WP.38
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 (vii): Indicators for measuring the quality impact of data editing and imputation

OVERVIEW ON EDITING AN IMPUTATION PRACTICES AT STATISTICS FINLAND

Supporting Paper

Prepared by Pauli Ollila, Statistics Finland

ABSTRACT

As the first part of the editing project at Statistics Finland a study of the practices of editing and imputation in different statistics of Statistics Finland together with possible needs for development is carried out. This process includes personal discussions with the researchers of some essential statistics of Statistics Finland together with studying additional documentation concerning editing and imputation in those statistics, mainly reports of the statistical auditing of 28 statistics during 2007 - 2009. The paper to be presented in the UNECE workshop is an overview on the results of these studies obtained by the end of July 2009.

Keywords: editing, imputation, quality

I. BACKGROUND

1. Due to administrative and personnel reasons the editing project of Statistics Finland began in Summer 2009. The first aim of the project is to study the practices of editing and imputation in different statistics of Statistics Finland together with possible needs for development. The main practice for this study is to have personal meetings between persons of some statistics and the editing project together with additional information on the topic, where different aspects of data collection, editing and imputation are dealt with. By the end of July 2009 there have been four meetings of this kind, and three memos covering the topic in statistics in question have been written. The target is to cover several statistics from different departments of Statistics Finland by the end of the year.

2. In addition, a specific E&I study based on existing reports is made. Since 2007 28 statistics have gone through the process of *statistical auditing* (situation at the end of July 2009). The reports covering these meetings and the background work are meant to describe all essential aspects concerning the statistics *(administrative information; planning and management; relations to the users of statistics; competence of personnel; information technology; data collection; data processing; distribution of results; documentation and archiving; monitoring, evaluation and development)*. Now these documents have been studied in order to find information on editing and imputation practices, mainly from the sections ‘data collection’ and ‘data processing’. It must be noted that **these reports are not written from the E&I point of view**, so some issues concerning E&I in those statistics might not be brought into the discussion. In addition, some of the descriptions of the older reports could have been altered lately, so the **reports are not a fully sufficient source of a study**. This paper is based on the study of the reports of statistical auditing.

II. EDITING AND IMPUTATION IN STATISTICS INCLUDED IN STATISTICAL AUDITING

Data Collection

Statistics in Auditing (28)	Data Collection									
	Interview by phone	Personal interview	Interview in places	Response by mail	Response via internet	Response via email	Register of StatFi	Data of StatFi	Other register /source	Web-sites
Social Statistics (6)										
Household Budget Survey		Y		Y			Y		Y	
Occupational Accidents									Y	
Income Distribution Survey	Y	Y					Y		Y	
Border Interview Survey			Y							
Use of Information and Communic. Technology	Y						Y			
Job Vacancy Survey	Y				Y		Y			
Population Statistics (4)										
Educational Structure of Population				Y	Y		Y		Y	
Population Structure							Y		Y	
Bankruptcies							Y		Y	
Causes of Death							Y		Y	
Prices and Wages (6)										
Index of Wage and Salary Earnings							Y	Y		
Producer Price Indices of Services					Y	Y		Y	Y	Y
Producer Price Indices					Y	Y		Y	Y	Y
Local Government Sector Wages and Salaries					Y	Y				
Finance of Housing Corporations				Y	Y				Y	
Prices of Dwellings									Y	
Economic Statistics (4)										
Travel Account Statistics								Y	Y	
Regional Accounts of Production and Empl. Stat.							Y	Y	Y	
Consumer Survey	Y									
Quarterly Statistics on the Finances of Municipalities					Y		Y			
Business Structures (4)										
Use of Information Technology in Enterprises				Y	Y		Y	Y		
Inquiry on Manufacturing Commodities	Y			Y	Y		Y			
Research & Development				Y	Y				Y	
Environmental Protection Expenditure in Industry				Y	Y	Y	Y			
Business Trends (4)										
Building Cost Index				Y	Y		Y	Y		Y
Statistics on Manufacturing & Trade Inventories				Y	Y		Y		Y	Y
Volume Index on Newbuilding									Y	
New Orders in Manufacturing				Y	Y		Y		Y	Y
NOTE: Aspects in reality existing in the statistics in question can be missing in the table (see the text).										

3. The special nature of the *Social Statistics* department (SS) concentrating on **surveys based on interviews** can be noticed clearly in the reports. Half of the audited SS statistics were based on telephone interviews. There are only few percents of personal interviews in the *Income Distribution Survey*, but all the interviews of the *Household Budget Survey* are personal. In addition the diaries of the HBS are sent by the respondent by post to Statistics Finland, and receipts can also be directed to StatFi for scanning. The nature of the *Border Interview Survey* is different with its groups of interviewers distributed to various cross-border points. The auditing included only two non-SS statistics which were carried out based on interviews, i.e. the *Consumer Survey* and the *Industrial Output Statistics*.

4. According to the reports, **enquiries by mail** are common especially in *the Business Trends* department (BT) and the *Business Structures* department *(BS).* The development during recent years can

be seen in utilising the **internet surveys** in the *Prices and Wages* department (PW) together with BT and BS. The systems which are applied vary: some statistics use XCOLA, some have their own applications and some systems are developed and maintained by external producers. As an alternative, some statistics in PW **use email** for sending prepared excel tables for the respondents, mainly in order to get price information.

5. **The use of registers of Statistics Finland** in the production of statistics is usual: as an independent source, attached together with the collected survey data at the observation level or merged into some other source of data. The Central Population Register (originated from the Population Register Centre) is in use in several statistics of Statistics Finland concentrating on persons and households, mostly in the *Population Statistics* (PS). The Register of Enterprises and Establishments is essential especially for the BT and BS departments.

6. **The data material collected by Statistics Finland** is reused in some statistics. In auditing, this kind of practice existed in the *Price of Dwellings Statistics*, the *Travel Account Statistics* and the *Regional Accounts of Production and Employment Statistics.* In the Prices and Wages department the data which are produced is utilised in various occasions.

7. **The external registers and data material** are used quite widely in departments. The sources vary from several registers and statistical data of different authorities to the data material of organisations and institutes (in all 42 sources counted in the auditing reports).

8. Some statistics utilise the registers and data materials when constructing the survey weights. **The websites of relevant sources** provide information for some statistics included in auditing, concentrating mainly on a few price values to be used in the calculation of indexes.

III. RECOGNISING ERRONEOUS VALUES

Statistics in Auditing (28)	Recognising Erroneous Values													
	Comm. of int.	Man. check	Softw. cond.	Logical checks	Ext-remes	Distr. & graph.	Prev err.	Prev. values	Non-valid v.	Est. param.	Prev par.	Other sour.	Input checks	Weight control
Social Stat. (6)														
Household B.S.	X	X	X	X	X	X	X		X	X		X		X
Occupat. Acc.											X			
Income Distr. S.	X	X	X	X	X	X	X	X	X	X	X	X		X
Border Interv.S.	X	X	X	X	X	X	X		X	X	X	X	X	X
Use of Inform. & Comm. Tech.	X		X	X						X	X			
Job Vacancy S.	X	X	X	X		X				X				
Popul. Stat. (4)														
Educ. Struct. of Population		X		X				X				X		
Population Str.		X		X				X						
Bankruptcies		X	X	X					X					
Caus.of Death		X	X	X			X		X			X		
Prices and Wages (6)														
Index of W&S Earnings		X		X				X	X			X		
Producer Price Ind. of Services		X	X	X				X	X	X	X	X		
Producer Price Indices		X	X	X				X	X	X	X	X		
Local Gov. Sector W&S		X	X	X					X	X	X			
Finance of Housing Corp.			X	X					X	X				
Prices of Dwell.		X	X	X					X					
Economic Statistics (4)														
Travel Acc. Stat.										X		X		
Reg. Accounts of Prod. & Empl. S.										X	X	X		
Consumer Survey	X	X	X	X						X	X			
Quart. St. on the Fin. of Municip.		X	X	X	X	X		X	X	X	X	X		X

Business Structures (4)														
Use of Inf.Tech. in Enterprises		X	X	X				X	X	X	X	X	X	
Inq. on Manuf. Commodities		X	X	X					X	X				
Research & Development		X	X	X					X	X	X			
Envir. Prot. Exp. in Industry		X	X	X						X	X	X		
Business Trends (4)														
Building Cost Index		X	X	X	X		X	X	X	X	X	X		
Stat. on Manuf. & Trade Inventories		X	X	X				X	X	X	X	X	X	
Volume Index on Newbuilding		X	X	X	X	X		X	X	X	X	X		
New Orders in Manufacturing		X	X	X			X	X	X	X	X	X		
NOTE: Aspects in reality existing in the statistics in question can be missing in the table (see the text).														

9. Almost all the audited surveys with interviews utilised the **comments of interviewers** for finding errors. Furthermore, the **manual check of the material** through the database, listings and/or questionnaires is ordinary in those statistics which have their own data collection. However, the resources allocated to this checking process vary clearly from one statistics to another. Challenging statistics can be found especially from the Business Trends and the Business Structures. The development of automatic mass check processes and in some cases mass editing is a future goal for some statistics.

10. The software used in the interviews (Blaise in Finland) and in the data collection via internet (e.g. XCOLA), the excel tables send via email and the input software for paper questionnaires all included possibilities to **make conditions and restrictions** for entering the data, in order to prevent giving erroneous information. This possibility is utilised well in the auditing statistics.

11. Excluding few statistics obtaining their material from other source, all the statistics in auditing have built **logical checks** in their data processing in order to find errors in the data.

12. The **study of extreme values** (minimum, maximum, some percentage or count usually from the upper end of values). This practice is common especially for the quantitative variables, and it often reveals outliers, which are clearly exceptional in respect of other values. The study of auditing reports reveals that the practice is not very usual among statistics, mainly concentrating on some surveys based on interviews. However, it is apparent that at this point not all statistics have mentioned this practice in the discussions, although they might use it.

13. The **distributions and relations of the variables** in the data can be studied in the form of **tables** and **graphics** in order to find errors. As before, it is probable that some simple tabulations of the data are not mentioned in the reports, possibly considered self-evident. On the other hand, graphical studies of the data are not widespread among the statistics in auditing, and the possibilities of this tool in the contexts of different statistics could be studied much more.

14. The statistics conducted previously provide valuable experiences on the quality of the data, and often there are variables or combinations of questions which appear to be technically problematic or the respondents / data providers having definitional or practical difficulties. Some statistics utilise the **knowledge of error types** from the previous data collections of the statistics in order to find new errors during the data processing.

15. The **variable values of the previous data collection** are used for error recognition rather broadly, when some of the observations remain at least two rounds. This practice is essential especially in the context of index calculation, but e.g. the *Population Structure Statistics* and the *Income Distribution Survey* control their data in this sense as well.

16. The software of data collection include restrictions and conditions, but many statistics set value limits or combinations in the phase of data processing for **revealing non-valid values** in the data.

17. **Estimated parameters and quality measures of the data** can also describe errors to some extent, and according to the auditing reports this practice is rather common in the statistics. The **previously estimated parameters** are studied frequently as well among the statistics in order to discover surprising changes.

18. Most of the statistics in the study control their results with the **results from other sources** (registers, statistics and other research organisations) and this practice is a tool for recognising oddities in the data.

19. The data, which come as questionnaires or other paper form, are stored either at Statistics Finland or by an external company. The **results of input are controlled** with random checks in order to assess the quality of entering the data.

20. In some cases combinations of big expanding survey weights and high variable values can cause instability even in the estimation of results at the overall level. This phenomenon is controlled in some surveys. The result can be an adjustment of the weights or the exceptional values (probably being outliers).

IV. DEALING WITH MISSING VALUES AND ERRORS

Statistics in Auditing (28)	Dealing with Missing Values and Errors											
	Missing val. not at all imputed	Value from the material	Values from other observat.	Values from other sources	Asking provider	Funct. correct.	Correction by reasoning	Mean value imputation	Regr imp.	Hot deck imp.	Cold deck imp.	Distrib. imput.
Social Stat. (6)												
Household B.S.		**X**		**X**		**X**	**X**	**X**				
Occupat. Acc.					**X**							
Income Distr. S.				**X**		**X**	**X**			**X**		
Border Interv.S.		**X**			**X**	**X**	**X**	**X**				**X**
Use of Inform. & Comm. Tech.							**X**					
Job Vacancy S.							**X**	**X**				
Popul. Stat. (4)												
Educ. Struct. of Population					X							
Population Str.			X			X						
Bankruptcies							X					
Caus.of Death		X			X							
Prices and Wages (6)												
Ind.of W&S Earn	X											
Producer Price Ind. of Services			X	X	X			X			X	
Producer Price I.				X	X			X			X	
Local Gov. Sector W&S							X				X	
Finance of Housing Corp.							X					
Prices of Dwell.							X					
Economic Statistics (4)												
Travel Acc. Stat.				X				X				
Reg. Accounts of Prod. & Empl. S.												
Consumer Survey							X					
Quart. St. on the Fin. of Municip.					X	X	X	X			X	
Business Structures (4)												
Use of Inf.Tech. in Enterprises		X				X	X					
Inq. on Manuf. Commodities				X							X	
Research & Development				X		X	X				X	
Envir. Prot. Exp. in Industry				X			X					
Business Trends (4)												

Building Cost I.				X	X	X	X	X		X		
Stat. on Manuf. & Trade Inventories				X	X	X	X					
Volume Index on Newbuilding				X			X					
New Orders in Manufacturing	X						X					
NOTE: Aspects in reality existing in the statistics in question can be missing in the table (see the text).												

21. It must be noted that the corrections mentioned in the table vary from very rare cases to very common practices. The data of the *Index of Wage and Salary Earnings* and the *Regional Accounts of Production and Employment Statistics* are not processed by the producers of these statistics.

22. The **data material based on paper form** can be the basis for corrections. The cases mentioned here are questionnaires or inquiries except the death certificates from the city administrative courts in the *Causes of Death Statistics*.

23. The **values obtained from another observation** are rarely used in corrections: in the *Population Structure* some new-born babies get values from their mother and in the *Producer Price Indices of Services* the enterprise which has quit will have price changes of the enterprise which the other has been merged into.

24. Many statistics in auditing have parts, where **values from other sources** are utilised. These are usually either price, product or benefit values, which can be strictly directed with other information of the observation. Some values are collected from the websites.

25. The practice of **asking the data provider** (e.g. enterprises, municipalities, communities, organisations and institutes) usually by phone, email or mail in some unclear cases is rather common in most statistics included in auditing. Exceptions are surveys based on persons, where there are usually no contacts after the interview or questionnaire has been finished.

26. The **functional correction**, where some other variable values of the observation strictly define the corrected value in the data processing, is used to some extent. This practice can be interpreted as deterministic imputation in this context. Often there are justified inference rules behind the correction.

27. The cases, which are not within functional correction and still to be edited, can be corrected by **reasoning from other information** of the observation. This kind of correction appears to some extent among different statistics. Sometimes the value must be defined, although there is not much information to base the decisions.

28. The **mean imputation** is used rather frequently in the statistics appearing in auditing. Often the mean values are calculated within some subgroup which is relevant in respect with imputation and the phenomenon to be studied.

29. The **regression imputation** is not used according to the auditing reports, though the method was used before in the *Income Distribution Survey*.

30. The **hot deck imputation**, i.e. fetching donor values among the other observations of the data, was mentioned in context of the *Income Distribution Survey* and the *Building Cost Index*, both using the nearest neighbour method.

31. Correspondingly the **cold deck imputation**, i.e. bringing the value of the previous data, was a practice (among other practices) in some statistics, where this kind of continuity exists.

32. The **distribution imputation,** i.e. the value definition based on the distribution of a categorical variable, was mentioned in the auditing report of the *Border Interview Survey*.

V. CONCLUSIONS

33. The study of 28 statistics taking part in the auditing process was interesting from the editing and imputation point of view. It clearly reveals that many statistics have developed good practices of data collection and processing, and some new tools and methodologies are introduced in production. However, there are needs for development as well, e.g. in the field of automated checking of the data, error recognition, and mass processing together with specific graphical tools for studying the data. In addition, most of the statistics included in auditing do not deal with editing and imputation as a process including different phases and well-defined practices for different cases in order to get a full picture of the quality of these operations and the traces of what have been done. One goal of the editing project is to provide tools and good practices in order to fulfil the needs of different statistical branches of Statistics Finland.