Harvard Data Science Review • Issue 5.3, Summer 2023

Reproducibility With Confidential Data: The Experience of BPLIM

Paulo Guimarães[1]

[1]Microdata Research Data Laboratory (BPLIM), Economics and Research Department, Banco de Portugal, Porto, Portugal

Published on: Jul 27, 2023

DOI: https://doi.org/10.1162/99608f92.54a00239

Background

The Banco de Portugal Microdata Laboratory (BPLIM) was established in 2016 with the primary goal of promoting external research on the Portuguese economy by making available data sets collected and maintained by Banco de Portugal (BdP). By making this information available to researchers from around the world, BdP aims to support the development of evidence-based policies and insights that can benefit the Portuguese economy and society. However, given that some of these data sets contain highly sensitive information, BPLIM had to implement a data access solution that preserved the confidentiality of the data.

The common approach by other research data centers that make confidential data available for research involves the provision of on-site access to accredited external researchers in a secure computing environment. However, in the case of BPLIM, this approach was deemed undesirable for two reasons. First, it would limit access to a handful of researchers who were able to come to the bank's premises. Second, there were still concerns that a breach of confidentiality might occur if individuals from outside the bank could gain access to original data sets that contained confidential information.

After an internal debate at the bank, it was decided that the solution to be adopted had to be based on the following principles:

- access free of charge and only for scientific purposes;
- all data should be analyzed on the servers of the bank;
- external researchers were granted remote access to the server;
- confidential data sets placed on the server had to always be perturbed/masked;
- researchers could always ask BPLIM staff to run their scripts on the original data.

The general workflow defined for data access at BPLIM was the following. Researchers must first submit a proposal containing a short description of the research project along with other documentation (for details, see https://bplim.bportugal.pt). Typically, within a week, a decision is taken. If the research project is approved and the external researchers are accredited, an account is opened on the BPLIM external server. External researchers gain access to a computing environment that does not allow users to transfer files to and from the server. They have access to a restricted area where standard software such as Stata, R, Julia, and Python are available. Since there is no connection to the internet, installation of specific packages must be requested from the staff. The data sets for the project are placed in a read-only folder. For the confidential data sets, what is placed in the account of the researcher are perturbed versions of the data (noise is added to the original data). The researcher implements all scripts based on the data he/she has available and produces the outputs required for the project. Since these outputs are obtained from the perturbed data, they do not contain valid results that can be used by the researcher. However, once researchers complete this task, they can ask BPLIM staff to rerun their scripts, this time using the original confidential data. For this process to be successful, BPLIM staff must

first run the scripts using the same data as the researcher to verify that the scripts written by the researchers reproduce exactly the outputs (typically graphs and/or tables). This process is done in a different server (BPLIM internal server). Only upon completion of this first step can BPLIM staff modify the scripts, this time to read the original data and regenerate the intended outputs. These outputs are then subject to standard output control checks for confidential data and delivered to the researcher. The duration of this process can vary based on the complexity of the analysis and the workload of the laboratory, with a potential time frame of up to a fortnight; however, it is typically completed within a span of 2 to 3 business days.

Main Thoughts

Over time we have come to realize that this somewhat cumbersome process of running the code thrice, first by the researcher on the perturbed data, second by BPLIM staff again on the perturbed data, and third, by BPLIM staff on the original data, was in fact an exercise on reproducibility. Even though the reproducibility check was on the perturbed data, it was already a very good assurance that a reproducibility check would hold on the original data.

We realized that a great deal of our work involved reproducing the results obtained by the researchers with the perturbed data and that led us to look at ways to improve our workflow. It became obvious that the process could be streamlined and would be more efficient if researchers adhered to the best practices on reproducibility. Hence, as part of our strategy, we have decided to raise awareness of our researchers to the need of implementing good practices in reproducible research. We have been doing this by several means. For example, we have held practical workshops designed to enhance the skills of our researchers. For these workshops, we invite leading experts to present best practices and recent developments on data analysis. We also provide direct advice to the researchers, prepare templates and documentation, and make available tools that facilitate the analysis of our data sets (particularly for more complex tasks such as building a panel or calculation of specific variables).

On the other end, it was also obvious that there was a margin for improvement in our work sequence. One possible improvement was the assurance that the computing environment used by the researcher on BPLIM's external server was identical to that used by BPLIM staff when reproducing the code. Thus, for the case of researchers that work with open source software, we have been incentivizing researchers to work with Singularity containers (for more information on these, see our GitHub repository: https://github.com/BPLIM/Containers). This facilitates our work because we are sure that our reproducibility check is implemented in the same self-contained environment that was used by the researcher. Researchers that use Stata can resort to containers, but in that case, it is easier to control the environment because we install all packages on a folder that is specific to each project and have developed tools that facilitate comparison of the Stata ado-files across environments (all tools are publicly available and can be found at https://github.com/BPLIM/Tools/tree/master/ados/General).

More recently, we have worked on shifting the burden of the reproducibility check to the researcher itself. We are developing an application that we are presently testing with a select number of researchers. The application is targeted mainly at researchers that use BPLIM's (perturbed) confidential data sets, but we hope to eventually convince other users to take advantage of it. In our GitHub repository we make available the source code (see https://github.com/BPLIM/ReplicationApp).

To illustrate, we provide a screenshot of the application in Figure 1:

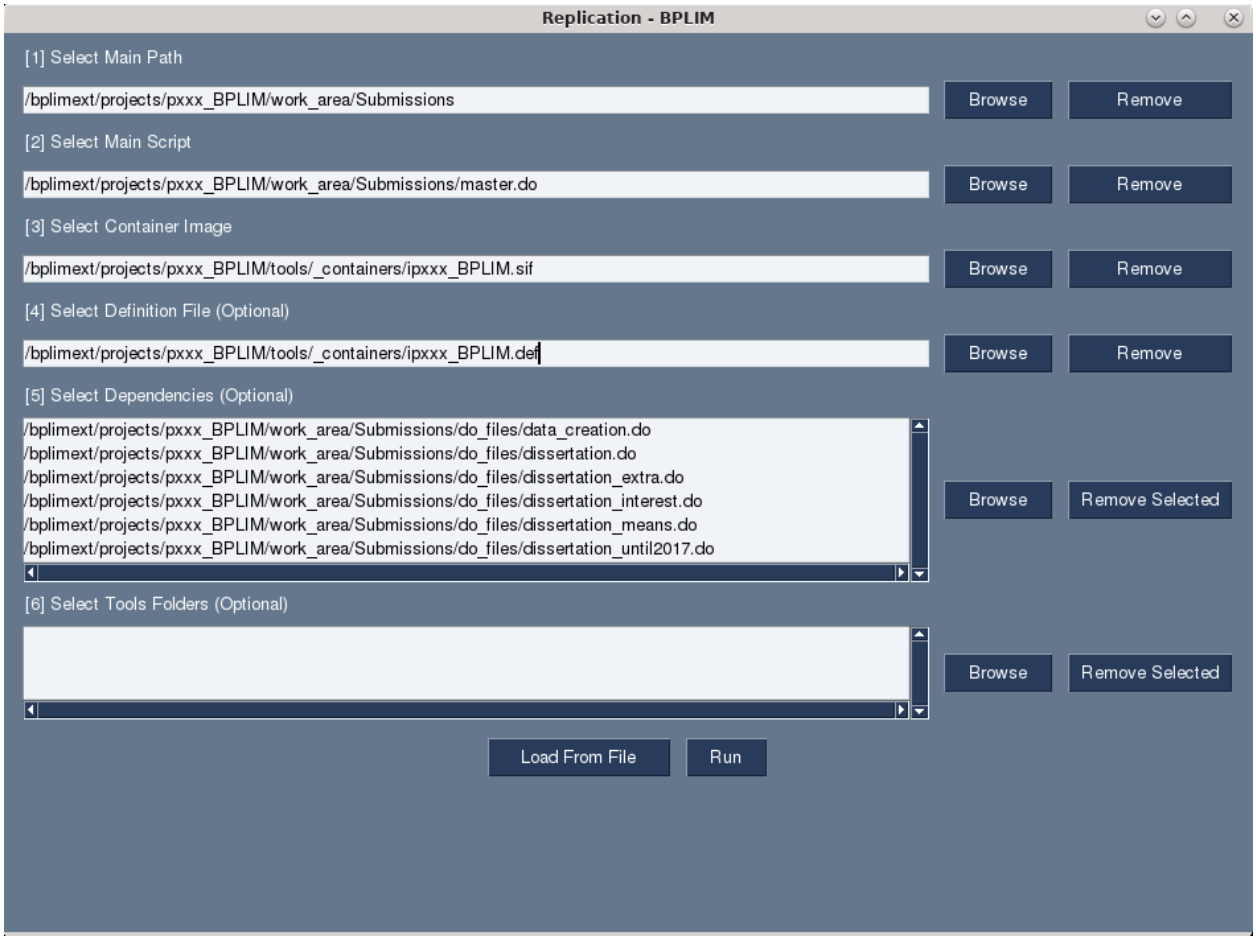

Figure 1. BPLIM Replication tool—selecting input files.

Before requesting a replication from BPLIM using the original data, the researcher must first validate his/her code by successfully submitting the scripts through BPLIM's Replication application. The process involves selecting the main script as well as all the required dependencies created by the researcher. The folder structure used by the researcher is replicated and the BPLIM data sets must be read from the (read-only) data folder. All intermediary output files must be created during the replication. (It is, however, possible to start from an intermediary output file. In that case, the intermediary file must have been validated in a prior run. BPLIM will then copy the file to the [read-only] data folder.)

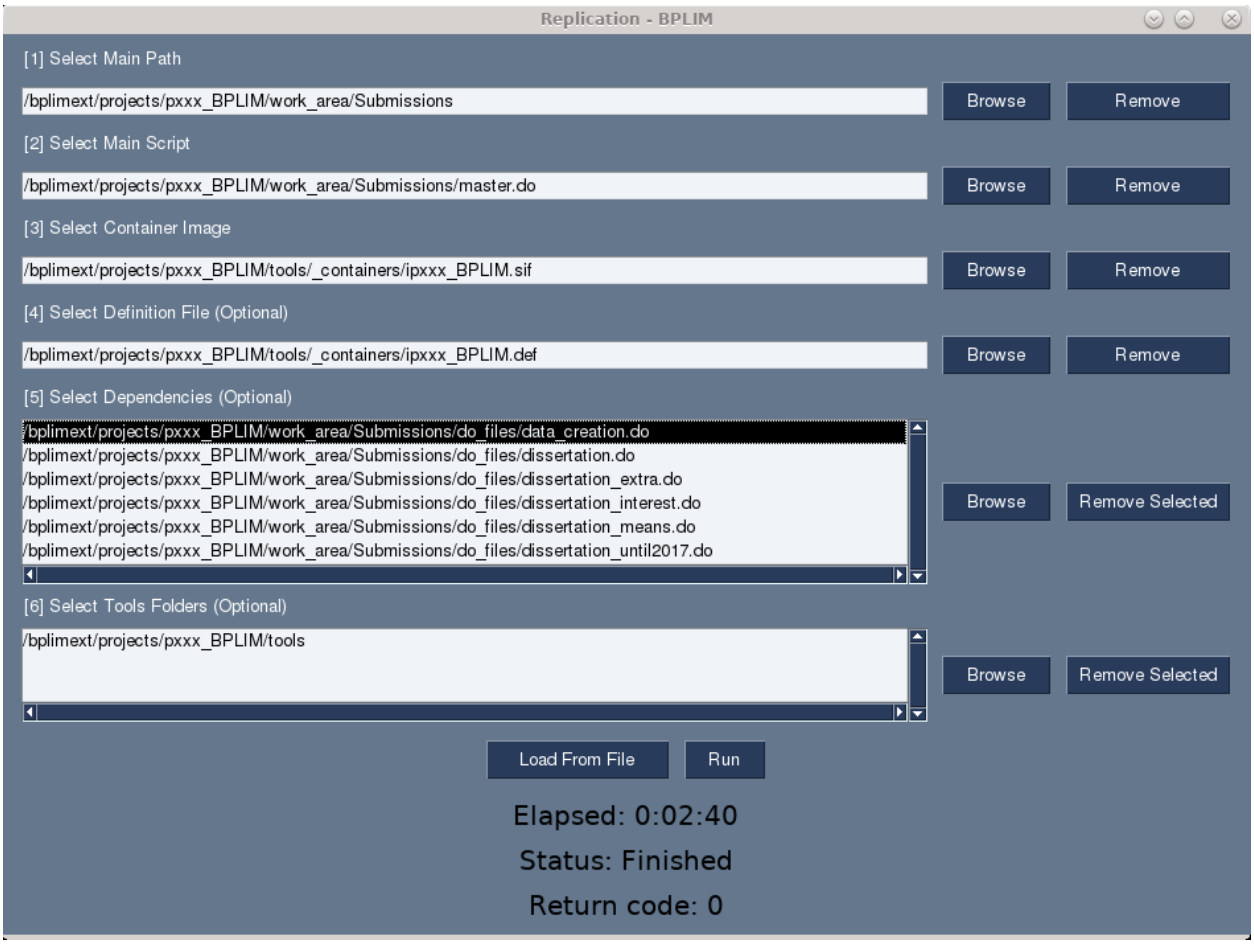

Figure 2. BPLIM Replication tool—finished task.

The researcher then uses the application to run the code (see Figure 2 for a completed run). The code must run from top to bottom and produce no errors. If the run is successful, then implementation on the original data requires only that BPLIM staff change the relative paths to the data folder and rerun the code. A side advantage of this process is that it automatically produces a replication package for the researcher. Stored in the folder are all replication scripts, the output files, as well as two additional files: one fully characterizing the software environment, and another JSON file containing a listing of all scripts used in the replication. If we add the definition file used to produce the container (or a listing of all packages and respective versions), then we have a full replication package (except for the data). This package is provided to the researcher, and BPLIM affords them the freedom to distribute it for public access if they so choose.

Conclusion

Our goal at BPLIM is to make sure that all researchers create their replication packages as an integrated part of their research process. The fact that we are the ones running the code on the original data should be seen as an opportunity to request that researchers make reproducible code while implementing their research.

In the ideal situation that we envision, researchers download a template for the definition file of the Singularity container, customize that template by adding and testing the packages they need, and share with us the definition file. Based on that definition file, we build the container for the project and make it available on our external server. The researcher then uses the container to implement the analysis and when he/she is ready to obtain results based on the original data, he/she must first validate the scripts using our application. The researcher can go through this process multiple times and each time a replication package will be created. Once he/she concludes the work, the researcher can choose to make the package publicly available or authorize its use. With a few modifications the package could be easily customized for submission to any data editor. Other researchers willing to replicate the results can request access to the data and use the replication package.

We are already implementing this solution for all new projects that use confidential data. However, we hope that over time we can convince all researchers at BPLIM to work with Singularity containers and go through the same validation steps that are needed for projects that deal with confidential data.

Projects that are implemented in BPLIM have additional advantages when it comes to reproducibility. First, because all BPLIM data sets are versioned and registered with a digital object identifier (DOI), we are sure that the original data is exactly identified. Second, the computing environment is stable, and the software packages used by researchers are specific to each project. Finally, if external researchers have used BPLIM confidential data, then there is an assurance that their code was reproduced at some point.

Ultimately, the reproducibility of scientific work depends on the researchers. Nevertheless, by integrating the creation of a replication package as part of the research process, we aim to enhance transparency and facilitate the efforts of third parties who may not have access to the original data but seek to verify the replicability of the results.

Acknowledgments

I express my gratitude to Miguel Portela, Gustavo Iglésias, Joana Pimentel, Marta Silva, Rita Sousa, and Sujiao Zhao for their valuable contributions to the implementation of the solutions and tools discussed in this work. Each of them has played a significant role in their respective capacities.

Disclosure Statement

Paulo Guimarães has no financial or non-financial disclosure to share for this article.