Register-based statistics production

Administrative data used for statistical purposes

Bo Sundgren

2010

Part 2

Creating, maintaining, and using statistical registers (and register systems)

- Determine the objectives: Which statistical needs are to be fulfilled by the register(s)?
- Define the desirable contents of the register(s) in terms of
  - primary and derived object types and populations
  - primary and derived variables and value sets (classifications)
  - time granularity (points, intervals)
  - time selection (cross-sectional status data, time series data)
- The inventory phase: Which sources are available, administrative and statistical?
- Editing and integration of the sources:
  - match, check, edit, redefine, derive, and reconcile objects, and synchronise them as regards times of reference
  - combine, check, edit, redefine, derive, and synchronise variables
  - adjust for missing data and other errors: estimates based on weights and calibration vs imputations
  - adjust for changes in definitions and level shifts in time series

Major phases in the design and data collection of register-based statistical surveys (cf traditional surveys)

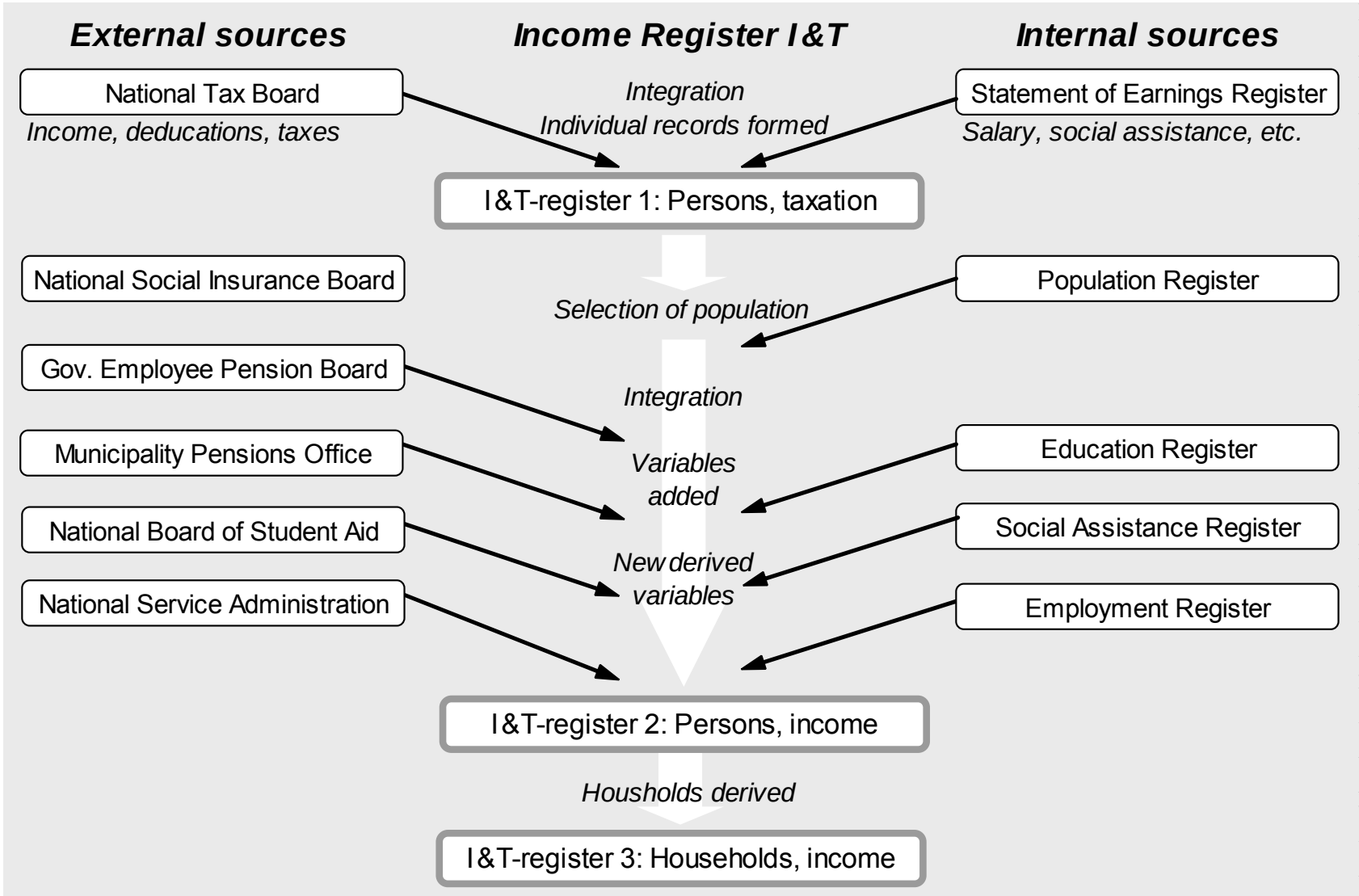
- Fundamental design: design of the (system of) registers, upon which (a large number of different) register-based surveys are based
- Fundamental data collection: data collection for the (system of) registers, upon which (a large number of different) register-based survey are based
- Supplementary design: design of each register-based survey as such, given the registers
- Supplementary data collection: combining and transforming data in the registers, possibly enhanced with some special data collection, with the needs of a particular, register-based survey in mind

Important differences between (sample) surveys and register-based statistics production

- In traditional (sample) surveys, the statistics producer is in full control of the design and data collection processes
- In register-based statistics production, the design and data collection processes are by and large beyond the control of the statistics producer – as regards the external, administrative sources
- The statistics producer is in general control of the statistical registers: how they are designed and created, and how they interact within the statistical system – the infrastructure for register-based statistics production
- The designer of a specific, register-based survey cannot influence the infrastructure: neither the administrative sources, nor the statistical registers, nor the system

An example of the estimation situation in register-based statistics production: estimation of incomes from the I&T Register

Chart 7.1 Data sources and register processing for the Income and Taxation Register



The I&T Register Example

- Assume that we want to estimate equalised disposble income (average disposable indome per consumption unit) of the households in different regions during a certain year
- In the chart on the previous page, six external sources are shown, but in fact there are about 30 administrative sources used
- The inventory work to find these sources and the communication with the responsible administrative authorities have a strong impact on the estimation situation
- The editing, first editing data from each source separately, and then consistency editing of all sources together, is also very important for the estimation situation
- cont'd

The I&T Register Example (cont'd)

- How the register population is defined and created, is fundamental for how the income estimates can be made
- If the population is defined as a calendar year population, the income sum will be greater than if the population is defined as the population at the turn of the year
- If the population by region is defined according to where persons are administratively registered by the Tax Board, or if actual addresses are used, will also influence the regional estimates
- The household unit in a register system is an objects type that is derived with administrative information; the way households are defined and created in the I&T-register is an essential part of the estimation
- Finally, derived variables may be created in different ways, e.g. the variable ”equalised disposable income”

Equalised disposable income

- Disposable (net) income — a sum of income from wage labour, benefits and losses from self-employment, property income, social transfers, regular inter-household cash transfers received and receipts for tax adjustment of which inter-household cash transfers paid, taxes on wealth and repayments for tax adjustment have been subtracted.
- Equalised income — total household income, which is divided by a sum of equivalence scales of all household members.
- Equivalence scale — a weight designated to a household member depending on his/her age to reflect the joint consumption of a household.
- Household — a group of persons living in the common main dwelling (at the same address), who share joint financial and/or food resources and whose members consider themselves to belong to the same household. Household can also consist of one member only.

The estimation process

- The estimation process mirrors the design process and the data collection process
- This is true both for (sample) surveys and register-based statistics production, but the interpretation of ”design process” and ”data collection process” differs in certain respects between the two modes of statistics production
- The purpose of the estimation process is to compensate for errors, uncertainties, and other shortcomings in the design and data collection processes, e.g.
  - sampling error (reflects sampling design)
  - measurement bias (reflects chosen measurement methods and design of measurement instrument)
  - non-response and missing data
  - other errors and uncertainties in collected data
- In register-based statistics production it is important that estimates and error compensations based on the same registers become consistent, even if they are done on different occasions

Defining populations of registers

- General methodology
  - Define the target population
  - Select the intended object set from the base register, giving the register population
  - Match against registers containing interesting variables
  - When receiving hits: import the variable values to the register which is created
  - When receiving mismatches: show missing values (item nonresponse
- Standardised populations created for general usage:
  - end of year version: suitable for annual stock statistics, such as the population on December 31
  - calendar year version: suitable for annual flow statistics, such as the population's income during a specific year
  - monthly/quarterly version: suitable for monthly/quarterly statistics

Important requirements on base registers

- A base register should contain time references, i.e. all events that affect the register's objects should be dated
  - dates of events (birth/deaths, moves, category changes...)
  - dates of registration/update
- A base register should have good coverage (neither overcoverage, nor undercoverage)
- Linkage variables should be of high quality
- Classification/spanning variables should be of high quality, otherwise there will be coverage errors in subpopulations (domains of interest)

Register matching

- When unique, officially authorised identities exist (like for persons and organisations in Sweden), and are used in registers involved, register matching (also called record linkage) is relatively easy
- Nevertheless, errors may occur, because of
  - errors in identities (not so common)
  - errors in references reflecting relations to other objects
  - coverage errors in the registers involved
- When unique, officially authorised identities do not exist (like in many countries), or are not used, more complex and error-prone matching has to take place
- Statistical matching is something else, where the purpose is to find *similar* objects for analytical purposes (or imputation)

Chart 5.7 Frame populations and annual registers

A. Frame population formed in Nov year 1 for years 1 and year 2

Enterprise id	Industry
Idnr 1	DE
Idnr 2	DB
Idnr 3	**DA**
Idnr 4	DC
-	-
-	-

B. Calendar year register formed in autumn year 2 regarding year 1

Enterprise id	Industry
-	-
Idnr 2	DB
Idnr 3	**DB**
Idnr 4	DC
Idnr 5	DG
-	-

C. Calendar year register formed in autumn year 3 regarding year 2

Enterprise id	Industry
-	-
-	-
Idnr 3	**DB**
Idnr 4	DC
Idnr 5	DG
Idnr 6	DC

Chart 5.8 Population definitions in different kinds of surveys

	Advantages	**Disadvantages**
Survey statistics, own data collection	Can be up-to-date	Significant problems with over- and undercoverage and errors in spanning variables if changes are reported late
Register-based statistics	Good coverage, more correct spanning variables	In certain cases, a long delay between the event to the statistics becoming available

A register population, created in the correct manner, has always better quality than the corresponding frame population, as it is based on more and better information.

Creating register variables and their values

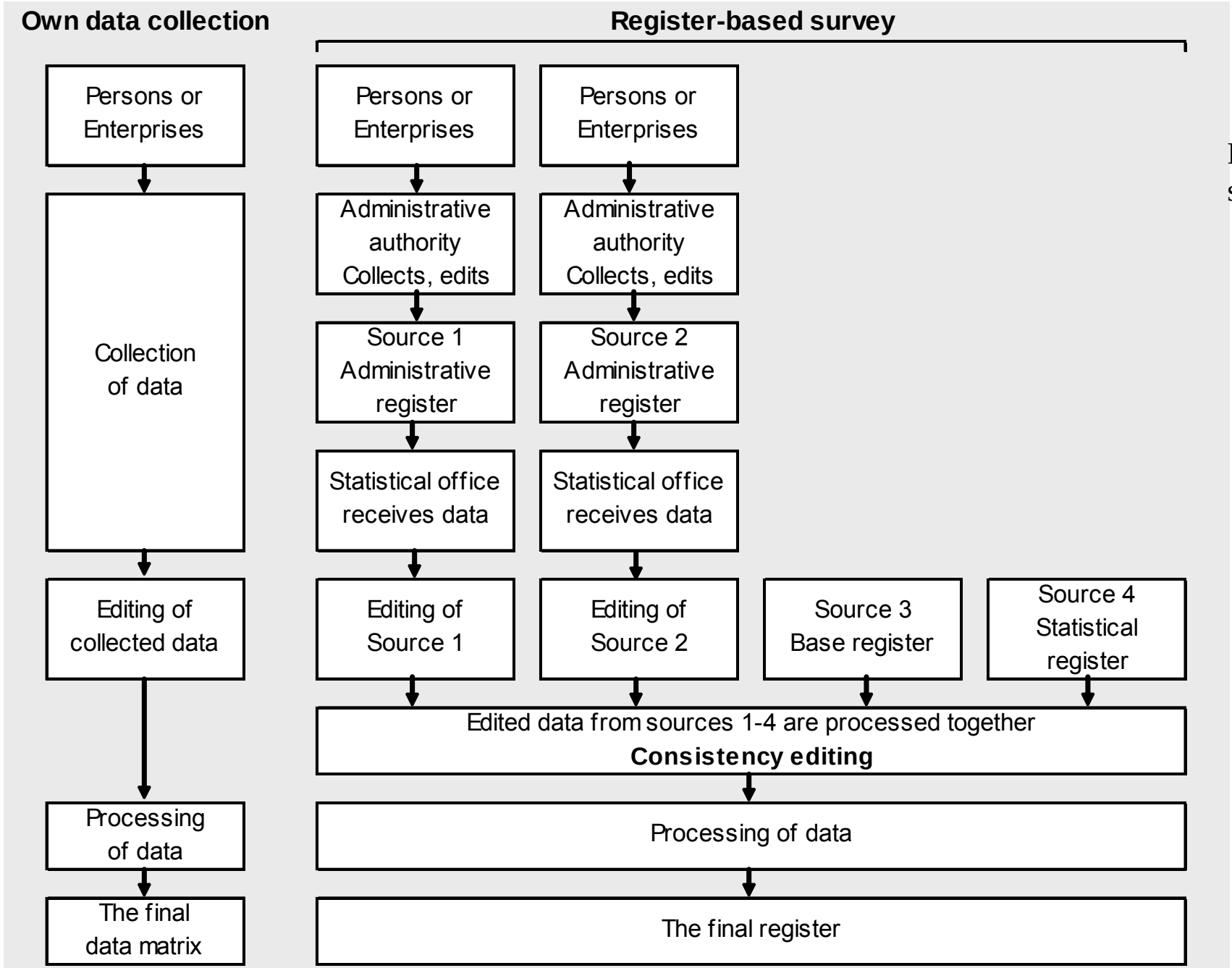
- When creating a statistical register, both objects and variables may come from different sources and need to be carefully checked and reconciled before they are accepted
- The checking and editing that has taken place in the source register, will have been done for other purposes, e.g. administrative purposes
- Derivation of variables (discussed before) and imputation of (missing or suspicious) values of variables are related but different phenomena: a derived variable is created for all objects in a register, whereas an imputed variable value is only formed for the objects in a register where values are missing (or deemed erroneous)

Editing processes in a register system

- Create a data matrix and combine all records that belong to the same object
- Check the register population
- Check that the data regarding a specific identity from different sources really refer to the same object
- Check that the data delivery from administrative sources are complete, both regarding objects and variables; differentiate between missing data, ”variable irrelevant for this object”, and true zero values
- Check variable values for ”obvious” errors
- Make sure that the editing process is documented

For more explanations and illustrative examples: see Wallgren & Wallgren, Chapter 6.

Chart 6.5 Editing in surveys with their own data collection and register-based surveys



Important aspects of data editing

- In many cases, a small number of huge errors destroy data – as a rule it is easy to find and correct these errors
- Use selective editing to find the most important errors first
- Capture knowledge and experiences from domain experts and use this information in documentation and software – neural networks an interesting possibility
- Automatic editing and imputation – pros and cons

The data editing process

- Also called ”data cleaning”
- Purposes of the data editing process:
  - making data ”processable”
  - eliminating simple processing errors, e.g. data entry errors
  - eliminating ”obvious” errors
  - identify suspected errors for further investigation or imputation
  - protect against errors by providing immediate feed-back to interactive respondents, interviewers, and data providers
  - contribute to quality control in the sense of keeping the quality under control; monitoring the quality of the data
- Data editing does not by itself improve the quality of the data – it may even lead to worse data, if the wrong measures are taken for suspected errors

Data editing in different stages of the statistics production process

- During data collection and data entry – this process may possibly be interactive
- As part of the preparation of the raw data that have been collected and entered into the production system → ”clean data”, ”final observation register” (before aggregation/estimation)
- After aggregation/estimation: comparing with previous repetition of ”the same” survey; sometimes called macro-editing, not to be mixed up with selective micro-editing

Different types of edits

- validation edits - to check the validity of basic identification of classificatory data item values
- logical edits - ensure that two or more data items do not have contradictory values
- consistency edits - check to ensure that precise and correct arithmetic relationships exists between two or more data items
- range edits - identify whether or not a data item value falls inside a determined acceptable range
- variance edits - involve looking for suspiciously high variances at the output edit stage.

Different editing strategies

- The book-keeper's strategy: everything should be correct
- The statistician's approach: suspected errors that would influence important estimates to be made should be investigated (first) – also called "selective editing"

Editing of registers

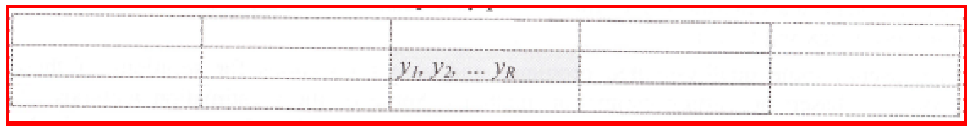
- A register may become used for a wide range of usages and estimation processes, many of which are not known when the register is planned and created
- This makes it difficult to determine which suspected errors are to be investigated (first), as required by the selective editing strategy, since one does not know for sure, which estimations are going to be made in the future, and which of them are to be considered as particularly important

Using weights in estimation

- Weights are used in the estimation process to compensate for uncertainties and errors in data in an efficient and unbiased way
- Weights are often created by stratum, defined by classification/spanning variables, and used for multiplying the observed values of summation/response variables
- Example: sampling
- Example: non-response (missing data)

Weights and calibration

- In a cell in a table, there are R observations from the register, and we want to estimate the cell total Y in the register population. This cell total must sometimes be adjusted for quality reasons.



With *survey samples*, estimates are made using formula (1) shown below. The design weights d_i depend on how the sample has been designed or allocated into different strata. The weights g_i in formula (1) are based on the auxiliary variables from statistical registers and are used to minimise sampling error and errors caused by nonresponse. Deville and Särndal (1992) introduced this method of estimation.

$$\hat{Y} = \sum_{i=1}^{r} d_i g_i y_i = \sum_{i=1}^{r} w_i y_i \qquad (1)$$

where r is the number of objects in the *sample* that responded in a particular cell

The weights d_i are the original weights before calibration, and the weights $d_i g_i = w_i$ are the weights *after* calibration using information about register totals of some auxiliary variables from statistical registers. With formula (1) weighted sums and weighted frequencies are calculated. When calculating mean values, the weighted sums are divided by weighted frequencies.

In register-based surveys, the weights $d_i = 1$ for objects without missing values, and $d_i =$ for objects with missing values. Estimates are here made by using formula (3):

$$\hat{Y} = \sum_{i=1}^{R} d_i g_i y_i = \sum_{i=1}^{R} w_i y_i \qquad (3$$

where R is the number of objects in the *register* in a particular cell

Using weights in register-based surveys

- How can weights be used in a register to make estimates?
- A number of examples will be shown

A salary register

Chart 7.2 A salary register where the observations have weights

Person (1)	Sex (2)	Age (3)	ISCO (4)	Level (5)	Salary (6)	Extent (7)	Salary Full (8)	Salary Class (9)	w_i (10)	$w_i \cdot \text{Salary}_i$ (11)	$w_i \cdot \text{Extent}_i$ (12)	$w_i \cdot \text{SalaryFull}_i$ (13)
PIN1	F	50–54	4190	2	14850	1.00	14850	14–14.9	1.028	15271.4	1.028	15271.4
PIN2	F	40–44	2330	4	16630	0.95	17505	17–17.9	1.031	17147.5	0.980	18049.8
PIN3	M	50–54	2492	4	17807	1.00	17807	17–17.9	1.083	19285.5	1.083	19285.5
PIN4	F	40–44	2330	4	1485	0.09	16500	16–16.9	1.031	1531.2	0.093	17013.5
PIN5	F	40–44	5133	2	6497	0.50	12994	12–12.9	1.031	6699.2	0.516	13398.4
PIN6	F	40–44	5131	2	14102	1.00	14102	14–14.9	1.031	14540.9	1.031	14540.9
PIN7	M	50–54	5131	2	858	0.06	14300	14–14.9	1.083	929.2	0.065	15487.3
...	...	...	...	...	...	...	...	...	...	...	...	...

- The register contains columns (1)-(10), and, when the estimates are carried out, columns (11)-(13) are temporarily formed
- The raw table below is formed by summing up the summation/response variables in register columns (10)-(13) for all combinations of the classification/ spanning variables in columns (2)-(6) and (9)

Basic raw table

Chart 7.3 A raw table that can be used to form many tables

Sex	Age	ISCO	Level	SalaryClass	$\sum w_i$	$\sum w_i \cdot \text{Salary}_i$	$\sum w_i \cdot \text{Extent}_i$	$\sum w_i \cdot \text{SalaryFull}_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
F	17–24	2330	4	12–12.9	42.52	429170	34.55	526165
F	17–24	2330	4	13–13.9	95.67	1293410	95.35	1297704
F	17–24	2330	4	14–14.9	42.52	201399	14.14	622852
F	17–24	2330	4	15–15.9	10.63	159444	10.63	159444
F	17–24	2330	4	16–15.9	53.15	163111	9.89	876942
...	...	...	...	...	...	...	...	...
M	60–64	8320	2	13–13.9	21.24	290107	21.23	290107
M	60–64	8320	2	14–14.9	10.62	149300	10.62	149300
M	60–64	9140	1	12–12.9	10.62	136422	10.62	136422
M	60–64	9140	1	13–13.9	10.62	71348	5.31	142697
M	60–64	9140	1	14–14.9	21.24	308040	21.23	308040

- By further aggregating this raw table in different ways, more tables can be formed for different purposes
- The variable ”salary” is used both as a classification/ spanning variable (column 5) and as a summation/ response variable (columns 7 and 9)

Example of aggregations from the basic raw table

Chart 7.4 Salary distribution by level and sex

A. Number of persons

Level:	1		2		3		4		5	
Salary	Wom.	Men	Wom.	Men	Wom.	Men	Wom.	Men	Wom.	Men
9–11.9	615	107	1823	484	83	32	31			
12–12.9	1138	108	2806	434	199	32	73			
13–13.9	2220	381	10382	1686	239	130	397	174		
14–14.9	560	162	9675	968	900	347	831	194		
15–15.9	114	54	4246	565	1719	533	911	228	21	
16–15.9	21	44	1709	651	1758	467	1293	454	10	
17–17.9			1389	520	1054	468	1675	576	124	
18–18.9	10	11	765	251	786	271	1729	721	114	11
19–19.9			196	122	487	229	1076	644	21	11
20–20.9			73	43	289	110	1492	882	31	21
21–22.9			21	22	237	66	550	567	62	44
23–25.9				11		22	238	412	238	250
26–29.9					10	11	114	205	186	163
30–34.9							52	151	10	99
35–39.9							155	230		44
40–125							145	492		33
Total	4677	869	33084	5758	7762	2717	10763	5930	817	675

B. Per cent

1		...	5	
Wom.	Men		Wom.	Men
13.2	12.3	...	**0.0**	0.0
24.3	12.5	...	**0.0**	0.0
47.5	43.8	...	**0.0**	0.0
12.0	**18.7**	...	**0.0**	0.0
2.4	**6.3**	...	**2.5**	0.0
0.4	**5.1**	...	**1.3**	0.0
0.0	**0.0**	...	**15.2**	0.0
0.2	**1.3**	...	**13.9**	1.6
0.0	**0.0**	...	**2.5**	1.6
0.0	**0.0**	...	**3.8**	3.2
0.0	**0.0**	...	**7.6**	6.4
0.0	**0.0**	...	29.1	**37.1**
0.0	**0.0**	...	22.7	**24.2**
0.0	**0.0**	...	1.3	**14.6**
0.0	**0.0**	...	0.0	**6.5**
0.0	**0.0**	...	0.0	**4.9**
100.0	100.0	...	100.0	100.0

7.4 ESTIMATION USING WEIGHTS – CALENDAR YEAR REGISTERS

Chart 7.8 Calendar year register for the population in a (small) municipality

Person	Existed 1/1 2005	Arrived during 2005 yyyymmdd	Ceased during 2005 yyyymmdd	Existed 31/12 2005	Weight = Time in the municipality, years
PIN1	Yes	-	20050517	No	136/365 = 0.37
PIN2	Yes	-	-	Yes	365/365 = 1.00
PIN3	No	20050315	20050925	No	194/365 = 0.53
PIN4	No	20050606	-	Yes	209/365 = 0.57
Total	**2**			**2**	**2.47**

The traditional way of calculating the average population for 2005 is to form the average value of the population on 1/1 in 2005 (2) and the population on 31/12 in 2005 (also 2). A more specific calculation, in which time in the municipality is used as weight, gives the average population during 2005 as 2.47 persons instead of the traditional measure of 2.

Flow and stock variables

Flow variables, such as value added of an enterprise, only relates to the values during the period of the year in which the enterprise was active, and therefore does not need to be weighted. A stock variable showing the level at a point in time, such as number of employees, must be weighted. The total value added in the region during 2004 was SEK 83 million, while the average number of employees was 112.5. Productivity is calculated as 83/112.5 = SEK 0.738 million per employee and year.

Chart 7.9 Calendar year register for 2004 for enterprises in a particular (small) region

Enterprise identity	Existed 1/1	Arrived	Ceased	Existed 31/12	Weight	Value added	Nr. of employees	Weight • Nr. empl.
EU1	Yes	-	20040630	No	0.50	10	30	0.50 • 30 = 15.0
EU2	Yes	-	-	Yes	1.00	42	45	1.00 • 45 = 45.0
EU3	No	20040401		Yes	0.75	31	70	0.75 • 70 = 52.5
Total					**2.25**	**83**		**112.5**

Weights and calibration of weights vs imputation

- Nonresponse – missing values
  - missing values in registers
  - make no adjustments, publish ”value unknown”
  - adjustment for missing values with weights
  - adjustment for missing values with imputation
  - missing values in a system of registers
- Estimation methods to correct for overcoverage
- Methods to correct for level shifts in time series

Weights and calibration of weights can be used as supplementary estimation methods in these three cases, and imputation can be used to adjust for missing values.

Different ways of publishing statistics with nonresponse

The Labour Force Survey 2001

Labour force category	000s	% of pop.
Employed	4 239	75.3
Unemployed	175	3.1
Not in labour force	1 218	21.6
Population aged 16–64	5 632	100.0

Note: The nonresponse rate in the Labour Force Survey is approximately 15%. The published estimates have been adjusted for the nonresponse.

Education Register 2001

Educational level	000s	% of pop.
Less than 9 yrs	755	11.8
Comp. school 9 yrs	939	14.7
Upper secondary 2 yrs	1 747	27.4
Upper secondary 3 yrs	1 142	17.9
University < 3 yrs	802	12.6
University ≥3 yrs	848	13.3
Postgraduate	48	0.7
Education unknown	**106**	**1.7**
Population aged 16–74	6 386	100.0

- Publish tables with one category ”value unknown”, not adjusting for missing values at all
- Use weights, which have been calibrated to reduce the effects of the missing values
- Impute values when values are missing

Adjustment for missing values with weights in a system of registers

Chart 8.13 Adjustment for missing values with weights in a system of registers

A. Population Reg.

Person	Sex	Age	d_i
PIN1	M	18	1
PIN2	F	72	1
PIN3	M	33	1
PIN4	M	62	1
PIN5	F	71	1
PIN6	F	26	1
PIN7	M	54	1
PIN8	M	67	1
PIN9	F	39	1
...	...	...	...

B. Education Register

PIN	Educ. level	$d_i\, g_i$
PIN1	Comp school 9 yrs	1.01689
PIN2	Less than 9 yrs	1.01689
PIN3	Upper 2nd 2 yrs	1.01689
PIN4	Upper 2nd 3 yrs	1.01689
PIN5	**Missing value**	**0**
PIN6	University ≥ 3 yrs	1.01689
PIN7	Postgraduate	1.01689
PIN8	**Missing value**	**0**
PIN9	Less than 9 yrs	1.01689
...	...	...

C. Employment Register 16–64 years

PIN	Industry	Educ. level	$d_i\, g_i$
PIN1	DM	Comp school 9 yrs	1.02930
-	-	-	-
PIN3	**Missing**	Upper 2nd 2 yrs	**0**
PIN4	DK	Upper 2nd 3 yrs	1.02183
-	-	-	-
PIN6	DB	University ≥ 3 yrs	1.02326
PIN7	DK	Postgraduate	1.02326
-	-	-	-
PIN9	DM	Less than 9 yrs	1.02930
...	...	...	...

- If each register is adjusted separately for missing values with weights, the weights for the same person will be different in the three registers
- If statistics from different registers that relate to the same population are to be consistent, weights must be calculated jointly, and the same weights must be used for all the registers
- This can be difficult to achieve
- Conclusion: adjustment for missing values with weights will cause problems for coordination and consistency within a register system

Adjustment for missing values with imputation in a system of registers

Chart 8.14 Adjustment for missing values in the Education Register using imputation

A. Actual register

Person	Sex	Age	Educational level	Random number
PIN1	M	18	Comp school 9 yrs	0.7771
PIN2	F	72	Less than 9 yrs	0.3168
PIN3	M	33	Upper 2nd 2 yrs	0.3096
PIN4	M	62	Upper 2nd 3 yrs	0.8667
PIN5	F	71	**Missing value**	**0.1749**
PIN6	F	26	University ≥ 3 yrs	0.4114
PIN7	M	54	Postgraduate	0.1605
PIN8	M	67	**Missing value**	**0.5536**
...	...	...	...	...

B. Data matrix for analysis

Sex	Age	Educational level	Educ. level imputed
M	18	Comp school 9 yrs	No
F	72	Less than 9 yrs	No
M	33	Upper 2nd 2 yrs	No
M	62	Upper 2nd 3 yrs	No
F	71	**Comp school 9 yrs**	Yes
F	26	University ≥ 3 yrs	No
M	54	Postgraduate	No
M	67	**Upper sec 3 yrs**	Yes
...	...	...	...

Chart 8.15 Adjustment for missing values in the Business Register using imputation

A. Actual register

Enterprise	Industry	Random number
LeU1	DB	0.0316
LeU2	DK	0.6444
LeU3	**Missing value**	**0.3978**
LeU4	DA	0.2846
LeU5	DK	0.2044
...	...	...

B. Data matrix for analysis

Industry	Industry imputed
DB	No
DK	No
DM	Yes
DA	No
DK	No
...	...

Chart 8.16
Adjustment for missing values in the Employment Register with imputation

A. Actual register

Person	Enter-prise	Industry	Random number Industry	Educational level	Random number Education
PIN1	LeU5	DK	0.2044	Comp school 9 yrs	0.7771
PIN2	-	-	-	Less than 9 yrs	0.3168
PIN3	LeU3	**Missing**	**0.3978**	Upper 2nd 2 yrs	0.3096
PIN4	LeU2	DK	0.6444	Upper 2nd 3 yrs	0.8667
PIN5	-	-	-	**Missing value**	**0.1749**
PIN6	LeU1	DB	0.0316	University ≥ 3 yrs	0.4114
PIN7	LeU5	DK	0.2044	Postgraduate	0.1605
PIN8	-	-	-	**Missing value**	**0.5536**
...	...	...	...	...	...

B. Data matrix for analysis

Indu-stry	Industry imputed	Educational level	Educ. level imputed
DK	No	Comp school 9 yrs	No
-	-	Less than 9 yrs	No
DM	Yes	Upper 2nd 2 yrs	No
DK	No	Upper 2nd 3 yrs	No
-	-	**Comp school 9 yrs**	Yes
DB	No	University ≥ 3 yrs	No
DK	No	Postgraduate	No
-	-	**Upper 2nd 3 yrs**	Yes
...	...	...	...

Adjustment for missing values in a system of registers: conclusions

- Adjustment for missing values should be done
- Adjustments must be coordinated
- Imputation is the most appropriate method
- Within the system in the example above, the Education Register is responsible for the nonresponse adjustment for the variable *Education*, the Business Register is responsible for the nonrespone adjustment for the variable *Industry*; other registers should use these adjustments

Estimation with combination objects: aggregation errors

Chart 9.1 Number of employed and wage sums in different registers

Register 1 – Persons

Person	Sex	Wage sum	1st Industry
PIN1	M	450 000	D
PIN2	F	210 000	D
PIN3	M	270 000	A

Register 2 – Job activities

Job	Person	Local unit	Wage sum	Industry	Sex
J1	PIN1	LU1	220 000	A	M
J2	PIN3	LU1	180 000	A	M
J3	PIN1	LU2	230 000	D	M
J4	PIN2	LU2	210 000	D	F
J5	PIN3	LU2	90 000	D	M

Aggregation

Aggregation

Register 3 – Local units

Local unit	Industry	Wage sum	Nr empl	Prop F
LU1	A	400 000	2	0.00
LU2	D	530 000	3	0.33

Three persons in Register 1, five persons in Register 3

Estimation with combination objects: many-to-one relations

Chart 9.4 Number of employed and wage sums in different registers

Register 1 – Persons

Person	Sex	Wage sum	1st Industry
PIN1	M	450 000	D
PIN2	F	210 000	D
PIN3	M	270 000	A

Aggregation

Register 2 – Job activities

Job	Person	Local unit	Wage sum	Industry	Sex
J1	PIN1	LU1	220 000	A	M
J3	PIN1	LU2	230 000	D	M
J4	PIN2	LU2	210 000	D	F
J2	PIN3	LU1	180 000	A	M
J5	PIN3	LU2	90 000	D	M

Chart 9.5 Number of employed and wage sums in different registers

Person	Sex	Local unit	Wage sum	Industry	Weight
PIN1	M	LU1	220 000	A	22/45 = 0.49
PIN1	M	LU2	230 000	D	23/45 = 0.51
PIN2	F	LU2	210 000	D	21/21 = 1.00
PIN3	M	LU1	180 000	A	18/27 = 0.67
PIN3	M	LU2	90 000	D	9/27 = 0.33

Chart 9.6 Employed by Industry

Industry	Number of employed
A	0.49 + 0.67 = 1.16
D	0.51 + 1.00 + 0.33 = 1.84
Total	3.00

Estimation with combination objects: consistency between different variables

9.2.5 Consistency between different variables

To ensure consistency when using the different multi-valued variables in the system, the weights for these should be included in the register that is responsible for each respective multi-valued variable. Everyone should then use these weights.

Certain registers contain many variables that need to fulfil certain consistency conditions, such as records in a profit and loss statement. Consistency is maintained if the same weights are used for all variables, or if sub-records are recalculated with different weights first, and then the totals and differences are calculated.

Estimation with combination objects: multi-valued variables – summary of recommendations

A variety of important variables in the register system are multi-valued. The current way of handling these variables can, in some cases, produce estimates with aggregation errors. By using combination objects and weights when estimating, these errors can be reduced. In this section, a series of different estimation problems with multi-valued variables is described, and suggestions are made for solutions to these problems. The above example on the change in Industry shows how relatively simple methods, such as using weights, can bring about quality improvements, even though the weights being used are not completely perfect.

Another important advantage with the estimation method presented in this section (Section 9.2) is that economic statistics for different kinds of enterprise units can be made consistent with each other – these inconsistencies are today a serious problem.

Documentation, metadata, and quality

- Registers (and systems of registers) should be accompanied by the same kind of documentation, metadata, and quality declarations as other statistical systems (e.g. surveys and censuses) and data collections (files, observation registers, databases)
- Examples:
  - The SCBDOK documentation template for processes and microdata
  - The quality declaration template of Statistics Sweden
- However, there are some special issues for registers:
  - data from many sources, some of which are outside the control of the statistics producer
  - complex chains of derivations, combinations, and tranformations of data (objects, variables, relationships, time)

Template for documentation of processes and microdata

SCBDOK 3.0	
0 General information 0.1 Subject matter area 0.2 Statistics area 0.3 Official statistics? 0.4 Responsibility 0.5 Producer 0.6 Mandatory response? 0.7 Secrecy 0.8 Destruction rules 0.9 EU regulation 0.10 Purpose and history 0.11 Users and usage 0.12 General approach to implementation 0.13 Planned changes	**1 Contents overview** 1.1 Observation characteristics 1.2 Statistical target characteristics 1.3 Outputs: microdata and statistics 1.4 Documentation and metadata **2 Data collection** 2.1 Frame and frame procedure 2.2 Sampling procedure (if applicable) 2.3 Measurement instruments 2.4 Data collection procedure 2.5 Data preparation
3 Final observation registers 3.1 Production versions 3.2 Archive versions 3.3 Experiences from the latest collection round	**4 Statistical processing and presentation** 4.1 Estimations: assumptions and formulas 4.2 Presentation and dissemination procedures
5 Data processing system	**6 Logbook**

The Quality Declaration Template of Statistics Sweden

Quality Declaration Template	
1 Contents 1.1 Statistical target characteristics 1.1.1 Objects and population 1.1.2 Variables 1.1.3 Statistical measures 1.1.4 Study domains 1.1.5 Reference time 1.2 Comprehensiveness	**2 Accuracy** 2.1 Overall accuracy 2.2 Sources of inaccuracy 2.2.1 Sampling 2.2.2 Coverage 2.2.3 Measurement 2.2.4 Non-response 2.2.5 Data processing 2.2.6 Model assumptions 2.3 Presentation of accuracy measures
3 Timeliness 3.1 Frequency 3.2 Production time 3.3 Punctuality	**4 Coherence especially comparability** 4.1 Comparability over time 4.2 Comparability over space 4.3 Coherence in general
5 Availability and clarity 5.1 Forms of dissemination 5.2 Presentation 5.3 Documentation 5.4 Access to microdata 5.5 Information services	

Statistical confidentiality

- According to the law: the publicity law does not apply to data that have been collected for statistical purposes; they are protected by the secrecy law
- Statistical confidentiality is in the self-interest of the statistics producer: to be able to get honest answers, and to be able to come back to the respondent
- Reidentification of sensitive information about individual objects (person, enterprises) is always a possibility, both on macro level (aggregated statistics) and micro level (anonymised data about single objects)
- Privacy is the typical concern of persons, economical interest is the typical concern of enterprises
- Reidentification of business data is relatively easy, especially in a small country with big enterprises, whereas reidentification of person data is relatively difficult, especially in sample surveys, but made easier by the existence of public registers
- In Sweden it is a crime to even try to reidentify statistical data
- Statistical confidentiality is best protected by a combination of legal, administrative, and technical measures

Conclusions

- Need for a new methodology for registers? A new approach?
- Registers, administrative data, and book-keeping of people and resources have always been part of the mainstream of official statistics – during the 20th century sample surveys came in as a new element of official statistics
- Many survey methods are relevant for registers, and vice versa, but some survey methods have to be reinterpreted and adapted before they can be applied to register-based statistics
- In register-based statistics there is a long and complex route from a wide range of data sources, many of which are outside the control of the statistics producer, via complex data collection and data transformation processes, to a wide range of final statistics for different purposes, at different times
- Design, data collection, data editing, and estimation processes in connection with register-based statistics production, have to cope with more complex multi-source and multi-purpose situations, and they also encounter more difficult challenges to ensure consistency between the wide range of statistics that could potentially be produced from a system of registers (→ micro rather than macro)