STATISTICAL COMMISSION and Working Paper No. 16
ECONOMIC COMMISSION FOR EUROPE English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic II: Propagation of knowledge to users

ON THE DESIGN OF THE SWISS CENSUS EDITING AND IMPUTATION PROCESS

Submitted by the Swiss Federal Statistical Office[1]

Contributed paper

Abstract

We will show the importance of the user information if new methodologies of editing and imputation (E&I) are discussed to be used in a census, in particular in the Swiss Census 2000. It is important to remark that the users are not methodologists but producers of statistical analysis.

I. INTRODUCTION

1. The use of callbacks and of register data is useful after a first process of recognition of non-response, invalid and incoherent data. However, the callbacks do not guarantee good answers and some problems may remain.

2. The main characteristic of registers is the link of the data to the specific unit (e.g. persons) and the facility to update data continuously. A sophisticated system could even guarantee the completeness of the registers. But the quality of the registers is normally not known because there is no other source to compare the data and they are not tested using samples.

3. These methods will hardly resolve all the existing problems with the data. For this reason another process should be considered to improve the quality of the data, now that up to now the subject matter statisticians relied heavily on register data. We will only consider methods based on the census data itself (hot deck methods).

II. POSSIBLE SCENARIOS

4. The use of deterministic corrections was already decided without evaluating other possibilities for the E&I process before the statistical methods unit could intervene. This caused a time shortage for proposing a new methodology. Before deciding on which E&I method will be used in our census, we consider three scenarios:

i) Use of deterministic corrections

---

[1] Prepared by Daniel Kilchmann.

ii) Use of existing E&I software
iii) Development of a new software for E&I

We discuss briefly advantages and disadvantages of these three scenarios.

Use of deterministic corrections

5. Deterministic corrections were used in our former censuses to improve the response rate and to correct invalid and incoherent data. On the one hand, subject matter specialists know very well how to use them and, on the other hand, they are supposed to be very efficient because the improvement of the response rate could already be observed. However problems with the quality of the resulting data are very difficult to see.

6. The use of deterministic corrections is often guided by the conviction of changing the data in a logical way, which should lead to results which are not too far from the reality. This leads to the hypothesis that common sense should lead to the best results for the majority of the failed records. But deterministic corrections are based on subjective decisions to decide which value is the "most probable" and in which circumstances (values of the other variables). This can give the impression that the corrected values are the true values, which can certainly not be proved nor can the opposite be proved. In this way, the impact of introducing a structural bias by deterministic corrections is supposedly minimised because of the "little" divergence of the true and the corrected values. However it is usually impossible to evaluate the quality and the impact of the deterministic corrections process. One reason is sequential processing which always starts from the point of view of one variable to be changed and suggests that the items of the related variables are true. This can lead to a cyclical process. Its advantage is that it is normally quite easy to explain the corrected values to non-statisticians by referring to the common sense.

Use of existing E&I software

7. E&I software is increasingly used to handle problems that remain after manual checking but for our Office the use of E&I software is new. Every E&I software has its underlying methodology, which can be quite complex for non-methodologists. The goals of E&I methods are to preserve the multivariate distribution of the population characteristics and to change as few data as possible (nearest neighbour, Fellegi-Holt, etc.). The imputed data is no longer strictly linked with the specific units and it is not easy to explain the imputed values to non-statisticians.

8. Some of the existing software for E&I does not treat all the variables at once because of computational limits or political decisions even if the method is based on this (NIM). Hence, groups of variables could be processed sequentially which implies that the processing becomes in some way deterministic too.

9. The implementation of existing E&I software needs also an adaptation of the software or more often a few interfaces to integrate the software into the existing computer system. Adaptations to the characteristics of the census can be cumbersome if the code is delivered, otherwise it is even impossible. The real costs of using existing software are not easy to estimate because of many unknowns.

10. The impact of the E&I process can be evaluated in general but analysing its behaviour on individual records needs a lot of time and work. For example, the NIM prototype writes logouts in flat files which are not very user friendly if a big data set was processed. An interface for analysing the logouts would be useful. Furthermore, it is not easy to explain the random behaviour to non-statisticians.

Development of a new software for E&I

11. Developing a new software needs a good knowledge of the E&I methods and of the needs of the census. Unfortunately the best methods have often restrictions that are incompatible with the needs of the census, like the exclusive use of qualitative or quantitative variables, restrictions on the number of variables to be processed in relation with the computational time used, etc.

12. Developing even a new E&I method needs years of research; for this reason we have refrained from developing a new E&I software. However mixtures of the development and deterministic corrections seem possible.

13. The decision which scenario will be used depends on the following criteria:

- Goals of the census
  - Analysing the characteristics of the population:
    - Are made and used by the statistical office on national and regional level (even per hectare in some cases)
    - Are used by the government (parliament, laws, taxes, planning and development etc.)
  - Register data:
    - Matching and updating of existing data
    - Build up new registers
    - Use for regional statistical offices for their own analyses
- Methods
  - Deterministic correction
  - E&I (non deterministic imputation)
- Quality

  The measurement of quality is one of the main problems of statistical offices. The lack of quality is often not obvious. For new methods it is essential to show that the quality is very good or at least better than the quality of existing methods.
- Availability of software

  Statistical offices for their own use develop most of the software. This means that updated or new software is timed with the use in these offices. Its dissemination seems to be of secondary importance.
- Skills and knowledge of people

  The users knowledge can be very different depending on its tasks in the office. In the Swiss Federal Statistical Office global experts are working on the decision process. Their knowledge was formed by former censuses and by working in different domains of the SFSO.
- Politics: acceptance of the methodology by users

  One of the main goals of SFSO is to furnish data to external users. The census becomes in this way user driven. This implies that SFSO must be able to explain the imputed data also to users with no statistical background. Methods that are more sophisticated than the common sense need a diplomatic introduction and could be objects of critics. Therefore the acceptance of methodologies that preserve the multivariate distribution is not assured.
- Time and milestones

  The dates of collecting and disseminating data were fixed before the planning of the E&I project was done. This means that the milestones of the E&I project are given by these dates. The decision about which methodology to use has to be taken early enough to allow the implementation of it in the whole process. This needs the information about implementation and testing new software.

Of course the performance of the E&I software influences this discussion too. If the processing is too slow the milestones could not be respected anymore.

- Costs

  The costs of every subproject of the census are limited with the goal of spending less money than planned. A very good E&I software could be rejected if its costs are higher than planned. For this reason it is essential being able to estimate the costs when the planning of the census is made.

14. These points lead to a communication network which often has to be established by methodologists because of their knowledge of statistical approaches of the E&I problem and because the need of changing the method does not seem obvious to non-methodologists. All the main points have to be evaluated and discussed by users of different knowledge from different domains.

15. The inputs of the decision procedure are the "values" of the criteria mentioned above.

III. ESSENTIAL INFORMATION

16. Three of the criteria of the decision process could be exchanged by methodologists of different National Statistical Offices or by the producers of E&I software:

- Methods
- Quality
- Availability

17. In the decision process at the Swiss Federal Statistical Office we had to find out that only information about the first point was available. But even for this point we could not find comparative studies where the power of E&I methods were compared to the power of deterministic correction. Users give this information a very high priority because a change of the methodology must increase the quality significantly. Comparative studies are normally understandable by decision-makers of any skills and knowledge in a National Statistical Office and could also be used to increase the acceptance by any user of a new methodology.

18. Given that a census has to furnish data for statistical analyses and for registers, we have to admit that none of the methods listed above can achieve both goals. For a statistical methods unit it is a hard task to convince the users of this fact and to show the advantages and disadvantages of the existing methods.

19. Quality measurement of E&I processes is always an open field for research (EUREDIT) although ISTAT has developed a software for testing purposes of E&I software. Again comparative studies could be used as guidelines for the quality.

20. These points cover already most of the aspects which influence the decision. Every National Statistical Office could work out such a comparative study by themselves but it seems easier to communicate the existing studies.

21. The information about availability of software is also essential because it is one of the starting points to evaluate the possible three scenarios listed in section II. For this reason, the Work Sessions on Statistical Data Editing are very important!

IV. CONCLUSION

22. A lack of information can cause a rejection of good software and the application of efficient methodologies in the E&I process. Theoretical aspects are certainly important for methodologist but it is

not their task to take the decision in National Statistical Offices. That is the reason why (practical) information for non-methodologists and non-statisticians should also be given in addition to theoretical and technical information.