A micro-based approach to ensure consistency among administrative sources and to improve population statistics

Un approccio basato sui microdati per garantire la coerenza tra fonti amministrative e per migliorare le statistiche demografiche

Gianni Corsetti, Sabrina Prati, Valeria Tomeo, Enrico Tucci

Abstract According to Istat modernisation programme the Registers Integrated System (RIS) is the core of the data production process. The RIS enables to enlarge the level of the analysis and the quality of the information collected linking at micro level the economic and social phenomena. In this view, the Base Register on Individuals and Households (BRIH) that identifies the usually-resident population will be enriched by the information collected by the other Base Registers (enterprises, addresses, activities) and by the thematic ones (Labour, Educational level, etc.). BRIH will be the common target for Continuous Census and Demographic statistics and for Social Statistics. To improve the accuracy of the population outputs in terms of timeliness, coverage and consistency it has been developed an information system of micro demographic accounting based on a longitudinal statistical register of individuals (Anvis). Anvis let us to continuously monitor population changes by linking the population stock with outflows and inflows. The new system allows us to adopt a multistate demographic approach for longitudinal models and for indicators based on change-of-state probabilities. Moreover, it let us to bring together all the events involving a given individual especially for the analysis of migration and migration trajectories.

Abstract *Secondo il programma di modernizzazione dell'Istat, il Sistema Integrato dei Registri (SIR) è il nucleo del processo di produzione dei dati. Il SIR consente di ampliare il livello delle analisi e la qualità delle informazioni raccolte integrando a*

[1] Gianni Corsetti, Istat; email: giacorsetti@istat.it
Sabrina Prati, Istat; email: prati@istat.it
Valeria Tomeo, Istat; email: tomeo@istat.it
Enrico Tucci, Istat; email: tucci@istat.it

livello micro i fenomeni economici e sociali. In questa prospettiva, il Registro Base su Individui e Famiglie (RBI), che identifica la popolazione abitualmente dimorante, sarà arricchito dalle informazioni raccolte dagli altri registri di base (imprese, indirizzi, attività) e da quelli tematici (manodopera, livello di istruzione, eccetera.). RBI sarà il punto di riferimento comune sia per il censimento permanente che per le statistiche demografiche e sociali. Per migliorare la qualità delle stime sulla popolazione in termini di tempestività, copertura e coerenza è stato sviluppato un sistema informativo di contabilità micro-demografica basato su un registro statistico longitudinale di individui (Anvis). Anvis ci permette di monitorare continuamente i cambiamenti della popolazione collegando lo stock di popolazione con gli eventi demografici. Attraverso Anvis sarà possibile, inoltre, realizzare analisi longitudinali seguendo l'approccio della demografia multistato e costruire indicatori basati sulle probabilità di cambiamento di stato. Infine, il sistema Anvis riunirà gli eventi riferiti ad un singolo individuo in modo da consentire studi sulle migrazioni e sulle traiettorie migratorie.

Key words: administrative data, population statistics, data integration, longitudinal database.

1 Introduction

This paper presents an overview of some results from the studies which Istat is carrying out in the context of the modernisation of Population and Social statistics. The object is to make the best use of the available administrative sources within the population statistics. New technologies and administrative changes allow to enhance the exploitation and the quality of data collected mainly due to the possibility to integrate different data sources. Istat has been developed a project of engineering and implementing an information system of micro demographic accounting based on a statistical population register of individuals Anvis (ANagrafic VIrtual Statistical database). Anvis makes it possible to continuously monitor population changes by linking the population stock with outflows and inflows. Fundamental requirements in Anvis information system are data integration and longitudinal consistency. For this reason, Istat needs to check and correct the system to improve data quality in terms of timeliness, coverage and consistency with other administrative sources. Italy supports the adoption of a Regulation on Demographic Statistics. We are aiming at improving the features of our systems according to the harmonization principle incorporated in the Regulation. ANVIS ensure the coherence and the consistence among the information requested by the Regulation on demographic statistics (in progress) and the information collected pursuant Regulation (EC) No 862/2007(of the European Parliament and of the Council on Community statistics) on migration and international protection and Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses. At the same time Anvis ensure the consistency of the Regulation to our National Legal requirements.

The new system will allows to achieve two research targets:

- A new longitudinal microdata system (MIDEA) for longitudinal models and techniques and demographic indicators based on change-of-state probabilities;
- An integrated system of administrative signals for measuring emigration flow.

2 The new longitudinal microdata system (Mi.DE.A-ANVIS)

The MIcro Demografic Accounting (Mi.DE.A) is a very large-scale linkage study which is going to be build up at Istat by using data of last Population Census and those of current administrative and statistical sources. These include all individual data available from the compulsory registration System (POPREG), such as Vital Events data (births, deaths, migration in or out Italy). Other data will be included in future steps, i.e marriages separations and divorces, or Health data collected by the National Health Service (Cedap, hospital admissions), or Household sample data.

There are some key differences among the MiDEA project and other longitudinal studies conducted in Europe as the Scottish Longitudinal Study (SLS) or as the England and Wales Longitudinal Study (LS). First of all the size of population involved into the Study. MiDEA is designed to include all the individual records captured in the last Population Census, around 60 million of records. Information for these people have been linked with vital events in order to follow the changing through time. In addition new records will be included or withdrawn each year using migration flows.

The ANagrafic VIrtual Statistical database (AN.VIS.) is one of the most relevant outcome of the project. It includes one records for each individual registered as resident according to census and flows data. An important aspect of the ANVISs individual data is the information on other household members (those who live in the same household as resulting at the time of last census or according to the variations collected by population registers).

The db [ANVIS] is a part of the Base Register on Individuals and Households (BRIH) in which individual Information can be merged with information of those living in the same household. Household data provide valuable additional information. In fact is the universe of populations and sub-populations for census and samples social surveys. This opens new opportunities and methodological challenges for the design of sample surveys. In perspective for each individual demographic information will be integrated with other socio-economic also in longitudinal perspective.

The starting point of the MiDEA are the Unit Records of Population Census at census time. The first step was to up-date census data until 31th December 2012 using Population register (POPREG) flows report entries and exits (Births, Deaths, Immigration and Emigration, Administrative adjustments); this process have been

repeated for each year in order to up-date data at 31th December 2016 (latest year available at moment).
At the end of each year we match Anvis and POPREG unit records in order to carry out corrections due to underreporting of changes on personal data. However, due to mistakes by the individuals or administrations concerned, errors occur especially related coverage issues of the demographic flows.
At the same time MiDEA-ANVIS ensure the consistency of the Regulation to our National Legal requirements.
The new longitudinal microdata system can improve the accuracy of currently released demographic indicators, especially the ones that are called incidence indicators because they require the calculation of the population at risk.
In the macro approach population at risk is calculated as an average between initial and final population, making uniform distribution hypothesis in the time of reference.
Longitudinal microdata system can link population stock and demographic events in the continuous time, so that it is possible a closer calculation of the population at risk in terms of person-time.
Person-time is an estimate of the actual time-at-risk – in years, months, or days – that all participants contributed to a population in a given reference period. A subject is eligible to contribute person-time only so long as that person does not yet have left population due to emigration or death (Figure 1).
Since ANVIS contains full information for birthdate and dates referred to any migratory events or death, we can use exact dates to calculate person-time.

Figure 1: Exposure to risk of the individuals in a reference year by type of demographic event

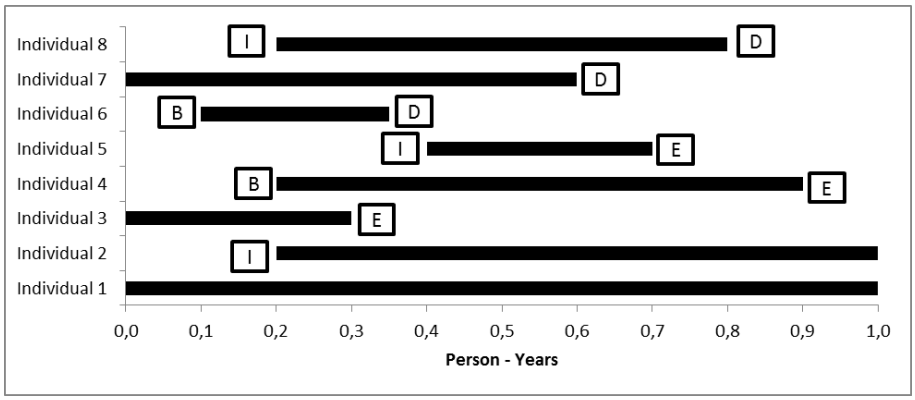
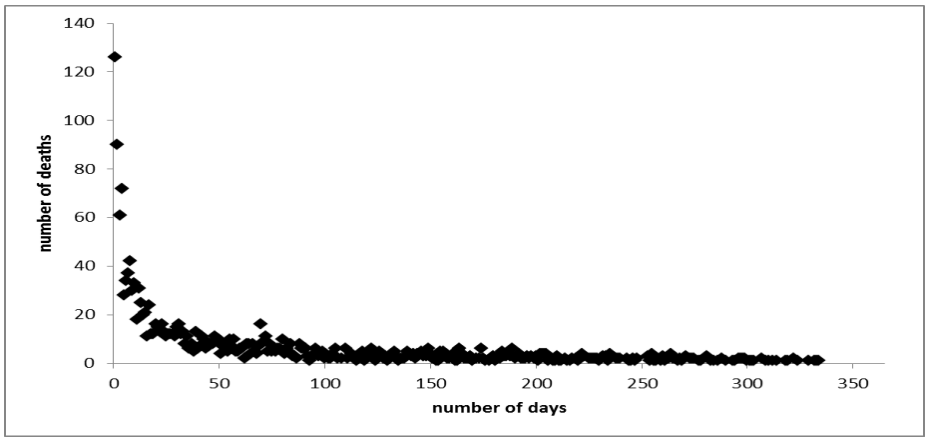


Table 1 shows age-specific death rates per thousand person-year by sex referred to 2016. They are compared with age-specific death rates per thousand inhabitant currently release by Istat and estimated with aggregated data (http://dati.istat.it/?lang=en&SubSessionId=1ca4016d-d5e1-4d13-a15c-a7ec98464094&themetreeid=-200).

Table 1: Age-specific death rates by sex and type of estimation – Year 2016

Age class	Male			Female		
	average population (A)	person-years (B)	relative difference	average population (A)	average population (B)	relative difference
0-4	0.7	0.8	14.3	0.6	0.7	16.7
5-9	0.1	0.1	-	0.1	0.1	-
10-14	0.1	0.1	-	0.1	0.1	-
15-19	0.3	0.3	-	0.1	0.1	-
20-24	0.4	0.4	-	0.2	0.2	-
25-29	0.4	0.4	-	0.2	0.2	-
30-34	0.5	0.5	-	0.2	0.2	-
35-39	0.7	0.7	-	0.4	0.4	-
40-44	1.1	1.1	-	0.7	0.7	-
45-49	1.8	1.7	5.6	1.1	1.1	-
50-54	3.0	3.0	-	1.8	1.8	-
55-59	5.0	5.0	-	2.8	2.8	-
60-64	8.0	8.0	-	4.3	4.2	2.3
65-69	13.2	13.1	0.8	7.1	7.0	1.4
70-74	21.4	21.3	0.5	11.7	11.7	-
75-79	35.7	35.4	0.8	20.8	20.6	1.0
80-84	65.9	65.4	0.8	42.8	42.5	0.7
85-89	123.7	122.9	0.6	87.3	86.7	0.7
90-94	218.9	216.3	1.2	168.5	166.5	1.2
95+	351	345.2	1.7	295.9	292.6	1.1
Total	10.0	10.0	-	10.2	10.2	-

Figure 2: Distribution of deaths in the first year of life by number of days between birth and death – Year 2016



Results are very close but differences increase in the first age class, where the uniform distribution hypothesis in the reference period is highly unlikely, because of the super mortality of the newborns in the very early period after birth, as shown in Figure 2.
This is a first attempt to fully exploit the potential of the new system under construction. Nevertheless we can underline the importance of longitudinal microdata in pursuing the goal of more consistently estimates.

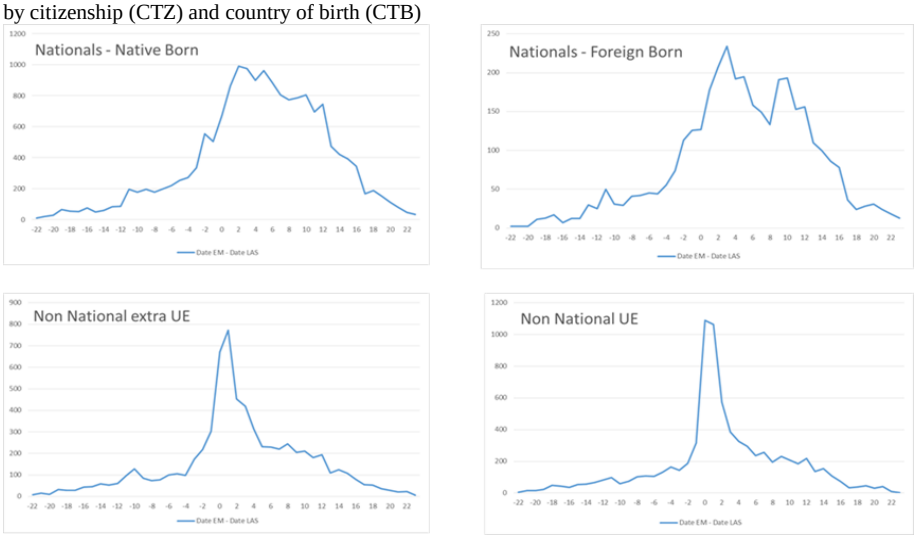
3 Building an integrated system of administrative signals for measuring emigration flow

The necessity for improvements in migration statistics is not an issue that is confined to a single country, with an increased international policy focus on the socio-economic impacts of demographic change. As a dominant driver of population growth, robust estimation of immigration and emigration is key to the production of consistent national and regional population projections for EU countries (Lanzieri, 2007) and the implementation of EU Regulation 862/2007 has provided a statutory basis for greater harmonization of international migration statistics in Europe (Boden and Rees, 2008). However, there is a certain asymmetry between data availability on immigration and emigration, for two main reasons: departures tend to be less well recorded than arrivals and it is difficult to count persons leaving the country from a statistical point of view because of their absence (United Nations, 2010).
In Italy, a total of 410 thousand people emigrated abroad from the Census date (9th October 2011) to 1th January 2016, while around 450 thousand people were deregistered ex officio from the population registers. The latter figure is included in the demographic balance as «deregistration for other reasons» (based on a macro approach) but it is excluded from the migration flows statistics (micro events). Nevertheless, it is likely that a deregistration refers to someone who left the country without informing the administrative accountable office of their departure. Among these 450 thousand, there were almost 300 thousand of deregistration that could be assimilated to emigrations since they are related to individuals that, after the date of cancellation, do not reappear at least for 12 months. In order to classify these deregistration as migrations flows, the date of occurrence of the emigration has to be estimated.
To this aim, we have used the information coming from other administrative sources (Labor and Education registers, Tax Returns register, Earnings, Retired, and Non-Pension Benefits registers, Permits to Stay archive) to derive a monthly Administrative Signal of presence in Italy.
As already mentioned, there have been 409,758 emigrations to abroad (EM) from the Population Census (9th October 2011) to 1th January 2016. Every individual emigration flow has been associated with the respective Last Administrative Signal (LAS) on the territory.

The underlying idea is that there is a relationship between the Date of emigration (Date EM) and the date of the Latest Administrative Signal (Date LAS). In particular, we have made the hypothesis that the difference between the mentioned dates is distributed normally.

Figure 3 shows that for non-nationals, the difference is strongly concentrated around 0 and 1 month, meaning that the majority of foreign citizens leaves the country very soon after they lose the signal of presence in Italy. The variability of the distribution is different for Nationals, meaning that, even without a signal of study or work, they can afford to stay longer in the country before emigrating abroad. In order to consider a deregistration ex officio as an emigration to abroad, the date of occurrence of the emigration (Date EM*) has to be imputed since the date of deregistration ex officio (Date DER) of a person has probably been postponed respect to the date of occurrence of the emigration (Date EM). To this aim, also the individual Date of deregistration ex officio has been associated with the respective Date of latest Administrative Signal on the territory (Date LAS).

Figure 3: Observed differences (in month) between Date of emigration and Latest Administrative Signal by citizenship (CTZ) and country of birth (CTB)



Observed Diff EMLAS distribution (by CTZ and CTB) has been used to estimate the date of occurrence of the emigration (Date EM*) for those deregistration carried out ex officio. In summary, the procedure was as follows:

1. Estimation of normal distribution parameters from observed data (Diff EMLAS = Date EM – Date LAS).
2. Computation of emigration data (Date EM* = Random Normal (Diff EMLAS) + Date LAS) for individuals deregistered ex officio by simulating Diff EMLAS with the normal distributions estimated in the previous step.

Figure 4: Observed differences (in month) between Date of deregistration and Latest Administrative Signal by citizenship (CTZ) and country of birth (CTB)

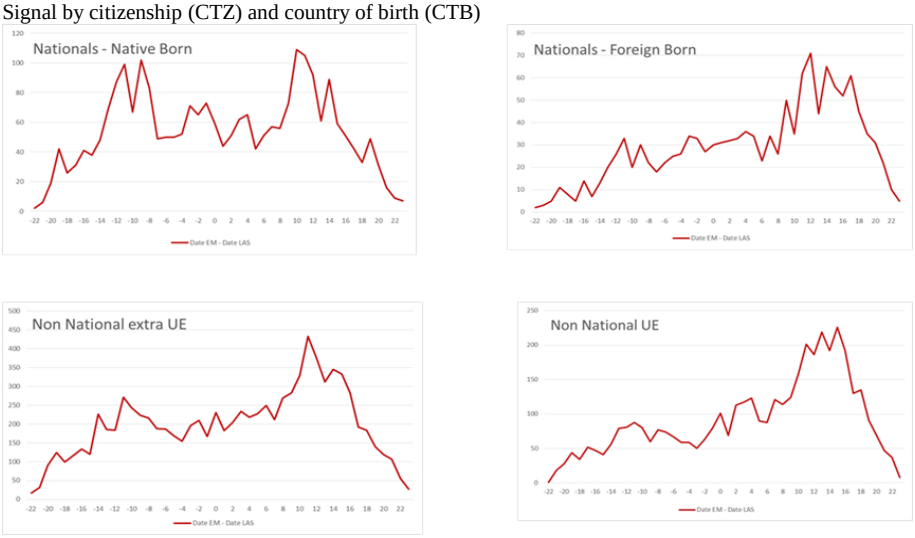
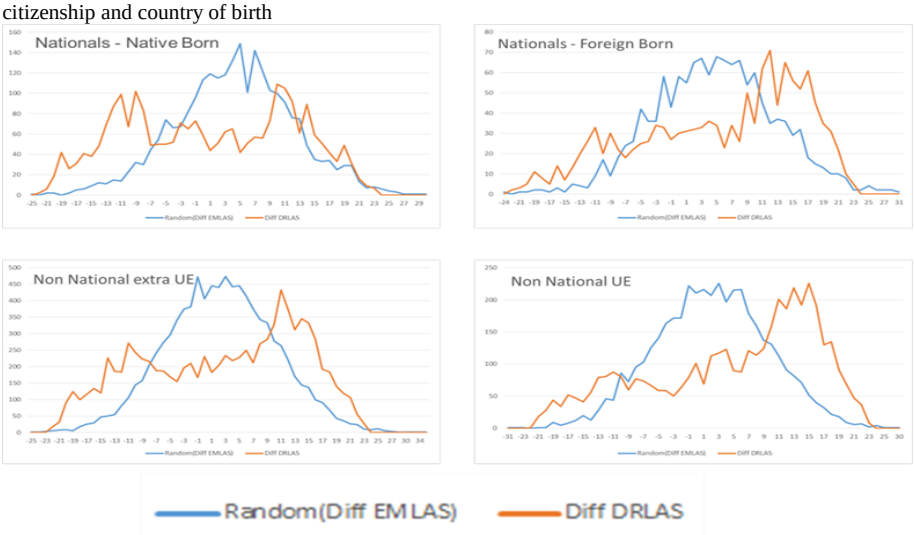


Figure 5: Estimated random normal differences (Diff EMLAS*) and observed ones (Diff DERLAS) by citizenship and country of birth



According to a first estimation of the emigration flows, the number of emigrants, almost doubled the number published by the official statistics. The latest figure for 2015, along with the number of immigrants during the same year, produces a negative net migration of -4 thousand while official statistics reported a positive net migration equal to +133 thousand units (Table 2).

An analysis by citizenship reveals that the Non Nationals report the largest impact on the emigration figures (from 45 thousand to 112 thousand). As a result the number of

Non National leaving the country is almost equal the number of Nationals. If we take into consideration that more than 20% of emigrants with Italian citizenship is born abroad, we can conclude that the majority of the people that leave the country has a foreign origin.

Table 2: Immigration, emigration flows and net migration observed and estimated (observed and estimated) by citizenship – Year 2015

Citizenship	***Immigration***	***Emigration***	***Net MIG***	***Emigration****	***Net MIG****
		Observed		***Estimated***	
Nationals	30,052	102,259	-72,207	147,300	-117,248
Non nationals	250,026	44,696	205,330	137,450	112,576
Total	280,078	146,955	133,123	284,750	-4,672

4 Conclusion

There is a lack of longitudinal database in Italy. An integrated period and cohort approach is a pillar of the MiDEA-ANVIS project. In order to produce information useful to the interpretation and the explanation of demographic phenomena, a deep change of strategies in organising basic statistics data is required, overcoming the traditional period approach at favour of cohort one.

Migration is the most difficult component of population change to evaluate, as if there is a comprehensive system which registers migration in our Country, either moves to or from the rest of the world, or moves within the Italian Municipalities.

The traditional concepts of resident or present population are less and less reliable to assess people usually or actually living in a Municipality, due to the increasing mobility of the population and the international migration (an international migrant is defined by the United Nations (UN) as someone who changes country of residence for 12 months or more).

Measuring migration movements of populations in different countries is a challenging task.

Administrative sources are an extremely powerful tool and a potential source of information for emigration statistics. The Italian case provided empirical evidence of the issues to be confronted and the challenges to the use of a data integration for improving the quantity and quality of data on emigration. A coherent and consistent database that contains detailed, up-to-date and accurate information allows to study migration through a longitudinal approach (Return Migration, Circular Migration).

References

1. Verhoef, Rolf, and Dirk J. van de Kaa. "Population registers and population statistics." Population Index (1987): 633-642.
2. Poulain, Michel, Anne Herm, and Roger Depledge. "Central population registers as a source of demographic statistics in Europe." Population 68.2 (2013): 183-212.http://www.rsc.org/dose/title of subordinate document. Cited 15 Jan 1999.
3. Myrskyla, Pekka. "New statistics made possible by the use of registers." Statistical Journal of the United Nations Economic Commission for Europe 16.2, 3 (1999): 165-180.