Guidelines for Coding Articles for Systematic Review

13 September 2018
Willson Gaul & Jon Yearsley
University College Dublin
email: willson.gaul@ucdconnect.ie

Purpose of the Project

The objective of this study is to determine how biological records data are currently being used in research. I am investigating this by reviewing studies covering the British Isles published in the four and a half years from 2014 to the present (June 2018). Specifically, I ask questions that can be grouped within a few main themes.

What is being studied in papers that use biological records data?

What questions are being studied? Are studies primarily focused on developing and testing methods for analyzing biological records data, or are they primarily focused on studying some ecological or biological question? Are the studies asking questions about individual species or communities? Are studies testing macro-ecological theory? What taxonomic groups are studied?

What types of biological records data are used?

Was the data "what, where, when" data, or was there additional data associated with each record (e.g. explicit non-detection data, survey effort, visit-specific weather or habitat data, life-stage information like flowering status, or photo or audio documentation)? Were the data collected as part of a semi-structured scheme (e.g. UK Butterfly Monitoring Scheme or a breeding bird atlas)? Did authors analyze data from open-access data repositories, or did they analyze data held by organizations they work for? What is the time span and spatial area covered by studies?

How are the biological records data used within the study?

Were the biological records data used as a response variable, as a predictor variable, or were they used in a facilitative role (e.g. identifying recorders to recruit for a subsequent study)? How was sampling bias addressed? Are some taxonomic groups more commonly used for some types of questions?

What analysis approaches were used?

Did researchers analyze biological records using predictions, statistical inference, or did they only use descriptive analyses? Did the type of data influence the analysis approach?

Who is using biological records data?

What authors and institutions perform the research? Were the authors associated with the institution that collected or supplied the data?

Instructions for Article Coders

1. Please take your time. Expect to spend at least 1 hour per article

The amount of time it takes to code an article will vary depending on the structure of the article and how clearly the authors write. Finishing coding an article in under 1 hour would be a pleasant surprise – in the beginning you can expect to spend at least an hour on each article as you figure out what kinds of clues to look for to code each category. Please take your time! Accurately coding each article is more important than coding lots of articles.

2. Fill in every column for each study

Some columns may require you to look beyond the study for information. For example, you may need to visit the website of a recording scheme to read their sampling protocols, or you may need to browse through a dataset in supplementary materials or on the Global Biodiversity Information Facility (GBIF) website to see what variables are included in the dataset.

3. Use a bit of detective work

I often scan back and forth through an article many times, re-reading important sentences that have information relevant to many different categories. Sections of a paper that often have a lot of relevant clues include the abstract, the last paragraph of the *Introduction* section, the first paragraphs of the *Methods* and *Results* sections, and captions for graphs.

4. You can assume all articles are suitable for the structured review

I have already evaluated each article for eligibility. However, if you think I made a mistake and a study shouldn't be eligible, please make a note of that and ask me specifically. The eligibility requirements are listed at the end of this document.

5. Focus on the biological records!

Assume for every category that I am interested in the biological records data and analyses done using the biological records data. Many studies include multiple different study questions, and they may make use of many different datasets. For example, a study might use biological records data about plants, experimental data collected by the authors in a greenhouse experiment, and weather data. This review is focusing on the characteristics and uses of the biological records data specifically. *Code categories with regard to the biological records data.* For example, the “data type” columns are asking about the data type for the biological records data, not the other data. In the example above, if the biological records did not include abundance information, than the category “data type – abundance” should be FALSE, even if the experimental data did have abundance information. The same thing applies to questions about the analysis methods – if a specific analysis or statistical test didn’t use biological records data, then don’t consider that test or analyses when coding.

Thanks!

- Willson

Column Heading Explanations

Basic Descriptive Info

qualifies
TRUE / FALSE the article qualify for the systematic review (see eligibility criteria at the end of this document)

authors
Names of all authors

publication
the name of the journal or conference proceedings in which the article was published. In the case of studies that were not published in an academic journal, this might be the name of an organization that published the report (e.g. “EPA” or “The Nature Conservancy”)

doi
digital object identifier. This is a unique code that can be used to search for the article online

year
year of publication

keywords
the keywords listed in the publication (not your best guess). These can often be found after the abstract or on a scroll bar on the right side of the article

institution of first author
the institution given for the first author in the author list. This can often be found in a footnote.

institution of last author
the institution given for the last author in the author list.

author associated with proximate data provider
TRUE / FALSE are any of the authors from the institution that provided the biological records data? Find this by looking at the author affiliations listed with the article and reading the methods section to find out how the data was acquired. You may need to Google the scheme or look at the author list of the publication that originally described the scheme.

proximate data source
the organization that most directly provided the biological records data to the authors. Note that this might not be the organization that collected the data. For example, if the data were collected by the “UK Butterfly Monitoring Scheme” but the methods state that the authors downloaded the data from “GBIF” (Global Biodiversity Information Facility), record “GBIF” in this column.

country of study
list all countries that the data and/or study cover

spatial extent of study
area that the study covers, either quantitatively (e.g. in square kilometers) or qualitatively (e.g. “British Isles”)

start year
the first year from which data was used for this study

end year
the last year from which data were used for the study

Study Question / Study Focus

Broad study question paradigm

Note: these categories are not mutually exclusive. There might be multiple goals of a study. Mark TRUE for all categories that are a focus of the study.

methodology development or analysis
TRUE / FALSE a major focus of the study is the development and presentation of a new analysis method, or testing how well a method works.

individual species question
TRUE / FALSE a major focus of the study is asking questions about individual species. It is possible to do this for multiple individual species. For example, if a study analyses the effect of temperature on the egg-laying date of 10 different bird species, mark TRUE in this column, because each outcome being measured (the egg-laying date of each of the 10 species) is an outcome about a single species.

community question
TRUE / FALSE a major focus of the study is asking questions about ecological communities (rather than the individual species in those communities). For example, studies of species richness, species diversity, or total biomass production would be marked TRUE in this column.

Response variables

response variable - species richness
TRUE / FALSE species richness (number of species) is an outcome / response variable that the study estimates or analyzes using biological records

response variable - diversity
TRUE / FALSE species turnover or community diversity is an outcome / response variable that the study estimates using biological records. Diversity as used here refers to beta diversity (turnover) or any diversity measure other than species richness, including any measure that accounts for both richness and species identity, abundance, evenness, genes, or traits (e.g. “Shannon diversity index”, “functional trait diversity).

response variable - distribution
TRUE / FALSE the spatial distribution or range of a species is an outcome / response that the study estimates or analyzes using biological records.

response variable - abundance

TRUE / FALSE the abundance of the study organism (e.g. number of individuals or percent cover of a plant species) is an outcome / response that the study estimates or analyzes using biological records.

response variable - phenology

TRUE / FALSE phenology (periodic life-cycle events) is an outcome / response that the study estimates or analyzes using biological records. For example, studies of the seasonal timing of bird egg-laying, the timing of flowers blooming, or the timing of migration would be marked TRUE in this column.

Study focuses

trends over time

TRUE / FALSE the study uses biological records data to estimate or analyze how something has changed over time. This will frequently be a change over time in one of the response variables listed above. For example, a study analyzing how a frog species has expanded its range in Great Britain over the last 100 years would be marked TRUE in this column and in the "distribution" column.

alien species focus

TRUE / FALSE a major focus of the study is about one or multiple alien or invasive species

taxonomic group

the name(s) of the group(s) of organisms that the study is about. Common names are fine (e.g. "birds"). If the study uses more than 3 different taxonomic groups, you can enter "many" in this column rather than listing all the groups.

terrestrial – marine – both

terrestrial / marine / both – The realm from which biological records were used. Freshwater systems are considered terrestrial for this category.

Data Structure

Note: ***The "data structure" columns are asking about whether the data had this structure, regardless of whether the study explicitly used or mentioned this structure****. If the data had any of these characteristics, mark TRUE in the appropriate column(s), even if it doesn't seem like the study used that information explicitly in the analysis. Coding these columns can often require some extra digging – for example if a study downloaded a bunch of data for an area from GBIF, you might have to look at the specific datasets to know if any of them were organized monitoring schemes, and then you might have to look up a monitoring scheme website and protocols to know whether the scheme specifies sampling effort (e.g. a standard transect length or time duration for surveys).*

When studies use multiple different data sources for biological records, code the data structure columns TRUE if they are true for any of the biological records data. For example, if a study uses both *opportunistic "what, where, when" records of frogs, but also uses data from a frog monitoring scheme in which volunteers count frogs for 30 minutes one night per week, then the "data structure – sampling effort known" and "data structure – organized data collection scheme" categories would both be TRUE, because at least some of the biological records data had that attribute.*

data structure – organized data collection scheme

TRUE / FALSE at least some of the biological records data come from an organized or structured monitoring scheme that specifies at least some standard elements of the data collection process (e.g. assigns specific locations, or requires a specific transect length or survey duration). For example, a study using data from the "Audubon Christmas Bird Count" would be marked TRUE in this category because that scheme tells the observers when and how to collect the data. This category pertains to organization of the data *collection* process, not to organization of the data *submission* process or *post-hoc* organization of data that was collected opportunistically. For example, the Wisconsin Odonata Survey (http://wiatri.net/inventory/odonata/) is a dedicated website for people to submit dragonfly observations, but it is *not* an organized data collection scheme because it does not specify a set of protocols for data collection – it's just a data entry website. Basically, this category is asking: is there some organization *telling people how to collect data*. Many studies integrate data from multiple sources, and sometimes only some of the data come from organized schemes. Code this category as TRUE if at least some of the data is a recognizable set of data from a known monitoring scheme. If the data have been aggregated so that it is no longer possible to distinguish whether any given record came from a monitoring scheme or from opportunistic sampling, then code this category as FALSE, even if some of the data were generated by a monitoring scheme originally. Basically, the criteria here is whether there are individual records that are assignable to specific monitoring schemes – if there are, then code this category as TRUE.

data structure - sampling effort known

TRUE / FALSE at least some of the biological records data come from sampling with measured or pre-specified sampling effort. This refers only to sampling effort information collected or specified as part of the data collection process. This does not refer to effort estimated *post-hoc (*e.g. using list length as a covariate to estimate sampling effort). Sampling effort may be reported either as the duration of sampling (e.g. 10 minute point count) or an area or distance sampled (e.g. data collected from a 1 km transect). The sampling effort could be known because it was either 1) reported with each observation (e.g. eBird data where people report the number of minutes they spent looking for birds) or 2) specified as part of the data collection protocol (e.g. a monitoring scheme in which transects are always 1 km, so the sampling effort is known to be 1 km even though the transect length isn't reported with each observation). This category is asking whether sampling effort information is *available* for the data, not whether sampling effort is reported in the article or study. So, a study that uses monitoring scheme data collected using a standardized transect length is TRUE in this category, even if the study's "Methods" do not report the transect length or any other measure of sampling effort. *Coding this category often requires reading the websites or original publications from the original data source* (e.g. read the UK Butterfly Monitoring Scheme website to find out if the scheme protocols specify a standardized sampling effort).

data structure - non-detection

TRUE / FALSE at least some of the biological records data includes absence or non-detection information, either explicitly or implicitly. Any sampling method that results in records of everything that is detected implicitly includes non-detection information (because things not recorded can be assumed to have not been detected). If the sampling method that generated biological records is known, and that sampling method includes complete recording of all detected species, then this category should be TRUE. If the sampling method that generated the data is unknown, and the data do not include "zero" values explicitly indicating absence, then this category should be FALSE. Clues include phrases like "complete checklists" or "all species seen were recorded" (which imply that things that were not recorded were in fact not seen). *Coding this category often requires detective work – you may need to read the websites of monitoring schemes or original publications from the original data source* in order to determine whether all species detected were recorded.

data structure - multiple datasets integrated for analysis

TRUE / FALSE the analysis combined biological records data from multiple different biological records datasets. This is only asking about whether multiple *biological records* datasets were integrated – this is not asking about whether e.g. climate or other non-biological records datasets were used.

Data Type

Note: In contrast to the previous four columns, ***the "data type" columns are asking whether the data type was used for analysis****. A data type might be available but not used because it wasn't relevant to the* *question or simply because the authors chose not to use it. If the data type was not used, mark FALSE in these columns, regardless of whether that data type was originally available with the data. If possible, looking at data presented in supplementary materials or elsewhere can help you figure out what additional information the study used along with the basic "what, where, when" information.*

When studies use multiple different data sources for biological records, code the data type columns TRUE if they are true for any of the biological records data. For example, if a study uses both opportunistic "what, where, when" records of moths, but also uses abundance data from a moth monitoring scheme in which volunteers count all moths caught in a light trap run for one night per week, then the "data type – non-detection" and "data type – abundance" categories would both be TRUE, because at least some of the biological records data had those attributes.

data type - what where when only
TRUE / FALSE at least some of the biological records data used contained only "what, where, when" information (species name, location, and date), but no information on abundance (counts of individuals), survey effort, habitat type, weather, or other information.

data type - abundance
TRUE / FALSE at least some of the biological records data used included abundance information (e.g. the number of individuals counted, or percent cover of plants).

data type - visit specific covariates
TRUE / FALSE at least some of the biological records data include additional information collected by the observer at the same time they collected the biological records data. For example, a butterfly monitoring scheme that asks recorders to report the temperature and wind speed at the time that they start looking for butterflies would be marked TRUE in this category.

data type - life stage
TRUE / FALSE at least some of the biological records data explicitly include information about some aspect of life stage (e.g. whether a plant was flowering or not).

data type - photo
TRUE / FALSE at least some of the biological records data include a photograph of the observed organism. For example, some online citizen science databases allows observers to upload a photo when they submit a record of a species.

data type - audio
TRUE / FALSE at least some of the biological records data include a sound recording of the observed organism.

data type - video
TRUE / FALSE at least some of the biological records data include a video recording of the observed organism.

data type - voucher of some kind

necessary for analysis

TRUE / FALSE vouchers can be either physical specimens (e.g. from a museum or herbarium) or digital vouchers (e.g. photos or sound recordings).

data type - gridded

TRUE / FALSE at least some of the biological records data is provided as gridded data (e.g. locations are given as 10 km squares in which a species was seen) rather than as point data (e.g. latitude/longitude coordinates given for a single point)

grid resolution

(only if TRUE in previous category) the size of the grid cells that data is presented in (e.g. 10 km). If multiple grid sizes were used, list all of them (e.g. “10 km; 1 km”). If the previous category is FALSE, mark NA here.

Analysis Approach

Note: Evaluate these “results type” categories ***only for analyses that use the biological records data.*** *Many papers do multiple analyses, some of which might not use any biological records data and are therefore not relevant to this review.*

results type - inference

TRUE / FALSE the results include statistical inference (including but not limited to hypothesis testing). This means that the data are being used to say something about a larger population (rather than just saying something about the specific data in hand). For example, a study that estimates the number of Curlews in Ireland and includes 95% confidence intervals is doing statistical inference. Evaluate this category *only for analyses that use the biological records data*. For example, a paper that does multiple analyses, one of which does inference on genetic samples collected by authors and another that does range prediction using biological records data would be coded FALSE for this category, because there was no inference *on the analysis using biological records data*. Common phrases that indicate that a study is doing statistical inference are "confidence intervals", “p-value”, “significantly different from”, and "posterior probability distribution".

results type - prediction

TRUE / FALSE results include prediction (e.g. reporting the modeled range or distribution of a species). Prediction could be in time or in space (or both). For example, results that just show a map of every 10km square where a species was seen are *not* doing prediction (just description), but results that show a map of the *expected* distribution of a species are doing prediction because they are predicting that the species is present, even in places where no observations have been collected. Evaluate this *only for analyses that use the biological records data*.

results type - descriptive only

TRUE / FALSE the results are descriptions of the data but do not include any inference or prediction. Examples include maps of where a species was seen or the percent of records that were the focal species (but no confidence intervals around those percents). Evaluate this *only for analyses that use the biological records data*.

Methods

biological records as response variable

TRUE / FALSE biological records data are used as a response (outcome) variable in analysis. For example, a study that uses habitat type, temperature, and rainfall to predict the distribution of blackbirds would be marked TRUE (the biological records of blackbird occurrence are the response variable, and the predictor variables are habitat type, temperature, and rainfall). Similarly, a study that uses biological records to estimate and compare species richness in different areas should be marked TRUE in this column because the number of species measured by biological records data is the response variable.

biological records as predictor variable

TRUE / FALSE biological records are used as a predictor variable in analysis. A study that uses observations from biological records to modify any other variable is using the biological records as a predictor (whether there is an explicit statement of a model or not). For example, a study that uses biological records of Golden Eagles to identify areas that have high potential for developing nature-based recreation programs would be marked TRUE in this column because it uses biological records (eagle occurrences) as a predictor variable, while the response (outcome) variable is the potential for nature-based recreation.

biological records in facilitative
role but not analyzed

TRUE / FALSE the study doesn't analyze biological records data directly, but rather uses it to facilitate some other aspect of the study. For example, a study that uses a biological records database of shorebirds to identify people who submit records from coastlines, and then asks those people to collect data on plastic trash on beaches should be marked TRUE because it uses the biological records data not for analysis but to facilitate finding volunteers for a new study.

spatially uneven sampling
corrected for

TRUE / FALSE the study addresses, corrects for, or accounts for spatially uneven sampling in some way. This includes methods to address uneven data density and spatial bias (preferential sampling of some areas). Addressing spatially uneven data density be as simple as “very few records were available from Scotland, so we restricted our analysis to England and Wales” or could be a complex statistical correction method. If there is anything at all that suggests that the authors changed their analysis in some way because of spatially uneven data, mark TRUE in this column. Note that sometimes studies will mention spatial issues in the data, but not correct for it in any way, in which case you would enter FALSE in this column.

prediction performance measure

(only if "results type - prediction" was marked TRUE) a list of all prediction performance measures used (e.g. AUC, TSS, RMSE). Look for this in the "Methods" section, often in a sentence like "Model performance was evaluated using the area under the receiver operating characteristic curve (AUC)." If the "results type - prediction" category is FALSE, mark NA here.

cross validation

TRUE / FALSE model testing was done using multiple different holdout datasets (as opposed to testing the model using the training data or using just a single randomly selected holdout data set). Hint: look for the phrase "cross validation" in the "Methods" section.

testing using simulation

TRUE / FALSE the analysis methods were tested using simulations in addition to (or instead of) using biological records data

testing using training data

TRUE / FALSE methods or models were tested using the training data (e.g. root mean square error is reported for the data used to estimate a linear regression line)

testing using data subset

TRUE / FALSE methods or models were tested using a hold-out subset of the data

testing using independent dataset

TRUE / FALSE methods or models were tested using an independently collected dataset

analysis method

a list of statistical analysis methods used for the main analysis (e.g. GLM, linear regression, Bayesian analysis). **This category will be coded only by wg for his personal reference. I do not plan to ask 2nd readers to code this.**

predictor variables

a list of the predictor variables (if any) used in the analysis (e.g. habitat type, weather, time of year, time of day). **This category will be coded only by wg for his personal reference. I do not plan to ask 2nd readers to code this.**

methodology focus

Fill if "methodology development/test" is TRUE. This would help inform questions and categories if later I decide to re-visit and code all methodology studies to explore what methodology studies are currently focusing on. I think that for such an analysis to be useful, it would have to be more detailed than just "testing spatial bias correction." I would want more categories to say exactly what the methodology is trying to do. **This category will be coded only by wg for his personal reference. I do not plan to ask 2nd readers to code this.**

Final check

coding DONE

TRUE / FALSE you have finished coding all categories for this article. Mark this as TRUE only after you have filled in something for every category (either TRUE, FALSE, NA, etc). If you have not yet filled in some category, mark FALSE in this column so you and I know to come back to this later. **Don’t send this to 2nd readers.**

other notes

This is a place for you to write any important comments that don’t fall under the other categories. **I do not intend to provide this column to 2nd readers, as I don’t want them to have an “escape” option that saves them from having to resolve difficult situations without making a decision about how to code them.**

Eligibility Criteria

Studies are deemed eligible if they are original research (no reviews or idea papers) published in the English language, use opportunistic biological data collected with non-standardized or semi-standardized designs, include (but are not necessarily limited to) data from Ireland or the UK, and the full text of the study is available through the University College Dublin library online platform, Google Scholar, Google search results, or ResearchGate. Grey literature is included and therefore peer review is not required. Studies using semi-standardized data collection schemes (e.g. UK Butterfly Monitoring Scheme) are included. Publications of data (e.g. atlases or data papers) are not considered eligible unless they included analysis of the data. Only studies using a sample size of greater than 20 are included; this sample size was chosen arbitrarily, mainly to exclude studies in which re-examination of museum specimens resulted in a taxonomic identification revision for one or a few specimens, thereby changing the known range of a species to include or exclude Ireland and the UK. Studies using museum data are considered eligible when the museum data used is similar in format to biological records data (e.g. "what, where, when" data); studies that used museum specimens only for taxonomic, genetic or morphology studies are excluded. Studies using only fossil records are not included. Studies using data from phenology networks are included; for the purposes of this review such data are considered biological records data with associated additional visit-specific data (e.g. the flowering status of plants). Studies for which all data was collected by the study authors are not considered biological records data for the purposes of this review and are excluded. The minimum required information in the data is a taxonomic name, a location, and date ("what, where, when"); additional information is permissible.

Eligibility checklist for inclusion

- original research
- English language
- used biological records data
- used (but was not necessarily limited to) data from Ireland or the UK
- data not collected entirely by study authors
- the full text of the study was available through the UCD library online platform, Google, GoogleScholar, or ResearchGate.
- study performed at least one new analysis of the data (data papers, biotic atlases and reports of previous analyses are not eligible). Descriptive analyses are considered analyses and are eligible.
- sample size of greater than 20 biological records used
- not exclusively fossil data