Geocoding process of the 9th Industry and services census data: sources and methods used

Antonella Balistreri, Simonetta Cozzi
Istat

Meeting of the Group of Experts on Business Registers
21-23 September 2015 - Brussels

Outline

- The 9th Industry and services census: general overview
- Sources and methodologies used for geocoding of census data
- Assignment of the enumeration areas of local units
- Main results and conclusions:
  - Productive areas
  - Quality
  - Overall

The 9th Industry and Services Census

Conducted in 2012, based on an extensive use of administrative sources and composed by three different surveys:

- ✓ the sample survey on Enterprises;
- ✓ survey of Non-profit institutions;
- ✓ survey of Public Institutions.

All surveys collect data for institutional units and local units, at municipality level.

Enterprises: Structural data obtained from the Statistical Business Register (Asia), in order to obtain statistical information and ensure the compatibility of the previous census of 2001.

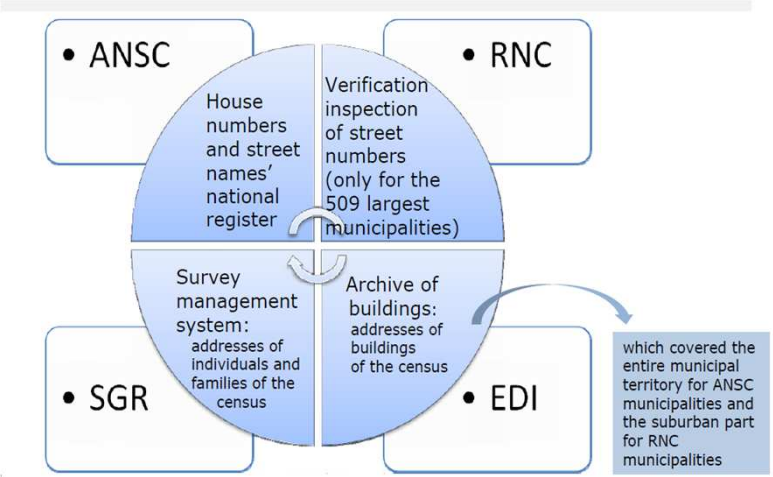
Public and Non-profit institutions surveys based on a pre-census list created by the National Institute of Statistics to integrate administrative archives and statistical sources.

Sources and methods used for geocoding of census data

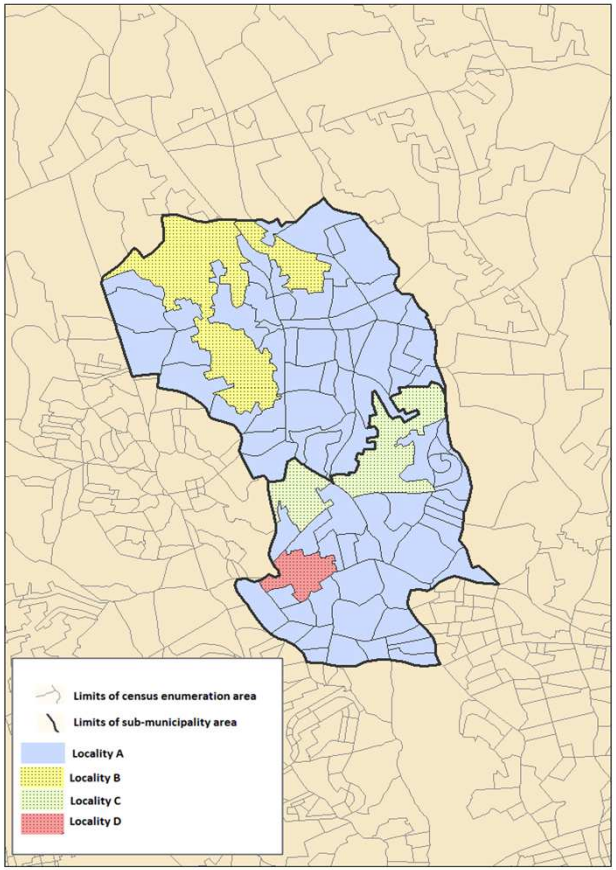
Several sources have been used to associate the census enumeration areas and census data.

- The main system adopted is "**Egon - Data quality**": a Web application that can handle a system of spatial data and allows to normalize and geocode data.
- **GIS** (Geographic Information System) is an information system that allows acquisition, recording, analysis, visualization of geographic data.

- Another source used to geocode is the **National Archive of urban street number** (ANNCSU), resulting from the integration of few archives. It's created and updated by ISTAT and Revenue Agency.



Census enumeration areas



The census enumeration areas (EA) cover the whole Italian territory, including special areas consisting of special geomorphological entities such as lakes, uninhabited mountains…

The whole country has been divided into census enumeration areas that are the minimum unit of detection of the municipality. The different aggregations of EA identify several partitions of the Italian territory:

- census enumeration areas;
- sub-municipality area (municipalities, districts, etc.);
- localities;
- municipalities.

Assignment of the census enumeration areas

Type of units	Number of local units
Non-profit institutions	347.602
Public institutions	109.358
Enterprises	4.806.014
Total	**5.262.974**

- Using of Egon-data quality
- Using of ANNCSU archive
- Geospatial coordinates trough GIS
- Assignment of residual enumeration areas

1 step: use of Egon-data quality

The first step has been the addresses' normalization of the local units of enterprises, public and non-profit institutions trough "Egon-data quality" software.

It provides to:

- normalize addresses;
- get enumeration area for 2001;
- get geographical coordinates.

To obtain the 2011 census enumeration areas, a transcoding table has been used.

Failed results of this first step derive from:

a) non-normalized addresses
b) geocoded addresses with low quality code,
c) EA not unique (multiple enumeration areas in 2011 for the same enumeration area in 2001).

2 step: use of ANNCSU: equivalence

The archive contains the 2011 census enumeration area for each address. The link was based on criteria of equivalence and similarity of the addresses, with or without civic number.

Through a procedure in **Oracle SQL**, string of census addresses, including their attribute (dug, denomination and civic number), have been compared with the addresses included in the archives used by ANNCSU. In order to increase efficiency, the links has been made only within the "municipalities" blocks.

COD_MUNIC	DUG	ADDRESS	NUM	ESP	EA
020017	VIA	DOTTORINA	151	BIS	83
020017	VIA	DOTTORINA	151	BIS A	37
020017	VIA	DOTTORINA	151	E	37
020017	VIA	DOTTORINA	151	P	83
020017	VIA	DOTTORINA	151	Q	83
020017	VIA	DOTTORINA	151	T	37
020017	VIA	DOTTORINA	151		37

For the same address, we can have different census EA. An EA can be attributed only if it is the unique possible.

For the remaining addresses, the search has occurred by subsequent approximations, removing some components of the specific string, such as street number or dug.

2 step: use of ANNCSU: similarity

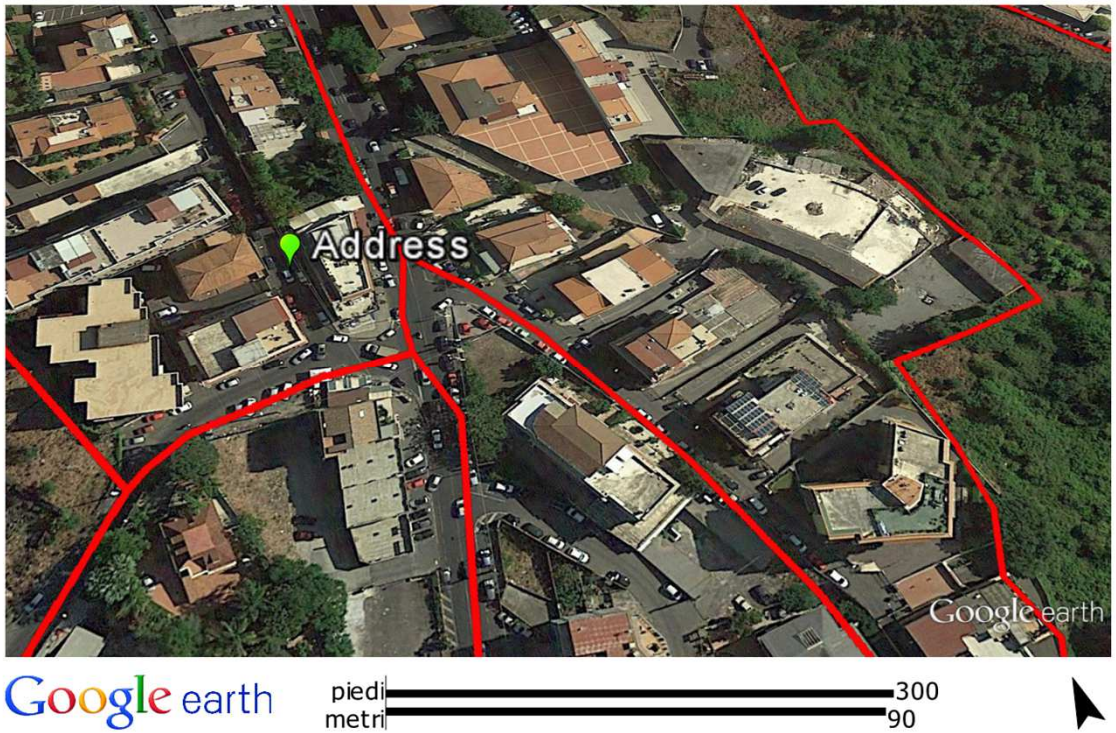
RELAIS (REcord Linkage At IStat) is a toolkit providing a set of techniques for dealing with record linkage projects.

Among comparison functions measuring the “similarity” between two fields, the similarity function **3-grams** was used.

This derives from Q-grams functions, generally used in approximate string matching by “sliding” a window of length q over the characters of a string s to create a number of 'q' length grams for matching. A match is then rated as number of q-gram matches within the second string, t, over possible q-grams. When two strings s and t have a small edit distance, they also have a large number of q-grams in common.

A valid match was considered using similarity function 3-grams with a selected threshold of 0.8

3 step: GIS



The geographical coordinates for the addresses, normalized but not assigned, has been projected on the GIS system, obtaining 2011 EA for overlapping addresses and geographical limits. Subsequent checks have shown that this kind of allocation is correct.

4 step: residual enumeration areas

For the residual local units without EA, the enumeration areas have been identified by:

- using the previous Census 2001: the enumeration area in 2001 assigned to the local unit, with the same fiscal code in the same municipality. If the typology of locality was compatible, it has been decided to assign the enumeration area of the previous census.
- using a re-proportioning procedure based on the structure and distribution of the enumeration area already assigned. In particular, in order to minimize mistakes in scarcely populated area of productive units, the remaining units have been assigned to areas with the greatest frequency of enterprises or institutions of the same type.

Reassignment for Productive areas

As 466 municipalities with no local units in their productive areas, some local units have been assigned to "productive areas" through the following criteria:

- similarity in ANNCSU archive
- linkage between the local units in 2011 and 2001 census, having EA in a productive area
- search of some keywords (“industrial”, “craft”, “commercial”, etc.) in the address of the 2011 census

At the end of this process, some local units in productive areas have been assigned to more than the half of the 466 municipalities.

	Enumeration areas “productive area”		Municipalities		Local units
	n.	%	n.	%	n.
EA without local units assigned	674	19.8	213	11.2	-
EA with local units assigned	2,728	80.2	1,689	88.8	63,995
Totale	3,402	100.0	1,902	100.0	63,995

Istat

Results and conclusions (1)

Typology of locality	Local units	
	n.	%
Urban center	4,776,026	90.7
Urban nucleus	84,877	1.6
Productive area	63,995	1.2
Extra-urban area	338,076	6.4
Total	**5,262,974**	**100.0**

In the Urban center there are proportionally more local units of public and non-profit institutions units than of enterprise ones.

More than 90% of enumeration areas are of high quality (Egon and Anncsu), only 5% is affected by a not robust assignment, because of several factors, such as the incompleteness or lack of addresses.

Typology of local unit	EGON	ANNCSU	CENS_2001	LOW_QUALITY	Total
	%	%	%	%	%
Enterprises	86.2	6.0	3.1	4.7	100.0
Public institutions	85.1	4.1	2.1	8.8	100.0
Non-profit institutions	88.6	3.5	2.6	5.3	100.0
Total	86.3	5.8	3.1	4.9	100.0

Results and conclusions (2)

Typology of local unit	Typology of locality	TOTAL OF LOCAL UNITS		LOW_QUALITY OF LOCAL UNITS	
		n.	%	n.	%
Enterprises	Urban center	4,351,463	90.5	207,510	91.0
	Urban nucleus	78,285	1.6	2,457	1.1
	Productive area	62,640	1.3	5,665	2.5
	Extra-urban area	313,626	6.5	12,374	5.4
	Total of Enterprises	**4,806,014**	**100.0**	**228,006**	**100.0**
Public institutions	Urban center	102,548	93.8	9,283	96.7
	Urban nucleus	1,267	1.2	87	0.9
	Productive area	268	0.3	11	0.1
	Extra-urban area	5,275	4.9	216	2.3
	Total of Public institutions	**109,358**	**100.0**	**9,597**	**100.0**
Non-profit institutions	Urban center	322,015	92.6	7,669	95.6
	Urban nucleus	5,325	1.5	179	1.0
	Productive area	1,087	0.3	105	0.6
	Extra-urban area	19,175	5.5	531	2.9
	Total of Non-profit institutions	**347,602**	**100.0**	**18,484**	**100.0**
Total		**5,262,974**		**256,087**	

With reference to the 5% of the EA allocated in LOW_QUALITY, in productive areas there is a greater presence of enterprises (2.5%) than of public or non profit institutions, mainly due to the specific activities carried out by this type of local units.

Thanks for your attention!