Readme

Description of relevant tables in the sqlite database file

For each table, a description will be provided. Note that where column names are repeated the descriptions for the columns will not be repeated.

# Tables describing the corpus

## Table name: allrefs

Short description

Table for all references (before de-duplication)

Long description

This table compiles all the references retrieved from the search strategy, along with automatically exported metadata, from both Web of Science and Scopus. Note that the source database is indicated in the "source\_database" column. Duplicates can appear with and between databases. A unique "analysis\_id" is assigned to each row (reference).

Column descriptions

analysis\_id : The unique record identifier for each unit of publication (i.e. when there are duplicates and before unique references are extracted, analysis\_id > duplicate\_id).

duplicate\_id : publications identified as duplicates by title and year matching will receive the same duplicate\_id

source\_database : the citation indexed database the record was retrieved from. If the record had duplicates identified from multiple databases, these will be separated by " | " (e.g. "Web Of Science | Scopus")

search\_substring : the search string used to retrieve the record. If multiple search strings returned the same record, numerous entries will be separated by " | "

database\_article\_id: the unique identifer given to the reference by its native source database

type: the type of publication

author: the authors of the publication

title : the text for the title of the publication

source\_title : title of the source of publication (e.g. journal, book)

year: the year the publication was published

volume: the volumne of the journal (if applicable)

number: the number of the issue (if applicable)

pages: which pages in the issue the publication appeared in

abstract : the abstract of the publication

keywords : author tagged keywords

keywords\_other : other keywords tagged by the database

publisher: the publisher of the publication

editor: the editor of the publication

language: the language of publication

affiliation: the affiliation (according to the protocol of the native citation indexed database, usually first author affiliation)

funding: any funding information

doi: the Digital Object Identifier

issn: the issn

isbn: the isbn

journal: the name of the publishing journal (if applicable)

research\_areas: The research areas attributed to the publishing journal according to the protocol of the source database

web\_of\_science\_categories: another form of research area classification from WOS

booktitle: if applicable

book\_group\_author: if applicable

book\_author: if applicable

organization: The organization publishing the publication

series: series of publication (if applicable)

da: date of publication

## Table name: uniquerefs

Short description

Table for all the unique references

Long description

This table contains only unique references that have an abstract that is not NA. Note that the function "extract\_unique\_references" from the R revtools package was used to decide which record to retain when duplication was identified. In this case, the most complete record was retained where duplicates were identified. Therefore in this dataset, the duplicates are removed, thus each row represents a unique publication.

Column description

Columns remain the same as for the "allrefs" table

# Tables describing ‘seen’ articles

## Table name: seen\_screen

Short description

Table for the reviewer screening choices

Long description

This table provides the screening decisions for the manually screened subset of the corpus. Note that this subset contains articles that were sampled semi-randomly as well as non-randomly (for more information see the description on the sample\_screen column). The include\_screen column giving the include/exclude (i.e. 1/0 respectively) decisions was used to train the relevance predictions from the machine learning model.

Note here that the number of articles that are classified as "test\_list" under the "sample\_screen" variable totals 87 as opposed to the 94 reported in the protocol. This is because after screening, some test list articles were determined to be irrelevant after revisions made to the eligibility criteria during reviewer training.

Column description

reviewer: The name of the primary reviewer to screen the article

reviewer\_2: The name of the second reviewer if a double blind screen was done.

sample\_screen: how the article was sampled in order to be screened. Some articles were test list articles, some were sampled semi-randomly for screening (by randomly sampling 1000 articles from each sub-string and then pooling the results), some were screened with priority according to a high predicted relevance as part of our active learning approach, and others were added after the coding stage in order to supplement article types that were under-sampled.

include\_screen: whether the decision was made to include (1) or exclude (0) the article

## Table name: seen\_coding

Short description

Table for the reviewer coding choices

Long description

This table provides all the reviewer coding decisions used to fit the mode. Each column represents a unique variable and label combination, and the reviewer decision is coded as 1 for being relevant and 0 for irrelevant. These decisions were used to train the relevance predictions from the machine learning model for the different metadata variables.

Column description

coder\_1: the name of the reviewer who coded the article

coder\_2: the name of the second reviewer who coded the article if the article was flagged for double coding.

*NB the rest of the column names following the coder\_2 column represent the different metadata variables that were coded in the coding sheet. For descriptions of the variables and labels for each variable, please see the coding instructions file. The column names are in the following format: variable\_name.label\_name if multiple labels are used, or variable\_name if just a binary 1/0 label. For example, for the variable "climate\_threat" and the label "Extreme\_weather" the corresponding column is named "climate\_threat.Extreme\_weather". Also note that the variable "ORO\_type" is decomposed into two variables with increasing granularity: oro\_branch which refers to whether the ORO aims to address mitigation (Mitigation), natural resilience (Nature) or societal adaptation (Societal) objectives, and oro\_any which refers specifically to which type of OROs are mentioned in the article. For all multi-label variables, multiple choice is possible.*

# Tables describing model predictions

## Table name: pred\_relevance

Short description

Table for relevance (i.e. screening)

Long description

This table provides the predicted relevance (lower, mean, upper and std) from the screening decision binary label model. Each row is a unique article (from uniquerefs) and is indexed by analysis\_id.

Column description

relevance\_mean: the mean predicted relevance

relevance\_std: the standard deviation of predicted relevance around the mean

relevance\_lower: the mean minus the standard deviation

relevance\_upper: the mean plus the standard deviation

## Table name: pred\_x (where x is a coded metadata variable name)

The rest of the model predictions follow this naming convention.

For binary labels (i.e. there was only a 1 or 0 outcome possible for the variable), the predictions table will have the following column names: analysis\_id, `0 – relevance – mean\_prediction`, `0 – relevance – std\_prediction`, `0 – relevance – lower\_prediction`, and `0 – relevance – upper\_prediction`.

For multi-label variables, this naming convention is followed, but columns are added for each unique label. For example, for the variable “climate\_threat” and label “Temperature”, there will be columns for `climate\_threat.Temperature - mean\_prediction`, `climate\_threat.Temperature - std\_prediction` and so on.

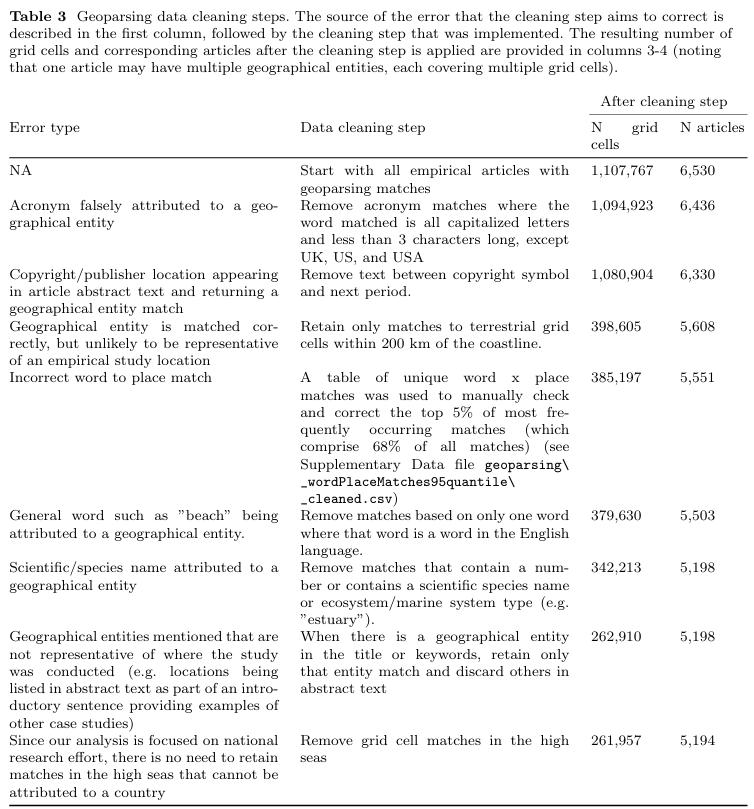
The following tables refer to the coded variable predictions:

pred\_adapt\_to\_threat, pred\_blue\_carbon, pred\_climate\_mitigation, pred\_climate\_threat, pred\_data, pred\_ecosystem\_type, pred\_marine\_system, pred\_method\_type, pred\_oro\_any\_mitigation, pred\_oro\_any\_nature, pred\_oro\_any\_societal, pred\_oro\_branch

The table "pred\_oro\_type\_long" combines and formats the predictions from pred\_oro\_any\_*x* into long format for easier plotting.

# Tables describing geoparsing results

These tables represent the geoparsing results that were cleaned for the largest sources of systemic bias, following the cleaning procedure described in the Supplementary Information of the associated article (Supplementary Table 3 summarizing these steps is reproduced below):



## Table name: geoparsed-text

Table description

This table contains the place information that was extracted from the text in each document using the geoparsing function in Mordecai. Each row represents a unique document and place combination.

Column description

Each row has the following unique identifier(s) that can be used to join/merge with other tables:

* "analysis\_id": A unique integer identifying the unique document
* "duplicate\_id": Analagous to analysis\_id -- don't worry about this one too much

The table contains the following relevant columns as metadata:

* "title": The title of the study
* "word": The word identified within the title/abstract text that indicates a place name
* "spans": Where in the text the word was located
* "country\_predicted", "country\_conf": which country the word was predicted in, and with what confidence
* "lat", "lon": the latitude and longitude of the centroid of the place
* "geonameid": the corresponding ID from the geonames database
* "place\_name": the corresponding place name from the geonames database

## Table name: grid\_df\_res2.5

Table description

This table contains the grid information that the geoparsed text locations and natural earth shape files are binned to.

Column description

Each row has the following unique identifier(s) that can be used to join/merge with other tables

* "grid\_df\_id": A unique integer identifying the grid cell.

The table contains the following columns as metadata:

* "LAT" and "LON": a unique latitude (LAT) and longitude (LON) corresponding to the unique grid cell (grid\_df\_id)
* "area\_km": The area of this grid cell in km2
* "is\_land": a logical vector indicating whether the grid cell is on land (True) or not (False)

## Table name: shp\_df\_natural-earth-shapes

Table description

This table contains the polygon information for the natural earth shapefiles. There are over 100 rows of metadata, so I will not go into all of them and just describe the most relevant.

Column description

Each row has the following unique identifier(s) that can be used to join/merge with other tables

* "shpfile\_id": A unique integer identifying the shapefile

The dataframe contains the following columns as metadata:

* 'area\_km2': the area in km2 of the shapefile

## Table name: shp\_grid\_df

Table description

A table containing the grid cells matching the coverage of each shapefile. If a shapefile did not have an extent and just a lat/lon, then the grid cell in which the centroid was closest to the lat/lon was matched.

Column description

Each row has the following unique identifier(s) that can be used to join/merge with other tables

* "shpfile\_id": A unique integer identifying the shapefile
* "grid\_df\_id": A unique integer identifying the grid cell.

## Table name: geoparsed-text\_shp\_df\_matches

Table description

This table contains information on which grid cells (grid\_df\_id) are covered by places (shp\_id) that were identified within documents (analysis\_id). These ids can be used to join with other tables to get more metadata if desired. For plotting, the most important metatdata variable is "cell\_weight" which is described in the introduction above.

Column description

Each row has the following identifier(s) that can be used to join/merge with other tables:

* grid\_df\_id
* analysis\_id
* shp\_id

Each row contains the following metadata:

* place: the place name
* cell\_weight: the proportional amount the place identified in the study contributes to that grid cell

## Table name: geoparsed-text\_grid-sums

Table description

Using the previous table (geoparsed-text\_shp\_df\_matches), this table has calculated the number of articles per unique grid cell (grid\_df\_id), both unweighted (n\_articles) and weighted (n\_articles\_weighted). Some metadata on the grid cell (LON, LAT, area\_km, is\_land) are also included for easy plotting

Column description

Each row has the following identifier(s) that can be used to join/merge with other tables:

* grid\_df\_id

Each row contains the following metadata:

* LAT, LON
* area\_km
* is\_land
* n\_articles
* n\_articles\_weighted