Assessing the quality of data on international migration flows in Europe: the case of undercounting

Maciej J. Dańko[1,*], **Arkadiusz Wiśniowski**[2,+], **Domantas Jasilionis**[1,+], **Dmitri A. Jdanov**[1], **and Emilio Zagheni**[1]

[1]Max Planck Institute for Demographic Research, 18057 Rostock, Germany
[2]Department of Social Statistics, School of Social Sciences, University of Manchester, M13 9PL Manchester United Kingdom.
[*]danko@demogr.mpg.de
[+]these authors contributed equally to this work

ABSTRACT

Undercounting is a serious data quality issue that can lead to a directional bias in migration statistics. It may be caused by a lack of legal requirements for reporting migration events between and within countries, or difficulties in enforcing such requirements. The main sources of information on undercounting are the metadata accompanying official statistics and expert opinions. However, metadata related to undercounting are very limited. Similarly, expert opinions may be arbitrary, elicited from few experts who might not know all the details of the migration data shared by different countries, or who might not take into account changes in methodologies or definitions, or retrospective updates of the data after censuses. This paper aims to develop a methodological solution for the assessment of undercounting in international migration data, and has three main objectives. First, the paper provides an overview of the available metadata and expert opinions on undercounting in European migration flows. Second, the study proposes a new, data-driven, year-specific, and duration-of-stay-adjusted approach to classifying undercounting that enables scientists, researchers, policymakers, and other users to combine information from various sources. The proposed methodological solution relies on bilateral migration data provided by Eurostat and the UN, as well as migration data provided directly by some national statistical institutes (NSIs), to compare flows, in the same direction reported by a given country with high-quality data reported by another set of countries. The duration-of-stay correction coefficients are calculated by using an optimization model, or are taken from previous migration models. We construct metadata and expert opinion scores and combine them into a single classification of undercounting. Third, the final outcome is a dynamic classification of undercounting for 32 European countries (2002–2019) that is easily accessible and, flexible, and that allows for changes to the underlying assumption via an online Shiny application.Our findings suggest that the highest level of undercounting in migration data are observed in the new EU member states, particularly in Bulgaria, Latvia, and Romania. However, we also show that for certain periods, there have been notable levels of undercounting in many other European countries, including in those countries that are traditionally assumed to maintain reliable population statistics.

Introduction

Background

International migration is a complex, heterogeneous, and hard-to-measure process. In recent decades, it has become an important force of population change affecting the socioeconomic and political domains in both the receiving and the origin countries[1,2]. Recently, migration waves in Europe have fueled heated debates regarding migration policies[3]. However, many international and national policy reports on migration are not based on reliable and representative evidence. A lack of trustworthy and comprehensive migration data has led to important research gaps, and to an increasing reliance on lay evidence in public domains.

One of the biggest challenges that arises when collecting demographic statistics is quantifying migration stocks and flows. At the national level, these data are usually collected in countries with developed statistical systems by national statistical institutes (NSIs) or/and other governmental agencies, including migration or foreigner registers. At the international level, the UN (United Nations), OECD (Organisation for Economic Co-operation and Development), and EUROSTAT (European Statistics) databases also provide quite comprehensive data series. However, these international repositories differ in terms of their focus and contents. The Eurostat database focuses on the EU member states and relies on national administrative sources, which include population registers, national surveys, censuses, border data collection systems and visas, residence permits, and/or work permits. The UN and OECD databases rely on a wider variety of data sources and indirect estimations, and usually provide flow estimates by citizenship or aggregated (total in- and out-) flows. In many cases, even this limited scope of migration data provided by the aforementioned international agencies is lacking detailed retrospective metadata and a thorough quality assessment.

In general, the quality of migration statistics varies over time and depends on the country's migration/registration procedures, the legal incentives for registering migration events, and the methodologies used by NSIs to measure migration. Unlike for the collection of vital statistics, universally recognized and standardized methods for the consistent and reliable collection of migration data are lacking. For technical and legal reasons, the quality of data on out-migration is often lower than that of data on in-migration. Therefore, it is challenging to generate reliable estimates for a single country without engaging in extensive data exchanges with other countries. Existing migration estimates may not be directly comparable across time and space, as they are obtained from different types of data and rely on various collection mechanisms and methods to generate statistical adjustments. The major problems associated with the quality of migration data can be classified into four groups[4–7]:

(1) Accuracy issues related to random, rather than systematic errors made in the data collection process. In population and migration registers, problems with data accuracy can arise due to random mistakes in the registration or de-registration process. The accuracy of survey data depends on both the chances of random mistakes occurring when the information is recorded, as well as the sampling error.

(2) Undercounting reflecting a non-systematic bias in migration estimates. In population and migration registers undercounting refers to arrival and departure registration failures. In survey-based data, it can occur due to non-response of individuals who experienced migration.

(3) Inconsistencies in coverage. This problem is a special case of undercounting that reflects systematic biases. Such biases can occurs as the results of the rules that govern the data collection process, which may exclude certain population segments, such as nationals who are return migrants or foreigners who are not counted in the official immigration and emigration statistics.

(4) Inconsistencies in the definition of an international migrant. This problem can occur when the national criteria used to define migration (e.g., the minimum duration of stay) deviate from the international (UN/Eurostat) standards. In many countries, these inconsistencies can also arise due to the changes in the international definition of migration. In the European Union,the most important change occurred around 2008, when the EU introduced the 12-month minimum duration-of-stay criterion (Reg (EC) 862/2007[8]).

To address the lack of reliable and internationally comparable migration data, several important projects have been implemented that aim to produce detailed modeling-based estimations of international migrant stocks and flows for developed countries (e.g.,[2,4,6,9–13]). These approaches rely on a wide range of input data and auxiliary information, including household and labor force survey data, official migration estimates, expert estimates, and various contextual parameters. More recently, innovative methods have been proposed to combine data from traditional sources and digital trace information (e.g.,[14–19]). One of the most comprehensive approaches for estimating international migration flows for Europe is the Integrated Model of European Migration (IMEM)[4]. The IMEM Bayesian modeling strategy focuses on estimating "true" latent migration flows and their uncertainty by combining estimates reported by the sending countries with the corresponding information reported by the receiving countries. The model accounts for various biases in migration data and sources of uncertainty, including expert judgments about the quality of official data, and imputes missing information by using covariates[4,20]. An important component of the migration flow models is accounting for data quality characteristics, such as undercounting. In the IMEM framework, Wisniowski et al.[20] developed a solution for converting expert evaluations into prior information for the further Bayesian modeling of European migration flows. However, estimates that rely solely on expert judgments can be at least partially biased depending on the qualifications of the experts who have been selected and successfully recruited[21].

The current study extends the prior work on this topic by developing more comprehensive formal criteria for data quality assessments. We propose a novel method that enables assessing the importance of one of the key sources of bias in migration statistics: namely, undercounting. The method introduces a score-based system that relies on information from estimated, year-specific, and duration-of-stay-adjusted bilateral flow ratios (based on a comparison of the flows reported by the countries under study and the same flows reported by the countries with high-quality data). We utilize data and accompanying metadata on bilateral flows involving 28 countries, including European Union and European Free Trade Association member countries and the United Kingdom, that are primarily sourced from the Eurostat database. Additionally, expert judgments from the IMEM and QuantMig projects, along with metadata information, are integrated into the scoring process, even for the countries that do not provide bilateral flows. Therefore, our method covers a total of 32 countries, while the bilateral flow data needed for the analysis are available for 28 of these countries.

The obtained scores and the country classification based on formal criteria can significantly improve the precision of model-based estimates of the bilateral migration flows within the European Union. An important outcome of this study is the development of a user-friendly and freely accessible online Shiny application. This application empowers users to produce undercounting scores by taking into account alternative expert opinion information and metadata. The app is a convenient and comprehensive tool for assessing undercounting in migration data, and thus offers valuable insights and analysis capabilities.

Undercounting as a key component in models for harmonizing migration

Bayesian inference for statistical models of migration that harmonize data from various sources can be used to account for data quality issues, reconcile the differences in the measurements of the same flows by different countries (bilateral flows), and estimate flows for which data are completely missing[22]. Such models typically assume Poisson, Poisson-lognormal, or other distributions for the migration counts[5], and include two nested blocks: (1) a measurement error model and (2) a predictive model (e.g.,[4,7,22]). The measurement error model (1) can correct the observed flows for differences in duration of stay, and can account for different levels of measurement accuracy, undercounting, and coverage biases. The predictive model (2) imputes missing data and predicts future migration by using covariates correlated with migration, such as geographic distances between countries and bilateral differences related to population size, language, trade, GDP/GNI, freedom of movement of workers, and migrant stocks. This model can also provide temporal smoothing (e.g.,[7,22]).

Prior studies have shown that including undercounting and duration-of-stay criterion in the Bayesian migration model is crucial. For example, the IMEM project estimated that the countries with a strong tendency to undercount emigration capture, on average, just, 45% of emigration flows[4]. The use of a consistent definition of a migrant across countries is important for ensuring an accurate count of migration events in a given unit of time. Most migration flow models assume a 12-month minimum duration of stay[4,7,22]. The failure to take into account durations of stay that are shorter or longer than 12 months may lead to the overcounting or the undercounting of the estimated flows. Indeed, there is a clear interplay in migration models between the parameters for undercounting and duration of stay, such that the misclassification of the duration of stay may lead to a substantial bias in the estimation of the undercounting parameters (e.g., unpublished results of Del Fava et al.[7]).

Assessments of undercounting in previous research

Expert opinions

A "ready-to-use" source of information on undercounting is the body of expert opinion that has proved to be useful in models for harmonizing migration flows data. Two of the best known classifications of undercounting were done for the IMEM (2002–2008, *Integrated Modeling of European Migration*[4]) and QuantMig (2009–2019, *Quantifying Migration Scenarios for Better Policy*[22]) projects. In Table 1, we show classifications of undercounting from both of these projects, and include numerical scores for each of them that are used later in the manuscript. We assign a value of zero to the lowest level of undercounting and a value of one to the highest level of undercounting. The first classification relies on the expertise of the IMEM project team, as well as on assessments of the data collection systems in Europe obtained for the MIMOSA[23,24] and THESIM projects[25]. The second classification (QuantMig Deliverable 6.3[22]) is not based solely on arbitrary expert opinion, but instead relies on the investigation of bilateral migration flow data through pairwise comparisons and rankings using the Bradley-Terry algorithm[26,27]. The algorithm assumes that there is an inherent "ability" assigned to each country, and that the probability of one country "beating" another country in a pairwise comparison depends on the differences in the countries' respective strengths. For example, a sending country beats a receiving country when the flows reported by the sending country are greater than the flows reported by the receiving country. In terms of ability, this means that the sending country is less likely than the receiving country to undercount migration. The algorithm takes into account that the emigration tends to be undercounted to a greater extent than immigration. However, it seems that this method does not account for differences in the duration of stay or coverage in the comparison of countries (these biases are reflected in the wider estimation framework of the QuantMig model). QuantMig uses the algorithm output to assign countries to the three undercount groups: excellent, low, and high; with the excellent group being assumed to have no group-specific undercounting. As was mentioned in the previous section, the inclusion of the duration of stay is important, as an incorrect assumption regarding the duration of stay can lead to an undercounting or an overcounting bias, independent of other factors.

The IMEM and QuantMig classifications are consistent for BE, BG, CH, DK, HR, LV, NL, PL, RO, and SK (here and below we use ISO2 codes for countries, see Table 1 for details). However these classifications differ for other countries, and are clearly contradictory for Italy (Table 1). These discrepancies may be attributable not only to changes in data quality, but also to differences in the methodology and/or metadata used in the undercounting assessments, the numbers of grouping classes, and the subjective intuitions of experts about the data.

Moreover, as well as being partially arbitrary, the expert opinions elicited in the IMEM project may no longer be current, as they covered the 2002-–2008 period only. This point is important for two reasons. First, after 2007, a new definition of migration was implemented in the EU migration statistics: i.e., the minimum duration-of-stay criterion was defined as 12 months (Reg (EC) 862/2007[8]). As this change prompted the NSIs to revise their data collection and processing mechanisms, it could have affected the quality of the migration estimates they produced, likely introducing additional uncertainty into the expert opinions, which were elicited after the implementation of the regulation[20].

Second, most of the national censuses took place in or close to 2011. Since the national censuses are often used by the NSIs to retrospectively update population estimates and migration flows, the assessments of migration data quality provided by the experts may not be valid.

Table 1. Summary of expert opinion assessments of undercounting for the years 2002-2008 in the IMEM model[4] and for the years 2009-2019 in the QuantMig (QM) model[22].

ISO2	Country	IMEM Score	IMEM Class	QM Score	QM Class
AT	Austria	0	Low	0.25	Low
BE	Belgium	0	Low	0.0	Excellent
BG	Bulgaria	1	High	1.0	High
CH	Switzerland	0	Low	0.0	Excellent
CY	Cyprus	0	Low	—	—
CZ	Czechia	1	High	—	—
DE	Germany	0	Low	—	—
DK	Denmark	0	Low	0.0	Excellent
EE	Estonia	1	High	0.25	Low
ES	Spain	0, 1*	Low, High *	0.25	Low
FI	Finland	0	Low	0.25	Low
FR	France	0	Low	0.25	Low
GR	Greece	1	High	—	—
HR	Croatia	1	High	1.0	High
HU	Hungary	1	High	—	—
IE	Ireland	0	Low	0.25	Low
IS	Iceland	0	Low	0.25	Low
IT	Italy	0	Low	1.0	High
LI	Liechtenstein	1	High	0.25	Low
LT	Lithuania	1	High	0.25	Low
LU	Luxemburg	0	Low	—	—
LV	Latvia	1	High	1.0	High
MT	Malta	1	High	—	—
NL	Netherlands	0	Low	0.0	Excellent
NO	Norway	0	Low	0.25	Low
PL	Poland	1	High	1.0	High
PT	Portugal	1	High	—	—
RO	Romania	1	High	1.0	High
SE	Sweden	0	Low	0.25	Low
SI	Slovenia	1	High	0.25	Low
SK	Slovakia	1	High	1.0	High
UK	United Kingdom	0	Low	0.25	Low

* – In the case of Spain, low levels of undercounting for immigration flows and high levels of undercounting for immigration flows are assumed; QuantMig does not provide classification of undercounting for countries for which data on bilateral flows are lacking. The IMEM Score and the QM Score include values for each class assigned by us (between zero and one) for ease of comparison. IMEM Score: Low = 0 and High = 1; QM Score: Excellent ("none to very low undercount") = 0, Low = 0.25, and High = 1.

Table 2. Immigration metadata related to undercounting and their classification. Metadata collected from Eurostat[28], Your Europe[29], and government websites (see QuantMig Deliverable 6.2[30] for details)

ISO2	Country	Reg. oblig.	Time limit	RM	Score	Class
AT	Austria	Yes	3 days		0.0	Low
BE	Belgium	Yes	90 days		0.0	Low
BG	Bulgaria	Yes	At arrival		0.0	Low
CH	Switzerland	Yes	14 days		0.0	Low
CY	Cyprus	Yes	7 days		0.0	Low
CZ	Czechia	Yes	90 days		0.0	Low
DE	Germany	Yes	3 months		0.0	Low
DK	Denmark	Yes	5 days	NC	0.0	Low
EE	Estonia	Yes	1 month	No sanctions	0.5	Medium
ES	Spain	Yes	No limit		0.5	Medium
FI	Finland	Yes	7 days	NC	0.0	Low
FR	France	No	—		1.0	High
GR	Greece	Yes	90 days		0.0	Low
HR	Croatia	Yes	2 days		0.0	Low
HU	Hungary	Yes	90 days		0.0	Low
IE	Ireland	No	—		1.0	High
IS	Iceland	Unk	Unk	NC	0.0	Low
IT	Italy	Yes	No limit		0.5	Medium
LI	Liechtenstein	Yes	Unk		0.0	Low
LT	Lithuania	Yes	7 days		0.0	Low
LU	Luxemburg	Yes	8 days		0.0	Low
LV	Latvia	Yes	90 days		0.0	Low
MT	Malta	Yes	1 month		0.0	Low
NL	Netherlands	Yes	5 days		0.0	Low
NO	Norway	Yes	8 days	NC	0.0	Low
PL	Poland	Yes	4 days		0.0	Low
PT	Portugal	No	—		1.0	High
RO	Romania	Yes	2 days		0.0	Low
SE	Sweden	Yes	7 days	NC	0.0	Low
SI	Slovenia	Yes	8 days		0.0	Low
SK	Slovakia	Yes	5 days		0.0	Low
UK	United Kingdom	No	—		1.0	High

Column names: Reg. oblig. – obligation of registration; RM – remarks; Fields: NC – Nordic country (higher data quality is assumed); Unk – unknown.

Table 3. Emigration metadata related to undercounting and their classification. Metadata collected from Eurostat[28], European Commission[31], and Eurostat[32], see also QuantMig Deliverable 6.2[30] for details.

ISO2	Country	DRO	DRO3rd	M3rd	AC	RM	Score	Class
AT	Austria	Yes	Yes	No			0.143	Low
BE	Belgium	No	Yes	Yes	Yes		0.500	Medium
BG	Bulgaria	No	No	Yes			0.857	High
CH	Switzerland	Yes	Unk	Unk	Yes		0.000	Low
CY	Cyprus	No	No	Yes			0.857	High
CZ	Czechia	No	Yes	No			0.857	High
DE	Germany	Yes	No	No	Yes		0.200	Low
DK	Denmark	Yes	Unk	Unk		NC	0.000	Low
EE	Estonia	Yes	Yes	No	Yes		0.100	Low
ES	Spain	Yes	No	No			0.286	Medium
FI	Finland	Yes	Yes	Yes	Yes	NC	0.000	Low
FR	France	No	No	No			1.000	High
GR	Greece	No	No	No			1.000	High
HR	Croatia	Yes	No	No			0.286	Medium
HU	Hungary	Yes	No	No			0.286	Medium
IE	Ireland	No	Unk	Unk			1.000	High
IS	Iceland	Unk	Unk	Unk		NC	0.000	Low
IT	Italy	Yes	No	No	Yes		0.200	Low
LI	Liechtenstein	Yes	Unk	Unk			0.000	Low
LT	Lithuania	Yes	Yes	Yes			0.000	Low
LU	Luxemburg	Yes	Yes	Yes	Yes		0.000	Low
LV	Latvia	Yes	Yes	Yes			0.000	Low
MT	Malta	No	No	No			1.000	High
NL	Netherlands	Yes	Yes	No	Yes		0.100	Low
NO	Norway	Yes	Unk	Unk		NC	0.000	Low
PL	Poland	Yes	No	No			0.286	Medium
PT	Portugal	No	Yes	No			0.857	High
RO	Romania	No	No	Yes			0.857	High
SE	Sweden	Yes	No	No		NC	0.286	Low
SI	Slovenia	Yes	Yes	Yes			0.000	Low
SK	Slovakia	Yes	No	Yes			0.143	Low
UK	United Kingdom	No	Unk	Unk			1.000	High

Column names: DRO – obligation of de-registration; DRO3rd – obligation of de-registration for third-country nationals; M3rd – monitoring third-country nationals; AC – corrections made by NSOs (old metadata[32]); Score – score based on weighted average (weights: DRO – 50%, DRO3rd – 10%, M3rd – 10%, and AC – 30%), Class – Exemplary three-level classification of the score using 0.25 and 0.75 as thresholds; Fields: RM – remarks; NC – Nordic country (higher data quality is assumed); Unk – unknown.

Metadata

The most basic and direct sources of information on undercounting at the country level are the metadata associated with the official migration statistics, as well as the additional information that may be provided by the NSIs. However, the availability of such metadata is limited. Indeed, the QuantMig project, which assessed the migration data quality in Europe between 2009 and 2019[30], was able to collect only a few useful variables for 2009–2019.

The collected immigration metadata contain information about the obligation to register and incomplete information about the time limits and sanctions for non-registration (Table 2). We assume that the countries without the obligation to register should be assigned a score of one (high level of undercounting), and should otherwise be assigned a score of zero. However, for countries that impose no time limits or sanctions, the score changes to 0.5 (medium level of undercounting). A score indicating a medium level of undercounting is assigned to Estonia (no sanctions), as well as to Spain and Italy (no time limits). While Iceland has no information on either the obligation to register or time limits[30], it is assumed that like all other Nordic countries, Iceland has good quality registers, hence a low level of undercounting.

The emigration metadata score is calculated as a weighted average with arbitrary (default) weights: 50% for the obligation to deregister, 30% for the presence of administrative corrections, and 10% each for the obligation to register and the monitoring of third-country nationals (Table 3). We assign relatively small weights to the last two variables, as third-country nationals are not currently included in our models. We use non-zero weights, because the obligation to deregister and the monitoring of this migration group may still indicate a better quality of the data collection *per se*.

Information on the presence of administrative corrections is provided in the Eurostat 2003 report[32]. Although the report encompasses years preceding 2003, which are largely beyond the scope of this study, this variable holds significant relevance as it indicates that certain countries had already established administrative correction procedures, which we can infer have been consistently applied since then. On the other hand, we do not know whether any of the countries for which no administrative adjustments are reported introduced them later, especially after the major methodological change around 2007. Therefore, we have decided to treat the lack of administrative corrections in 2002 as missing entries. Again, while there are no metadata for undercounting in Iceland, we assume that as a Nordic country, Iceland is likely to have a good registry quality, and hence low levels of undercounting.

A key question is to what extent we can rely on the classification of undercounting based on the metadata. There are two important issues to consider. First, the metadata are based on reports that take into account only very narrow time frames; i.e., they usually reflect the situation around the time the report was created. For example, the metadata provided in the Eurostat 2015 report[28] may not describe the most recent or much earlier periods. Similarly, the metadata published by Eurostat in 2022[33,34], Your Europe in 2021[29], and the European Commission in 2019[31] may be relevant for recent years only. Second, even when metadata are available, they are still insufficient to adequately describe current issues with data quality, including undercounting. Despite these limitations, metadata remain invaluable sources of information when assessing and classifying the data quality of migration statistics.

Data-driven method for classifying undercounting

Motivation

Here, we present a novel approach that aims to minimize the arbitrariness of the undercounting classification, while taking into account how it has changed over time and the differences in the duration-of-stay criteria used in various countries. Our approach also enables users to combine new and previous undercounting classifications based on expert judgments and metadata. While some of the proposed parameter choices rely on our best (albeit arbitrary) subjective judgments, the provided software (*UndercountMigScores*[35]) permits users to adjust the model parameters to account for the impact of various inputs and assumptions.

Bilateral migration flows ratio model

The approach presented here is analogous to the method proposed by Poulain[36,37] and reviewed by de Beer[38]. Bilateral flows give researchers the opportunity to look at the same origin–destination-specific flows from the perspectives of both the sending and the receiving countries. Poulain[36] developed a method that looks for time-invariant correction factors that adjust both immigration and emigration flows with the goal of obtaining a consistent set of migration flows. The correction factors for immigration and emigration countries were obtained by using a constrained optimization algorithm that minimized the differences between these two available data sets. In our approach, the correction factors are specific to the duration-of-stay criteria, rather than being immigration- or emigration-specific.

The bilateral migration flow ratios are constructed by taking a flow from country X to a group of countries with high-quality data reported by country X and dividing it by the same flow reported by the reference countries. As a default set of the reference countries Y, we have selected the Nordic countries (Denmark, Finland, Sweden, Norway, and Iceland), Belgium, the Netherlands, and Switzerland. These countries have been widely recognized by experts for having high-quality data. For example, these countries have been identified by both the IMEM and QuantMig projects as having low or very low levels of undercounting. Additionally, as indicated by our model, they exhibit the lowest levels of undercounting when compared to various reference country sets. Because the minimum duration of stay used in the definition of international migration may differ from country to country (Figure 1), the flows reported by each country need to be adjusted accordingly.

Formally, the undercounting ratio $U^E_{X,Y,t}$ for emigration data from country X to a set of countries Y in year t can be defined as follows:

$$U^E_{X,Y,t} = \frac{\sum_c M\,(X_t \to Y_{c,t}, X_t)\,R_{X_t}}{\sum_c M\,(X_t \to Y_{c,t}, Y_{c,t})\,R_{Y_{c,t}}}, \tag{1}$$

where $M\,(X_t \to Y_{c,t}, X_t)$ is the emigration flow from country X to country Y_c reported by country X in year t; $M\,(X_t \to Y_{c,t}, Y_{c,t})$ is the immigration flow from country X to country Y_c reported by country Y_c in year t; R_{X_t} corrects for the duration of stay for emigration in country X in year t, whereas $R_{Y_{c,t}}$ corrects for the duration of stay for immigration in country Y_c in year t. The immigration bilateral flow ratios are calculated analogously. The correction is designed to reduce flows for stays of less than 12 months and to increase flows for permanent stays.

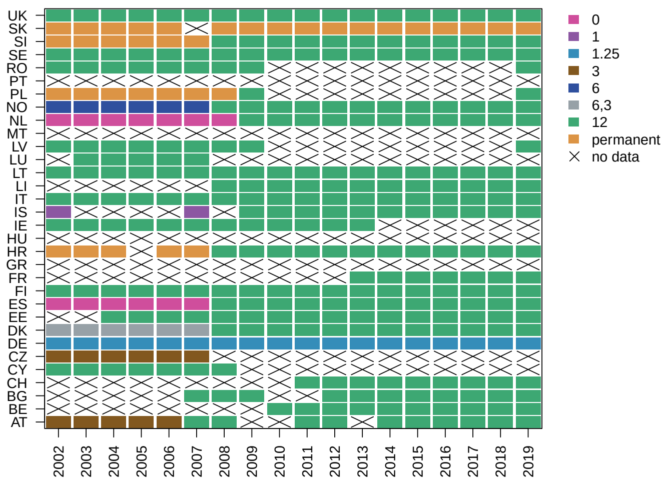
The correction coefficients, denoted by R, are a set of parameters that increase monotonically with the duration of stay. These parameters are obtained through a process of constrained optimization aimed at minimizing undercounting across all European countries from 2002 to 2019. Specifically, the R coefficients for each duration class (zero, three, and six months and permanent) are represented on a cumulative scale and estimated using the L-BFGS-M algorithm, which is a modification of the BFGS quasi-Newton method[39]. To address local minima problems, we use a robust approach that involves starting from multiple random points. For countries with durations other than those mentioned above, linear interpolation is employed. For countries with a duration of stay of 12 months, the correction coefficient is set to one.

By default, the optimization process is performed using the squared difference between the reported flows of a given country X and the flows reported by countries with high-quality data Y_c. For emigration data, the objective function used in the optimization is defined in Equation 2.

$$H^E(R) = \sum_{X,Y,t} \Big(\sum_c M\,(X_t \to Y_{c,t}, X_t)\,R_{X_t} - \sum_c M\,(X_t \to Y_{c,t}, Y_{c,t})\,R_{Y_{c,t}}\Big)^2 \tag{2}$$

The Shiny app also offers the option to use $|logU|$ (where U is defined in Equation 1) as an objective function, but the default option (Equation 2) is considered more suitable for our purposes as it is more sensitive to the magnitude of differences between flows. The correction coefficients can be estimated jointly (which is the default in the app) or separately for each type of migration. Alternatively, the correction coefficients can be obtained directly from previous models, such as the IMEM model[4] or other models summarized by Willekens[5].

(**a**) Duration of stay for immigration flows.



(**b**) Duration of stay for emigration flows.

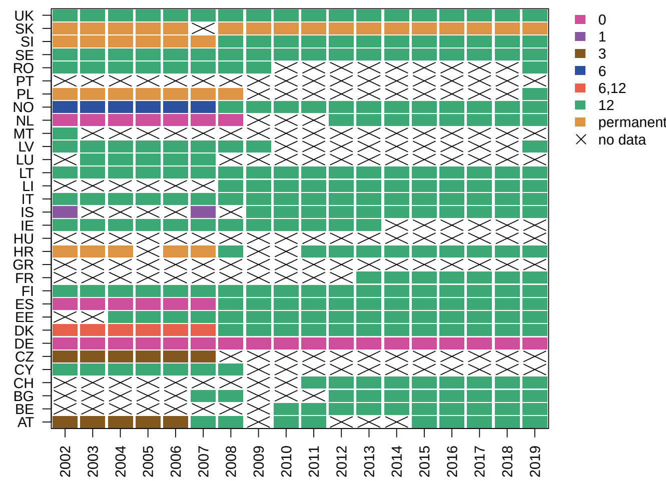


Figure 1. Duration of stay and availability of bilateral flows in the data used. (**a**) immigration (**b**) emigration. In the Danish (DK) data, "6, 3" means three months for immigration from Switzerland (CH) and six months for immigration from other countries, and "6, 12" means 12 months for emigration to Sweden (SE) or Finland (FI) and six months for emigration to other countries. In the case of Germany (DE), "1.25" is a mean duration of stay among different federal states.

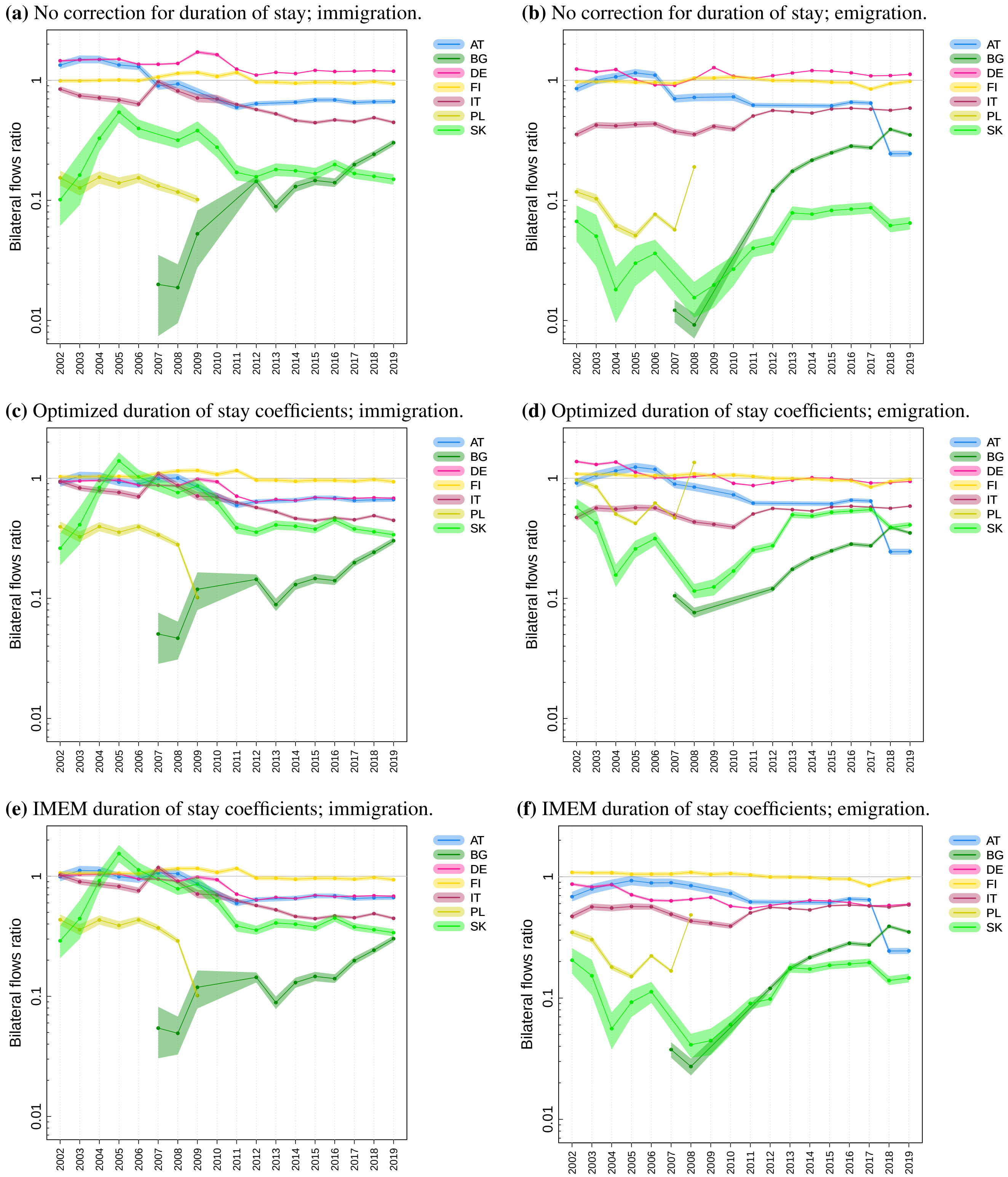


Figure 2. Selected bilateral flow ratios for immigration and emigration data. The ratio is calculated by dividing flows from country X to a group of countries with good data quality (the reference group of countries) reported by country X by the flows in the same direction reported by the reference group of countries. Ratios higher than one indicate the overcounting of emigration flows, while ratios lower than one indicate the undercounting of emigration flows. The lower the ratio, the higher the level of undercounting. The 95% confidence intervals are calculated using the percentile bootstrap method. The figure was generated using *UndercountMigScores*[35] at the default settings, but without PCA imputations. The columns show the results for the immigration and the emigration data, while the rows show the results for different sets of duration-of-stay correction coefficients.

Obtaining normalized undercounting scores

Since bilateral flow ratios U (quotients, see equation 1) are placed on the multiplicative scale, it is convenient to refer to their logarithms. The logarithm of bilateral flow ratios adjusted for the duration of stay ranges from - ∞ to + ∞. Values greater than zero are considered overcounted, zero denotes no undercounting or overcounting, and negative values represent undercounting. These values are then projected to a discretized 0—1 scale (e.g., for a classification with five categories, the values we have chosen are zero, 0.25, 0.5, 0.75, and one), as these thresholds are particularly useful in the next steps of our model. For flexibility, we use categorizations based on evenly spaced thresholds or quantiles (default in our Shiny app).

The projection is not trivial as overcounting problems often arise. Generally, overcounting occurs when the reported migration flows exceed the actual number of migrants who leave or enter a country, which may happen for various reasons, such as double counting, reporting errors, irregular migration, or coverage differences. In our case, overcounting arises if a country reports more migration than the reference group ($\log U > 0$). However, it is unlikely that a pure overcounting class exists, as none of the countries have perfect data quality. As overcounting and undercounting issues may occur at the same time, disentangling them can be challenging.

To address this issue, we propose two options. The first option is to combine the overcounting class with the lowest undercounting class, which is the default setting in our Shiny app. The second option is to treat overcounting as a separate class. The results can be directly used to classify countries in migration models, or, after numerical representation (as described above), combined with expert opinion and metadata scores.

Combining multiple sources of information on undercounting

The proposed procedure for combining undercounting from different sources requires scoring; i.e., a numerical representation of the classifications in the range $(0,1)$. Scores obtained from various sources are then combined by using a weighted average. By default, we have chosen a set of weights that reflect our subjective assessment of the relative importance of each factor: 20% for the expert opinion scores, 10% for the metadata scores and 70% for the model scores. We have chosen relatively small weights for metadata scores because of their limited availability. Future work in this area may collect more precise metadata directly from the NSIs, which can lead to the updating of these weights. Currently, our Shiny app offers users considerable flexibility in setting the weights and testing the sensitivity of the resulting scores to these assumptions.

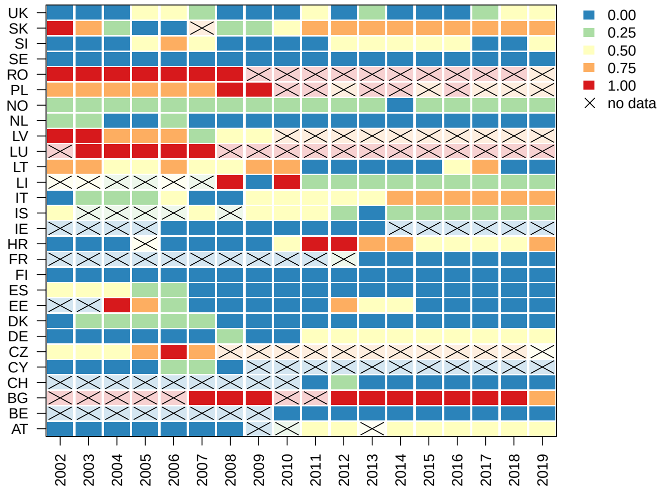
As was previously mentioned (in the expert opinions section), after 2007, there was a significant change in the processes of migration data collection and organization in the European Union following the implementation of Regulation (EC) 862/2007[8] which harmonized the definition of the duration of stay for all EU countries. Furthermore, migration statistics could have been retrospectively updated by using the 2011 round of censuses. In addition, the IMEM and the QuantMig expert opinions refer to separate time periods (2002—2008 and 2009—2019, respectively). For this reason, we have split the contributions of the IMEM and QuantMig expert opinions, metadata, and models into two periods. This cut-off is defined by a flexible threshold parameter (set to 2009).

Results

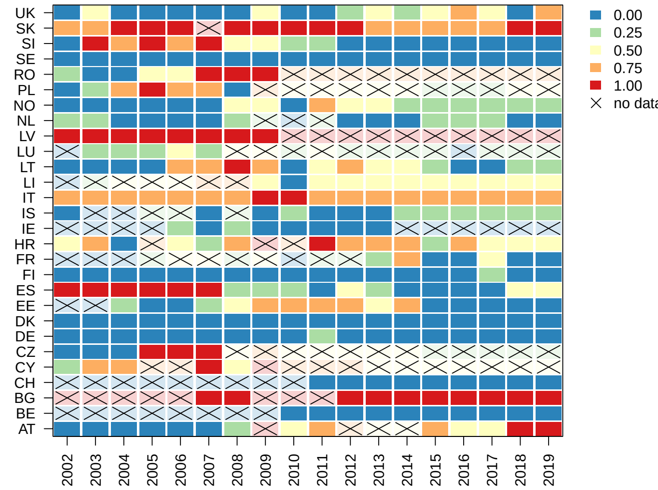
In Figure 2, we present six examples of bilateral flow ratios for Austria (AT), Bulgaria (BG), Germany (DE), Finland (FI), Italy (IT), Poland (PL), and Slovakia (SK). The first row (Figures 2a and 2b) shows bilateral flow ratios that are not corrected for the duration of stay, for the default model settings. We observe that there are many cases of overcounting in AT (before 2007) and DE (whole range of years) that are especially visible in the immigration data, whereas there is considerable undercounting in BG, PL, and SK. As these ratios are uncorrected for the duration-of-stay criteria, the observed levels of undercounting and overcounting result at least in part from the differences in how these criteria are applied in the compared countries. For example, AT has a three-month duration until 2006 for both immigration and emigration, while DE has a very short stay (from two weeks to two months) for immigration and even a “zero” duration for emigration. A shorter duration of stay than the standard 12-month criterion leads to the overcounting of migration flows. The opposite situation is observed in BG, PL, and SK. All of these countries use the permanent duration of stay to define a migrant (PL until 2008, BG until 2009, and SK for the whole range of years).

The effect of the duration of stay is mitigated when the correction for duration (factors R_X and R_Y in Eq. 1) is introduced (Figures 2c, 2d, 2e, and 2f). After the correction, there is no overcounting in the first years in AT, while in DE, there is overcounting only for emigration in the 2002—2004 period, and there is undercounting for immigration (2011–2019, Figures 2c, and 2e). Similarly, for the BG, SK, and PL data, the level of undercounting is substantially reduced after the correction. Both IT and FI use the 12-month definition for the entire 2002—2019 period. The introduction of the correction factors slightly reduces the level of undercounting for IT, as some of the reference countries have a duration of stay other than 12 months. Interestingly, the bilateral flow ratios corrected by our optimization model and IMEM model estimates of the duration-of-stay parameters seem to be almost identical for the immigration data (Figures 2c and 2e), but they differ considerably for the emigration data (Figures 2b and 2d).

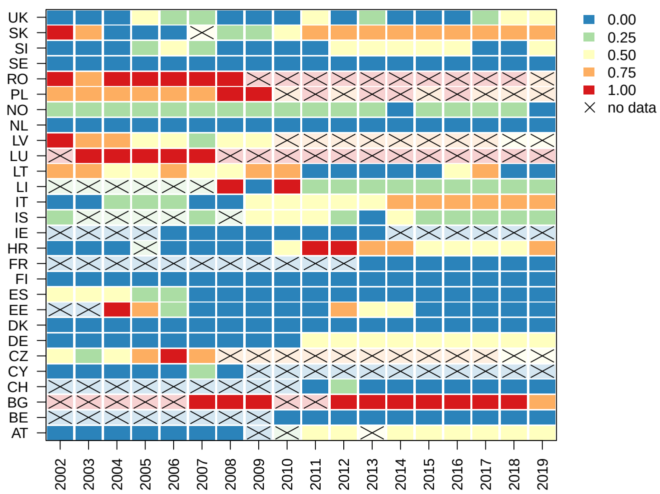
(a) Optimized duration-of-stay coefficients for immigration.



(b) Optimized duration-of-stay coefficients for emigration.



(c) IMEM duration-of-stay coefficients for immigration.



(d) IMEM duration-of-stay coefficients for emigration.

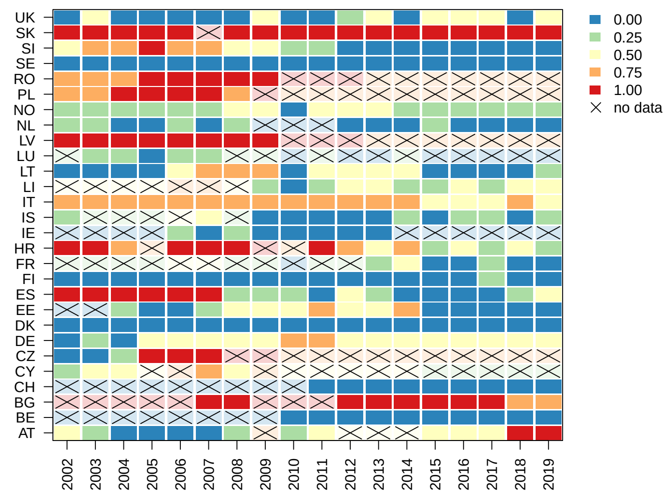


Figure 3. Undercounting scores predicted by the model calculated by projecting bilateral flows ratios into discretized 0–1 scale (quantile method). The figure is generated using *UndercountMigScores*[35] at the default settings. The columns show the results for the immigration and the emigration data, while the rows show the results for different sets of duration-of-stay correction coefficients. It is not possible to calculate some results (denoted as "X") due to the lack of country-specific flows in the considered country or in the reference countries. To fill these gaps, the model offers PCA imputations, which are shown in lightened colors in the figure.

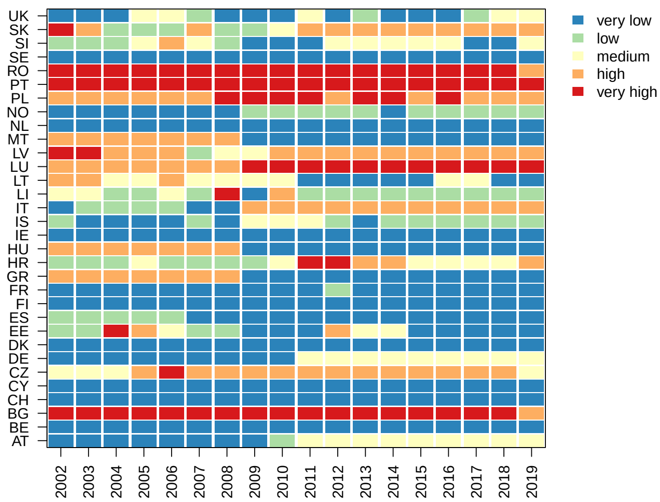
The optimization approach seems to reduce the undercounting issues more than the IMEM estimates; that is, the flows ratios are relatively closer to one.

Figure 3 shows the undercounting scores that result from projecting the bilateral flow ratios into a discreticized 0—1 scale. In this representation, we have opted for equally spaced intervals, which we consider to be the simplest solution. These results cover almost the entire range of Eurostat countries that provide bilateral migration flows in 2002—2019. They are qualitatively in line with the results presented in Figure 2. The scores for the immigration data obtained from the optimization corrections (Figure 3a) and the correction derived from the IMEM model (Figure 3c) are very similar. For the optimized correction parameters, the countries with the highest levels of undercounting include BG, RO (Romania), and LU (Luxembourg), followed by PL, LV (Latvia, except 2007-–2010; Figure 3a), SK (except 2004-–2010), HR (Croatia, in 2011-–2014 and 2019), and CZ (Czech Republic, in 2005–2007). Interestingly, IT has higher levels of undercounting since 2014. As expected, the Nordic countries, Belgium (BE), Switzerland (CH) and the Netherlands (NL) are characterized by very low levels of undercounting. It is, however, somewhat surprising that Cyprus (CY), Ireland (IE) and Slovenia (SI) also belong to this group. CY uses the

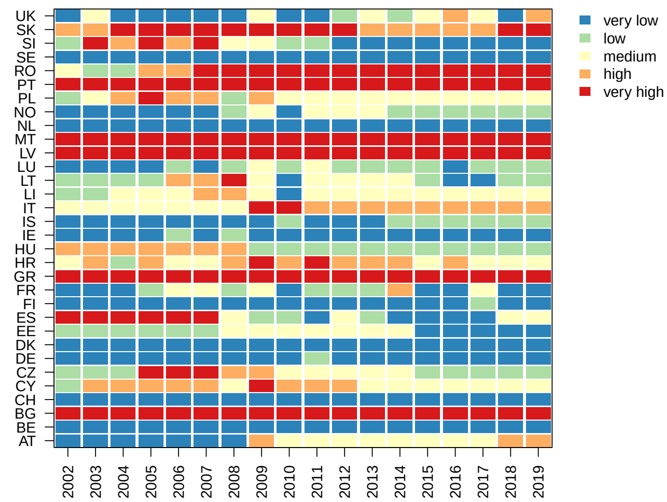
Passenger Survey as a source of data on migration. Even if the level of accuracy of the survey is low, there appears to be only a small undercounting bias in the migration estimates. IE and SI use different data sources (survey and register data, respectively). The levels of overcounting of immigration flows are found to be high in both countries (classified as low levels of undercounting in our model). These overcounting issues should be clarified by collecting more metadata from the NSIs.

The levels of undercounting are expected to be higher for the emigration data than for the immigration data because of the widespread problem of the failure emigrants to deregister[4]. Similarly to Figure 2, we observe that compared to the scores for the immigration data, the scores for the emigration data differ more between the optimized and the IMEM-based corrections for duration of stay(Figures 3b and 3d). The optimization model seems to correct more strongly for undercounting and overcounting problems. As in the case of the immigration data, the highest levels of undercounting are observed for BG, LV, RO (since at least 2007), SK, CZ (2005-–2007), and ES (Spain, 2002-–2007), followed by for PL (mainly for the IMEM correction), IT, and HR. Interestingly, the flow ratios for CY no longer show low levels of undercounting, while SI is found to have high and very high levels of undercounting in 2003–2007.

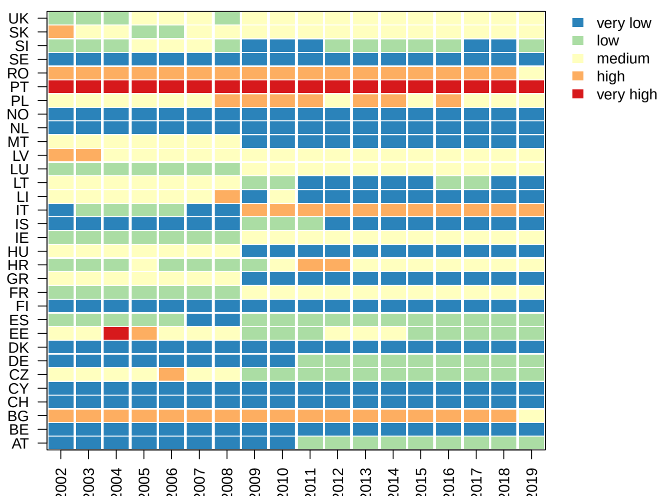
(a) Immigration. Default weights.



(b) Emigration. Default weights.



(c) Immigration. Alternative weights.



(d) Emigration. Alternative weights.

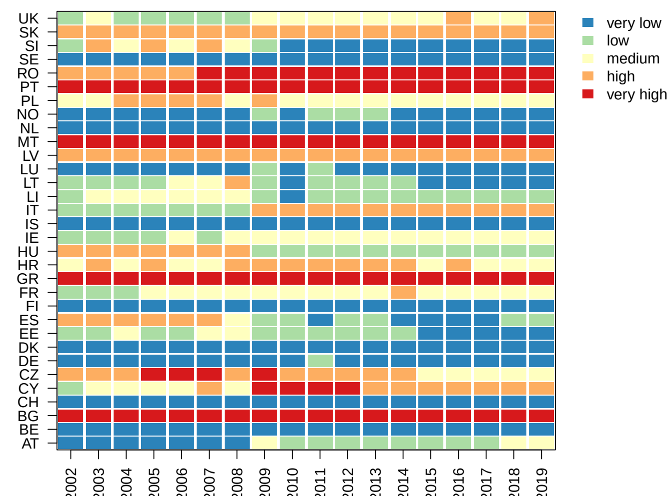


Figure 4. Classification of undercounting based on the weighted average of different undercounting scores. Optimized duration-of-stay coefficients. The figure is generated using *UndercountMigScores*[35] at the default settings. The columns show the results for the immigration (left) and the emigration data (right), while the rows show the results for different sets of mixing weights. Default weights (**a** and **b**) assume that expert opinions = 20%, metadata = 10%, and the model = 70%; alternative weights (**c** and **d**) assume that all weights are equal.

Figure 4 presents the undercounting classification for the combined (weighted average) undercounting scores obtained from the metadata, the expert opinions, and our model. We consider two cases. Figures 4a and 4b show the default mixing weights (20% for the expert opinion scores, 10% for the metadata scores, and 70% for the model scores), while Figures 4c and 4d show an alternative approach in which all weights are equal. While the results based on expert opinions and the results of our bilateral flow ratios model are quite similar, the scores based on the metadata are different, which leads to large differences between the upper and the lower panels in Figure 4. This is especially visible for the immigration data (Figure 4a and 4c). Our default selection of mixing weights (Figure 4a and 4b) results in similar undercounting scores, as shown in Figure 3. This is expected, because the share of the model scores in the total undercounting score is 70%, which is much larger than the 10% share for the metadata scores.

All panels of Figure 4 include additional countries that are not present in the previous figures due to missing data. These countries include GR (Greece), Hungary (HU), Malta (MT), and Portugal (PT), and their classification is entirely based on metadata and expert opinions. This also applies to cases in which only data for certain years are missing. Our Shiny app offers imputations for missing bilateral flow ratios.

Discussion

Producing reliable data on international migration stocks and flows remains one of the key challenges for national statistical institutes. Considering the high and increasing demand for migration data, international agencies such as the UN, the OECD, and EUROSTAT provide a wide range of migration estimates that rely on a variety of data sources and estimation methods. Unfortunately, the seemingly harmonized available data often lack documentation and a thorough assessment of their quality. Our study provides a comprehensive analysis of international migration flows in Europe, with a particular focus on undercounting over the past two decades. Our results are key inputs for statistical models (similar to the IMEM[4] or QuantMig[22] models) that harmonize migration flows for the European Union.

It is evident that misclassification of undercounting used in migration models can introduce bias into the estimated migration flows. If the level of undercounting is overestimated (i.e., is wrongly classified as “very high”) for a specific year in a given country, the model may exaggerate the estimated true migration flows. Conversely, underestimating the level of undercounting can lead to an underestimation of the fitted flows. Additionally, estimation bias can be extended to other countries and years with the same level of undercounting, thereby increasing the overall bias.

Indeed, undercounting is a serious issue that could lead to wrong conclusions being drawn and speculations being made about the true levels and directions of international migration. Establishing migrant data exchange systems between national statistical offices (currently functioning in the Nordic countries; there are also instances of data exchanges between, for instance, Poland and Germany, see[40,41]), and developing more comprehensive modeling approaches by incorporating expert information and measures of uncertainty, are two potential strategies for overcoming the limitations of migration data[42].

The main contribution of this study is to extend prior approaches to assessing the quality of migration data that are based on expert knowledge by establishing more objective and data-driven criteria that are able to account for differences in duration of stay. First, the proposed approach that we tested on the data from 32 European countries relies on the outcomes from the bilateral flow ratio model. This model compares the same migration flows reported by the country under study and by a set of “gold standard” countries with reliable register-based data. Second, metadata and expert opinions are used to supplement the data from which the country- and period-specific undercounting score is derived. The obtained final classification of undercounting integrates information from different classifications, and can inform crucial parameters for statistical models that harmonize existing migration flows and impute missing migration flows[4,6,22].

The assessment of undercounting provided in this study can be compared to the corresponding results of the expert-based assessments in the IMEM[4] and QuantMig[22] projects. In both projects, unlike in our estimates, the expert opinions assumed that the level of undercounting of immigration and emigration data are the same. Although there are differences in the classification methodologies and the numbers of assigned undercounting classes used, we found that the scores reported using the IMEM/QuantMig expert opinions are generally consistent with the annual average scores reported using the bilateral flow ratio model proposed in our study. However, this consistency usually concerns either immigration only or emigration only. For example, for Lithuania (LT) and Estonia (EE), the bilateral flows ratio model and IMEM/QuantMig expert opinion assessments coincide only for immigration (Figure 3a). In the case of emigration (Figure 3b), the model shows elevated levels of undercounting (high or moderate) during the 2006-2014 and 2008–2014 periods. The opposite pattern can be observed in Slovenia (SI), as the experts assume a high level of undercounting before 2009 and a low or moderate level of undercounting thereafter, whereas our immigration flow ratio model indicates a low level of undercounting before 2009 and a moderate level of undercounting thereafter.

Significant disparities can be seen between the expert-based and the metadata-based scores. As was discussed in the metadata section, it is important to note that metadata-based classifications currently have limited applicability, as metadata often rely on reports that capture narrow time frames only, or provide insufficient information to adequately describe all aspects of data quality concerns. Consequently, we believe that expert assessments tend to be more accurate than the corresponding assessments based on the available metadata. However, the main problem that arises with expert opinions is that they usually provide only a single classification for the entire period under consideration. Thus, such a classification is not sensitive to changes in the methodology and in the quality of the data. A similar limitation is also attributable to metadata that lack annual information. Moreover, the expert opinions of IMEM/QuantMig imply that there are nearly identical undercounting classifications for immigration and emigration data sources. Indeed, good quality and detailed metadata can complement the expert opinions by adding the time-specific context. A key advantage of our Shiny app is that it offers users the option to combine scores derived from metadata, expert opinions, and our bilateral flow ratio model.

The metadata are collected separately for the immigration (Table 2) and the emigration (Table 3) data sources. Unfortunately, they are collected mainly by analyzing the most recent reports (apart from the information on "auto-correction" in emigration flows made by the NSIs, which was taken from Eurostat 2003 report[32]). According to this report, of the countries under investigation, Bulgaria (BG), Cyprus (CY), th Czech Republic (CZ), France (FR), Greece (GR), Ireland (IE), Malta (MT), Portugal (PT), the Slovak Republic (SK), Slovenia (SI), Spain (ES), and the United Kingdom (UK) did not have a registration system covering international migration before 2002 (Austria introduced a registration system in 2002). Some of these countries may have already had registration or deregistration obligations and sanctions (neither of which are included in the report). This situation has changed according to the most recent metadata (see QuantMig Deliverable 6.2[30] for an overview), which show that BG, CZ, MT, SK, SI, and ES now use population or migration register data to estimate migration and impose a legal obligation to register. While the use of a single score for the entire 2002–2019 period is inevitable in the absence of year-specific metadata, it is an obvious oversimplification. However, even recently collected metadata are very limited and may not fully address the problems in national data collection systems.

The classifications based solely on the currently available metadata are found to be inconsistent with the opinions of the IMEM/QuantMig experts and the results of our bilateral flow ratio model for most countries. The most striking examples of misclassification are observed in the immigration data of Bulgaria (BG) and Romania (RO). In both countries, registration upon arrival is obligatory and the time limits for registration are tight; thus, the levels of undercounting in these countries are assumed to be low (Table 2). Both the expert opinions and the results of our model (Figures 3a and 3c) show the opposite classification.

In general, the approach and analysis presented in this article is merely a first step toward developing a more comprehensive framework for assessing the quality of international migration data in European countries. The current approach is dependent on the availability and the completeness of the data and information provided by Eurostat. In some cases, it is possible that the observed problems are attributable to other issues, such as discrepancies in the administrative or harmonization procedures used to transmit the data from national statistical offices to Eurostat. Thus, the proposed bilateral flow ratio model can be fully applied only to countries that submit bilateral migration data. However, even these data should be treated with caution due to the possible influence of other reporting problems, such as the systematic exclusion or undercounting of specific subpopulations like returning nationals.

Our work has confirmed that collecting more precise and detailed data on bilateral flows and filling in gaps in meta-information are key to gaining a better understanding of the overall quality and comparability of international migration data[5].In particular, more efforts to collect information about the procedures used by the national statistical institutes are needed before the data are submitted to Eurostat and other statistical agencies. One potential way to improve the metadata is to survey experts in national statistical offices who are responsible for migration data collection, asking them to provide precise (and retrospective) information on the national definitions that are currently used, and on how these definitions have changed over time. For example, these experts could provide insights and help to document the process of implementing EU Regulation 862/2007[8] on using the 12-month minimum duration-of-stay criterion. However, the validity of the metadata cannot be established by using only a formal approach together with a one-off survey. All of the available data sources on migration registration procedures that are used for official statistics purposes should be employed to validate the official migration data. In addition, comprehensive information about turning points in migration patterns, especially those resulting from changes in procedures and definitions, should be documented. The metadata built by using such an approach would provide the most coherent picture of data quality, as measured by the extent to which they match the existing data and knowledge. Our approach and the freely available Shiny application provide an opportunity to test the potential impact of alternative expert-based assumptions and new metadata on the final classification of migration undercounting.

References

1. Abel, G. J. & Cohen, J. E. Bilateral international migration flow estimates for 200 countries. *Sci. Data* **6**, 1–13, DOI: 10.1038/s41597-019-0089-3 (2019).
2. Azose, J. J. & Raftery, A. E. Estimation of emigration, return migration, and transit migration between all pairs of countries. *Proc. Natl. Acad. Sci.* **116**, 116–122, DOI: 10.1073/pnas.1722334116 (2019). https://www.pnas.org/doi/pdf/10.1073/pnas.1722334116.
3. Castles, S. Migration policies are problematic–because they are about migration. *Ethn. racial studies* **40**, 1538–1543 (2017).
4. Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W. F. & Bijak, J. Integrated modeling of European migration. *J. Am. Stat. Assoc.* **108**, 801–819, DOI: 10.1080/01621459.2013.789435 (2013).
5. Willekens, F. Evidence-based monitoring of international migration flows in Europe. *J. Off. Stat.* **35**, 231–277, DOI: 10.2478/jos-2019-0011 (2019).
6. Del Fava, E., Wiśniowski, A. & Zagheni, E. Modelling international migration flows by integrating multiple data sources. *SocArXiv* DOI: 10.31235/osf.io/cma5h (2019).
7. Del Fava, E., Wiśniowski, A., Dańko, M. J. & Zagheni, E. Modelling international migration flows by integrating multiple data sources. *(In preparation)* (2023).
8. The European Parliament & the European Council. Regulation (EC) No 862/2007 of the European Parliament and of the Council. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32007R0862 (2007).
9. Raymer, J. & Willekens, F. (eds.) *International Migration in Europe: Data, Models and Estimates* (John Wiley & Sons, Ltd, 2008). https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470985557.
10. Bijak, J. *Forecasting international migration in Europe: a Bayesian view (Springer Series on Demographic Methods and Population Analysis, 24)* (Springer Dordrecht, 2010).
11. Raymer, J., de Beer, J. & van der Erf, R. Putting the pieces of the puzzle together: Age and sex-specific estimates of migration amongst countries in the eu/efta, 2002-2007. *Eur. J. Popul.* **27**, 185–215, DOI: 10.1007/s10680-011-9230-5 (2011).
12. Abel, G. J. & Sander, N. Quantifying Global International Migration Flows. *Science* **343**, 1520–1522, DOI: 10.1126/science.1248676 (2014).
13. Martí, M. & Ródenas, C. Migration estimation based on the labour force survey: An EU-15 perspective. *Int. Migr. Rev.* **41**, 101 (2007).
14. Cesare, N., Lee, H., McCormick, T., Spiro, E. & Zagheni, E. Promises and pitfalls of using digital traces for demographic research. *Demography* **55**, 1979–1999 (2018).
15. Alexander, M., Polimis, K. & Zagheni, E. The impact of hurricane maria on out-migration from puerto rico: Evidence from facebook data. *Popul. Dev. Rev.* **45**, 617–630, DOI: 10.1111/padr.12289 (2019). https://onlinelibrary.wiley.com/doi/pdf/10.1111/padr.12289.
16. Wiśniowski, A. Migration forecasting using new technology and methods. In McAuliffe, M. (ed.) *Research Handbook on International Migration and Digital Technology*, 376–392 (Edward Elgar Publishing, 2021).
17. Alexander, M., Polimis, K. & Zagheni, E. Combining social media and survey data to nowcast migrant stocks in the united states. *Popul. Res. Policy Rev.* **41**, 1–28, DOI: 10.1007/s11113-020-09599-3 (2022).
18. Rampazzo, F., Bijak, J., Vitali, A., Weber, I. & Zagheni, E. A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom. *Demography* **58**, 2193–2218, DOI: 10.1215/00703370-9578562 (2021). https://read.dukeupress.edu/demography/article-pdf/58/6/2193/1428663/2193rampazzo.pdf.
19. Hsiao, Y., Fiorio, L., Wakefield, J. & Zagheni, E. Modeling the bias of digital data: An approach to combining digital with official statistics to estimate and predict migration trends. *Sociol. Methods & Res.* DOI: 10.1177/00491241221140144 (2023). https://doi.org/10.1177/00491241221140144.
20. Wiśniowski, A. *et al.* Utilising expert opinion to improve the measurement of international migration in Europe. *J. Off. Stat. (JOS)* **29**, 583–607 (2013).
21. Bijak, J. & Wiśniowski, A. Bayesian forecasting of immigration to selected european countries by using expert knowledge. *J. Royal Stat. Soc. Ser. A (Statistics Soc.* **173**, 775–796, DOI: 10.1111/j.1467-985X.2009.00635.x (2010).

22. Aristotelous, G., Smith, P. W. F. & Bijak, J. *Technical report: Estimation methodology. QuantMig Project Deliverable 6.3.* (The Hague: Netherlands Interdisciplinary, Demographic Institute (NIDI-KNAW)/University of Groningen, 2022).
23. Kupiszewska, D. & Wiśniowski, A. *Availability of Statistical Data on Migration and Migrant Population and Potential Supplementary Sources for Data Estimation. MIMOSA Project Deliverable 9.1A* (The Hague: Netherlands Interdisciplinary, Demographic Institute (NIDI-KNAW) / University of Groningen, 2009).
24. Van der Erf, R. F. *Typology of data and feasibility study. MIMOSA deliverable 9.1B* (The Hague: Netherlands Interdisciplinary Demographic Institute, 2009).
25. Poulain, M., Perrin, N. & Singleton, A. (eds.) *THESIM: Towards Harmonised European Statistics on International Migration* (Louvain-la-Neuve: UCL Presses, 2006).
26. Bradley, R. A. & Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39**, 324 (1952).
27. Turner, H. & Firth, D. Bradley-terry models in r: The bradleyterry2 package. *J. Stat. Softw.* **48** (2012).
28. Eurostat. Demographic statistics: a review of definitions and methods of collection in 44 european countries. https://ec.europa.eu/eurostat/documents/3859598/6851536/KS-GQ-15-002-EN-N/, DOI: doi:10.2785/717072 (2015).
29. Your Europe. Residence formalities. https://europa.eu/youreurope/citizens/residence/documents-formalities/reporting-presence/index_en.htm (2021).
30. Mooyaart, J., Dańko, M. J., Costa, R. & Boissonneault, M. *Quality assessment of European migration data. QuantMig Project Deliverable 6.2.* (The Hague: Netherlands Interdisciplinary, Demographic Institute (NIDI-KNAW)/University of Groningen, 2021).
31. European Commission. Residence formalities. https://ec.europa.eu/home-affairs/what-we-do/policies/legal-migration/fitness-check_en (2019).
32. Eurostat. Demographic statistics: Definitions and methods of collection in 31 european countries. https://ec.europa.eu/eurostat/ramon/statmanuals/files/KS-CC-03-005-EN.pdf (2003).
33. Eurostat. Immigration (migr_immi). Reference Metadata in Euro SDMX Metadata Structure (ESMS). https://ec.europa.eu/eurostat/cache/metadata/en/migr_immi_esms.htm (2022).
34. Eurostat. Metadata and quality report on european demographic and migration statistics. population (national level) (demo_pop). Reference Metadata in Euro SDMX Metadata Structure (ESMS). https://ec.europa.eu/eurostat/cache/metadata/en/demo_pop_esms.htm (2022).
35. Dańko, M. J. Undercountmigscores v0.8.1. assessing the level of undercounting in the internationalmigration flows reported by Eurostat. *GitHub repository* DOI: 10.5281/zenodo.6612951 (2023).
36. Poulain, M. Confrontation des statistiques de migrations intra-européennes: Vers plus d'harmonisation? *Eur. J. Popul. / Revue européenne de Démographie* **9**, 353–381, DOI: 10.1007/BF01265643 (1993).
37. Poulain, M. & Dal, L. *Estimation of flows within the intra-EU migration matrix. Report for the MIMOSA project* (Netherlands Interdisciplinary Demographic Institute, The Hague, 2008).
38. de Beer, J., Raymer, J., van der Erf, R. & van Wissen, L. Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe. *Eur. J. Popul. / Revue européenne de Démographie* **26**, 459–481, DOI: 10.1007/s10680-010-9220-z (2010).
39. Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM J. on Sci. Comput.* **16**, 1190–1208, DOI: 10.1137/0916069 (1995). https://doi.org/10.1137/0916069.
40. Nowak, L. *et al.* Poprawa jakości i dostępności statystyki migracji zagranicznych. Raport finalny, Projekt Nr 12. Central Statistical Office, Poland (2007).
41. Kostrzewa, Z., Nowak, L., Szałtys, D. & Witkowski, J. Kierunki doskonalenia statystyki migracji zagranicznych ludności. *Wiadomości Statystyczne* **5**, 1–26 (2010).
42. Willekens, F., Massey, D., Raymer, J. & Beauchemin, C. International migration under the microscope. *Science* **352**, 897–899 (2016).

Acknowledgements

We thank Emanuele del Fava for suggestions and comments on an earlier version of this article. Arkadiusz Wiśniowski acknowledges funding received from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870649, the project Future Migration Scenarios for Europe (FUME).

Data availability

The data used to estimate bilateral flow ratios are sourced from Eurostat and National Statistical Offices. Metadata and expert opinions can be accessed through the cited reports and papers. To replicate the results, the software *UndercountMigScores*[35] can be utilized.

Author contributions statement

E.Z. and M.J.D initiated the project, M.J.D developed the methods, performed calculations, plot figures, created the software and wrote first draft. M.J.D., A.W., D.J., D.A.J., and E.Z. tested software and wrote subsequent versions of the draft. All authors contributed to the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.