Population Dynamics Lab
peer-reviewed, citable, reproducible demographic insights

Increasing the Lifespan of Software for Demographic Analysis

Egor Kotov [1,2], and **Esther Denecke** [1,3]

[1] Max Planck Institute for Demographic Research, [2] Universitat Pompeu Fabra, [3] University of Rostock

Many researchers face challenges with computational reproducibility. For instance, running analysis code written just a year earlier can be problematic. Even if it worked flawlessly and gave the expected results earlier, it might fail due to errors now (see Figure 1). These issues are typically due to the use of newer versions of analysis software. Software updates are essential for introducing new features, fixing bugs, improving security, and compatibility with other updated software. Consequently, researchers have to switch to updated analysis tools over time, which can prevent them from running older code. This impacts the reproducibility of scientific findings, as other researchers may face difficulties testing published methods in new situations or with different data.

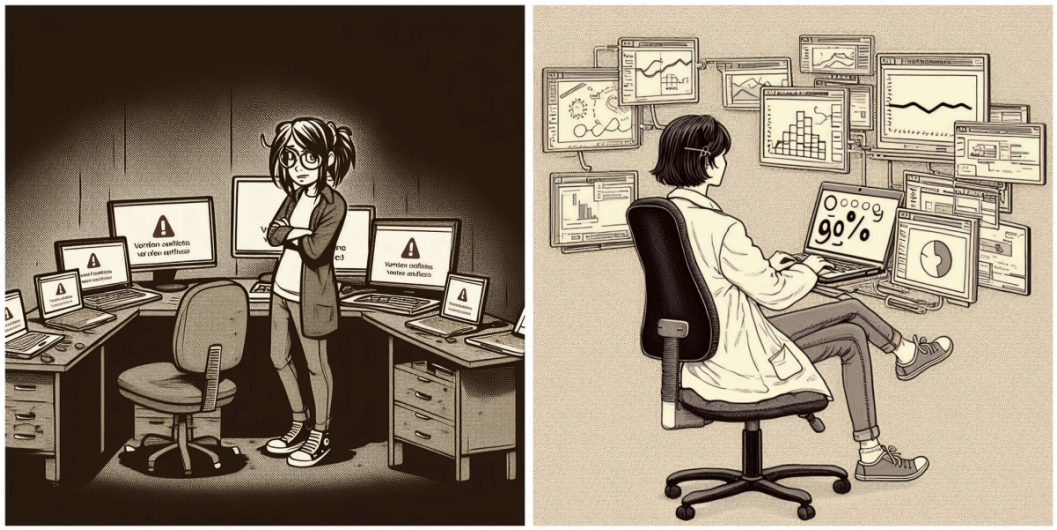


Figure 1: Reproducibility struggles. Left: failing to reproduce the exact software setup on multiple computers. Right: running countless software setups on a single laptop with the help of containers. Image source: Microsoft Image Creator powered by DALL-E 3. Generated with AI · 25 October 2023

In this note we aim to introduce the demographic community to containers, a software solution to achieve computational reproducibility. Containers are a popular tool from software engineering (Cito & Gall, 2016) and already used in other scientific fields, for

THE DOWNLOAD

OPEN ACCESS

Citation:

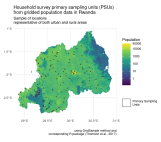
Kotov, E., & Denecke, E. (2024). Increasing the Lifespan of Software for Demographic Analysis. https://doi.org/10.6069/7JXS-6C18

Published: 2024-04-17

License: CC-BY 2.0

Template based on LaPreprint for Typst by Rowan Cockett (MIT License).

Related Denominator



Expanding the Lifespan of Software for Demographic Analysis with Containers: An Application of Spatial Sampling Introduction Software, such as specific R packages, evolve over time, which may prevent older analysis code from working as expected. For example, default values for arguments in a function can change. Therefore, for computational reproducibility, knowing which specific R and package versions were used to run the analysis is crucial. One popular solution in R... read more

example in education (LeBeau et al., 2021) and psychological research (Wiebels & Moreau, 2021). Computational reproducibility refers to achieving the exact same results when using the original analysis code and data. This practice is crucial for maintaining transparency and trust in scientific research.

An ideal scenario involves having a computer with necessary software for specific analysis. However, packaging this system in a zip archive is inconvenient, requiring time-consuming backups and restores for each project. Using a separate computer with specific software for each project is economically unfeasible. Containers solve these issues (see Figure 1), functioning like a dedicated computer for each project, ready for instant use without waiting for the software to be installed or restored from a backup.

To illustrate, we focus on the scenario of reusing a published methodology accompanied by an R package (Thomson et al., 2018) that is no longer available from the CRAN R package repository. Thomson et al.'s 2017 GridSample method is a more accurate alternative to traditional census-based sampling, particularly in areas where census data is outdated or unreliable. By using gridded population datasets, GridSample allows for more representative survey samples, enhancing the accuracy and reliability of demographic studies.

To apply the method from GridSample, we created a container with the correct R version and necessary packages, mimicking the 2017 R environment. To access and run the example online without installing anything, open the GitHub repository and click the "Launch Binder" launch binder button. The RStudio running from a container will open in a web browser in a few moments. To reproduce the example, (1) open the "main.Rmd" file in the bottom right files panel by clicking on it, then (2) click the "Run -> Run all" button in the top middle. Once the analysis finishes, the result is a sample of locations representative of both urban and rural areas.For those interested in the technical details and trying to create similar repositories to run containers, please refer to the related Denominator and the comments inside the configuration files in the GitHub repository.

Computation & Reproducibility

All code necessary to implement the methods and reproduce the figures and results in Increasing the Lifespan of Software for Demographic Analysis has been archived as of publication on April 17, 2024 by the Population Dynamics Lab: https://github.com/Population-Dynamics-Lab/grid-sample-containerized.

The original repository maintained by Egor Kotov can be found here: https://github.com/e-kotov/grid-sample-containerized. Note: this repository is maintained by Egor Kotov and may differ from that originally used to produce the results in this publication.

References

Cito, J., & Gall, H. C. (2016). Using docker containers to improve reproducibility in software engineering research. *Proceedings of the 38th International Conference on Software Engineering Companion*, 906–907. https://doi.org/10.1145/2889160.2891057

LeBeau, B., Ellison, S., & Aloe, A. M. (2021). Reproducible Analyses in Education Research. *Review of Research in Education*, *45*(1), 195–222. https://doi.org/10.3102/0091732X20985076

Thomson, D. R., Stevens, F. R., Castro, M. C., & Tatem, A. J. (2018,). *GridSample: Tools for Grid-Based Survey Sampling Design. R package version 0.2.2*. https://cran.r-project.org/package=gridsample

Thomson, D. R., Stevens, F. R., Ruktanonchai, N. W., Tatem, A. J., & Castro, M. C. (2017). GridSample: An R package to generate household survey primary sampling units (PSUs) from gridded population data. *International Journal of Health Geographics*, *16*(1), 25. https://doi.org/10.1186/s12942-017-0098-4

Wiebels, K., & Moreau, D. (2021). Leveraging Containers for Reproducible Psychological Research. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211017853. https://doi.org/10.1177/25152459211017853