UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (iv): Evaluation and feedback

Editing and evaluation of statistics based on administrative microdata – example by Norway

Prepared by Aslaug Hurlen Foss and Ane Seierstad, Statistics Norway, Norway

I. Introduction

1. The use of administrative data in the production of statistics is increasing in Statistics Norway. It is therefore necessary to developed methods for editing, techniques for evaluation and IT-systems to handle this.

A. What is administrative data?

2. Administrative data is information on micro level units based on files from local administrative systems. It may also be called register data. This type of data is collected and processed for administrative purposes and thereafter passed on to the statistical agency. The data is thus secondary in nature. The data may not be designed for the statistical target and has therefore often to be processed before it can be used for statistics. Administrative data can for example be information on social clients, bank transactions, investment funds or employment. The statistics based on administrative data are strongly dependent on ID-identification: person identification number, enterprise identification number, housing/location identification number and other kinds of identification. The ID-identifications make it possible to add information from the basis registers: population register, business register and dwelling register. The ID is important for controlling the population and for editing. Some statistics, like the 'register-based employment statistics' are based on several administrative data merged together by the ID-numbers. Good ID identification plays a key role for the quality of the statistics.

Example of administrative data

Enterprise id	Person id	Start date	Stop date	Working hours
918273645	987654321	01.01.2015	31.01.2015	40
918273645	123456789	01.01.2015	31.01.2015	40
918273645	728163548	15.01.2015	31.01.2015	23

B. The role of local IT-systems

3. There exist usually several IT-systems that handle administrative data which the local administrative unit can choose to use. It is often very beneficial to have good contact with the IT-suppliers of the systems. Together it can be decided on which controls can be situated in the local IT-system. It is of everyone's interest that information is entered correctly into the administrative systems. Most controls are of absolute character: unlikely combination or missing values. Some statistics arrange regular meetings with the IT-suppliers to discuss changes in what should be reported to Statistics Norway and the possibility to include new controls of the data in the local It-system.

C. Data collection

4. Data is collected as files reported to Statistics Norway, either as an attachment in the government dialog system called Altinn or reported directly electronic to Statistics Norway. When the files are downloaded, some statistics run different controls before accepting the files. The ID-identification of the reporting unit must be correct and there may be some absolute controls of main variables. Some controls may just give a warning and to give the statistician an opportunity to check the data and report later.

D Outline of this paper

5. Methods for editing administrative micro-data are outlined in section 2. Evaluation of administrative micro-data is described in section 3. Statistics Norway's general It-system, Driller, to edit administrative micro-data will be described in section 4. Finally; future works with administrative microdata are outlined in section 5.

II. Methods for editing administrative micro-data

A. Macro-editing methods on aggregated data

6. Standard macro methods for editing and evaluating can often not be used directly on the data. It is necessary to aggregate micro data to a higher level before using macro-editing methods. This can be the sum of numeric values as well as sum of units or categorical variables. This is especially for applying methods like Hidiroglou-Berthelot-method and other quartile methods. For example "register based labour statistics" has to be aggregated from "person" level to enterprise level before using macro-editing methods.

Example of aggregated administrative data

Enterprise id	**Number of employees**	Sum working hours
918273645	3	103

When a suspicious unit is found on aggregated level the population for this unit has to be controlled.

B. Numeric variables

7. In administrative micro-data numeric values are rarely changed, except for missing values and "thousand error". Missing values are often controlled in the data collection phase; if it is discovered in the editing phase it is often sent back to the reporting unit. Instead, some statistics ask that 'mistakes' are corrected for the next period and choose to impute the value by the last reported value.

C. Categorical variables

8. Categorical and dating variables are often very important in administrative micro-data. Categorical values can for example be controlled by comparing with an approved coding list and combination of appropriated categories. The editing of categorical variables can often be based on expert knowledge of the subject. For example change of classification of industry for one enterprise can be very important for the statistics and it is often based on expert knowledge of the subject.

D. Population controls

9. The most important and time consuming activity is controlling the population. The population is unknown, unlike the population of reporting units which is known. Some populations change from period to period while others are more stable. The population of social clients are much more unstable than the working force in an enterprise. There can be different kinds of "duplicate populations" that need to be managed depending on the statistics. There may be also a "missing population"; that would be in the case where just part of the population is reported. This can be controlled by classical method like for example Hidiroglou-Berthelot-method on aggregated level, which controls that the variations are within a normal range between periods.

A classification of different kinds of "duplicate populations" can be useful:

a) Reporting unit sends inn data twice with equal population units and variables.
b) Reporting unit sends inn data twice with unequal population units or variables.
c) Two reporting unit sends inn data with information on the same population.
d) Two reporting unit sends inn data with information on the same population but for different periods.

An automatization of this process may improve the efficiency of the statistical production. However, the automatiszation should be controlled to ensure the quality. Situations may occur, that where not thought and need to be handled manual by subject matter specialists. If possible, the automatc procedures should later be adjusted to handle this situation. An automatization with a quality evaluation is much less time consuming than manual editing of the population. It also ensures equal treatment of all population units.

For the situations a) and b) where one reporting unit sends data twice or more, a rule can be made that the latest data are the data that are going to be used. This can be expanded by including a report on the difference between the two reported datasets, to help evaluate the correctness of this rule. The report should be on aggregate level with information on the difference of the population size and the change of variables.

For the case c) when two reporting units send information on the same population it is usually solved in different ways depending on the statistics. One population unit is manually set to the status "active", to be included in the statistics and the other as "inactive". If the population unit is stable in the statistics, one of the reporting units may be asked to continue reporting it and the other reporting unit not. Instead of regarding the units as "active" or "inactive" it is possible to assign weights to the every unit. Then the units that are not going to be used are assigned the weight zero.

Example of data with doublets with weight 0 and 1

Unit id	Variable 1	Reporting id	Weight
123456789	40 000	918	1
728163548	23 000	918	1
728163548	24 000	846	0

Instead of choosing one of the units it is possible to use a weighted mean of the two population unit, see Zhang 2012. However, this solution functions just for numerical values and not for categorical values. The weight for the equal unit can be set to 0.5 and we can then get a weighted sum of the two observation.

Example of data with doublets with weight 0.5

Unit id	Variable 1	Reporting id	Weight
123456789	40 000	918	1
728163548	23 000	918	0.5
728163548	24 000	846	0.5

Results of data with doublets with weight 0.5

Unit id	Variable 1	Reporting id 1	Reporting id 2
123456789	40 000	918	
728163548	23 500	918	846

In the case d) where two reporting units send data with information on the same population, but for different periods, we want to have the sum. This can be done by setting the weight to 1 and compute the sum of the value. However, for counting the population, the weight has to be 0.5

Example of data with doublets for two different periods

Unit id	Variable 1	Reporting id 1	Weight	Weight population
123456789	40 000	918	1	1
728163548	23 000	918	1	0.5
728163548	24 000	846	1	0.5

Results of data with doublets with doublets for two different periods

Unit id	Variable 1	Reporting id 1	Reporting id 2
123456789	40.000	918	
728163548	47.000	918	846

For categorical variables, one has to choose which unit the categorical variables should be taken from. In cases were the categorical values arc missing, the categorical value can be a mixture of the two units.

With using weights the solution for handling the populations becomes generic and flexible. Now most statistics control the population manually or by IT-solutions that are tailor-made to the statistics. In the future it would be a great advantage to develop generic solutions for population controls.

E. From administrative micro-data to statistical data

10. For some administrative data the gap between the statistical definition and what is reported is huge. Data has then to be processed to fit the statistical definition. Often administrative data from several sources has to be integrated and variables from population register have to be added, see Zhang 2012 for general theory on the topic or Fosen 2011 for practical example from "Register-based employment statistics".

III. Evaluation of statistics based on administrative micro-data

11. Generally in Statistics Norway there are very few statistics that evaluate the effect of editing data. In 2004/2005 Statistics Norway implemented a project that among other things resulted in SAS-macros to evaluate the effect of editing. The SAS-macros were based on the comparison of original and edited dataset, and it is possible to have a more specific evaluation on numeric and categorical variables separately. The programs were available for all statistics, but not integrated with the general editing-system. It was used in courses in statistical editing, but beyond this it was hardly integrated in any production of statistics. The reasons for this are probably that the management did not emphasise that this was useful a process in statistical production, the lack of knowledge about methods of evaluation and that there has been no easy access to software for the users. Recently the management has become aware of the usefulness of evaluation as a tool to improve efficiency and data quality. They have therefore asked that this is included in the plans for the modernisation of the general editing system. However, administrative micro-data have a different character than survey data and therefore the evaluating has to be adopted to suit the needs of administrative micro-data.

12. Evaluation is usually done having a given population and evaluating the changes done to the variables. It is then important to have a system that records all the changes to the variables. In administrative micro-data the population is not given and the largest change to the data, is the change of

the population. In addition to recording all changes done to the variables, all changes to the population should be recorded as well, in a way that makes it possible to do an evaluation. The occurrence of duplicate reporting units should be recorded and classified in different categories. Then one should calculate how this changes the total population size. Finally the changes of the variables on aggregated level should be calculated.

Draft of report for population changes

Classification of duplicate units	Number	Population sum of changes	Variable sum of changes
a) Duplicate reporting unit, equal population and variables			
b) Duplicate reporting unit, unequal population or variables			
c) Two reporting units send in data with information on the same population			
d) Two reporting units send in data with information on the same population but for different periods			

13. When the evaluation of the population change is done, the evaluation of the editing of variables for fixed population can be done. The report on changes of population and variables should be made into one report on the total changes of the data. The report should be used to monitor the overall quality and to suggest quality improvements that have grate advantages. Specific evaluation reports on aggregates, like industry and other classifications, are often wanted. It is especially important that id identification of the population units and reporting units are stored so that it is possible to identify problematic observations. Because of the huge amount of data it is very convenient to have the possibility to aggregate and drill down to lower levels easily. It is not possible to contact the population units, just the reporting unit. Therefore it is often very useful to aggregate to the level of reporting unit. The aggregated report can be used to follow the quality of the data from the reporting unit or use it as feedback if it is possible. Some statistics wants to have feedback from reporting unit on a special population unit. This can however sometimes be difficult because of the statistical law on confidentiality. The communication has to be done in a secure way, which ensures that information on population units is just accessible by the authorized persons.

14. For evaluating categorical value, the numeric values associated to the category are usually of main interest. The categorical variable may have several categories. One way to do this analysis is then to make a two-way table with the categorical variable before and after editing and in the cells have sum of the numeric value.

IV. Driller: General IT-system for editing administrative micro-data

A. ISEE- Integrated System for Editing and Estimation

15. ISEE is a general IT system under development in Statistics Norway. The first modules were built in 2004: Dynarev for micro-editing, Struktur for estimation and Pris for some price indexes. The system is built on four important principles; metadata driven processes, general functionality, self-service and reuse. The original idea was to integrate estimation with editing (Zhang, 2008). However, few changed their production system to integrate editing with estimation. The reason for this was probably lack of management and the design of the modules. In 2014, 134 surveys used the general application Dynarev for micro-editing. When starting to use Dynarev each survey has to build their own view of data and choose appropriate controls. For macro-editing the statistics have to develop their own systems in

SAS or other software. Dynarev was original built for small surveys and a choice was made that data should have vertical table structure. This data structure is inconvenient for big datasets because the performance then is very slow.

B. Driller

16. In 2012 Statistics Norway therefore built a new general editing system with a traditional data table format. This new application is called Driller, and was created to meet the demands of administrative micro-data. In 2014 25 surveys used Driller to edit their survey and one of the surveys has 18 million records each month. The database is built in Adobe Flex (Oracle) and the view for the users is in the browser FireFox. Groovy/Grails are used as server application between the client and database. The last year one has tried parallelisation and partitioning the data to increase the performance of Driller. In an administrative module every survey has to define the variables and which of them that are going to be used in aggregation. In this module the micro editing controls has also to be set up. Driller and Dynarev are linked so it is possible to view and use both applications with the same data.

Functionality of the system is: view to the data, search and filter, aggregation and drill down, automatization and rule updating and reports on evaluation.

17. *View to the data:* The large amount of data makes it difficult to view all in one window. A module has therefore been built, where the user can set up different views to look at the data. This means that the user can specify which variables that are going to be viewed and which variables are aggregated or sorted. These views are stored and are reused each new period.

18. *Search and filter:* An important functionality is to search for special units and to filter groups of data. This gives the possibility to study more closely observation of special interest.

19. *Aggregation and drill down:* Data-editing of administrative micro-data involves checking data on different aggregated levels. Driller was designed to be an efficient tool in this process. Within this system, the user can easily define on what level of aggregation he/she wants to check the data, and drill down to lower levels to find the problematic groups or observations.

20. *Automatization and rule updating:* The user can define edit rules that can be run automatically or by the user every time new data is entered into Driller. The edit rules are constructed in the same way as controls, but with the addition that it can be assigned values when the edit rule is true. However, too many automatic edit rules reduce the efficiency of the system.

21. *Reports on evaluation:* A first version of a report module is created for the largest survey. The report can be used to keep track of how many observations that were changed due to the automatic edits, or give figures on different aggregated levels before and after editing. In that way the same checks can be executed every time, even if the editing of a specific number might be done on the basis of manual consideration. The report module is also built so that it is possible to drill down from a higher level to lower levels to check where the problem might originate.

V. Conclusion and future work

22. A new project to modernise the ISEE system in general is just initialized. One of the main issues is how to build an even better editing system for administrative micro-data data. We want to look at the possibilities for improving the report module for evaluation and facilitate generic population controls. One other goal is to build a system in which one can make use of macro editing in an efficient way. For administrative micro-data this means using data on aggregated level. One suggestion is to build a library of methods or a toolbox for macro-editing and other methods used in production of statistics. The idea to construct a toolbox with methods has also been suggested by Statistics New Zealand (2014) in the paper “Methodology Architecture”. The IT-architects at Statistics Norway are now a discussing the use of the software R in developing the methods in the library, and other software may handle the communication

between the library and the databases. With using program R to construct the methods in the library it would be easy to share programs with others statistical agency.

References:

Fosen, Johan (2011). Register-based employment statistics. A case of microintegration. WP4.1. of ESSnet Data Integration

Seyb, Allyson and Darnbrough, Jeni (2014). Methodology Architecture. Statistics New Zealand

Wallgren, A. and B. Wallgren (2007), Register-based statistics – Administrative data for statistical purposes, John Wiley and Sons, Chichester.

Zhang, L.-C., Faldmo, M. I., and Lien, O. K. (2008). ISEE - Integrert System for Editering og Estimering. Paper in Norwegian presented to the 24th Nordic Meeting of Statisticians. Reykjavik.

Zhang Li-Chun (2012) Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica, Volume 66, Issue 1, pages 41–63, February 2012