WORKING PAPER No 8

ENGLISH ONLY*

**STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Joint UNECE/EUROSTAT/FAO/OECD Meeting on Food and Agricultural Statistics in Europe**
**(Geneva, 2-4 July 2003)**

**STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

**FOOD AND AGRICULTURAL ORGANISATION (FAO)**

**ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD)**

# AN INTEGRATED TOOL FOR MANAGING AGRICULTURAL SURVEYS AND DATA: OUR BEST PRACTICE APPLICATION

Supporting paper submitted by ISTAT, Italy**

[*] Paper posted on Internet as submitted by the country.

[**] Prepared by Ms. Aurora De Santis, Ms. Federica Piersimoni and Mr. Stefano Salvi, ISTAT, Italy.

## Abstract

1. The purpose of this paper is to show the effort produced by the Italian National Institute of Statistics (NSI) in improving efficacy and efficiency in data production of a complex sampling structure surveys conducted to gather and manage information about the agricultural sector. In order to achieve this, two projects have been started. The first one, AGAIN, is an user-friendly, flexible and timesaving tool set up using a SAS/AF® module (release 8.00). It ensures high interactivity and high-resolution graphic images. It also enables to support estimation, editing and report on quality indicators. At the moment agricultural surveys implemented with an AGAIN system tool are: monthly and annual slaughtering of red meat, monthly slaughtering of poultry meat, fishery, monthly and annual surveys on milk dairies and fruit and berry plantations surveys. A new version suitable for surveys collecting a lot of variables (e.g. survey on the structure of agricultural holdings) and allowing more people to work on the same survey data at the same time has been released. For this reason it is considered as a solution of old problems in sample design and data production such as the use of auxiliary information and the detection of possible errors in the raw data. The second project, called ASIA, is a relational data base containing all data collected by Italian NSI on agricultural holdings since 1990. This database will became the base of a data warehouse on the Italian agricultural sector.

Keywords: data warehouse, data processing, system architecture

## I. INTRODUCTION

2. In the last decades, agriculture has covered a marginal role in the Italian economy giving a little contribution in terms of the added value of the Country. Recently, phenomena such as the ESB and the foot-and-mouth disease, or the production of the genetically modified organisms (GMO), have recently caused a new interest in agriculture sector: it does not represent only an economic sector but mainly the origin of the food chain.

3. Not only the relationship between agriculture and environment but also the recent evolution in the concept of quality in statistics (Eurostat 2000), produced a critical analysis on the national system of agricultural statistics in particular from a methodological, managerial and technological point of view.

4. With regard to the methodological revision, a crucial choice has been the adoption, where possible, of the sampling scheme instead of the census one.

5. In addition it has been introduced the Computer Assisted Telephone Interviewing (CATI) technique for some surveys (on milk and diaries, on slaughtering and on the use of pesticides) traditionally conducted by mail. These innovations provided very good results in terms of human, material and financial savings and in terms of response rate and data dissemination timeliness.

6. Concerning the managerial aspect, a great investment has been done for rationalizing statistical production in the agricultural sector encouraging the use of administrative data. The integration of administrative archives with the Statistical Archive of Active Farms (ASIA - Agriculture) is one of the most relevant tasks in this context.

7. Technological innovations have also been introduced in the new system of agricultural statistics in terms of use of modern communication and standardized data processing tools. Since 2001 the Agricultural Service has activated an his own web-site entirely dedicated to the produced agricultural statistics (De Santis 2001).

8. In the same period an interactive, user-friendly and flexible computerized system, named AGAIN (Analysis and Automatic Management of Surveys), has been set up.

9. In the following paragraph the architecture of the system we are moving towards will be presented.

## II. THE "DESIDERATA" ARCHITECTURE

10. The underlying idea to the scheme above presented, is that, nowadays, a complete integration of all the phases of a statistical production process is a pre-requisite to guarantee a high level of quality, in terms of products and services, of the obtained results.

11. Each module is strongly correlated with the others, and they all together contribute to the application of principles of good practice in the statistical production process.

12. The functions that will be described are mainly two: the data processing and the data warehousing.

13. To start with, an overview about some advantages of a computerized and standardized system able to support data processing will be given.

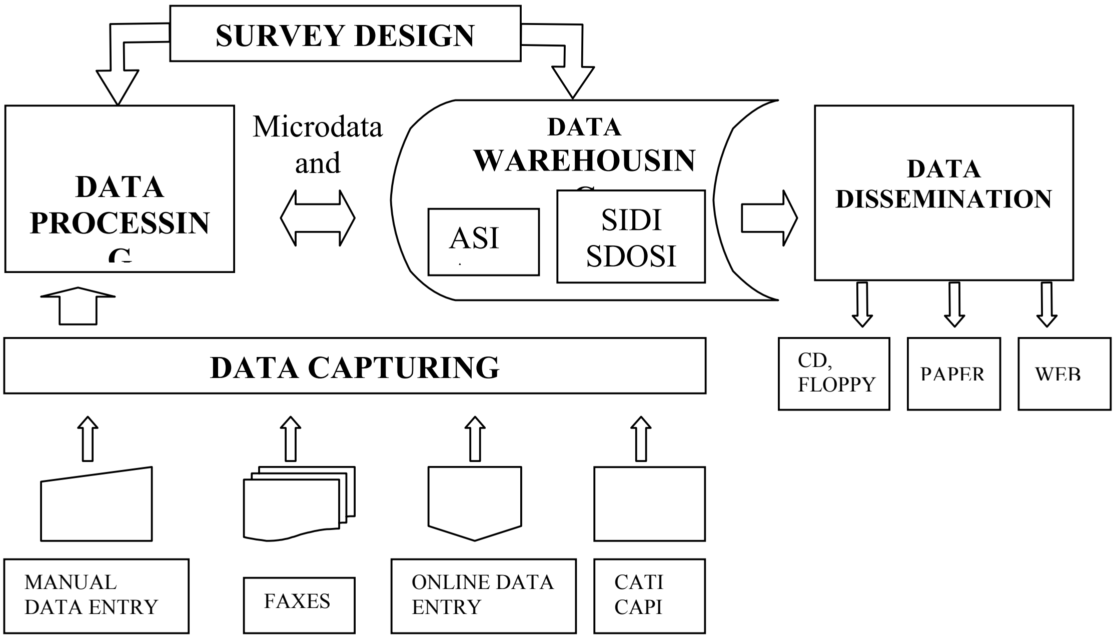
14. Even if the procedures need to be clearly defined and specified in advance it gives many useful opportunities such as:

- the procedures can be documented and reproduced;
- time and resources devoted to the process are reduced;
- the personnel skills are stimulate and improved.

15. The second key function in the architecture of an integrated statistical production system is data warehousing. It represents a centralized collection of microdata, macrodata and metadata set up according to a kit of specified criteria and objectives. It might contain data coming from different sources (mainly surveys and administrative archives). It can be also organized in independent sub-systems each one related to specific topics. Since it represents the base for any data dissemination, a data warehouse needs to satisfy some requisites that include being completely integrated, adaptable during the time and stable. In addition, it needs to be equipped with metadata and, whenever possible, with performance indicators.

16. In paragraphs 3 and 4 an overall description of the two implemented modules of the architecture described will be given: the data processing application named AGAIN and the data warehousing module named ASIA.

Figure 1: Scheme of an integrated system supporting data production



## III. AGAIN: OUR BEST PRACTICE APPLICATION THAT SUPPORTS STATISTICAL DATA PRODUCTION

17. The main features of AGAIN are its portability (since it only requires having SAS license with the module AF) and adaptability.

18. The software architecture has been designed to support editing and estimation procedures and provide documentation on corrections, missing data integration and estimation process (Benedetti R., Martino L & Salvi S. 2001). It also includes some basic utilities such as a module for handling archives including the possibility of data recording and data viewing, a module for importing external files useful in managing administrative data and a module for exporting ASCII file useful, for example, in preparing monthly files for CATI interviewers.

19. The main menu of the AGAIN tool implemented for some surveys, conjunctural or annual, are shown in figure 2.

20. Some important modules implemented in the AGAIN system are described in the following sub-paragraphs.

### 3.1 Estimation

21. This module supports all the procedures necessary to get estimates of the totals of each variable of interest, as well as the relative sampling errors, using the Horvitz-Thompson estimator. It also permits adjustments of the basic sampling design weights by total non-response. These weights can be further adjusted in AGAIN by means of calibration (Deville and Särndal 1992), so that the current survey estimates are consistent both with those produced by other surveys and the reference frame.

22. Weights can be computed by strata or by groups (obtained collapsing original strata).

23. The procedure also provides the opportunity to record the variables' selection carried out, useful in case of phenomena that show seasonal trends.

24. The estimation module also proposes a list of rules that have to be satisfied. They basically include range and consistency checks. It might be applied after interactive correction has been carried out. Units that violate rules can be corrected by a deterministic procedure or by a macro-edit module.

25. Statistical tables in the format used for internal publications and for transmission to Eurostat can be obtained at the end of the process.

### 3.2 Outliers detection

26. AGAIN enables to detect and highlight the most influential potentially wrong data by a set of procedures completely integrated. The aim is selecting a restricted set of elementary data to edit minimizing the number of interventions and maximizing the efficiency of final estimates. This macro-editing approach, has the advantage of strongly reducing time and resources devoted to data editing and increasing, in the meanwhile, the final estimates precision (Hidiroglou and Berthelot 1986).

27. The first two procedures (scatter and histograms) that are made available by the system are based on the comparison of the current value of the unit with a reference one. Such a value can be the annual data reduced to a monthly figure, the data referred to the same month in the preceding year or to the preceding month of the same year. The scatter plots each couple of elementary data (the current data and the reference one) making the comparison easier.

28. AGAIN’s second data editing method displays the histograms of the normalized variations in the interval [-1, 1]. By selecting one histogram at a time, and altering its limits simply moving the horizontal slide bars, the operator isolates the elements on the distribution codes.

29. Points falling outside the main cloud of data or lying in the tails of the distributions are the ones probably affected by errors.

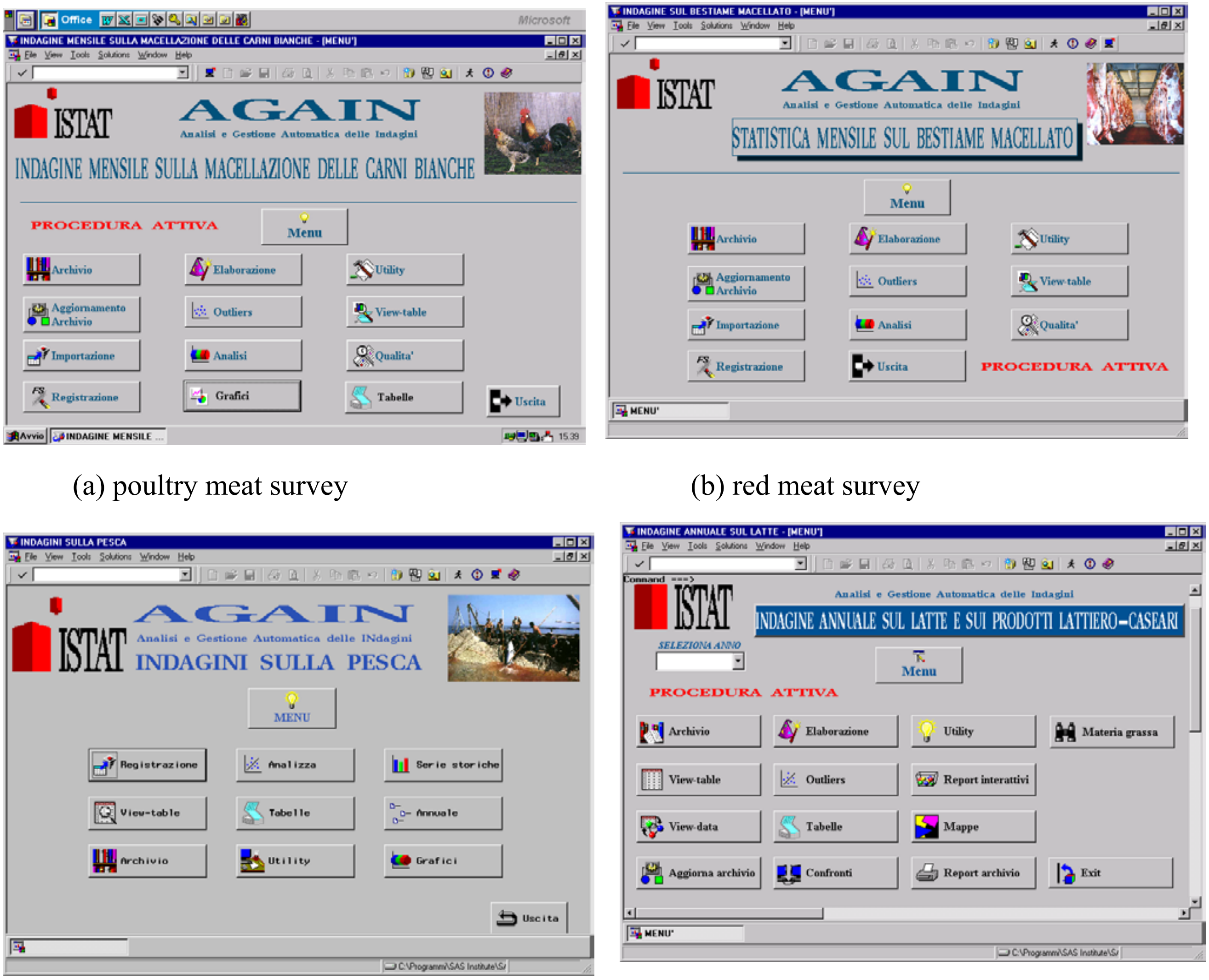
30. These two methods should be used jointly, because they have different information content: by scatters it is possible to detect ‘large outliers’, by examining histograms the ‘small’ ones can be identified.

31. Whatever criteria has been used, these points can be highlighted by clicking on them. In this way they are contemporaneously pointed out in any other graphical representation of the module so that user can do cross-check of data looking at the position of the unit respect to many different criteria.

32. Outliers' detection procedures include also a third macro-edit module that gives the opportunity of selecting units that largely contribute to variations between current and reference values. Afterwards outliers detected can be shown in the same table with other units coming from the preceding selections.

33. Once units potentially affected by errors have been selected, they can be manually corrected after editing or they can be treated as total missing responses and imputed with automatic procedures (Benedetti R., Espa G. & Piersimoni F. 2001).

Figure 2: *Main menu of AGAIN system tool for different survey: (a) poultry meat, (b) red meat, (c) fishery, (d) milk dairies.*



(a) poultry meat survey

(b) red meat survey

(c) fishery survey

(d) annual milk dairies survey

### 3.3 Quality report

34. The main feature of a high quality statistical production is represented by the repeatability. This property requires that everybody is able to reproduce the process and obtain the same results. For this reason, methods that have been used to get estimates, corrections applied and any other useful information on how the process has been carried out should be made available. For satisfying this need all the modules maintain memory on what it has been done. In addition a specific module has been set up with the aim at reporting about error correction and providing indicators quantifying corrections introduced by variables, by strata and by single unit. It is extremely useful since it documents the statistical process and it gives an idea on how much the final data are far away from raw data.

## IV. ASIA AGRICULTURE: THE DATA-WAREHOUSE TOWARDS WE ARE MOVING

35. Recently a large effort has been done by the Italian NSI to set up a general data warehouse on enterprises and farms. It is constituted by the Statistical Information System on Enterprises (SISSI - Capasso et al. 2000), a multidimensional and multifunctional structure that integrates all the official information available on enterprises and farms and economic issues in general. Since its first release, it helped reducing duplication in data collection and rationalizing surveys design process among all agencies involved in statistical production. It is structured in relational databases periodically updated when surveys are conducted. Currently SISSI includes the Statistical Archive of Active Enterprises (ASIA), an integrated set of raw and processed data coming from national surveys.

36. ASIA archive has been set up complying with Eurostat Regulation n. 2186/93 regarding to European statistical harmonization.

37. The archive provides basic information on enterprises and farms and it is updated with data resulting from all the Economic Department surveys and with data coming out from other databases. Its updating requires a continuous work of standardization of definitions and classifications, comparison of methodologies and contents both in time and among different sources.

38. The purpose is obtaining an updated frame of units complete with variables of stratification and other indicators such as income, number of employee, etc…

39. ASIA is going to become the reference universe for the Italian NSI economic sample surveys. In particular ASIA Agriculture contains data relaing to farms and, in a near future, also to that units which activities are in the fishery and in the forestry sectors.

40. The first step of the project has been merging records coming from different archives (Agricultural Census and administrative ones) for only one region and, in a second step, applying this method to all the other regions. The state of the art of the experimental results indicates that about 71.2% of the records matched (45.800 on 64.225), the other ones (about 10.000) unmatched.

41. A lot of works must be done. First of all, it is necessary to include, among the structural variables, those ones which are typical for farms (e.g. Utilized Agriculture Area, Total Agriculture Area, etc…).

42. Secondly the unmatched records must be re-processed with probabilistic matching methods after a normalization procedure on the principal matching variables such as address or name.

43. In addiction the system includes two documentation tools whose development is still in progress. Both of them contain metadata and quality indicators whose aim is helping final users in making a correct interpretation of data. The first one, SIDI, contains information on methodological features provided for any single survey information (on sampling design, estimators, methods of integration and so on). The second tool, SDOSIS, contains definitions, classifications and other additional information needed to better understand data content.

## V. CONCLUSIONS

44. The Italian National Institute of Statistics has started a challenging activity aiming at improving quality in the agriculture statistical production process. Two of the emerging results in this context have been the decisions of setting up a survey automatic processing system, such as AGAIN and developing a data warehouse such as ASIA Agriculture. In a close future AGAIN system will follow two main directions. On one hand the system could be adapted to the other surveys by adding some modules. On the other hand it could be generalized and made applicable to any survey. The first hypothesis would not require a great effort but its the *minimum scenario*. The second hypothesis could require a large investment of time and human resources but it could come out with a more complete and flexible tool, the *maximum scenario*.

45. Whatever the scenario will be the system will have to include a link with ASIA Agriculture data warehouse; only in this way it will be possible to implement and conduct surveys in order to analyze data from a territorial and sectorial point of view.

## References

Benedetti R., Espa G. & Piersimoni F. (2001): Available methods, techniques and software for survey data editing. Proceedings of the Conference on agricultural and environmental statistical applications in Rome, contributed paper. Rome: 5-7 June 2001.

Benedetti R., Martino L & Salvi S. (2001): Un sistema interattivo per la valutazione ed il miglioramento della qualità delle indagini. Paper submitted for the intermediate conference of SIS Society "Processes and statistical methods of evaluation". Roma, 4-6 2001.

Capasso G., Del Mondo G., Vignola L. (2000): Un sistema informativo per la produzione e l'integrazione delle statistiche sulle imprese (progetto SISSI). In GIUSTI A. (eds.) Ingegnerizzazione del processo di produzione dei dati statistici, 181-191. Padova: Cleup.

De Santis A. (2001): On the opportunity of a web-site in Agricultural Service. Proceedings of the Conference on agricultural and environmental statistical applications in Rome, contributed paper. Rome: 5-7 June 2001.

Deville J. C., Särndal C. E. (1992): Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376-382.

Eurostat (2000): Assessment of quality in statistics. Item 4 of the agenda "Standard quality report". Doc. Eurostat/A4/Quality/00/General/Definition. Luxemburg: 4-5 Aprile 2000.

Hidiroglou M.A., Berthelot, J.-M. (1986): Statistical editing and imputation for periodic business survey, *Survey Methodology*, 12, 73-83.

-----