RESEARCH ARTICLE

Frontiers of Biogeography
the scientific journal of
the International Biogeography Society

What have biological records ever done for us? A systematic scoping review

Willson Gaul[1]*, Dinara Sadykova[2], Ellie Roark[3], Hannah J. White[1], Lupe León-Sánchez[3], Paul Caplat[3], and Jon M. Yearsley[1]

[1] School of Biology and Environmental Sciences, Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland; [2] School of Biological Sciences, Queen's University Belfast, BT9 5DL Belfast, UK; [3] Dublin, Ireland; * Corresponding author: Willson Gaul: willson.gaul@ucdconnect.ie

Abstract

Biological records provide biodiversity information over large spatial and temporal scales. Our systematic scoping review of biological records from the well-recorded region of the United Kingdom (UK) and Ireland revealed that over half of all studies using biological records were studying species distributions (134 of 253 studies) and/or temporal trends (139 of 253 studies). A minority of studies (61 of 253) focused on methodological questions, while most studies used biological records with existing methods as tools for answering biological and ecological questions. However, only 31 of 253 studies tested models using independent data. Most studies (154 of 253) integrated multiple biological records datasets, showing that biological records hold a largely untapped potential for independently testing conclusions by withholding some of those datasets for use as independent test data. Our results provide guidance for data providers and researchers interested in more effectively collecting and using biological records.

Highlights

- Our scoping review showed that most studies using British and Irish species occurrence records focused on ecological or biological questions, rather than developing or testing methodology.
- We found that different types of biological records data were used to study different questions, and it follows that data providers should identify priority uses in order to provide appropriate data.
- More structured biological records data were generally analyzed with more powerful methods (e.g., statistical inference).
- Although most studies used multiple biological records datasets, few studies attempted to validate results using independent data.
- We suggest that future studies using multiple biological records datasets withhold at least one of those datasets for use as independent validation data.

Keywords: biodiversity, biological records, citizen science, GBIF, Ireland, opportunistic data, scoping review, United Kingdom

Introduction

Biological records are "what, where, when" observations from citizen science or other sources that record the presence of a species in a particular place and at a particular time (Isaac and Pocock 2015). Additional information may be associated with the basic "what, where, when" data: the name of the person who collected the data, the survey methods, the duration or spatial extent of the survey during which the records, photos, or sound recordings were collected. National and international data centers, including the Global Biodiversity Information Facility (GBIF), aggregate and disseminate biological records data, with the mission of making biodiversity data available for the scientific community, governments, and non-governmental organizations (National Biodiversity Network 2015).

We are interested in questions such as the following. To what extent are efforts to share biological records data successful? Are the data widely used by researchers not affiliated with the groups that collect and provide the data? What questions are studied using biological records? Some data providers are explicitly interested in providing data that will be used in applied settings.

e-ISSN: 1948-6596 https://escholarship.org/uc/fb doi:10.21425/F5FBG48607

For example, informing decision-making is a "key objective" of the Irish National Biodiversity Data Centre (NBDC): "The [NBDC] facilitates and promotes the use of biodiversity data to inform public policy and decision-making through analysis, interpretation and reporting" (National Biodiversity Data Centre n.d.). In order to provide data that meet the needs of users, data providers need to know how biological records are likely to be used and what characteristics of the data enable or inhibit studying particular questions or using particular techniques of analysis. For example, governments may wish (or be required) to report on multiple different biodiversity questions, including concerning species distributions, population sizes, and temporal trends for individual species (Department of Arts Heritage and the Gaeltacht 2017). But data needs are not necessarily the same for all study questions. Studying temporal trends in species population abundance may require different types of data than studying species distributions. Knowing which types of data are commonly used to study different types of questions will be useful to data providers who wish to facilitate studies of specific types of questions.

Similarly, researchers downloading and using biological records data, especially those using biological records for the first time, can gain insight into how to conduct their own studies by knowing which types of data and which methods of analysis are used or avoided by the community of researchers working with biological records. Identifying biological and ecological study questions that are commonly studied with biological records but for which methodological development is not particularly active can give a "sense of the field" about areas where methods for analyzing biological records are largely settled. Similarly, identifying particular study questions that are rarely addressed with certain types of data could highlight fundamental underlying challenges – challenges either to be avoided, or perhaps challenges to be tackled with methodological development.

Analyses using statistical inference to estimate relationships between variables, along with associated uncertainty (e.g., confidence intervals or *p*-values), are arguably the most powerful way to answer questions using data. But statistical inference is not the only useful way of studying questions. Prediction can be a powerful and useful tool, and some research focuses on high-quality prediction of response variables, without explicitly attempting to interpret the effects or statistical significance of predictor variables (Fink et al. 2010). Purely descriptive studies of observations or descriptive statistics are more limited, but can still be useful, for example, in assessments of invasive species (Millane and Caffrey 2014). However, we assume that a researcher who has adequate data to do prediction or statistical inference would not present only descriptive results. We therefore tested whether some types of biological records were more frequently used than others in purely descriptive analyses, which would indicate that characteristics of those data limit the strength of conclusions. If some types of biological records do limit analyses to being purely descriptive, why would researchers still use those data? To explore this, we investigated potential trade-offs by looking at whether some types of data were associated with studies that covered longer temporal or spatial extents.

Identifying new or unusual uses of biological records can help data providers anticipate and prepare for new data needs and opportunities for the future. Digital innovations have transformed how citizen scientists submit vouchers of specimens with biological records (August et al. 2015). Biological records now include everything from literature records, to opportunistic observations submitted via mobile phone apps (Sullivan et al. 2009), to abundance counts from standardized transects in citizen science monitoring schemes (Van Swaay et al. 2008). Biological records can be generated by remote machine-based observations (e.g., camera traps or automated acoustic recorders), sometimes with the aid of citizen scientists to process the large amounts of data generated (Swanson et al. 2015). There is even potential for citizen scientists to use eDNA to produce biological records, which may be especially useful for taxa that are difficult to survey otherwise (Biggs et al. 2015). Biological records data may include photographs or audio recordings (Vellinga and Planqué 2015), information about survey effort, or nothing more than "what, where, when" observations. How are researchers making use of these new types of biological records data?

The United Kingdom (UK) and Ireland have some of the most intensive biodiversity recording schemes in the world (Meyer et al. 2016), and therefore offer an excellent opportunity to systematically study the use of "what, where, when" data. National biodiversity data centers in Ireland and the UK have well-developed infrastructure for collating and disseminating data[1]. However, despite relatively intense recording effort, the biases and gaps common to all biological records are present in UK and Irish records (Isaac and Pocock 2015). The uses of biological records from the UK and Ireland therefore provide insight about the current "best case" scenario in terms of biological recording; to the extent that these records are used to study biological and ecological questions with powerful methods (i.e., prediction and statistical inference), the structure of these records can serve as a goal for data collectors and aggregators in other regions. On the other hand, identifying characteristics of these records that limit their use can inform recorders and data aggregators elsewhere (and in the UK and Ireland) about what is needed to make data more useful in the future.

Methods

Literature search, review process, and data reliability

We undertook a systematic scoping review (Arksey and O'Malley 2005) of original research published since 2014 that used biological records from Ireland and/or the UK. We limited the review to studies

[1] www.biodiversityireland.ie, last accessed 02/06/2020; https://nbnatlas.org/, last accessed 02/06/2020

published since 2014 in order to focus on the most recent literature and to keep the number of reviewed studies below 300, which we estimated would be a manageable number given the time-intensive review process. We generated a list of potentially relevant studies by searching Web of Science, Scopus, ProQuest, the GBIF website, and GoogleScholar (using Harzing 2007; see Appendix S1 for search terms). One researcher evaluated each study for inclusion eligibility according to criteria described in Gaul, Roark, and Yearsley (2020). For eligible studies, one researcher coded variables describing basic descriptive information (e.g., temporal and spatial extent), study questions, data type, and analytical approach (Table 1). To assess the reliability of codings, a second researcher (hereafter "coder") coded the variables for a subset of 20% of the eligible studies (40 studies). Agreement between the two coders was evaluated using Krippendorff's *alpha,* which measures agreement while accounting for agreement by chance (Gamer et al. 2012, Krippendorff 2013). We only conducted statistical analyses for variables with Krippendorff's *alpha* values above 0.67 and with relative frequencies above 0.14 for the least common category of the variable, following rule-of-thumb guidelines from Krippendorff (2013;Table S1). For variables with insufficient variation to estimate coding reliability (variables with relative frequencies below 0.14 for the least common category) we did not perform statistical analyses, but we do comment on them because the rare categories of these variables are interesting from a horizon-scanning perspective. A brief description of the variables that were reliably coded according to Krippendorff's *alpha* and that we used in analyses is given in Table 1. Variables that

Table 1. Descriptions of variables that were reliably coded and used in subsequent analyses. For clarity of presentation, some variables are grouped by theme (indicated in italic font), and we refer to those variables in the text using the theme.

Theme or Variable	Description
Study Question	This describes the ecological or biological focus of the study. We categorized study questions as being about: *abundance*, *species distribution*, *phenology*, *species richness*, *temporal trends*, or "*other*".
Data Type	This describes different types of biological records data. We categorized each study based on whether the biological records data: came from an *organized monitoring scheme*; included explicit *non-detection* information; included explicit *sampling effort* information; included *visit-specific covariates* (e.g., wind speed at the time of the survey); or whether the study included *multiple biological records datasets.* We also evaluated whether studies used biological records that included *photographs, audio* or *video recordings, physical voucher specimens,* or *life stage* information.
Analytical Approach	This describes the broad analytical approach used for biological records. We categorized each study as using one of the following three analytical approaches: *statistical inference, prediction,* or *descriptive only. Statistical inference* indicates analyses that included the estimation of uncertainty or significance (e.g., *p*-values or confidence intervals). *Prediction* indicates analyses that use predicted values of a response variable with no associated uncertainty estimates (e.g., model-predicted probability of a species occurring at a location). *Descriptive only* indicates analyses that used neither statistical inference nor prediction, but presented results descriptively through, e.g., narrative descriptions of patterns, maps of observed values, or point estimates of descriptive statistics.
Testing procedure	Indicates whether the study tested their analyses using *cross-validation*, a "holdout" *subset* of the data, or *independent data.*
Author associated with data provider	Indicates whether any of the authors were affiliated with the institution that provided the data
Temporal extent	The time period (in years) covered by the study.
Methodology development or analysis	Indicates whether a major focus of the study was development of a new analysis method or testing how well a method works.
Proximate data source	The source from which the biological records data were obtained by the authors of each study. Note this is not necessarily the source that generated the data (e.g., if a study downloaded data from GBIF, but the data were produced by a non-profit citizen science organization, the proximate data source would be GBIF).

were not reliably coded, or that were too rare for us to evaluate their reliability, are listed in Table S1 and described fully in Gaul, Roark, and Yearsley (2020). Analyses were conducted in R version 3.5 (R Core Team 2018). Data and review protocols are provided in Gaul, Roark and Yearsley (2020). Scripts to conduct all analyses are available in Gaul (2020) or from GitHub[2]. A list of reviewed studies is in Appendix S2.

Who used biological records?

In order to determine how often studies using biological records were conducted by researchers who were not affiliated with the data provider, we coded whether each study had at least one author who was affiliated with the data-providing institution.

Data sources

To investigate how studies acquired biological records data, we counted the number of studies that got data from each proximate data source.

Temporal extent of studies

We tested whether the temporal extent of studies differed based on *data type* by fitting a linear model with the natural-log transformed temporal extent of studies (in years) as the response. We summarized the *data type* variables using multiple correspondence analysis (MCA;Mair and de Leeuw 2019) to reduce dimensionality and produce uncorrelated predictor variables, because three of our *data type* variables indicating structured data (sampling effort, non-detection, and data from an organized monitoring scheme) were correlated. As predictors in the linear model, we used the first two MCA dimensions (accounting for 85.9% of the variance in the MCA, Fig. S1), indicating whether studies used more structured biological records *data types* (dimension one, eigenvalue = 2.44) and whether studies used multiple biological records datasets (dimension two, eigenvalue = 0.99). We assessed the significance of *data types* for predicting temporal extent by using a likelihood ratio test to compare a model containing the two MCA dimensions to an intercept-only model.

Ecological and biological study questions

We assessed which ecological or biological *study questions* were addressed using biological records. A study could address more than one *study question* using biological records. We tested for differences in the number of studies investigating each *study question* by looking at whether there was overlap of 95% bias-corrected accelerated bootstrap confidence intervals around the mean number of studies investigating each *study question*.

Development of methodology across different study questions

We expected to find more methodological development studies for some types of ecological and biological questions than for others. To test for differences in the proportion of methodological studies across different study questions, we used logistic regression with a binary response variable indicating whether the study addressed a methodological question and six predictor variables (listed in Table 2) indicating the ecological or biological *study questions*. We tested whether the full model including all six predictor variables was better than an intercept-only null model using a likelihood ratio test. After performing the likelihood ratio test comparing the null and full models, we performed exploratory variable selection using AIC. We tested the prediction performance of the full, null, and AIC top-ranked models using McFadden's pseudo-R^2 (Domencich and McFadden 1975) and the area under the receiver operating characteristic curve (AUC).

What data types are used for each study question?

We asked whether different biological records *data types* were used to study different ecological or biological *study questions*. We used a random permutation procedure (Manly 2007) that broke the association between *data type* and *study questions* to

Table 2. The relationship between ecological or biological *study question* and whether or not studies developed or tested methodology. *Study questions* are listed, along with the number of studies that were and were not methodological. Coefficient estimates (with 95% confidence intervals), *z*-statistics, and *p*-values for each term are from a logistic regression testing whether the probability of a study developing or testing methodology depended on the ecological or biological *study question*.

Study question	**Methodological**	**Not methodological**	**Coefficient estimate (95% CI)**	**z**	***P*-value**
(intercept)	na	na	-1.12 (-2.02, -0.25)	-2.51	0.01
abundance	19	41	0.36 (-0.40, 1.09)	0.95	0.34
species distribution	25	109	-0.52 (-1.26, 0.22)	-1.38	0.17
phenology	7	29	-0.47 (-1.49, 0.43)	-0.98	0.33
species richness	8	34	-0.35 (-1.32, 0.52)	-0.76	0.45
temporal trends	36	103	0.33 (-0.36, 1.04)	0.92	0.36
other	9	15	0.61 (-0.61, 1.82)	0.99	0.32

[2] https://github.com/wgaul/systematic_review, last access on 06/02/2020

test the null hypothesis of no difference in the types of data used to study different biological or ecological questions. The permutation procedure and associated residual and test statistic calculations are described in Figure S2.

For variables indicating rare *data types* (including photo, audio, and video data) we could not estimate the reliability of our coding of the variables (Table S1). We therefore did not include those rare data types in the formal permutation analysis. Instead, we reported the percentage of studies that used each rare data type and discussed the implications of the rarity.

Analytical approach and data type

To investigate whether different biological records *data types* were used with different *analytical approaches*, we used random permutations to test a null hypothesis of no difference in the *data types* used with different *analytical approaches*. We classified the *analytical approach* of each study into one of three categories: “statistical inference”, “prediction”, or “descriptive”. We defined “statistical inference” narrowly as statistical inference that included estimation of uncertainty or confidence (including but not limited to hypothesis testing). We defined prediction as predicting point estimates (e.g., from a model), but without including estimates of uncertainty or confidence with the predictions. For example, a map of the predicted probability of a species occurring in each grid square of a raster would be categorized as “prediction” if there was no measure of the uncertainty associated with the predicted probabilities. Finally, we defined descriptive analyses as those that did not include any estimation of uncertainty or prediction to new data or parts of data space (e.g., maps of observations or point estimates of summary statistics are descriptive only). We chose to differentiate between prediction and statistical inference because biological records data often contain a wide variety of sampling biases, including spatial and temporal biases, which make many observations non-independent. Non-independent data, such as when observations are spatially auto-correlated, are particularly problematic when estimating coefficients and their associated uncertainty (Beale et al. 2010, and references therein), in part because it can be unclear what the effective sample size is (Lennon 2000). Model predictions (i.e., point estimates of the response variable) are less affected by spatial auto-correlation (Thibaud et al. 2014, but see Guélat and Kéry 2018), and therefore we expected that researchers might chose to do prediction without estimating uncertainty or confidence and without interpreting coefficient estimates when the data were particularly “messy”.

The permutation procedure we used to investigate whether different biological records *data types* were used with different *analytical approaches* was similar to that used for the analysis of the relationship between *data type* and *study questions*, but using variables describing *analytical approach* instead of variables describing *study questions* (Fig. S2).

Testing on independent data

We estimated the proportion of studies that tested models using cross-validation (in which data are split into sets used for either model training or testing) and the proportion that tested on independent data.

Results

Literature search, review process, and data reliability

The search returned 2,695 potentially relevant studies, of which 253 (9.4%) were eligible for inclusion in this review. Of the studies that we deemed eligible for inclusion in this review, 53 (20.9%) were returned more than once by the search. Sixteen variables met our criteria for inclusion in further analyses (Krippendorff’s *alpha* greater than 0.67 and relative frequency of at least 0.14 for the least common category, Table S1). Five variables had relative frequencies of the least common category high enough to estimate *alpha*, but they had *alpha* values less than or equal to 0.67 and were therefore deemed unreliable.

Who used biological records?

Coder agreement was low when determining whether a study had an author associated with the proximate data provider (Table S1), so we did not further analyze the association between authors and data providers.

Data sources

Most studies used traditional data sources such as biodiversity data centers, taxon-specific monitoring schemes, and natural history museums (Table S2). There were many data sources that were only used by one study, including some non-traditional data sources, such as the media sharing platforms Flickr and YouTube.

Temporal extent of studies

Studies using multiple biological records datasets covered longer temporal extents than studies that did not (Fig. 1, overall significance of *data type* on temporal extent $p < 0.0001$, $F_{2, 198} = 41.32$, Adjusted $R^2 = 0.29$, regression coefficient estimate for MCA dimension two = 0.71, p < 0.0001, $F_{1, 199} = 46.38$). Studies using more structured biological records *data types* covered shorter temporal extents than studies that used less structured *data types* (regression coefficient estimate for MCA dimension one = -0.63, p < 0.0001, $F_{1, 199} = 36.26$). We removed from analysis 52 studies for which we could not determine the temporal extent, leaving 201 studies for analysis.

Ecological and biological study questions

Over half the studies in this review analyzed species distribution (proportion of studies = 0.53, 95% bootstrap CI [0.46, 0.58]) and/or temporal trends (proportion = 0.55, 95% bootstrap CI [0.48, 0.61]), which was higher than the proportion analyzing any other *study question* (Fig. 2). We found no difference in the proportion of studies analyzing abundance (proportion = 0.24, 95% bootstrap CI [0.18, 0.29]),

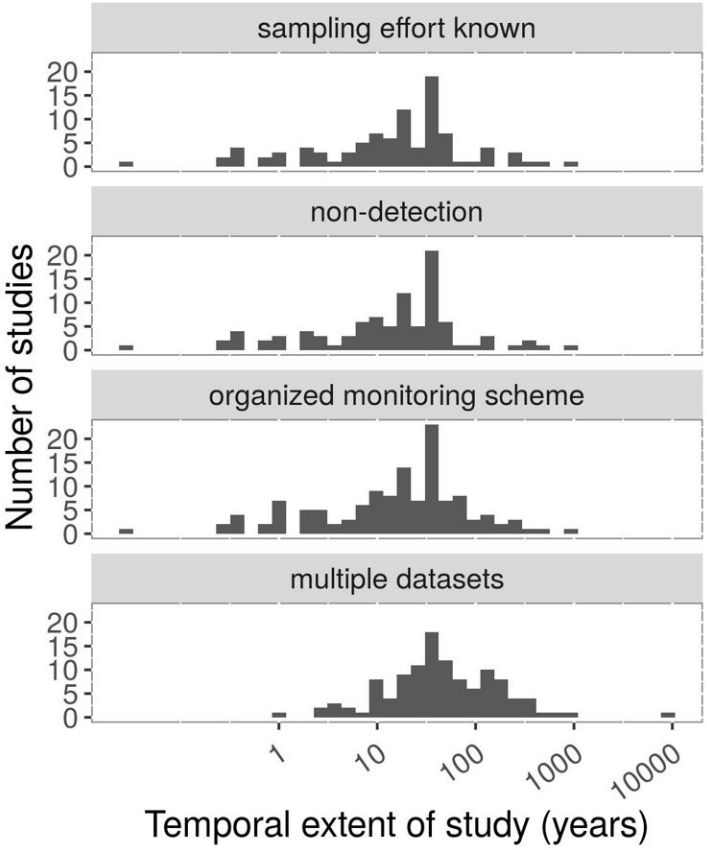

Fig. 1. The distribution of temporal extents (in years) covered by studies using four different biological records *data types* (panels). Studies using multiple biological records datasets covered longer temporal extents. Note the logarithmic scale of the horizontal axis.

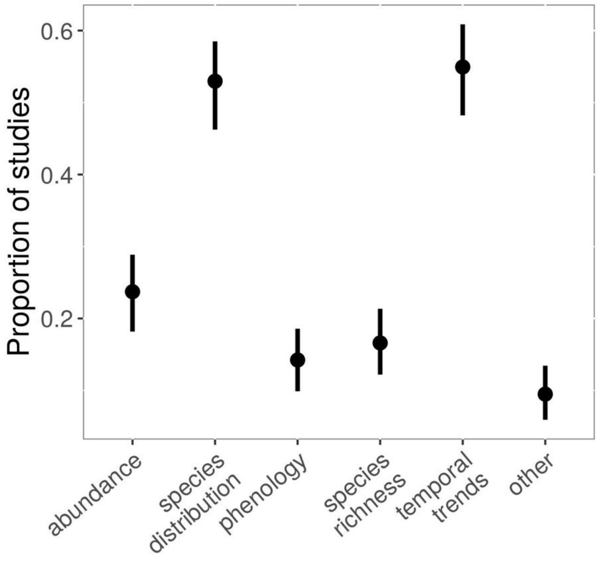

Fig. 2. The proportion of studies that analyzed different ecological or biological *study questions*. Studies of species distributions and temporal trends were the most common uses of biological records. Points show proportion of studies, vertical lines show 95% bootstrap confidence intervals. Proportions in this plot add up to more than one because some studies performed multiple analyses and studied multiple questions.

species richness (proportion = 0.17, 95% bootstrap CI [0.12, 0.21]), or phenology (proportion = 0.14, 95% bootstrap CI [0.1, 0.19]). The proportion of studies focused on species diversity (proportion = 0.06) or alien species (proportion = 0.11) were too low for us to estimate coder agreement using Krippendorff's *alpha*.

Development of methodology across different study questions

Studies in this review primarily used biological records to study ecological and biological questions rather than developing or testing methodology for analyzing biological records (proportion of studies that developed or tested methodology = 0.24, 95% bootstrap CI [0.19, 0.29]). Despite our expectations, there was no evidence that the probability of a study developing or testing methodology depended on the ecological or biological *study question* (χ^2_6 = 9.66, *p* = 0.14, Table 2). Our models, including the model ranked highest according to AIC, had low prediction performance (null model AUC = 0.5; full model AUC = 0.62, McFadden's pseudo-R^2 = 0.034; top-ranked model AUC = 0.58, McFadden's pseudo-R^2 = 0.017).

What data types are used for each study question?

Data type and *study question* were not independent (χ^2 = 29.1, df = 6, p <= 0.0001, 10^4 permutations, Fig. 3). This non-independence is primarily explained by abundance studies and species distribution studies (Fig. 3). Studies of species distributions used multiple biological records datasets more than studies addressing other *study questions* (Fig. 3). Structured data (e.g., data from organized monitoring schemes or with sampling effort or non-detection information) were used more frequently for abundance studies and less frequently for studies of species distributions than would be expected if there were no relationship between *data type* and *study question* (Fig. 3). Multiple datasets were used less often than expected for abundance studies (Fig. 3). There was no evidence that particular *data types* were associated with phenology, species richness, or "other" *study questions*.

Analysis of voucher specimens, either physical or digital, was rare: 5.1% of studies used photos, 0.4% of studies used videos, 2.8% of studies used audio recordings, 9.5% of studies used life stage (e.g., phenology) information, and 9.1% of studies used physical voucher specimens.

Analytical approach and data type

There was strong evidence that *data type* was associated with *analytical approach* ($\chi^2 = 21.6$, df = 6, $p < 0.0001$, 10^4 permutations, Fig. 4). This was primarily explained by studies using statistical inference and studies using only descriptive results (Fig. 4). Structured data (e.g., data from organized monitoring schemes or with sampling effort or non-detection information) were analyzed with statistical inference more frequently than expected by chance, and they were analyzed descriptively less frequently (Fig. 4). Multiple biological records datasets were used less often than expected for statistical inference and more often than expected for descriptive-only analyses.

Testing on independent data

The proportion of studies that tested models using cross-validation was only 0.07 (18 of 253 studies), and the proportion that tested on independent data was 0.12 (31 of 253 studies). Those proportions were too low to estimate the reliability of our coding using Krippendorff's *alpha*, but if our coding is accurate, it would mean that a large majority of studies did not validate their results on independent data.

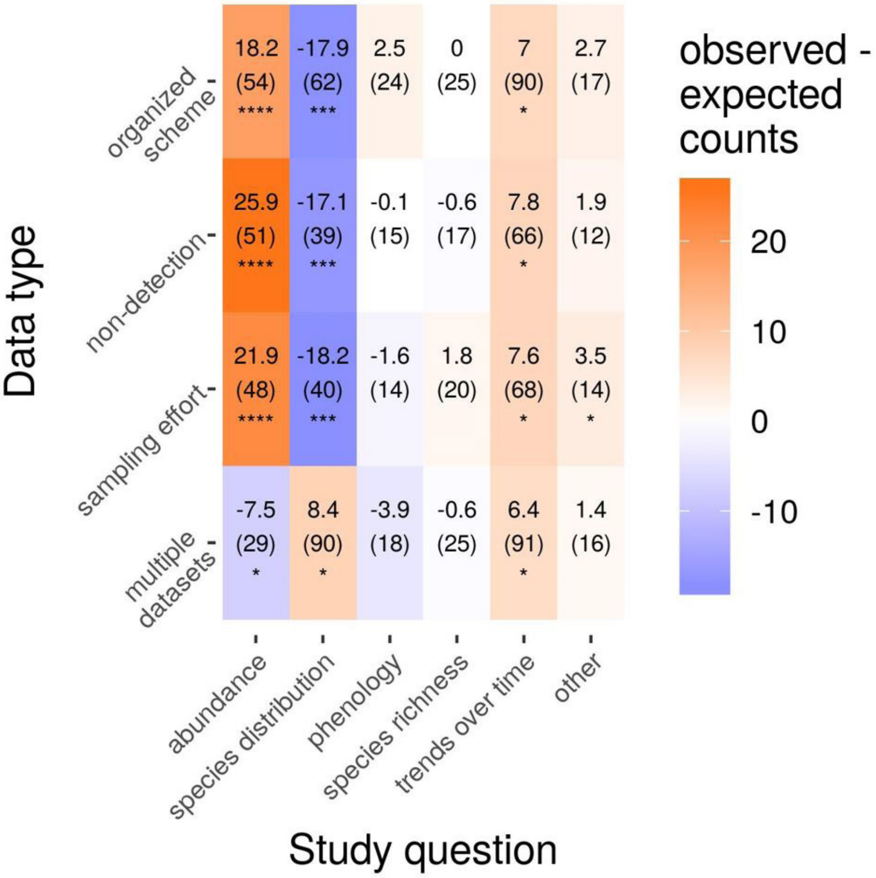

Fig. 3. Residuals from a permutation test showed that different biological records *data types* were used to address different *study questions*. Numbers printed in each grid cell are residuals (observed minus expected counts, outside parentheses) and observed counts of studies (inside parentheses). Asterisks indicate significance levels (* p < 0.1, ** p < 0.01, *** p < 0.001, **** p < 0.0001). Big absolute values of residuals and small permutation significance levels provide evidence for an association between *data type* and *Study Question*.

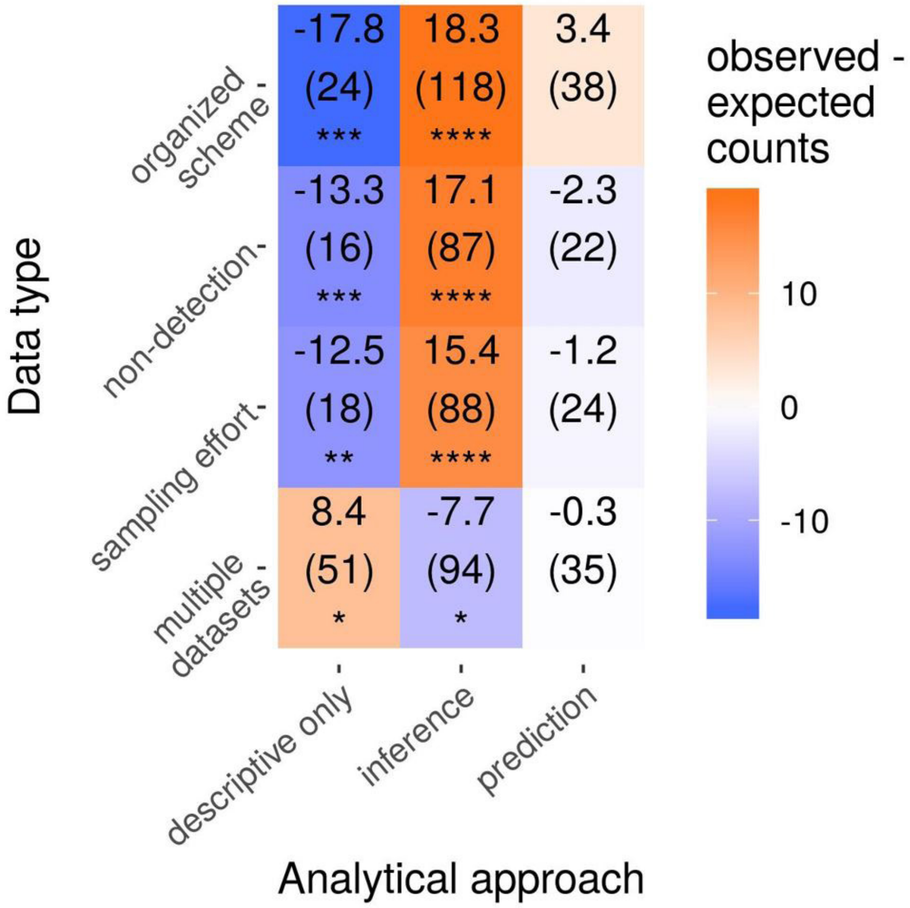

Fig. 4. More structured *data types* (non-detection, sampling effort, and data from organized monitoring schemes) were analysed with statistical inference more frequently and with purely descriptive methods less frequently than would be expected by chance. Studies using multiple different biological records datasets were more likely to use purely descriptive analyses than would be expected by chance. Residuals, observed counts, and asterisks indicating significance level are as for Fig. 3.

Discussion

Completeness and representativeness of the search

The low proportion of eligible studies that were found by more than one of our search methods indicated that none of our search methods searched the entire literature. Even combining multiple search methods, we did not approach a complete search. Our search produced a sample of convenience from the "population" of all literature. Our results are not as generalizable as they would be if we had a true random sample from the literature. However, we do not know of any way to achieve a truly random sample from the literature. Some reviews try to achieve a complete census (i.e., find all relevant studies), by combining database searches such as ours with methods including "snowballing", which searches the references of already-discovered studies (Pham et al. 2014), hand searching important journals, and soliciting input from disciplinary experts (Arksey and O'Malley 2005). Given the widespread use of biological records, it may not be possible to achieve a complete census of studies using UK and Irish biological records, even using methods such as snowballing. Nevertheless, our review provides a useful snapshot of how researchers use biological records. A review of a more limited question, for example, of studies that used biological records to assess the outcomes of habitat restorations, could come closer to achieving a comprehensive census.

We included grey literature in our review because we expected that biological records would be used in the grey literature with direct relevance for society. The GBIF website and GoogleScholar were our main sources for finding grey literature, though they are known to not identify all relevant grey literature (Haddaway et al. 2015). Our search failed to find some obviously important grey literature, including the UK State of Nature 2016 report (Hayhow et al. 2016), and there are almost certainly other important studies that our search missed, from both the grey and peer-reviewed literature. Similarly, our analyses treated each eligible study equally, without taking into consideration the quality of the studies, including sample size or the rigour of analytical methods used in the studies (Lortie 2014).

Reliability of codings

Measuring the agreement between two coders using Krippendorff's *alpha* allowed us to assess the reliability of our data. Much ecological and biological research treats explanatory or predictor data as having been measured without error (e.g., in regression analyses;Austin 2007). Reviews such as ours are challenging because the data are generated by human coders following guidelines that are almost never complete enough to cover all possible cases. Some reviews can answer their questions by searching texts for specific words (e.g., McCallen et al. 2019), which produces unambiguously "correct" data. But such an approach is not applicable when assessing concepts that are not easily encapsulated by particular words. Coding many of our variables required some degree of judgment by the human coders, inherently introducing the possibility for noise or variation in the measurement of the data. Added to this is the ever-present possibility of error in data collection (e.g., a coder might mis-type the data into the spreadsheet). Krippendorff's *alpha* assessed the agreement between two coders, and therefore served as a quantitative measure of how reliable the data are: when coders agreed more often than expected by chance, it indicated that the data were showing a reliable signal, and another researcher could expect to get similar results when following our protocol. For most of our variables, agreement between the coders was good but less than perfect. For example, coders were in good but not perfect agreement about whether the *analytical approach* produced only descriptive results (Krippendorff's *alpha* = 0.8, Table S1). The imperfect agreement between coders is not a reason to mistrust the data. We used a binary decision threshold to determine whether or not a variable was reliable enough to use in analyses (see Methods). We did not subsequently propagate the estimated error in the predictor data through our models to influence our measures of uncertainty or significance, but nevertheless our binary decision threshold based on the measured data reliability is an important step for reviews such as ours in which the data are generated by an inherently subjective coding process. We note briefly here that it is unwise to "fix" disagreements between coders before conducting subsequent analyses, because there is then no way to assess the reliability of the new "fixed" data without repeating the coding process (Krippendorff 2013).

Determining the spatial extent of studies was surprisingly difficult. Only a minority of studies clearly reported the spatial extent in quantitative terms (e.g., km^2). More commonly, studies included a statement such as "we collected... records from different locations within the UK..." (Hart, Nesbit and Goodenough 2018). The spatial extent in these cases is less clear – it might be reasonable to assume that the spatial extent of Hart, Nesbit and Goodenough (2018) is the spatial area of the UK, because the study question seemed to encompass that extent. But it is unlikely that the data covered that entire extent. (We do not mean to single out Hart, Nesbit and Goodenouh (2018) in this respect – we simply use that study to demonstrate a common way in which spatial extent is ambiguously reported). Even more difficult were cases in which the study question was not necessarily limited to any particular spatial area, but depended on the data available. For example, Leighton, Hugo, Roulin and Amar (2016) searched the internet for photographs of focal taxa and used the images to assess geographic patterns of colour morphs. One of their focal taxa, the barn owl (*Tyto alba*), is widely distributed across the globe. It was not clear to us whether we should have considered the spatial extent of the study to be the potential global distribution of barn owls (which was arguably the spatial extent being considered at the start of the data search), the area of a convex polygon encompassing the locations of the 347 data points they ultimately used in their analysis (which we would have had to calculate), or some other value. Future reviews that wish to evaluate the spatial extent of studies will need a more detailed and explicit protocol than ours for determining the spatial extent of studies.

More structured data types allowed stronger analytical approaches

Using statistical inference generally provides more confidence in conclusions than using purely descriptive analyses. Our results showed that combining data from multiple biological records datasets reduced the frequency with which studies performed statistical inference (Fig. 4). So why would researchers use multiple different datasets in a study? One possible reason is that studies using multiple different datasets covered longer temporal extents than single-dataset studies (Fig. 1). There was, therefore, a trade-off between using data from multiple sources to cover longer time spans and the inferential rigour of the analyses.

The greater use of statistical inference with structured compared to unstructured biological records *data types* underscores the importance of routinely recording non-detection data and survey effort data with biological records (Sullivan et al. 2009, Tingley and Beissinger 2009, Isaac and Pocock 2015). New biological recording schemes should collect non-detection and survey effort data, and existing schemes should consider modifying data submission procedures to record non-detection information. Input from social scientists and experts in human/technology interaction may be helpful to recordings schemes attempting to collect more complex data (e.g., survey effort information) while still attracting and retaining volunteer recorders.

Tailoring biological records data to different study questions

Researchers' use of different *data types* for studying different ecological or biological *study questions* is an important consideration for data collectors and aggregators. For example, a biodiversity data center that aims to provide evidence for evaluating species population abundance will need to focus on collating structured data (e.g., data with survey effort and non-detection information, and data from organized monitoring schemes). On the other hand, a data center that wants to provide information about spatial patterns of species distribution, perhaps for identification of priority conservation areas, may wish to focus on collating a wide range of different datasets and presenting them in a common, standardized format, even if the data lack non-detection or survey effort information.

Studies of species distributions and temporal trends were more likely to use multiple biological records datasets, perhaps because researchers felt that multiple datasets added important information to distribution studies. Or it may be that integrating data from multiple sources into a single model is easier or has been better studied for distribution models (e.g., Pacifici et al. 2017). However, all types of biological or ecological *study questions* were studied by at least a few analyses that integrated multiple separate biological records datasets, highlighting the importance of standardized formats for sharing biological records data (Wieczorek et al. 2012).

Most studies did not develop methodology... but also did not validate results

The fact that less than a third of studies developed or tested methodology suggests that most studies in this review used biological records as a tool for answering biological or ecological questions. However, we found that rigorous assessments of analyses using biological records (e.g., through cross-validation or testing on independent data) were rare. Our results are similar to those from a review of species distribution model ensembles that found that only 13 out of their 224 reviewed studies (5.8%) tested model performance on independent data (Hao et al. 2019).

Tests of model performance are overly optimistic when testing uses the same data as model training (Hastie et al. 2009; Bahn and McGill 2013; Roberts et al. 2017). It is particularly important that conclusions drawn from models trained with biological records – which are opportunistic and therefore not a random sample from the population of interest – are tested on independent or semi-independent data. Given the rarity of cross-validation and tests on independent data, it seems likely that a majority of studies in this review overestimated the success of their models and analyses (in terms of the models' ability to explain or predict the data, measured using e.g., root mean squared error, R^2, or AUC), which may lead to unwarranted confidence in using biological records.

We urge future studies using biological records to use cross-validation or test on independent data as standard procedure. The existence of multiple biological records datasets that could be used as independent test data may be a particular strength of biological records. We found that most studies in this review (61%) used multiple biological records datasets, which means most studies already have datasets that could potentially be withheld from model fitting for use as independent test data. However, testing on independent data will not be insightful if the data are bad; all of the criteria that researchers consider when choosing appropriate data for model fitting should be applied when choosing independent test data, including making sure the data are of high quality and adequately represent the population of interest. Methods for spatial block cross-validation (Roberts et al. 2017) and testing on independent data (Elith et al. 2006, El-Gabbas and Dormann 2017) are now widely documented and are available in software packages (Valavi et al. 2018). Only when most studies use rigorous validation in independent data will we know whether methods for analyzing biological records are truly ready to use for answering ecological and biological questions.

Untapped potential of digital voucher specimens and non-traditional data sources

Given the large and rapidly growing archive of digital voucher specimens (e.g., eBird users uploaded 135,000 audio recordings to the Macaulay Library in 2019;eBird, 2019), the analysis of digital vouchers associated with biological records seems poised to be a rich area for new research. We were surprised to find only one study in this review that used data from the bird song archive xeno-canto (Petrusková et al. 2015), despite the fact that xeno-canto contains tens of thousands of recordings of birds from the UK and Ireland (Vellinga 2020). A few studies made innovative use of digital data that was not necessarily collected or stored as "biological" data, including photos from Google Images (Leighton et al. 2016) and Flickr (Petrusková et al. 2015, Jeawak et al. 2017), and video and audio recordings from YouTube, Vimeo, and SoundCloud (Petrusková et al. 2015).

Digital voucher records are often opportunistically collected, so analyses of them will face similar challenges as analyses using more traditional "what, where, when" biological records, including biases in data and non-standardized sampling methods. Existing techniques for addressing these challenges when analyzing traditional biological records could be extended to analyses of digital vouchers. For example, the spatial distribution of bird song dialects, which Petrusková et al. (2015) analyzed descriptively with maps of observations, could potentially be modelled using predictive species distribution modelling methods, as could animal colour morphs sampled using online photos (Leighton et al. 2016).

An unanswered question: Who uses biological records?

Surprisingly, low coder agreement (Table S1) prevented us from determining how frequently biological records were used by researchers unaffiliated with the data provider. Pearce-Higgins et al. (2018) warned that unfamiliarity with details of a dataset could lead to misuse if data are used without input from data collectors and providers. This danger can be mitigated by providing guidelines for analyzing particular datasets (Strimas-Mackey et al. 2020), but whether it is worthwhile for a data provider to write such guidelines will depend on how often their data are used by unaffiliated researchers.

Conclusions, implications for data providers, and opportunities for researchers using biological records

Given the extensive biological recording and well-developed infrastructure for collating and sharing biological records in the UK and Ireland, uses of British and Irish biological records show the current state of the field for research using biological records. Researchers and biological records data providers in other geographic areas can use results from this review to anticipate needs (such as the need for survey effort and non-detection data to enable strong analyses, and the opportunity to collect "digital vouchers") and avoid pitfalls (such as a failure to test on independent data).

Biodiversity data aggregators and data providers should identify what questions they expect their data to be used for because this will inform what data types will be most useful. Data providers can facilitate the most common uses of data by providing guidance targeted at researchers studying species distributions or temporal trends (which together accounted for over half the studies in this review). For example, eBird provides guidelines and a tutorial with R code demonstrating how to use eBird data for species distribution modelling and studying relative abundance (Strimas-Mackey et al. 2020). Data providers could further facilitate species distribution studies – which are used in spatial and conservation planning – by finding, digitizing, and putting into a standard format diverse biodiversity datasets. In contrast, studies of species abundance might be better facilitated by data providers organizing monitoring schemes or coordinating the collection of structured biological records that include non-detection and sampling effort information.

Most studies in this review treated existing methods for analyzing biological records as adequate tools for answering biological and ecological questions. Only a minority of studies developed or tested methodology. At first glance, this appears to be good news for researchers considering using biological records in their own research: the biases and non-standardized data collection methods characteristic of biological records do not seem to interfere with the use of biological records for studying ecological and biological questions. However, studies in this review rarely tested models on independent data or with cross-validation. We caution that confidence in using biological records for studying ecological or biological questions may be based on unrealistically optimistic measures of model performance derived from testing models on the same data that was used for model fitting. Fortunately, this shortcoming is easy to overcome, as over half of the studies in this review used multiple different biological records datasets. We suggest that researchers using multiple datasets reserve one or more high quality datasets to use as independent test data.

Acknowledgements

This publication emanated from research supported in part by a grant from Science Foundation Ireland under grant number 15/IA/2881. We thank Casey Patmore and Davis Gaul for advice on early versions of the instructions to coders. We are grateful to two anonymous reviewers whose comments helped us improve the manuscript. The authors declare no conflicts of interest.

Author Contributions

W.G. and J.Y. conceived the idea for this study, W.G. and E.R. collected the data, W.G., J.Y. and D.S. conducted the analyses, W.G. led the writing with commenting and reviewing by J.Y., E.R., D.S., H.W., P.C. and L.L.S.. All authors approved the final manuscript.

Data Accessibility Statement

Data, review protocols, and scripts to conduct all analyses are available at https://doi.org/10.5281/zenodo.3760678 (data and review protocols), and https://doi.org/10.5281/zenodo.3760737 or https://github.com/wgaul/systematic_review (scripts).

Supplementary Materials

The following materials are available as part of the online article from https://escholarship.org/uc/fb

Table S1. The reliability of coding for variables as measured using Krippendorff's alpha.

Table S2. Data sources used by eligible studies in this review, and the number of studies that used each data source.

Figure S1. The position of each of the original data type categories on the first two MCA dimensions.

Figure S2. Schematic demonstrating the random permutation procedure used to test for independence between data types and analysis approach and between data types and study questions.

Appendix S1. Search terms used to find potentially relevant studies.

Appendix S2. List of reviewed studies.

References

Amano T., Lamming J.D.L. & Sutherland W.J. (2016) Spatial gaps in global biodiversity information and the role of citizen science. BioScience, 66, 393–400.

Arksey H. & O'Malley L. (2005) Scoping studies: towards a methodological framework. International Journal of Social Research Methodology: Theory and Practice, 8, 19–32.

August T., Harvey M., Lightfoot P., Kilbey, D., Papadopoulos, T. & Jepson, P. (2015) Emerging technologies for biological recording. Biological Journal of the Linnean Society, 115, 731–749.

Bahn V. & McGill B.J. (2013) Testing the predictive performance of distribution models. Oikos 122, 321–331.

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J. & Elston, D.A. (2010) Regression analysis of spatial data. Ecology Letters, 13, 246-264.

Boakes E.H., McGowan P.J.K., Fuller R.A., Chang-qing, D., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology 8, e1000385.

Canty A. & Ripley B. (2015) boot: bootstrap R (S-Plus) functions. R package. https://cran.r-project.org/package=boot

Department of Culture Heritage and the Gaeltacht (2017) National biodiversity action plan 2017-2021. https://www.cbd.int/doc/world/ie/ie-nbsap-v3-en.pdf [accessed 12 August 2020].

Domencich, T.A. & McFadden, D. (1975) Urban travel demand: a behavioral analysis. North-Holland Publishing Company, Amsterdam, the Netherlands.

eBird (2019) eBird 2019 – Year in review. Web article available at https://ebird.org/news/ebird-2019-year-in-review. Viewed 17 Feb 2020.

Efron B. & Tibshirani R.J. (1993) An introduction to the bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). Chapman and Hall, New York, USA.

El-Gabbas A. & Dormann C.F. (2017) Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography 41, 1161-1172.

Elith J., Graham C., Anderson R., et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29, 129–151.

Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B., Munson, M.A., Hooker, G., Riedewald, M., Sheldon, D. & Kelling, S. (2010) Spatiotemporal exploratory models for broad-scale survey data. Ecological Applications, 20, 2131-2147.

Gamer M., Lemon J. & Singh I.F.P. (2012) irr: various coefficients of interrater reliability and agreement. R package. https://cran.r-project.org/package=irr

Gaul, W. (2020) wgaul/systematic_review: release for final author review before journal submission (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.3760736.

Gaul, W., Roark, E. & Yearsley, J. (2020) Data for "What have biological records ever done for us? A systematic scoping review" (Version 1.0.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3760678.

Guélat, J. & Kéry, M. (2018) Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods in Ecology and Evolution, 9, 1614-1625.

Haddaway, N.R., Collins, A.M., Coughlin, D. & Kirk, S. (2015) The role of google scholar in evidence reviews and its applicability to grey literature searching. PLoS ONE, 10, e0138237.

Harzing A.W. (2007) Publish or perish. Digital resource available at https://harzing.com/resources/publish-or-perish.

Hastie T., Tibshirani R. & Friedman J. (2009) The elements of statistical learning: data mining, inference and prediction (2nd ed.). Springer, New York, USA.

Hao T., Elith J., Guillera-Arroita G. & Lahoz-Monfort J.J. (2019) A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Diversity and Distributions, 25, 839–852.

Hart, A.G., Nesbit, R. & Goodenough, A.E. (2018) Spatiotemporal variation in house spider phenology at a national scale using citizen science. Arachnology, 17, 331–334.

Hayhow, D.B., Burns, F., Eaton, M.A., et al. (2016) State of nature 2016. The State of Nature partnership. DOI: 10.13140/RG.2.2.14263.93602

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

Jeawak S.S., Jones C.B. & Schockaert S. (2017) Using flickr for characterizing the environment: an exploratory analysis. Leibniz International Proceedings in Informatics, LIPIcs. Dagstuhl, Germany: Schloss Dagstuhl—Liebniz-Zentrum fuer Informatik.

Krippendorff K. (2013) Content analysis: an introduction to its methodology (3rd ed.). SAGE, Los Angeles, USA.

Leighton G.R.M., Hugo P.S., Roulin A. & Amar A. (2016) Just Google it: assessing the use of Google images to describe geographical variation in visible traits of organisms. Methods in Ecology and Evolution, 7, 1060–1070.

Lennon, J.J. (2000) Red-shifts and red herrings in geographical ecology. Ecography, 23, 101-113.

Lortie, C.J. (2014) Formalized synthesis opportunities for ecology: systematic reviews and meta-analyses. Oikos, 123, 897–902.

Mair P. & De Leeuw J. (2019) Gifi: multivariate analysis with optimal scaling. R package version 0.3-9. https://CRAN.R-project.org/package=Gifi

Manly B.F.J. (2007) Randomization, bootstrap, and Monte Carlo methods in biology (3rd ed.). Chapman & Hall/CRC, Boca Raton, USA.

McCallen, E., Knott, J., Nunez-Mir, G., Taylor, B., Jo, I. & Fei, S. (2019) Trends in ecology: shifts in ecological research themes over the past four decades. Frontiers in Ecology and the Environment, 17, 109-116.

Meyer C., Weigelt P. & Kreft H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecology Letters, 19, 992–1006.

National Biodiversity Data Centre (n.d.) Introducing the National Biodiversity Data Centre. The National Biodiversity Data Centre, Waterford, Ireland. https://www.biodiversityireland.ie/wordpress/wp-content/uploads/NBDC-Overview-WEB.pdf [accessed 12 August 2020].

National Biodiversity Network (2015) Collecting and sharing biological data to educate and inform: NBN Strategy 2015-2020. https://nbn.org.uk/wp-content/uploads/2015/10/NBN-Strategy-2015-2020-Aug-2015.pdf [accessed 2 June 2020].

Pacifici K., Reich B.J., Miller D.A.W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. & Collazo, J.A. (2017) Integrating multiple data sources in species distribution modeling: a framework for data fusion. Ecology, 98, 840–850.

Pearce-Higgins J.W., Baillie S.R., Boughey K., et al. (2018) Overcoming the challenges of public data archiving for citizen science biodiversity recording and monitoring schemes. Journal of Applied Ecology, 55, 2544–2551.

Petrusková T., Diblíková L., Pipek P., Frauendorf, E., Procházka, P. & Petrusek, A. (2015) A review of the distribution of Yellowhammer (Emberiza citrinella) dialects in Europe reveals the lack of a clear macrogeographic pattern. Journal of Ornithology, 156, 263–273.

Pham, M.T., Rajić, A., Greig, J.D., Sargeant, J.M., Papadopoulos, A. & McEwen, S.A. (2014) A scoping review of scoping reviews: advancing the approach and enhancing the consistency. Research Synthesis Methods, 5, 371–385.

R Core Team (2018) R: A language and environment for statistical computing. https://www.r-project.org/.

Roberts D.R., Bahn V., Ciuti S., et al. (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40, 913–929.

Strimas-Mackey M., Hochachka W., Ruiz-Gutierrez V., Robinson, O.J., Miller, E.T., Auer, T., Kelling, S., Fink, D. & Johnston, A. (2020) Best Practices for Using eBird Data (Version 1). https://doi.org/https://doi.org/10.5281/zenodo.3620739

Sullivan B.L., Wood C.L., Iliff M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009) eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation, 142, 2282–2292.

Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C. & Guisan, A. (2014) Measuring the relative effect of factors affecting species distribution model predictions. Methods in Ecology and Evolution, 5, 947-955.

Tingley M.W. & Beissinger S.R. (2009). Detecting range shifts from historical species occurrences: new perspectives on old data. Trends in Ecology and Evolution, 24, 625–633.

Valavi R., Elith J., Lahoz-Monfort J.J. & Guillera-Arroita G. (2019) block CV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution, 10, 225-232.

Van Swaay C.A.M., Nowicki P., Settele J. & Van Strien A.J. (2008) Butterfly monitoring in Europe: methods, applications and perspectives. Biodiversity and Conservation, 17, 3455–3469.

Vellinga W.P. & Planqué R. (2015) The xeno-canto collection and its relation to sound recognition and classification. Working notes of CLEF 2015 – Conference and labs of the evaluation forum; 8-11 Sept 2015; Toulouse, France.

Vellinga W. (2020) Xeno-canto – Bird sounds from around the world. Xeno-canto Foundation for Nature Sounds. Occurrence dataset https://doi.org/10.15468/qv0ksn accessed via GBIF.org on 22 April 2020.

Wieczorek J., Bloom D., Guralnick R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012) Darwin core: an evolving community-developed biodiversity data standard. PLoS ONE, 7, e29715.

Submitted: 06 June 2020
First decision: 28 August 2019
Accepted: 3 July 2020

Edited by Robert J. Whittaker