gosset: An R package for analysis and synthesis of ranking data in experimental agriculture

Kauê de Sousa[1,2], Jacob van Etten[1], David Brown[3,4], Jonathan Steinke[1,5]

Abstract

Appropriate data management and analysis are necessary to produce practical information from agronomy and crop science experimental data. There is also an ongoing trend advocating for programmatic tools that supports reproducible workflows in scientific research. Recently developed approaches for data analysis and synthesis based on ranking data lack of customized analytical software to facilitate data science workflows. We present the R package gosset, which provides functions to support data preparation, modelling, validation, and results presentation with rank-based models, such as Plackett-Luce and Bradley-Terry. We demonstrate the functionality of the package with a case of on-farm evaluations of common bean (*Phaseolus vulgaris* L.) varieties in Nicaragua.

data-driven agriculture, Plackett-Luce model

Required Metadata

Current code version

Ancillary data table required for subversion of the codebase. Kindly replace examples in right column with the correct information about your current code, and leave the left column as it is.

Table 1: *Code metadata (mandatory)*

Nr.	Code metadata description	Please fill in this column
C1	Current code version	0.4.003
C2	Permanent link to code/repository used for this code version	https://github.com/AgrDataSci/gosset
C3	Code Ocean compute capsule	
C4	Legal Code License	MIT
C5	Code versioning system used	Git
C6	Software code languages, tools, and services used	R
C7	Compilation requirements, operating environments & dependencies	
C8	If available Link to developer documentation/manual	https://agrdatasci.github.io/gosset/
C9	Support email for questions	k.desousa@cgiar.org

The permanent link to code/repository or the zip archive should include the following requirements: README.txt and LICENSE.txt.

Source code in a src/ directory, not the root of the repository.

Tag corresponding with the version of the software that is reviewed.

*Documentation in the repository in a docs/ directory, and/or README*s*, as appropriate.*

Motivation and significance

Introduce the scientific background and the motivation for developing the software.

Explain why the software is important, and describe the exact (scientific) problem(s) it solves.

Indicate in what way the software has contributed (or how it will contribute in the future) to the process of scientific discovery; if available, this is to be supported by citing a research paper using the software.

Provide a description of the experimental setting (how does the user use the software?).

Introduce related work in literature (cite or list algorithms used, other software etc.).

Reproducible and efficient workflows are fundamental in scientific research (Lowndes et al. 2017). Data analysis workflows roughly include the following stages: (1) Data preparation and cleaning, (2) modelling and validation, and (3) results presentation. It is also common that those stages would be repeated iteratively until a solution is found which satisfies the initial objectives. Every of these stages presents different difficulties and constraints for the researchers. Digital tools can facilitate the tasks within those stages, but it is necessary to choose the right tool, if any, for the intended work. While not frequently used, experimental approaches using ranking data are being applied for the evaluation of crop varieties. Recently developed rank-based approaches for on-farm experimentation, such as the tricot methodology (Van Etten et al., 2019), required customized tools for all the aforementioned stages. On the other, new rank-based data synthesis approaches also required tailored tools to facilitate all the involved stages. Along with those experiences, we developed the R package gosset, supporting several activities in the analysis of experiments in agronomy and crop science.

Software description

Describe the software in as much as is necessary to establish a vocabulary needed to explain its impact.

The R package gosset provides functionality supporting the data analysis workflows in experimental agriculture, especially with rank-based approaches in on-farm experimental agriculture. Typically, this data analysis workflow includes (1) data management and preparation, (2) modelling and (3) results visualization and presentation.

Software Architecture

Give a short overview of the overall software architecture; provide a pictorial component overview or similar (if possible). If necessary provide implementation details.

Software Functionalities

Present the major functionalities of the software.

Data management and preparation

When data from agricultural experiments is not in ranking format, it should be transformed to be used as inputs into R packages for the analysis of ranking data. For instance, the Plackett-Luce model (Luce, 1959;

Plackett, 1975) is implemented in the R package as PlackettLuce, which requires the data to be formatted as ranking matrix. Another example is the Bradley-Terry model (Bradley & Terry, 1952), implemented in the package BradleyTerry2 (Turner & Firth, 2012) and requires the input data to be formatted as paired comparisons. For these cases, gosset provides the functions rank_numeric and rank_binomial. The function rank_numeric transforms a set of numeric values into an ordinal ranking, considering if higher numeric values should be ranked first or not. The function rank_binomial transforms data in rankings format into pairwise comparisons, as required by the package BradleyTerry2. Additionally, gosset provides the function rank_tricot, for the case when the experimental data is generated from trial established using the Triadic Comparison of Technologies (tricot) approach (van Etten et al. 2019).

Modelling

The gosset package provides complementary functions for validation of models Bradley-Terry, Plackett-Luce, Generalized Linear and Generalized Nonlinear. The function pseudoR2 computes goodness-of-fit measures such as Cragg-Uhler (Cragg & Uhler, 1970) and McFadden's R2 (McFadden, 1973). The function AIC computes the Akaike Information Criterion (Akaike, 1974).

Visualization

Sample code snippets analysis (optional)

Illustrative Examples

Provide at least one illustrative example to demonstrate the major functions.

Optional: you may include one explanatory video that will appear next to your article, in the right hand side panel. (Please upload any video as a single supplementary file with your article. Only one MP4 formatted, with 50MB maximum size, video is possible per article. Recommended video dimensions are 640 x 480 at a maximum of 30 frames/second. Prior to submission please test and validate your .mp4 file at http : $//elsevier - apps.sciverse.com/GadgetVideoPodcastPlayerWeb/verification$. This tool will display your video exactly in the same way as it will appear on ScienceDirect.).

```
library(gosset)
```

Impact

This is the main section of the article and the reviewers weight the description here appropriately

Indicate in what way new research questions can be pursued as a result of the software (if any).

Indicate in what way, and to what extent, the pursuit of existing research questions is improved (if so).

Indicate in what way the software has changed the daily practice of its users (if so).

Indicate how widespread the use of the software is within and outside the intended user group.

Indicate in what way the software is used in commercial settings and/or how it led to the creation of spin-off companies (if so).

Conclusions

Set out the conclusion of this original software publication.

Conflict of Interest

No conflict of interest exists: We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

References

Please add the reference to the software repository if DOI for software is available.

Zenodo repository: https://doi.org/10.5281/zenodo.6339989

Lowndes, Julia S. Stewart, Benjamin D. Best, Courtney Scarborough, Jamie C. Afflerbach, Melanie R. Frazier, Casey C. O'Hara, Ning Jiang, and Benjamin S. Halpern. 2017. "Our Path to Better Science in Less Time Using Open Data Science Tools." *Nature Ecology & Evolution* 1 (6). https://doi.org/10.1038/s41559-017-0160.

van Etten, Jacob, Kauê de Sousa, Amílcar Aguilar, Mirna Barrios, Allan Coto, Matteo Dell'Acqua, Carlo Fadda, et al. 2019. "Crop variety management for climate adaptation supported by citizen science." *Proceedings of the National Academy of Sciences* 116 (10): 4194–99. https://doi.org/10.1073/pnas.1813720116.