Appendix S3 - The effect of dataset size on parameter estimates

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

Alexander Zizka, Alexandre Antonelli, Daniele Silvestro

The density of records might affect to estimate the sampling rate and bias weights (i.e. "How many records are needed to get reliable estimates?". To test the sensitivity of sampbias to the number of records in the input dataset, we ran two simulation studies:

1) As an example for a biased data set, we used the empirical example dataset with occurrence records of mammals from the island of Borneo. We down sampled the records by the factors 0.5, 0.25, 0.1, 0.01, 0.001 times the original number of records (i.e. 3131, 1566, 626, 62, and 6 records for all of Borneo) and ran a separate sampbias analyses on each of these down sampled datasets and compared the estimated posterior weights for each biasing factor. We replicated this analyses 5 times (A total of 25 analyses). The estimates of the bias weights were similar across analyses, suggesting that sampbias can robustly quantify sampling bias even with low record densities (Fig. 1).
2) We simulated seven data sets with 100,000, 50,000, 10,000, 1,000, 100, and 10 records randomly distributed across Borneo (no accessibility bias) and ran a separate sampbias analyses for each of these data sets, and replicated the analyses five times (a total of 35 analyses). The estimates of the bias weights were similar across analyses, suggesting that sampbias is robust against false positives in bias detection even in datasets with low record density (Fig. 2).

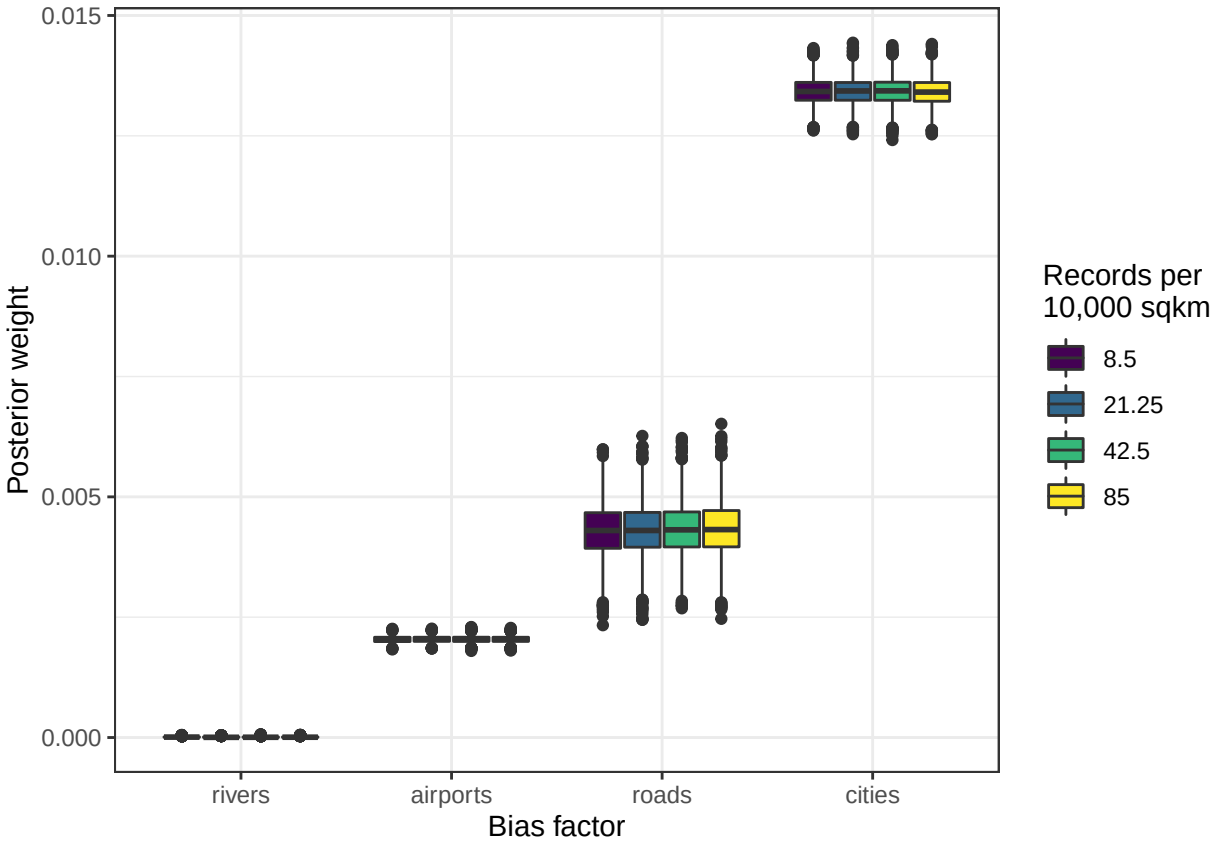

Figure S3.1: The bias weights (w) defining the effects of each bias factor estimated from data sets with differing density of occurrence records across the study area. Datasets generated by down sampling the empirical example dataset of mammal occurrences on Borneo.

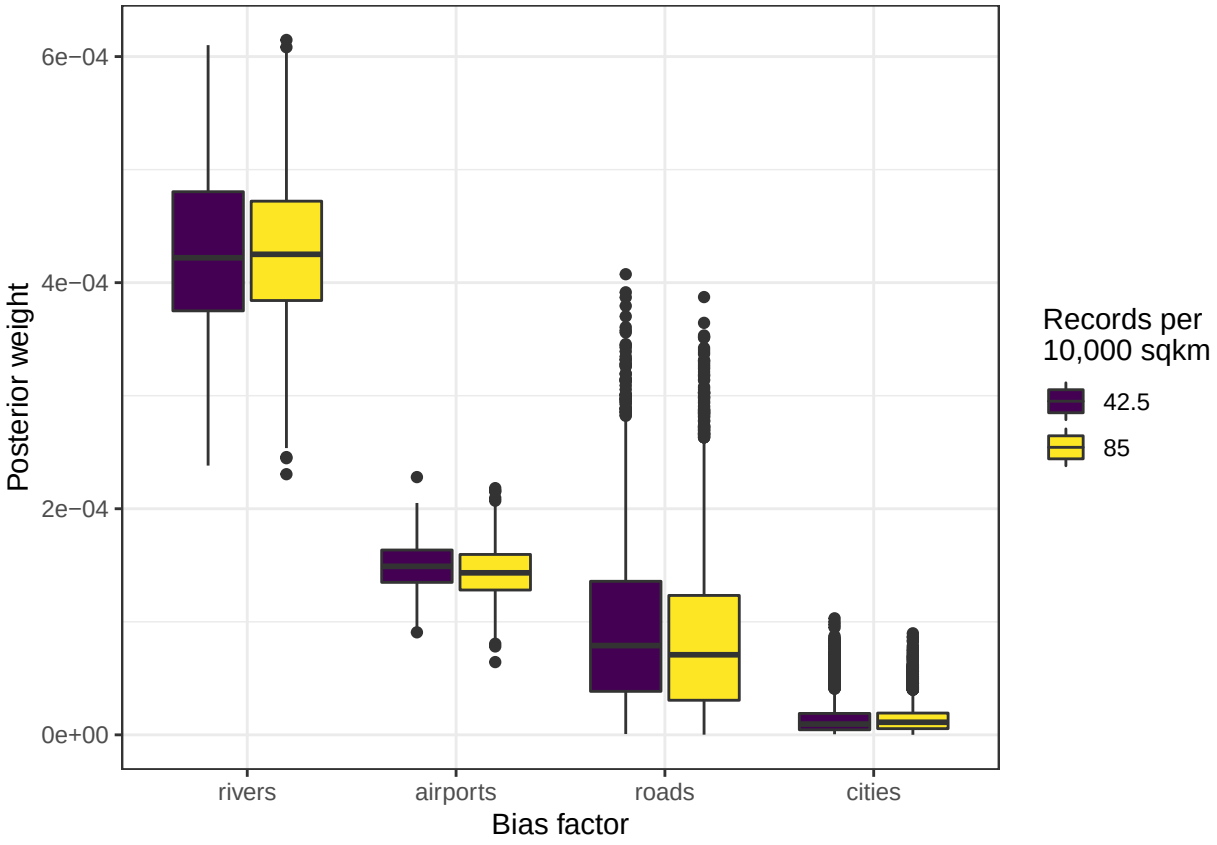

Figure S3.2: The bias weights (w) defining the effects of each bias factor estimated from data sets with differing density of occurrence records across the study area. Datasets generated by random simulation of records across Borneo (no bias).