BANCO DE ESPAÑA
Eurosistema

BELab
BANCO DE ESPAÑA
Eurosistema

Directorate General Economics, Statistics and Research

08/03/2022

BELab Operating Documentation

Output control guidelines

BELab. Banco de España
Statistics Department

CONTENTS

1 Introduction **2**

2 Output Control Principles **3**

2.1 Anonymisation principles **3**

2.2 Principle of verifiability **4**

2.3 Principle of reproductibility **5**

2.4 Principle of reasonable use of resources **6**

2.5 Principle of responsibility **6**

3 Publication control principles **7**

1 Introduction

This document is part of BELab's operating manuals and sets out a series of rules, recommendations and best practices to ensure that the work carried out by external researchers using BELab microdata is disseminated securely.

The Banco de España Data Laboratory (BELab) provides access to high-quality Banco de España microdata. The data provided to BELab are treated responsibly and in accordance with all the legal and internal rules on data processing and disclosure.

The external researchers who work with Banco de España microdata are responsible for ensuring that the results of their calculations do not contravene data confidentiality requirements. When the results of their research are extracted in accordance with the procedures in place in BELab's secure environment, researchers must ensure that the calculation results obtained do not contain any data that may be traced to individual observation units or any of the statistical units comprising the BELab catalogue (individual firms, consolidated groups, etc.), providing sufficient relevant information for BELab staff to be able to perform the final verification, called "output control".

Compliance with these confidentiality requirements for data extraction is also checked by BELab staff during the Output Control process. If these requirements are not satisfied, extraction – and thus publication – of the calculation results will not be possible.

This document aims to make it easier for visiting researchers to comply with the requirements stipulated for extracting their results and for using them in publications. BELab staff extract the results and then make them available to the researchers outside the controlled system. A closer involvement by researchers in the process will lead to a more agile review of data and an improvement in the use of BELab resources. It will also help to enhance the confidence of data producers and providers, who will see how their information is used securely and responsibly. Therefore, depending on the database used, output could also be controlled by data providers, as a third protective layer (the researcher, BELab staff and data providers).

BELab reserves the right to amend, supplement or expand the following principles and rules during the Output Control process, if necessary.

The following sections set out the principles and rules governing Output Control and the rules on publications.

2 Output Control Principles

2.1 Anonymisation principles

The following rules aim to facilitate compliance with data confidentiality regulations. All results intended to be extracted from BELab's secure environment must be fully anonymised. This means that no individual agent may be identified, directly or indirectly (by deduction), taking into account the possible ways in which third-party users might re-identify the microdata. External researchers are at all times responsible for ensuring that their results meet the following full anonymisation criteria:

1. **Non-extraction of identifiers:** Identifiers cannot be included in the results to be extracted or in the codes that generate them.

2. **Non-extraction of microdata:** No result may contain microdata. This means that data subsets, tables, charts, codes or log files containing microdata may not be extracted. Consequently, the extraction of minimum and maximum values and of quantiles strictly corresponding to one item of microdata or which do not meet criterion number 7 is not permitted.

3. **Minimum number of observations:** All the results to be extracted should be based on at least three different observations. This applies both to aggregate results (averages, medians, etc.) and to charts and tables (at least three observations per cell/information node). The simplest way to demonstrate compliance with this criterion is to always generate the frequency table associated with each result.

4. **Degrees of freedom:** Regression models must be calculated with at least ten observations and must also have at least ten degrees of freedom.

5. **Degrees of freedom are calculated as:** Number of observations – number of estimated parameters (variables) – other model restrictions.

6. **Dominance rule (p%):** To be certain that it is not possible to identify any entity, even when the minimum of three observations is met, it is necessary to ensure that the largest observation does not exceed 85% of the total weight of the value analysed, or of any other weighting used. This prevents indirect identification of any high weight observations.

   For example, to calculate total sales in a specific sector in a specific year we only have three firms. The total volume is €100 million, of which €90 million relate to the largest firm and €5 million to each of the others. In this case, the largest firm is potentially identifiable owing to its contribution to the total value.

7. **Confidentiality in multiple tables, control of differences:** If the results are calculated based on a G population, but are subsequently recalculated for an X subset of G, the rules explained above must be met for observations of the

difference. Otherwise, the individual observations could be identified on the basis of the differentiation.

For example, we have a table with all the firms in a given sector and another with the firms in that sector that exceed an X volume of sales. We would have to create a third table with the firms that do not reach such X volume and check that the confidentiality criteria are met in that table; otherwise, the firms could be identified by differentiation.

8. **Dichotomous (0-1) categorical variables (dummies):** If the averages of these variables are calculated, there must be at least three observations for each category (three observations with 0 and three with 1).

9. **Treatment of zeros and missing values:** Zeros are permitted in regressions and descriptive statistical analysis, provided they do not represent missing values in dichotomous and categorical variables. In descriptive statistics, missing values will not be taken into account for determining the number of different observations used. If missing values are imputed, the number of imputed and observed observations should be reported.

2.2 Principle of verifiability

The Output Control process involves considerable time and effort by the BELab Team. To optimise the use of the Data Laboratory resources and minimise wait time after the data extraction is requested, external researchers must comply with the following rules:

1. **Master File:** A master file must be created that contains all the relevant information on the research project and calls all of the sub-programs used.

2. **Log file:** The log file function must be activated for each program code. The registration must commence before the description of the research project's content and before the first line of the calculation code.

3. **Order and structure within the code:** The code must be structured in a visually clear manner, such that the individual blocks (header, individual analytical stages, etc.) can be visually distinguished. Loops must have indentations. Long programmes and analytical steps must be divided into smaller code files, e.g. ("0_master.do", "1_data_preparation.do", "2_descriptive_analysis.do" and "3_regressions.do").

4. **Comments on programme codes:** The programme code must have sufficient comments to ensure that even people who are unfamiliar with the project are able to understand it within a reasonable timeframe.

5. **Clear names for output files**: The names of all the output files must start with the same name as the programme used to generate the file and should be numbered logically.

6. **Clear names for variables:** All the assigned names must be as informative as possible and be used consistently. Labels and brief descriptions of the variables must be provided for all the data generated by the user and for all the data originating externally. If (categorical) variables are created or modified, the corresponding value labels should be assigned to these values.

7. **Specification of compliance with anonymisation rules:** Researchers must include a code justifying compliance with the anonymisation requirements described above. Thus, they should, for example, provide frequency tables, model descriptions, or any other element evidencing compliance with the rules for the output extraction requested.

8. **Re-assessment of output:** In the case of a request for extraction of a previously revised code on which minor changes have been made, these must be specifically reflected in the new request. Where possible, researchers should only submit for assessment the programme elements that have been modified.

2.3 Principle of reproductibility

To check compliance with the anonymisation requirements specified in principle 3, all the results of the calculations presented for review and extraction must be reproducible. Researchers must comply with the following rules:

1. **Program code reproductibility:** All results of calculations must be generated without problems by a "0_master.do" program that is able to be executed without errors and must contain all the analysis programs used throughout the project. This program must commence with the uploading of the original data provided by BELab. The program must always execute the same steps and produce exactly the same results as those presented for review. In the "0_master.do" file, every program call must be followed by a brief description of the sub-program's content.

2. **Software reproductibility:** All computer software used to generate the calculation results should be clearly described at the beginning of the "0_master.do" file (name and version number). Together with the analysis software version number, the names of all the packages used (e.g. R, Python, Octave) are to be included.

3. **Data reproductibility:** All BELab data subsets used to generate the results of the calculations must be clearly described at the beginning of the "0_master.do" file (DOI if available, variables and year). Any dataset used originating from external data providers must also be described in the "0_master.do" file.

4. **Reproductibility of output to be published:** All the results of the calculations whose extraction is sought should be able to be found quickly, easily and clearly in the output produced by the analysis programs. To this end, the code used to generate the results to be extracted and published must be clearly listed. In this connection, BELab recommends creating a "master_yyymmdd_publication.do"

document file only containing the last steps for generating the calculations, tables and charts comprising the output whose extraction is requested.

2.4 Principle of reasonable use of resources

As a general rule, elements whose extraction from BELab is requested are only to be directly used in a publication. For this reason, visiting researchers must respect the principle of reasonable use of resources, especially at the time of deciding which results they wish to extract. The number of elements to be presented must be consistent with what is normally expected in the sphere of an empirical scientific article.

In general, visiting researchers must take into account the following rules:

1. **Exploratory data analysis cannot be part of the output to be extracted.** Only analyses that may be published directly can be presented for review. The task of selecting results worthy of publication is part of the work to be conducted by researchers in BELab.

2. **Maximum number of lines in the output.** BELab does not establish, a priori, a maximum number of code lines to be reviewed during the Output Control process, but it reserves the possibility of doing so if researchers do not use the Laboratory's resources reasonably. This entails being prudent in the quantity of output requested, program use, execution times, use of sessions, etc.

2.5 Principle of responsibility

Visiting researchers are responsible for ensuring compliance with all the principles and rules set out in this document. Failure to comply with these rules will entail BELab’s refusal to deliver the calculation results. Researchers must respect the following rules:

1. **Checking all calculation results for the purpose of publication:** Before researchers ask BELab to review an output they wish to extract, they must first check that the Output Control principles have been applied. Once that has been done, they will ask the BELab team to review their data. They will place the output requested in the /Out/Output folder of their project and indicate, as detailed in these guidelines, the elements required to carry out their review.

2. **Functioning of the code:** If the program code contains syntax or other errors, BELab will leave the codes uncorrected and ask the researchers to correct them.

3. **Output control format:** Program results and codes will only be accepted for output control if they are editable and are presented as unformatted or .csv files. Charts must have a read-only (static) format and be presented in .jpeg or .png format.

3 Publication control principles

The following rules aim to help researchers comply with the publication control rules ("publication control") more easily.

1. **Review by BELab of all publications:** No information relating to research projects developed at BELab may be published until it has been reviewed and authorised. Authorisation may be withheld if the results to be published do not meet BELab's criteria.

2. **Copy of papers:** Researchers are responsible for delivering a copy of the published papers prepared by them containing the research results of the analyses conducted during their stay at BELab. Researchers who fail to do so will be disqualified from the future use of BELab until the papers published as a result of previous research conducted at BELab are furnished.

3. **Referencing sources:** Researchers undertake to mention the ultimate source of the data in any publication resulting from this study as indicated in the respective guidelines for each database.

4. **Referencing charts and tables:** All charts and tables should be referenced as follows: "Source: BELab. Banco de España Data Laboratory, <name of the set of microdata used from the BELab catalogue (if appropriate, with the common abbreviation)>, <period during which the microdata were used>, own calculations.".

5. **Specification of type of data access:** Each publication must specify the type of access the researcher had to the data, e.g. in-person access from a dataroom (indicate whether Madrid or Barcelona), remote access or mixed access.

6. **Specification of datasets used, use of DOI:** All datasets used in the research project must be cited, indicating the name and, where appropriate, the DOI.