Addressing the need for interactive, efficient, and reproducible data processing in ecology with the datacleanr R package

JournalArticle (Originalarbeit in einer wissenschaftlichen Zeitschrift)

ID 4659525
Author(s) Hurley, Alexander G.; Peters, Richard L.; Pappas, Christoforos; Steger, David N.; Heinrich, Ingo
Author(s) at UniBasel Peters, Richard ;
Year 2022
Title Addressing the need for interactive, efficient, and reproducible data processing in ecology with the datacleanr R package
Journal PLoS ONE
Volume 17
Number 5
Pages / **Article-Number** e0268426
Mesh terms Data Analysis; Programming Languages; Reproducibility of Results; Software; Workflow
Ecological research, just as all Earth System Sciences, is becoming increasingly data-rich. Tools for processing of ”big data” are continuously developed to meet corresponding technical and logistical challenges. However, even at smaller scales, data sets may be challenging when best practices in data exploration, quality control and reproducibility are to be met. This can occur when conventional methods, such as generating and assessing diagnostic visualizations or tables, become unfeasible due to time and practicality constraints. Interactive processing can alleviate this issue, and is increasingly utilized to ensure that large data sets are diligently handled. However, recent interactive tools rarely enable data manipulation, may not generate reproducible outputs, or are typically data/domain-specific. We developed datacleanr, an interactive tool that facilitates best practices in data exploration, quality control (e.g., outlier assessment) and flexible processing for multiple tabular data types, including time series and georeferenced data. The package is open-source, and based on the R programming language. A key functionality of datacleanr is the ”reproducible recipe”-a translation of all interactive actions into R code, which can be integrated into existing analyses pipelines. This enables researchers experienced with script-based workflows to utilize the strengths of interactive processing without sacrificing their usual work style or functionalities from other (R) packages. We demonstrate the package's utility by addressing two common issues during data analyses, namely 1) identifying problematic structures and artefacts in hierarchically nested data, and 2) preventing excessive loss of data from 'coarse,' code-based filtering of time series. Ultimately, with datacleanr we aim to improve researchers' workflows and increase confidence in and reproducibility of their results.
Publisher Public Library of Science
ISSN/ISBN 1932-6203
edoc-URL https://edoc.unibas.ch/92737/
Full Text on edoc No;
Digital Object Identifier DOI 10.1371/journal.pone.0268426
PubMed ID http://www.ncbi.nlm.nih.gov/pubmed/35551557
ISI-Number MEDLINE:35551557
Document type (ISI) Journal Article