Reconciling Estimates of Demographic Stocks and Flows through Balancing Methods

Marco Di Zio, Istat – The Italian National Institute of Statistics, dizio@istat.it

Marco Fortini, Istat – The Italian National Institute of Statistics, fortini@istat.it

Diego Zardetto, Istat – The Italian National Institute of Statistics, zardetto@istat.it

Abstract

In the near future, the Italian population census will be the result of the integration of administrative and survey data. This will enable Istat to deliver official population size estimates more frequently than it happened before through traditional censuses. Census-based estimates of population counts ('stocks') should be consistent with information about demographic events ('flows') available from municipal civil registries. In particular, the Demographic Balancing Equation (DBE) should be fulfilled, which states that final population counts $P^{(t+1)}$ must be equal to the starting population counts $P^{(t)}$ plus the sum of natural increase N (i.e. births minus deaths) and net migration M (i.e. immigrants minus emigrants): $P^{(t+1)} = P^{(t)} + N + M$. In practice, due to sampling and non-sampling errors, the DBE will not be trivially satisfied. Therefore, suitable methods must be investigated to obtain consistent final estimates. These methods should simultaneously adjust both the initial estimates of population counts and the rough civil registry figures, in such a way that the resulting macrodata exactly fulfill the DBE. In addition, more reliable initial estimates should possibly undergo smaller adjustments. We formalize the problem of ensuring time and space consistency of demographic estimates as a constrained optimization problem. Given initial, rough estimates of stocks and flows entering the demographic balancing equations 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. To solve the problem, we propose to exploit methods that are commonly adopted for balancing large systems of national accounts. Experiments on real data 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 rough ones, and the observed efficiency gain seems robust against misspecification of reliability weights.

Keywords: Macro-integration, demographic balancing equation, data integration, administrative data, census estimates.

1. Introduction: Motivation of the Work and 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 and to integrate 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 concerns an in-depth rethinking of 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. This will enable Istat to deliver official population size estimates more frequently than it happened before through traditional censuses, very likely on a yearly basis.

These estimates should be consistent with available information about demographic events. In particular, population size estimates at different reference times should fulfill 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:

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

In practice, 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].

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

[1] Though a final decision has not yet been made, estimated population counts will likely arise from a Dual System Estimation approach, based on the linkage between the BRI (*first capture*) and a dedicated area sampling survey (*second capture*).

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 fulfill the DBE.

The DBE can be exploited 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 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.

In order to solve this problem, we propose to use methods that 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 characterized by different degrees of reliability and have to be adjusted to satisfy a large set of accounting identities. Within their seminal paper about balancing problems in National Accounts (Stone et al., 1942), 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 (Stone et al., 1942), essentially derived from the generalized least-squares method, is so computationally demanding that cannot be applied to any large-scale balancing problem of practical interest. As a viable alternative, an iterative constrained optimization approach is proposed in (Byron, 1978), 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 (Eurostat, 2003) 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. Second, we study how to solve the problem along the lines of

(Stone et al., 1942) and (Byron, 1978), 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 (R Core Team, 2018).

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 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 ("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 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, $\mathrm{M}_{\mathrm{ij}} = \mathrm{F}_{\mathrm{ji}} - \mathrm{F}_{\mathrm{ij}}$:

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

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. $\widetilde{M}_{ij}$), *balanced estimates* with a *circumflex hat* (e.g. $\widehat{M}_{ij}$), and - for later usage - *true values* with *no hat*, (e.g. M_{ij}).

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

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

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

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

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

The solution of problem (8) results in time and space consistent estimates of population size, natural increase and migration.

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. Hence, as anticipated, (Stone et al., 1942) closed form solution is computationally infeasible, and our R prototype uses instead a dedicated implementation of the iterative Conjugate Gradient algorithm.

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.

3. Simulation Study

In this section we discuss a simulation study designed to investigate the behavior of balanced estimates in a more general setting than the ideal one of (Theil, 1961).

We start with official demographic figures $(P^t,\ N,\ F,\ M)$ of administrative Italian regions (NUTS 2) in 2015, so that $\mathrm{k} = 20$. From these data, we compute P^{t+1} using the DBE: this way the set $(P^{t+1},\ P^t,\ N,\ F,\ M)$ *exactly* fulfills the DBE *by construction*. Then, we use such figures as *ground-truth* and perturb them to generate *raw estimates*. The simulation goes as follows:

A. We assume that the *Natural Increase* and the *Population counts* at time t are *known without errors*, i.e. $\tilde{N} = N$ and $\tilde{P}^t = P^t$. Therefore, to prevent them from being changed by the balancing algorithm, we set their *reliability weights* to infinite:

$$R[N_i] = R[\tilde{P}_i^t] \rightarrow \infty \tag{9}$$

B. We generate the vector of raw estimates of Population Counts $\tilde{P}^{t+1}$ by adding to P^{t+1} a *Gaussian noise* with a *relative bias* β and a *coefficient of variation* α:

$$\tilde{P}_i^{t+1} = \mathcal{N}\left((1+\beta)P_i^{t+1},\ (\alpha P_i^{t+1})^2\right) \tag{10}$$

C. We generate the perturbed Migration Flows Matrix $\tilde{F}$ from a *Negative Binomial* distribution centered around F with a *relative bias* γ and *dispersion parameter* δ:

$$\tilde{F}_{ij} = \mathcal{NB}\left(\mu = (1+\gamma)F_{ij},\ v = \mu + \delta\mu^2\right) \tag{11}$$

D. We derive the perturbed Net Migration Matrix $\widetilde{M}$ from $\tilde{F}$ as computed in (11):

$$\widetilde{M} = \tilde{F}^t - \tilde{F} \tag{12}$$

E. For the *reliability weights* of $\tilde{P}^{t+1}$, $\tilde{F}$ and $\widetilde{M}$, we deliberately assume a *naïve model*, setting their value to the reciprocal of the corresponding raw estimate:

$$R[\tilde{\cdot}] = 1/(\tilde{\cdot}) \tag{13}$$

F. Lastly, we compute *balanced estimates* $\left(\hat{P}^{t+1},\ \hat{F},\ \widehat{M}\right)$ by solving the constrained optimization problem (8).

We repeated all the steps above S times ($S = 5{,}000$) and compared the resulting Monte Carlo distributions of *raw estimates*, *balanced estimates* and *ground-truth figures*. For evaluation, we used standard global accuracy measures:

- **MARB** (Mean Absolute Relative Bias): the average over regions of absolute values of Monte Carlo estimated relative biases (see equations (14))
- **MRRMSE** (Mean Relative Root Mean Squared Error): the average over regions of absolute values of Monte Carlo estimated relative square roots of MSEs (see equations (15))

For instance, setting for notational convenience $t \overset{\text{def}}{=} 0$ and $t+1 \overset{\text{def}}{=} 1$, the accuracy measures for the *balanced population counts* $\widehat{P}_i^1$ have been computed as follows:

$$RB_i = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\widehat{P}_i^{1\,(s)} - P_i^1}{P_i^1}\right) \qquad MARB = \frac{1}{k}\sum_{i=1}^{k}|RB_i| \tag{14}$$

$$RRMSE_i = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\frac{\widehat{P}_i^{1\,(s)} - P_i^1}{P_i^1}\right)^2} \qquad MRRMSE = \frac{1}{k}\sum_{i=1}^{k} RRMSE_i \tag{15}$$

We have studied 5 different simulation scenarios: S1, …, S5. The main features of these scenarios are summarized in Table 1. Note that, to make the presentation of simulation results easier, in Table 1 different scenarios have been highlighted using different colors.

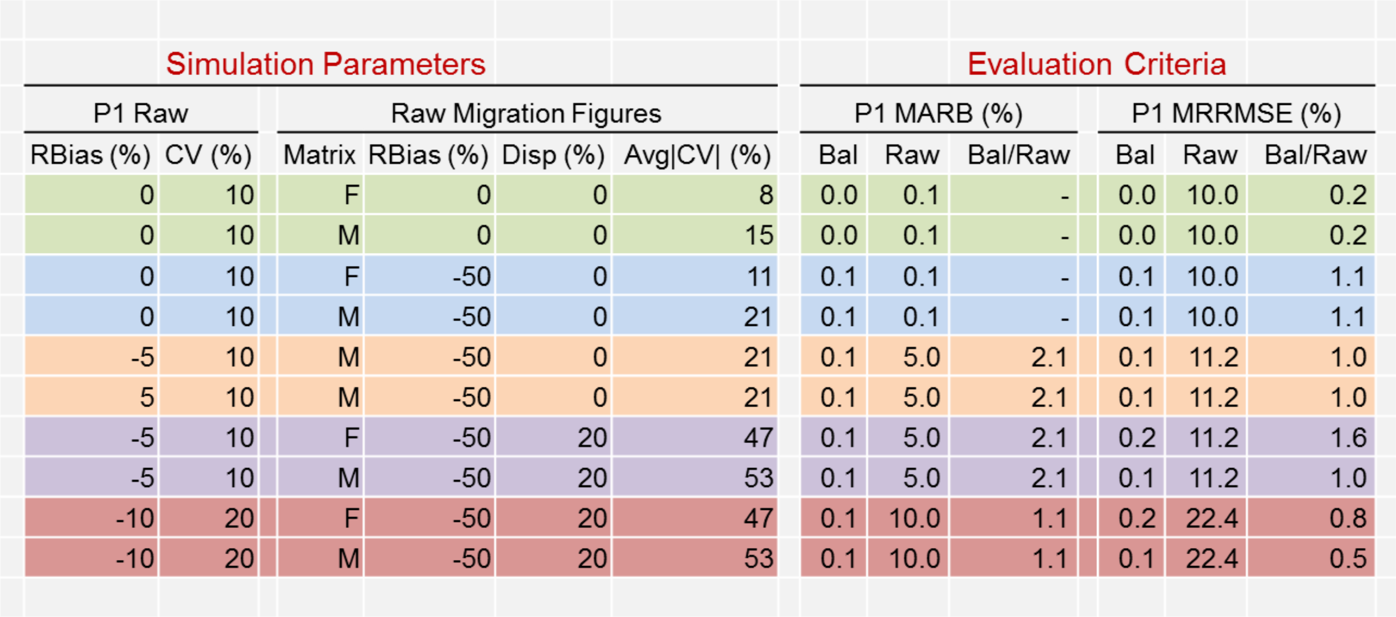
Table 1. The investigated simulation scenarios

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

Note also that simulation scenarios S1, …, S5 have to be intended as a *hierarchy*, in the sense that each scenario actually *adds* its main features to those characterizing the previous one (e.g. S4 switches on overdispersion in perturbed migration flows, but both migration flows an population counts $\widetilde{P}_i^1$ are already biased owing to S3).

Within each scenario, two combinations of the simulation parameters $(\beta,\ \alpha,\ \gamma,\ \delta)$ defined in (10) and (11) have been investigated. The simulation parameters and the corresponding simulation results expressed in terms of MARB(%) and RRMSE(%) for the *balanced population counts* $\widehat{P}_i^1$ are reported in Table 2. Note that the rows of Table 2 have been consistently highlighted with the same colors that have been used in Table 1 to differentiate the simulation scenarios. This makes it straightforward to visually link a given combination of simulation parameters of Table 2 to the simulation scenario it belongs to.

Table 2. Main results of the Monte Carlo simulation (5 scenarios, 10 combinations of simulation parameters, 5,000 runs for each combination)



Simulation Parameters						Evaluation Criteria					
P1 Raw		Raw Migration Figures				P1 MARB (%)			P1 MRRMSE (%)		
RBias (%)	CV (%)	Matrix	RBias (%)	Disp (%)	Avg|CV| (%)	Bal	Raw	Bal/Raw	Bal	Raw	Bal/Raw
0	10	F	0	0	8	0.0	0.1	-	0.0	10.0	0.2
0	10	M	0	0	15	0.0	0.1	-	0.0	10.0	0.2
0	10	F	-50	0	11	0.1	0.1	-	0.1	10.0	1.1
0	10	M	-50	0	21	0.1	0.1	-	0.1	10.0	1.1
-5	10	M	-50	0	21	0.1	5.0	2.1	0.1	11.2	1.0
5	10	M	-50	0	21	0.1	5.0	2.1	0.1	11.2	1.0
-5	10	F	-50	20	47	0.1	5.0	2.1	0.2	11.2	1.6
-5	10	M	-50	20	53	0.1	5.0	2.1	0.1	11.2	1.0
-10	20	F	-50	20	47	0.1	10.0	1.1	0.2	22.4	0.8
-10	20	M	-50	20	53	0.1	10.0	1.1	0.1	22.4	0.5

The left panel of Table 2 reports simulation parameters used to generate raw values of *population counts* ($\tilde{P}^1$) and *migration figures* (either $\tilde{F}$ or $\widetilde{M}$): ‘RBias’ columns indicate β and γ respectively, ‘Disp’ indicates δ, and ‘Avg|CV|’ indicates the average CVs (in absolute value) of migration figures resulting from a given choice of $(\gamma,\ \delta)$. The right panel of Table 2 reports the MARB(%) and RRMSE(%) for the *balanced* and the *raw estimates* of P^1 (columns ‘Bal’ and ‘Raw’ respectively), as well as the ratios of the corresponding accuracy measures (column ‘Bal/Raw’). The main results of Table 2 can be summarized as follows:

- Despite we injected *substantial bias* inside *raw estimates* of population counts ($\tilde{P}^1$) and migration figures ($\tilde{F}$ and $\widetilde{M}$), *balanced estimates* of population counts ($\hat{P}^1$) *are always nearly unbiased*: balancing removed *at least* 98% of the original bias.
- In all simulation scenarios, *balancing dramatically increased the efficiency of P^1 estimates*: the MSE of balanced estimates $\hat{P}^1$ is only ~1% of raw estimates’ one.

4. Ongoing Work and Conclusions

In this paper we showed that balancing methods can ensure *consistency* between estimated population counts and demographic figures derived from municipal civil registries (births, deaths and migrations). Consistency promotes credibility in published statistics, thus enhancing the reputation of the NSI. However, our balancing approach also determines *improved estimates of population counts*:

besides gaining consistency, they exhibit *lower bias and variance* as compared to rough ones, and the *accuracy* gain seems robust against misspecification of reliability weights. Simulation evidence has been obtained under two fundamental assumptions (see point A of Section 2): (1) errors affecting civil registry figures of births and deaths are negligible; (2) high-quality estimates of population counts at time $\mathrm{t} = 0$ are available. Condition (1) is surely realistic in Italy. Condition (2) has far reaching practical consequences. Indeed, as shown in our simulation, unbiased estimates $\tilde{P}^t$ induce balanced estimates $\hat{P}^{t+1}$ which are *still* nearly unbiased. This would allow Istat to produce balanced estimates on a yearly basis without need to revise ever again any already disseminated estimates:

$$\tilde{P}^0 \xrightarrow{\text{balance}} \hat{P}^1 \xrightarrow{\text{balance}} \cdots \xrightarrow{\text{balance}} \hat{P}^n \tag{16}$$

Let us conclude with a hint at ongoing work. NSIs need to publish population counts by domains that cross territory with ‘sex’, ‘age classes’, and so on. Our method can produce consistent estimates in this context as well. When covariates like ‘sex’ and ‘age classes’ 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; (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. Fortunately, we can leverage *reliability weights* to prevent “illicit migrations” from being generated within the balanced solution. As illicit cells have 0 *raw counts*, we must just to let $R[\cdot] \to \infty$, and the corresponding *balanced counts* will still be 0.

5. References

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

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

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

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.