Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-112-RC2, 2020

Open Access Earth System Science Data Discussions

Interactive comment on "The *fortedata* R package: open-science datasets from a manipulative experiment testing forest resilience" *by* Jeff W. Atkins et al.

Anonymous Referee #2

Received and published: 12 November 2020

General comments:

The fortedata R package is an easy to use platform to access the FoRTE project datasets, which are logically organized and of general value. I have no reservations that the package will be useful as the project matures and see no issues with its main functionality. Nevertheless, I have four main areas of critique to bring the package to publication quality: (1) the limited documentation within the R package, (2) the usefulness of the vignettes, (3) the lack of clarity on present and future data availability, and (4) individual comments broken down by dataset under the "Specific comments" header below.

For publication, the package should include additional annotation, description of data, and the fleshing out of package documentation within R. As an example, the fd_inventory() help file lacks description other than "Raw inventory table" and a note on collection using a Haglof Postex Inventory Unit. A few extra steps do get the user to the metadata table with a description of variables, but I see value in fleshing out the description here to the level done in the data paper. There are also notes like that in the help file for fd_canopy_structure_summary() that simply says "For now this is pretty basic." And, there are instances like in the help file for fd_subplots() where the listed columns don't match those of the dataset. Some documentation is missing from the submitted ESSD preprint and SI. There are functions in the package that are not described such as forte_colors() and plot_metadata(). Pointing out these issues is not to nitpick these particular details, but I think these small details add up to make the package feel in development rather than a final product for outside users.

Similarly, the vignettes (https://fortexperiment.github.io/fortedata/articles/index.html) are promising and needed to familiarize outside users with the package and data but only two of seven vignettes plot or manipulate the data. The other vignettes are descriptive of the project or just show the datasets being called. Vignettes are not needed to show the dataset being called. Rather, a handful of more illustrative uses of the package would be better as vignettes. Even the more fleshed out vignette (https://fortexperiment.github.io/fortedata/articles/fd_inventory_vignette.html) only shows a density plot by replicate. Again, the vignette doesn't really help the user parse the dataset much since there is no breakdown by species, subplot, etc. Certainly the group has made more interesting and complete visualizations/analyses that can be shown like your Figure 3. These vignettes don't help the outside user get a head start on analyzing these data.

Information on anticipated release schedules for new data (and for which variables) is needed, or at least a statement about where that information can be found once known. Will additional sites be added? What additional data will be released for the

years 2019 and 2020? The selling point of near real-time data availability through the R package is compelling, but Figure 2 does not convince that there is real-time data being released (e.g., no spectrometry or photosynthesis data since mid-2018). Are these data to be made available following separate publication or do they not exist? As a naïve user, I need to know if what I am analyzing is complete. The R package is still a suitable platform to distribute these data but these details are necessary to make the data publicly usable and not just publicly available.

Specific comments:

The following are comments broken down by dataset: fd_inventory: Explain/note data with missing date information, species codes marked ???? (unidentifiable species?) for DBH inventory. What is the column "tag" in the fd_inventory data set? It lacks a description in the SI. This appears to maybe be an index column but there are a few errors in the numbering. "tag" 2236-2244 have an erroneous 9 in front of them, it appears. Should be DP II instead of PD II caliper, presumably.

fd_soil_respiration: 2,791 observations are in the dataset when loaded through the R package. Again there are missing timestamp values (1,622 or over half the data, which even if the date is available is notable).

fd_leaf_spectrometry: Was this dataset exported as only the head of the data? Only 8 rows of data import using the R package when 7,155 are expected. There is an additional column "tree_id" that is not in the format of USDA PLANTS species codes and not defined under Table S6 in the SI.

fd_litter: Is "MISC" code equivalent to "MIX" as defined in L184 or is it actually Mikania scandens? What are the codes "SWD" and "FAGRE", these don't appear to be USDA PLANTS codes? Is there a reason to not just use a column of the actual litter mass rather than the intermediate columns for bag mass and bag+litter mass?

fd_hemi_camera: Again a mismatch between reported observations and the number

in the R package dataset.

fd_canopy_structure: Again a mismatch between reported observations and the number in the R package dataset. In the associated Table S10 variables are separated by periods instead of underscores as in the actual data and the other SI tables. There are additional undescribed variables such as the skew and kurtosis intensity missing from Table S10.

L158: fd_plot_metadata() is not how the function appears in the R package. The help file in the R package does not describe how to use it to get the metadata properly.

fd_metadata(table = "fd_inventory") returns the whole metadata tibble.

Technical corrections: No major issues found.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-112, 2020.