Automated Filtering May Remove a Large Fraction of Records, But May Have Little Impact on Downstream Analyses

Alexander Zizka ‡

‡ Philipps-University Marburg, Marburg, Germany

Corresponding author: Alexander Zizka (alexander.zizka@biologie.uni-marburg.de)

Abstract

Species occurrence records provide the basis for many analyses in biodiversity research. They often derive from georeferenced specimens deposited in natural history collections or visual observations, such as those obtained through various mobile applications. Given the rapid increase in availability of such data, the control of quality and accuracy constitutes a particular concern. Automated flagging has emerged as a feasible way to identify potentially problematic records and correct them when possible, or, if a correction is not feasible, remove them from downstream analyses (automated filtering). A large number of flagging criteria exist, implemented in various software tools - reaching from simple flags, such as identifying zero coordinates, to more complex flags based on external data. These flags are widely applied for data cleaning in large-scale biodiversity analyses, yet their effect is rarely quantified.

Here, I present the rationale behind the automated flags included in the CoordinateCleaner R package (https://docs.ropensci.org/CoordinateCleaner/). CoordinateCleaner implementents functions to identify recurrent errors in biodiversity data compiled from different sources, and provides a user-friedly way to apply those functions in a pipeline to any dataset that can be loaded into R. Furthermore, I illustrate how many records these flags remove when used as filters across 18 taxa of plants, animal, and fungi from tropical America (Zizka et al. 2020b), and show results on the effect of automated filtering on the accuracy of automated conservation assessments using cleaned species occurrence records from biological collections (Zizka et al. 2020a).

The results show that, on average, almost half of the records in most datasets are flagged as potentially problematic, with large variation across taxonomic groups. Only around 5% of records was identified as erroneous in the strict sense, but a much larger proportion (20%) as unfit for common downstream analyses. Interestingly, automated filtering had little effect on the accuracy of automated red listing despite the large amount of records removed. This discrepancy may be caused by an interaction of data quality with sampling

bias and the overall under-sampling of species ranges. Automated record filtering can help in identifying problematic records, but requires customization of tests and thresholds to the taxonomic group and geographic area under focus. The results stress the importance of thorough recording and exploration of the meta-data associated with species records for biodiversity research.

Keywords

CoordinateCleaner, red listing, automated filters, data quality, GBIF, collection specimen

Presenting author

Alexander Zizka

Presented at

TDWG 2022

Conflicts of interest

References

- Zizka A, Silvestro D, Vitt P, Knight T (2020a) Automated conservation assessment of the orchid family with deep learning. Conservation Biology 35 (3): 897-908. https://doi.org/10.1111/cobi.13616
- Zizka A, Antunes Carvalho F, Calvente A, Rocio Baez-Lizarazo M, Cabral A, Coelho JFR, Colli-Silva M, Fantinati MR, Fernandes M, Ferreira-Araújo T, Gondim Lambert Moreira F, Santos NMC, Santos TAB, dos Santos-Costa RC, Serrano F, Alves da Silva AP, de Souza Soares A, Cavalcante de Souza PG, Calisto Tomaz E, Vale VF, Vieira TL, Antonelli A (2020b) No one-size-fits-all solution to clean GBIF. PeerJ 8 https://doi.org/10.7717/peerj.9916