Methods and IT tools for statistical production

Content of web area

To send information and/or updates on methods
and IT tools please write to softgen@istat.it

Last update: **November 2022**

INDEX

Methods and IT tools for statistical production

The Repository of the methods and IT tools for statistical production contains:

- **statistical methods**
- **generalized IT tools**

used at Istat for the production of statistical outputs.

The Repository is organized by **phases of the statistical production process**, in accordance with the Generic Statistical Business Process Model (GSBPM) Version 5.1.

PHASES OF THE PROCESS

DESIGN	METHODS	TOOLS
COLLECT	METHODS	TOOLS
PROCESS	METHODS	TOOLS
ANALYSE	METHODS	TOOLS

The phases of the production process are the "gates" of access to information and specific materials. Only the phases Design, Collect, Process, Analyse, i.e., phases for which methods and tools are currently available, are taken into account.

The PHASES pages contain the most important information about the sub-processes within the selected phase.

Methods and tools are grouped by sub-processes. Methods and tools relevant to more than one sub-process are attributed to the prevailing sub-process.

The METHODS pages contain lists of documents that describe statistical methods. The documents are Istat standards (or European Statistical System standards), or methodological manuals, publications and reports drawn up by Istat researchers (also in collaborations). For a given sub-process, the documents are displayed in reverse chronological order (newest to oldest) and are available in PDF format or by linking to external sites.

The TOOLS pages contain lists of generalized IT tools that implement statistical methods. For a given sub-process, the tools are listed in alphabetical order. For each tool a general description showing the functions implemented is provided together with an information sheet.

Tools developed by Istat can be downloaded for free. How to find the tools not owned by Istat is also shown.

Latest updates

- ReGenesees (R evolved Generalised software for sampling estimates and errors in surveys) – Published the **release 2.2** of the full-fledged R software for design-based and model-assisted analysis of complex sample surveys. The new function **'svyDelta'** computes estimates and sampling errors of complex Measures of Change derived from two not necessarily independent samples. November 7, 2022
- STATMATCH – Published the release 1.4.1 of the R package for data integration through statistical matching and, as a by-product, the ability to impute missing values in a data set. June 23, 2022
- R2BEAT (Multistage Sampling Design and PSUs selection) – Published the 1.0.2 release of the generalized software for multivariate optimal allocation for different domains in one and two stages stratified sample design. February 2, 2021

To stay informed on the latest updates of methods and IT tools, you must read the privacy statement and provide an e-mail address by writing to: softgen@istat.it.

DESIGN phase

Design frame and sample

The activities related to the design of the list and the sampling methodology refer to sub-process 2.4 "*Design frame and sample*" of GSBPM. More precisely:

- Construction of the selection frame of the target population, containing, for each unit of the population, all the identifying information needed for the contact, any auxiliary variables used to define the sample design (stratification variables, identifying variables of any selection stage);
- Design of the sampling design which, on the basis of the objectives specified in Phase 1 "Specify Needs" and of the operating and cost constrains, allows obtaining estimates as accurate as possible.

Design of the selection list

The characteristics of the sample list are essential for the correct definition of the sampling design.

The list should meet quality criteria in terms of refreshing, coverage and accuracy of the information contained in it. From a theoretical point of view the selection list should ideally have the following requirements:

- it is composed only of the units belonging to the population of interest at the reference time of the survey;

- it includes all units of the population only once;

- it contains the most updated data for identifying variables (name and address) and for any descriptive information (other relevant structural data) of the units.

Possible situations of departure from the ideal list are:

- *under-coverage*, which occur when some elements of the population are not contained in the list and cannot, therefore, be included in the sample;

- *over-coverage*, when some elements of the list are non-existent and / or do not belong to the population of interest;

- *duplication of some units*, when some elements of the population are enumerated more than once in the list;

- *clusters of units*, when some elements of the list contain clusters of elements of the population.

Planning and implementation of the sample design

The planning of the sample design consists firstly of the following activities:

- definition of the sampling scheme, performed on the basis of the cost related to the chosen data collection technique and the information contained in the selection list (multistage sampling selection, stratified sampling). The choice of a multistage design generally derives from the need to concentrate the sample locally in order to limit the cost of interview in case of survey using a direct mode of administration of the questionnaire (face to face interview). The choice of a stratified sample design has the purpose of improving the precision of the estimates and guaranteeing planned sample size for the domains of estimate. The division of the units of the population in strata is carried out using auxiliary variables contained in the list and related to the variables under investigation.

On the basis of the adopted sampling scheme, the following steps may be undertaken:

- **choice of stratification criteria** (choice of variables, choice of the number of strata, definition of the criterion of strata construction);

- **choice of the probabilistic method for the selection of the sample units** (selection with equal probabilities, selection with unequal probabilities). For designs in two or more stages the selection of the primary sampling units (units of the first stage) is generally done with probability proportional to a measure of size suppository correlated with the variables under investigation.

- **determination of sample size** for the different stages of selection and **allocation of the sample** between the strata on the basis of the sampling error admitted for the main estimates, in relation to the reference domains and subclasses of the population. Since surveys are usually designed for the production of a variety of estimates for different domains of interest, it is necessary to use approaches that address the problem in a global perspective of determining the optimal sample size in the presence of a high number of objectives and constraints.

DESIGN phase – METHODS

- **DESIGN FRAME AND SAMPLE**

Generalized Framework for Defining the Optimal Inclusion Probabilities of One-Stage Sampling Designs for Multivariate and Multi-domain Surveys
2015
Survey Methodology, 41.

MEMOBUST – Handbook on Methodology of Modern Business Statistics:
Statistical Registers and Frames
Sample Selection
2014
MEETS ESSnet MEMOBUST

References

Falorsi, S., C. De Vitiis, L. Di Consiglio, C. Ceccarelli, A. Cutillo, e C. Rinaldelli. 2008. "Strategia di campionamento e precisione delle stime". In "L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc)". *Metodi e Norme*, N. 37. Roma: Istat.

De Vitiis, C., L. Di Consiglio, S. Falorsi, e R. Ranaldi. 2006. "Il disegno campionario della nuova indagine e la fase di estrazione". In "La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione". *Metodi e Norme*, N. 32. Roma: Istat.

De Vitiis, C., e S. Falorsi. 2006. "Il piano di campionamento". In "Il sistema di indagini sociali multiscopo. Contenuti e metodologia delle indagini". *Metodi e Norme*, N. 31. Roma: Istat.

De Vitiis, C., e A. Pareto. 2006. "Strategia di campionamento e livello di precisione delle stime". In "L'indagine campionaria sulle nascite: obiettivi, metodologia e organizzazione". *Metodi e Norme*, N. 28. Roma: Istat.

Cicchitelli, G., A. Herzel, e G.E. Montanari. 1992. *Il campionamento statistico.* Bologna: Il Mulino.

Särndal, C.E., B. Swensson, and J. Wretman 1992. *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Bethel, J. 1989. "Sample Allocation in Multivariate Surveys". *Survey Methodology*, Volume 15, N. 1: 47-57.

Chromy, J. R. 1987. "Design Optimization with Multiple Objectives". In *Proceedings of the Survey Research Methods Section*: 194-199. American Statistical Association.

Cochran, W.G. 1977. *Sampling Techniques.* New York: J. Wiley & Sons.

DESIGN phase – TOOLS

- **DESIGN FRAME AND SAMPLE**

FS4 (First Stage Stratification and Selection in Sampling). FS4 is a generalized software for first stage stratification and selection in sampling related to two or more stages, implemented completely in R and with a GUI (Graphical User Interface).

MAUSS-R (Multivariate Allocation of Units in Sampling Surveys – version R with Java interface). Software for the determination of the sample allocation in the multivariate case and for multiple domains of estimate for single selection stage samples.

Multiway Sample Allocation R-package implementing the sample allocation for one stage multi-way stratified (simple or with varying inclusion probabilities) sampling and for incomplete stratified sampling designs. The allocation allows to control the expected sampling errors of the estimates of many parameters and several reference subpopulations.

R2BEAT (Multistage Sampling Design and PSUs selection) R-package implementing the multivariate optimal allocation for different domains in one and two stages stratified sample design.

SamplingStrata. Optimal stratification of sampling frames for multipurpose sampling surveys.

FS4 (First Stage Stratification and Selection in Sampling)

Description

FS4 is an open source generalized software for first stage stratification and selection in sampling related to two or more stages, implemented completely in R language and with a GUI (Graphical User Interface).

It merges two data frames (whatever PSU – Primary Sampling Unit – population and allocation data frames) and then, computes stratification and selection of a fixed number of sample PSUs per stratum using the Sampford's method (unequal probabilities, without replacement, fixed sample size), implemented by the UPsampford function of R package "sampling".

As for stratification, the function carries out, for each estimation domain, the computation of a size threshold for a given size PSU. PSUs with measure of size exceeding a calculated threshold are identified as SR (Self Representative) and each constitutes a stratum on its own. The remaining NSR (Non Self Representative) PSUs are ordered by measure of size and divided into strata having sizes approximately constant to the corrected threshold and with PSUs having sizes as homogeneous as possible. For each PSU, FS4 determines the size of final sample units on the basis of the size planned in input.

The main features of FS4 are the following:

- it is a flexible software allowing the user to:
  - choose the measure of size to define stratification and inclusion probability for PPS (Probability Proportional to Size);
  - insert as input a single or plural (variable between domains) planned size of final sample units to observe in each PSU;
  - launch the procedure in two separate steps, so to observe an initial output and on the basis of this modify the input parameters;
- it is an open source package released under EUPL licence;
- it has a user friendly GUI that also allows to non-expert R users to use it, implemented in Tcl/Tk by means the R "tcltk" package, although for expert R users the StratSel function from R command line can be used directly.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	2.4 Design frame and sample 4.1 Create frame and select sample
Programming language:	R
Language of the GUI:	EN
Keywords:	stratification, selection, sampling, PSU
Contact:	name: Raffaella Cianchetta email: cianchet@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

R (version ≥ 3.0.1)

R packages: tcltk2, svMisc, plyr and sampling

COPYRIGHT

Copyright 2013 Istat

Licensed under the European Union Public Licence (EUPL), version 1.0 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://ec.europa.eu/idabc/eupl.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 07/07/2020

- FS4 Version 2.0 – Windows binaries
- FS4 Version 2.0 – Package source

INSTALLATION

Install the downloaded package from within R as follows:

> install.packages(path_to_file, repos = NULL)

where the character path_to_file is the path to the .zip or .tar.gz file you downloaded.

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Reference manual – FS4 v. 2.0

User manual – FS4 v. 2.0

MAUSS-R (Multivariate Allocation of Units in Sampling Surveys)

Description

MAUSS-R is an open source generalized software for the determination of the sample allocation, developed using the R language and provided with a Java interface to facilitate the selection of data and parameters of interest.

The reference methodology is an extension of the Neyman allocation method to the case of several variables and adopts as a method of solving a generalization of the Bethel's proposal (1989).

The generalized Bethel algorithm allows to calculate the sample allocation for stratified sampling designs: the overall sample size and the allocation among the different strata is determined starting from the constraints imposed in the survey on the accuracy of the estimates of interest.

The software in the current version allows the determination of the sample allocation in the multivariate case and for multiple domains of estimate for single selection stage samples.

The interface allows:

- changing the default values of the processing parameters (such as the minimum number of units per stratum);
- the management of the file containing the constraints of precision and the storing of different versions of the constraints and the obtained results;
- the execution of the calculation of the optimal solution;
- the comparison with the proportional allocation and equal;
- viewing reports on population, the results and the comparison of the results obtained changing the constraints.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.0
GSBPM code:	2.4 Design frame and sample
Programming language:	R, Java
Language of the GUI:	IT, EN
Keywords:	sampling design, optimal allocation, sample size, Bethel
Contact:	name: Maria Teresa Buglielli email: bugliell@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

R (version ≥ 2.7.0)

Only for graphical interface: Java 2 Runtime Environment (version ≥ 6.0)

COPYRIGHT

Copyright 2013 Istat

Licensed under the European Union Public Licence (EUPL), version 1.0 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://ec.europa.eu/idabc/en/document/7330.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 02/09/2013

MAUSS-R version 1.1

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

User manual – MAUSS-R v. 1.1

Multiway Sample Allocation

Description

R-package Multiway Sample Allocation allows to allocate the sample of multi-way stratified sampling design.

The multi-way stratification is achieved by combining the variables defining the marginal cells of a multi-way contingency table. In the context of the sampling theory the marginal cells are identified by the categories of the variables defining the interest domain partitions being independent of each other. For example, the stratification obtained by combining the categories of two partitions into domains of interest defines a two-way stratification. A partition is denoted as independent or marginal when the domains cannot be got as aggregation of other domains defined in other partitions. The allocation of the sample in the dependent domains is done by aggregating the sample sizes of the marginal domains. In case of multi-way stratified designs, the sample size is fixed for each stratum. In case of incomplete stratified sampling design the sample size is fixed for each domain of interest.

The Multiway.Sample.Allocation package defines the allocation of the sample in accordance with the precision constraints of the estimates for:

- different parameters of interest (total) (multivariate problem) at sub-population or reference domain level (multi-domain issue);
- one stage stratified sampling designs simple or with varying inclusion probabilities variables within the multi-way strata, in which the sample size is fixed at the stratum level; incomplete stratified sampling designs in which the sample size is fixed at the domain level (except for rounding to the above/below integer);
- varying inclusion probability designs where the sample size is fixed (approximately) at the domain level (the performance of the allocation process subject to computational constraints associated with the population size and the number of the estimates to be considered).

The algorithm performing the allocation (Falorsi and Righi, 2015) is an extension of the Chromy (1987) and Bethel (1989) algorithm. The algorithm solves an optimization problem formalized according to a general expression of the variance of the estimates (Falorsi and Righi, 2015) depending on:

- the superpopulation model used to define input parameters (mean and variance of the variable defining the parameters of interest);
- the implemented sampling design.

The main output of the package is the inclusion probability of the population units.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	2.4 Design frame and sample
Programming language:	R
Keywords:	sample allocation, multi-way stratification, incomplete stratified sampling design
Contact:	name: Paolo Righi email: parighi@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

R (version ≥ 3.1.1), R-package MASS.

COPYRIGHT

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 21/11/2016

- MULTIWAY Version 1.0 – Windows binaries
- MULTIWAY Version 1.0 – Package source

INSTALLATION

Install the downloaded package from within R as follows:

> install.packages(path_to_file, repos = NULL)

where the character path_to_file is the path to the .zip or .tar.gz file you downloaded.

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Reference manual - Multiway Sample Allocation v. 1.0

OTHER DOCUMENTATION

De Vitiis, C., P. Righi, and M. D. Terribili. 2016. "*Incomplete Stratified Sampling design for the University graduates' vocational integration survey*". *Rivista di Statistica Ufficiale/Review of official statistics*, N. 2/2016: 87-98.

Falorsi, P. D., and P. Righi. 2016. "A flexible tool for defining optimal sampling designs." *The Survey Statistician*, N. 73: 21-31.

Falorsi, P. D., and P. Righi. 2015. Generalized Framework for Defining the Optimal Inclusion Probabilities of One-Stage Sampling Designs for Multivariate and Multi-domain Surveys. *Survey Methodology*, Volume 41, N. 1: 215-236.

Falorsi, P. D., and P. Righi. 2008. A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation. *Survey Methodology*, Volume 34, N. 2: 223-234.

R2BEAT (Multistage Sampling Design and PSUs selection)

Description

Multivariate optimal allocation for different domains in one and two stages stratified sample design. R2BEAT extends the Neyman (1934) – Tschuprow (1923) allocation method to the case of several variables, adopting a generalization of the Bethel's proposal (1989). R2BEAT develops this methodology but, moreover, it allows to determine the sample allocation in the multivariate and multi-domains case of estimates for two-stage stratified samples. It also allows to perform Primary Stage Units selection.

R2BEAT easily manages all the complexity due to the optimal sample allocation in two-stage sampling design and provides several outputs for evaluating the allocation. Its name stands for "R 'to' Bethel Extended Allocation for Two-stage". It is an extension of another open-source software called Mauss-R (Multivariate Allocation of Units in Sampling Surveys), implemented by ISTAT researchers. Mauss-R determines the optimal sample allocation in multivariate and multi-domains estimation, for one-stage stratified samples.

Information

Status:	validated
Author:	Istat
License:	EUPL
GSBPM Codification:	2.4 Design frame and sample 4.1 Create frame and select sample
Programming language:	R
Keywords:	sampling, two-stages, optimal allocation, multivariate, multi domain, stratified
Maintainer Contact:	Andrea Fasulo email: fasulo@istat.it

Software and documentation

TECHNICAL SPECIFICATIONS

The R2BEAT package needs R (version 2.15.0 or more) and depending packages: plyr, sampling, deytools.

COPYRIGHT

Copyright 2021 Istat

Licensed under the European Union Public Licence (EUPL), version 1.2 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: https://joinup.ec.europa.eu/collection/eupl/eupl-text-eupl-12. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 02/02/2021

- R2BEAT Version 1.0.2 – Windows binaries
- R2BEAT Version 1.0.2 – Package source

INSTALLATION

Install the downloaded package from within R as follows:

> install.packages(path_to_file, repos = NULL)

where the character path_to_file is the path to the .zip or .tar.gz file you downloaded.

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Reference manual – R2BEAT1.0.2

OTHER DOCUMENTATION

R2BEAT website

Presentation at UROS2019 – R2BEAT 1.0.1

SamplingStrata

Description

When designing a sampling survey, usually constraints are set on the desired precision levels regarding one or more target estimates (the Y's). If a sampling frame is available, containing auxiliary information related to each unit (the X's), it is possible to adopt a stratified sample design. For any given stratification of the frame, in the multivariate case it is possible to solve the problem of the best allocation of units in strata, by minimizing a cost function subject to precision constraints (or, conversely, by maximizing the precision of the estimates under a given budget). The problem is to determine the best stratification in the frame, i.e., the one that ensures the overall minimal cost of the sample necessary to satisfy precision constraints. The X's can be categorical or continuous; continuous ones can be transformed into categorical ones. The most detailed stratification is given by the Cartesian product of the X's (the atomic strata). A way to determine the best stratification is to explore exhaustively the set of all possible partitions derivable by the set of atomic strata, evaluating each one by calculating the corresponding cost in terms of the sample required to satisfy precision constraints. This is unaffordable in practical situations, where the dimension of the space of the partitions can be very high. Another possible way is to explore the space of partitions with an algorithm that is particularly suitable in such situations: the genetic algorithm. The R package SamplingStrata, based on the use of a genetic algorithm, allows to determine the best stratification for a population frame, i.e., the one that ensures the minimum sample cost necessary to satisfy precision constraints, in a multivariate and multi-domain case.

The optimization of the sampling design starts by making the sampling frame available, defining the target estimates of the survey and establishing the precision constraints on them. It is then possible to determine the best stratification and the optimal allocation. Finally, we proceed with the selection of the sample.

Formalizing, these are the required steps:

1. analysis of the frame data: identification of available auxiliary information;

2. manipulation of auxiliary information: in case auxiliary variables are of the continuous type, they must be transformed into a categorical form;

3. construction of atomic strata: on the basis of the categorical auxiliary variables available in the sampling frame, a set of strata can be constructed by calculating the Cartesian product of the values of all the auxiliary variables;

4. characterization of each atomic stratum with the information related to the target variables: in order to optimise both strata and allocation of sampling units in strata, we need information on the distributions of the target variables (means and standard deviations);

5. choice of the precision constraints for each target estimate, possibly differentiated by domain;

6. optimization of stratification and determination of required sample size and allocation in order to satisfy precision constraints on target estimates;

7. analysis of the resulting optimized strata;

8. association of new labels to sampling frame units, each of them indicating the new strata resulting by the optimal aggregation of the atomic strata;

9. selection of units from the sampling frame with a stratified random sample selection scheme;
10. evaluation of the found optimal solution in terms of expected precision and bias.

Information

Status:	validated
Author:	Istat
Licence:	GPL-2 | GPL-3
GSBPM code:	2.4 Design frame and sample 4.1 Create frame and select sample
Programming language:	R
Keywords:	optimal stratification, sample design, sample allocation, genetic algorithm
Contact:	name: Marco Ballin email: ballin@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

R (version ≥ 2.15.0)

COPYRIGHT

Copyright 2016 Istat

Licensed under the GNU General Public License (GPL), version 2 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://www.gnu.org/licenses/licenses.en.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 10/03/2020

- SAMPLINGSTRATA Version 1.5-1 – Windows binaries

- SAMPLINGSTRATA Version 1.5-1 – Package source

INSTALLATION

Install the downloaded package from within R as follows:

> install.packages(path_to_file, repos = NULL)

where the character path_to_file is the path to the .zip or .tar.gz file you downloaded.

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Reference manual – SamplingStrata v. 1.5-1

Vignettes:

- Optimization of sampling strata with the SamplingStrata package
- Spatial sampling with SamplingStrata
- Use of models in SamplingStrata

OTHER DOCUMENTATION

SamplingStrata website

Barcaroli, G. 2014. "SamplingStrata: An R Package for the Optimization of Stratified Sampling". *Journal of Statistical Software*, Volume 61, Issue 4: 1-24.

Ballin, M., and G. Barcaroli. 2013. "Joint determination of optimal stratification and sample allocation using genetic algorithm". *Survey Methodology*, Volume 39, N. 2: 369-393.

COLLECT phase

Create frame and select sample

The activities related to the creation of the list and the selection of the sample refer to sub-process 4.1 "*Create frame and select sample*" of GSBPM.

The creation of the frame consists in the construction of the archive of units belonging to the target population. The selection of the sample consists in the identification of the sample units on the basis of the sampling scheme.

For a given iteration of the collection, the creation of the frame and the selection of the sample are made on the basis of the specifications defined in the sub-process 2.4 "*Design frame and sample*".

Data collection

Collection of data comes after a quite complex set of activities aimed at designing the survey questionnaire used to observe the different aspects of the phenomena to be measured.

What above stated is represented in the GSBPM by a set of sub-processes that belong to different phases of the model, from phase 1 to phase 4, as described here after.

Phase 1 "*Specify needs*". Sub-processes 1.1 to 1.5 are involved for this phase. They allow to identify the survey needs and to translate them into concepts that have to be comprehensible to respondents. Besides, concepts should be easily measurable and therefore convertible in statistical variables that will be designed in the following Phase 2. At this stage, it is very important to check whether current data sources (for example administrative data) could meet survey needs in order to reduce the number of variables to collect. In this way both cost and response burden can be reduced.

Phase 2: "*Design*". Once Phase 1 is over, activities described in sub-processes 2.1, 2.2 and 2.3 can be carried on. Survey variables, specified in Phase 1, are used to design the tabulation plan (also useful for the Dissemination phase) that will also allow to define derived variables as well as any statistical classifications that will be used in the collection phase. At this point, the creation of the questionnaire can start: statistical variables can be "translated" into survey questions, relations among variables and variable's characteristics into questionnaire's paths and/or checking rules. Institutional metadata system should be taken into account when designing survey variables, in order to use existing definitions or to update the system with new ones. This will foster national and international standardisation processes and the re-use of any "element" from previous or similar surveys.

Variables design (sub-process 2.2) should be run in parallel with sub-process 2.3 "*Design Collection*" that determines the most appropriate collection method(s) and instrument(s). This is because variables design and, therefore, questions' structure and wording, strictly depend on the mode used to collect data. The same is true for the design of checking rules in case computer assisted techniques are used (example: CATI, CAPI, CAWI, described in Phase 3). With these modes, data can be validated while they dare collected. It is therefore advisable to design checking rules in such a way to solve the greatest number of inconsistencies (major or more frequent inconsistencies), paying attention, at the same time, to the fluency of the interview. In other words, a balance between data quality and respondent burden should be respected: a too

high number of checking rules can increase response burden and negatively affect total or partial non response rate.

Phase 3 "*Build*": The sub-process 3.1 "*Reuse or build collection instruments*" describes the activities to reuse or build the collection instruments to be used during the next Phase 4 "*Collect*". The collection instrument is generated or built based on the design specifications created during the "*Design*" phase. Data collection may use one or more modes to receive the data, e.g., personal or telephone interviews, paper, electronic or web questionnaires. Collection instruments may also be data extraction routines used to gather data from existing data sources. In this last case instruments of data exchange like EDI (*Electronic Data Interchange*) or XBRL (*eXtensible Business Reporting Language*) can be used for information exchange between Istat and reporting units (enterprises or public organizations that provide data in the form of administrative data sets or registers). This sub-process also includes preparing and testing the contents and functioning of that instrument, e.g., testing the questions in a questionnaire. It is recommended to consider the direct connection of collection instruments to the statistical metadata system, so that metadata can be more easily captured in the collection phase. Connection of metadata and data at the point of capture can save work in later phases, like for instance the Dissemination one. Capturing the metrics of data collection (paradata) is also an important consideration in this sub-process.

Computer assisted techniques play an increasingly important role in the data collection phase. The main features of these modes, **CADI** (*Computer Assisted Data Imputing*), **CAPI** (*Computer Assisted Personal Interviewing*), **CATI** (*Computer Assisted Telephone Interviewing*), **CAWI** (*Computer Assisted Web Interviewing*) is represented by the possibility of performing the editing phase during the data collection, allowing the collection of only valid data. **CADI** differs from the others techniques because checking rules are only used to limit keying errors or to support the revision during the data entry of paper questionnaires.

Another distinctive feature of all computer assisted techniques, except **CADI**, is the customisation of the electronic questionnaire: questions wording can be personalised according to respondent's characteristics (name, gender) or to answers to previously asked questions or to already available information (previous wave of the survey). In this way, the questionnaire appears more friendly and the respondents cooperation is enhanced. How many and what type of checking rules can be implemented in the electronic questionnaire depends on the technique. Main differences are between CATI/CAPI and CAWI:

- **CATI** and **CAPI** are interviewer administered and, therefore, it is possible to use a greater number of checks than for CAWI. Besides, relying on the fact that the interviewer is well trained on both technical aspects and survey content, it is also possible to use a greater number of blocking checking rules, that require to solve the inconsistency before proceeding with the interview;

- As to **CAWI**, which is instead self-administered, a smaller number of checks should be used. Besides, checking rules should be more like warnings than blocking checks, because they should advise respondents about possible inconsistencies among data, letting them free to solve or not the "error" before proceeding with new questions. This is to avoid the abandoning of the interview before its very end.

It is important to remind here, that the set of checking rules should be such to guarantee a balance between data quality and response burden, whatever technique is used.

Besides, at this point, it is also useful to remind, that the questionnaire design phase has to take into account the technique chosen for collection of data, in order to design the electronic

questionnaire in the most suitable way. The electronic questionnaire should be deeply tested to check for its compliance with technical specifications and for its usability and fluency.

Phase 4 "*Collect*": Once collection instruments have been built and tested, the collection of data can start. Activities described in sub-processes 4.2 "*Set up collection*", 4.3 "*Run collection*" and 4.4 "*Finalise collection*" are involved: from interviewers training to the storage of collected data into suitable electronic environment for further processing (Phase 5 "*Process*").

COLLECT phase – METHODS

- **CREATE FRAME AND SELECT SAMPLE**

MEMOBUST – Handbook on Methodology of Modern Business Statistics:
Statistical Registers and Frames
Sample Selection
2014
MEETS ESSnet MEMOBUST

- **DATA COLLECTION**

La modernizzazione delle indagini via Web sulle imprese - Pratiche Raccomandate per il disegno dei questionari
2015
Gdl Armonizzazione dei questionari di impresa, Istat

MEMOBUST – Handbook on Methodology of Modern Business Statistics:
Questionnaire Design
Data Collection
2014
MEETS ESSnet MEMOBUST

Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System
2006
European project QDet

References

Murgia, M., and A. Nunnari. 2012. "Improving the quality of data collection: minimum requirements for generalised software independent from the mode". *Seminar on New Frontiers for Statistical Data Collection*. UNECE, Geneva, 31 October - 2 November 2012.

Budano, G. (A cura di). 2008. "*La progettazione e lo sviluppo informatico del sistema Capi sulle forze di lavoro*". *Metodi e Norme*, N. 36. Roma: Istat.

Lo Conte, M., e S. Prati (A cura di). 2006. "*L'indagine campionaria sulle nascite: obiettivi, metodologia e organizzazione*". *Metodi e Norme*, N. 28. Roma: Istat.

COLLECT phase – TOOLS

- **DATA COLLECTION**

Blaise
Computer Assisted Data Collection Software.

Blaise

Description

Blaise is a software tool for computer assisted data collection developed by Statistics Netherlands. It consists of several modules to perform data collections functions:

- the scheduling of telephone contacts for CATI surveys (*CATI Call Management System)*
- the questionnaire administration (*Data Entry*)
- the assisted coding of free text answers (*Coding*)

The calls scheduling is managed by an ad hoc module called *CATI Call ManagementSystem*. Generally speaking, it selects the telephone numbers (sample units) to be called in each survey day according to a call priority that depends on previous contact results and on the values of the scheduling parameters settled by the survey manager.

Scheduling parameters indicate how many times each phone number has to be tried before abandoning it. In particular, they indicate how many times not definitive contact results (busy, no one answer, appointment) have to be tried during the survey period before they get a definitive contact results (interviews, refusal , definitive interruptions).

The *Data Entry* module is an user interface for the questionnaire administration suitable for any kind of computer assisted data collection techniques. It allows to manage the screen layout, any question format and to organise the questionnaire in sections. Besides, it is possible to use interactive editing to manage consistency, coherency and routing checks. Error messages can be customised thus helping the interviewer and/or the respondent in filling in the questionnaire.

The assisted *Coding* module, based on sub-strings algorithms, supports three kinds of coding:

- Hierarchical coding: it looks for the code following the tree structure of the classification, from the highest levels (branches) to the lowest ones (leaves). It can be used only for hierarchical classifications.
- Alphabetical or trigram coding: it looks for the code directly using the input text, making partial matches based on sub-strings of texts called bigram or trigrams or based on the alphabetical order.
- Mixed coding: it looks for code making an alphabetical or trigram search inside the selected classification level.

Information

Status:	validated
Author:	Statistics Netherlands
GSBPM code:	3.1 Build collection instrument
	4.2 Set up collection
	4.3 Run collection
	4.4 Finalise collection
	5.2 Classify and code
Keywords:	CAPI, CATI, CADI, scheduler, assisted coding

Software and documentation

To get the Blaise software as well as the methodological and technical documentation, please contact Statistics Netherlands.

OTHER DOCUMENTATION

Degortes, M., S. Macchia, and M. Murgia. 2003. "The usage of BLAISE in an integrated application for the Births Sample Survey in CATI mode". In *Proceedings of the 8th International BLAISE Users Conference (IBUC 2003)*. International Blaise Users Group, Copenhagen, 21st-23rd May 2003.

PROCESS phase

Data integration

Record linkage

Record linkage is an important process for the integration of data coming from different sources. The purpose of record linkage is to identify the same real world entity that can be differently represented in data sources, even if unique identifiers are not available or are affected by errors.

In official statistics, record linkage is needed for several applications: for instance,

- to enrich the information stored in different data sets;
- to create, update and de-duplicate a frame;
- to improve the data quality of a source;
- to measure a population amount by capture-recapture method;
- to check the confidentiality of public-use microdata.

The complexity of the whole linking process relies on several aspects. The lack of unique identifiers requires sophisticated statistical procedures, the huge amount of data to process involves complex IT solutions, constraints related to a specific application may require the solution of difficult linear programming problems.

Statistical matching

The goal of statistical matching (sometimes named as data fusion) is the integration of two or more data sources referring to the same population with the aim of exploring the relationships between variables that are not jointly observed in the same data source. The sources to be integrated are composed of different non-overlapping units; as usually happens when data from several sample surveys are integrated. The typical situation of statistical matching is the one in which there are two data sources A and B; variables X and Y are available in A, variables X and Z are observed in B; the objective is to study the relationship between Y and Z by exploiting the common information in X. The objective of statistical matching can be macro or micro; in the first case the interest is in one or more parameters that summarize the relationship between Y and Z (correlation coefficient, regression coefficient, contingency table, *etc*.); in the second case the result of integration is a synthetic data set in which all the variables of interest, X, Y and Z are present.

The objectives of matching can be achieved by means of a parametric or nonparametric approach, or a mixture of them (mixed methods).

The parametric approach requires the specification of a model and the estimation of the related parameters. In absence of auxiliary information, it is generally assumed the conditional independence of Y and Z given the common variables X. This assumption is rather strong and unfortunately in the typical situation of the matching it is not testable.

Nonparametric methods are usually applied when you have a micro objective. In this case hot-deck imputation methods are frequently used. They aims at imputing missing variables in the data set chosen as recipient (for instance A) by using the observed values in the data set (B) chosen as

donor. The donor unit for a given unit in A is the most similar observation in B in terms of the values of the common variables X. The mixed approach is composed of two steps: 1) a parametric model is assumed, parameter estimation is performed, and imputation is carried out; 2) a hot-deck imputation procedure is applied, it makes use of imputed values for choosing the donor observation.

It is worthwhile noting that an alternative approach based on the quantification of the uncertainty inherent the estimation of a particular parameter can be used. This approach does not require the conditional independence assumption nor of auxiliary information on non-identifiable parameters, i.e., those related to relations between Y and Z. The study of uncertainty does not lead to a point estimate but to a set of plausible estimates. The set of parameter estimates referring to Y and Z is composed of those consistent with the estimates obtainable from the data at hand, i.e., parameters concerning the couples (Y, X) and (Z, X).

The application of the matching data from complex surveys poses additional problems. In such circumstances it is required to take into account the sampling design as well as the other methodologies used to deal with nonsampling errors (coverage and non-response).

Coding of textual answers

The coding activity is generally performed in case the survey questionnaire contains textual variables that refer to official classifications that allow for national and/or international data comparability. Example of this kind of variables are Economic Activity (NACE), Occupation, Education, Places (of birth, of residence, *etc*.).

Coding means to assign a unique code to a textual answer according to a classification scheme. The level of detail of the matched code depends on the survey aims and/or the dissemination needs. Coding can be performed manually or trough automated systems. Manual coding can be performed only at the end of the data collection phase, while if automated systems are used, it can be run during or after collection of data: in the first case it is called assisted coding (on-line coding) while in second case automated coding (batch coding).

With reference to GSBPM, coding belongs to the sub-process 5.2 "*Classify and code*" of the Phase 5 "*Process*" that includes those activities that are necessary to make data ready for the analysis (Phase 6 "*Analyse*"). Obviously, in case of assisted coding some of the activities of sub-process 5.2 can start before Phase 4 "Collect" ends, improving the timeliness of data delivery.

Coding is, in general, a very hard activity of the survey process. Besides, if it is manually performed it is also difficult to standardise, because coding results strictly depend on coders. Despite coders are well trained about criteria and principles of each official classification, coding is influenced by the cognitive process of each coder that might lead to different (subjective) interpretations and, therefore, different coding of the same textual answer.

The use of specialised coding software can produce a considerable saving of time and resources and will also guarantee a higher standardisation level of the coding process, increasing the expected quality of the coding results.

As already said, computer assisted coding can be distinguished in "automatic coding" and in "assisted coding". They differs in terms of aims and coding process:

- **Automatic coding**: the coding software analyses and codes, on the basis of a reference dictionary, a data file containing all the textual answers collected during the collection

phase (batch coding). The aim is to look for and to assign a single code to each textual answer according to quality thresholds;

- **Assisted coding**: the coding software is an interactive instrument, that aids the coder/respondent in coding the textual answer. The aim is to offer the user a wider set of possible matching codes among which to choose the correct one.

The key point of any coding system, automated or assisted, is the implementation of the informative basis that represents the reference dictionary containing codes and texts of the official classification and enriched with textual answers collected by Istat surveys (and correctly coded). In order to be processed by a software, the reference dictionary has to undergo a number a standardisation operations aimed at producing analytic, synthetic and not ambiguous descriptions. Besides, in general, the richest the dictionary the higher the coding rate.

Generally speaking, coding systems varying according to the algorithm used to match the textual answers with the dictionary descriptions. They can be classified as follows:

- dictionary algorithms: they look for exact matches on the bases of key words (or groups of key words);

- weighting algorithms: they look for partial or exact matches on the basis of similarity functions among texts that assign weights to each word according to its informative content;

- sub-strings algorithms: they look for partial or exact matches processing portions of texts (bigrams or trigrams).

Besides, for what concern assisted coding, there are three possible methods to consult (to navigate) the reference dictionary:

- tree search: it navigates inside the classification hierarchical structure, from the higher branch to the lowest one (leave) that represents the most detailed code (highest number of digits) that can be assigned to a textual answer;

- alphabetic search: it navigates inside the entire dictionary looking for the definition which is equal or the most similar to the textual answer to be coded;

- mixed mode search: it makes an alphabetic search inside the selected classification branch.

Data collection technique highly influences the choice of the searching method. A special distinction is among interviewer administered and self-administered modes. For the latter, where respondents are not trained on classifications and coding like interviewers are, it is extremely important to provide a coding system that is user friendly and guarantees high quality results.

The quality of coding activity is highly influenced by the update of both the dictionary content and the matching rules (training phase). It is advisable to perform the training phase periodically, in general after the coding of textual answers collected by a survey. To this aim, after a coding application, it is important to:

- verify the quality of the coded cases;

- use the not coded cases to update the coding application (dictionary and checking rules);

- highlight eventual lacks of the classification used.

Per la valutazione della qualità delle due modalità di codifica, è possibile utilizzare i seguenti indicatori:

Indicators for assisted and automated coding can be used to evaluate the performance of the coding phase:

Automated coding indicators:

- efficacy/coding rate: ratio of "number of coded texts" to "total number of texts to be coded";
- accuracy: ratio of "number of correctly coded texts" to "number of coded texts";
- efficiency: unitary coding time.

Assisted coding indicators:

- average time to assign a single code;
- coherence among each collected textual description and the assigned code.

Detection and treatment of measurement errors and imputation of partial non-responses

Partial non-responses (PNR) and measurement errors are specific types of non-sampling errors which are generally treated in the editing and imputation phase. In the present context, a measurement error is defined as a discrepancy between "true value" and observed value of a variable in a statistical unit. Discrepancy could be originated in any phase of the measurement process (data collection, coding, storing, *etc*.). Since NRPs and measurement errors may seriously compromise the accuracy of the target estimates, they should be prevented by means of suitable strategies. However, even though efforts are made to limit the impact of non-sampling errors, a proportion of collected data are typically affected by partial non-responses and measurement errors. Thus, use of editing and imputation methods are necessary.

In the context of GSBPM, the two sub-processes 5.3 "Review and validate" e 5.4 "Edit and impute" concern input data validation and editing and imputation respectively. However, often in the real contexts, these two sub-processes are hardly distinguishable.

Detection of measurement errors

Methods for error detection can be classified according to the different error typologies. A first important distinction is between **systematic errors** and **random errors**.

An error is called **systematic** if it depends on structural problems in the statistical production process such as defects in the questionnaire design or in the storing system. They usually determine deviation from the true values "in the same direction" for one or more variables of interest. Systematic errors are generally treated by means of deterministic rules base on the knowledge of the error mechanism. A typical systematic error for quantitative variables is the unity measure error.

An error is called **random** if it depends on stochastic elements that are not identifiable. In contrast with the case of systematic errors, a deterministic approach is generally not appropriate for random errors.

An important typology of errors involves errors causing "**out of range**" values, that is, values that do not belong to a known set of acceptable values. A similar typology involves errors determining **inconsistencies** between different observed variables. These errors are detected by applying **coherence rules (edit)** on data. Consistency errors that are believed to be "not influent" are

generally detected via automatic methods based on some "general" principles. In this context, a popular approach is the **minimum change principle**. According to this principle, erroneous values are detected in such a way that for each record the minimum number of items is to be changed in order for the record to pass all the edits. The **Fellegi-Holt** methodology implements the minim change principle. It has been originally developed for categorical variables and later extended to numerical variables.

Another important class of errors includes errors which determine **outliers** in data. Outliers are observations whose behavior deviates from the one typical for most observations. Usually, technics for outlier identifications assume, at least implicitly, a model for the "typical" data and try to identify deviations from the model. Methods for outlier detection are often used to identify **influential errors**. However, the concept of influential error is to be distinguished from the one of outlier. In fact the latter is related to a model assumed for the data, while the former depends on the population parameter to be estimated. Specifically, outlier may not be caused by influential errors, and influential errors may not cause outliers.

Correction of errors and imputation of partial non-responses

Once errors have been detected, incorrect values have to be replaced (imputed) by correct or nearly correct values. Imputation is also used for partial non-responses. Imputation is commonly used for a number of technical and practical reasons. First, released data must be complete and coherent at micro level. Furthermore, imputation allows the users using standard methods and software on the final dataset.

A lot of imputation methods are available in many statistical packages both proprietary and free. A frequent classification of imputation methods is in terms of **parametric methods** (e.g., **regression** type methods), relying on some explicit assumptions on the probability distribution generating the data, and **non-parametric methods**, generally based on weaker not explicit assumptions (e.g., **hot-deck, nearest neighbor donor**).

Sometimes imputation methods are called **deterministic** when different applications on the same dataset provide same outputs, and stochastic, if the outcomes are characterized by a certain level of variability.

The choice of the imputation method depends on the analyses that have to be performed on data. For instance, If the quantity of interest is a linear quantity such as a mean or a total, a deterministic method may be appropriate, while, if also are distributional characteristics (such as the ones involving the second moment of the data distribution), a stochastic imputation method is generally preferable.

Weighting, estimation and sampling error evaluation

The activities concerning the production of target estimates and the evaluation of sampling errors refer to sub-process 5.6 "*Calculate weights*" and 5.7 "*Calculate aggregates*" of GSBPM.

Production of the estimates of interest

Each estimation method is based on the principle that the subset of the units of the population included in the sample must also represent the complementary subset consisting of the remaining

units of the population. This principle is generally achieved by assigning to each unit included in the sample a weight that can be seen as the number of population elements represented by that unit.

The sample surveys carried out by Istat are large-scale surveys that have the purpose of providing a large number of estimates of population parameters such as counts, totals, proportions, averages, etc.

The estimation of the parameters of the population can be made using two different approaches:

- **Methods based on the direct approach** using values of the variable of interest observed on the sample units belonging to the domain of interest. They are the standard methods used by Istat and by all the main National Institutes of Statistics to produce survey estimates.

- **Methods based on the indirect approach** that make use values of the variable of interest observed on sample units belonging to a wider domain containing the domain of interest and/or other survey occasions. They are, usually, used for particular estimation problems, such as those associated with the generation of estimates referring to domains in which the sample size is too small for the production of estimates using direct methods.

Direct methods

In general, for the estimation a total the following two operations should be performed:

1. computation of the weight to be assigned to each unit included in the sample;

2. calculation of the estimates as weighted sums of the values of the target variables using with weights determined in step 1.

The weight given to each unit is obtained according to a procedure divided in several steps:

1. the *starting weight* of each sample unit, named *direct weight*, is calculated according to the sampling design, as the reciprocal of the inclusion probability;

2. the starting weight is adjusted in order to account for non-response, obtaining the *base weight*;

3. correction factors of the base weight basis are computed to take into account equality constraints between some known parameters of the population and the corresponding sample estimates;

4. the *final weight* is obtained as the product between the base weight and the correction factors.

The class of estimators corresponding to the operations described above is known as *calibration estimators*, since both the adjustment to correct for non-response and the weight correction to achieve consistency with known population parameters is obtained by solving a constrained minimization problem. In details we want to minimize the distance between the weight before and after the calibration phase.

The main problem for the choice of the estimation method is to find an estimator satisfying:

- *efficiency* criteria in terms of sample variance and bias due to the presence of *total and partial non-responses*, frame under-coverage;

- external and internal *coherence*. The *external consistency* of the estimates arises whenever known totals are available from external sources. Estimates of the total produced by the survey should generally match or not deviate too much from the known values of these totals. The *internal consistency* of the estimates is achieved when all the estimates of the same aggregate coincide with each other. This result can be obtained using a unique system of weights.

Calibration estimators meet the above criteria since:

- they yield, generally, more efficient estimates than those obtained using direct estimators; the higher the correlation between the auxiliary variables and the target variables the greater the efficiency;

- they are approximately design unbiased;

- they produce estimates of totals that coincide with the known values of these totals;

- they mitigate the non-response bias effect;

- they reduce the bias due to the under-coverage of the frame from which the sample is selected.

Calibration estimators are used for calculating the final weights for most social and business surveys carried out by Istat.

Indirect methods

Indirect estimation methods are used by Istat to give clear responses to the growing need of local governments for accurate information for small geographical areas, or more generally domains, called small areas. Sample surveys conducted by Istat are, however, designed to provide reliable information for the main aggregates of interest for planned domains defined at design stage and may not be able to respond appropriately to the production of estimates for larger level of detail.

The solution adopted in the past by Istat to obtain estimates at unplanned domain level, was to increase the sample size without changing the sampling strategy, i.e. without modifying neither the sampling design nor the estimator. However over-sampling, besides rising collection costs and increasing the difficulty of organizational issues, implies the increase of non-sampling errors due to the difficulty to handle with too large sample sizes. In addition, the over-sampling is a partial solution to the problem of small area estimation, since not being able to increase the sample size over a certain threshold makes it possible to provide reliable estimates only for a subset of the small areas of interest.

For these reasons, Istat makes use of estimation methods based on:

- the use of auxiliary information, related to the phenomena under study, known at small areas level;

- the adoption (implicit or explicit) of statistical models linking the variable of interest observed in the small area with the values of the same variable related to a larger area containing the small area of interest and/or related to other survey occasions.

An important problem related with these methods is that they are based on models and, therefore, the properties of the results depend on the validity of the model assumptions. Since models are never expected to perfectly match reality, these estimators introduce unmeasurable bias which may arise serious concerns about their use.

Assessment of sampling errors

For the evaluation of the sampling errors of the estimates, Istat usually uses approximated variance estimation methods. In fact, for most of the estimation procedure an analytical expression of the estimator variance is not available, since:

- ISTAT surveys are carried out using complex sampling designs, generally based on multiple selection stages, on the stratification of the units, and on without repetition selection scheme with varying selection probabilities among units;

- estimates are determined through the use of calibration estimators which are non-linear functions of the sample information.

The estimation methods of the sample variance generally used in Istat are based on the method of linearization of Woodruff (1971) which provides estimates of the sample variance in the case where the estimators used are non-linear functions of the sample data.

This variance estimation methods are implemented by Istat in the generalized software GENESEES and ReGenesees, which feature a user friendly interface and are currently used to estimate the sampling errors of the estimates produced by Istat surveys.

In addition, by means of these software, it is possible to compute important analysis statistics which provide useful tools to evaluate the adopted sampling design. In particular it is possible to evaluate:

- the overall efficiency of the sampling design;

- the impact on the efficiency of the estimates due to stratification, the number and type of selection stages and weighting effect.

It is important to point out that a generalized variance errors representation is also provided. It is a summary obtained making use of regression models that relate the estimates with the corresponding sampling errors. These models provide important summary information on sampling errors and are disseminated together with the tables reporting the estimate values.

PROCESS phase – METHODS

- **DATA INTEGRATION**

MEMOBUST – Handbook on Methodology of Modern Business Statistics
Micro-Fusion
2014
MEETS ESSnet MEMOBUST

State of the art on statistical methodologies for data integration
Methodological developments
2011
ESSnet on Data Integration

Old and new approaches in statistical matching when samples are drawn with complex survey designs
2010
in Proceedings of the SIS Conference, Padua

ISAD Work packages and executive summary
2008
ESSnet Statistical Methodology – Area ISAD (Integration of Survey and Administrative Data)

Metodi statistici per il record linkage
2005
Metodi e Norme, N. 16, Istat

References

D'Orazio, M., M. Di Zio, and M. Scanu. 2006. "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints". *Journal of Official Statistics - JOS*, Volume 22, N. 1: 137-157.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006. *Statistical Matching: Theory and Practice*. Chichester: J. Wiley & Sons.

CODING OF TEXTUAL ANSWERS

Metodi e software per la codifica automatica e assistita dei dati
2007
Tecniche e strumenti, N.4, Istat

References

Vicari, P. (A cura di). 2008. "L'ambiente di codifica automatica dell'ATECO 2007. Esperienze effettuate e prospettive". *Metodi e Norme*, N. 41. Roma: Istat.

Macchia, S., M. Murgia, L. Mazza, G. Simeoni, F. Di Patrizio, V. Parisi, R. Petrillo, e P. Ungaro. 2005. "Una soluzione per la rilevazione e codifica della Professione nelle indagini CATI". *Contributi*, N. 11. Roma: Istat.

Macchia, S., and M. D'Orazio. 2001. "A system to monitor the quality of automated coding of textual answers to open questions". *Research in Official Statistics*, Volume 4, N. 2: 5-19.

DETECTION AND TREATMENT OF MEASUREMENT ERRORS AND IMPUTATIONOF PARTIAL NON-RESPONSES

MEMOBUST – Handbook on Methodology of Modern Business Statistics
Statistical Data Editing
Imputation
2014
MEETS ESSnet MEMOBUST

A Contamination Model for Selective Editing
2013
Journal of Official Statistics. Volume 29, Issue 4, Pages 539-555.

Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys
2007
European project EDIMBUS

Results of the EUREDIT project
Euredit – The Development and Evaluation of New Methods for Editing and Imputation
2003
European project EUREDIT

References

De Wall, T., J. Pannekoek, and S. Sholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Hoboken, N.J.: J. Wiley & Sons.

Di Zio, M., U. Guarnera, and R. Rocci. 2007. "A mixture of mixture models for a classification problem: the unity measure error". *Computational Statistics and Data Analysis*, N. 51: 2573-2585.

Di Zio, M., U. Guarnera, and O. Luzi. 2005. "Editing Systematic Unity Measure Errors Through Mixture Modelling". *Survey Methodology*, Volume 31, N. 1: 53-63.

Pannekoek, J., and T. De Waal. 2005. "Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project". *Journal of Official Statistics - JOS*, Volume 21, N. 2: 257-286.

Särndal, C. E., and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. New York: J. Wiley & Sons.

Chambers, R., A. Hentges, and X. Zhao. 2004. "Robust automatic methods for outlier and error detection". *Journal of the Royal Statistical Society*, Series A (Statistics in society), Volume 167, Issue 2: 323-339.

Little, J., and D. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: J. Wiley & Sons.

Chen, J., and J. Shao. 2000. "Nearest Neighbor Imputation for Survey Data". *Journal of Official Statistics*, Volume 16, N. 2: 113-131.

Schafer, J. L. 2000. *Analysis of Incomplete Multivariate Data*. New York: Chapmann and Hall/CRC.

Latouche, M., and J.M. Berthelot. 1992. "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys". *Journal of Official Statistics*, Volume 8, N. 3: 389-400.

Little, R.J.A. 1988. "Missing-data adjustments in large surveys". *Journal of Business & Economic Statistics*, Volume 6, N. 3: 287-296.

Rubin, D. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley & Sons.

Hidiroglou, M.A., and J.M. Berthelot. 1986. "Statistical editing and imputation for periodic business surveys." *Survey Methodology*, Volume 12, N. 1: 73-83.

Kalton, G., and D. Kasprzyk. 1982. "Imputing for missing survey responses". In *Proceedings of the section on Survey Research Methods*, American Statistical Association.

Fellegi, P.I., and D. Holt. 1976. "A systematic approach to automatic edit and imputation". *Journal of the American Statistical Association*, Volume 71, Issue 353, Application Section: 17-35.

- **WEIGHTING, ESTIMATION AND SAMPLING ERROR EVALUATION**

MEMOBUST – Handbook on Methodology of Modern Business Statistics
Weighting and Estimation
2014
MEETS ESSnet MEMOBUST

Results of the SAE project
SAE – Small Area Estimation
2012
ESSnet SAE

Riponderazione
2005
Note metodologiche, Istat

Stime ed Errori
2005
Note metodologiche, Istat

References

Falorsi, S., C. De Vitiis, L. Di Consiglio, C. Ceccarelli, A. Cutillo, e C. Rinaldelli. 2008. "Strategia di campionamento e precisione delle stime" In "*L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc)*". *Metodi e Norme*, N. 37. Roma: Istat.

De Vitiis, C., e S. Falorsi. 2006. "La procedura di stima e la valutazione degli errori campionari". In "*Il sistema di indagini sociali multiscopo. Contenuti e metodologia delle indagini*". *Metodi e Norme*, N. 31. Roma: Istat.

De Vitiis, C., e A. Pareto. 2006. "Strategia di campionamento e livello di precisione delle stime". In "*L'indagine campionaria sulle nascite: obiettivi, metodologia e organizzazione*", *Metodi e Norme*, N. 28. Roma: Istat.

Rao, J.N.K. 2003. *Small Area Estimation*. New York: J. Wiley & Sons.

Cicchitelli G., A. Herzel, e G.E. Montanari. 1992. *Il campionamento statistico.* Bologna: Il Mulino.

Deville, J.C., and C.E. Särndal. 1992. "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, Volume 87, Issue 418, Theory and Method: 376-382.

Särndal, C.E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Cochran, W.G. 1977. *Sampling Techniques.* New York: J. Wiley & Sons.

PROCESS phase – TOOLS

- **DATA INTEGRATION**

RELAIS (REcord Linkage At IStat)

Toolkit for dealing with record linkage projects.

StatMatch

R package for data integration through statistical matching and, as a by-product, the ability to impute missing values in a data set.

- **CODING OF TEXTUAL ANSWERS**

CIRCE (Comprehensive Istat R Coding Environment)

Automated coding system for textual answers.

- **DETECTION AND TREATMENT OF MEASUREMENT ERRORS AND IMPUTATION OFPARTIAL NON-RESPONSES**

Banff
Edit and imputation system for numeric and continuous variables.

CANCEIS (CANadian Census Edit and Imputation System)

Edit and imputation system based on the Nearest-neighbour Imputation Methodology (NIM). The NIM allows the simultaneous hot-deck imputation of numeric and categorical variables based on a single donor. The basic entity that is processed (the unit) can be composed of one or more sub-units of lower hierarchical level.

CONCORDJava (CONtrollo e CORrezione dei Dati version with Java interface)
Integrated system to check and correct the data (imputation). One of the modules (SCIA) implements the Fellegi-Holt Methodology to treat the inconsistencies between the values of qualitative variables.

SeleMix (Selective editing via Mixture models)

R package to treat quantitative data, which aims to identify a set of units affected by errors which potentially influence the interest estimates (selective editing).

- **WEIGHTING, ESTIMATION AND SAMPLING ERROR EVALUATION**

EVER (Estimation of Variance by Efficient Replication)

R package for calibration, estimation and sampling error assessment in complex sample surveys, based on replication methods .

ReGenesees (R evolved Generalised software for sampling estimates and errors in surveys)
R-based software system for design-based and model-assisted analysis of complex sample surveys.

RELAIS (REcord Linkage At IStat)

Description

RELAIS (REcord Linkage At IStat) is a toolkit providing a set of techniques for dealing with record linkage projects.

The purpose of record linkage is to identify the same real world entity that can be differently represented in data sources, even if unique identifiers are not available or are affected by errors. In statistics, record linkage is needed for several applications, including: enriching the information stored in different data-sets; de-duplicating data-sets; improving the data quality of a source; measuring a population amount by capture-recapture method; checking the confidentiality of public-use micro data. In fact, record linkage can be seen as a complex process consisting of several phases involving different knowledge areas; moreover, several different techniques can be adopted for each phase. We believe that the choice of the most appropriate technique not only depends on the practitioner's skill but, most of all, it is application specific.

Moreover, in some applications, there is no evidence to prefer a given method to others or of the fact that different choices, at some linkage stage, could bring to the same results. This is why it is reasonable to dynamically select the most appropriate technique for each phase and to combine the selected techniques for building a record linkage work-flow of a given application. RELAIS is a toolkit relying on these ideas.

The principal features of RELAIS are:

- It is designed and developed to allow the combination of different techniques for each of the record linkage phases, so that the resulting work-flow is actually built on the basis of application and data specific requirements.

- It has been developed as an open source project, so several solutions already available for record linkage in the scientific community can be easily re-used. It is released under the EUPL license (European Union Public License).

- It has been implemented by using two languages based on different paradigms: Java, an object-oriented language, and R, a functional language. This choice depends on our belief that a record linkage process is composed of techniques for manipulating data, for which Java is more appropriate, and of calculation-oriented techniques for which R is a preferable choice. The choice of Java and R is also in line with the open source philosophy of the RELAIS project.

- It has been implemented using a relational database architecture, in particular it is based on a MySQL environment that is also in line with the open source philosophy of the RELAIS project.

The RELAIS project aims to provide record linkage techniques easily accessible to not-expert users. Indeed, the developed system has a GUI (Graphical User Interface) that on the one hand permits to build record linkage work-flows with a good flexibility. On the other hand it checks the execution order among the different provided techniques whereas precedence rules must be controlled.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	5.1 Integrate data
Programming language:	R, Java
Language of the GUI:	EN
Keywords:	data integration, probabilistic record linkage, string comparators, blocking/sorting/indexing, deduplication, open source software
Contact:	name: Luca Valentino email: luvalent@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

Java SE Development Kit (version ≥ 13)

R (version ≥ 3.4.0)

R packages: ROI, ROI.plugin.clp, slam, RODBC

MySQL Server (version ≥ 5.0)

MySQL Connector/ODBC (version ≥ 5.0)

COPYRIGHT

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 10/03/2020

RELAIS version 3.1

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

User manual – RELAIS v. 3.1

OTHER DOCUMENTATION

Cibella, N., M. Scannapieco, L. Tosco, T. Tuoto, and L. Valentino. 2012. "Record Linkage with RELAIS: Experiences and Challenges". *Estadistica Española*, Volume 54, N. 179/2012: 311-328

Cibella, N., G.L. Fernandez, M. Guigò, F. Hernandez, M. Scannapieco, L. Tosco, and T. Tuoto. 2009. "Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences". In *Proceedings of the New Techniques and Technologies for Statistics (NTTS) Conference 2009.* Eurostat, Brussels, 18-20 February 2009.

Cibella, N., M. Fortini, M. Scannapieco, L. Tosco, and T. Tuoto. 2009. "Theory and practice of developing a record linkage software". In *ESSnet-ISAD Workshop, Insights on Data Integration Methodologies*. Eurostat, Vienna, 29-30 May 2008.

Cibella, N., M. Fortini, M. Scannapieco, L. Tosco, and T. Tuoto. 2007. "RELAIS: Don't Get Lost in a Record Linkage Project". In *Research Conference of the Federal Committee on Statistical Methodology.* FCSM, Arlington, 5-7 November 2007.

Fortini, M., P.D. Falorsi, C. Vaccari, N. Cibella, T. Tuoto, M. Scannapieco, and L. Tosco. 2006. "Towards an Open Source Toolkit for Building Record Linkage Workflows". In *IQIS International workshop. IQIS,* Chicago, 30 June 2006.

StatMatch

Description

StatMatch is an add-on package for the R environment to carry out data integration through statistical matching and, as a by-product, to impute missing values in a data set.

The package contains functions that implement methods of matching and functions to support matching (calculation of the distances, etc.). There are three functions dedicated to the application of nonparametric methods of matching at the micro level:

- NND.hotdeck: mimimum distance donor selection; implements many distance functions; allows the definition of donation classes. It is possible to introduce a constraint for avoiding the selection of a donor more than once (constrained matching).

- RNDwNND.hotdeck: random selection of the donor in fixed or "moving" classes. In the latter case, it is possible to select at random a donor among the k closest ones; random choice of a donor among those with a distance less than a given threshold, etc. The selection of the donor can be done with probability proportional to a weighing variable. Different distance functions are implemented.

- rankNND.hotdeck: select the donor closer based on the distance between the percentiles of the empirical cumulative distribution of a continuous variable available in both the data set. In estimating the empirical cumulative distribution it is possible to account for different weights assigned to units. The empirical cumulative distribution may be estimated in given classes of units.

These functions can be used to impute the missing values in a data set through the corresponding hot-deck methods.

There is only one function mixed.mtc that allows to implement methods of matching in parametric macro or mixed framework. The function assumes that the variables X, Y and Z are distributed according to a multivariate normal distribution. The estimation of the parameters can be done via the maximum likelihood method or by applying a method based on sample estimates of the parameters of interest (means and variances).

Two functions implement methods for matching at the macro level in the presence of data from complex surveys. These functions are based on a series of calibrations of the weights associated to the units in the source data sets according to the methods suggested by Renssen (1998). At this time the application of these methods is limited to the case where Y and Z are both categorical and the goal consists in estimating the contingency table Y vs. Z. In particular, the function harmonize.x harmonizes the marginal/joint distribution of the chosen X variables so to make it coherent between the two initial data sets; then the function comb.samples estimates the table Y vs. Z through the methodologies proposed by Renssen, with or without an additional data source C where Y and Z or X, Y and Z are jointly observed.

Finally the functions Frechet.bounds.cat and Fbwidths.by.x allow the exploration of uncertainty when all the variables (X, Y, and Z) are categorical. The first function estimates the ranges of uncertainty bounds for all the cells in the contingency table Y vs. Z. In the presence of many common variables X, the function Fbwidths.by.x permits to identify the variables most related to Y and Z, and that allow a reduction of the width of the uncertainty bounds.

Information

Status:	validated
Author:	Istat
Licence:	GPL-2	GPL-3
GSBPM code:	5.1 Integrate data 5.4 Edit and impute
Programming language:	R
Keywords:	statistical matching, data fusion, hot deck imputation, uncertainty analysis
Contact:	name: Marcello D'Orazio email: madorazi@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

R (≥ 2.7.0)

R packages: proxy, lpSolve, survey, ggplot2..

In some cases, it may be necessary to have the additional R packages: Hmisc, MASS, mipfp, clue, RANN.

COPYRIGHT

Copyright 2016 Marcello D'Orazio

Licensed under the GNU General Public License (GPL), version 2 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://www.gnu.org/licenses/licenses.en.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 01/03/2022

- STATMATCH Version 1.4.1 – Windows binaries
- STATMATCH Version 1.4.1 – Package source

INSTALLATION

Install the downloaded package from within R as follows:

> install.packages(path_to_file, repos = NULL)

where the character path_to_file is the path to the .zip or .tar.gz file you downloaded.

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Reference manual – StatMatch v. 1.4.1

OTHER DOCUMENTATION

StatMatch website

CIRCE (Comprehensive Istat R Coding Environment)

Description

CIRCE is a software package based on R aimed at automatically coding textual variables according to official classifications. It is a generalised software system with respect to the language and the classification used. CIRCE replaces Actr v3, that has been adopted by Istat since 1998, because it was no more commercialised and maintained from its producers and because it was not compatible with the software platform used in Istat (Windows 7, Windows Server 2008).

To prevent lower quality results, the same matching algorithms of ACTR v3 has been developed for CIRCE.

Being an R package it is portable to different platform with no need of compilation. This made it possible to have one single package running on both Windows and Linux operating systems. It can be used on Windows environment through an User Graphical Interface and on web through a "call" to a web service.

CIRCE belongs to the weighting algorithms category and manages three types of coding procedures:

- automated coding, for set of records (batch coding);
- interactive coding, to analyse coding results of single record (a GUI is provided to coders);
- web coding, a web service for single record coding. In this case is currently available a web service dedicated to the identification of the activity code (in Italian language) accessible through the page: https://www.istat.it/it/archivio/17888.

Notwithstanding the type of procedure, the coding phase is performed in two consecutive steps:

1) standardization of texts, called parsing;

2) matching of parsed texts.

The parsing step is a quite sophisticated phase of text standardisation totally customisable, that provides (till now) 14 different functions such as characters mapping, deletion of trivial words, definition of synonymous, suffixes removal, etc.. The parsing aims at removing grammatical or syntactical differences in order to make equal two different descriptions but with the same semantic content.

The second step is the matching phase. The parsed response is compared with the parsed descriptions of the informative base. If this search returns a perfect match or direct match, then a unique code is assigned, otherwise the software uses an algorithm to find the best partial matches, providing an indirect match.

CIRCE is developed by Istat. This will make it easier adding or changing its functionalities with respect to standardization parings and/or matching steps.

Please note: for the moment, both CIRCE user guide (Manuale Utente.pdf) and its GUI are in Italian. English versions will eventually be provided in the future.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	5.2. Classify and code
Programming language:	R, VB.NET
Language of the GUI:	IT
Keywords:	automated coding, weighting coding algorithms
Contact:	name: Laura Capparucci email: capparuc@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

- R (version ≥ 3.1.1).

- Windows (version ≥7).

- Microsoft Framework .net 4 (only for the graphical user interface).

COPYRIGHT

Copyright 2016 Istat

Licensed under the European Union Public Licence (EUPL), version 1.0 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://ec.europa.eu/idabc/eupl.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 28/07/2016

CIRCE version 1.0

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

User manual – CIRCE v. 1.0

OTHER DOCUMENTATION

Macchia, S., P. Giovani, D. Perrone, M. Degortes, e L. Mazza. 2007. "Metodi e software per la codifica automatica dei dati". *Tecniche e strumenti*, N. 4. Roma: Istat.

Cuccia, F., S. De Angelis, A. Laureti Palma, S. Macchia, S. Mastroluca, e D. Perrone. 2005. "La codifica delle variabili testuali nel 14° Censimento Generale della Popolazione". *Documenti*, N. 1. Roma: Istat.

Macchia, S., and M. D'Orazio. 2001. "A system to monitor the quality of automated coding of textual answers to open questions". *Research in Official Statistics*, Volume 4, N.2: 5-19.

De Angelis, R., S. Macchia, e L. Mazza. 2000. "Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale". *Quaderni di ricerca – Rivista di statistica Ufficiale*, N. 1.

Banff

Description

Banff is a generalized system for editing and imputing survey data based on the SAS architecture. Banff was developed by Statistics Canada to process numeric and continuous variables. It uses consistency rules (edit) that must be expressed in linear form.

Banff has a modular structure: each module corresponds to a particular sub-function of the general structure of an edit and imputation process of quantitative variables:

- edit specification;
- check edits for consistency and redundancy;
- error localization;
- detection of outlier values;
- imputation.

The error localization module uses the *Chernikova algorithm* based on the *minimum change principle* or Fellegi-Holt paradigm. For each record that fails at least one edit, the algorithm identifies the minimum number of fields to change (impute) so that the record passes all the rules. In general, the Fellegi-Holt paradigm is considered appropriate to treat stochastic errors.

Banff implements several imputation methods:

- Deterministic imputation

  It checks if there is one and only one value that, once assigned to the field to impute, allow the record to pass all the edits.

- Donor imputation

  The nearest neighbour record (according to a specific distance function) to the current failed record is chosen among the potential donors, i.e. units that pass all the edits. All required fields are imputed by transferring the corresponding values from the nearest neighbour record.

  It is important to note that a potential donor will be actually chosen as the donor, if the imputed values are such that the imputed record pass the user-specified post-imputation edits.

- Estimator Imputation

  Values to be imputed are obtained through modeling or observed data. Examples are mean imputation and regression imputation.

Banff also provides outputs that allow the user to analyze the impact of the editing process on the data (for example, the list of redundant rules or the failure frequency of the erroneous record).

Information

Status:	decommissioned
Author:	Statistics Canada
GSBPM code:	5.3 Review and validate 5.4 Edit and impute
Keywords:	editing for numerical variables, error localization, minimum change principle, nearest neighbour donor

Software and documentation

To get the Banff software as well as the methodological and technical documentation, please contact Statistics Canada.

Only for Istat staff: contact Francesco Dell'Orco.

CANCEIS (CANadian Census Edit and Imputation System)

Description

CANCEIS is an edit and imputation system developed by Statistics Canada. It was designed based upon the Nearest-neighbour Imputation Methodology (NIM) and was first used to perform edit and imputation on census data. CANCEIS performs minimum change nearest-neighbour imputation and deterministic imputation. The NIM allows the simultaneous hot-deck imputation of numeric and categorical variables based on a single donor. The NIM identifies donors for the entire household, not only for the individuals. For each failed household, the NIM identifies a set of potential donors (nearest neighbours) which are as similar as possible to the failed household to be imputed. For each nearest neighbour, the smallest subsets of variables which, if imputed, allow the imputed record to pass the edits, are identified. One of those imputation actions is randomly selected by giving a better chance to those imputation actions that are simultaneously closer to the failed household and the potential donor.

Information

Status:	validated
Author:	Statistics Canada
GSBPM code:	5.3 Review and validate 5.4 Edit and impute
Keywords:	editing, nearest-neighbour donor imputation, numeric and categorical variables

Software and documentation

To get the CANCEIS software as well as the methodological and technical documentation, please contact Statistics Canada.

OTHER DOCUMENTATION

Bruzzone, S., A. Manzari, M. Pappagallo, e A. 2007. "Indagine sulle Cause di Morte: Nuova procedura automatica per il controllo e la correzione delle variabili demo-sociali". *Documenti*, N. 6. Roma: Istat.

Manzari, A., and A. Reale. 2001. "Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology". In *ISI World Statistics Congress Proceedings 53rd Session.* International Statistical Institute, Seoul, 2001.

CONCORDJava (CONtrollo e CORrezione dei Dati version with Java interface)

Description

CONCORDJava is an *open source* software for editing and imputing data. The application integrates software previously developed and used in Istat: SCIA, RIDA and GRANADA.

The application, currently released in beta edition for the part about the deterministic corrections, is available for download in Italian and English.

The different methods in the software are implemented in separate modules:

- **SCIA** (Sistema di Controllo e Imputazione Automatici):

  Editing and imputation of qualitative variables by applying the Fellegi-Holt methodology. First the system identifies the minimum number of fields to impute for each failed record and then it imputes ensuring the imputed record will pass all edits.

- **RIDA** (Ricostruzione dell'Informazione con Donazione Automatica):

  It performs imputation of both categorical and continuous variables through donor of minimum distance. Preliminary operations are:

  - the classification of units in passed and failed;
  - their recording into two separate files;
  - the identification of values to be imputed using a preset character (error character).

- **GRANADA** (Gestione delle Regole per l'ANAlisi dei DAti):

  It imputes qualitative and quantitative variables according to the deterministic approach, in other words applying the rules of the type IF [error condition] THEN [corrective action]. Using this module you can also check data according to edits which admit logical and arithmetic operators (and therefore valid for qualitative and quantitative variables).

Preparatory to the various steps is the phase of definition of the variables, i.e. of the fields of the record to be checked, and the edits.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	5.3 Review and validate 5.4 Edit and impute
Programming language:	Fortran, Java
Language of the GUI:	EN, IT
Keywords:	localization, imputation, nearest neighbour donor, Fellegi-Holt
Contact:	name: Maria Teresa Buglielli email: bugliell@istat.it

Software and documentation

Minimum size hardware

256 Mb RAM

30 Mb on C:/

SOFTWARE DEPENDENCIES

Java 2 Runtime Environment (version ≥ 6.0)

COPYRIGHT

Copyright 2014 Istat

Licensed under the European Union Public Licence (EUPL), version 1.1 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://ec.europa.eu/idabc/eupl.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 10/06/2014

CONCORDJAVA version 2.2

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

User manual – CONCORDJava v. 2.2

SeleMix (Selective editing via Mixture models)

Description

Selemix is an R package to treat quantitative data, which aims to identify a set of units affected by errors which potentially influence the estimates of interest (selective editing).

The underlying methodology is based on particular latent class models known in the literature as contamination models. Specifically, it is assumed that "true" data (that is not affected by errors), possibly in logarithmic scale, are independent realizations of a multivariate Gaussian distribution with the mean vector which can in turn be expressed as a linear combination of a set of covariates not contaminated. The "intermittent" nature of the error mechanism is captured by Bernoulli variables that have the role of indicators for the occurrence of error on each unit. Moreover, the error is assumed additive and associated with a Gaussian vector with zero mean and variance covariance matrix proportional to the variance covariance matrix that characterizes the distribution of data without errors. The explicit modeling of the not contaminated data distribution and the error mechanism allow to obtain the distribution of actual data conditionally on the observed data. On the basis of this distribution estimates of the true values, and therefore of errors, are obtained. For each unit, a score is calculated in terms of the difference (possibly weighted with sample weight) between the predicted value and the observed value. So all units are sorted (in descending order) according to their score. Assuming that the parameter of interest is an average or total population, selection of observation to be reviewed as interactive, is made considering the estimated error that remains in the data after interactive editing. The number of units selected according to this criterion relies on a user-specified threshold that is related to the accuracy of the estimate of interest.

In the following, the main functions of the package SeleMix are described:

- **ml.est:** this function performs the maximum likelihood estimates of the parameters of a contamination model by ECM algorithm and it provides the expected values of the "true" data for all units that were used for the estimation. Also it returns, to each unit, the posterior probability of occurrence of the error and the flags of classification as outliers – no outlier calculated on the basis of a threshold for the probability of error specified by the user. It requires the specification of the type of model assumed for the true data (normal or lognormal) and some technical parameters for the algorithm ECM. It requires the specification of the type of model assumed for the true data (normal or lognormal) and some technical parameters for the algorithm ECM.

- **pred.y**: on the basis of a set of contamination model parameters, and a set of observed data, it calculates the expected values of the corresponding real data. Missing values for the variables response as well as are allowed, but not for covariates.

- **sel.edit**: it performs Selective Editing. On the basis of a set of observed data and the corresponding predictions for the true data, it selects the units required for interactive editing. It requires in input the wanted accuracy threshold and, if present, the sample weights associated with the units. It provides the score for each unit and the corresponding rank.

Given the opportunity to use the features of the package also in the presence of incomplete data, it can also be used as a tool for robust imputation of multivariate Gaussian data.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	5.3 Review and validate 5.4 Edit and impute
Programming language:	R
Keywords:	latent class models, selective editing, influential error
Contact:	name: Maria Teresa Buglielli email: bugliell@istat.it

Software and documentation

COPYRIGHT

Copyright 2013 Istat

Licensed under the European Union Public Licence (EUPL), version 1.1 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://ec.europa.eu/idabc/eupl.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 12/12/2013

- SELEMIX Version 0.9.1 – Windows binaries

- SELEMIX Version 0.9.1 – Package source

INSTALLATION

Install the downloaded package from within R as follows:

> install.packages(path_to_file, repos = NULL)

where the character path_to_file is the path to the .zip or .tar.gz file you downloaded.

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Reference manual – SeleMix v. 0.9.1

Vignettes – SeleMix v. 0.9.1

OTHER DOCUMENTATION

Barcaroli, G., and D. Zardetto. 2012. "Use of R in Business Surveys at the Italian National Institute of Statistics: Experiences and Perspectives". In *Proceedings of the 4th International Conference of Establishment Surveys (ICES IV).* American Statistical Association, Montréal, 11-14 June 2012.

EVER (Estimation of Variance by Efficient Replication)

Description

The EVER package (the acronym stands for: Estimation of Variance by Efficient Replication) is mainly intended for calculating estimates and standard errors in complex surveys. Variance estimation is based on the extended DAGJK (Delete-A-group Jackknife) technique proposed by P. S. Kott.

The advantage of the DAGJK method over the traditional jackknife is that, unlike the latter, it remains computationally manageable even when dealing with "complex and big" surveys (tens of thousands of PSUs arranged in a large number of strata with widely varying sizes). In fact, the DAGJK method is known to provide, for a broad range of sampling designs and estimators, (near) unbiased standard error estimates even with a "small" number (e.g. a few tens) of replicate weights.

Besides its peculiar computational efficiency, the DAGJK method takes advantage of the strong points it shares with the most common replication methods. As a remarkable example, EVER is designed to fully exploit DAGJK's versatility: the package provides the user with a user-friendly tool for calculating estimates, standard errors and confidence intervals for estimators defined by the user themselves (even non-analytic). This functionality makes EVER especially appealing whenever variance estimation by Taylor linearisation can be applied only at the price of crude approximations (e.g. poverty estimates).

Main Statistical Functions

- Delete-A-Group Jackknife replication

- Calibration of replicate weights

- Estimates and Sampling Errors (standard error, variance, coefficient of variation, confidence interval, design effect) for:
    - Totals
    - Means
    - Absolute and relative frequency distributions
    - Ratios between totals
    - Multiple regression coefficients
    - Quantiles

- Estimates and Sampling Errors for user-defined Complex Estimators (even non-analytic)

- Estimates and Sampling Errors for Subpopulations (Domains)
    - All the analyses above can be carried out for arbitrary domains

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	5.6 Calculate weights 5.7 Calculate aggregates
Programming language:	R
Keywords:	calibration, estimation, replication variance, complex surveys, complex estimators, Delete-A-Group Jackknife, R
Contact:	name: Diego Zardetto email: zardetto@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

R (version ≥ 2.5.1)

COPYRIGHT

Copyright 2013 Istat

Licensed under the European Union Public Licence (EUPL), version 1.1 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://ec.europa.eu/idabc/eupl.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 16/06/2020

- EVER Version 1.3 – Windows binaries

- EVER Version 1.3 – Package source

INSTALLATION

Install the downloaded package from within R as follows:

> install.packages(path_to_file, repos = NULL)

where the character path_to_file is the path to the .zip or .tar.gz file you downloaded.

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Reference manual – EVER v. 1.3

OTHER DOCUMENTATION

EVER website

ReGenesees (R evolved Generalised software for sampling estimates and errors in surveys)

Description

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys) is a full-fledged R software for design-based and model-assisted analysis of complex sample surveys.

This system is the outcome of a long-term research and development project, aimed at defining a new Istat standard for calibration, estimation and sampling error assessment in large-scale sample surveys.

Main Statistical Functions

- Complex Sampling Designs
    - Multistage, stratified, clustered, sampling designs
    - Sampling with equal or unequal probabilities, with or without replacement
    - "Mixed" sampling designs (i.e. with both self-representing and non-self-representing strata)
- Calibration
    - Global and partitioned (for factorizable calibration models)
    - Unit-level and cluster-level weights adjustment
    - Homoscedastic and heteroscedastic models
    - Linear, raking and logit distance functions
    - Bounded and unbounded weights adjustment
    - Multi-step calibration
    - Calibration on multiple regression coefficients
    - Consistent trimming of calibration weights
- Basic Estimators
    - Horvitz-Thompson
    - Calibration Estimators
- Variance Estimation
    - Multistage formulation
    - Ultimate Cluster approximation
    - Collapsed strata technique for handling lonely PSUs
    - Taylor-linearization of nonlinear "smooth" estimators
    - Generalized Variance Functions method
- Estimates and Sampling Errors (standard error, variance, coefficient of variation, confidence interval, design effect) for:
    - Totals
    - Means
    - Absolute and relative frequency distributions (marginal, conditional and joint)

- o Ratios between totals
- o Shares and ratios between shares
- o Multiple regression coefficients
- o Quantiles
- o Population variance and standard deviation of numeric variables
- o Measures of Change derived from two not necessarily independent samples
- Estimates and Sampling Errors for Complex Estimators
  - o Handles arbitrary differentiable functions of Horvitz-Thompson or Calibration estimators
  - o Complex Estimators can be freely defined by the user
  - o Automated Taylor-linearization
  - o Design covariance and correlation between Complex Estimators
- Estimates and Sampling Errors for Subpopulations (Domains)

System Architecture

The ReGenesees system has a clear-cut two-layer architecture. The application layer of the system is embedded into an R package named **ReGenesees**. A second R package, called **ReGenesees.GUI**, implements the presentation layer of the system (namely a Tcl/Tk GUI). Both packages can be run under Windows as well as under Mac, Linux and most of the Unix-like operating systems.

While the **ReGenesees.GUI** package requires the **ReGenesees** package, the latter can be used also without the GUI on its top. This means that the statistical functions of the system will always be accessible by users interacting with R through the traditional command-line. On the contrary, less experienced R users will take advantage from the user-friendly mouse-click graphical interface.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	5.6 Calculate weights 5.7 Calculate aggregates
Programming language:	R
Language of the GUI:	EN
Keywords:	calibration, estimation, variance estimation, complex surveys, complex estimators, automated linearization, R
Contact:	name: Diego Zardetto email: zardetto@istat.it

Software and documentation

SOFTWARE DEPENDENCIES for package ReGenesees

R (≥ 2.14.0)

SOFTWARE DEPENDENCIES for package ReGenesees.GUI

R (≥ 2.14.0)

R packages: ReGenesees, tcltk2, RODBC and svMisc

COPYRIGHT

Copyright 2015 Istat

Licensed under the European Union Public Licence (EUPL), version 1.1 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://ec.europa.eu/idabc/eupl.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

DOWNLOAD

Release date: 09/09/2021

Package **ReGenesees** (statistical engine)

- ReGenesees Version 2.2 – Windows binaries
- ReGenesees Version 2.2 – Package source

Package **ReGenesees.GUI** (graphical user interface)

- ReGenesees Version 2.2 – Windows binaries
- ReGenesees Version 2.2 – Package source

INSTALLATION

Install the downloaded package from within R as follows:

> install.packages(path_to_file, repos = NULL)

where the character path_to_file is the path to the .zip or .tar.gz file you downloaded.

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Reference manual -ReGenesees v. 2.2

Reference manual – ReGeneseesGUI v. 2.2

OTHER DOCUMENTATION

ReGenesees website

Fallows, A., M. Pope, J. Digby-North, G. Brown, and D. Lewis. 2015. "A Comparative Study of Complex Survey Estimation Software in ONS". *Romanian Statistical Review*, Issue 3: 46-64.

Zardetto, D. 2015. "ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys". *Journal of Official Statistics*, Volume 31, N. 2: 177-203.

Zardetto, D. 2013. "ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Errors Assessment in Complex Sample Surveys". In *Proceedings of the 7th International Conferences on New Techniques and Technologies for Statistics (NTTS)*. Eurostat, Brussels, 5-7 March 2013.

Barcaroli, G., and D. Zardetto. 2012. "Use of R in Business Surveys at the Italian National Institute of Statistics: Experiences and Perspectives". In *Proceedings of the 4th International Conference of Establishment Surveys (ICES IV)*. American Statistical Association, Montréal, 11-14 June 2012.

ANALYSE phase

Prepare draft outputs

This sub-process is where the data are transformed into statistical outputs. It includes the production of additional measurements such as indices, trends or seasonally adjusted series, as well as the recording of quality characteristics.

A. Composite indices

A composite index is a mathematical combination (or aggregation as it is termed) of a set of indicators that represent the different dimensions of a phenomenon to be measured.

Constructing a composite index is a complex task. Its phases involve several alternatives and possibilities that affect the quality and reliability of the results. The main problems, in this approach, concern the choice of theoretical framework, the availability of the data, the selection of the more representative indicators and their treatment in order to compare and aggregate them.

In particular, we can summarize the procedure in the following main steps:

1. Defining the phenomenon to be measured. The definition of the concept should give a clear sense of what is being measured by the composite index. It should refer to a theoretical framework, linking various sub-groups and underlying indicators. Also the model of measurement must be defined, in order to specify the relationship between the phenomenon to be measured (concept) and its measures (individual indicators). If causality is from the concept to the indicators we have a reflective model – indicators are interchangeable and correlations between indicators are explained by the model; if causality is from the indicators to the concept we have a formative model – indicators are not interchangeable and correlations between indicators are not explained by the model.
2. Selecting a group of individual indicators. The selection is generally based on theory, empirical analysis, pragmatism or intuitive appeal. Ideally, indicators should be selected according to their relevance, analytical soundness, timeliness, accessibility and so on. The selection step is the result of a trade-off between possible redundancies caused by overlapping information and the risk of losing information. However, the selection process also depends on the measurement model used: in a reflective model, all the individual indicators must be intercorrelated; whereas in a formative model they can show negative or zero correlations.
3. Normalizing the individual indicators. This step aims to make the indicators comparable. Normalization is required before any data aggregation as the indicators in a data set often have different measurement units. Therefore, it is necessary to bring the indicators to the same standard, by transforming them into pure, dimensionless, numbers. Another motivation for the normalization is the fact that some indicators may be positively correlated with the phenomenon to be measured (positive polarity), whereas others may be negatively correlated with it (negative polarity). We want to normalize the indicators so that an increase in the normalized indicators corresponds to increase in the composite

index. There are various methods of normalization, such as re-scaling (or Min-Max), standardization (or z-scores) and 'distance' from a reference (or index numbers).

4. Aggregating the normalized indicators. It is the combination of all the components to form one or more composite indices (mathematical functions). This step requires the definition of the importance of each individual indicator (weighting system) and the identification of the technique (compensatory or non-compensatory) for summarizing the individual indicator values into a single number. Different aggregation methods can be used, such as additive methods (compensatory approach) or multiplicative methods and unbalance-adjusted functions (non-compensatory or partially compensatory approach).

5. Validating the composite index. Validation step aims to assess the robustness of the composite index, in terms of capacity to produce correct and stable measure, and its discriminant capacity (Influence Analysis and Robustness Analysis).

B. Seasonal adjustment of time series

Seasonality can be defined as the systematic intra-year movement caused by various factors, e.g. weather changes, calendar, vacation or holidays and usually consists of periodic, repetitive and generally regular and predictable patterns in the level of a time series. Seasonality can be influenced also by production and consumption decisions made by economic agents taking into account several factors like endowments, their own expectations, as well as preferences and the production techniques available in the economy.

Differently, cyclic pattern presents non fixed rises and falls and its fluctuation length is usually not shorter than 2 years.

The overlap of this two kinds of fluctuations (seasonal and cyclic) in a time series could provide some problems for short term (monthly or quarterly) variation interpretation, above all when the seasonal component is highly represented in the observed data. For this reason, in order to measure cyclical changes, short term variations are computed from seasonal adjusted series. In turn, seasonal adjustment is the process of seasonal and calendar effects removal from a time series. This process is performed by means of analytical techniques that break down the series into components with different dynamic features. These components are unobserved and have to be identified from the observed data based on an ex-ante assumptions on their expected behavior. Broadly speaking, seasonal adjustment includes the removal of both within-a-year seasonal movements and the influence of calendar effects (such as the different number of working days, or Easter and moving holidays).

Notice that calendar effects are not constant among different countries or economic sectors, so that, time series which include them are not comparable each other. For this reason, generally, calendar effects are removed together with seasonal component in the seasonal adjusted series, so that, it is possible to better catch the yearly variation (computed with respect to the same period of the previous year), as well as, the mean yearly variation. Moreover, together with the seasonal adjusted series, can be also produced time series corrected only for calendar effects.

Once removed the repeated impact of these effects, seasonally adjusted data highlight the underlying long-term trend and short-run innovations in the series.

Seasonal adjustment approaches

All the seasonal adjustment methods are based on the assumption that each time series, Yt (with a time index t = 1,2,…T), can be decomposed into three different unobserved components:

- A **trend-cycle (CT$_t$)** component representing long-run movement of the series (like those associated to business cycles). It generally depends on structural conditions like institutional situations, technological and demographic trends or patterns of civil and social organization.

- A **seasonal component (S$_t$)** representing the intra-year (monthly, quarterly) fluctuations.

- An **irregular component (I$_t$)** representing the short term fluctuations that are not systematic and, to a certain extent unpredictable, e.g. uncharacteristic weather patterns.

Although the series may be decomposed in different ways, generally two main approaches consistent with the European guideline (Eurostat 2015), are considered:

1. Arima Model Based (AMB) approach, developed among the others by Burman (1980), Box, Hillmer and Tiao (1978) and Hillmer and Tiao (1982), based on the assumption that there exists a statistical parametric model (ARIMA) representing the probabilistic structure of the stochastic process connected to the observed time series. Time series assumed to be a finite part of a particular realization of a stochastic process. The linear filters used in this approach depend, consequently, on the features of the time series considered. This kind of approach is adopted in the TRAMO-SEATS (Time series regression with ARIMA noise, missing observations and outliers and Signal Extraction in ARIMA time series - TS) procedure developed by Gómez and Maravall (1996).

2. Filter Based Approach (FLB), a non-parametric or semiparametric approach, which, differently from AMB approach, does not require to hypothesize a statistical model representing the series. Indeed, it is based on an iterative application of several linear filters on the series based on central moving averages. These procedures are referred to as ad hoc because the filters are chosen according to empirical rules, not taking into account the probabilistic structure of the stochastic process generating the series. To this approach belong the classical methods of the X-11 (X11) family: from the first X11 and X11-Arima (X-11A) to the more recent X-12-ARIMA (X-12A) (Findley et al. 1998) and X-13-ARIMA-SEATS (X-13AS) (Findley, 2005) which include several improvements over the previous versions; among which, the most remarkable is the use of reg-Arima models aimed at pre-treating the data and at improving the forecasting performances of the series, that in turn translates into an improvement of the moving average symmetric filters employed and, generally, into a higher stability of the estimated seasonal factor.

In both cases data are pre-treated for selecting the decomposition scheme to be applied to the time series (additive, multiplicative, log-additive, etc.). Moreover, some deterministic effects like outliers or calendar effects are removed. This pre-treated series is the input of the following step whose output is the seasonal adjusted series (SA). Once the seasonal adjusted series is obtained there is a last step in which some elements identified in the pre-treatment phase and related to the trend-cycle components (like level shift) or to the irregular component (like additive outlier or temporary changes) are included back; while are taken out from the final series the calendar effects and the seasonal outliers.

Apply disclosure control

The primary function of a public statistical system is to produce official statistics for its own country. In fact, the Legislative Decree no.322 of September 6, 1989, constituting the National Statistical System (Sistan), cites: "*The official statistical information is provided to the country and to international organizations through the National Statistical System*" (Article 1, paragraph 2), and "*the data processed in the framework of statistical surveys included in the national statistics program are the heritage of the community and are distributed for purposes of study and research to those who require them under the rules of the present decree, subject to the prohibitions contained in art. 9* "concerning statistical confidentiality (art. 10 paragraph 1).The Legislative Decree no.322 / 1989, also states that *"the data collected in the context of statistical surveys included in the national Statistical Programme cannot be communicated or disseminated to any external, public or private entity, nor to any office of public administration except in aggregate form so as to be unable to derive any reference to identifiable individuals."* In any case, the data cannot be used in order to re-identify the parties concerned.

General principles concerning the protection of data confidentiality, are established by the REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. In the framework of European legislation, the Code of Ethics and Good Conduct attached to the Code on the protection of personal data (Legislative Decree no. 4. 196 of 2003, amended by Legislative Decree no. 101 of 2018 containing provisions for the adaptation of national legislation to EU Regulation no. 679/2016), has been replaced by the Ethical Standards for processing related to statistical or scientific research purposes carried out within the National Statistical System (Gazzetta Ufficiale no. 11 of 14 January 2019). These Standards defines the concept of identifiability of a statistical unit, in terms of possibilities, through the use of reasonable means, to establish a significantly likely relationship between a combination of values referred to a certain set of variables and the identity of a respondent. It also specifies the means that can reasonably be used to identify the respondents, such as, for example, economic resources, time, linkage with nominative archives or other sources, etc..

The translation of the concepts set out by the law into operational rules, from a statistical point of view, requires a preliminary identification of the statistical units subject to identification risk and thus a precise definition of what constitutes a breach of confidentiality. The subsequent quantification of the probability of disclosure allows us to define the most suitable techniques to ensure data protection. By defining as *intruder* a person who has an interest in breaching the confidentiality of the released data, the intrusion occurs when the intruder is able to match, with a certain degree of certainty, the information released to the respondent. The release of statistical information in no case involves the variables that uniquely identify the person such as tax code, name or company name, address, etc.). The confidentiality problem arises for the variables that allow the intruder to narrow down a subset of units in order to pursue her purposes. To limit the risk of disclosure, National Statistical Institutes may modify data (for example by using perturbative techniques), or reduce the degree of details (*e.g. municipality replaced by district or region*). The application of the protection techniques, both for the dissemination of tables and for

the communication of elementary data, leads to a reduction or a change in the information content of the data.

The breach of confidentiality in the disseminated tables

Tabular data are mostly used by the National Statistical Institutes for the dissemination of aggregate data. The concept of a breach of confidentiality does not depend upon the type of product used for the dissemination. The definition of "confidential" information also includes data as defined by the articles 9 and 10 in the Regulation UE n. 679/2016 of the European Parliament and the Council, while public variables are not considered confidential (public variables refer to information present in public registers, lists, records, documents or sources accessible to anyone). In order to release a table an initial assessment is needed and regards the information to be published: if they are not confidential then protections are not applied. Data protection for tabular data involves several steps. The first stage defines which tables are to be processed and what their characteristics are. Then we define the risk rule or the criterion to determine whether or not a cell is at risk of a breach of confidentiality. The final phase concerns the implementation of procedures for the protection of confidentiality.

Magnitude tables

The risk rules for magnitude tables are those based on the size of the cells (threshold or frequency rule), and those based on concentration measurements (dominance rule and ratio rule). The threshold rule is widely used by the Istat and states that a cell is sensitive if the number of units contained therein is less than an n value (threshold) fixed a priori. In order to apply this rule to magnitude tables, the relative frequency tables are needed. There is no univocal criterion for identifying the threshold value that will depend on the assumed scenario of intrusion and the data processed.

According to the dominance rule [(n, k)-dominance] a cell is at risk if the first n contributors hold a proportion of its total value higher than the threshold k% fixed a priori. The level of protection that should be applied to the table depends the two values n and k. Based on the statistical units involved and the desired levels of security, it is possible to define the parameters by identifying a maximum allowable concentration.

The ratio rule (p-rule) relies on the accuracy with which the value of the first contributor can be estimated, assuming that the second contributor attempts the breach. The cell is considered at risk if the relative error is less than a p threshold fixed a priori.

In case of tables with possible data of opposite sign, the rules based on concentration measurements become meaningless. However, their application is possible by using the absolute values of the contributors.

A breach of confidentiality is more difficult to happen when negative magnitudes are possible. The general recommendation is to assign parameters to the risk functions with less stringent values than those adopted in the case all magnitudes are non-negative.

In the case of sample tables, the evaluation of the risk of disclosure has to take into account the sampling plan. Each cell value is an estimate achieved extending the sample value to the reference population because the true value of the population is not available. A weak risk of violation characterize cells with survey's weight greater than the unit. In this context a breach of confidentiality appears unlikely. However, a careful assessment of the disclosure risk is necessary. In fact, in some cases the most representative units (dominant) are included in the sample with certainty.

Frequency tables

Frequency tables are mainly used to represent social phenomena and census data. The criterion adopted to determine whether or not a cell is at risk for this type of tables is based on the size of the cells. Indeed the risk rules based on measures of concentration cannot be applied.

Statistical protection of tables

After the cells at risk are selected, it is necessary to modify the table to make anonymous its information content. There are many techniques of data protection and they range from the merging of originary cells to methods based on modifications of the original data, to the introduction of missing values (suppressions). Perturbation techniques modify data so that they cannot trace back to the original values. In this case the structure of the table remains unchanged but the additivity between internal and marginal values is not always guaranteed. The methods mostly used by Istat are the amendment of classification variables in order to induce the merging of the original cells and the introduction of missing values.

The technique based on the introduction of missing values consists in the suppression of the cells at risk. The distribution of the treated cells (primary suppression) has to ensure that the table is adequately protected. When it does not happen, additional missing values (secondary suppressions) applied to non-risk cells are needed. Several algorithms have been proposed by the literature on Statistical Disclosure Control in order to select secondary suppressions. Currently the HiTas algorithm , available in some generalized software such as Tau-ARGUS, is adopted by the Istat.

The breach of confidentiality in the release of elementary data

Elementary data can be defined as the end product of a statistical survey after the phases of design, implementation, monitoring and correction. The elementary data in the dissemination phase is an archive of records each containing all the validated information (usually a subset of those recorded) related to statistical units. These variables, as well as in the case of aggregate

data disseminated through tables, can be classified as variables usable for re-identification (hereinafter key variables) and confidential variables

The methods of protection fall into three main categories:

- o global recoding of variables reduces the detail of some variables (e.g. the age in the five-year classes rather than annual classes);
- o local suppression of information erases features that make some records at risk of disclosure;
- o data perturbation involves different methods to weaken the association between key and confidential variables .

Among the initiatives involved in the "protected" release of elementary data, Microdata Files for Research (MFR), public use files (mIcro.STAT) and the Laboratory for the Analysis of the ELEmentary data (ADELE) have to be mentioned. MFR files for social or business statistical surveys are specifically designed for scientific research needs. The release of these files is subject to the fulfilment of certain requirements relating both to the organization that requests to access the protected microdata and to the characteristics of its research project. MIcro.STAT files are public use files, obtained from the respective MFR file, properly treated in terms of confidentiality and downloaded directly from Istat website.

The ADELE Laboratory, active since 1999, is a so-called Research Data Centre (RDC) that is a "safe" place which can be accessed by researchers to make their own statistical analysis on elementary data produced by the National Statistics Institution in compliance with the confidentiality principles. The ADELE Laboratory provides to external "expert" users the ability to analyze the elementary data of the Istat surveys, by shifting the verification of confidentiality on the output of the analyses rather than on their input (as in the case of files for research purposes and public use files). The protection of confidentiality for the statistical analyses carried out at the ADELE Laboratory is ensured in several ways:

- o legally; the user subscribes a form in which he commits himself to respect specific rules of conduct;
- o physically, through the control of the working environment; the Laboratory is located at the Istat headquarters and input, output and access to the external network are inhibited to users;
- o statistically; all results are checked before the release.

ANALYSE phase – METHODS

- **PREPARE DRAFT OUTPUTS**

Computation and evaluation of composite indices

Methods for constructing composite indices: one for all or all for one?
2013
Rivista Italiana di Economia Demografia e Statistica, Volume LXVII n. 2.

Seasonal adjustment of time series

Relazione del team tecnico incaricato della definizione di metodi standard per la destagionalizzazione di serie storiche con metodi implementati in diversi strumenti IT (TS, X12-Arima, X13-Arima-Seats, JDemetra)
2015
Gdl per la definizione di standard per l'Istat, Istat

Destagionalizzazione di serie storiche con metodologia Arima model based (AMB) implementata nel software JDemetra+
2015
Gdl per la definizione di standard per l'Istat, Istat

ESS guidelines on seasonal adjustment (2015 Edition)
2015
Manuals and Guidelines, Eurostat

References

Computation and evaluation of composite indices

Massoli, P., M. Mazziotta, A. Pareto, e C. Rinaldelli. 2014. "Indici compositi per il BES". *Giornate della Ricerca*. Istat, 10-11 Novembre 2014.

Massoli, P., M. Mazziotta, A. Pareto, e C. Rinaldelli. 2013. "Metodologie di sintesi sperimentali per i domini del BES". *XXXIV Conferenza Scientifica Annuale dell'Associazione Italiana di Scienze Regionali (AISRe)*. AISRe, Palermo, 2-3 Settembre 2013.

Mazziotta, M., e A. Pareto. 2011. "Un indice sintetico non compensativo per la misura della dotazione infrastrutturale: un'applicazione in ambito sanitario". *Rivista di Statistica Ufficiale/Review of official statistics*, N. 1/2011: 63-79.

Mazziotta, C., M. Mazziotta, A. Pareto, e F. Vidoli. 2010. "La sintesi di indicatori territoriali di dotazione infrastrutturale: metodi di costruzione e procedure di ponderazione a confronto". *Rivista di Economia e Statistica del Territorio*, 1/2010: 7-33.

Organisation for Economic Co-operation and Development - OECD. 2008. "Handbook on Constructing Composite Indicators. Methodology and user guide". *OECD Publications*. OECD.

Aureli Cutillo, E. 1996. *Lezioni di statistica sociale. Parte II. Sintesi e graduatorie*. Roma: CISU.

Delvecchio, F. 1995. *Scale di misura e indicatori sociali*. Bari: Cacucci.

Silvio-Pomenta, J.F. 1973. "Typological study using the Wroclaw Taxonomic Method (A study of regional disparities in Venezuela)". *Social science project on human resources indicators, Study XXVIII.* United Nations Educational, Scientific and Cultural Organization (UNESCO).

Harbison, F.H., J. Maruhnic, and J.R. Resnick. 1970. *Quantitative Analyses of Modernization and Development*. New Jersey: Princeton University Press.

Seasonal adjustment of time series

Eurostat. 2015. "ESS Guidelines on Seasonal Adjustment (2015 Edition)". *Manuals and Guidelines*. Eurostat.

Grudkowska, S. 2015. *JDemetra+ Reference Manual, v. 0.1*. Narodowy Bank Polski.

Eurostat. 2014. "ESS Handbook for Quality Reports (2014 Edition)". *Manuals and guidelines*. Eurostat.

Findley, D.F. 2005. "Some Recent Developments and Directions in Seasonal Adjustment". *Journal of Official Statistics,* Volume 21, N. 2: 343-365.

Giovannini, E., and D. Piccolo. (Editors). 2000. " Seasonal adjustment procedures – experiences and perspectives. Proceedings of International Conference, 9-10 June 1998, Rome, Italy". *Annali di Statistica Istat*, Serie X, Volume 20. Roma: Istat

Findley, D.F., B.C. Monsell, W.R. Bell, M.C. Otto, and B.C. Chen. 1998. "New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program". *Journal of Business & Economic Statistics,* Volume 16, N. 2: 127-152.

Gómez, V., and A. Maravall. 1996. *Programs TRAMO and SEATS, Instruction for User (Beta Version: September 1996)*. Banco de España.

Chen, C., and L. Liu. 1993. "Joint Estimation of Model Parameters and Outlier Effects in Time Series". *Journal of the American Statistical Association,* Volume 88, Issue 421, Theory and Methods: 284-297.

Hillmer, S.C., and G.C. Tiao. 1982. "An ARIMA-Model-Based Approach to Seasonal Adjustment". *Journal of the American Statistical Association,* Volume 77, Issue 377, Theory and Methods: 63-70.

Burman, J.P. 1980. "Seasonal Adjustment by Signal Extraction". *Journal of the Royal Statistical Society*, Series A (General), Volume 143, Issue 3: 321-337.

Box, G.E.P., S.C. Hillmer, and G.C. Tiao. 1978. "Analysis and Modeling of Seasonal Time Series". In Zeller, A. (Editor). *Seasonal Analysis of Economic Time Series*: 309-334. NBER

Box, G.E.P., G.M. Jenkins, and G.C. Reinsel. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

- **APPLY DISCLOSURE CONTROL**

La produzione di MFR e mIcro.STAT in breve – Indagini su famiglie e individui
2017
Linee guida, Istat

Disclosure protection of non-nested linked tables in Business Statistics
2009
Supporting paper, Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain

Metodologie e tecniche di tutela della riservatezza nel rilascio dell'informazione statistica
2004
Metodi e Norme, N. 20, Istat

References

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Sculte Nordholt, K. Spicer, and P.P. De Wolf. 2012. *Statistical Disclosure Control*. London: J. Wiley & Sons.

ANALYSE phase – TOOLS

- **PREPARE DRAFT OUTPUTS**

COMIC (COMposite Indices Creator)

The COMIC software computes composite indices, by various aggregation methods, and evaluates their robustness

Ranker
Software system based on Visual Basic language for the analysis and benchmarking of results produced by different methods of statistical summary on composite indicators available in the literature.

- **APPLY DISCLOSURE CONTROL**

ARGUS
Generalized software for applying disclosure control.

COMIC

Description

The COMIC software computes composite indices, by various aggregation methods, and evaluates their robustness. COMIC calculates composite indices relevant to different periods of time.

COMIC deals with:

- different format acquisition (.csv or .xls or .txt or SAS) for different individual indicators available for each unit, as specified by users;
- standardization/normalization of the individual indicators;
- different aggregation methods;
- output display in both tables and graphics;
- comparison of the results;
- sensitivity and robustness analysis to evaluate the composite indices.

COMIC has implemented the following aggregation methods:

- Mean indices 0-1. Arithmetic mean of individual indicators normalized by min-max method;
- Mean z-scores. Arithmetic mean of standardized individual indicators;
- Jevons Index. Geometric mean of individual indicators computed as index numbers;
- Mazziotta-Pareto Index (MPI). Arithmetic mean of standardized individual indicators; a penalty is subtracted from the mean to take into account the variability of the individual indicators (MPI with negative penalty);
- Adjusted Mazziotta-Pareto Index (AMPI). Arithmetic mean of the individual indicators normalized by the min-max method; a penalty is subtracted from the mean to take into account the variability of the individual indicators (AMPI with negative penalty);
- Geometric Mean Index (IMG). Geometric mean of the individual indicators normalized by the min-max method.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	6.1 Prepare draft outputs
Programming language:	SAS
Language of the GUI:	IT
Keywords:	composite indices, normalization, aggregation, ranking, sensitivity, robustness
Contact:	name: Pierpaolo Massoli email:pimassol@istat.it

Software and documentation

SOFTWARE DEPENDENCIES

SAS for Windows (SAS base)

COPYRIGHT

DISCLAIMER

DOWNLOAD

Release date: 27/02/2017

COMIC version 1.0

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

Manuale utente - COMIC v. 1.0

OTHER DOCUMENTATION

Massoli, P., M. Mazziotta, A. Pareto, and C. Rinaldelli 2015. "COMIC: a tool for composite indices evaluation". *Conference Dealing with complexity in society: from plurality of data to synthetic indicators.* Università di Padova, Padova, 17-18 Settembre 2015.

Ranker

Description

Ranker is a software system, for the analysis and benchmarking of results produced by different composite indicators.

Ranker is a prototype of a generalized tool, which allows to:

- acquire in standard format (.xls) values of the different individual indicators available for each unit (eg. geographical areas);
- compute, for each unit, one or more implemented methods;
- display the values and rankings of each method both in tables and graphics;
- compare the rankings of the different methods.

Ranker has implemented eight methods:

- Mazziotta-Pareto Index (MPI) (De Muro et al. 2010);
- Wroclaw (Wroclaw);
- Arithmetic Mean of the z-scores (M1Z);
- Ranking (Grad.RNK);
- Relative indices (IR);
- Arithmetic Mean of index numbers (ANIM);
- Geometric Mean of index numbers (GNIM);
- Squared Mean of index numbers (QNIM).

The data analysis is based on five different phases:

1. The organization and the reading of data. The acquired data must be organized in tabular form, in a matrix that describes each territorial unit (in line) according to the value of the individual indicators (in column). For each indicator considered, it is necessary to provide, in addition, the polarity, distinguishing those that describe an effect "positive" with respect to the dynamics of the phenomenon and those that, on the contrary, are related in the reverse direction and to which corresponds a decreasing ranking of the territorial units;

2. The normalization. This step is intended to obtain indicators purified by the specific units of measurement, which are of equal width (eg. between 0 and 100);

3. The aggregation. After loading the matrix with individual indicators and after their standardization, the application makes the processing of all the methods mentioned above in order to calculate the relative composite indicators and describe each territorial unit;

4. The visualization. Through the application you can view both the values obtained and the rankings derived from them. Besides the description of the numerical values in tabular form, Ranker enables the production of maps that graphically represent the results obtained by each methods;

5. The evaluation of the methods. The software enables you to evaluate comparatively the results produced by different methods: in particular, the impact of the choice of the

method on the final result (the ranking).A first instrument is given by the matrix of rank correlation obtained comparing different methods. A second instrument used is the matrix plot, which can be produced on the values of the indicators, or better, on the rankings obtained. A final tool is given by the scatter plot calculated on pairs of selected methods.

Istat offers the availability on choosing the online version (i.ranker) or the desktop one. The online version implements the first five methods above described and offers the ability to view the values and the ranking resulting from the application of each method both in tabular and graphical layout. The desktop version implements all methods but only allows the tabular display of results.

Information

Status:	validated
Author:	Istat
Licence:	EUPL-1.1
GSBPM code:	6.1 Prepare draft outputs
Programming language:	Visual Basic
Language of the GUI:	IT
Keywords:	composite indicators, normalization, aggregation, ranking
Contact:	name: Marco Broccoli email: broccoli@istat.it

Software and documentation

COPYRIGHT

Copyright 2014 Istat

Licensed under the European Union Public Licence (EUPL), version 1.0 or subsequent. You may not use this work except in compliance with the Licence. You may obtain a copy of the Licence at: http://ec.europa.eu/idabc/eupl.html. Unless required by applicable law or agreed to in writing, software distributed under the Licence is distributed on an "AS IS" basis, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licence for the specific language governing permissions and limitations under the Licence.

DISCLAIMER

Istat assumes no responsibility for the results arising from use of the instrument that is inconsistent with the methodological guidance contained in the documentation available.

ONLINE SYSTEM

i.ranker

DOWNLOAD

Release date: 17/02/2015

RANKER version 1.2

TECHNICAL AND METHODOLOGICAL DOCUMENTATION

User manual – Ranker v. 1.2

ARGUS

Description

Istat participated in the European project, CASC (Computational Aspects of Statistical Confidentiality) whose goal was the development of the software ARGUS for the protection of the data confidentiality in the release of statistical information.

The ARGUS software consists of two modules Mu -ARGUS, for elementary data, and Tau-ARGUS, for tables, and contains many of the methods of protection proposed in the literature.

Information

Status:	validated
Author:	CASC
GSBPM code:	6.4 Apply disclosure control
Keywords:	confidentiality, risk of identification, value at risk, suppression of value, risk rule

Software and documentation

To get the ARGUS software as well as the methodological and technical documentation, please contact CASC.