**Title**

A cautionary note on the use of hypervolume kernel density estimators in ecological niche modeling

**Authors**

Huijie Qiao[1*], Luis E. Escobar[2,3*], Erin E. Saupe[4], Liqiang Ji[1] and Jorge Soberón[5]

**Affiliations**

[1]*Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, 100101 China.*

[2]*Department of Veterinary Population Medicine, University of Minnesota, St. Paul, Minnesota, 55108 USA.*

[3]*Minnesota Aquatic Invasive Species Research Center, University of Minnesota, St. Paul, Minnesota, 55108 USA;*

[4]*Department of Geology & Geophysics, Yale University, New Haven, Connecticut, 06511 USA.*

[5]*Biodiversity Institute and Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas, 66045, USA.*

*Corresponding author: Luis E. Escobar, Phone: (+1) 315-313-8584. Email: lescobar@umn.edu. Address: Veterinary Diagnostic Laboratory, University of

Minnesota, 1365 Gortner Avenue, St. Paul, MN 55108.

## Running title

Kernel density methods in ecological niche modeling

## Abstract

Blonder *et al.* (2014) introduced a new multivariate kernel density estimation (KDE) method to infer Hutchinsonian hypervolumes to model ecological niches. The authors argued their KDE method matches or outperforms several methods for estimating hypervolume geometries and for conducting species distribution modeling. Further clarification, however, is appropriate with respect to the assumptions and limitations of KDE as a method for species distribution modeling. Using virtual species and controlled environmental scenarios, we show that KDE both under- and over-estimates niche volumes depending on the dimensionality of the dataset and the number of occurrence records considered. We suggest KDE may be a viable approach when dealing with large sample sizes, limited sampling bias, and only a few environmental dimensions.

**Key words**

Hutchinsonian hypervolumes; ecological space; multivariate kernel density estimation; virtual species; minimum volume ellipsoid; niche

## Introduction

In a recent contribution, Blonder *et al.* (2014) introduced a new hypervolume multivariate kernel density estimation (KDE) method to delineate Hutchinsonian hypervolumes (Hutchinson, 1957, 1978) in high-dimensional ecological space. A hypervolume, in this formulation, is defined by a set of points within an *n*-dimensional environmental or ecological space that reflects suitable values of these *n* variables. According to the authors, KDE outperforms several methods for estimating hypervolume geometries and for conducting species distribution modeling (SDM).

Blonder *et al.* (2014) argued KDE is useful for fitting observed occurrences to environmental values, and for recognizing clusters or holes in occurrence datasets within environmental space. Here, we show that KDE recognizes clusters or holes in occurrence datasets only when occurrence data are numerous and when the dimensionality of the environmental space is not too large. Indeed, the KDE method may have difficulty identifying holes, gaps, and/or clusters in environmental space with limited occurrences, since in this case the method tends to produce broad niche estimates that smooth out

these clusters and holes. Caution is warranted when applying KDE in high-dimensional space because of the curse of dimensionality (Hastie *et al.*, 2009, sections 2.5 and 6.3) and the empty space phenomenon (Silverman, 1986, section 4.5). That is to say, as dimensionality increases, the number of samples required to estimate a shape with accuracy will also increase dramatically.

In situations where KDE is able to recognize correctly clusters or holes in occurrence datasets, we argue that doing so is useful only to the extent that the realized niche (RN) is sought and not the fundamental niche (FN). Blonder and colleagues indicated their KDE method estimates 'holey' Hutchinsonian hypervolumes without *a priori* reason to assume that a hypervolume (or niche) should be normally or uniformly distributed in multiple dimensions (Fig. 1e in Blonder *et al*. 2014). We argue, however, that traditional Hutchinsonian hypervolumes would not fit tightly to available occurrence data, especially if one seeks the FN (a point also noted by Blonder and colleagues). Empirical and theoretical arguments suggest the FN has a convex shape (Birch, 1953; Maguire, 1973; Austin *et al.*, 1984; Colwell & Rangel, 2009; Araújo & Peterson, 2012; Drake, 2015), and, consequently, convex-hulls or ellipsoids (multivariate Gaussian shapes) may often be the simplest proxy (Peterson *et al.*, 2011).

Our argument is theoretical and emphasizes choosing the appropriate method for a particular application: if the RN is desired, the KDE method of Blonder *et al.* (2014) may be a good candidate, assuming low occurrence

density and high dimensionality does not prevent its practical application (Franklin, 2005; Hastie *et al*., 2009, sections 2.5 and 6.3). However, if the FN is to be estimated, the KDE method may not be ideal.

If the KDE method functions as Blonder and colleagues propose, producing strict estimates of the environmental space occupied by a species, model transferability to different regions or time periods – a common goal in SDM – will be limited. For example, say available occurrences for a species are distributed in temperatures of 15°, 16°, 17°, 19°, and 20°C. In this scenario, ignoring potential suitability for the species at 18°C, the 'hole' in the series, may be biologically unrealistic. As is likely the case in this simplistic example, many environmental holes in occurrence data may be due to biases in sampling, the availability of existing environmental conditions, and/or biological constraints, and do not reflect real suitability requirements.

Based on the considerations above, we re-evaluated the experiments of Blonder *et al.* (2014) using diverse FN shapes, including range-boxes (RB; Birch, 1953), convex-hulls (CH; Godsoe, 2010; Qiao *et al.*, 2015), and minimum-volume ellipsoids (MVE; Maguire, 1973; Qiao *et al.*, 2015), which have been previously invoked and employed in ecological studies. This reassessment identifies those tools that best fit with a particular and diverse set of research questions, and provides users with a rich source of information for selecting model approaches.

## Methods

### 1. KDE performance

We illustrate the functionality of the KDE method using a virtual environmental space, ***E***, composed of 10,000 unique random observations in two dimensions. Different configurations and densities of occurrences were sampled from a virtual FN within this environmental space, defined as a range box (red rectangle in Figs. 1, S1, S2). Note that the FN is easily estimated with virtual species based on controlled occurrence data, but observed occurrences from real species will most likely capture the RN and not the FN, which is constrained by biotic interactions, accessibility, and the available environment (Peterson *et al.* 2011). Within this virtual FN, we collected independent occurrence datasets of three different sample sizes ***m*** (***m*** = 10, 100, and 1000) and four different sampling configurations: (i) evenly distributed or unbiased (Figs. 1.a, S1.a, S2.a), clustered or biased (Figs. 1.b, S1.b, S2.b), (ii) absent from the center of the FN or "holey" (Figs. 1.c, S1.c, S2.c), and (iv) distributed in two distinct environmental clusters (Figs. 1.d, S1.d, S2.d). We repeated the sampling process 10 times to capture variation. Using these 120 sampling datasets (i.e., three sample sizes ***m*** x four sample configurations x 10 replicates), we estimated the virtual FN using the KDE approach and assessed the quality of these estimates based on the resulting type I (i.e., false presence, or incorrect rejection of a true null hypothesis) and type II error (i.e., false absence, or the failure to reject a false null hypothesis). Error was quantified as

the number of observations in the virtual space that were incorrectly predicted.

**2. Comparison of KDE to other algorithms**

We created three virtual FN configurations — RB, CH, and MVE — to explore quantitatively the performance of different modeling algorithms in estimating FNs. To create these virtual FNs, we first generated ***e*** uncorrelated virtual environmental variables (with ***e*** taking one of four possible values, ***e*** = 2,4,6, and 8), to create the environmental space, ***E***, composed of 10,000 unique random observations. Environmental values in ***E*** ranged between 0 and 1 in each of the eight dimensions (Fig. S3). Next, we selected 10 random observations (***N***) in ***E*** to define the vertices of the FNs under three shape hypotheses, ***N*** = RB, CH, and MVE. Environmental values used to define ***N*** were constrained between 0.2 and 0.8 to avoid potential novel environmental conditions (Fig. S3).

The environmental observations inside each of these virtual niches were regarded as the species' presences. For each virtual FN (i.e., RB, CH and MVE), we collected independent occurrence datasets of three sample sizes, ***m*** (***m*** = 10, 100, and 1000) in ***e*** environmental dimensions (***e*** = 2,4,6, and 8). This sampling process was repeated 10 times to generate random replicates of species' occurrences, which resulted in 360 simulations from the combination of three FN shape hypotheses x three sample sizes (***m***) x four environmental dimensions (***e***) x 10 random replicates.

To model the virtual FNs, we used the methods proposed by Blonder and colleagues (2014), including RB, CH, MVE, and KDE. As in Blonder *et al*. (2014), KDEs were inferred using a Silverman bandwidth estimator (Silverman, 1986, section 4.5) and a quantile threshold of 0.5. Note that smaller bandwidths (i.e., larger thresholds) will lead to smaller hypervolumes. As aptly noted by Blonder and colleagues, analyses that have few observations (***m***/***e*** < 10, as a rough guideline) will be sensitive to the choice of bandwidth.

Following Blonder *et al.* (2014), we used the volume of the niche to explore the amount of ***E*** predicted by the models. We compared the volume of the 'estimated niche' (***n***) with the known 'true' volume (***N***) of the virtual FN. Niche size measured as volume, however, may be insensitive to type I error, such that the 'true niche' and the 'estimated niche' may yield similar volumes but have minimal or no environmental overlap. To avoid this problem, we evaluated all models using sensitivity (Eq. S1) and specificity (Eq. S2) based on omission error (Fielding & Bell, 1997), and the Jaccard index (Eq. S3) based on comparisons between the known (***N***) and estimated (***n***) niche volumes (Jaccard, 1912; Godsoe, 2014).

## Results

### 1. KDE performance

The KDE method tended to overestimate the FN and extend beyond the

occurrence data (i.e., the RN) when using small sample sizes. The severity of this overestimation, however, varied depending on the sample configuration (Figs.S1, S2). KDE identified the 'hole' (black box; Figs. S1c, S2c) only under the largest sample size (Fig. 1). The 'clusters' were identified with sample sizes over 100 (Figs. 1, S2), but in these instances, KDE estimates extended significantly beyond the FN and the RN. In general, type I error decreased and type II error increased when more occurrences were used for model calibration (Fig.2).

**2. Comparison of KDE to other algorithms**

In most cases, the KDE algorithm overestimated the volume of the true FN when the shape of the niche was defined as RB (Fig. S4.a), a result congruent with that of Blonder and colleagues (2014). The RB and CH algorithms returned the most variable niche volume estimates, with consistent underestimation of FN volumes. These two algorithms, however, obtained the highest Jaccard similarity values between the estimated and observed RB FN (Fig. 3.a), particularly in high-dimensional environmental space. When the virtual FN was defined as CH, MVE algorithm got the highest Jaccard similarity values in the low dimension (***e***=2), KDE was the winner in the middle dimension (***e*** = 4), and RB won in the high dimension (***e***=6, 8, Fig. 3.a). When the virtual FN was defined as MVE, we failed to replicate the results of Blonder and colleagues, who found that MVE consistently overestimated niche

193 volumes (Fig. 4c in Blonder *et al.*, 2014, our figure Fig. S4.c).

194 Overall, method performance varied as a function of the ‘true shape’ of the
195 virtual niche. That is to say, the RB method performed best when the true
196 shape was RB, and so forth. In general, CH tended to underestimate true
197 niche volumes. Similarly, MVE and RB underestimated true niche volumes
198 using small sample sizes, but overestimated niche volumes using larger
199 sample sizes. KDE tended to underestimate volumes of niches in high
200 dimensionality and overestimate volumes of niches in low dimensionality (Fig.
201 S4).

202 All methods performed well in terms of specificity and sensitivity using
203 large sample sizes (***m*** = 100, 1000). Results for smaller sample sizes (***m*** = 10),
204 however, were more variable. When considering sensitivity, KDE performed
205 well, as this method tends to generate broader niche estimates (Fig. S5).
206 Broader niche estimates, however, will generate more opportunities for type I
207 error, resulting in lower specificity values. Indeed, the KDE method performed
208 worst in terms of specificity for small sample sizes, whereas the CH method
209 performed best. Overall, the CH method performed well in terms of specificity,
210 but performed poorly when considering sensitivity (Fig. S6). As dimensionality
211 increased, the KDE method exhibited decreased sensitivity but increased
212 specificity, and underestimated the true volume of the niche. In other words,
213 estimates were constrained severely in high dimensions.
214

## Discussion

Our results suggest that accuracy on niche estimations depends on the research question and particularities of the data. A complex algorithm, such as KDE, may function best when the goal is to fit models tightly to available data and avoid environmental interpolation across ‘holes’ in environmental space. These are often desirable features when exploring the occupied area or RN of a species, or the distribution of non-living organisms (e.g., when mapping potential wildfires). KDE, however, is sensitive to both sample size and environmental dimensionality. Contrary to the claims of Blonder and colleagues, KDE may overestimate niche volumes in low dimensions and constrict niche volume estimates in high dimensions. We found that as dimensionality increases, specificity increases as sensitivity decreases (Drake 2015; Figs. S5, S6).

The MVE algorithm performed best when the target shape is ellipsoid in nature, which is often hypothesized to be the true shape of species’ FNs (Hutchinson, 1957; Maguire, 1973; Brown, 1984; Drake, 2015). The CH method tended to generate narrow niche estimates relative to the KDE method, as reflected in the specificity and sensitivity values. The CH algorithm may be suitable when the goal is to estimate suitable environmental conditions allowing environmental interpolation, but avoiding prediction of suitable conditions in novel environments.

The analyses conducted herein support the idea that there is often not a

single ‘best’ algorithm or method that fits with all ecological applications and data configurations for estimating species’ niches (Guillera-Arroita *et al.*, 2015; Qiao *et al.*, 2015). As is now common practice in phylogenetics, we propose that the best niche model should be selected from a variety of model hypotheses, based on its fit to the nature of the data and the specific research question (Diniz-Filho *et al.*, 2015).

## Abbreviations

**CH** Convex-hull.

***E*** The environmental variables forming an environmental space.

***e*** The number of environmental variables used to create the environmental space.

**FN** Fundamental niche.

**KDE** Multivariate kernel density method.

***N*** ‘True’ virtual niche.

***n*** Estimation of the ‘true’ fundamental niche from a sample of occurrences.

**MVE** Minimum-volume ellipsoid.

***m*** Number of occurrences collected from the environmental space used to estimate ***N***, the true virtual niche.

**RN** Realized niche.

**RB** Range-box.

**SDM** Species distribution modeling.


## Acknowledgements

We thank editors Richard Field and Antoine Guisan and two anonymous reviewers for invaluable suggestions that improved this manuscript. HQ was supported by National Natural Science Foundation of China (A New Method to Predict the Species Distributions, 31100390). LEE was supported by the Minnesota Environment and Natural Resources Trust Fund, the Minnesota Aquatic Invasive Species Research Center, and the Clean Water Land and Legacy. JS was partially supported by NSF grant 1208472


## References


Araújo, M.B. & Peterson, A.T. (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology,* **93**, 1527-1539.

Austin, M.P., Cunningham, R.B. & Fleming, P.M. (1984) New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. *Plant Ecology*, **55**, 11-27.

Birch, L.C. (1953) Experimental background to the study of the distribution and abundance of insects: III. The relation between innate capacity for

increase and survival of different species of beetles living together on the same food. *Evolution*, **7**, 136-144.

Blonder, B., Lamanna, C., Violle, C. & Enquist, B.J. (2014) The n-dimensional hypervolume. *Global Ecology and Biogeography*, **23**, 595-609.

Brown, J.H. (1984) On the relationship between abundance and distribution of species. *American Naturalist*, **124**, 255-279.

Colwell, R.K. & Rangel, T.F. (2009) Hutchinson's duality: The once and future niche. *Proceedings of the National Academy of Sciences USA,* **106**, 19651-19658.

Diniz-Filho, J.A.F., Rodrigues, H., Telles, M.P.D.C., Oliveira, G.D., Terribile, L.C., Soares, T.N. & Nabout, J.C. (2015) Correlation between genetic diversity and environmental suitability: taking uncertainty from ecological niche models into account. *Molecular Ecology Resources*, **15**, 1059-1066.

Drake, J.M. (2015) Range bagging: A new method for ecological niche modelling from presence-only data. *Journal of The Royal Society Interface*, **12**, 10.1098/rsif.2015.0086.

Fielding, A. & Bell, J. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38-49.

Franklin, J. (2005) The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, **27**: 83-85.

Godsoe, W. (2010) I can't define the niche but I know it when I see it: A formal link between statistical theory and the ecological niche. *Oikos*, **119**, 53-60.

Godsoe, W. (2014) Inferring the similarity of species distributions using species' distribution models. *Ecography*, **37**, 130-136.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R. & Wintle, B.A. (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, **24**, 276-292.

Hastie, T.R. Tibshirani, & Friedman J. (2009). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Second Edition. Springer *New York*.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall. *London*.

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415-427.

Hutchinson, G.E. (1978) *An introduction to Population Ecology*. Yale University Press, New Haven.

Jaccard, P. (1912) The distribution of the flora in the alpine zone. *New Phytologist*, **11**, 37-50.

Maguire, B.Jr (1973) Niche response structure and the analytical potentials of its relationship to the habitat. *American Naturalist*, **107**, 213-246.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M.B. (2011) *Ecological Niches and Geographic Distributions*. Princeton University, Princeton and Oxford.

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

## Biosketch

Research interests of the team include invasion ecology, virtual ecology, and the evaluation of ecological niche modeling methods in ecology and epidemiology.

## Figures

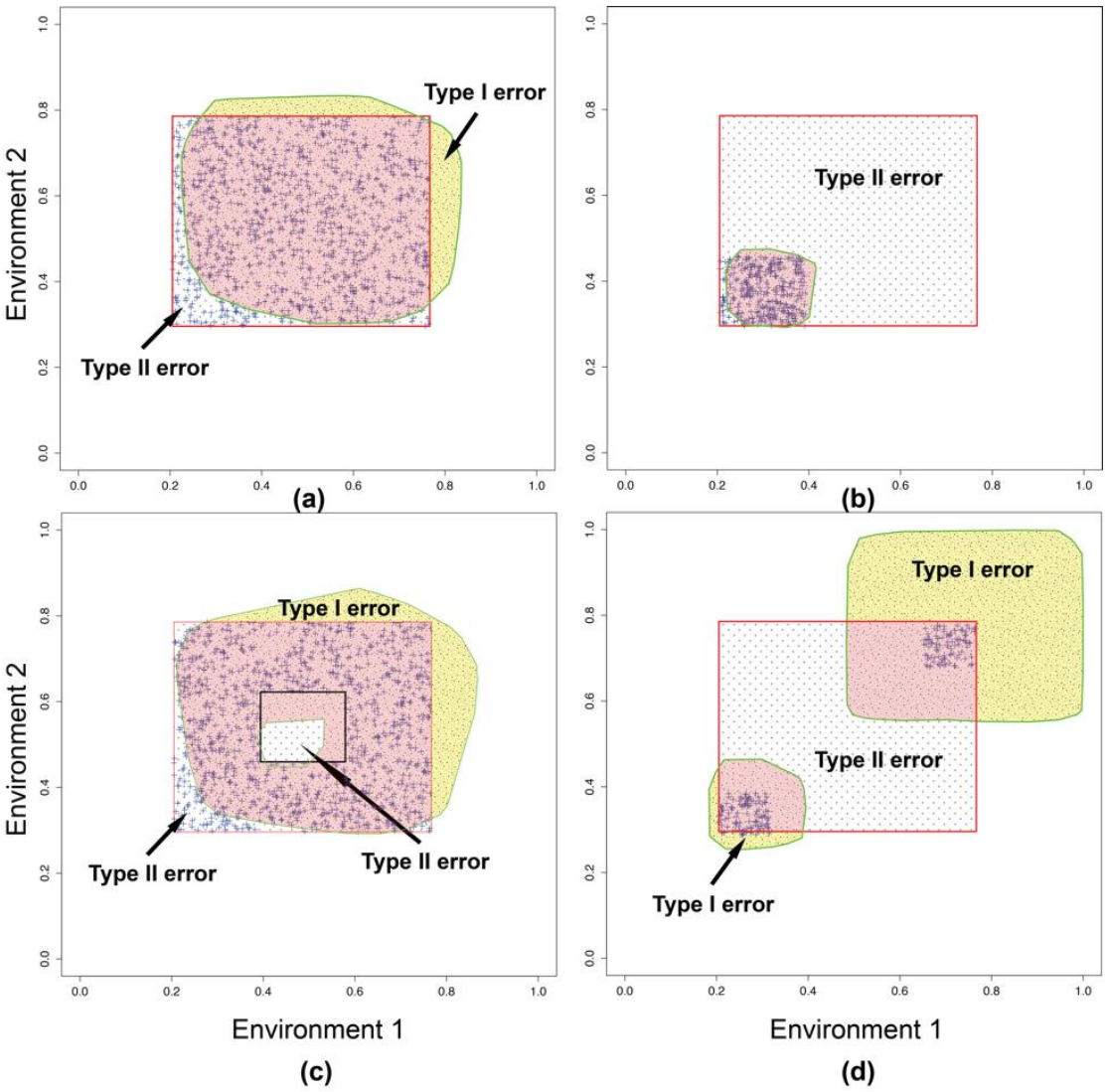

Type I error
Type II error
(a)
Type II error
(b)
Type I error
Type II error
Type II error
Environment 2
(c)
Type I error
Type II error
Type I error
Environment 1
(d)

**Figure 1. Type I and type II errors resulting from the KDE method using small sample sizes of *m* = 1000.** The red rectangle denotes a virtual fundamental niche (FN), while the blue points represent unbiased (a), biased (b), "holey", as indicated by the black rectangle (c), and (d) two-clustered observations of the virtual FN. The green polygons are the estimated niche from the KDE method based on the blue observations. The overlap (pink) of the virtual FN and the estimated niche is the portion of virtual FN correctly predicted by the KDE method. The yellow area (b and d) outside of the virtual FN denotes type I error resulting from the KDE method. The white area with cross-shading denotes type II error resulting from the KDE method. Note that abundant occurrences reduce type I error at the cost of increased type II error.

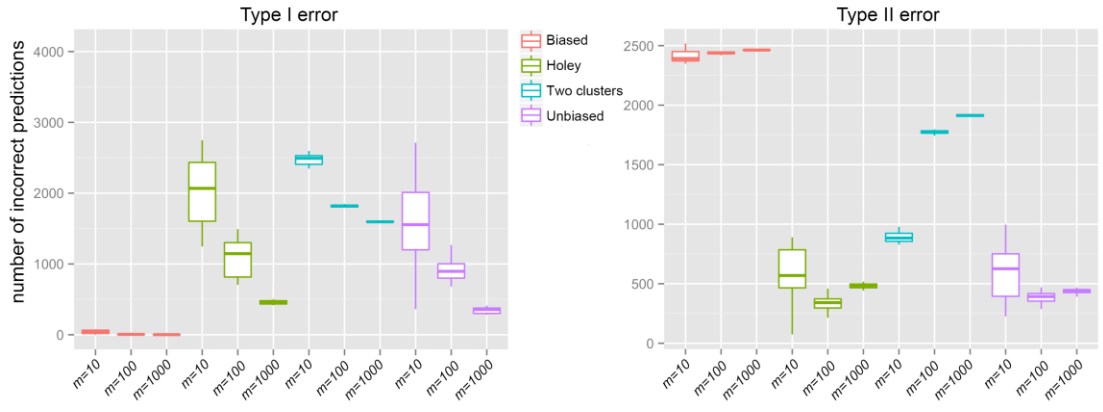
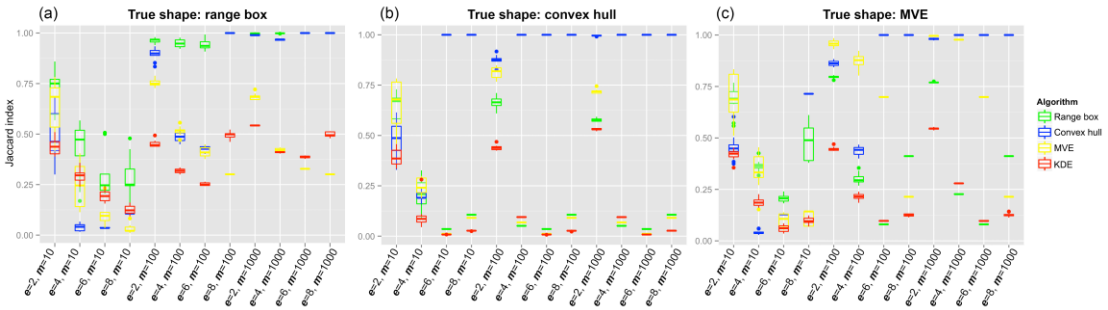



**Figure 2. Type I and II error for different sampling configurations estimated using the multivariate kernel density estimation (KDE) method.** Left: type I error based on comparisons of the 'true' and estimated niche under unbiased (purple; Figs. 1.a, S1.a, S2.a), biased (red; Figs. 1.b, S1.b, S2.b), holey (green; Figs. 1.c, S1.c, S2.c), and two-clustered (blue; Figs. 1.d S1.d, S2.d) sampling configurations. Estimates are based on ten sampling replicates of 10, 100, and 1000 occurrences (***m***). Right: type II error from the same study

355 design. The *y*-axis indicates the number of false observations (left; type I error)

356 and the number of false negatives (right; type II error).

357




358

359 **Figure 3. Jaccard index for each modeling method based on the different**

360 **virtual fundamental niche shapes.** Fundamental niches (FN) were

361 represented as single hypercubes or range boxes (a), convex-hulls (b), and

362 ellipsoids (c). To estimate these 'true' FNs, we used four modeling methods:

363 range-box (RB; green), convex-hull (CH; blue), minimum-volume ellipsoid

364 (MVE; yellow), and multivariate kernel density estimation (KDE; red). Each

365 boxplot represents the Jaccard index of the niche according to 10 independent

366 subsamples of observations (***m*** = 10, 100, 1000) collected randomly in a 2- to

367 8-dimensional dataset (***e***). Boxes closer to the top indicate better predictions (***n***)

368 in the form of high similarity or overlap between estimated (***n***) and 'true' virtual

369 fundamental niches (***N***).

370

371

372