Supporting Information

Table S1 Selected R tools that can help with missing data problems and their potential application for use in biodiversity research.

R packages	Applications	Useful functions
Exploring missing data		
naniar	visualizing/exploring the missing data pattern	*mcar_test* – Little's missing completely at random (MCAR) test *vis_miss* – plot the missing data for all variables
occAssess	measure of the potential for bias in taxonomic, temporal, spatial, and environmental dimensions	*assessEnvBias* - assess whether data are sampled from a representative portion of environmental space in the spatial domain of interest *assessSpatialBias* – assess whether data resemble a random distribution in the geographic space of interest for inference *assessSpatialCov* – assess whether a representative portion of the spatial domain of interest has been sampled and whether the same portion of geographic space has been sampled over time
sampbias	a Bayesian approach to estimate how sampling rates vary as a function of proximity to one or multiple bias factors	*calculate_bias* - calculating the bias effect of sampling bias due to geographic structures, such as the vicinity to cities, airports, rivers and roads
Subsampling		
base	Base R functions	*sample* - sample data with predefined inclusion probabilities specified with the prob argument
sampling	draw random samples using different sampling schemes	*balancedcluster* – selects a balanced cluster sample according to defined auxiliary variables *strata* - stratified sampling with unequal probabilities.

spatialEco	spatial data manipulation and modelling	*stratified.random* - creates a stratified random sample of an sp class object *stratified.distance* - draws a minimum, and optional maximum constrained, distance sub-sampling
spThin	Spatial thinning of species occurence records	*thin* - returns a dataset with the maximum number of records for a given thinning distance
terra	spatial data manipulation and processing	*spatSample* – sample a SpatRaster, SpatVector or SpatExtent objcy
Imputation		
agTrend	modelling regional trends with missing data	*mcmc.aggregate* - a zero-inflated, nonparameteric model with a definable observation model, augmenting missing values before calculating regional abundances
INLA/ inlabru	fitting Bayesian models, especially useful for spatial models via its spatial mesh	*inla/bru* - fit a Bayesian model using Integrated Nested Laplace approximation *predict* – draw predictions from the fitted model, where the prediction data frame can be a SpatialPointsDataFrame object
LORI	imputation of missing count data	*lori* – impute missing count data using a large covariate set, including interactions, with a LASSO penalty
mice	multiple imputation by chained equations	*mice* – multiple imputation method that will generate plausible values for any missing data – in the response and in any covariates
Rjags JAGS nimble	fitting Bayesian models allowing for missing values in the response	*Jags/runMCMC* - fitting Bayesian models allowing for imputation of missing values in the response during model fitting (options available for missing values in covariates too)
rtrim	functions to calculate annual indices and trends of abundances	fit a GLM imputing missing values based on mean site and year effects, with optional covariates
Weighting		

survey srvyr	range of functions for analysis of data from complex surveys, including fitting models with weights	*Svyglm* – generalized linear models with survey weights *postStratify* – function for post-stratification to match the joint distribution of the variables of the population
svrep	Analysis of replicate/boostrapped survey weights	*svyby_repwts* – compare estimates from different sets of weights
twang	functions to estimate propensity scores and weights	*ps* - gradient boosted trees to predict non-response from covariates *bal.table* – compare covariate values between sample and population

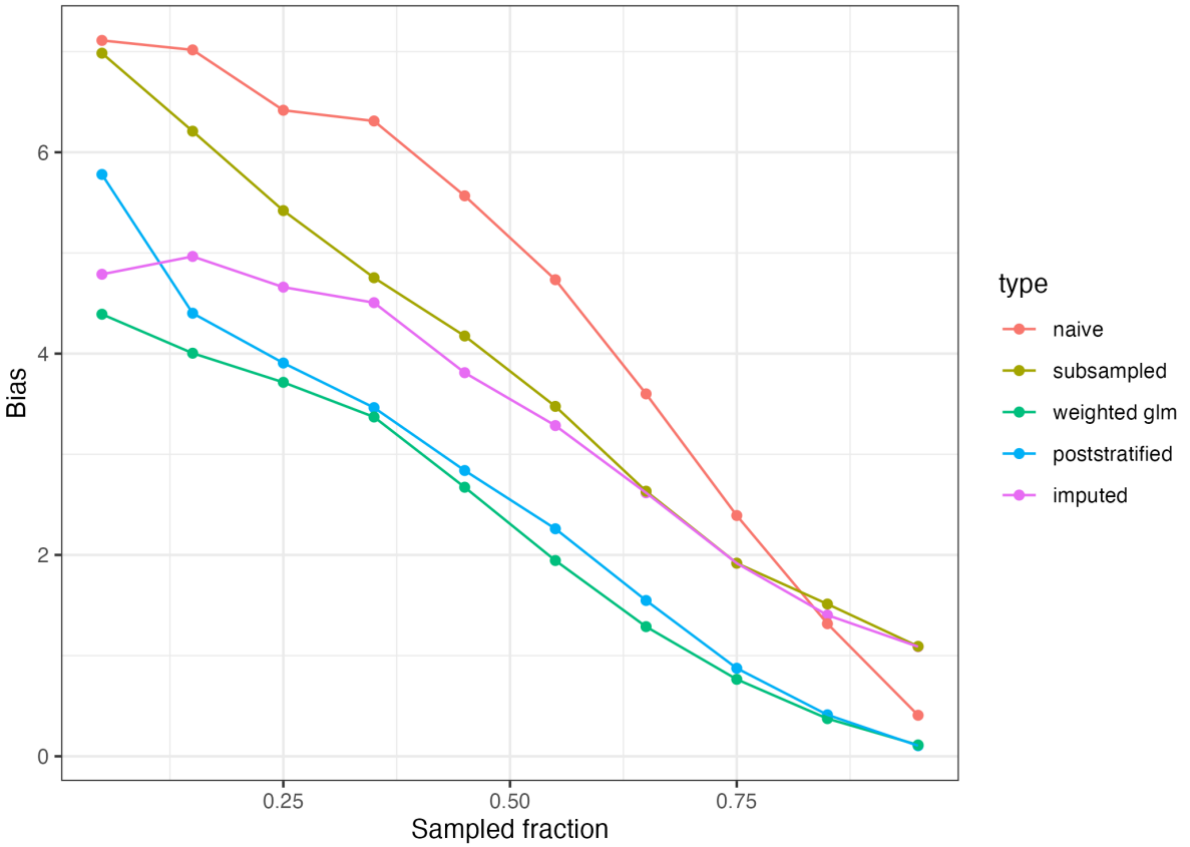

Fig. S1 The ability of missing data solutions to adjust for bias in biodiversity data.

We assumed a landscape of 400 cells and that a covariate affected both species abundance and the likelihood of a cell being sampled. We vary the fraction of the cells that were sampled. In contrast to Fig 5A (main text), we assumed that the species abundance was affected by an additional covariate that did not affect sampling; this variable was not included in any of the analysis. The models to estimate the parameter of interest (mean abundance) were: naive (no correction); subsampled (cells were subsampled along the covariate gradient to reduce the sampling bias), weighted (two methods: weighted glm using the svyglm function, and weighted by poststratification, using postStratify, both in the survey package) and imputed (using JAGS to impute NAs in the response). Points show the mean bias (difference between model prediction and truth) across 100 independent runs.