Abstracts

SDC-User Workshop 2022

Using rtauargus package to protect tabular data in a complete R workflow

Julien Jamme and Natahanael Rastout**,** INSEE

Taking advantage of a new release of the package during the summer with a complete English documentation, we'd like to present in more details how we proceed to protect data from one business survey (maybe ICT), and how we deal with non-nested hierarchies, multiple linked tables, etc.

An approach to resolve feasibility problems due to the frozen cell problem in τ-Argus Modular

Michel Reiffert, Destatis

When using cell suppression methods to protect tables in a late stage of the dissemination process of a statistic, one often has to take into account the suppression pattern of disseminated tables released earlier. Suppression status of identical cells should be consistent across linked tables. Even if not all tables are released at the same time. In practice, this often leads to infeasibility problems, especially in multi-dimensional hierarchical tables, where the number of sub-tables that have to be considered are large.

We introduce an approach that allows those infeasibility problems to be resolved in an iterative process, in which certain relaxed versions of the original optimization problems are solved. From those solutions, it is possible to draw conclusions about the nature of the underlying infeasibility problem. Depending on the level of accepted disclosure risk, different relaxations are considered in the iterative process.

The suggested approach will hopefully strongly reduce the number of tables, for which the existing Modular method implemented in τ-Argus terminates abnormally. Our workshop contribution will present a first implementation of the approach and we will discuss possible ways to integrate it in future versions of τ-Argus, if testing the methodology in a proof of concept leads to positive results.

Primary and secondary protection of statistical confidentiality of census data in Poland – practical issue

Tomasz Klimanek, Andrzej Młodak, Tomasz Józefowski, Statistical Office in Poznań

The purpose of the paper is to present the basic assumptions and problems concerning protection of data collected during the National Population and Housing Census 2021 in Poland. We will describe forms and the scope of census data to be released and key requirements for the protection of data against unit identification and the disclosure of sensible information. Various non-perturbative and perturbative methods and tools of statistical disclosure control for microdata and tabular data will be

discussed. In this regard we will highlight problems encountered when applying these methods, including those resulting from of the limitations of software used in the process. These include, among others, primary and secondary suppression of tabular data and identification of quasi-identifiers in microdata. Our case study will be based on a special test file reflecting the content and distribution of real key census variables.

GaussSuppression: An R package for Tabular Data Suppression

Øyvind Langsrud, Daniel Lupp and Hege Marie Bøvelstad, Statistics Norway

The package GaussSuppression is designed to be a flexible tool for cell suppression using Gaussian elimination. The method shows promising results in practice, striking a balance between computational complexity and number of suppressed cells. The algorithm is built around a dummy matrix that relates input data to output data. We refer to this matrix as the model matrix. Internally, this matrix is constructed based on the specifications of the tables to be published. The package provides a lot of flexibility, due to its dependency on the function ModelMatrix from the SSBtools package: hierarchies can be specified in several ways and need not be tree-shaped, and several linked tables can be specified by a model formula in R. Primary suppression can be done by using built-in functions (with support for common heuristics such as small counts and dominance rules) or custom functions supplied by the user. The package also implements k-disclosure suppression, a novel method for cell suppression in frequency tables aimed directly at protecting against disclosure of sensitive information related to statistical units. In addition, more advanced options such as singleton handling and specification of forced and hidden cells are supported. Aggregation from microdata can also be done automatically. The whole process from input to output can be performed with a single function call. Our presentation will provide an overview of the package and its features, as well as demonstrate examples of its use in production at Statistics Norway.

Solution for applying SDC to Eurostat Outward Foreign Affiliates Statistics (OFATS) deliverables involving use of TAU-Argus

Karina Dineen, Tim Linehan, NSI (CSO Ireland)

As the geographical groupings in this series contain non-nested hierarchies, and there is also a non-standard primary suppression requirement, a standard TAU-Argus implementation could not be used for the application of primary and secondary suppression.
I, Karina Dineen, and my colleague, Tim Linehan, developed a solution for applying statistical disclosure control to the Business Demography deliverables based on the method presented by Virgili and Franconi ("Disclosure protection of non-nested linked tables in business statistics." http://www.istat.it/it/files/2013/12/Franconi_Virgili_wp.36.e.pdf) where non-nested hierarchies are changed to a series of nested hierarchies and protection is applied through the use of a-priori files. Currently we are not using commercial secondary solvers but rather the standard free solvers available with TAU-Argus. This solution has led to a streamlined SDC process with significantly less time required for checking when compared to the previous manual-checking based approach.

SDC methods for wage data explorer

Hans Haraldsson, Statistics Iceland

Statistics Iceland will be making population wage data available to the public through a data explorer or dashboard. The data will include several variables that are predictive of wages, such as business sector, occupation, job experience, age and gender. The data explorer will allow members of the public to explore the distribution of wages by up to 3 or 4 variables.

The variables in question can often form unique or rare combinations with high risk of identity disclosure and possibly unacceptable levels of inferential disclosure. Three potential solutions will be studied (the categorical variables are not considered sensitive and it is not necessary to avoid attribute disclosure).

One solution is to show only summary tables while merging categorical variables and binning numerical variables to meet criteria for k-anonymity and l-variety for every possible table. While this is a relatively simple solution to execute there is high risk that cells will be unbalanced giving misleading results (e.g. age within bins may be differently distributed for men and women producing an illusory gender effect).

The second solution is to replace the data with synthetic data. While this approach prevents identity and attribute disclosure it has two serious drawbacks. The first is that the validity of results rests heavily on the ability of the models used to synthesize the data to recreate relationships between variables. The second is potential lack of face validity.

The third solution is to replace a relatively small proportion of the data with synthetic values, i.e. the wage variable for all persons with sample unique combinations of values on 3 or 4 categorical variables and a small randomly selected proportion of values on all numerical variables. This approach prevents positive identification of individuals while avoiding the problem of unbalanced cells and places less weight on the models used to synthesize data. However, an unacceptable level of inferential disclosure for some members of the population is a possibility.

Use of multilevel grids to release protected grid data from the French 2021 Census

Julien Jamme and Clément Guillo, INSEE

In the context of releasing census data on a 1km² grid, Insee chose to protect these data with a method of geographical aggregation based on *quadtree* approach applied on grid data. This method allows to preserve confidentiality while detailing at the finest level. The confidentiality rule is here not to release proper data with less than 11 households in a cell. Nevertheless, total of population in each cell is disclosed without any protection. In the case of grid data, one strategy may be to assemble contiguous cells into larger polygons (e.g.larger rectangles and cells) so that each polygon meets the threshold. The new polygons can be obtained by aggregation, i.e by grouping the cells until the threshold is reached, or by disaggregation, starting from the largest cell and splitting it until it is no longer possible to cut it without going below the threshold. These methods have the advantage of preserving the additivity but also allow to avoid the creation of "false zeros". In addition the threshold rule is respected by construction. The areas first obtained by disaggregation are called the "natural" areas and they can be of different sizes depending on when we stop splitting. However, it is possible to obtain more accurate information if we continue to divide the resulting cells in smaller cells and if we decide to hide the information from the cells that are below the threshold. Obviously, it is not sufficient to hide only the cells below the threshold since the information can be found by geographical differentiation comparing the finer level of detail to a coarser level of detail.

Therefore, we have to hide the information contained in another cell at the same level. This process requires disseminating information on several grids ("composite" grid) corresponding to different levels of detail, which is equivalent to disseminating information on the same grid with cells having different shapes and sizes. In this way, we are able to protect data without hiding any information. At Insee, this method was originally implemented to release very sensitive grid data such as income and other data from tax sources and has been in use for several years. The ideal dissemination of such data does not consist in a single 1km² grid, but in a whole set of grids of different tile sizes, which will be released by Insee. The new contours introduced with this method, combined with other administrative contours can generate statistical disclosure through geographical differentiation. A package R called *diffman* makes the detection of these potential disclosures easier.