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

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 "accessibility 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. While several bias factors exist, here defined as anthropogenic or natural features that facilitate human accessibility (e.g. roads, rivers, airports, cities), quantifying their effect on occurrence data remains difficult. Here we present *sampbias*, an algorithm and software to quantify the effect of accessibility bias in species occurrence datasets. *Sampbias* uses a Bayesian approach to estimate how sampling rates vary as a function of proximity to one or multiple bias factors. The results are comparable among bias factors and datasets. *Sampbias* is implemented as a user-friendly R package. We demonstrate the use of *sampbias* on a dataset of mammal occurrences from the Indonesian island of Borneo, showing a high biasing effect of cities and a moderate effect of roads and airports.

Keywords

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

Background

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, and macroecology. However, these datasets are are typically not collected systematically and rarely include information on collection effort. Instead, they are often compiled from a variety of sources (e.g. scientific expeditions, census counts, genetic barcoding studies, and citizen-science observations), therefore often subject to sampling bias (Meyer et al. 2016).

The number of occurrence available in such datasets is likely 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). Geographic sampling bias—the fact that sampling effort is spatially biased, rather than equally distributed over the study area—is prevalent in all non-systematically collected datasets of species distributions. Many factors can can cause sampling bias, including socio-economic factors (i.e. national research spending, history of scientific research; www.bio-dem.surge.sh, Meyer et al. 2015, Daru et al. 2018), political factors (armed conflict, democratic rights; Rydén et al. 2019), and physical accessibility (i.e. distance to a road or river, terrain conditions, slope; Yang et al. 2014, Botts et al. 2011). Especially physical accessibility is omnipresent as a biasing factor (e.g. Lin et al. 2015, Kadmon et al. 2004, 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 are made in or near cities, along roads and rivers, and near other human settlements. Relatively fewer observations are expected to be available 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 geographic sampling bias in biodiversity research are likely substantial (Rocchini et al. 2011, Barbosa et al. 2013, Yang et al. 2013, Kramer-Schadt et al. 2013, Shimadzu and Darnell 2015, Meyer et al. 2016). While the presence of geographic sampling bias is broadly recognized (e.g. Kadmon et al. 2004), and approaches exist to account for it in some analyses—for instance for species-richness estimates (Engemann et al. 2015) species distribution models (Beck et al. 2014, Varela et al. 2014, Warren et al. 2014, Boria et al. 2014, Fourcade et al. 2014, Fithian et al. 2015, Stolar and Nielsen 2015, Monsarrat et al. 2019), occupancy models (Kery and Royle 2016), or abundance estimates (Shimadzu and Darnell 2015)—few attempts have been made to explicitly quantify the bias (Hijmans et al. 2000, Kadmon et al. 2004) or to discern among different sources of bias (Fithian et al. 2015, Fernández and Nakamura 2015, Ruete 2015), and to our knowledge, no tools exist for comparing the strength of accessibility bias among bias factors or datasets. We define as *bias factors* any anthropogenic or natural features that facilitate human accessibility and sampling, such as roads, rivers, airports, and cities.

While it is unrealistic to expect that accessibility in biodiversity data will ever disappear, it is crucial that researchers realise the intrinsic bias associated with the 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 underlying data. Finally, a quantification of accessibility bias can help researchers to target their sampling efforts.

Here, we present *sampbias*, a probabilistic method to quantify accessibility bias in datasets of species occurrences, in a way that is comparable across datasets. *Sampbias* is implemented as user-friendly R-package and uses a Bayesian approach to address three questions:

1) How strong is the accessibility bias in a given dataset?
2) How important are different bias factors in causing this bias?
3) How is accessibility bias distributed in space, i.e. which areas are a priority for targeted sampling?

Sampbias is implemented in R (R Core Team 2019), based on commonly used packages for data handling (`ggplot`, Wickham 2009, `forcats`, 2019, `tidyr`, Wickham and Henry 2019, `dplyr`, Wickham et al. 2019, `magrittr`, Bache and Wickham 2014, `viridis`, Garnier 2018), handling geographic information and geo-computation (`raster`, Hijmans 2019, `sp`, Pebesma and Bivand 2005, Bivand et al. 2013) and statistical modelling (`stats`, R Core Team 2019). *Sampbias* offers an easy and largely automated means for biodiversity scientists and non-specialists alike to explore bias in species occurrence data and 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.

Methods and Features

General concept

Under the assumption that organisms exist across the entire area of interest, we can expect the number of sampled occurrences in a restricted areas, such as a single biome, to be distributed uniformly in space (even though, of course, the density of individuals and the species diversity may be heterogeneous). With *sampbias* we assess to which extent variation in sampling rates can be explained by distance from bias factors.

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 global 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 the sampling rates across space, which are comparable between different gazetteers (e.g. comparing the biasing effect of roads and rivers), different taxa (e.g. birds *vs.* flowering plants) and different data sets (e.g. specimens *vs.* human observations).

Distance calculation

Sampbias uses gazetteers of the geographic location of bias factors to generate a regular grid across the study area (the geographic extent of the dataset). For each grid cell i, we then compute a vector $X_i(j)$ of minimum distances (straight aerial distance, "as the crow flies") to each bias factor $j \in B$. The resolution of the grid defines the precision of the distance estimates, for instance a 1x1 degree raster will yield approximately a 100 km precision at the equator. Due to the assumption of homogeneous sampling and a computational trade-off between the resolution of the distance raster and the extent of the study area (for instance, a 1000 m resolution for a global dataset would lead to the generation of grid for which distance calculation will become computationally prohibitive in most practical cases), *sampbias* is best suited for local or regional datasets at high resolution (c. 100 – 10,000 m).

Quantifying accessibility bias using a Bayesian framework

We describe the observed number of sampled occurrences S_i within each cell i as the result of a Poisson sampling process with rate λ_i. We model the rate λ_i as a function of a parameter q, which represents the expected number of occurrences per cell in the absence of biases, i.e. when $\sum_{j=1}^{B} X_i(j) = 0$. Additionally, we model λ_i to decrease exponentially as a function of distance from bias factors, such that increasing distances will result in a lower sampling rate. For a single bias factor the rates of cell i with distance X_i from a bias is:

$$\lambda_i = q \times \exp\left(-wX_i\right)$$

where $w \in \mathbb{R}^+$ defines the steepness of the Poisson rate decline, such that $w \approx 0$ results in a null model of uniform sampling rate q across cells. In the presence of multiple bias factors (e.g. roads and rivers), the sampling rate decrease is a function of the cumulative effects of each bias and its distance from the cell:

$$\lambda_i = q \times \exp\left(-\sum_{j=1}^{B} w_j X_i(j)\right) \qquad (1)$$

where a vector $\mathbf{w} = [w_1, ..., w_B]$ describes the amount of bias attributed to each specific factor.

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

$$P(q, \mathbf{w}|\mathbf{S}) \propto \prod_{i=1}^{N} Poi(S_i|\lambda_i) \times P(q)P(\mathbf{w}) \qquad (2)$$

where the likelihood of sampled occurrences S_i within each cell $Poi(S_i|\lambda_i)$ 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 N cells considered. We used exponential priors on the parameters q and $\mathbf{w}$, $P(q) \sim \Gamma(1, 0.01)$ and $P(\mathbf{w}) \sim \Gamma(1, 1)$, respectively.

We summarize the parameters by computing the mean of the posterior samples and their standard deviation. We interpret the magnitude of the elements in $\mathbf{w}$ as a function of the importance of the individual biases. We note, however, that this test is not explicitly intended to assess the significance of each bias factor (for which a Bayesian variable selection method could be used), particularly since several bias factors might be correlated (e.g. cities, and airports). Instead, these analyses 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.

We summarize the results by mapping the estimated sampling rates (λ_i) across space. These rates represent the expected number of sampled occurrences for each grid cell and provide a graphical representation of the spatial variation of sampling rates. Provided that the cells are of comparable size, the estimated rates will be comparable across data sets, regions, and taxonomic groups. Analyzing different regions, biomes, or taxa in separate analyses allows to account for differences in over sampling rates, which are not linked with bias factors. For instance, the unbiased sampling rate q is expected to differ between a highly sampled clade like birds and under-sampled groups of invertebrates, but their sampling biases ($\mathbf{w}$) might be similar across the two groups.

Example and Empirical analysis

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. 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).

To exemplify the use and output of *sampbias*, we downloaded the occurrence records of all mammals available from the Indonesian island of Borneo (n = 6,262, GBIF.org 2016), and ran *sampbias* using the default gazetteers as shown in the example code below, to test the biasing effect of the main airports, cities and roads in the dataset. The example dataset is provided with *sampbias*. We found a strong effect of cities on sampling intensity, a moderate effect of roads and airports and negligible effect of rivers (Fig. 1). All models predict a low number of collection records in the centre of Borneo (Fig. 2), which reflects the original data, and where accessibility means are low (Figure S1 in Appendix S1). The empirical example illustrates the use of *sampbias*, for detailed analyses or a smaller geographic scale, higher resolution gazetteers, including smaller roads and rivers and a higher spatial resolution would be desirable. Results might change with increasing resolution, since roads and rivers might have a stronger effect on higher resolutions (facilitating most the access to their immediate vicinity), whereas cities and airports might have a stronger effect on the larger scale (facilitating access to a larger area).

```
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 = 0.05,
                                 buffer = 0.5)


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


#project in space
proj <- project_bias(example.out)
map_bias(proj)
```

Data accessibility

Sampbias is available under a GPL-3 license from https://github.com/azizka/sampbias, and includes the example dataset as well as a tutorial (Appendix S2) and a summary of possibly warnings produced by the package (Appendix S3).

Author contributions

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

177 # Figures

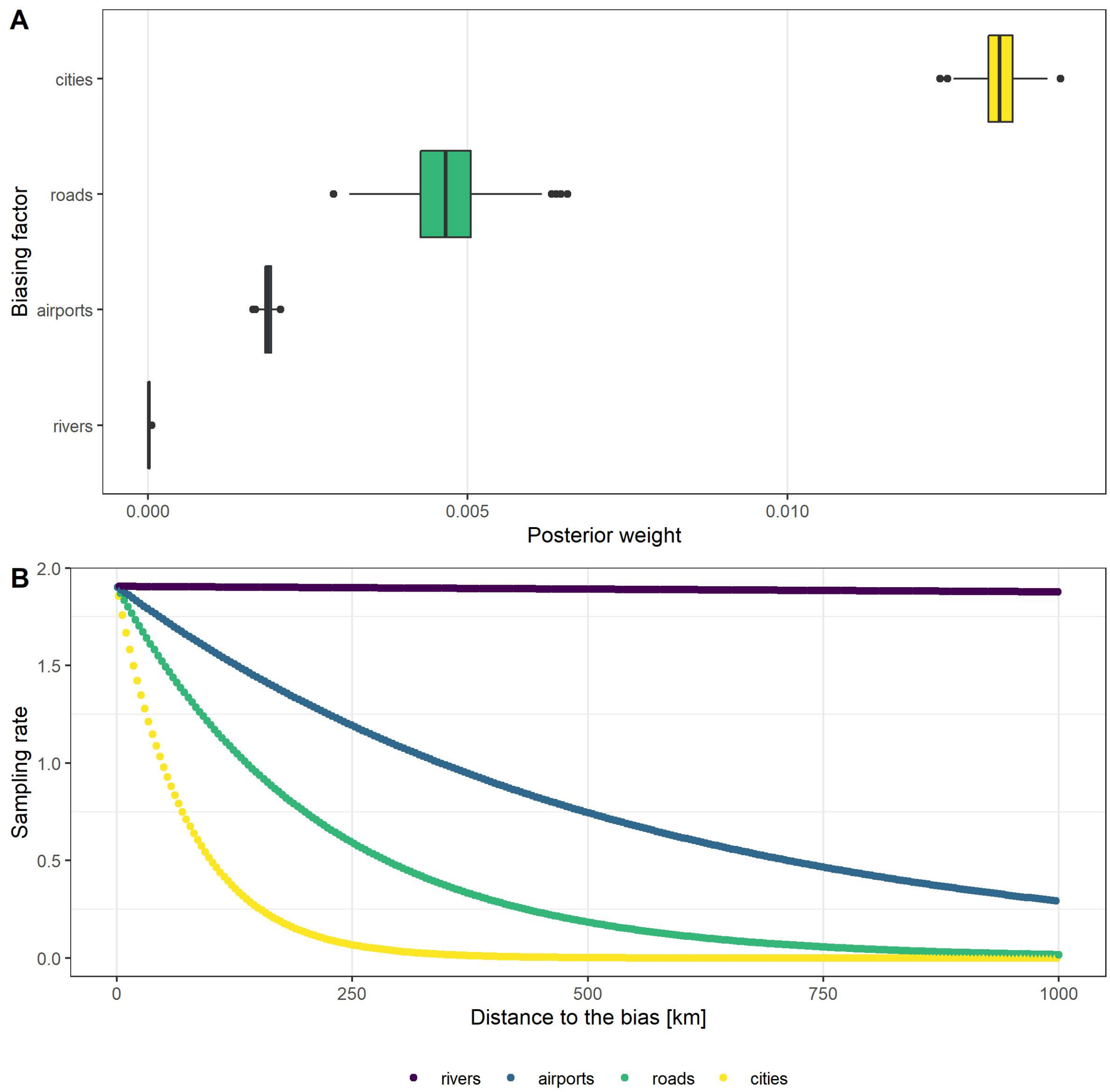

Figure 1: Results of a *sampbias* analysis. A) bias weights (w) defining the effects of each bias factor, B) the sampling rates as function of distance to the closest instance of each biasing factor (i.e. expected number of occurrences) given the inferred *sampbias* model. At the study scale of 0.05 degrees (5 km) *sampbias* finds the strongest biasing effect for the proximity of cities and roads.

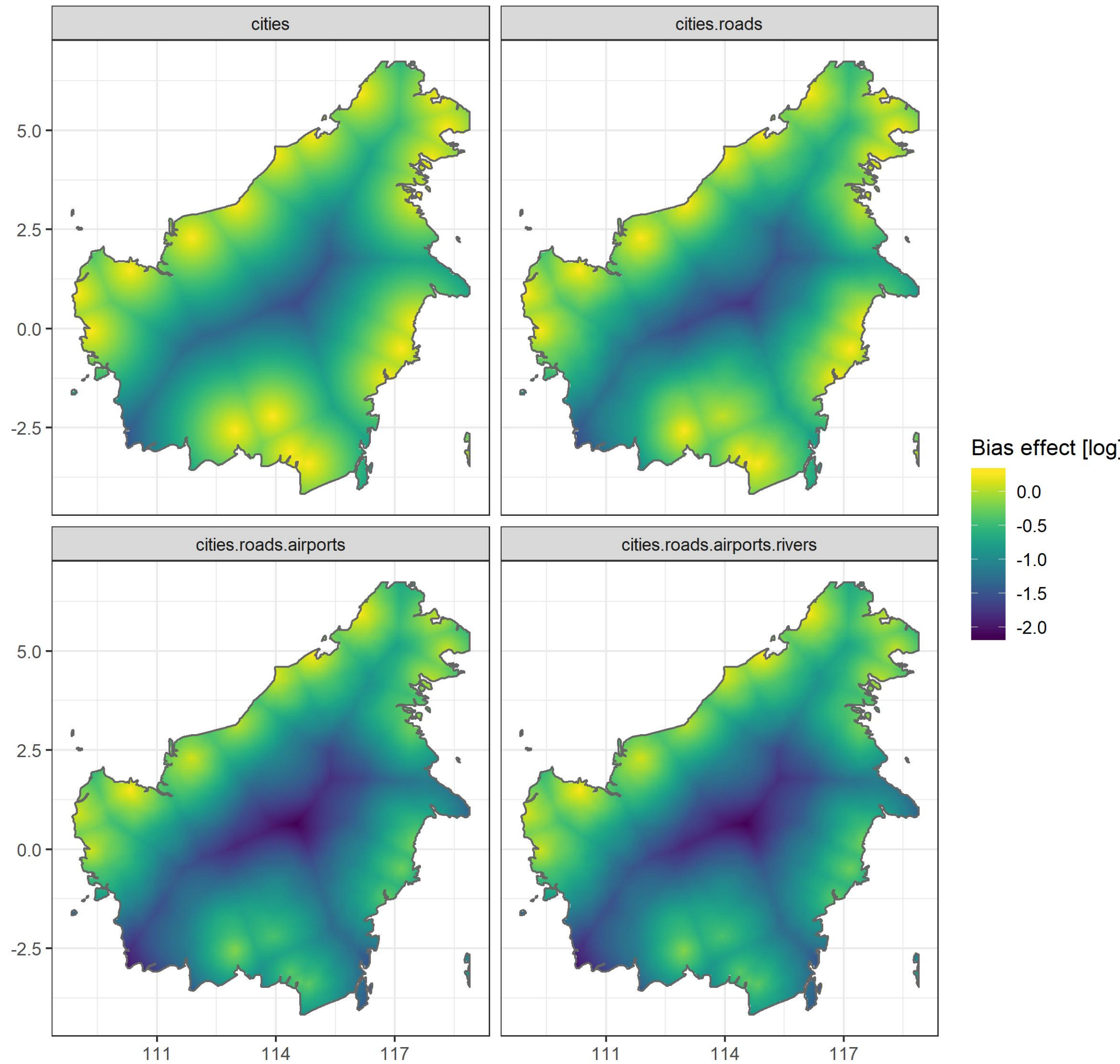

Figure 2: The spatial projection of the estimated sampling rates in an empirical example dataset of mammal occurrences on the Indonesian island of Borneo from www.gbif.org. The colours show the projection of the sampling rates (i.e. expected number of occurrences per cell) given the inferred extitsampbias model. The highest undersampling is in the center of the island.

Supplementary material

Appendix S1 - Supplementary Figure S1

Appendix S2 - Tutorial running sampbias in R

Appendix S3 - Possible warnings and their solutions

References

Bache, S. M. and Wickham, H. 2014. magrittr: A Forward-Pipe Operator for R.

Barbosa, A. M. et al. 2013. Species-people correlations and the need to account for survey effort in biodiversity analyses. - Diversity and Distributions 19: 1188–1197.

Beck, J. et al. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. - Ecological Informatics 19: 10–15.

Bivand, R. S. et al. 2013. Applied spatial data analysis with R, Second edition. - Springer.

Boakes, E. H. et al. 2010. Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. - PLoS Biology 8: e1000385.

Boria, R. A. et al. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. - Ecological Modelling 275: 73–77.

Botts, E. A. et al. 2011. Geographic sampling bias in the South African Frog Atlas Project: Implications for conservation planning. - Biodiversity and Conservation 20: 119–139.

Daru, B. H. et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. - New Phytologist 217: 939–955.

Engemann, K. et al. 2015. Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. - Ecology and Evolution 5: 807–820.

Fernández, D. and Nakamura, M. 2015. Estimation of spatial sampling effort based on presence-only data and accessibility. - Ecological Modelling 299: 147–155.

Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. - Methods in Ecology and Evolution 6: 424–438.

Fourcade, Y. et al. 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: e97122.

Garnier, S. 2018. viridis: Default color maps from 'matplotlib'.

GBIF.org 2016. (08 September 2016) GBIF occurrence download.

Hijmans, R. J. 2019. geosphere: Spherical Trigonometry.

Hijmans, R. et al. 2000. Assessing the geographic representativeness of Genebank collections: The case of Bolivian wild potatoes. - Conservation Biology 14: 1755–1765.

Isaac, N. J. B. and Pocock, M. J. O. 2015. Bias and information in biological records. - Biological Journal of the Linnean Society 115: 522–531.

Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. - Ecological Applications 14: 401–413.

Kery, M. and Royle, J. A. 2016. Applied hierarchical modeling in ecology - Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models. - Academic Press, Elsevier.

Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. - Diversity and Distributions 19: 1366–1379.

Lin, Y.-p. et al. 2015. Uncertainty analysis of crowd-sourced and professionally collected field data used in species distribution models of Taiwanese moths. - Biological Conservation 181: 102–110.

Meyer, C. et al. 2015. Global priorities for an effective information basis of biodiversity distributions. - Nature Communications 6: 8221.

Meyer, C. et al. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. - Ecology Letters 19: 992–1006.

Monsarrat, S. et al. 2019. Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. - Ecography 42: 125–136.

Pebesma, E. J. and Bivand, R. S. 2005. Classes and methods for spatial Data: the sp Package. - R News 5: 21–41.

R Core Team 2019. R: A language and environment for statistical computing.

Rocchini, D. et al. 2011. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. - Progress in Physical Geography: Earth and Environment 35: 211–226.

Ruete, A. 2015. Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. - Biodiversity Data Journal 3: e5361.

Rydén, O. et al. 2019. Linking democracy and biodiversity conservation: Empirical evidence and research gaps. - Ambio in press.

Shimadzu, H. and Darnell, R. 2015. Attenuation of species abundance distributions by sampling. - Royal Society Open Science 2: 140219.

Stolar, J. and Nielsen, S. E. 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. - Diversity and Distributions 21: 595–608.

Vale, M. M. and Jenkins, C. N. 2012. Across-taxa incongruence in patterns of collecting bias. - Journal of Biogeography 39: 1744–1744.

Varela, S. et al. 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. - Ecography: 1084–1091.

Warren, D. L. et al. 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: 334–343.

Wickham, H. 2009. ggplot2 - Elegant graphics for data analysis. - Springer.

Wickham, H. 2019. forcats: Tools for working with categorical variables (Factors).

Wickham, H. and Henry, L. 2019. tidyr: Tidy messy data.

Wickham, H. et al. 2019. dplyr: A grammar of data manipulation.

Yang, W. et al. 2013. Geographical sampling bias in a large distributional database and its effects on species richness-environment models. - Journal of Biogeography 40: 1415–1426.

Yang, W. et al. 2014. Environmental and socio-economic factors shaping the geography of floristic collections in China. - Global Ecology and Biogeography 23: 1284–1292.