

Secure Access to Sensitive Microdata: The BPLIM Approach

WORKSHOP ON DATA SECURITY AND PRIVACY



BANCO DE PORTUGAL
EUROSYSTEM

Monday, 14 September 2026

PAULO GUIMARÃES, GUSTAVO IGLESIAS, RITA SOUSA

Data Analyst

Banco de Portugal Microdata Research Laboratory - BPLIM



BANCO DE
PORTUGAL
EUROSISTEMA



Overview



01 | BPLIM Overview

02 | Pseudo-data Workflow

03 | Pseudo-data Tools

04 | Future Challenges



BPLIM

Brief Overview



- ❖ **Established:** 2016, Banco de Portugal (Porto branch)
- ❖ **Role:** Microdata Research Laboratory
- ❖ **Mission:**
 - ✓ Support internal and external research on the Portuguese economy using microdata
- ❖ **Main challenges**
 - ✓ Balancing data confidentiality with research access
 - ✓ Promoting credible and impactful research



BPLIM

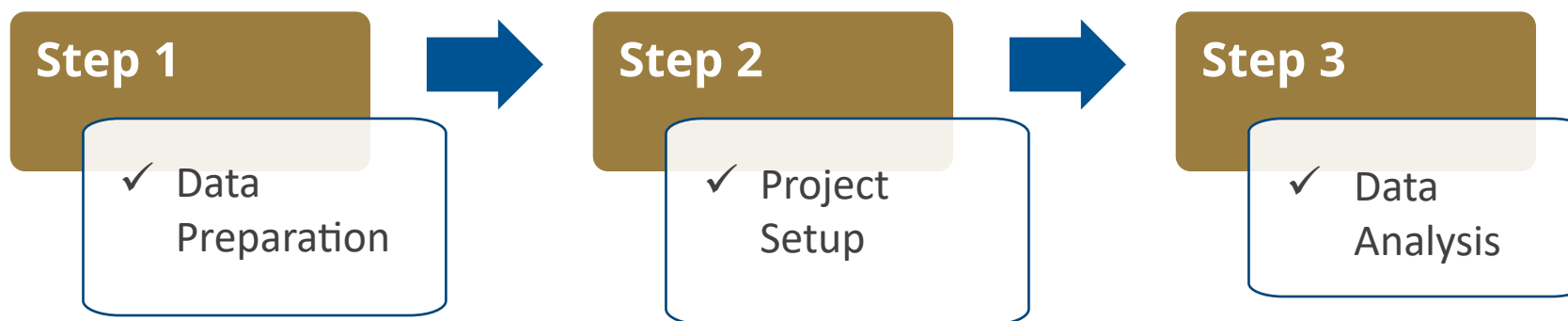
Levels of Confidentiality



- ❖ Type of access depends on the **level of confidentiality** of data: **low**, **medium** and **high**.
- ❖ **Internal researchers** have access to **all datasets**, **pseudo-anonymized**.
- ❖ **External researchers** have access to:
 - ✓ **Pseudo-anonymized** data, if the dataset is classified as **low** level of confidentiality;
 - ✓ **Modified data**, if the dataset is classified as **medium** or **high** level of confidentiality;
 - ✓ For data with a **high** level of confidentiality, access is only possible through **pseudo-data**.



- ❖ When researchers work with **high confidentiality** data, we prepare and make available **pseudo-data**
- ❖ The process involves **three steps**:





BPLIM

Step1: Pseudo-data Structure and Preparation



- ❖ Researchers identify the **sensitive micro datasets** and **linking variables**.
- ❖ BPLIM generates a **sample of pseudo-anonymized IDs**.
- ❖ The **metaxl tool** ([GitHub](#)) is then used to **extract metadata** and **descriptive statistics**.
- ❖ The sample of IDs and metadata are used as inputs to the **dummyfi tool**, which generates the code for producing the pseudo-data.



BPLIM

metaxl: Metadata Management in Stata



❖ **Purpose:**

- ✓ **Extracts metadata** from the dataset in memory to an Excel file
- ✓ **Optionally adds** descriptive statistics to the Excel file
- ✓ **Applies** metadata from an Excel file to the dataset in memory
- ✓ **Identifies** inconsistencies and corrects problems in metadata

❖ **Stata Journal** article: [metaxl](#): *A Package of Tools to Handle Metadata*



BPLIM

dummyfi: Generate scripts for pseudo-data



- ❖ **Dummyfi** is a package that uses the **information on the metafiles** and the sample of **IDs** to **produce Stata scripts** that generate pseudo-data:
 - ✓ All **variables** are **generated randomly**
 - ✓ **Continuous variables** assume values between **p5** and **p95**
 - ✓ **Categorical values** are generated with **similar frequencies** to original values
 - ✓ **Proportion of missing, zeros, negative** values and **time-invariant** observations is maintained
 - ✓ **Panel structure** is preserved
 - ✓ The **scripts** to generate the dummy datasets **can be edited/modified** as desired
 - ✓ **Dummyfi** is available in [GitHub](#)



❖ **Pseudo-data projects** are set in **BPLIM's External Server**:

- ✓ BPLIM prepares **original** and **pseudo datasets**
- ✓ Researchers select a **container** (built according to researchers' needs) and **tools**
- ✓ BPLIM **sets up an account** for **remote access**
- ✓ **Original** and **pseudo-datasets** share **same metadata**
- ✓ **Pseudo-data** and **tools** are placed as **read-only** on the **external server**
- ✓ **Original data** is placed in an account on the **internal server**

bplimext (\\sipe-bplim-int\er

Name

initial_dataset
results
tools
work_area



BPLIM

Step3: Data Analysis



- ❖ Researchers **develop and test their code** using the pseudo-data.
- ❖ Once ready, they **create a replication package** using the [Replication App](#).
- ❖ They request **BPLIM** to **run the scripts** on the original data.
- ❖ BPLIM staff **executes the code** on the **original data**.
- ❖ **Results** are **shared** with researchers following **output control**.

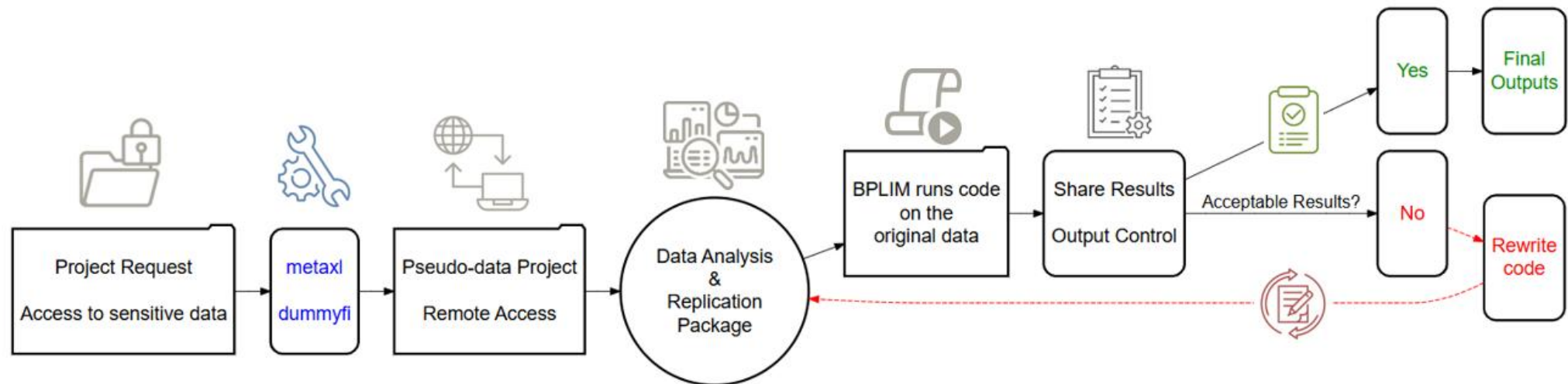


BPLIM

Workflow with highly sensitive data



1. **Researcher** views **pseudo-data** and develops the code.
2. **BPLIM** runs the code on the **original data** and performs **confidentiality checks**.
3. **Researcher** receives the approved results.





Future Challenges



- ❖ **Preserve relationships and dependencies** between variables
- ❖ **Extend support for other statistical languages** (such as R and Julia)
- ❖ Provide **large micro databases** in **parquet** format
- ❖ **Increase efficiency and speed** in creating and manipulating large micro databases (with tools like **DuckDB** and **Polars**)
- ❖ **Explore SDC methodologies for highly sensitive and panel microdata.**



Comments - Questions

Rita Sousa



rclsousa@bportugal.pt



linkedin.com/in/rita-sousa-84379310

Thank you for your attention!

