

Register-based statistics production

Administrative data used for statistical purposes

Bo Sundgren

2010

Part 1

Basic definition of a register

(register in a strict and narrow sense)

- A register is **an authorised, up-to-date list of all objects belonging to a certain population**
- The objects are uniquely identified by an authorised **identifier**, such as person number for persons, organisation number for enterprises and other organisations, etc
- In addition to the identifier, a register may contain **additional basic and up-to-date information** about the objects, such as name (not necessarily unique) and location and other **contact information**, e.g. address and telephone number

Cf. Wallgren&Wallgren, page 4. But many other definitions are used throughout the book, often vaguely or implicitly.

Extended register

(register in a broader sense)

- An authorised, up-to-date list of all objects belonging to a certain population
- The objects are uniquely identified by an authorised identifier
- In addition to the identifier, a register may contain additional basic and up-to-date information about the objects, such as name (not necessarily unique) and location and other contact information, e.g. address and telephone number
- **Furthermore**, an extended register may contain
 - links to other registers and data sources, as well as
 - additional information from those other sources

Administrative registers and statistical registers

- An **administrative register** is a register used for **administrative purposes**, e.g. by a government agency or an enterprise
- A **statistical register** is a register used for **statistical purposes**, e.g. by a statistical agency or an enterprise
- A statistical register may be created from
 - one or more administrative registers,
 - sometimes in combination with information from other sources, such as
 - other **administrative sources**, e.g. administrative databases, or
 - other **statistical sources**, e.g. surveys and statistical databases
- A basic function of a statistical register is to serve as a (sampling) **frame** for (sample) surveys
- Extended statistical registers are also useful by themselves as **sources for statistics production**

Sources of data used by a statistical agency

- censuses
- surveys
 - sample surveys and total surveys
 - repeated surveys and ad hoc surveys
- administrative systems
 - administrative registers
 - other administrative data collections, e.g. databases
- statistical registers
- archives
 - data already collected and stored by the statistical agency or by somebody else

About 97-99% of the data used by Statistics Sweden come from administrative sources. The cost of a value collected from an administrative source is about 1% of the cost of collecting it by a survey. Cf the virtual census of the Netherlands: 3 M€ vs 300 M€.

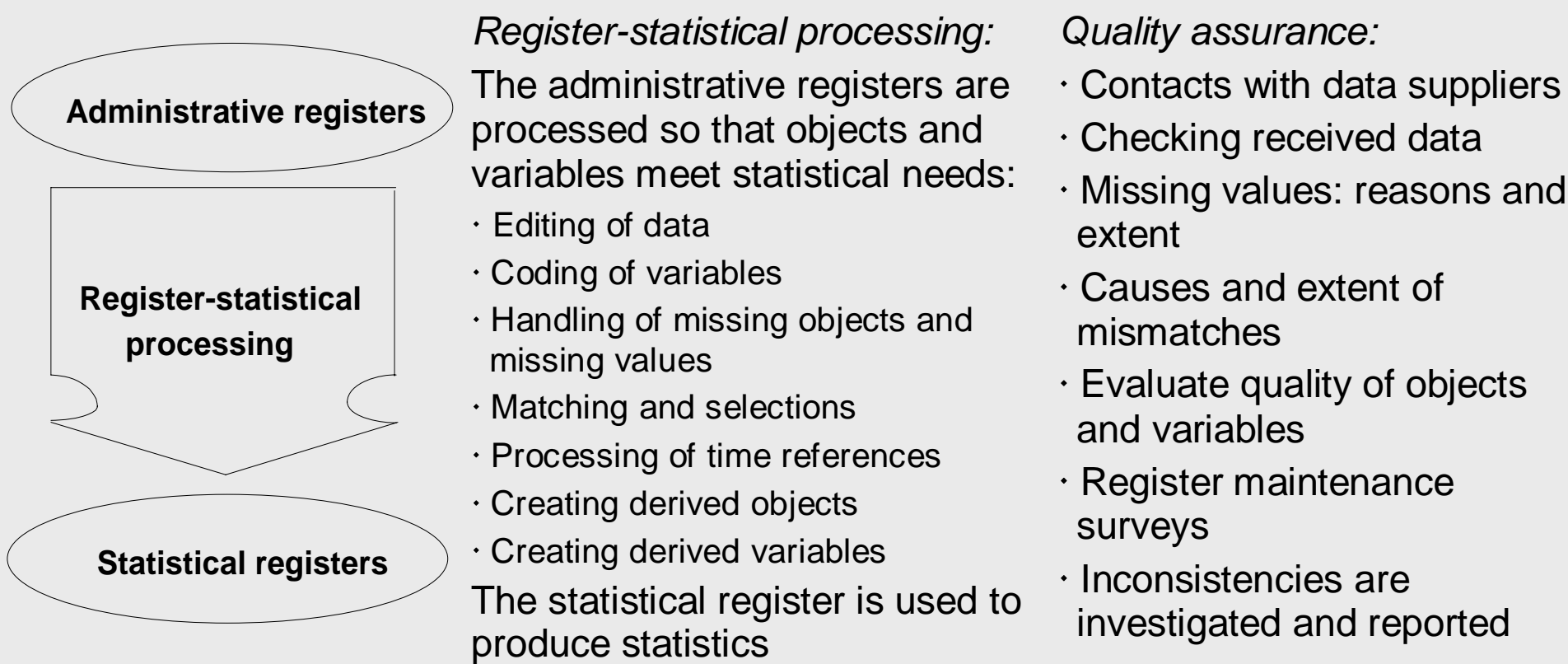
Chart 1.1 Four principles on how to use administrative data

1. A statistical office should have access to administrative registers kept by public authorities. This right should be supported by law as the protection of privacy.
2. These administrative registers should be transformed into statistical registers. Many sources should be used and compared during this transformation.
3. All statistical registers should be included in a coordinated register system. This system will ensure that all data can be integrated and used effectively.
4. Consistency regarding populations and variables are necessary for the coherence of estimates from different register-based surveys.

2. These administrative registers should be transformed into statistical registers. Many sources should be used and compared during this transformation.

Wallgren&Wallgren, page 5:

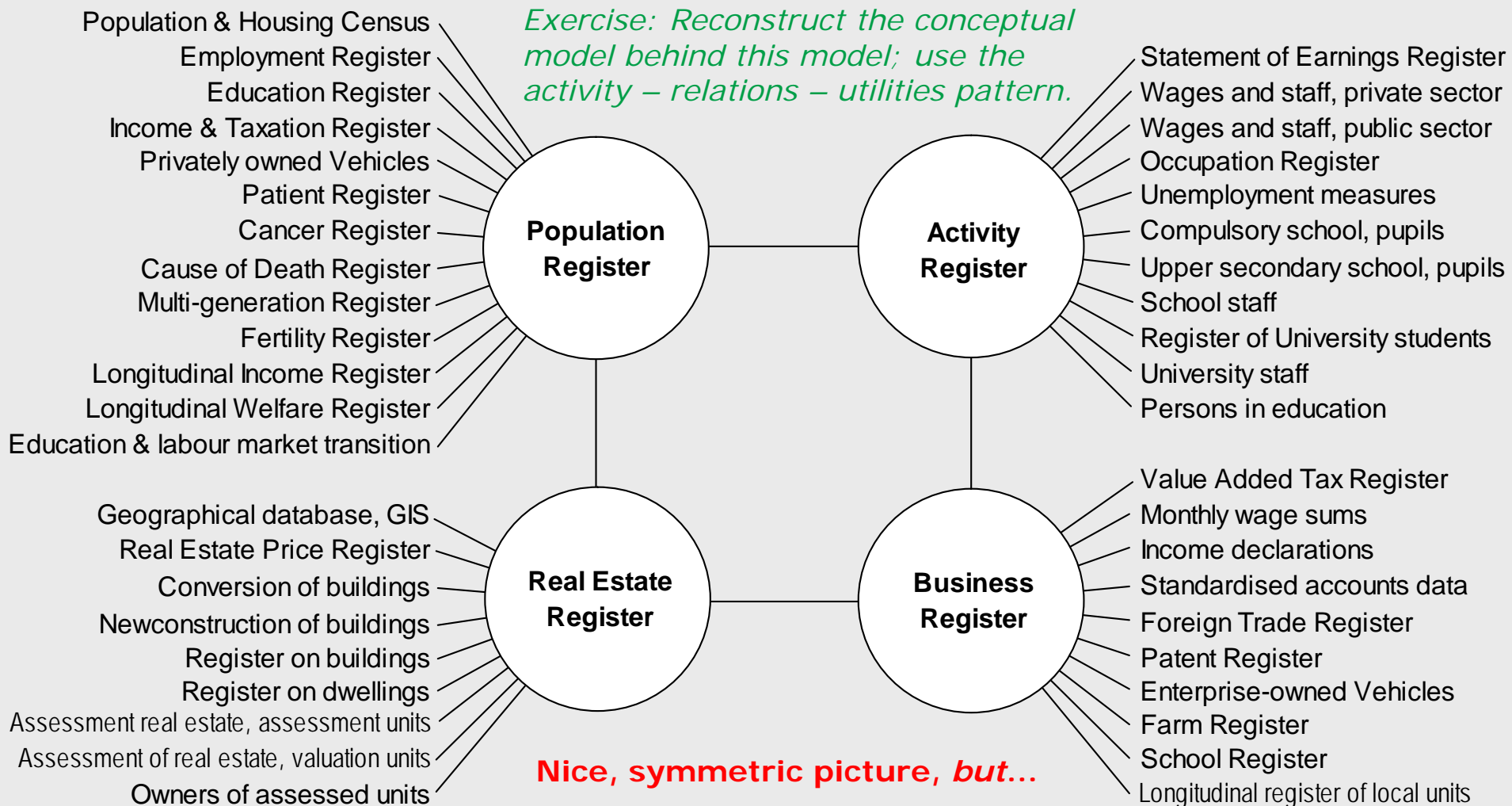
Chart 1.3 From an administrative register to a statistical register



3. All statistical registers should be included in a coordinated register system. This system will ensure that all data can be integrated and used effectively.

Wallgren&Wallgren Page 30: (Note: A very broad and fuzzy register concept is used here.)

Chart 2.10 A system of statistical registers – registers by object type and subject field



Note: This model is not a model of Statistics Sweden's system, the model shows general possibilities.

4. Consistency regarding populations and variables are necessary for the coherence of estimates from different register-based surveys.

Wallgren&Wallgren, Page 32:

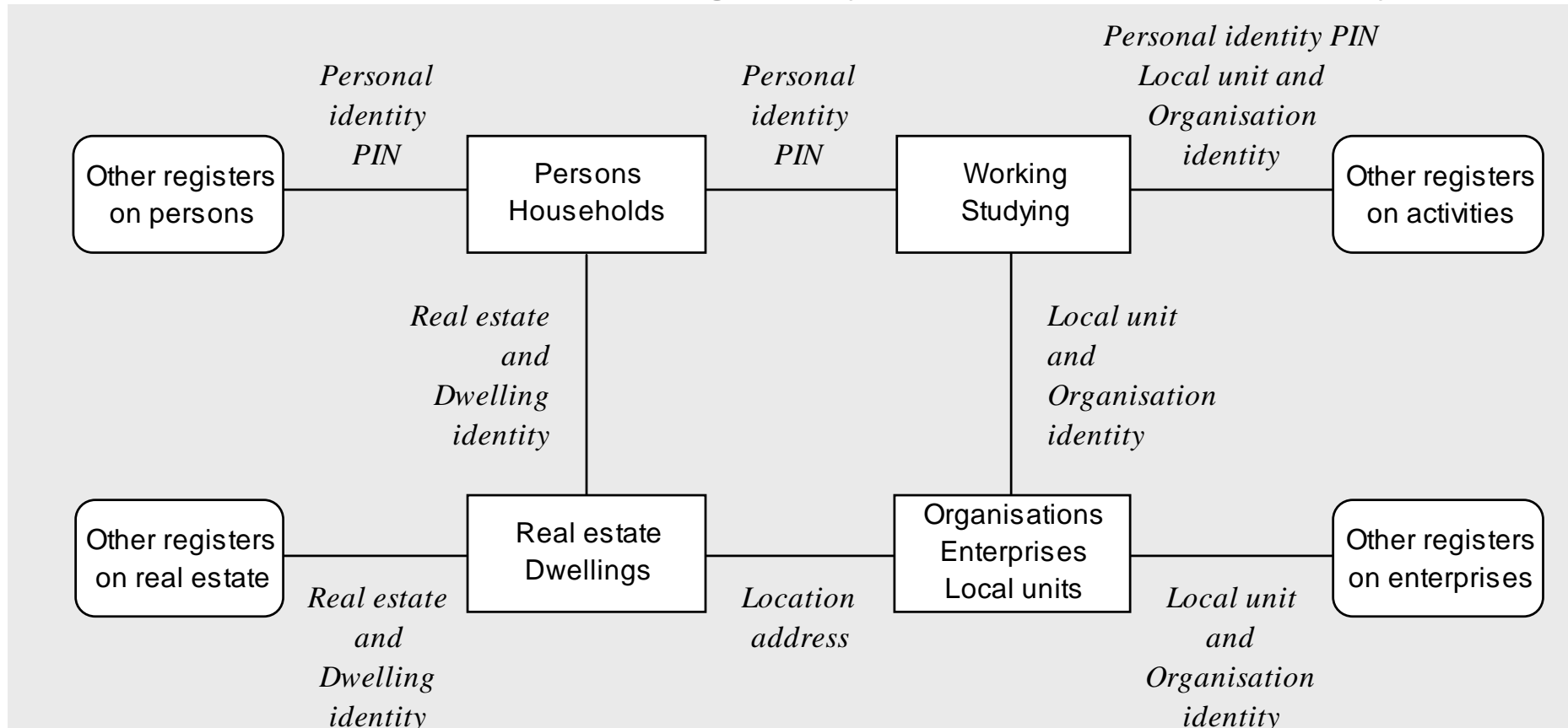
Chart 2.13 Register-based statistics for one small municipality in Sweden 2003

Population		Employment		Education					Income Yearly earned, \$ thousands				
Age	Number	Em- ployed	Not em- ployed	Com- pulsory	Upper secon- dary	Post- secon- dary	Post- gra- duate	Not known	0	1–14	15–29	30–44	45+
0–15	1416	-	-	-	-	-	-	-	-	-	-	-	-
16–19	387	69	318	306	71	0	0	10	118	265	4	0	0
20–24	293	207	86	44	219	26	0	4	12	130	128	23	0
25–34	764	616	148	79	469	210	0	6	20	133	388	202	21
35–44	937	782	155	142	558	226	2	9	27	128	440	270	72
45–54	1002	847	155	259	510	225	4	4	14	90	501	318	79
55–64	1042	713	329	420	413	199	6	4	21	166	502	288	65
65+	1199	40	1159	333	168	78	3	617	3	552	535	90	19

"Register system" or "Register-based statistical system"

- A **national statistical system** could be defined as a system of statistical data about socio-economic conditions and developments in a country
- A **system** always consists of related (coordinated, integrated) subsystems, parts, or components
- The **related data collections** in a statistical system could come from different sources: censuses, surveys, registers, (other) administrative sources
- In a **register-based statistical system**, registers play an important role as a *"backbone"* and coordination tool within the statistical system

Chart 2.5 A conceptual model of a register system of statistics on society



(Wallgren & Wallgren page 22)

Exercise: Reconstruct the conceptual model behind this model; use the activity – relations – utilities pattern.

Base registers = "important registers"

Chart 2.6 The characteristics of a base register

1. Defines important object types.
2. Defines important object sets or standardised populations.
3. Contains links to objects in other base registers.
4. Contains links to other registers that relate to the same object type.
5. Is important for the system as a whole – which is why it is essential for them to be of high quality and be well-documented.
6. Is important as a sampling frame.
7. Can be used for demographic statistics regarding persons, activities, real estate or enterprises.

In the same way that age distribution and births and deaths in a population of persons are described, it should be possible to describe age distribution and births and deaths among jobs, buildings or local units. Birth dates and death dates must be available in the base register so that demographic statistics can be produced.

(Wallgren&Wallgren, Page 23)

Backbone of the Swedish register system

Population register – exists and works well; person, household?

Activity register – exists; note: relational objects

Business register – exists; object type: (part of) organisation

Real estate register – exists; dwellings underway

Chart 2.8 The relation between registers on persons, activities and enterprises

Population Register – Persons	
Person	Wage sum
PIN1	450 000
PIN2	210 000
PIN3	270 000

Activity Register – Jobs			
Job	Person	Local unit	Wage sum
J1	PIN1	LU1	220 000
J2	PIN3	LU1	180 000
J3	PIN1	LU2	230 000
J4	PIN2	LU2	210 000
J5	PIN3	LU2	90 000

The Activity Register contains the bi-variate distribution and the Business and Population Registers contain marginal distributions

Business Register – Local units	
Local unit	Wage sum
LU1	400 000
LU2	530 000

Conceptual models and data models (cf previous course)

Chart 3.5 A database on individuals with three database tables

A. Conceptual database model

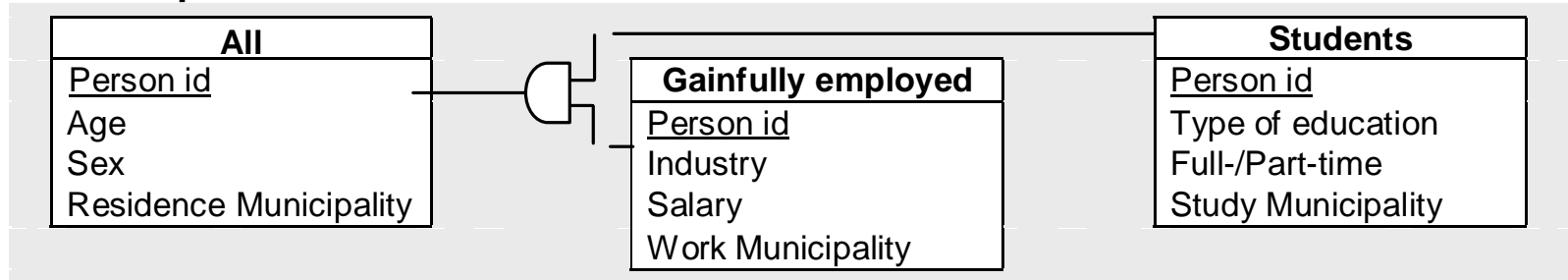


Chart 3.5 A database on individuals with three database tables

B. Example of content in the database

All

Person id	Age	Sex	ResMun.
PIN1	20	F	0586
PIN2	23	M	0586
PIN3	31	M	0586
PIN4	32	F	0586
PIN5	33	M	0586
PIN6	40	F	0586
PIN7	59	F	0586
PIN8	65	M	0586
PIN9	71	F	0586

Gainfully employed

Person id	Industry	Salary	WorkMun.
PIN2	G	52 000	0586
PIN3	G	287 000	0580
PIN4	A	193 000	0586
PIN6	D	291 000	0586
PIN7	D	314 000	0580

Students

Person id	Educ.Type	Full/Part-time	StudMun.
PIN1	AdultEduc	100	0586
PIN2	Univ	100	0580
PIN5	Univ	100	0580

Example

Chart 3.6 Two data matrices for different statistical purposes

A. Data matrix: Employment register

Person	Age	Sex	Emp- loyed	Industry	Salary
PIN1	20	F	No	null	0
PIN2	23	M	Yes	G	52 000
PIN3	31	M	Yes	G	287 000
PIN4	32	F	Yes	A	193 000
PIN5	33	M	No	null	0
PIN6	40	F	Yes	D	291 000
PIN7	59	F	Yes	D	314 000
PIN8	65	M	No	null	0
PIN9	71	F	No	null	0

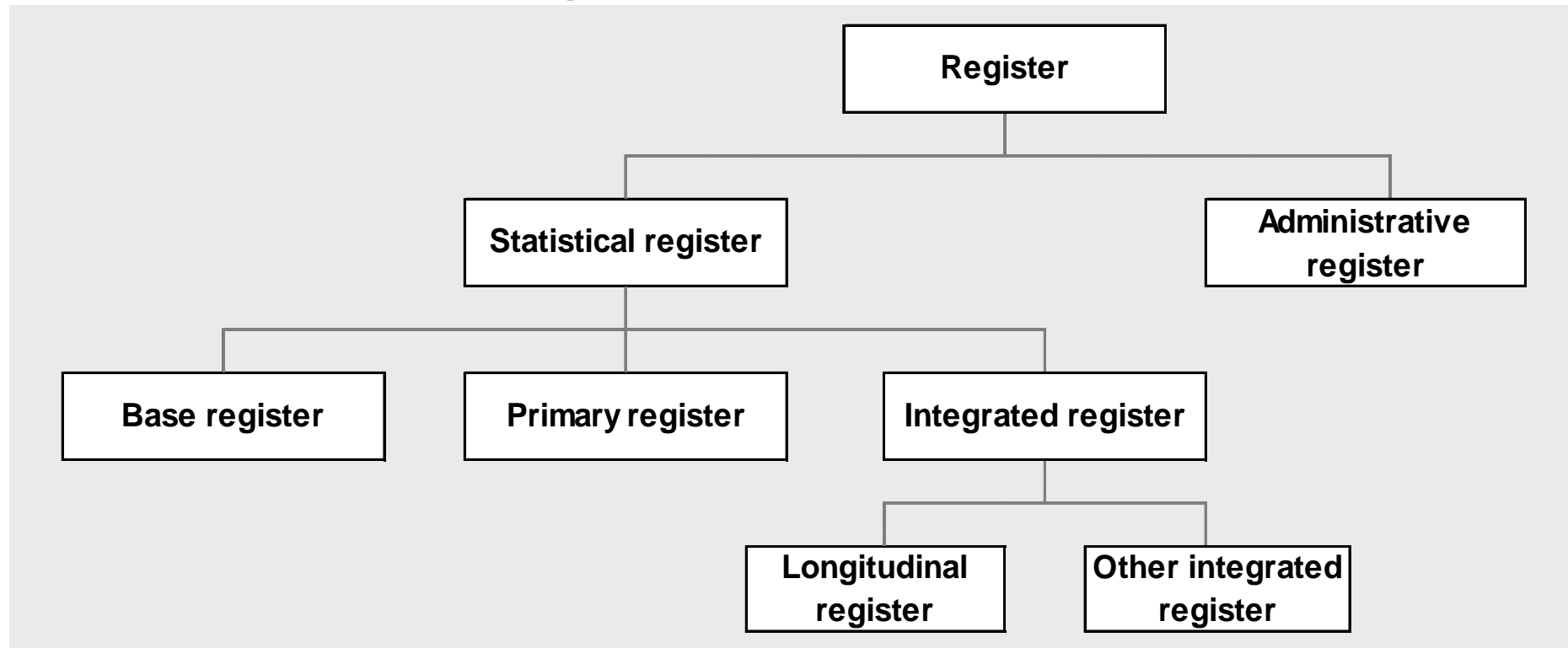
B. Data matrix: Commuting register

Person	ResMun.	WorkMun.	StudMun.	Com- muting
PIN1	0586	null	0586	0
PIN2	0586	0586	0580	1
PIN3	0586	0580	null	1
PIN4	0586	0586	null	0
PIN5	0586	null	0580	1
PIN6	0586	0586	null	0
PIN7	0586	0580	null	1
PIN8	0586	null	null	0
PIN9	0586	null	null	0

Exercise: Reconstruct the conceptual models behind the two data matrices!

data matrix = relational data table

Chart 3.8 Different kinds of registers



- **Base register**: **important** object type, **important** population, **links** to other objects or to same objects in other register, **important** as a frame, **important** for the system as a whole, can be used for demographic statistics
- **Primary register**: directly based on at least one administrative register
- **Integrated register**: created from other statistical registers only
- **Longitudinal register**: integrated register, where it is possible to follow objects over time

Part 1: Extra material

Bo Sundgren
2010

Registers and time

- **Current stock register**: updated with all information on currently active/live objects; used as frame for (sample) surveys
- **Register referring to a specific point in time**, e.g. turn of the year, updated *after* the point in time when reports about "all" events have arrived; used for register-based surveys
- **Calendar year register**: containing all objects that have existed at any point during a specific year; used for register-based surveys
- **Events register**: all events during a specific period
- **Historical register**: all events, used for longitudinal surveys
- **Longitudinal register**: all events during a time period

1. The *current stock register* ...is used as frame population for sample surveys or censuses.
2. The *register referring to a specific point in time*, ... Is used for register-based surveys.
3. The *calendar year register* .. It is used as register populations for register-based surveys.

Chart 3.9 Calendar year register for 2002

Object identity	Existed 1/1	Added	Ceased to exist	Existed 31/12	Other variables
Idnr 1	Yes	-	20020517	No	...
Idnr 2	Yes	-	-	Yes	...

4. The *events register* for a specific period, It is used in register-based surveys.

Chart 3.10 Events register for 2002 regarding change of address

Object identity	Address 1/1	Date of change of address	New address
Idnr 1	Address 11	20020517	Address 21

5. The *historical register* It is used for longitudinal surveys.

Chart 3.11 Historical register regarding change of address

Object identity	From address	Date of change of address	To address
Idnr 1	Born	19670517	Address 1
Idnr 1	Address 1	19810606	Address 2

6. A *longitudinal register* for a period of time

Chart 3.12 Longitudinal register for 2000–2002

Object identity	Existed 1/1/2000	Added	Ceased to exist	Income 2000	Income 2001	Income 2002
Idnr 1	Yes	-	20010517	183 450	97 600	-
Idnr 2	Yes	-	-	273 500	281 360	258 340

Derived objects

Example 1: Household

- <http://stats.oecd.org/glossary/detail.asp?ID=1255>
- "A household is a small group of persons who share the same living accommodation, who pool some, or all, of their income and wealth and who consume certain types of goods and services collectively, mainly housing and food." (SNA 4.132 [4.20]).

Derived objects

Example 2: Part of organisation

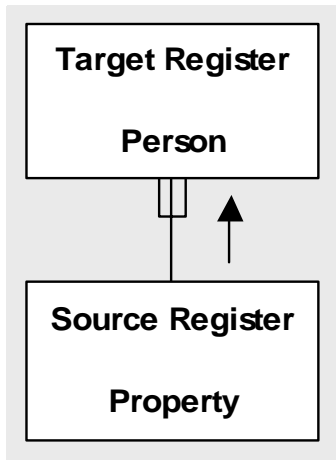
- “An **organisation** is a unique framework of authority within which a person or persons act, or are designated to act, towards some purpose.” (OECD)
- “An **enterprise** is an institutional unit in its capacity as a producer of goods and services; an enterprise may be a corporation, a quasi- corporation, a non-profit institution, or an unincorporated enterprise.” (SNA)
- “An **institutional unit** is an economic entity that is capable, in its own right, of owning assets, incurring liabilities and engaging in economic activities and in transactions with other entities.” (SNA)
- Eurostat: The Council Regulation ((EEC), No. 696/93 of 15 March 1993) on statistical units for the observation and analysis of the production system in the Community lays down a list of **eight (types of) statistical units**:
 - the **Enterprise**;
 - the **Institutional Unit**;
 - the **Enterprise Group**;
 - the **Kind-of-activity Unit** (KAU);
 - the **Unit of Homogeneous Production** (UHP);
 - the **Local Unit**;
 - the **Local Kind-of-Activity Unit** (local KAU);
 - the **Local Unit of Homogeneous Production** (local UHP).

Different kinds of variables

- Stocks and flows
 - a **stock variable** is measured *at one specific time*, and represents a quantity existing at that point in time, which may have been accumulated over time
 - a **flow variable** is measured over an interval of time and would be measured *per unit of time*
- Quantities and qualities
 - a **quantitative variable** has values that can be summarised in a meaningful way
 - a **qualitative variable** has values (codes) that cannot be summarised in a meaningful way, although the codes may sometimes be numerical (sex = 1, 2)
- Open values (free text) vs predefined codes
- Single-valued vs multi-valued variables

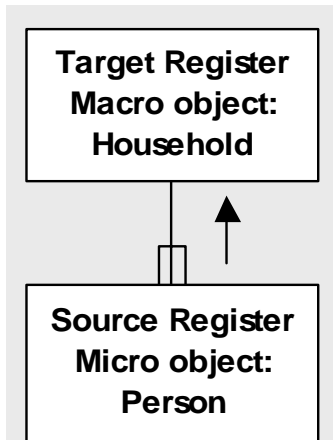
Derived variables

1. *Variables derived by grouping values and dividing into class intervals*
2. *Variables derived by arithmetic operations using variables in the data matrix*
3. *Variables derived by adjoining*



This involves creating a derived variable in a register using variables from another register. The objects in the first register can be linked to objects in the second register in a *one-to-one* relationship or a *one-to-many* relationship. This means that every object in the source register can be linked to one or many objects in the target register. Using this relationship, variables in the source register can be adjoined to the objects in the target register.

4. *Variables derived by aggregation*



This involves creating a derived variable in a register using variables from another register. The objects in the source register can be linked to the objects in the target register using a *many-to-one* relationship. One or many objects in the source register can be linked to one object in the target register. It is possible to aggregate values, in a way that is relevant for the survey, for the *micro objects* in the source register that is linked to the respective *macro object* in the target register.

Chart 3.14A The relations between persons, activities and local units

Register 1 – Persons

Person	Sex
PIN1	M
PIN2	F
PIN3	M

Register 2 – Job activities

Job	Person	Local unit	Wage sum
J1	PIN1	LU1	220 000
J2	PIN3	LU1	180 000
J3	PIN1	LU2	230 000
J4	PIN2	LU2	210 000
J5	PIN3	LU2	90 000

Register 3 – Local units

Local unit	Industry
LU1	A
LU2	D

Chart 3.14B Wage sums for persons and local units created by aggregation

Register 1 – Persons

Person	Sex	Wage sum
PIN1	M	450 000
PIN2	F	210 000
PIN3	M	270 000

Register 2 – Job activities

Job	Person	Local unit	Wage sum
J1	PIN1	LU1	220 000
J2	PIN3	LU1	180 000
J3	PIN1	LU2	230 000
J4	PIN2	LU2	210 000
J5	PIN3	LU2	90 000

Aggregation

Aggregation

Register 3 – Local units

Local unit	Industry	Wage sum
LU1	A	400 000
LU2	D	530 000

Chart 3.14C Industry and sex as derived variables for jobs created by adjoining

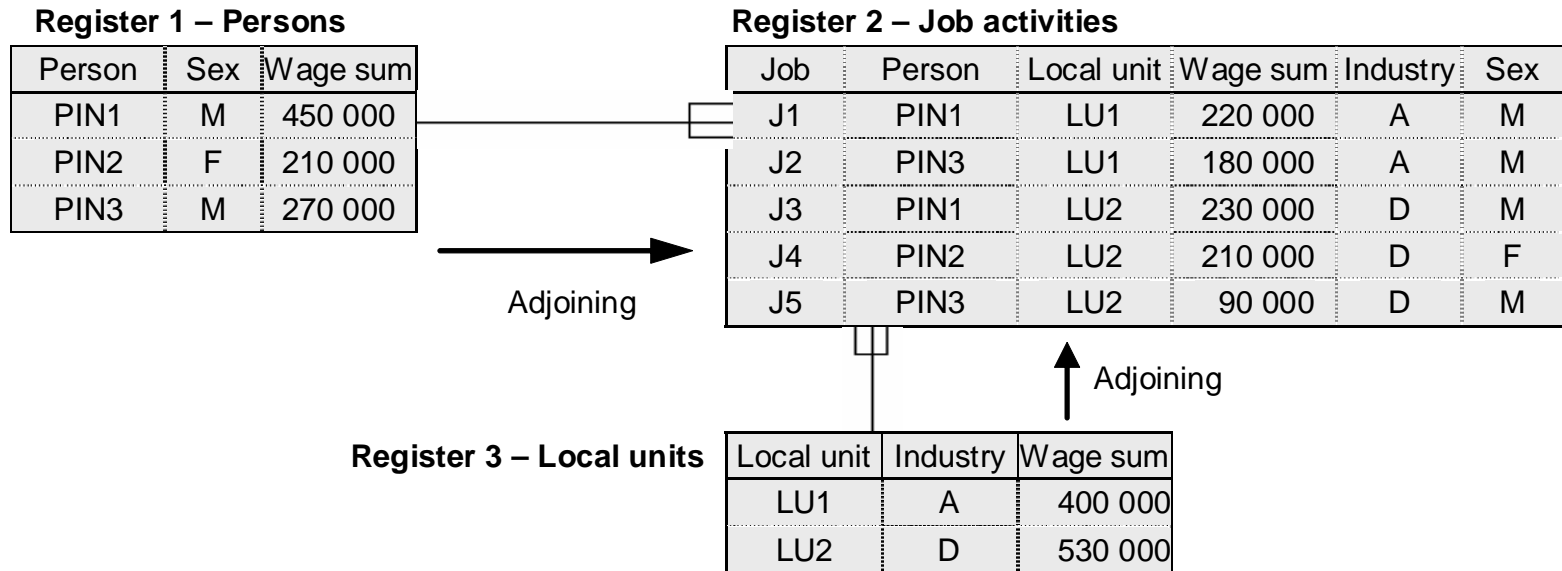
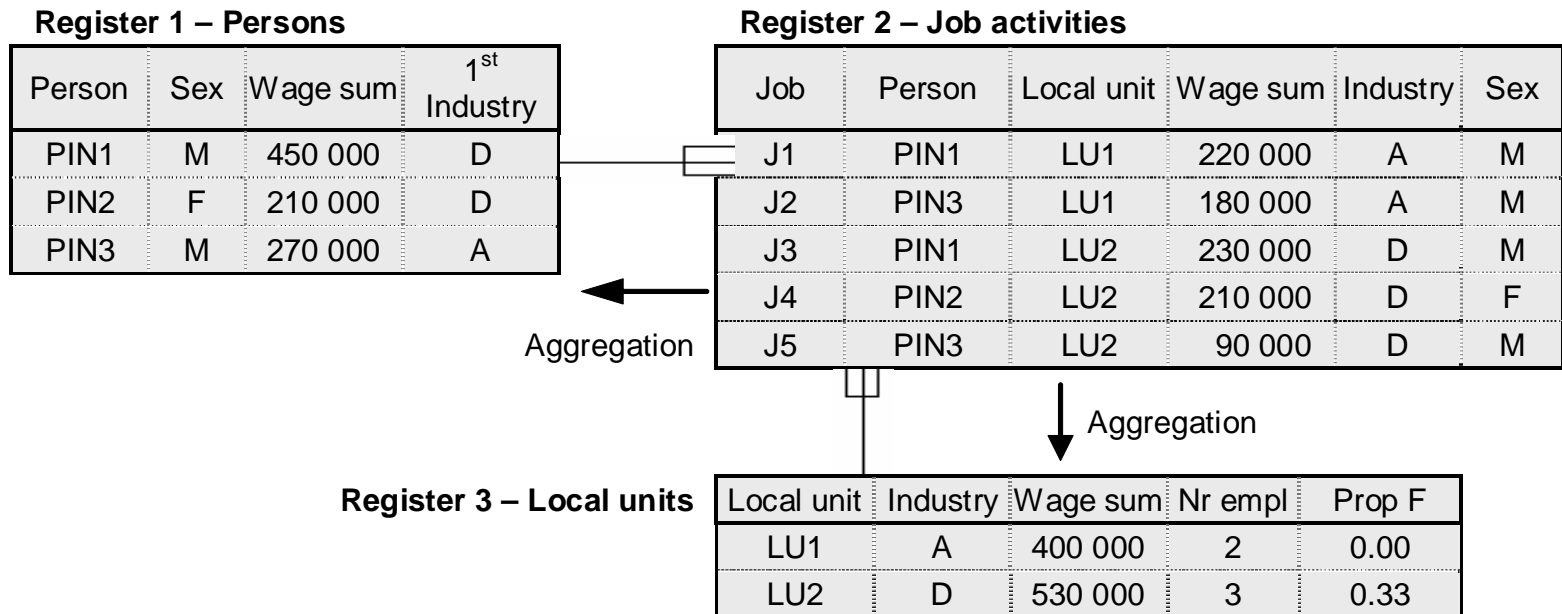


Chart 3.14D

Industry, number of employees and proportion of females as derived variables – by aggregation



Variables with different origins

- **Local primary variable**: coming directly from an administrative register
- **Locally derived variable**: derived from other variables in the same register
- **Imported primary variable**: coming from another statistical register with identical objects
- **Imported derived variable**: derived variable from another statistical register with identical objects

Variables with different functions

- identifying variable (primary key)
- reference variable (foreign key, link)
- communication/location variable
- time variable: reference time, reporting time, ...
- statistical variable: classification/spanning variable, summation/response variable, ...
- metadata/technical variable: source of value, computed/corrected value, weight, comment, ...

Variables used for matching/linking

- Linking identical objects in different registers or database tables, e.g. linking different variables about the same person in different registers, using the person number
- Linking objects in the same register, or in different registers, which are related to each other in a certain way, e.g. linking a person to his/her spouse (via person number), or his/her dwelling (via dwelling identifier)

Chart 3.17 A register's primary role in the system

Type of register	Types of variables	Role and responsibility
Base register	<i>Local primary variables:</i> Identifying variables Communication variables Reference variables Time references	Receive administrative data Create object sets Define objects Create some basic spanning variables Produce demographic statistics
Primary register	<i>Local primary variables:</i> Identifying variables Statistical variables	Receive administrative data Create the actual statistical variables
Integrated register	<i>Imported variables:</i> Identifying variables Statistical variables Locally derived variables, adjoined and aggregated variables	Create new information without data collection Compile information from different fields Compile information from different time periods

Chapter 4

Page 63

Note: What is not included in a register system may be included in a register-based statistical system!

Chart 4.2 Similarities and differences between the different types of survey

Sample survey	Census	Register-based survey
Not included in register system	Included in register system – can be used for other register-based surveys	
Uses the register system to define populations and as a source for variables		
Sample design, estimation, measures of uncertainty	System-based thinking and coordination with other register-based surveys are important	
Own data collection – produce own questionnaires	Uses others' administrative registers	
Editing – can contact respondents	Editing – can contact register-providing authority	
Nonresponse – reminders, when to stop data collection?	Mismatch related to missing values or undercoverage	
Quality flaws – sampling errors, measurement errors	Quality flaws – measurement errors	Quality flaws – relevance errors, lack of comparability
Small tables – cannot give estimates for small groups	Presentation – large tables with many cells	

The objects and population(s) of a register

- Object type and object instance
- Object set, set of object instances:
 - population: "all" objects of a type at a time
 - domain, subpopulation: a subset of a population, often defined by a crossclassification of variables, classification variables
- Population of interest: the "ideal" population, sometimes rather vaguely defined
- Target population: the population decided in the design process to be aimed at in practice, precisely defined, in an operational way
- Register population: the set of objects actually obtained/registered

Example: Population register of Sweden

- Population of interest: People permanently living in Sweden on December 31



Relevance error

- Target population: People registered by the appropriate agency as living in Sweden on December 31



Coverage error

- Register population: People actually registered as living in Sweden on December 31, according to information obtained up to January 31 next year

Creating a statistical register

- **Determine the objectives:** Which statistical needs are to be fulfilled by the register? Examples:
 - cross-sectional status data
 - events and changes
 - time series data (macro-level)
 - longitudinal studies
- **Define the desirable contents** of the register in terms of objects and variables
- **The inventory phase:** Which sources are available, administrative and statistical?
- **Editing and integration of the sources:**
 - match, check, edit, redefine, and reconcile objects, and synchronise them as regards times of reference
 - combine, check, edit, redefine, and synchronise variables

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 registers
 - 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 authorized 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 authorized 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
ldnr 1	DE
ldnr 2	DB
ldnr 3	DA
ldnr 4	DC
-	-
-	-

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

Enterprise id	Industry
-	-
ldnr 2	DB
ldnr 3	DB
ldnr 4	DC
ldnr 5	DG
-	-

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

Enterprise id	Industry
-	-
-	-
ldnr 3	DB
ldnr 4	DC
ldnr 5	DG
ldnr 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.