Influence of different data cleaning solutions of point-occurrence records on downstream macroecological diversity models

Petra Fuehrding-Potschkat[1], Holger Kreft[1], and Stefanie Ickert-Bond[2]

[1]University of Göttingen Faculty of Forest Sciences and Forest Ecology
[2]University of Alaska Fairbanks

October 14, 2021

Abstract

Digital point-occurrence records from the Global Biodiversity Information Facility (GBIF) and other repositories enable a wide range of research in macroecology and biogeography. However, data errors may hamper immediate use. Manual data cleaning is time-consuming and often unfeasible, given that the databases may contain thousands or millions of records. Automated data cleaning pipelines are therefore of high importance. This study examined the extent to which cleaned data from six pipelines using data cleaning tools (e.g., the GBIF web application, different R packages) affect downstream species distribution models. In addition, we assessed how the pipeline data differ from expert data. From 13,889 North American Ephedra observations in GBIF, the pipelines removed 31.7% to 62.7% false-positives, invalid coordinates, and duplicates, leading to data sets that included between 9,484 (GBIF application) and 5,196 records (manual-guided filtering). The expert data consisted of 703 thoroughly handpicked records, comparable to data from field studies. Although differences in the record numbers were relatively large, stacked species distribution models (sSDM) from the pipelines and the expert data were strongly related (mean Pearson's r across the pipelines: 0.9986, versus the expert data: 0.9173). The ever-stronger correlations resulted from occurrence information that became increasingly condensed in the course of the workflow (from individual occurrences to collectivized occurrences in grid cells to predicted probabilities in the sSDMs). In sum, our results suggest that the R package-based pipelines reliably identified invalid coordinates. In contrast, the GBIF-filtered data still contained both spatial and taxonomic errors. However, major drawbacks emerge from the fact that no pipeline fully discovered misidentified specimens without the assistance of expert taxonomic knowledge. We conclude that application-filtered GBIF data will still need additional review to achieve higher spatial data quality. Achieving high-quality taxonomic data will require extra effort, probably by thoroughly analyzing the data for misidentified taxa, supported by experts.

Hosted file

`M1 Main Document 21-09-20 linenumbers.docx` available at https://authorea.com/users/441045/articles/541642-influence-of-different-data-cleaning-solutions-of-point-occurrence-records-on-downstream-macroecological-diversity-models