Austrian Journal of Statistics
2025, Volume 54, 50–54.
http://www.ajs.or.at/
doi:10.17713/ajs.v54i3.2068

Did We Practice what We Preached?

Torsten Hothorn
Universität Zürich

Abstract

Reproducibility of statistical simulations is crucial but proved being a challenge in its own right. Recently, lack of reproducibility of important simulation studies stimulated developments of reporting guidelines and specific protocols trying to improve on this situation. The problem, of course, is not new and issues regarding reproducibility of numerical results, for example statistical analyses or simulations, have long been known. Documented lack of progress regarding reproducibility in the past decade naturally leads to the question if problem awareness was not powerful enough to lead to improved reproducibility. As a benchmark case, I tried to reproduce a simulation study in a so far unpublished manuscript by Fritz Leisch and myself. The results show that, time and again, the devil is in the details and much self-discipline and extensive record-keeping and documentation are mission critical.

Keywords: reproducible research, simulations, Sweave, literate programming.

1. Introduction

When I was invited to contribute to this special issue, I was reminded of two papers I had the pleasure to work on together with Fritz while we both taught statistics at Universität München. One was written as a joint contribution to the "Validation in Bioinformatics and Molecular Medicine " special issue of *Briefings in Bioinformatics* edited by Anne-Laure Boulesteix (Hothorn and Leisch 2011). In this paper, we commented on issues in reproducibility of statistical analyses based on our experience as authors, referees, and editors. Literate statistical reports, using the Sweave system, were discussed as a technical tool to improve long(er)-term reproducibility of numerical results.

The second manuscript we had been working on did not see daylight until finally being published in this special issue (Leisch and Hothorn 2025). The manuscript bundles Fritz' and Bettina's **flexmix** package (Grün and Leisch 2025) with the **multcomp** package (Hothorn, Bretz, and Westfall 2025) for developing simultaneous inference procedures on contrasts of regression coefficients in mixture models. We submitted the manuscript to *Psychometrika* and, after a rejection, to *The American Statistician*, also without success. The second rejection came in around the time when Fritz left Munich for Vienna and I was on my way to Zurich in 2011. The main criticism from the referees concentrated around lack of proper literature review. After the second rejection, somehow, we both had lost interest and energy for this topic, so the manuscript went into the bottom drawer.

The question I wondered about was how well Fritz and I had managed to implement the advice we gave to others (for example in Hothorn and Leisch 2011) in our own work. Would it be possible for me to reproduce the results of a never published manuscript? In the following, I'll describe the necessary steps towards reproducing results obtained by Fritz and myself 15 years ago.

2. Steps towards reproducing results

2.1. Accessing source files

I was not able to find the subversion repository containing the relevant source files in my archive. The reason, beside bad record keeping, was that we had used Fritz' subversion server running in Munich and the machine was, of course, long gone. However, Fritz ran a tidier ship than I did and it was possible to access the last version of the relevant repository from his backup files in Vienna. Bettina Grün provided me with the last version of this repository.

The repository contained all LaTeX and Sweave source files as well as R code implementing simulation experiments. A data source for the example presented and the simulation results were also part of the archive. In addition, Fritz had kept a PDF file containing the initial version submitted to *Psychometrika*. So, in theory, all relevant material was present.

2.2. Re-compiling the manuscript

After locally updating the **flexmix** and **multcomp** packages, running the code in the Sweave files failed. Fritz had added a file `multcomp.R` to his **flexmix** package but did not export the functionality back then. The file contains a new generic `flxglht` with a corresponding method, in summary 61 lines of R code. He planned to make it available once the paper had been published. The current maintainer of **flexmix**, Bettina Grün, found the file in the subversion repository hosting the **flexmix** package (version 2.3-20 of **flexmix** now exports this functionality).

After adding and sourcing this additional file, it was possible to tangle and weave the three Sweave files. The second step of compiling the LaTeX document almost (after fixing some minor LaTeX hickups) worked out of the box as well.

2.3. Re-running simulations

Before we consider reproducibility of data analyses and simulation experiments, we take a step back and ask what can be reasonably expected from a manuscript nobody touched for almost 15 years.

As a benchmark, we have a look at reproducibility material accompanying the R handbook by Everitt and Hothorn (2006). An R add-on package called **HSAUR** (Hothorn and Everitt 2022) was first published on the Comprehensive R Archive Network (CRAN) 2006-02-01. All analyses discussed in this book are contained in package vignettes, one for each chapter of the book. Over time, 21 revisions of this package were published, most of them triggered by new CRAN requirements. Some were triggered by changes in other packages which were only noticed because the CRAN check services evaluate the R code contained in these vignettes on a daily basis on a large number of platforms. Such constant maintenance is required for all CRAN packages. The process ensures that problems, such as API changes, do not go unnoticed and are addressed in reasonable time.

No such process was performed for the material underlying the manuscript Fritz and I wrote around 2010 and put in the bottom drawer after two failed attempts to publish the material. In contrast to **HSAUR**, which lists 20 packages it depends upon, the simulations and data analyses presented in the manuscript "only" involve two packages, **flexmix** and **multcomp**.

However, both packages, first published 2003-07-09 and 2002-06-20, underwent extensive changes over the years. Bugs were fixed, new functionality added and existing functionality improved. In summary, there is ample potential for things going sideways when it comes to reproducibility of "old" research software.

As another benchmark, it might be helpful to reflect on the reproducibility of recent submissions to the *Journal of Statistical Software*. As an initial check, JSS' editors-in-chief install the submitted software and evaluate the submitted reproducibility material. The probability of this exercise resulting in exactly the numerical and graphical results presented in the submitted manuscript is well below one. Hothorn and Leisch (2011) reported similar results, based on the practical experience of the first author as "reproducibility research editor" of *Biometrical Journal*. Following an initiative of Leonhard Held, the editor of *Biometrical Journal* in the late 2000s, a team of Reproducible Research Editors independently evaluates software code submitted along manuscripts and checks to what extend the reported results are independently reproducible. Results of such exercises are often quite sobering. Other journals followed, some early (*Biostatistics*) and some only recently (*The Journal of the American Statistical Association*), but the expectation that accepted manuscripts are independently reproducible, given data and code, is still naive and far fetched (Peng 2025).

Unfortunately, the situation only improves marginally for published simulation studies. Luijken, Lohmann, Alter, Claramunt Gonzalez, Clouth, Fossum, Hesen, Huizing, Ketelaar, Montoya, Nab, Nijman, Penning de Vries, Tibbe, Wang, and Groenwold (2024) tried to replicate eight highly cited statistical simulations and were only able to independently and fully replicate three of them. As mitigation measures, guidelines for the design, analysis, and reporting of simulation studies are emerging in different disciplines (Williams, Yang, Lagisz, Morrison, Ricolfi, Warton, and Nakagawa 2024; Siepe, Bartoš, Morris, Boulesteix, Heck, and Pawel 2024).

In summary, the prior odds of the code Fritz (mostly) and I wrote being reproducible was rather low. However, it turned out that the data analyses for the PhD student data were exactly reproducible out-of-the box. The results obtained from re-running Fritz' code on the dataset he stored in the subversion repository resulted in the same estimates, standard errors, and multiplicity-adjusted p-values as reported in the PDF file initially submitted to *Psychometrika*.

As for the simulation results, things turned out to be a bit more complex. From what was documented in the PDF manuscript and the subversion directory, the exact versions of R and the **flexmix** and **multcomp** packages used to generate the numerical results back in 2010/2011 were not entirely clear. The citation of R gave the year (2009), but not the version. The simulation results were saved in binary format and the textual output of the simulation runs (`.Rout` files) were part of the repository, however, the output of `sessionInfo()` was missing. From the time-stamp of the binary files containing the simulation output (2011-03-14), I guessed that R version 2.12.2 published 2011-02-25 might have been used to perform the simulation experiments. This information is crucial, because the way random numbers are generated depends on the exact version of R. Thankfully, run in a recent version of R, the command `RNGversion("2.12.2")` resets the clock to spring 2011, at least for the generation of random numbers. It goes without saying that the simulation code explicitly initialised the seed, so at least I was able to regenerate the simulation datasets of the original simulation.

Unfortunately, running the simulation code failed. Fritz knew that sometimes fitting the model using `flexmix()` would fail. The manuscript devotes a section on this issue and explains how we tried to limit the number of such failures in our simulations. The code checked for and recorded the number of such failures, however, only after inspecting the `components` slot of the object returned by `flexmix()`. When I ran the simulations, this slot was no longer available in the right dimensions after a failed attempt to fit the model to some simulation dataset. I added three lines of code checking if this slot contains suitable information. If this is not the case, I count this particular simulation run as a failure. In

order to achieve the intended ten thousand simulation runs, Fritz repeated data generation and model fitting after a failed attempt, printing a star in the transcript file for each failure. Out of forty thousand new simulation runs, 21 failed and were replaced. When Fritz ran the code, this was the case for 13 of the simulation runs.

The main text of Leisch and Hothorn (2025) presents the original simulation results. The appendix of Leisch and Hothorn (2025) shows the differences to simulation results obtained from re-running the initial simulation code in fall 2024. The initial results were not exactly reproducible, but the observed differences turned out all very minor and well below Monte-Carlo error. In this respect, we can report that the simulation results are partially but not exactly reproducible.

3. Summary

With more effort, for example a re-installation of R 2.12.2 and package versions of that time, we would probably come closer to the original results. However, because the new results are close enough to the old results and we obtain the same qualitative conclusions, there is no need for investing more time.

This small reproducibility experiment demonstrates that the devil is in the detail. Even given simulation code and freely available implementations of the relevant models and estimation procedures, exact reproducibility is hard to achieve unless the exact computational environment is re-created. Lack of precise information on software versions and maybe even a precise specification of the platform used to generate the results is a problem and this information should therefore be stored. LaTeX source files containing tables and thus all relevant numbers as well as R output should be saved as `tex.save` or `Rout.save` files for later comparison with updated material, for example using `diff`.

Overall, we have succeeded in reproducing numerical results obtained more than 15 years ago. In a fast changing world, this is a remarkable result. It also demonstrates that sufficient tools to achieve reproducibility have been around for more than two decades. However, it took, and probably still takes, much dedication to invest into long-term reproducibility of simulation results.

An important, maybe the most important, factor is long-term stability of research software, that is, the R system with its **flexmix** and **multcomp** add-on packages in our small case study. Simulation code itself is often very dense and, at least conceptually, recoverable from the description of the data generating process in case the code was lost. This does not apply to the implementations of the statistical procedures under test. Long-term package maintenance in the CRAN universe, which Fritz co-founded in the late 1990s, is the key issue here. The contribution of stable and high-quality research software to statistical science, and empirical research in general, must not be underestimated and should therefore be incentivised and rewarded more prominently.

References

Everitt BS, Hothorn T (2006). *A Handbook of Statistical Analyses Using R*. Chapman & Hall/CRC Press, Boca Raton, Florida, USA. ISBN 1–58488–539–4.

Grün B, Leisch F (2025). ***flexmix**: Flexible Mixture Modeling*. doi:10.32614/CRAN.package.flexmix. R package version 2.3-20.

Hothorn T, Bretz F, Westfall P (2025). ***multcomp**: Simultaneous Inference in General Parametric Models*. doi:10.32614/CRAN.package.multcomp. R package version 1.4-28.

Hothorn T, Everitt BS (2022). ***HSAUR**: A Handbook of Statistical Analyses Using R (1st Edition)*, 1st edition. doi:10.32614/CRAN.package.HSAUR. R package version 1.3-10.

Hothorn T, Leisch F (2011). "Case Studies in Reproducibility." *Briefings in Bioinformatics*, **12**(3), 288–300. doi:10.1093/bib/bbq084.

Leisch F, Hothorn T (2025). "Simultaneous Inference in Finite Mixtures of Regression Models." *Austrian Journal of Statistics*.

Luijken K, Lohmann A, Alter U, Claramunt Gonzalez J, Clouth FJ, Fossum JL, Hesen L, Huizing AHJ, Ketelaar J, Montoya AK, Nab L, Nijman RCC, Penning de Vries BBL, Tibbe TD, Wang YA, Groenwold RHH (2024). "Replicability of Simulation Studies for the Investigation of Statistical Methods: The RepliSims Project." *Royal Society Open Science*, **11**(1), 231003. doi:10.1098/rsos.231003.

Peng RD (2025). "Tooling for Reproducible Research: Considerations for the Past and Future of Data Analysis." *Austrian Journal of Statistics*.

Siepe BS, Bartoš F, Morris TP, Boulesteix AL, Heck DW, Pawel S (2024). "Simulation Studies for Methodological Research in Psychology: A Standardized Structure for Planning, Preregistration, and Reporting." *Psychological Methods*. doi:10.1037/met0000695.

Williams C, Yang Y, Lagisz M, Morrison K, Ricolfi L, Warton DI, Nakagawa S (2024). "Transparent Reporting Items for Simulation Studies Evaluating Statistical Methods: Foundations for Reproducibility and Reliability." *Methods in Ecology and Evolution*, **15**(11), 1926–1939. doi:10.1111/2041-210X.14415.

Affiliation:

Torsten Hothorn
Institut für Epidemiologie, Biostatistik und Prävention
Universität Zürich
Hirschengraben 84, CH-8001 Zürich, Schweiz
Email: Torsten.Hothorn@R-project.org

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2024-10-15
Accepted: 2025-03-13