Why record linkage?

According to Fellegi (1997)*, the development of tools for data integration is due to the intersection of these facts:

- occasion: construction of big data bases
- tool: computer
- need: new informative needs


Book of life

Dunn (1946)** describes record linkage in this way:

...each person in the world creates a book of life. The book starts with the birth and ends with the death. Its pages are made up of all the principal events of life. Record linkage is the name given to the process of assembling the pages of this book into one volume. The person retains the same identity throughout the book. Except for advancing age, he is the same person...

Applications

There are many record linkage applications. The objectives are the most diverse.
Let us analyze some of them

- Applications described on scientific articles
- Applications detected during a former ESSnet
- Other applications?

Objectives

Some objectives for record linkage

- To have joint information on two or more variables observed in distinct data sources
- To "enumerate" a population
- To substitute (parts of) surveys with archives
- To create a "list" of a population
- Other official statistics objectives (to study the risk of identification of the released micro data)

Example 1 – analysis of mortality

Problem: to analyze jointly the "risk factors" with the event "death".

- The risk factors are observed on ad hoc surveys (e.g. those on nutrition habits, work conditions, etc.)
- The event "death" (after some months the survey is conducted) can be taken from administrative archives

These two sources (survey on the risk factors and death archive) should be "fused" so that each unit observed in the risk factor survey can be associated with a new dichotomous variable (equal to 1 if the person is dead and zero otherwise).
Example 1 – other analyses

Other important analyses on linked data sets performed in recent times:

**Linked longitudinal employer-employee data set** (Statistics New Zealand) based on linking administrative data held in the NZ Inland Revenue Department’s tax system and Statistics New Zealand’s list of NZ businesses (analysis of job and worker flows, employment tenure, multiple jobholding and business demography)

Example 1 – other analyses

Statistical longitudinal census dataset (ABS) An important feature of the Census Data Enhancement (CDE) project is the formation of a Statistical Longitudinal Census Dataset (SLCD) by bringing together data from the 2006 Census with data from the 2011 Census and future Censuses to build a picture of how society moves through various changes: which groups are affected by different types of change and in what way. The non-identifying grouped numeric code will be used in conjunction with characteristics such as age, sex, geographic region and country of birth to link records from the 5% SLCD to the 2016 Census and future Censuses using probabilistic record linkage techniques. Name and address information will not be used in the linkage process and will not be available for the 5% SLCD dataset as they are deleted at the end of Census processing.

Example 2 – to enumerate a population

Problem: what is the number of residents in Italy?

Often the number of residents is found in two steps, by means of a procedure known as "capture-recapture". This method is usually applied to determine the size of animal populations.

- Population census
- Post enumeration survey (some months after the census) to evaluate Census quality and give an accurate estimate of the population size

Italy - in 2001 "Indagine di Copertura del Censimento"
Example 2 – to enumerate a population

The result of the comparison between Census and post enumeration survey is a 2x2 table:

<table>
<thead>
<tr>
<th></th>
<th>Observed post enumer. survey</th>
<th>Non observed post enumer. survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed in Census</td>
<td>( n_{oo} )</td>
<td>( n_{on} )</td>
</tr>
<tr>
<td>Non observed in Census</td>
<td>( n_{no} )</td>
<td>??</td>
</tr>
</tbody>
</table>

Example 2 - to enumerate a population

For short, for any distinct unit it is necessary to understand if this unit was observed

1) both in the census and in the PES
2) only in the census
3) only in the PES

These three values allow to estimate (with an appropriate model) the fourth value.

Example 3 – surveys and archives

Problem: is it possible to use jointly administrative archives and sample surveys?

At the micro level this means: to modify the questionnaire of a survey dropping those questions that are already available on some administrative archives (reduction of the response burden)

E.g., for enterprises:
Social security archives, chambers of commerce, ...
**Example 4 – Creation of a list**

**Problem:** what is the set of the active enterprises in Italy?

In Istat, ASIA (Archivio Statistico delle Imprese Attive) is the most important example of a creation of a list of units (the active enterprises in a time instant) “fusing” different archives.

It is necessary to pay attention to:

- Enterprises which are present in more than one archives (deduplication)
- Non active enterprises
- New born enterprises
- Transformations (that can lead to a new enterprise or to a continuation of the previous one)

---

**Example 5 - Privacy**

**Problem:** does it exist a "measure" of the degree of identification of the released microdata?

In order to evaluate if a method for the protection of data disclosure is good, it is possible to compare two datasets (the true and the protected ones) and detect how many modified records are “easily” linked to the true ones.

---

**ESSnet “Integration of surveys and administrative data”**

In a former ESSnet, different data integration projects are described:

- EU-SILC
- Business demography data
- SBS
- Business Register
- Population and Housing Census
- Population statistics
- LFS and TUS

A survey on the use of data integration procedures has been set up (eligible respondents: NSIs in the ESS). We have details on 37 data integration projects (as an average, 9 data sources per project are integrated; one of them uses 40 data sources!)

Some results from the survey

<table>
<thead>
<tr>
<th>Main area of interest</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>To carry out a business survey</td>
<td>8</td>
</tr>
<tr>
<td>To carry out a social survey</td>
<td>6</td>
</tr>
<tr>
<td>To produce a population or housing census</td>
<td>6</td>
</tr>
<tr>
<td>To produce a business census</td>
<td>1</td>
</tr>
<tr>
<td>To produce a population archive/register</td>
<td>7</td>
</tr>
<tr>
<td>Other</td>
<td>10</td>
</tr>
</tbody>
</table>

Objectives

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional census activities (for instance post-enumeration surveys)</td>
<td>9</td>
</tr>
<tr>
<td>Reduction of costs and response burden</td>
<td>23</td>
</tr>
<tr>
<td>Enhancement of editing and imputation processes</td>
<td>11</td>
</tr>
<tr>
<td>Analysis of statistical relations</td>
<td>16</td>
</tr>
<tr>
<td>Microsimulation policies</td>
<td>4</td>
</tr>
<tr>
<td>Register/archive maintenance</td>
<td>14</td>
</tr>
<tr>
<td>Improvement of estimation methods (weighting, imputation, small area estimators)</td>
<td>18</td>
</tr>
<tr>
<td>Set up of a sampling frame (for instance improvement of coverage)</td>
<td>17</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
</tr>
</tbody>
</table>

What kind of data sources have been used?

<table>
<thead>
<tr>
<th>Main area of interest</th>
<th>Archive / register</th>
<th>Sample</th>
<th>Census</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>To carry out a business / social survey</td>
<td>58.8%</td>
<td>28.0%</td>
<td>5.9%</td>
<td>7.3%</td>
</tr>
<tr>
<td>To produce a population / business census</td>
<td>79.7%</td>
<td>6.1%</td>
<td>12.2%</td>
<td>2.0%</td>
</tr>
<tr>
<td>To produce a population / business register</td>
<td>78.7%</td>
<td>17.9%</td>
<td>0.0%</td>
<td>4.4%</td>
</tr>
<tr>
<td>Other</td>
<td>49.4%</td>
<td>24.7%</td>
<td>2.5%</td>
<td>21.4%</td>
</tr>
<tr>
<td>Total</td>
<td>65.7%</td>
<td>28.4%</td>
<td>4.9%</td>
<td>11.0%</td>
</tr>
</tbody>
</table>
ESSnet “Integration of surveys and administrative data”

In what way laws on privacy have influenced the data integration process?

<table>
<thead>
<tr>
<th>Problems</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit identifiers have to be cancelled in some data sets</td>
<td>13</td>
</tr>
<tr>
<td>Some data sets can provide only aggregate data</td>
<td>7</td>
</tr>
<tr>
<td>One or more archives/samples useful for integration goals can not be used</td>
<td>5</td>
</tr>
<tr>
<td>Linking of some groups of administrative data is prohibited by law</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
</tr>
</tbody>
</table>

Other experiences?

In your opinion, is there any other record linkage objective that can be included in the previous list?
Real life examples are welcome!

What’s in common, what’s not

Common aspects of the previous applications

- The previous examples compare two (or more) data sets (sources, archives, surveys,...)
- The objective is always micro: to link records in the two files that are available in the two data sets
- Creation of a list: the procedure stops at the unit identification step
- Analysis, enumeration of a population: unit identification is only a step for the estimation of “aggregates”
- Imputation, privacy protection, analysis, enumeration of a population: if the variables used for linking the records (common variables X) are “unstable”, a fundamental role is played by the “probability of correct link
Record Linkage: definition

- T. Belin, D. Rubin (1995, JASA): The term “record linkage” refers to the use of algorithmic techniques for the record identification in different data bases that refer to the same unit.

- Synonyms: Exact Matching / Computer matching

Gu et al. While epidemiologists and statisticians speak of record linkage, the same process is called entity heterogeneity, entity identification, object isomerism, instance identification, merge/purge, entity reconciliation, list washing and data cleaning by computer scientists and others.

Record Linkage – simple case

I have two data sets A and B, (e.g. on individuals). Every record possesses a unique unit identifier (e.g. a PIN) which is not affected by errors. It is possible to neglect the other variables, and link the records with the same PIN.

<table>
<thead>
<tr>
<th>Person identity number</th>
<th>Dwelling identity number</th>
<th>Sex</th>
<th>Urbanization level</th>
<th>Race code</th>
<th>Person status of employment</th>
<th>Age group</th>
</tr>
</thead>
<tbody>
<tr>
<td>271183-12238</td>
<td>0100913</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>

Unemployment register

<table>
<thead>
<tr>
<th>Person identity number</th>
<th>Name</th>
<th>Surname</th>
<th>Sex</th>
<th>Nationality</th>
<th>Start unemployment date</th>
<th>End unemployment date</th>
</tr>
</thead>
<tbody>
<tr>
<td>271183-12238</td>
<td>Āgārs</td>
<td>Āgs</td>
<td>M</td>
<td>Ķīgars (-iete)</td>
<td>04-gen-10</td>
<td>&lt;null&gt;</td>
</tr>
</tbody>
</table>
Record Linkage - intermediate case

A unique identifier does not exist, or cannot be used.
The other variables are jointly able to identify the unit (together they play the role of an identifier). Furthermore they are not affected by errors or missing items.

Record Linkage - difficult case

A unique identifier does not exist, or cannot be used.
The other variables are jointly able to identify the unit (together they play the role of an identifier). Anyway there can be differences in the answers for:

- Errors
- Missing answers
- Correct answers with a different codification/structure
- Changes due to time.
Record Linkage – difficult case

<table>
<thead>
<tr>
<th>Name</th>
<th>Surname</th>
<th>ZIP code</th>
<th>Date of birth</th>
<th>Place of birth</th>
<th>Marital status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>Smith</td>
<td>SW20 ORQ</td>
<td>18/05/70</td>
<td>London</td>
<td>single</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Surname</th>
<th>ZIP code</th>
<th>Date of birth</th>
<th>Place of birth</th>
<th>Marital status</th>
</tr>
</thead>
<tbody>
<tr>
<td>M.</td>
<td>Smith</td>
<td>WS20 ORQ</td>
<td>18 May 1970</td>
<td>Wimbledon</td>
<td>married</td>
</tr>
</tbody>
</table>

Record Linkage - characteristics

If a unique identifier exists, or if a set of variables not effected by errors can play the role of an identifier, the integration problem is straightforward.

We consider the record linkage problem in the case there is not a unique identifier and the other variables can play the role of an identifier but are reported with error.

Record Linkage - characteristics

As a matter of fact, record linkage methods are due to:
1) Bad data "maintenance" (lack of a correct identifier – positive examples: Finland, Netherlands)
2) "pathology" in data (errors and modifications)

In this case, the problem is that
- Some true matches are not detected
- Some false matches are interpreted as true matches

Objective of a record linkage procedure: to minimize these linkage errors
Other experiences

In your record linkage examples, do you have an identifier?
Is the identifier “good” in both the data sets?
If a unique identifier does not exist, what is the quality of the common variables in the data sets to link?

Bibliography

Data integration manual – Statistics New Zealand

Bibliography