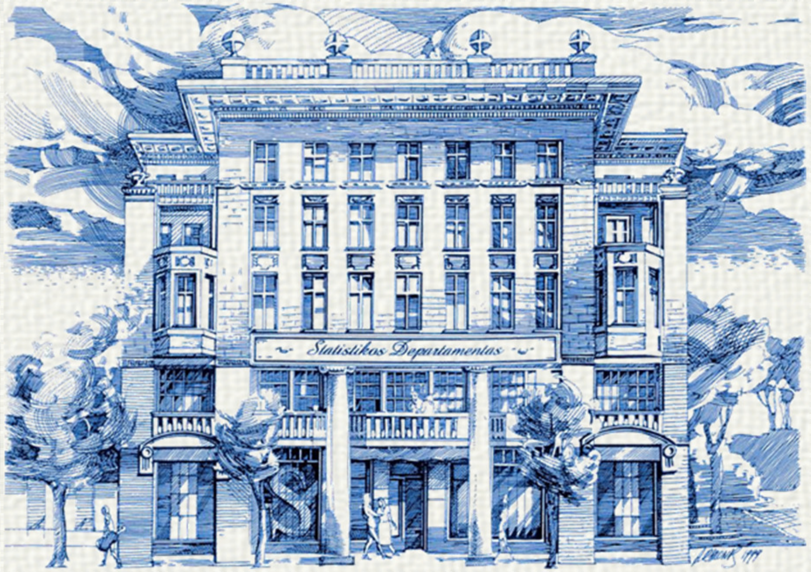
STATISTICS
LITHUANIA

EDITING PROCEDURES IMPLEMENTED IN SBS



Antanina Valiuliene

Enterprise Statistics Division,

Statistics Lithuania

25–29 June 2012

EDITING LEVELS

- ❑ MICRO DATA EDITING
  - ❑ DATA FROM STATISTICAL QUESTIONNAIRE (hereafter – SQ) – the editing is performed in the application
  - ❑ DATA FROM ADMINISTRATIVE SOURCES – the editing procedure is performed automatically
- ❑ MACRO DATA EDITING – QUALITY CHECKS
  - ❑ Consistency checks
  - ❑ Year-to-year variations checks
  - ❑ Ratio checking

EDITING PROCEDURE OF DATA FROM SQ

- ❑ DATA ENTERING APPLICATION IS REALIZED IN ORACLE SOFTWARE
- ❑ DATA EDITING IS REALISED THROUGH THE EDIT RULES
- ❑ EDIT RULES ARE APPLIED IN THE ELECTRONICAL QUESTIONNAIRES TOO. FEWER RULES ARE APPLIED IN THE e-SQ THAN IN THE DATABASE APPLICATION;
- ❑ EDITING IS PERFORMED WHEN ALL DATA FROM STATISTICAL QUESTIONNAIRE IS ENTERED INTO THE DATABASE (OR e-SQ IS FILLED IN)
- ❑ ALL THE RULES HAVE A STATUS. IT MIGHT BE MANDATORY (hard) OR IGNORED (soft)

  (if a rule is ignored, the explanation why data does not meet the condition is requested);

EDITING RULES

❑ EDITING RULES APPLIED IN THE DATABASE APPLICATION

❑ VALIDITY RULE – IF THE VALUE CORESPONDS WITH A PARTICULAR CLASIFICATOR OR SATISFIES SOME STRICT RULE, IT IS VERIFIED (enterprise ID code, NACE code, sign of value (+/-))

❑ MISSED VALUES – IT IS CHECKED WHETHER THE REQUIRED VALUES ARE PROVIDED

(if the number of employees is provided, the hours worked as well as wages and salaries should be fulfilled)

❑ LOGICAL RULE – IT IS CHECKED WHETHER THE INDICATORS DO NOT CONTRADICT EACH OTHER.

(if purchases of goods and services for resale plus changes in stocks of goods and services for resale in the same condition as received > 0, then turnover from trade or services should be >0 (with the exception in activities NACE code 3513, 3514, 3522, 3523); if the activity code is industry, then turnover from industry should be > 0)

EDITING RULES (cont.)

- ❑ MATHEMATIC RULE - IT IS CHECKED WHETHER THE SUM OF VALUES BY SUBCLASSIFICATION EQUALS TO TOTAL

- ❑ RULE OF LIMITS – IS APLIED FOR TWO OR MORE VALUES WHICH MIGHT BE LINKED BY RATE AND SHOULD NOT EXCEED SOME LIMITS

  (this rule is applied for the employees’ data which are related and have some limits)

- ❑ HISTORICAL RULE – WHEN DATA FROM THE REFERENCE AND PREVIOUS PERIODS IS COMPARED   (the stocks by sub-classification are checked with data from the previous period. Historical data is showed for information in the application)

ADDITIONAL EDITING PROCEDURE

- ❑ ADDITIONAL EDITING:
  - ❑ number of employees and wages and salaries is compared with admin data;
  - ❑ comparability of investments in tangible goods is checked with data from the quarterly survey of investments

- ❑ THE SYSTEM OF EDITING RULES HAS TO BE OPTIMAL – RULES HAVE TO BE TESTED; EXCEPTIONAL RULES SHOULD BE REMOVED;

- ❑ EDITING PROCESS SHOULD BE MEASURED

  (indicators might be such as: number of units edited; number of values edited; number of indicators edited)

- ❑ MONITORING OF EDITING PROCESS SHOULD BE PERFORMED.

EDITING PROCEDURE OF THE ADMINISTRATIVE DATA

❑ DATA EDITING PROCEDURE OF ANNUAL FINANCIAL STATEMENT IS AUTOMATED AND IMPLEMENTED BY SAS SOFTWARE

❑ A CORRECT RECORD MUST PASS AN EDIT RULE, A LINEAR EQUALITY BASED ON ACCOUNTING IDENTITIES

❑ INCORRECT RECORDS MUST BE CORRECTED USING MATHEMATICAL METHODS

❑ DEVELOPMENT OF EDITING METHODS HAS BEEN A CONTINOUS PROCESS

EDITING METHODS

- THE FOLLOWING EDITING METHODS ARE APPLIED FOR THE ANNUAL FINANCIAL STATEMENT:

  - EDIT RULE
  - SIGN CHECKING
  - LOCATING THE ERROR
  - OUTLIER DETECTION
  - RE-SCALING
  - DONOR BASED EDITING

EDITING METHODS (2)

EDIT RULE

- ❑ THE EDIT RULE DETERMINES WHETHER A RECORD IS CORRECT OR NOT. IT IS A LOGICAL CONDITION OR A RESTRICTION TO THE VALUE OF A DATA ITEM WHICH MUST BE MET IF THE DATA IS TO BE CONSIDERED CORRECT
- ❑ *EXAMPLE*:

PROFIT AND LOSS ACCOUNT			
VARIABLE	***EDIT RULE***	***SIGN***	***NAME OF THE VARIABLE***
X_1	+	+	SALES
X_2	-	+	COST OF SALES
Y_1	Sum(X_1: X_2)	+/-	GROSS PROFIT (LOSS)
X_3	-	+	OPERATING EXPENSES
Y_2	Sum(X_1: X_3)	+/-	OPERATING PROFIT (LOSS)
X_4	+	+/-	OTHER ACTIVITIES
X_5	+	+	INCOME OF FINANCIAL AND INVESTING ACTIVITIES
X_6	-	+	EXPENSES OF FINANCIAL AND INVESTING ACTIVITIES
Y_3	Sum (X_1: X_6)	+/-	ORDINARY PROFIT (LOSS)
X_7	+	+	EXTRAORDINARY GAIN
X_8	-	+	EXTRAORDINARY LOSS
Y_4	Sum(X_1: X_8)	+/-	PROFIT (LOSS) BEFORE TAXATION
X_9	-	+	CORPORATE INCOME TAX
Y	Sum (X_1: X_9)	+/-	NET PROFIT (LOSS)

EDITING METHODS (3)
EDIT RULE

- ❑ VARIABLES FROM X_1 TO X_9 MUST SATISFY A LINEAR EQUATION

$$Y = \sum_{i=1}^{9} X_i$$

- ❑ THEN THE VALUE OF ERROR IS CALCULATED

$$e = \sum_{i=1}^{9} X_i - Y$$

- ❑ A RECORD IS CONSIDERED CORRECT IF THE VALUE OF ERROR IS EQUAL TO ZERO

$$e = 0$$

EDITING METHODS (4)
SIGN CHECKING

- VARIABLES WHICH CAN HAVE EITHER A POSITIVE OR A NEGATIVE VALUE ARE CHECKED WHETHER THE SIGN IS CORRECT
- *EXAMPLE*:
  - VARIABLE X_4 CAN HAVE EITHER A POSITIVE OR A NEGATIVE VALUE. THE SIGN IS CHANGED

$$X_4^* = -X_4$$

  - THEN THE EDIT RULE IS APPLIED TO CHECK IF A RECORD WITH A CHANGED SIGN OF THE CORRESPONDIG VARIABLE IS CORRECT OR NOT

$$\left(\sum_{i=1}^{3} X_i + X_4^* + X_5 - X_6 + \sum_{i=7}^{9} X_i\right) - Y = 0$$

  - IF THE EDIT RULE HOLDS TRUE, THE VALUE OF VARIABLE X_4 IS CHANGED INTO VALUE X_4^*

EDITING METHODS (5)
LOCATING THE ERROR

- ❑ FOR INCORRECT RECORDS THE ERROR IS LOCATED TO A CERTAIN PART OF THE ANNUAL FINANCIAL STATEMENT (SET OF VARIABLES) BY THE USE OF SUBTOTALS

EXAMPLE:

USING THE SUBTOTALS (Y_1, Y_2, Y_3, Y_4) CERTAIN ERRONEOUS VARIABLES ARE DETECTED.

FOR THE FOLLOWING SUBTOTALS THE FOLLOWING EQUATIONS ARE APPLIED:

$$\sum_{i=1}^{2} X_i = Y_1 \qquad \sum_{i=1}^{3} X_i = Y_2 \qquad \sum_{i=1}^{6} X_i = Y_3 \qquad \sum_{i=1}^{8} X_i = Y_4$$

THUS, THE FOLLOWING EQUATIONS CAN BE DERIVED:

$$Y_1 - \sum_{i=3}^{9} X_i = Y \qquad Y_2 + \sum_{i=4}^{9} X_i = Y \qquad Y_3 + \sum_{i=7}^{9} X_i = Y \qquad Y_4 - X_9 = Y$$

THE FOLLOWING CONDITIONS CAN BE TESTED:

$$Y_1 - \sum_{i=3}^{9} X_i - Y = 0 \qquad Y_2 + \sum_{i=4}^{9} X_i - Y = 0 \qquad Y_3 + \sum_{i=7}^{9} X_i - Y = 0 \qquad Y_4 - X_9 - Y = 0$$

IF THE FIRST CONDITION IS NOT TRUE THEN IT IS ASSUMABLE THAT THE ERROR IS LOCATED IN VARIABLES X_1, X_2. IF THE SECOND CONDITION IS NOT TRUE THEN IT IS ASSUMABLE THAT THE ERROR IS LOCATED IN VARIABLES X_1, X_2, X_3 AND SO ON

EDITING METHODS (6)
OUTLIER DETECTION

❑ THE INCORRECT SET OF VARIABLES IS COMPARED TO THE DISTRIBUTION OF CORRESPONDING VARIABLES OF THE CORRECT RECORDS IN THE RESPECTIVE ACTIVITY. THIS METHOD IS USED TO IDENTIFY AND CORRECT BIG ERRORS FOR ONE PARTICULAR VARIABLE;

❑ IN THIS METHOD THE VALUES ARE PRESENTED IN RELATION TO TURNOVER (X_1);

❑ FOR THE INCORRECT SET OF VARIABLES A RATIO IS CALCULATED:

$$S_i = \frac{X_i}{X_1}$$

❑ THE CORRESPONDING RATIOS ARE CALCULATED FOR THE CORRECT SET OF VARIABLES

EDITING METHODS (7)
OUTLIER DETECTION

- ❑ DISTRIBUTIONS OF ALL RATIOS ARE CALCULATED AND 1st (D1) AND 9th (D9) DECILES ARE SELECTED AS THRESHOLD VALUES. SUSPECT VALUES OUTSIDE THIS TARGET RANGE MAY CONTAIN AN ERROR

- ❑ THE RELATIVE ERROR IS CALCULATED:

$$S_e = \frac{e}{X_1}$$

- ❑ IF A VALUE OF RATIO IS OUTSIDE THE TARGET RANGE, IT IS TESTED WHETHER THE VALUE FITS INSIDE THE TARGET RANGE AFTER ADJUSTING IT BY THE ERROR *e*

EDITING METHODS (8)
OUTLIER DETECTION

- ❑ WHEN THE VALUE OF ERROR IS POSITIVE, FOR THE NEGATIVE VARIABLES THE FOLLOWING CONDITIONS ARE TESTED:

$$S_i < D_1(S_i) \qquad AND \qquad D_1 \leq S_i + S_e \leq D_9$$

- ❑ WHEN BOTH CONDITIONS ARE TRUE, THE ERROR IS ADJUSTED TO THAT PARTICULAR VARIABLE:

$$X_i^* = X_i + e$$

- ❑ FOR POSITIVE VARIABLES THE FOLLOWING CONDITIONS ARE TESTED:

$$S_i > D_9(S_i) \qquad AND \qquad D_1 \leq S_i - S_e \leq D_9$$

- ❑ WHEN BOTH CONDITIONS ARE TRUE, THE ERROR IS ADJUSTED TO THAT PARTICULAR VARIABLE:

$$X_i^* = X_i - e$$

EDITING METHODS (9)
OUTLIER DETECTION

- ❑ IF THE VALUE OF ERROR IS NEGATIVE, FOR NEGATIVE VARIABLES THE FOLLOWING CONDITIONS ARE TESTED:

$$S_i > D_9(S_i) \qquad AND \qquad D_1 \leq S_i + S_e \leq D_9$$

- ❑ FOR POSITIVE VARIABLES THE FOLLOWING CONDITIONS ARE TESTED:

$$S_i < D_1(S_i) \qquad AND \qquad D_1 \leq S_i - S_e \leq D_9$$

EDITING METHODS (10)
RE-SCALING

- ❑ WHEN A RECORD CONTAINS A RELATIVELY SMALL ERROR, THE INCORRECT SET OF VARIABLES IS RE-SCALED
- ❑ ALL REMAINING INCORRECT RECORDS ARE DIVIDED INTO TWO GROUPS DETERMINED BY THEIR RELATIVE ERROR. A RELATIVE ERROR OF ± 5% OF TURNOVER IS USED AS A THRESHOLD
- ❑ THE INCORRECT SET OF VARIABLES IS MULTIPLIED BY A SCALING FACTOR TO THE LEVEL OF THE RECORD. THE ERROR IS DISTRIBUTED TO ALL VARIABLES BELONGING TO THE INCORRECT SET
- ❑ THE SCALING FACTOR IS THE ERROR DIVIDED BY THE SUM OF THE INCORRECT SET OF VARIABLES (*E*):

$$k = \frac{e}{\sum_{i \in E} |X_i|}$$

- ❑ EVERY INCORRECT VARIABLE IS MULTIPLIED BY THE SCALING FACTOR:

$$X_i^* = (1-k) \cdot X_i, \ \ when \ \ X_i \geq 0$$
$$X_i^* = (1+k) \cdot X_i, \ \ when \ \ X_i < 0$$

EDITING METHODS (11)
DONOR BASED EDITING

- ❑ FOR THE REST OF THE INCORRECT RECORDS CONTAINING AN ERROR BIGGER THAN ± 5% OF A RELATIVE ERROR, A DONOR UNIT IS DETERMINED AND THE INCORECT PART IS ESTIMATED BY A DATA STRUCTURE OF A VARIABLE OF A DONOR UNIT

- ❑ TWO TYPES OF DONORS ARE USED:
  - ❑ PAST INFORMATION OF THE SAME ENTERPRISE
  - ❑ NEAREST NEIGHBOUR

EDITING METHODS (12)
DONOR BASED EDITING

❑ PAST INFORMATION OF THE SAME ENTERPRISE

- ❑ PAST INFORMATION IS USED AS A PRIMARY DONOR
- ❑ THE INCORRECT SET OF VARIABLES IS ESTIMATED BY A DATA STRUCTURE OF THE SAME ENTERPRISE FROM THE PREVIOUS YEAR

$$X_i^* = Y \cdot \frac{X_i^{past}}{Y^{past}}$$

where *Y* – a corresponding subtotal of an annual financial statement

STATISTICS
LITHUANIA

EDITING METHODS (12)
DONOR BASED EDITING

- ❑ NEAREST NEIGHBOUR
  - ❑ NEAREST NEIGHBOUR IS USED AS THE DONOR IF PAST INFORMATION IS NOT AVAILABLE. IT IS SELECTED FROM A GROUP CONSISTING OF CORRECT RECORDS IN THE RESPECTIVE ACTIVITY. THE DISTANCE MEASURE BETWEEN TWO VARIABLES IS:

$$D_i = MIN\left\{\sum_{k\in F}\left|X_{ik} - X_{ik}^{near}\right|\right\}$$

where ***F*** – set of variables selected for comparison

  - ❑ THE INCORRECT SET OF VARIABLES IS ESTIMATED BY A DATA STRUCTURE OF A NEAREST NEIGHBOUR

$$X_i^* = Y \cdot \frac{X_i^{near}}{Y^{near}}$$

where ***Y*** – a corresponding subtotal of an annual financial statement

QUALITY CHECKS(1)

- ❑ YEAR-TO-YEAR VARIATIONS (where real growth and inflation rate are taken into account; the boundaries are used as provided in the Eurostat manual;)

  - ➢ THE FOLLOWING CONDITIONS BY NACE CLASS (OR GROUP) FOR VARIABLE V12110 (TURNOVER) ARE TESTED:

$$\text{lower boundary} < \text{V12110 variation} < \text{upper boundary}$$

If the following conditions are not true then records at the enterprise level must be corrected or confirmed. For activities with less then three enterprises no year-to-year variations are checked.

THE VARIATION FOR TURNOVER OF YEAR T IS CALCULATED:

$$\text{V12110 variation} = \left(1 + \frac{V12110_t - V12110_{t-1}}{V12110_{t-1}}\right) * 100$$

QUALITY CHECKS(2)

- THE LOWER AND UPPER BOUNDARIES ARE CALCULATED:

$$\text{lower boundary} = \frac{82}{1+\frac{4}{\sqrt{n_{t-1}}}} * (1+\text{real growth}) * (1+\text{inflation rate})$$

$$\text{upper boundary} = \frac{122}{1+\frac{4}{\sqrt{n_{t-1}}}} * (1+\text{real growth}) * (1+\text{inflation rate})$$

where n – number of enterprises by Nace class.

For turnover characteristics, a suitable standard interval is [82%; 122%] as provided in the Eurostat manual. Real growth and inflation rate are used at the national level.

QUALITY CHECKS(3)

- ❑ EXAMPLE:

NACE	V11110_2010	V11110_2009	V12110_2010	V12110_2009	V12110 variation	lower boundary	upper boundary	ERROR
J62	1056	1001	1070526897	912000536	117	77	145	0
J620	1056	1001	1070526897	912000536	117	77	145	0
J6201	523	459	548713389	442085104	124	73	152	0
J6202	238	253	235901676	218246396	108	69	161	0
J6203	35	32	66177865	47133007	140	51	219	0
J6209	260	257	219733967	204536029	107	69	161	0
J63	180	183	202896466	231476553	88	67	166	0
J631	152	160	177676966	152763878	116	66	169	0
J6311	94	90	145339335	128141156	113	61	183	0
J6312	58	70	32337631	24622722	131	58	190	0
J639	28	23	25219500	78712675	32	47	236	1
J6391	5	5	7223149	7666011	94	31	358	0
J6399	23	18	17996351	71046664	25	44	250	1

- ❑ YEAR-TO-YEAR VARIATIONS ARE ONLY CHECKED FOR THE ANNUAL DATA SERIES

QUALITY CHECKS (4)

- ❑ RATIO CHECKING (list of ratios and standard intervals is used as provided in the Eurostat manual)

Example:

NACE	V13310_V16130_ 2010	V13310_V16130_ 2009	V13310_V16130 variation	lower boundary	upper boundary	ERROR
N	22628,13	24213,85	93	83	131	0
N77	22065,35	25518,48	86	79	137	0
N771	21000,81	23089,33	91	71	152	0
N7711	25586,90	26898,48	95	69	158	0
N7712	11277,13	12624,94	89	62	175	0
N772	20624,97	20622,17	100	76	142	0
N7721	5232,84	8279,45	63	67	160	1
N7722	8265,73	8437,64	98	61	178	0
N7729	27445,19	25507,55	108	73	148	0
N773	23370,92	28794,22	81	71	151	0

The final interval for this ratio:

$$\left[\frac{85}{1+\frac{4}{\sqrt{n_{t-1}}}} * (1+\text{inflation rate}); \frac{118}{1+\frac{4}{\sqrt{n_{t-1}}}} * (1+\text{inflation rate}) \right]$$