Statistics Finland

Q2010 Training Courses

Editing and imputation

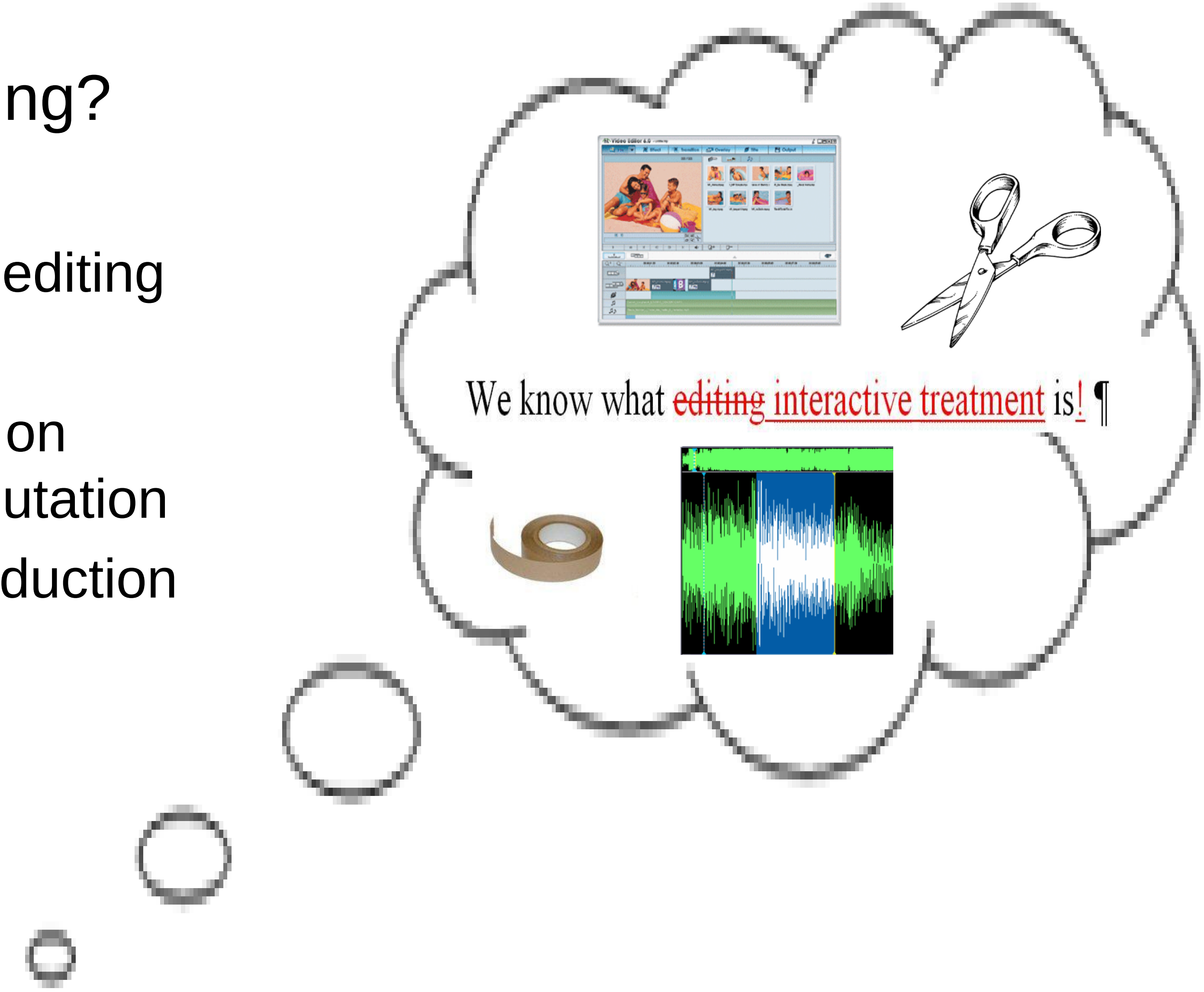
Editing and imputation

- Time
  - Monday 3 May 2010
  - 09:30-17:00
- Venue
  - Statistics Finland
  - Työpajankatu 13, FI-00580, Helsinki, Kalasatama
- Instructors
  - Prof Seppo Laaksonen (University of Helsinki, part 2)
  - Dr Pauli Ollila (Statistics Finland, part 1)

1 What is editing?

Structure

- Statistical data editing
- Basic concepts
- Different views on editing and imputation
- Editing and production of statistics



We know what ~~editing~~ interactive treatment is! ¶

STATISTICAL DATA EDITING

- EDIMBUS project: *Detection of missing, invalid or inconsistent values.*
- EUREDIT project / Ray Chambers: *Editing is the process of detecting errors in statistical data*
- UNECE glossary (editing procedure): *The process of detecting and handling errors in data. It usually includes three phases:*
  - *the definition of a consistent system of requirements,*
  - *their verification on given data, and*
  - *elimination or substitution of data which is in contradiction with the defined requirements.*

Basic concepts by EDIMBUS

Editing: Detection of missing, invalid or inconsistent values.

Adults	2
Children	3
Size of hh	6

Interactive treatment: Computer aided manual treatment of values flagged as erroneous during editing usually directly performed on the computer and assisted by implemented edit rules.

Children from register:
John, Sally, Ted, Ann

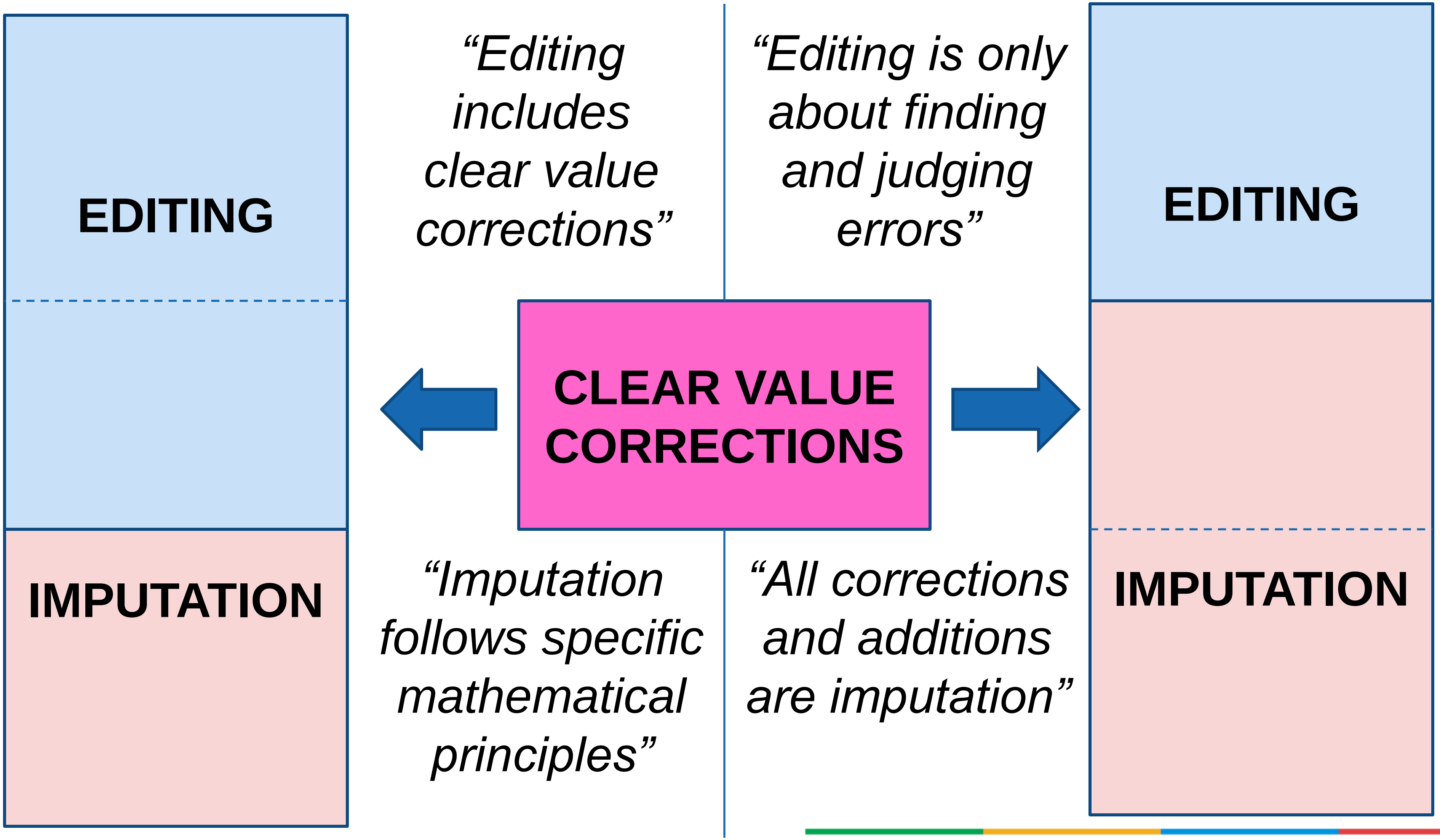
Adults	2
Children	4
Size of hh	6

Imputation: Imputation is the treatment of data used to treat problems of missing, invalid or inconsistent values identied during editing. This is done by substituting estimated values for the values flagged during editing and error localization.

Adults	_ <- 2
Children	_ <- 3
Size of hh	5

Imputing the most common alternative

Different views on editing and imputation



EDITING

IMPUTATION

“Editing includes clear value corrections”

“Imputation follows specific mathematical principles”

CLEAR VALUE CORRECTIONS

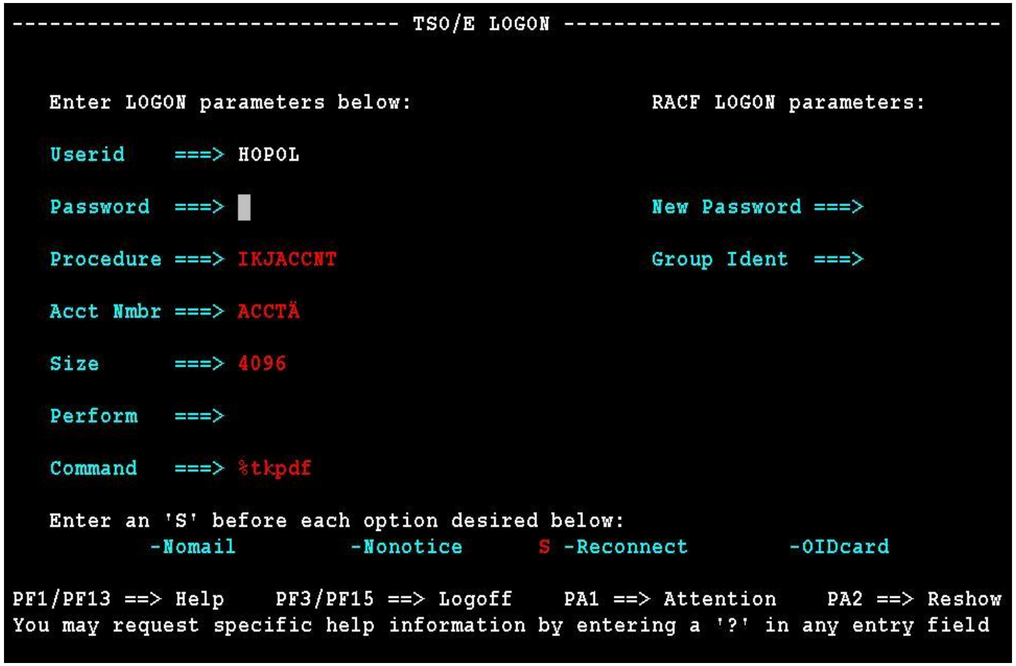
“Editing is only about finding and judging errors”

“All corrections and additions are imputation”

EDITING

IMPUTATION

Editing and production of statistics



```
------------------------------ TSO/E LOGON -----------------------------------


   Enter LOGON parameters below:                   RACF LOGON parameters:

   Userid    ===> HOPOL

   Password  ===>                                  New Password ===>

   Procedure ===> IKJACCNT                         Group Ident  ===>

   Acct Nmbr ===> ACCTÄ

   Size      ===> 4096

   Perform   ===>

   Command   ===> %tkpdf

   Enter an 'S' before each option desired below:
           -Nomail         -Nonotice      S -Reconnect        -OIDcard

PF1/PF13 ==> Help    PF3/PF15 ==> Logoff    PA1 ==> Attention    PA2 ==> Reshow
You may request specific help information by entering a '?' in any entry field
```

- Editing and correcting of statistical data often appears to be developed gradually by experience, being some kind of ”ad hoc” practice.
- Some practices can be crystallised to a quick and effective form, but some may be resource-intensive and time-consuming, not using newest techniques and methods.
- Evaluation and possible systematization of editing, correction and imputation processes might be beneficial when considering the quality, speed and resource use of statistical production

Work time spent for editing and imputation in statistics (%)

STATISTICS DEPARTMENT	Missing	0 - 10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	ALL
Population Statistics	2	23	7	1	4	4	1	0	9	51
Social Statistics	0	9	4	1	2	0	2	0	0	18
Prices and Wages	1	11	3	2	3	1	1	2	0	24
Economic Statistics	8	8	4	2	0	3	3	0	3	31
Business Trends	0	11	3	2	2	0	1	5	0	24
Business Structures	2	13	1	5	2	1	0	1	12	37
ALL	13	75	22	13	13	9	8	8	24	185

In this presentation, the results describing the basis and the practices of editing and imputation at Statistics Finland are from the internal survey of E&I practices of Statistics Finland, conducted in January 2010.

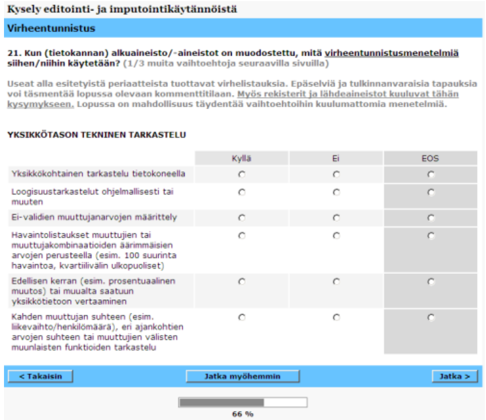
2 Different contexts of editing

Structure

- Data types
- Type of data in making statistics at Statistics Finland
- Statistics with no unit-level processing
- Combining different data sets

Data types

Survey data



Kysely editointi- ja imputointikäytännöistä

Virheentunnistus

21. Kun (tietokannan) alkuaineisto/-aineistot on muodostettu, mitä virheentunnistusmenetelmiä siihen/niihin käytetään? (1/3 muita vaihtoehtoja seuraavilla sivuilla)

Useat alla esitetyistä periaatteista tuottavat virhelistauksia. Epäselviä ja tulkinnanvaraisia tapauksia voi täsmentää lopussa olevaan kommenttitilaan. Myös rekisterit ja lähdeaineistot kuuluvat tähän kysymykseen. Lopussa on mahdollisuus täydentää vaihtoehtoihin kuulumattomia menetelmiä.

YKSIKKÖTASON TEKNINEN TARKASTELU

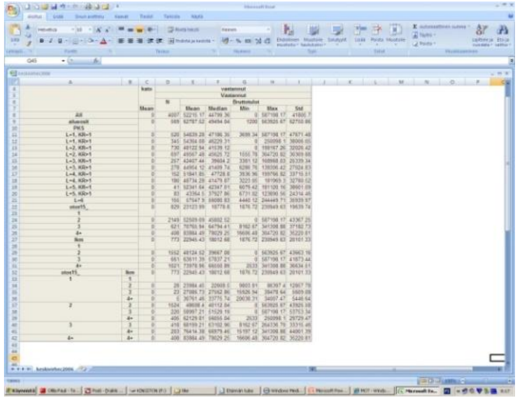
	Kyllä	Ei	EOS
Yksikkökohtainen tarkastelu tietokoneella	○	○	○
Loogisuustarkastelut ohjelmallisesti tai muuten	○	○	○
Ei-validien muuttujanarvojen määrittely	○	○	○
Havaintolistaukset muuttujien tai muuttujakombinaatioiden äärimmäisien arvojen perusteella (esim. 100 suurinta havaintoa, kvartiilivälin ulkopuoliset)	○	○	○
Edellisen kerran (esim. prosentuaalinen muutos) tai muualta saatuun yksikkötietoon vertaaminen	○	○	○
Kahden muuttujan suhteen (esim. liikevaihto/henkilömäärä), eri ajankohtien arvojen suhteen tai muuttujien välisten muunlaisten funktioiden tarkastelu	○	○	○

- Created for the purposes of one or several statistics
- Planning and controlling possible (data collection, E&I processing etc.)
- Changes can be done

Register

- Administrational purposes as the basis
- Different definitions, classifications etc.
- Rigid structures, not easily modified

Source data



- Often collected for other purposes, including different classifications etc.
- Various collection methods
- E&I processes and quality aspects may not be well known

Type of data in making statistics at Statistics Finland

STATISTICS DEPARTMENT	SUR	REG	SOU	SUR REG	SUR SOU	REG SOU	SUR REG SOU	ALL
Population Statistics	0	12	7	4	4	9	15	51
Social Statistics	1	2	2	10	1	1	1	18
Prices and Wages	0	1	1	2	12	0	8	24
Economic Statistics	4	0	8	4	4	1	11	32
Business Trends	0	2	0	10	0	1	9	22
Business Structures	4	1	2	4	1	5	21	38
ALL	9	18	20	34	22	17	65	185

Statistics with no unit-level processing (and editing)

- **Collecting statistics** utilises statistics and tabulations from several sources, and after gathering information the required form of the statistics is reached.
- **Strict processing statistics** is based on one or more data (statistical data, external source data or register), which are used strictly without changes in order to make the statistics.
- **Calculation model statistics** lean on existing, already edited data and/or tabulations/statistics in such way that with using them one can realise a mathematical or statistical calculation model required by the statistics.

STATISTICS DEPARTMENT	Unit-level processing	No processing of units	All
Population Statistics	44	7	51
Social Statistics	15	3	18
Prices and Wages	22	2	24
Economic Statistics	24	7	31
Business Trends	22	2	24
Business Structures	32	5	37
ALL	**159**	**26**	**185**

Combining different data sets

- can cause various problems (non-matches, incoherences, contradictions, differing concepts & classifications etc.)

Including statistics with unit processing **STATISTICS DEPARTMENT**	No answer	Only one unit level data set	Data set combinations done	More than one data set, no combinations	**All**
Population Statistics	2	17	15	10	44
Social Statistics	0	6	8	1	15
Prices and Wages	3	6	10	3	22
Economic Statistics	0	6	16	2	24
Business Trends	0	12	5	5	22
Business Structures	1	7	10	14	32
ALL	6	54	64	35	159

3 Editing strategy: criteria and planning

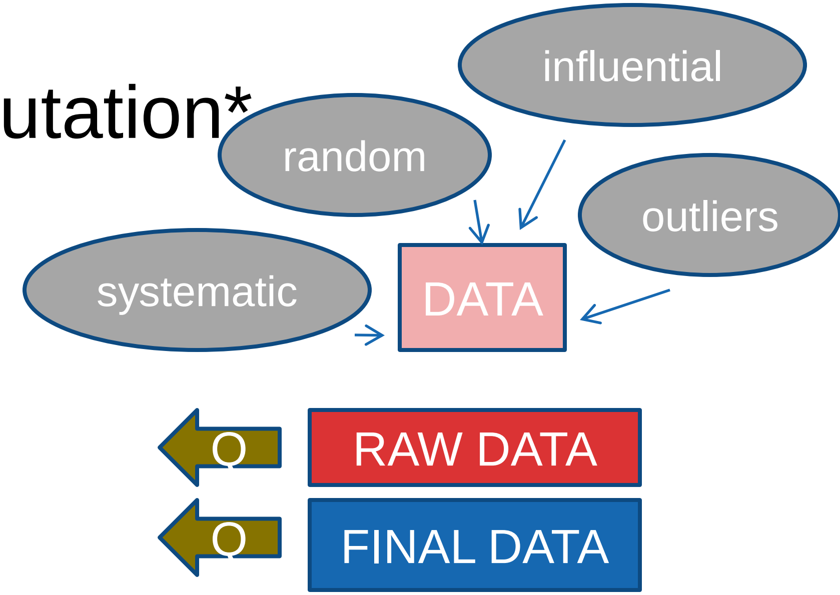
Structure

- Main objectives of editing and imputation
- Micro and macro editing
- General rules for building an E&I strategy
- Notions of the editing and imputation process flow
- Phases and procedures of editing and imputation
- Four phases of editing
- Micro and macro E&I interaction
- Key elements for the design of editing and imputation
- Recommendations for planning the E&I strategy

Main objectives of editing and imputation*

1. **identify error sources** in order to provide information for future improvements of the survey process
2. provide information about **the quality of the incoming/outgoing data**
3. identify and treat **the most significant errors**
4. when needed, provide **complete and consistent (coherent) individual data**

* by *Granquist*



Company	Turnover	R&D
VeryBig Ltd.	23 482 272	20 542 221
Tiny & Bros.	51 221	45

286099	10666	8880.28	68.0326
47005	6157	5979	11.1776
347108	9036	7257.03	82.5401
66294	7084	6957.15	15.7643
237600	10788	9489.48	56.4997
43214	5976	5792.72	10.2761
69597	7566	7026.12	16.5496

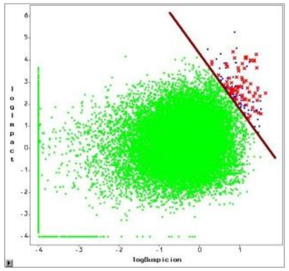
Micro and macro editing

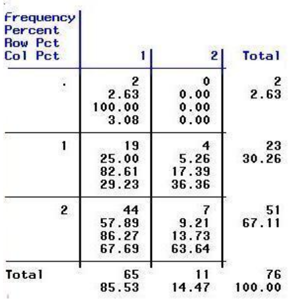
- Micro editing: *finding errors by inspecting individual observations, carried out at the record or questionnaire level.*

menot	tulot		avust
		valtio	kunta
257764	27110		
740000	36500	164908	538592
103021	28034		
334193	31048	70675	232470

- Macro editing: *subsamples or the entire sample are checked together, i.e. an important part of the data is edited simultaneously, based usually on statistical procedures and models*

2008	2009	2010
4778	5201	2714





```
Frequency |
Percent   |
Row Pct   |
Col Pct   |       1|       2|  Total
----------+--------+--------+
        . |       2|       0|      2
          |    2.63|    0.00|   2.63
          |  100.00|    0.00|
          |    3.08|    0.00|
----------+--------+--------+
        1 |      19|       4|     23
          |   25.00|    5.26|  30.26
          |   82.61|   17.39|
          |   29.23|   36.36|
----------+--------+--------+
        2 |      44|       7|     51
          |   57.89|    9.21|  67.11
          |   86.27|   13.73|
          |   67.69|   63.64|
----------+--------+--------+
Total           65       11      76
             85.53    14.47  100.00
```

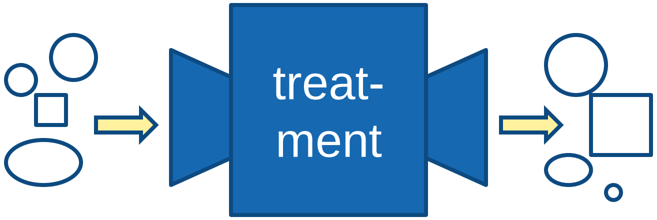
EDIT RULE (or EDIT): a restriction to the values of one or more data items that identifies missing, invalid or inconsistent values, or that points to data records that are potentially in error.

```
proc VERIFYEDITS edits =
" NETINC1 + NETINC2 = NETINCYEAR;
INC1 - EXP1 = NETINC1;
INC2 - EXP2 = NETINC2; INC1 >= 0;
INC2 >= 0; EXP1 >= 0; EXP2 >= 0;"
imply = 50 extremal = 10 acceptnegative ; run;
```

(from software Banff / SAS)

General rules for building an E&I strategy

1. Identify and eliminate **errors that are evident and easy to treat with sufficient reliability** provide information about the quality of the incoming/outgoing data



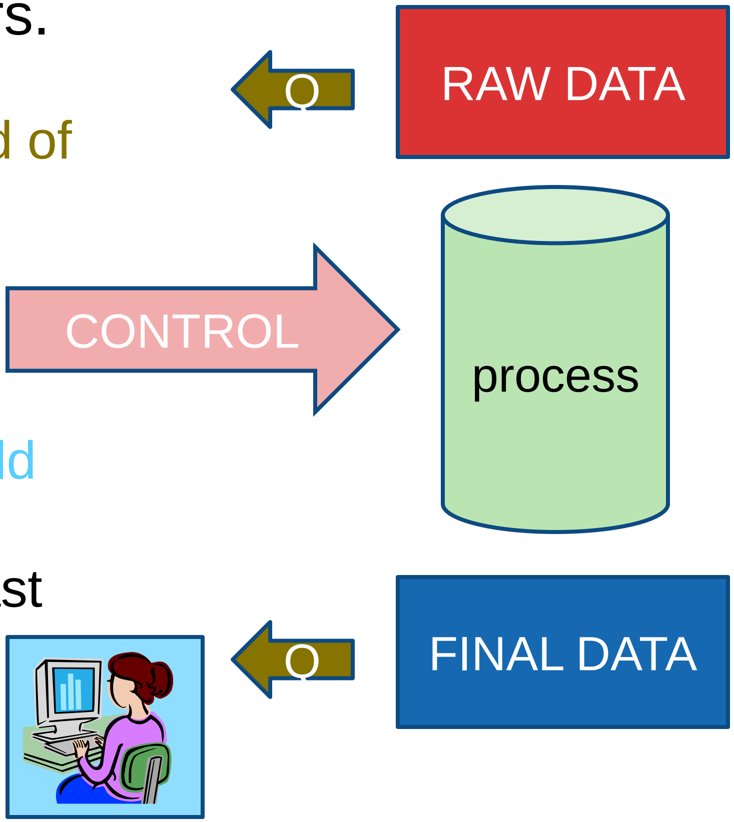
2. Select and treat with great care **influential errors**, carefully inspect influential observations when needed, provide **complete and consistent (coherent) individual data**; automatically treat the **remaining non-influential errors**

3. Check the final output to see if there are **influential errors undetected** in the previous phases or introduced by the procedure itself.

Notions of the editing and imputation process flow

E&I process is defined as a **process with a parameterization**. The implementation of the E&I process therefore requires a set of parameters.

1. The data quality at the beginning and at the end of the E&I process must be assessed.
2. The E&I process has to be designed and executed in a way that allows for control of the process.
3. The data quality at the end of the process should satisfy the needs of the users.
4. The process should be as simple, cheap and fast as possible.



Phases and procedures of editing and imputation

Editing & imputation phases by EDIMBUS	Editing and imputation phases presented here
	1) E&I during data collection / acquisition
1) Initial E & I	*2) Initial E&I*
2) Interactive / automatic E & I	*3) "Actual" micro E & I*
3) Macro E & I	*4) Macro E & I*

A phase should be reproducible and repeatable using a phase specific set of parameters. Every phase can be divided into four procedures.

Procedures

1. Detection of erroneous data
2. Decision about the treatment
3. Treatment
4. Control of the treatment

Four phases of editing

	Closing the data collection / acquisition	Getting the first data base	Data base ready for calculation of results	Final data
Procedures	Editing during data collection / acquisition	Initial editing	”Actual” micro editing	Macro editing
Error detection	***INSTANT EDITS***	***ROUTINE EDITS***	***UNIT-BASED EDITS***	***MACRO EDITS***
Decision about the treatment*	*- recontact* *- other source* *- reasoning*	*- automatic directing* *- mass correction after basic study*	*- judging the importance of unit and error* *- interactive/automatic*	*- substance decisions based on results, distributions and aggregates*
Treatment*	*- getting value* *- approximate* *- no treatment*	*- automatic correction program* *- solution for many units at once*	*- several ways of correcting and imputation*	*- several ways of correcting and imputation*
Control of treatment*	*- comparisons* *- consistency checks*	*- running routine checks again*	*- further unit checks* *- consistency checks* *- preliminary results*	*- macro-level calculations* *- comparisons*

* examples

3 Editing strategy: criteria and planning

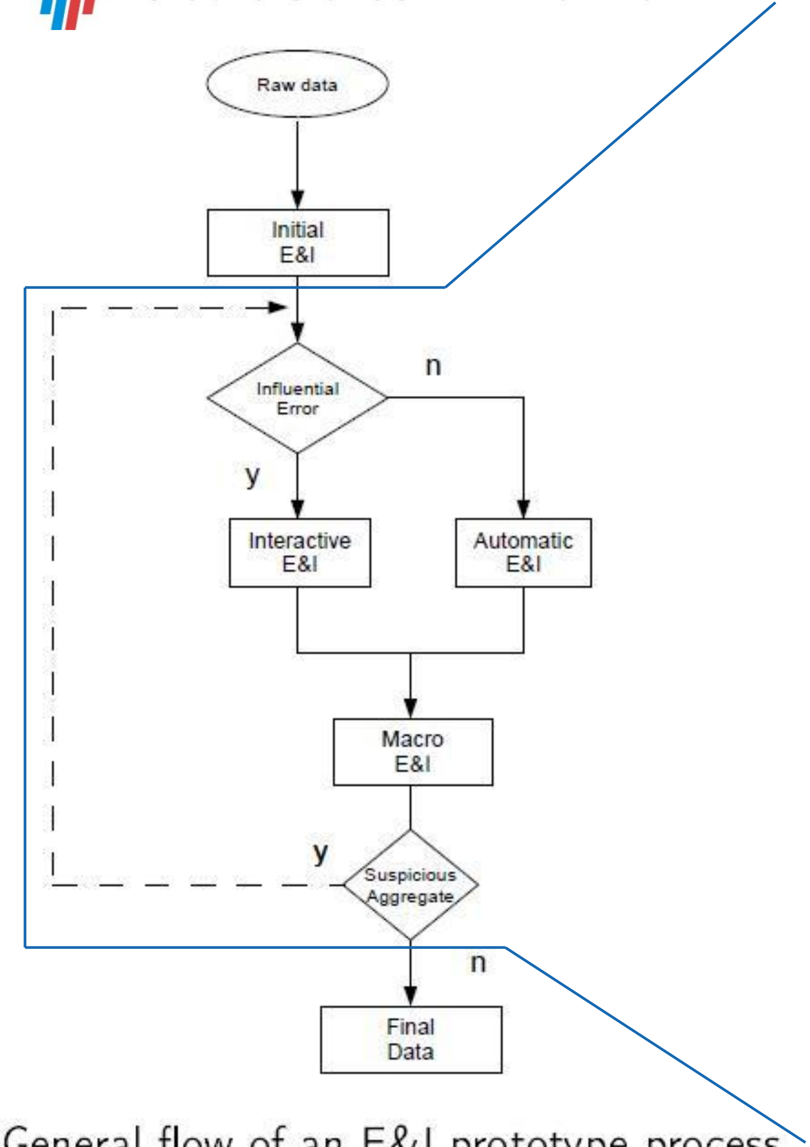
Including statistics with unit processing **STATISTICS DEPARTMENT**	Studying and/or processing data during data collection / acquisition	Getting data at once or no studying / processing	**All**
Population Statistics	20	24	44
Social Statistics	8	7	15
Prices and Wages	15	7	22
Economic Statistics	21	3	24
Business Trends	18	4	22
Business Structures	30	2	32
ALL	112	47	159

Statistics Finland

3 Editing strategy: criteria and planning

Including statistics with unit processing **STATISTICS DEPARTMENT**	Initial editing conducted	No initial editing	No answer	All
Population Statistics	13	27	4	44
Social Statistics	9	6	0	15
Prices and Wages	14	6	2	22
Economic Statistics	17	7	0	24
Business Trends	12	10	0	22
Business Structures	21	11	0	32
YHTEENSÄ	**86**	**67**	**6**	**159**

Micro and macro E&I interaction



General flow of an E&I prototype process

- Errors detected and located
- Importance of error and observation
- Decisions of their treatment
- Indicators (impact on estimators, results based on edit rules)
- Drill-down from macro to individual
- Loop-backs
- Slow, inefficient and costly E&I practices may occur

Edimbus project

Key elements for the design of E & I*

- **Survey characteristics:** type of survey (Short-Term Statistics, Structural Statistics, Economic Censuses), survey size (number of units, number of variables)
- **Survey objectives:** target parameters (totals, means, ratios, covariances, etc.), level of detail of released data (micro data, aggregates)
- **Available auxiliary information:** historical micro / aggregate data, administrative data, data from other surveys
- **Available resources:** human, time, financial resources, available equipment (software, infrastructures, etc.)
- **Applied methods and their integration:** the method applied in one phase may have an effect on the choice of the methods to be applied in other phases of the E&I process

** from ”Recommended practices for editing and imputation in cross-sectional business surveys”*

Recommendations for planning the E&I strategy (1)

- The E&I process should be designed as a **part of the whole survey process**. The **overall management of the E&I process and the interfaces with the other survey sub-processes** should be considered. For each phase, the specific aim, the required quality, the expected inputs and outputs, the starting point, the possible loop-backs, the ending point and the parameters should be described. For each phase the resources and time needed to implement, test, execute and document it should be planned.
- The E&I process should **minimize the changes to the data**. In other words, **data consistency should be obtained** by changing as few observed data as possible.

Recommendations for planning the E&I strategy (2)

- **Edit rules** should be designed in collaboration with subject matter specialists and should be based on the analysis of previous surveys. **Consistency and non-redundancy** of edits should be verified. Edits should be **designed cautiously in order to avoid over-editing**.
- The appropriate **flags**, the **documentation** including indicators and **archiving** should be part of the design.
- **Systematic errors** should be detected and treated first
- Resources should concentrate on **influential errors** (including nonrespondents). Selective editing, outlier detection and detection of influential observations are means to this purpose.

4 Testing, tuning and monitoring the strategy

Structure

- Evaluation criteria for testing error detection methods
- Two alternatives for comparative evaluation of error detecting methods
- Some aspects on testing the strategy

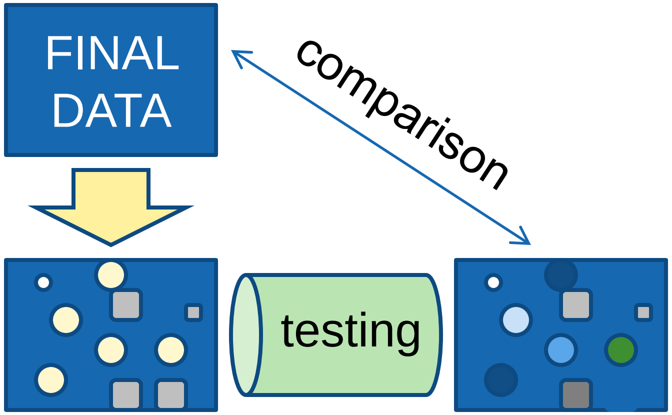
Evaluation criteria for testing error detection methods*

Criteria	***Indicators***
1) The ability of the detection procedure to find the **maximum number of errors** that are present in the dataset	*The number of correct errors found divided by the total number of errors in the dataset*
2) The ability of the detection method to find the **most influential errors**. As these errors have a **substantial impact on the final survey estimates,** it is important that they are located through the error detection mechanism.	*The number of influential errors found divided by the total number of influential errors present.*
3) Although the method must be capable of finding all errors present in the data (item 1), this should not be achieved at the cost of **flagging items erroneously to be in error**.	*The total number of incorrect errors found divided by the total number of flagged items*

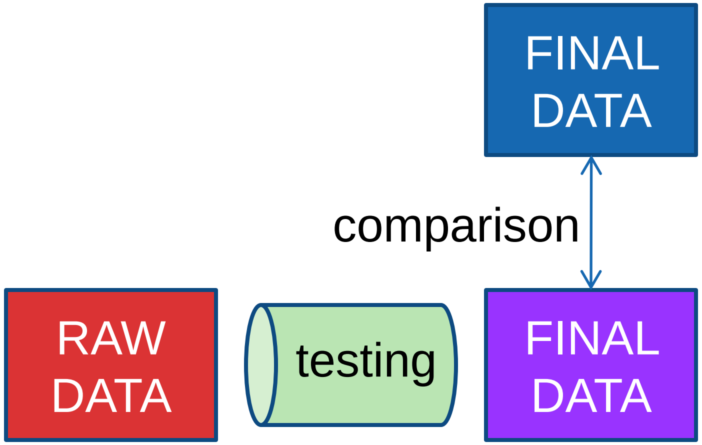
* *Euredit project*

Two alternatives for comparative evaluation of error detecting methods*

- The results obtained with the current E&I process are considered the true dataset and inconsistencies and missing values are articially introduced (by simulation) in it, using some sort of error or missing data mechanism. These mechanisms should be developed to approximate reality as much as possible



- Apply the different techniques to the original raw data and to compare the results with the data relying on the gold standard E&I process, consisting e.g. in the use of external information sources, call-backs and subject matter specialists knowledge.



* *Edimbus project*

Some aspects on testing the strategy

- *When the methods are tested in practice, it is good to check the realisation in order to get to the problems occurring. Only observed and/or processed data can be the basis for that work (no ”true” data available)*

- *Indicators e.g. rate of editing errors, number of edited values, number of missing values, E&I impact on statistics to be published*

- *Once the E&I process is implemented in the actual survey process, only slight changes should be made to monitoring and tuning in order to avoid structural breaks in the time series. The monitoring of the phases may be included in a general assessment framework for longitudinal evaluation”.*

5 Error types and tools for recognition

Structure

- Missing values
- Nonresponse mechanisms
- How full should the data be?
- Systematic errors
- Technical editing at the unit level
- Influential errors
- Model editing at the unit level
- Outliers
- Macro editing
- Random errors

Missing values

- Surveys: questions the respondent did not answer
- Non-survey data collection: values the data provider could not obtain / form
- Sometimes unusable information is discarded (preferably “flagged”)

2	1		1957	257764	27110			
2	2	1962	1964	740000	36500	164908	538592	
2	1		1928	103021	28034			
2	1		1967	480115	149120			
3	2	1982	1984	281755	38494	70675	172586	
3	2	1953	1957					
3	2							
1	3	1982	1983	336079	40954	141350	113800	48000
5	2	1958	1958	6374558	310784	839263	5224511	
5	3	1957	1961	2287934	1116649	523422	472000	183712
1	2	1960	1960	334193	31048	70675	232470	
1	2							
3	4	1923	1934	493835	37469	141350	180000	28600
3	4							
3	2	1907	1909	868700	87172	130748	668245	
3	2							
2	2	1916	1916	1478787	60535	442245	865342	110655
4	2	1910	1912	3168614	290287	1093472	1782940	

ITEM NONRESPONSE

Item nonresponse is easily observed, but the reasons causing the missing values are of special interest. **Nonresponse mechanisms** can be modelled and utilised when deciding the imputation methods.

Nonresponse mechanisms

- **Missing completely at random (MCAR).** The missing data are said to be missing completely at random if the fact that a certain item is missing does not depend on the missing nor the observed data. This means that MCAR is like simple random sampling. No missing data adjustment is therefore needed, but the assumption of a MCAR nonresponse mechanism is quite unrealistic.
- **Missing at random (MAR).** In this case the nonresponse mechanism is random *conditional on the observed covariates*. Therefore, imputation methods using the relevant observed covariates as auxiliary information can reduce nonresponse bias. This means that MAR is like simple random sampling within the classes determined by the relevant covariates.
- **Not missing at random (NMAR).** If the nonresponse mechanism depends on unobserved data, such as variables outside the survey or the target variable itself, it is said to be NMAR. In this instance the nonresponse bias cannot be reduced by imputation as the respondents and the nonrespondents differ from each other, even after conditioning on the covariates. For example, if - in a survey on income – the households with relatively high and/or low income have more nonresponse, there may be no way to reduce the resulting bias.

How full should the data be?

- The practices vary depending on the use of the data
- Usually essential item nonresponse is treated
- Sometimes “full matrix” is required (e.g. by Eurostat)

Statistics with unit processing	Pop. Stat.	Soc. Stat.	Prices & Wag	Econ. Stat.	Busin. Trends	Busin. Struct.	ALL
No answer	2	0	2	0	0	0	4
No imputation / derivation	7	4	2	4	2	2	21
Some imputation / deriv.	28	7	17	14	17	24	107
All imputed / derived	0	2	1	2	2	5	12
No missing values	7	2	0	4	1	1	15
ALL	44	15	22	24	22	32	159

Systematic errors

- A systematic error is an error that is reported consistently over time by responding units.*
- The reasons are usually misunderstandings, erroneous interpretations in coding or technical problems (e.g. “thousand” error, wrong classification, error in data collection software [jump error])
- Systematic errors are not always seen easily, and often one must have some idea of the mechanisms existing in the data.

Our turnover is
3 282 548
thousand euros

** Edimbus project*

Observing systematic errors

- Sometimes the problem can be modelled by using prior knowledge (e.g. finite mixture models)
- Also methods detecting outliers can reveal systematic errors
- **Fatal edit:** Identies data errors with certainty. Examples are an economic activity that does not exists in a list of acceptable economic activities and balance edits. Also known as hard edit.
- **Ratio edit:** An edit rule determining the acceptable bounds for a ratio of two variables (e.g. thousand-errors via turnover/number of workers).
- Systematic errors must be cleaned before processing the random errors, which requires the data as bias-free as possible.
- If systematic error mechanisms exist in the survey/data collection process (e.g. by questionnaire, interviewer education, coding, processing), improvements must be made. One can develop specified methods for observing that in the future.

Technical editing at the unit level

Statistics with unit-level processing	Pop. Stat. (44)	Soc. Stat. (15)	Pric. & Wages (22)	Econ. Stat. (24)	Busin. Trends (22)	Busin. Struct. (32)	ALL (159)
Unit-level examination with a computer	19	10	17	23	18	29	116
Logical checks using a program or otherwise	37	13	8	21	13	25	117
Defining non-valid variable values	31	12	8	14	11	19	95
Listing extreme values of variables	13	11	9	10	11	24	78
Comparing with previous or other values	34	10	14	22	13	23	116
Ratio of values of two variables or different time points, other functions	16	8	5	13	4	19	65

Influential errors

- Influential errors are errors in values of variables that have a significant influence on publication target statistics for those variables.
- An influential observation is an observation that has a large impact on a particular result of a survey, i.e. a statistic (e.g. in business surveys).
- That value may be correct or not and, in this latter case, it can generate an influential error.
- Two main approaches for error recognition
  - *Interactive editing:* computer aided manual editing after the data capturing process, which is usually assisted by automatic editing
  - *Editing based on models* (e.g. selective editing), where models which take the most important phenomena into account indicate the influence of the observation and error and give information for the further operations (manual or automatic treatment)

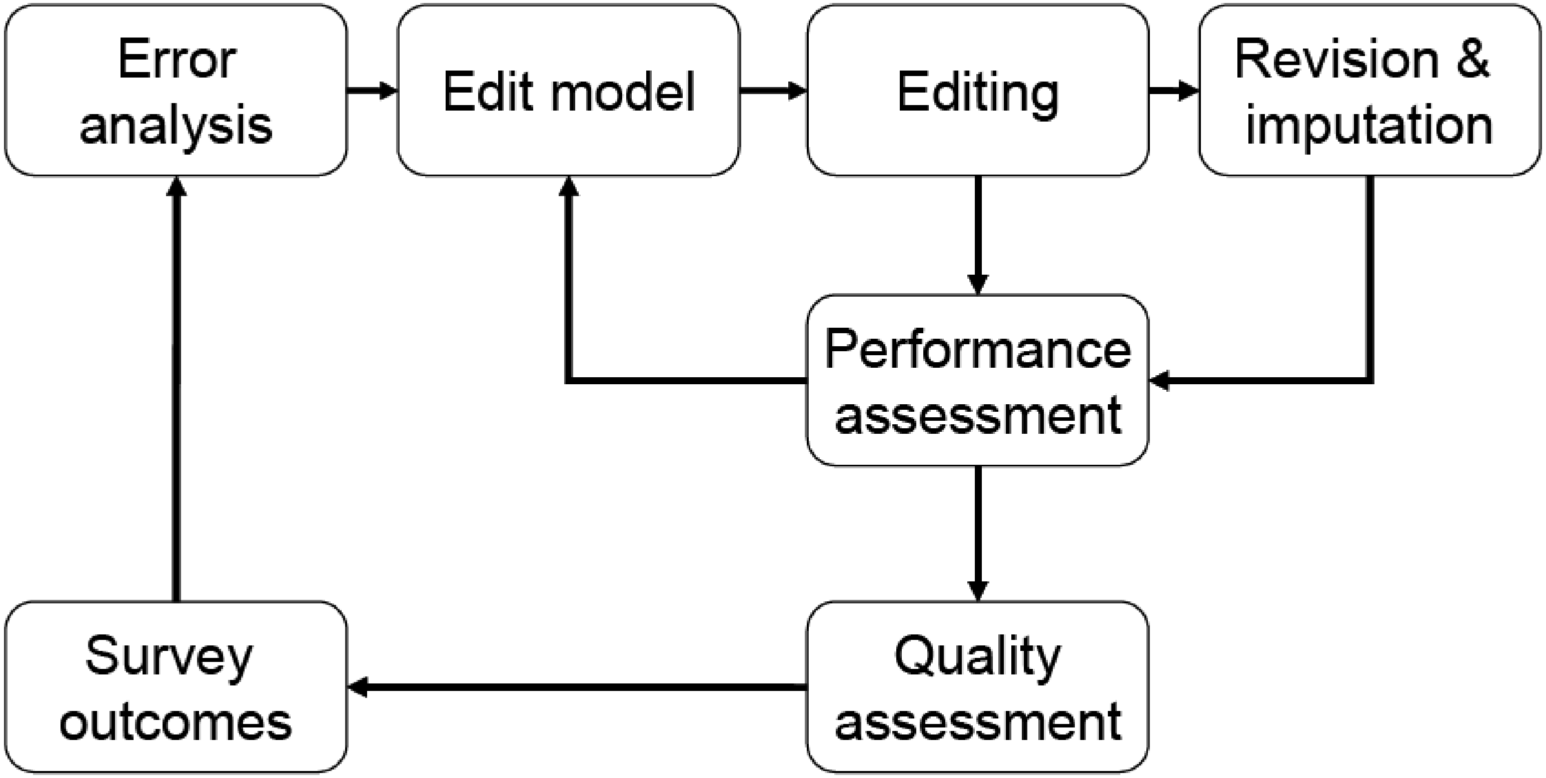
Influential errors, edit types

- **Fatal edit (hard edit):** Identies data errors with certainty
- **Query edit (soft or statistical edit):** An edit rule whose failure indicates an error with probability less than 1. For example, a value that, compared to historical data, seems suspiciously high.
- **The use of edit rules** (assuming influential errors will violate edit checks):
  - Select records that fail edits.
  - For each of these records, estimate the amount of changes for variables involved in failed edits needed to make the records satisfy edit constraints.
  - Use the estimated amounts of changes to build a score function to prioritize records to be manually reviewed.

Selective editing

- ***Selective editing*** is based on the assumption that the raw data can be divided into observations which require detailed inspection *(critical stream)* and observations which do not require that *(non-critical stream)*.
- In editing at the micro level there will be computational operations, which model some variables so that they can obtain predicted values based on the data available.
- With utilising those one can define *score functions*, which include two components: *influence* to the estimator and *risk* to be erroneous.
- For one variable a *local score function* can be calculated, and correspondingly one can create an observation-level *global score function*.
- A *cut off value* is set to decide when a record should be treated. That is, a given record is suspicious if the value of a score function exceeds.

Selective Editing
(Statistics New Zealand)



Some aspects on selective editing

- If score functions cannot be applied to some observations, those observations should have a special study.
- It is important to priorise errors and variables. These priorities should be put into score functions, which have to include both risk and influence components.
- The threshold levels of the score functions must be chosen accurately. If the survey process changes, the whole system must be reassessed.
- The predicted values should be tested at least in some subpopulations.

At its best selective editing improves quality, reduces workload, speeds up production and saves costs. Several statistical offices currently study the possibilities to implement selective editing, and in some offices the method has been taken into production of a few statistics. Statistics Sweden has developed a module (SELEKT) for carrying out selective editing.

Model editing at the unit level

Statistics with unit-level processing	Pop. Stat. (44)	Soc. Stat. (15)	Pric. & Wages (22)	Econ. Stat. (24)	Busin. Trends (22)	Busin. Struct. (32)	ALL (159)
Defining the certainty of different variables to be right in the case of conflicting variables (reliability weight, minimum change Fellegi-Holt -principle)	6	3	2	0	6	0	17
Comparing modelled value and observed value	0	1	4	8	1	1	15
Modelling variable values / observations risk to be erroneous (e.g. selective editing)	1	1	1	0	0	0	3
Finding problematic values with defining the importance of the observation or so called sensitivity function (reveals the effect of the observation to the estimate)	0	5	12	0	7	6	30

Macro editing

- ***Macro editing*** concentrates on utilising the distributions and preliminary estimates (possibly in different subpopulations) in order to find errors in the data (moving away from the unit level).
- Problematic observations can be found by e.g. studying extremities or creating a sensitivity function revealing the influence to the estimates.
- The basic idea behind the *aggregate method* is that the calculated estimates are compared with the estimates from other sources. An essential practice is to compare results with the previous results, possibly as a time series.
- The *graphical study* together with the current results is an efficient way for finding errors.

Macro editing requires statistical skills, subject matter knowledge and good information on all previous stages of the survey and the E&I process. (Edimbus project)

Outliers

- An **outlier** is an observation which is not fitted well by a model for the majority of the data. For instance, an outlier may lie in the tail of the statistical distribution or far away from the center of the data.
- The model constructed for evaluating the deviation is based on the distributions of the population, not the sample, and the models can vary from one subpopulation to another, targeted to one or more variables (univariate / multivariate outliers).
- Outlier ⟺ influential observation. *Sensitivity curve* as a tool for evaluation.
- Usually outliers are identified via macro editing.
- **Representative outliers:** correct observations which may have similar units in the population.
- **Non-representative outliers:** either incorrect observations whose true values would not show up as outlying, or unique, but correct values, in the sense that one should not extrapolate them to other observations in the population.

Tools for finding outliers

- The *score functions* of selective editing are useful also when searching outliers. They can be marked with dichotomic flag variables or specific robustness weights, which are continous variables between 0 and 1.
- *Robust estimates* are important in this context, because they are not sensitive to exceptional or erroneous values appearing in the data.
- **Univariate methods:** e.g. distance from the median with a threshold and the quartile range with and without weighting, winsorised and trimmed means.
- **For periodic data:** in the Hidiroglou-Berthelot method the ratio of two consecutive values forms the basis for a measure including an interval. The observation is an outlier outside the interval.
- **Residuals in a regression model based on robust estimation** can reveal outliers.
- **Multivariate methods** are often based on the Mahalanobis distance. Together with graphical plot figures (ellipse) they can reveal also rather complex outliers.
- The robustness of many of these methods can be adjusted by a *tuning constant* constructed to the method. It is beneficial to concentrate especially on the choice and testing of this constant, because different data require different adjustments).
- Graphical studies are very useful in detecting outliers.

Macro editing

Statistics with unit-level processing	Pop. Stat. (44)	Soc. Stat. (15)	Pric. & Wages (22)	Econ. Stat. (24)	Busin. Trends (22)	Busin. Struct. (32)	ALL (159)
Studying distributions and cross-tabulations	32	15	6	15	6	23	97
Information from calculating preliminary estimates (e.g. mean, total, correlation, deviation)	23	14	10	15	7	26	95
Controlling the joint effect of survey weights and exceptional values	0	5	4	0	1	5	15
Comparing with estimates from previous occasion(s), valid limits for estimates (e.g. time series)	15	11	15	18	10	26	95
Using graphical methods	8	8	5	13	7	15	56
Studying aggregated data	25	6	19	19	17	28	114
Comparing with other possible data	28	10	8	18	7	27	98

Random errors

- **Random errors** are not caused by a systematic reason, but by accident. Usually they appear due to in-attention by respondents, interviewers and other processing staff during the various phases of the survey cycle.
- In the statistical context the expectation of a random error is typically zero (not always), and often they are non-important.
- However, these errors can lead to inconsistent records because some edit rules are violated.
- **Deterministic checking rules** state which variables are considered erroneous when the edit rules are violated in a certain record, for instance because that variable is less reliable. Often the deterministic checking rules are coupled with deterministic imputations.
- Sometimes this yields consistency problems.

Fellegi-Holt principles

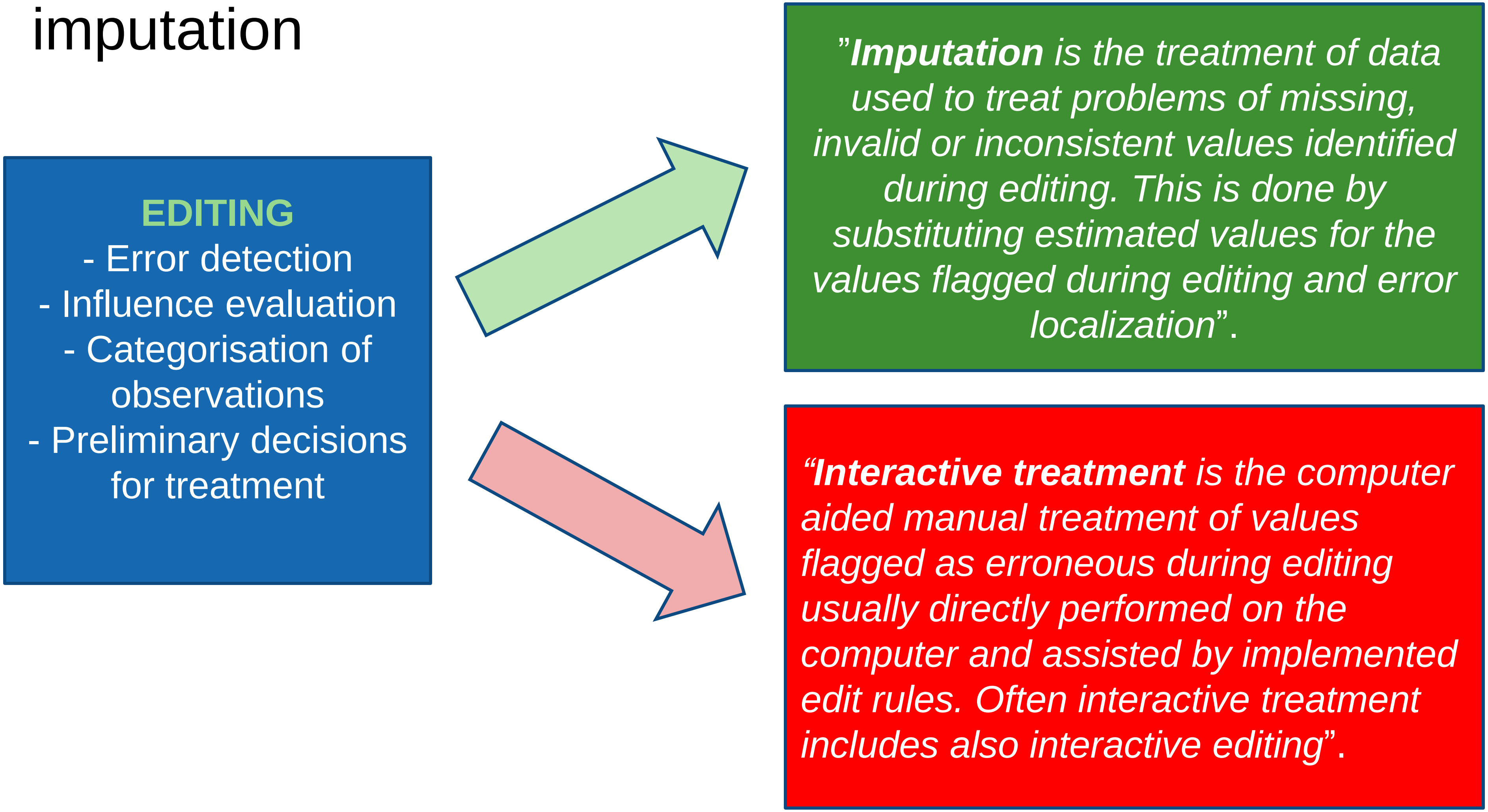
- There are several general guiding principles for the localization of the erroneous elds in an inconsistent record. The best-known and most-used of these general guiding principles is the Fellegi-Holt paradigm. This paradigm is, in fact, only one of three principles for automatic edit and imputation proposed by Fellegi and Holt.

    1. The data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields);
    2. As far as possible the frequency structure of the data file should be maintained;
    3. Imputation rules should be derived from the corresponding edit rules without explicit specication.

Reliability weight

- **Reliability weight** is a measure of confidence in the value of this variable.
- Variables that are generally correctly observed are given a high reliability weight;
- Variables that are often incorrectly observed are given a low reliability weight.
- A reliability weight of a variable corresponds to the error probability of this variable, i.e. the probability that its observed value is erroneous.
- The higher the reliability weight of a variable, the lower its error probability.

6 Editing as the basis for error correction and imputation



EDITING
- Error detection
- Influence evaluation
- Categorisation of observations
- Preliminary decisions for treatment

*”**Imputation** is the treatment of data used to treat problems of missing, invalid or inconsistent values identified during editing. This is done by substituting estimated values for the values flagged during editing and error localization*”.

*“**Interactive treatment** is the computer aided manual treatment of values flagged as erroneous during editing usually directly performed on the computer and assisted by implemented edit rules. Often interactive treatment includes also interactive editing*”.

Interactive treatment

- Some errors are directed to the interactive treatment, where the examiner (subject matter expert) tries to find the origins of the error with call-backs, programs, unit level inspection and possibly with a paper questionnaire study.
- Interactive treatment is called ”manual treatment” in many contexts.
- Experienced examiner can provide clear and important quality improvements to the data with interactive treatment, but in some cases (e.g. new workers) there is a danger for creative editing, where the correction principles are subjective and not justified by the contents of the data and subject matter.
- Interactive treatment is resource and time intensive. Ineffective and inaccurate error recognition can bring a lot of less important observations to the detailed inspection, and this can cause over-editing.
- Extensive manual treatments can cause problems for keeping the consistency of the results.

Treatment types (not imputation)

Statistics with unit-level processing	Pop. Stat. (44)	Soc. Stat. (15)	Pric. & Wages (22)	Econ. Stat. (24)	Busin. Trends (22)	Busin. Struct. (32)	ALL (159)
Getting contact to the respondent and asking the value or getting it from the paper questioinnaire of the postal enquiry	27	5	17	20	16	30	115
Fetching the previous value (cold-deck)	6	2	13	11	8	20	60
Getting the value from another observation or another source	12	5	13	14	14	25	83
Getting the real value by reasoning based on the information of the observation in question	27	7	8	21	13	27	103
Correcting automatically with program lines including conditions or based on a list of erroneuos values (e.g. ‘america’ = ‘United States’)	37	8	6	14	10	18	93
Correcting automatically based on risk functions (e.g. selective editing)	0	0	1	0	6	0	7