Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (i): New methods for protection of tabular data or for other types of results from table and analysis servers

Metadata driven application for aggregation and tabular protection

Prepared by Andreja Smukavec, Statistical Office of the Republic of Slovenia

Metadata driven application for aggregation and tabular protection

Andreja Smukavec*

* Statistical Office of the Republic of Slovenia, 54 Litostrojska ulica, 1000 Ljubljana, Slovenia, andreja.smukavec@gov.si

Abstract: In modern official statistics production short timeliness of the disseminated results is more and more demanded; therefore, the Statistical Office of Slovenia (SURS) started with the internal project the final aim of which is to relieve survey managers, methodologists and the IT sector of repeated work. The final product of the project will be a metadata driven application for aggregation and tabular protection. The application will be built on the basis of the so-called metadata driven principle, meaning that there is one general program code which is then for the particular survey parameterized through the (process) metadata tables. The metadata approach causes some additional work for the first time, but for the next time points the metadata will be possible to reuse. For tabular protection the German SAS-Tool[1] will be implemented and used whenever possible. In cases of too complex table sets, the metadata driven application will not be a suitable solution, but it will still help us to construct the input files for Tau Argus[2], a software program designed to protect statistical tables. At the same time precision requirements will be implemented and used as additional rule for primary sensitivity of aggregated data.
The paper describes the metadata driven procedure and application, the dilemmas and trade-offs of such an approach.

1 Introduction

It is quite a common practice at SURS that the same data are tabulated more than once. The survey manager gives the instructions for tabulation to the IT sector, to the survey methodology department for estimating standard errors and to the methodologist for confidentiality treatment, who applies tabular protection using Tau Argus. Therefore, SURS started the internal project Standardization of Data Treatment in 2011 to find a more rational solution. Although the project is constructed of two parts, one part is dedicated to editing and imputations, this paper will talk only about the second part of the project, which is dedicated to aggregation, tabulation, sampling error estimation and confidentiality treatment of tables. The main goal of the project is automation of the data production process and making it more transparent. The final output of the project will be a metadata driven application with a user-friendly interface. A metadata driven application means that all the parameters for the particular survey and the corresponding reference period are provided in a special so-called metadata database, for which MS Access is used

[1] http://www.cros-portal.eu/sites/default/files//PS1%20Poster%207.pdf
[2] http://neon.vb.cbs.nl/casc/tau.htm

at the moment, but it will be transferred to ORACLE database by the end of 2013. Behind the application a general code will run, which will read data from the metadata database and execute the required procedure. A general code means that it should never be changed for the needs of a particular survey, and it will be written in SAS. A part of the general code for aggregation, simple tabulation and sample error estimation has already been written. SAS code for preparation of the input files for Tau Argus (a tool for secondary suppression) and PX-Edit (a tool for publication) is still missing and is being written at this moment. The current output of the application are Excel files, where all the needed statistics are written with denotations for lower degree of precision and primary sensitivity due to selected safety rules. The final aim of the project is to have subject-matter personnel run all or a part of data production fully automated with this application. It will be much more work for the first time as the whole metadata database will have to be filled, but for the next reference periods the transfer from previous time points will be possible. We are aware that not all surveys are appropriate for such an approach, but maybe at least a part of production could be automated.

The part of the project that is dedicated to editing and imputations has mainly been finished, the metadata database has been already transferred into ORACLE environment, the general SAS code has been already built-in the application, which is in the testing phase at this moment. Some bugs have been discovered and have to be fixed and of course some upgrades are planned for the future, but the first version of the application written in DOT.NET technology is available. The production is planned for 2014. The second part of the project started to evolve this year.

At this moment SURS does not have any automated procedures for the confidentiality treatment of tables; sometimes batch files are used, but usually the methodologist for confidentiality treatment applies cell suppression using Tau Argus. We plan to incorporate the German SAS-Tool for secondary suppression of the tables in the application. It is also a metadata driven application with general SAS code and Tau Argus for secondary suppression, which will simplify the transition to our new system as SAS and Tau Argus have been already a part of our data production. Because Tau Argus covers statistical disclosure control only for aggregates, additional instructions will have to be used for other statistics. All required statistics will be written into one special table with their corresponding statuses for precision and confidentiality. If the subject-matter methodologist is satisfied with the results, the tabulation and publication of the data will be done. By joining all the processes from editing raw data to publication of the data in this application, all the metadata information will be available in one place.

Currently the outputs to Excel are:

- estimated statistics;
- estimated sampling errors and coefficients of variation;

- statistics with primary sensitivity status and precision statuses.

We are in the testing phase at the moment, comparing sensitivity statuses obtained by our general SAS code to the results from Tau Argus; of course, this is done only for statistics which can be protected with Tau Argus – frequencies and sums of quantitative variables.

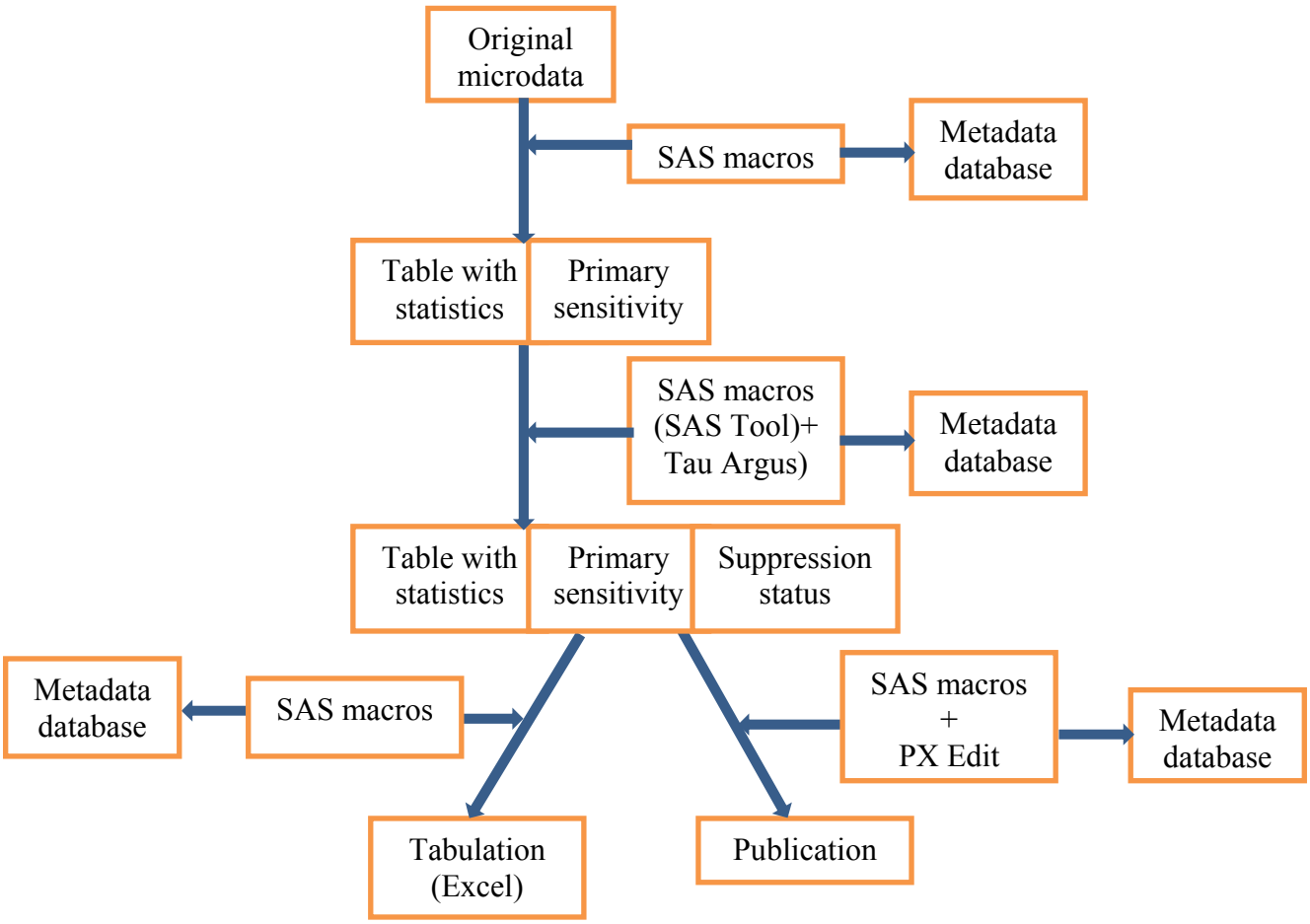

Fig 1. Process of tabulation and aggregation

1.1 Safety rules for determining the primary sensitivity of data

Currently SURS uses the following safety rules for tabular protection:

- Threshold – the minimum number of contributors that the statistic is safe.
- Dominance rule – a statistic is sensitive if one or two contributors are dominant.
- p%-rule – a statistic is sensitive if the second largest contributor assesses the largest one too precisely.
- Zero unsafe – all zero statistics are unsafe (optional).
- Holding feature – each respondent can contribute data to more than one cell in a table.

In very rare cases the request rule (some data providers demand the protection) is also used, but it has not been implemented in the general SAS code yet. All other features have been already incorporated, including holdings.

Different types of statistics were defined for this application. The most commonly used and corresponding safety rules are:

Type	Description	Safety rules
01	Number of reporting units with specific property	threshold
02	Proportion of reporting units with specific property	threshold
03	Sum of continuous variable	threshold, dominance rule, p%-rule, zero unsafe
04	Average of continuous variable	threshold, dominance rule, p%-rule, zero unsafe
05	Ratio of sums of two continuous variables	A sum in numerator or denominator is unsafe according to the rules for statistic type 03.

Tab 1. Safety rules for different types of statistics

We also have a parameter in the metadata database called "*complement"*, which checks primary sensitivity of the statistic, which is the complement of the demanded statistic.

We also designed a code list for primary sensitivity statuses, which are situated in the table with statistics:

Type	Status for primary sensitivity
10	threshold t
21	dominance rule (1,k)
22	dominance rule (2,k)
	…
31	p%-rule
	…

Tab 2. Statuses for primary sensitivity in the table with statistics

2 Metadata database and general SAS code

There will be one metadata database for both parts of the project to avoid duplicating parts that are the same. At the starting point the whole metadata database was constructed in MS Access; only for editing and imputations it has already been transferred to ORACLE, for the tabulation and aggregation the transition to ORACLE will be done by the end of 2013.

The metadata database for tabulation and aggregation is constructed of several metadata tables:

- Table of derived and dummy variables needed for the aggregation (domains and statistics which are not present in the original micro-data set).
- Table of demanded statistics and their description - averages, sums, frequencies, ratios, etc. Statistics can be derived variables or variables from the original micro-data file.
- Table of domains for which statistics have to be calculated. The domains define the dimensions of the tables (at this moment it is possible to use up to 10 variables for one domain).
- Metadata for safety rules.
- Metadata for sampling error estimation.
- Metadata for tabulation.

The general code calculates all statistics for all domains, because it is not a time-consuming operation. All calculations are written in a table with statistics, but only the statistics which are defined by the metadata for tabulation are written

into the Excel output. The condition for this application is that each statistic is calculated from one micro-data file, which has to contain the variable *reference period*. If more than one micro-data file is the source for statistic, then those micro-data files have to be joined before running the application. As such a situation usually occurs, this will be the only part of the code that changes from survey to survey.

The general SAS code (SAS macros) reads the metadata database, creates the table with statistics from micro-data and tabulates. It is formed by three different SAS macros:

- Preparation of the micro-data file – macro that adds derived and dummy variables to the micro-data file.
- Macro that computes all needed statistics, marks the primary sensitive ones and adds the corresponding standard error and coefficient of variation.
- Macro for tabulation into Excel.

The general code for the calculation of statistics (sum, number of units, average, etc.) and simple tabulation to Excel has already been written, also the part for calculating standard errors is almost finished.

We need to incorporate the SAS code from the SAS-Tool, write a code for preparation of the input files for Tau Argus from the table with aggregates and write a code for preparation of special format files for publication on our data portal. We still have not added the metadata for construction of the input files for Tau Argus and SAS-Tool to our metadata database, also metadata for creating special format files for publication on SURS's data portal are missing.

2.1 Boundaries table in the metadata database

The Boundaries table contains the metadata for safety rules and criteria for the denotation of the statistics with lower precision. These metadata are used for determination of statistic's status, which indicates whether a statistic is primary sensitive, less precise or imprecise. It also contains the needed metadata for creating input files (tabular data format) for Tau Argus. We also add a parameter for safety margin in case of threshold (in case of dominance rule in p%-rule upper and lower protection levels are derived from values and parameters of the rule). A parameter which determines how zero values are treated (usually they should be all sensitive) and corresponding safety margin depend on statistic, therefore these two parameters exist in the metadata table of demanded statistics.

We plan to use the precision requirements also in the confidentiality treatment of the statistics, therefore we add in one of the metadata tables the parameter how criteria for the denotation of the statistics with lower precision measurements (SURS uses

three levels of precision according to the value of coefficient of variation – precise, less precise and imprecise data) influence the statistic's primary sensitivity statuses.

We included four possibilities:

- Precision requirements and safety rules are used.
- Only precision requirements are used.
- Only safety rules are used.
- Neither precision requirements nor safety rules are used.

In case that both measurements are taken into account, we decided for the following rule:

- If the statistic is imprecise and primary sensitive, then we set its status to safe.
- If the statistic is less precise and primary sensitive, then it is still primary sensitive.

At SURS imprecise data are usually determined by coefficient of variation larger than 30%.

Content of the Boundaries metadata table:

SURVEY_ID	Survey code
REF_PERIOD	Reference period
MICRODATA	Table with micro-data
BOUND1	Boundary for distinction between precise and less precise data
BOUND2	Boundary for distinction between less precise and imprecise data
SAMPLE_TYPE	Type of sample
THRESHOLD	Minimum number of correspondents in the statistic.
DOM_N	Parameter *n* for *(n,k)*-dominance rule
DOM_K	Parameter *k* for *(n,k)*-dominance rule
NP_N	Parameter *n* for *(n,p)*-rule
NP_P	Parameter *p* for *(n,p)*-rule
HOLD_IND	Holding indicator, the variable which defines associated units
SAFE_MARG	Safety margin for the threshold

3 Output – table with statistics

The main output of general code is the table with statistics and statuses for primary sensitivity and precision. All statistics for all domains are in this table. For now this table exists in SAS environment, but we plan to transfer it to ORACLE environment.

The table has the following structure:

MICRODATA	Table with micro-data
REF_PERIOD	Reference period
DOM1	Variable that determines the first variable in the domain definition
DOM_VAL1	Value of the first domain variable
…	
DOM10	Variable that determines the tenth variable in the domain definition
DOM_VAL10	Value of the tenth domain variable
STAT_CODE	Code of the statistics
STAT_DESC	Description of the statistics
TYPE	Type of statistics
STAT_VAL	Estimated value of the statistic
NO_UNIT	Number of contributions for the statistic
STAT_SE	Standard error of statistics
STAT_CV	Coefficient of variation
STAT_DIS	Estimated value, formatted for dissemination
W_UNIT_NO	Weighted number of correspondents
NO_UNIT_HOLD	Number of units according to the holding indicator
W_NO_UNIT_HOLD	Weighted number of correspondents according to the holding indicator. In case of no holdings, the value is the same as W_UNIT_NO.
PRIM_SENS_CODE	Code from the code list for primary sensitivity (see Table 2.)

For now only primary sensitivity statuses, calculated from our own general SAS code, are added to statistics. Cell suppression is still not automated. The application currently allows the tabulation of data with statuses for primary sensitivity and precision to Excel. That is good first information on information loss and tables' quality concerning definition. The plans for the future are that after adequate results, the user will construct the input files (files with tabular data and corresponding metadata files) for Tau Argus and fill the metadata information for cell suppression in the metadata database. Then the SAS-Tool will be used for cell suppression and final statuses for dissemination will be written back to the table with statistics; only then the final tabulation and dissemination of the data will be done.

4 Conclusion

This application will be a great benefit for SURS; some work has already been done, but a lot of development is still needed, such as writing the missing general SAS code and upgrading the metadata database with tables for confidentiality treatment and publication. After taking care of the "core engine", we will also need to develop:

- The graphical user interface with functionalities of aggregation and tabulation.
- A system for all the corresponding code lists, which are very important for publication and tabulation.

The subject-matter personnel have accepted the new solution openly as they would like to be more independent. We expect some difficulties at the introduction of the application, but our experience has shown very good results so far.

We know that some simplifications will have to be made due to the metadata driven approach and that the best suppression pattern is found with individual approach, but the application will enable better transparency and overview over the whole data production, and consequently the risk of errors will be reduced. We believe that with this application we can relieve the employees of regular routine work, as we are aware of the current financial situation and future policy plans, which involve further personnel reduction and consequently higher workload of staff. If we want to maintain the quality of work on the same level, we need to find a long-term solution. We believe that such a metadata driven application can be a part of the solution.

References

Seljak R., Blažič P.(2011). The 2nd European Establishment Statistics Workshop, Neuchâtel, Switzerland: *Sampling Error Estimation – SURS practice.*

Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Lenz R., Naylor J., Schulte Nordholt E., Seri G., de Wolf P. (2010). *Handbook on Statistical Disclosure Control.*

Brandt M., Franconi L., etc.: *Guidelines for the checking of output based on microdata research.*

Hundepool A., van de Wetering A., Ramaswamy R., de Wolf P., Giessing S., Fischetti M., Salazar J., Castro J. (2011). *Tau Argus User's Manual.*

Giessing S. (2009). *Techniques for Using Tau-Argus Modular on Sets of Linked Tables.*

Giessing S., Schmidt K.. *A SAS-Tool for Managing Secondary Cell Suppression on Sets of Linked Tables by Tau-Argus Modular.*

Seljak R., Smukavec A., Stanek M. (2013). *Aplikacija za agregacijo in oceno standardnih napak.*