Multiple data sources and freely available code is critical when investigating species distributions and diversity: a response to Knouft (2018)

Tad Dallas,[1,2,3]* Robin R Decker[1,2] and Alan Hastings[1,2]

[1]Center for Population Biology University of California Davis, CA 95616, USA

[2]Department of Environmental Science and Policy University of California Davis, CA 95616, USA

[3]Centre for Ecological Change University of Helsinki Finland, FI-00014, USA

*Correspondence: E-mail: tad.a.dallas@gmail.com

Abstract

A recent comment from Knouft (2018) has suggested that our original article (Dallas *et al.* 2017) was an 'inappropriate application of biodiversity data'. Here, we affirm our results, and address the more general point about biodiversity data use.

Keywords

abundant-centre hypothesis, climatic niche, eBird, forest Inventory and Analysis, geographic range, macroecology.

Ecology Letters (2018)

A recent paper suggested that the relationship between a species geographic range or climatic niche centre was largely unrelated to population density (Dallas *et al.* 2017), a prevailing biogeographical pattern that is at the foundation of many ecological hypotheses (Sagarin & Gaines 2002). Knouft (2018) is concerned that the data used for assessing *distance–abundance* relationships in fish species – which accounted for less than 5% of examined species – suffered from biases and were therefore unsuitable for use, suggesting that *distance-abundance* relationships may apply for freshwater taxa. The main concerns of Knouft (2018) were that the data used (1) may include non-native or stocked fish species, (2) does not reflect the actual range of species, (3) represent pseudoreplicated samples.

First, fish species stocked to support recreational fisheries certainly pose an issue for detecting *distance–abundance* relationships, in much the same way differential fishing pressure could drive down certain populations. However, the claim that baitfish introductions and stocking are the reason for the lack of *distance–abundance* relationships observed is premature, as there are many causal pathways to reach our conclusions, and we also observed a pronounced lack of support in species not typically subjected to stocking or take. Incorporating species traits and land-use changes into the study of species abundance patterns represents an interesting future step, as it allows researchers to determine the relative effects of climate and other factors (e.g. habitat fragmentation, human-mediated transport, etc.).

Second, Knouft (2018) suggest that the narrow sampling of fish species could result in the lack of observed *distance–abundance* relationships. This is a concern, which we attempted to address (see supplement of Dallas *et al.* (2017)) by quantifying geographic range and climatic niche centroids, using species occurrence data from the Global Biodiversity Information Facility, relating species geographic range size and occurrence number to *distance–abundance* relationship slope to determine the potential effect of sampling or geographical bias, and acquiring data from BirdLife International on migratory status to examine the effect of bird migratory status on *distance–abundance* relationships. Geographic range estimation of populations embedded in a metapopulation, where much of the range of inhospitable, is a clear concern – and a point raised in Knouft & Page (2011) – but the calculating range both in terms of sampled populations and GBIF records accounts for this effect as well as possible.

Lastly, Knouft (2018) expressed concern that we used multiple samples of population density from the same lakes. This potentially stems from a lack of clarity in the original article. When sites were repeatedly sampled, we took the mean value for each unique latitude and longitude coordinate. This procedure was used for all data sources. However, we recognise that multiple samples can come from the same lake, but have slightly different geographic coordinates. We explore this in the supplement, where we compare aggregation of samples by rounding geographic coordinates to quantify the number of unique localities. We show that (1) pseudo-replication did not take place, and (2) the number of species for which sufficient data were available did not change substantially when aggregating data at coarser scales.

Ecological theory built on a small number of observational points – like many macroecological relationships – should be evaluated with the best possible data. Our effort combined data from governmental surveys, citizen science efforts, published literature estimates, and museum specimens to provide the most comprehensive test of *distance–abundance* relationships. While we agree with Knouft (2018) that biodiversity data needs to be used appropriately, we also believe it needs

to be used. We have made every possible effort to programmatically access and clean data, combine multiple data streams of different quality, and provide all code to reproduce our original results (https://doi.org/10.6084/m9.figshare.5023232.v2) and the results of this supplemental analysis (https://doi.org/10.6084/m9.figshare.6444608). This will hopefully enable researchers to revisit these analyses once more or better quality data are available. In summary, we believe our original findings are robust and represent a good example of how biodiversity data from multiple sources can be combined to provide thorough tests of existing ecological theory.

ACKNOWLEDGEMENTS

DATA ACCESSIBILITY STATEMENT

R code is available on figshare to reproduce the original analyses at https://doi.org/10.6084/m9.figshare.5023232 and to create the additional analyses https://doi.org/10.6084/m9.figshare.6444608. Data are available for eBird data (Sullivan *et al.* 2009), EPA-EMAP data (https://www.epa.gov/emap/), NAWQA data (Knouft & Anthony (2016); https://water.usgs.gov/nawqa), Forest Inventory and Analysis data (Woudenberg *et al.* 2010) (https://www.fia.fs.fed.us/), and the mammal community database Thibault *et al.* (2011). While authors should cite the original data sources, we also provide data used in the analyses and analytic code.

REFERENCES

Dallas, T., Decker, R.R. & Hastings, A. (2017). Species are not most abundant in the centre of their geographic range or climatic niche. *Ecol. Lett.*, 20, 1526–1533.

Knouft, J.H. (2018). Appropriate application of information from biodiversity databases is critical when investigating species distributions and diversity: a comment on dallas et al. (2017). *Ecol. Lett.* 21, 1119–1120

Knouft, J.H. & Anthony, M.M. (2016). Climate and local abundance in freshwater fishes. *Royal Soc. Open Sci.*, 3, 160093.

Knouft, J.H. & Page, L.M. (2011). Assessment of the relationships of geographic variation in species richness to climate and landscape variables within and among lineages of north american freshwater fishes. *J. Biogeogr.*, 38, 2259–2269.

Sagarin, R.D. & Gaines, S.D. (2002). The "abundant centre" distribution: to what extent is it a biogeographical rule? *Ecol. Lett.*, 5, 137–147.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009). ebird: A citizen-based bird observation network in the biological sciences. *Biol. Cons.*, 142, 2282–2292.

Thibault, K.M., Supp, S.R., Giffin, M., White, E.P. & Ernest, S. (2011). Species composition and abundance of mammalian communities. *Ecology*, 92, 2316.

Woudenberg, S.W., Conkling, B.L., OConnell, B.M., LaPoint, E.B., Turner, J.A. & Waddell, K.L. (2010). The forest inventory and analysis database: Database description and users manual version 4.0 for phase 2.Gen. Tech. Rep. RMRS-GTR-245. Fort Col lins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.