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

AG Hurley[1], *RL Peters*[2,3], *C Pappas*[4,5], *D Steger*[1,6], *I Heinrich*[1,7]

1 - GFZ, Potsdam, DE | 2 - WSL, Birmensdorf, CH | 3 - Ghent Uni, Ghent, BE | 4 - UTEQ, Montreal, CA | 5 - UQ, Montreal, CA | 6 - HU, Berlin, DE | 7 - DAI, Berlin, DE

hurley@gfz-potsdam.de
@aglhurley
the-hull.github.io/datacleanr

Background and Issue

Ecology is increasingly data intensive, posing new technical and logistical challenges[1]. Yet, few tools fully cater to best practices in analyses and publishing of "big data"[2] - often from different sources and types.

Existing tools are frequently:

- method-specific
- closed-source / costly
- complex and bespoke (code)
- not always interoperable
- interactive, but not reproducible
- reproducible, but not flexible or interactive

Solution

Diligent processing of large datasets will require interactive tools[3,4] to guarantee that any idiosyncrasies or patterns are identified and properly addressed (e.g., in deeply nested data), while enabling best practices, enhancing workflows, and ensuring reproducibility.

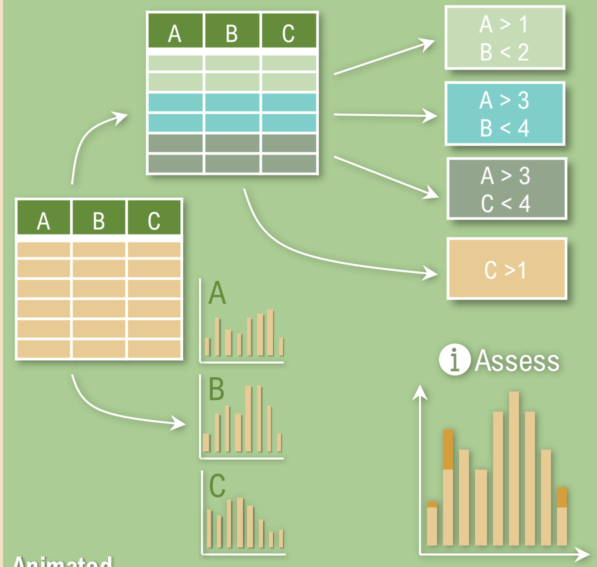
datacleanr provides **interactive** processing of **tabular**, **temporal** and **spatial** data. It facilitates best practices in exploration, data cleaning and outlier handling. Reproducibility is ensured by generating a "**code recipe**" that repeats all interactive steps. This recipe can be slotted into existing QA / QC or analyses pipelines. With it's **intuitive GUI**, *datacleanr* can be used by a wide audience.

Implementation: 4 Intuitive Modules

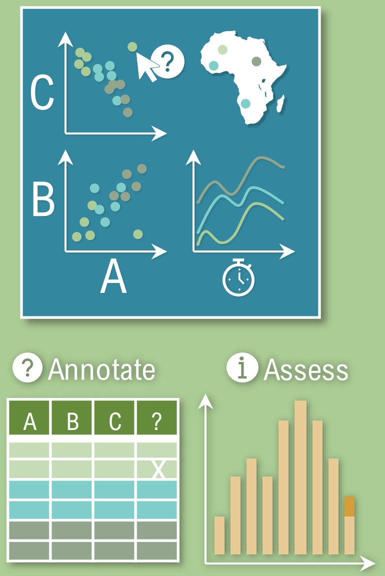
1. Define nesting groups and explore
2. Filter (value / statistics-based)
3. Visualize and annotate conspicuous data
4. Extract reproducible code recipe

1. Grouping and Exploration

2. Filtering Statements



3. Visualization and Annotation



4. Extract Reproducible Recipe

④ # Meta info -----------
Load data

① Group data by 🟩🟦▨

② Filter data (across groups)

③ Add annotations ?

④ Save data_annotations
Save data_clean

Set Folders and Naming

Send to RStudio

» Slot into Pipeline »

Animated examples: https://git.io/JkKp7 https://git.io/JkKpb https://git.io/JkKpj https://git.io/JkKhJ

Development and Details

- Open-source (R[5], shiny[6], plotly[7], etc.)
- Datasets up to approx. 2 million observations
- Modular code - easy to maintain and extend
- Unit tests for important features
- Use feedback, improve performance, add features

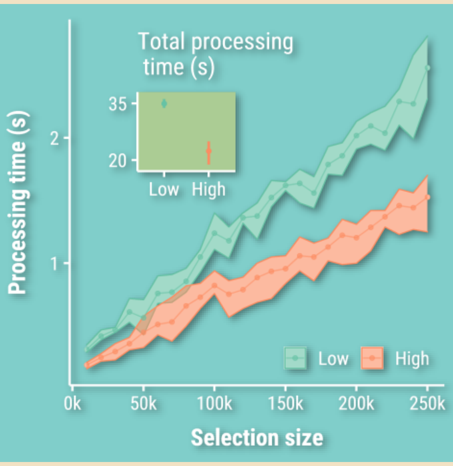


FIGURE 1: Performance tests for interactive visualization (scatter plot) with successive selection of 10k points (totaling 250k) showed efficient and comfortable processing times even at improbable selection scales. The tests were repeated 3 times (points: mean; bands: min/max) on low and high CPU power settings on a mid-end, mobile workstation.

Conclusion

datacleanr is our contribution toward free, open-source tools for reproducibly processing ecological and Earth System data. Through its design, we hope it will enable scientists across programming skill levels to deal with their big and messy data transparently and efficiently.

Sources

[1] Schimel, D., & Keller, M. (2015). *Oecologia*, *177*(4), 925–934. doi:10.1007/s00442-015-3236-3
[2] Farley, S. S., *et al*. (2018). BioScience, 68(8), 563–576. doi:10.1093/biosci/biy068
[3] Beilschmidt, C., *et al*. (2017). *Datenbank-Spektrum*, *17*(3), 233–243. doi:10.1007/s13222-017-0266-5
[4] Binnig, C., *et al*. (2015). Towards interactive data exploration. In *Real-time business intelligence and analytics* (pp. 177–190). Springer.
[5] R Core Team. (2020). https://www.R-project.org/
[6] Chang, W., *et al*. (2020). https://CRAN.R-project.org/package=shiny
[7] Sievert, C. (2020). https://plotly-r.com

General: Font Awesome, https://fontawesome.com/license