Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2017) **26,** 1066–1070

A Journal of Macroecology

Global Ecology and Biogeography

CORRESPONDENCE

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

ABSTRACT

Blonder *et al.* (2014, Global Ecology and Biogeography, 23, 595–609) introduced a new multivariate kernel density estimation (KDE) method to infer Hutchinsonian hypervolumes in the modelling of ecological niches. The authors argued that their KDE method matches or outperforms several methods for estimating hypervolume geometries and for conducting species distribution modelling. Further clarification, however, is appropriate with respect to the assumptions and limitations of KDE as a method for species distribution modelling. Using virtual species and controlled environmental scenarios, we show that KDE both under- and overestimates niche volumes depending on the dimensionality of the dataset and the number of occurrence records considered. We suggest that KDE may be a viable approach when dealing with large sample sizes, limited sampling bias and only a few environmental dimensions.

Keywords

Ecological space, Hutchinsonian hypervolumes, minimum volume ellipsoid, multivariate kernel density estimation, niche, virtual species.

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 modelling (SDM).

Blonder *et al.* (2014) argued that 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 only recognizes clusters or holes in occurrence datasets 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 & Friedman, 2009, sections 2.5 & 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 accurately estimate a shape 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 *et al.* 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 *et al.*). Empirical and theoretical arguments suggest that 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 choice of 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 & Friedman, 2009, sections 2.5 & 6.3). However, if the FN is to be estimated, the KDE method may not be ideal.

If the KDE method functions as Blonder *et al.* propose, producing strict estimates of the environmental space occupied by a species, transferability of the model 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 its likely to be 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

DOI: 10.1111/geb.12492

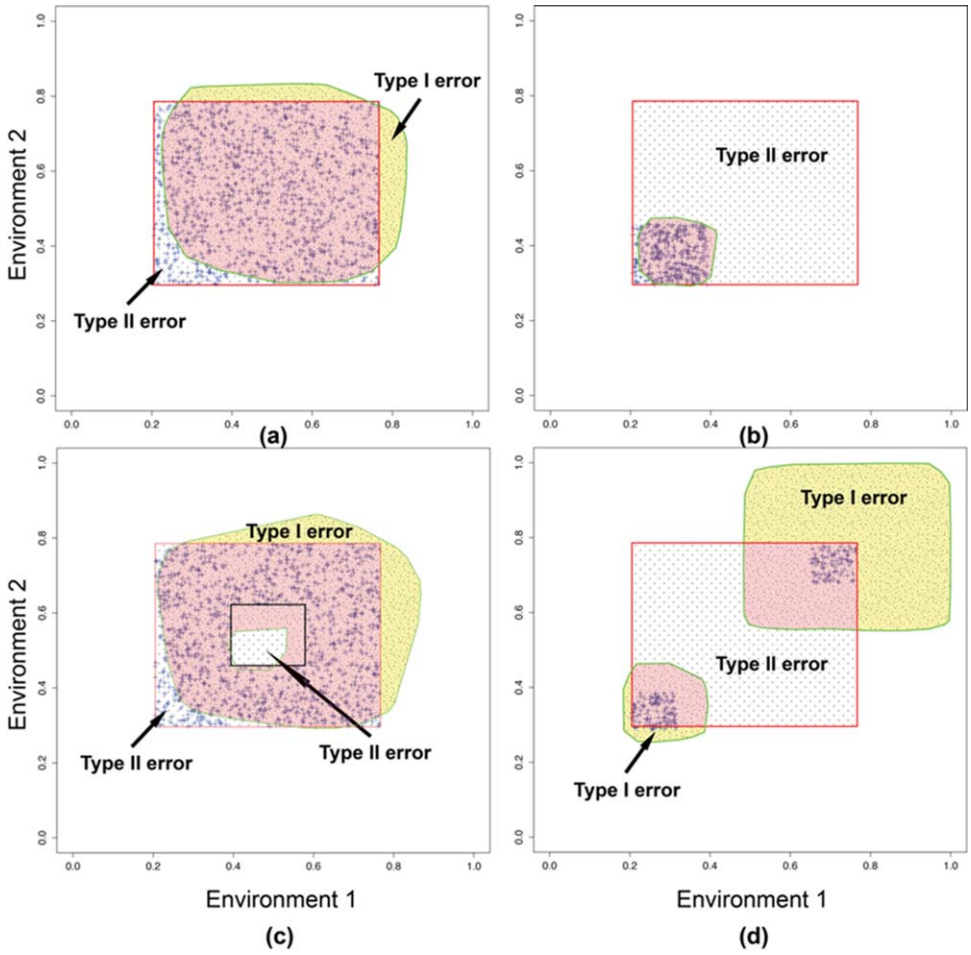

Figure 1 Type I and II errors resulting from the KDE method using small sample sizes of $m = 1000$. The (red) rectangle denotes a virtual fundamental niche (FN), while the (blue) points represent unbiased (a), biased (b), 'holey', as indicated by the inner (black) rectangle in (c), and (d) two-clustered observations of the virtual FN. The (green) polygons are the estimated niche from the KDE method based on the (blue) observation points. The overlap (pink) of the virtual FN and the estimated niche is the portion of virtual FN correctly predicted by the KDE method. The shaded (yellow) area (in b and d) outside the virtual FN denotes Type I error resulting from the KDE method. The white area with dotted 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.

provides users with a rich source of information for selecting model approaches.

METHODS

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 Fig. 1 and Figs. S1 & S2 in the Supporting Information]. 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 for three different sample sizes m ($m = 10$, 100 and 1000) and four different sampling configurations: (1) evenly distributed or unbiased (Figs. 1a, S1a & S2a), (2) clustered or biased (Figs. 1b, S1b & S2b), (3) absent from the centre of the FN or 'holey' (Figs. 1c, S1c & S2c), and (4)distributed in two distinct environmental clusters (Figs. 1d, S1d & S2d). We repeated the sampling process 10 times to capture variation. Using these 120 sampling datasets (i.e. 3 sample sizes $m \times 4$ sample configurations $\times$ 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.

Comparison of KDE with other algorithms

We created three virtual FN configurations – RB, CH and MVE – to explore quantitatively the performance of different modelling 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 species presences. For each virtual FN (i.e. RB, CH and MVE), we collected independent occurrence datasets for 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 3 FN shape hypotheses $\times$ 3 sample sizes (m) $\times$ 4

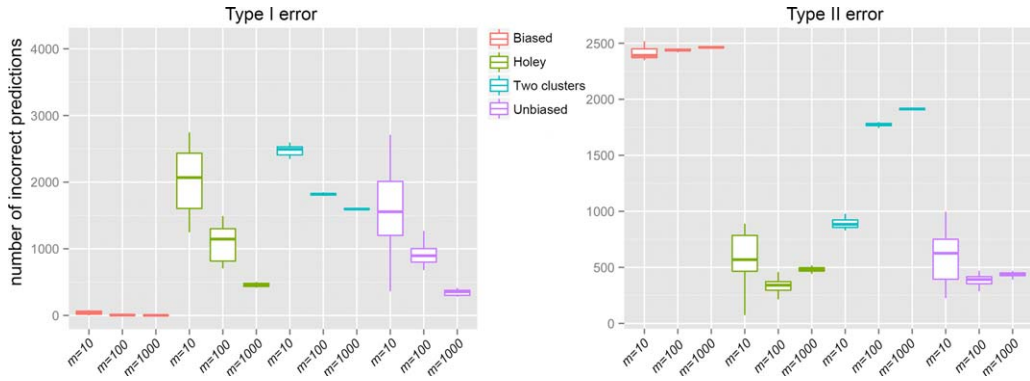

Figure 2 Type I and II error for different sampling configurations estimated using the multivariate kernel density estimation (KDE) method. Left: Type I error based on comparisons of the 'true' and estimated niche under unbiased (purple; Figs 1a, S1a & S2a), biased (red; Figs 1b, S1b & S2b), holey (green; Figs 1c, S1c & S2c), and two-clustered (blue; Figs 1d, S1d & S2d) sampling configurations. Estimates are based on 10 sampling replicates of 10, 100 and 1000 occurrences (m). Right: Type II error from the same study design. The y-axis indicates the number of false observations (left; Type I error) and the number of false negatives (right; Type II error).

environmental dimensions (e) × 10 random replicates.

To model the virtual FNs, we used the methods proposed by Blonder *et al.* (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 *et al.* (2014), 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 (equation S1) and specificity (equation S2) based on omission error (Fielding & Bell, 1997), and the Jaccard index (equation S3) based on comparisons between the known (N) and estimated (Blonder *et al.* (2014) niche volumes (Jaccard, 1912; Godsoe, 2014).

RESULTS

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).

Comparison of KDE with 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. S4a), a result congruent with that of Blonder *et al.* (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. 3a), particularly in high-dimensional environmental space. When the virtual FN was defined as CH, the MVE algorithm got the highest Jaccard similarity values in the low dimension ($e = 2$), KDE performed best in the middle dimension ($e = 4$), and RB in the high dimension ($e = 6$, 8, Fig. 3a). When the virtual FN was defined as MVE, we failed to replicate the results of Blonder *et al.* (2014), who found that MVE consistently overestimated niche volumes (Fig. 4c in Blonder *et al.*, 2014; our Fig. S4c).

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

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

DISCUSSION

Our results suggest that accuracy of 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

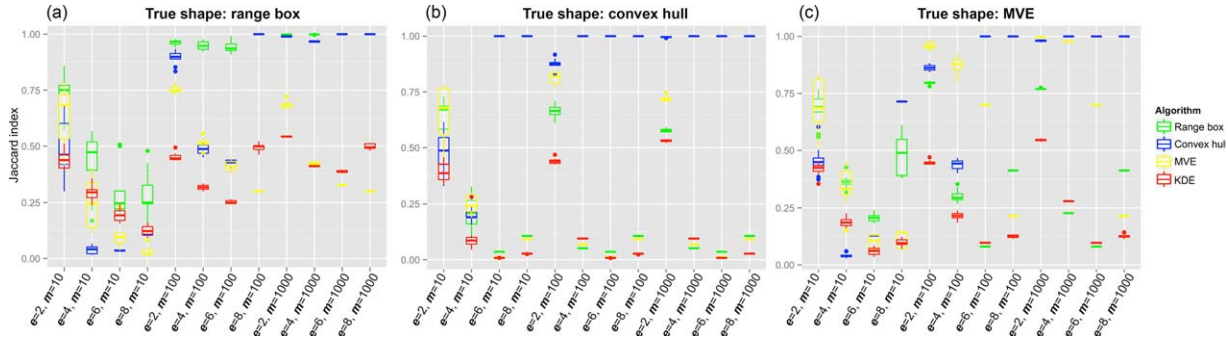

Figure 3 Jaccard index for each modelling method based on the different virtual fundamental niche shapes. Fundamental niches (FN) were represented as single hypercubes or range boxes (a), convex-hulls (b), and ellipsoids (c). To estimate these 'true' FNs, we used four modelling methods: range box (RB; green), convex hull (CH; blue), minimum-volume ellipsoid (MVE; yellow), and multivariate kernel density estimation (KDE; red). Each boxplot represents the Jaccard index of the niche according to 10 independent subsamples of observations ($m = 10$, 100, 1000) collected randomly in a two- to eight-dimensional dataset (e). Boxes closer to the top indicate better predictions (n) in the form of high similarity or overlap between estimated (n) and 'true' virtual fundamental niches (N).

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).

ACKNOWLEDGEMENTS

We thank the editors Richard Field and Antoine Guisan and two anonymous referees for invaluable suggestions that improved this manuscript. H.Q. was supported by the National Natural Science Foundation of China (A New Method to Predict the Species Distributions, 31100390). L.E.E. 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. J.S. was partially supported by NSF grant 1208472. Research interests of the team include invasion ecology, virtual ecology, and the evaluation of ecological niche modeling methods in ecology and epidemiology.

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

[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, MN 55108, USA,* [3]*Minnesota Aquatic Invasive Species Research Center, University of Minnesota, St Paul, MN 55108, USA,* [4]*Department of Geology & Geophysics, Yale University, New Haven, CT 06511, USA,* [5]*Biodiversity Institute and Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KN 66045, USA*

*Correspondence: Luis E. Escobar, Veterinary Diagnostic Laboratory, University of Minnesota, 1365 Gortner Avenue, St Paul, MN 55108, USA.
E-mail: lescobar@umn.edu

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. *The 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.T. & Friedman, J, (2009) *The elements of statistical learning. Data mining, inference and prediction*, 2nd edn. Springer, New York.

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, CT.

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

Maguire Jr, B. (1973) Niche response structure and the analytical potentials of its relationship to the habitat. *The 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, NJ.

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.

Silverman, B.W. (1986) *Density estimation for statistics and data analysis*. Chapman and Hall, London.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site:

Figure S1 Type I and II errors resulting from the multivariate kernel density estimation method using small sample sizes of $m = 10$.
Figure S2 Type I and II errors resulting from the kernel density estimation method using sample sizes of $m = 100$.
Figure S3 First two dimensions of the eight-dimensional virtual fundamental niches.
Figure S4 Volume of the niche models.
Figure S5 Sensitivity of each method based on different virtual fundamental niche shapes.
Figure S6. Specificity of each method based on different virtual niche shapes.
Appendix S1 R functions to generate minimum-volume ellipsoids, including the sample data and the experiments used in this contribution.
Appendix S2 Equations used for model evaluation.

Editor: Antoine Guisan

doi: 10.1111/geb.12492