Methodology for the Estimation of Annual Population Stocks by Citizenship Group, Age and Sex in the EU and EFTA Countries

Jakub Bijak and Dorota Kupiszewska
Central European Forum for Migration and Population Research (CEFMR)
ul. Twarda 51/55, 00-818 Warsaw, Poland
E-mail: j.bijak@cefmr.pan.pl, d.kupisz@cefmr.pan.pl
www.cefmr.pan.pl

Keywords: population estimates, stocks by citizenship, Europe, missing data, cohort-wise methods, fitting methods

Received: March 9, 2008

The paper addresses selected computational issues related to the challenge of dealing with poor statistics on international migration. Partial results of the on-going Eurostat-funded project on "Modelling of statistical data on migration and migrant population" (MIMOSA) are presented. The focus is on the data on population stocks by broad group of citizenship, sex and age. After a brief overview of the main problems with data on population by citizenship for 31 European countries (27 European Union countries, Iceland, Liechtenstein, Norway and Switzerland), a range of computational methods is proposed including cohort-wise interpolation, cohort-component projections, cohort-wise weights propagation and proportional fitting methods, as well as other, auxiliary methods. The algorithm for choosing the best method for estimating missing data on population stock by broad citizenship (nationals, foreigners – EU27 citizens, foreigners – non EU27 citizens), five-year age group (up to 85+) and sex on 1st January 2002–2006 is proposed and illustrated by examples of its application for selected countries.

Povzetek: Opisane so različne metode za ocenjevanje demografskih podatkov.

1 Introduction

The deficiencies of statistical information on migration-related variables, such as population flows or stocks, are well-known and widely discussed in the literature [1, 7]. The aim of the paper is to contribute to the works on dealing with these shortcomings and to propose a set of computational methods, as well as an algorithmic procedure of selecting the best one, for the estimation of population stocks as of 1st January in a breakdown by sex, age group and broad citizenship category, for the countries for which information is unavailable or incomplete.

The study was carried out within a Eurostat-funded project on "Modelling of statistical data on migration and migrant population" (MIMOSA). It covers 31 European countries, of which 27 belong to the European Union (as per 1st January 2007), and further four – to the EFTA (Iceland, Liechtenstein, Norway and Switzerland). The period of interest is 2002–2006. The citizenship groups under study are: nationals, European Union (EU27) foreigners and non-EU27 foreigners, while the age groups are five-year, with the last, open-ended category being 85 years or more.

Generally, the proposed estimation methods aim to combine data from different sources (population census, vital statistics, data on acquisition of citizenship, specific surveys, etc.). In principle, the data that are already available are not modified (for example, in order to harmonise definitions, or for any other reason), unless in the case of inconsistencies between the sources. In the latter cases, the demographic data, provided to Eurostat by national statistical institutes (NSIs), are given priority.

Apart from the Introduction, the paper is structured in four sections. Section 2 contains summary information on the availability and quality of the 2002–2006 data on population stocks for 31 countries under study. In Section 3, the proposed methodology for estimating population stocks by sex, age and citizenship groups is discussed. This section presents such tools as estimation of data in single years of age from five-year age-groups, cohort-wise interpolation of population stocks, cohort-component projections, cohort-wise propagation of weights, proportional fitting, as well as other, auxiliary methods. Subsequently, Section 4 contains recommendations concerning the procedure of selecting appropriate estimation methods for each of the countries under study, presented in the form of a decision algorithm and accompanied by several illustrative examples for selected countries. The discussion is summarised in Section 5.

The study is based on the data available in the Eurostat databases, supplemented by additional information obtained from national statistical institutes, whenever required and feasible. Throughout the paper, the abbreviation 'NSI' is used to denote the national statistical institute of the respective country, 'JMQ' stands for the Joint Questionnaire on International Migration Statistics (hereafter: Joint Migration Questionnaire) of Eurostat, UN Statistical Division, UN Economic Commission for

Europe, the Council of Europe and the International Labour Office. 'LFS' depicts the Labour Force Survey.

2 Availability of the 2002–2006 data on population stocks for 31 European countries

Annual statistics on usually resident population by citizenship, sex and age are collected by Eurostat from the NSIs via the Joint Questionnaire on International Migration, together with migration flow data. Population statistics for 37 European countries, collected through the JMQ are checked and subsequently loaded into Eurostat's on-line database, NewCronos. The data are located under the *Population and Social conditions* theme, in the *International Migration and Asylum* domain (MIGR), tables *migr_st_popctz* (population by sex and citizenship) and *migr_st_popage* (population by age group, citizenship and sex). The data for 2000–2006 come from the following tables in the 2000–2006 JMQs:

- Table 7a (for 2000–2003, 8a): Usually resident population by citizenship and age, both sexes;
- Table 7b (for 2000–2003, 8b): Usually resident population by citizenship and age, males;
- Table 7c (for 2000–2003, 8c): Usually resident population by citizenship and age, females.

A detailed analysis of statistics on population stocks by citizenship provided by the 31 countries covered the JMQs for the reference period 2002–2006. Selected results of the analysis of the data availability for particular countries are summarised in Table 1, providing an overview of the situation for all 31 countries. The information on the lack of data, marked as 'not available' in Table 1, was based on the information provided in the JMQ or on information obtained during the THESIM project[1]. Other missing data were marked as 'not provided to Eurostat'. In addition to missing data, a number of other problems were detected, for example the presence of provisional data, some citizenship categories only, broad age groups, or a different reference date than 1st January.

Data on total population stock on 1st January, not disaggregated by citizenship, are also collected by Eurostat within the framework of the Annual Demographic Statistics data collection. These data, disaggregated by sex and age, are located under the *Population and social conditions* theme, in the *Demography* (DEMO) domain of the database, table *demo_pjan*. The results of the review of the availability of these data for the years 2002–2006 revealed that the data on total population stock by sex and five-year age group (up to 85+) are available for all 31 countries, with the following exceptions: there is no 2006 data by age for Belgium and Italy, while for Romania the highest age group in 2004 data is 80+.

[1] Research project *THESIM: Towards Harmonised European Statistics on International Migration*, funded by the European Commission through the Sixth Framework Programme and executed by a research consortium led by Groupe d'étude de démographie appliquée (GéDAP), Université Catholique de Louvain.

In addition to the annual data, Eurostat also collects and disseminates statistics on population by citizenship, sex and age obtained by the countries during population censuses. Like other statistics, the census data are located under the *Population and social conditions* theme, in the *Census* (CENS) domain of the database, table *cens_nsctz*. Unlike annual population figures, the census data on population by citizenship, sex and age are available for almost all 31 countries, with the notable exceptions of the United Kingdom, Germany and Malta.

A supplementary source of data on population stock by citizenship is the Labour Force Survey. However, the availability of data from the LFS in the Eurostat database is very limited and the reliability of data is probably not high, due to the nature of the data source. By definition, the LFS statistics are estimates and thus bear certain errors, which can be relatively high for disaggregated categories (e.g., for population broken down by age, sex and citizenship groups). However, some use of the LFS data could be considered as an alternative to the proposed methods in the countries where data on total nationals and total foreigners are missing.

In the Eurostat database, the LFS tables are located under the *Population and social conditions* theme, the *Labour market* (LABOUR) domain, in the table with population data containing the nationality dimension (population by sex, age groups, nationality and labour status, table *lfsa_pganws*). However, the table does not contain data on the level of individual countries of citizenship and only data on total population and on nationals could be useful for this project. Estimates of the 2002–2006 stock of the EU27 citizens cannot be prepared using the LFS tables in the Eurostat database. These considerations need to be taken into account when proposing computation methods for the current study.

3 Proposed methods of estimating population stocks by citizenship, sex and age

The current section presents a theoretical background of the methods proposed for the calculations of the missing elements in population stocks by age, sex and citizenship group. After a brief summary of the notation, the following methods are discussed: interpolation of five-year into one-year age groups, regarded as a preparatory method (Section 3.2), followed by cohort-wise interpolation of population stocks (3.3), cohort-component projections, traditionally used in demography (3.4) and cohort-wise weights propagation (3.5). Further, Section 3.6 describes selected proportional fitting methods, which category encompasses three approaches, depending on the availability of information: the proportional adjustment, direct proportional fitting and iterative proportional fitting. Section 3 concludes by presenting some auxiliary methods for dealing with the Unknown categories, and for the estimation of missing elements of age distributions (3.7).

Country	2002	2003	2004	2005	2006
Austria	+	+	+	+	-
Belgium	+	x	-	-	-
Bulgaria	-	-	-	-	-
Cyprus	dref	-	-	-	-
Czech Republic	+	+	+	+	+
Denmark	+	+	+	+	+
Estonia	na	na	na	na	-
Finland	+	+	+	+	+
France	-	-	-	-	-
Germany	for, broad age, ±agesex , i	±age, i	for, i	i	p, i
Greece	-	-	i	for, ±sex	for
Hungary	for	for	+	+	+
Ireland	p, ±ctz, broad age, dref	p, ±ctz, broad age, dref	p, ±ctz, broad age, dref	p, ±ctz, broad age, dref	p, ±ctz, broad age, dref
Italy	dref	-	-	-age	-age
Latvia	-age	+	+	+	+
Lithuania	-	-ctz	-ctz	+	+
Luxembourg	-	-	tot, nat	±ctz, ±age, ±sex	±ctz, ±age, ±sex
Malta	-	-	-	-	-
Netherlands	+	+	+	+	+
Poland	dref	-	-	-ctz	-
Portugal	p, for, -age	p, for	-	-	-
Romania	dref	-	+	+	+
Slovakia	-	-	for	i	i
Slovenia	+	+	+	+	+
Spain	+	-	p	+	+
Sweden	+	+	+	+	+
United Kingdom	-	±ctz, dref	±ctz, dref, a70	±ctz, broad age, dref	-
Iceland	+	+	-	-	-
Lichtenstein	-	-	-	-	-
Norway	+	+	+	+	+
Switzerland	+	+	+	+	+

+ data provided to Eurostat; **-** data not provided to Eurostat; **-age** no disaggregation by age; **-ctz** no disaggregation by citizenship; **±age** age disaggregation only for a few citizenship categories; **±agesex** disaggregation by age not provided for Males and Females; **±ctz** data provided for a few citizenship categories; **±sex** disaggregation by sex provided for a few citizenship categories only; **a70** age provided only until 70 years, with the open-ended group 70+; **broad age** data disaggregated by broad age groups; **dref** reference date different than 1st January; **for** data provided for foreigners only; **i** data inconsistency problems; **na** data not available; **nat** data provided for nationals; **p** provisional data; **x** problems detected in the data sent by the NSI, **tot** data provided for Total.

Table 1: Availability of data on population stock by citizenship, sex and age in the JMQ, 31 countries, as of 1st January 2002–2006.

3.1 Notation and basic concepts

Throughout the paper, the notation used for population variables follows a common convention presented below. In all cases, the superscript n indicates one of the three broad groups of citizenship: nationals, EU27 foreigners or non-EU27 foreigners, abbreviated as *N*, *EU* and *nEU*, respectively, thus reflecting the composition of the European Union as of 1st January 2007. The non-EU27 group includes also the stateless persons. An abbreviation *FOR* is used for all foreign population (EU27 and non-EU27 together). For the transparency of presentation, the country index is skipped, as all calculations proposed in the paper are always country-specific. The variables in question are as follows:

Stock variables:

$P^n(x, t)$ - Population in broad citizenship group n, in the age of x years on 1st January, year t.

$P^n(x, c)$ - Population in broad citizenship group n, in the age of x years at the census date c.

Event variables:

$B^n(t)$ - Births during calendar year t in citizenship group n;

$D^n(x, t)$ - Deaths of persons aged x years, belonging to citizenship group n, during calendar year t;

$I^n(x, t)$ - Registered immigration of persons in citizenship group n, aged x years, during calendar year t, regardless of the country of origin;

$E^n(x, t)$ - Registered emigration of persons in citizenship group n, aged x years, during calendar year t, regardless of the country of destination;

$R^n(x, t)$ - Outcome of the regularisation of the status of formerly irregular residents (cf. [4]) aged x, in year t, by definition referring only to foreigners, $n \in \{EU, nEU\}$, thus with $R^N(x, t) \equiv 0$;

$S^n(x, t)$ - Statistical adjustment (official correction) of the size of population aged x, in year t, due to the reasons other than regularisations;

$A^n(x, t)$ - Acquisitions of citizenship by the population aged x, in year t, by definition referring only to foreigners, $n \in \{EU, nEU\}$, with $A^N(x, t) \equiv 0$.

Unless noted otherwise, age is reported in years *reached* during a given calendar year, and thus the *events* in question (deaths, migrations, citizenship changes, etc.) correspond to parallelograms with vertical sides on the Lexis diagram. An illustration of the relevant concepts on a Lexis plane is shown in Figure 2, in Section 3.4.

Whenever necessary, the index denoting sex is added as an additional subscript $g \in \{m, f\}$ for males and females, respectively, e.g. $P_f^n(x, t)$ refers to female population stock, and ${D_m}^n(x, t)$ to deaths among males. In order to distinguish five-year age groups, an additional left-hand side subscript '5' is added. For example, ${}_5{P_m}^n(x, t)$ refers to male population belonging to broad citizenship group n which was in the age of $[x, x+5)$ years on 1st January of year t. The same principle applies to almost all event variables (D, I, E, R, S and A), with a clear exception of B.

In some instances, for the clarity of presentation, the summation of a particular variable over a given index is indicated by an asterisk in a respective place, e.g. $A^{nEU}(*, t) = \Sigma_x A^{nEU}(x, t)$ refers to all acquisitions of citizenship by non-EU27 foreigners in year t, regardless of age. Similarly, $I^*(x, t) = \Sigma_n I^n(x, t)$ denotes all immigrants aged x, in year t, irrespective of their citizenship, and $D^*(*, t) = \Sigma_n \Sigma_x D^n(x, t)$ refers to all deaths registered in year t, without respect to nationality or age. It has to be noted that in several cases the summation over n involves only two components, e.g. $n \in \{EU, nEU\}$ for $R^n(x, t)$ and $A^n(x, t)$.

3.2 Interpolation of five-year age groups into one-year groups

Among the preparatory steps for the estimation of missing data, the most frequent problem concerns disaggregation of five-year age groups of population (or events) into single years. This has to be performed in order to enable cohort-wise interpolations, cohort-component projections with yearly steps, or cohort-wise weights propagation, as described in Sections 3.3, 3.4 and 3.5.

If auxiliary information is available from a different source (e.g. from a census, from the previous or next year, etc.), the population size or the number of events can be disaggregated using a 'Prorating' method [11, p. 5-61], whereby the relative distribution from the auxiliary source is imposed on the data in question. The obtained distribution might need to be further adjusted to marginal totals, by means of proportional fitting procedures, described in Section 3.6.

If the data on population stocks by sex, broad citizenship group and five-year age group ${}_5P^n(x, t)$ are available and the stocks by sex and one-year age group $P^*(x, t)$ are also known, then, assuming no other information about the distribution by single years, we can estimate the missing distributions for particular citizenship groups proportionally, that is as: $P^n(x+i, t) = {}_5P^n(x, t) \cdot P^*(x+i, t) / / {}_5P^*(x, t)$. This is an example of the application of the direct proportional fitting described in Section 3.6.2.

If none of the above information is available, the proposed methodology is to use the well-known interpolative four-term third-difference solution of Karup and King [11, p. 5-65]. For each five-year group, the disaggregation into fifths is done via applying multiplicative coefficients to the global value of this group and the neighbouring ones. Different multipliers are used for the first group, the middle groups and the last group, as set forth in Table 2. For example, if we want to split a middle five-year group with population N_i into five single-year groups n_1, n_2, n_3, n_4, n_5, then:

$$n_1 = 0.064\, N_{i-1} + 0.152\, N_i - 0.016\, N_{i+1},$$
$$n_2 = 0.008\, N_{i-1} + 0.224\, N_i - 0.032\, N_{i+1}, \text{ etc.}$$

When Karup-King multipliers are used, the condition $N_i = n_1 + n_2 + n_3 + n_4 + n_5$ is automatically fulfilled.

	First group, N_0			Middle groups, N_i			Last group, N_K		
	N_0	N_1	N_2	N_{i-1}	N_i	N_{i+1}	N_{K-2}	N_{K-1}	N_K
First fifth	+0.344	–0.208	+0.064	+0.064	+0.152	–0.016	–0.016	+0.112	+0.104
Second fifth	+0.248	–0.056	+0.008	+0.008	+0.224	–0.032	–0.032	+0.104	+0.128
Third fifth	+0.176	+0.048	–0.024	–0.024	+0.248	–0.024	–0.024	+0.048	+0.176
Fourth fifth	+0.128	+0.104	–0.032	–0.032	+0.224	+0.008	+0.008	–0.056	+0.248
Last fifth	+0.104	+0.112	–0.016	–0.016	+0.152	+0.064	+0.064	–0.208	+0.344

Source: [11], Table C-1, p. 5-69.

Table 2: Coefficients for the Karup-King interpolation formula.

As an alternative to the Karup-King interpolation, the six-term fifth-difference interpolative formulae of Sprague or Beers can be applied, which however use information from more surrounding groups. Methodological details can be found in Shryock et al. [11]. In our case, the Karup-King interpolation is recommended for the sake of simplicity.

For variables depicting non-vital events, like migration or citizenship acquisitions, the estimates for particular *cohorts* can be obtained from two neighbouring period-age estimates yielded by the Karup-King formula, split equally by half. For the first cohort, we can assume that a half of the relevant period-age events concern the cohort in question, while for the last, open-ended cohort, we can add up the period-age estimate for the open-ended group and a half of the events concerning the age group immediately preceding the last one. The underlying rationale is an assumption that non-vital events are equally spread over the period-age squares of the Lexis diagram. In any case, the estimates for the eldest cohorts would be close to zero for all practical migration-related applications.

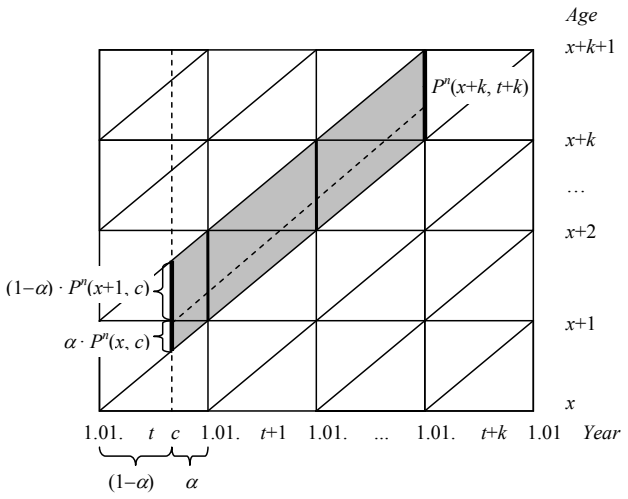
Regardless of the method, if the disaggregation is performed on data broken down by sex or citizenship, the final estimate might need to be obtained by proportional fitting methods (described in Section 3.6), in order to ensure the summation to available marginal totals.

3.3 Cohort-wise interpolation of population stocks

Given the information on the age structures of the population for two non-adjacent moments of time, a simple idea to obtain the missing figures for in-between moments would be to apply interpolation techniques. In this case, we propose cohort-wise interpolation for all cohorts apart from the youngest and oldest one, which are discussed separately. Overall, this method requires much less information on input than the cohort-component projections presented in the next section, but it requires information about population both before and after the moment for which the estimates are to be done. The interpolative approach is recommended for the cases where (a) the span between two points with available data is not wide (say, two-three years), and (b) no information on the distribution of vital and migratory events by citizenship is available.

In practical applications, as the ones described in Section 4, it often happens that the data are available for year t from the census conducted at time c ($t \le c < t+1$), and for 1st January of the year $t+k$, not immediately following the census. Such situations can be put in a general framework illustrated on a Lexis diagram in Figure 1, where α denotes the fraction of a year remaining after the census until 31st December. Figure 1 and the methodology proposed below cover also the situations when data come from other sources than the census, and the situations when the reference date of the data for year t is 1st January. In the latter case it suffices to set $\alpha = 1$.

For the cohorts already existent at the census date c, the interpolation can follow various patterns, but an arithmetic and geometric pattern of growth [3, 10] will be the most frequent choices. As noted by Rowland [10, p. 50], "under arithmetic growth, successive population totals differ from one another by a constant amount [, while] under geometric growth they differ by a constant ratio". For short-period interpolations, both approaches should yield similar results, although this is an empirical issue, and there is no convincing argument in favour of either of them. Hence, a selection of appropriate methods should rely on case-specific judgements.



Source: Own elaboration.

Figure 1: Cohort-wise interpolation of population stocks: a general idea.

It has to be noted that the cohort aged x completed years on 1st January $t+k$ was split at the census date between two age groups: the younger one (aged x completed years) and the older (aged $x+1$), as shown in Figure 1. Therefore, the interpolative estimate of $P^n(x, t+i)$ depends on $P^n(x, c)$, $P^n(x+1, c)$ and $P^n(x, t+k)$.

Given the above, the formula for an interpolative estimate of population sizes belonging to a particular age group $x+i$ and citizenship group n, assuming the linear pattern of change, is as follows:

$$P^n(x+i, t+i) = (k-i) / (k-1+\alpha) \cdot [\alpha \cdot P^n(x, c) + (1-\alpha) \cdot \cdot P^n(x+1, c)] + (i-1+\alpha) / (k-1+\alpha) \cdot P^n(x+k, t+k), \quad (1a)$$

while for the geometric change:

$$P^n(x+i, t+i) = \{[\alpha \cdot P^n(x, c) + (1-\alpha) \cdot P^n(x+1, c)]^{k-i} \cdot \cdot P^n(x+k, t+k)^{i-1+\alpha}\}^{1/(k-1+\alpha)}. \quad (1b)$$

For the youngest and oldest cohorts, for which interpolation as proposed above is not possible, a simplified solution is proposed. In such cases, we suggest to take the average *shares* (proportions) of the sizes of the respective age groups in the total population, calculated from the data available for neighbouring periods, weighted by the distance between the available data points and estimation point.

In order to ensure consistency of the results and summation of the age-specific estimates to the marginal totals by sex or citizenship group, whenever available, the estimates have to be adjusted by the means of proportional fitting, presented in Section 3.6.

The framework presented above can be easily generalised to a much less frequent situation with interpolation between two censuses – in such case, a fraction β of a year between the 1st January of the year of the second census and the second census date, c', should be additionally accounted for. However, the estimates obtained in such cases would be only very approximate, due to a usually large time span between the censuses.

It should be noted that an identical solution as shown above in (1a), or in (1b) can be used for extrapolating cohort sizes *beyond* the available data points, in whichever direction. In either case, it would suffice to put an appropriate integer $i \leq 0$ for the backward extrapolation (in particular, following the example from Figure 1, set $i = 0$ to obtain values for the beginning of the census year), or $i > k$ for the forward extrapolation.

The methods discussed above resemble to some extent the ones presented in the *Human Mortality Database Methods Protocol* [15], with the exception of the oldest age groups, where the quoted study suggests more sophisticated extinct cohort and survivor ratios approaches. Direct application of the methods proposed by Wilmoth et al. [15] would be, however, difficult. This is not because of computational reasons, but rather due to the lack of yearly estimates of deaths, births and migratory events broken down by citizenship groups, which has been listed at the beginning of the current section as a precondition for selecting cohort-wise interpolation method.

3.4 Cohort-component projections

As concerns projections, let us denote by $X^n(x, t)$ a sum of all event variables *not* related to the natural change of population stocks (i.e. all but births and deaths), thus:

$$X^n(x, t) = I^n(x, t) - E^n(x, t) + S^n(x, t) + \Sigma_{k \in \{EU, nEU\}} A^k(x, t), \text{ for } n = N; \quad (2a)$$

$$X^n(x, t) = I^n(x, t) - E^n(x, t) + S^n(x, t) + R^n(x, t) - A^n(x, t), \text{ for } n \neq N. \quad (2b)$$

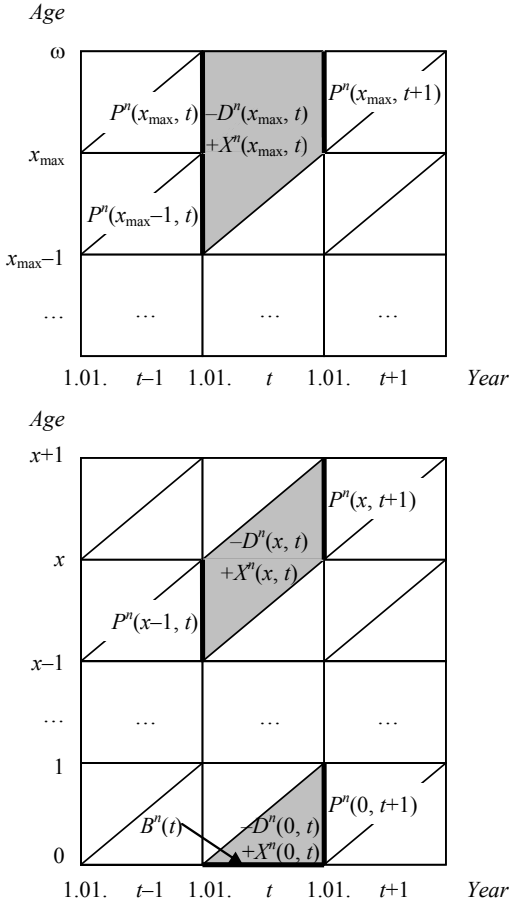
Given (2a) and (2b), the population accounting equations for each broad citizenship group are:

$$P^n(0, t+1) = B^n(t) - D^n(0, t) + X^n(0, t); \quad (3a)$$

$$P^n(x, t+1) = P^n(x-1, t) - D^n(x, t) + X^n(x, t), \text{ for } x \in \{1, 2, ..., x_{\max}-1\}; \quad (3b)$$

$$P^n(x_{\max}, t+1) = [P^n(x_{\max}-1, t) + P^n(x_{\max}, t)] - D^n(x_{\max}, t) + X^n(x_{\max}, t). \quad (3c)$$

In (3c), $x_{\max}$ stands for the highest (open-ended) age group for which information is available. Note also that deaths and other event variables in age group $x_{\max}$ refer to the trapezoid on the Lexis diagram rather than to a parallelogram, while for age group 0 – to a right triangle, as shown in Figure 2.



Source: Own elaboration.

Figure 2: Relationships between population stocks P^n, and events B^n, D^n and X^n on a Lexis diagram.

Under the assumptions presented above, the projection is made following the equations (3a), (3b) and (3c) for consecutive years, on the basis of information available for single-year age groups, decomposed from the five-year groups, if needed.

Note that the default citizenship of a newborn child can differ between the countries, either following the *ius soli* principle, whereby a child acquires the citizenship of the country of birth, or *ius sanguinis*, according to which a child inherits the citizenship of its parent(s), or finally a mixture of those two, for example differentiating between the generations of migrants, taking into account the length of stay in the country, etc. The general rules are as follows:

a) *Ius sanguinis*

If the child gets citizenship of any of the parents, then $B^n(t)$ in equation (3a) may be assumed to be roughly proportional to $P^n(t)$. If the child acquires citizenship of the mother and we have no separate estimate of fertility for nationals and foreigners, then $B^n(t)$ may be assumed to be roughly proportional to $P_f^n(t)$. If the estimates of fertility by broad citizenship and age of mother exist then a better estimate may be obtained using the formula:

$$B^n(t) = B^*(t)\, \Sigma_x f^n(x)\, P_f{}^n(x,\, t) \,/\, \Sigma_{k,\,x} f^k(x)\, P_f{}^k(x,\, t), \qquad (4)$$

where $f^n(x)$ denotes age-specific fertility rates for women in age group x, belonging to the group of citizenship n. If the estimates of fertility are available by broad citizenship group, but not by the age of mother, the formula (4) would have to be modified, so as the summation over age reflects only the female population aged 15–49 years.

b) *Ius soli*

If the child automatically acquires the citizenship of a given country, then the balance equation for the youngest age group, (3a), becomes, depending on the citizenship in question:

$$P^N(0,\, t{+}1) = B^*(t) - D^N(0,\, t) + X^N(0,\, t), \text{ for } n = N; \qquad (3a')$$
$$P^n(0,\, t{+}1) = X^n(0,\, t) - D^n(0,\, t), \text{ for } n \neq N. \qquad (3a'')$$

In mixed cases, it is recommended to project one part of births according to formulas for *ius soli* and another part according to the *ius sanguinis* principle.

Note also that losses of citizenship are not accounted for, as they in most instances concern persons in reality either already living abroad, or emigrating (and counted in E). For acquisitions of citizenship, we assume that non-nationals fall in the category of nationals upon naturalization, in order to count the same people only once, regardless of the number of citizenships they have.

If the breakdown by citizenship group of all variables referring to vital and migratory events can be assumed proportional to the citizenship structure of the population at the beginning of each year, then the projection methodology can be often *de facto* simplified to proportional adjustment / decomposition, whereby the citizenship distribution of the considered cohort in the previous year would directly apply to all cohorts except the first and the last one in each year. In particular, this situation applies if the following four conditions hold:

1. Total population by age, ${}_5P^*(x,\, t)$, is known for successive years, but the citizenship structure is missing;
2. We may assume that the distribution of deaths and migration flows by broad citizenship is the same as the citizenship composition of the population;
3. Acquisitions of citizenship may be ignored;
4. There was no regularization, or it may be ignored.

In such cases, the projection equation (3b) combined with proportional fitting is equivalent to proportional decomposition of ${}_5P^*(x,\, t)$ by citizenship group described in Section 3.6.1. The estimations can be performed using the formula:

$${}_5P^n(x,\, t) = {}_5P^*(x,\, t) \cdot {}_5P^n(x{-}1,\, t{-}1) \,/\, {}_5P^*(x{-}1,\, t{-}1). \qquad (5)$$

The first and the last cohort may be disaggregated using the citizenship composition of the first and last age group in the previous year. In such cases, the following formulas apply:

$${}_5P^n(0,\, t) = {}_5P^*(0,\, t) \cdot {}_5P^n(0,\, t{-}1) \,/\, {}_5P^*(0,\, t{-}1), \text{ or:} \qquad (6a)$$
$${}_5P^n(x_{\max},\, t) = {}_5P^*(x_{\max},\, t) \cdot {}_5P^n(x_{\max},\, t{-}1) \,/\, {}_5P^*(\, x_{\max},\, t{-}1). \qquad (6b)$$

3.5 Cohort-wise weights propagation

In some cases, too much information on the age-sex-citizenship distribution of the components of population change is missing, which renders projections too dubious with respect to the number of assumptions that need to be made. In practice, in such instances the only reliable information comes from the population census and from annual population stocks available in the DEMO domain of the NewCronos database. Hence, the proposed procedure is as follows.

For the census population, apply the structure by citizenship, taken from each five-year age group, to the respective single-year age groups (i.e. from age group 0–4 to single ages 0, 1, …, 4; from 5–9 to 5, 6, …, 9 etc.). Let $w^n(x,\, c) = P^n(x,\, c) \,/\, P(x,\, c)$ denote the age-specific shares ('weights') of citizenship group n in the census.

Further, set α as a fraction of the calendar year before the census date. It is implicitly assumed that the census population in single-year age groups can be divided between 'older' and 'younger' cohorts using the α and $(1{-}\alpha)$ partition. For the census date, use the following formula to calculate the share of citizenship group n in the cohort that was aged x years on 1st January of the census year:

$$w^n(x{+}\alpha,\, c) = [(1{-}\,\alpha) \cdot P^n(x,\, c) + \alpha \cdot P^n(x{+}1,\, c)] \,/$$
$$/\, [(1{-}\,\alpha) \cdot P^*(x,\, c) + \alpha \cdot P^*(x{+}1,\, c)], \text{ for } x < x_{\max}; \qquad (7a)$$
$$w^n(x_{\max}{+}\alpha,\, c) = P^n(x_{\max},\, c) \,/\, P^*(x_{\max},\, c). \qquad (7b)$$

For the 1st January of the census year assume that the weights $w^n(x,\, t) = w^n(x{+}\alpha,\, c)$. For the 1st January of the year following the census year ($t > c$), assume in turn:

$$w^n(x,\, t) = w^n(x{-}1{+}\alpha,\, c), \text{ for } 0 < x < x_{\max}; \qquad (8a)$$

$$w^n(x_{\max}, t) = [P^n(x_{\max}-1, c)\,(1-\alpha) + P^n(x_{\max}, c)] / \\ / [P^*(x_{\max}-1, c)\,(1-\alpha) + P^*(x_{\max}, c)]. \quad (8b)$$

For the youngest age group assume $w^n(x, t) = w^n(0, c)$, or alternatively that the shares are the same as the shares of citizenship group n in the births during the census year, so as: $w^n(0, t) = B^n(t-1) / B^*(t-1)$. For consecutive years calculate:

$$w^n(x, t) = w^n(x-1, t-1), \text{ for } x = 1, \ldots, x_{\max}-1; \quad (9a)$$
$$w^n(x_{\max}, t) = [P^n(x_{\max}-1, t-1) + P^n(x_{\max}, t-1)] / \\ / [P^*(x_{\max}-1, t-1) + P^*(x_{\max}, t-1)]; \quad (9b)$$
$$w^n(0, t) = w^n(0, t-1), \quad (9c)$$

or, as an alternative to (9c): $w^n(0, t) = B^n(t-1) / B^*(t-1)$.

Subsequently, calculate populations for all years using the above shares and total populations (available e.g. from DEMO), as: $P^n(x, t) = P^*(x, t) \cdot w^n(x, t)$. Finally, aggregate single-year age groups into five-year ones.

3.6 Proportional fitting methods

In the proportional fitting methods, the general task is to estimate $P_g{}^n(x, t)$, i.e. the elements of a three-dimensional cube (with the dimensions being sex, age and citizenship). The choice of a particular method depends on which marginal information (cube's edges or faces) is known and if an initial estimate of the cube elements are available. Below, the examples of frequent situations are presented. The formulas have been given for population by single years of age but the analogical formulas apply to population by five-year age group. For more information on multi-proportional techniques see for example the studies of Willekens [12, 13], Willekens et al. [14], Rees [8] and Norman [6]. Note that proportional fitting methods presented below are known under various names in the scientific literature.

In addition to a potential application as the main estimation method, proportional fitting may be used, in almost all the countries for which estimations are needed, as the final stage of the estimation procedure, in order to adjust the initial estimates to known aggregates or marginal totals. The initial estimate might be obtained for example using interpolation or projection, or assumed to be the same as at some different time (e.g. the same as at the census date). Such an initial estimate has to be subsequently adjusted for example to the known total population size by age and sex.

3.6.1 Proportional adjustment / decomposition

Among the proportional fitting methods, the simplest one can be applied to situations, when a population can be directly disaggregated by a variable (sex, age or citizenship), according to the pattern observed in an auxiliary source. In general, the idea is the same as in the Prorating method [11, p. 5-61] mentioned in Section 3.2.

For example, if the aggregates $P_g{}^*(x, t)$ and an initial estimate of the citizenship structure $P'_g{}^n(x, t)$ are known, then the final estimate $P''_g{}^n(x, t)$ may be obtained as:

$$P''_g{}^n(x, t) = P'_g{}^n(x, t) \cdot P_g{}^*(x, t) / P'_g{}^*(x, t). \quad (10)$$

In particular, if one wants to estimate the breakdown by citizenship using the citizenship structure taken from the census, $P_g{}^n(x, c)$, then (10) becomes: $P_g{}^n(x, t) = P_g{}^*(x, t) \cdot P_g{}^n(x, c) / P_g{}^*(x, c)$.

3.6.2 Direct proportional fitting

The estimation problem becomes slightly more complicated, if one wants to estimate $P_g{}^n(x, t)$, but does not have any initial estimate of it. One possible situation is that at least some fragments of the data cube (faces and/or edges) are available and provide coherent information (sum up to the same totals). In such cases, the most straightforward solution is provided by a direct proportional fitting method, whereby the missing elements (i.e. the inside of the cube) can be obtained by taking simple proportions to all available marginal totals.

For example, let the available data consist of known $P_g{}^*(x, t)$ and $P^n_{\bullet}(*, t)$, i.e. the age-sex face and the citizenship edge of the age-sex-nationality cube. Then, the sought-for $P_g{}^n(x, t)$ can be estimated as:

$$P_g{}^n(x, t) = P_g{}^*(x, t) \cdot P^n_{\bullet}(*, t) / P^*_{\bullet}(*, t) \quad (11)$$

In practical applications discussed in Section 4, this option was used rather infrequently, because there usually are some initial estimates of the population structures, for example from the census. Willekens et al. [14, p. 97] noted that general formulae of a form akin to (11) for a one face – one edge problem, as well as similar closed-form solutions for the cases with three edges or two faces are the solutions of the entropy-maximisation problems in research tasks aimed at reconstructing the elements of a three-dimensional arrays, given the available marginal totals.

3.6.3 Iterative proportional fitting

In a general case, a closed-form solution (11) may not exist due to possible incoherence between various data at hand. Such problems call for a multi-step iterative proportional fitting (IPF) method, whereby the solutions are sought step-wise, through iterative adjustments of their successive approximations to marginal totals available from the faces or edges of the data cube. In particular, this method can be used for adjusting the preliminary joint distributions to the known marginal distributions.

For example, let the initial estimate of the citizenship structure $P_g{}'^n(x, t)$ be known, as well as the sex-age face and the citizenship edge of the data cube, respectively $P_g{}^*(x, t)$ and $P^n_{\bullet}(*, t)$. By the IPF algorithm, the initial estimates are iteratively corrected by proportional adjustment. An additional superscript (k) in $P_g{}^{(k)\,n}(x, t)$ denotes the iteration step (for $k \geq 1$). The starting value $k = 1$ defines also the initial estimate of the joint sex-age-citizenship distribution, $P_g{}^{(1)\,n}(x, t) = P_g{}'^n(x, t)$. Subsequent steps are computed as follows:

$$P_g{}^{(2k)\,n}(x, t) = P_g{}^{(2k-1)\,n}(x, t) \cdot P_g{}^*(x, t) / P_g{}^{(2k-1)\,*}(x, t); \quad (12a)$$
$$P_g{}^{(2k+1)\,n}(x, t) = P_g{}^{(2k)\,n}(x, t) \cdot P^n_{\bullet}(*, t) / P^{(2k)\,n}_{\bullet}(*, t). \quad (12b)$$

The procedure defined by (12a) and (12b) is repeated iteratively till some convergence criterion is achieved. For example, the estimates yielded by consecutive steps

should differ by no more than by an arbitrarily-selected small number ε. More details of the method have been discussed by Willekens [13, pp. 69–71], Willekens et al. [14], Rees [8] and Norman [6].

Although the IPF method is purely mechanical, its main advantage is that it does not require any additional information (such as data on vital events or migration) or excessive labour resources, and the obtained results (in terms of joint distributions by all variables under study) are automatically coherent with marginal distributions of particular variables. Moreover, under some general assumptions, the IPF estimates can be interpreted from a statistical viewpoint as joint probability distributions obtained using the maximum likelihood or entropy maximisation methods [2, pp. 83–97; after: 13, p. 70].

3.7 Auxiliary methods

Among the auxiliary methods proposed in the current study, the foremost one is the decomposition of the *Unknown* category wherever it appears (i.e., with respect to age, citizenship, or even sex, as in the case of Greece for 2005). The universal solution proposed in such cases is a proportional disaggregation: population belonging to the *Unknown* category is broken down proportionally to the existing, well defined categories (citizenship groups, age groups, etc.) and the resulting parts are attached to these categories. For example, if total population P consists of n well-defined groups $P_1, \ldots, P_n$, and the *Unknown* category, P_{unk}, such that $P = \Sigma_i P_i + P_{unk}$, where $i = 1, \ldots, n$, then the following corrections apply:

$$P'_j = P_j + P_{unk} \cdot P_j / \Sigma_i P_i = P_j (1 + P_{unk} / \Sigma_i P_i) \text{ , for all } j, \text{ with } i = 1, \ldots, n. \quad (13)$$

If some elements of age structures are missing (e.g. tails of respective age distributions, or a breakdown into five-year groups given the availability of broader ones), we may either use a structure from a different year or fit a mathematical function to available data. For example, we can assume that foreign population stocks are a double-exponential function of age, as originally proposed for the intensity of migration flows by Rogers and Castro [5, 9]. The number of foreign population aged x, $\phi(x)$, would then be given by the following equation:

$$\phi(x) = c + a_1 \cdot \exp(-\alpha_1 \cdot x) + a_2 \cdot \exp\{-\alpha_2 \cdot (x - \mu_2) + - \exp[-\lambda_2 \cdot (x - \mu_2)]\}. \quad (14)$$

The parameters c, a_1, α_1, a_2, α_2, λ_2 and μ_2 can be estimated separately for each sex, for example using the ordinary least squares method (OLS) on the basis of the data for the available age groups (for example, below 65 years of age). Technically, the calculations can be done in a spreadsheet (e.g. MS Excel) using a solver-like tool, controlling for sensitivity of the algorithm to the choice of initial input values. Based on the obtained parameter estimates, formula (14) yields approximations of $\phi(x)$ for the remaining age groups. The last, open-ended group (85+) can be obtained by subtraction of all other figures from the total. To avoid negative numbers in the 85+ category, appropriate constraints should be set during the estimation procedure.

In either case, when adjustment to broader age groups is needed in order to ensure summation to respective totals (e.g. for functional age groups), it can be done via proportional fitting presented in Section 3.6.

4 Estimating population stock for EU27 and EFTA countries

The current section briefly summarises the algorithm for the selection of an appropriate method of computations for a given country (Section 4.1), followed by a brief illustration of the proposed approach employed for the 31 countries under study, and a selection of the results (4.2).

4.1 Procedure for selecting an estimation method

In the light of the overview of data availability presented in Section 2 and the methodological discussion presented in Section 3, it is suggested to inspect the following general options of data availability, in order to apply the relevant data estimation procedures:

Option 1. All the required data are available in the Eurostat database

Whenever all data are available in the Eurostat database, the following five-step procedure is recommended:

1. Organize the data in a database;
2. Verify the data (perform data validation and internal consistency checks);
3. Deal with the *Unknown* categories (if applicable);
4. Calculate the required aggregates;
5. Check the results.

This option includes cases when there is a need for combining data from various parts of the Eurostat database (e.g. in DEMO and in JMQ), and the cases where there is an 'Unknown' category, which has to be disaggregated proportionally among the well specified categories, as described in Section 3.7 on 'Auxilliary methods'.

Option 2. Some of the data missing in the Eurostat database can be obtained from the respective NSI or from other sources

In this case, two situations are possible:

Option 2a. All the missing data may be obtained without contacting the NSI

If all the missing information is publicly available, for example from the NSI webpage, it should be downloaded and combined with the Eurostat data. Such an overall dataset should be then subject to a procedure described under Option 1, points 2. through 5.

Option 2b. Some (or all) missing data are (or are suspected to be) available either from the NSI or from other sources

If some missing information is downloadable from sources like the NSI webpage, it should be collected and merged with the Eurostat data. Nonetheless, there are cases when data are not publicly available but it can be suspected that either some, or even all the missing information is in the possession of the NSI. In such case, the undertaken actions should be as follows:

1. Contact the NSI in order to obtain the missing data. If successful, proceed as in **Option 2a**;
2. For the data that are still unavailable, but can be estimated, proceed as in **Option 3**;
3. For the data that are still not available and cannot be estimated, look at **Option 4**.

Option 2 includes cases when data from various national sources has to be combined, for example aggregated data obtained using the component method and data on citizenship composition from the register of foreigners.

Option 3. Some data are not available anywhere but can be estimated

Even if some auxiliary data can be collected from whichever source, in many instances the available information can be still incomplete. In a vast majority of such cases, the missing information can be still estimated, either following the methodological guidelines and techniques outlined in Section 3, or by means of more straightforward and easy-to-apply solutions. The order of undertaken actions is then as follows:

1. Organize the available data (from Eurostat and other sources);
2. Verify the data (perform data validation and internal consistency checks);
3. Deal with the *Unknown* categories (if applicable);
4. Calculate the required aggregates for the available data;
5. For each year for which data are missing select the best method to estimate missing data;
6. Collect supplementary data needed for the estimations;
7. Estimate the missing data;
8. Check the results.

For the estimations (item 7), various methods can be used, depending on the range and type of the missing information and data availability. In general, five broad groups of methods can be distinguished here, following the outline presented in Section 3:

a. Proportional fitting methods (Section 3.6);
b. Cohort-wise weights propagation (Section 3.5);
c. Cohort-wise interpolation of population stocks (Section 3.3);
d. Cohort-component projections (Section 3.4);
e. Other solutions, not listed above, or combined approaches.

The methodology of interpolating the five-year into one-year age groups, presented in Section 3.2, as well as some methods described in Section 3.7 should be treated as auxiliary to all the remaining ones rather than constituting separate estimation methods *per se*.

Option 4. Data are not available, and no or only very rough estimates can be produced

In principle, this should be a very infrequent option. If no information is available that would enable estimation under Option 3, none or only very rough approximations can be performed, such as for example the 50-50 division of all foreigners into the EU27 and non-EU27 categories.

Under Option 3, in all the cases where several methods could be alternatively applied, preference is given to the more straightforward ones, and definitely to the ones having less judgemental elements, thus less potential sources of error. This approach conforms to the *Occam's razor* principle, stating that "entities are not to be multiplied beyond necessity"[2], which in this case means that the proposed models should not be more sophisticated than necessary, due to various possible sources of error. For example, given complete data on stocks by age and citizenship group for two moments of time (e.g. from successive population censuses), if the data on flows (I and E), natural change (B and D) and citizenship acquisitions (A) are not available by citizenship and/or age, and require estimation, then the intermediate values are recommended to be calculated using cohort-wise interpolation rather than projection. In the former case, the only source of possible error is the composition of population as such, whereas in the latter, judgemental assumptions on the relevant distributions of all components of the balance equation are likely to result in higher uncertainty of the ultimate results, which within the deterministic framework of the project is impossible to assess.

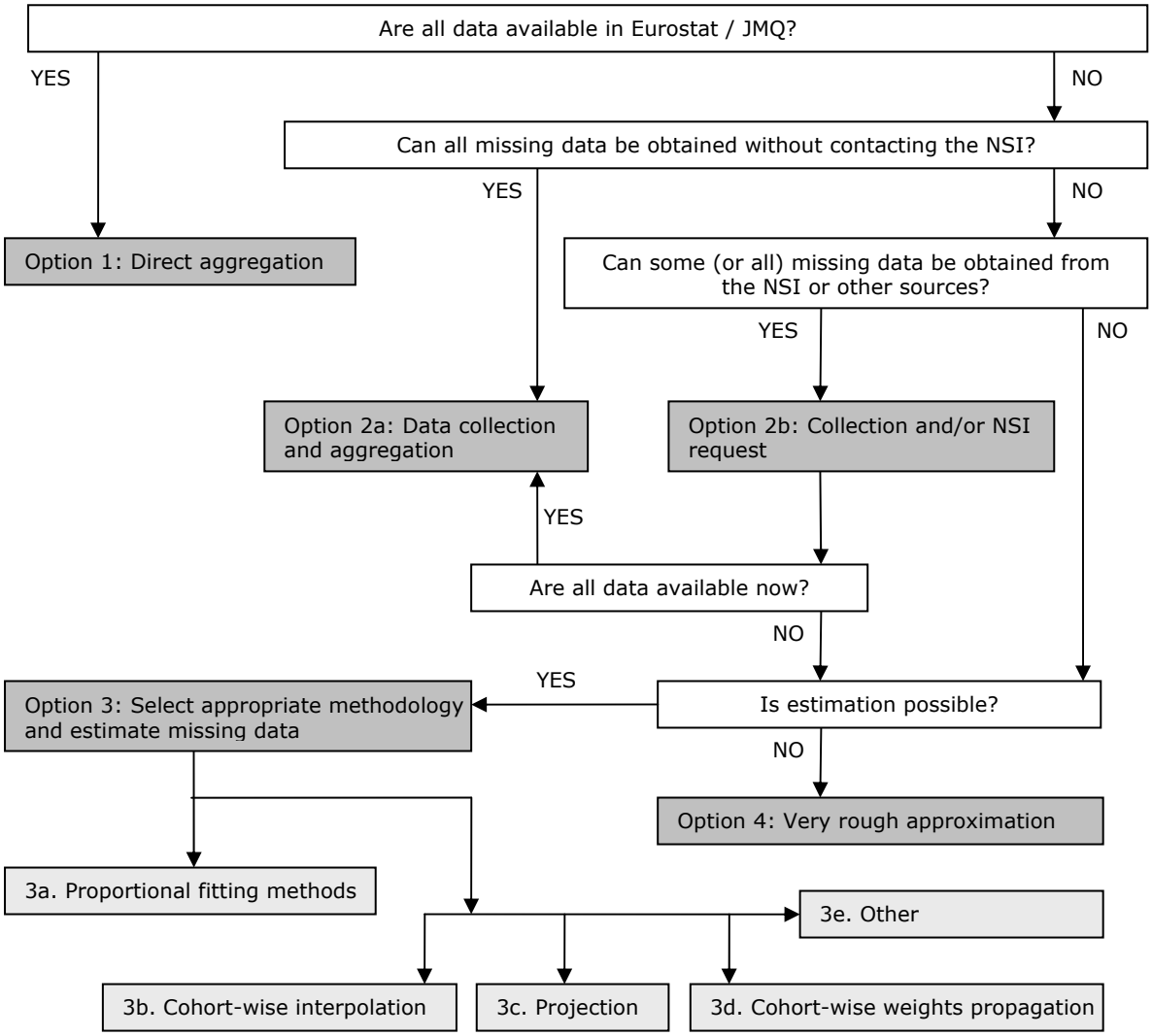
Figure 3 presents a decision tree summarising the procedure for selecting the estimation methodology, taking into account all the above options.

4.2 Application of the methodology, examples, selected results

The decision tree presented in Figure 3 has been used to select the best estimation method for each of the 31 EU and EFTA countries, accounting for the availability of data in the Eurostat database (either on-line or in the JMQs), in the NSI databases, and at other sources. It turned out that complete data needed to estimate population by broad group of citizenship, sex and age on 1st January 2002–2006 were available in the JMQs for nine countries: Austria, the Czech Republic, Denmark, Finland, Hungary[3], Norway, Slovenia and Sweden.

[2] After: 'Ockham's razor', in: *Encyclopædia Britannica Online*, http://www.britannica.com/eb/article-9056716, accessed on 21st May 2007.

[3] For Hungary, data on total population and on the number of Hungarian citizens were not always provided in the JMQ and therefore not available in the migration part of the Eurostat database. However, data on total population were available in the demographic part of the Eurostat database and the number of Hungarian citizens could be calculated directly as a difference between total population and total foreigners, the latter taken from the JMQ.



Source: Own elaboration.

Figure 3: Decision algorithm for obtaining population stocks by broad citizenship group, sex and age.

For additional four countries it was possible either to collect all the missing data from the NSI websites (Belgium and Iceland), or to get them by contacting directly the NSI (Lichtenstein and Switzerland).

For the remaining 18 countries some estimations were necessary. The method that proved to useful in the largest number of cases was some sort of proportional fitting (one of the three versions presented in Section 3.6). It was used as the main method for estimating population by broad citizenship in Cyprus, France, Germany, Greece, Italy, Latvia, Luxembourg, Malta, Slovakia, Spain and the UK. In all cases the total population was assumed to be as reported by the NSI in their demographic statistics, while the citizenship structure was taken from varied sources, for example the JMQ data for the same year, data taken from the NSI website (Italy), the census data (Cyprus, France), the data for another year (Romania, Spain), the LFS data (Cyprus, France) or the data from the register of foreigners (Germany) (see also examples below).

The cohort-wise interpolation method was used for Ireland, Lithuania and Portugal. For Bulgaria, Estonia and Poland, where only data from the census were available, the cohort-wise weight propagation was applied. For Cyprus, Lithuania, Luxembourg and Poland it was originally planned to use a projection method, however it was decided that it would require too many assumptions that would be difficult to justify, and that the final result would not be reliable enough to justify the additional effort required when using this method.

Estimations done for Romania do not fit any of the above groups. They involved simple combination of data coming from various sources.

Below, more details about the estimation procedures are provided for selected countries. In doing so, we have tried to give an example for each estimation method. The resulting numbers in terms of the estimated citizenship structures of the populations of 18 European countries (all being EU Member States) on 1st January 2006, are presented in Table 3.

Country	Total	Nationals	EU27 foreigners	Non-EU27 foreigners
Bulgaria	7 718 750	7 693 214	3 855	21 681
Cyprus	766 414	678 114	52 217	36 084
Estonia	1 344 684	1 082 605	3 961	258 118
France	61 166 822	58 208 155	1 148 691	1 809 976
Germany	82 437 995	75 148 846	2 448 113	4 841 036
Greece	11 125 179	10 165 903	180 282	778 994
Ireland	4 209 019	3 779 755	295 165	134 099
Italy	58 751 711	56 081 197	538 853	2 131 661
Latvia	2 294 590	1 837 832	5 527	451 231
Lithuania	3 403 284	3 370 422	1 962	30 900
Luxembourg	469 086	280 938	171 876	16 273
Malta	404 962	392 850	7 022	5 090
Poland	38 157 055	38 115 920	18 660	22 476
Portugal	10 569 592	10 293 686	80 039	195 867
Romania	21 610 213	21 584 220	6 058	19 935
Slovakia	5 389 180	5 368 255	12 289	8 636
Spain	43 758 250	39 755 741	1 326 128	2 676 381
United Kingdom	60 393 100	56 990 704	1 365 190	2 036 807

Source: Own calculations based on the Eurostat and NSI data.

Table 3: Estimated population by broad group of citizenship in 18 EU countries, as of 1st January 2006.

As concerns particular examples: in **Germany**, data on foreigners come from two different sources. The component method (*Bevölkerungsfortschreibung*), based on the last traditional German census of 25th May 1987, is used by the NSI to produce annual figures on total population, total nationals and total foreigners, as well as nationals and foreigners by sex and age. The other source is the Central Register on Foreigners which contains data on foreigners by citizenship, sex and age. The total numbers of foreigners and their sex and age structures differ between both sources. In order to obtain a single set of estimates, the total number of German citizens, the total number of foreigners, as well as the age structures of Germans and foreigners were taken, following the NSI procedures, from the *Bevölkerungsfortschreibung* data. The distribution of foreigners into EU27 and non-EU27 foreigners was done in proportion to their shares in respective age groups according to the data from the Central Register of Foreigners. Thus, all in all, the proportional decomposition method was used.

In **Latvia**, no joint distribution of population by citizenship and age was available for 1st January 2002, only the structures by age and by citizenship separately. However, the full joint distribution was available for 2003. The iterative proportional fitting method was selected to deal with this case. The joint distribution by citizenship group and age on 1st January 2003 was taken as the starting point for estimating the 2002 structure of population, which was then iteratively adjusted to the known marginal totals.

Lithuania is an example for the application of a cohort-wise interpolation method. In this country, the joint distribution of population by sex, age and citizenship was available for the Census date (6th April 2001), as well as for 1st January 2005. The cohort-wise interpolation, as described in Section 3.3, was used to obtain the initial estimates of males and females on 1st January 2002, 2003 and 2004. In the next step those initial estimates where proportionally adjusted to the known numbers of males and females by age, taken from the Eurostat demographic database.

In **Bulgaria**, annual data on population by citizenship were not available. The only information on citizenship structure came from the census of 1st March 2001. There are also annual data on population by age and sex prepared by the NSI using the component method, available from the Eurostat database. The estimates of annual 2002–2006 population by citizenship, sex and age were prepared using the cohort-wise weight propagation method. The census data were used as the starting point for calculating the initial shares (weights) of citizenship groups in each age cohort. These shares were iteratively propagated forward as described in Section 3.5 and the resulting weights were combined with the available data on population by sex and age to calculate the required joint distribution by citizenship, sex and age.

5 Conclusion

As it can be seen from the country-specific overview of problems with data on population stocks by age, sex and citizenship, there is no universal solution for estimating the missing pieces of information in the European countries under study. Nevertheless, depending on the availability of data at hand, either in the Eurostat / JMQ, or in the respective national statistical institutes, several estimation procedures can be proposed and applied, as mentioned in Sections 3 and 4.

The methods and algorithm we proposed for this purpose do not, however, consider the issue of the harmonisation of the data and definitions, as mentioned in Section 1. More work would be needed in order to recalculate the population stocks into a common definition (cf. [1, 7]), and make them consistent with the (also re-estimated) statistics on migration flows. These very important research tasks are still to be performed in the subsequent tasks of the MIMOSA research project, of which the current study forms a part.

Acknowledgement

The paper was prepared within the framework of the research project on “Modelling of statistical data on migration and migrant population” (MIMOSA), commissioned by Eurostat (contract no. 2006/S 100-106607/EN; Lot 2) to the Netherlands Interdisciplinary Demographic Institute (NIDI) and conducted jointly by NIDI, Central European Forum for Migration and Population Research (CEFMR), Groupe d'étude de démographie appliquée, Université Catholique de Louvain (GéDAP UCL) and Southampton Statistical Sciences Research Institute of the University of Southampton (S3RI). We are grateful to our colleagues for their comments and discussions. Special credits go to Peter Ekamper from NIDI for his assistance in the computational part of the current study.

References

[1] Bilsborrow, R., Hugo, G., Oberai, A.S. and Zlotnik, H. (1997). *International migration statistics. Guidelines for improving data collection systems*. International Labour Organisation, Geneva.

[2] Bishop, Y.M.M., Fienberg, E.F. and Holland, P.W. (1975). *Discrete multivariate analysis*. MIT Press, Cambridge, MA.

[3] Calot, G. and Sardon, J.-P. (2003). *Methodology for the calculation of Eurostat's demographic indicators*. Eurostat Working Papers and Studies, Population and Social Conditions 3/2003/F/n° 26. Eurostat, Luxembourg.

[4] Cangiano, A. (2008). Foreign Migrants in Southern European Countries: Evaluation of Recent Data. In: J. Raymer, F. Willekens (eds.), *International Migration in Europe: Data, Models and Estimates*. John Wiley, Chichester, pp. 89–114.

[5] Castro, L.J. and Rogers, A. (1983). Patterns of Family Migration: Two Methodological Approaches. *Environment and Planning A*, 15 (2), pp. 237–254.

[6] Norman, P. (1999). *Putting Iterative Proportional Fitting on the Researcher's Desk*. Working Paper 99/03. School of Geography, University of Leeds.

[7] Poulain, M., Perrin, N. and Singleton, A. (eds.) (2006), *THESIM: Towards Harmonised European Statistics on International Migration*, Presses Universitaires de Louvain, Louvain-la-Neuve.

[8] Rees, P. (1994). Estimating and projecting the populations of urban communities. *Environment and Planning A*, 26 (11), pp. 1671–1697.

[9] Rogers, A. and Castro, L.J. (1981). *Model Migration Schedules*. IIASA Report RR-81-30. International Institute for Applied Systems Analysis, Laxenburg.

[10] Rowland, D.T. (2006). *Demographic methods and concepts*. Oxford University Press, Oxford.

[11] Shryock, H.S., Siegel, J.S. and Associates (1993). Interpolation: Selected General Methods. In: D.J. Bogue, E.E. Arriaga, D.L. Anderton and G.W. Rumsey (eds.), *Readings in Population Research Methodology. Vol. 1: Basic Tools*. Social Development Center / UN Population Fund, Chicago, pp. 5-48–5-72.

[12] Willekens, F. (1977). *The Recovery of Detailed Migration Patterns from Aggregate Data: An Entropy Maximizing Approach*. IIASA Report RR-81-30. International Institute for Applied Systems Analysis, Laxenburg.

[13] Willekens, F. (1982). Multidimensional Population Analysis with Incomplete Data. In: K.C. Land and A. Rogers (eds.), *Multidimensional Mathematical Demography*. Academic Press, New York, pp. 43–111.

[14] Willekens F.J., Pór A. and Raquillet R. (1981). Entropy, multiproportional, and quadratic techniques for inferring patterns of migration from aggregate data. In: A. Rogers (ed.), *Advances in Multiregional Demography*. IIASA Report RR-81-6. International Institute for Applied Systems Analysis, Laxenburg, pp. 83–124.

[15] Wilmoth, J.R., Andreev, K., Jdanov, D., Glei, D.A., with the assistance of C. Boe, M. Bubenheim, D. Philipov, V. Shkolnikov and P. Vachon (2005). *Methods Protocol for the Human Mortality Database, version 4*. Department of Demography, University of California, Berkeley. Available from: http://www.mortality.org/Public/Docs/MethodsProtocol.pdf (accessed on 17th May 2007).