Guide for Researchers

Banco de Portugal's Microdata Research Laboratory (BPLIM)

2024-07-24

Contents

1 General Remarks

The Banco de Portugal Microdata Research Laboratory (BPLIM) is part of the Economics and Research Department (DEE) of Banco de Portugal (BdP) and was created with the objective of facilitating researchers' access to and use of micro datasets about the Portuguese economy. Primarily, these are datasets owned by BdP, but there are other datasets supplied by third parties. Data can only be accessed on BPLIM's **External Server** or in the internal infrastructure of the bank. The External Server has **remote access** capabilities and is intended for use by external researchers. Internal researchers (those with a contractual link to BdP) have access to a scalable computational platform (**Pitagoras**) and to a data repository created by BPLIM, which can be used without restrictions.

2 Data Available for Researchers

2.1 What are the characteristics of BPLIM's datasets?

All BPLIM datasets created for research purposes are stripped of elements that allow for direct identification of companies, banks, or individuals. Whenever possible, the datasets contain **unique unit identifiers** common across datasets: examples of these are *tina* – the tax identification number anonymized for companies – and *bina* – the bank identification number anonymized.

By default BPLIM datasets are made available in *Stata* format. Larger datasets may be made available in *parquet* format. Data is stored in an efficient way that minimizes file size and follows BPLIM's naming convention. **Labels** are applied to all variables and value labels to all categorical variables. Whenever possible, labels can be displayed in Portuguese and English.

All datasets are accompanied by a **Manual** that contains all relevant information regarding the data. The data manuals, the metafiles[1] and citation information for the different data extractions are available on **GitHub**. A metafile that contains additional descriptive statistics for each dataset can be obtained once researchers are given access to the server. Please refer to the **External Server Guide** on how to access this information.

BPLIM datasets may also have companion script files that calculate additional variables or harmonized variables to guarantee comparability over time and across datasets. Datasets are updated regularly based on a **data extraction** ("data freeze") at a specific point in time, and a versioning system is applied to reflect any changes to the data set. Most datasets have an associated **Digital Object Identifier** (DOI).

[1] All metafiles are created with the *metaxl* Stata command.

2.2 What data are available?

The complete **list of datasets**, including a short description and the access conditions, is available in the **BPLIM Datasets Guide**. On **BPLIM´s website** you will find a list of the latest version of the datasets available for external researchers, along with a link to the respective documentation.

3 Data Access

3.1 Who can gain access to the data?

Access is restricted to BPLIM accredited researchers who intend to utilize the data for scientific purposes. Individuals affiliated with BdP are classified as **Internal Researchers** and have unrestricted access to all datasets maintained by BPLIM. Those not affiliated with BdP are considered **External Researchers**, and their access is subject to several restrictions. **BPLIM Datasets Guide** summarizes the access restrictions for each dataset.

3.2 How can researchers request access to the data?

Internal researchers are provided access to a data repository maintained by BPLIM. The repository makes available data, which can be used unreservedly for research and policy activities without any formality. However, data made available strictly for policy should not be used for research because as it may have lower quality, are not documented, and may not be reproducible. If Internal Researchers are working with external co-authors or if they need to use or link other micro datasets not available in the data repository they will need to submit a project to BPLIM.

External Researchers must always submit a project. The **project proposal** must: (1) contain a short description of the research project; (2) identify all participants involved in the project along with their affiliations, and include a curriculum vitae (CV) for each; and (3) specify the datasets, timeframe, and variables required. All external researchers with access to the data are required to sign a *confidentiality agreement*. If the project consists of a master or doctoral dissertation then the supervisor(s) has to be identified and must also sign the *confidentiality agreement*. BPLIM staff can collaborate with the researcher(s) to identify the required datasets and, if necessary, construct a customized dataset. A copy of the required documents can be found in **BPLIM's website**.

3.3 Approval process

Upon submission of all required documentation and verification that it conforms to BPLIM rules, the project will be evaluated to ensure that it addresses a legitimate research question. **Compliance with the guidelines** is crucial for recurring external researchers (those who have already participated in BPLIM projects). Once the project is approved, the researcher

will receive an email notification with the **user credentials and instructions** for accessing the data. Summary information about the project and researchers will be posted on BPLIM's website.

3.4 How can the data be accessed?

When applying for a project, researchers must specify if they plan to access an internal account, in *Pitagoras*, or an external account in the External Server.

Accounts open at ***Pitagoras*** can only be accessed at the installations of BdP (**"on-site access"**) either at Lisbon or Porto. Internal researchers can log into *Pitagoras* from their terminal using their network login credentials. External researchers will be provided with a login and password for *Pitagoras* and granted access to a terminal where it is technically restricted to transfer, download, copy, paste, or print any data. BPLIM projects at *Pitagoras* are placed in a specific folder containing all projects, with users having access only to their designated project folder.

For more details on accessing BPLIM projects in *Pitagoras*, please refer to the *BPLIM Pitagoras Manual*. Due to a limited number of terminals available, external researchers must book their visits well in advance.

If the account is on the **External Server**, then the data must be accessed remotely (**"remote access"**) using a secure connection. BPLIM uses the ***NoMachine*** software for this purpose. With this connection, it is not possible to exchange files between the external server and the local computer. For more details on using the External Server, please refer to the **External Server Guide**.

In special circumstances explained below, the external researcher may be granted indirect access to the data (**"surrogate access"**, also known as, **"remote execution"**). With surrogate access, there is no need for the external researcher to have an account, as BPLIM staff (or an internal researcher) will act as a proxy for data access. This means that BPLIM staff (or an internal researcher) will execute the scripts written by the external researcher and share the outputs after disclosure control.

3.5 What type of anonymization is applied to BPLIM's datasets?

When BPLIM makes its datasets accessible to researchers, it uses several different strategies to anonymize the data. The type of anonymization depends on the specific data and the user. BPLIM uses **four levels of anonymization**:

- **Level 1** - All information that could lead to the direct identification of statistical units (firms/banks/individuals) is omitted, and unique identifiers (e.g., NIF, bank ID) undergo a 1-to-1 transformation to new identifiers that are specific to the project. Level 1 datasets will contain **"__A__"** in the name.

- **Level 2** - in addition to Level 1, the values of variables containing sensitive information will be replace by modified values, which are random values that exhibit some correlation with the original values. The file name of a Level 2 dataset will contain **"_P_"**, and the labels of the modified variables will reflect this information.
- **Level 3** - in addition to Level 2, variables may be sorted randomly and independently to break the link between the observations. Level 3 datasets will be identified with **"_R_"**.
- **Level 4** - a subset of the data is generated randomly (pseudo-data), respecting only the metadata and the time structure of the original data. Level 4 datasets will be identified with **"_D_"** in their name.

Datasets of **Level 2, 3, or 4** are generically designated as **modified** datasets.

Level 4 datasets are the only ones that researchers are allowed to use **outside of the bank computing environment**, because the values generated for this level are fictitious.

3.6 What determines the type of anonymization applied to the data set?

BPLIM data meant to be used by **Internal researchers** are always anonymized at Level 1. The exception is if the Internal Researcher is accessing the data through the External Server. In that case, Internals Researchers have the same access conditions as External Researchers.

Datasets made available to **External Researchers** are subject to a **confidentiality classification** as follows: **low**, **medium**, or **high**. If the data are classified as **low**, then the data is anonymized at Level 1. Datasets classified as **medium** may be anonymized at Level 2 or 3, depending on the risk of identification. For datasets classified with a **high** level of confidentiality, External Researchers may only have access to Level 4 data.

3.7 How can researchers work with *modified* data?

Modified datasets serve only the purpose of facilitating the creation of scripts that manipulate/analyze the data. Results of analysis performed on "modified" datasets **are not valid for research purposes**. However, external researchers can always request to have their scripts run on the original datasets. This rule applies whether the access is "on-site" or "remote". Researchers working with **modified** data should use BPLIM's **Replication App**. For instructions on how to use the Replication App, please refer to the **Replication App User Guide**. While not strictly enforced, use of the BPLIM's Replication App, will ensure that the replication on the original data is implemented correctly and in a much more timely manner. We strongly encourage use of the BPLIM's Replication App.

Researchers working with Level 4 data (**pseudo-data**) in their personal computers, will receive a package along with instructions to create the pseudo-data. For instructions on how to work with pseudo-data, please refer to the **BPLIM Guide to Working with Pseudo-Data**.

3.8 How are projects involving co-authorship between Internal Researchers and External Researchers handled?

Projects where internal and external researchers have access to the data are designated **"mixed projects"**.

If the mixed project is implemented in the **External Server** and only data with low level of confidentiality is used, then the distinction is irrelevant as all researchers are treated as external and the data is anonymized (Level 1).

However, if the data needed for the project is classified at a higher confidentiality level, external researchers can only access **modified data** or, in the most restrictive cases, **pseudo-data**, but never the original data. In such mixed projects, where the external researcher does not have access to the original, the internal researcher is responsible for ensuring that the information shared with their external co-authors complies with the confidentiality requirements associated with the data.

In these cases, BPLIM will open a second **"parallel" account** with access only for the internal researcher(s) and place all the original (anonymized) data there. It will be the responsibility of the internal researcher to execute all scripts on the original data stored in this "parallel" account.

Furthermore, it will be their responsibility to ensure that external researchers do not have any access to confidential data. Specifically, external co-authors must not access the project account containing the original data, the internal co-author's desktop computer, or any logs that may contain confidential information.

In a mixed project all interaction with BPLIM should be done via the internal researcher.

3.9 How can an External Researcher gain surrogate access to a dataset?

The *BPLIM Dataset Guide* lists the datasets that can be accessed in surrogate mode by an external researcher that **does not have an internal co-author**. In that case the external researcher needs to submit a detailed project to BPLIM. The project will be evaluated **according to the relevance of the topic** to the research agenda of BdP. Only projects deemed relevant will be granted surrogate access. If BPLIM decides to support the project, external researchers will be assigned a data expert at BPLIM, who will collaborate with them to prepare and run scripts on the original data. However, the coding itself remains the ultimate responsibility of the external researcher, and BPLIM will not validate or certify the scripts written by the researcher. External researchers are encouraged to work closely with BPLIM staff to ensure they achieve the intended results and are also encouraged to discuss their findings with BPLIM staff. It is highly recommended that external researchers initiate their project with a short-stay visit in BPLIM, during which time they can discuss their research with BPLIM staff and gain a thorough understanding of the data complexities. Additional visits throughout the project are also encouraged. All outputs shared with the external researcher are subject to the usual disclosure restrictions.

3.10 Transfer of external files

Internal researchers working in *Pitagoras* can freely copy files to and from their accounts. Thus, they are free to place external files in their *Pitagoras* accounts. However, if the external data needs to be merged with BPLIM datasets using an anonymized key, then the internal researcher must be working in a BPLIM project account at *Pitagoras*. They will also need to fill in an application for using an external dataset (see below). BPLIM will anonymize the external datasets using the same linking key as the one used for the BPLIM datasets. Note that BPLIM identifiers (eg: *tina* and *bina*) are specific to a project and are not valid to link files exchanged between accounts.

If the account is shared with external researchers - a **"mixed project"** - the internal researcher must ensure that external researchers are allowed access to the external dataset and that they do not gain undue access to confidential data. At the request of the internal researcher, BPLIM will anonymize/modify the external datasets intended for use in "mixed projects".

External researchers may also request that **external data files** be placed in their accounts. BPLIM staff will assist if there is a need to merge external datasets with BPLIM datasets. External datasets typically contain aggregated data, but it may be possible to add external datasets with finer granularity. BPLIM staff will assess if the addition of the external datasets increases the risk of identification of individual observations. In such cases, additional measures will be undertaken to ensure that the confidentiality of the data is preserved once external files are merged with existing BPLIM datasets. These situations will be analyzed case by case and discussed with BPLIM staff.

All external datasets provided to BPLIM should be in a **Stata or CSV format** and an **External Datasets Form** must be filled in. In the form, researchers are required to explain the data provenance, provide a justification for its use, and identify the key variables that enable linking the external dataset with BPLIM's datasets. The researcher must also certify that all researchers with access to the account are authorized to use the data. It is the responsibility of the researcher to ensure the external files can be legitimately used for that purpose.

4 Statistical Software

4.1 Statistical Software

When researchers apply for a new project, they will need to specify the software to be used. Available options are **Stata**, **R**, **Julia**, and **Python**. BPLIM provides the researcher with a default list of **external packages/ados** for each software. If researchers require additional packages, they must specify the package, its source, and version.

For each project, BPLIM creates a **container** with the software and packages required for the project. If researchers require **additional external packages** during the project, they should send a **request via email** to BPLIM.

Once the project account is set up, researchers will have access to a **Linux environment** where they can use the container. **Templates** for writing code are also provided in the account, and researchers should strive to adhere to the conventions outlined in these templates as much as possible.

4.2 BPLIM Tools

BPLIM staff has developed several **Stata packages** to assist researchers in their tasks. Some of these tools are tailored for use with BPLIM datasets, while others have broader utility. To promote transparency in coding and versioning, BPLIM makes all tools available on **Github**. Users are welcome to suggest improvements or add their own contributions. Tools with general applicability can be installed directly from Github on any internet-connected machine.

5 Output Extraction

5.1 Can researchers transfer files from the server?

External researchers are **never allowed to transfer files to/from BPLIM's accounts**. This policy also applies to internal researchers accessing the external server. However, internal researchers have the flexibility to transfer files to and from their accounts in *Pitagoras*, including data and output logs.

When internal researchers collaborate with external co-authors, it becomes their responsibility to ensure that all files shared with external co-authors comply with BPLIM's data security and confidentiality policies. This ensures that no sensitive or restricted information is improperly disseminated.

For further details, please refer to the **Output Control Guide**.

5.2 What restrictions apply to the release of output logs?

As a general principle, BPLIM will not verify the "correctness" of scripts used by researchers to generate logs and expressly disclaims responsibility for any errors or inaccuracies in researchers' code. This responsibility solely rests with the researcher.

Output logs should never contain information that discloses individual dataset records; they should only include **aggregate-level information**. Therefore, listings of individual records, tables with cells derived from manipulation of three or fewer observations, statistical measures with standard errors of zero, minimum and maximum values, etc., are prohibited.

Plain text files (including Latex and comma or tab separated values) are the preferable formats, although other formats may be accepted provided that the data content can be easily verified. **Graphical outputs** should be generated in ".png" format.

In mixed projects, the internal researcher is responsible for ensuring adherence to these principles. BPLIM will assist in this process upon request from the internal researcher. For projects involving only external researchers, BPLIM staff will verify the conformity of all outputs.

Output disclosure control depends on staff availability and may take longer in periods of high workload. BPLIM staff will only answer requests for **output extraction sent by email** and will take the necessary time to ensure that all **confidentiality requirements are safeguarded**. Researchers should keep their output requests to a minimum and, whenever feasible, these requests should be of **final outputs**.

For further details, please refer to the **Output Control Guide**.

5.3 What happens if the researcher violates the rules?

BPLIM assumes that all researchers operate in good faith and should endeavor to adhere to the provisions outlined in the signed **"Declaration on Confidentiality and Use of Data"**. In cases where a researcher engages in behavior deemed inappropriate, BPLIM reserves the right to terminate or suspend all projects involving the researcher and the institution to which he/she is affiliated.

6 Replicability and Data Archiving

6.1 Replicability of work

Researchers working at BPLIM have all the necessary conditions to ensure that their work is reproducible. All **BPLIM datasets are versioned** and can be exactly recreated based on archived extractions. Researchers have access to a **Singularity container** specific to their project, including the respective definition file. This means that the computing enviroment is also replicable. All external files, including external datasets and scripts, are stored in the project folder. Researchers can utilize BPLIM's **Replication App** to verify the replicability of results and to generate a **replication package**. Ultimately, it is up to the researcher to garantee that his/her work is reproducible.

6.2 Replicability by third parties

If requested, BPLIM will work with third-parties such as **data editors**, **certification services**, or **individual researchers** to provide conditions for replication of results of individual projects.

The replication process consists of opening an account for the "replicator" and providing him/her with access to the same conditions as the researcher(s). Projects by internal researchers may have to be replicated in "surrogate" mode.

To **facilitate replication**, researchers should unequivocally identify all datasets used in the analysis, ideally using the **Replication App**. All other needed files should be provided by the "replicator".

BPLIM will work directly with data editors or certification services (e.g. Cascad) to evaluate the best approach to implement their replication protocol. In the case of individual researchers willing to act as "replicators" they will have to go through the standard process of submitting a research project.

6.3 Archiving

By default, once a project is closed BPLIM will keep a **copy of all syntax files** (e.g. text files with "do", "R", "py", and "jl" extensions) plus all files found in the "*initial_dataset*" and the "*tools*" folders.

A copy of the syntax files will be sent to the researcher, who should verify that the list is complete. All other files will be deleted unless the researcher explicitly requests that certain external file(s) are archived. Ideally, the researcher should use the Replication app and **save the replication package** created by the application.

Only in exceptional and well justified circumstances will BPLIM agree to archive intermediate data files created by the researchers. It is the responsibility of the researcher to ensure that all files needed for proper replication of the results are archived with the project.

Archives are kept for **ten years** since the closing of the project. BPLIM may archive a project that has been **inactive for more than one year**.

7 Publication of Research Papers

7.1 Citations and research outputs

All BPLIM datasets should be cited according to the information provided in the data manual. If available, the **Digital Object Identifier** (DOI) should be referenced.

When the topic analyzed by the external researcher(s) bears special relationship to BdP's statutory tasks, researchers are encouraged to **discuss their results** with BdP staff. In this case, BdP may ask to see a copy of the work, prior to any public release, and may provide suggestions regarding the research project.

Moreover, in situations where data access is granted based on the relevance of the topic (**"surrogate access"**) researchers are not only encouraged to discuss their results with BdP staff but must also seek BdP **approval prior to any public release** of their results.

As soon as available, researchers are **required to send to BPLIM a copy of all research outputs** (working paper, conference proceedings, paper, thesis, etc.) related to the project.

8 Contacting BPLIM

8.1 How can I contact BPLIM?

The preferred way to contact BPLIM is through email: **BPLIM@bportugal.pt**. For projects already ongoing the subject line should always include the project reference (eg: p###_Surname). If necessary you can contact us at:

Address:

Banco de Portugal

Microdata Research Laboratory

Rua do Almada, 71

4050-036 Porto

Portugal