

ESSnet on Data Integration

Mauro Scanu

Istat

Central Unit on User Needs, Integration and Territorial Statistics

scanu@istat.it

ESSnet DI: contents and objectives

Partners: Italy, Netherlands, Spain, Poland, Switzerland, Norway

Duration: 24 months

- WP1. Building and maintaining common knowledge in record linkage, statistical matching and micro integration processing by reviewing the state-of-the-art in the literature and NSIs' practices, and setting up of a repository of technical documents.
 - WP2. Developing methods in some specific domains.
 - WP3. Developing tools extending the existing library of applications for data integration supported by appropriate documentation.
 - WP4. Fostering knowledge transfer by the development of a case study and associated recommendations on representative problems in data integration in the ESS.
 - WP5. Building sustainable capacity through targeted training and active communication with non-participating NSIs.
 - WP6 Management.
-

Data integration methods

- Let A and B be two sets (sources) of data.
- Aim: “integration” of the two sources
- What does “integration” mean?

Let us consider a first distinction: micro and macro

Micro: the aim is to find the records in the two sources that refer to the same unit

Macro: the aim is the construction of “parameters” (e.g. a contingency table) for two (or more) variables never jointly observed, but observed distinctly in the two data sources

Data sources

A: consists of n_A records

B: consists of n_B records

Some of the variables (X) are observed in both A and B (common variables)

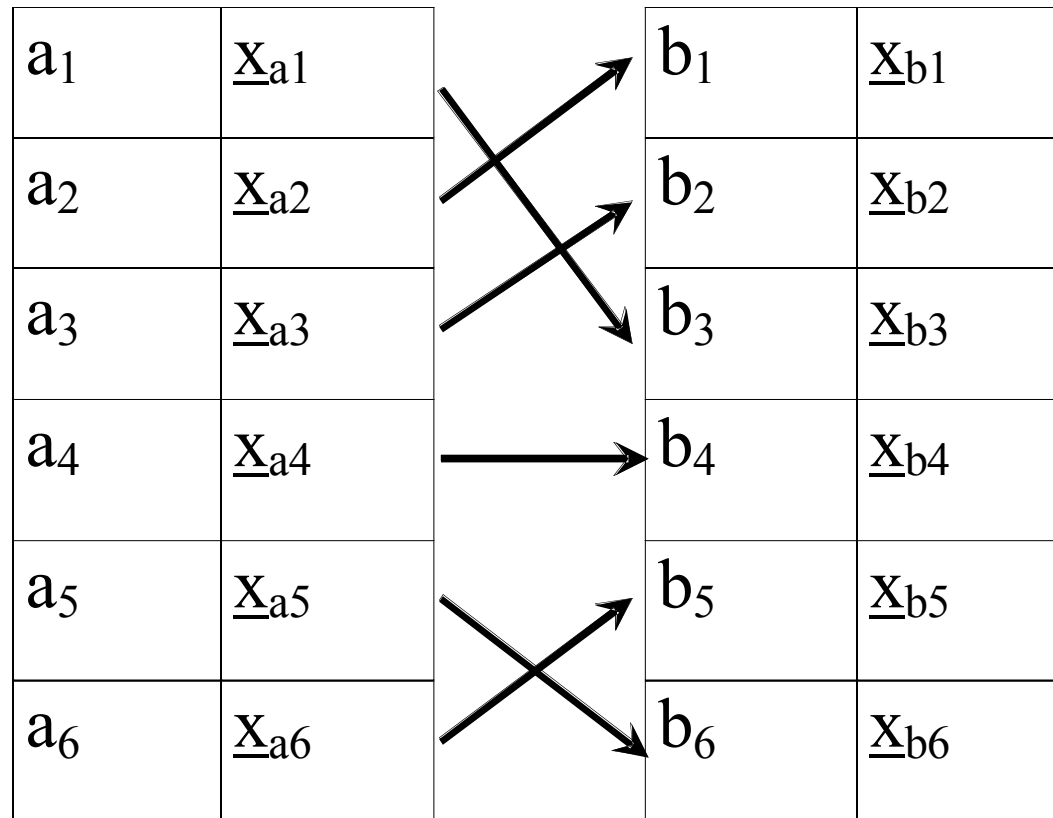
A	
<u>X</u>	<u>Y</u>
<u>X</u> _{a1}	<u>Y</u> _{a1}
<u>X</u> _{a2}	<u>Y</u> _{a2}
...	...
<u>X</u> _{an}	<u>Y</u> _{an}

B	
<u>X</u>	<u>Z</u>
<u>X</u> _{b1}	<u>Z</u> _{b1}
<u>X</u> _{b2}	<u>Z</u> _{b2}
...	...
<u>X</u> _{bv}	<u>Z</u> _{bv}

Micro integration

Record detection
is performed by
means of the
common
variables X

There exist
different kinds of
micro integration



Statistical methods for integration

It is useful to distinguish the methodologies according to the available input (i.e. the data sets to integrate)

Input	Method
Two data sets that observe (partially) overlapping groups of units	Record linkage
Two independent samples, without any unit in common	Statistical matching

Record linkage

Input: two data sets that observe overlapping groups of units

Problem:

Lack of a unique and correct identifier

S1: SHOE SHOPS (OHIO)			
ID	Name	Address	Telephone
S1.1	Rugged Boot The	4901 W Broad St Prairie Twp OH 43228	614-878-0569
S1.2	Rugged Boot The	4788 Columbus Park Lewis Center OH 43035	740-548-7463
S1.3	Springshod Footware	2300 E Kemper Rd Sharonville OH 45241	513-771-1175

MATCH

MATCH

S2: SHOE SHOPS (OHIO)			
ID	Name	Address	Telephone
S2.1	Springshod Footware	2300 E Kemper Rd Cincinnati OH 45241-6501	513-771-1175
S2.2	Springshod Footware	8969 Kingsridge Drive Dayton, OH	937-312-0506
S2.3	The Rugged Boot	4901 W Broad St Columbus, OH	800-605-2668

Alternative: presence of a set of variables that (jointly) allows the detection of records

Attention: variables can have “problems”!

Aim: highest number of correct linkages, lowest number of wrong linkages

When an identifier is missing

Probabilistic record linkage:

Methods based on statistical decision making. The first attempts (Fellegi-Sunter 1969) associated to the likelihood ratio test

Output: not only a set of linked records. It is possible to have also the probability that a pair of records is an actual match. This is extremely useful for statistical analysis on linked data sets

Statistical matching

What happens if the variables we want to analyze jointly are in two distinct **sample** surveys?

- Let A and B be two independent sample surveys of size n_A and n_B , drawn from the same population.
- Some variables X are observed in both the samples
- Variables Y are observed only in A
- Variables Z are observed only in B.

The aim of statistical matching is to draw information on (X;Y;Z), or at least on the variables that are not jointly observed (Y;Z)

Statistical matching

It is very unlikely that the two sample surveys observe the same units, hence record linkage cannot be applied.

Sample	Y_1	...	Y_Q	X_1	...	X_P	Z_1	...	Z_R
A	y_{11}^A	...	y_{1Q}^A	x_{11}^A	...	x_{1P}^A			
	y_{a1}^A	...	y_{aQ}^A	x_{a1}^A	...	x_{aP}^A			
	$y_{n_A 1}^A$...	$y_{n_A Q}^A$	$x_{n_A 1}^A$...	$x_{n_A P}^A$			
B				x_{11}^B	...	x_{1P}^B	z_{11}^B	...	z_{1R}^B
				x_{b1}^B	...	x_{bP}^B	z_{b1}^B	...	z_{bR}^B
				$x_{n_B 1}^B$...	$x_{n_B P}^B$	$z_{n_B 1}^B$...	$z_{n_B R}^B$

Statistical matching

Frequently this problem is solved mimicking a record linkage problem: instead of matching records referring to the same unit, I try to link records referring to “similar” units, where similarity is measured on the common variables X .

For this reason, very often this problem is tackled by means of imputation procedures, as the hot-deck ones.

ATTENTION: statistical matching and record linkage are completely different problems.

Once the linked data set is obtained

Micro integration processing

Is the integrated data set useful for statistical purposes?

The problem is that there can be incoherencies that should be solved.

The project will investigate what editing and imputation procedures can turn out useful and produce a good integrated data set

Possible areas of interest

Integration is not useful in itself. It is a step that is needed mainly for statistical analyses, but also for:

Statistical disclosure control: eg Skinner and Shlomo for RL and Fienberg et al for SM

Imputation and editing

Small area (e.g. data pooling?)

Ecological inference (King)

Data pooling (ABS)

Data pooling: constructing estimates across collections

Geoff Lee, Jonathon Khoo, Anil Kumar, James Chipperfield, Julia Chessman and Russell Lim.
Australian Bureau of Statistics

Abstract: Users of official statistics are becoming more sophisticated, requiring estimates for small sub-populations. This causes a problem for sample surveys that have broad design parameters which do not support the level of detail required. The Australian Bureau of Statistics has begun to investigate the feasibility of creating estimates using data obtained from different statistical collections. The purpose of this work is to develop a framework to evaluate whether gains can be obtained by pooling data from different collections for particular applications. We also will investigate the key technical and practical issues with estimates constructed this way. This paper reports on the proposed approach to data pooling and describes our initial findings from pooled labour force estimates for Australia's Indigenous people, a relatively small subpopulation.

Connections between SAE and DI?

In the next two years, how can the two ESSnets interact?

- 1) Definitions of problems: can RL or SM be useful for small area estimations?
- 2) Sharing solutions/methods: can the tools developed for SM and RL be already useful for small area estimations? Viceversa?
- 3) Tackling new common problems together: are there unexplored areas where we can work together?

Possible outcomes

- joint documents for the two ESSnets
 - Presentations on the ESSnet-DI workshop (Madrid, November 2011)
-