gosset: An R package for analysis and synthesis of ranking data in agricultural experimentation

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

[1] Department of Agricultural Sciences, Inland Norway University of Applied Sciences, 2318 Hamar, Norway
[2] Digital Inclusion, Bioversity International, Parc Scientifique Agropolis II, 34397, Montpellier Cedex 5, France
[3] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB, Wageningen, The Netherlands
[4] Digital Inclusion, Bioversity International, 30501, Turrialba, Costa Rica
[5] Humboldt University Berlin, Berlin, Germany

[*] Correspondence should be addressed to: kaue.desousa@inn.no

Abstract

Appropriate data management and analysis are necessary to produce practical information from agricultural experimentation data. There is also an ongoing trend advocating for programmatic tools that supports reproducible workflows in scientific research. We developed the R package gosset, providing functionality to support analysis workflows with rank-based models, such as Plackett-Luce and Bradley-Terry. The gosset package facilitates data preparation, modelling and results presentation stages. We demonstrate the functionality of the package with a case of on-farm evaluations of common bean (*Phaseolus vulgaris* L.) genotypes in Nicaragua.

data-driven agriculture, Plackett-Luce model

Required Metadata

Current code version

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	kaue.desousa@inn.no

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.).

Participatory experimentation approaches have been increasingly applied in agricultural research (De Roo, Andersson, and Krupnik 2019). While collecting data in ranking format is uncommon in general agricultural research settings, it is often collected in participatory experiments (Coe 2002). Recently developed approaches for on-farm experimentation, such as the tricot methodology are based on the collection of data in ranking format (van Etten, Beza, et al. 2019). On the other hand, newly proposed approaches for synthesis of crop variety evaluation data largely depend on the analysis of ranking data (Brown et al. 2020). The analysis of ranking data requires the use of appropriate statistical models such as Plackett-Luce (Luce 1959; Plackett 1975) or Bradley-Terry (Bradley and Terry 1952). Functionality for fitting Bradley-Terry and Plackett-Luce models are available in R with the packages BradleyTerry2 and PlackettLuce respectively (H. Turner and Firth 2012; H. L. Turner et al. 2020). However, extended functionality was required for the entire data science workflow, which usually includes: (1) Data preparation and cleaning, (2) modelling and validation, and (3) results presentation. For (1) gosset provides functions for converting and preparing data into ranking or pairwise format required by the packages PlackettLuce and BradleyTerry2 respectively. For (2), gosset provides functions for model selection and validation using cross-validation. In the case of (3), enhanced functionality for plotting model results is provided by the gosset package.

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 analysis workflows in agricultural experimentation, especially with rank-based approaches. The package is available in The Comprehensive R Archive Network (CRAN) and can be installed by executing `install.packages("gosset")`.

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.

The R package gosset is structured following the guidelines described in the manual for creating R add-on packages (Team 2022). This structure basically consist of files DESCRTIPTION, LICENSE, NAMESPACE and NEWS, and directories data, dev, docs, inst, man, R, and vignettes. The package functions were developed following the S3 methods style (Team 2022) and are contained in the R sub-directory.

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 and Terry 1952), implemented in the package BradleyTerry2 (H. Turner and Firth 2012) and which requires the input data to be formatted as paired comparisons. For these cases, gosset provides the functions `rank_numeric` and `rank_binomial` respectively. 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 collected from tricot trials. In those trials, farmers rank the technology under evaluation (e.g., crop variety) stating which is the best and which is the worst from a set of three (van Etten, Beza, et al. 2019).

Modelling

The gosset package provides the following functions to support model selection and validation of Bradley-Terry and Plackett-Luce models. The function `pseudoR2` computes goodness-of-fit measure McFadden's pseudo-R2 (McFadden et al. 1973). The `AIC` function computes the Akaike Information Criterion(Akaike 1974). The `kendallTau` function computes the Kendall-tau rank correlation coefficient (Kendall 1938).

Visualization and results presentation

`plot`

`worth_map`

`worth_map`

`regret`

`reliability`

Illustrative Examples

We demonstrate the functionality of the gosset package using the trial dataset "nicabean", which consists of trial data collected in Nicaragua following the tricot approach. We use the Plackett-Luce model implemented in the R package *PlackettLuce*(H. L. Turner et al. 2020). We use climate data as model covariates to investigate the effect fo climate factors on the performance of common bean (*Phaseolus vulgaris* L.) genotypes. For obtaining the climate data, we use the *nasapower* package(Sparks 2018). Climatic indices were computed with the *climatrends* package(de Sousa, van Etten, and Solberg 2020).

First, the required packages and data are loaded.

```
library("gosset")
library("PlackettLuce")
library("climatrends")
library("nasapower")

data("nicabean", package = "gosset")

dat <- nicabean$trial

covar <- nicabean$covar

traits <- unique(dat$trait)
```

Make a PlackettLuce rank using the function `rank_numeric`

```
R <- vector(mode = "list", length = length(traits))

for (i in seq_along(traits)) {

  dat_i <- subset(dat, dat$trait == traits[i])

  R[[i]] <- rank_numeric(data = dat_i,
                         items = "item",
                         input = "rank",
                         id = "id",
                         ascending = TRUE)
}
```

Kendall correlation between traits using kendallTau()

Worth map

Plackett-Luce tree using environmental data

Model selection using crossvalidation

pseudoR2

```
regret
```

```
reliability
```

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).

Reproducible and efficient workflows are fundamental in scientific research (Lowndes et al. 2017). The gosset package provides functionality that was not previously available from other R packages and which enabled scientific studies based on the analysis of ranking data. This functionality enables making the entire workflow to be reproducible and more efficient. The utility of the gosset package has been demonstrated by enabling studies based on the analysis of ranking data. For instance, both van Etten, de Sousa, et al. (2019) and Sousa et al. (2021) applied the Plackett-Luce model in combination with recursive partitioning (H. L. Turner et al. 2020; Zeileis, Hothorn, and Hornik 2008). In both studies, the gosset package supported the data preparation, model validation and results presentation tasks. The gosset package is freely available and can be downloaded from CRAN https://cran.r-project.org/package=gosset.

Conclusions

Set out the conclusion of this original software publication.

We described the functionality of the R package gosset to support the synthesis and analysis of ranking data. The package provide functions not available in existing R packages for analyzing ranking data. We provided an illustrative example covering the main functionality across the stages involved in the analysis workflow.

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

We acknowledge Olga Spellman (Science Writing Service of the Alliance of Bioversity International and CIAT) for English editing of this manuscript

References

Zenodo repository for the gosset package: https://doi.org/10.5281/zenodo.6339989

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

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6): 716–23.

Bradley, Ralph Allan, and Milton E Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39 (3/4): 324–45.

Brown, David, Inge Van den Bergh, Sytze de Bruin, Lewis Machida, and Jacob van Etten. 2020. "Data Synthesis for Crop Variety Evaluation. A Review." *Agronomy for Sustainable Development* 40 (4): 1–20.

Coe, Richard. 2002. "Analyzing Ranking and Rating Data from Participatory on-Farm Trials." In.

De Roo, Nina, Jens A Andersson, and Timothy J Krupnik. 2019. "On-Farm Trials for Development Impact? The Organisation of Research and the Scaling of Agricultural Technologies." *Experimental Agriculture* 55 (2): 163–84.

de Sousa, Kauê, Jacob van Etten, and Svein Ø. Solberg. 2020. *Climatrends: Climate Variability Indices for Ecological Modelling*. https://CRAN.R-project.org/package=climatrends.

Kendall, M. G. 1938. "A new measure of rank correlation." *Biometrika* 30 (1-2): 81–93. https://doi.org/10.1093/biomet/30.1-2.81.

Luce, R. Duncan. 1959. *Individual Choice Behavior.* Individual Choice Behavior. Oxford, England: John Wiley.

McFadden, Daniel et al. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior."

Plackett, R. L. 1975. "The Analysis of Permutations." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24 (2): 193–202. https://doi.org/10.2307/2346567.

Sousa, Kauê de, Jacob van Etten, Jesse Poland, Carlo Fadda, Jean-Luc Jannink, Yosef Gebrehawaryat Kidane, Basazen Fantahun Lakew, et al. 2021. "Data-Driven Decentralized Breeding Increases Prediction Accuracy in a Challenging Crop Production Environment." *Communications Biology* 4 (1): 1–9.

Sparks, Adam H. 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. https://doi.org/10.21105/joss.01035.

Team, R Core. 2022. "Writing r Extensions." R Foundation for Statistical Computing Vienna, Austria. https://cran.r-project.org/doc/manuals/r-release/R-exts.html.

Turner, Heather L, Jacob van Etten, David Firth, and Ioannis Kosmidis. 2020. "Modelling rankings in R: the PlackettLuce package." *Computational Statistics*. https://doi.org/10.1007/s00180-020-00959-3.

Turner, Heather, and David Firth. 2012. "Bradley-Terry Models in R: The BradleyTerry2 Package." *Journal of Statistical Software* 48 (9): 1–21. https://www.jstatsoft.org/v48/i09/.

van Etten, Jacob, Eskender Beza, Lluís Calderer, Kees Van Duijvendijk, Carlo Fadda, Basazen Fantahun, Yosef Gebrehawaryat Kidane, et al. 2019. "First Experiences with a Novel Farmer Citizen Science Approach: Crowdsourcing Participatory Variety Selection Through on-Farm Triadic Comparisons of Technologies (Tricot)." *Experimental Agriculture* 55 (S1): 275–96. https://doi.org/10.1017/S0014479716000739.

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.

Zeileis, Achim, Torsten Hothorn, and Kurt Hornik. 2008. "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics* 17 (2): 492–514. https://doi.org/10.1198/106186008X319331.