22/10/2018
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 will 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 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 (as we will see later, "regions" can actually be any population partition, e.g. territory∗sex∗age classes) at times t and $t+1$, as well as initial estimates of births, deaths and natural increase occurred for each region between time t and $t+1$:

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

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

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

[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 $(k+1)^{th}$ row and column of F represent migrations from and to any territory *outside* the nation, thus $k+1$ 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 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, $M_{ij} = F_{ji} - F_{ij}$:

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

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 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:

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

where $\hat{P}^{(t+1)}$, $\hat{P}^t$, $\hat{B}$, $\hat{D}$, $\hat{F}$ and $\hat{M}$ 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:

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

The constraints acting on problem (8) are, of course, the area-level DBEs, plus structural constraints expressing the relation 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 'NUTS 3' $*$ 'sex' $*$ '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 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).

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 $\mathrm{t}+1$. 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].

[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 $\mathrm{t}+1$.

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.

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.

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

R Core Team (2018), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

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