WP. 6
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (i): Automated editing and imputation and software applications

IDEAS ON EDITING OF STATISTICAL REGISTERS

Supporting Paper

Prepared by Li-Chun Zhang, Statistics Norway

Work Session on Statistical Data Editing
(Neuchatel, Switzerland, 5-7 October 2009)

Ideas on editing of statistical registers

Li-Chun Zhang[1]

Abstract
The Integrated System for Editing and Estimation (ISEE) is one of the most important generalized systems for statistical data processing at Statistics Norway. It is designed to help the implementation of the so-called top-down approach to editing. In its present form the ISEE is primarily directed at business sample surveys. In this paper we report some of the ongoing conceptual work aimed at the editing of statistical registers. We argue that the ISEE philosophy and methodological foundations can be extended to this area of statistical production, but there are also special challenges of both technical and theoretical nature that need to be resolved.

I. Introduction

The development of generalized systems for statistical data processing is an important measure undertaken at Statistics Norway, in order to achieve the strategic goals it has set for statistical production. A basic principle is to simplify, to improve and to re-use the production processes. The implementation of the various generalized systems aims to provide more efficient use of resources, greater flexibility in production means, as well as better quality insurance. Integrated System for Editing and Estimation (ISEE) is such a generalized system. It consists of 3 core applications: DynaRev for microdata editing, Struktur for estimation of population aggregates, and Pris for calculation of price indices. The key feature is the complete integration of editing and estimation processes. For instance, inside the ISEE it is possible to examine instantly the final estimate of a population total due to any changes made to the data. One is thus in a much better position to implement the so-called top-down approach to editing (e.g. Grankvist and Kovar, 1997; Grankvist, 1997). However, in its present form the ISEE has its primary applications in processing of business sample surveys, where it is being used in the production of about 60-70 statistics. Below we shall report some ongoing work which investigates the possibilities of more extensive use of ISEE in different fields of statistical production. Firstly, in Section (II.) we outline the IT and methodological framework for the present ISEE (Zhang *et al.*, 2007); then, in Section (III.) we consider in particular the potentials as well as challenges in developing an ISEE for statistical registers.

II. IT and methodological framework

a. Overview of IT architecture

There are mainly two types of objects in ISEE: a common database and a set of applications. Different kinds of units are organized and maintained in the common database, including the

[1]Statistics Norway. E-mail: lcz@ssb.no

various edits and all the changes made to the data. The observation data are organized both at the unit level and some pre-defined aggregated levels. The main applications can be divided into three groups:

- Core applications DynaRev, Struktur and Pris. On the one hand, being fully integrated, the various functionality can be arranged rather freely to optimize the system performance. For instance, logical edits and microdata editing are located in Dynarev, while statistical edits and case weighting are located in Struktur. On the other hand, the data administration ensures that the 3 applications can be run separately for particular purposes, without necessarily having to invoke each other. This helps to increase the flexibility in use.

- User-administered interfaces such as Screen Builder, Weights Administration, *etc.* These are applications that can be individually formed by the users. For instance, the Screen Builder allows the user to define the edits without having to learning the program codes, and allows one to set up the screen look that is suitable for the observation data of concern.

- Standardized interfaces that communicate with data outside of the ISEE. For instance, the Data Collection Interface channels data from different sources and formats to a standard solution supported by the common database. Other data that are transferred between the ISEE and the general data environment through such standardized interfaces include the various registers, samples, dissemination data as well as the metedata systems.

b. Statistical methodological functionality for editing

The core applications of ISEE contain functionality for editing as well as various estimation and calculation procedures. Here we concentrate on the process of editing.

The functionality should allow us to pursue the top-down approach to editing. Under such a perspective, correction of microdata is no longer the primary concern, and the effects of microdata correction should be judged based on the final statistical outputs. This is the conceptual interpretation of the top-down approach. Moreover, not all the errors in data are equally important for the results. We call an observation or a value critical if it has relatively strong impact on the statistics. Not all the critical values need to be outliers, let alone mistaken. Nevertheless, in order to reduce the cost of editing, one must focus primarily on critical values rather than outliers *per se*. This is the operational interpretation of the top-down approach.

In ISEE there are mainly two types of functionality for editing: *edit* and *assignment*. An edit is a procedure used to check the data. The result of an edit is the marking (or flagging) of values and/or units, either because they have failed the check or because they require further editing. An assignment is a procedure used to give (or change) values to the variables. An assignment is only allowed to take place after one or several edits. It is not a necessary action following the edits, but only one of the options. In addition there are meta-edits (i.e. checks or analysis of the edits) and para-data for monitoring of the production processes, such as

- Log: records of changes made to the data over time, or other aspects of the editing process.

- Report: Summary or juxtaposition of the various edits or assignments applied.

- Analysis of process efficiency (ANOPE): Summary of the measures undertaken in editing against the final results.

When it comes to the edits in ISEE we differ among 3 types:

- A *non-statistical* edit checks the data unit by unit, or record by record. All kinds of fatal errors are detected by non-statistical edits. Some examples of non-statistical edits are:
  - Initial automated status edits, such as recorded, missing, out-of-scope, *etc.*.
  - Logical edits on a single variable, e.g. Age ¡ 0, and so on.
  - Consistence edits on fatal errors in combination of variables, e.g. Age = 14 and Marital status = Married, and so on.
- A *statistical* edit must be run on a number of observations (or all of them). All edits aimed at detecting critical values and outliers are statistical edits. Some examples are:
  - Graphic edit, e.g. a scatter plot of the current values against the previous values.
  - HB-method, which is a non-parametric edit for detecting outliers.
  - Regression-based diagnostics for outliers and/or critical values.
- An *aggregated* edit is not directed at any single observation (or unit). Instead it identifies a group of values (or units) which together may have a strong impact on the results, or constitute an outlier at an aggregated level. Often the groups can be motivated from the natural hierarchical structure in the data, such as design strata, elementary aggregations for price index, clusters of units, *etc.*.

We also distinguish among 3 types of assignment:

- A *non-statistical* assignment is based only on information about the concerned observation unit. For example, changes made after contacting the owner of the data, imputation of missing values from register sources after unit-level linkage, and so on.
- A *statistical* assignment must be based on an assumed statistical distribution underlying the variables of concern and, thus, other units than the target one for the assignment. Typical examples are neighbor imputation, regression imputation, hot-deck imputation, *etc.*.
- A *consistence* assignment gives or changes one or several values in order to achieve consistency in the data. It can be applied to different variables of the same unit, but also units underlying a certain aggregation. Characteristic of a consistence assignment is that the logical relationship among the variables are taken into account.

III. Challenges for editing of statistical registers

As mentioned before, the ISEE currently has its primary applications in processing of business sample surveys. In Table 1 the statistics produced at Statistics Norway are classified according to, respectively, the source and type of the target variables of interest, and the present ISEE is

placed in this larger context. A question that is being investigated at the moment is whether and how the ISEE can be extended to register-based statistics. Below we summarize some of the issues that have been raised as well as the tentative conclusions that have been reached.

Table 1: Areas of statistical production by data source and type of interest variable

	Type of interest variable	
Data Source	Social	Economic
Sample Survey		(ISEE present)
Statistical Register		

The top-down approach is also preferred in register-based statistics. Limited resources in reality implies that in editing one has to prioritize. A top-down perspective forces one to consider the relative importance of various objects, even if it is impossible to sort them in a strict order.

Editing and estimation are parallel processes in register- and sample-based statistical production. The philosophy of integration of the two remains desirable. However, different areas of statistical production (Table 1) may require different technical and methodological solutions. It is, for instance, unclear at this stage whether a single system can efficiently handle statistical registers of both social and economic data.

Register-based statistical production calls for its own theories. For instance, imputation by which one obtains a complete data matrix for tabulation is more attractive than case-weighting as the mode of production. Evaluation of uncertainty in register totals, often at a rather detailed aggregation level, requires conceptualization of the different sources of randomization as well as sensible empirical modeling. Also, theories need to be developed for dealing with register-specific errors, such as relevance errors, unit errors, microdata inconsistency across different registers, *etc.*.

Modifications and extensions of established doctrines and approaches are needed for editing of statistical registers. Take for instance the consistence edit and the associated consistence assignment. Under the well-known Fellegi-Holt model of editing (1976), fewest possible changes are made to the data so that consistence can be achieved. Such an approach seems practically sensible provided the errors occur more or less equally likely at different places, which typically is not the case when it comes to microdata inconsistency across different registers. Some form of plausibility-weighted error localization may be more suitable in this respect.

Take another example of statistical edits for detecting potential critical values. The richness of linked register data means that register-based statistics are often highly multivariate as well as multi-purpose. In such situations standardized approaches are currently lacking in order to define critical values and to form statistical edits accordingly.

Finally, an obvious matter of concern is the sheer amount of records in a statistical register that need to be processed. In order to be able to instantly examine the results due to changes made to the data, one needs efficient methods of tabulation and aggregation. As another example, a common technique of editing registers is drilling. Formally this amounts to repeated application of nested and/or overlapping aggregated edits. Hierarchically structured data seems therefore necessary for efficient computing. Technical database solutions must take this into account.

References

Fellegi, I.P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, vol. **71**, pp. 17-35.

Grankvist, L. (1997). The new view on editing. *International Statistical Review*, vol. **65**, pp. 381-387.

Grankvist, L. and Kovar, J.G. (1997). Editing of survey data: How much is enough? In *Survey Measurement and Process Quality*, Eds. L. Lyberg *et al.* New York: Wiley, pp. 415-435.

Zhang, L.-C., Faldmo, M. I., and Lien, O. K. (2007). ISEE - Integrert System for Editering og Estimering. Paper in Norwegian presented to *the 24th Nordic Meeting of Statisticians.* Reykjavik.