Generalizing multivariate data analysis protocols through package development in R

An open source and data experience from Archaeology

Andreas Angourakis, Verònica Martínez Ferreras,
and Josep M. Gurt Esparraguera

ERAAUB, Department of History and Archaeology, University of Barcelona, Barcelona, Spain. | contact: Andreas Angourakis andros.spica@gmail.com

Overview

- We recently created two **R packages**: cerUB and biplot2d3d.
- cerUB: analysis of archaeometric data on archaeological ceramics.
- biplot2d3d: to plot the result of ordination methods (*e.g.*, biplot).
- We explain the main **steps** towards creating an R packages based on applied R code used in previous/ongoing research.

Our motivation

Many techniques can inform on the origin and **production technology** of archaeological ceramics. However, studies often do not integrate more than one **data source** or do it only through textual description and argument rather than **statistics**. We consider this to be an obstacle for more **informative** data analyses, the **comparability** of datasets, and **reproducibility**.

What we have done

- Defined **four protocols for multivariate analysis of ceramics**, including methods that can integrate numerical and categorical variables.
- Selected the **statistical methods**, identified the **R packages** available to apply those methods, and produced the **R code** required.
- Realized that our efforts could be **generalized and offered as R packages** for other researchers to use/develop, the same way we were allowed to do with previous contributions.
- Created, documented and released **two R packages** (GitHub+Zenodo)
- Published a demonstrative paper [1].

The cerUB package

- **an R package focused on applying multivariate statistics on archaeometric analysis of archaeological ceramics**.
- **Target data**: geochemical compositions (*e.g.*, XRF readings), range of firing temperature (inferred through XRD readings), and petrographic observations (*e.g.*, thin-section analysis through optical microscopy).
- **Protocols**: encapsulate and order the execution of several processing and analysis techniques. Main end product is an ordination object (*e.g.*, result of Principal Components Analysis).
- Protocols offer **comparable results** and combine numeric and categorical data (***mixed-mode*** approach). Special thanks to the contributions of open-source contributions of [2] and [3].

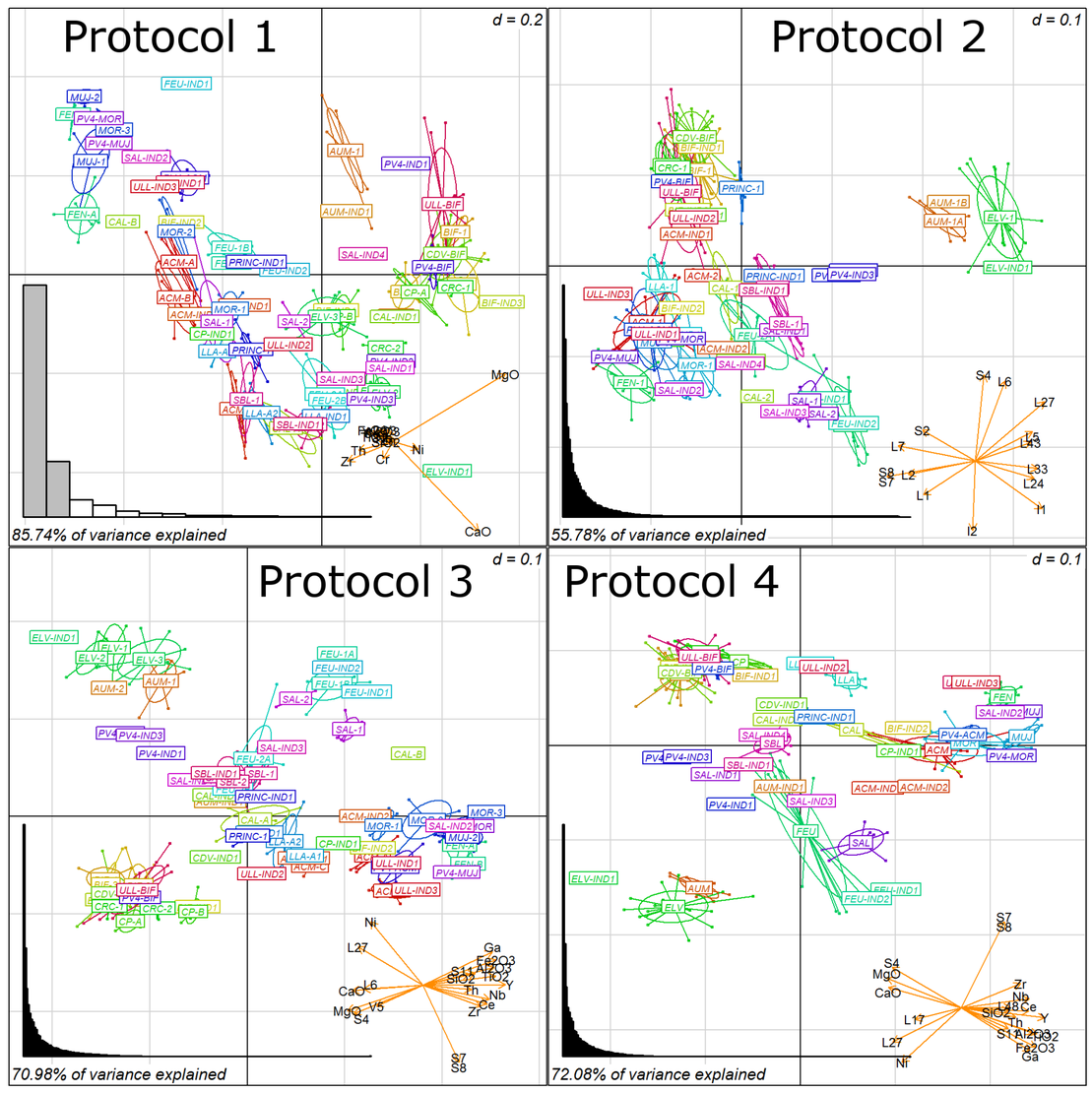
The biplot2d3d package

- **an R package dedicated to creating highly-customizable biplots** (*i.e.*, the most common representation of ordination methods results, where: points=observations and arrows=variables).
- Enables **customizing** points, arrows styles, groups of points, annotations, etc. Includes functions to create both **2D** and **3D** biplots with similar styles. Relies mainly on the ade4 and rgl packages.

Further information

The cerUB package has a **tutorial** online, showing how to install and use the package following the analyses in the reference article [1]. There you can also download this **poster**. The source code for both packages is on A. Angourakis' GitHub profile (Andros-Spica).

https://andros-spica.github.io/cerUB_tutorial



Example of the application of the four protocols offered in cerUB, using biplot2d3d for plotting results. Data concerns XRF-WD, XRD, and petrographic OM observations on a sample of 236 **wine Roman amphorae** found in 15 pottery workshops and 3 coastal shipwrecks in Catalonia. This demonstrative case study was published [1] and the dataset is included in the cerUB package.

Building R packages

Our workflow was not ideal. Much work was done indirectly, aiming at achieving our research goals. We were always learning new methods, features, and resources (and continue to do so). Based on our experience, these are the steps for creating R packages out of a research piece:

- Get some experience with the **R language** or collaborate with someone that already has it. Logically, the required experience is proportional to the complexity of goals.
- Search the Internet for any work addressing the same or similar goals. Do not be afraid to re-use R code; that is what **open-source** is about!
- Get a **minimum working procedure** with clear visual results (*e.g.*, tables, plots, maps).
- Create **more** procedures and improve them with **new features and options**.
- When and *if* your scripts start repeating code fragments or getting longer than 500 lines, it is time to create **functions** and placing them on separate .R files.
- At this point, you will be ready to start building an R package. We strongly recommend following this online book [4], which explains how to combine the devtools package with **RStudio**, and **GitHub**.

References

[1] A. Angourakis, V. Martínez Ferreras, A. Torrano, and J. M. Gurt Esparraguera.
Presenting multivariate statistical protocols in R using Roman wine amphorae productions in Catalonia, Spain.
Journal of Archaeological Science, 93:150–165, may 2018.

[2] P. Filzmoser, K. Hron, and C. Reimann.
Principal component analysis for compositional data with outliers.
Environmetrics, 20(6):621–632, sep 2009.

[3] S. Pavoine, J. Vallet, A.-B. Dufour, S. Gachet, and H. Daniel.
On the challenge of treating various types of variables: application for improving the measurement of functional diversity.
Oikos, 118(3):391–402, mar 2009.

[4] H. Wickham.
R packages.
http://r-pkgs.had.co.nz, 2015.