An operational workflow for producing periodic estimates of species occupancy at large scales

[1]Robin J. Boyd, [1] Thomas A. August, [1] Robert Cooke [1] Mark Logie, [1] Francesca Mancini, [1] Gary D. Powney, [1] David B. Roy, [1]Katharine Turvey, [1] Nick J. B. Isaac

[1] UK Centre for Ecology and Hydrology, MacLean Building, Benson Lane, Crowmarsh Gifford, Wallingford OX10 8BB, UK

Corresponding author: Robin J. Boyd (robboy@ceh.ac.uk)

Short title: From raw data to high-level summaries of biodiversity change

Keywords: biodiversity; Essential Biodiversity Variable; citizen science; occupancy model; species distributions

Abstract

Policy makers require high-level summaries of biodiversity change. However, deriving such summaries from raw biodiversity data is a complex process involving several intermediary stages. In this paper, we describe a workflow for generating annual estimates of species' occupancy at national scales from raw species occurrence data, which can be used to construct a range of policy-relevant biodiversity indicators. We describe the workflow in detail: from data acquisition, data assessment and data manipulation, through modelling, model evaluation, application and dissemination. At each stage, we draw on our experience developing and applying the workflow for almost a decade to outline the challenges that analysts might face. These challenges span many areas of ecology, taxonomy, data science, computing and statistics. In our case, a key output of the workflow is annual estimates of occupancy, with measures of uncertainty, for over 5,000 species in each of several defined "regions" (e.g., countries, protected areas, etc.) of the United Kingdom from 1970-2019. This product corresponds closely to the notion of a species distribution "Essential Biodiversity Variable" (EBV). Throughout the paper, we note where the workflow can be adapted to other situations (e.g., geographic regions or data types). We also highlight areas where the workflow can be improved; in particular, we suggest incorporation of methods to diagnose biases in the species occurrence data, to understand whether and to what extent these bias downstream products, and to mitigate them if needed. Finally, we compare the data products generated using our workflow to the first generation of species distribution EBVs and the "idealized" product as defined by others. Going forward, we hope that this paper can act as a template for research groups around the world seeking to develop similar data products.

Introduction

Information on the status of biodiversity and trends thereof is needed to monitor progress towards biodiversity targets and evaluate the effectiveness of conservation action. The rudiments of this information are primary (raw) data, but policy makers require high-level summaries such as indicators. The route from raw data to biodiversity indicator is not straightforward because the data typically derive from disparate sources and are heterogeneous in terms of sampling protocol, extent and resolution (grain size). To bridge this gap, the Group on Earth Observations Biodiversity Observation Network (GEO BON) conceptualised Essential Biodiversity Variables (EBVs; Pereira et al., 2013) as intermediary products that synthesise the available information in a common spatial, temporal and taxonomic framework. Several categories of EBV have been characterised to summarise the major dimensions of biodiversity and biodiversity change: Genetic composition, Species populations (abundance or distribution), Species traits, Community composition, Ecosystem structure, and Ecosystem function (Pereira et al., 2013). Taken together, these EBVs form a key component of a global information infrastructure for biodiversity (Peterson and Soberón, 2018). For example, EBV-type data products underpin multinational biodiversity syntheses, such as the IPBES Global Assessment, Global Biodiversity Outlook, and the Biodiversity Indicators Partnership dashboard, and are increasingly being used at national and local levels (Vihervaara et al., 2017).

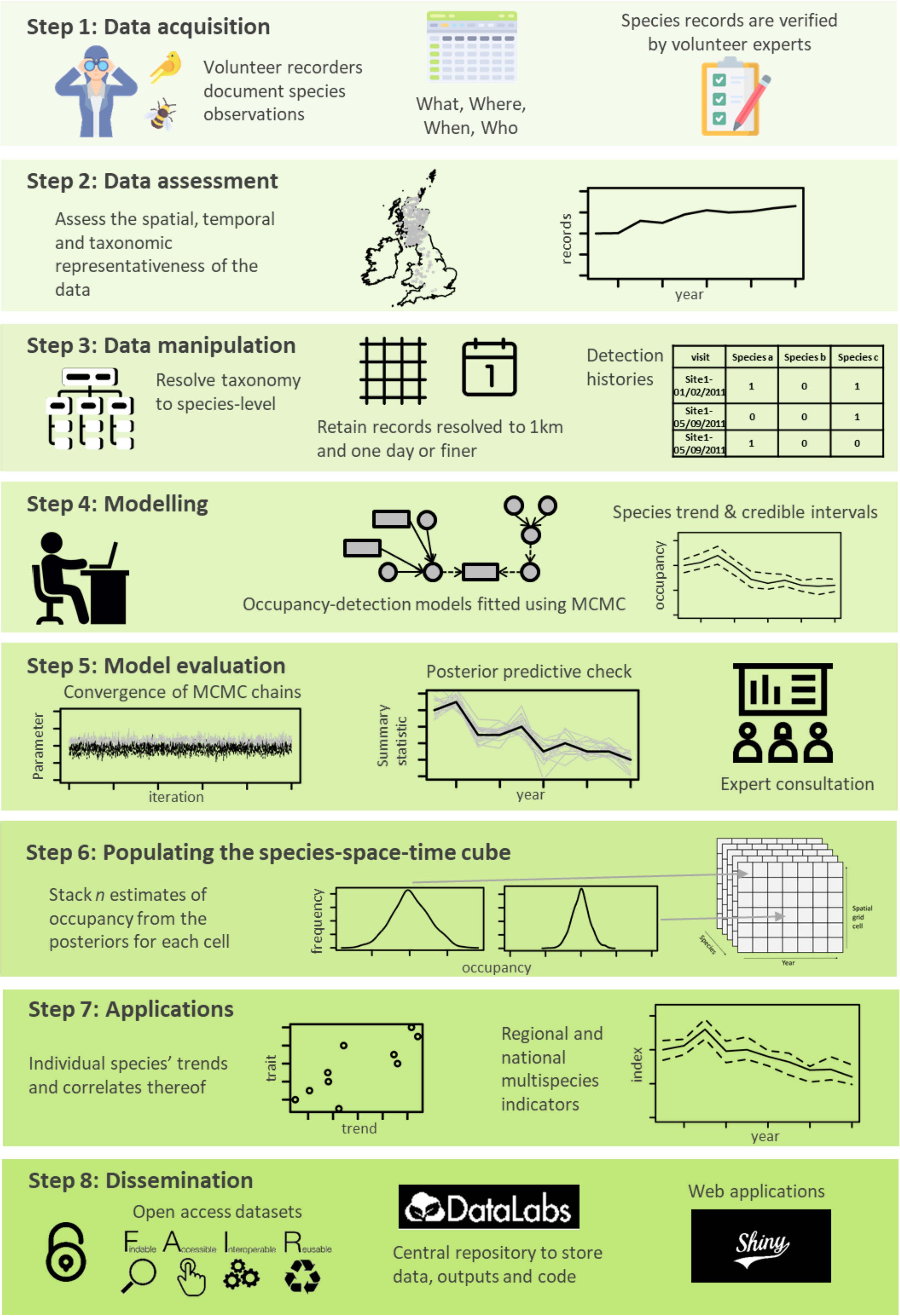
Species population EBVs characterise species' populations along the axes of taxonomy, space, and time (Jetz et al., 2019; Kissling et al., 2018). One way to view species population EBVs is as three-dimensional grids in which each cell denotes the status of some species' population in some spatiotemporal unit – the species-space-time cube (Fig. 2; Jetz et al., 2019; Kissling et al., 2018; Schmeller et al., 2017). Within each cell, population status may be quantified using one of two state variables: abundance, i.e., an index of the number of individuals present; or occurrence, i.e., whether at least one individual is present (or the probability thereof). The choice of state variable determines the specific category of EBV; that is, the species abundance or species distribution EBV, respectively. Abundance is often the preferred measure of species' population status (occurrence simply being a summary of abundance), but data on species' abundances are expensive and complicated to collect. Hence, for most taxa, places, and time periods – and therefore most cells in the species-space-time cube – occurrence is the only feasible measure of species' populations.

Populating the species-space-time cube with information on species' occurrences requires data and models. Structured monitoring data are the gold standard but are not available for most taxa in most parts of the world. Instead, analysts must rely on unstructured, presence-only, data of the types held in biological collections or collected through citizen science initiatives. These data are available for more cells in the species-space-time cube than structured data, but not all cells; and the data are typically heterogeneous (Robin J Boyd et al., 2022a). Hence, modelling is required. Several types of model might be considered: correlative habitat suitability models (Amini Tehrani et al., 2021); deductive habitat suitability models, which are based on expert advice about habitat associations (e.g., https://mol.org/indicators/habitat); or models with a temporal component that estimate changes in species' occupancy (Outhwaite et al., 2020). These models vary in their suitability for the populating the spatial and temporal axes of the species-space-time cube, and the optimal choice is not always clear.

Moving beyond the choice of data and model, there is a growing literature on the multitude of steps required to create, evaluate and disseminate species distribution EBVs and derivatives such as biodiversity indicators. Kissling et al. (2018) and Jetz et al. (2019) proposed high-level workflows for developing species population EBVs (distribution and abundance). Hardisty et al. (2019) produced the "Bari Manifesto" comprising ten steps for producing interoperable EBVs of all categories. Rapacciuolo et al. (2021) proposed four general steps for mitigating the unstructured nature of community-contributed (or citizen science) data and using them to create indicators. These contributions have provided a framework for constructing species distribution and other EBVs. However, as noted by Fernández et al. (2020), "At present, fully operational workflows that facilitate the automated and widespread production of EBVs are missing".

In this paper we describe the steps of an operational and relatively mature workflow for generating periodic estimates of species occupancy over large spatial and temporal extents—a species distribution EBV—from presence-only species occurrence data. Our paper is not intended to be a review of the field of EBVs, which is covered elsewhere (e.g., Jetz et al., 2019; Kissling et al., 2018). Rather, we document the decision-making process at each so they can be replicated and adapted by others. The paper is aimed at scientists working on national biodiversity infrastructures and researchers developing biodiversity indicators. The details and examples pertain to our experience working with citizen science groups and government agencies in the UK, but the general principles are widely applicable. For each of the eight steps (Figure 1), we begin with a general statement of the challenges that analysts might face, before describing the details of how we overcome them. The steps in our workflow relate to, but are not directly equivalent to, the frameworks cited above: we have chosen a structure that best encapsulates the practicalities of what we do. Having described each step in our workflow, we explain how these are implemented on a regular basis to update the EBV. We then discuss the “ideal” vs “minimal” requirements for species distribution EBVs [see Jetz et al. (2019) and Kissling et al. (2018)], noting which of the ideal criteria our product satisfies, and finally, the limitations of our current workflow and priorities for future development.

The workflow



Step 1: Data acquisition

Volunteer recorders document species observations

What, Where, When, Who

Species records are verified by volunteer experts

Step 2: Data assessment

Assess the spatial, temporal and taxonomic representativeness of the data

Step 3: Data manipulation

Resolve taxonomy to species-level

Retain records resolved to 1km and one day or finer

Detection histories

visit	Species a	Species b	Species c
Site1-01/02/2011	1	0	1
Site1-05/09/2011	0	0	1
Site1-05/09/2011	1	0	0

Step 4: Modelling

Occupancy-detection models fitted using MCMC

Species trend & credible intervals

Step 5: Model evaluation

Convergence of MCMC chains

Posterior predictive check

Expert consultation

Step 6: Populating the species-space-time cube

Stack *n* estimates of occupancy from the posteriors for each cell

Step 7: Applications

Individual species' trends and correlates thereof

Regional and national multispecies indicators

Step 8: Dissemination

Open access datasets

Central repository to store data, outputs and code

Web applications

Figure 1. A schematic representation of our workflow as applied in the UK. In this case study, the raw data are biological records provided by taxon-specific schemes and societies, and the downstream products include regional and national indicators. Icons from Flaticon. MCMC denotes Markov Chain Monte Carlo methods used to fit the occupancy-detection models.

Raw data acquisition

The first task when constructing a species distribution EBV is to obtain reliable data on species' occurrences for as many cells in the species-space-time cube as possible. Many data types might be considered: preserved specimens from museums and herbaria (Jönsson et al., 2021), observational data documenting sightings of some taxon (Sullivan et al., 2014), and more modern forms of monitoring such as passive (e.g., acoustic) sensors and eDNA (August et al., 2015), amongst others. These data types have different properties, which has important implications for how they are treated downstream.

Data sources vary in terms of their reliability. For example, records from preserved specimens are generally reliable in terms of taxonomic identity but lack precise information on where and when they were collected. On the other hand, community-contributed data (e.g. from eBird) often come with precise information on where and when they were collected, but are more likely to contain misidentifications. Many data providers have procedures to identify dubious records: GBIF—a global data aggregator—flags records with various spatial, temporal and taxonomic issues; eBird (Sullivan et al., 2014) flags "unusual" records which are then reviewed by regional experts; and iNaturalist designates only those records which have been photographed and accepted by the community as "research grade". Analysts should consider the reliability of the available records when deciding whether they are suitable for further analysis.

In our workflow, we use observational species occurrence data. These data comprise information on the four "Ws" of biological recording: What was seen, Where, When and by Whom (Isaac and Pocock, 2015). Whilst providing the same information (the four "Ws"), the data derive from disparate sources such as structured surveys, atlas projects and mass participation projects aiming to engage audiences with a range of expertise. Hence, they comprise a mix of "opportunistic" records, checklists and inventories as well as structured monitoring with a defined protocol and repeated sampling of the same location between years (Pocock et al., 2015).

In the UK we are fortunate in that biological recording has a wide taxonomic coverage: there are more than 80 schemes and societies, each focussing on the compilation and review of records for a taxonomic group of interest (Baker et al., 2021; Pocock et al., 2015). Through collaboration with these schemes, we have access to >24 million records for >10,000 species of bryophyte, lichen, insect and non-insect invertebrate (noting that many species are removed downstream; see Data manipulation). It should be noted that we treat the data from each scheme, and hence for each taxonomic group (e.g., bees, bryophytes, spiders, etc.), independently (hereafter "datasets"). Treating the datasets in this way has several advantages, which we describe throughout.

Data assessment

Constructing a species distribution EBV is a matter of statistical inference: the analyst does not possess data on all cells in the species-space-time cube, so must instead rely on a sample (Fig. 2A, B). If this sample is not representative of the spatial, temporal and taxonomic dimensions of the cube, or sampling was heterogeneous in those cells with data, then there is a risk that the resultant occupancy estimates will be biased. It is therefore crucial to assess the representativeness of the available data.

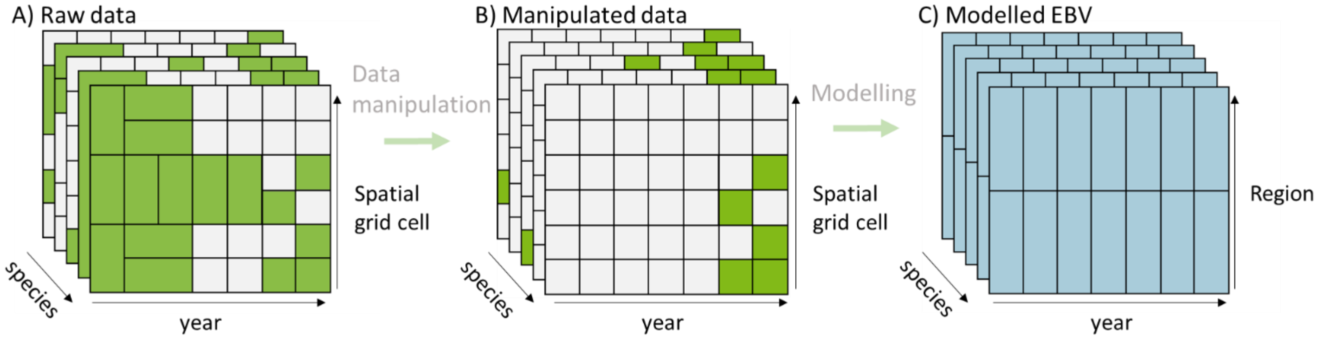


Figure 2. Progression of the species-space-time cube through various stages in the workflow. Grey cells indicate a lack of information, green cells indicate that data is available, and blue cells indicate that information on species' occupancy has been inferred through statistical modelling. Cube A represents the raw data. Note that data are available for many cells, but that the cells vary in size, which indicates variable spatial and temporal resolutions. Cube B represents the EBV-ready dataset (sensu Kissling et al., 2018), which is obtained after the data manipulation stage. At this step, spatially and temporally imprecise data have been removed, which is reflected by a common cell size, but also by the fact that fewer cells are populated. Cube C represents the modelled/derived EBV (sensu Kissling et al., 2018). We use occupancy-detection models to infer information on species' occupancy in every sampled cell in B, then calculated the proportion of those cells that are occupied in each "region" (e.g. country within UK).

Several tools are emerging to assess biases and other uncertainties in species occurrence data (Boyd et al., 2021; Robin J Boyd et al., 2022b; Zizka et al., 2019, 2021). One example is the R package occAssess, which takes a dataset and returns several heuristics indicating the potential for spatial, temporal, taxonomic and environmental biases (Boyd et al., 2021; Box 1). Whilst data-driven heuristics are useful, they are not a substitute for a thorough consideration of how such biases might impact on the estimates of species' distributions and how they change over time. New "risk-of-bias" assessments, first developed in medicine and related areas, are now being considered in ecology (Boyd et al., 2022): we see assessments of this nature as an important component of EBV workflows in future.

Box 1. Computer infrastructure and codebase.

Our workflow is underpinned by a virtual research environment —"DataLabs" [https://datalab.datalabs.ceh.ac.uk/]— accessible via a web browser. DataLabs is a tailorable, cloud-based research platform that supports end-to-end analysis and increases collaboration by allowing users (e.g., scientists, practitioners and stakeholders) to share working environments, code, data and visualizations from anywhere in the world. It is a realisation of the 'data science lab' concept introduced in Hollaway et al. (2020). DataLabs uses JASMIN (the Natural Environment Research Council's high performance computing facility) as the host computing platform, giving researchers seamless access to computer clusters, while taking advantage of the cloud scalability (Salama et al., 2022).

Inputs to and outputs of the analyses are stored on what we call the "Object Store". The Object Store is an S3 compatible object-based storage system that manages data as objects referenced by a globally unique identifier, with attached metadata, and underpins JASMIN and DataLabs. These objects exist in a single flat domain, allowing the Object Store to scale out much more easily than a traditional shared file system. The Object Store therefore ensures that our raw data and data products are easy to store, locate and access through DataLabs. All data products are archived on the Object Store, currently totalling >2,100GB across 55 model runs.
Models are fitted on DataLabs, and, where necessary, computer clusters are used to speed up processing. Clusters can be created within DataLabs for smaller tasks, or JASMIN's cluster facility – LOTUS – for larger jobs. LOTUS has direct access to the object store and vice versa, so data do not need to be copied between them manually.

Our workflow sits on an extensive codebase comprising four R packages, which are openly available on GitHub. The first is *occAssess*: this package facilitates screening of the species occurrence data for obvious potential biases (i.e., step 2 in our workflow). Next is *sparta* (August et al., 2020b), which contains all the code needed to prepare data for, and run, the occupancy models (i.e., steps 3-5 in our workflow), as well as other methods (e.g. Frescalo: Hill, 2012). s*parta* also contains helper functions for viewing model outputs. The third package, *BRCindicators* (August et al., 2022), contains methods for combining individual species' outputs to generate composite indicators (i.e., step 7 in our workflow). This package has been designed to work with the outputs of *sparta,* but also supports other data formats. The final package is *wrappeR* (Boyd et al., 2022c), which wraps around the BRCindicators package and a range of functions to streamline the processing of occupancy model outputs into multispecies indicators (i.e., steps 6 and 7 in our workflow).

Using DataLabs, JASMIN, the Object Store and the R packages listed above, we have developed an extensive pipeline that is collaborative, repeatable, efficient and FAIR.

Risk of bias assessments will reveal where mitigating action must be taken. This could include manipulating the data (e.g., thinning; Inman et al., 2021), attempting to correct for the biases statistically (van Strien et al., 2019), leveraging additional probability (random) samples where they are available (Isaac et al., 2020), redefining the extent and/or resolutions of the species-space-time cube to better reflect data availability (Pescott et al., 2019), or simply acknowledging that the data do not permit inference and proceeding with descriptive statistics instead. These steps generally come under the Data Manipulation and Modelling stages below.

Initial ROBITT assessments have revealed a lack of representativeness along all axes of the species-space-time cube. This can be explained at least in part by the fact that most of the scheme data were collected opportunistically, i.e., without a formal protocol. A corollary is that the distribution of sampling effort has varied across taxa, space and time in a non-random fashion (Pescott et al., 2019). For example, recorders have tended to preferentially sample accessible areas and rare species, and sampling intensity has generally increased over time (Isaac and Pocock, 2015). We outline steps that have been taken to try and mitigate these biases, and where additional action is required, below.

Data manipulation

Having assessed the raw data for biases, the next step is to prepare those data for modelling. This data manipulation step includes harmonisation to common spatial, temporal and taxonomic resolutions, cropping the data to the desired extents in those dimensions, and other types of (dis)aggregation and filtering. Note that for simplicity we present our workflow as linear, but it will likely be necessary to revisit the data assessment stage if the data are modified appreciably at this stage (e.g. if the data are substantially coarsened or reduced in extent).

Our choice of resolution is informed by several factors. First, we consider the resolution(s) at which the data were recorded. Second, we consider the trade-off between coverage (the proportion of each dimension in the species-space-time cube for which we have data) and resolution (Rapacciuolo et al., 2021). Finally, we consider assumptions related to our modelling framework; for example, the spatial and temporal resolution at which it is reasonable to assume population closure. At present, we unify the raw data at the species-level (with some exceptions due to taxonomic separation difficulties), 1km (British Ordnance Survey grid) and day-level resolution. This involves discarding imprecise records and duplicates—both true duplicates, i.e. multiple records of the same observation, and records that become duplicates at the specified resolution, i.e. same date, species and 1km grid square (henceforth “site”). However, acknowledging that the ecological and data generation processes differ among taxonomic groups, it may be preferable to move beyond our “one size fits all” approach in the future. One option would be to choose the spatial or temporal resolutions that result in the most even coverage (Box 3; Jönsson et al., 2021; Pescott et al., 2019) for each group. However, scale effects mean that estimates made at different resolutions are not directly comparable, so workflow design faces a trade-off between generality and specificity.

Having discarded imprecise and duplicate records, we organise the remainder of the data into “detection histories”: dataframes indicating whether each species was recorded on each visit (a unique combination of site and date). This step has three purposes: 1) to reverse engineer the survey structure (i.e., visits to some place on some day); 2) to infer non-detections of each species (what Rapacciuolo et al. 2021 referred to as “borrowing strength across taxa”); and 3) to approximate sampling effort per visit using the list length, i.e., the number of species recorded on that visit (Franklin 1999; Szabo et al., 2010; van Strien et al., 2013). Arranging the data in this way is possible because we treat the records for each taxonomic group as a combined dataset.

For many species, there is simply not enough data to estimate a trend in its distribution. A key question, therefore, is how to select which species should be taken forward to modelling in a way that introduces the fewest additional biases in the resulting data product. In the past, we used a subjective cut-off of 50 observations (Outhwaite et al., 2019). More recently, we have adopted thresholds based on the properties of those datasets that produce estimates with acceptable precision (Pocock et al., 2019). Specifically, we set thresholds for the number of observations in the most frequently observed years and the number of observation events that did not produce an observation of the focal species. An alternative approach would be to retain all species, even those which are likely to have low precision, to be transparent about our lack of knowledge about these species in downstream applications. Understanding the strengths and weaknesses of these choices is a priority as we further develop our workflow. Further research is required to explore whether these “rules of thumb” are transferrable, whether they are applicable to all taxa, or whether alternative selection criteria would be preferable.

In addition to the taxonomic filters described above, we also remove data from poorly sampled portions of the species-space-time cube. We exclude sites visited in one year only, since these cannot inform on changes in status over time (Isaac et al., 2014). It has been proposed to exclude all lists with fewer than a certain number of species recorded (Kamp et al., 2016). This and other filtering techniques are designed to amplify the signal:noise ratio in the data, but filtering also has the potential to amplify spatial biases in the set of locations sampled, which are typically large (Hughes et al., 2020).

Biases introduced through data manipulation could, in some cases, be mitigated by thinning, i.e., the removal of data from well-sampled as opposed to poorly sampled portions of the species-space-time cube. Thinning might also be used to address class imbalance (i.e., the ratio between detections and non-detections; Steen et al., 2020), or to reduce variation in sampling intensity over time (Hickling et al., 2006). Questions remain about the relative merits of retaining or removing data in the ways described above, and the optimal strategy will depend on the extent of the biases in the available data.

Modelling

For most, if not all, datasets, it is highly unlikely that reliable estimates of distributional trends at large (e.g., national) scales can be inferred from raw data alone. Rather, it is common to use statistical modelling to infer species' occupancy (or occurrence) in sites and years in which they were not observed. In our workflow we derive the "modelled and derived EBV" (Kissling et al., 2018; Fig. 2C) by fitting a type of occupancy-detection model to the detection histories described above. However, in some situations the available data will violate the assumptions of this model, in which case alternatives are available (Pescott et al., 2019). We expand on this point in the discussion.

We use single species multi-season occupancy-detection models (Altwegg and Nichols, 2019), in which each year is considered one "season". Other formulations, such as multispecies (Guzman et al., 2021; Ruiz-Gutierrez et al., 2010) and dynamic occupancy-detection models (Van Strien et al., 2013)—which explicitly describe colonisation and extirpation—might be preferable in some circumstances. The model structure comprises two hierarchically coupled Generalized Linear Models: the first, the state sub model, describes species' occupancy (i.e., presence vs absence); the second, the detection sub model, describes the data generation process. The key advantage of using occupancy-detection models is that they can, in the right circumstances, mitigate for uneven detectability (Royle, 2006).

State sub model

The state sub model describes the proportion of sites occupied (occupancy) in a given region and year for the focal species. In general, we make separate estimates for each country of the UK by including a year effect for each of these "regions", as well as for the UK as a whole (see the Applications section for extensions of this principle). The year effects are estimated using a random walk prior (Outhwaite et al., 2018), which reflects the fact that the occupancy status of most sites does not change from year to year. The state sub model also includes a random site effect (intercept) to allow for variation in occupancy status among grid squares (Isaac et al, 2014).

Our models are fitted to data from the subset of sites for which records are available (Fig. 1). As noted above, the spatial coverage of the data is not representative of geographic or environmental space in the UK, which limits the degree to which our occupancy estimates can be described as nationally or regionally representative. We are currently exploring options to address the issue of unrepresentative sampling locations (e.g., by including environmental covariates). We briefly review these in Box 3.

Detection sub model

The detection sub model describes the probability that the focal species is detected given that it is present. The probability of detection is clearly contingent on sampling effort (Franklin, 1999), which must be accounted for. Ideally, we would have visit level meta-data to provide a proxy for sampling effort, e.g., time spent searching (Sullivan et al., 2014). However, the only data we currently have available are the number of species recorded from the focal taxonomic group, i.e., the list length (Franklin, 1999). If list length is a reasonable proxy for sampling intensity, including it as a covariate will improve model performance (Isaac et al., 2014). Parameterising the list length effect as a monotonic function (Szabo et al., 2010) is appropriate when the majority of records derive from checklists, in which zeros in the detection history represent genuine non-detections. However, in opportunistic datasets, zeros often represent selective reporting; it is therefore more appropriate to treat categories of list as distinct data types (Van Strien et al., 2013), which allows for the possibility that detection might be highest on short lists (e.g., if sampling is strongly preferential). Specifically, we estimate the difference in the logit of the probability of detection for lists of length 2-3 and 4+ relative to lists of length 1 (Outhwaite et al., 2019). Whilst this parameterisation is the most appropriate for opportunistic data, further work is required to explore the sensitivity of results to the choice of boundaries between categories of list, particularly for speciose groups and where there are strong gradients in species richness.

Uneven sampling effort is only one source of heterogeneity in the data generation process. For any one species, there are at least three additional factors that might influence the probability of being observed and reported on a given list. Heterogeneity among observers is a particular source of concern for citizen science datasets (Lewandowski and Specht, 2015). These “observer effects” are usually discussed in the context of expertise in ecology (knowing where to look) and taxonomy (recognising what you see). A less-appreciated form of observer effects is variation in the probability that an observation will be reported. Selective reporting arises from of the tendency of some observers to record opportunistically, i.e., when something interesting or unusual is spotted. This behaviour would lead to an under-recording of commonly-encountered species (August et al., 2020a). Accounting for observer identity has been shown to improve the performance of spatial distribution models (Johnston et al., 2018), so incorporation of observer effects in our workflow is desirable. At present we are hindered by the fact that observer identities are not regularised in most of the scheme datasets. The increased adoption of online recording technologies (e.g. iNaturalist – www.inaturalist.org; iRecord; www.brc.ac.uk/irecord) offers the potential for a solution in the longer term. A second important source of heterogeneity in detection probability is the observation date (most species have seasonal life-history). van Strien et al. (2013) addressed this problem by modelling the phenology of detection as a quadratic function of Julian date. Our experience is that the parameters of this function are not mutually identifiable; hence, we have explored modelling phenology using a Gaussian distribution, in which the mean and standard deviation of detection dates are estimated. The Gaussian function is suitable for many species with annual life-cycles, but not for long-lived or multi-voltine species, in which case a different formulation is required, perhaps involving splines (Crainiceanu et al., 2005) or via additional levels of the hierarchy (Direnzo et al., 2021). Finally, detection is more likely on sites with abundant populations: ignoring this variation can lead to biased estimation in occupancy models (Royle and Nichols, 2003).

Many datasets we encounter have few repeat visits to the same site on different dates in the same year, which are necessary for estimating detection probabilities. There has been some debate about whether it is appropriate to model detectability in this situation, or whether it is better to estimate occupancy naively (i.e., assuming detectability = 1; Guillera-Arroita et al., 2014; Welsh et al., 2013). This decision on how to proceed depends on what the analyst considers to be useful information. Where repeat visits are few, estimates of occupancy are likely to be uncertain because the model does not know whether non-detections reflect absences or low detectability (i.e., multiple samples from the joint posterior of the parameters might fit similarly well). On the other hand, estimating occupancy naively will introduce a bias, especially if there is variation in detection probabilities over time (Isaac et al 2014). We have chosen to estimate detectability but acknowledge this may introduce biases where there is heterogeneity in site selection, recorder behaviour and detectability (see above), and that it interacts with the a priori removal of species that are likely to produce uncertain trends. In future, we plan to assess the sensitivity of our outputs to these methodological decisions.

Model fitting

We fit the occupancy-detection models to the detection histories in a Bayesian framework using Markov Chain Monte Carlo (MCMC) implemented in JAGS (Plummer, 2003) via the R package *sparta* (August et al., 2020b). In Outhwaite et al. (2019) we ran each model on three chains for 20,000 iterations with a burn in of 10,000 iterations and a thinning rate of three. In our most recent set, we used 32,000 iterations with a burn in of 30,000 and a thinning rate of six: the longer burn-in and higher thinning rate leads to improved mixing of the MCMC chains. These values were chosen to balance the trade-off between computation time and convergence, recognising that for some species there is insufficient data to achieve convergence for all parameters. Priors and hyperpriors are set to be uninformative (see Outhwaite et al 2018 for details) with two exceptions: a) the random walk in the state sub model (see above); and b) detection probability for single-species lists is set to have a prior mean of 0.12 (if recording was unselective and all species were equally detectable, then the probability of being recorded on a single species list would be 1/*n*, where *n* is the species richness of the average site).

Model evaluation

Having fitted statistical models to populate the species-space-time cube (Fig. 1C), the next step is to evaluate the performance of those models. Common measures of model performance include uncertainty and goodness-of-fit (i.e., the plausibility of the model given the data; MacKenzie and Bailey, 2004). Goodness-of-fit is typically evaluated using the data to which the model was fitted (training data). However, it is often useful to assess the degree to which a model matches independent data, or other forms of evidence.

For some species, there is insufficient information in the data to derive useful measures of change for all regions. Notwithstanding the *a priori* exclusion criteria described above, it is sometimes useful to exclude these species *a posteriori*. Several tools are available to assess this information content. One is the degree to which the parameter estimates from the MCMC chains have converged upon a common distribution. The Gelman-Rubin “Rhat” (Gelman and Rubin, 1992) is a convenient measure of convergence. A related measure of information content is the precision of the occupancy and trend estimates, which captures the degree to which the data have overcome the minimally informative prior. We assess precision and convergence, but do not exclude species based on these criteria: we reason that this will not bias downstream applications and that it is more transparent to propagate the uncertainty. In other situations, it might be preferable to remove species based on these metrics.

Goodness-of-fit is typically evaluated by comparing some fit statistic (e.g. X^2) describing the discrepancy between the predictions and observations with those from a reference distribution (Warton et al., 2017). The reference distribution is calculated by simulating many datasets under the model and calculating the equivalent fit statistics; that is, calculating the fit statistics that would be obtained if the model is a perfect representation of the system. Reference distributions may be constructed via bootstrapping for models analysed using classical inference (MacKenzie and Bailey, 2004), or as a natural by-product of the MCMC algorithm for models analysed in a Bayesian framework (Gelman et al., 1996; Royle et al., 2007). The latter approach, often called a “posterior predictive check”, can be used to calculate the posterior probability that the model provides a better fit to the simulated data than the observations (Kéry and Royle, 2016): this is often called a Bayesian P-value.

We have used the Bayesian P-value to evaluate our models in the past (Outhwaite et al., 2020). However, it has been shown to have limited ability to detect a lack of fit (Wright et al., 2019), and provides no information on which components of the model fit well or poorly (Warton et al., 2017). Residual plots, constructed for both the occupancy and detection components of the model (e.g., Warton et al., 2017), provide a promising alternative for diagnosing lack of fit.

Precision and goodness-of-fit are useful measures of model performance, but where the available data contain unmodelled heterogeneity (as in our case), neither necessarily indicates a model’s accuracy. For this reason, it would also be useful to consider independent model evaluation using either independent data or elicitation of expert opinion. In the past we have asked data providers whether model outputs are plausible (Powney et al., 2019), but formalised expert elicitation would be more objective (Mukherjee et al., 2018).

Populating the species-space-time cube

For each species, the software used to fit our models (see box 1) produces large samples from the posterior distribution of each parameter. For most applications, occupancy in each year for each region and species are the quantities of interest. We therefore extract 1,000 samples of the posterior distributions of occupancy for each species and region of interest to populate the final species-space-time cube (Fig. 1C). By retaining 1,000 samples, we can produce a point estimate (usually mean occupancy) and credible intervals for each cell of the cube, whilst allowing parameter uncertainty to be propagated, e.g. to multispecies indicators.

Applications

Having populated the species-space-time cube, the next step is to apply the cube for scientific research and to inform policy (Jetz et al., 2019). Here, we focus on the applications for which our workflow was designed; these all involve the estimation of *temporal trends* in species’ occupancy.

Species trends

Estimates of occupancy for each species in each year across some spatial domain can be extracted from the species-space-time cube. These can be used to calculate temporal change as mean annual growth rates or linear trends. Species-level trends are useful for identifying correlates of range contractions and expansions (Bowler et al., 2021; Powney et al., 2014), tracking the spread of invasive species and their effects of native taxa (Roy et al., 2012), and conducting species Red List assessments (Maes et al., 2015), amongst other applications.

Multispecies indicators

Species' occupancy or trends thereof can be "averaged" over some set of taxa to produce multispecies indicators. For many applications, the geometric mean is a sufficient summary statistic (Outhwaite et al., 2020). More complex methods propagate the uncertainty from the individual species' trends, can handle missing values and can incorporate smoothing (Freeman et al., 2020; Soldaat et al., 2017); these are now preferred for national biodiversity indicators in the UK.

We have produced indicators for several taxonomic groupings and regions. These include the UK indicator of pollinating insects (JNCC, 2021a), an index of "priority species" at both UK (JNCC, 2021b) and England levels, and an index of terrestrial occupancy for ~2000 species in Scotland. Multispecies indicators of this type are perhaps the most important derivative of the species-space-time cube because they can be used to track progress towards biodiversity targets.

Comparing trends

As described above, our models include terms for regions within the UK. This provides a flexible way to assess regional variation in trends of specific groups, or to evaluate the impact of differing land management strategies (e.g., comparing between land cover types, or between grid squares inside vs outside protected areas; Cooke et al., in prep.). In this way, our data products can be tailored to spatio-temporal resolutions that are most useful for decision-makers and policy creation (Jetz et al., 2019) without the need to go back to the raw data.

Functional diversity

Occupancy estimates can be combined with species' trait data to estimate patterns of functional diversity in space or time. Using this approach, Greenop et al. (2021) assessed changes in pollination and pest control functions in Great Britain. This example demonstrates the potential of our workflow to inform on policy-relevant questions about ecosystem health, or to provide for other EBV categories (community composition, ecosystem functioning) (Pereira et al., 2013).

Dissemination

The final step in our workflow is to disseminate the outputs of the preceding stages to the relevant audiences. This might include policy makers, collaborators and the wider scientific community. We suggest that the dissemination stage should follow two general principles. First, data products should be FAIR: findable, accessible, interoperable and reusable (Wilkinson, 2016). Ensuring EBV data sets are FAIR means that they can be can easily be found and accessed by others, they use common standards that allow them to be combined with other EBV data sets, and they have appropriate metadata data describing the data and how it was generated (Box 2). Second, data products should be tailored to the target audience, the details of the use-case (e.g., species trends or multispecies indicators), and in an appropriate format (e.g., data and code versus interactive visualisation).

Box 2. Metadata.

We save EBVs at several points in our workflow (Fig. 2). Each time data are saved, metadata are stored in .rdata or .rds format. The model outputs for each species include metadata embedded as attributes in the R object (see the sparta package Box 1). These metadata include the name of the species modelled; the temporal and spatial coverage; the regions modelled and the quantity of data available in each region for the focal species; the model type, parameters and BUGS code; the Sparta version used; the date the model was fitted and the modeller who fitted it; the R session information; and provenance. Provenance is a free text field used to capture the rationale for the model run and/or to summarise data acquisition steps.

In addition to the species-level metadata, we store metadata at the "run" level (i.e. for all species in a taxonomic group). These files are generated, and updated, using the createMetadata function in the wrappeR package (Box 1), which summarises the input and output files from the Object Store (Box 1). The run-level metadata propagates metadata from the species level, where applicable, as well as summarising higher level metadata, such as the number of species modelled. Propagation of metadata means that the EBV data products shown in Fig. 2 retain information about the raw data from which they were derived and the model configuration, thus being reproducible. Run-level metadata is subsequently used by functions in the BRCindicators package (Box 1) to create multispecies indicators based on the latest model outputs.

Working in a large team, the primary audience for our datasets is people within the same organisation (UK Centre for Ecology and Hydrology). For this audience, we have built a computing environment that facilitates collaborative working (Box 1) supported by comprehensive metadata (Box 2). This computing environment allows controlled access by the wider scientific community via shared Notebooks in DataLabs (Box 1; Hollaway et al., 2020). We have also published occupancy trends for 5,293 species under an open government license with an accompanying data paper (Outhwaite et al., 2019; Outhwaite et al., 2019).

For non-technical audiences—such as staff in government agencies, NGOs and some members of the schemes who supplied the raw data—we have developed R Shiny web applications deployed via DataLabs (Box 1). These allow those users to browse outputs graphically without needing to download the underlying data.

We typically share our data products with policy-makers through reports. Examples include the triennial State of Nature reports and national biodiversity indicators. Typically, these reports document multispecies indicators for a taxonomic group and region of interest. The figures and the underlying data (indicator values plus uncertainty for each year, etc.) are made openly accessible (e.g., JNCC 2021a).

A priority for future development is to standardise our dissemination formats and improve interoperability. Hardisty et al. (2019) developed a road map for achieving the vision of interoperable EBVs in what they called the "Bari Manifesto". We cannot act on the Bari Manifesto unilaterally; rather, we must work with the EBV community to develop data standards. We would also like to develop APIs (Application Programming Interfaces) that provide access to our EBV datasets, in much the same way as existing APIs provide access to raw observations (e.g., iNaturalist and GBIF).

Implementing the workflow

It is likely that users will want to update their species-space-time cube and downstream products over time. In our case, we implement most stages of the workflow annually. This reflects the fact that we are funded to produce national indicators of species' distributions annually. However, we have neither the resources nor the data to update every taxonomic group each year. Typically, a handful of the ~30 taxonomic groups are updated, so for most groups the data are a few years out of date. Resourcing constraints mean that the Data assessment and Model evaluation steps are implemented less frequently, but we are working to change this.

Discussion

We have provided a step-by-step description of the workflow that underpins our EBV-type data product and indicators of species' distributions in the UK. This description spans the entire process – starting with the acquisition of raw data, through data assessment and data manipulation, modelling and model evaluation, and finally application and dissemination. Throughout, we hope to have given an honest appraisal of the strengths of our workflow and where it can be improved in future. As such, we hope that this paper will be a useful resource for other groups/organisations seeking to construct similar products.

Applicability of the occupancy-detection model

Our workflow is built around the occupancy-detection model, but this will not be the most appropriate method in all situations (Box 3). If the available data are severely biased at fine scales, then it will be necessary to work at coarser resolutions at which those biases are less evident (Pescott et al., 2019). However, the occupancy-detection model assumes that species' occupancy at each site does not change within "closure periods" (here one year); as the definitions of the site and closure period become coarser, this assumption becomes less tenable. Likewise, it becomes less realistic to suppose that repeat visits to a site pertain to the same location. Where alternative analytical approaches are deemed more appropriate, the general structure of our workflow will still apply, but the detail will differ.

Comparison with the first generation of species distribution EBVs

Our data products differ from the first generation of species distribution EBVs in two key ways. First, our product pertains to >5,000 species, much more than most. Second, the majority of the first generation EBVs were constructed using correlative or deductive species distribution models (SDMs) that lack any temporal component (Amini Tehrani et al., 2021; Fernández et al., 2020; Velásquez-Tibatá et al., 2018; also see e.g., https://portal.geobon.org/ebv-detail?id=5 and https://mol.org/indicators/habitat). In contrast, our occupancy models are temporally explicit. Both types of model *can* populate the spatial and temporal dimensions of the species-space-time cube: for occupancy models we would need to include spatial terms as SDMs do; projecting static SDMs to new time periods is possible by assuming a space-for-time substitution.

Comparison with the "ideal" species distribution EBV

Kissling et al. (2018) set out seven criteria for the "ideal" species distribution EBV; the data products produced using our workflow in the UK satisfy some but not all of these. The first is that both presence and absence data should be used. Technically, we work with presence-only data and infer non-detections based on assemblages of co-recorded species. The second criterion is that the EBV should be global in extent: our previous data products do not meet this criterion because we focused on the UK. The third criterion is that the EBV should have a fine spatial resolution; this is true of our EBV-ready dataset (Fig. 1B) but not our model derived EBV (Fig. 1C) which is coarsened during the modelling stage (by summing occupancy states across sites within each region). Our EBV satisfies the next two ideal criteria: we provide a continuous long-term time series spanning several decades, and the temporal resolution (annual) is sufficient to capture the focal species' population dynamics. We are some way toward the criterion about taxonomic and ecological representation: we include a large pool of species (>5,000) spanning bryophytes, lichens, insects and non-insect invertebrates. However, there are other groups, such as mammals, birds, herpetofauna and vascular plants – for which occupancy data are available but which are not currently included in our EBV. We do satisfy the final criterion—that taxonomic dictionaries should be updated according to published checklists (e.g., species aggregates and synonyms). In our workflow taxon names align with the UK Species Inventory (UKSI; https://www.nhm.ac.uk/our-science/data/uk-species.html), which standardises checklists for more than 70,000 species and integrates with the GBIF backbone taxonomy. In summary then, the data products that we have produced in the UK fall somewhere between the minimal and ideal products as defined in Kissling et al. (2018).

Whilst our previous data products do not constitute ideal species distribution EBVs, they are closer than most of the first generation, and it might be possible to get closer still using the existing pipeline. It would be relatively simple, for example, to increase the spatial and taxonomic extents where data are sufficient. Our current extents and resolutions reflect the aims of our research group.

It is worth pointing out that the "ideal" species distribution EBV is likely unattainable, as acknowledged by Kissling et al. (2018). First, there are trade-offs between criteria. For example, working at a 1km spatial resolution precludes inclusion of species for which such precise data are not available. The stipulation that species distribution EBVs should be global and temporally explicit is also optimistic, given current data availability (Hughes et al., 2020; Peterson and Soberón, 2018). For the foreseeable future, species distribution EBVs will be most useful if constrained in spatial or taxonomic domains, and/or if coarse resolutions are employed.

Conclusion

To tackle the ongoing biodiversity crisis, data products are needed that are accurate, synthetic, synoptic, and interoperable. Our workflow and paper represent a step towards this ambition, yet as noted, substantial challenges remain (Box 3). We hope that research groups around the world will adopt our workflow, but consider these challenges, which are likely to be more acute in regions sampled less comprehensively than the UK (e.g. Boyd et al., 2022a). In this way, we can make progress towards a better understanding of global biodiversity change.

Box 3. Outstanding questions and priorities for future development of the workflow

There remain outstanding questions at all stages in our workflow, the most pressing of which are outlined below.

Can we statistically correct for a wider range of biases in the species occurrence data? At present, we construct our EBV using an occupancy-detection model that does not mitigate all biases. Options to improve the models include extra terms in the state (e.g. environmental covariates) and observation sub models (especially if accompanied by additional metadata from data providers); weighting or thinning the data; and integration with structured datasets (where available), amongst others.

How do we evaluate model adequacy? Implementing statistical fixes for data biases is one thing; assessing whether these were successful is another. Model evaluation is particularly difficult where the comparison data are biased, because a model with similar biases will appear to fit the data better than an unbiased one. More work is needed to understand which goodness-of-fit measures are most effective, and to establish best practices for leveraging independent information (e.g. from experts or structured data).

What are the optimal species inclusion criteria and are they generalizable? For some species the data are so few that we can say little about their distributions. In this situation there are two options: 1) ignore the poorly-recorded species and focus on those with more data; or 2) accept the uncertainty and include all species to maximise taxonomic coverage. At present we drop species based on the “rules of thumb” described in the text, but it might be preferable to take a different approach in other circumstances.

Is the one-size-fits-all approach appropriate? We estimate occupancy for each species at the same resolutions and extents using the same model. This “one-size-fits" all approach is relatively simple, easy to implement and produces comparable outputs. However, questions remain about whether more bespoke models that capture taxonomic idiosyncrasies might be more appropriate, and how best to combine the outputs of such models.

Acknowledgements

This work has been supported by the Natural Environment Research Council award number NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability. We are grateful to the many volunteer-run recording schemes for making their data available, and for constructive comments on the modelling framework. We would also like to thank Diana Bowler for her helpful comments on the manuscript.

References

Altwegg, R., Nichols, J.D., 2019. Occupancy models for citizen-science data. Methods Ecol. Evol. 10, 8–21. https://doi.org/10.1111/2041-210X.13090

Amini Tehrani, N., Naimi, B., Jaboyedoff, M., 2021. Modeling current and future species distribution of breeding birds as regional essential biodiversity variables (SD EBVs): A bird perspective in Swiss Alps. Glob. Ecol. Conserv. 27, e01596. https://doi.org/10.1016/j.gecco.2021.e01596

August, T., Fox, R., Roy, D.B., Pocock, M.J.O., 2020a. Data-derived metrics describing the behaviour of field-based citizen scientists provide insights for project design and modelling bias. Sci. Rep. 10, 1–13. https://doi.org/10.1038/s41598-020-67658-3

August, T., Powney, G., Outhwaite, C., Harrower, C., Hill, M., Hatfield, J., Mancini, F., Isaac, N., 2020b. sparta: Trend Analysis for Unstructured Data. R package version 0.2.18.

August, T.A., Harvey, M.C., August, T.O.M., Harvey, M., Lightfoot, P., Kilbey, D., Papadopoulos, T., Jepson, P., 2015. Emerging technologies for biological recording. Biol. J. Linn. Soc. https://doi.org/10.1111/bij.12534

August, T.A., Powney, G.D., Outhwaite, C.L., Hatfield, J., Logie, M.W., Isaac, N.J.B., 2022. BRCindicators: creating multispecies biodiversity indicators.

Baker, E., Drury, J.P., Judge, J., Roy, D.B., Smith, G.C., Philip, A., 2021. The Verification of Ecological Citizen Science Data : Current approaches and future possibilities Presenting author 2011–2013. https://doi.org/10.3897/biss.5.75506

Bowler, D.E., Klaus-, D.E., Conze, J., Suhling, F., Baumann, K., Benken, T., Bönsel, A., Bittner, T., Drews, A., Günther, A., Isaac, N.J.B., Petzold, F., 2021. Winners and losers over 35 years of dragonfly and damselfly distributional change in Germany 1353–1366. https://doi.org/10.1111/ddi.13274

Boyd, R., Powney, G., Carvell, C., Pescott, O.L., 2021. occAssess: An R package for assessing potential biases in species occurrence data. Ecol. Evol. https://doi.org/10.1002/ece3.8299

Boyd, Robin J, Aizen, M.A., Prado, L.F.-, Fontúrbel, F.E., Francoy, T.M., Martinez, L., Morales, C.L., Ollerton, J., Pescott, O.L., Powney, G.D., Mauro, A., Reto, S., Eduardo, S., Carvell, C., 2022a. Inferring trends in pollinator distributions across the Neotropics from publicly available data remains challenging despite mobilization efforts. Divers. Distrib. 1–12. https://doi.org/10.1111/ddi.13551

Boyd, Robin J., Cooke, R., August, T.A., Powney, G.D., Mancini, F., Logie, M.W., Isaac, N.J.B., 2022. wrappeR: A wrapper for producing biodiversity indicators.

Boyd, Robin J, Powney, G.D., Burns, F., Danet, A., Duchenne, F., Grainger, M.J., Jarvis, S.G., Martin, G., Nilsen, E.B., Porcher, E., Stewart, G.B., Wilson, O.J., Pescott, O.L., 2022b. ROBITT: A tool for assessing the risk-of-bias in studies of temporal trends in ecology. Methods Ecol. Evol. 2022, 1–11. https://doi.org/10.1111/2041-210X.13857

Crainiceanu, C.M., Ruppert, D., Wand, M.P., 2005. Bayesian analysis for penalized spline regression using WinBUGS. J. Stat. Softw. 14. https://doi.org/10.18637/jss.v014.i14

Direnzo, G. V, Miller, D.A.W., Grant, E.H.C., 2021. Ignoring species availability biases occupancy estimates in scale occupancy models. Methods Ecol. Evol. 1–15. https://doi.org/10.1111/2041-210X.13881

Fernández, N., Ferrier, S., Navarro, L.M., Pereira, H.M., 2020. Essential biodiversity variables: Integrating in-situ observations and remote sensing through modeling. Remote Sens. Plant Biodivers. 485–501. https://doi.org/10.1007/978-3-030-33157-3_18

Franklin, D.C., 1999. Evidence of disarray amongst granivorous bird assemblages in the savannas of northern Australia, a region of sparse human settlement. Biol. Conserv. 90, 53–68. https://doi.org/10.1016/S0006-3207(99)00010-5

Freeman, S.N., Isaac, N.J.B., Besbeas, P., Dennis, E.B., Morgan, B.J.T., 2020. A Generic Method for Estimating and Smoothing Multispecies Biodiversity Indicators Using Intermittent Data. J. Agric. Biol. Environ. Stat. https://doi.org/10.1007/s13253-020-00410-6

Gelman, A., Meng, X.L., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin. 6, 733–807.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7, 457–511.

Greenop, A., Woodcock, B.A., Outhwaite, C.L., Carvell, C., Pywell, R.F., Mancini, F., Edwards, F.K., Johnson, A.C., Isaac, N.J.B., 2021. Patterns of invertebrate functional diversity highlight the vulnerability of ecosystem services over a 45-year period. Curr. Biol. 1–8. https://doi.org/10.1016/j.cub.2021.07.080

Guillera-Arroita, G., Lahoz-Monfort, J.J., MacKenzie, D.I., Wintle, B.A., McCarthy, M.A., 2014. Ignoring imperfect detection in biological surveys is dangerous: A response to "fitting and interpreting occupancy models." PLoS One 9. https://doi.org/10.1371/journal.pone.0099571

Guzman, L.M., Johnson, S.A., Mooers, A.O., Gonigle, L.K.M., 2021. Using historical data to estimate bumble bee occurrence : Variable trends across species provide little support for community-level declines. Biol. Conserv. 257, 109141. https://doi.org/10.1016/j.biocon.2021.109141

Hardisty, A.R., Michener, W.K., Agosti, D., Alonso García, E., Bastin, L., Belbin, L., Bowser, A., Buttigieg, P.L., Canhos, D.A.L., Egloff, W., De Giovanni, R., Figueira, R., Groom, Q., Guralnick, R.P., Hobern, D., Hugo, W., Koureas, D., Ji, L., Los, W., Manuel, J., Manset, D., Poelen, J., Saarenmaa, H., Schigel, D., Uhlir, P.F., Kissling, W.D., 2019. The Bari Manifesto: An interoperability framework for essential biodiversity variables. Ecol. Inform. 49, 22–31. https://doi.org/10.1016/j.ecoinf.2018.11.003

Hickling, R., Roy, D.B., Hill, J., Fox, R., Thomas, C., 2006. The distributions of a wide range of taxonomic groups are expanding polewards. Glob. Chang. Biol. 450–455. https://doi.org/10.1111/j.1365-2486.2006.01116.x

Hill, M.O., 2012. Local frequency as a key to interpreting species occurrence data when recording effort is not known 195–205. https://doi.org/10.1111/j.2041-210X.2011.00146.x

Hollaway, M.J., Dean, G., Blair, G.S., Brown, M., Henrys, P.A., Watkins, J., 2020. Tackling the Challenges of 21st-Century Open Science and Beyond: A Data Science Lab Approach. Patterns 1, 100103. https://doi.org/10.1016/j.patter.2020.100103

Hughes, A., Orr, M., Ma, K., Costello, M., Waller, J., Provoost, P., Zhu, C., Qiao, H., 2020. Sampling biases shape our view of the natural world. Ecography (Cop.). 1–11. https://doi.org/10.1111/ecog.05926

Inman, R., Franklin, J., Esque, T., Nussear, K., 2021. Comparing sample bias correction methods for species distribution modeling using virtual species. Ecosphere 12. https://doi.org/10.1002/ecs2.3422

Isaac, N.J.B., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O.L., Schmucki, R., Simmonds, E.G., O'Hara, R.B., 2020. Data Integration for Large-Scale Models of Species Distributions. Trends Ecol. Evol. 35, 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Isaac, N.J.B., Pocock, M.J.O., 2015. Bias and information in biological records. Biol. J. Linn. Soc. 115, 522–531. https://doi.org/10.1111/bij.12532

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P., Roy, D.B., 2014. Statistics for citizen science: Extracting signals of change from noisy ecological data. Methods Ecol. Evol. 5, 1052–1060. https://doi.org/10.1111/2041-210X.12254

Jetz, W., McGeoch, M.A., Guralnick, R., Ferrier, S., Beck, J., Costello, M.J., Fernandez, M., Geller, G.N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F.E., Pereira, H.M., Regan, E.C., Schmeller, D.S., Turak, E., 2019. Essential biodiversity variables for mapping and monitoring species populations. Nat. Ecol. Evol. 3, 539–551. https://doi.org/10.1038/s41559-019-0826-1

JNCC, 2021a. Indicator D1c. Status of pollinating insects.

JNCC, 2021b. Indicator C4a. Status of priority species - distribution.

Johnston, A., Fink, D., Hochachka, W.M., Kelling, S., 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods Ecol. Evol. 9, 88–97. https://doi.org/10.1111/2041-210X.12838

Jönsson, G.M., Broad, G.R., Umner, S.S., 2021. A century of social wasp occupancy trends from natural history collections : spatiotemporal resolutions have little effect on model performance. https://doi.org/10.1111/icad.12494

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., Donald, P.F., 2016. Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. Divers. Distrib. 22, 1024–1035. https://doi.org/10.1111/ddi.12463

Kéry, M., Royle, J.A., 2016. Applied hierarchical modelling in ecology: analysis of species distribution, abundance and species richness in R and BUGS. Academic press, London.

Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., Guralnick, R.P., Isaac, N.J.B., Kelling, S., Los, W., McRae, L., Mihoub, J.B., Obst, M., Santamaria, M., Skidmore, A.K., Williams, K.J., Agosti, D., Amariles, D., Arvanitidis, C., Bastin, L., De Leo, F., Egloff, W., Elith, J., Hobern, D., Martin, D., Pereira, H.M., Pesole, G., Peterseil, J., Saarenmaa, H., Schigel, D., Schmeller, D.S., Segata, N., Turak, E., Uhlir, P.F., Wee, B., Hardisty, A.R., 2018. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biol. Rev. 93, 600–625. https://doi.org/10.1111/brv.12359

Lewandowski, E., Specht, H., 2015. Influence of volunteer and project characteristics on data quality of biological surveys. Conserv. Biol. 29, 713–723. https://doi.org/10.1111/cobi.12481

MacKenzie, D.I., Bailey, L.L., 2004. Assessing the fit of site-occupancy models. J. Agric. Biol. Environ. Stat. 9, 300–318. https://doi.org/10.1198/108571104X3361

Maes, D., Isaac, N.J.B., Harrower, C.A., Collen, B., van Strien, A.J., Roy, D.B., 2015. The use of opportunistic data for IUCN Red List assessments. Biol. J. Linn. Soc. 115, 690–706. https://doi.org/10.1111/bij.12530

Mukherjee, N., Zabala, A., Huge, J., Nyumba, T.O., Adem Esmail, B., Sutherland, W.J., 2018. Comparison of techniques for eliciting views and judgements in decision-making. Methods Ecol. Evol. 9, 54–63. https://doi.org/10.1111/2041-210X.12940

Outhwaite, C., Chandler, R., Powney, G., Collen, B., Gregory, R.D., Isaac, N.J.B., 2018. Prior specification in Bayesian occupancy modelling improves analysis of species occurrence data. Ecol. Indic. 93, 333–343. https://doi.org/10.1016/j.ecolind.2018.05.010

Outhwaite, C., Powney, G., August, T., Chandler, R., Rorke, S., Pescott, O.L., Harvey, M., Roy, H.E., Fox, R., Roy, D.B., Alexander, K., Ball, S., Bantock, T., Barber, T., Beckmann, B.C., Cook, T., Flanagan, J., Fowles, A., Hammond, P., Harvey, P., Hepper, D., Hubble, D., Kramer, J., Lee, P., MacAdam, C., Morris, R., Norris, A., Palmer, S., Plant, C.W., Simkin, J., Stubbs, A., Sutton, P., Telfer, M., Wallace, I., Isaac, N.J.B., 2019. Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK, 1970-2015. Sci. data 6, 259. https://doi.org/10.1038/s41597-019-0269-1

Outhwaite, C.L., Gregory, R.D., Chandler, R.E., Collen, B., Isaac, N.J.B., 2020. Complex long-term biodiversity change among invertebrates, bryophytes and lichens. Nat. Ecol. Evol.

https://doi.org/10.1038/s41559-020-1111-z

Outhwaite, C.L., Powney, G.D., August, T.A., Chandler, R.E., Rorke, S., Pescott, O., Harvey, M., Roy, H.E., Fox, R., Walker, K., Roy, D.B., Alexander, K., Ball, S., Bantock, T., Barber, T., Beckmann, B.C., Cook, T., Flanagan, J., Fowles, A., Hammond, P., Harvey, P., Hepper, D., Hubble, D., Kramer, J., Lee, P., MacAdam, C., Morris, R., Norris, A., Palmer, S., Plant, C., Simkin, J., Stubbs, A., Sutton, P., Telfer, M., Wallace, I., Isaac, N.J.B., 2019. Annual estimates of occupancy for bryophytes, lichens and invertebrates in the UK (1970-2015). https://doi.org/https://doi.org/10.5285/0ec7e549-57d4-4e2d-b2d3-2199e1578d84

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science (80-.). 339, 277–278. https://doi.org/10.1126/science.1229931

Pescott, O.L., Humphrey, T.A., Stroh, P.A., Walker, K.J., 2019. Temporal changes in distributions and the species atlas: How can British and Irish plant data shoulder the inferential burden? Br. Irish Bot. 1, 250–282. https://doi.org/10.33928/bib.2019.01.250

Peterson, A.T., Soberón, J., 2018. Essential biodiversity variables are not global. Biodivers. Conserv. 27, 1277–1288. https://doi.org/10.1007/s10531-017-1479-5

Plummer, M., 2003. JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling.

Pocock, M.J.O., Logie, M.W., Isaac, N.J.B., Outhwaite, C.L., August, T., 2019. Rapid assessment of the suitability of multi-species citizen science datasets for occupancy trend analysis. bioRxiv 813626. https://doi.org/10.1101/813626

Pocock, M.J.O., Roy, H.E., Preston, C.D., Roy, D.B., 2015. The Biological Records Centre: A pioneer of citizen science. Biol. J. Linn. Soc. 115, 475–493. https://doi.org/10.1111/bij.12548

Powney, G.D., Carvell, C., Edwards, M., Morris, R.K.A., Roy, H.E., Woodcock, B.A., Isaac, N.J.B., 2019. Widespread losses of pollinating insects in Britain. Nat. Commun. 1–6. https://doi.org/10.1038/s41467-019-08974-9

Powney, G.D., Rapacciuolo, G., Preston, C.D., Purvis, A., Roy, D.B., 2014. A phylogenetically-informed trait-based analysis of range change in the vascular plant flora of Britain. Biodivers. Conserv. 23, 171–185. https://doi.org/10.1007/s10531-013-0590-5

Rapacciuolo, G., Young, A., Johnson, R., 2021. Deriving indicators of biodiversity change from unstructured community-contributed data. Oikos 1–15. https://doi.org/10.1111/oik.08215

Roy, H.E., Adriaens, T., Isaac, N.J.B., Kenis, M., Martin, G.S., Brown, P.M.J., Hautier, L., Frost, R., Roy, D.B., Comont, R., Zindel, R., Vlaenderen, J. Van, Lane, B., Gifford, C., 2012. declines of native European ladybirds. Divers. Distrib. 18, 717–725.

Royle, J.A., 2006. Site occupancy models with heterogeneous detection probabilities. Biometrics 62, 97–102. https://doi.org/10.1111/j.1541-0420.2005.00439.x

Royle, J.A., Kéry, M., Gautier, R., Schmid, H., 2007. Hierarchical spatial models of abundance and occurrence from imperfect survey data. Ecol. Monogr. 77, 465–481. https://doi.org/10.1890/06-0912.1

Royle, J.A., Nichols, J.D., 2003. Estimating abundance from repeated presence-absence data or point counts. Ecology 84, 777–790. https://doi.org/10.1890/0012-9658(2003)084[0777:EAFRPA]2.0.CO;2

Ruiz-Gutierrez, V., Zipkin, E.F., Dhont, A., 2010. Occupancy dynamics in a tropical bird community : unexpectedly high forest use by birds classified as non-forest species. J. Appl. Ecol. 47, 621–630. https://doi.org/10.1111/j.1365-2664.2010.01811.x

Salama, M., Blair, G., Brown, M., Holloway, M., 2022. Virtual Labs for Collaborative Environmental Data Science (No. EGU22-10719)., in: Copernicus Meetings.

Schmeller, D.S., Mihoub, J.B., Bowser, A., Arvanitidis, C., Costello, M.J., Fernandez, M., Geller, G.N., Hobern, D., Kissling, W.D., Regan, E., Saarenmaa, H., Turak, E., Isaac, N.J.B., 2017. An operational definition of essential biodiversity variables. Biodivers. Conserv. 26, 2967–2972. https://doi.org/10.1007/s10531-017-1386-9

Soldaat, L.L., Pannekoek, J., Verweij, R.J.T., van Turnhout, C.A.M., van Strien, A.J., 2017. A Monte Carlo method to account for sampling error in multi-species indicators. Ecol. Indic. 81, 340–347. https://doi.org/10.1016/j.ecolind.2017.05.033

Steen, V.A., Tingley, M.W., Paton, P., Elphick, C., 2020. Spatial thinning and class balancing : Key choices lead to variation in the performance of species distribution models with citizen science data. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13525

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J.W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W.M., Iliff, M.J., Lagoze, C., La Sorte, F.A., Merrifield, M., Morris, W., Phillips, T.B., Reynolds, M., Rodewald, A.D., Rosenberg, K. V., Trautmann, N.M., Wiggins, A., Winkler, D.W., Wong, W.K., Wood, C.L., Yu, J., Kelling, S., 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biol. Conserv. 169, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Szabo, J.U.K.S., Vesk, P.E.A. V, Baxter, P.E.W.J.B., Possingham, H., 2010. Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. Ecol. Appl. 20, 2157–2169.

Van Strien, A.J., Van Swaay, C.A.M., Termaat, T., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. J. Appl. Ecol. 50, 1450–1458. https://doi.org/10.1111/1365-2664.12158

van Strien, A.J., van Swaay, C.A.M., van Strien-van Liempt, W.T.F.H., Poot, M.J.M., WallisDeVries, M.F., 2019. Over a century of data reveal more than 80% decline in butterflies in the Netherlands. Biol. Conserv. 234, 116–122. https://doi.org/10.1016/j.biocon.2019.03.023

Velásquez-Tibatá, J., Olaya-Rodríguez, M., López-Lozano, D., Gutiérrez, C., González, I., Londoño-Murcia, y M.C., 2018. BioModelos: a collaborative online system to map species distributions. bioRxiv 432617. https://doi.org/10.1101/432617

Vihervaara, P., Auvinen, A.P., Mononen, L., Törmä, M., Ahlroth, P., Anttila, S., Böttcher, K., Forsius, M., Heino, J., Heliölä, J., Koskelainen, M., Kuussaari, M., Meissner, K., Ojala, O., Tuominen, S., Viitasalo, M., Virkkala, R., 2017. How Essential Biodiversity Variables and remote sensing can help national biodiversity monitoring. Glob. Ecol. Conserv. 10, 43–59. https://doi.org/10.1016/j.gecco.2017.01.007

Warton, D.I., Stoklosa, J., Guillera-Arroita, G., MacKenzie, D.I., Welsh, A.H., 2017. Graphical diagnostics for occupancy models with imperfect detection. Methods Ecol. Evol. 8, 408–419. https://doi.org/10.1111/2041-210X.12761

Welsh, A.H., Lindenmayer, D.B., Donnelly, C.F., 2013. Fitting and Interpreting Occupancy Models. PLoS One 8. https://doi.org/10.1371/journal.pone.0052015

Wilkinson, M.D., 2016. Comment : The FAIR Guiding Principles for scientific data management and stewardship 1–9. https://doi.org/10.1038/sdata.2016.18

Wright, W.J., Irvine, K.M., Higgs, M.D., 2019. Identifying occupancy model inadequacies: can residuals separately assess detection and presence? Ecology 100, 1–9. https://doi.org/10.1002/ecy.2703

Zizka, A., Antonelli, A., Silvestro, D., 2021. Sampbias, a Method for Quantifying Geographic Sampling Biases in Species Distribution Data. Ecography (Cop.). 44, 25–32. https://doi.org/10.1111/ecog.05102

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., Antonelli, A., 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10, 744–751. https://doi.org/10.1111/2041-210X.13152