Methods in Ecology and Evolution

MS MATILDA BROWN (Orcid ID : 0000-0003-2536-8365)

Article type : Research Article

Editor : Natalie Cooper

Hyperoverlap: detecting biological overlap in *n*-dimensional space

M.J.M. Brown[1], B.R. Holland[1], G.J. Jordan[1]

Corresponding author: M.J.M. Brown; matilda.brown@utas.edu.au

[1]School of Natural Sciences, University of Tasmania, Hobart, Australia

Running title

Hyperoverlap: detecting biological overlap

Hyperoverlap: detecting biological overlap in *n*-dimensional space

Abstract and keywords

1. Comparative biological studies often investigate the morphological, physiological or ecological divergence (or overlap) between entities such as species or populations. Here, we discuss the weaknesses of using existing methods to analyse patterns of phenotypic overlap and present a novel method to analyse co-occurrence in multidimensional space.

2. We propose a 'hyperoverlap' framework to detect qualitative overlap (or divergence) between point data sets and present the HYPEROVERLAP R package which implements this framework, including functions for visualisation. HYPEROVERLAP uses support vector machines (SVMs) to train a classifier based on point data (such as morphological or ecological data) for two entities. This classifier finds the optimal boundary between the two sets of data and compares the predictions to the original labels. Misclassification is evidence of overlap between the two entities. We demonstrate the theoretical and practical advantages of this method compared to existing approaches (e.g. single-entity hypervolume models) using the bioclimatic data extracted from global occurrence records of conifers.

3. We find that there are instances where single-entity hypervolume models predict overlap, but there are no observations of either entity in the shared hypervolume. In these instances, hyperoverlap reports non-overlap. We show that our method is stable and less likely to be affected by sampling biases than current approaches. We also find that hyperoverlap is particularly effective for situations involving entities with a small number of data points (e.g. narrowly endemic species) for which single-entity models cannot be reliably constructed.

4. We argue that overlap can be reliably detected using HYPEROVERLAP, particularly for descriptive studies. The method proposed here is a valuable tool for studying patterns of overlap in multidimensional space.

Keywords ecospace, hyperoverlap, hypervolume, machine learning, morphospace, overlap, support vector machines.

Main Text

Introduction

Many ecological and evolutionary questions revolve around the study of overlap: Do two species (or any other entity, Table 1) overlap in terms of climatic requirements? How have particular entities diverged (or converged) over evolutionary time? Under what conditions could two entities coexist? These questions of biological overlap are central to a broad range of studies including taxonomy (Rissler & Apodaca, 2007), investigating broad-scale evolution of climatic envelopes (Donoghue & Edwards, 2014), niche partitioning (Peterson *et al.*, 2013), predicting the spread of invasive species (Guisan *et al.*, 2014) and palaeoclimatic estimation (Mosbrugger & Utescher, 1997). Many of these studies have been made possible because of the relatively recent development of large online databases such as the Global Biodiversity Information Facility (GBIF), the Plant Trait Database (TRY; Kattge *et al*. 2020) and WorldClim (Fick & Hijmans, 2017), which make large amounts of biological data publicly available.

We can use hypervolume concepts to analyse patterns of overlap between sets of point data in multidimensional space (e.g. Blonder et al, 2018). Current hypervolume approaches first map each observation in an *n*-dimensional space, where the dimensions are the chosen variables. These approaches then create a multidimensional object (a “hypervolume”) that encloses the observations, often allowing for error. The hypervolume is then assumed to represent the set of phenotypes or environments occupied by the entity. Hypervolume concepts were first used to describe the ecological niche (see Holt, 2009) but they are broadly applicable to any multidimensional space and have been utilised in several other fields (e.g. morphometry, Sidlauskas, 2008; functional traits, Díaz *et al.*, 2016). However, such hypervolume-type studies typically seek to *predict* the distributions of entities, and require *a priori* assumptions about the distribution or shape of the hypervolume, so methods developed for this purpose may not be suited to answer questions which require qualitative inference of overlap.

Many hypervolume-based algorithms in ecology model the occupied region of a single entity – overlap detection is a by-product of this application. In this paper, we use the term ‘single-entity method’ to refer to any which constructs individual models for each entity and then compares them to analyse overlap. Joint species models (e.g. Pollock *et al.*, 2014; Ovaskainen *et al.*, 2016) are an emerging tool to incorporate biotic interactions into niche models but require absence data as well as presence data so they are not considered further here.

In this paper, we discuss the limitations of detecting biological overlap using single-entity methods and argue that all single-entity solutions to this problem share similar theoretical problems. We present the 'hyperoverlap' framework – a novel application of a machine learning classifier to detect overlap between point data sets sampled from hypervolumes in *n*-dimensional space. We also present an R package that implements this analytical framework – HYPEROVERLAP (see https://github.com/matildabrown/hyperoverlap). To highlight the conceptual novelty of our approach, we compare the performance of HYPEROVERLAP with the most comparable single-entity approach: Blonder's 'HYPERVOLUME' algorithms (hereafter referred to as HYPERVOLUME to distinguish the R package from more general uses of the term hypervolume). We analyse a real-world example (the ecological ranges of genera of conifers) to demonstrate the advantages of our method and discuss the caveats that should be considered when using the hyperoverlap framework.

Current approaches

The geometry of the hypervolume may be measured in several ways, depending on *a priori* expectations about the shape of the hypervolume. A plethora of increasingly sophisticated algorithms have been developed to model this hypervolume, either directly or indirectly, and measure overlap between the estimated hypervolumes occupied by two entities. Although earlier approaches were computationally and/or conceptually limited to low-dimensional analyses (e.g. Broennimann *et al.*, 2012), several recent methods allow direct analysis in *n*-dimensional space.

Machine learning methods are used extensively to analyse landscape-scale, multidimensional data. In explicit hypervolume models, machine learning classifiers are used to predict the habitat suitability of each pixel in a landscape (e.g. MaxEnt; Phillips *et al.*, 2006); to classify points in ecological space as 'in' or 'out' of the modelled niche (e.g. `hypervolume_exclusion_test`; Blonder *et al.*, 2018); or to define the boundary of the niche in n-dimensional space (e.g. `hypervolume_svm`; Blonder *et al.*, 2018). Once described by an appropriate model, two hypervolumes may be compared and the volumes of the overlapping and unique regions can be measured (see `hypervolume_set`, `hypervolume_overlap_statistics` functions; Blonder *et al.*, 2018).

Current methods of describing hypervolumes vary in the geometric model used. The simplest of these methods is the *n*-dimensional convex hull, implemented in the GEOMETRY R package (R Core Team, 2014; Habel *et al.* 2015). However, many biological hypervolumes are not convex. Similarly, the hypervolumes simulated by NICHEA software (Qiao *et al*., 2016) are constrained to ellipsoids, and so are not broadly applicable to the non-convex or irregular data common encountered in ecological problems. Dynamic range boxes (DYNRB PACKAGE; Schreyer *et al.* 2015) have the advantage that they do not assume normally or elliptically distributed data, but the authors acknowledge that correlated variables must be removed during pre-processing (Junker *et al.* 2016). Because many biological variables are strongly correlated, this method is limited in the variables that can be analysed. Blonder's HYPERVOLUME package (Blonder *et al.*, 2014; Blonder *et al.*, 2018) includes a range of functions for hypervolume modelling and comparison. In both HYPERVOLUME AND HYPEROVERLAP the shape of the hypervolume is not defined by *a priori* expectations, so we have used HYPERVOLUME as a standard to evaluate the performance of HYPEROVERLAP. Additionally, these methods use the same machine learning classifier (SVM), so conflicting results will be driven by conceptual rather than algorithmic differences.

Weaknesses of using single-entity methods for detecting overlap in multidimensional space

Reliable results from single-entity methods depend on meeting several assumptions, many of which are unlikely to hold for landscape-scale datasets (Jarnevich *et al.*, 2015). The most commonly violated of these assumptions is that the records are an unbiased sample of the biological range. Satisfying this assumption requires even sampling from the entire geographic and ecological and/or phenotypic range of an entity. For almost all entities in GBIF, occurrence sampling is substantially biased in geographic space (Boakes *et al.*, 2010; Beck *et al.*, 2014), with strong biases towards roads and urbanised areas, and especially strong biases towards rare species (Stolar and Nielsen, 2015).

Thus, real-world entities are often represented by sampled data that are irregular, holey, discontinuous, or include outliers (Blonder, 2016). Outliers are often treated as noise by modelling algorithms – which are designed to filter out noisy data – but sampling effort, habitat fragmentation, and the geographical distribution of suitable habitat can each cause real occurrence records to appear as outliers. Highly restricted, often endangered entities with geographic outliers are often high priorities for conservation management but are also most likely to be misrepresented by these models. Adjusting the model-fitting parameters to ensure that every occurrence record is included in the model predictions (i.e. a 0% omission threshold) can result in severe extrapolation. In species distribution modelling, this means that conditions well outside the observed hypervolume are predicted to be suitable (Escobar *et al.*, 2018). This means that the choice of omission threshold may falsely inflate or decrease observations of overlap between entities.

These issues are unavoidable when attempting to resolve the complex problem of accurately modelling the hypervolume from sampled point data, and there is no universal best approach (Qiao *et al.*, 2015). However, we suggest that the detection and description of the observed overlap between two hypervolumes can be achieved by comparing the point data for entities directly and is thus a simpler task than explicitly modelling the hypervolume.

Hyperoverlap conceptual framework

We propose a qualitative method for detecting multidimensional overlap. There are three possible qualitative relationships between points sampled from two hypervolumes: nested, overlapping, or non-overlapping (Fig. 1). If the observations from each hypervolume can be perfectly separated by a decision boundary (Table 1), we cannot identify a shared region and the entities do not overlap. If this decision boundary does not exist, the entities overlap (with misclassified points occupying the shared region). If we assume that all observations of an entity are within the hypervolume, this principle can be applied to samples of point data (but see *Caveats and Limitations*).

The HYPEROVERLAP algorithm finds the optimal separating hyperplane between two entities using SVMs based on point data and calculates the number of points belonging to each entity on either side of this boundary. If there are no misclassified points, we infer that the hypervolumes for the entities do not overlap (Fig 1a, but see *Caveats and Limitations*). If at least one point is misclassified (Fig.1b), the two entities overlap. If no boundary can be found (Fig.1c), one hypervolume is 'nested' within the other (see *Terminology*).

If there is a single hyperplane (of *n*-1 dimensions) which perfectly separates the observations from each hypervolume, the entities are linearly separable (Fig. 2a). For entities which cannot be separated using a linear plane but occupy distinct regions of space (Fig. 2b), a kernel function (Scholkopf & Smola, 2002) can be used to find a curvilinear decision boundary. Polynomial kernel functions are preferred because other functions (e.g. sigmoidal or Gaussian) can create complex decision boundary shapes that are likely to overfit the classifier (Fig. 2c). The order of the polynomial kernel function constrains the complexity of the decision boundary. Potential concerns about the biological meaningfulness of this boundary may be addressed by visualisation (functions provided in the HYPEROVERLAP package).

Sketch of the HYPEROVERLAP algorithm

Before implementing The HYPEROVERLAP workflow, it is important to pre-process data to exclude duplicate, incomplete or erroneous records, and to ensure that the dimensions are comparable (see Blonder, 2018). A support vector machine (SVM) is then trained on the data using the E1071 package (Meyer *et al.*, 2018). This creates a fitted linear model that is used to predict the labels of the input data. If the model correctly classifies every point (i.e. the entities can be separated by the linear hyperplane) the function returns the result (non-overlap) and the coordinates of the decision boundary. If there are misclassified points, SVMs are trained using polynomial kernels of increasing complexity, each time evaluating the number of misclassified points until a separating hyperplane is found. If such a hyperplane is not found, the result ('overlap') is returned.

Finding the decision boundary for non-overlapping entities is fast (typically milliseconds) but can be much slower if the entities overlap. To prevent excessive searching, the algorithm does not attempt a non-linear kernel if the linear result is that the two entities are nested, or if a certain number of points representing significant overlap are misclassified. This parameter is user-defined (see `stoppage.threshold`; package documentation).

Machine learning classifiers are typically trained with the aim to correctly predict the labels of unknown data. Various caveats about relative and absolute sample sizes apply to SVMs when they are used to automate identification in this way. However, these caveats are not relevant to HYPEROVERLAP, which does not use SVMs in a predictive fashion. Instead, HYPEROVERLAP uses the SVM classifier as a descriptive tool and so overfitting is prevented by setting constraints on the shape of the decision boundary. This can be verified using visualisation of the decision boundary (in three or fewer dimensions) or visualisation of the data using ordination (in four or more dimensions) using functions in the HYPEROVERLAP R package (see *Appendix S1* in Supporting Information for example).

Theoretical advantages of HYPEROVERLAP

Dimensionality and sample size

The hyperoverlap algorithm considers the data for two entities simultaneously, unlike other hypervolume methods (e.g. HYPERVOLUME, BLONDER ET AL., 2018; NICHEA, Qiao *et al*., 2016). It is often difficult or impossible to use single-entity methods to fit models to very small samples, and thus to investigate many relevant problems (e.g. those involving threats to endangered species). This problem affects all methods which fit individual models to entities. However, the most relevant sample size for HYPEROVERLAP is *total* sample size for the pair of entities. As a result, this approach can be effective with sample sizes as small as 1 for one of the entities –provided that the number of observations of the other entity is at least moderately large (see *Evaluation: Results; Case Study 2*). However, care should be taken when analysing two very small entities, as discussed in *Caveats and Limitations*.

Computational effort

Conventional measurement of overlap from single-entity models require two phases; initial modelling, then pairwise comparison of models. Unless the number of entities is very small, conventional memory constraints demand that these models are written to disk and re-read for comparison, separating these two phases. Hyperoverlap builds models using the paired data, so does not require this storage step. Computational effort is further reduced by constraints on the shape of decision boundary; the decision boundary produced by HYPEROVERLAP is constrained to linear and low-degree polynomial kernels (unlike the edges of the hypervolumes modelled using HYPERVOLUME).

Evaluation

Methods

To evaluate the performance of HYPEROVERLAP, we compared parallel results between HYPEROVERLAP and HYPERVOLUME for 71 conifer genera (2485 pairs). Conifers are an ideal group for this because the group is diverse with regard to ecological and distributional range (e.g. *Pinus* occurs across the Northern Hemisphere; *Wollemia* is only found in one gorge near Sydney, Australia; Farjon & Filer, 2013) and because species of conifers have well-defined bioclimatic ranges (Brodribb & Hill, 1999). The data are geographic point records for each genus of conifer used by Larcombe *et al.* (2018). We extracted climatic data for each point record from WorldClimV2 at 30” (approximately 1km^2) resolution and used DISMO (Hijmans *et al.*, 2015) to build the values for three variables which are known to correlate to physiological stresses in conifers. These variables were mean minimum temperature of the coldest month (mint.cm) reflecting frost tolerance (Sakai & Larcher, 2012); mean temperature of the warmest quarter (at.warmq) reflecting growing season temperature (Prentice et al., 1992); and mean precipitation of the driest quarter (p.dryq) reflecting drought tolerance (Mackey, 1994). Although HYPEROVERLAP has been developed for *n*-dimensional analyses, using only three dimensions for evaluation allowed the results to be inspected directly, without requiring ordination. We also conducted analyses using two additional variables (mean precipitation of the warmest and wettest quarters, respectively) to assess computational performance in higher dimensional space.

Precipitation records (p.dryq) were transformed to an approximately normal distribution by taking the fourth root and all variables were *z*-transformed to the global (-90° to 90° latitude) mean and standard deviation of each variable. We compared the overlap/non-overlap results, computational time and stability of the two methods (HYPEROVERLAP and HYPERVOLUME). To evaluate stability, each overlap detection function was run ten times (a larger number of runs was not computationally feasible). We then compiled and compared the results from each method. For each entity pair that gave conflicting results, we visually inspected the data to assess the accuracy of each method.

Runtimes are given for scripts run on an Intel i7-8700k CPU.

Results

Overlap detection

Hyperoverlap detected 1134 non-overlapping pairs of entities (of 2485 pairs; Fig. 3). Of these non-overlapping pairs, 1082 (95%) could be separated with a linear decision boundary, and only 52 (2.1%) required a curvilinear hyperplane (polynomial kernel function) to identify ecological non-overlap. The number of non-overlapping pairs identified by hypervolume varied with run, ranging from 1076 to 1092 (see *Computational Time and Stability*).

There were differences in the results given by different methods. hyperoverlap reported 133 non-overlaps (5.5% of the 2415 pairs excluding *Wollemia*) that were classified as overlaps by hypervolume (see *Case Study 1*), and 33 overlaps (1.4% of total) where hypervolume reported non-overlap (see *Case Study 2*). Visualisation confirmed the status of all the non-overlaps identified by hyperoverlap that were reported as overlaps by hypervolume. There was no discernible pattern in these conflicts; they do not cluster by taxonomic group or sample size (Fig. 4). In addition, while Hyperoverlap satisfactorily created models to compare *Wollemia* with each other genus, Hypervolume could not produce a hypervolume for this taxon because, with only two unique points in ecospace, it was not possible to build a model in three dimensions. Although *Wollemia* cannot be included in comparisons of stability or computation times between hypervolume and hyperoverlap, it should be noted that the small number of points for this entity is not an artefact of sampling effort. These data represent the entire range of this genus at this spatial resolution.

Computational time & stability

At the default parameters `(cost=1000, kernel=”polynomial”, kernel.degree=5, stoppage.threshold=0.4)`, the mean total runtime (all pairwise comparisons) for hyperoverlap was 228 minutes (range 212-239 minutes). The results from hyperoverlap were exceptionally stable; the results for identifying overlap versus non-overlap, shape, polynomial order and number of misclassified points were identical in all 10 runs. When the algorithm was constrained to linear decision boundaries, the average runtime was 85 minutes.

At default parameters, computation of HYPERVOLUME results took 16 minutes. However, these results were less stable than those produced by HYPEROVERLAP. Qualitative results (overlap/non-overlap) were inconsistent for 109 pairs of entities (4.5%). Increasing the `samples.per.point` parameter by a factor of 100 reduced this instability to 38 pairs (1.6%) but increased the average runtime to 327 minutes.

Preliminary tests in five-dimensional ecospace (adding mean precipitation of the warmest and wettest quarters) emphasised the computational advantage of HYPEROVERLAP in higher dimensions; the average runtime at default parameters was 147 minutes for HYPEROVERLAP and 855 minutes for HYPERVOLUME.

Case Study 1: Dacrycarpus and Cupressus

The comparison of *Dacrycarpus* (555 unique points in ecospace) and *Cupressus* (133 points) illustrates the main reason for the observed conflicting results between HYPEROVERLAP and HYPERVOLUME (points in orange and red, Fig. 3). HYPERVOLUME finds that these entities overlap (Fig. 5b; overlap shown in green), but HYPEROVERLAP finds that the points of each entity occupy distinct regions of ecospace. This can be verified by visualisation of the decision boundary (Fig. 5a). The region of overlap found by HYPERVOLUME is the result of small but non-trivial extrapolation by the model-building algorithm; none of the original observations are within this region of apparent overlap. This extrapolation effect was observed for all entity pairs for which HYPEROVERLAP detected non-overlap, but HYPERVOLUME reported overlap (80% of total conflicts). If our goal is to predict potential overlap, then this extrapolation may be sensible. However, if we are aiming to identify regions of multidimensional space *occupied by both entities*, we suggest that the result given by HYPEROVERLAP is more accurate.

Case study 2: Metasequoia

Metasequoia (representing the single species, *M. glyptostroboides*) is a narrowly endemic genus of conifers with only three unique points in ecospace at our sampling resolution. Its native range is limited to a small region of Hubei Province, China, although fossils indicate that it was previously widespread (LePage *et al.*, 2005). This entity proved the most problematic for HYPERVOLUME; for over 20% of pairs involving *Metasequoia* (15 pairs) the results for HYPEROVERLAP and HYPERVOLUME were in conflict. Although the conflicting result for *Metasequoia* and *Cathaya* is a case of false separation like those discussed in *Case Study 1,* all the other conflicts represent cases in which HYPERVOLUME finds a false separation between *Metasequoia* and the other entity. In these latter cases, HYPEROVERLAP identified overlap, and visualisation shows that the region occupied by *Metasequoia* is deeply nested within the hypervolume occupied by the other entity (Fig. 6). It is not clear what is driving this anomalous result from HYPERVOLUME, but large differences in sample size may contribute.

Caveats and limitations

The first obvious limitation of the hyperoverlap framework is that while it effectively *detects* overlap or non-overlap, it does not *measure the amount* of overlap. The overlapping region may be studied by visualisation or inspection of misclassified points, but to measure its volume or calculate a similarity index between the two entities would require the edges of each entity to be defined. This would then invoke the assumptions and challenges associated with single-entity models that this framework was designed to circumvent. However, the shared hypervolume may be modelled based on misclassified points using existing methods.

There are certain theoretical situations where entities do not overlap but cannot be separated using the HYPEROVERLAP algorithm (see Fig. 7 for examples). Although some of these situations may be biologically plausible, we did not find evidence of any in this study. However, such cases may be identified by using the visualisation functions in the HYPEROVERLAP package.

HYPEROVERLAP is also subject to many caveats that apply to the use of hypervolume concepts. Incomplete records cannot be placed in hyperspace so must be excluded or otherwise augmented (see Blonder 2014). Although SVMs handle high dimensionality well, care should be taken when comparing entities that are both highly restricted in multidimensional space. The extreme case is that if the total number of unique points for a pair of entities is lower than *n+1*, where *n* is the number of dimensions, the two entities can always be separated perfectly with a linear hyperplane. Curvilinear separation is not recommended for small total sample sizes.

It should also be noted that observations represent points in time as well as space; the occupation of morphological or ecological space by an entity is dynamic and is likely to change through time – the fossil record of conifers shows evidence of major changes in ecological occupation during the Cenozoic (Macphail, 2007). A significant caveat is that hyperoverlap does not directly identify overlap between pairs of hypervolumes, instead it identifies overlap between observations sampled from those hypervolumes. Thus, there will be a false identification of non-overlap if there are no observations from the true region of intersection. Other methods deal with this issue mainly by padding each point, in effect extrapolating the range of each entity. However, this solution is problematic, as discussed above (*Case Study 1*). In any case, no approach can fully overcome poor sampling. In particular, care should be taken when using databased occurrence records, which are likely to include some erroneous observations. Visualisation of results and expert knowledge of the entities concerned are both vital to using HYPEROVERLAP and to identify errors such as those illustrated in Fig. 7.

Extensions to HYPEROVERLAP

Here, we have focused on overlap versus non-overlap, rather than exploring the question of nested hypervolumes, but this type of relationship can also be explored using the hyperoverlap framework. This has several possible applications in studying recent changes in hypervolumes, including phenological shifts and detection of ecological range expansion in invasive species. Although this conceptual extension has not been tested, it is a promising avenue for further research and potential inclusion in future versions of the HYPEROVERLAP R package.

Conclusions

The hyperoverlap framework presented here has potential applications in many disciplines – although the concepts underpinning this method have been used widely within ecology, they are not specific to this field. HYPEROVERLAP can be used to investigate ecological and evolutionary partitioning, palaeoclimatic conditions, taxonomy and historical changes in ecology or morphology.

For many biological questions, it is not necessary to model the underlying hypervolume to evaluate overlap. By comparing the space occupied by entities without explicitly describing the geometry of the underlying hypervolumes, fewer assumptions are required to be met and results can be more accurate and reliable than existing methods, as demonstrated clearly for our real-world example (conifers). The approach is particularly effective when the set of entities to be compared is very large and includes entities with a small number of occurrences relative to the dimensionality of the analysis (e.g. species with highly restricted distributions), or when there are potential complex interactions between variables. The HYPEROVERLAP R package provides a user-friendly, intuitive machine-learning method to detect overlap in *n*-dimensional space, and is an additional tool to use in analyses of many biological datasets.

References

Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, **19**, 10-15. doi: 10.1016/j.ecoinf.2013.11.002

Blonder, B. (2016) Do hypervolumes have holes? *The American Naturalist*, **187**, E93-E105. doi: 10.1086/685444

Blonder, B. (2018) Hypervolume concepts in niche‐and trait‐based ecology. *Ecography*, **41**, 1441-1455. doi: 10.1111/ecog.03187

Blonder, B., Lamanna, C., Violle, C. & Enquist, B.J. (2014) The *n*‐dimensional hypervolume. *Global Ecology and Biogeography*, **23**, 595-609. doi: 10.1111/geb.12146

Blonder, B., Morrow, C.B., Maitner, B., Harris, D.J., Lamanna, C., Violle, C., Enquist, B.J. & Kerkhoff, A.J. (2018) New approaches for delineating n-dimensional hypervolumes. *Methods in Ecology and Evolution*, **9**, 305– 319. doi: 10.1111/2041-210x.12865

Boakes, E.H., McGowan, P.J., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, **8**, e1000385. doi: 10.1371/journal.pbio.1000385

Brodribb, T. & Hill, R. (1999) The importance of xylem constraints in the distribution of conifer species. *New Phytologist*, **143**, 365-372. doi: 10.1046/j.1469-8137.1999.00446.x

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.J., Randin, C., Zimmermann, N.E., Graham, C.H. & Guisan, A. (2012) Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography*, **21**, 481-497. doi: 10.1111/j.1466-8238.2011.00698.x

Díaz, S., Kattge, J., Cornelissen, J.H., Wright, I.J., Lavorel, S., Dray, S., Reu, B., Kleyer, M., Wirth, C. & Prentice, I.C. (2016) The global spectrum of plant form and function. *Nature*, **529**, 167. doi: 10.1038/nature16489

Donoghue, M.J. & Edwards, E.J. (2014) Biome shifts and niche evolution in plants. In: *Annual Review of Ecology, Evolution, and Systematics*, pp. 547-572. doi: 10.1146/annurev-ecolsys-120213-091905

Escobar, L.E., Qiao, H., Cabello, J. & Peterson, A.T. (2018) Ecological niche modeling re-examined: A case study with the Darwin's fox. *Ecology and Evolution*, **8**, 4757-4770. doi: 10.1002/ece3.4014

Farjon, A. & Filer, D. (2013) *An atlas of the world's conifers: an analysis of their distribution, biogeography, diversity and conservation status*. Brill, Leiden. doi: 10.1163/9789004211810

Fick, S.E. & Hijmans, R.J. (2017) WorldClim 2: new 1‐km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, **37**, 4302-4315. doi: 10.1002/joc.5086

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C. & Kueffer, C. (2014) Unifying niche shift studies: insights from biological invasions. *Trends Ecol Evol*, **29**, 260-9. doi: 10.1016/j.tree.2014.02.009

Habel, K., Grasman, R., Gramacy, R.B., Stahel, A. & Sterratt, D.C. (2015). Package ‘geometry’. Available online at: http://cran.r-project.org/web/packages/geometry/index.html

Kattge, J, Bönisch, G, Díaz, S, et al. (2020) TRY plant trait database – enhanced coverage and open access. *Glob Change Biol*, **26**, 119-188. doi: 10.1111/gcb.14904

Hijmans, R.J, Phillips, S., Leathwick, J. and Elith, J. (2011), Package ‘dismo’. Available online at: http://cran.r-project.org/web/packages/dismo/index.html.

Holt, R.D. (2009) Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 19659-65. doi: 10.1073/pnas.0905137106

Jarnevich, C.S., Stohlgren, T.J., Kumar, S., Morisette, J.T. & Holcombe, T.R. (2015) Caveats for correlative species distribution modeling. *Ecological Informatics*, **29**, 6-15. doi: 10.1016/j.ecoinf.2015.06.007

Junker, R. R., Kuppler, J., Bathke, A. C., Schreyer, M. L., & Trutschnig, W. (2016). Dynamic range boxes–a robust nonparametric approach to quantify size and overlap of n-dimensional hypervolumes. *Methods in Ecology and Evolution*, **7**(12), 1503-1513. doi: 10.1111/2041-210x.12611

Larcombe, M.J., Jordan, G.J., Bryant, D. & Higgins, S.I. (2018) The dimensionality of niche space allows bounded and unbounded processes to jointly influence diversification. *Nature Communications*, **9**, 4258. doi: 10.1038/s41467-018-06732-x

LePage, B.A., Yang, H. & Matsumoto, M. (2005) The evolution and biogeographic history of Metasequoia. *The geobiology and ecology of Metasequoia*, pp. 3-114. Springer, Dordrecht. doi: 10.1007/1-4020-2764-8_1

Leslie, A. B., Beaulieu, J. M., Rai, H. S., Crane, P. R., Donoghue, M. J., & Mathews, S. (2012). Hemisphere-scale differences in conifer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, **109**(40), 16217-16221. doi: 10.1073/pnas.1213621109

Mackey, B.G. (1994) Predicting the potential distribution of rain‐forest structural characteristics. *Journal of Vegetation Science*, **5**, 43-54. doi: 10.2307/3235636

Macphail, M. (2007) *Australian palaeoclimates: Cretaceous to Tertiary a review of palaeobotanical and related evidence to the year 2000*. CRC LEME, Bentley.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., Lin, C.-C. & Meyer, M.D. (2018) Package ‘e1071’. Available online at: https://cran.r-project.org/web/packages/e1071/index.html

Mosbrugger, V. & Utescher, T. (1997) The coexistence approach—a method for quantitative reconstructions of Tertiary terrestrial palaeoclimate data using plant fossils. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **134**, 61-86. doi: 10.1016/s0031-0182(96)00154-x

Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2016) Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, **7**, 428-436. doi: 10.1111/2041-210x.12502

Peterson, M.L., Rice, K.J. & Sexton, J.P. (2013) Niche partitioning between close relatives suggests trade-offs between adaptation to local environments and competition. *Ecol Evol*, **3**, 512-22. doi: 10.1002/ece3.462

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231-259. doi: 10.1016/j.ecolmodel.2005.03.026

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A. & McCarthy, M.A. (2014) Understanding co‐occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**, 397-406. doi: 10.1111/2041-210x.12180

Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A. & Solomon, A.M. (1992) A Global Biome Model Based on Plant Physiology and Dominance, Soil Properties and Climate. *Journal of Biogeography*, **19**, 117-134. doi: 10.2307/2845499

Qiao, H., Soberón, J. & Peterson, A.T. (2015) No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, **6**, 1126-1136. doi: 10.1111/2041-210x.12397

Qiao, H., Peterson, A.T., Campbell, L.P., Soberón, J., Ji, L. & Escobar, L.E. (2016) NicheA: creating virtual species and ecological niches in multivariate environmental scenarios. *Ecography*, **39**, 805-813. doi: 10.1111/ecog.01961

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rissler, L.J. & Apodaca, J.J. (2007) Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (Aneides flavipunctatus). *Systematic Biology*, **56**, 924-942. doi: 10.1080/10635150701703063

Sakai, A. and Larcher, W. (1987) *Frost Survival of Plants. Responses and Adaptation to Freezing Stress*, Springer‐Verlag, New York.

Scholkopf, B. & Smola, A.J. (2002) Kernels. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press. doi: 10.7551/mitpress/4175.003.0005

Schreyer, M., Trutschnig, W., Junker, R.R, Kuppler, J., Bathke, A., Parkinson, J.H. & Kutil, R. (2018). Package ‘dynRB’. Available online at: https://cran.r-project.org/web/packages/dynRB/index.html

Sidlauskas, B. (2008) Continuous and arrested morphological diversification in sister clades of characiform fishes: a phylomorphospace approach. *Evolution: International Journal of Organic Evolution*, **62**, 3135-3156. doi: 10.1111/j.1558-5646.2008.00519.x

Soberon, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics,* **2**, 1–10. doi: 10.17161/bi.v2i0.4

Stolar, J. & Nielsen, S.E. (2015) Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, **21**, 595-608. doi: 10.1111/ddi.12279

Tables

Table 1. Key terms in the hyperoverlap framework

decision boundary	The hypersurface which best separates the data of a pair of entities in *n*-dimensional space. May be linear or non-linear.
entity	Any group of individuals to be compared, as per Broennimann *et al.* (2012). Not limited to species; other examples may be genera, families, native or invasive populations or any other statistical population.
explicit hypervolume method/model	Any method or model which explicitly describes the geometry of the hypervolume. Examples include convex hulls (Habel *et al.* 2015), dynamic range boxes (Schreyer *et al.* 2015) and the HYPERVOLUME package (Blonder *et al.*, 2014; Blonder *et al.*, 2018).
hyperplane	An *n*-1-dimensional subspace of an *n*-dimensional space.
hypervolume	A contiguous *n*-dimensional region in *n*-dimensional space.
kernel	A function that transforms the original, *n*-dimensional data into higher dimensional space in such a way that a hyperplane can be fitted to the data (see Scholkopf & Smola, 2002).
nested hypervolumes	A qualitative relationship between two hypervolumes where one entity occurs entirely within the region of space occupied by the other entity.
overlap	The observed intersection in *n*-dimensional space of the hypervolumes occupied by two entities, where the dimensions represent biological variables.
single-entity method	An approach to overlap detection which constructs

	individual models for each entity, then measures overlap of these models.
support vector machine (SVM)	A machine learning classifier that finds the maximal-margin separating hyperplane within classes (see Scholkopf & Smola, 2002)

Figure captions

Figure 1. *There are three possible relationships between two hypervolumes. Points sampled from two hypervolumes (top panels) can be used to train a classifier, find the optimal decision boundary (dashed line), and identify misclassified points (highlighted in yellow). The possible relationships are: the hypervolumes do not intersect (a); the hypervolumes intersect (b) or one hypervolume is contained within the other (c). This concept can be easily visualised in two or three dimensions but can be generalised to any* n-*dimensional space.*

Figure 2. *Decision boundaries generated using different kernel functions. A linear kernel (a) always produces a linear decision boundary, a polynomial kernel (b) may produce a curvilinear decision boundary and a Gaussian kernel (c) can produce a complex decision boundary which does not reflect the underlying biology.*

Figure 3*. Pairwise comparison of climatic distributions of conifer genera (grouped phylogenetically) using* hyperoverlap*. A fully labelled version of this figure is available online (See Appendix S2). Phylogeny from Leslie et al.* (*2012).*

Figure 4. *Conflicting results between* hyperoverlap *and* hypervolume*, with entities ordered phylogenetically (a) and by number of unique points in hyperspace (b).*

Figure 5. *The ecological occupation of* Dacrycarpus (blue) *and* Cupressus (red)*. These entities can be separated by a single linear hyperplane using* hyperoverlap *(a), but* hypervolume *predicts a region of overlap, shown in green (b).*

Figure 6. *The ecological occupation of* Metasequoia *and* Taxus*. The occurrences of* Metasequoia *(position indicated by arrows) are nested within the region occupied by* Taxus*, but the models produced by* hypervolume *do not intersect, despite obvious visual overlap.*

Figure 7*. Two possible relationships between two entities for which* hyperoverlap *would be expected to falsely detect overlap. The pattern shown in (a) could be caused by a combination of biological thresholds (e.g. enzyme thermal tolerances) and competitive exclusion. In (b), biological, geographic or other factors could cause the hypervolume geometry to be holey or otherwise very complex. In both cases, the* hyperoverlap *decision boundary (shown by dotted line in (a)) cannot separate the two entities when constrained to a polynomial kernel. However, these scenarios can be resolved using visualisation.*

Acknowledgements

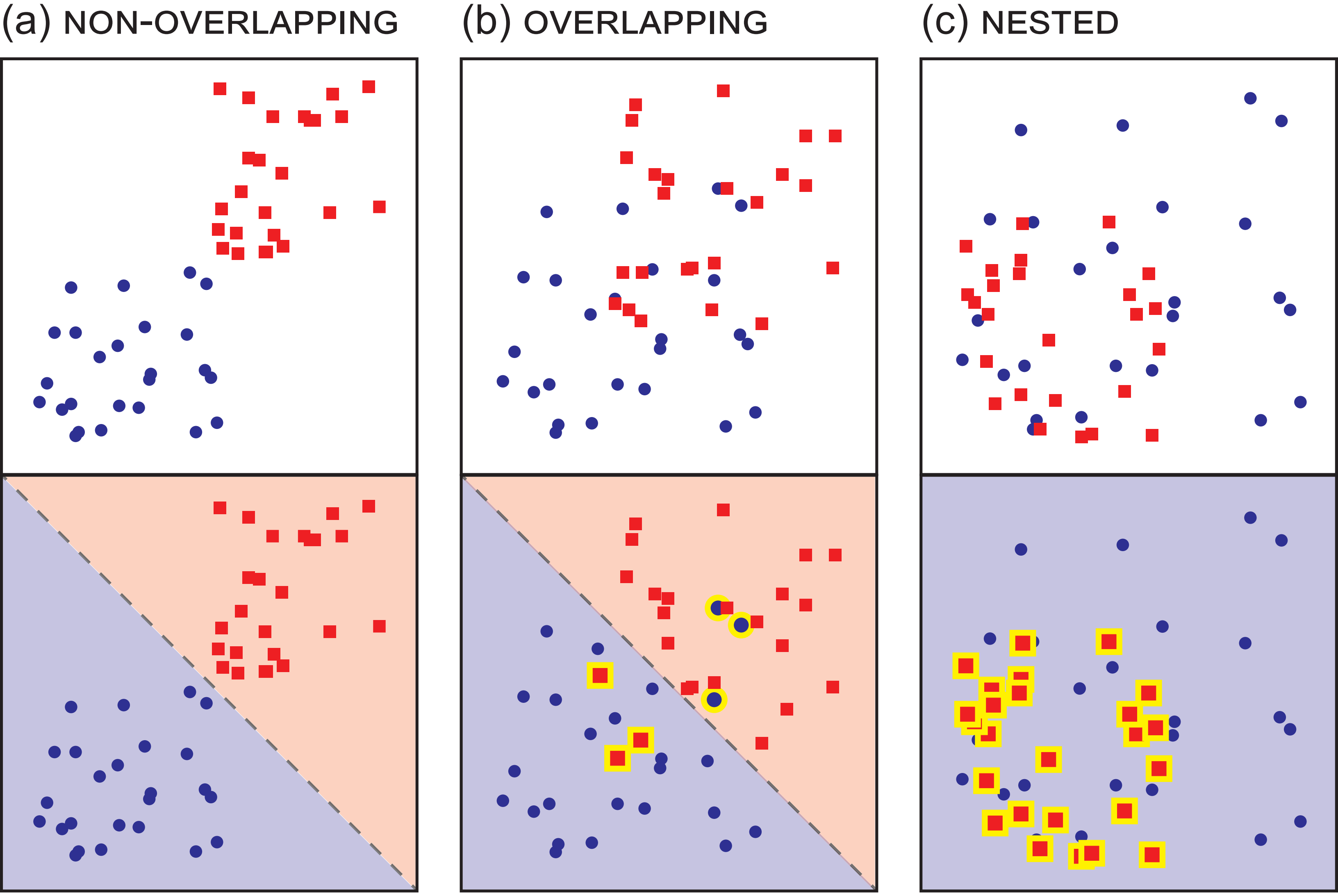
We thank Tim Brodribb, Kate Johnson and Michael Charleston for their assistance in preparing this manuscript. This research was supported by an Australian Government Research Training Program (RTP) Scholarship, ARC Discovery Project DP160100809 and the Westpac Scholars Trust.

Author contribution statement

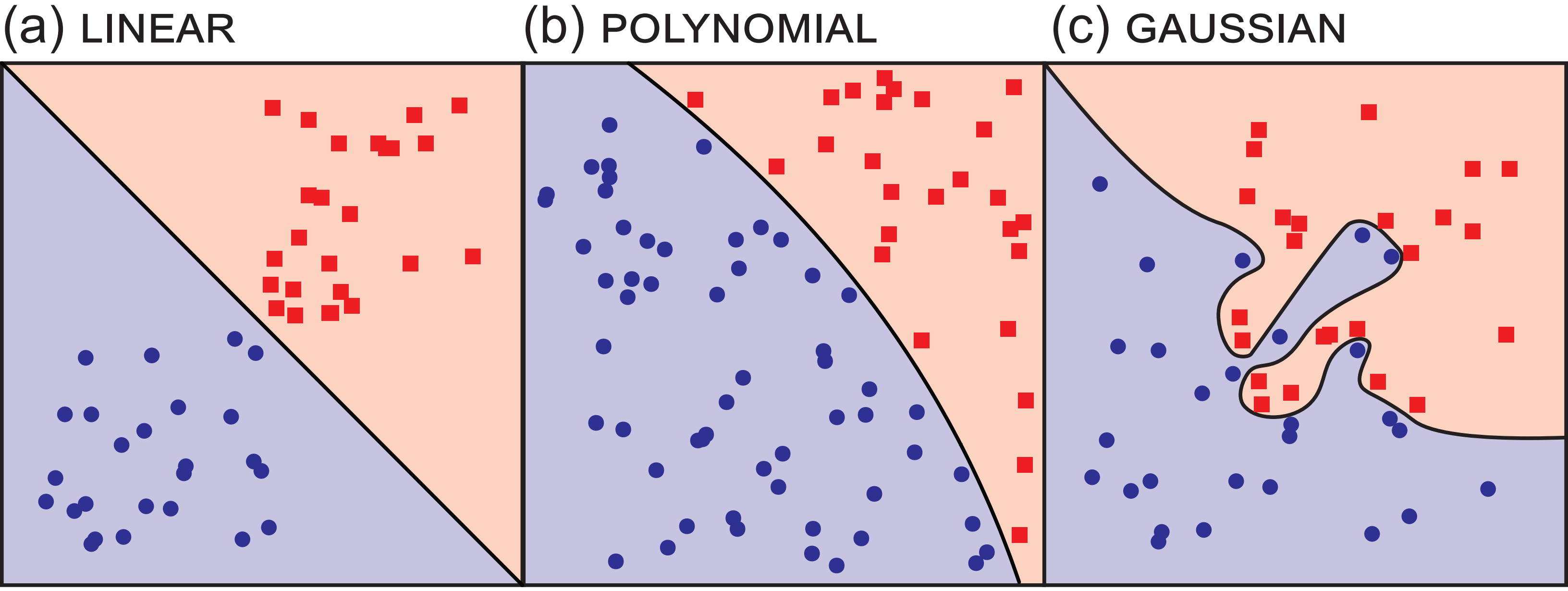
All authors contributed to methodology; MJMB led the R package development; MJMB and GJJ led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility statement

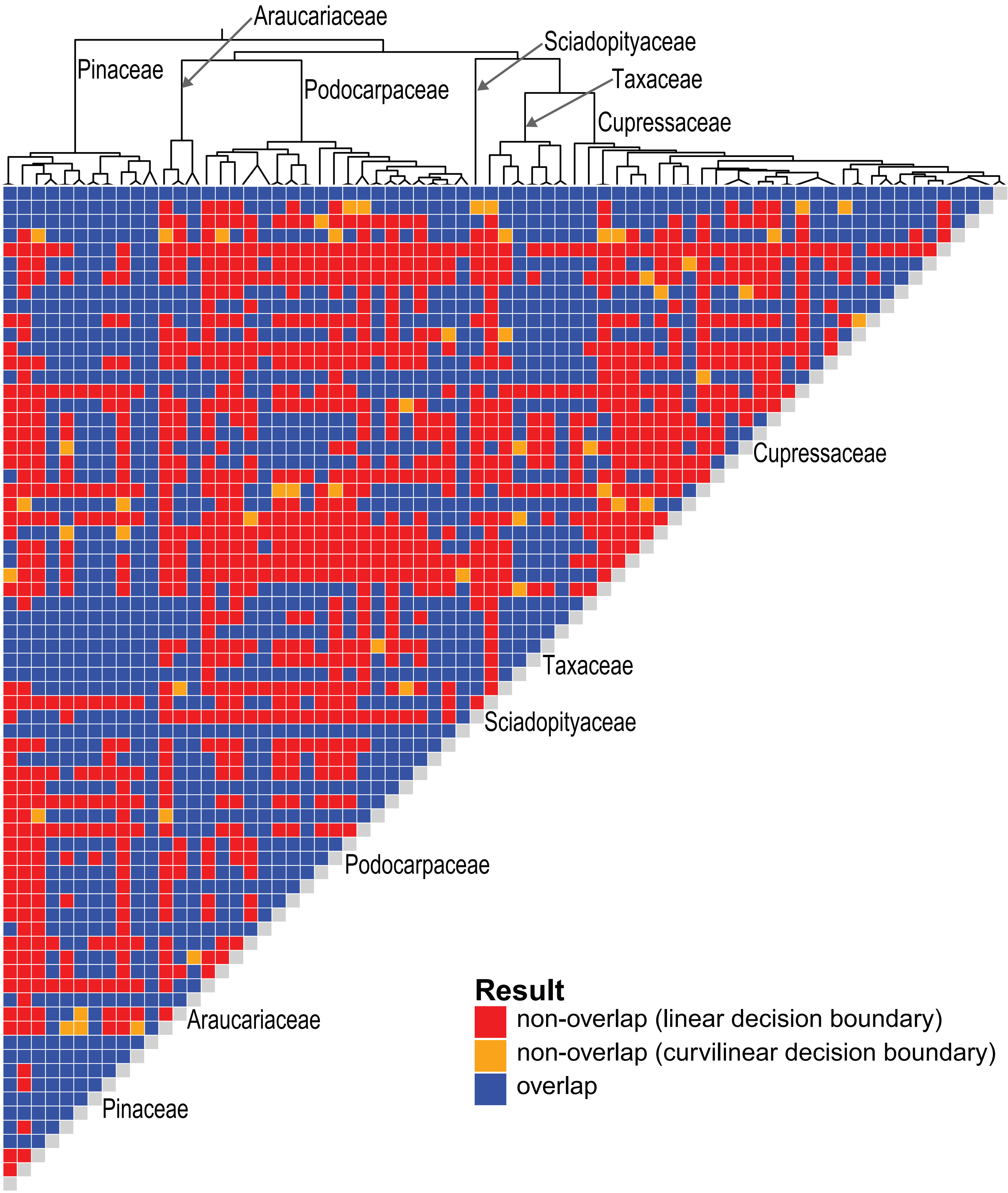
The R package code, conifer data and script used to for evaluations are available on GitHub and deposited in Zenodo (doi: 10.5281/zenodo.3628890).

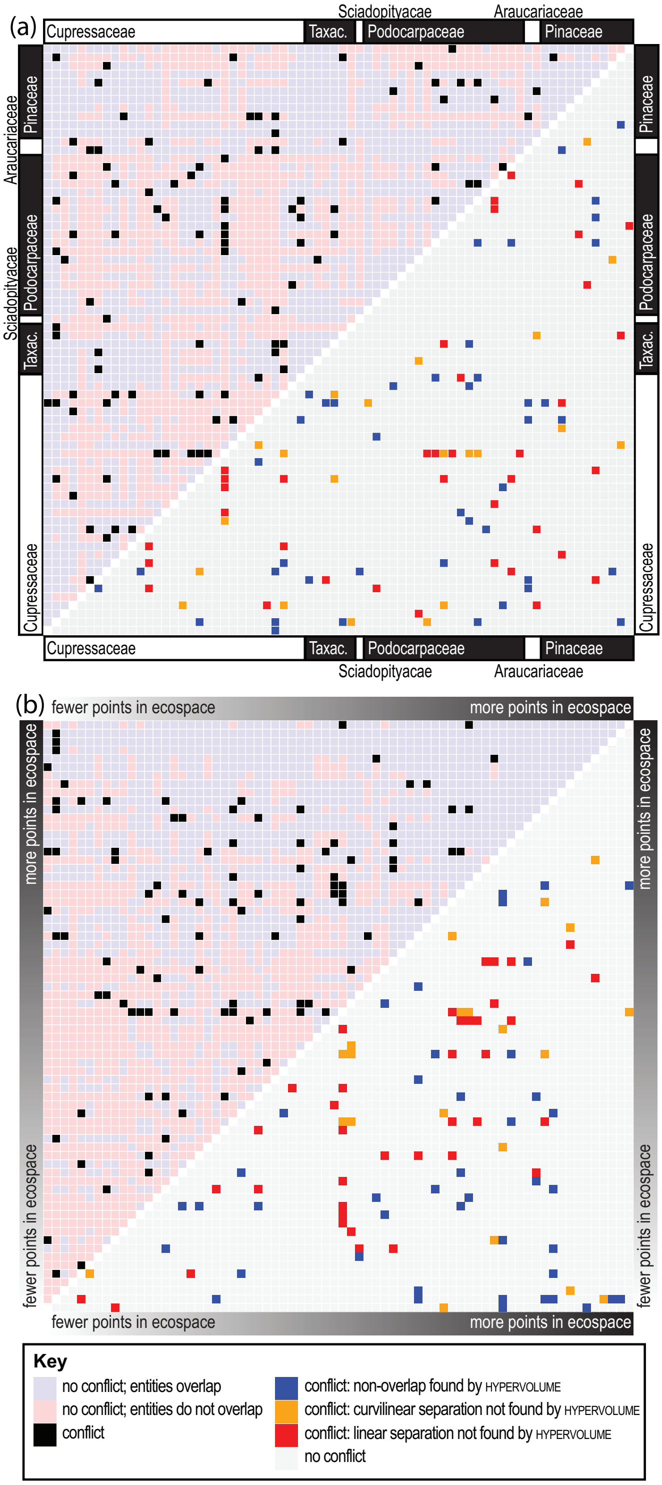


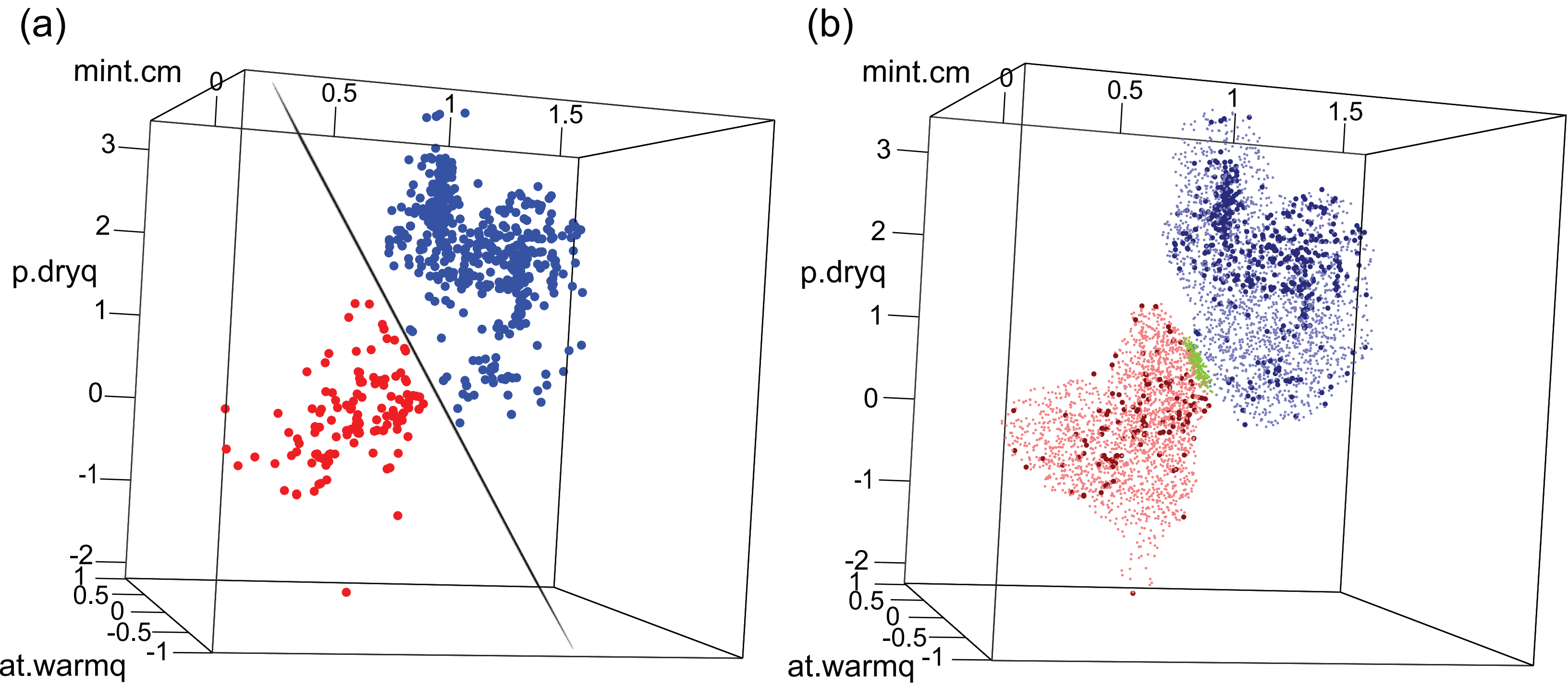
mee3_13363_f1.eps



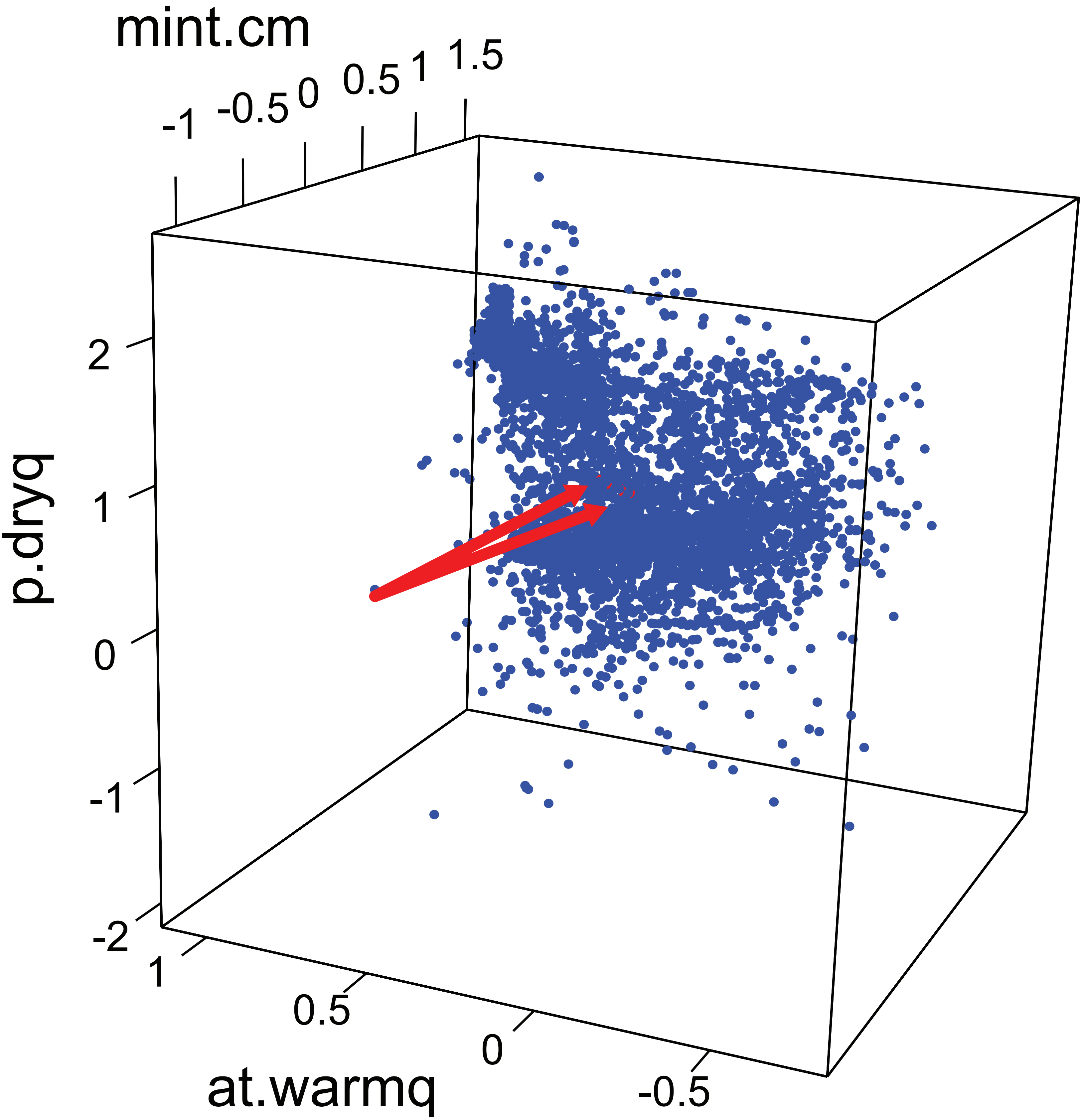
mee3_13363_f2.eps


Araucariaceae
Sciadopityaceae
Pinaceae
Taxaceae
Podocarpaceae
Cupressaceae
Cupressaceae
Taxaceae
Sciadopityaceae
Podocarpaceae
Araucariaceae
Pinaceae
Result
🟥 non-overlap (linear decision boundary)
🟧 non-overlap (curvilinear decision boundary)
🟦 overlap

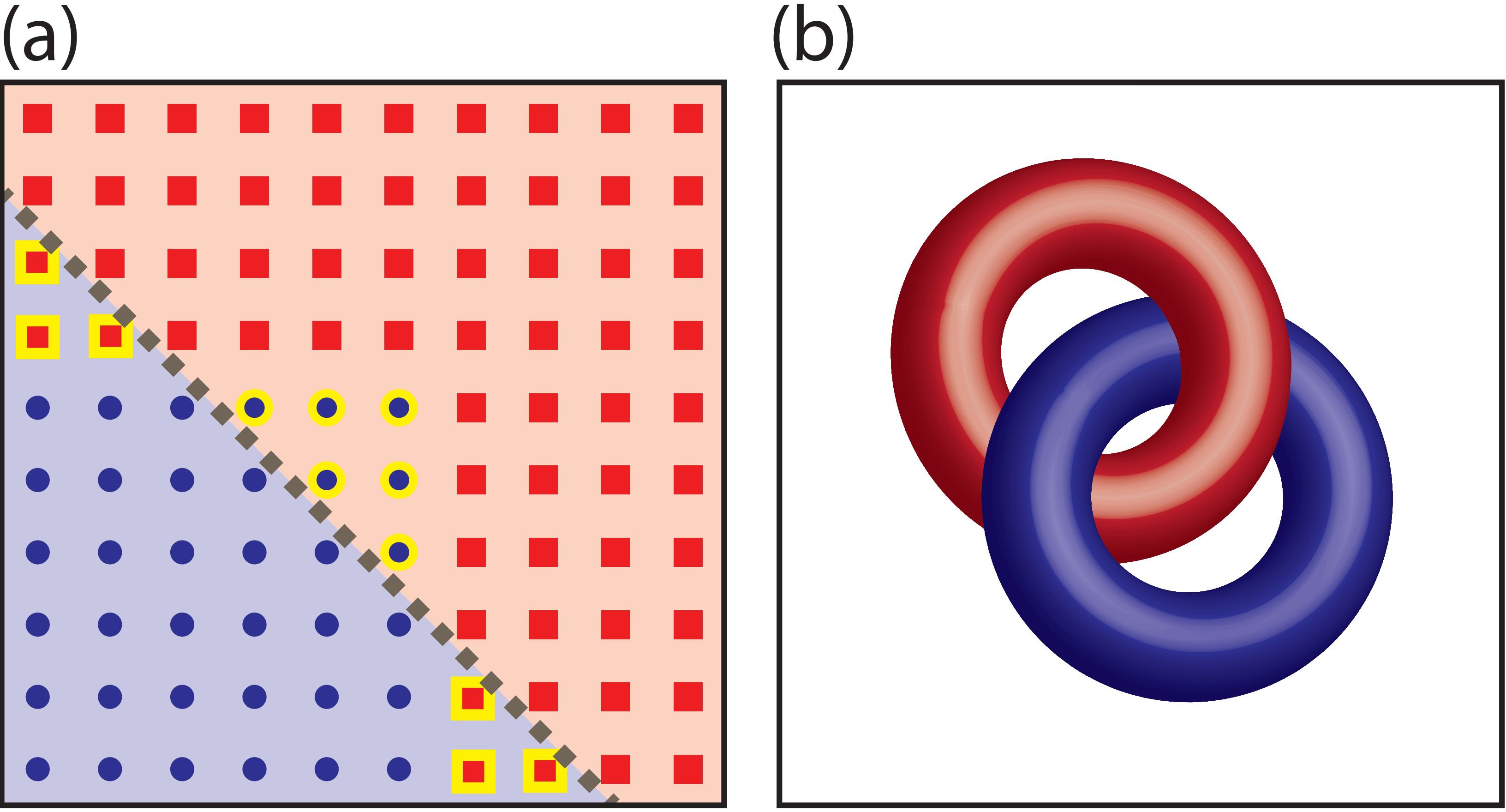

(a)
Cupressaceae
Taxac.
Sciadopityacae
Podocarpaceae
Araucariaceae
Pinaceae
(b)
fewer points in ecospace
more points in ecospace
Key
no conflict; entities overlap
no conflict; entities do not overlap
conflict
conflict: non-overlap found by HYPERVOLUME
conflict: curvilinear separation not found by HYPERVOLUME
conflict: linear separation not found by HYPERVOLUME
no conflict



mee3_13363_f5.eps



mee3_13363_f6.eps



mee3_13363_f7.eps