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

Alexander Zizka1,2,3, Alexandre Antonelli3,4,5, Daniele Silvestro3,4

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. For each cell , we compute a vector of minimum distances to each source of bias .

## Quantifying accessibility bias using a Bayesian framework

We describe the observed number of sampled occurrences within each cell as the result of Poisson sampling process with rate . We model the rate as a function of a constant , which represent the expected number of occurrences per cell in the absence of biases, i.e. when . Additionally, is modeled to decrease exponentially as a function of distance from sources of bias, such that increasing distances will result in a lower sampling rate. For a single source of bias the rates of cell with distance from a bias is:

where defines the steepness of the Poisson rate decline, such that results in a null model of uniform sampling rate across cells.

In the presence of multiple bias predictors, the sampling rate decrease is a function of the cumulative effects of each bias and its distance from the cell:

where a vector describes the amount of bias attributed to each specific predictor.

To quantify the amount of bias associated with each predictor, we jointly estimate the parameters and in a Bayesian framework. We use Markov Chain Monte Carlo (MCMC) to sample these parameters from their posterior distribution:

where the likelihood of sampled occurrences within each cell is the probability mass function of a Poisson distribution with rate per cell defined as in Eqn. (1). The likelihood is then multiplied across the cells considered. We used exponential priors on the parameters and , and , respectively.

We summarize the parameters by computing the mean of the posterior samples and their 95% credible intervals. We interpret the magnitude of the elements in as a function of the importance of the individual biases. We note however although that this test is not explicitly intended to assess the significance of each bias predictor (for which a Bayesian variable selection could be used), particularly since several sources of bias might b correlated (e.g. cities, and airports). Instead, these aanlyses can be use to quantify the expected amount of bias in the data that can be predicted by single or multiple predictors in order to identify under-sampled and unexplored areas.

# Running sampbias

A default *sampbias* analysis can be run with few lines of code in R. The main function calculate\_bias 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 <- calculate\_bias(x = example.in, res = 0.1)  
  
# summarizing the results  
summary(example.out)  
plot(example.out)  
  
#project in space  
proj <- project\_bias(example.out)  
map\_bias(proj)

# 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. )

# 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).

# Comparison to other software

# Todo

re-run empirical analysis

test units

# Data accessibility

The software presented here is available under a GPL-3 license. The *sampbias* R package is available <https://github.com/azizka/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 S2).

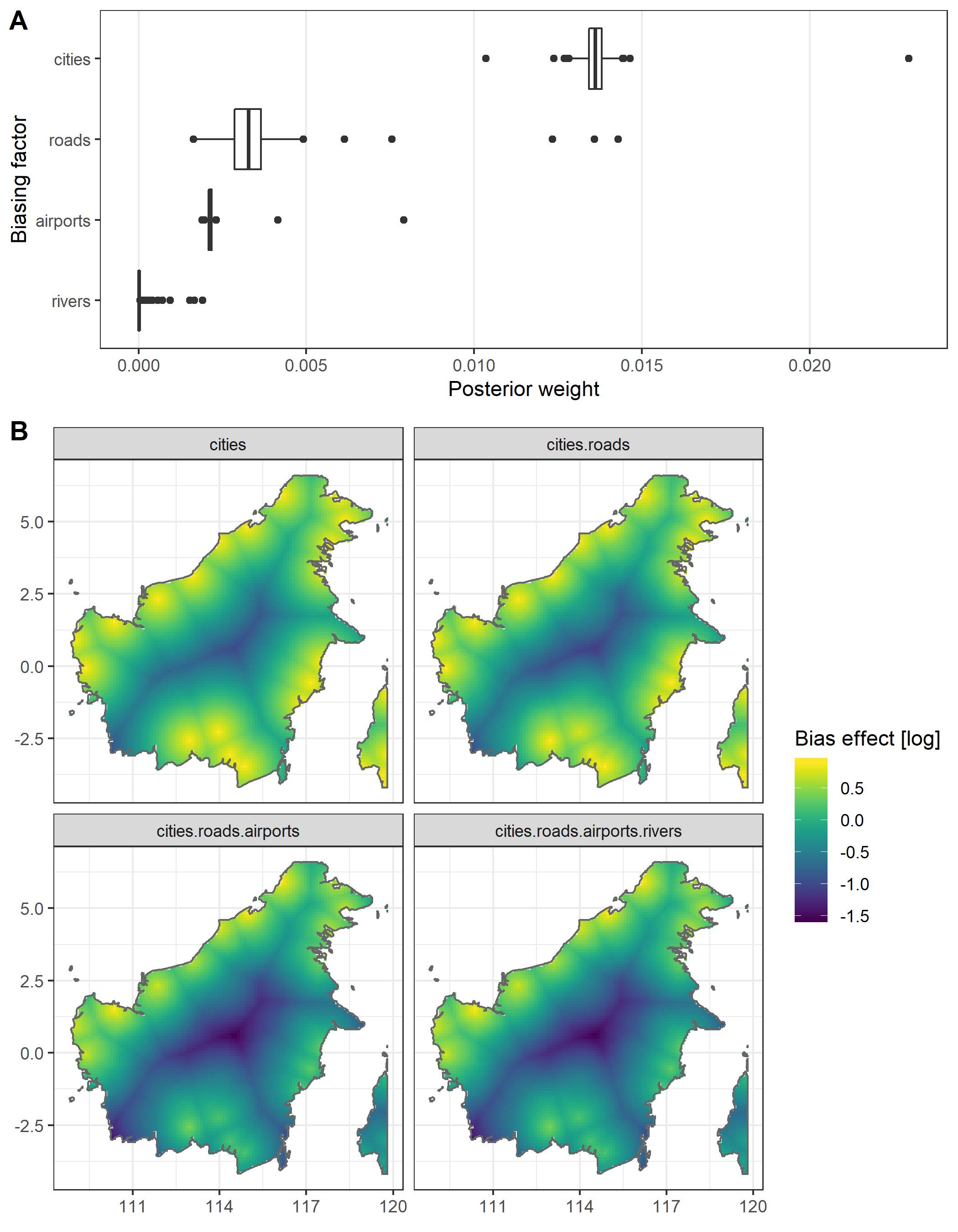
# Acknowledgements

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.

# Figures



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. A) bias weights, B) projection of the expected number of occurrences given the sampbias model. finds the strongest biasing effect for cities.

# Supplementary material

Appendix S1 - Tutorial running sampbias in R

Appendix S2 - 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>.