BALANCING METHODS FOR ENSURING TIME AND SPACE CONSISTENCY OF DEMOGRAPHIC ESTIMATES IN THE ITALIAN INTEGRATED SYSTEM OF STATISTICAL REGISTERS

Marco Di Zio[1], Marco Fortini[1] and Diego Zardetto[1]

Keywords: Demographic Balancing Equation, Data Integration, Data Reconciliation, Statistical Registers.

1. INTRODUCTION: MOTIVATION OF THE WORK AND PROPOSED METHODS

The Italian National Institute of Statistics (Istat) is currently investing resources for changing in depth its production processes, striving to overcome its traditional “stovepipe” production model based on the *vertical* integration of different survey-specific tasks. As a result of this modernisation effort, the new Istat production model is expected to rely, instead, on a *horizontal* composition of tasks, e.g. by integrating as much as possible administrative data and survey data concerning related topics. The backbone of the envisioned production system will be 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 storing data gathered from disparate sources about people usually residing in Italy.

One of the most important outputs of this new statistical production system is concerned with the population census. In the near future, the Italian population census will no longer be a complete enumeration survey, but rather result from the integration of administrative and survey data. In this respect, official population size estimates are expected to be delivered more frequently than it happened before through traditional censuses.

These estimates should be consistent with the available information about demographic events. In particular, population size estimation at different reference times should fulfil the demographic balancing equation (DBE), which states that the final population counts are equal to the starting population counts plus the sum of natural increase and net migration:

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

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

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

Each component of the DBE will be estimated independently; in particular birth, death and migration figures will be obtained from administrative data released by municipal civil registries, while population size estimates at subsequent reference times are planned to be derived from the BRI, thereby hinging upon integrated administrative data and sample survey data.

[1] Istat – Directorate for Methodology and Statistical Process Design

Taking into account sampling and non-sampling errors affecting all the involved data, the DBE will *not* be trivially satisfied. Therefore, suitable methods must be investigated in order to obtain consistent final estimates. These methods should simultaneously adjust both (i) the initial estimates of population sizes and (ii) the rough civil registry figures, in such a way that the resulting data exactly fulfil the DBE.

It is worthwhile to stress that one can leverage the DBE to jointly enforce (i) the *time consistency* of estimated population counts referred to subsequent points in time, as well as (ii) the *space consistency* between natural increase figures, net migration figures and population size estimates referred to different geographic areas.

As for the time consistency goal, the reference dates of any two subsequent production-stable releases of the BRI seem natural candidates to play the role of (t) and (t + 1) within the DBE. As for the space dimension, intuitively it would be desirable to leverage the DBE to achieve consistency between natural increases, net migrations and population size estimates at the *finest* possible territorial level. Nevertheless, the computational complexity of the required adjustments is, of course, expected to *increase* with the cardinality of the adopted territorial classification. At the moment, we guess that NUTS 3 regions (i.e. provinces) could be a good trade-off for Italy. Moreover, we are studying the technical feasibility of simultaneously enforcing the DBE for a nested hierarchy of geographical areas.

In order to solve this problem, we propose to use methods which are commonly adopted inside National Statistical Institutes (NSI) for balancing large systems of national accounts. Indeed, the National Accounts divisions of most NSIs routinely use independent initial estimates that are characterised by different degrees of reliability and have to be adjusted in order to satisfy a large set of accounting identities. An important reference paper about balancing problems in National Accounts is [1]. In this paper, the authors explicitly recognize the impact of measurement errors on initial estimates, and suggest the idea that less reliable initial estimates should undergo larger adjustments. Unfortunately, the closed form solution proposed in [1], essentially derived from the generalized least-squares method, is so computationally demanding that it cannot be applied to any large-scale balancing problem of practical interest. As a viable alternative, an iterative constrained optimization approach is proposed in [2], which exploits the conjugate gradient algorithm. This approach is computationally efficient, even for very large matrices, and is currently adopted as a standard inside the National Accounts division of Istat (see [3] and references therein).

The contribution of this work is threefold. First, we formalize the problem of ensuring the time and space consistency of demographic estimates as a constrained optimization problem (see Section 2 for an illustrative example). Second, we study how to solve the problem along the lines of [1] and [2], by suitably restating the models and algorithms introduced in those classical papers. Third, we offer an empirical evaluation of our approach on simulated and real demographic data, using a dedicated software prototype developed in R [4].

2. The constrained optimization approach

We formulate the problem of finding consistent demographic estimates as a constrained optimization task. 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 which are *balanced*, i.e. satisfy all the DBEs, and 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 DBEs defined for all the areas. 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 ("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 the natural increase occurred for each region between time *t* and *t* + 1:

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

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:

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

Note that: (i) the (*k* + 1)th row and column of *F* represent migrations from and to any territory *outside* the nation, and (ii) 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*, $M_{\text{ij}} = F_{\text{ji}} - F_{\text{ij}}$:

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

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

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

Furthermore, let us assume we can attach to each *atomic* initial estimate involved in (3) (4) and (5) a measure of *reliability*, *R*. 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 notational convenience, let us specify the reliability measures R of objects (3) (4) and (5) through their reciprocals, which in turn could be regarded as measures of *alterability*, *A*. For instance, we will indicate the reliability measure $R(m_{ij})$ of a generic element m_{ij} of the net migration matrix *M* as:

$$R(m_{ij}) = \frac{1}{A_{ij}^{[M]}} \qquad (7)$$

Note that any zero values occurring within $A^{[\cdot]}$ will signal *absolute reliability* (as $R(\cdot) \to \infty$) and thus *prevent* the corresponding initial atomic estimates from being altered.

Lastly, let us denote the raw estimates with a tilde ($\tilde{\cdot}$), the balanced estimates with a circumflex hat ($\widehat{\cdot}$), and the true values with no hat.

Given (3), (4), (5) and (7), we define the objective function, *L*, for our constrained optimization problem as follows:

$$\begin{aligned} L\left(\hat{P}^{(t+1)}, \hat{P}^{t}, \widehat{N}, \widehat{F}, \widehat{M}\right) \\ &= \sum_{i=1}^{k} \frac{\left(\hat{P}_i^{(t+1)} - \tilde{P}_i^{(t+1)}\right)^2}{A_i^{[P^{(t+1)}]}} \\ &+ \sum_{i=1}^{k} \frac{\left(\hat{P}_i^{(t)} - \tilde{P}_i^{(t)}\right)^2}{A_i^{[P^{(t)}]}} \\ &+ \sum_{i=1}^{k+1} \sum_{j=1}^{k+1} \frac{\left(\widehat{M}_{ij} - \widetilde{M}_{ij}\right)^2}{A_{ij}^{[M]}} \\ &+ \sum_{i=1}^{k+1} \sum_{j=1}^{k+1} \frac{\left(\widehat{F}_{ij} - \widetilde{F}_{ij}\right)^2}{A_{ij}^{[F]}} \end{aligned} \qquad (8)$$

where $\hat{P}^{(t+1)}$, $\hat{P}^{t}$, $\widehat{N}$ and $\widehat{F}$ are the *final* (i.e. *adjusted* and *balanced*) estimates we are looking for.

Therefore, the constrained optimization problem we propose to solve along the lines of [1] and [2] is the following:

$$\left.\begin{aligned}
&Argmin\; L\big(\hat{P}^{(t+1)}, \hat{P}^{t}, \hat{N}, \hat{F}, \hat{M}\big) \\
&\text{subject to:} \\
&\hat{P}_i^{(t+1)} = \hat{P}_i^{(t)} + \hat{N}_i + \textstyle\sum_{j=1}^{k+1} \hat{M}_{ij} && \text{for } i = 1, \dots, k \\
&\hat{M}_{ij} = \hat{F}_{ji} - \hat{F}_{ij} && \text{for } i, j = 1, \dots, k+1
\end{aligned}\right\} \quad (9)$$

The solution of problem (9) results in time and space consistent estimates of population size, natural increase and migration.

Problem (9) involves $2(k + 1)^2 + 3k$ unknowns and $(k + 1)^2 + k$ 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. Hence, as anticipated, Stone's closed form solution is computationally infeasible, and our R prototype uses instead a dedicated implementation of iterative Conjugate Gradient algorithm.

It is worthwhile to investigate the statistical properties of the balanced (final) estimates of population stocks and flows. As shown in [5], they are BLUE if:

(1) Errors affecting raw (=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, the general properties of balanced estimates are no longer under theoretical control.

3. SIMULATION STUDY

In this section we show the results of a simulation study aiming at obtaining empirical evidences on the behavior of balanced estimates in a setting different form the one previously introduced.

First we assume that *Natural Increase* and the *Population counts* at time t are known without errors, i.e., $\tilde{N} = N$ and $\tilde{P}^t = P^t$. Then we start with official demographic figures (P^t, N, F) of administrative Italian regions (NUTS 2) in 2015, so that K = 20, P^t is obtained via the BDE. These figures are as ground-truth and we perturb them to generate raw estimates. The count estimates $\tilde{P}^{t+1}$ are obtained by adding a Gaussian noise with a given relative bias β and coefficient of variation α to P^{t+1}, i.e.,

$$\tilde{P}_i^{t+1} = \mathrm{N}\left((1+\beta)P_i^{t+1},\, (\alpha P_i^{t+1})^2\right)$$

The perturbed migration flows $\tilde{F}$ are generated from a negative binomial distribution centered around F with a given relative bias γ and dispersion parameter δ, i.e,

$$\tilde{F}_{ij} = \mathrm{NB}\left(\mu = (1+\gamma)F_{ij},\ v = \mu + \delta\mu^2\right)$$

The perturbed migration matrix $\widetilde{M}$ is derived as $\widetilde{M} = \tilde{F}' - \tilde{F}$.

Since the population counts and the natural increases are known, their reliability measure is set to infinite (they will not be changed). For the other variables we assume a simple "non-informative" model, i.e., weights equal to the inverse of the observed value, $R(\tilde{\cdot}\ ) = 1/R(\tilde{\cdot})$. The balanced estimates $\left(\hat{P}^t, \hat{F}, \widehat{M}\right)$ are obtained by solving the system (7).

Those steps are repeated S times (S=5000) and raw estimates, balanced estimates and ground-truth figures are compared. Indicators evaluating the relative bias (*RB*) and the relative root mean square error (*RRMSE*) are computed for each region by averaging the differences over the S simulations. Then global accuracy measures are obtained by averaging *RB* and *RRMSE* values over the k regions, namely: the mean absolute relative bias (*MARB*) and the mean relative root mean squared error (*MRRMSE*). For instance, for the parameter $\tilde{P}^{t+1}$ we have

$$RB_i = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{P}_i^{t+1^{(s)}} - P_i^{t+1}}{P_i^{t+1}}\right) \quad , \quad RRMSE_i = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{P}_i^{t+1^{(s)}} - P_i^{t+1}}{P_i^{t+1}}\right)^2}$$

and

$$MARB = \frac{1}{k}\sum_{i=1}^{k}\left|RB_i\right| \quad , \quad MRRMSE = \frac{1}{k}\sum_{i=1}^{k} RRMSE_i$$

We have studied 5 different simulation scenarios (see Table 1) and explored 10 combinations of simulation parameters (see Table 2). The results expressed in terms of MARB(%) and RRMSE(%) are reported in Table 3.

Table 1. Simulation scenarios

S1	No Bias
S2	Only Migration Bias
S3	Both P1 and Migration Biases
S4	Overdispersed Migrations
S5	High Bias - High Variance

Table 2. Simulation parameters

	P1 Raw		Raw Migration Figures			
	RBias (%)	CV (%)	Matrix	Rbias (%)	Disp (%)	Avg|CV| (%)
P1	0	10	F	0	0	8
P2	0	10	M	0	0	15
P3	0	10	F	-50	0	11
P4	0	10	M	-50	0	21
P5	-5	10	M	-50	0	21
P6	5	10	M	-50	0	21
P7	-5	10	F	-50	20	47
P8	-5	10	M	-50	20	53
P9	-10	20	F	-50	20	47
P10	-10	20	M	-50	20	53

Table 3 MARB and MRRMSE

	P1 MARB (%)			P1 MRRMSE (%)		
	Bal	Raw	Bal/Raw	Bal	Raw	Bal/Raw
P1	0.0	0.1	-	0.0	10.0	0.2
P2	0.0	0.1	-	0.0	10.0	0.2
P3	0.1	0.1	-	0.1	10.0	1.1
P4	0.1	0.1	-	0.1	10.0	1.1
P5	0.1	5.0	2.1	0.1	11.2	1.0
P6	0.1	5.0	2.1	0.1	11.2	1.0
P7	0.1	5.0	2.1	0.2	11.2	1.6
P8	0.1	5.0	2.1	0.1	11.2	1.0
P9	0.1	10.0	1.1	0.2	22.4	0.8
P10	0.1	10.0	1.1	0.1	22.4	0.5

Despite the substantial bias inside raw estimates of $\tilde{P}^{t+1}$ and $\tilde{P}^{t}$, balanced estimates are almost unbiased: balancing removes around 98% of the original $\tilde{P}^{t+1}$ bias.

In all scenarios, balancing also dramatically increases the efficiency of $\tilde{P}^{t+1}$ estimates: the MSE of balanced estimates is around 1% of raw estimates' one.

4. Conclusions and Future works

Balancing methods can jointly ensure *time* and *space* consistency of estimated population counts and demographic figures (e.g. migrations). This promotes credibility in published statistics, thus enhancing the reputation of the NSI.

The empirical results, under realistic assumptions, suggest that balancing results in improved estimates of population counts:

- ✓ Balanced estimates exhibit lower bias and variance
- ✓ Efficiency gains seem robust against misspecification of reliability weights

Next studies will be dedicated to analyse other scenarios, and to investigate the effects of balancing on the estimates of migration flows.

Finally, we will perform a thorough time-series analysis of estimated population counts obtained by *repeatedly* solving balancing problems in a chain

$$\boldsymbol{P}^{t_s} \xrightarrow{\text{BALANCE}} \widehat{\boldsymbol{P}}^{t} \xrightarrow{\text{BALANCE}} \widehat{\boldsymbol{P}}^{t+1} \xrightarrow{\text{BALANCE}} \cdots \xrightarrow{\text{BALANCE}} \widehat{\boldsymbol{P}}^{t+n}$$

to analyse the behavior of a sequence of balanced estimates.

5. Issues for discussion

- ✓ What is your general opinion about the proposed approach?
- ✓ Are there any alternative methods used in NSIs for dealing with this problem?
- ✓ It would be desirable to compute a measure of accuracy of the estimates obtained according to the balancing method. Is there any result to be used to this aim?
- ✓ The method modifies initial estimates to fulfil the BDE. Is it possible to identify hypotheses under which the final estimates are more accurate than the initial ones?

References

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

[2] Byron, R., The estimation of large social account matrices, *Journal of the Royal Statistical Society A* (1978), vol. 141(3), pp. 359-367.

[3] Eurostat Leadership group SAM, *Handbook on Social Accounting Matrices and Labour Accounts* (2003), Luxembourg: Eurostat Secretariat Unit E3, European Commission.

[4] R Core Team, *R: A language and environment for statistical computing* (2016), R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[5] Theil, H., *Economic Forecasts and Policy* (1961), North Holland Publishing Company.