UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (iii): Software tools and international collaboration

Usage of external software tools at SURS - experiences and lessons learned so far

Prepared by Rudi Seljak, Andreja Smukavec, Igor Kuzma, Statistical Office of the Republic of Slovenia, Slovenia

I. Introduction

1. In the framework of the modernisation of the statistical production NSIs are in the recent years constantly trying to move their production towards the more wide usage of the generalised software solutions. Development of such solutions is certainly a long-term job, usually associated with large inputs, both in terms of material costs and human resources. It is clear that such development cannot be afforded by every statistical institution and that such development is very difficult or even impossible for small NSIs where constitution of the large development teams can rarely be afforded. In such situations the possibility of usage of the software applications that were developed in another (usually larger) institution is a very tempting option. These applications can be taken and used under different conditions. They can be of commercial nature (bought in the market), free for use or even open source products.

2. It is somehow illusory to expect that the usage of the software that was developed by an external institution will be introduced with no costs. There is always a certain input needed to adjust the sharable application for usage in your statistical production. There is always some kind of a trade-off between the input needed for your own production on one hand and the expected loss of flexibility with comparison to the solution that would be produced in the "domestic workshop" on the other hand. When deciding between two options, own development or sharing, all the pros and cons of both options should be considered and explored.

3. The Statistical Office of the Republic of Slovenia (hereinafter SURS) can certainly be considered as a small institution and therefore it is understandable that sharing of the already developed application is in our office a commonly used practice. In the paper we present some of the most widely used "sharable applications", the practice of their inclusion in our statistical system and the experiences with their usage. We will summarise these experiences in the last part of the paper and draw some general conclusions. We will also present our experiences and considerations with the reverse situation when SURS could provide its own products to be shared by other institutions.

II. Usage of external applications at SURS – few examples

A. Banff

4. Banff is a collection of SAS procedures, developed at Statistics Canada with the aim to support editing and imputation procedures. It is a flexible system that allows users to use each SAS procedure independently or as a linked system. Banff procedures accept inputs that can be created by the user (outside the Banff procedures) or inputs that are created by another Banff procedure. Banff is designed to deal only with the numeric and continuous data and with the linear edit rules. Banff is free of charge

only for Statistics Canada employees. Other organizations may receive a free limited-time evaluation copy or can purchase it. It has very good documentation and the external users can also get the support of the Stat Canada team of programmers and methodologists.

5. Banff procedures offer a wide range of functionalities for the field of Editing and Imputations, like: verification and analyses of edit rules; detection of outliers; error location; imputation of data by using several different methods (deterministic and stochastic). Banff also offers the possibility of mass imputation procedure, which can estimate the values of the whole record by using the donor approach where the donor is chosen from the "clean set of records".

6. SURS started to test possibilities for usage of Banff in 2008, at the same time that development of the general system for edit and imputation started. Although at the beginning also the testing for its usage for the ad-hoc procedures and analyses was carried out, in this way of software usage in fact never really came to life. However, Banff was very successfully incorporated into our general application for editing and imputation, called SOP[1]. As already described in previous papers (see Seljak 2009; Seljak and Blažič, 2011) SOP is a modular, metadata driven application, which is based on a set of SAS macros for data processing and on a set of processing rules (process metadata), which parameterise processing for the particular survey and for a particular survey instance. So far several of these modules, covering different parts of the statistical process (e.g. detection of outliers, deterministic corrections, imputations, aggregation, standard error estimation…) were already developed and are also already introduced into statistical production. Banff is at the moment incorporated into two of these modules: the module for detection of outlier and the module for imputations.

B. Calmar

7. CALMAR is a SAS macro program that was developed by INSEE in the early 1990s and implements the calibration methods developed by Deville and Särndal (1992). The program calculates new sampling weights by using four calibration methods, corresponding to four different distance functions. It is a free software that can be downloaded from the INSEE website. Although it is free, the user doesn't get access to the source code of the SAS macro, but can download only the executable version of it. Since Calmar is a SAS macro, its usage is very much connected to the usage of the SAS software itself. The first condition that must be fulfilled, if someone wants to use this macro, is that the used SAS software package includes the SAS IML[2] module. The user must also be aware that there are different Calmar versions on disposal for different environments. There is for instance a different version for the 64 bit Windows OS than for the 32 bit Windows. Also the SAS version 9.4 requires a completely new version of Calmar that can be downloaded from the INSEE website.

8. With the proper parameterisation the Calmar SAS macro can be used inside any SAS program. Basic inputs for the macro are two SAS datasets. The first one contains micro-data with the initial weights, auxiliary variables and domain variables. The second dataset provides information on margin population values. These values can be population counts for a chosen domain or population totals for the chosen (continuous) auxiliary variables. Totals of the auxiliary variables can only be provided on the level of the whole population in question. The key output of the procedure is the dataset where for each unit the new, calibrated weight is provided. Additional information on the procedure and its performance is written into the SAS output window.

9. At SURS Calmar is used as the main tool when the calibration procedures are in question. We are successfully using it for more than 10 years, so its usage is now a very common practice. It is most commonly used in social surveys where the advanced calibration techniques (e.g. raking) are a more usual practice. It is for instance successfully used in the EU-SILC survey, the survey which is very much dependent on administrative sources and such procedures can significantly contribute to the precision of the statistical estimates. Since the calculation of weights, including calibration procedures, is at SURS at the moment still not "covered" with the general software solution, also Calmar is used mostly in ad-hoc programs, designed for the needs of a particular survey. For the future it is planned to develop a general

[1] SOP is a Slovenian acronym for Statistical Data Processing.

[2] IML is a SAS module, designed to support work with matrices.

solution also for this part of the statistical process (as part of the SOP application) and Calmar will then be incorporated into this application.

C. Tau Argus

10. Tau Argus is a software application designed to protect statistical tables. It is the result of several European projects. Tau Argus was initiated as part of the Fourth Framework SDC-project and in its current form has been developed as part of the CASC project that was partly sponsored by the EU under the contract number IST-2000-25069. The CASC (Computational Aspects of Statistical Confidentiality) project is part of the Fifth Framework of the European Union. The main part of Tau Argus has been developed at Statistics Netherlands by Aad van de Wetering and Ramya Ramaswamy (who wrote the kernel) and Anco Hundepool (who wrote the interface).The purpose of Tau Argus is to protect tables against the risk of disclosure, i.e. the accidental or deliberate disclosure of information related to individuals from a statistical table. This is achieved by modifying the table so that it contains less detailed information. Partly funded by Eurostat, a project was started in 2012 to rewrite at that time the most recent version of τ-ARGUS to an open source version with the possibility to be run on a Windows platform as well as on a Linux/Unix platform. The cell suppression, CTA and rounding methods are able to use Open Solvers as in the original version Tau Argus uses payable LP-solvers for cell suppression. SURS intends to introduce the open-source version into its production in the next 2 years.

11. SURS uses Tau Argus for more than 10 years. At the moment it is mainly used by statistical disclosure control experts. In order to reduce the SDC experts' workload in some cases the Tau Argus is also run also by survey managers. This practice is mostly implemented in cases of few and simple table sets. To support the survey managers, thorough and very detailed instructions are written in the Slovenian language. To widen this practice of transition of the operative management of Tau Argus to the subject-matter side, SURS started an internal project a few years ago. The aim of the project is to implement the metadata-driven procedure SAS-Tool for protection of linked tables into the statistical process. The metadata-driven procedure SAS-Tool is an application of Destatis (German Federal Statistical Office) and is a combination of SAS macros and Tau-Argus. As metadata entry requires a lot of knowledge on tabular data protection, the metadata are prepared by SDC experts, but for the next recurrences of the survey the application is run by survey managers.

D. Demetra

12. Demetra is a software tool aiming at supporting the seasonal adjustment procedures. The first version of Demetra was developed by Eurostat at late 90s. As it turned out, this tool is not flexible enough to assure sufficiently comparable results among the EU member states, it was decided that a new tool, Demetra+, should be developed. The National Bank of Belgium developed Demetra+ at the request of Eurostat. The Eurostat-European Central Bank (ECB) high-level group of experts on seasonal adjustment steered the development work. In 2009, the same group produced the European Statistical System (ESS) Guidelines on Seasonal Adjustment. The aim of creating the software was to offer an open-source tool that reflects the ESS Guidelines. The first version of Demetra+ was not an open-source, later JDemetra+ was written in Java and is available in open source form. The modules, which form JDemetra+, can be adapted by IT teams or new modules that can be plugged into the original software in a transparent way can be developed.

13. All described applications offer two different methods for seasonal adjustment, TRAMO/SEATS[3] and X-12-ARIMA[4], which are the two most commonly used methods for seasonal adjustment. Demetra+ software doesn't suggest a particular guided process for seasonal adjustment but offers several options for its users.

14. Demetra+ was incorporated into the statistical process of SURS two years ago. Before that the old version of Demetra was used for seasonal adjustment at SURS for more than 10 years. The whole time since seasonal adjustment has been incorporated in the statistical process at SURS, survey managers

[3] TramoSeats is a seasonal adjustment method developed by Victor Gomez and Agustin Maravall (Bank of Spain).
[4] X-12-Arima is a seasonal adjustment program developed by the US Census Bureau.

are responsible for running regular production each month or quarter, depending on the specificities of their time series. General instructions and guidelines for working with Demetra+ were prepared, where all possibilities are described, e.g.: if bad statistics occur or temporary outliers are set for some time series, the models are sent to the experts for seasonal adjustment, who check and change the models if necessary. Following the Guidelines for seasonal adjustment, the annual reviews of the models are necessary and are done by experts for seasonal adjustment for all the time series, defined at SURS. Introduction of the open-source software JDemetra+ into the statistical process at SURS will be carried out in the next couple of years.

III. Experiences with usage of the shared application

15. As described above, SURS have long years' experience of usage of different kinds of shared software tools in its production. Due to the diversity of its characteristics and its role in the statistical process it is difficult to provide some feedback on a general level. What follows is hence more a "potpourri "of reflections from the experiences of usage of the particular applications, but which could still indicate also reflections on a more general level.

16. Banff is certainly a product that has a very comprehensive, well organised and well-structured documentation. Since Banff procedures can be used as any other SAS procedures, for a "SAS oriented" organisation as SURS is, it is a very flexible and easy to use programming tool. There were also some attempts to introduce Banff as a tool for the end users (subject-matter statisticians), but it turned out that its functionalities and many features are a bit too advanced for most of these users. Therefore we decided to keep it mostly as a programming tool, especially as a very important part of a general application for editing and imputations. So in fact the end users are using Banff, but through another application that makes (some of) the functionalities of Banff easier to use for a wider group of users. To make Banff functionalities really applicable for all the surveys, few adjustment were needed. The most 'problematic' was the fact that the Banff procedures are designed only for numerical variables. Therefore some of the procedures needed certain adjustment when they were incorporated into a general application. This was, for instance, the case with the imputation procedures (e.g. procedure for donor imputation), which is in fact the key Banff functionality that is used in a general solution. In order to really support all the survey needs, some additional programming of these procedures was carried out.

15. Calmar is also a SAS based product (SAS macro) and as such very easy to use in our production. It is designed in such way that it is quite easy to include it in any SAS program. Of course, input data sets have to follow the prescribed structure, but it is quite straightforward and easy to implement. The main problem of Calmar is the lack of really good and comprehensive documentation in English. Also the messages that appear in case of errors are in French, but this can be now, thanks to better and better computer translators, easily handled even for non-French speaking users. We also encountered some problems with the usage of the inappropriate Calmar version (e.g. when moving to a 64 bit version of Windows). Namely, in such cases the program doesn't provide the message that would give you information on the wrong version, but only provides the SAS code error message. But when the problem is correctly recognised, it can very easily be solved, since INSEE provides all the needed versions on their webpage.

16. Tau Argus was in our office not recognised as a tool that could be used by the end user and was so far mainly, with few exceptions, used by statistical disclosure control (SDC) methodologists. Especially in the first years of usage it had a lot of bugs and crashed many times during the execution. But anyway, for the SDC specialist it is still quite easy to use the tool. The biggest challenge is usually to define all the links between tables and apply appropriate protection to the set of tables. The latest developments, especially the inclusion of the above mentioned SAS-Tool, automated tabular data protection and simplified the work with tau Argus to that level that, at least for less complex tables, it can be carried out by the survey managers.

17. When the usage of Demetra is in question, there is a big difference in its usage in the first years and recently. In the first years of usage the software still contained many bugs and deficiencies. Through the usage of many different users, their feedback and especially their complaints the software gradually improved and it became much more reliable. In general it could be stated that Demetra+ is much more

user-friendly than Demetra and also most of the bugs from Demetra were removed in Demetra+. On the other hand Demetra+ is very different in comparison to Demetra, therefore its introduction represented a huge change for the final users and a lot of education was needed. But again on the other hand this change offered a perfect chance for the renovation of the seasonal adjustment process at our organisation. It's also worth mentioning that SURS is a part of The Seasonal Adjustment User Group (SAUG) that will collaborate with the Seasonal Adjustment Center of Excellence (SACE) to assure the quality and the promotion of JDemetra+, (JDemetra+ is now officially recommended as the software for the seasonal adjustment of official statistics by Eurostat).

IV. Sharing of SURS products

18. Although SURS can by any standard be considered as a small institution and its possibilities of own developments are quite limited, there are few cases where our products could be shared with other institutions. One of such examples is the STAGE application.

19. STAGE is a Web-GIS application for cartographic visualisation and dissemination of geospatial statistical data, i.e. statistical data integrated to spatial information. It consists of a web mapping application and a download service linked to the spatial database. STAGE is designed as an INSPIRE compliant application regarding the network services and the metadata descriptions. STAGE is based on the Google Maps platform. The application enables the user to present statistical data while selecting the spatial level, content and time reference. In STAGE geospatial statistical data for over 300 statistical variables in variously available time series are presented for up to 10 levels of geospatial data, i.e. from administrative units (cohesion and statistical regions, municipalities and settlements) down to grid cells of 10km x 10km, 5km x 5km, 2.5km x 2.5km, 1km x 1km, 500m x 500m and 100m x 100m size.

20. STAGE stands for statistics & geography but figuratively it also suggests the play of the world being monitored, interpreted and visualised by statistics and further communicated to the experts or general public. The wide range of statistical data presented on various levels of spatial units makes the STAGE a powerful tool for monitoring the past development or a particular phenomenon and suggesting the future trends. Development of Stage of co-financed by Eurostat and it was developed in the cooperation with Geodetic Institute of Slovenia. The application can be accessed through the website: http://gis.stat.si/en/.

20. The open data policy of the STAGE makes no restrictions on the data applicability and favourites no user group in their advocating any political, social or business views. Experts or the general public can thus easily evaluate the implementation of various national, regional and local policies. By opening the geospatial data infrastructure, SURS follows its commitment to disseminate statistics with direct impact on effective operating of the private or public sector and to improve the geospatial statistical literacy of the users. To enable wider application of the STAGE or its components, SURS decided to share the application under the European Union Public Licence. This makes the STAGE an open source product that can be used by any user.

21. Another SURS product that was recently developed and that could also be interesting for external users is the above mentioned SOP application for statistical data processing. SOP was developed inside SURS with certain financial support of Eurostat in last stage of development. The basic idea behind this project was the strategic, gradual transition of our institution from the stove-pipe oriented production, where IT solutions are customised for the purposes of the certain survey, towards the more centralised production, based on the "process oriented" IT products. Such transition is only possible if the proper general IT solutions, covering certain parts of the statistical process, are in place. Development of such generic solutions was in fact the first goal of our project and such solutions were for several statistical processes developed at that stage. These solutions are in fact SAS macros, which are developed on the basis of the metadata driven principle and could be through the set of process metadata easily parameterised for the needs of different surveys. At the second stage these basic solutions were then linked into a more integrated system, designed to be used by the final user, which is usually the subject matter statistician.

22. The SOP application can hence be seen as a three-layer application. The first layer consists of the set of (metadata driven) SAS macros that are the core of the whole application and take care of the actual data processing. The second layer presents the process metadata database, where all the process rules for all the surveys and all survey instances are stored and maintained. The third layer consists of the (.Net) graphical interfaces, which enable easy and user friendly insertion and editing of process metadata and also running of the statistical process itself.

23. When the possibility of sharing of the SOP application was internally discussed in our office we considered three options:

- Sharing only the conceptual design and conceptual architecture of the application.
- Sharing only the SAS macros.
- Sharing the whole application.

24. For each of these options there are several issues to be considered. The first option is probably the least “problematic” from our side. What is still needed, that such sharing would be possible, is to improve the documentation of the whole system and especially the documentation on the basic concepts that the application is developed on. Otherwise comprehensive and detailed documentation exists, but is at the moment available only in the Slovene language, therefore translation of this documentation into English would be required.

25. The second option seems also still quite feasible from our side. The SAS macros that would be shared are in fact developed in such a way that could be easily adjusted for usage in another institution. The basic condition for this is of course that this other institution uses SAS and in cases of some SAS macros also Banff. Also the technical and methodological documentation on these solutions should be a bit supplemented and translated. It also has to be mentioned that in the case of sharing it would be from our side difficult to ensure regular methodological and technical support.

26. The third option would have otherwise offered the most, but is also the most problematic from the point of view of portability. Some of the issues that come with this option are:

- Contrary to the above mentioned SAS macros, the input that would be needed to adjust the application for the specifics of another environment would be much higher, especially because of the following:
    - The database of process metadata is linked with some of our infrastructure databases (e.g. Metadata repository) that are quite specific and would require a considerable amount of adjustments.
    - The management of user rights is customised to our system of domain users.
    - All the graphical interfaces are designed in the Slovene language. Additional translation would be needed also for this element of application.
- Documentation for the whole application is much more comprehensive than the one for the SAS macros and its translation and adjustment for the external user would also require a considerable amount of time.
- In case of sharing, the external user would probably, at least in the initial phase of usage, need a lot of help and support. This would be difficult to ensure with our limited resources.

V. Conclusions

27. Usage of the software applications that were developed by another institution(s) has become a more and more common practice in the statistical community in the recent years. This is partly due to the fact that for many of the institutions (especially for the small ones) only by such software sharing they can fulfil the more and more pronounced needs for modernisation and harmonisation, and partly also due to the more and more successful international cooperation in this community. These products can be of different nature (can be purchased on the market, can be free for use, open source), determining the conditions of their usage. They can be developed in different environments (SAS, R, Java…), often determining their flexibility and width of usage. Furthermore they can be produced through a very different development process. They can be the result of a planned strategy of a certain institution to provide a (commercial) product for wider use; they can be the result of large international projects (e.g.

"FP projects"); they can simply be the result of the development in a certain institution and later the decision that the developed product can be shared with other institutions.

28. SURS has now already a long-term practice of using different kinds of "externally developed software applications" and many of them have successfully been included into the regular statistical production. Based on this long-term practice, the following main advantages and benefits of software sharing can be pointed out:

- Significant reduction of the development resources that would be spent for the "home made products".
- Usage of more advanced statistical methods that would (with our limited resources) otherwise be very difficult to implement.
- Usage of harmonised procedures and methods. This is especially true in case of products that were developed through the joint projects of several NSIs and other statistical organisations.

29. Despite the undoubted benefits of software sharing there are also a few shortcomings that should also be considered in view of such usage. By our experiences these can be summarised as follows:

- The flexibility of the "external application" is usually lower than it would be in case of the "own development".
- There is always certain input needed to adjust the "external application" for the usage in our own production.
- The input needed for adjustment is very different for differently shared applications and is usually connected to its "origin reason".
- The product documentation can sometimes be a weak point and can hinder its usage and implementation.

30. SURS also has some experience with the reverse situation when the possibility of offering our products to the wider statistical community was considered. Despite the fact that generally we certainly are in favour of such practice, there are several issues that are always on the table when this option is considered:

- Although there would be no need for the adjustment of the product itself, there is a significant amount of work that should be adjusted for the "external users".
- Installation of the software application to another environment would certainly require certain support from our side and this can be problematic due to our very limited (human) resources.
- The software applications, at least in the first phase of its usage, still contain certain amount of deficiencies and hidden errors. It is difficult to ensure sufficient support for the external users in case that such an error occurs.

References

Banff Support Team: Functional Description of the Banff System for Edit and Imputation System, Statistics Canada, Quality Assurance and Generalized Systems Section Technical Report.

Eurostat (2014): The ESS Vision 2020; Available at: http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/8d97506b-b802-439e-9ea4-303e905f4255

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. (2009), “New Application for the Slovenian EU-SILC Data Editing”, Presented at the UNECE Work Session on Statistical Data Editing, Neuchatel, Switzerland, 5-7 October, 2009

Seljak, R., Blazic, P. (2011), “Sampling error estimation – SORS practice”, Presented at the 2nd European Establishment Statistics Workshop, Neuchatel, Switzerland, 12-14 September, 2011

Stage application; Available at: http://gis.stat.si/en/