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

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

**Summary**

Meteorological records captured by weather stations and data loggers can accumulate large amounts of digital information, generating sizeable data projects and complex data analysis for non-expert users. Local and regional efforts to create weather station networks, as well as projects oriented to the study of microclimates and smart agriculture, have increased the use of weather stations in the last couple of decades (Estévez et al., 2011; Lembrechts et al., 2021; Hachimi et al., 2022). HOBO (ONSET, United Kingdom) is among the most popular weather stations and data loggers. A HOBO graphic user interface software exists but is incompatible with post-process data analysis and big data. To close this gap, we have developed a series of algorithms to combine and manipulate several csv files, allowing the removal of redundant data and the summarization of the meteorological data by time and date, identification of sensor failures and impossible values, and calculation of the summary statistics.

**Statement of need**

HoboR is an R package (R Core Team, 2024) for efficiently processing large datasets obtained from weather stations and data loggers. We developed multiple tools designed to import, curate, and summarize weather data in csv format. Packages to analyze weather data exist in Turbo Pascal (Pickering et al., 1994), and current R packages to analyze weather data captured by satellites include NASA Power and rnoa (Sparks 2018, Chamberlain and Hocking, 2023). HoboR allows users to load csv files into a data frame and is automatically adaptable to HOBO sensor configurations e.g. temperature, residual humidity, precipitation, leaf wetness, allowing the removal of duplicate entries, summarizing the data by time intervals (minutes, hours, and/or days), and subset files by date ranges. The package can also identify and address common data quality and accuracy issues related to sensor failures, out-of-range entries and time zone discrepancies, and correct data formats.

Developing software to automate the processing of data collected by weather stations and data loggers can facilitate the analysis of local weather and microclimate patterns and projects aimed at correlating meteorological data with epidemiological processes, ecological species composition, and smart agriculture, among other biological sciences (Hachimi et al., 2022, Dahl et al., 2023; Nikolauo et al., 2023; Wu et al., 2023 ). Automating these tasks in hoboR enhances accuracy and significantly reduces the time and effort required for data handling and management, increasing the reproducibility of the analysis. Conversely, traditional spreadsheet interfaces pose an inherent challenge when handling data from large and complex studies, making managing and curating these datasets time-consuming and error-prone. In many cases, the spreadsheet-based interfaces might not be able to handle an entire dataset at once or allow for the reduction of redundant data. The integration of advanced algorithms and user-friendly software makes hoboR 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.

HOBOR tools seamlessly facilitate data manipulation, merging, and summarization, thereby reducing curation time prior to downstream analysis and modeling. Weather station data can be logged at various time intervals for different types of sensors, including rain gauges, temperature, relative humidity (RH), and radiation, among others. The main functions of hoboR implement dynamic interpretation programming, enabling the processing of spreadsheets independently for any number of sensors, and can transpose data into a range of initial column structures. Among the challenges in recording meteorological data are the various errors that occur during data collection e.g. the sensors can be saturated, debris blocking the sensors, battery replacement, and malfunctioning sensors or loggers. These issues can result in multiple entries that might be challenging and time-consuming to detect, correct, and curate in tabular data. HoboR was tested on field data processing thousands of entries, facilitating loading the csv files with variable header column order and dimensions, processing, and summarizing the data. The summary statistics can be rounded to the nearest minute, hour, or day. Outputs include the minimum, maximum, mean, and standard deviation and other functions that can help summarize the data by time intervals and range of dates. Additionally, we provide functions that can help identify unrealistic values and a framework to calibrate and correct the variation across data loggers (Fig. 1).

A diagram of a computer

Description automatically generated

Fig. 1) A workflow illustrating the steps of the recommended HOBOR package: data parsing, summary and subset of entries, quality checking, and summary statistics results. Optional calibration steps for HOBO data loggers. Discontinuous lines are optional; solid lines represent the recommended pipeline for HOBO data analysis.

**Example**

The hoboR package includes a test dataset collected in Brookins, Oregon, between August and December 2021 (Fig. 2). We provided a partial dataset using different weather stations and data loggers to test the package. The code is reproduced below.

```R

library(hobor)

# Standard Analysis

# Add the PATH to your sites for weather data (from hobo)

path = ("./site\_12")

files <- hobinder(path, header = T, skip = 1) # loading all hobo files

cleaned <- hobocleaner(files, format = "yymd") # remove duplicate entries

sum <- hobotime(cleaned, summariseby = “5 mins”, na.rm = T) # rounds data every 5 minutes

summa <- hoborange(sum, start = "2022-08-04", end = "2022-12-31") # select a time range

summary <- meanhobo(summa, summariseby = "1 day", na.rm = T) # get the summary statistics by "24 h"

# Quality check

impossiblevalues(cleaned, showrows = 6) # show impossible values

sensorfailures(cleaned, condition = ">", threshold = c(50, 3000, 101), opt = c("Temp", "Rain", RH) # flag impossible values to NA

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, reshape, lubridate, plyr and ggplot2. These dependencies should be installed automatically when dependencies = TRUE is set in the command used to install the

package.

```R

> if (!require("devtools")) \\

> install.packages("devtools")\\

> devtools::install\_github("leboldus\_lab/hobor", dependencies = TRUE)

```

**Authors contributions**

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 the post-processing of weather stations and data loggers. Niklaus J. Grunwald participated in manuscript preparation and funding. Jared M. LeBoldus supervised the project, manuscript preparation, and funding.

**Acknowledgements**

Grant money No. 1234567890

**Disclosure**

Authors declare no conflict of interest.

**References**

@article{Estevez2011,

title = {Guidelines on validation procedures for meteorological data from automatic weather stations},

journal = {Journal of Hydrology},

volume = {402},

number = {1},

pages = {144-154},

year = {2011},

issn = {0022-1694},

doi = {https://doi.org/10.1016/j.jhydrol.2011.02.031},

url = {https://www.sciencedirect.com/science/article/pii/S0022169411001594},

author = {J. Estévez and P. Gavilán and J.V. Giráldez},

keywords = {Quality control, Validation, Agrometeorology, Weather network},

}

@article{Lembrechts2021,

author = {Lembrechts, Jonas J. and Lenoir, Jonathan and R. Scheffers, Brett and De Frenne, Pieter},

title = {Designing countrywide and regional microclimate networks},

journal = {Global Ecology and Biogeography},

volume = {30},

number = {6},

pages = {1168-1174},

keywords = {biogeography, climate change, data loggers, location selection, microclimate, microweather, sensors, temperature},

doi = {https://doi.org/10.1111/geb.13290},

url = {https://onlinelibrary.wiley.com/doi/abs/10.1111/geb.13290},

eprint = {https://onlinelibrary.wiley.com/doi/pdf/10.1111/geb.13290},

year = {2021}

}

@Article{Hachimi2023,

Author = {Hachimi, Chouaib El and Belaqziz, Salwa and Khabba, Saïd and Sebbar, Badreddine and Dhiba, Driss and Chehbouni, Abdelghani},

Title = {Smart Weather Data Management Based on Artificial Intelligence and Big Data Analytics for Precision Agriculture},

Journal = {Agriculture},

Volume = {13},

Year = {2023},

Number = {1},

Article-Number = {95},

url = {https://www.mdpi.com/2077-0472/13/1/95},

issn = {2077-0472},

doi = {10.3390/agriculture13010095}

}

@Manual{R2024,

title = {R: A Language and Environment for Statistical Computing},

author = {{R Core Team}},

organization = {R Foundation for Statistical Computing},

address = {Vienna, Austria},

year = {2024},

url = {https://www.R-project.org/},

}

@article{Pickering1994,

author = {Pickering, N. B. and Hansen, J. W. and Jones, J. W. and Wells, C. M. and Chan, V. K. and Godwin, D. C.},

title = {WeatherMan: A Utility for Managing and Generating Daily Weather Data},

journal = {Agronomy Journal},

volume = {86},

number = {2},

pages = {332-337},

doi = {https://doi.org/10.2134/agronj1994.00021962008600020023x},

url = {https://acsess.onlinelibrary.wiley.com/doi/abs/10.2134/agronj1994.00021962008600020023x},

eprint = {https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2134/agronj1994.00021962008600020023x},

}

@article{Sparks2018,

doi = {10.21105/joss.01035},

url = {https://doi.org/10.21105/joss.01035},

year = {2018},

publisher = {The Open Journal},

volume = {3}, number = {30},

pages = {1035},

author = {Adam H. Sparks},

title = {nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R},

journal = {Journal of Open Source Software} }

@Manual{Chamberlain2024,

title = {rnoaa: 'NOAA' Weather Data from R},

author = {Scott Chamberlain and Daniel Hocking},

year = {2024},

note = {R package version 1.4.0,

https://github.com/ropensci/rnoaa (devel)},

url = {https://docs.ropensci.org/rnoaa/ (docs)},

}

@article{Dahl2023,

author = {Dahl, Mathilde Borg and Kreyling, Juergen and Petters, Sebastian and Wang, Haitao and Mortensen, Martin Steen and Maccario, Lorrie and Sørensen, Søren J. and Urich, Tim and Weigel, Robert},

title = {Warmer winters result in reshaping of the European beech forest soil microbiome (bacteria, archaea and fungi)—With potential implications for ecosystem functioning},

journal = {Environmental Microbiology},

volume = {25},

number = {6},

pages = {1118-1135},

doi = {https://doi.org/10.1111/1462-2920.16347},

url = {https://enviromicro-journals.onlinelibrary.wiley.com/doi/abs/10.1111/1462-2920.16347},

eprint = {https://enviromicro-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.16347},

year = {2023}

}

@article{Nikolaou2023,

title = {Improved daily estimates of relative humidity at high resolution across Germany: A random forest approach},

journal = {Environmental Research},

volume = {238},

pages = {117173},

year = {2023},

issn = {0013-9351},

doi = {https://doi.org/10.1016/j.envres.2023.117173},

url = {https://www.sciencedirect.com/science/article/pii/S0013935123019771},

author = {Nikolaos Nikolaou and Laurens M. Bouwer and Marco Dallavalle and Mahyar Valizadeh and Massimo Stafoggia and Annette Peters and Kathrin Wolf and Alexandra Schneider},

keywords = {Relative humidity, Spatiotemporal modeling, Machine learning, External validation, Exposure assessment, Environmental epidemiology},

}

@article{Wu2023,

author = {Wu, Carrie and Powers, John M and Hopp, David Z and Campbell, Diane R},

title = "{Effects of experimental warming on floral scent, display and rewards in two subalpine herbs}",

journal = {Annals of Botany},

pages = {mcad195},

year = {2023},

month = {12},

issn = {0305-7364},

doi = {10.1093/aob/mcad195},

url = {https://doi.org/10.1093/aob/mcad195},

eprint = {https://academic.oup.com/aob/advance-article-pdf/doi/10.1093/aob/mcad195/55122486/mcad195.pdf},

}

@article{Garrett2022,

author = "Garrett, K.A. and Bebber, D.P. and Etherton, B.A. and Gold, K.M. and Plex Sulá, A.I. and Selvaraj, M.G.",

title = "Climate Change Effects on Pathogen Emergence: Artificial Intelligence to Translate Big Data for Mitigation",

journal= "Annual Review of Phytopathology",

year = "2022",

volume = "60",

number = "Volume 60, 2022",

pages = "357-378",

doi = "https://doi.org/10.1146/annurev-phyto-021021-042636",

url = "https://www.annualreviews.org/content/journals/10.1146/annurev-phyto-021021-042636",

publisher = "Annual Reviews",

issn = "1545-2107",

type = "Journal Article",

}