

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (ii): Generalized software packages for statistical data editing, their evaluation

**DOWNLOADING THE DATA EDITING PROCESSES FROM THE MAINFRAME TO
CLIENT/SERVER PLATFORM**

Submitted by the Statistical Office of the Republic of Slovenia¹

Contributed paper

I. INTRODUCTION

1. The Statistical Office of the Republic of Slovenia (SORS) is a centrally organised government institution. It carries out more than 300 regular and development surveys according to the National Program of Statistical Surveys. Most of the statistical production is carried out on a mainframe platform, provided by the Governmental Centre of Informatics (GCI). Details on SORS technological infrastructure are given in Katic and Kozjek (1997). There are several reasons why SORS decided to develop a step-by-step migration project, which should end up with final migration of existing surveys and registers from mainframe to user-friendly client/server based environment. The most important reasons are:

- better accessibility to data,
- increased efficiency,
- better quality of statistical products,
- lower production costs,
- timeliness,
- technical independence of SORS from the Governmental Centre for Informatics (GCI).

2. SORS has already developed in-house competence for some PC LAN-based systems, like Oracle, SAS, TPL, Blaise and MS Office. A statistical information system with a "backbone" consisting of these components proved to be suitable for the step-by-step migration to the client/server (C/S) environment. In the existing environment there are a lot of different operation systems, tools and programs used and there is an increasing demand for the standardisation of processes. The feasibility study performed by experts from SCB - Statistics Sweden and SORS (SORS and SCB, 1997) recommended that the number of tools to support the processes on a new platform should be decreased in order to enable and facilitate standardisation.

II. TARGETS OF MODERNISATION

3. During the implementation of the migration project, several processes will be going on simultaneously. These overlapping processes will end with the following goals achieved:

- SORS is functioning as a user-oriented office
- all statistical registers and all main surveys at SORS are migrated to a client/server environment

¹ Prepared by Pavle Kozjek.

- more efficient statistical methods are used; most full-coverage surveys are replaced by sample surveys,
- quality of all parts of statistical production is better, with more efficient monitoring,
- technical and methodological standardisation.

Based on the outcome and findings of the feasibility study, the MGS (Modernisation of Government Statistics) program was launched, following the proposed strategies and guidelines to reach the goals.

III. DATA EDITING AT SORS: CURRENT SITUATION

4. Data editing is a relatively important part of SORS statistical production process. We have a good basis to become a register-oriented statistical office, but currently there is still a large amount of raw data collected directly from respondents. There are some (deterministic) imputations made in the phase of data collection and capture editing, but in most cases imputation processes are separated and done later, using the SAS system.

5. Most data editing processes are carried out on the mainframe platform, where there are three main types of data editing applications:

- Cobol batch programs
- PL/I batch programs
- G-VS (Godar-Vega Stat) applications for interactive data editing

All batch programs represent the old-fashioned cyclic data editing, producing lists of errors, which are manually corrected and re-entered into the mainframe file. The main deficiencies of mainframe batch data editing applications are:

- time consumption
- process is "isolated" (from data source and from the internal data end-users: statisticians, subject-matter people)
- no feedback (just error corrections)
- overediting
- problems with documentation and maintenance.

6. Interactive G-VS system is more effective and successful tool, and was relatively well accepted amongst data editing staff and application developers. It is based on Rapid RDBMS and had introduced some important standards to SORS data editing processes. But the system can not be transferred to the C/S environment and there's a year 2000 problem, which is still not completely solved. The system support and maintenance are very limited.

7. On the other hand, SORS is already using a system that supports data editing on the C/S platform. The Blaise system for integrated survey processing is currently used mainly for household surveys. Details on system are given in Bethlehem (1995). In most cases, data entry and editing are integrated in computer assisted processes, and often the multi-mode approach is used (combinations of CAPI, CATI and CADI). More details on integrating survey processes with Blaise and multi-mode surveys are given in Heuvelmans, Kerssemakers and Winkels (1997), Schou (1995) and De Jong (1996). Although there are some problems (data transmission, in-house management), significant improvements in timeliness and data quality are obvious with CAI surveys.

8. In general, the data editing system at SORS currently does not meet the requirements of a modern and efficient system, and data editing is one of the important issues of the MGS program.

IV. TRANSFERRING THE DATA EDITING PROCESSES FROM MAINFRAME TO C/S ENVIRONMENT

9. The MGS project plans a step-by-step movement towards PC-based statistics in a C/S environment, including data editing processes. Our main goals in a new environment are:

- better efficiency and functionality
- integration in the survey processes
- satisfied users
 - application developers (development and maintenance)
 - data editors (operational work)
 - end users of 'clean' data sets
- safety of data.

Some surveys, mainly from the social-demographic area, are already supported by the Blaise applications, running on LAN in a mixed DOS/Windows environment. They only need to be completely transferred to the NT platform. But greater number of surveys are depending more or less on a mainframe environment: large and complex economic surveys, time series and longitudinal surveys, using registers, databases and reference files, maintained on a mainframe.

10. The question is: can the editing of those surveys' data be shifted to LAN and supported by the Blaise system as well, or do we need to look for a complementary solution when defining new editing processes on a C/S platform? To answer this question, one of the MGS pilot projects is explicitly focused on defining data editing processes in a new C/S environment. The monthly Industrial Production Index Survey was chosen as a typical representative of a complex economic survey. Blaise 4 Windows is used as a data editing tool, running on a LAN Windows NT platform, and Visual Basic as a system to build the end-user interface.

V. PILOT PROJECT ON DATA EDITING

11. The monthly Industrial Production Index is one of the important surveys at SORS. It's a longitudinal survey with complex checking rules, different reference files, and demanding survey management and administration. The development of a pilot data editing application began in June 1998, using prototyping methods and JAD (Joint Application Development). The first prototype of the new application was already presented, discussed and evaluated by developers and end-users, and a short analysis was done to define further development.

12. A similar approach was used with first more complex household surveys at SORS, where the pilot projects were considered as a good indicator for general development. An important point is that the good cooperation between EDP and subject matter staff was continued later, with the maintenance of applications.

13. The main criteria for evaluating how the new approach meets the objectives are:

- Better efficiency of the new system. At the first phase it could be evaluated by comparing the old and the new (pilot) data editing application of the same survey.
- Quality of final data sets.
- Integration of new applications into the SORS information system.
- User orientation: rapid, standardised and controlled development and maintenance of applications (application developers); easy and user friendly work with applications (data editors); well-defined and documented complete data editing process (statisticians).
- Data safety and security. Mainframe safety measures cannot be directly copied to LAN, so the new LAN system for data backup and archiving should prevent the loss of data. Access should be restricted only to authorised users to maintain data confidence.

14. First an evaluation was done with the analysis of the prototype of the pilot project. We expect to obtain a more detailed view of different aspects of the new approach with the implementation of the pilot project. Parallel production work of old and new applications should clearly show the advantages and

disadvantages of the new solution, and feedback from all involved staff will be the key for the final evaluation.

VI. INITIAL OUTCOME AND EVALUATION OF THE PILOT PROJECT

15. The first prototype of the Industrial Production Index survey was presented, discussed and evaluated by representatives of departments involved in the data editing process. In the first phase, development was focused on a data editing application and its functionality. User interfaces and integration into the C/S environment will be discussed in the next phase. After the first analysis of the situation, two main conclusions were pointed out:

- From the functional point of view, the Blaise system is able to replace the old data editing applications. All the important features of the old system can be maintained, and some new ones added.
- Replacing the mainframe editing system (batch applications and G-VS system) with Blaise on a C/S platform is much more than just replacing production tools - it means a change of platform and complete environment, with all its complexity.

16. According to these findings, our previous experience with the Blaise system and other users' experience, the Blaise system is remaining a serious candidate as a general data editing tool on a C/S platform at SORS. But implementation should be performed step by step, considering the complex problems of platform change. The major issues of building a new data editing system on a new platform seem to be:

- **In-house survey management and administration system.** Blaise itself does not provide the end-user menu system, to manage and control the editing process without knowing the LAN PC environment. There is the separate Maniplus system available to help administration, but we already started to build our own system, using Visual Basic.
- **A LAN archiving and back-up system, which should be an adequate replacement for the mainframe HMS system.** It is not an easy task to provide network reliability that could be compared to the mainframe systems.
- **Need for major reorganisation.** In some surveys, data entry and editing will be integrated; in others, new methods and techniques of data entry will be used, like EDI and scanning (instead of high-speed data entry). All these processes should be reorganised and integrated in the SORS information system.
- **Education.** Many SORS employees are used to work on a mainframe only; so basic PC education would be necessary to begin in a C/S environment. There is no company in Slovenia where Blaise courses are organised, so we have to build education mainly on our knowledge and internal courses.

17. The above-mentioned problems are more related to the general question of changing platform than directly to the choice of the data editing system. However, the decision about the editing system on a new platform is extremely important. There are reasons why to continue with Blaise and extend its use:

- in-house knowledge and experience;
- efficient, standardised and cost-effective solutions with methodological background;
- one system used for all data editing (on a micro level);
- data entry and editing integration.

But there are also hesitations and reasons against general use of the system:

- Blaise (and its database) is above all designed for interactive data entry and editing processes;
- compatibility with the other systems (no ODBC yet);
- lack of built-in administration possibilities;
- parts of the system are not user-friendly.

Most deficiencies seem to be removed in the next version of the system, and the rest depend very much on our own approach to the internal organisation of processes. Visual Basic seems to be suitable to build the systems (shell) around data editing applications, and development is continued.

18. Bearing in mind the complexity of changes, and knowing that the majority of employees in the SORS input division has little or no experience working with PCs, the process of transferring data editing to a C/S platform has to be executed carefully step by step, maintaining the existing applications until they are

replaced by reliable new solutions. G-VS system will still be maintained, and old batch data editing applications will be modified and adjusted to work for the time necessary.

VII. CONCLUSION

19. Since the pilot project, the Industrial Production Index survey contains the main critical aspects of generalising the data editing system to a wide range of different surveys; it is expected to be a good indicator to enable decisions about further development. If we get what we need, the development should continue the same way and increase. On the other hand, an additional complementary solution will be necessary to support data editing processes on a new platform. There are other systems on the market, but few of them are specialised to support different data editing processes, and we don't have very much knowledge or experience about them. Adding a new system would also increase the number of software systems at SORS (and the number of licenses to pay), and increase standardisation problems. Therefore, if we meet some special needs, it would probably be wise to look for some in-house development and modifications of existing tools, to obtain smooth and efficient data editing processes for all kinds of data.

20. In any case, the pilot project on data editing will add new knowledge to survey processing at the Statistical Office of the Republic of Slovenia and perhaps could also be useful to the other institutions confronting the same problems.

REFERENCES

- Bethlehem, J., A Control Centre for Computer-assisted Survey Processing. Report, Voorburg, Statistics Netherlands, 1995
- De Jong, W.A.M., Designing a Complete Edit Strategy; Combining Techniques, Research paper no. 9639, Voorburg, Statistics Netherlands, 1996
- Heuvelmans F., Kerssemakers F., Winkels J., Integrating Surveys with Blaise III, Proceedings of the Fourth International Blaise Users Conference, Paris, 1997.
- Jenkins, A. M., Readings in Research in Management Information Systems. Baltimore, MD: Information Systems Research Center, University of Baltimore, 1998.
- Katiæ, M; Kozjek, P., Technological infrastructure used for data editing at Slovenian Statistics, Statistical Data Editing - Methods and Techniques, Volume 2, UN - Statistical Commission and Economic Commission For Europe, New York and Geneva, 1997.
- Lundeberg, M., Sundgren, B., Advancing Your Business: People and Information Systems in Concert, EFI, Stockholm School of Economics, 1996.
- Schou, R., Developing a Multi-Mode Survey System, Essays on Blaise, Third International Blaise Users' Conference, Helsinki, 1995
- SORS and Statistics Sweden, Feasibility study on the architecture of information systems and related equipment issues, Study results, September