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

Tad Dallas[a,b,c,*], Robin R Decker[a,b] and Alan Hastings[a,b]

[a] *Center for Population Biology, University of California, Davis, CA 95616*

[b] *Department of Environmental Science and Policy, University of California, Davis, CA 95616*

[c] *Centre for Ecological Change, University of Helsinki, Finland FI-00014*

*Corresponding author: tad.a.dallas@gmail.com

Article type: Technical comment

Running title: No support for distance-abundance pattern

Data accessibility: *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.

Keywords: Geographic range, climatic niche, eBird, Forest Inventory and Analysis, macroecology, abundant-center hypothesis

Words in text: 650

Abstract

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

A recent paper suggested that the relationship between a species geographic range or climatic niche center 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) do not reflect the actual range of the 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 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 recognize 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

We sincerely thank all individuals associated with the collection and curation of the open data on species abundances (FIA, EPA-EMAP, MCDB, eBird), species occurrences (GBIF), species traits (Pantheria, fishbase, Amniote database), and phylogenies (bird and mammal supertrees) used in this manuscript. The study has been supported by the TRY initiative on plant traits (http://www.tryŋdb.org).

References

Dallas, T., Decker, R.R. & Hastings, A. (2017). Species are not most abundant in the centre of their geographic range or climatic niche. *Ecology Letters*, 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). *Ecology Letters*.

Knouft, J.H. & Anthony, M.M. (2016). Climate and local abundance in freshwater fishes. *Royal Society Open Science*, 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. *Journal of Biogeography*, 38, 2259–2269.

Sagarin, R.D. & Gaines, S.D. (2002). The "abundant centre" distribution: to what extent is it a biogeographical rule? *Ecology Letters*, 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. *Biological Conservation*, 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, 2316–2316.

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 Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station.*

Supplemental Material

Tad Dallas[a,b,c,*], Robin R Decker[a,b], Alan Hastings[a,b]

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

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

[c] Centre for Ecological Change, University of Helsinki, Finland

[*] tad.a.dallas@gmail.com

Removing potentially pseudo-replication

In the main text of this comment, we address the claim of non-independence in species density estimates due to the same locality being sampled multiple times. This was, in part, our mistake for not being clear in the methods section of the original article. Here, we examine the effect of spatial resolution and multiple sampling to estimate how many species still satisfy our criteria of at least 10 unique sampled sites. Code to reproduce the main text analyses is available at `https://doi.org/10.6084/m9.figshare.5023232.v2`, and the supplemental analyses contained here at `https://doi.org/10.6084/m9.figshare.6444608`.

Table S1: The effect of binning multiple abundance measures by geographic coordinates. The total number of occurrences (n) represents locations that have been sampled multiple times, presenting a potential pseudoreplication issue. We can find unique localities with the precision of either 3 (n_3), 2 (n_2), or 1 (n_1) decimal degrees by rounding latitude and longitude coordinates and taking the mean species density value for non-unique localities.

Species	Total occurrences (n)	Fine (n_3)	Moderate (n_2)	Coarse (n_1)
Alosa pseudoharengus	19	7	7	7
Ambloplites rupestris	20	20	20	20
Ameiurus natalis	39	39	39	37
Ameiurus nebulosus	128	85	84	82
Anguilla rostrata	44	44	44	44
Campostoma anomalum	59	12	12	12
Campostoma oligolepis	211	25	25	24
Catostomus catostomus	38	5	5	5
Catostomus commersoni	171	26	26	26
Cottus carolinae	143	27	27	24
Cottus cognatus	36	7	6	6
Couesius plumbeus	1217	128	128	117
Cyprinella analostoma	479	110	109	97
Cyprinella spiloptera	69	28	28	27
Cyprinella venusta	151	150	150	143
Cyprinus carpio	80	13	13	13
Enneacanthus gloriosus	102	30	30	29

Enneacanthus obesus	100	100	100	98
Erimyzon oblongus	47	11	11	11
Esox americanus	84	82	82	75
Esox lucius	80	80	80	78
Esox niger	36	36	36	35
Etheostoma blennioides	25	7	7	7
Etheostoma caeruleum	594	91	90	84
Etheostoma flabellare	88	88	88	81
Etheostoma olmstedi	27	9	9	8
Fundulus diaphanus	165	36	36	36
Fundulus olivaceus	116	15	15	15
Ictalurus punctatus	502	54	54	50
Lepistoseus oculatus	25	25	25	24
Lepistoseus osseus	142	120	120	118
Lepomis auritus	139	28	27	27
Lepomis gibbosus	21	10	10	10
Lepomis macrochirus	102	102	102	97
Lota lota	28	8	8	8
Luxilus cornutus	12	4	4	4
Margariscus margarita	39	39	38	33
Micropterus dolomieu	300	63	63	59
Micropterus salmoides	55	55	55	54
Morone americana	987	122	122	110
Moxostoma duquesnei	43	43	43	42
Moxostoma erythrurum	29	9	9	8
Notemigonus crysoleucas	15	5	5	5
Notropis bifrenatus	104	24	24	24

Noturus exilis	23	5	5	5
Oncorhynchus mykiss	274	46	46	44
Osmerus mordax	539	105	103	93
Perca flavescens	61	61	61	59
Percina nigrofasciata	18	6	6	5
Percina sciera	34	6	6	6
Phoxinus eos	22	8	8	8
Phoxinus neogaeus	45	9	9	8
Pimephales notatus	180	29	29	29
Pimephales promelas	20	6	5	5
Pomoxis nigromaculatus	24	24	24	24
Rhinichthys atratulus	134	36	35	32
Salmo salar	78	13	13	13
Salmo trutta	312	30	30	29
Salvelinus fontinalis	480	88	88	79
Salvelinus namaycush	138	30	30	27
Semotilus atromaculatus	22	22	22	22
Semotilus corporalis	43	19	18	18
Stizostedion vitreum	17	4	4	4

The influence of distance measure used

While not mentioned in the comment, some researchers are concerned that the use of Euclidean distance in niche space could have influenced our overall findings. For thoroughness, we re-analyzed our data using Mahalanobis distance instead of Euclidean distance, finding no change in our results (Figure S1). This is either because the distance measure doesn't strongly influence the overall relationship, or because our niche axes were based on a PCA decomposition of 56 climatic covariates, and the first two axes are orthogonal. As a consequence, covariance structure is nearly zero. Apart from not influencing our results, we found both distance measures were highly correlated, suggesting the choice of distance measure is unlikely to influence our overall conclusions (Figure S2).

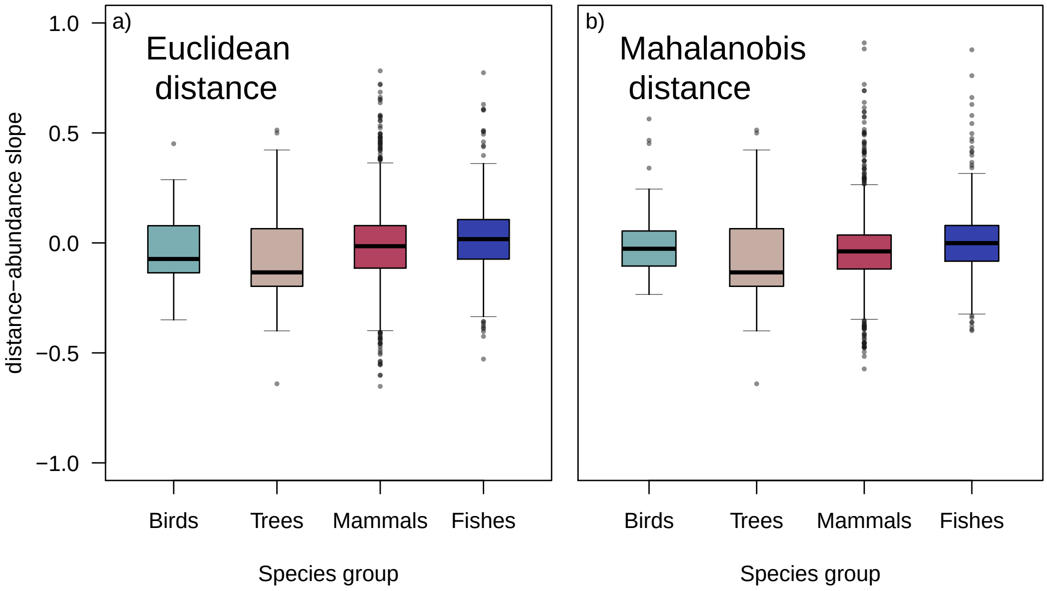

Figure S1: The relationship between distance from the niche centroid and species population density for four groups of species, using either Euclidean distance (left panel) or Mahalanobis ditance (right panel). The use of distance metric did not influence our failure to detect significant *distance-abundance* relationships.

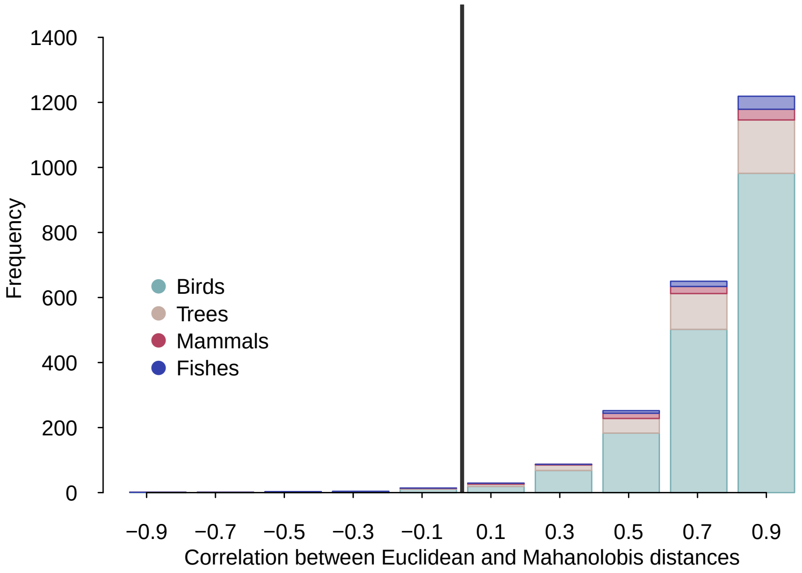

Figure S2: For each species, we calculated the correlation between distances from niche centroids calculated as Euclidean and Mahalanobis distance and calculated the correlation between distance meausures for each species. These relationships tended to be quite positive and near 1, suggesting that the two metrics were strongly related.