

fossilbrush

Flannery-Sutherland, Joseph T.; Raja, Nussaïbah B.; Kocsis, Ádám T.; Kiessling, Wolfgang

DOI:

[10.1111/2041-210X.13966](https://doi.org/10.1111/2041-210X.13966)

License:

Creative Commons: Attribution (CC BY)

Document Version

Publisher's PDF, also known as Version of record

Citation for published version (Harvard):

Flannery-Sutherland, JT, Raja, NB, Kocsis, ÁT & Kiessling, W 2022, 'fossilbrush: An R package for automated detection and resolution of anomalies in palaeontological occurrence data', *Methods in Ecology and Evolution*, vol. 13, no. 11, pp. 2404-2418. <https://doi.org/10.1111/2041-210X.13966>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.





Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

RESEARCH ARTICLE

fossilbrush: An R package for automated detection and resolution of anomalies in palaeontological occurrence data

Joseph T. Flannery-Sutherland¹  | Nussaïbah B. Raja²  | Ádám T. Kocsis²  |
Wolfgang Kiessling² 

¹School of Earth Sciences, University of Bristol, Bristol, UK

²GeoZentrum Nordbayern, Department of Geography and Geosciences, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

Correspondence

Joseph T. Flannery-Sutherland
Email: jf15558@bristol.ac.uk

Handling Editor: Dr. Samantha Price

Abstract

1. Fossil occurrence databases are indispensable resources to the palaeontological community, yet present unique data cleaning challenges. Many studies devote significant attention to cleaning fossil occurrence data prior to analysis, but such efforts are typically bespoke and difficult to reproduce. There are also no standardised methods to detect and resolve errors despite the development of an ecosystem of cleaning tools fuelled by the concurrent growth of neontological occurrence databases.
2. As fossil occurrence databases continue to increase in size, the demand for standardised, automated and reproducible methods to improve data quality will only grow. Here, we present semi-automated cleaning solutions to address these issues with a new R package *fossilbrush*. We apply our cleaning protocols to the Paleobiology Database to assess the prevalence of anomalous entries and the efficacy and impact of our methods.
3. We find that anomalies may be effectively resolved by comparison against a published compendium of stratigraphic ranges, improving the stratigraphic quality of the data, and through methods which detect outliers in taxon-wise occurrence stratigraphic distributions. Despite this, anomalous entries remain prevalent throughout major clades, with often more than 30% of genera in major fossil groups (e.g. bivalves, echinoderms) displaying stratigraphically suspect occurrence records.
4. Our methods provide a way to flag and resolve anomalous taxonomic data before downstream palaeobiological analysis and may also aid in the automation and targeting of future cleaning efforts. We stress, however, that our methods are semi-automated and are primarily for the detection of potential anomalies for further scrutiny, as full automation should not be a substitute for expert vetting. We note that some of our methods do not rely on external databases for anomaly resolution and so are also applicable to occurrences in neontological databases, expanding the utility of the *fossilbrush* R package.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

KEYWORDS

chronostratigraphy, data cleaning, fossil occurrence, palaeobiology database, Sepkoski Compendium, stratigraphic density

1 | INTRODUCTION

Biological occurrence databases record the present and historical spatial distributions of organisms across the tree of life, enabling assessment of the drivers of biodiversity across a wide variety of geographical and taxonomic scales (Zizka et al., 2019). These databases are critical to predicting the impacts of habitat change on species distribution and future ecosystems, particularly to inform conservation efforts (Bell-Damarow et al., 2019; Stephenson & Stengel, 2020). Such efforts are strengthened by incorporating organismal distribution records in the geological past from fossil occurrence databases (Dietl, 2019; Kiessling et al., 2019), including Triton (Fenton et al., 2021), FAUNMAP (Graham & Lundelius, 2010), PaleoReefs (Kiessling & Krause, 2022), Neotoma (Williams et al., 2018), the Paleobiology Database (PBDB; Alroy et al., 2008), FRED (Clowes et al., 2021) and NOW (The NOW Community, 2020). Alongside geographical and systematic data, fossil occurrence databases record the geological ages of their entries, permitting connection of palaeontological and neontological information in the recent past (Eduardo et al., 2018) and analysis of diversity dynamics, macroecological structuring and biogeography in deep time (Peters & McClennen, 2016).

The proliferation of occurrence databases has promoted development of an ecosystem of tools for database access and cleansing, typically as packages for the R programming environment (R Core Team, 2022). Errors in taxonomy and spatial coordinates inevitably accrue due to improper data imputation and changing taxonomic convention in the primary literature (Grenié et al., 2022). Tools for coordinate cleaning cover both fossil and modern occurrence databases (Zizka et al., 2019), but available tools for spell checking of taxonomic names and taxonomic harmonisation are usually limited to extant species (Norman et al., 2020; Ribeiro et al., 2022). Consequently, these are unsuited for resolution of fossil taxonomy and fossil data cleaning is typically done in isolation without support from standardised tools. As well as taxonomic and coordinate errors, fossil occurrence databases may contain erroneous geological ages arising from increasing refinements to chronostratigraphic time-scales, improved lithostratigraphic constraints for fossil-bearing formations, or typographical errors during data imputation. Finally, fossil occurrences present the unique problem where misidentification of an occurrence may induce an anomalous record in the stratigraphic range of a taxon.

The PBDB is one of the most well-known fossil occurrence databases, with over 1.5 million occurrences of >45,000 taxa. Its data underpin seminal works on the quality of the fossil record and the broadest patterns of biotic change throughout the Phanerozoic (Alroy et al., 2008), yet the presence of taxonomic, spatial and stratigraphic errors leads some authors to condemn the PBDB altogether

(Prothero, 2015). Manual cleaning of the entire database is rendered infeasible by its scale, and such an effort would quickly become outdated with addition of new data and changing taxonomic and stratigraphic conventions in the primary literature. New databases may also emerge in the future, for example, through application of automated data extraction and machine-learning methods to digitised publications (Kopperud et al., 2018; Peters et al., 2014) or from dark data in museum collections (Allmon et al., 2018; Marshall et al., 2018), and these may also contain inconsistencies necessitating revision. As such, cleaning of fossil occurrence data is an ever-present challenge requiring automated, standardised and reproducible solutions capable of scaling to large datasets, and which tackle the unique challenges posed by the stratigraphic component of the data.

Here, we present several data-cleaning considerations and tools to resolve misspellings and harmonise taxonomic schemes. These operate independently of any external databases and can be applied to any occurrence dataset that contains higher taxonomic information. Tools to address spatial coordinate errors already exist, but accurate geological ages are required for calculation of accurate palaeocoordinates and we provide methods to update chronostratigraphic ages for fossil occurrence datasets. Resolution of taxonomic and stratigraphic errors in isolation is predicated on expert knowledge. In cases where taxonomic misidentification results in stratigraphic range anomalies, however, error detection can be predicated on outlier detection, permitting development of parameterised, automated and reproducible techniques. We develop three methods to flag and resolve outliers in taxon stratigraphic ranges which we have made available within a new R package, *fossilbrush*. We apply our cleaning protocols to the PBDB to demonstrate how the methods can be applied in a standardised, reproducible fashion, then examine the efficacy of anomaly detection and their impact on the database.

2 | MATERIALS AND METHODS

2.1 | Resolving database inconsistency

2.1.1 | Chronostratigraphic harmonisation

Occurrence chronostratigraphy and analytical strategy should conform to the same chronostratigraphic time-scale. The Paleobiology Database currently uses International Chronostratigraphic Time Scale 2013 ages (Cohen, 2013), yet studies often use more recent chronostratigraphic schemes for data analysis, which may lead to erroneous division of occurrences into successive, temporally discrete partitions. For example, the base of the Triassic sits at 252.17 Ma in ICS 2013 compared to 251.9 Ma as of 2022. This is a pertinent example as it coincides with the end-Permian mass extinction where

well-resolved and correctly binned data are critical to determining extinction magnitude and timing. A Geologic Time Scale 2020 (Gradstein et al., 2020) is the most recently published comprehensive chronostratigraphic standard and is being adopted both analytically (e.g. Marshall et al., 2021; Metcalfe & Crowley, 2021) and in databasing efforts (e.g. Fenton et al., 2021). We provide a lookup table (sample in Table S1) and function `chrono_scale()`, enabling quick revision of chronostratigraphic dates in existing fossil occurrence datasets. While the lookup table will inevitably accrue errors as chronostratigraphic time-scales receive refinement, the majority of the dates are expected to remain valid, and a user can either update the lookup table in R or supply a new one entirely to `chrono_scale()` ensure that their data and analytical strategy conform to the latest chronostratigraphic standard.

2.1.2 | Taxonomic harmonisation

Spelling variations and inconsistent higher classifications risk incomplete representation of a taxon by its occurrences, leading to underestimation of spatiotemporal ranges and overestimation of diversity. We provide a modular taxonomic harmonisation workflow through the function `check_taxonomy()`, with options to report flagged issues or resolve them automatically. Our workflow is applicable to any occurrence dataset with hierarchically organised taxonomic information, including composite datasets compiled from multiple databases, and ensures that all classifications are internally consistent so that taxa are fully represented by all their occurrences at any given taxonomic level. Following previous recommendations (Grenié et al., 2022), our taxonomic harmonisation function initially standardises taxon name formatting as a prerequisite to improve detection of inconsistent spelling using fuzzy string methods (see Supplementary Information). Instances of inconsistent higher classifications can be displayed graphically for user inspection using our `plot_taxa()` function and automatically resolved within `check_taxonomy()` if desired.

2.2 | Stratigraphic outlier detection

2.2.1 | Independent database comparison

Here, we flag taxon and occurrence age ranges against a reference database with similar basic data, using notation (Figure 1) where 0 and 1 denote error versus validity for the FAD (left) and LAD (right) of a range (R): 1R1 = valid, 0R1 = FAD outside of reference, 1R0 = LAD outside of reference, 0R0 = FAD and LAD both exceed reference, 00R = older than reference, R00 = younger than reference, 000 = unrecorded in the reference database. Occurrence ages can be checked individually against the reference database, with the magnitude of age discrepancies helping to inform whether they are potentially anomalous. Stage to substage-level discrepancies may mark discoveries whose FADs and LADs genuinely supersede

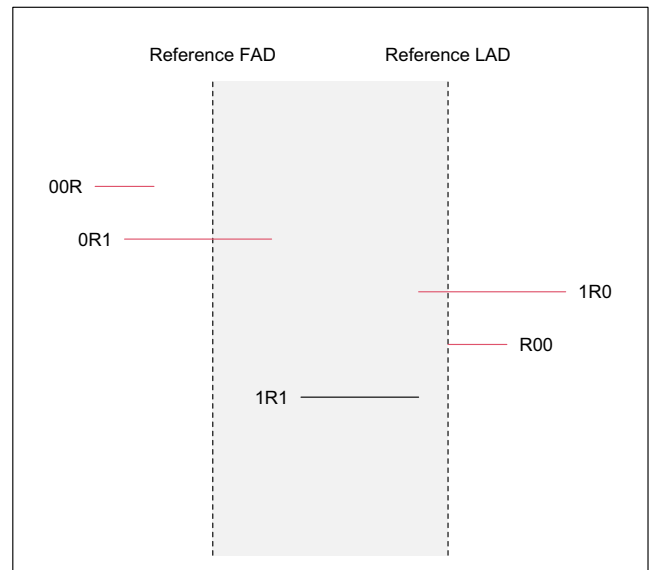


FIGURE 1 Schematic representation of the stratigraphic error flagging codes used in the *fossilbrush* R package. Dotted lines demarcate the reference range of a taxon and solid lines the age range of occurrences assigned to that taxon. Red lines indicate occurrences with stratigraphic age anomalies relative to the reference range, while black lines are occurrences with valid stratigraphic ranges. See text for the definitions of the error codes.

their reference range while those which overlap with their reference range (0R1, 0R0 and 1R0) have a partially plausible stratigraphic range and may reflect cases where dating imprecision for the parent stratigraphic unit results in erroneous range extension. On the other hand, occurrences which fall outside the reference range (00R and R00) are more likely to reflect taxonomic error, particularly if that error is substantial (epoch-level or greater). This is a common issue for wastebasket taxa where superficial taxonomic work can result in backward range extension of relatively recent taxa by hundreds of millions of years, for example the brachiopod *Lingula* or bivalve mollusc *Ostrea* (Plotnick & Wagner, 2006).

Comparison against a reference database is implemented in the `flag_ranges()` function, with the option to tag occurrence age discrepancies greater than a user-defined threshold to quickly distinguish between cases of stratigraphic versus taxonomic error. This is a useful exploratory procedure but is limited to taxa common to both databases. Additionally, occurrences in the PBDB are organised into collections, in idealised terms a discrete community from a single bedding plane (i.e. a precise point in space and time), but in reality a time-averaged assemblage bearing the stratigraphic uncertainty of its parent lithological unit. Fossil occurrences in the same collection must possess the same stratigraphic age range, but resolving occurrence ages individually may violate this property, so our flagging procedure instead uses the reference database to determine if a collection range age is plausible. The procedure retrieves the available reference ranges for any taxa in the collection, then determines if a threshold proportion of those ranges overlap. If the overlap falls below the threshold, then the entire collection is treated as bearing a high prevalence of stratigraphic and taxonomic anomalies. If

a common interval exceeding the threshold can be identified, the collection age should fall within this interval to be considered stratigraphically plausible and can be modified according to this robust consensus of calibrations. We implement this procedure in the function `revise_ranges()` with a default, arbitrary threshold of 75% overlap. While a more stringent threshold could be used, a more relaxed threshold better accommodates erroneous occurrences in some collections when searching for a plausible interval. As above, any discrepancies between original and revised ages of occurrences within a collection can then be examined to determine whether they represent plausible range extension, taxonomic error or stratigraphic error.

2.2.2 | Occurrence density distribution

As reference databases may still contain errors and are unavailable for many groups, we detect and resolve stratigraphic anomalies using the density of fossil occurrences through time. We treat fossil occurrence age uncertainties as uniform probability density distributions on their point-wise ages (Zhang et al., 2016). Combining the uniform density distributions for a set of taxon occurrences gives a composite density distribution of all possible pointwise observations of that taxon through time (Figure 2). At face value, a density distribution represents how palaeontologists have assigned stratigraphically varied fossil observations with associated age uncertainties to a single morphospecies. The probability of fossil observations through time and so the idealised shape of their density distribution is expected to follow biological principles. A taxon is expected to originate in a limited area, expand its range and population size, then decline until extinction, producing a bell curve (Figure 2) with symmetry dependent on the rapidity of expansion and extinction (Foote, 2007; Foote et al., 2007). This unimodal distribution may conceivably become multimodal in several ways. First, boom-bust cycles in population size may be detectable through intensive bed-wise sampling in finely resolved or recent sedimentary sections (Kidwell, 2015; Kidwell & Flessa, 1995), but we assume here that such signals are unlikely to manifest in temporally coarser macrofossil occurrence databases. Second, heterogeneous sampling through time may introduce peaks and troughs into an idealised observation record (Barido-Sottani et al., 2019). Third, density distributions for stratigraphically distinct taxa may be conflated into a single multimodal density record due to superficially similar morphologies (Figure 3a; an invalid morphospecies or wastebasket taxon) or through taxonomic misclassification of individual occurrences (Figure 3d). We distinguish between multimodality arising from sampling effects versus taxonomic error based on the assumption that, as sampling probability increases exponentially with time (Stadler, 2010), short durations between peaks are more likely to represent sampling variation (Figure 3b) while longer durations are more likely to represent conflation of distinct taxonomic records (Figure 3d), with average genus longevities within the clade to which a given taxon belongs helping to distinguish between these cases. Taxonomic misidentifications resulting in stratigraphic

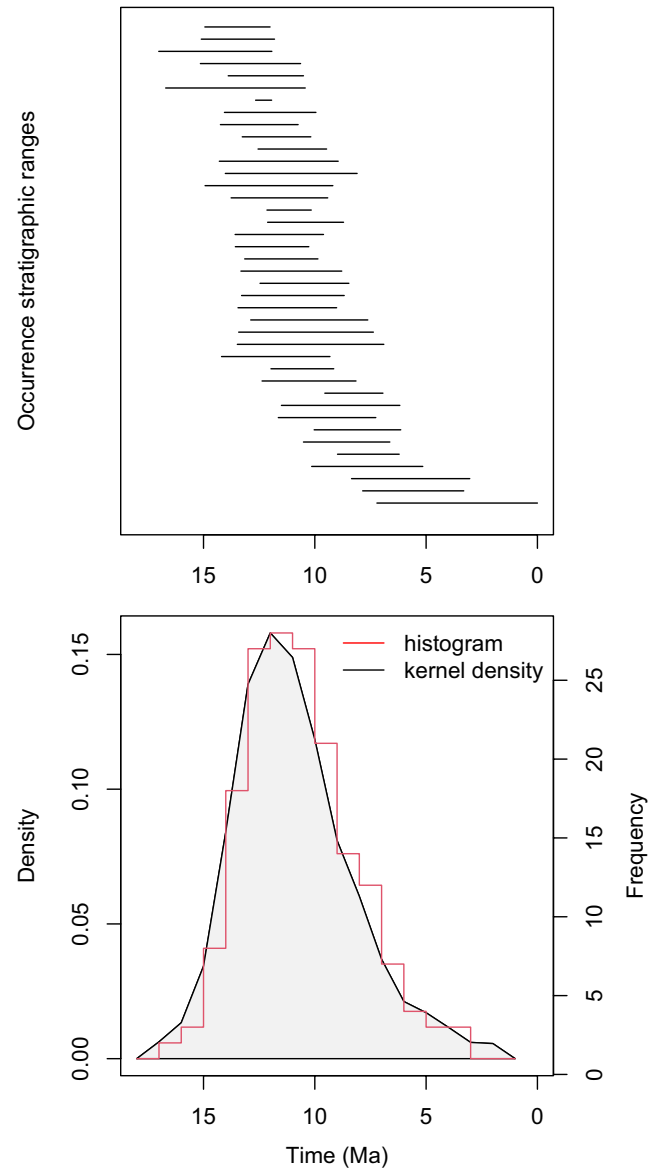


FIGURE 2 Construction of an occurrence stratigraphic density distribution from simulated data, displaying the idealised bell-curve expected under the biological sequence of origination, population expansion, population decline and extinction.

anomalies are expected to be concentrated towards the ends of a taxon's stratigraphic range, leading to long tails in its density distribution (Lazarus et al., 2012). Alternatively, long tails may arise from high stratigraphic uncertainty in a handful of occurrences. As such, the density distribution of macrofossil observations provides information on the stratigraphic plausibility and consistency of a morphospecies and so a means of detecting outliers.

2.2.3 | Interpeak thresholding

Interpeak thresholding ascertains whether interpeak durations in a multimodal occurrence density distribution more likely reflect incomplete sampling versus taxonomic error. As short overlaps in

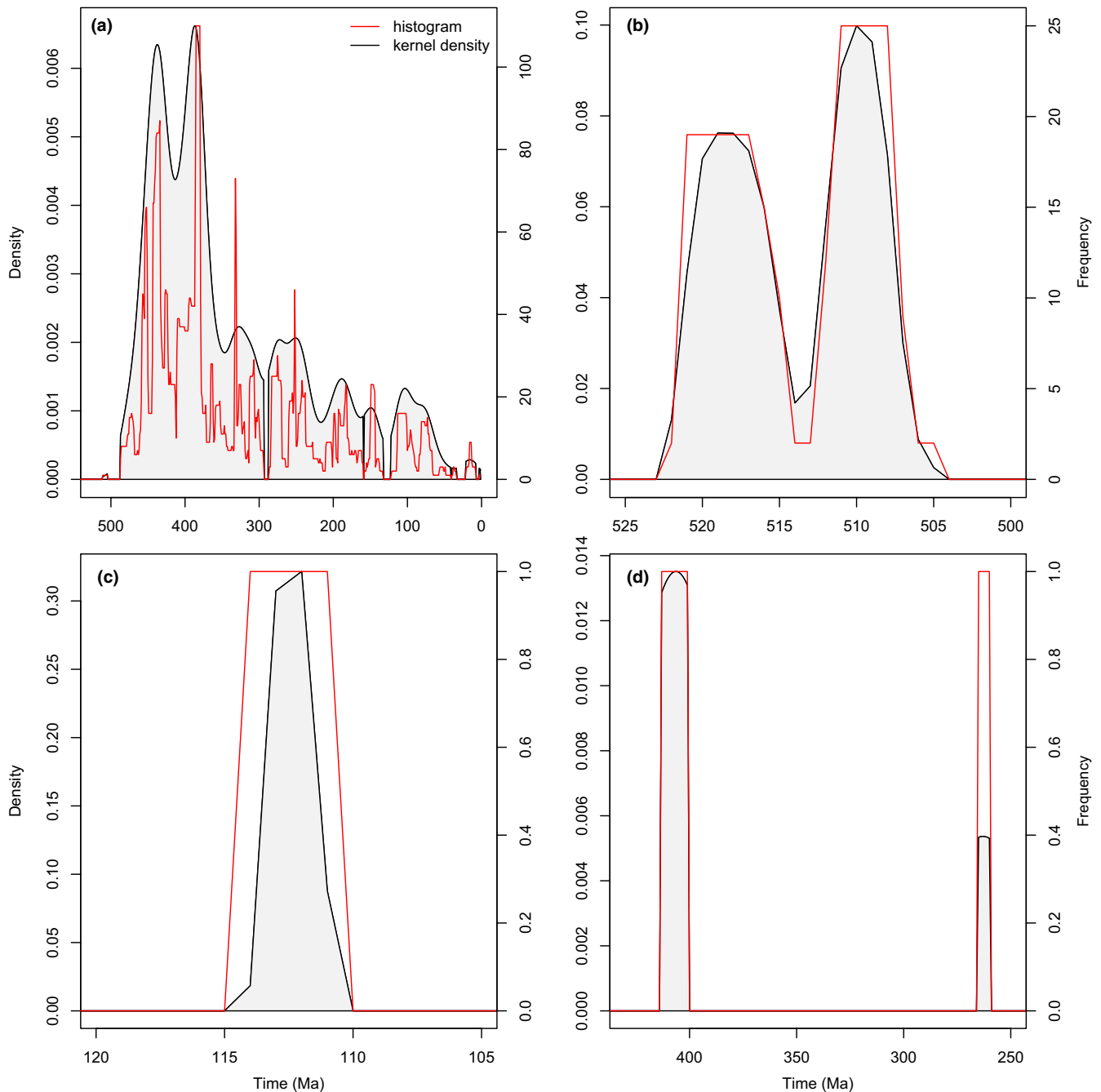


FIGURE 3 Empirical stratigraphic density distributions of Paleobiology Database (PBDB) taxa. (a) A complex, multi-peaked distribution for the brachiopod *Lingula*, reflecting its frequent misuse throughout the fossil record for various linguliform taxa. (b) A bimodal distribution resulting from patchy sampling of the valid Cambrian arthropod taxon *Waptia*. (c) The expected density distribution for a stratigraphically plausible taxon, here the Cretaceous turtle *Caririemys*. (d) A bimodal density distribution which is stratigraphically suspect given the gap between the peaks and lack of intervening sampling, resulting from the re-use of the name *Nanochilina* for unrelated taxa.

occurrence ranges or minor fluctuations in sampling may produce small peaks in density (Figure 4a), a density distribution is first smoothed within a local window to reduce noise, then the local peak (i.e. a density value in excess of both of its immediate neighbours) taken as significant if its value exceeds the mean plus standard deviation of the local window (Figure 4b). If one significant peak is detected, the distribution conforms to the unimodal expectation and all occurrences are considered valid. For multimodal distributions, interpeak durations are sequentially examined forwards in time with

an interpeak duration below a threshold gap length taken to indicate a peak split by poor sampling. Otherwise, the peaks are considered to represent separate taxa and their separation point in time defined as the interpeak nadir. This equates to dividing a stratigraphically suspect morphospecies into a sequential set of stratospecies. As outlined above, the gap threshold value may be informed by the expected taxonomic duration for the taxon in question. Careful choice is critical here: too broad and anomalies will not be detected, too short and valid morphospecies will be erroneously subdivided,

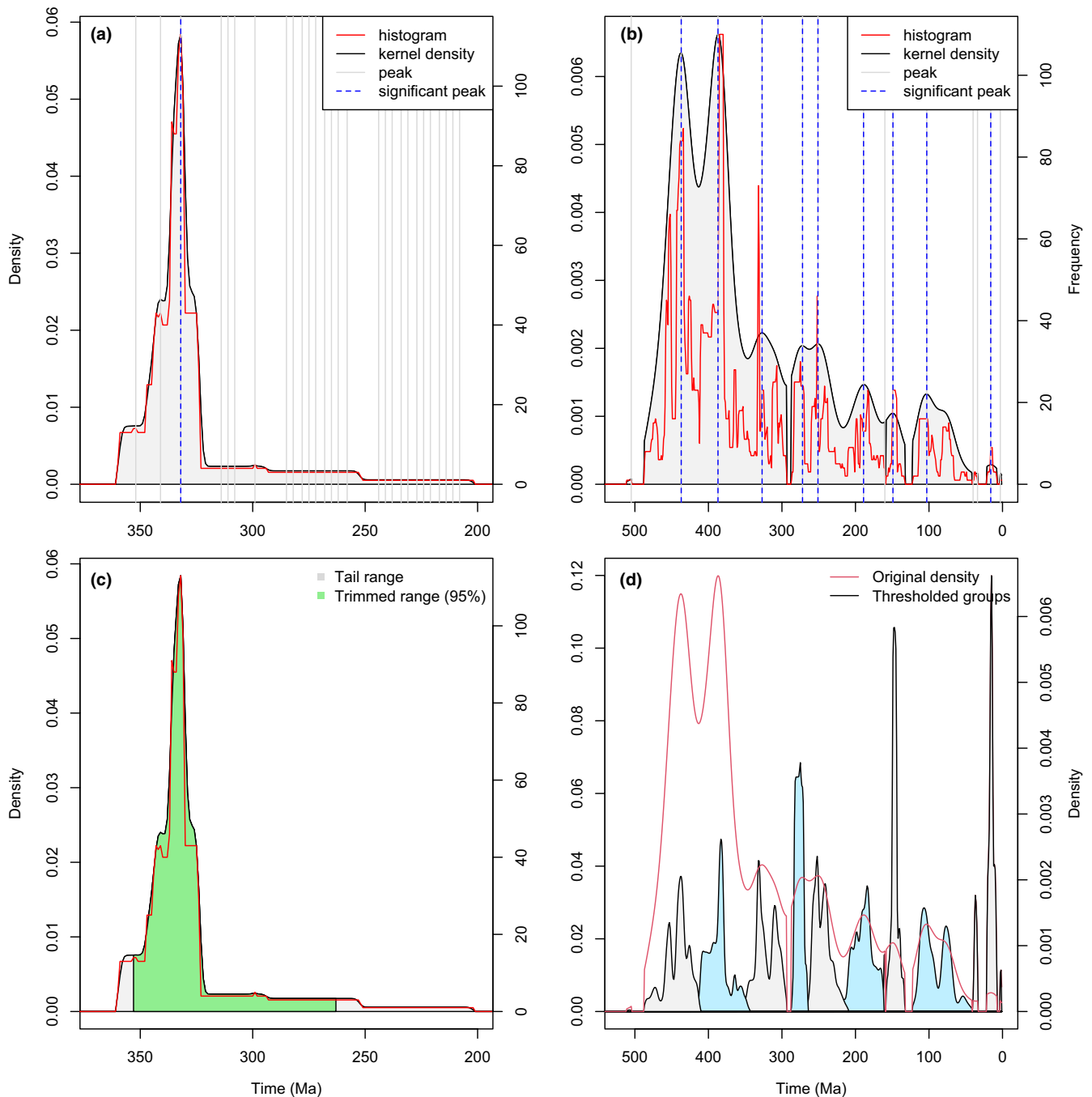


FIGURE 4 Peak detection, interpeak thresholding and Pacmacro tail trimming methods applied to empirical occurrence density distributions. (a, b) Density distribution, local peaks and significant peaks for the Carboniferous coral *Lithostrotion* and the brachiopod *Lingula* respectively. (c) Tail detection and trimming by Pacmacro for *Lithostrotion*, with the trimmed range highlighted in green. (d) Interpeak thresholding for *Lingula*, dividing its occurrences into nine stratigraphically coherent groups of linguliform morphologies.

including in the likely rare cases where population-level dynamics manifest as peaks and troughs in occurrence density distributions. As interpeak thresholds should relate to the durations of entire taxa on macroevolutionary time-scales; however, these are expected to encompass peaks and troughs present at population-level time-scales. In turn, average taxon duration provides one potential means of selecting a suitable interpeak threshold (see *Application to the PBDB*), although this will then depend upon the taxonomic composition of an occurrence dataset and how well the sampled taxon durations

reflect true taxon durations. Similarly, the interpeak threshold may also vary through time due to the varying coarseness of the chronostratigraphic intervals on which occurrence ages are based.

2.2.4 | Pacmacro

We develop an analogous method to Pacman profiling, used for detecting stratigraphic errors in microfossil occurrence data from

sediment cores (Lazarus et al., 2012). Pacman profiling scrutinises the density of microfossil occurrences versus core depth (i.e. time) to minimise errors at the tails of a range, for example due to reworking of fossils in the sediment or through conflation of ancestors and descendants within anagenetic lineages. As Pacman profiling is linked to well-constrained age models, microfossil occurrences can be treated as point-wise observations rather than bearing stratigraphic uncertainties and a simple percentage trim is applied to the tails of a taxon density distribution in the core. Under the assumption that errors will be primarily distributed towards the ends of the distribution, this procedure preferentially targets the noisy tail portions of a set of occurrences, naturally improving the signal to noise ratio (Lazarus et al., 2012). Using a percentage trim has the further benefits of locally adapting to the density distribution at the tail ends of data and freedom from any assumptions about the shape of distribution (Lazarus et al., 2012). A nonparametric approach is preferable as macrofossil density distributions show varied shapes (Figure 3) but applying a percentage trim directly may be inappropriate as macrofossil occurrences are not pointwise observations. Instead, Pacmacro flags anomalously long tails in a stratigraphic density distribution based on the proportion of the total stratigraphic duration they represent, then trims the tail stratigraphic proportion from the taxon range (Figure 4c). We applied our procedure with tail proportions of 1%–90%, then used breakpoint analysis with the *segmented* R package (Vito & Muggeo, 2008) on the relationship between tail proportion and the numbers of flagged taxa to identify a suitably conservative value (Figure S2). Under a conservative tail density proportion of 5%, we suggest that stratigraphic proportions of 40%–60% are suitable for detecting anomalously long tails and implement this procedure using the function `pacmacro_ranges()`. Ranges trimmed by Pacmacro can then be supplied to `flag_ranges()` to identify anomalous occurrences, and suspicious density distributions flagged by either our Pacmacro or interpeak thresholding methods plotted using the function `plot_dprofile()` thresholding for further inspection.

2.3 | Application to the PBDB

2.3.1 | Data acquisition and standardisation

Here, we apply our data harmonisation and stratigraphic flagging procedures to the PBDB to examine the structure of anomalies in the database and demonstrate how our protocols can be applied and documented in a transparent, reproducible and standardised manner. We downloaded the entire PBDB on 15/06/21 and updated occurrence chronostratigraphy to GTS 2020. Next, we scanned for formatting irregularities using `check_taxonomy()`. Suprageneric synonyms were identified and resolved manually due to the higher frequency of false positives, genus-level synonyms below a conservative q-gram threshold of 0.2 automatically resolved to the more frequent name and inconsistent higher taxonomic schemes automatically resolved using default settings. For anomaly flagging,

we used the Sepkoski Compendium, a database of stratigraphic ranges for >36,000 Phanerozoic marine genera (Sepkoski, 2002). While this precluded checking of terrestrial occurrences and plants, marine animals comprise the majority of PBDB entries, and the Sepkoski Compendium remains taxonomically useful despite some accumulated errors as it was based on secondary literature such as the Treatise of Invertebrate Palaeontology and uses interval-based epoch to substage-level dating which can be updated to a modern chronostratigraphic standard. The Sepkoski Compendium was downloaded using the `fetch()` function of the *chronosphere* R package (Kocsis & Raja, 2019). Minor typographical errors in interval notations were corrected manually without any alteration to the interval ranges themselves, chronostratigraphy updated to GTS 2020, then spelling errors and higher taxonomy checked as above. Finally, the Sepkoski Compendium was appended to the cleaned PBDB dataset and `check_taxonomy()` reapplied to align the higher taxonomy between both databases.

2.3.2 | Anomaly resolution and analysis

Stratigraphic anomalies were flagged against the Sepkoski Compendium using `flag_ranges()`, then collection ages revised by `resolve_ranges()` with the default consensus threshold of 75%. Pacmacro was applied to detect extended stratigraphic ranges using `pacmacro_ranges()`, with the default tail density proportions of 5% and a tail stratigraphic proportion of 40%. Further anomalous occurrences were then identified using `flag_ranges()` and the Pacmacro-trimmed stratigraphic ranges. Finally, the stratigraphic consistency of taxonomic names was queried using `threshold_ranges()`. We calculated mean and median class-wise genus durations from the PBDB and the Sepkoski Compendium, with the distribution of average durations indicating that 15 Ma is a suitably relaxed default (Figure S3); this value was used where class-specific durations were unavailable. To determine the potential impact of our data cleaning procedures on empirical palaeobiological analysis, we calculated range-through diversity, and second-for-third speciation and extinction rates at 5 Ma intervals through the Phanerozoic using the `divDyn()` function from the *divDyn* R package (Kocsis et al., 2019). The results of all our analyses are available in the electronic supplement (Flannery Sutherland, Raja, et al., 2022).

3 | RESULTS

3.1 | Data imputation errors

Our `check_taxonomy()` function flagged 67,629 name formatting irregularities in the 1,526,026 database entries (Table 1). Flagged irregularities become increasingly common as taxonomic level decreases, arising at higher levels from PBDB-specific formatting for missing higher taxonomy (e.g. NO_CLASS_SPECIFIED), and at genus level from frequent inclusion of bracketed subgenus names.

TABLE 1 Counts of data imputation anomalies in the Paleobiology Database (PBDB; $n = 1,526,026$)

	Phylum	Class	Order	Family	Genus
Non-letter characters	2747	15,413	20,601	13,276	7172
Incorrect word count	0	0	0	0	7020
Potential synonyms	2	2	38	210	885
Cross-rank homonyms	3	5	6	3	11
Inconsistently classified	0	1	3	32	199

The majority of suprageneric synonyms are taxonomically distinct (e.g. Homocrinidae–Holocrinidae), with q-gram distances typically >0.2 , that is, $<80\%$ similarity. Otherwise, synonyms mostly reflect inconsistent spelling or use of subclades or superclades at formal taxonomic ranks (e.g. Pyrotheridae–Pyrotheriidae, Bothriolepididae–Bothriolepidae). While these latter cases are not true synonyms, it is still inappropriate to use, for example, a subclass classification at the class level when the class level classification is also present (e.g. Actinopterygi–Actinopterygii). The number of genuine synonyms rises substantially at the genus level, primarily from spelling errors and assonance (e.g. *Allonnia*–*Allonia*, *Sichuanolenus*–*Szechuanolenus*, *Drepanochilus*–*Drepanocheilus*). The PBDB dynamically generates a coherent taxonomic scheme so cases of inconsistent higher classifications are unsurprisingly infrequent at just 235 instances, the majority of which are homonyms between distant clades or instances of missing higher classifications for some occurrences.

3.2 | Anomaly resolution using the Sepkoski compendium

Prior to resolving occurrence ages collection wise, 24.6% of PBDB taxa with entries in the Sepkoski Compendium had stratigraphic distributions concordant with their reference ranges. $0R0$ anomalies are most prevalent (Table 2), indicating frequent range overextension by their occurrence records. Taxa fully outside of their ranges ($00R$, $R00$) are roughly equal while $0R1$ anomalies are more common than $1R0$ anomalies. Conversely, the proportion of valid occurrences is substantially greater than the proportion of valid taxa and occurrence-wise $0R0$ anomalies are least common (Table 2), which is expected given that occurrence age uncertainties should typically be smaller than taxon age ranges.

Resolution using the Sepkoski Compendium substantially reduced taxon-wise and occurrence-wise error prevalence (Figure 5a,b). The increase in the proportion of valid occurrences is modest, but the proportion of valid taxa more than doubles and $0R0$ anomalies indicative of high age imprecision are virtually eliminated (Table 2). Unexpectedly, taxon-wise $00R$ anomalies decrease only slightly while $R00$ anomalies increase slightly, despite substantial reductions in the prevalence of all other error types for both taxa and occurrences and the high stratigraphic plausibility of collections after revision. Of the 216,568 collections in the download, 68.4% went unchecked as they did not contain any taxa present in the Sepkoski Compendium, highlighting a limitation of the reference database approach. Of those which could be checked, 10.8% did not

TABLE 2 Prevalence of taxon-wise and occurrence-wise anomalies in the Paleobiology Database (PBDB)

	Error type prevalence (%)					
	$00R$	$0R1$	$0R0$	$1R0$	$R00$	$1R1$
Sepkoski						
Taxa (pre)	13.7	17.2	23.6	9.1	11.9	24.6
Taxa (post)	12.9	6.6	9.9	4.7	12.9	52.9
Occs (pre)	5.6	6.3	1.1	6.0	7.5	73.5
Occs (post)	3.6	1.3	0.1	1.3	5.2	88.5
Pacmacro						
Taxa	12.1	10.7	70.0	2.0	5.5	–
Occs	24.5	36.1	2.0	38.1	25.5	–

meet the consensus threshold for revision and 2.4% retained their original age, while 86.5% had their age revised with the vast majority displaying 100% consensus (Figure S4). Taxon-wise FAD and LAD anomalies show similar distributions (Figure 5b) in contrast to occurrence-wise anomalies, where LAD anomalies are more prevalent overall but FAD anomalies in the range of a few million years are disproportionately more common (Figure 5d).

While our consensus revision method returns less precise ages during the Carboniferous and Permian, it substantially improves median collection age precision throughout the Cambrian to Devonian and during the Triassic and achieves median age precision comparable to the unrevised ages throughout the remainder of the Phanerozoic in the absence of any lithostratigraphic information (Figure 6). Age precisions show a greater interquartile range after revision, indicating that the method performs variably between collections. Nonetheless, the 25% quartile on precision falls below that for unrevised age precision in virtually each time bin (Figure 6), demonstrating that improved stratigraphic resolution is always achieved for at least some collections.

3.3 | Density methods

In all, 2487 genera (3.7%) showed long tails in their occurrence density distributions, with distinct differences in taxon-wise and occurrence-wise anomalies when flagged against Pacmacro-trimmed ranges. Occurrence-wise $0R0$ anomalies are the least prevalent, with relatively even balances of $00R$ to $R00$ and $0R1$ to $1R0$ anomalies (Table 2). Consequently, $0R0$ anomalies indicative of two-tailed stratigraphic density distributions are the most

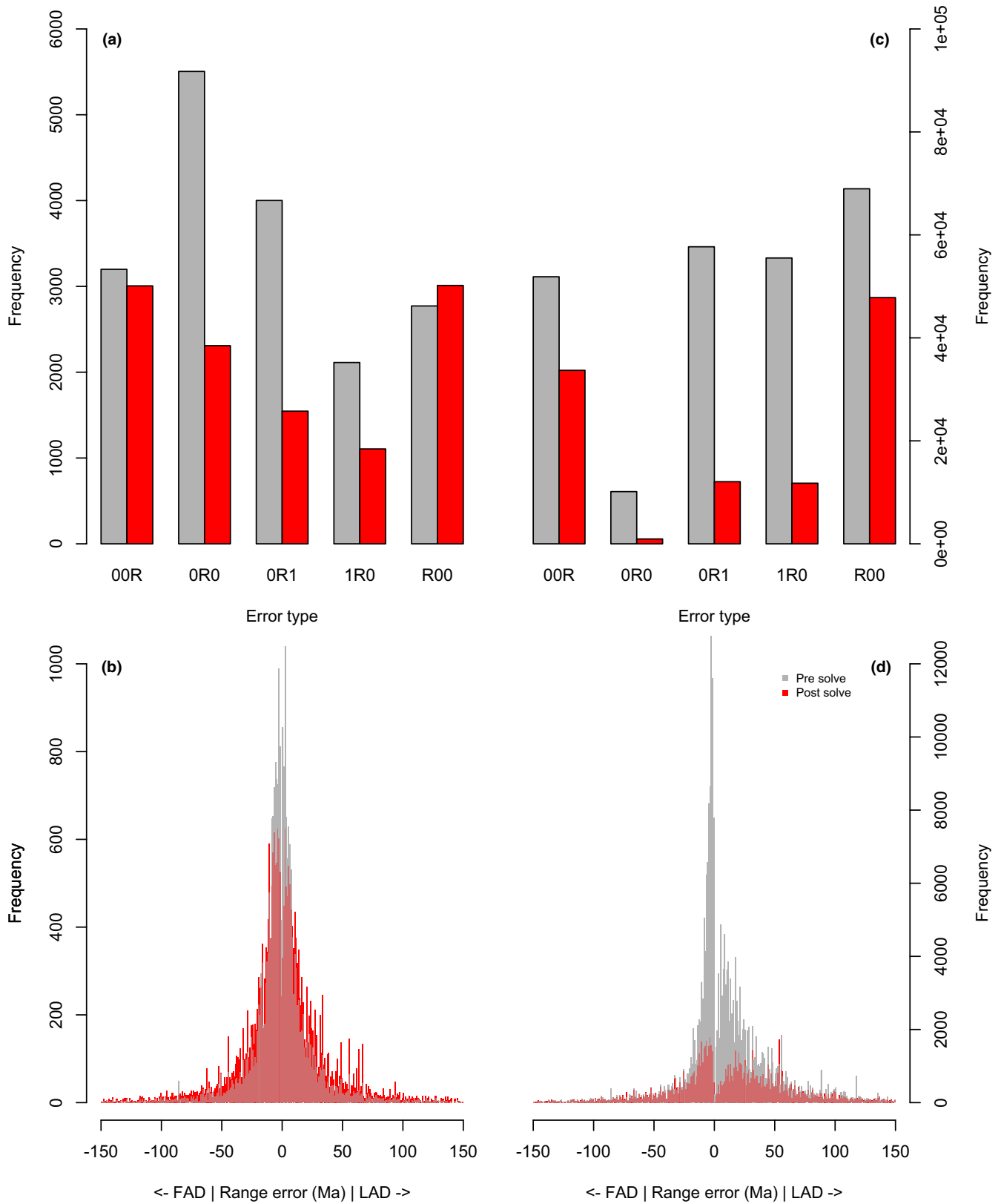
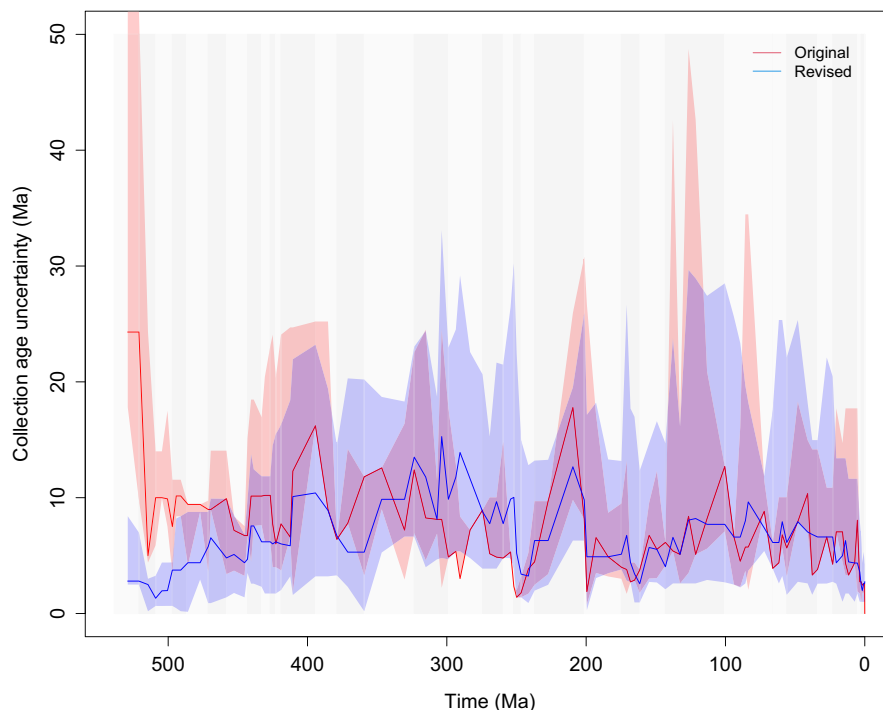


FIGURE 5 Impact of collection age revision using the Sepkoski compendium. (a) Taxon-wise anomaly prevalence. (b) Taxon-wise distribution of FAD and LAD anomalies. (c) Occurrence-wise anomaly prevalence. (d) Occurrence-wise distribution of FAD and LAD anomalies. See text for the definitions of the error codes.

FIGURE 6 Collection age uncertainty throughout the phanerozoic, before and after collection age revision using the Sepkoski compendium. Solid lines indicate the median and the shaded regions the interquartile range. Grey bars demarcate the epochs of the Phanerozoic.



prevalent among taxa, reflecting the relatively even distribution of occurrence-wise FAD and LAD anomalies, but anomalies on taxon FADs (ORR, OR1) are disproportionately more common than anomalies on taxon LADs (Table 2). As an example, we draw attention to *Lithostrotion*, a colonial rugose coral abundant during the Visean (Early Carboniferous). Its density distribution contains multiple local peaks, but our method picks out the single true peak (Figure 4a), marking it as stratigraphically coherent morphospecies. Rugose corals went extinct during the Permo-Triassic mass extinction (Wang et al., 2018), but records of *Lithostrotion* persist to the end of the Triassic (Figure 4a). Anomalously, long tails were successfully flagged for the genus and its range truncated to within the Palaeozoic by the 5% tail trim (Figure 4c). While its range in the Sepkoski Compendium is more conservative (346–326 Ma), our method still removes highly erroneous portions of the occurrence record in this case, returning a range that is concordant for rugose corals and the family Lithostrotionidae (Carboniferous–Late Permian; Wang et al., 2018).

Using class-specific interpeak thresholds, 10,733 genera were split stratigraphically (16.1%). The number of splits shows a positive relationship with original genus duration (Figure S5), demonstrating that longer stratigraphic durations are less likely to show plausible occurrence density distributions, and more likely to be split into greater numbers of groups. We highlight several examples here, chiefly *Lingula*, a brachiopod reported throughout the Phanerozoic and the so-called 'living fossil' given its supposed antiquity. In reality, *Lingula* is a recent genus, with older occurrences representing multiple different genera which convergently evolved shell morphologies adapted for burrowing (Emig, 2003). This is reflected by its occurrence density record which contains multiple significant peaks (Figure 4b), split by our thresholding method into

nine stratigraphically distinct groups of linguliform occurrences (Figure 4d). Similarly, the two peaks for *Nanochilina* are separated by nearly 250 Ma (Figure 4d) and demonstrates a case where a homonym has resulted in a stratigraphically implausible density record. Conversely, the Cambrian arthropod *Waptia* shows two peaks in its density record (Figure 4b) which fall within the interpeak threshold, indicating a stratigraphically coherent, taxonomically valid record split by sampling artefacts. This is supported by its limited fossil distribution but distinctive morphology and apomorphies, with identifiable specimens derived primarily from the Burgess Shale of British Columbia (Taylor, 2002), and from older occurrences in the Marjum Shale and Wheeler Formation of Utah (Briggs et al., 2008).

4 | DISCUSSION

4.1 | Outliers and anomalies in PBDB occurrences

The even balance of taxon-wise FAD and LAD anomalies (Figure 5b) shows that there is no particular bias in the PBDB towards either mode of range overestimation at the genus level. Instead, the disproportionate frequency of small FAD anomalies versus the greater total prevalence of occurrence-wise LAD anomalies (Figure 5d) suggests overestimation of FADs relates more strongly to stratigraphic imprecision, and overestimation of LADs to taxonomic misidentification. This is supported by independent flagging of outliers against Pacmacro-trimmed ranges. The even distributions of occurrence-wise FAD and LAD anomalies flagged against Pacmacro-trimmed ranges and the resulting prevalence of OR0 anomalies and two-tailed anomalies on taxon stratigraphic density distributions is congruent

with the lack of bias towards predominant FAD or LAD overestimation. Similarly, the greater prevalence of occurrence-wise FAD anomalies relative to the Sepkoski Compendium is matched by the disproportionate frequency of FAD anomalies relative to Pacmacro-trimmed ranges.

Taxonomic representation in the PBDB is highly disproportionate, so we normalise class-wise numbers of flagged genera by the total numbers of class-wise genera to investigate proportional error rates under each detection method. The distribution of error proportions versus clade size indicates that comparison against the Sepkoski Compendium is the most stringent detection procedure and Pacmacro profiling the most conservative in its approach to identifying anomalies (Figure S6; Tables S2 and S3). Under each method, the highest proportional error rates typically occur in classes with relatively small numbers of genera, reflecting the patchiness of their fossil records. Plants in particular show stratigraphically suspect occurrence densities with a high prevalence of long stratigraphic tails and splitting by interpeak

thresholding (Table S3), corroborating previous concerns over their poor representation in the PBDB (Cleal & Thomas, 2010; Silvestro et al., 2015). Nonetheless, these relatively poorly sampled clades contribute only a small amount of the overall anomalies in the database. More concerning is the repeated flagging within well-sampled, taxonomically diverse clades comprising the bulk of PBDB occurrences which, despite their lower proportion of anomalies, contribute the most to the overall error pool. Major clades containing thousands of genera, including gastropods, chondrichthyans, ostracods, brachiopods, bivalves, cephalopods, osteichthyans and tetrapods show range error proportions >50% prior to resolution against the Sepkoski Compendium. Collection age resolution had a positive impact on class-wise error proportions, but these remain in the range of 20%–50% for major clades. Density methods also reveal moderate class-wise error proportions, with division of stratigraphically suspect occurrence density distributions by interpeak thresholding of 20%–30%, and anomalous stratigraphic tails in the range of 2%–9% (Tables S2

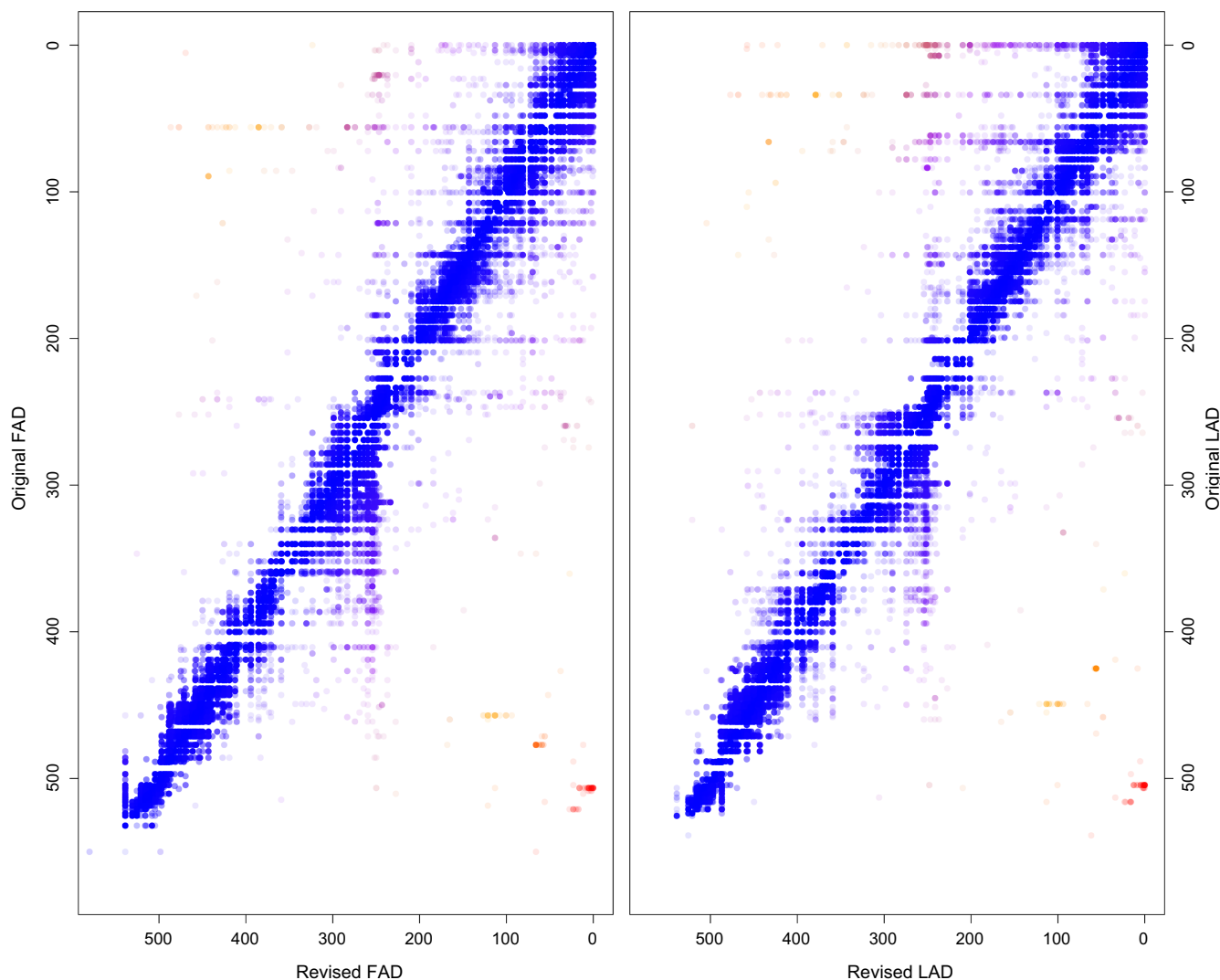


FIGURE 7 Cross-plots of pre- and post-revision collection FADs and LADs throughout the Phanerozoic. Warmer colours indicate increasing discrepancy between original and revised ages.

and S3). Long tails and range anomalies call into question the validity of individual occurrence identifications while the impact of interpeak thresholding calls into question the validity of entire genera with stratigraphically suspect occurrence density distributions. Conversely, the discrepancy between the number of taxon-wise and occurrence-wise anomalies flagged against the Sepkoski Compendium suggests that a relatively small number of occurrence-wise anomalies are responsible for the alarming prevalence of taxon-wise range anomalies in the PBDB, while the vast majority of occurrences are otherwise valid.

4.2 | Phanerozoic diversity dynamics and error structure

Temporal structuring in discrepancies between original and revised FADs and LADs (Figure 7) highlights forward-smearing of a number of Cambrian and Ordovician-aged collections and of variably aged Palaeozoic FADs and LADs to revised ages coincident with the Permo-Triassic boundary, back-smearing of FADs at approximately 65Ma coincident with the end-Cretaceous mass extinction, back-smearing of LADs around 35Ma at the end of the Eocene, and back-smearing of present-day LADs. The magnitudes of some of these revisions are on the order of hundreds of millions of years, suggesting that frequent misidentification of age-diagnostic taxa in younger assemblages may produce 'relict' collections of superficially greater

antiquity. Bands of back-smearing appear to coincide at least partially with major phases of turnover, highlighting their effect as taxonomic watersheds between discrete assemblages of taxa on broader temporal scales (Muscente et al., 2018; Rojas et al., 2021). For example, there is virtually no revision of collection ages across the Permo-Triassic boundary, highlighting how the event creates a clear division between Palaeozoic and Mesozoic morphospecies and assemblages. This watershed effect may be responsible for the marked grid-wise pattern of age revision in the post-Palaeozoic portion of the dataset, with the increasing taxonomic misidentification of increasingly prevalent 'modern' morphospecies driving the greater degree of age restructuring. This could also conceivably be tied to variation in stage and substage length throughout the Phanerozoic, with coarser bin lengths inducing greater age revisions. Consequently, suitable thresholds for the interpeak method and tail proportions for Pacmacro may also range through time as well as between clades due to varying precision of Phanerozoic stages.

There are clear differences in the tempo of diversification throughout the Phanerozoic as a result of our taxonomic and stratigraphic revision (Figure 8). Most notably, diversity accrues more rapidly in the wake of the Late Devonian mass extinction before a trend of Late Permian decline up to the end-Permian mass extinction in contrast to the pre-cleaning trend of continued increase. While these results remain inaccurate due to the known impact of spatial sampling bias and regional heterogeneity on global estimates of diversity, speciation and extinction rates through geological time

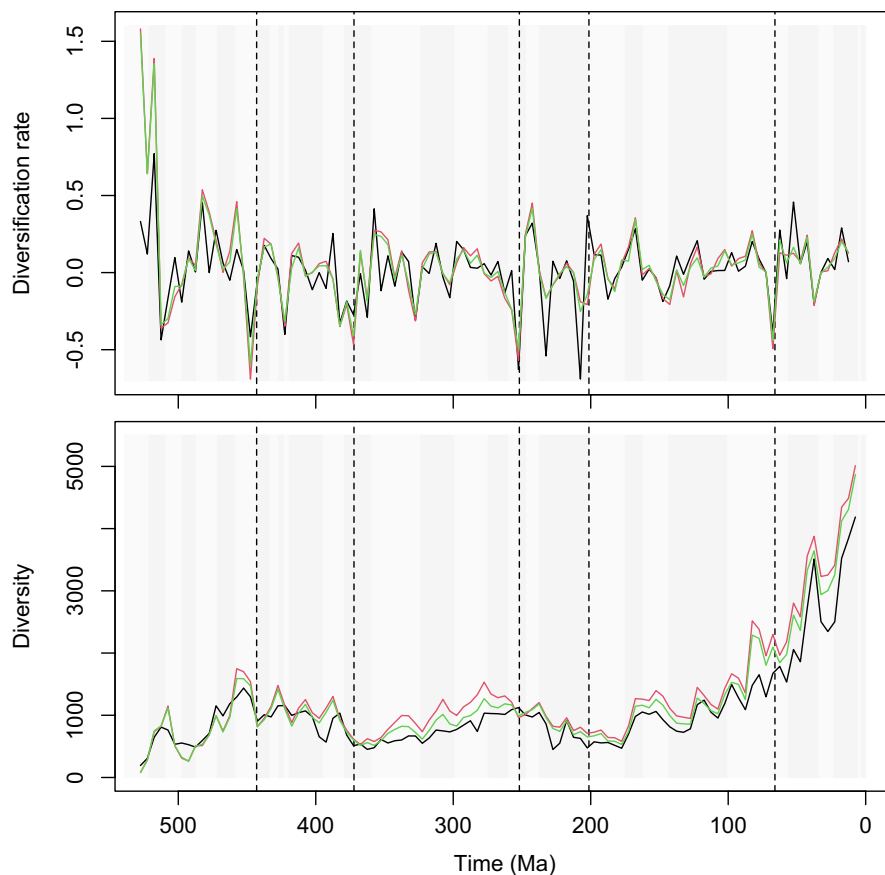


FIGURE 8 Diversification rate (origination minus extinction) and diversity through the Phanerozoic, before and after anomaly detection and resolution. Black lines are uncorrected, red lines from data with stratigraphic ages corrected using the Sepkoski Compendium and green lines from data further split using interpeak thresholding. Dashed lines mark the Big Five mass extinctions. Grey bars demarcate the epochs of the Phanerozoic.

(Close et al., 2020; Flannery Sutherland, Silvestro, & Benton, 2022), they nonetheless demonstrate the impact of taxonomic and stratigraphic revision using our methods. Set against the context of the entire PBDB, the anomalies we detect and resolve using purely statistical methods have only a modest impact on patterns of diversity (Figure 8). This corroborates previous analysis of the impact of taxonomic revision on PBDB data (Wagner et al., 2007) and comparison of broad patterns of diversity before and after taxonomic revision of the Sepkoski Compendium additionally found that higher level trends remained stable despite a high prevalence of errors, as their distribution throughout the entire database was random (Adrain & Westrop, 2000). Recovered error prevalence within individual clades may differ if detection were followed by manual vetting by taxonomic specialists, however, and so the impact of error resolution on downstream analysis is expected to be greater for smaller taxonomic subsets of fossil occurrence data.

5 | CONCLUSIONS

Robust palaeobiological research is predicated on high-quality data, yet techniques which address errors in fossil occurrences databases are underdeveloped. We add to existing tools for resolving errors in geographical coordinate data with methods for standardisation and cleaning of taxonomic names, and for the unique challenges presented by the stratigraphic component of fossil occurrence data within the new R package *fossilbrush*. Our multi-step name cleaning routing, covering consistent formatting, detection of homonyms and inconsistent higher classifications and the re-use of names at different taxonomic levels, is provided as a single R function *check_taxonomy()* and does not rely on any external databases for checking. As such, it can be applied to any dataset with taxonomic information present, and with any number of levels in the recorded taxonomic hierarchy and so we anticipate that its utility will extend beyond application solely to fossil occurrence datasets.

Our most stringent error detection method utilises a reference database (here the Sepkoski Compendium) to assess entire assemblages of fossils, querying the validity of occurrences, taxa and collections. We find that collection ages can be effectively revised using plausible consensus ages of their taxa, often improving their stratigraphic precision in the process, but anomalies in taxon ranges relative to the Sepkoski Compendium are frequent in the PBDB. FAD anomalies may relate to stratigraphic imprecision, while LAD anomalies may arise more from taxonomic misidentification, informing where future cleaning efforts based on expert knowledge should be targeted. We also provide conceptual advances on how occurrences records may be treated as observation densities which incorporate stratigraphic uncertainty, along with how the properties of these density distributions may be used to flag anomalous occurrences and stratigraphically suspect taxa.

Finally, we stress that while our methods appear to function effectively and scale well to the challenges presented by a large occurrence database, the occurrence density methods rely on outlier

detection, a statistical solution, rather than drawing upon expert knowledge where a definitive solution can be achieved. As such, they are best applied in concert so that occurrences which are repeatedly flagged by each method may be confidently assessed as erroneous.

AUTHOR CONTRIBUTIONS

Nussaibah Raja and Wolfgang Kiessling devised and issued the PBDB data-cleaning challenge. Joseph Flannery-Sutherland developed the methods and code selected as the winning submission. All authors helped to refine this submission into the *fossilbrush* R package and wrote and commented on the manuscript.

ACKNOWLEDGEMENT

We thank the two anonymous reviewers whose comments helped improve the quality of this manuscript.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest posed by this work.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.13966>

DATA AVAILABILITY STATEMENT

The *fossilbrush* R package may be downloaded from the CRAN repository at <https://cran.r-project.org/web/packages/fossilbrush/index.html>. All data, code, supplementary figures and tables are available on Zenodo at <http://doi.org/10.5281/zenodo.6572920> (Flannery Sutherland, Raja, et al., 2022).

ORCID

Joseph T. Flannery-Sutherland  <https://orcid.org/0000-0001-8232-6773>

Nussaibah B. Raja  <https://orcid.org/0000-0002-0000-3944>

Ádám T. Kocsis  <https://orcid.org/0000-0002-9028-665X>

Wolfgang Kiessling  <https://orcid.org/0000-0002-1088-2014>

REFERENCES

- Adrain, J., & Westrop, S. (2000). An empirical assessment of taxic palaeobiology. *Science*, 289, 110–112. <https://doi.org/10.1126/science.289.5476.110>
- Allmon, W., Dietl, G., Hendricks, J., & Ross, R. (2018). Bridging the two fossil records: Paleontology's 'big data' future resides in museum collections. In G. Rosenberg & R. Clary (Eds.), *Museums at the forefront of the history and philosophy of geology: History made, history in the making* (Vol. 535, pp. 35–44). Geological Society of America Special Papers. [https://doi.org/10.1130/2018.2535\(03\)](https://doi.org/10.1130/2018.2535(03))
- Alroy, J., Aberhan, M., Bottjer, D., Foote, M., Fürsich, F., Hendy, A., Holland, S., Ivany, L., Kiessling, W., Kosnik, M., Marshall, C., McGowan, A., Miller, A., Olszewski, T., Patzkowsky, M., Wagner, P., Bonuso, N., Borkow, P., Brenneis, B., & Clapham, M. (2008). Phanerozoic trends in the diversity of marine invertebrates. *Science*, 321, 97–100. <https://doi.org/10.1126/science.1156963>
- Barido-Sottani, J., Pett, W., O'Reilly, J., & Warnock, R. (2019). FossilSim: An R package for simulating fossil occurrence data

- under mechanistic models of preservation and recovery. *Methods in Ecology and Evolution*, 10, 835–840. <https://doi.org/10.1111/2041-210X.13170>
- Bell-Damarow, J., Brenskelle, L., Barve, N., Soltis, P., Sierwald, P., Bieler, R., LaFrance, R., Arino, A., & Guralnick, R. (2019). Research applications of primary biodiversity databases in the digital age. *PLoS ONE*, 14, e0215794. <https://doi.org/10.1371/journal.pone.0215794>
- Briggs, D., Lieberman, B., Hendricks, J., Halgedahl, S., & Jarrard, R. (2008). Middle Cambrian Arthropods from Utah. *Journal of Paleontology*, 82, 238–254. <https://doi.org/10.1666/06-086.1>
- Cleal, C., & Thomas, B. (2010). Botanical nomenclature and plant fossils. *Taxon*, 59, 261–268. <https://doi.org/10.2307/27757068>
- Close, R., Benson, R., Saupe, E., Clapham, M., & Butler, R. (2020). The spatial structure of Phanerozoic marine animal diversity. *Science*, 368, 420–424. <https://doi.org/10.1126/science.aay8309>
- Clowes, C., Crampton, J., Bland, B., Collins, K., Prebble, J., Raine, J., Strogon, D., Terezow, M., & Womack, T. (2021). The New Zealand Fossil Record File: A unique database of biological history. *New Zealand Journal of Geology and Geophysics*, 64, 62–71. <https://doi.org/10.1080/00288306.2020.1799827>
- Cohen, K., Finney, S., Gibbard, P., & Fan, J. (2013). The ICS International Chronostratigraphic Chart. *Episodes*, 36, 199–204.
- Dietl, G. (2019). Conservation palaeobiology and the shape of things to come. *Philosophical Transactions of the Royal Society B*, 374, 20190294. <https://doi.org/10.1098/rstb.2019.0294>
- Eduardo, A., Martinez, P., Gouveia, S., Santos, F., deAragao, W., Morales-Barbero, J., Kerber, L., & Liparini, A. (2018). Extending the paleontology–biogeography reciprocity with SDMs: Exploring models and data in reducing fossil taxonomic uncertainty. *PLoS ONE*, 13, e0194725. <https://doi.org/10.1371/journal.pone.0194725>
- Emig, C. (2003). Proof that lingula (Brachiopoda) is not a living-fossil, and emended diagnoses of the family Lingulidae. *Carnets de Geologie*, CG2003_L01, 1–8.
- Fenton, I., Woodhouse, A., Aze, T., Lazarus, D., Renaudie, J., Dunhill, A., Young, J., & Saupe, E. (2021). Triton, a new species-level database of Cenozoic planktonic foraminiferal occurrences. *Nature Scientific Data*, 8, 160. <https://doi.org/10.1038/s41597-021-00942-7>
- Flannery Sutherland, J., Raja, N., Kocsis, A., & Kiessling, W. (2022). Electronic supplement for Flannery Sutherland et al. 2022 – fossilbrush: An R package for automated detection and resolution of errors in palaeontological occurrence data. *Zenodo*, <https://doi.org/10.5281/zenodo.6572920>
- Flannery Sutherland, J., Silvestro, D., & Benton, M. (2022). Global diversity dynamics in the fossil record are regionally heterogeneous. *Nature Communications*, 13, 275. <https://doi.org/10.1038/s41467-022-30507-0>
- Foote, M. (2007). Symmetric waxing and waning of marine invertebrate genera. *Paleobiology*, 33, 517–529. <https://doi.org/10.1666/06084.1>
- Foote, M., Crampton, J., Beu, A., Marshall, B., Cooper, R., Maxwell, P., & Matcham, I. (2007). Rise and fall of species occupancy in Cenozoic fossil mollusks. *Science*, 318, 1131–1134. <https://doi.org/10.1126/science.1146303>
- Gradstein, F., Ogg, J., Schmitz, M., & Ogg, G. (2020). *Geologic Timescale 2020* (1st ed.). Elsevier. <https://doi.org/10.1016/C2020-1-02369-3>
- Graham, R., & Lundelius, E. (2010). *FAUNMAP II: New data for North America with a temporal extension for the Blancan, Irvingtonian and early Rancholabrean*. FAUNMAP II Database, version 1.0, <https://ucmp.berkeley.edu/faunmap/index.html>
- Grenié, M., Berti, E., Carvajal-Quintero, J., Sagouis, A., & Winter, M. (2022). Harmonizing taxon names in biodiversity data: A review of tools, databases, and best practices. *Methods in Ecology and Evolution*. <https://doi.org/10.1111/2041-210X.13802>
- Kidwell, S. (2015). Biology in the Anthropocene: Challenges and insights from young fossil records. *Proceedings of the National Academy of the United States of America*, 112, 4922–4929. <https://doi.org/10.1073/pnas.1403660112>
- Kidwell, S., & Flessa, K. (1995). The quality of the fossil record: Populations, species and communities. *Annual Review in Ecology and Systematics*, 26, 269–299. <https://doi.org/10.1146/annurev.earth.24.1.433>
- Kiessling, W., & Krause, M. C. (2022). PARED – An online database of Phanerozoic reefs. <https://www.paleo-reefs.pal.uni-erlangen.de>
- Kiessling, W., Raja, N., Roden, V., Turvey, S., & Saupe, E. (2019). Addressing priority questions of conservation science with palaeontological data. *Philosophical Transactions of the Royal Society B*, 374, 1788. <https://doi.org/10.1098/rstb.2019.0222>
- Kocsis, Á., & Raja, N. (2019). Chronosphere: Earth system history variables. <https://doi.org/10.1111/2041-210X.13161>
- Kocsis, Á., Reddin, C., Alroy, J., & Kiessling, W. (2019). The R package divDyn for quantifying diversity dynamics using fossil sampling data. *Methods in Ecology and Evolution*, 10, 735–743. <https://doi.org/10.1111/2041-210X.13161>
- Kopperud, B., Lidgard, S., & Liow, L. (2018). Text-mined fossil biodiversity dynamics using machine learning. *Proceedings of the Royal Society B: Biological Sciences*, 286, 20190022. <https://doi.org/10.1098/rspb.2019.0022>
- Lazarus, D., Weinkauff, M., & Diver, P. (2012). Pacman profiling: A simple procedure to identify stratigraphic outliers in high-density deep-sea microfossil data. *Paleobiology*, 38, 144–161. <https://doi.org/10.1666/10067.1>
- Marshall, C., Finnegan, S., Clites, E., Holroyd, P., Bonuso, N., Cortez, C., Davis, E., Dietl, G., Druckenmiller, P., Eng, R., Garcia, C., Estes-Smargiassi, K., Hendy, A., Hollis, K., Little, H., Nesbitt, E., Roopnarine, P., Skibinski, L., Vendetti, J., & White, D. (2018). Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. *Biology Letters*, 14, 20180431. <https://doi.org/10.1098/rsbl.2018.0431>
- Marshall, J., Lakin, J., Troth, I., & Wallace-Johnson, S. (2021). UV-B radiation was the Devonian-Carboniferous boundary terrestrial extinction kill mechanism. *Science Advances*, 6, eaba0768. <https://doi.org/10.1126/sciadv.aba0768>
- Metcalfe, I., & Crowley, J. (2021). Upper Permian and Lower Triassic conodonts, high-precision U-Pb zircon ages and the Permian-Triassic boundary in the Malay Peninsula. *Journal of Asian Earth Sciences*, 199, 104403. <https://doi.org/10.1016/j.jseas.2020.104403>
- Muscante, A., Prabh, A., Zing, H., Eleish, A., Meyer, M., Fox, P., Hazen, R., & Knoll, A. (2018). Quantifying ecological impacts of mass extinctions with network analysis of fossil communities. *Proceedings of the National Academy of the United States of America*, 115, 517–522. <https://doi.org/10.1073/pnas.1719976115>
- Norman, K., Chamberlain, S., & Boettiger, C. (2020). taxadb: A high-performance local taxonomic database interface. *Methods in Ecology and Evolution*, 11, 1153–1159. <https://doi.org/10.1111/2041-210X.13440>
- Peters, S., & McClennen, M. (2016). The Paleobiology Database application programming interface. *Paleobiology*, 42, 1–7. <https://doi.org/10.1017/pab.2015.39>
- Peters, S., Zhang, C., Livny, M., & Re, C. (2014). A machine reading system for assembling synthetic paleontological databases. *PLoS ONE*, 9, e113523. <https://doi.org/10.1371/journal.pone.0113523>
- Plotnick, R., & Wagner, P. (2006). Round up the usual suspects: Common genera in the fossil record and the nature of wastebasket taxa. *Paleobiology*, 32, 126–146. [https://doi.org/10.1666/0094-8373\(2006\)032\[0126:RUTUSC\]2.0.CO;2](https://doi.org/10.1666/0094-8373(2006)032[0126:RUTUSC]2.0.CO;2)
- Prothero, D. (2015). Garbage in, garbage out: The effect of immature taxonomy on database compilations of North American fossil mammals. *New Mexico Museum of Natural History and Science Bulletin*, 68, 257–264.
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ribeiro, B., Velazco, S., Guidoni-Martins, K., Tassarolo, G., Jardim, L., Bachman, S., & Loyola, R. (2022). bdc: A toolkit for standardizing, integrating and cleaning biodiversity data.

- Methods in Ecology and Evolution*, 13, 1421–1428. <https://doi.org/10.1111/2041-210X.13868>
- Rojas, A., Calatayud, J., Kowalewski, M., Neuman, M., & Rosvall, M. (2021). A multiscale view of the Phanerozoic fossil record reveals the three major biotic transitions. *Communications Biology*, 4, 309. <https://doi.org/10.1038/s42003-021-01805-y>
- Sepkoski, J. (2002). A compendium of fossil marine animal genera. *Bulletins of American Paleontology*, 363, 1–560.
- Silvestro, D., Cascales-Minana, B., Bacon, C., & Antonelli, A. (2015). Revisiting the origin and diversification of vascular plants through a comprehensive Bayesian analysis of the fossil record. *New Phytologist*, 207, 425–436. <https://doi.org/10.1111/nph.13247>
- Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267, 396–404. <https://doi.org/10.1016/j.jtbi.2010.09.010>
- Stephenson, P., & Stengel, C. (2020). An inventory of biodiversity data sources for conservation monitoring. *PLoS ONE*, 15, e0242923. <https://doi.org/10.1371/journal.pone.0242923>
- Taylor, R. (2002). A new bivalved arthropod from the Early Cambrian Sirius Passet fauna, North Greenland. *Palaeontology*, 45, 97–123. <https://doi.org/10.1111/1475-4983.00229>
- The NOW Community. (2020). *New and old worlds database of fossil mammals (NOW)*. Licensed under CC BY 4.0. <https://nowdatabase.org/now/database/>, <https://doi.org/10.5281/zenodo.4268068>
- Vito, M., & Muggeo, R. (2008). segmented: An R package to fit regression models with broken-line relationships. *R News*, 8, 20–25.
- Wagner, P., Hendy, A., & Kiessling, W. (2007). The effects of taxonomic standardization on sampling-standardized estimates of historical diversity. *Proceedings of the Royal Society B: Biological Sciences*, 274, 439–444. <https://doi.org/10.1098/rspb.2006.3742>
- Wang, X., Yao, L., & Lin, W. (2018). Permian rugose corals of the world. In S. Lucas & S. Shen (Eds.), *The Permian timescale* (Vol. 450, pp. 165–184). Geological Society of London Special Publications. <https://doi.org/10.1144/SP450.13>
- Williams, J., Grimm, E., Blois, J., Charles, D., Davis, E., Goring, S., Graham, R., Smith, A., Anderson, M., Arroyo-Cabrales, J., Ashworth, A., Betancourt, J., Bills, B., Booth, R., Buckland, P., Curry, B., Giesecke, T., Hausmann, S., Jackson, S., & Latorre, C. (2018). The Neotoma Paleocology Database: A multi-proxy, international community-curated data resource. *Quaternary Research*, 89, 156–177. <https://doi.org/10.1017/qua.2017.105>
- Zhang, C., Stadler, T., Klopstein, S., Heath, T., & Ronquist, F. (2016). Total-evidence dating under the fossilized birth–death process. *Systematic Biology*, 65, 228–249. <https://doi.org/10.1093/sysbio/syv080>
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C., Edler, D., Farooq, H., Herdean, A., Arizo, M., Scharn, R., Svantesson, S., Wengstrom, N., Zizka, V., & Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, 10, 744–751. <https://doi.org/10.1111/2041-210X.13152>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Flannery-Sutherland, J. T., Raja, N. B., Kocsis, Á. T., & Kiessling, W. (2022). *fossilbrush*: An R package for automated detection and resolution of anomalies in palaeontological occurrence data. *Methods in Ecology and Evolution*, 13, 2404–2418. <https://doi.org/10.1111/2041-210X.13966>