Received: 26 March 2018 | Accepted: 21 December 2018

DOI: 10.1111/2041-210X.13152

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases

Alexander Zizka[1,2,3] | Daniele Silvestro[1,2,4] | Tobias Andermann[1,2] | Josué Azevedo[1,2] | Camila Duarte Ritter[1,2,5] | Daniel Edler[1,2,6] | Harith Farooq[1,2,7,8] | Andrei Herdean[1] | María Ariza[9] | Ruud Scharn[2,10] | Sten Svantesson[1] | Niklas Wengström[1] | Vera Zizka[11] | Alexandre Antonelli[1,2,12]

[1]Department of Biological and Environmental Sciences, University of Gothenburg, Göteborg, Sweden; [2]Gothenburg Global Biodiversity Centre, Göteborg, Sweden; [3]German Center for Integrative Biodiversity Research (iDiv), Leipzig, Germany; [4]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland; [5]Department of Eukaryotic Microbiology, University of Duisburg-Essen, Essen, Germany; [6]Integrated Science Lab, Department of Physics, Umeå University, Umeå, Sweden; [7]Departamento de Biologia & CESAM, Universidade de Aveiro, Aveiro, Umeå, Portugal; [8]Faculty of Natural Sciences at Lúrio University, Universidade de Aveiro, Pemba, Mozambique; [9]Natural History Museum, University of Oslo, Oslo, Norway; [10]Department of Earth Sciences, University of Gothenburg, Göteborg, Sweden; [11]Faculty of Biology, University Duisburg-Essen, Essen, Germany and [12]Gothenburg Botanical Garden, Göteborg, Sweden

Correspondence
Alexander Zizka
Email: alexander.zizka@idiv.de

Funding information
A.A. and A.Z. are supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024 to A.A.). DS received funding from the Swedish Research Council (2015-04748). A.A. is further supported by the Swedish Research Council, the Swedish Foundation for Strategic Research, a Wallenberg Academy Fellowship, the Faculty of Sciences at the University of Gothenburg, and the David Rockefeller Center for Latin American Studies at Harvard University. C.D.R. is financed by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil: 249064/2013-8).

rOpenSci Resources
The software package [CoordinateCleaner], developed as part of this research effort, was extensively reviewed and approved by the rOpenSci project (https://ropensci.org). A full record of the review is available at: [https://github.com/ropensci/CoordinateCleaner]

Handling Editor: Tiago Quental

Abstract

1. Species occurrence records from online databases are an indispensable resource in ecological, biogeographical and palaeontological research. However, issues with data quality, especially incorrect geo-referencing or dating, can diminish their usefulness. Manual cleaning is time-consuming, error prone, difficult to reproduce and limited to known geographical areas and taxonomic groups, making it impractical for datasets with thousands or millions of records.
2. Here, we present CoordinateCleaner, an R-package to scan datasets of species occurrence records for geo-referencing and dating imprecisions and data entry errors in a standardized and reproducible way. CoordinateCleaner is tailored to problems common in biological and palaeontological databases and can handle datasets with millions of records. The software includes (a) functions to flag potentially problematic coordinate records based on geographical gazetteers, (b) a global database of 9,691 geo-referenced biodiversity institutions to identify records that are likely from horticulture or captivity, (c) novel algorithms to identify datasets with rasterized data, conversion errors and strong decimal rounding and (d) spatio-temporal tests for fossils.
3. We describe the individual functions available in CoordinateCleaner and demonstrate them on more than 90 million occurrences of flowering plants from the Global Biodiversity Information Facility (GBIF) and 19,000 fossil occurrences from the Palaeobiology Database (PBDB). We find that in GBIF more than 3.4 million records (3.7%) are potentially problematic and that 179 of the tested contributing

datasets (18.5%) might be biased by rasterized coordinates. In PBDB, 1205 records (6.3%) are potentially problematic.
4. All cleaning functions and the biodiversity institution database are open-source and available within the CoordinateCleaner R-package.

KEYWORDS
biodiversity institutions, data quality, fossils, GBIF, geo-referencing, palaeobiology database (PBDB), R package, species distribution modelling

1 | INTRODUCTION

The digitalization of biological and palaeontological collections from museums and herbaria is rapidly increasing the public availability of species' geographical distribution records. To date, more than 1 billion geo-referenced occurrence records are freely available from online databases, such as the Global Biodiversity Information Facility (GBIF, www.gbif.org), BirdLife International (www.birdlife.org) or other taxonomically, temporally, or spatially more focused databases (e.g. http://www.paleobiodb.org, http://bien.nceas.ucsb.edu/bien). Together, these resources have become widely used in ecological, biogeographical and palaeontological research and have greatly facilitated our understanding of biodiversity patterns and processes (e.g. Díaz et al., 2016; Zanne et al., 2014).

Most biodiversity databases are composed of, or provide access to, a variety of sources. Hence, they integrate data of varying quality, often compiled and curated at different times and places. Unfortunately, the available meta-data, for example on the nature of the records (museum specimen, survey, citizen science observation), the collection method (GPS record, grid cell from an atlas project) and collection-time, varies and often meta-data are missing. As a consequence, data quality in online databases is a major concern, and has limited their utility and reliability for research and conservation (Anderson et al., 2016; Chapman, 2005; Gratton et al., 2017; Yesson et al., 2007).

In the case of species occurrence records for extant taxa, problems with the geographical location constitute a major concern. In particular, erroneous or overly imprecise geographical coordinates can bias biodiversity patterns at multiple spatial scales (Maldonado et al., 2015). Common problems include (a) occurrence records assigned to country or province centroids due to automated geo-referencing from vague locality description, (b) records with switched latitude and longitude, (c) zero coordinates due to data entry errors, (d) records from zoos, botanical gardens or museums, (e) records based on rasterized collections and (f) records that have been subject to strong decimal rounding (Table 1, Gueta & Carmel, 2016; Maldonado et al., 2015; Robertson, Visser, and Hui, 2016; Yesson et al., 2007). Records affected by these issues can cause severe bias depending on the research question and the geographical scale of analyses (Graham et al., 2008; Gueta & Carmel, 2016; Johnson & Gillingham, 2008).

In addition to spatial issues, the temporal information (i.e. the year of collection) associated with occurrence records can be erroneous. In the case of fossil occurrences, temporal information includes the age of the specimen typically defined by the stratigraphic range of the sampling locality. Although sampling biases (and their temporal and spatial heterogeneity) are arguably the most severe issue in the analysis of the fossil record (Foote, 2000; Xing et al., 2016), overly imprecise or erroneous fossil ages, data entry errors or taxonomic uncertainties can negatively affect the reliability of the analysis (Varela, Lobo, & Hortal, 2011). While large-scale analyses of the fossil record appear resilient to error in the data (Adrain & Westrop, 2000; Sepkoski, 1993), the inclusion of erroneous data is likely to generate non-negligible biases at smaller temporal and taxonomic scales.

Manual cleaning is possible, but time-consuming and limited to the taxonomic and geographical expertise of individual researchers. It is thus generally not feasible for datasets that comprise thousands or millions of occurrence records. Furthermore, manual cleaning — often based on poorly documented and thus irreproducible ad hoc decisions — can add subjectivity and, in the worst case, bias. These issues call for standardized data validation and cleaning tools for large-scale biodiversity data (Gueta & Carmel, 2016).

2 | DESCRIPTION

Here, we present CoordinateCleaner, a new software package for standardized, reproducible and fast identification of potential geographical and temporal errors in databases of recent and fossil species occurrences. CoordinateCleaner is implemented in R (R Core Team, 2018) based on standard tools for data handling and spatial statistics (Allaire et al., 2018; Arel-Bundock, 2018; Becker, Wilks, Brownrigg, Minka, Deckmyn, 2017; Bivand & Lewin-Koh, 2017; Bivand & Rundel, 2018; Chamberlain, 2017; Hester, 2017; Hijmans, 2017a,b; Pebesma & Bivand, 2005; Varela, Gonzalez Hernandez, & Fabris Sgarbi, 2016; Wickham, 2011, 2016; Wickham, Danenberg, & Eugster, 2017; Wickham & Hesselberth, 2018; Wickham, Hester, & Chang, 2018; Xie, 2018). See the online documentation available at https://ropensci.github.io/CoordinateCleaner for an in-depth description of methods and simulations. The main features of the package are listed below.

2.1 | Automatic tests for suspicious geographical coordinates or temporal information

CoordinateCleaner compares the coordinates of occurrence records to reference databases of country and province centroids, country capitals, urban areas, known natural ranges and tests for plain zeros, equal

longitude/latitude, coordinates at sea, country borders and outliers in collection year. The reference databases are compiled from several sources (Central Intelligence Agency, 2014; South, 2017, and www.naturalearthdata.com/). All functions available in CoordinateCleaner are summarized in Table 1 and each of them can be customized with flexible parameters and individual reference databases.

2.2 | A global database of biodiversity institutions

A common problem are occurrence records matching the location of biodiversity institutions, such as zoological and botanical gardens, museums, herbaria or universities. These can have various origins: records from living individuals in captivity or horticulture, individuals that have escaped horticulture near the institution, or specimens without collection coordinates that have been erroneously geo-referenced to their physical location (e.g. a museum). To address these problems we compiled a global reference database of 9,691 biodiversity institutions from multiple sources (Botanic Gardens Conservation International, 2017; GeoNames, 2017; Global Biodiveristy Information Facility, 2017; Index Herbariorum, 2017; The Global Registry of Biodiversity Repositories, 2017; Wikipedia, 2017) and geo-referenced them using the ggmap and opencage R-packages (Kahle & Wickham, 2013; Salmon, 2017). Where automatic geo-referencing failed (c. 50% of the entries), we geo-referenced manually using Google Earth Pro (Google Inc, 2017) or information from the institutions web-pages, if available. We acknowledge that this database might not be complete, and have set up a website at http://biodiversity-institutions.surge.sh/ where scientists can explore the database and submit additions or corrections. See https://ropensci.github.io/CoordinateCleaner/articles/Background_the_institutions_database.html for a detailed description of the database.

TABLE 1 Geographical and temporal tests implemented in the CoordinateCleaner package

Test function	Level	Flags	Main error source	GBIF (%)	PBDB (%)
cc_cap	REC	Radius around country capitals	Imprecise geo-referencing based on vague locality description	1.1	–
cc_cen	REC	Radius around country and province centroids	Imprecise geo-referencing based on vague locality description	1.8	1
cc_coun	REC	Records outside indicated country borders	Various, e.g. swapped latitude and longitude	–	–
cc_dupl	REC	Records from one species with identical coordinates	Various, e.g. duplicates from various institutions, records from genetic sequencing data	–	–
cc_equ	REC	Records with identical lon/lat	Data entry errors	1.6	1
cc_gbif	REC	Radius around the GBIF headquarters in Copenhagen	Data entry errors, erroneous geo-referencing	0	0
cc_inst	REC	Radius around biodiversity institutions	Cultivated/captured individuals, data entry errors	0.8	0
cc_iucn	REC	Records outside external range polygon	Naturalized individuals, data entry errors	–	–
cc_outl	REC	Geographically isolated records of a species	Various, e.g. swapped latitude and longitude	–	–
cc_sea	REC	Records located within oceans	Various, e.g. swapped latitude and longitude	0.1	–
cc_urb	REC	Records from within urban areas	Cultivated individuals, old records	–	–
cc_val	REC	Records outside lat/lon coordinate system	Data entry errors, e.g. wrong decimal delimiter	0	0
cc_zero	REC	Plain zeros in the coordinates and a radius around (0/0)	Data entry errors, failed geo-referencing	1.6	0.01
cd_ddmm	DS	Over proportional drop of records at 0.6	Erroneous conversion from dd.mm to dd.dd	4.1% datasets	–
cd_round	DS	Decimal periodicity or over proportional number of zero decimals	Rasterized or rounded data	18.5% datasets	–
cf_age	FOS/REC	Temporal outliers in fossil age or collection year	Various	–	–
cf_equal	FOS	General time validity	Data entry errors	–	0
cf_range	FOS	Overly imprecise age ranges	Lack of data	–	3.3
cf_outl	FOS	Outliers in space-time	Data entry error	–	2.1

REC, record-level; DS, dataset-level; FOS, fossil-level; dd.mm, degree minute annotation; dd.dd, decimal degree annotation; GBIF, Global Biodiversity Information Facility; PBDB, Paleobiology Database.

2.3 | Algorithms to identify conversion errors and rasterized data

Two types of potential bias are unidentifiable on record-level if the relevant meta-data are missing: (A) coordinate conversion errors based on the misinterpretation of the degree sign (°) as decimal delimiter and (B) occurrence records derived from rasterized collection designs or subjected to strong decimal rounding (e.g. presence/absence in 100 × 100 km grid cells). This may be particularly problematic for studies with small geographical scale, which need high precision, and if the erroneous records have been combined with precise GPS-records into datasets of mixed precision. CoordinateCleaner implements two novel algorithms to identify these problems on a dataset-level (a dataset in this context can either be all available records or subsets thereof, for instance from different contributing institutions). The tests assume that datasets with a sufficient number of biased records show a characteristic periodicity in the statistical distribution of their coordinates or coordinate decimals.

To detect coordinate conversion bias (A), we use a binomial test together with the expectation of a random distribution of the coordinate decimals in the dataset (implemented in the `cd_ddmm` function). If we consider a dataset of coordinates spanning several degrees of latitude and longitude, we can expect the distribution of decimals to be roughly uniform in range [0, 1). In the case of a conversion error, the coordinate decimal cannot be above 0.59 (because one degree only has 59 min). Thus, conversion errors tend to inflate the frequency of coordinates with decimals <0.6. We use two tests to identify this bias. First, we use the fraction of coordinate decimals below 0.6 to fit a binomial distribution with parameter $q = 0.59^2$ (which assumes uniformly distributed decimals). This yields estimates of (a) a *p*-value accepting or rejecting the hypothesis of a uniform distribution and (b) the parameter $\hat{q}$, which best explains the empirical distribution of decimals below and above 0.6. The first test is therefore given by the *p*-value that can be used to reject the hypothesis of a uniform distribution when smaller than a given threshold. The second test is based on the relative difference $(r = (\hat{q} - q)/q)$ between the estimated frequency of decimals below 0.6 ($\hat{q}$) and the expected one (q). Thus any $r > 0$ indicates a higher-than-expected frequency of decimals smaller than 0.6. We flag a dataset as biased, if the *p*-value is smaller than a user-defined threshold (by default set to 0.025) and *r* is larger than a user-defined threshold (by default set to 1).

To detect rasterized sampling bias (B), we test for the regular pattern in the sample coordinates caused by a rasterized sampling (or strong decimal rounding). This test involves three steps, which are implemented in a single function (`cd_round`). First, the algorithm amplifies the pattern by binning the coordinates and then calculates the autocorrelation among the number of records per bin as the covariance of two consecutive sliding windows. This step generates a vector **x** of autocorrelation values.

Second, we identify outliers of high autocorrelation within **x**, which we interpret as points of high sampling frequency, that is the nodes of the sampling raster. Using a second sliding-window **x** of size 10, where $\mathbf{x}_k = \{x_k, x_{k+1}, \ldots, x_{k+9}\}$, we flag a point x_{k+i} as highly autocorrelated when

$$x_{k+i} > Q_{75}(\mathbf{x}_k) + I_{25}^{75}(\mathbf{x}_k) \times T$$

where Q_{75} is the 75% quantile of $\mathbf{x}_k$, I_{25}^{75} is its interquartile range, and *T* is a user-set multiplier defining the test sensitivity. Third, we compute the distance (in degrees) between all flagged outliers and identify *D* as the most common distance. A dataset is then flagged as potentially biased if *D* is within a user-defined range (by default between 0.1 and 2 degrees) and the number of outliers spaced by a distance *D* exceeds a user-defined value (by default set to 3).

We optimized all default settings based on simulations to obtain high sensitivity for datasets of variable size and geographical scale. The `cd_ddmm` and `cd_round` functions succeeded to identify bias A) and bias B) in simulated datasets with more than 100 records and more than 50 individual sampling locations data respectively (https://ropensci.github.io/CoordinateCleaner/articles/Background_dataset_level_cleaning.html). Both functions include optional visual diagnostic output to evaluate the results for flagged datasets, which we recommend to guide a final decision, especially for dataset with few records, or geographically restricted extent.

2.4 | Spatio-temporal tests for fossil data

Problems with inaccurate or overly imprecise temporal information are exacerbated in fossils. In particular, insufficient data, taxonomic misidentification, homonyms (names with same spelling but referring to different taxa) and data entry errors can cause very imprecise or wrong ages. CoordinateCleaner includes functions to identify fossils with (a) an unexpectedly large age range ($r = a_{max} - a_{min}$), (b) an unexpected age, and (c) an unexpected location in space-time in a given dataset. To identify (a) and (b) we use an interquartile-based outlier test implemented in the `cf_range` function, so that a fossil *i* in a dataset is flagged if

$$r_i > Q_{75}(\mathbf{r}) + I_{25}^{75}(\mathbf{r}) \times M$$

where $Q_{75}(\mathbf{r})$ is the 75 quartile age range (a) or age (b) across all records in the set, $I_{25}^{75}(\mathbf{r})$ is the interquartile range of **r** and *M* is a user-defined sensitivity threshold (by default set to 5).

To identify C) we test for outliers in a linear combination of range standardized geographical and temporal distances, based on a random sampling between minimum and maximum ages implemented in the `cf_outl` function. We calculate for each fossil *i* the mean scaled temporal and spatial distances to all other records in the set, t_i and s_i respectively. To compare temporal and spatial distances, which are otherwise expressed in different units (Myr and km), we rescale the temporal distances to the range of spatial distances. We use the sum of mean scaled distances ($t_i + s_i$) to identify temporal and spatial outliers, based on interquantile ranges as above:

$$t_i + s_i > Q_{75}(\mathbf{t}+\mathbf{s}) + I_{25}^{75}(\mathbf{t}+\mathbf{s}) \times Q$$

where and *Q* is a user-set sensitivity threshold (five by default). The test is replicated *n* times, where each replicate uses a randomly sampled age within the age range of *i*. Records are flagged if they have been identified as outlier in a fraction of *k* replicates, where *n* and *k* user-defined parameters (by default set to 5 and 0.5 respectively). The `cf_range` and `cf_outl` function can identify outliers across entire datasets or on a per-taxon base.

3 | RUNNING CoordinateCleaner

CoordinateCleaner includes three wrapper functions: `clean_coordinates`, `clean_dataset` and `clean_fossils` which combine a set of tests suitable for the respective data. `clean_coordinates` is the main function and creates an object of the S3-class 'spatialvalid', which has a summary and plotting method. Flagged occurrence records can easily be identified, checked or removed before further analyses. We provide two tutorials demonstrating how to use CoordinateCleaner on recent and fossil datasets and multiple short examples on the package at https://ropensci.github.io/CoordinateCleaner/. A reproducible minimal example is:

```
library(CoordinateCleaner)

exmpl <- data.frame(species = letters[1:10],

                    dataset = c("test1", "test2"),

                    decimallongitude = runif(250) * 180,

                    decimallatitude = runif(250) * 90)

flags <- clean_coordinates(exmpl) #record-level tests

summary(flags)

plot(flags)

flags.ds <- clean_dataset(exmpl) #data set-level tests

flags.ds
```

Alternatively, eah cleaning function can be called individually, for instance in pipelines based on the magrittr pipe (`%>%`).

4 | EMPIRICAL EXAMPLE

We demonstrate CoordinateCleaner on occurrence records for flowering plants available from GBIF (*c*. 91 million geo-referenced records; Global Biodiversity Information Facility, 2017, accessed 02 Feb 2017) and the Palaeobiology Database (PBDB, *c*. 19,000 records; PBDB, 2018 accessed 26 Jan 2018). We chose GBIF and PBDB as examples because they are large and widely used providers of biodiversity data. We stress that both platforms put substantial efforts in identifying problematic records and acquiring meta-data to increase data quality, and that we consider their data as having generally high quality and improving. We ran the `clean_coordinates`, `clean_fossils` and `clean_dataset` wrapper functions with all tests recommended in our tutorials, except those that are dependent on downstream analyses (Table 1). We used a custom gazetteer with a 1-degree buffer for `cc_sea`, to avoid flagging records close to the coastline (available in the package with `data('buffland')`). For computational efficiency, we divided the GBIF data into subsets of 200K records.

`clean_coordinates` flagged more than 3,340,000 GBIF records (3.6%), the majority due to coordinates matching country centroids, zero coordinates and equal latitude and longitude (Table 1). Figure 1a shows the number of occurrence records flagged per 100 × 100 km grid-cell, globally. Concerning the fossil data from PBDB, `clean_fossils` flagged 1,205 records (6.3%), mostly due to large uncertainty in dating and unexpected old age or distant location. These flags might include records where a precise dating was not possible, records with low taxonomic resolution, homonyms or problems during data entry. Figure 1b shows the number of fossil records flagged per 100 × 100 km grid-cell, globally.

On the dataset-level, we retrieved 2,494 individual datasets of flowering plants from GBIF, mostly representing data from different publishers (e.g. collections of specific museums). These datasets varied considerably in the number of records (from 1 record to 16 million) and geographical extent (<1 degree to global). We limited the tests to 641 datasets with at least 50 individual sampling locations to test for bias in decimal conversion (function `cd_ddmm`, Table 1) and 966 datasets with more than 100 occurrence records for the rasterization bias (function `cd_round`, Table 1). `clean_dataset` flagged 26 (4.1%) datasets as biased towards decimals below 0.6 (potentially related to ddmm to dd.dd conversion) and 179 datasets (18.5%) with a signature of decimal periodicity (potential rounding or rasterization). The high percentage of datasets with biased decimals was surprising and these might include datasets with clustered sampling. Since the value of such data for biological research is strongly dependent on follow up analyses we recommend to use a case-by-case judgement based on the desired precision, diagnostic plots and meta-data for a final decision on the flagged datasets. In general, not all flagged records and datasets are necessarily erroneous: our tests only indicate deviations from common and explicit assumptions. Flagged data may require further validation by researchers or exclusion from subsequent analyses.

5 | COMPARISON TO OTHER SOFTWARE

To our knowledge, few other tools exist for standardized data cleaning, namely the scrubr (Chamberlain, 2016) and biogeo (Robertson et al., 2016) R packages. Additionally, the modestR package (García-Roselló et al., 2013) implements a graphical user interface and includes cleaning of GBIF data based on habitat suitability. Some of the basic functions performed by CoordinateCleaner overlap with

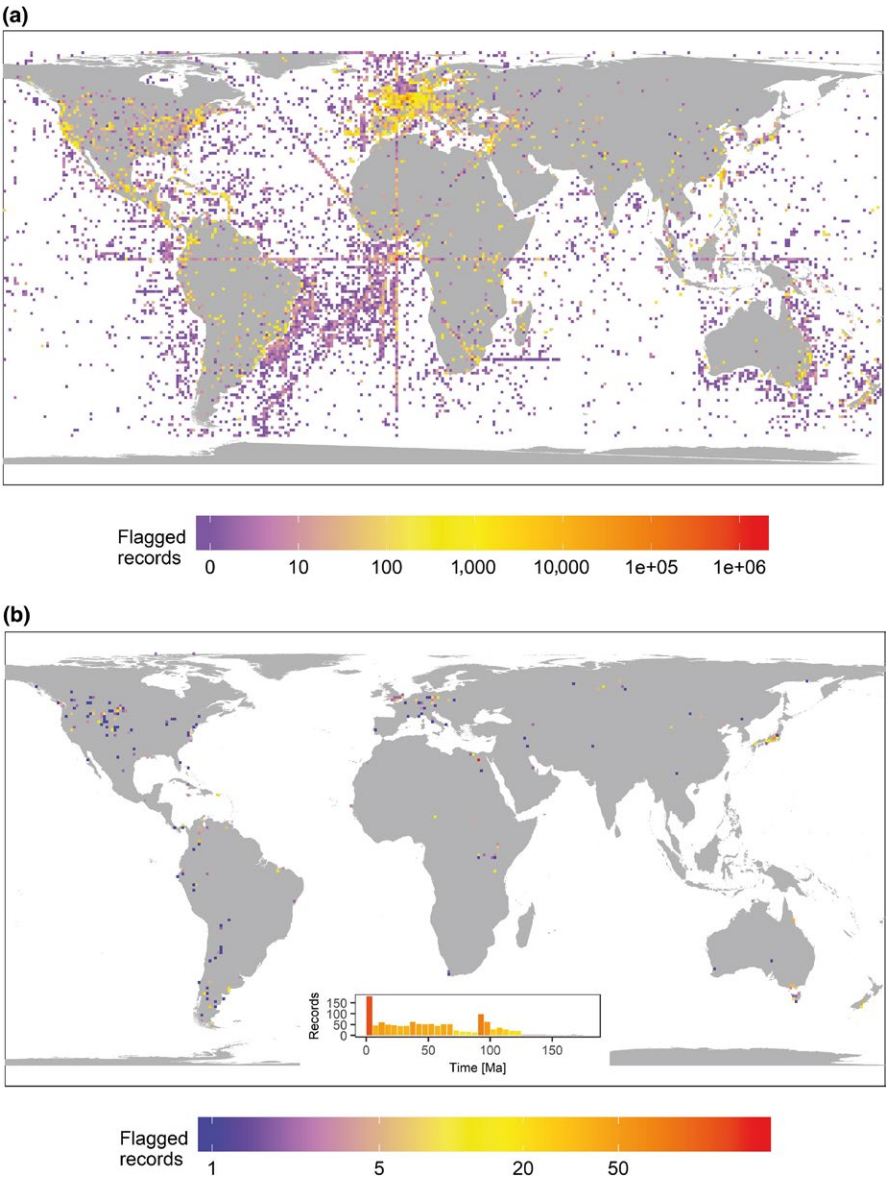

FIGURE 1 The number of species occurrence records flagged by CoordinateCleaner in empirical datasets, per 100 × 100 km grid cell. Warmer colours indicate more flagged records. (a) Flowering plants from the Global Biodiversity Information Facility (*c.* 91M; Global Biodiversity Information Facility, 2017) (b) Angiosperm fossils from PBDB (*c.* 19,000; PBDB, 2018). Note the logarithmic scale

these packages, however, CoordinateCleaner provides a substantially more comprehensive set of options, including novel tests and data (see https://ropensci.github.io/CoordinateCleaner/articles/Background_comparison_other_software for a function-by-function comparison of CoordinateCleaner, scrubr and biogeo).

Primarily, CoordinateCleaner adds the following novelties as compared to available packages: (a) A unique set of tests for problematic geographical coordinates, tailored to common but often overlooked problems in biological databases and not restricted to specific organisms, (b) A global, geo-referenced database of biodiversity institutions, to identify records from cultivation, zoos, museums, etc., (c) Novel algorithms to identify problems not identifiable on record-level, for example errors from the conversion of the coordinate annotation or low coordinate precision due to rasterized data collection, (d) Tests tailored to fossils, accounting for problems in dating and (e) Applicability to large datasets. These features in combination with their user-friendly implementation and extensive documentation and tutorials, will render CoordinateCleaner a useful tool for research in biogeography, palaeontology, ecology and conservation.

In general, no hard rule exists to judge data quality for biogeographical analyses – what is 'good data' depends largely on downstream analyses. For instance, continent-level precision might suffice for ancestral range estimation in some global studies, whereas species distribution models based on environmental data can require a 1-km precision. The objective of CoordinateCleaner is to automate the identification of problematic records as far as possible for all scales, with default values tailored to large datasets with millions of records and thousands of species. Nevertheless, some researcher judgement will always be necessary to choose suitable tests, specify appropriate thresholds, and avoid adding bias by cleaning. In the worst case, automatic cleaning could bias downstream analyses by information loss caused by overly strict filtering, exacerbating sampling bias by false outlier removal, and over-confidence in the cleaned data. In most cases, however, CoordinateCleaner speeds up the identification of

problematic records and common problems in a datasets for further verification. In some cases, disregarding flagged records might be warranted, but we recommend to carefully judge, and verify flagged records when possible, especially for the outlier and dataset-level tests. We provide an extensive documentation to guide cleaning and output interpretation (https://ropensci.github.io/CoordinateCleaner).

ACKNOWLEDGEMENTS

We thank all GBIF and PDBD administrators and contributors for their excellent work. We thank Sara Varela, Carsten Meyer and an anonymous reviewer for helpful comments on an earlier version of the manuscript, and rOpenSci, Maëlle Salmon, Irene Steves and Francisco Rodriguez-Sanchez for helpful comments on the R-code, as well as Juan D Carrillo for valuable feedback on the tutorial for cleaning fossil records.

AUTHORS' CONTRIBUTIONS

A.Z. developed the tools and designed this study. D.S. and A.Z. designed and implemented the dataset-level cleaning algorithms. D.E. developed the website for contributing to the biodiversity institutions database. A.Z., T.A., J.A., C.D.R., H.F., A.H., M.A., R.S., S.t.S., N.W. and V.Z. contributed data to the biodiversity institutions database. A.Z. wrote the manuscript, with contributions from A.A., D.S., T.A., J.A., D.E., H.F. and V.Z. All authors read and approved the final version of the manuscript.

DATA ACCESSIBILITY

The code of CoordinateCleaner is open source and has been reviewed by rOpenSci. The package is available as R-package from the CRAN repository (stable, https://cran.rstudio.com/web/packages/CoordinateCleaner/index.html) and GitHub (developmental, https://github.com/ropensci/CoordinateCleaner). The biodiversity institutions database is part of the package under a CC-BY license. Cleaning pipelines for occurrence records from GBIF and fossils from PBDB are available from https://ropensci.github.io/CoordinateCleaner, (https://doi.org/10.5281/zenodo.2539408) and from CRAN as part of the package.

ORCID

Alexander Zizka https://orcid.org/0000-0002-1680-9192
Daniele Silvestro https://orcid.org/0000-0003-0100-0961
Andrei Herdean https://orcid.org/0000-0003-2143-0213

REFERENCES

Adrain, J. M., & Westrop, S. R. (2000). An empirical assessment of taxic paleobiology. *Science*, *289*(5476), 110–112. https://doi.org/10.1126/science.289.5476.110

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., … Chang, W. (2018). rmarkdown: Dynamic Documents for R. Retrieved from https://cran.r-project.org/package=rmarkdown

Anderson, R. P., Araújo, M., Guisan, A., Lobo, J. M., Martínez-Meyer, E., Peterson, T., & Soberón, J. (2016). *Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling - Are species occurrence data in global online repositories fit for modeling species distributions? The case of the Global Biodiversity Information Facility (GBIF)*.

Arel-Bundock, V. (2018). countrycode: Convert country names and country codes. Retrieved from https://cran.r-project.org/package=countrycode

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., Deckmyn, A. (2017). maps: Draw geographical maps. Retrieved from https://cran.r-project.org/package=maps

Botanic Gardens Conservation International. (2017). BGCI - List of Botanic Gardens. Retrieved 20 November 2017, from https://www.bgci.org/

Bivand, R., & Lewin-Koh, N. (2017). maptools: Tools for reading and handling spatial objects. Retrieved from https://cran.r-project.org/package=maptools

Bivand, R., & Rundel, C. (2018). rgeos: Interface to geometry engine - open source ('GEOS'). Retrieved from https://cran.r-project.org/package=rgeos

Central Intelligence Agency. (2014). The world Factbook. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/

Chamberlain, S. (2016). scrubr: Clean biological occurrence records. Retrieved from https://cran.r-project.org/package=scrubr

Chamberlain, S. (2017). rgbif: Interface to the global 'Biodiversity' information facility API. Retrieved from https://cran.r-project.org/package=rgbif

Chapman, A. D. (2005). *Principles and methods of data cleaning - primary species and species occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen*, Copenhagen.

Díaz, S., Kattge, J., Cornelissen, J. H. C., Wright, I. J., Lavorel, S., Dray, S., … Gorné, L. D. (2016). The global spectrum of plant form and function. *Nature*, *529*(7585), 167–171. https://doi.org/10.1038/nature16489

Foote, M. (2000). Origination and extinction components of taxonomic diversity: general problems. *Paleobiology*, *26*(4), 74–102. https://doi.org/10.1666/0094-8373(2000)26[74:OAECOT]2.0.CO;2

García-Roselló, E., Guisande, C., Gonz, J., Heine, J., Pelayo-Villamil, P., Manjarrés Hernandéz, A., … Granado-Lorencio, C. (2013). ModestR: A software tool for managing and analyzing species distribution map databases. *Ecography*, *36*, 1202–1207. https://doi.org/10.1111/j.1600-0587.2013.00374.x

GeoNames. (2017). www.geonames.org.

Global Biodiveristy Information Facility. (2017). List of data publishers. Retrieved from www.gbif.org/publisher/search

Global Biodiversity Information Facility. (2017). Magnoliopsida. https://doi.org/10.15468/dl.wquvxb

Google Inc. (2017). Google Earth Pro, 7.1.7.2606.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Peterson, A. T., Loiselle, B. A., & The Nceas Predicting Species Distribution Working Group (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, *45*, 239–247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E., & Kühl, H. (2017). A world of sequences: Can we use georeferenced nucleotide databases for a robust automated phylogeography? *Journal of Biogeography*, *44*(2), 475–486. https://doi.org/10.1111/jbi.12786

Gueta, T., & Carmel, Y. (2016). Ecological Informatics Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological Informatics*, *34*, 139–145. https://doi.org/10.1016/j.ecoinf.2016.06.001

Hester, J. (2017). covr: Test coverage for packages. Retrieved from https://cran.r-project.org/package=covr

Hijmans, R. J. (2017a). geosphere: Spherical trigonometry. Retrieved from https://cran.r-project.org/package=geosphere

Hijmans, R. J. (2017b). raster: Geographic data analysis and modeling. Retrieved from https://cran.r-project.org/package=raster

Index Herbariorum. (2017). http://sweetgum.nybg.org/science/ih/.

Johnson, C. J., & Gillingham, M. P. (2008). Sensitivity of species-distribution models to error, bias, and model design: An application to resource selection functions for woodland caribou. *Ecological Modelling*, *3*, 143–155. https://doi.org/10.1016/j.ecolmodel.2007.11.013

Kahle, D., & Wickham, H. (2013). GGMAP: Spatial visualization with ggplot2. *The R Journal*, *5*(1), 144–161. Retrieved from http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., … Antonelli, A. (2015). Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases. *Global Ecology and Biogeography*, *24*(8), 973–984. https://doi.org/10.1111/geb.12326

PBDB. (2018). The data were downloaded from the Paleobiology Database on 26 January 2018, using the group names magnoliophyta, magnoliopsida, angiospermae. Paleobiology database. Retrieved from www.paleobiodb.org

Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, *5*(2). Retrieved from https://cran.r-project.org/doc/Rnews/

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

Robertson, M. P., Visser, V., & Hui, C. (2016). Biogeo: An R package for assessing and improving data quality of occurrence record datasets. *Ecography (Cop)*, *39*, 394–401. https://doi.org/10.1111/ecog.02118

Salmon, M. (2017). OPENCAGE: Interface to the OpenCage API. Retrieved from https://cran.r-project.org/package=opencage

Sepkoski, J. J. (1993). Ten years in the library: New data confirm paleontological patterns. *Paleobiology*, *19*(1), 43–51. https://doi.org/10.1017/S0094837300012306

South, A. (2017). rnaturalearth: World map data from natural earth. Retrieved from https://github.com/ropenscilabs/rnaturalearth

The Global Registry of Biodiversity Repositories. (2017). www.grbio.org. Retrieved from www.grbio.org

Varela, S., Gonzalez Hernandez, J., & Fabris Sgarbi, L. (2016). paleobioDB: Download and Process Data from the Paleobiology Database. Retrieved from https://cran.r-project.org/package=paleobioDB

Varela, S., Lobo, J. M., & Hortal, J. (2011). Using species distribution models in paleobiogeography: A matter of data, predictors and concepts. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *310*(3–4), 451–463. https://doi.org/10.1016/j.palaeo.2011.07.021

Wickham, H. (2011). testthat: Get started with testing. *The R Journal*, *3*, 5–10. Retrieved from https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org/

Wickham, H., Danenberg, P., & Eugster, M. (2017). roxygen2: In-line documentation for R. Retrieved from https://cran.r-project.org/package=roxygen2

Wickham, H., & Hesselberth, J. (2018). pkgdown: Make static HTML documentation for a Package. Retrieved from https://cran.r-project.org/package=pkgdown

Wickham, H., Hester, J., & Chang, W. (2018). devtools: Tools to make developing R packages easier. Retrieved from https://cran.r-project.org/package=devtools

Wikipedia. (2017). List of zoos by country. Retrieved 20 November 2016, from https://en.wikipedia.org/wiki/List_of_zoos_by_country

Xie, Y. (2018). knitr: A general-purpose package for dynamic report generation in R. Retrieved from https://yihui.name/knitr/

Xing, Y., Gandolfo, M. A., Onstein, R. E., Cantrill, D. J., Jacobs, B. F., Jordan, G. J., … Linder, H. P. (2016). Testing the biases in the rich cenozoic angiosperm macrofossil record. *International Journal of Plant Sciences*, *177*(4), 371–388. https://doi.org/10.1086/685388

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., … Culham, A. (2007). How global is the global biodiversity information facility? *PLoS ONE*, *11*, e1124. https://doi.org/10.1371/journal.pone.0001124

Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A., FitzJohn, R. G., … Beaulieu, J. M. (2014). Three keys to the radiation of angiosperms into freezing environments. *Nature*, *506*(7486), 89–92. https://doi.org/10.1038/nature12872

How to cite this article: Zizka A, Silvestro D, Andermann T, et al. COORDINATECLEANER: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol Evol*. 2019;10:744–751. https://doi.org/10.1111/2041-210X.13152