Ecography

**ECOG-03187**
Blonder, B. 2017. Hypervolume concepts in niche- and trait-based ecology. – Ecography doi: 10.1111/ecog.03187

## Supplementary material

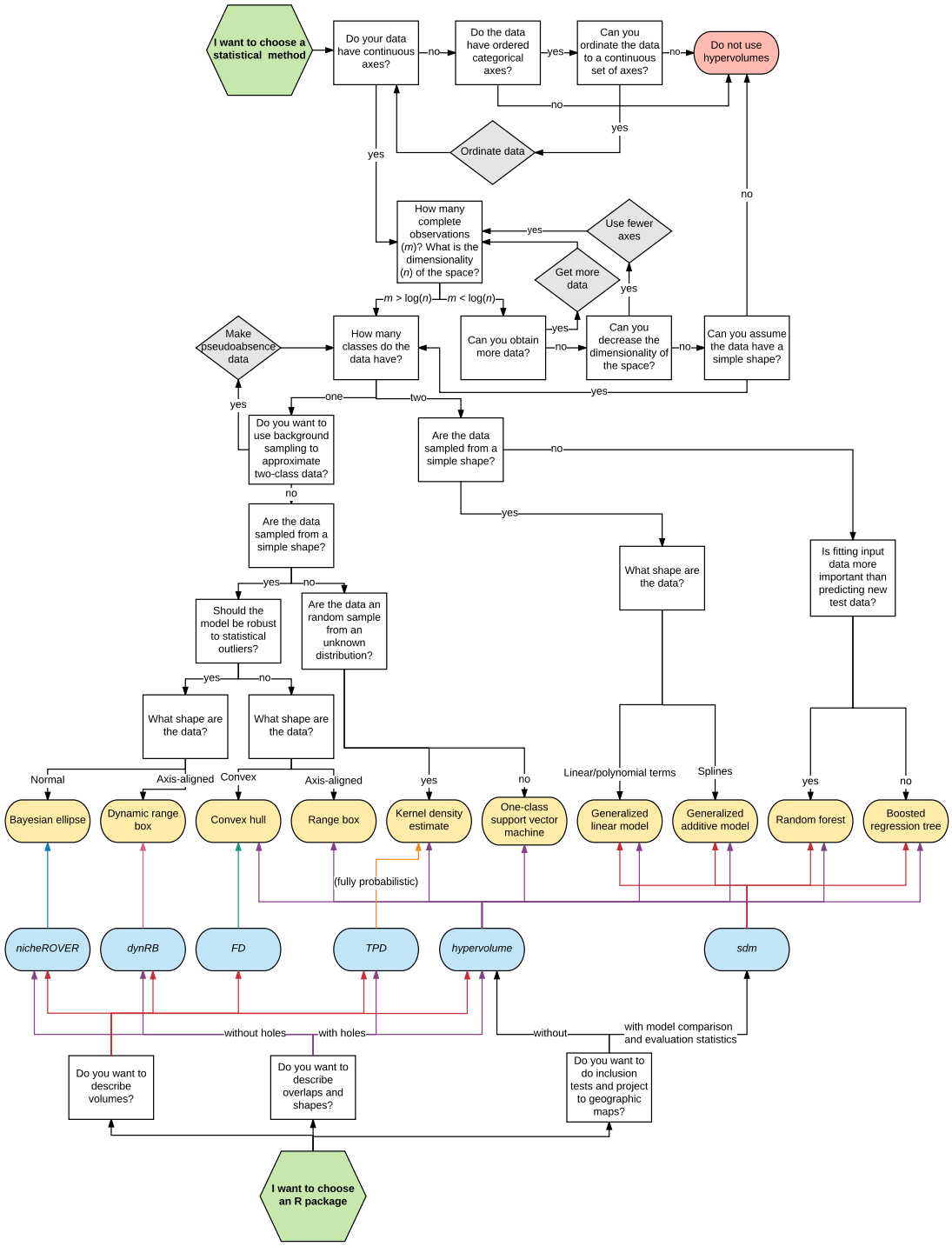
**Table A1.** Comparison of selected statistical methods. Distributional assumptions indicate the proposed structure of the data. Free parameters indicate investigator-determined parameters whose choice will influence model assumptions. Computational parameters indicate nuisance values that can influence model convergence. Invariances indicate data transformations that will not affect results. Sensitive to outliers indicates whether the method's results will be changed by inclusion of a small number of points far away from the rest of the data set. Recommended minimum sample size is my guidance on how many data points (*m*) are needed to conduct an analysis in an *n*-dimensional space. Recommended dimensionality is guidance for how large *n* can be before computational time costs become excessive.

| Biological states | Method | Distributional assumptions | Free parameters | Computational parameters | Invariances | Outlier sensitive | Recommended minimum sample size (*m*) | Recommended maximum dimensionality (*n*) |
|---|---|---|---|---|---|---|---|---|
| One-class | Bayesian ellipse | Normal | Prior distribution parameters (lambda, kappa, psi), degrees of freedom | None | Rotation, translation | No | m > log(n) | None |
| One-class | Convex hull | Convex | None | None | Rotation, translation | Yes | m > n | n < 10 |
| One-class | Range bagging | Convex | # of base learners, fraction of records per bootstrap | None | Rotation, translation | No | m > n | n < 10 |
| One-class | Dynamic range box | Rectangular | None | Number of steps | Rotation, translation, scaling | No | None | None |
| One-class | Gaussian mixture model | Normal mixture | # of mixture components | None | Translation | No | None | None |
| One-class | Density-based clustering | None | Minimum cluster size, neighborhood distance, outlier factor, maximum noise factor | None | Rotation, translation | No | None | None |
| One-class | kernel density estimation | None | Threshold, bandwidth | Number of points / grid resolution | Rotation, translation, log-transformation (sometimes) | No | m > log(n) | n < 8, (hypervolume R package ) n<4 (TPD R package) |
| One-class | one class support vector machine | None | Kernel type and degree, regularization coefficient | Tolerance parameter, loss function parameter | Rotation, translation | No | m > log(n) | None |
| One-class | range box | Rectangular | None | None | Translation | Yes | None | None |
| One-class | Malahanobis | Normal | None | None | Rotation, translation | No | None | None |
| One-class | Ecological niche factor analysis | Normal | None | None | Rotation, translation | No | None | None |
| Two-class | boosted regression tree | None | Family, learning rate, bag fraction, tree complexity | Tolerance parameter | Rotation, translation | No | m > log(n) | None |
| Two-class, quantitative | generalized additive model | None | Family | Fit algorithm, tolerance parameters, iteration parameters | Rotation, translation | No | m > log(n) | None |
| Two-class, quantitative | generalized linear model | None | None | Tolerance parameters, iteration parameters | Rotation, translation | No | m > log(n) | None |
| Two-class, quantitative | random forest | None | # of trees, # of variables sampled per split, minimum / maximum node number, cutoff parameter | None | Rotation, translation | No | m > log(n) | None |
| Two-class, quantitative | two class support vector machine | None | Kernel type and degree, regularization coefficient | Tolerance parameter, loss function parameter | Rotation, translation | No | m > log(n) | None |

**Table A2.** Comparison of software packages available to measure hypervolumes. Packages are scored for their ability to perform a range of different tasks; blank if functionality is absent, - if partially present, and + if present. Tasks include size, calculation of a metric of size; overlap, calculation of a metric of intersection between multiple hypervolumes in n-dimensional (+) or only geographic space (-); holes, modeling and detection (+) or modeling (-) of empty regions, projection/prediction, the prediction of hypervolume values at test points; fit statistics, the calculation of predictive accuracies; inference, the calculation of likelihoods or confidence intervals; visualization, the plotting of hypervolume functions.

| R package | Estimation methods | Volume | Overlap | Holes | Projection/prediction | Fit statistics | Inference | Visualization |
|---|---|---|---|---|---|---|---|---|
| adeHabitatHS | ecological niche factor analysis | | | | | | | |
| dismo | boosted regression tree, generalized additive model, generalized linear model, MaxEnt, random forest, quantile-based range box (BIOCLIM) | | - | | + | + | + | + |
| dynRB | dynamic range box | + | + | | | | | |
| FD | convex hull | + | + | | + | | | |
| hypervolume | kernel density estimation, support vector machine, convex hull, range box | + | + | + | + | | | + |
| mclust | Gaussian mixture model | | | | | + | + | + |
| nicheROVER | Bayesian ellipse | + | + | | + | + | + | + |
| sdm | boosted regression tree, generalized additive model, generalized linear model, Mahalanobis, random forest, support vector machine | | | | + | + | + | + |
| TPD | kernel density estimation | + | + | - | - | | | + |

**Figure A1.** Potential decision flowchart for hypervolume analyses. Starting points are indicated as green hexagons, key questions as white rectangles, data manipulation actions as gray diamonds, statistical methods as yellow ovals, R packages as blue ovals, and a stop point as a red oval. Other methods (**Table S1**) and packages (**Table S2**) are also useful but not shown here for visual clarity.



**Text A1. Guidelines on choosing a statistical estimation method**.

Among two-class methods, several are regression-type models using different component functions, which include generalized linear models (Austin et al. 1990), generalized additive models (Guisan et al. 2002), multivariate adaptive regression splines (Leathwick et al. 2006), and to an extent, maximum entropy (Guillera-Arroita et al. 2014). Others rely on a combination of classification algorithms and model combination approaches to generate ensemble predictions for class membership. These include boosted regression trees (Elith et al. 2008) and random forests (Breiman 2001). These methods have been surveyed and extensively compared elsewhere (Bahn and McGill 2013, Elith et al. 2006).

There are also several available one-class methods. The simplest of these methods involve fitting a simple geometric shape to the data. BIOCLIM models fit an axis-aligned box (Busby 1991), while principal component analyses and minimum volume ellipsoids (Green 1974) fit a rotated ellipse. However all these methods are sensitive to outlying points and give only binary output. These methods have been extended to cases that are robust against outliers: for example distance-based methods, which find points within a certain distance of the data centroid (e.g. Mahalanobis (Farber and Kadmon 2003) or outlying mean index (Dolédec et al. 2000) and ecological niche factor analysis (Hirzel et al. 2002)), dynamic range boxes, which are based on estimating empirical cumulative distribution functions for quantiles of ranges (Junker et al. 2016), and Bayesian probabilistic ellipses (Jackson et al. 2011, Swanson et al. 2015).

There are also one-class methods available that do not assume the data have a simple geometry. Convex hulls have been used to fit the smallest convex shape that wraps around all the data points, i.e. a rubber band or shrink-wrap fit to the data

(Cornwell et al. 2006). However this method is also sensitive to outliers, a problem potentially solved by fitting plateau-like splines to data instead (Brewer et al. 2016). There are several alternate approaches based on geometrical interpretation of probability distribution quantiles. Gaussian mixture models fit multivariate normal distributions to unknown subgroups within the data, while kernel density estimation allows an arbitrary probability distribution to be applied to every data point (Blonder et al. in press, Blonder et al. 2014, Carmona et al. 2016). These methods yield more complex fits and can be converted to geometrical shapes by applying a threshold to the distribution at a chosen volume quantile. However, kernel density estimation methods may be less robust in high dimensionalities because of the difficulty in estimating probability density functions compared to methods based on estimating cumulative distribution functions (Blonder et al. in press, Junker et al. 2016). There are also several machine learning methods that can be given geometrical interpretations as hypervolume functions that optimize classification accuracy subject to regularization parameters. These include support vector machines (Blonder et al. in press, Drake and Bossenbroek 2009, Drake et al. 2006, Guo et al. 2005), density-based clustering (Qiao et al. 2015, Sander et al. 1998), range bagging algorithms (Drake 2015), and a range of other ensemble learning (Drake and Beier 2014) and pattern-recognition algorithms (Maher et al. 2014).

Once the general type of model has been chosen, the desired types of analysis may further determine the method used. Investigators interested in evaluating $h(x)$ and projecting to geographic coordinates should choose one of the methods commonly used for species distribution modeling. Those interested in comparing hypervolume similarity should use one of the several methods that can calculate volumes and overlaps. Those

interested in describing complex variation in shape should use a method that can detect holes (Blonder 2016).

There are also several practical issues that distinguish methods. Some methods are computationally much faster than others, although all yield slower performance as dimensionality increases. For example, range boxes can generally be calculated in milliseconds, while convex hulls or boosted regression trees can require seconds, and high-dimensional kernel density estimates can require hours. If the investigator must calculate a large number of hypervolumes (e.g. for null modeling (Gotelli 1996)) then some of these methods will become impractical. Additionally, some methods are easy to use in R software packages and come with a range of visualization and analysis tools while others provide a more limited interface. Lastly, methods vary in the number of parameters that the investigator must either fit or choose. In some cases, models are parameter-free (convex hull, range box) while in other cases there are a large number of parameters that are fit from the data (all regression and machine learning methods). Some methods have only a small number of free parameters controlling the algorithm (e.g. kernel density estimation, boosted regression trees, support vector machines) that can either be chosen based on biological knowledge or fit from the data. Most methods have a small number of nuisance computational parameters that determine the accuracy of the modeling results and that can generally be left to default values. Parsimony in number of free parameters cannot necessarily guide the choice of method, because each method also has strong distributional assumptions that are important in determining whether the modeling approach is suitable for the data.

## Supporting References


Austin, M. et al. 1990. Measurement of the realized qualitative niche: environmental niches of five Eucalyptus species. — Ecol. Monogr. 60: 161-177.

Bahn, V. and McGill, B. J. 2013. Testing the predictive performance of distribution models. — Oikos 122: 321-331.

Blonder, B. 2016. Do Hypervolumes Have Holes? — Am. Nat. 187: E93-E105.

Blonder, B. et al. in press. New approaches for delineating boundaries for n-dimensional hypervolumes. — Meth. Ecol. Evol.

Blonder, B. et al. 2014. The n-dimensional hypervolume. — Glob. Ecol. Biogeogr. 23: 595-609.

Breiman, L. 2001. Random forests. — Mach. Learn. 45:

Brewer, M. J. et al. 2016. Plateau: a new method for ecologically plausible climate envelopes for species distribution modelling. — Meth. Ecol. Evol. 7: 1489-1502.

Busby, J. 1991. BIOCLIM-a bioclimate analysis and prediction system. — Plant Protect. Quart.

Carmona, C. P. et al. 2016. Traits without borders: integrating functional diversity across scales. — Trends Ecol. Evol. 31: 382-94.

Cornwell, W. K. et al. 2006. A trait‐based test for habitat filtering: convex hull volume. − Ecology 87: 1465-1471.

Dolédec, S. et al. 2000. Niche separation in community analysis: a new method. — Ecology 81: 2914-2927.

Drake, J. M. 2015. Range bagging: a new method for ecological niche modelling from presence-only data. — J. Roy. Soc. Inter. 12: 20150086.

Drake, J. M. and Beier, J. C. 2014. Ecological niche and potential distribution of Anopheles arabiensis in Africa in 2050. — Malaria J. 13: 213.

Drake, J. M. and Bossenbroek, J. M. 2009. Profiling ecosystem vulnerability to invasion by zebra mussels with support vector machines. — Theor. Ecol. 2: 189-198.

Drake, J. M. et al. 2006. Modelling ecological niches with support vector machines. — J. Appl. Ecol. 43: 424-432.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. — Ecography 29: 129-151.

Elith, J. et al. 2008. A working guide to boosted regression trees. — J. Anim. Ecol. 77: 802-813.

Farber, O. and Kadmon, R. 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. — Ecol. Model. 160: 115-130.

Gotelli, N. J. G. 1996. Null models in ecology. — Smithsonian Institution Press.

Green, R. H. 1974. Multivariate niche analysis with temporally varying environmental factors. — Ecology 55: 73-83.

Guillera-Arroita, G. et al. 2014. Maxent is not a presence–absence method: a comment on Thibaud et al. — Meth. Ecol. Evol. 5: 1192-1197.

Guisan, A. et al. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. — Ecol. Model. 157: 89-100.

Guo, Q. et al. 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. — Ecol. Model. 182: 75-90.

Hirzel, A. H. et al. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? — Ecology 83: 2027-2036.

Jackson, A. L. et al. 2011. Comparing isotopic niche widths among and within communities: SIBER–Stable Isotope Bayesian Ellipses in R. — J. Anim. Ecol. 80: 595-602.

Junker, R. R. et al. 2016. Dynamic range boxes – a robust nonparametric approach to quantify size and overlap of n-dimensional hypervolumes. — Meth. Ecol. Evol. 7: 1503-1513.

Leathwick, J. et al. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. — Ecol. Model. 199: 188-196.

Maher, S. P. et al. 2014. Pattern-recognition ecological niche models fit to presence-only and presence–absence data. — Meth. Ecol. Evol. 5: 761-770.

Qiao, H. et al. 2015. Marble Algorithm: a solution to estimating ecological niches from presence-only records. — Sci. Rep. 5: 14232.

Sander, J. et al. 1998. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. — Data Min. Knowl. Disc. 2: 169-194.

Swanson, H. K. et al. 2015. A new probabilistic method for quantifying n‐dimensional ecological niches and niche overlap. − Ecology 96: 318-324.