Tilastokeskus

Editing and imputation methods in Finnish SBS

Tilastokeskus

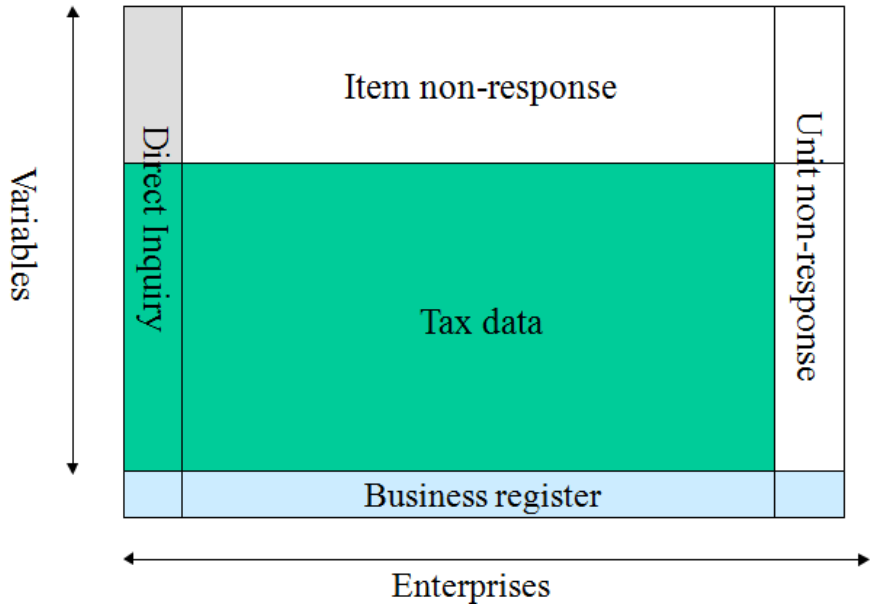
Editing and Imputation

- General overview of main SBS data sources
- General remarks
- E&I process
- Unit and item non-response imputation
- Using flags in the E&I process
- Initial E&I
- Imputation, unit non-response
- Imputation, item non-response
- Interactive treatment

Tilastokeskus

General Overview of main SBS data sources

- Main data sources
  - Business register, 300 000 units, 40-50 variables
  - Income tax files, 270 000 units, about 350 variables
  - Direct inquiry, 5000 units (FSS) + 1200 units (SRA), about 100 variables
- Auxiliary/additional information
  - Value-added tax data
  - Official financial statements



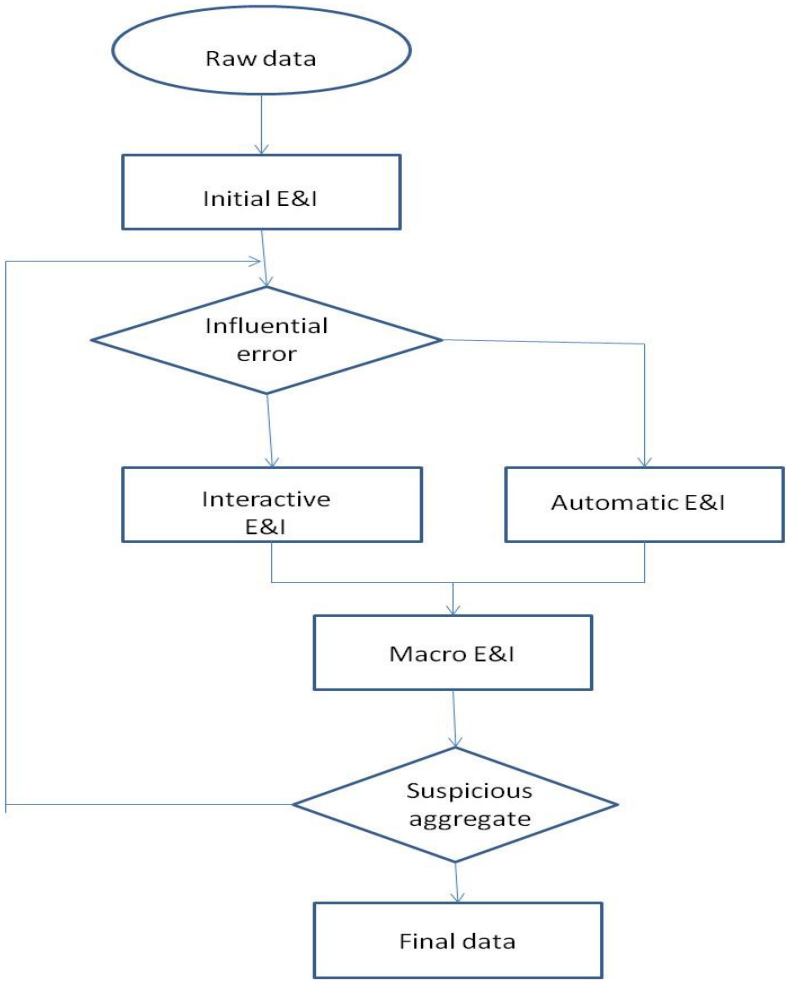
Tilastokeskus

General remarks

- Three main sources of data
  - Business register (Frame)
  - Tax data (Administrative data, accounting data)
  - Direct inquiry (breakdown of turnover and purchases, investments)
- Two interlinked E&I processes
  - Administrative data
  - Direct inquiry data
- Editing is mainly based on variable groups and their relationship
  - Editing rules on tax data (logical rules on tax variables)
  - Editing rules on accounting data (accounting identities)

Tilastokeskus

E&I Process



Tilastokeskus

Using flags in the E&I process

- In automatic E&I processes, the errors and missing values found are flagged
- Also treatments made are flagged by type of treatment
- Flagging means that a separate variabl is made, where the information about error/missing value and their treatment stored
- Flagging in Finnish SBS is done via separate quality variables for different variable groups (TU,KU,SS,VA,VE,OP)
- For example, TU, refers to variables about the revenues of the company and can have values of 1 to 99

Example: Quality variables in FSS

Variable	Description
TU	Income
KU	Costs
SS	Income statement
VA	Assets
VE	Debts
OP	Equity

Example: Quality codes

Code	Description
1	Unit is valid
21	An outlier is detected and treateded
31	A small error in completeness found and scaled
41	Imputed unit by COLD-DECK
51	Imputed unit by HOT-DECK
99	Erroreuos unit, undefined

Tilastokeskus

Using flags in the E&I process, benefits

- Flagging the errors, missing values and corrections
  - Efficient monitoring of the process
  - Analysis of the process
    - Continuous development
  - Technical benefits, selection of group within the process
  - Quality indicators
  - Quality variables for users

Tilastokeskus

Initial E&I

- Error detection by edit rules, logical edits
- Errors are detected for variable groups
- Exact corrections by replacement of values from different source
- Flagging of errors, corrections

Example: exact corrections			Source2
+	Turnover	1 000	
+	Variation in stocks	100	100
+	Manufacturing for own use	100	100
+	Other operating income	100	100
-	Materials and services	-200	-200
-	Personnel costs	-200	-500
-	Depreciation	-100	-500
-	Other expenses	-100	-100
+/-	Financial income and expenses	0	0
+/-	Satunnaiset tuotot ja kulut	100	100
-	Change in cum. accel. depreciation	0	0
+	Change in untaxed reserves	0	0
=	Profit (sum of subvariables)	500	100
=	Profit (stated, from tax data)	100	100

Tilastokeskus

Initial E&I – Outlier detection

- For determining the variable inside variable group, that contains the error

- Calculation of ratio of variable y to the correlated total Y, R = y / Y
- Define the boundaries that the ratio cannot exceed as distance of quantile from median multiplied by multiplier(2 in case of SBS Finland)
- If ratio variable is outside the boundaries, variable is corrected to fulfill the failed edit

Tilastokeskus

Initial E&I – Outlier detection

Example: Initial editing and imputation			Ratio
+	Turnover	1000	
+	Variation in stocks of finished and semifinished goods	100	0,1000
+	Manufacturing for own prodeuce	100	0,1000
+	Other operative income	100	0,1000
-	Raw materials and services	-200	-0,2000
-	Staff expenses	-500	-0,5000
-	Depreciation and reduction in value	-10000	-10,0000
-	Othes operating expences	-100	-0,1000
+/-	Financial income and expences	0	0,0000
+/-	Extraordinary items	100	0,1000
-	Change in cumulative accelerated depreciation	0	0,0000
+	Change in untaxed reserves	0	0,0000
=	Profit (loss) of the financial year (sum of subsets)	-9400	
=	Profit (loss) of the financial year (stated, from Tax authority)	100	

OUTLIER!

INVALID!

Tilastokeskus

Initial E&I – Outlier detection

Example: Initial editing and imputation			Ratio
+	Turnover	1000	
+	Variation in stocks of finished and semifinished goods	100	0,1000
+	Manufacturing for own prodeuce	100	0,1000
+	Other operative income	100	0,1000
-	Raw materials and services	-200	-0,2000
-	Staff expenses	-500	-0,5000
-	Depreciation and reduction in value	-500	-10,0000
-	Othes operating expences	-100	-0,1000
+/-	Financial income and expences	0	0,0000
+/-	Extraordinary items	100	0,1000
-	Change in cumulative accelerated depreciation	0	0,0000
+	Change in untaxed reserves	0	0,0000
=	Profit (loss) of the financial year (sum of subsets)	100	
=	Profit (loss) of the financial year (stated, from Tax authority)	100	

OK! VALID!

Tilastokeskus

Initial E&I – Re-scaling

- All the variables in erroneus variable group are corrected by multipliers that balances the variables to the summary variable

- Calculate the difference between sum of sub-variables and their corresponding total
- Calculate the ratio of the difference to the total, R = difference / sum1+sum2
- If this ratio is below the threshold of re-scaling, multiply all the sub-variables and the summary variable with the ratio, y = y * R

Tilastokeskus

Initial E&I – Re-scaling

	Example: Initial editing and imputation		Scaled
+	Turnover	1050	**1024,39**
+	Variation in stocks of finished and semifinished goods	105	**102,44**
+	Manufacturing for own prodeuce	105	**102,44**
+	Other operative income	105	**102,44**
-	Raw materials and services	-210	**-204,88**
-	Staff expenses	-525	**-512,20**
-	Depreciation and reduction in value	-525	**-512,20**
-	Othes operating expences	-105	**-102,44**
+/-	Financial income and expences	0	**0,00**
+/-	Extraordinary items	105	**102,44**
-	Change in cumulative accelerated depreciation	0	**0,00**
+	Change in untaxed reserves	0	**0,00**
=	Profit (loss) of the financial year (sum of subsets)	105	**102,44**
=	Profit (loss) of the financial year (stated, from Tax authority)	100	**102,44**
	Difference:	5	

OK! Valid!

Coefficient=5/(105+100) 0,024390244

1,024390244

0,975609756

Tilastokeskus

Initial E&I, donor imputation

- On erroneous variable groups
- Donor imputation to be explained later in presentation

Tilastokeskus

Initial E&I, benefits

- Automatic treatment before selective editing and interactive treatment
  - Cost-effectiveness, all systematic errors are automatically corrected
  - Quality
    - Systematic corrections to data
    - Edit rules satisfied
- Continuous development, rules from interactive treatment adopted for automatic treatment
- Outlier detection
  - Efficient corrections for large errors
- Re-scaling
  - Efficient corrections for small errors

Tilastokeskus

Selective editing

- Selective editing is used to separate influental observations from non-influental observations
  - Influental -> Interactive treatment
  - Non-influental -> automatic treatment / no treatment
- Influental means units that have high contribution to the estimates at the level of usage
  - Define the levels of usage (for example NACE 2-digit level)
- Local scores are at first calculated on invidual variables of the observation
- Global score is calculated from local scores
- Observations with global score over the set limit are considered influential

Tilastokeskus

Selective editing, benefits

- Determination of influential errors
- Improves cost-efficiency
- Improves quality
- Offers a way to prioritize observations for interactive editing

Tilastokeskus

Imputation, unit non-response, cold deck

- Donor imputation, previous year (cold deck)
  - For profit and loss account
    - Turnover for statistical year from BR (VAT)
    - Calculation of percentage change in turnover for the obs
    - Retrieval of observations previous year data, if found
    - All data is multiplied by the percentage change in turnover
    - Estimation (regression analysis) of structural changes in variables by activity classes
    - Applying the activity class multipliers to variables
  - Balance sheet items are copied as such from previous year

Tilastokeskus

Imputation, unit non-response , cold deck

Example: Imputing from historical data			Regression coefficient	
Enterprise id: 10101	2008	2009	for the subset	2009
Turnover	1000			
Salaries	-500			
Other costs	-300			
Financial incomes	200			
Financial expences	-100			
Result of the financial year	300			

- Only the 2008 Financial statement is known
- The estimation of turnover from VAT data is known for 2008 and 2009. Let us assume that the turnover for unit decreased by 20%

Tilastokeskus

Imputation, unit non-response , cold deck

Example: Imputing from historical data			Regression coefficient	
Enterprise id: 10101	2008	2009	for the subset	2009
Turnover	1000	800		
Salaries	-500	-400		
Other costs	-300	-240		
Financial incomes	200	160		
Financial expences	-100	-80		
Result of the financial year	300	240		

- The turnover and each subset will be multiplied by the estimated growth rate (-20%)

Tilastokeskus

Imputation, unit non-response , cold deck

Example: Imputing from historical data Enterprise id: 10101	2008	2009	Regression coefficient for the subset	2009
Turnover	1000	800	1,00	
Salaries	-500	-400	1,10	
Other costs	-300	-240	1,00	
Financial incomes	200	160	0,50	
Financial expences	-100	-80	1,50	
Result of the financial year	300	240		

- The regression growth rates are calculated from the valid data for each subset by the Profit and Loss Account

Tilastokeskus

Imputation, unit non-response , cold deck

Example: Imputing from historical data			Regression coefficient	
Enterprise id: 10101	2008	2009	for the subset	2009
Turnover	1000	800	1,00	800
Salaries	-500	-400	1,10	-440
Other costs	-300	-240	1,00	-240
Financial incomes	200	160	0,50	80
Financial expences	-100	-80	1,50	-120
Result of the financial year	300	240		80

- Each subset will be grossed by the regression coefficient
- The unit is imputed, valid and flagged as ”41” (fixed by data from last year)

Tilastokeskus

Imputation, unit non-response, hot deck

- Donor imputation, nearest neighbour (hot deck)
  - Basic principle is finding a similar sized observation in the same activity class, from which items of profit and loss account and balance sheet is received
  - The donor is searced via distance measure

$$D_{ij} = \sum_{i \in V} \left| \log(x_{ik}) - \log(x_{jk}) \right|$$

where D = distance measure

x_{ik} = value for missing unit

x_{jk} = value for the donor

V = vector of variables for which the distance is calculated

(In case of SBS: turnover and personnel)

Tilastokeskus

Imputation, unit non-response , hot deck

- Distance calculation is done with turnover and personnel, logarithmic change applied
- Distance measure uses auxiliary information, which must be known for all units
  - Turnover primarily from Business register, secondary from VAT data
  - Number of personnel from Business register
  - Donors are searched by branch
    - Primary by national 5-digit level, secondary by 3-digit or 1-digit level
    - Donors are searched if atleast 50 units are found
    - If not, then more aggregated level is chosen
- Also a small random term is applied (to prevent needless duplicate donors)
- Ratio of the variables relative to turnover is calculated from donor
- Variables for recipient is calculated by multiplying turnover by this ratio
- Balance sheet items are copied as such

Tilastokeskus

Imputation, unit non-response , hot deck

Example: Nearest neighbour		Donor	Recipient	Recipien
+	Turnover	1 000	500	500
+	Variation in stocks	100		50
+	Manufacturing for own use	100		50
+	Other operating income	100		50
-	Materials and services	-200		-100
-	Personnel costs	-500		-250
-	Depreciation	-100		-50
-	Other expenses	-100		-50
+/-	Financial income and expenses	0		0
+/-	Satunnaiset tuotot ja kulut	100		50
-	Change in cum. accel. depreciation	0		0
+	Change in untaxed reserves	0		0
=	Tilikauden tulos	500		250
		coefficient	0,5	

Tilastokeskus

Imputation, item non-response (1/2)

- Done for direct inquiry items (breakdown of turnover and costs)
- A separate model for each sub-item and each principal activity using information from the direct data collection
- Simple linear regression model with one explanatory variable

$$S_i = a_i T$$

- Dependent variable S_i is the sub-item of turnover (or costs)
- Explanatory variable T is turnover (or costs)
- All variables are ratios of variable divided by personnel
- Regression is done hierarchically for all nace classes, from 5-digit level to 1 digit level

Tilastokeskus

Imputation, item non-response (2/2)

- Outliers are removed from the model
- Coefficients (a_i) are balanced to sum up to 1
- Results are coefficients by NACE 5-digit class
- Observations to be imputed have turnover and costs from tax data
- All sub-items (breakdown of turover or costs) are imputed by multiplying turnover (or costs) by appropriate coefficient

Tilastokeskus

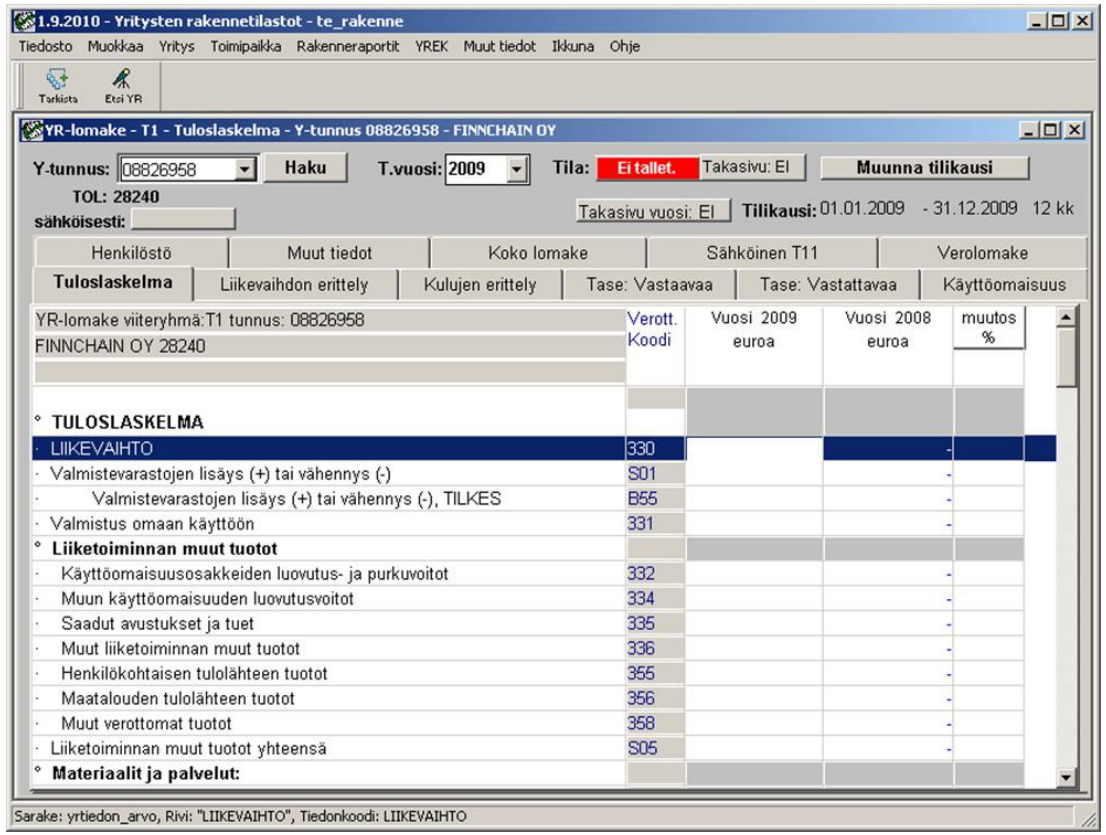
Imputation, benefits

- Mass imputation means complete data in the data base for observations
  - No structural non-responses
  - Easy to use
  - Data can be used and distributed on observation level
  - No need for weights and estimation to totals
  - Variables sum up to total

Tilastokeskus

Interactive treatment

- For influential observations
- Edit rules to point out errors
- Mainly with the help of official financial statements (in pdf)
- 1% of enterprises but 90% of the turnover
- Still very labour intensive
- 8 persons in interactive treating of Financial Statemenst Statistics in Finland
- But (almost) no burden for enterprises!!



The contribution of valid and imputed units

(Year 2009)

[redacted]