Sampbias, a method to evaluate geographic sampling bias in species distribution data

Alexander Zizka[1,2,3], *Alexandre Antonelli*[3,4,5], *Daniele Silvestro*[3,4]

1. German Center for Integrative Biodiversity Research, University of Leipzig, Leipzig, Germany
2. Naturalis Biodiversity Center, Leiden University, Leiden, The Netherlands
3. Gothenburg Global Biodiversity Centre, University of Gothenburg, Gothenburg, Sweden
4. Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden
5. Royal Botanical Gardens Kew, Richmond, Surrey, United Kingdom

Abstract

Georeferenced species occurrences from public databases have become essential to biodiversity research and conservation, but have limitations. Geographically biased sampling is a widely recognized issue that might severely affect analyses. Especially "roadside bias", i.e. differences in sampling intensity among localities caused by differences in accessibility for humans is ubiquitous and might differ in strength among taxonomic groups and datasets. Yet, no general methodology exists to quantify the effect of roadside or other sources of bias on a dataset level. Here we present *sampbias*, a novel algorithm and software to estimate the biasing effect of accessibility (by roads, rivers, airports, cities, or any user-defined structures) in species occurrence datasets. *Sampbias* is based on a null model of even sampling and assesses whether instead sampling probability decays exponentially with distance. The results are comparable among biasing factors and datasets. *Sampbias* is implemented as a user-friendly R package, and Shiny app. We exemplify the use of *sampbias* on a dataset of mammal occurrences from the Indonesian island of Borneo, downloaded from www.gbif.org. *Sampbias* offers an efficient and largely automated means for biodiversity scientists and non-specialists alike to explore bias in species occurrence data. The output of *sampbias* may be used to identify priorities for further collection or digitalization efforts, provide bias surfaces for species distribution modelling, or assess the reliability of scientific results based on publicly available species distribution data.

Keywords

Collection effort, Global biodiversity Information Facility (GBIF), Presence only data, Roadside bias, Sampling intensity

Introduction

Publicly available datasets of geo-referenced species occurrences, such as provided by the Global Biodiversity Information Facility (www.gbif.org) have become a fundamental resource in biological sciences, especially in biogeography, conservation, macroecology, and systematics. However, because these datasets are "presence-only" data, they rarely include information on collection effort. Instead they are typically not collected systematically and often compiled from a variety of sources (e.g. scientific expeditions, census counts, genetic barcoding studies, and citizen-science observations), thus becoming subject to collection biases (Meyer, Weigelt, and Kreft 2016). That is the number of data points available is biased by factors other than species' presence or abundance, including the under-sampling of specific taxa ("taxonomic bias", e.g., birds *vs.* nematodes), specific geographic regions ("geographic bias", i.e. easily accessible *vs.* remote areas), and specific temporal periods ("temporal bias", i.e. wet season *vs.* dry season, Isaac and Pocock 2015; Boakes et al. 2010). While these biases are broadly recognized, and approaches exist to account for them in some analyses (for instance for species-richness estimation (Engemann et al. 2015) species distribution modelling (Stolar and Nielsen 2015; Beck et al. 2014; Fithian et al. 2014; Warren et al. 2014; Boria et al. 2014; Varela et al. 2014; Fourcade et al. 2014), occupancy modelling (Kery and Royle 2016), or abundance estimations (Shimadzu and Darnell 2015)), few attempts have been made to discern among different sources of bias or to compare the strength of bias among datasets (but see Ruete 2015).

Geographic sampling bias, the fact that sampling effort is spatially biased, rather than equally distributed over a given study area is prevalent in all non-systematically collected datasets of species distributions. Many factors can affect sampling effort, such as socio-economic factors (i.e. national research spending, history of scientific research; Meyer et al. 2015, @Daru2018) and political factors (armed conflict, democratic rights; Rydén et al. 2019) or physical accessibility (i.e. distance to a road or river, terrain conditions, slope; Yang, Ma, and Kreft 2014; Botts, Erasmus, and Alexander 2011). Especially physical accessibility is omnipresent as a biasing factor (e.g. Lin et al. 2015; Engemann et al. 2015), across spatial scales, and the term "roadside bias" has been coined for it. In practice, this means that most species observations (occurrence points) are made in or near cities, along roads and rivers, and near other human settlements (such as airports). Less observations come from the middle of a tropical rainforest or from a mountain top. Interestingly, since the observation of different taxonomic groups has different challenges, geographic sampling bias and the effect of accessibility may differ among taxonomic groups (Vale and Jenkins 2012).

The implications of not considering spatial collection bias in biodiversity research are likely to be substantial (Meyer, Weigelt, and Kreft 2016; Rocchini et al. 2011; Shimadzu and Darnell 2015; Yang, Ma, and Kreft 2013; Kramer-Schadt et al. 2013; Barbosa, Pautasso, and Figueiredo 2013). While it is unrealistic to expect that spatial biases in biodiversity data will ever disappear, it is crucial that researchers realise the intrinsic biases associated with the biodiversity data they are dealing with. This is the first step towards estimating to which extent these biases may affect their analyses, results, and conclusions drawn from such data. Therefore, it is advisable for any study dealing with species occurrence data to assess the strength of accessibility bias in the underlaying data.

Here, we present *sampbias*, a novel method to quantify accessibility bias in individual datasets of species occurrences, in a way that is comparable across datasets. *Sampbias* is implemented as an R-package. Specifically, *sampbias* uses a null-model of random sampling to address two questions:

1) How strong is the accessibility bias in a given dataset?

2) How important are different means of human accessibility, such as to airport, cities, rivers or roads, in causing this bias?

3) How is sampling bias distributed in space, i.e. which areas are a priority for targeted sampling?

Description

General concept

Under the assumption that organisms exist across the entire area of interest, we can expect the number of sampled occurrences to be distributed uniformly in space (even though, of course, the density of individuals and the species composition may be heterogeneous). [[*I think we should acknowledge here that this assumption is valid when looking at a geographically restricted area, eg within a tropical forest. Of course different biomes eg forest, alpine, oceanic will result in different carrying capacity*]] With *sampbias* we assess if a set of occurrences significantly departs from a null uniform distribution and whether these discrepancies between expected and observed distributions can be explained by distance from factors that potentially bias their sampling probability (e.g. distance from cities or roads).

Sampbias works on a user-defined scale, and any dataset of multi-species occurrence records can be tested against any geographic gazetteer (reliability increases with increasing dataset size). Default large-scale gazetteers for airports, cities, rivers and roads are provided with *sampbias.* Species occurrence data as downloaded from the data portal of GBIF can be directly used as input data for sampbias. The output of the package includes measures of bias effect, which are comparable between different gazetteers (e.g. comparing biasing effect of roads and rivers), different taxa (e.g. birds *vs.* flowering plants) and different data sets (e.g. specimens *vs.* human observations).

CoordinateCleaner is implemented in R (R Core Team 2019) based on standard tools for spatial statistics: ggplot2 (Wickham 2009), geosphere (Hijmans 2019), maptools (Bivand and Rundel 2019), raster (Hijmans 2019), sp (Pebesma and Bivand 2005; Bivand, Pebesma, and Gomez-Rubio 2013), and viridis (Garnier 2018).

Distance calculation

Sampbias uses gazetteers of the geographic location of bias sources (e.g. roads) to generate a grid across the study area (the geographic extent of the dataset) for each gazetteer and then calculates the distance ("as the crow flies") [[*meaning?*]] of the midpoint of each grid cell to the closest cell containing an instance of the gazetteer. We then use these distance grids to sample the distribution of distances in the observed dataset and the null distribution in a reference dataset of equal size with randomly distributed records (the null model). The resolution of the grid defined the precision of the distance estimates, for instance a 1x1 degree raster will yield approximately a 100km precision at the equator.

Quantifying accessibility bias using maximum likelihood

Given the placement of a particular bias sources in the area of interest and assuming a uniform distribution of samples, the probability of a sampled occurrence located at a distance d from the closest bias source is a function of the amount of available area at that distance. That is, the larger the area located at distance d from a bias source the more samples we expect. For simplicity we discretize the area of interest in a number of grid cells and indicate with $f(x)$ the function describing the number of available grid cells at any distance x, for $0 < x < \max(x)$, where $\max(x)$ is the maximum observed distance between a cell and the clostest bias source. The function $f(x)$ is therefore calculated based on the distances of each grid cell from its closest bias source.

The distribution of samples, in the absence of bias, should therefore represent a random sample from $f(x)$ and reflect its shape, i.e. d $f(x)$ (Fig. 1). However, in the presence of a bias, we expect the probability of finding occurrences to decrease with increasing distance from a bias source. This in turn will alter the resulting distribution of samples that no longer match the expected distribution. Here, we model the effect of sampling bias by assuming that the probability of sampling an occurrence decreases exponentially with increasing distance (Fig. 1), following the function $b(x, l) = l \exp(-lx)$, where l is the rate parameter. Under these assumption the expected distribution of samples is given by $g(x, l) \propto f(x)b(x, l)$.

[[*I wonder if we shouldn't use " $b(x, l) = \exp(-lx)$ " so the Y-value is always 1 at a distance of 0. That would require normalizing the likelihood based on the integral of the curve though, as it will be different from 1*] get the dsitribution from all grid cells and normalize by the number of all available grid cells, not downsampled to the same number of points ignore the starting at 1/intercept

a possion likelihood where the rate i two paramters, 1. speed of decae 2. how high is the bias at distance zero

$b(x, l) = \exp(-lx)$ $q * exp(-lambdax) / \int_0^\infty (q * exp(-lambdax))$

Average the bias by Akaike wheights, or likelihood

The rate parameter l describes the strength of the bias effect. When l is large the expected probability of sampling occurrences decreases very quickly as you move away from a bias source (Fig. 1). In contrast, when the l parameter is small, the resulting exponential distribution effectively becomes more and more similar to a uniform distribution (Fig. 1), indicating that increasing distance from e.g. a city does not affect the sampling probability. We treat l as an unknown variable and estimate it using maximum likelihood from the data based on the probability density function described by $g(x, l)$. This essentially means finding the value of l that best explains any discrepancies between the expected distribution ($f(x)$) and the observed occurrences. [[*I think this is in fact a posterior probability where f(x) is the empirical prior and b(x,l) the likelihood and we do a maximum a posteriori optimization*]]

Once we have an estimated value of l, we can infer the expected accessibility bias as a function of distance using $b(x, l)$. Since the function $b(x, l)$ describes the exponentially decreasing sampling probability in relation to distance, we can define a standardized bias function as: $B(x, l) = 1 - b(x, l)/b(0, l)$, [[*Actually I think this is effectively equivalent to what I wrote* *above* $b(x, l) = \exp(-lx)$]] where the level of bias is set to 0 at distance 0 from the bias source. The standardized bias function tends asymptotically to 1 as the distance tends to infinity. However, since in any area $max(x) << \infty$, for small values of l (i.e. little or no bias), $B(x, l)$ will look essentially like a uniform distribution with values very close to 0. Large values of l (i.e. strong bias) will instead result in a curve that quickly approaches 1 with increasing distance. The values provided by the standardized bias function can be interpreted as the proportion of occurrences that are missing from the sample, compared to the observed samples at distance 0. Thus, if for a given estimated l we have $B(50, l) = 0.20$, we can expect that at 50 Km from e.g. a road the number of occurrences per grid cell will be about 80% of the occurrences sampled at distance 0 from the road, with 20% missing due to sampling bias.

When running *sampbias*, we typically test different sources of biases, such as roads, cities, airports, etc. For each factor an independent expected distribution ($f(x)$) is computed and a parameter l is estimated from $g(x, l)$. These estimates can be use to produce maps showing the intensity of potential biases across the area based on the standardized bias function. The bias values obtained from different sources can then be averaged in each grid cell to produce a map showing the combined effects of all sources. [[*yeah this part is still weird. Basically, if rivers are not explaining anything they should count nothing toward the combined estimate, whereas they do now. I think we could try to see if the max likelihoods are comparable and weight the average by that.*]]

Running sampbias

A default *sampbias* analysis can be run with few lines of code in R. The main function `SamplingBias` creates an object of the class `"sampbias"`, for which the package provides a plotting and summary method. Based on a `data.frame` including species identity and geographic coordinates, sampbias provides a bias effect estimate for each gazetteer and an average bias. Additionally some options exist to provide custom gazetteers, custom distances for the bias estimation, a custom grain size of the analysis, as well as some operators for the calculation of the bias distances. A tutorial on how to use sampbias is available with the package and in the electronic supplement of this publication (Appendix S1).

```
library(sampbias)

# a data table with species identify, longitude, and latitude
example.in <- read.csv(system.file("extdata", "mammals_borneo.csv", package = "sampbias'
    sep = "\t")

# running sampbias
example.out <- SamplingBias(x = example.in, res = 0.1)

# summarizing the results
summary(example.out)
plot(example.out)
```

For data exploration we implemented the basic functionalities in a shiny app as graphical user interface (Fig. 2). Analyses can be run based on a tab separated .txt file with occurrence information including the column headers "species", "decimallongitude" and "decimallatitude", as for instance files downloaded from www.gbif.org, using custom gazetteers. A tutorial on how to use the *sampbias* GUI is available online (https://ropensci.github.io/sampbias/) and in the electronic supplement of this publication (Appendix S2).

Empirical example

To exemplify the use and output of sampbias, we downloaded the occurrence records of all mammals available from the Indonesian island of Borneo (**???**), and quantify the biasing effect of airports, cities and roads in the dataset. **Something on the results, also add a table** (Fig. 3)

Assumptions and future prospective

Two assumptions of *sampbias* are a equal sampling of occurrence records across the study area as null model and an exponential increase of the biasing effect with distance from the gazetteers. We considered both acceptable approximations for the purpose of the package, but future expansions of *sampbias* could relax these assumptions, for instance by allowing other distance decay functions, such as gamma or Weibull distributions, and by changing the sampling scheme of the background points. The first steps towards these goals are already implemented in the current version of *sampbias* with the option to limit background points to a convex hull around the dataset or limiting background points to terrestrial surface. [[*I don't understand this part*]]

A practical limitation of *sampbias* is the trade-off between the resolution of the grid for the distance calculation and the geographic extent of the dataset. For instance, a 100m resolution for a global dataset would lead to the generation of grid for which distance calculation will become computationally prohibitive in most practical cases, Hence, *sampbias* is best suited for local or regional datasets at high resolution (c. 100 – 10,000m) or continental datasets at low resolution (c. 10 – 100km).

Todo

re-run empirical analysis

test units

Data accessibility

The software presented here is available under a GPL-3 license. The *sampbias* R package and the source code for the shiny app are available via https://github.com/azizka/sampbias, the shiny app can be accessed at https://azizka.shinyapps.io/sampbias/. The R package includes an example dataset as well as vignettes detailing the use of the R package, the use of the shiny app and possibly warnings produced by the package (Appendix S3).

Acknowledgements

We thank the organizers of the 2016 Ebben Nielsen challenge for inspiring and recognizing this research. We thank all data collectors and contributors to GBIF for their effort.

Author contributions

All authors conceived of this study, AZ and DS developed the statistical algorithm, AZ and DS wrote the R-package and AZ the Shiny app, AZ and DS wrote the manuscript with contributions from AA.

229 # Figures

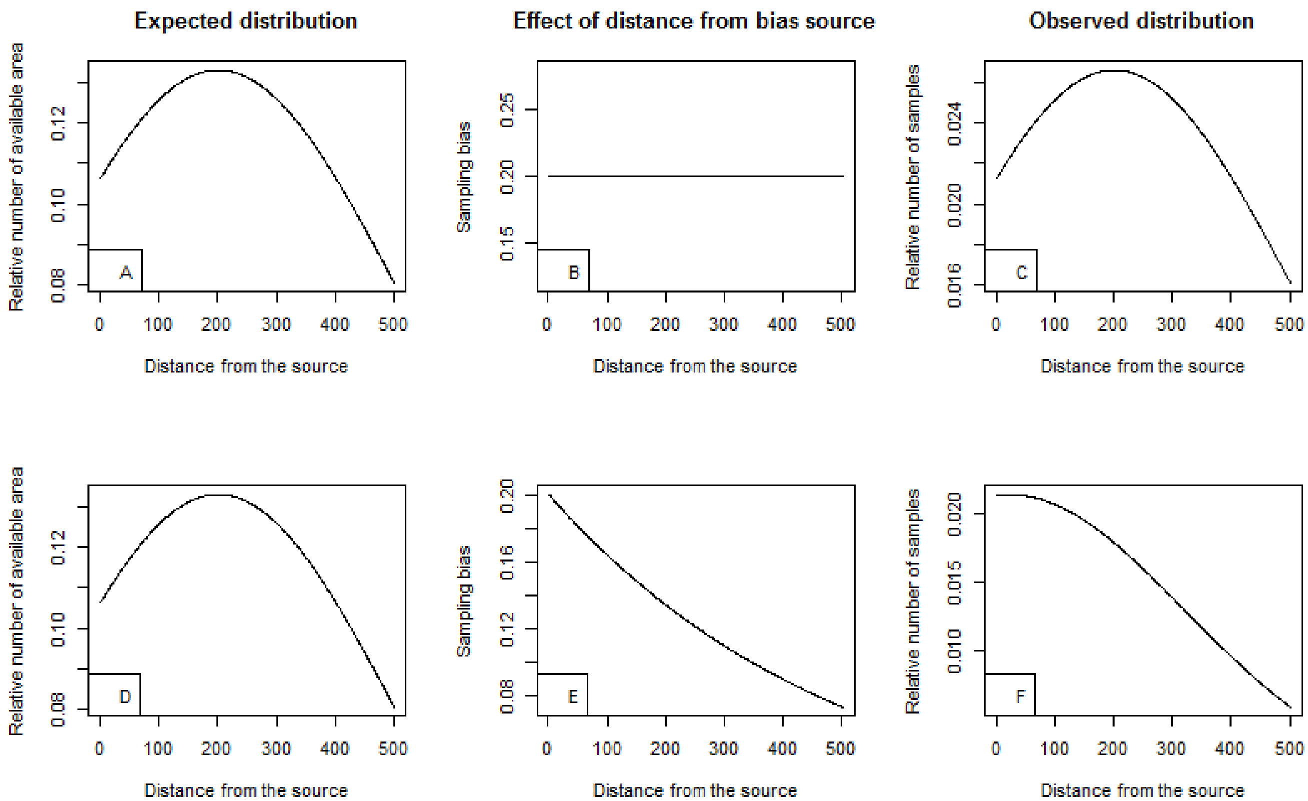

Figure 1: Schematic representation of the *sampbias* algorithm. The upper panel (A-C) shows a case without accessibility bias: the observed distribution of distances from the geographic feature of interest matches the expected distribution and hence the bias function is flat. The lower panel (D-F) shows a case with bias: the expected and observed distribution differ and the bias is best represented by an exponential function.

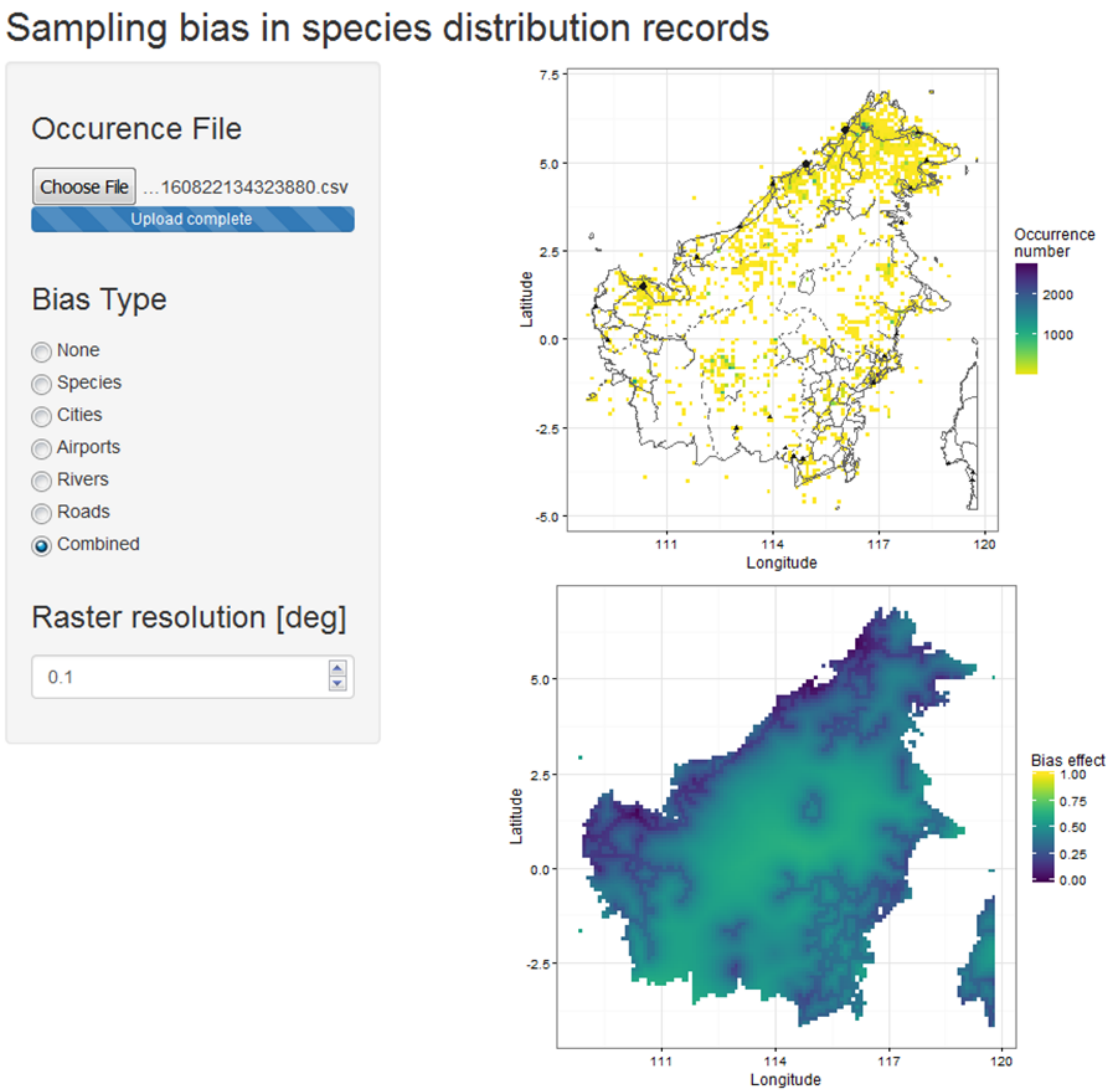

Figure 2: The interface of the sampbias shiny app. With a navigation panel on the left and visualization of the results on the right.

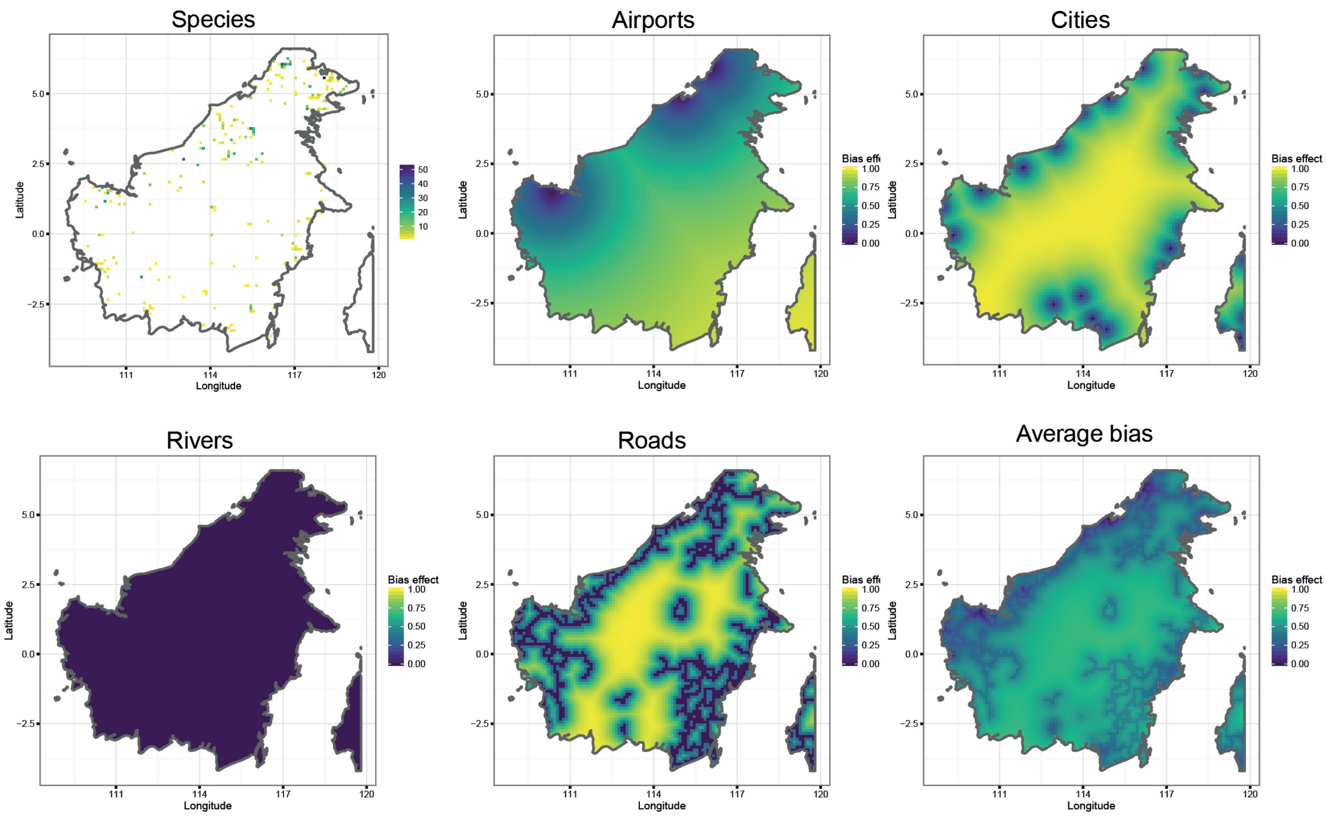

Figure 3: The spatial projection of the accessibility bias in an empirical example dataset of mammal occurrences on the Indonesian island of Borneo from www.gbif.org. *Sampbias* finds a strong biasing effect by roads, cities and lower by airports, but, surprisingly, no effect of rivers.

Supplementary material

Appendix S1 - Tutorial running sampbias in R

Appendix S2 - Tutorial running the sampbias shiny app

Appendix S3 - Possible warnings and their solutions

References

Barbosa, A. Márcia, Marco Pautasso, and Diogo Figueiredo. 2013. "Species-people correlations and the need to account for survey effort in biodiversity analyses." *Diversity and Distributions* 19 (9): 1188–97. https://doi.org/10.1111/ddi.12106.

Beck, Jan, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. 2014. "Spatial bias in the GBIF database and its effect on modeling species' geographic distributions." *Ecological Informatics* 19: 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Bivand, Roger, and Colin Rundel. 2019. "rgeos: Interface to Geometry Engine - Open Source ('GEOS')." https://cran.r-project.org/package=rgeos.

Bivand, Roger S., Edzer J. Pebesma, and Virgilio Gomez-Rubio. 2013. *Applied spatial data analysis with R, Second edition*. New York, USA: Springer.

Boakes, Elizabeth H., Philip J K Mcgowan, Richard A Fuller, Ding Chang-Qing, Natalie E Clark, Kim O Connor, and Georgina M. Mace. 2010. "Distorted views of biodiversity: Spatial and temporal bias in species occurrence data." *PLoS Biology* 8 (6): e1000385. https://doi.org/10.1371/journal.pbio.1000385.

Boria, Robert A., Link E. Olson, Steven M. Goodman, and Robert P. Anderson. 2014. "Spatial filtering to reduce sampling bias can improve the performance of ecological niche models." *Ecological Modelling* 275 (March): 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012.

Botts, Emily A., Barend F N Erasmus, and Graham J. Alexander. 2011. "Geographic sampling bias in the South African Frog Atlas Project: Implications for conservation planning." *Biodiversity and Conservation* 20 (1): 119–39. https://doi.org/10.1007/s10531-010-9950-6.

Daru, Barnabas H, Daniel S Park, Richard B Primack, Charles G Willis, David S Barrington, Timothy J S Whitfeld, Tristram G Seidler, et al. 2018. "Widespread sampling biases in herbaria revealed from large-scale digitization." *New Phytologist* 217 (2): 939–55. https://doi.org/10.1111/nph.14855.

Engemann, Kristine, Brian J Enquist, Brody Sandel, Brad Boyle, Peter M Jørgensen, Naia Morueta-Holme, Robert K Peet, Cyrille Violle, and Jens-Christian Svenning. 2015. "Limited sampling hampers 'big data' estimation of species richness in a tropical biodiversity hotspot." *Ecology and Evolution* 5 (3): 807–20. https://doi.org/10.1002/ece3.1405.

Fithian, William, Jane Elith, Trevor Hastie, and David a. Keith. 2014. "Bias correction in species distribution models: pooling survey and collection data for multiple species." *Methods in Ecology and Evolution* 6 (4): 424–38. https://doi.org/10.1111/2041-210X.12242.

Fourcade, Yoan, Jan O. Engler, Dennis Rödder, and Jean Secondi. 2014. "Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias." *PLoS ONE* 9 (5): e97122. https://doi.org/10.1371/journal.pone.0097122.

Garnier, Simon. 2018. *viridis: Default Color Maps from 'matplotlib'*. https://cran.r-project.org/package=viridis.

Hijmans, Robert J. 2019. "geosphere: Spherical Trigonometry." https://cran.r-project.org/

package=geosphere.

Isaac, Nick J B, and Michael J O Pocock. 2015. “Bias and information in biological records.” *Biological Journal of the Linnean Society* 115 (3): 522–31. https://doi.org/10.1111/bij.12532.

Kery, Marc, and J Andrew Royle. 2016. *Applied Hierarchical Modeling in Ecology - Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models*. Amsterdam: Academic Press, Elsevier.

Kramer-Schadt, Stephanie, Jürgen Niedballa, John D. Pilgrim, Boris Schröder, Jana Lindenborn, Vanessa Reinfelder, Milena Stillfried, et al. 2013. “The importance of correcting for sampling bias in MaxEnt species distribution models.” *Diversity and Distributions* 19 (11): 1366–79. https://doi.org/10.1111/ddi.12096.

Lin, Yu-pin, Dongpo Deng, Wei-chih Lin, Rob Lemmens, Neville D Crossman, Klaus Henle, and Dirk S Schmeller. 2015. “Uncertainty analysis of crowd-sourced and professionally collected field data used in species distribution models of Taiwanese moths.” *Biological Conservation* 181: 102–10. https://doi.org/10.1016/j.biocon.2014.11.012.

Meyer, Carsten, Holger Kreft, Robert P Guralnick, and Walter Jetz. 2015. “Global priorities for an effective information basis of biodiversity distributions.” *Nature Communications* 6 (e1057): 8221. https://doi.org/10.1038/ncomms9221.

Meyer, Carsten, Patrick Weigelt, and Holger Kreft. 2016. “Multidimensional biases, gaps and uncertainties in global plant occurrence information.” *Ecology Letters* 19: 992–1006. https://doi.org/10.1111/ele.12624.

Pebesma, Edzer J., and Roger S. Bivand. 2005. "Classes and methods for spatial data in R." *R News* 5 (2). https://cran.r-project.org/doc/Rnews/.

R Core Team. 2019. "R: A Language and Environment for Statistical Computing." Austria, Vienna: R Foundation for Statistical Computing. https://www.r-project.org/.

Rocchini, Duccio, Joaquín Hortal, Szabolcs Lengyel, Jorge M Lobo, Alberto Jiménez-Valverde, Carlo Ricotta, Giovanni Bacaro, and Alessandro Chiarucci. 2011. "Accounting for uncertainty when mapping species distributions: The need for maps of ignorance." *Progress in Physical Geography: Earth and Environment* 35 (2): 211–26. https://doi.org/10.1177/0309133311399491.

Ruete, Alejandro. 2015. "Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps." *Biodiversity Data Journal* 3 (1): e5361. https://doi.org/10.3897/BDJ.3.e5361.

Rydén, Oskar, Alexander Zizka, Sverker C Jagers, Staffan I Lindberg, and Alexandre Antonelli. 2019. "Linking democracy and biodiversity conservation: Empirical evidence and research gaps." *Ambio*, June. https://doi.org/10.1007/s13280-019-01210-0.

Shimadzu, Hideyasu, and Ross Darnell. 2015. "Attenuation of species abundance distributions by sampling." *Royal Society Open Science* 2 (4): 140219. https://doi.org/10.1098/rsos.140219.

Stolar, Jessica, and Scott E. Nielsen. 2015. "Accounting for spatially biased sampling effort in presence-only species distribution modelling." *Diversity and Distributions* 21 (5): 595–608. https://doi.org/10.1111/ddi.12279.

Vale, Mariana M., and Clinton N. Jenkins. 2012. "Across-taxa incongruence in patterns of collecting bias." *Journal of Biogeography* 39 (9): 1744–4. https://doi.org/10.1111/j.1365-2699.2012.02759.x.

Varela, Sara, Robert P. Anderson, Raúl García-Valdés, and Federico Fernández-González. 2014. "Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models." *Ecography*, no. September 2013: 1084–91. https://doi.org/10.1111/j.1600-0587.2013.00441.x.

Warren, Dan L., Amber N. Wright, Stephanie N. Seifert, and H. Bradley Shaffer. 2014. "Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern." *Diversity and Distributions* 20 (3): 334–43. https://doi.org/10.1111/ddi.12160.

Wickham, Hadley. 2009. *ggplot2 - Elegant Graphics for Data Analysis*. New York: Springer. https://doi.org/10.1007/978-0-387-98141-3.

Yang, Wenjing, Keping Ma, and Holger Kreft. 2013. "Geographical sampling bias in a large distributional database and its effects on species richness-environment models." *Journal of Biogeography* 40 (8): 1415–26. https://doi.org/10.1111/jbi.12108.

———. 2014. "Environmental and socio-economic factors shaping the geography of floristic collections in China." *Global Ecology and Biogeography* 23 (11): 1284–92. https://doi.org/10.1111/geb.12225.