HOBOR: An R package to summarize and manipulate weather station data.

Alcalá Briseño R.I., Carson A.R., Lan Y-H, Peterson E, NJ, Grunwald, and LeBoldus, J.M.

January 19, 2024

Summary

HOBOR is an R package for efficiently processing large datasets obtained from HOBO (ONSET, United Kingdom) weather stations and data loggers. I developed multiple tools designed to curate and summarize weather data in various formats. Packages to analyze existing weather data captured by satellites are available in R, including NASA Power and rnoa (Sparks 2018, Chamberlain and Hocking, 2023). Similar packages have not been developed for weather stations and data loggers. HOBOR allows users to load CSV files into a tibble format, eliminate duplicates, summarize data by time interval (minutes, hours, and/or days), and subset files by date ranges. The package can also address common data quality and accuracy issues related to sensor failures, identify out-of-range entries and time zone discrepancies, and correct data formats. Despite its name, HOBOR is adaptable to other weather station output formats with a similar data structure.

Weather station data can be logged at a variety of time intervals from different types of sensors, including rain gages, relative humidity sensors (RH), and light meters. HOBOR main functions implement dynamic interpretation programming, allowing the processing of spreadsheets independently for any number of sensors, and can adjust to a range of initial column structures. Among the difficulties in recording and collecting data are the errors that occur when replacing batteries, downloading data, and from malfunctioning sensors or loggers. These issues can create multiple entries that might be challenging and time-consuming to detect and curate in tabular data interfaces.

HOBOR tools seamlessly facilitate data manipulation, merging, and summarization, reducing curation time prior to downstream analysis and modeling. HOBOR was tested using csv files with hundreds to thousands of entries, facilitating the post-processing, loading csv files with variable header column order and dimensions, and summarizing data within seconds. The summary statistics can be rounded to the nearest minute, hour, or day. Outputs include minimum, maximum, mean, and standard deviation of the data. Additional functions can help to identify and replace unrealistic values and correct the variation across data loggers.

Statement of need

Developing automated software for preprocessing weather station and data logger information

may facilitate the analysis of epidemiological surveillance and microbiome across disciplines

(Dahl et al., 2023; Nikolauo et al., 2023; Wu et al., 2023 ). Traditional spreadsheet

interfaces pose a challenge when handling data from large and complex studies. The management and curation of these datasets are time-consuming and error-prone if done by hand. In many cases, the spreadsheet-based interfaces might not be able to handle an entire dataset at once. Automation of these tasks in HOBOR enhances accuracy and significantly reduces the time and effort required for data handling and management. The integration of advanced algorithms and user-friendly software makes it accessible to both experienced researchers and program beginners, addressing the current potential for implementing weather variables in 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 : loads multiple CSV files some columns 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 : clean duplicate entries and column names format from loaded CSV files, need to specify the data format e.g., "ymd"
* meanhobo : summary statistic (min, max, mean, and standard deviation) of the weather variables summarize by time intervals
* horange : selects a range of dates
* hobotime : aggregates data by time in minutes, hours or days
* impossiblevalues : shows the maximum and minimum values and displys a number of values
* sensorfailures : detect the sensor failures and impossible values based on specific threshold measurements
* timestamp : gets an interval snapshot from selected date and time frequency on a range of days
* calibration : collect controlled weather data and calculate the variability of HOBO data loggers and base the calibration
* correction.test : use the result of the calibration process to test the accuracy of the data logger by selecting a threshold difference
* – correction : correct the experimental data using the weather variable of interest
* horrelation : display the correlation between variables in plot format
* testhobolist : if calibration() or correction() do not compute, test the list of hobo dataframes checks data viability
* samplingrates : calculate the total of samples collected
* summarybates : summarises the weather data by sample collection

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 millions of

entries. The test was carried out on a Dell PC (2016, 8 GB, Intel 5), a MacBook Pro (2022, 16 GB RAM, M2) and Ubuntu 22, Linux (2022, 128 Gb, RTX A4000) . The code and results are reproduced below:

```R

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

cleaned <- hobocleaner(files, format = "ymd")

summary <- meanhobo(cleaned, summariseby = "1 day", na.rm = T)

# data quality assessment

hobotime(cleaned, summariseby = "5 mins", na.rm = T) # rounds data every 5 minutes

hoborange(cleaned, start = "2022-08-04", end = "2022-08-10") # select a time window

impossiblevalues(cleaned, showrows = 3) # show impossible values

sensorfailures(cleaned, condition = ">", threshold = c(50, 3000, 101), opt = c("Temp", "Rain", RH)

timestamp(cleaned, stamp = "2022-08-05 00:01", by = "24 hours", days = 100, na.rm = TRUE, plot = TRUE) # shows the trends by time range .

InstallationL:

This package requires R version 4.3.0 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 Lan collected the data, wrote code implemented in the package, and assisted in the user-functionality of the code functions. Ebba Peterson assisted in best practices for post-processing. Jared M. 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