February 2018

RULES FOR RELEASING THE RESULTS OF DATA PROCESSING

Extract from:
Laboratory for Microdata Analysis (ADELE)
User guide

For further information: adele@istat.it.

The aims of the Laboratory

The ADELE Laboratory meets the needs of statistical analysis for scientific research purposes when the use of microdata is required, and they are not available in other forms.
Within the Laboratory, the security of data and statistical confidentiality are ensured by controlling both the environment and the results carried out by the users of the Laboratory.

When all processing related to the research project are concluded, the output to be released will be assessed, as far as it concerns statistical confidentiality, by the ADELE Laboratory staff.

What can and what cannot be obtained

Some specific rules for the most frequent output disclosure are listed below:

1) **Descriptive statistics and tables to support statistical models**
   Each value reported in tables or descriptive statistics in general must refer to at least 10 statistical units.
   In details:
   - descriptive statistics that report precise data on the individual units (for example maximum and minimum for continuous variables) cannot be released;
   - mode, minimum and maximum can be released if the modes that identify them are assumed by at least 10 units;
   - quantiles: the median is considered releasable if referring to a distribution of at least 50 units; the other quantiles are not released except for special cases to be agreed;
   - means, ratios and indicators: these outputs must be presented in their disaggregated form (for example for means and ratios separate numerator and denominator should be presentes; for means of dichotomous variables also the complement should be presented, etc.); each element must be accompanied by the number of units (at least 10) that contribute to determine its value; this also applies to any additions, also to be explicitly presented (for example, if an indicator shows the value of 95%, it should be possible to verify that even 5% corresponds to at least 10 units; dichotomous variables, etc.);
   - intensity tables: users must specify the number of units (at least 10) that contribute to determine the value of each cell (match the relative frequency table);
   - frequency tables: tables with a cell number of less than 10 unweighted units are not released in any case.

2) **Charts on variables**
   Charts on non-continuous variables must be accompanied by the corresponding table of the values they represent; this will be evaluated as specified in the previous point. Charts on continuous variables must be saved as images and deprived of the abscissa values.

3) **Regressions**
   The following output can be released:
   a) (p-1) *estimated parameters*, where p is the number of regressors, when all the conditions specified below are verified:
      ✓ the total number of observations must exceed the number of explanatory variables of at least 100 units;

- ✓ among the explanatory variables the presence of at least one variable for which the operations of sum, difference, product and quotient make sense is necessary;
- ✓ the observations on all data must refer to at least 100 different units of analysis.

b) Diagrams on the correct model specification:

1. the histogram of residues, deprived of the abscissa values;
2. the diagram of the density of residues, deprived of the abscissa values;
3. the Q-Q plot of residues, deprived of the abscissa and ordinate values;
4. the P-P plot of residues;
5. the diagram of the ranks of the residues vs. the ranks of the predicted values of the variable to be explained;
6. the diagram of the ranks of the residues vs. the ranks of an explaining variable;

c) Statistics on the adaptation and correct specification of the mode:

1. The statistics expressed by a scalar;
2. The statistics expressed by a vector, having a size not greater than the number of parameters estimated, that is (p-1). Only the conventional level of significance is released from the obscured regressor (0.005, 0.01, 0.025, 0.05, 0.1).

However are excluded from the release:

1. residuals of regression;
2. "aforesaid" values of the variable to be explained.

4) Factorial analysis and models with structural equations

The following output can be released:

1. Model parameters,
2. the (possible) correlation matrix among factors,
3. standard errors and statistics on the significance of the model parameters,
4. communality and specificity for each variable,
5. factorial scores referring to units of analysis that are not individuals, families or enterprises,
6. statistics on the goodness of the model, expressed by a scalar,
7. the scree plots relating to the eigenvalues of the covariance / correlation matrices observed,
8. diagrams of relational models between manifest and latent variables.

5) Analysis by principal components

The output listed below can be released:

1. eigenvalues,
2. the following statistics:
   a) variance explained by the factor axes,
   b) matrix (p x k) of the relative contributions (squares of the cosines) of the points-variable,
   c) matrix (p x k) of the absolute contributions of the points-variable,
   d) matrix (p x k) of the coordinates of the points-variable,

where p is the number of variables and k is the number of eigenvalues which, organized in non-decreasing order, accumulate a fraction of the total variability not exceeding 85%,

3. scree plot of the eigenvalues,
4. diagrams of the projection of the variable points on the factorial plans.

6) Analysis of correspondences

The output listed below can be released:

1) eigenvalues,
2) the following statistics:

a) inertia explained by the factorial axes,
b) matrix (p x k) of the relative contributions (cosine squares) of the mode-points (column and/or row),
c) matrix (p x k) of the absolute contributions of the mode-points (column and / or row),
d) matrix (p x k) of the coordinates of the mode-points (column and / or row), where p does not exceed the total number of modes and k is the number of eigenvalues which, ordered in non-decreasing order, accumulate a fraction of the total inertia not exceeding 85%,
e) test values, expressed by scalars, on the significance of each additional mode (in the analysis of multiple correspondences),

3) scree plot of the eigenvalues,
4) diagrams of the projection of the row and / or column point-modes on the factorial plans.

With regard to the units of analysis, for any type of processing, the values observed and the statistics that do not comply with the rules on "Descriptive statistics and tables" remain excluded from the release.

Rules for the output presentation

- Producing results without weight normalization is strongly discouraged; however, for assessment purposes, the users are invited to communicate if standardized (normalized) weights are used and how (whether it is a normalization to the total population or to a specific sub-population);
- The output **volume** may be itself a reason for refusal to disclosure: the output to be released should be minimal and correspond to the output included in the work to be disseminated; as an indication, a maximum of 30 pages is suggested (~ 60Kb in ASCII text format);
- In order to enable assessment by the Laboratory Staff, the output is preferred in a text file, either Word or Excel, but not in the format of the owner of the statistical applications used; in case of descriptive statistics or tables the excel format should be used;
- The output should be organized in order to be released as it is, with no need of changes by the Laboratory staff during assessment; in case of no releasable output some further appointments should be agreed in order to make it releasable.;
- The output should be clearly and completely documented according to the "Output record" (*cf.* Annex 1), in there the following information should be specified; the purpose, the analysis modes, the output file's name and contents, treatments performed on the original dataset and the possible (sub)populations under analysis, the meaning of each variable (the definition of the derived

variables should be included), as well as any other useful information intended to a correct interpretation of the output file. The output record should be enough to understand it (reference to other sources such as, for example, the used syntax files is not permitted);

- **Release of intermediate output (**i.e. not concluding the work**) is not permitted**.

When a project is concluded the users are required to optionally answer a short questionnaire aimed at assessing the service from the user's point of view. The form doesn't include personal data, and the collected information are exclusively used to produce reports on the service quality and are not in any way disseminated in connection to the users' personal data. The questionnaire will be transmitted to the user via email at the end of the project.

ANNEX 1

OUTPUT RECORD

DATA USED

Specify, among the data provided, those actually used in the processing for which release is requested: indicate the name and the reference period of the survey(s) used and specify any external data files used in the processing.

VARIABLES/INDICATORS

Report the name and a brief description of the variables used. In the case of variables not present in the original data bases (reclassifications made by the user, external variables etc.) in addition to the name and description, report the meaning of the methods assumed (or the construction method, especially in the case a variable takes values as a function of other variables).

TRANSFORMATIONS OPERATED ON VARIABLES

For each variable supplied by the laboratory and subjected to transformation, indicate the function used to obtain its transformation. For each variable created by the user, indicate in detail the construction procedure.

OUTPUT FILES

Report the name and structure (example: Excel file with a sheet per each considered year) of the output files for which release is requested, providing a summary description of the content.

PROCESSING CARRIED OUT

Describe each elaboration carried out, providing a brief but complete description.

It is useful to associate a denomination to each processing and report it in the output file, so as to ensure an unequivocal identification and interpretation.

FILTERS ON UNITS

For each processing (or processing group) specify the filters applied to the starting population and the number of observations involved.

Note that it is necessary to specify exactly the effective number in each processing, even in the case of reductions in the number due to the presence of missing values in one or more of the used variables.

WEIGHT SYSTEM

Specify the weight system you used and if this varied in the different processes.

If standardized (normalized) weights were used, specify whether the normalization was with respect to the total population or specific subpopulations.

Note that if you request the release of weighted output, the same must also be presented in the unweighted version to allow evaluation.

NOTES

Report any other information you may consider useful for a correct interpretation of the output files.

The applicant: _____ Date: ___/___/_____

N.B.: the output description should be sufficient to understand it; reference to other sources (i.e. the used syntax files)is not allowed.