The use of continuity patterns in administrative data to define the usually resident population in Italy

Economic Commission for Europe
Conference of European Statisticians Group of Experts on Population and Housing Censuses
4-6 October 2017- Geneve

Angela Chieppa
Raffaele Ferrara
Gerardo Gallo
Valeria Tomeo

ISTAT - Italian National Institute of Statistics
chieppa@istat.it, gegallo@istat.it
Reference context: where are we going?

Source: Unece (2011)
The Exploration of administrative data for population counts…Istat is working on!

- A specific task-force of Istat is in charge of investigating how to handle administrative data for inferring usually resident population counts

- This presentation shows goals, approach in investigation and first population groups when using administrative data

- Key idea of the work is the use of a continuously improving process that combines knowledge discovery from databases with expert/domain guidance
GENERAL FRAMEWORK AND STARTING POINTS

• EU Regulations on *usually resident* population

• ISTAT Data Base where administrative data are integrated (SIM)

• Experts choice of qualified administrative sources and related attributes, relevant when managing the Population Register

• *Knowledge discovery from databases* (KDD) approach
“Usual residence” means the place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage.

The following persons alone shall be considered to be usual residents of a specific geographical area:

(a) those who have lived in their place of usual residence for a continuous **period of at least 12 months** before the reference time; or
(b) those who arrived in their place of usual residence during the 12 months before the reference time **with the intention of staying** there for at least one year.

Where the circumstances described in point (a) or (b) cannot be established, "usual residence" shall mean **the place of legal or registered residence.**
Currently Istat handles an Integrated System of Microdata (SIM) which is continuously updated with the aim of supporting the statistical production processes both for social and economic statistics.

- SIM makes it possible to build many different “thematic database” for statistical purposes.
- Each object or unit (person; economic unit; places; relations) has been assigned an unique ID number, which remains the same for different tables/time periods.
- Logical and physical relations have been designed, regarding time and spaces, between different sources.

We integrate the administrative sources with records of population registers

Populations Registers:
ANVIS (a virtual national population register built up in ISTAT for statistical purposes) at 1st of January of each year (2012, 2013, and 2014)

Administrative data sources:
Employment Register, Education Register, Income and Taxation Register, Occupation Register, Compulsory school/pupils Register, Register of University Students and Statement of Earnings Register, Permits to stay at 1st of January of each year (2012, 2013, and 2014)

• Records have been evaluated in a period of 24 months: i.e. from January 2012 to December 2013 according to EU definition
A DB has been built with qualified administrative data from 2012 to 2014

- Some sources include information about the geographical location at municipality level
- The different time references (starting and ending date, monthly or quarterly information) have been standardized to achieve a presence/absence monthly scheme
- Other attributes (i.e. position in job, type of employment contract) have been made comparable within different sources
TIME-WINDOW AND CONTINUITY OF RECORDS

• In each time-window a person could have one or more signals, in the same local area or in different ones

• How more time-window length is, more signals a person has (due to migration, changing in job, and so on)

• Once a time-window is defined, throughout the presence/absence monthly scheme available for the majority of signals, patterns of continuity could be investigated
Reference time period and records of presences from January 1\textsuperscript{st} 2012 to December 31\textsuperscript{st} 2013

<table>
<thead>
<tr>
<th>Time period from January 2012 to December 2013</th>
<th>Type of presence in labor and education registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

Strong records

Weak records

No usable
Proceeding scheme and counts of population groups (000) as of the year 2012

<table>
<thead>
<tr>
<th>Step I</th>
<th></th>
<th></th>
<th></th>
<th>Step II</th>
<th></th>
<th></th>
<th>Step III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population register</td>
<td>Labor-education register</td>
<td>Permits to stay archive</td>
<td>Group</td>
<td>Count</td>
<td>Retired-non pension-benefits registers</td>
<td>Group</td>
<td>Count</td>
</tr>
<tr>
<td>61,068</td>
<td>37,704</td>
<td>3,378</td>
<td>A</td>
<td>36,618</td>
<td>20,764</td>
<td>E</td>
<td>14,485</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>-</td>
<td>B</td>
<td>24,450</td>
<td>yes</td>
<td>F</td>
<td>9,965</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>-</td>
<td>C</td>
<td>1,086</td>
<td>no</td>
<td>-</td>
<td>G</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>-</td>
<td>D</td>
<td>351</td>
<td>yes</td>
<td>H</td>
<td>3,026</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>-</td>
<td></td>
<td>no</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table legend:
- Eligible as usual residents
- Possible usual residents

Possible under-coverage

We focus on this group by using the records on three consecutive years

Possible over-coverage
Longitudinal counts on people with records of presence in Italy on three consecutive years

<table>
<thead>
<tr>
<th>Records in the Labor-education registers</th>
<th>Records in the Population registers</th>
<th>Strong records</th>
<th>Weak records</th>
<th>Profile of groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td>165,4</td>
<td>76,0</td>
<td>People with records of presence in 2012 recorded in the population register by the next two years</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>33.8%</td>
<td>15.7%</td>
<td>People with records of presence in 2012 NO recorded in the population register by the next two years</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>182,5</td>
<td>89,7</td>
<td>People with records of presence only in 2012 and 2013 without any registration in population registers</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>37.3%</td>
<td>18.6%</td>
<td>Total amounts</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>142,0</td>
<td>317,6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>29.0%</td>
<td>65.7%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>489,90</td>
<td>483,30</td>
<td></td>
</tr>
</tbody>
</table>

Do the individuals who record in the population registers show specific characteristics and place of settlement in comparison to those who do not record?
Foreign people recorded in the population registers in the time period from 1.1.2012 to 1.1.2014

- With strong records (150,000 Individuals)
- With weak records (70,000 Individuals)

Values per 1,000 Foreigners:
- 131 to 667 (235)
- 71 to 131 (583)
- 45 to 71 (927)
- 29 to 45 (1433)
- 0 to 29 (4914)

Values per 1,000 foreigners:
- 160 to 1,000 (127)
- 80 to 160 (306)
- 40 to 80 (752)
- 20 to 40 (1293)
- 0 to 20 (5614)
Foreign people NO recorded in the population registers in the time period from 1.1.2012 to 1.1.2014

<table>
<thead>
<tr>
<th>with strong records</th>
<th>with weak records</th>
</tr>
</thead>
<tbody>
<tr>
<td>(153,000 Individuals)</td>
<td>(86,500 Individuals)</td>
</tr>
</tbody>
</table>

Values per 1,000 foreigners:
- 280 to 881 (111)
- 115 to 280 (268)
- 56 to 115 (606)
- 28 to 56 (1222)
- 0 to 28 (5885)

Values per 1,000 foreigners:
- 500 to 1,000 (72)
- 230 to 500 (130)
- 110 to 230 (291)
- 40 to 110 (723)
- 0 to 40 (6876)
Italian people recorded in the population registers in the time period from 1.1.2012 to 1.1.2014

with strong records (15,000 individuals)

with weak records (6,100 individuals)
Italian people NO recorded in the population registers in the time period from 1.1.2012 to 1.1.2014

- With strong records (29,000 Individuals)
- With weak records (3,200 Individuals)

Values per 1,000 Italians:
- 2.2 to 83.5 (59)
- 0.9 to 2.2 (256)
- 0.6 to 0.9 (347)
- 0.3 to 0.6 (948)
- 0 to 0.3 (6482)

Values per 1,000 Italians:
- 2.7 to 38.5 (16)
- 1.2 to 2.7 (68)
- 0.6 to 1.2 (148)
- 0.3 to 0.6 (256)
- 0 to 0.3 (7604)
Exploratory analysis to find significative clusters of individuals

- Multiple Correspondence Analysis (MCA) is used to explore associations between categories of available variables that can offer useful understanding to specific clusters of individuals.

- The variables considered in this step are gender, age, citizenship, type of employment, type of signals of presence in administrative archives and, finally, the possible recording in population registers.

Is there any specific cluster emerging in the under-covered population? When studying the possible recording in official register of the population, are there specific clusters emerging?
Two important discriminant factors emerge from the MCA, which account for 73.9% of explained variance.

The discrimination measures shows that:

- the first dimension is related to variables *Gender* and *Age*: demographic profiles.
- The variable which represents different longitudinal group *Profiles* has a very large discrimination value on the second: longitudinal profiles.
- Type of employment has a very large value on both dimensions, indicating discrimination in both the first and second dimensions; the same happens for Citizenship.
MCA: dimension 1

Demographic profiles

Conference of European Statisticians Group of Experts on Population and Housing Censuses
4-6 October 2017 - Geneve
Since the goal of this first analysis is to explore association between categories and not objects (=individuals). Categories of different variables are close if they belong to the same objects.
FINAL REMARKS AND NEXT STEPS

• All administrative sources can be used to improve quality of the base Population Register
  ➢ to complete the process of identification of critical subgroups
  ➢ longitudinal data of labour records provide useful information on specific groups when considering type employment/source of data and demographic characteristics of individuals (we need more depth analysis to improve continuity patterns of presence)

• Geographical analysis result to be significant when profiling subgroups

• A combined population census might use administrative records with the aim to support the survey ... improving quality and coverage of data collection (i.e. informative packages to support municipal population census in order to involve immigrant associations, unions and cultural mediators for census purposes)
.... Thank you ....