STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 15
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

EXPERIENCES FROM FINLAND, ESPECIALLY ON IMPUTATION TECHNIQUES

Submitted by Statistics Finland[1]

Contributed paper

I. INTRODUCTION

1. Statistics Finland has *no unified framework* for editing and imputation techniques such as GEIS in Canada but a new quality checklist has been under preparation for about one year. The first version of the checklist has just been published. It will be continuously updated. The concept of this manual was derived from the 1998 Statistics Canada Quality Guidelines. Our manual, of course, follows the Finnish statistical production style, not that of Canada. There are some themes not expressed in the Canadian manual. The structure for each theme is as follows: (i) scope and purpose including concepts, definitions; (ii) principles and methods; (iii) recommendations and guidelines; and (iv) references. Each section is rather short, 1-6 pages. One section is devoted to editing and imputation. It contains a general background and details on both subjects.

2. This quality guideline will hopefully help to harmonize and integrate our statistical production system which is rather at variance at the moment. Fairly unique terms, concepts and methods have been applied in each survey, although the integration has become more common during recent years. This is especially true concerning the editing systems which are different in each survey/statistical area. The situation with imputation methods is somewhat different because, 10 years ago, these were very rarely used, except for logical imputation which may be used as a part of the editing process in most surveys. Earlier, imputation techniques were used, for example, in income surveys for some income items, and as a comparison method for reweighting the whole income distribution to estimate it as correctly as possible (Laaksonen 1991). Recently, imputation methods have encountered a renaissance, mostly in business surveys including wage and salary surveys. The paper presents in more detail experiences since that renaissance.

II. QUALITY GUIDELINES

3. Our first quality guideline was published in 1987 and focused on the accuracy of surveys and statistics and their various components. The guideline was followed in some statistical units but not in general. We can still find today some publications with the quality assessment annex. In certain fields, such as the household budget surveys and the latest environmental opinion poll, a special issue on quality

[1] Prepared by Seppo Laaksonen.

has been published. These special issues have been prepared as a result of the needs of internal and external researchers who want to know more explicitly the quality (accuracy or error components) of the survey. To publish such a document is not an overall principle in Statistics Finland, because no system to follow the quality guideline was established.

4. In 1996, we began to apply total quality management (TQM). Its crucial elements are leadership, strategy, people, partnerships and processes, and it will encompass innovation and learning. It is believed that this approach will advance traditional data quality work as well. In order to succeed in this work, a special position of quality agent was established. The quality agent serves as coordinator for quality issues in general. Many new activities have emerged. Consultants from Westat in the United States have assisted in this work. This model is very similar to the Swedish one, where these activities were started about 2 years earlier. The Westat consultants, who are also experts in survey techniques, have visited Finland several times, and they have given much advice and training courses focused on statistical process control.

5. The latest step was to prepare the first draft guideline on statistical product quality including some process factors. At present, it is difficult to say to what extent we have succeeded concerning process-quality, and even more so on how well we will succeed in the new product-quality effort. There are some success stories, especially in narrow statistical fields, but the integration at the statistical office level has not been successful. This paper will present some example cases of our attempts to improve and develop imputation techniques, and explain why improvements in this methodology have been important.

III. IMPUTATION STORIES FROM STATISTICS FINLAND

6. Imputation methods (Laaksonen 2000) have been classified into four main categories:

(a) *Use of available/complete cases*, when any missing items have not been imputed.

(b) Deductive or *logical imputation,* when there is a known function (identity equation) between certain observed values and missing values.

(c) *Model-donor imputation*, where imputed values are derived from a (behavioural) model, that is, imputed values may be non-observable in a real life situation.

(d) *Real-donor imputation,* where imputed values are derived from a set of observed values, from a real donor respondent. Note that the methods in group (b) may provide a real value as well, but this is not derived directly from a real donor.

The first of these categories is not a real imputation method, but instead a course of action, or a baseline. The same classification is used in our draft quality guideline.

7. Traditionally, alternative (a) has been mainly used in our statistics, except that logical techniques have been applied at the editing stage of the survey process. Unfortunately, we have not surveyed how much and with which methods logical imputation has been used; perhaps not all solutions have been very objective. We plan to make an evaluation of such techniques in the near future if this will be accepted by our management body. The use of available cases means that the number of completed records (sample size) will be reduced especially in multivariate analyses.

8. Until recently, imputation techniques have not been used for ‘core variables’. Instead, in many cases some components of ‘core’ variables have been imputed in order to provide ‘a full accounting.’ Laaksonen (1992) gives an example on the income survey in which some income components were imputed. Using register data we were able to know with high probability who could have such incomes. On average, the influence of such components was not great except for certain socio-economic groups. Without such imputation, even poorer households would have been obtained.

9. There have been many similar cases to the one mentioned above. It is very typical that certain components y_p, for example, a farmer's income and production costs, or total wages, are not correctly reported in survey data. In some cases, we know the sum of these components ($y=Ó y_p$), but in other cases we know nothing, or one component, say y_s. This situation is difficult for a statistician because the use of available cases obliges to drop out all such units. If such missing data does not concern a group of units, and is thus a more or less random process, it is easier to overcome. This situation is more common in household surveys. In establishment surveys, there are often bigger units and if data are missing for these, without adjustments the bias may be relatively high. This is the *first reason* why the use of imputation in Finland has become more common in recent years. An example is the occasional and periodical items of wages (Laaksonen 2000).

10. The *second reason* is more complex. It concerns new business surveys which have been redesigned especially due to the needs of the European Union (EU). Why the EU? There are two factors:

(i) Finland started to use the EU VAT system in 1994 and, respectively, much new register data became available. We have traditionally used our registers and other administrative records, and this new approach has been used both in the Business Register (BR) and in statistics production. The information helped us to improve the coverage of the BR, and to find more quickly some useful data for statistics production.

(ii) Many business surveys have been harmonized. The harmonization integrated some previous surveys, and created new ones, such as Structural Business Statistics (SBS) and new needs for short-term economic statistics.

11. These factors have been one reason for using imputation: register data is used only for some parts of the data, so that any real data will not be collected. Below are two examples:

• The new design for SBS has been fairly simple since 1995. Full data on the enterprises with more than 20 employees will be collected, the data on the smaller enterprises are based on registers. This leads to mass imputation for all enterprises, including even the smallest (micro) enterprises at micro level. The data for this group are more limited than before when the manufacturing enterprises with more than 5 employees were surveyed completely. For EU purposes the data are reasonable, but maybe not for econometricians who wish to continue their longitudinal series' since 1995. It should be mentioned that the comparable manufacturing data from 1974 to 1994 have been greatly used in micro econometric research (e.g., Ilmakunnas et al 1999, Laaksonen and Teikari 1999, Vainiomäki and Laaksonen 1999).

• Short-term business indicators are very important for business people, meaning that their freshness sometimes is more important than their accuracy. Of course, it is ideal if the first results could be published quickly and would be of high quality. A standard way to solve this problem has been a simple automated survey for a small number of enterprises, covering completely large enterprises, but based on samples for small enterprises. These statistics are not basically measuring the level of businesses but their dynamics. Thus, the changes in business life are of greatest interest. This has been measured with certain indicators (indices). When using a panel approach, such dynamics may be analyzed to a certain extent, for example, so that the changes of the same business units have been measured on a yearly basis.

12. The redesign of some short-term statistics is currently under development. More attention has been paid to imputation. For the largest enterprises a small-scale simple survey is still used, but the rest of the data are picked up from the VAT and other taxation files. It is interesting that there are different time schedules for obtaining these data. Our main outcome variable, turnover, is available in most cases 1-5 months later than wages, salaries and taxes paid. The latter variables, on the other hand, are not available

for all enterprises but for these, where we have a more or less full historical time series at individual enterprise level. Therefore, we have developed two types of imputation models and imputation solutions (Piela and Laaksonen 2000) for filling up missing items, and for estimating the short-term indicators needed. This work has so far focused on retail and wholesale statistics which have been traditionally extremely important for the economy. The results are promising but we have to develop further and automate these techniques.

13. The *third reason* for imputation concerns complex missing data patterns in standard surveys when a user needs to make multivariate analysis (multi-tables, models) afterwards. This type of analysis has become more common due to the increased needs of research-oriented users. The missing item rates are not very high for any individual variables, maybe only 5-20%, but in a certain multivariate analysis the missing data rate may even be 50%. Respectively, in most models, the number of observations will be awkwardly diverse. The only 'good' solution is to impute missing values to a certain extent, for example, so that the multivariate missing data rate would be less than 10%. This sort of imputation should have as small as possible an influence on the further multivariate analysis. This requires that the so-called 'anticipated preservation rate', $\mathrm{PR} = \dfrac{\sum_k |\hat{y}_k - y_k|}{\sum_k y_k}$, would be low (in the formula $\hat{y}_k$ is the imputed value and y_k the corresponging real (anticipated) value). These principles are, to some extent, used in several surveys. Much work in this area has been done for the European Community Innovation Survey II. Eurostat was helping in this work with a special tailored SAS program which also covered outlier-detecting tools and other robusting techniques. The greatest effort was made to harmonize imputation techniques including the nearest neighbour method using entropy measure, and ratio-based regression imputation.

IV. CONCLUSIONS

14. Our approach to developing edit and imputation techniques has been so far 'experimental'. That is, we have implemented it on statistical processes that have been redesigned for general reasons, not for performing better edit and imputation techniques. Some developments have been implemented in each statistical process, but some are still under further examination. These methods, mostly imputation methods, have been presented in national seminars where the participants are mainly methodologists. In a few cases, more detailed and theoretical experiments have been presented to an international audience.

15. It is not easy to introduce into daily statistical production these mostly positive experiences using new techniques, because the use of the new techniques requires knowledge and user-friendly software. There are also administrative obstacles. An additional reason is the problem of resources. However, for example, the basic form of the *regression-based nearest neighbor hot deck imputation* method (Laaksonen 2000) is used without any problems.

16. A broader implementation of new editing and imputation techniques could now be undertaken because the first version of the quality guideline has been issued, and there are some favourable pilot studies. We have not, however, explicitly carried out pilot projects for integrating best editing techniques and imputation methods under the same process. This needs to be done in some complex statistical processes. After that, we will have a reasonable material and experience to start a more general implementation of editing and imputation techniques. But, this requires support from the top management including the provision of resources. It is not fully clear how well methodological practices may be implemented using the newest techniques. All new methods and techniques should be discussed in-depth with the people involved. Training courses should be given, and coaching should be made available.

Finally, the first good practice should be achieved, and continuously developed. Good documentation should be available for internal and external users. However, these aspects are very little discussed in the draft quality guideline.

17. We are interested in learning from the experiences of other countries in their good practices in order to avoid as many mistakes as possible.

References

Ilmakunnas, P., Laaksonen, S. and Maliranta, M. (1999). Enterprise Demography and Job Flows. In: Alho, J. (ed). Statistics, Registries and Science. Experience from Finland. Helsinki. Pp. 73-88.

Laaksonen, S. (1999). Weighting and Auxiliary Variables in Sample Surveys. In: G. Brossier and A-M. Dussaix (eds). "Enquêtes et Sondages. Méthodes, modèles, applications, nouvelles approches". Dunod. Paris. pp. 168-180.

Laaksonen, S. and Teikari, I. (1999). Analysis of Effects of Reconstructed Business Units on Employment and Productivity. Longitudinal study using synthetic units of Finnish manufacturing. In: Biffignandi, S. (Ed.). Analysis of Enterprise Data. Physica Verlag. Springer.

Laaksonen, S. (1996) (ed). *International Comparisons on Nonresponse*. Statistics Finland Research Reports 219. Helsinki.

Laaksonen, S. (1992). Adjustment Methods for Non-response and their Application to Finnish Income Data. *Statistical Journal of Economic Commission for Europe of United Nations* 9, 125-137.

Laaksonen, S. (1991). Adjustment for Non-response in Two-year Panel Data: Applications to Problems of Household Income Distribution. *The Statistician* 40, 153-168.

Piela, P. and Laaksonen, S. (2000). Mass Imputation When Missingness Varies Awkwardly Over Time. Draft Paper for the *International Conference on Establishment Surveys 2*. Buffalo June 17-21, 2.

Rubin, D. (1987). *Multiple Imputation in Surveys*. John Wiley & Sons.

Särndal, C-E. (1996). For a Better Understanding of Imputation. In: *Laaksonen, S. (ed.). International Perspectives on Non-response*. Research Reports 219, 7-22.

Vainiomäki, J. and Laaksonen, S. (1999). Technology, Job Creation and Job Destruction in Finnish Manufacturing. *Applied Economics Letters*, 81-88.