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
(Helsinki, Finland, 5 to 7 October 2015)

Topic (v): Practicum: Case Studies and Software

The application for statistical processing at SURS

Andreja Smukavec[*], Rudi Seljak[**]

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

[**] Statistical Office of the Republic of Slovenia, 54 Litostrojska ulica, 1000 Ljubljana, Slovenia, rudi.seljak@gov.si

Abstract: Modernisation of the statistical production systems is demanding and certainly a long-term task, which includes development of the new IT solutions as well as a radical change of the whole production system at the institutional level. In recent years statistical institutions have been putting a lot of their resources into the projects aiming at moving from domain oriented systems to generalised, process oriented solutions.
The Statistical Office of the Republic of Slovenia (SURS) began this modernisation a few years ago by launching a large infrastructure project. The aim of the project has been to develop a so-called metadata driven MetaSOP application for execution of one or more parts of the statistical process (e.g. data editing, aggregation, standard estimation, tabular data protection and tabulation) for most of the surveys. The application has been gradually being developed and the first version of the application was introduced into the regular statistical production in the beginning of 2015.
The paper describes the dilemmas and trade-offs of such a solution, presents the capacity of the application achieved so far with focus on statistical disclosure control for tables and sketches the plans for further development.

1 History

Statistical data processing has always been a demanding, time consuming and consequently very expensive task. A lot of resources have to be spent on data processing, especially for the processes such as data validation, statistical data editing, aggregation and standard error estimation. Furthermore, confidentiality has become a very important issue in recent decades, introducing statistical disclosure control as another very demanding part of the statistical process. On the other hand, there is also

a constant pressure for budget cuts, which is of course in evident contradiction with the above-mentioned demands. Hence, the official statisticians are increasingly facing the challenge of producing confidential statistics of high (or at least sufficient) quality with the significantly reduced resources.

To at least partly reduce the gap between the above-mentioned demands, in recent years a lot of effort has been put into the rationalization of the statistical process. One fact that definitely acts in favour of these efforts is the enormously rapid development in the IT area, meaning the development of hardware equipment as well as the development of a wide range of software tools, which are at disposal to a larger and larger extent. So there is no surprise that also in the area of the official statistics in recent years a lot of effort has been made in the direction of efficient use of all these new tools and applications in order to make the whole production cycle less burdensome and most particularly less expensive.

At the Statistical Office of the Republic of Slovenia (hereinafter SURS) systematic work in this area began some seven years ago when the first prototype system for the modernized data processing was built. The prototype was a metadata-driven application consisting of a few modules which aimed at "covering" the different parts of the statistical process (e.g. data validation, data correction and imputation, aggregation and standard error estimation, tabulation). Metadata were saved in MS Access databases and general SAS macros were written for reading process metadata from MS Access and executing particular statistical processes. It was impossible to control and manage the inserted metadata in MS Access and metadata were scattered all over the network as survey managers were predominantly in charge of filling in and fixing metadata; therefore, SURS decided to establish an internal project in 2012, the aim of which was:

- to build one metadata database for all surveys and all instances in ORACLE environment;
- to create user-friendly graphical interfaces (.NET application) for management of process metadata;
- to connect the system with the metadata repository, where the data on surveys and survey instances are stored.

SURS was aware that "survey-dependent", ad-hoc SAS programs for preparing input microdata tables in proper form are still needed, but the amount of work decreases drastically, as it is easy to adapt and reuse the metadata for next survey instances.

The project was split into two parts: one part focused more on editing and imputations and the other more on standard error estimation, tabulation and tabular data protection. SURS's employees from three different areas are included in the project team: experts from general methodology, IT experts and subject-matter specialists.

The application for imputations and editing was built first and put into production in the beginning of 2015. The graphical interfaces for other statistical processes were parallely built and added to the application. The second part will start with the production in October 2015. The most complicated statistical process was tabular data protection as we added the German application SAS-Tool, developed by the Federal Statistical Office of Germany (Destatis), to our application. The SAS-Tool is a metadata-driven application for tabular data protection of linked tables, where cell suppression method is used.

2 Statistical processes

2.1 Imputations and editing

In the first version of the system for data processing, the input table – which has to be a SAS (work) table – consists of all the microdata to be processed. The output of this module is also a SAS table. Therefore, we always need a small ad-hoc program which transfers the output table back to the microdata database in the case of data editing and imputations. A simplified schematic presentation of the functioning of a module for "data editing" is presented in the following figure:

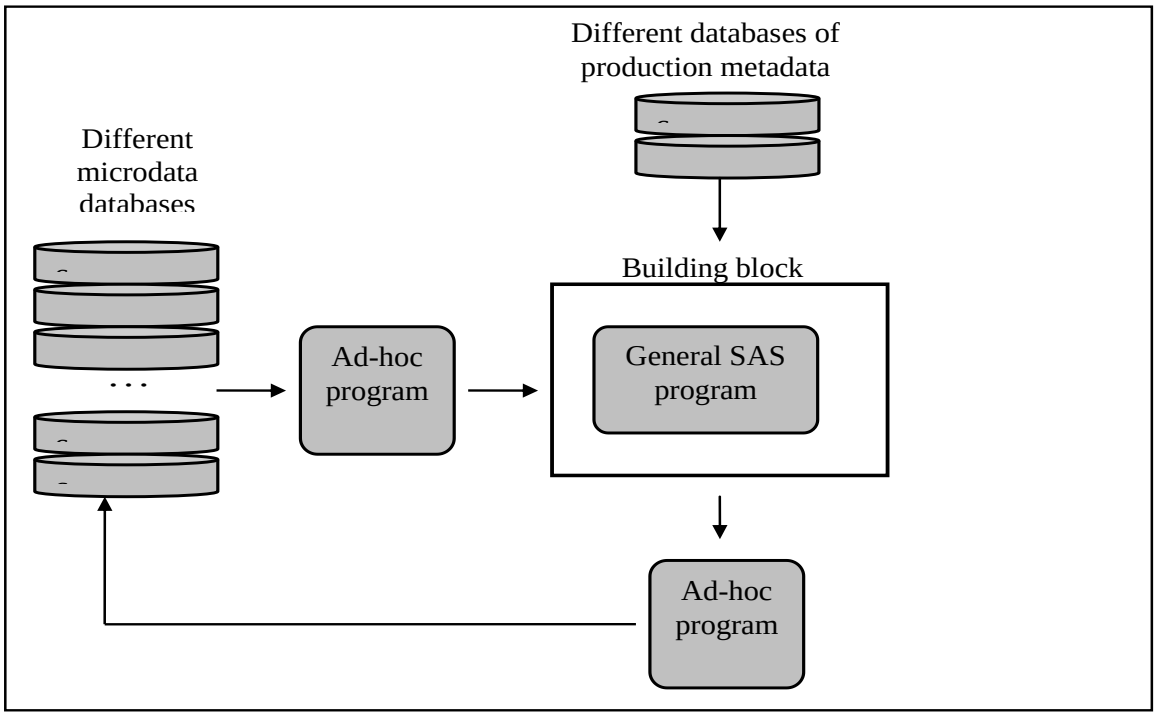

Fig 2.1 Schematic presentation of the functioning of the module for "data editing"

2.2 Aggregation and tabulation

In this paragraph we will focus on standard error estimation and tabulation, with the emphasis on statistical disclosure control for tabular data. The system is slightly different. The input table for the calculation of statistics is also a SAS (work) table, which consists of all final, cleaned microdata; the output is written to one specific database, where all statistics' values with respect to the domains' values are listed. The users have the possibility to choose and calculate seven types of statistics: number and share of units, total and average, ratio of two totals, percentile and modus. Domains can have up to 10 domain variables (comparable to 10-dimensional tables); we added the option to calculate marginal values as they are usually required.

For the calculation of standard errors, we have to set two thresholds (two different values of coefficient of variation), which classify each statistic as precise, less precise or imprecise.

For tabular data protection, we first have to set the safety rules. Users are allowed to set the threshold, dominance rule and p%-rule. It is also possible to set the group identifier, which defines the holding indicator. This option is used very often in business statistics. For example, a number of 'reporting units' (the lower level of unit)

within a cell might belong to an ‘enterprise group’ (higher level), so the level at which the safety rules are applied clearly matters. General SAS macros calculate primary sensitivity also for the cases where sampling weights and holdings are present; of course, with the assumption that the sampling weight for each individual unit is always the same.

General SAS code for the aggregation calculates statuses for precision and primary sensitivity and adds them to the statistic’s value into a special database with calculated statistics for all domain values. After the aggregation we have only primary sensitivity statuses in the database with the calculated statistics, so we cannot yet create tables for publication. As this database is not easily readable and users do not know how many statistics are unsafe or have lower degree of precision, we need to perform tabulation. There are three different types of tabulation possible in the application:

- Excel format (the most advanced option as the user has the possibility to influence the format of the table) is the most user-friendly form;
- plain text format (.csv), suitable for SURS’s publication tool (PX-Edit);
- plain text format (.tab), suitable for the Tau-Argus software (SAS-Tool).

When we perform the tabulation for the first time, we create only Excel format tables, to find out if we need to perform secondary suppression of additional (secondary) cells in the tables. The application returns a warning to inform the user that tables are not safe yet.

In the case of primary sensitive cells in additive tables, we need to apply secondary suppression to prevent the recalculation of the sensitive (primary) cells in additive tables. The tabular data protection for linked tables is done by the SAS-Tool, which calls Tau-Argus for secondary suppression of each individual table. Tau-Argus is the software developed by Statistics Netherlands as part of the CASC project that was partly sponsored by the EU. The SAS-Tool is based on general SAS macros, developed by Destatis. They read the Excel file, where all information (metadata) about tables we want to protect, is saved. The secondary suppression method is then applied on a set of tables in such a way that the transfer of the suppression pattern between linked tables is done with the help of history files. At the end of the procedure, SAS macros check the consistency of the suppression pattern between tables. The SAS-Tool uses a special format of files (.tab), so we added this format as one of the possibilities for the tabulation in the application. An expert for statistical disclosure control for tabular data needs to be involved to prepare the Excel file with necessary metadata, as the SAS-Tool is not a simple tool. One of the reasons is that the tables for publication very often differ from the tables for tabular data protection, which need to be additive in the case of using Tau-Argus, run by the SAS-Tool application. Actually, only statistics “total” and “number of units” are appropriate for Tau-Argus, but these two types of statistics are also the most commonly used for a dissemination. The present version of our application does not apply secondary suppression on tables with other types of statistics, it only calculates the primary sensitivity.

The confidentiality statuses of secondary suppression are added to the database with statistics. After that output tables with confidential pattern and denotations for lower degree of precision are created; they are also suitable for publication. A simplified schematic presentation of the functioning of the module for “tabulation” is presented in the following figure:

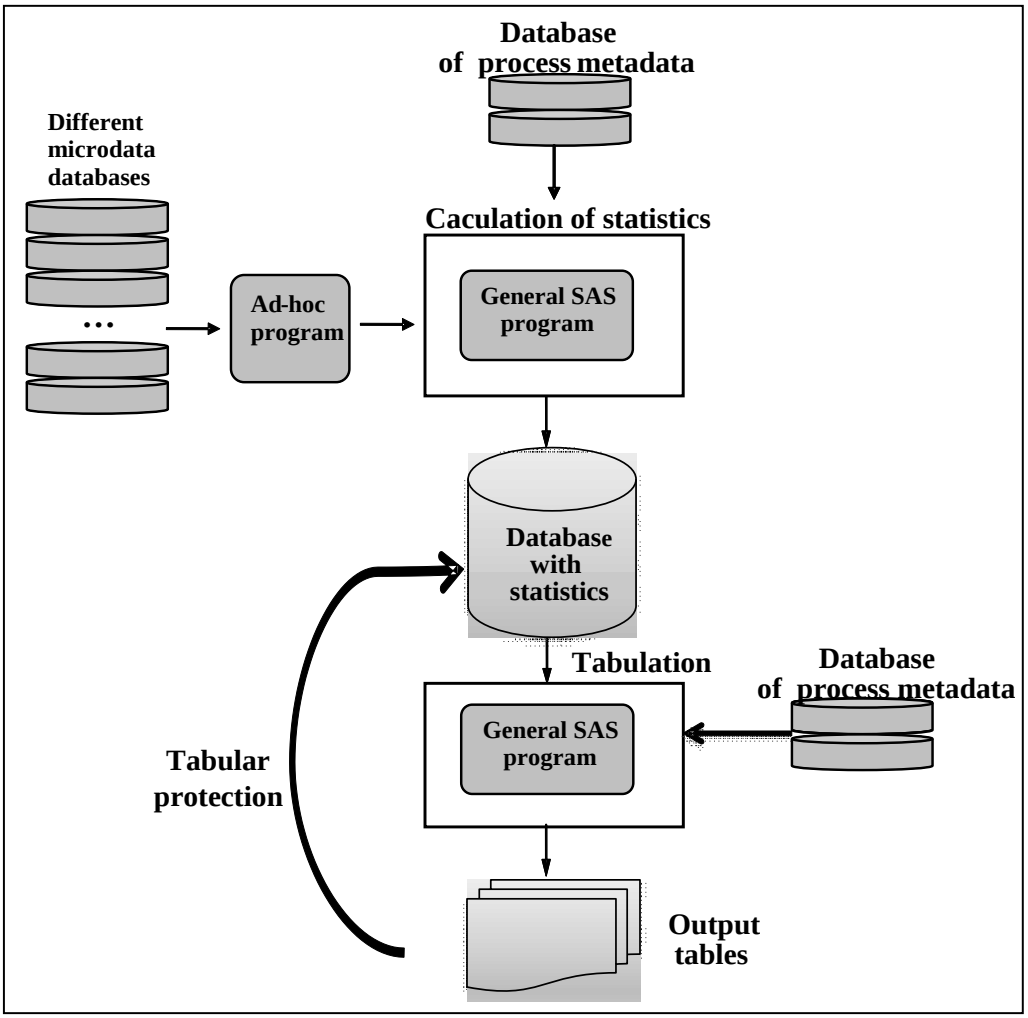

Fig 2.2 Schematic presentation of functioning of the module for "tabulation"

3 MetaSOP application

The basis of the new application for management is the ORACLE database for process metadata, where all the information about editing, standard error estimation, tabulation and tabular data protection for each survey instance is located in one place.

From the user's point of view the .NET graphical interfaces are the central point of the whole system. Through these interfaces the user selects the survey, selects the survey instance, inserts or edits the process metadata and also runs the particular statistical process.

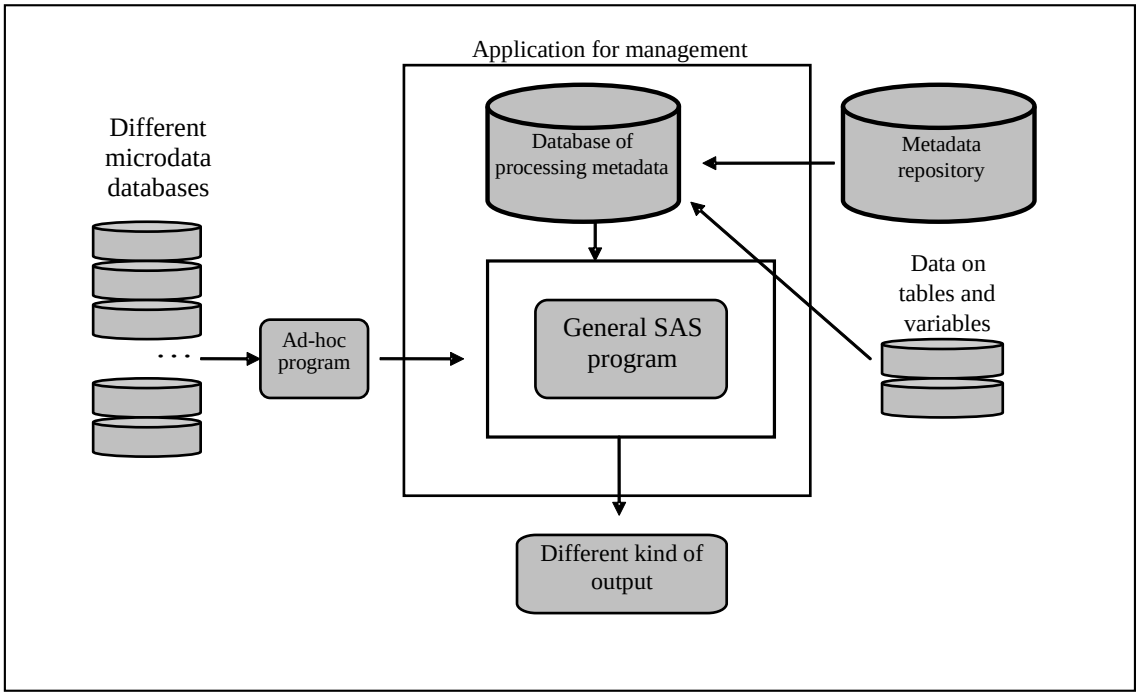

Fig 3.1 Schematic presentation of the modernised system

The application for management of process metadata, called MetaSOP, is hence the key output of the project. It is a Windows Presentation Foundation (WPF) application developed in .NET C# language, for which the user has the same password and username as for logging on the computer. That is necessary as the application accesses also the surveys' microdata databases and the user is not allowed to access microdata databases of all surveys.

Fig 3.2 Starting window of the MetaSOP application

There are quite a large number of the graphical interfaces. They are all written in the Slovenian language; here we present just one, where the user determines the phases (statistical processes) which have to be carried out for the chosen survey instance.

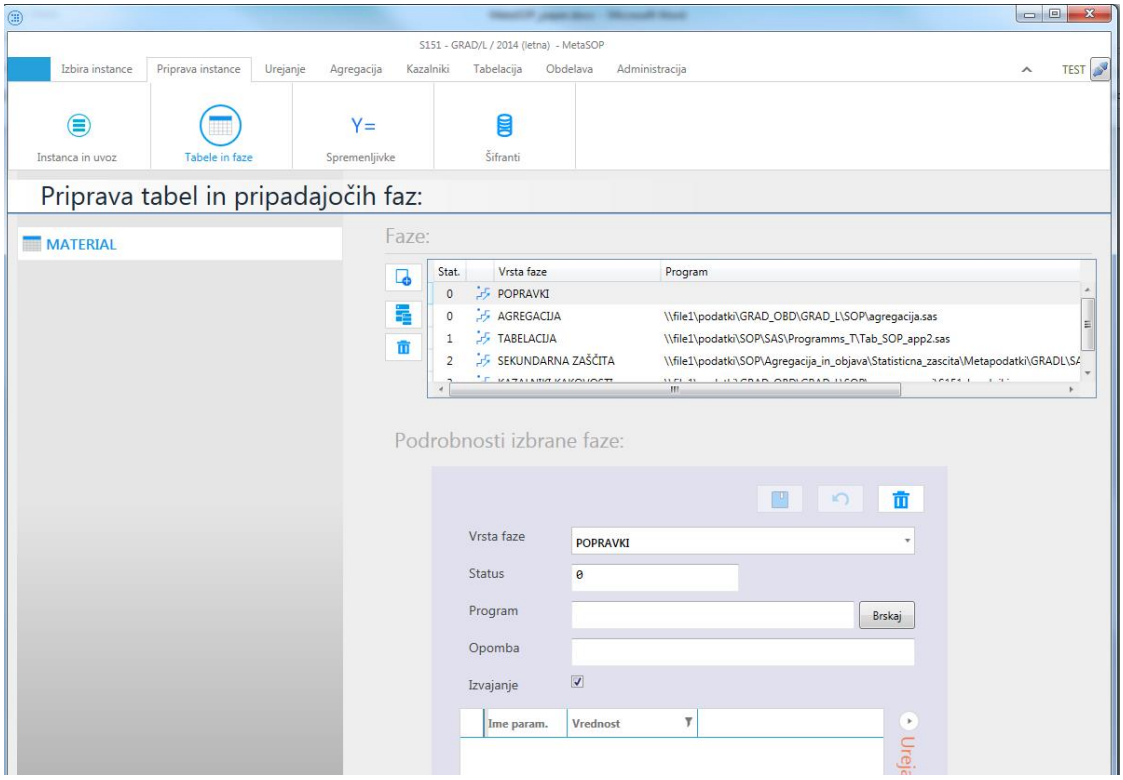

Fig 3.3 Determination of statistical processes for the chosen survey instance

The user can select the following statistical processes: editing, imputations, data validation, aggregation (standard error estimation and determination of unsafe statistics), secondary suppression and tabulation. There is an ad-hoc SAS program needed for all processes. In the case of secondary suppression the ad-hoc SAS program reads the metadata from the Excel file, which is saved in the ORACLE database for process metadata, executes the tabular data protection for linked tables and writes the suppression statuses into the database with statistics. After the tabular data protection is done, the user can create tables with suppressed confidential cells and all necessary labels for imprecise and less precise data. Below is one example of a 3-dimensional table (Excel format) from the construction field, where the MetaSOP (confidential cells labelled "z") was used.

DWELLING TYPE / INVESTOR		TYPE OF WORK					
		TOT	2	3	4	6	7
TOT	TOT	1924960368	902804552	16689032	410198617	13877178	581390989
	1	134320361	66787334	1309641	16976230	1168048	48079108
	2	1790640007	836017218	15379391	393222387	12709131	533311881
1	TOT	657718885	322277342	z	118235160	13877178	z
	1	130153104	65207060	z	z	1168048	z
	2	527565781	257070282	z	z	12709131	z
11	TOT	197382877	98999977	z	19543762	1455782	z
	1	120491019	z	z	z	z	z
	2	76891857	z	z	z	z	z
111	TOT	118363049	65244467	z	z	z	z
	1	100520910	53801983	z	z	z	z
	2	17842138	11442485	z	z	z	z
1110	TOT	118363049	65244467	z	z	z	z
	1	100520910	53801983	z	z	z	z
	2	17842138	11442485	z	z	z	z

Fig 3.4 Excel format of a 3-dimensional table

4 Conclusion

When we started to introduce the new application for editing and imputations to the regular production of the statistical surveys, we regularly collected the feedbacks from the survey methodologists that were faced with the new way of data processing. The main advantages and main drawbacks of the new approach, as perceived by them, can be summarised as follows.

Main advantages:

(a) The subject-matter personnel are much more independent of the IT department, which was previously in charge of the technical execution of the processes.
(b) The rules for the data processing can very quickly be changed through the centralised system of process metadata. This makes the whole data processing cycle much more flexible.
(c) Since the user can run the procedures several times in short time, it is now easier to check the feasibility of different methods for data processing.

Main drawbacks:

(a) In the process of the insertion of the metadata expressions there is a high risk of syntax errors. As the consequence, the application cannot be executed or is executed with the wrong parameterization.
(b) The subject-matter specialists need to learn some new skills (SAS expressions), which is sometimes a problem in the reality of the very burdensome statistical production.
(c) If an error occurs during the execution of the procedure, the technical staff must be contacted and if they are not available, the process execution can stop for some time.

Development of a totally new system for statistical production is certainly a big step forward for SURS. We firmly believe that the project outcomes will help us to build a new, modernised system of management of statistical data processing. Movement from stove-pipe to centralised methodological and IT solution is the final goal of these

developments. The central point of the modernised system is the metadata driven application, which is on one hand flexible in the sense that it can be plugged to different microdata environments, while on the other hand introduces very centralised management of the process metadata.

The first part of the application (editing, imputations) was put into production in the beginning of 2015, while for the second part (aggregation, tabulation) the production is planned in the beginning of October 2015. We have tested our application on some surveys (EU-SILC, CIS, etc.) and it worked well. Of course, we still have a long way ahead of us. We need to train our employees, bugs are expected and plans exist for upgrading the application with additional functionalities, but we believe that the MetaSOP application is a step in the right direction for SURS.

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.*

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.*

Smukavec A. (2013), UNECE Work session on statistical data confidentiality, Ottawa, Canada: *Metadata driven application for aggregation and tabular protection.*

Smukavec A., Seljak R. (2015), UNECE Work session on Modernisation of Statistical Production, Geneva, Switzerland: *Modernisation of statistical processing at SURS.*