitability (i,j), the volume $V_{i,j}$ weighted by i (j).

Matrices M_V, M_A and M_B can be related in some convenient mathematical expression to quantify niche overlap of species A and B. A possibility is

$$wJ = \frac{\sum_{i>0}\sum_{j>0}\left(M_A(i,j)+M_B(i,j)\right)}{2\sum_i\sum_j\left(M_A(i,j)+M_B(i,j)\right)-\sum_{i>0}\sum_{j>0}\left(M_A(i,j)+M_B(i,j)\right)}$$

Table 1. Matrices of volumes for the data in Fig. 2. Entry (i,j) of M_V is the volume of the intersection of regions with suitability i for species A and j for species B. ($M_V(0,0)$ is arbitrary, here total volume = volume of the bounding box is 100). Matrix M_V^b is similar to M_V but considering a binary response. Note that the volume of the union of the convex hulls is (100–31.046), while the volume of the intersection is 15.283. Matrix M_A (M_B) is the matrix of volumes weighted by the suitability for species A (B).

M_V

$i\backslash j$	0	0.5	1
0	31.046	20.037	2.235
1/3	16.459	5.401	2.98
2/3	13.058	5.95	0.069
1	1.882	0.882	0

M_V^b

$i\backslash j$	0	1
0	31.046	22.272
1	31.399	15.283

M_A

$i\backslash j$	0	0.5	1
0	0	0	0
1/3	5.486	1.8	0.993
2/3	8.705	3.967	0.046
1	1.882	0.882	0

M_B

$i\backslash j$	0	0.5	1
0	0	10.019	2.235
1/3	0	2.701	2.98
2/3	0	2.975	0.069
1	0	0.441	0

which can be written as the ratio of weighted volumes,

$$wJ = \frac{\sum_{i>0}\sum_{j>0}(i+j)\times M_V(i,j)}{2\sum_i\sum_j(i+j)\times M_V(i,j)-\sum_{i>0}\sum_{j>0}(i+j)\times M_V(i,j)}.$$

It is noteworthy that if suitability is binary predicted (0/1), M_V is a square order 2 matrix, as M_V^b binary matrix in Table 1, and wJ becomes

$$\text{Jaccard} = \frac{M_V^b(1,1)}{M_V^b(0,1)+M_V^b(1,0)+M_V^b(1,1)}$$

which is the Jaccard index for the volume measure, as the numerator is the volume of the intersection of the two convex hulls and the denominator is the volume of the union. We call wJ the weighted Jaccard overlap index.

Another possibility of a different nature is to consider the matrices M_A and M_B as vectors of length $k_A \times k_B$ and to compute the cosine of the angle between the two vectors, i.e.

Table 2. Overlap indices for the pattern represented in Fig. 2. *wJ* refers to the weighted Jaccard volumes index; cos refers to the cosine index.

wJ	Jaccard	cos	Pianka
0.2293	0.2216	0.1535	0.3192

Doi: 10.1111/jvs.12188

$$\cos = \frac{\sum_i \sum_j \left(M_A(i,j) \times M_B(i,j) \right)}{\sqrt{\sum_i \sum_j \left(M_A(i,j) \right)^2} \sqrt{\sum_i \sum_j \left(M_B(i,j) \right)^2}} = \frac{\sum_i \sum_j ij \left(M_V(i,j) \right)^2}{\sqrt{\sum_i \sum_j i^2 \left(M_V(i,j) \right)^2} \sqrt{\sum_i \sum_j j^2 \left(M_V(i,j) \right)^2}}$$

Notice that, when discarding $M_V(i,j)$ in the above expression (i.e. making it constant), cos becomes the cosine of the angle between the vectors of suitability for species A and B, which is Pianka overlap index (Pianka 1973).

The values for *wJ*, Jaccard, cos and Pianka indices corresponding to the regions of iso-suitability of two species represented in Fig. 2, are given in Table 2.

Both *wJ* and cos range between 0 (no overlap) and 1 (total overlap). While *wJ* relates volumes of iso-suitability regions weighted by the suitability of species A and B, cos measures the similarity of the two vectors regardless the magnitude of their components.

Conclusion

Combinatorial mathematical tools such as data depth can be used to improve geometric procedures to explore and interpret the niche. Moreover, these tools define a finite number of measurable iso-suitability regions that allow us to consistently evaluate niche overlap, bringing a closer link between ecology and modelling.

Acknowledgements

This work was supported by the Portuguese Foundation for Science and Technology (FCT) through the projects PEst-OE/AGR/UI0239/2011, CEF (Centro de Estudos Florestais) under FEDER/POCI and PTDC/AAC-AMB/113394/2009. D Alagador was funded by a FCT post-doctoral fellowship (SFRH.BPD.51512.2011), integrated in the European Regional Development Fund Integrated Program IC&DT N°1/SAESCTN/ALENT-07-0224-FEDER-001755.

References

Broennimann, O., Fitzpatrick, P.B., Pearman, M.C., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.-J., Randin, C., (...) & Guisan, A. 2012. Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography* 21: 481–497.

Busby, J.R. 1991. BIOCLIM – A bioclimate analysis and prediction system. In: Margules, C.R. & Austin, M.P. (eds.) *Nature conservation: cost effective biological surveys and data analysis*, pp. 64–68. CSIRO, Melbourne, AU.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., (...) & Phillips, S.J. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.

Fukuda, K. & Rosta, V. 2005. Data depth and maximum feasible subsystems. In: Avis, D., Hertz, A. & Marcotte, O. (eds.) *Graph theory and combinatorial optimization*, pp. 37–67. Springer, Berlin, DE.

Hutchinson, G.E. 1957. Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22: 415–427.

Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes. In: Longmore, R. (ed.) *Atlas of elapid snakes of Australia*, pp. 4–15. Australian Government Publishing Service, Canberra, AU.

Pianka, E.R. 1973. The structure of lizard communities. *Annual Review of Ecology and Systematics* 4: 53–74.

Pianka, E.R. 2000. *Evolutionary ecology*, 6th edn. Benjamin/Cummings - Addison Wesley Longman, San Francisco, CA, US.

Schoener, T.W. 1968. The Anolis lizards of Bimini: resource partitioning in a complex fauna. *Ecology* 49: 704–726.

Silva, P.C., Cerdeira, J.O., Martins, M.J. & Monteiro-Henriques, T. 2014. Data depth for the uniform distribution. *Environmental and Ecological Statistics* 21: 27–39.

Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter* 5: 1–34.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. & Kadmon, R. 2007. A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* 13: 397–405.

Tukey, J.W. 1975. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians* 2: 523–531.

Walker, P.A. & Cocks, K.D. 1991. HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters* 1: 108–118.

Warren, D.L., Glor, R.E. & Turelli, M. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62: 2868–2883.