EUROPEAN CENTRAL BANK

EUROSYSTEM

PILOT PROJECT ON RESEARCH ACCESS TO CSI COLLECTED UNDER COUNCIL REGULATION 2533/98

CRITERIA FOR OUTPUT CONTROL

1. Introduction and legal basis

This document sets out a series of rules, recommendations, and best practices to ensure that the work carried out by external researchers using Eurosystem microdata is disseminated securely. They are based Art 8(1) and 8(3) of Council Regulation 2533/99 and in particular on Art 2(3) Guideline ECB/1998/NP28[1].

2. Output Control Principles

The following rules aim to facilitate compliance with data confidentiality regulations:

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 entails refraining from extracting subsets of data, as well as tables, graphs, codes, or log files that contain microdata themselves. Consequently, the extraction of minimum and maximum values 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.

[1] Art 2(3): "*all published data to cover at least three economic agents, and that when one or two make up a sufficiently large proportion of any observation, the data shall be arranged in such a way as to prevent their indirect identification*".

5. Dominance Rule (%p): It is necessary to ensure that the largest observation does not exceed 85% of the total weight of the analysed value or any other weighting used.

Example: For calculating total sales in a specific sector for a particular year, let's consider only 3 companies. The total volume is 100 million euros, with the following composition: 90 million from the largest company, and 5 million from each of the others. In this case, the largest company is potentially identifiable due to its contribution to the total value.

6. 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.

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.

7. Dichotomous (0-1) Categorical Variables (Dummies): When calculating averages of these variables, there should be a minimum of three observations for each category (three observations with 0 and three with 1).

8. 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.

3. Publication Control Principles

The following guidelines aim to assist researchers in more easily complying with the publication control standards ('publication control').

1. Copy of Publications: It is the responsibility of researchers to provide a copy of the published works that they produce and that contain research results from the analyses conducted on the accessed datasets.

2. Referencing sources: The researcher commits to mentioning the ultimate data source in any publication resulting from this study, as indicated in the respective guide of each database.

3. Referencing charts and tables: All charts and tables should be referenced as follows: “Source: ECB, <name of the set of microdata used>, <period during which the microdata were used>, own calculations.”

4. Specification of type of data access: Each publication must specify the type of access the researcher had to the data, e.g., remote execution, on-site access (location), remote access or mixed access.