04/03/2021
Author: Diego Zardetto

Project ‘Stocks & Flows’

Section 1 – Background Information

The Italian National Institute of Statistics (Istat) is currently striving to radically change its production processes. The backbone of the envisioned new production system is the ‘Integrated System of Statistical Registers’ (ISSR), namely a system of connected registers that will be used as reference for all the statistical programs carried out by Istat. A pivotal role within the ISSR will be played by the ‘Base Register of Individuals’ (BRI), a comprehensive statistical register that will integrate and store data gathered from disparate sources about people usually or temporarily residing in Italy.

One of the most important expected outputs of the new statistical production system is the modernization of the Italian population census. Traditional population censuses have been conducted in Italy every ten years up to 2011, and their outcomes have been routinely used to correct municipal civil registries once-in-a-decade. Starting from 2018, the next Italian population census will no longer be a complete enumeration survey, but rather result from a *twofold* large scale sample survey that will be carried out each year. Istat has named this new census design ‘Permanent Census’. The Permanent Census involves two simultaneous sample surveys: the ‘A’ survey and the ‘L’ survey. The L component relies on a list sample: its main objective is to observe variables that are either of insufficient quality or not available at all in the BRI. The A component is instead based on an area sample: it is designed to provide yearly estimates of the under-coverage and over-coverage rates of the BRI, evaluated at national and local levels for different sub-population profiles (defined by variables like ‘sex’, ‘age class’, ‘nationality’).

The new production system will enable Istat to deliver official population size estimates more frequently than it happened before through traditional censuses, very likely on a yearly basis. *Raw* estimates of population counts will result from the integration of the BRI with the A component of the Permanent Census. To this end, a dual estimation system could be adopted, based on the linkage between the BRI (first capture) and the A sample survey (second capture). Eventually, *official* estimates of population counts will be derived from a macro-integration procedure that will simultaneously adjust both raw population size estimates (*stocks*) and raw civil registry figures (*flows*), in such a way that the resulting data exactly fulfill the Demographic Balancing Equation (DBE).

The DBE states that the population counts at time $\mathrm{t}+1$ must be equal to the population counts at time t plus the sum of the natural increase and the net migration occurred between t and $\mathrm{t}+1$:

$$P^{(t+1)} = P^{(t)} + N + M \tag{1}$$

where the natural increase, N, is the difference between births and deaths, and the net migration, M, is the difference between immigrants and emigrants:

$$\begin{cases} N = B - D \\ M = I - E \end{cases} \tag{2}$$

In Italy, raw estimates of stocks and flows entering the DBE will be obtained *independently*. Birth, death and migration figures will be provided by municipal civil registries, while population size estimates at subsequent reference times will be derived from the BRI and the Permanent Census. Therefore, owing to sampling and non-sampling errors affecting raw estimates of stocks and flows, the DBE will *not* be trivially satisfied. To solve this problem, Istat made the decision to rely on methods that are commonly adopted inside National Statistical Institutes (NSIs) for balancing large systems of national accounts[1]. A macro-integration procedure has been implemented accordingly, which will ensure consistency between official (i.e. *adjusted*) estimates of demographic stocks and flows. The next section provides a concise description of this procedure.

Section 2 – Problem Formulation

We formalize the task of finding a system of consistent estimates of demographic stocks and flows as a constrained optimization problem. This is accomplished along the lines of (Stone et al., 1942) and (Byron, 1978), by suitably reformulating the models and algorithms introduced in those classical papers.

Given *initial* estimates of all the aggregates entering the demographic balancing equations (1) defined for all the geographic areas of a given territorial level, we search for *final* estimates that are *balanced*, i.e. (i) satisfy all the DBEs, and (ii) are *as close as possible* to the initial estimates. Therefore, the objective function to be minimized is an appropriate distance metric between final and initial estimates, while the constraints acting on the final estimates are the area-level DBEs. Moreover, we adopt a *weighted* distance metric such that aggregates whose initial estimates are more *reliable* will tend to be changed less.

Let us suppose we have initial estimates of the population size of k Italian regions $\mathrm{U_i}$ (as we will see later, “regions” can actually be any population partition, e.g. territory$*$sex$*$age classes) at times t and $\mathrm{t} + 1$, as well as initial estimates of births, deaths and natural increase occurred for each region between time t and $\mathrm{t} + 1$:

$$\begin{cases} P^{(t)} = \left(P_1^{(t)}, \dots, P_k^{(t)}\right)' \\ P^{(t+1)} = \left(P_1^{(t+1)}, \dots, P_k^{(t+1)}\right)' \\ B = (B_1, \dots, B_k)' \\ D = (D_1, \dots, D_k)' \\ N = (N_1, \dots, N_k)' \end{cases} \tag{3}$$

Moreover, let us suppose we have initial estimates of the *Migration Flows Matrix* F, whose generic element $\mathrm{F_{ij}}$ equals the number of people who *moved* from region i to region j between time t and $\mathrm{t} + 1$:

$$F = \begin{pmatrix} 0 & F_{1,2} & \cdots & F_{1,k} & F_{1,k+1} \\ F_{2,1} & 0 & \cdots & F_{2,k} & F_{2,k+1} \\ \cdots & \cdots & 0 & \cdots & \cdots \\ F_{k,1} & F_{k,2} & \cdots & 0 & F_{k,k+1} \\ F_{k+1,1} & F_{K+1,2} & \cdots & F_{k+1,k} & 0 \end{pmatrix} \tag{4}$$

[1] Indeed, the National Accounts divisions of most NSIs routinely make use of independent initial estimates, which:
(i) are characterized by different degrees of reliability (as is also the case of demographic stocks and flows);
(ii) have to be adjusted to satisfy a large set of accounting identities (as is the system of DBEs associated to any partition of the overall population into estimation domains).

Note that the $(\mathrm{k}+1)^{\text{th}}$ row and column of F represent migrations from and to any territory *outside* the nation, thus $\mathrm{k}+1$ means *"abroad"*. Note also that matrix F is not, in general, symmetric nor antisymmetric.

Let us indicate with M the *Net Migration Matrix*, whose generic element $\mathrm{M_{ij}}$ equals the count of people who *immigrated* in region i from region j *minus* the count of people who *emigrated* from region i to region j, $\mathrm{M_{ij}} = \mathrm{F_{ji}} - \mathrm{F_{ij}}$:

$$M = \begin{pmatrix} 0 & M_{1,2} & \cdots & M_{1,k} & M_{1,k+1} \\ -M_{1,2} & 0 & \cdots & M_{2,k} & M_{2,k+1} \\ \cdots & \cdots & 0 & \cdots & \cdots \\ -M_{1,k} & -M_{2,k} & \cdots & 0 & M_{k,k+1} \\ -M_{1,k+1} & -M_{2,k+1} & \cdots & -M_{k,k+1} & 0 \end{pmatrix} \tag{5}$$

Note that matrix M is *antisymmetric* and actually equal to minus twice the antisymmetric part of F:

$$\begin{cases} M = -M^{t} \\ M = F^{t} - F = -2F^{A} \end{cases} \tag{6}$$

Furthermore, let us assume we can attach to each *atomic* initial estimate involved in (3) (4) and (5) a measure of *reliability,* $R \in [0, \infty)$. These reliability measures could be either based on proper statistical measures (e.g. proportional to inverse estimated variances) or derived from an assessment made by subject matter experts. For instance, we will indicate the reliability measure of a generic element $\mathrm{M_{ij}}$ of the Net Migration Matrix M as $R[M_{ij}]$. Note that $R[\cdot] \to \infty$ will signal *absolute reliability*, and thus *prevent* the corresponding initial atomic estimates from being altered.

Lastly, let us denote *raw estimates* with a *tilde* (e.g. $\widetilde{M}_{ij}$) and *balanced estimates* with a *circumflex hat* (e.g. $\widehat{M}_{ij}$). Given (3), (4), and (5), we define the objective function, L, for the constrained optimization problem as follows:

$$\begin{aligned} L&\left(\hat{P}^{(t+1)}, \hat{P}^{t}, \hat{B}, \widehat{D}, \widehat{N}, \hat{F}, \widehat{M}\right) \\ &= \sum_{i=1}^{k} \left(\hat{P}_i^{(t+1)} - \tilde{P}_i^{(t+1)}\right)^2 R\left[\tilde{P}_i^{(t+1)}\right] + \sum_{i=1}^{k} \left(\hat{P}_i^{(t)} - \tilde{P}_i^{(t)}\right)^2 R\left[\tilde{P}_i^{(t)}\right] \\ &+ \sum_{i=1}^{k} \left(\hat{B}_i - \tilde{B}_i\right)^2 R\left[\tilde{B}_i\right] + \sum_{i=1}^{k} \left(\widehat{D}_i - \widetilde{D}_i\right)^2 R\left[\widetilde{D}_i\right] + \sum_{i=1}^{k} \left(\widehat{N}_i - \widetilde{N}_i\right)^2 R\left[\widetilde{N}_i\right] \\ &+ \sum_{i=1}^{k+1} \sum_{j=1}^{k+1} \left(\hat{F}_{ij} - \tilde{F}_{ij}\right)^2 R\left[\tilde{F}_{ij}\right] + \sum_{i=1}^{k+1} \sum_{j=1}^{k+1} \left(\widehat{M}_{ij} - \widetilde{M}_{ij}\right)^2 R\left[\widetilde{M}_{ij}\right] \end{aligned} \tag{7}$$

where $\hat{P}^{(t+1)}$, $\hat{P}^{t}$, $\hat{B}$, $\widehat{D}$, $\hat{F}$ and $\widehat{M}$ are the final (i.e. adjusted and balanced) estimates we are looking for. Function L is simply the (squared) weighted Euclidean distance between the vectors of raw and balanced estimates of stocks and flows.

Note that the objective function (7) involves both gross and net migrations flows, and both gross and net natural flows, as they are all very significant demographic statistics that we would like to modify the least during the balancing procedure.

Therefore, the constrained optimization problem we propose to solve is the following:

$$\begin{cases} Argmin\ L\left(\widehat{P}^{(t+1)}, \widehat{P}^{t}, \widehat{B}, \widehat{D}, \widehat{N}, \widehat{F}, \widehat{M}\right) & \\ \text{subject to:} & \\ \widehat{P}_i^{(t+1)} = \widehat{P}_i^{(t)} + \widehat{N}_i + \sum_{j=1}^{k+1} \widehat{M}_{ij} & \text{for } i = 1, \dots, k \\ \widehat{N}_i = \widehat{B}_i - \widehat{D}_i & \text{for } i = 1, \dots, k \\ \widehat{M}_{ij} = \widehat{F}_{ji} - \widehat{F}_{ij} & \text{for } i, j = 1, \dots, k+1 \end{cases} \qquad (8)$$

The constraints acting on problem (8) are, of course, the area-level DBEs, plus structural constraints expressing the relation between births, deaths and natural increase, and the antisymmetry of the Net Migration Matrix. The solution of problem (8) results in time and space consistent estimates of population counts, natural flows, and migration flows.

Problem (8) involves $2(\mathrm{k}+1)^2 + 3\mathrm{k}$ unknowns and $(\mathrm{k}+1)^2 + \mathrm{k}$ linear constraints. If we were to consider as "regions" the partitions determined by cross-classifying $\text{'NUTS 3'} * \text{'sex'} * \text{'5 years age classes'}$, we would need to handle approximately $35{,}000{,}000$ unknowns. For problems of this size the closed form solution proposed in (Stone et al., 1942), which is essentially derived from the generalized least squares method, is so computationally demanding that it cannot be applied in practice. As a viable alternative, an iterative constrained optimization approach is proposed in (Byron, 1978), which exploits the Conjugate Gradient algorithm. The iterative Conjugate Gradient algorithm is indeed computationally very efficient and proved a perfect fit for the stocks and flows reconciliation task (8). To fully automate the solution of this task, we implemented a dedicated software system, based on R (R Core Team, 2018).

Coming to the statistical properties of the balanced (i.e. final) estimates of population stocks and flows, (Theil, 1961) has shown that they are BLUE if:

(1) *Errors* affecting raw (i.e. initial) estimates are *uncorrelated* and have *zero mean*;

(2) *Reliability weights* are equal to *inverse variances* of raw estimates.

When the above assumptions do not hold, e.g. because raw estimates are *biased* or reliability weights are *misspecified*, the general properties of balanced estimates are no longer under theoretical control. Yet, of course, they can still be investigated through Monte Carlo simulations. This task has been tackled in (Di Zio, Fortini, Zardetto, 2018). Experiments on real Italian data reported therein suggest that, under reasonable assumptions, the proposed approach determines improved estimates of population counts: *besides gaining consistency, they exhibit lower bias and variance as compared to raw ones, and the observed accuracy gain seems robust against misspecification of reliability weights*.

NSIs need to publish population counts by domains that cross territory with 'sex', 'age class', 'nationality' and so on. The macro-integration method described here can produce consistent estimates in this context as well. Indeed, when covariates like 'sex', 'age class' and 'nationality' are introduced, we can still write down *generalized DBEs* constraining cell counts of the corresponding N-way classification at subsequent times t and $t + 1$. However, these covariates bring into play:

(i) a *more abstract notion of migration flows*, e.g. people can "migrate" from a given 'age class' to the subsequent one or from one 'nationality' to another;

(ii) *new structural constraints* (i.e. "illicit migrations"), e.g. since people cannot get younger, they can only get stuck in their original 'age class', move to the next one, or die[2].

Fortunately, we can leverage *reliability weights* in equation (7) to prevent "illicit migrations" from being generated within the balanced solution. In fact, since illicit cells have 0 *raw counts*, all we have to do is to let $R[\cdot] \to \infty$ and the corresponding *balanced counts* will still be 0.

Section 3 – Downstream Effects on the Base Register of Individuals

In practice, *raw* estimates of population stocks in equations (7) and (8):

$$\begin{cases} \tilde{P}^{(t)} = \left(\tilde{P}_1^{(t)}, \dots, \tilde{P}_k^{(t)}\right)' \\ \tilde{P}^{(t+1)} = \left(\tilde{P}_1^{(t+1)}, \dots, \tilde{P}_k^{(t+1)}\right)' \end{cases} \tag{9}$$

which must be fed as input to the balancing procedure, will be computed from released BRI versions referred to times t and $t + 1$. Each released version of the BRI – referred to a generic time τ – will contain under-coverage and over-coverage corrected weights associated to *individual* records:

$$d_q^{(\tau)} \qquad \mathrm{q} = 1, \dots, \mathrm{N}_{\mathrm{BRI}}^{(\tau)} \tag{10}$$

and raw population counts for each balancing cell (=subpopulation) $\mathrm{U_i}$ will be obtained by simply adding the weights of BRI individuals belonging to the cell:

$$\tilde{P}_i^{(\tau)} = \sum\nolimits_{q \in U_i} d_q^{(\tau)} \qquad \mathrm{i} = 1, \dots, \mathrm{k} \tag{11}$$

Once the balancing procedure has been successfully executed, output *balanced* estimates of population stocks will be available. These balanced estimates can easily be exploited to *adjust* individual weights of the BRI in such a way that estimates of population counts derived from *adjusted BRI weights* fulfill all the DBEs in equation (8):

[2] Note that structural constraints arising from variable 'age class' can actually be greatly simplified by trading variable 'age class' for variable 'class of cohort'. When using cohorts the only residual constraint is that the modality of variable 'class of cohort' cannot change between t and t + 1.

$$\begin{cases} d_q^{(\tau)} \xrightarrow{\text{BALANCING}} w_q^{(\tau)} & \text{q = 1, ... , } \mathrm{N}_{\mathrm{BRI}}^{(\tau)} \\ \sum_{q \in U_i} w_q^{(\tau)} = \widehat{P}_i^{(\tau)} & \text{i = 1, ... , k} \end{cases} \qquad \text{such that:} \tag{12}$$

In what follows, we will show how this appealing result can be achieved in practice.

Let us start with a very important distinction between population stocks referred to times t and t + 1, reported in (9) and involved in equations (7) and (8).

(1) At the time the balancing procedure takes place $\mathrm{t}^{\mathrm{BAL}} > \mathrm{t} + 1$, *Istat will have already disseminated to the external audience official population counts referred to time* t. Therefore:

(i) The balancing procedure taking place at time $\mathrm{t}^{\mathrm{BAL}} > \mathrm{t} + 1$ will be performed in such a way that these official population estimates will *not* be altered:

$$\widehat{P}_i^{(t)} \equiv \widetilde{P}_i^{(t)} \qquad \mathrm{i} = 1, \ \dots \ , \mathrm{k} \tag{13}$$

which will simply be obtained by letting $R\left[\widetilde{P}_i^{(t)}\right] \to \infty$.

Note that this will allow Istat to produce balanced estimates on a yearly basis without the need to revise ever again any already disseminated official population counts.

(ii) We can assume BRI weights referred to time t to have *already been adjusted* by a balancing procedure run *the year before* and involving stocks and flows referred to times t – 1 and t.

(2) At the time the balancing procedure takes place $\mathrm{t}^{\mathrm{BAL}} > \mathrm{t} + 1$, Istat will have *not* disseminated official population counts referred to time t + 1 *yet*, however *a released version of the BRI referred to time* t + 1 *will be available, along with under-coverage and over-coverage corrected (but still raw) individual weights*. Therefore:

(iii) After successfully executing the balancing procedure, we will use its outputs to *adjust* individual weights of the BRI referred to time t + 1 in such a way that estimates of population counts derived from the BRI will henceforth be consistent and fulfill all the DBEs:

$$d_q^{(t+1)} \xrightarrow{\text{BALANCING}} w_q^{(t+1)} \qquad \mathrm{q} = 1, \ \dots \ , \mathrm{N}_{\mathrm{BRI}}^{(t+1)} \tag{14}$$

(iv) Istat will use the *post-balancing-adjusted individual weights* $w_q^{(t+1)}$ in (14) to compute *official estimates of population counts for arbitrary estimation domains* $\mathrm{D_e}$:

$$\widehat{P}_e^{(t+1)} = \sum_{q \in D_e} w_q^{(t+1)} \qquad \mathrm{e} = 1, \ \dots \ , \mathrm{E} \tag{15}$$

The way to obtain *post-balancing-adjusted* individual weights $w_q^{(t+1)}$ appearing in equation (14) is straightforward. It will only take to post-stratify raw individual weights $d_q^{(t+1)}$ using the balanced population estimates $\widehat{P}_i^{(t+1)}$ as calibration benchmarks:

$$w_q^{(t+1)} = d_q^{(t+1)} \left[\frac{\widehat{P}_i^{(t+1)}}{\widetilde{P}_i^{(t+1)}} \right] = d_q^{(t+1)} \left[\frac{\widehat{P}_i^{(t+1)}}{\sum_{q \in U_i} d_q^{(t+1)}} \right] \qquad \forall\, q \in U_i \tag{16}$$

$$\text{for} \quad i = 1, \dots, k \quad \text{and} \quad q = 1, \dots, \mathrm{N}_{\mathrm{BRI}}^{(t+1)}$$

Despite equation (16) is formally the solution of a calibration problem (which uses the unbounded linear distance, one single auxiliary variable whose values are identically equal to 1, and calibration domains identified by the $\mathrm{U_i}$ cells of the balancing partition), its expression is so simple that a trivial PL/SQL Data Base procedure will be enough to accordingly update BRI weights in a real production setting.

Now suppose that under-coverage and over-coverage corrected weights (10) associated to individual records of the BRI were constructed in such a way that *all the individuals belonging to the same household share the same weight*. This property would ensure consistent estimates of individual-level and household-level aggregates computed from the BRI[3]. Should this be the case, it would be desirable that *post-balancing-adjusted* individual weights (14) retain that same property. In order to achieve such a goal, the simple formula (16) would *not* in general be suitable. More specifically, formula (16) would produce constant individual weights within households *only if* the balancing cells (=subpopulations) $\mathrm{U_i}$ *do not* cut-across households; of course, this would *not* be the case whenever individual-level variables like ‘sex’ or ‘age class’ are involved as classification variables in the balancing procedure.

For arbitrary settings – i.e. whatever might be the choice of balancing cells (=subpopulations) $\mathrm{U_i}$ – the goal of obtaining *post-balancing-adjusted* individual weights that are *constant within households* can be easily achieved by means of standard calibration software available in Istat, e.g. ReGenesees[4] (Zardetto, 2015). Despite the size of RBI (~$6{\cdot}10^7$ rows) may seem at first sight prohibitively large when compared to typical survey samples for which calibration procedures are routinely carried out in Istat, actually this will not pose any serious computational or technical issue. In general, however, the mathematical relation between weights $w_q^{(t+1)}$ and $d_q^{(t+1)}$ will be more complex than (16) and not expressable in analytic closed-form.

[3] For instance, the estimated number of people living in households of 4 members would be equal to 4 times the estimated number of households of 4 members.

[4] Indeed ReGenesees provides calibration facilities that allow cluster-level weights adjustments.

References

Stone, R., Champernowne, D.G., and Meade, J.E. (1942), The Precision of National Income Estimates, *Review of Economic Studies,* vol. 9 (2), pp. 111-125.

Theil, H. (1961), Economic Forecasts and Policy, North Holland Publishing Company.

Byron, R. (1978), The Estimation of Large Social Account Matrices, *Journal of the Royal Statistical Society A*, vol. 141(3), pp. 359-367.

Zardetto, D. (2015), ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys, in *Journal of Official Statistics*, Vol. 31, n. 2, 2015, pag. 177-203, De Gruyter, DOI 10.1515/JOS-2015-0013.

Di Zio, M., Fortini, M., Zardetto, D. (2018), Reconciling Estimates of Demographic Stocks and Flows through Balancing Methods, in *Proceedings of the 9th European Conference on Quality in Official Statistics (Q2018)*, Krakow, Poland, 26-29 June 2018, pag. 1-10. Available at: https://www.q2018.pl/wp-content/uploads/Sessions/Session%2025/Diego%20Zardetto/Session%2025_Diego%20Zardetto.docx

R Core Team (2021), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/