HoboR: An R package to manipulate weather stations data

Alcalá Briseño RI, Carson AR, Lang S, Peterson E, and LeBoldus, J.

January 19, 2024

Summary

`HoboR` is an R package for efficiently processing extensive datasets obtained from HOBO weather stations and data loggers. I developed multiple tools designed for streamlined weather data supporting various weather station formats. Existing satellite weather data manipulation packages are available in R, such as NASA Power and rnoa (Sparks 2018, Chamberlain and Hocking, 2023), but not for weather stations and data loggers. `HoboR` facilitate users to load CSV files into a tibble format, eliminate duplicates, summarize data by time intervals (minutes, hours, and days), subset files by date ranges, and address common data quality and accuracy issues due to sensor failures identifying out-of-range entries, time zone discrepancies, and correcting data tools. Additionally, the package incorporates guidelines for weather data analysis (REF), advocating for adherence to standard practices in handling weather variables. Despite its name, `HoboR` is adaptable to other weather station formats sharing a similar data structure.

Weather station data can be logged by the minute at any point in time and from different types of sensors such as rain, relative humidity (RH), light, and more. `HoboR` main functions implement dynamic interpretation programming, allowing the processing of spreadsheets independently for any number of sensors, adjusting to diverse initial column structures. Among the difficulties in recording and collecting data are the errors that occur when replacing batteries, downloading data, and malfunctioning. These issues create multiple entries that might be challenging and time-consuming to handle in tabular data interfaces. These tools seamlessly facilitate data manipulation, merging, and summarizing, promoting reproducibility and freeing up time for further analysis and modeling.

`HoboR` was tested on log files with hundreds to thousands entries, facilitating the post-processing of weather station and data loggers, loading csv files regardless of the header column order and dimensions, and summarised within seconds, the summary statistics can be rounded to minutes, hours and days, yielding minimums and maximum, mean and standard deviation of your data. Additional functions can help to identify and replace impossible values and correct the variation within your loggers. As a proof of concept applied to these weather data, we implement a couple of functions to calculate disease

trends of the sudden oak death epidemiology affecting tanoek (*Notholitocarpus densiflorus*) in the Pacific Northwest.

Statement of need

Developing automated software for preprocessing weather stations and data logger information may facilitate the analysis of epidemiological surveillance, microbiome, and multiple disciplines (Dahl et al., 2023; Nikolauo et al., 2023; Wu et al., 2023). Traditional spreadsheet interfaces pose a challenge in handling extensive and complex studies that are difficult to manage, time-consuming to organize, error-prone if done by hand, and might not handle whole datasets. By automating these tasks, `hoboR` enhances accuracy and significantly reduces the time and effort required for data preparation, leaving more time for robust epidemiological modeling. The integration of advanced algorithms and user-friendly software makes it accessible to both experienced researchers and program beginners, addressing the current potential of implementing weather variables for plant pathology and disease ecology for effective management (Garrett et al., 2023).

To our knowledge, no packages in R are available online for the analysis of weather station and data logger files. A graphic user interface for HOBO exists but is incompatible with data postprocessing and data manipulation.

Package workflow

The workflow of the `HoboR` package consists of three consecutive steps and seven assisting functions:
– `hobinder`: Load multiple csv files regardless of the order and number of columns from a single directory; the files must come from the same weather station or data logger model.
– `hobocleaner`: Averages duplicate entries from the large csv file.
– `meanhobo`: Summary statistic (min, max, mean, and standard deviation) for the different weather station and data logger sensors
– `hobotime`: Allows aggregating your data by minutes, hours, or days.
– `horange`: Allows to parse your data by date ranges,
– `impossiblevalues`: Identify the min and max values in the data set, the user should consider what are the minimum and maximum values for the region.
– `NAsensorfailures`: Allows to replace with NAs impossible values in your data set using logical statements.
– `timestamp`: Select a time and give the same time interval for a select length
– `horrelation`: Plot what weather variable correlates among them.
– `hoboplot`: Plot weather variable trends.

Example

A test dataset is provided with the `HoboR` package. This data set was collected in China Creek, Brookins, Oregon, between August to December 2021 (Fig. 1). We tested in partial datasets from different weather stations and data loggers and a full dataset of million entries. The test was carried out on a MacBook Pro (2022, 16 GB RAM, M2) and a Dell PC (2016, 8 GB, Intel 5). The code and results are reproduced below:

```
library(hoboR)
# Add the PATH to your sites for weather data (from hobo)
path = ("Documents/site_1")
files <- hobinder(path, header = T, skip = 1) # loading all hobo files

# remvoe duplicate entries
cleaned <- hobocleaner(files, format = "ymd")

# get the summary statistics by time ("5 min", "1 hour"', "1 day")
summary <- meanhobo(cleaned, summariseby = "1 day",  na.rm = T)

# data quality assesment
hobotime(cleaned, summariseby = "5 mins", na.rm = T) # rounds data every 5 minutes
horange(cleaned, start = "2022-08-04", end = "2022-08-10") # select a time window
impossiblevalues(cleaned, showrows = 3) # show impossible values

# flag impossible values to NA
NAsensorfailures(cleaned, condition = ">", threshold = c(50, 3000, 101), opt = c("Temp", "R
# get the same date by time frame
timestamp(cleaned, stamp = "2022-08-05 00:01", by = "24 hours", days = 100, na.rm = TRUE, p
```

Fig. 1) shows the output variables using plotweather() function.

Installation

This package requires R version 4.1.3 or later. It also requires the following packages: data.table, dplyr, ggplot2, lubridate, plyr, purrr. These dependencies should be installed automatically when dependencies = TRUE is set in the command used to install the package.

```
> if (!require("devtools")) \\
> install.packages("devtools")\\
> devtools::install_github("leboldus_lab/hoboR", dependencies = TRUE)
```

Authors contribution

Ricardo I. Alcalá Briseño developed the original version of the package, maintained the package, wrote the documentation, debugged the code, and wrote the manuscript. Adam R. Carson collected the data, wrote code implemented in the package, and debugged the code. Sky Lang collected the data and assisted in the user-functionality of the code functions. Ebba Peterson assisted in best practices for post-processing. Jared LeBoldus supervised the project and participated in the manuscript drafting process.

Acknowledgements

Grant money No. 1234567890

References

Sparks A (2018). "nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R." The Journal of Open Source Software, $*\ 3\ *$ $(30), 1035. doi : 10.21105/joss.01035 < https : //doi.org/10.21105/joss.01035 >$.

Chamberlain, S., Hocking, D. (2023). rnoaa: 'NOAA' Weather Data from R (Version 1.4.0). Retrieved from https://CRAN.R-project.org/package=rnoaa

Garrett et al., 2023 https://doi.org/10.1146/annurev-phyto-021021-042636

Dahl et al., 2023, https://doi.org/10.1111/1462-2920.16347

Nikolauo et al., 2023, https://doi.org/10.1016/j.envres.2023.117173

Wu et al., 2023, https://doi.org/10.1093/aob/mcad195