

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (ii): Editing data from multiple sources

DATA EDITING ACROSS SOURCES IN "MUNICIPALITY-STATE REPORTING"

Supporting Paper

Prepared by Dina Rafat, Statistics Norway, Norway

I. INTRODUCTION

1. Municipality-State Reporting (MSR) is a national information system in Norway that provides information on municipal and county municipal activities, including economy, schools, health, culture, the environment, social services, public housing, technical services and transport and communication. MSR started as a project in 1995 and went over to operating phase in 2002. It has a goal of bringing relevant, reliable and comparable information on municipalities, which give basis for better governmental and municipal control and better dialogue between municipalities and citizens.

2. The data collection done by Statistics Norway makes it possible to publish indicators that combine data from many sources, for example data on accounts and data on services and personnel. In this paper we will discuss the suggested approach to data editing across sources in "Municipality-State-Reporting". Section II describes the infrastructure of MSR and the way different types of key indicators are formed. The use of SAS/Insight for revealing the types of errors in the data and for developing automatic controls will be offered in section III. The conclusions from the project work are presented in section IV. Section V specifies the requirements to functionality that an overall editing system has to fulfill. Some challenges connected to the IT solutions are presented and possible future actions are discussed.

II. MSR INFRASTRUCTURE AND KEY INDICATORS

3. There are 19 working groups under Municipality-State Reporting with the responsibility of reviewing each year the reporting arrangements and key indicators for certain areas. The majority of the groups are led by Statistics Norway (SN) with members from municipalities, ministries and other interested organizations. One group is responsible for accounts and issues connected to it, while the other 18 deal with the service and personnel reporting related to certain functions in the accounts.

4. Together with Municipality-State Reporting a new structure of municipality accounts was introduced. Accounts are divided into services, which are called *functions*. One MSR requirement is that collection of service data (for example emission from a sewage system, number of kindergarten places etc.) must follow the same service areas as accounts, so that it is possible to put them together to form key indicators. For each function the accounts are separated by *type*, which represent a kind of income or expenditure.

5. The data are published on 3 different levels:

Level 1 - selected key indicators are calculated for administrative and political leaders and other interested that need an overview of municipality key figures.

Level 2 - detailed key indicators provide deeper information on certain areas of municipality activities.

Level 3 - background data are for those users who conduct their own analysis.

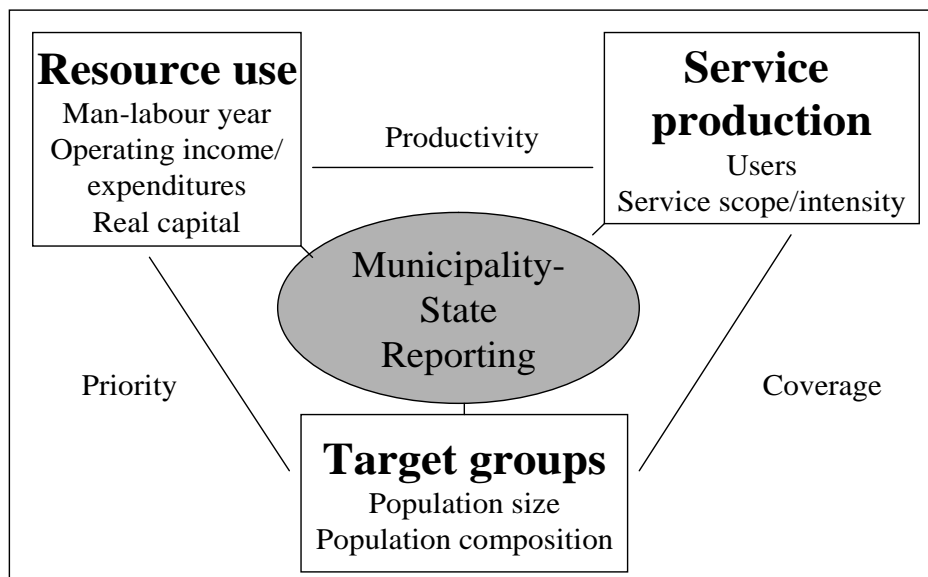
6. The indicators on level 1 and 2 are ratios that consist of a numerator and a denominator, based on the data on level 3. Numerator and denominator can come from different sources.

7. Key indicators are divided into the following categories:

- **Priority** - resource use in relation to service target group (for example, percentage of municipality investments in the school sector);
- **Coverage** of needs - service production in relation to service target group (for example, the ratio of kindergarten places to the number of children in the kindergarten age in municipality);
- **Productivity/unit cost** - resource use in relation to service production (for example, cost of a place in a nursing home);
- Detailed service indicators - supplementary information that can illustrate municipality's expenditures or priorities (for example, requirements for sewage system or quality of a certain service).

8. Figure 1 illustrates how data from different sources can be linked together to form different types of MSR indicators.

Figure 1. Publishing principles for MSR key indicators.



9. One of the main aims of introducing MSR was to simplify the information flow from municipalities to the State. It was done by creating a common data delivery system via MSR. The following are ways of collecting MSR data:

- File extraction from municipality systems (accounts, child welfare authorities and social services);
- Electronic questionnaires (service data and personnel data);
- External data (other data that do not come directly via MSR system).

10. Account data come directly from municipality systems, while the service data are collected mainly by questionnaires. Personnel data come both from a separate MSR-questionnaire as well as being drawn from registers. Population data are drawn from SN's demographic statistics.

11. Statistics based on MSR data is published twice a year: preliminary (unaudeted) figures on the 15th of March and edited figures on the 15th of June.

12. Several systems are used for data editing, depending on the type of data. Most of questionnaire data are edited in one system; there is also a separate system for accounts and yet another system for individual data. External data are edited to a little extent and linked together with the other types of data in the production of key indicators. There is little coordination in editing work between different areas and a need for editing across multiple sources has gradually become evident. A small group was set up to assess the need and importance of editing across different areas.

13. The next section illustrates the approach taken by this project group and some findings of the work.

III. EDITING METHOD

14. Three service areas were chosen for this project:

- Dental health service;
- Secondary education;
- Sewage service.

Within each area some key indicators were selected that combine service data and account data or population data.

15. The idea of the project was to conduct full data editing of the chosen key indicators for the year 2004, i.e. analysis of the indicators, development of controls, conducting controls, finding errors in the data sources, communicating the errors between the area departments within SN and giving feedback to the municipalities.

A. Data Analysis

16. SAS/Insight was used for graphic analysis. The visual presentation helped us see the distribution of certain variables, data pattern and deviation from normal. After we had become familiar with the data material, we used SAS/Insight to perform the analysis and data editing, i.e. find out what types of errors exist in the data, identify the threshold values for each indicator and develop controls.

17. We have analyzed the key indicators on level 2 and background figures on level 3, where denominator and numerator are collected from different sources.

18. Box plot was used to analyze the key indicators on level 2. It showed the distribution of the indicator and marked the median and outliers.

19. A simple ratio model was used for level 3 figures, where we looked at the numerator as a response variable and the denominator as an explanatory variable, for example, we assumed that a municipality's college expenditures (response) are dependent on the number of students in the municipality (predictor).

20. We looked at the following model:

$$Y = \beta X + \varepsilon, \text{ where}$$

Y -response variable;

X - predictor variable;

ε - part of Y that can not be explained by the model with an assumed normal distribution;

β - parameter that specify the model and that must be estimated.

21. SAS/Insight gives us an X-Y plot of numerator against denominator and an estimated model (regression line). We get a visual picture of possible deviant values (outliers). By deviant unit we mean

that the observation has a "big" distance to the regression line. Residual r_i is the difference between the observed Y-value and the predicted Y-value from the estimated model.

22. The so-called studentized residual r_{ii} is used to detect the deviant units and is defined as a ratio between r_i and the estimated standard error s_i of r_i :

$$r_{ii} = \frac{r_i}{s_i}$$

Criterion $|r_{ii}| > 2$ is recommended to pick the units that deviate strongly from the rest of the sample.

B. Types of errors in the data material

23. Based on our analysis of the chosen areas we have identified 3 types of errors in the data material:

- Zero in the numerator, but not in the denominator and vice versa;
- "Thousand" error;
- Deviant observation.

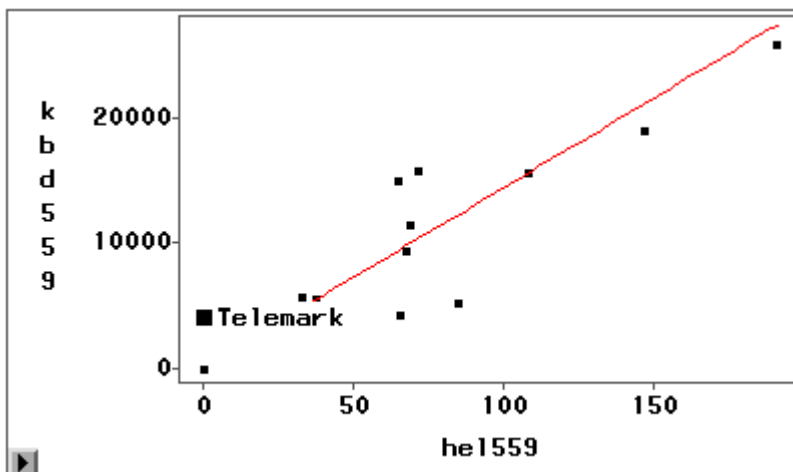
24. In addition there were many cases of missing values either in the denominator or the numerator. There were also cases of partial non-response, when some parts of the questionnaire remained unfilled even when the form was delivered.

25. In the analysis and when carrying out the controls it was important to separate missing values (that municipalities did not report) from the reported value "zero".

Example 1. "Zero" in the denominator but not in the numerator

26. Figure 2 shows the ratio model for key indicator "Corrected gross operating expenditure for the subject "woodwork" per student". As long as there is no activity on the service side it is not necessarily wrong if both denominator and numerator are "zero": if there are no students, there are no expenses. But in this example we can also see one municipality that have expenditure for the service but no students - Telemark municipality. This error would not be detected by analyzing the key indicator on level 2. One has to look at the denominator in relation to the numerator to discover this type of errors. In this particular case the control of student numbers in SN's register showed an error in the extraction of numbers from the register and indicated the necessity of examining the routines in SN closer.

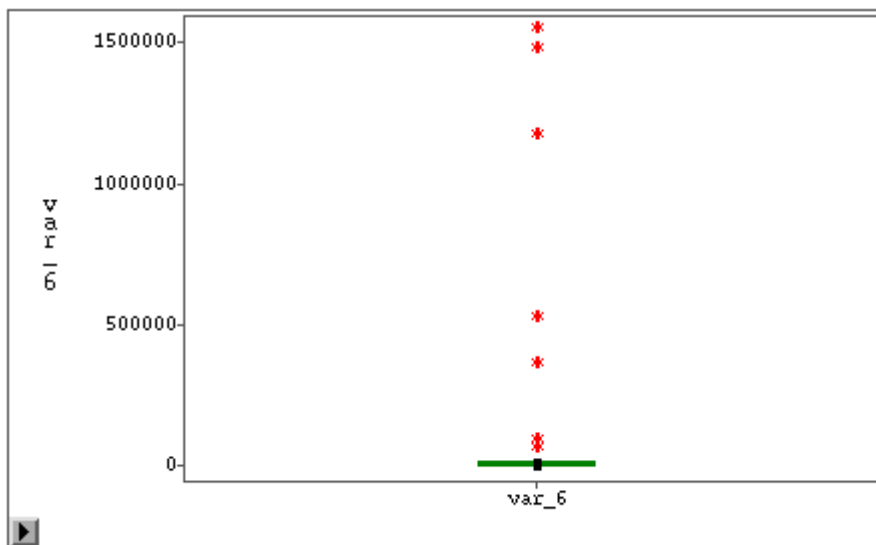
Figur 2. Model fit for corrected gross operating expenditure on the subject "woodwork" against number of students in this subject.



Example 2. "Thousand error"

27. This term means that the value is reported in NOK instead of thousand NOK. Figure 3 shows the box plot for the indicator "Fee per inhabitant connected to municipal drainage service". Here we can see that 5 relatively small municipalities report values which are ca. 1000 times higher than the median. In these cases SAS/Insight can be used to set a marginal value for a certain indicator that can be used further in automatic controls.

Figur 3. Box plot for key indicator "Fee per inhabitant connected to municipal drainage service".



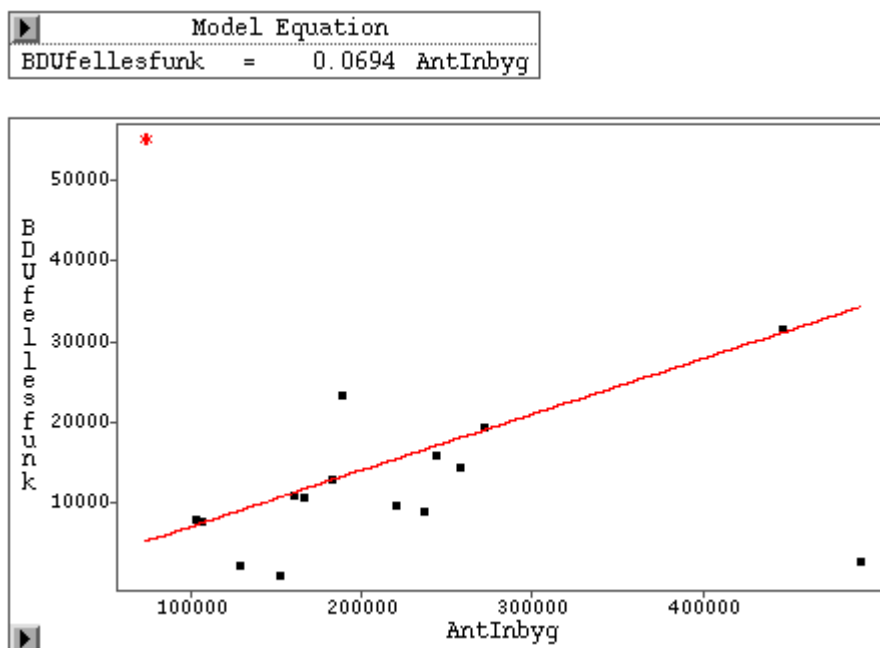
Example 3. Deviant observations

28. After elimination of observations with missing values and "thousand errors" we can conduct the model fit described in section III (A). Observations with the studentized residual $|r_{ii}| > 2$ are regarded as deviant. Professional assessment should also be used before one contacts the municipality.

29. Figure 4 shows gross operating expenses for common functions in dental health service plotted against the number of inhabitants. (Key indicator "Gross operating expenses, common functions, per inhabitant in NOK").

30. One municipality (marked with *) had a studentized residual equal to 12,65. Contact with the municipality proved an error in the accounts and the new value was reported. It turned out that several services were posted on the wrong function.

Figure 4. Model fit for the indicator "Gross operating expenditure, common functions in dental health service per inhabitant in NOK"



IV. CONCLUSIONS

31. The comparison of ordinary revision done for each working area with editing across sources has shown that observations with "zero" in denominator but not in the numerator and vice versa and deviant observations are not discovered by ordinary revision. It means that a great deal of errors is not uncovered before the different sources are put together. "Thousand errors" were for the most part detected by the editing of the separate working areas.

32. In addition, we have also seen cases when the municipality needs SNs competence to find out where the error lies. We can not expect that municipalities discover and correct the errors themselves. Good communication between municipalities and Statistics Norway is therefore important.

33. The need for editing across multiple sources will be biggest in the first years because of the necessity to train the municipalities to see the connection between the service figures and accounts, so that they can report correct figures. In that way the need for editing across sources will be reduced in the long-run.

34. Therefore, it would be worthwhile for Statistics Norway to use extra resources for including data editing across different areas into the MSR system.

V. RECOMMENDATIONS

35. Today's MSR system is quite fragmented and is made to solve specific problems like data entry, editing or production of key indicators. On the editing side several systems are used. In order to conduct MSR editing across areas in an effective way we consider it necessary to coordinate these different systems into a more homogeneous structure.

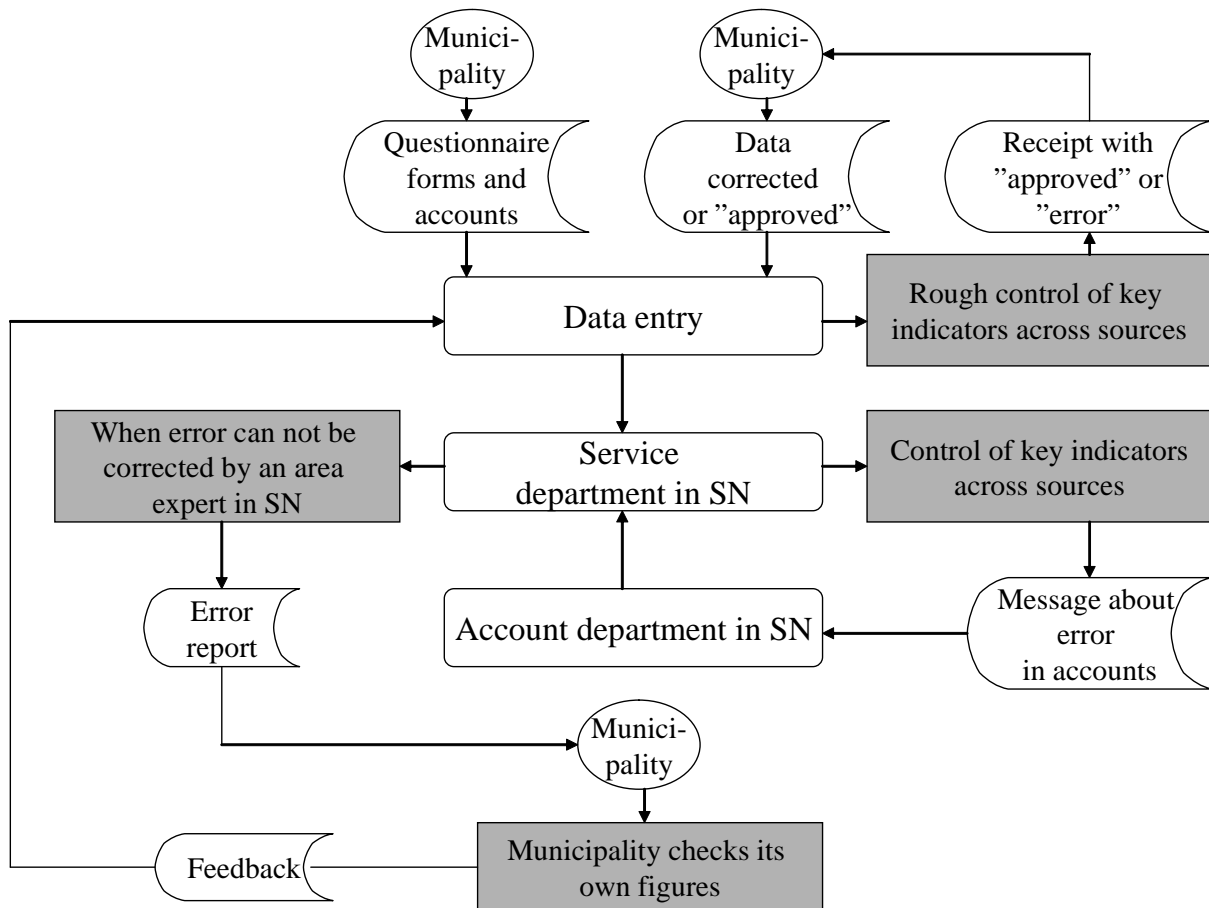
36. There are several given conditions we have to take into consideration when assessing the implementation of MSN data editing across multiple sources. The most important restrictions are:

- We can not influence the time of municipality reporting;
 - Publishing times are given: 15.03 and 15.06.
37. The external conditions also have consequences for the MSR work internally:
- Already existing high workload during the editing period;
 - Short deadlines.
38. Because of these conditions we see a strong need in an atomized routines and coordination with the ordinary editing in MSR.

A. Specifications for MSR data editing across multiple sources

39. This chapter specifies point by point the requirements for functionality that we mean should be fulfilled in order to implement the MSR data editing across sources in an effective way. Figure 5 gives an overview of the activities that editing across multiple areas should include.

Figure 5. Overview of the activities for editing across multiple sources in MSR.



40. Reporting/data entry:

- On-line reporting, where municipalities report directly to SN's server;
- In that case the control of questionnaire forms and accounts can be conducted already at the data entry point, both ordinary editing and across multiple sources;
- Automatic feedback (on the screen) in case of missing values or positive outcome of the controls;
- All types of data (from questionnaires, registers or external) must have the same data format as early in the process as possible. Preferably at or right after data-entry point;

- The aim of having most of the controls at data-entry is to detect the errors as soon as possible, which is labour-saving for both the municipality and SN. The figures will be correct at the earlier stage of the process and data editing can be reduced in the long run;
- Documentation of the control results that were conducted at data-entry point, which gives the possibility to follow up the process;
- Due to a short time span from data-entry to the first publishing we do not recommend to follow up municipalities at this point.

41. **Data editing system:**

- Data editing across multiple sources should use the same editing system as ordinary revision, i.e. there must be a common portal, common editing system for all types of MSR data (service data, account data, personnel data);
- The editing system should contain datasets on different detailing levels: key indicators on level 2, background data on level 3, and even more detailed figures where it is necessary (for example, more detailed account data);
- One should conduct statistical analysis within the editing system (model fit etc.)
- Flexibility: the data editing system should be able to create specific controls for certain MSR areas;
- Continuous loading of data from data-entry, i.e. figures must be updated in the editing system at all times.

42. **Data editing process:**

- Data editing across multiple sources should be conducted in several steps: 1.) missing values; 2.) "zero" in denominator and not in numerator and visa versa; 3.) "thousand" error; 4.) deviant observations.
- Sequence in analyzing and conducting controls: The linked dataset is analyzed to develop controls. After conducting controls we get a list of errors. It is not necessary to analyze and develop controls each year. Once they are developed they can be used every year with supplements.
- It should be easy to develop and program controls for those who use the editing system, so that one does not depend on IT expertise.
- User-friendly and automatized.
- The list of errors is generated automatically.
- Internal SN errors and "thousand" errors are corrected automatically.

43. **Communication between departments in SN:**

- The list of errors can be formed both by control type and by municipality;
- The list of controls is generated by the service area, so that the feedback from controls across sources can be sent together with ordinary controls done by the same department. (We do not recommend that a common error list is generated for all the areas, because it will be unwieldy).

44. **Communication with municipalities:**

- Easy to put together error list from ordinary editing and revision across sources, so that municipalities get the list for a certain area at the same time.
- Error list is sent to one e-mail address per municipality, preferably to a special MSR contact.
- Feedback to municipalities should contain all the details about assumed error and calculations behind the key indicator, which can help municipalities to find out where the error lies.
- Feedback from the municipality should be marked in the dataset, so that it is documented what the municipality chose to do: keep the existing figure or report a new value.

- After received feedback from municipalities the figures should be updated as soon as possible in the editing system.

45. The findings from this work have proved that data editing across sources in Municipality-State Reporting is important and highlighted the complexity of this task. The work has resulted in increased awareness and attention from management side. Two new projects have been created connected to the challenge: one that looks at the long-term possibilities, and another one that concentrates on data editing across sources with existing solutions.
