“Towards a national statistical register of persons based on administrative sources. Characteristics and potentials for the official statistics”
Roberta Vivio*, Valentina Talucci**, Grazia Petraccone**

1 Introduction

It is rather common to use administrative data in the official statistics thanks to their economic and technical advantages.

It is indeed more economic to use administrative archives than to gather data through traditional sample surveys or censuses; the outputs are more quickly issued and the information can be issued on very small areas. No biases such as the memory effect, the interviewer effect, the question formulation effect, the reticence to answer, etc. can be found.

Only few answers are missing in the archive while the correction process is simpler and the information on sensitive variables (e.g. income) more accurate.

On the other hand, the non-primary variables are often of a lower quality. In addition the “administrative” definition and the “statistical” definition of the units generally differ; but in case they do not differ, the administrative reference universe is different from the statistical one.

Nonetheless, it is now impossible not to use administrative data.

In Italy the statistical potentials of the General Government’s informative data have greatly increased in the past years: there are many excellent archives, the new hardware and software technologies and the processing networks have become much more safe and secure; the theme contents have widened; and the offer coverage has increased. The job of the National Statistical Institutes’ statisticians has become much easier thanks to the new norms that regulate the access and use of such archives.

Istat can count on its statistical research and development programme to use administrative data for economical and socio-demographic statistics.

This paper refers to the project prepared to “Identify and acquire sources on socio-demographic units” and to the relative construction of a first national register on persons based on administrative sources. It will show not only some of the factors behind the importance of the archive, but also some analysis points regarding the quality of the archive.

The quality problems linked with the products and production processes will be dealt with in future documents.
2 The Register project and its importance

The idea behind the project was to bridge the informative gap Istat’s demographic statistics have, constructing a micro-data archive on all persons in Italy, or a national statistical register of persons. Hopefully it will imply important consequences in the official statistics.

The first step was to study the possibilities of the administrative data as only way possible. Hence, it involved analysing the informative systems of the General Government\(^2\) and the archives fed by data travelling on them (especially those on persons managed by the Central Government). The statistical requisites of each system were evaluated and the laws regulating them studied before checking which one had already been used in the *Piano Statistico Nazionale* and how.

As some archives were selected, contacts with their holder institute were activated – in case it had not yet been done in virtue of already existing work protocols – to obtain sample data on which to carry out a quality tests on the sources. Finally a prototype of the “new” demographic register was planned and implemented with the archives identified and hypothesised other demographic registers to be built in the future.

Here following are some of the possible uses and advantages of the Register hypothesised in the test, whether using it alone (by raw data) or together with other archives (by matched data):

- Survey design and planning
- Direct extraction of samples of persons
- Estimation of the relations among variables and application of sample surveys
- Calculation of demographic indicators
- Improvement of the coverage
- Data gathering for auxiliary variables
- Control and calculation of the units missing in the traditional surveys
- Production costs’ and human resources’ savings and decrease in data publication schedule
- Adherence to and obligations of the European regulation on migrations\(^3\)
- Estimation of the longitudinal models or relation among data variables
- Constitution of a base for the “Statistical System of Persons”, integrating several specialised registers together with a base Register
- Etc.

While some of these were tested, this paper will briefly present just one example of calculation of the demographic indicators (see Table 6) of the population structure.

3 Characteristics of the primary source

At the basis of such project, a tax archive was to be used despite it had never been considered for demographic purposes, that is, the *Tax Authorities Database of*

\(^2\) Among the main ones: Sister, Sigmater (National cadastral informative system), Fisconline, Entratel, Siatel AGEA; INA SAIA.

\(^3\) Regulation of the European Parliament and of the Council on Community statistics on migration and international protection
physical persons. Because it appears, according to some studies, to meet by itself almost all the construction requisites of the register of persons, it was decided to use it as primary input. Moreover its intrinsic quality, its adherence to the population data and the possibility it offers to be read demographically all support such decision.

3.1 Content
The “Tax Authorities' Database of Physical Persons”, or also called AT, is one of the tax informative System’s archives part of the Tax Authority and realised by the Ministry of Finances. The Presidential Decree\(^4\) of 29 September 1973 on the Verifications of Income Taxes, which defines the structure and content of the tax register, states that: “…the AT is to gather and order nationally data and information drawn from the tax returns and reports presented to the tax administration offices and relative verifications, as well as data and information that could be important for tax purposes (art. 1).
“…the AT is to register all physical persons, legal persons, enterprises, associations and any other organisation of persons or goods having no legal status, and which the data and information gathered according to art. 1 refer to, or which requested a fiscal code number as per art. 3” (art. 2)
“…every tax payer must request a fiscal code number” (art.3.).

3.2 The feeding process
The AT not only encloses the tax returns filled in by physical persons (model 730, model 740/74, sole model, etc) but also manages tax reports with the taxpayers, interacts with municipal registry offices, integrates other administrative sources (VAT tax payers’ special model, etc). Nowadays, it is fully computerised, a process that was started more than twenty years ago: upon setting it up, the Tax Authority imported in it all data contained in the paper forms used back then. While some of these no longer even exist, all the other forms are now directly obtained on-line. Graph 1 shows how the way to obtain the variable Fiscal code Number has changed over time: on-line for residents in younger age classes, and up to 40% for residents aged 60 and over.

3.3 Time reference of the information
The frequency with which each input is entered in the archive differs: while some are entered periodically (e.g. every year as regards the income tax returns), others just once (e.g. attribution of a fiscal code number). In addition, due to the multi-channel modality of archive feeding, the archive changes either because new units are entered or the already present unit variables change. Hence, it evolves day after day since it is fed on a continuous basis. However, such feeding registers a certain delay between the moment in which an event takes place and the moment in which the unit or information is entered in the database. Consequently, some units result as missing in the database at time \(T_0\): events that take place before \(T_0\) but not yet entered (or drawn) in the database within \(T_0\). For example, the entering of births in the database can require some days; therefore, a child born within \(T_0\) could be not present in the archive at time \(T_0\).

\(^4\) Presidential decree nr 605 of 29 September 1973, Provisions relative to the Tax registry
same goes for deaths of persons aged over 100 that could be not registered within $T_0$ due to missing documentation.

Graph 1 – The fiscal code number source of per age class of the person

3.4 Units and variables

Physical persons represent the units of interest for the project, while the fiscal code number and personal data its variables:

- Fiscal code number
- Name, surname and gender
- Place and date of birth
- Municipality and fiscal address
- Municipality and address of legal residence (if different from the fiscal domicile)
- Source of fiscal code number, fiscal address and residence.

3.5 The statistical aggregates of population

For each municipality, it is possible to identify three populations in the Tax Authority’s archive and that for:

- The **total population**, which includes Italians and/or foreigners, “Residents” and “non-Residents but holder of a fiscal code number$^5$”.
- The **aligned population**, which includes “Residents whose personal data are aligned both in the tax archive and in the municipal archive”.
- The **non-aligned population**, that is, the rest of the total and the aligned populations.

Compared to the resident population (registry’s target), the contents of the aligned population make it more similar to the resident one surveyed by Istat through its traditional surveys, the difference being two types of unit:

---

$^5$ In Italy, one must by law indicate its fiscal code number in almost all administration, etc; therefore, this code is attributed to non-resident citizens who need it upon request.
– Resident persons not yet aligned,
– Persons not yet entered in the database.

The non-aligned population contributes too in the importance of the sources as it includes:
– “Persons not residing in the municipality, born abroad and who have a fiscal code number”;
  e.g. in 2007, as Romania entered the European Union and its citizens became part of the latter group, more than 150 thousand fiscal code numbers were issued to them between January and October (during 2006 the amount was 66 thousand).

4 Analysis of the source’s quality

Quality involves a multi-dimension concept where the importance of each vector changes according to the user: “...the totality of features and characteristics of a product or service that bear on its ability to satisfy a given need.....”

The six quality vectors are: Relevance, Accuracy, Comparability, Coherence, Timeliness and Punctuality, Accessibility and Clearness.

- Relevance, or capacity of the statistical information produced and/or diffused to satisfy the real needs of the users.
- Accuracy, or closeness between estimate produced for a parameter (characteristics of the population studied) and real value unknown. Here, it measures the closeness of the source to the population of interest. Comparability, or degree of comparability of the statistical information over time and space; a comparable data enables to compare between geographical areas, it contributes in stepping up the length of a historical series relative to available data, it influences the different concepts and/or classifications as regards the possibility to compare data from the same source over time or between various sources. Here, the source can be used for historical analyses, despite some limits (see classification of the place of birth)
- Coherence; or adequacy of the statistics to be combined in various ways and for various uses; in addition, it also refers to the possibility to use together statistics produced for different primary uses.
- Timeliness and Punctuality, while the first refers to the period between the moment to which the data refer and the moment in which they became available to the users, the second is the period between the publication date of statistical information and the date planned. To respect both means to be able to meet the ever-growing need to describe statistical phenomena as soon as they occur (actuality of the statistical data) especially because of the increasing weight of some types of data; to respect them means to be able to respect a data publication calendar for the users. In our case, it is still in its quantification phase
- Accessibility and Clearness; while the first one is the easiness with which users can access statistical information produced, the second is the availability of the information linked with data and assistance provided by the institution.

The next step in analysing the information available in the AT was to analyse the definitions and classifications of each variable, and the statistical aggregates contained.

Accuracy was the vector to which most attention was given.

Based on calculation of evaluation indicators, archive coverage errors were evaluated, measuring the closeness between resident population (target population) and benchmark sources’ population. Finally, demographic indicators were calculated to be used as endogenous controls of the source.
4.1 Quality of each variable examined

The on-line acquisition process that characterises the tax informative system ensures by itself that each variable examined is accurate:

- The fiscal code numbers are codes that the Ministry possesses and issues to all citizens, ensuring their uniqueness on the national territory. Hence this variable meets the highest level of quality.
- The fiscal domicile address is drawn from fiscal administrative acts (income tax returns) obtained on-line.
- The sources for the fiscal code numbers, for the fiscal domiciles and for the legal residences all relate to the acts from which they derive.
- The personal data variables (sex, name, surname, year and place of birth) are checked by means of a reciprocal control process between the Ministry of Finances and the Resident Population Registries that own such data. The municipal registry offices indeed communicate to the Ministry of Finances data on the population that resides in their municipality in order to obtain their fiscal code number. Subsequently, the data drawn from both archives are first compared and then aligned one to the other. Though, in case of previous residence in another municipality, the latter could have validated the data, the adherence level still reaches 100%. All these controls allow selecting a subset of residents in an Italian municipality and present in the tax registry. Therefore the quality of this variable is indirectly certified by the owner institute.
- The legal residence is drawn (only if different from the fiscal domicile) during the personal data alignment process.

The missing answers’ and out of range errors are rather negligible.

Coherence errors: the process for editing errors is currently in progress.

Territorial comparability lacks in the case of variable Place of Birth: it isn’t submit to updating processes when administrative changes of the territories take place (fusions/splitting up). Hence, they require an ad hoc editing and calculation treatment.

The errors from missing answers and out of range present negligible values.

4.2 Quality of the aggregated data

4.2.1 Benchmark sources

The purpose for analysing the survey unit “person” was to quantify the Closeness and Similarity (quality analysis for vector Accuracy) between the distribution of the Italian Population in the Tax Authority’s archive and the population in Istat sources. The control sources used to analyse the quality of the aggregated data included:

- The resident population based on official statistics from administrative source (POSAS survey on 1 January 2006 and 2007 – Population per gender and civil status)
- The resident population based on census survey (XIV Population Census 2001 per gender, age and place of birth).
It was necessary to compare to the Census 2001 despite it being rather dated, as it is the only one available that allows analysing the distribution of the population per place of birth: the current statistics (POSAS) indeed do not register such data. However, the Regulation of the European Parliament and of the Council on Community statistics on migration and international protection no longer admits such informative gap, requiring instead population statistics per place of birth.

4.2.2 The Sample

The sample analysed wasn’t randomizes, choosing March 2007 as reference date, “person” as survey units, 3.8 millions of records as number\(^6\), and as variable for each person:

- Fiscal code number
- Name, surname and gender
- Date and place of birth
- Municipality and fiscal address
- Municipality and address of legal residence (if different from the fiscal address)
- Source of the fiscal code number, of the fiscal address and the legal residence.

4.3 Accuracy

The distances between target population and benchmarks were measured by means of the following five comparison couples:

- Tax registry population (Tax Authority, extraction of July 2006) vs. resident population based on official statistics from administrative sources (POSAS Survey – Istat, January 2006);
- Total tax registry population (Tax Authority, population extracted in July 2006 and retro-dated to December 2001) vs. resident population based on official statistics from census source (XIV Population Census - Istat, 2001);
- Tax registry population aligned with the registries (Tax Authority, extraction in March 2007) vs. resident population based on official statistics from administrative sources (POSAS survey - Istat, January 2007);
- Tax registry population aligned with the registries (Tax Authority, extraction in March 2007 and retro-dated to December 2001) vs. resident population based on official statistics from census source (XIV Population Census - Istat, 2001);
- Total tax registry population (Tax Authority, extraction of July 2006) vs. population aligned in the registries (Tax Authority, extraction in March 2007).

The characteristics to focus on are place of birth and age.

4.4 Evaluation indicators

The matching between the distribution couples was measured by means of evaluation indicators, used according to the nature of the variables, that is:

\(^6\) Roma, Padova, Salerno, Firenze e Bologna, Monopoli, Verretto, Capri
→ Simple distance index

\[ d = \sum_{i=1}^{k} |y_i - y_i^*| \]

Where

\( y_i \) = absolute frequency of modality \( i \)-th of variable \( Y \) according to which the population of the Tax Authority source was distributed,
\( y_i^* \) = absolute frequency of modality \( i \)-th of variable \( Y \) according to which the population of the Istat source was distributed
\( k \) = \( k \)-th modality of variable \( Y \)

→ Percentage simple distance index

\[ d^* = \frac{1}{N} \sum_{i=1}^{k} |y_i - y_i^*| \times 100 \]

Where

\( N \) = total number of units observed in Istat population.

→ Simple relative dissimilarity index

\[ z_1 = \frac{1}{2} \sum_{i=1}^{k} |f_i - f_i^*| \]

Where

\( f_i \) = relative frequency of modality \( i \)-th of variable \( Y \) according to which the population of the Tax Authority source was distributed,
\( f_i^* \) = relative frequency of modality \( i \)-th of variable \( Y \) according to which the population of the Istat source was distributed
\( k \) = \( k \)-th modality of the variable according to which the population was distributed.

The index \( z_1 \) goes from 0 in case of minimum dissimilarity to 1 in case of maximum dissimilarity.

→ Simple relative dissimilarity index per ordered characters

\[ z'_1 = \frac{1}{k-1} \sum_{i=1}^{k} |F_i - F_i^*| \]

Where

\( F_i \) = accumulated relative frequency of modality \( i \)-th of variable \( Y \) according to which the population of the Tax Authority source was distributed
\( F_i^* \) = accumulated relative frequency of modality \( i \)-th of variable \( Y \) according to which the population of the Istat source was distributed

The index \( z'_1 \) goes from 0 in case of minimum dissimilarity to 1 in case of maximum dissimilarity.

→ Correlation coefficient

\[ r_{yy} = \frac{\sigma_{yy}}{ \sigma_y \sigma_{y^*}^{1/2} } \]

Where

\( \sigma_{yy} \) = co-deviance of \( Y \) and \( Y^* \)
\( (\sigma_y \sigma_{y^*})^{1/2} \) = geometrical average of the deviances of \( Y \) and \( Y^* \).
4.5 Analysis of accuracy in relation to the variable place of birth

The first analysis (Table 1) involved calculating the simple distance index (d), the percentage simple distance index \((d^*)\) and the simple relative dissimilarity index \((z_1)\) between the two couples of population\(^7\): aligned tax registry population and resident population based on Census 2001; total tax registry population and resident population based on Census 2001.

First, the values of the indicators (in particular \(d^*\) and \(z_1\)) confirm the hypothesis according which the aligned population aggregate is most similar to the resident population (target population). The percentage distance equals indeed 26.2% between total population and census population, but drops to 8.4% between aligned population and census population for Capri. The same trend was registered in the other municipality: from 19.5% to 2.5% for Monopoli; from 33% to 6.5% for Rome, etc.

The value of the relative dissimilarity index drops when calculated for the aligned population.

Second, the closeness of the distribution per municipality of birth results to be very good; hence, thanks to the accuracy of the source for variable Place of birth, it could be used as proxy of the target “resident population per place of birth” from traditional surveys. In our case, they could help meet the obligations of Regulation of the European Parliament and of the Council on Community statistics on migration and international protection, according which population statistics per place of birth must be provided. In Italy, the only source available at the moment is the census.

Table 1 – Distance index and dissimilarity index between the distributions of the POPULATION BY MUNICIPALITY OF BIRTH according to the Tax Registry 2007 and to the XIV General Population Census 2001.

<table>
<thead>
<tr>
<th>Municipality</th>
<th>(d)</th>
<th>(d^*)</th>
<th>(z_1)</th>
<th>(d)</th>
<th>(d^*)</th>
<th>(z_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligned tax registry VS. Census</td>
<td>(a)</td>
<td></td>
<td></td>
<td>Total tax registry VS. Census</td>
<td>(b)</td>
<td></td>
</tr>
<tr>
<td>Simple distance index</td>
<td>Percentage distance index</td>
<td>Relative simple dissimilarity index</td>
<td>Simple distance index</td>
<td>Percentage distance index</td>
<td>Relative simple dissimilarity index</td>
<td></td>
</tr>
<tr>
<td>Bologna</td>
<td>60,311</td>
<td>16.2</td>
<td>0.076</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Capri</td>
<td>593</td>
<td>8.4</td>
<td>0.031</td>
<td>1,848</td>
<td>26.2</td>
<td>0.090</td>
</tr>
<tr>
<td>Firenze</td>
<td>41,166</td>
<td>11.6</td>
<td>0.055</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Monopoli</td>
<td>1,154</td>
<td>2.5</td>
<td>0.009</td>
<td>9,097</td>
<td>19.5</td>
<td>0.020</td>
</tr>
<tr>
<td>Padova</td>
<td>317,844</td>
<td>155.1</td>
<td>0.079</td>
<td>375,966</td>
<td>183.5</td>
<td>0.080</td>
</tr>
<tr>
<td>Roma</td>
<td>166,663</td>
<td>6.5</td>
<td>0.027</td>
<td>842,154</td>
<td>33.1</td>
<td>0.125</td>
</tr>
<tr>
<td>Salerno</td>
<td>3,790</td>
<td>2.7</td>
<td>0.014</td>
<td>26,328</td>
<td>19.1</td>
<td>0.032</td>
</tr>
<tr>
<td>Verdetto</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

(a) Population aligned in the registry (Tax Authority, population extracted in March 2007 and retro-dated to December 2001) compared with the resident population based on the official statistics from census source (XIV Population census - Istat, 2001)

---

\(^7\) The aligned population data are not available for Bologna and Firenze. In order to make the distribution couples (one of 2006 and one of 2001) comparable, children born between 2001 and 2006 were subtracted and the age for each person recalculated at 2001. Data on persons who died in that same period – from Authority source – are not available either; hence, the results contain this bias.
(b) Total population in the registry (Tax Authority, population extracted in July 2006 and retro-dated to December 2001) compared with the resident population based on the official statistics from census source (XIV Population census - Istat, 2001)

This analysis was repeated but the variable Place of Birth was changed into 5 modalities: (a) same municipality of residence; (b) another municipality in the same province; (c) another province in the same region; (d) another region; (e) abroad. It confirmed the results of the first analysis.

Table 2– Distance index and dissimilarity index between the distributions of the POPULATION BY PLACE OF BIRTH (same as that of residence, another municipality in the same province; another province in the same region; another region; abroad) according to the Tax Registry 2007 and to the XIV General Population Census 2001.

<table>
<thead>
<tr>
<th>Municipality</th>
<th>d (a)</th>
<th>d* (a)</th>
<th>z₁ (a)</th>
<th>d (b)</th>
<th>d* (b)</th>
<th>z₁ (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bologna</td>
<td>42,077</td>
<td>11.3</td>
<td>0.052</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Capri</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Firenze</td>
<td>32,656</td>
<td>9.2</td>
<td>0.045</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Monopoli</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Padova</td>
<td>17,842</td>
<td>8.7</td>
<td>0.039</td>
<td>50,270</td>
<td>24.5</td>
<td>0.125</td>
</tr>
<tr>
<td>Roma</td>
<td>138,170</td>
<td>5.4</td>
<td>0.021</td>
<td>699,046</td>
<td>27.4</td>
<td>0.123</td>
</tr>
<tr>
<td>Salerno</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>Verretto</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

(a) Population aligned in the registry (Tax Authority, population extracted in March 2007 and retro-dated to December 2001) compared with the resident population based on the official statistics from census source (XIV Population census - Istat, 2001)

(b) Total population in the registry (Tax Authority, population extracted in July 2006 and retro-dated to December 2001) compared with the resident population based on the official statistics from census source (XIV Population census - Istat, 2001)

4.6 Analysis of accuracy in relation to the variable Age

In the third analysis, the simple percentage distance index (d*), the simple relative dissimilarity index (z₁) and the simple relative dissimilarity index were calculated per ordered characters (z₁').

The data were compared with the Census 2001 (Table 3), based on the distribution of the populations per five-year age classes. While the closeness between populations' couples is better for the aligned population, it is even more so when considering as benchmark the current demographic statistics (table 4).

In this context too, the distances d* between target population and benchmark rise from 6.9% in Padova to 11.6% in Bologna. The similarity between populations is very high: the simple relative dissimilarity index is always under 0.06.

Hence, thanks to the accuracy of the source for the variable age, it could be used as proxy of the "Resident population per municipality and per age" of traditional analyses.
Based on the current statistics (Table 4), closeness in absolute value (d) is surely higher for the target population; d* equals 1.9% in Firenze, 2.2% in Capri, 2.3% in Bologna, and 8.3% in Salerno.

In the fourth analysis (table 5), the correlation coefficient was also calculated and compared with the populations distributed per each age year. The current demographic statistics were used here as benchmark. The correlation
coefficient is particularly high (more than 0.98) for all municipalities analysed. The percentage distance amounts to 5.7% for Roma and 8.4% for Salerno: 146 thousand units to be recuperated, 11 thousand more compared to the benchmark.

Table 5– Distance index and dissimilarity index between the distributions of the POPULATION BY FIVE-YEAR AGE CLASSES according to the Tax Registry 2007 and to Posas.

<table>
<thead>
<tr>
<th>Municipality</th>
<th>Balanced distance index</th>
<th>Percentages distance index</th>
<th>Relative simple dissimilarity index</th>
<th>Relative simple dissimilarity index</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bologna</td>
<td>8,720</td>
<td>2.3</td>
<td>0.033</td>
<td>0.010</td>
<td>0.999</td>
</tr>
<tr>
<td>Capri</td>
<td>215</td>
<td>2.9</td>
<td>0.013</td>
<td>0.003</td>
<td>0.997</td>
</tr>
<tr>
<td>Firenze</td>
<td>7,423</td>
<td>2.0</td>
<td>0.010</td>
<td>0.003</td>
<td>0.999</td>
</tr>
<tr>
<td>Monopoli</td>
<td>1,823</td>
<td>3.6</td>
<td>0.006</td>
<td>0.002</td>
<td>1.000</td>
</tr>
<tr>
<td>Padova</td>
<td>10,086</td>
<td>4.7</td>
<td>0.012</td>
<td>0.003</td>
<td>0.997</td>
</tr>
<tr>
<td>Roma</td>
<td>146,660</td>
<td>5.7</td>
<td>0.027</td>
<td>0.006</td>
<td>0.988</td>
</tr>
<tr>
<td>Salerno</td>
<td>11,374</td>
<td>8.4</td>
<td>0.034</td>
<td>0.016</td>
<td>0.988</td>
</tr>
<tr>
<td>Verretto</td>
<td>n.c.</td>
<td>n.c.</td>
<td>n.c.</td>
<td>n.c.</td>
<td>n.c.</td>
</tr>
</tbody>
</table>

(a) Population aligned in the registry (Tax Authority, population extracted in March 2007 and retro-dated to December 2001) compared with the resident population based on the official statistics from census source (XIV Population census - Istat, 2001)

(b) Total population in the registry (Tax Authority, population extracted in July 2006 and retro-dated to December 2001) compared with the resident population based on the official statistics from census source (XIV Population census - Istat, 2001)

4.7 **Analysis of accuracy in relation to some demographic indicators**

Table 6 reports some demographic indicators calculated on both sources (Istat and tax registry). It represents another test made possible by the nature of the variables, in addition to the test conducted with the statistical evaluation indicators. Thanks to the extreme similarity among values, the source analysed could have a double value: registry input and real diffusion process input.

5 **Critical points**

All these analyses have led to some results but have also revealed some critical aspects too.

One such aspect is the time lag between the moment in which the data are updated and when they occurred, such as the delay with which some registries communicate births, an event that activates the attribution of a fiscal code number. This of course affects the time concordance between AT data and Istat statistical data on the first years of life. However, this problem occurs in two of the 8 sample municipalities (Salerno and Rome). The data in the other 6 municipalities are updated even as regards the younger age classes.

Such archive-updating problem regards also the “deaths”, as the first databank constructed and set up only involved the basis of the resident population. This lack of communication regards elderly persons (aged 75 and over) who died about 10/15 years ago and because of the non-automation of the registries until that date do not appear cancelled in the information provided.
Another such critical element, intrinsic to the cross-control system of the tax authority with the resident population registries, is that quality problems present in the personal data archives are projected in the A.T. database. However it is possible to find solutions to these problems by interrelating the resident population registries with the Tax Authority, and by using at the same time and together other sources of personal information.

6 Conclusions
The source analysed represents an excellent input for the national registry of persons, though to be used with caution due to the experimental aspect of these analyses.

The accuracy of the distributions by place of birth and by age in relation to the Census benchmark and current statistics is very high, just as the accuracy of the demographic indicators. Finally, the source contains the variable Place of birth currently available only from census source.

Though this paper does not report the first (positive) analyses of the territorial variables (residence address and municipality of birth), these tests are important. The availability of a quality territorial archive (potentially) not only provides a non-negligible support for future population census but allows also for theme developments (geo-reference and territorial reading of the phenomena).

Hence, though the information from the tax archive is by itself important, with the possibility of having an individual archive with an identification key of the persons (fiscal code number), this register represents the first step to the realisation of a more complex and important system that will provide absolutely innovative statistical information.

This registry of persons still in project will help the official statistics improve the production of information and reduce the costs. These objectives can be reached, reducing the data object of direct surveys and implementing data produced through their integration with other administrative sources at a micro-data level.
<table>
<thead>
<tr>
<th>Municipality</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bologna</td>
<td>175,194</td>
<td>197,447</td>
<td>372,641</td>
<td>44.7</td>
<td>49.2</td>
<td>47.1</td>
</tr>
<tr>
<td>Capri</td>
<td>3,447</td>
<td>3,653</td>
<td>7,100</td>
<td>42.4</td>
<td>45.1</td>
<td>43.8</td>
</tr>
<tr>
<td>Firenze</td>
<td>171,807</td>
<td>192,533</td>
<td>364,340</td>
<td>44.3</td>
<td>48.6</td>
<td>46.6</td>
</tr>
<tr>
<td>Monopoli</td>
<td>23,583</td>
<td>24,448</td>
<td>48,031</td>
<td>39.4</td>
<td>41.7</td>
<td>40.6</td>
</tr>
<tr>
<td>Padova</td>
<td>95,412</td>
<td>105,907</td>
<td>201,319</td>
<td>43.5</td>
<td>48.1</td>
<td>46.0</td>
</tr>
<tr>
<td>Roma</td>
<td>1,225,681</td>
<td>1,344,834</td>
<td>2,570,515</td>
<td>42.2</td>
<td>45.4</td>
<td>43.9</td>
</tr>
<tr>
<td>Salerno</td>
<td>67,354</td>
<td>74,132</td>
<td>141,486</td>
<td>43.3</td>
<td>46.9</td>
<td>45.2</td>
</tr>
<tr>
<td>Verretto</td>
<td>174</td>
<td>173</td>
<td>347</td>
<td>41.7</td>
<td>46.4</td>
<td>44.1</td>
</tr>
</tbody>
</table>

Population aligned in the registry. Tax Authority, population extracted in March 2007 and retro-dated to December 2001

<table>
<thead>
<tr>
<th>Municipality</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bologna</td>
<td>174,708</td>
<td>199,035</td>
<td>373,743</td>
<td>45.2</td>
<td>49.6</td>
<td>47.7</td>
</tr>
<tr>
<td>Capri</td>
<td>3,513</td>
<td>3,734</td>
<td>7,247</td>
<td>42.8</td>
<td>45.5</td>
<td>44.2</td>
</tr>
<tr>
<td>Firenze</td>
<td>171,544</td>
<td>195,357</td>
<td>366,901</td>
<td>44.6</td>
<td>49.0</td>
<td>46.9</td>
</tr>
<tr>
<td>Monopoli</td>
<td>24,257</td>
<td>25,589</td>
<td>49,846</td>
<td>39.6</td>
<td>41.9</td>
<td>40.7</td>
</tr>
<tr>
<td>Padova</td>
<td>99,314</td>
<td>111,671</td>
<td>210,985</td>
<td>43.3</td>
<td>47.8</td>
<td>45.7</td>
</tr>
<tr>
<td>Roma</td>
<td>1,192,792</td>
<td>1,354,060</td>
<td>2,546,852</td>
<td>41.6</td>
<td>45.2</td>
<td>43.5</td>
</tr>
<tr>
<td>Salerno</td>
<td>63,005</td>
<td>71,815</td>
<td>134,820</td>
<td>41.4</td>
<td>45.2</td>
<td>43.4</td>
</tr>
<tr>
<td>Verretto</td>
<td>163</td>
<td>164</td>
<td>327</td>
<td>42.4</td>
<td>47.0</td>
<td>44.7</td>
</tr>
</tbody>
</table>

Resident population based on official statistics. Posas survey - Istat, January 2006
Bibliography


Della Rocca G., Di Zio M., Luzi O., Simeoni G., () “DEA (Indices for Data Editing Assessment) – Sistema per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori di SIDI”. Istat.