EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

STATISTICAL DISCLOSURE CONTROL FOR THE GENERAL PUBLIC DISTRIBUTION OF MULTIDIMENSIONAL CUBES: AN EXPERIMENT AT THE FRENCH SERVICE OF AGRICULTURAL STATISTICS

Michael Levi-Valensin – SSP/French Ministry of Agriculture
michael.levi-valensin@agriculture.gouv.fr

MINISTÈRE DE L'AGRICULTURE ET DE LA SOUVERAINETÉ ALIMENTAIRE
Liberté
Égalité
Fraternité

Background

Suppression methods (masking information in frequency and magnitude tables) are widely used in the French public statistical service according to the principle of the EUROPEAN STATISTICS CODE OF PRACTICE

Primary suppression rules: minimum frequency (3 units) and dominance (the dominant unit must not represent more than 85% of the aggregate value), a heuristic widely used in statistical services

Secondary secret generated by the publication of subtotals or by hierarchical levels is **more tricky to handle**. Several ways to mask other cells. You don't want to lose too much information !

Quality Assurance Framework of the European Statistical System

Principle 5 – Statistical Confidentiality and Data Protection

The privacy of data providers, the confidentiality of the information they provide, its use only for statistical purposes and the security of data are absolutely guaranteed.

Tools

- **τ-Argus,** the reference software, do not integrate into R processing chains.
- Perturbative methods (rounding, cellkey…) not appropriate for national heuristics
- R library **sdctable** from Statistics Austria more handy !
  https://sdctools.github.io/sdcTable/articles/sdcTable.html

Three steps:

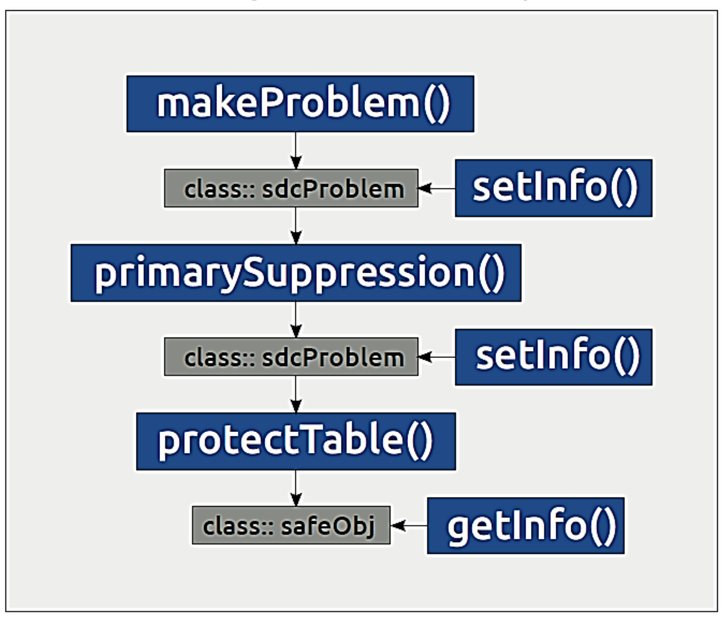
1. Create the hypercube using the **makeProblem()** function with the list of the dimensions, the names of the table variables, and the numerical variable.

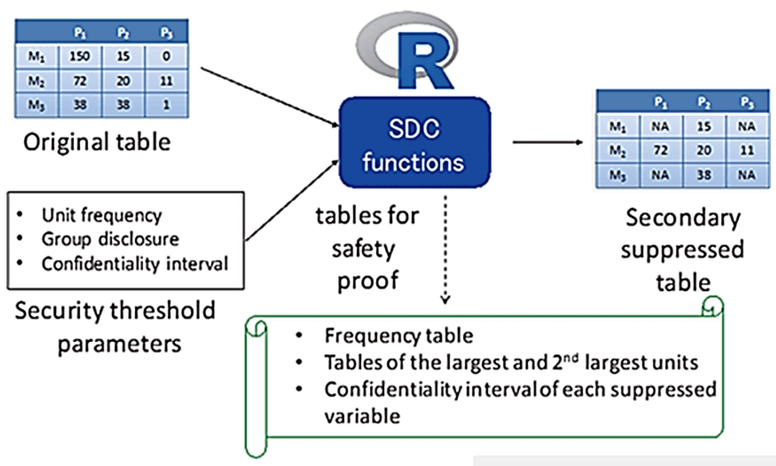
2. Apply primary secret rules to the hypercube using the **primarySuppression()** function
3. Apply **secondary secret** rules to the object using the **protectTable()** function.

level	name
@	France entière
@@	0
@@@	011
@@@@	01177
@@@@@	01177001
@@@@@	01177002

Define hierarchical levels (geography, nomenclatures) with the library sdcHierarchies

Methods from τ-Argus are available (HYPERCUBE, OPT, HITAS), as well as a specific algorithm SIMPLEHEURISTIC





A status is given to the cell after primary and secondary suppression : cells with the status « u » or « x » are masked

- "u": cell is primary suppressed and needs to be protected
- "x": cell has been secondary suppressed
- "s": cell can be published
- "z": cell must not be suppressed

Application to interactive tables on the public dissemination website

The library **sdctable** has been used to produce *hypercubes*, interactive tables with several dimensions the user can choose and several hierarchical levels, on the French website Agreste

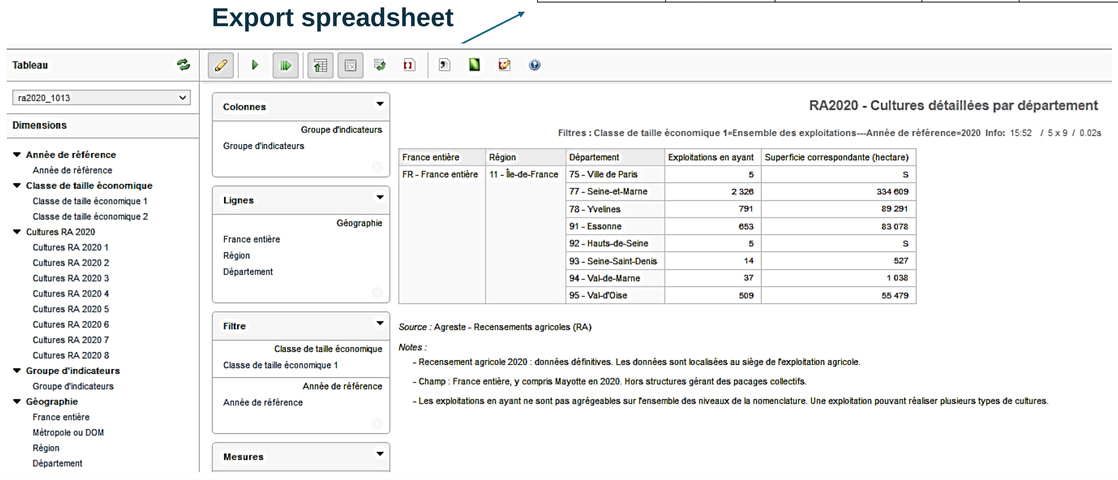
For example, the agricultural census 2020
https://agreste.agriculture.gouv.fr/agreste-web/disaron/RA2020_1013/detail/
All the variables and all the hierarchical levels of the hypercube are loaded.

Well chosen cells are masked with the value "s" : such as in departments with few agricultural exploitations

RA2020 - Cultures détaillées par département
Agreste - Recensements agricoles (RA)
Accéder au tableau interactif

France	Region	Department	Agricultural exploitations	Area (hectare)
FR - France entière	11 - Ile-de-France	75 – City of Paris	s	
FR - France entière	11 - Ile-de-France	77 - Seine-et-Marne	2326	334609
FR - France entière	11 - Ile-de-France	78 - Yvelines	791	89291
FR - France entière	11 - Ile-de-France	91 - Essonne	653	83078
FR - France entière	11 - Ile-de-France	92 - Hauts-de-Seine	s	
FR - France entière	11 - Ile-de-France	93 - Seine-Saint-Denis	14	527
FR - France entière	11 - Ile-de-France	94 - Val-de-Marne	37	1038
FR - France entière	11 - Ile-de-France	95 - Val-d'Oise	509	55479

Export spreadsheet



Additional targets and constraints

- ✓ **Cell protection**: Some cells may be allowed to be published and not masked. Use of the function **change_cellstatus** to protect the cell before secondary suppression. *For example, at the municipal level in the 2020 Agricultural Census, secret no longer applies, among other criterias, to* ***areas under cereals and oilseed crops, permanent crops, and grasslands.***
- ✓ **Protection of zeros :** cells with a zero value must be published.
- ✓ **Weighted data:** For survey data, integer weights are used to count the frequency and the dominance. The algorithm needs integer weights to sort the values.

Conclusions

Successfully applied in about ten interactive tables of the agricultural census

Advantages of the use of the R library **sdctable** for putting files under secret:

- Same working environment as for preparing the data to be disseminated in a processing chain.
- Having predefined functions to format data files and hierarchical files before global processing.
- Keeping a record of the processing and being able to easily adapt the scripts as needed.

Drawbacks:

- It takes a long time and high capacity to build hierarchical files and to run the functions, too high with all the French municipalities and the detailed livestock or crops.
  We must reduce the number of modalities
- With huge hypercubes with many hierarchies, the dominance rule needs a long processing time
- The library is actually used by few countries and relies on only one person
- Residual risks of disclosure (approximate disclosure, logical links between variables...)

Pursuant to two principles of the **Quality Assurance Framework**

Principle 7: Sound Methodology.

Sound Methodology underpins quality statistics. This requires adequate tools, procedures and expertise.

Principle 10: Cost effectiveness.

Resources are used effectively.

Indicator 10.2: The productivity potential of information and communications technology is being optimised for the statistical processes.