# The State of Migration and Mobility Data in Europe: a Systematic Assessment

M. J. Dańko[1,2], E. Del Fava[1], D. Jasilionis[1], D. A. Jdanov[1,3], and E. Zagheni[1]
[1] Max Planck Institute for Demographic Research, Rostock, Germany
[2] Department of Public Health, University of Copenhagen, Copenhagen, Denmark
[3] National Research University Higher School of Economics, Moscow, Russia
Email: danko@demogr.mpg.de

**Abstract**
The article assesses, in a systematic way, the availability and quality of migration and mobility data in Europe, with a focus on bilateral migration flows. It identifies and discusses the main issues associated with commonly used administrative data sources. These data problems can be grouped into four categories: definition of migrants, undercounting, coverage, and accuracy. An important component of statistical data is metadata. The article emphasizes the importance of metadata reliability and proposes methods for its improvement. Reliable metadata is an important element for the building of formal data quality criteria which are crucial for the modeling of migration flows. Systematic classification of data quality issues is an important task for producing reliable evidence base for research and policymaking. Ignoring potential biases and misinterpretations of problematic data may lead to misleading conclusions and recommendations in the area of international migration. This article is also intended to lay the foundations for the development of a broader project, the Human Migration Database, which aims at producing migration estimates of the highest quality, by incorporating all available data and metadata within a solid statistical framework.

**Keywords**
Migration flows, quality assessment, Eurostat, coverage, undercounting

## Introduction

In addition to national statistical offices, there are several international sources of migration data, including Eurostat, the United Nations international migration database, the OECD International Migration database, and the World Bank global bilateral migration database. While being important repositories of international migration estimates, these databases suffer from several limitations. Some of these databases are at least partial copies of each other, others provide only migration stocks or total migration counts, but lack information on bilateral flows. For European countries, the main and most complete source for data and metadata on international flows is the Eurostat Database, based on annual official data from national statistical institutes in Europe.

The methodology of data collection, and thus data quality, varies considerably by country and can change with time. The main goal of this paper is to assess the quality of available migration data for European countries as well as related metadata, with the main focus on the information collected and reported by the Eurostat. Thus, the geographical scope of the current report is concentrated on the European Union and few other European countries. This work is intended to lay the foundations for the development of a broader project, the Human Migration Database, which aims at producing migration estimates of the highest quality, by incorporating all available data and metadata within a solid Bayesian statistical framework.

## Main problems of available migration data coming from administrative sources

The data sources for the Eurostat database are national administrative sources, which include population registers, national surveys, censuses, border data collection systems and visas, residence permits, and/or work permits. In general, the quality depends on the country's migration/registration procedures, the legal incentives for registering the migration event, and the methodologies used by national statistical offices to measure migration. Administrative data may come from one or more registers, using different approaches and procedures for data collection.

Different data types can be combined, and different estimation methods can be applied. For example, censuses are sometimes merged with register data and can help to improve the migration estimates (e.g., Lithuania (LT) merges register data with the Population and Housing Census 2011). Importantly, approaches towards data collections may change in time leading to improvements or deteriorations in the quality of data. One of the biggest challenges is that bilateral flows that include information on previous and next country of residence are rarely recorded: often only total flows without disaggregation by country of previous or next residence are available (Figure 1).

The issues of administrative data can be classified into four categories: **(i)** definition of duration of stay, **(ii)** undercounting, **(iii)** coverage, and **(iv)** accuracy.

First, the minimum duration of stay defines a migrant, as it is the minimum length of time the migrating individual must reside inside or outside the country to be officially classified as an international migrant. This definition can vary not only among different countries, but also within a country by population group. For example, during the period 1998-2007 Denmark used a different definition of duration of stay for migrants coming from Nordic, European, and other countries. The definition of duration of stay can also change with time. The biggest change in the methodology occurred around 2008, when the migrant definition of minimum duration of stay was set by the EU to 12 months (Reg (EC) 862/2007). However, in some countries the duration definition did not change, rather *ad hoc* methodologies were implemented to follow Eurostat requirements (e.g., in Austria (AT)). Such re-estimation can potentially lead to a bias. It is also noteworthy that, with the change of length of stay policy, many countries stopped posting bilateral flows via Eurostat. Instead, only the total flows were published by Eurostat (see Germany (DE), Poland (PL), Czechia (CZ), and Luxembourg (LU) in Figure 1).

Second, the undercounting bias can be considered as a consequence of individual choices, occurring when people fail to communicate their relocation from one country to the other, i.e., people do not register when they in-migrate or, more often, do not de-register when they out-migrate. Probably, the most striking example of such behavior is Poland (PL, Figure 2A), where the emigration from Poland to Germany is highly underestimated. This problem also applies to other Eastern European countries, for example, Slovakia (SK) seems to have a similar scale of under-counting as Poland. In Lithuania, the undercounting of emigration has declined following administrative measures introduced in 2010, which consisted of a requirement for all residents to make regular compulsory health insurance contributions. The exceptionally high undercounting rates for PL in PL-DE migration data are likely the result of not only lack of de-registrations, but also the different definition of migration: PL considers migrants those with the intention of permanently moving, whereas there is no length of stay criteria in DE (1998-2008). See Figure 2 for more examples of potential undercounting.

Third, the impact of the population coverage reflects a systematic bias due to the rules that govern the data collection process, which may exclude certain population segments, such as nationals who are return migrants, or foreigners not being counted in the official immigration and emigration counts. For example, one country can have reliable de-registration of foreigners, but unreliable de-registration of nationals. Subpopulations such as asylum seekers, nomad populations, military personnel, homeless people, as well as some geographic areas may not be included in the migration data. As with other quality parameters, coverage may change over time. For instance, Belgium (BE) started to include asylum seekers in 2010 and further improved the quality of such data in 2011. Czechia (CZ) improved the quality of national migration data in 2011 and foreigner's data in 2013 (however, Eurostat metadata does not specifically provide information on these changes and improvements).

Finally, accuracy refers to the random, rather than systematic, error made in the data collection process and depends on the characteristics of the data sources used for the official migration information. For the registers, data accuracy refers to the chance of making random mistakes in the registration or de-registration process. Instead, the accuracy of survey data depends both on the chance of random mistakes when recording the information, as well as the sampling error.

**Main problems of the available metadata**

In general, metadata related to international migration must provide information on exact data sources, ways of data collection, and data processing methods. Metadata should also include year-specific changes in the definitions, methodology of data collection, and assessment of data quality. Our assessment of national and international data repositories suggests that such information on international migration are often incomplete, imprecise or even missing. For example, the metadata provided by Eurostat in many cases refers only to the most recent year except for the information on duration of stay, which includes year-specific records for many countries.

**Towards systematic classification of methodological data characteristics**

Systematic classification of data quality issues is an important task for producing reliable evidence base for research and policy making. Ignoring potential biases and misinterpretations of problematic data may lead to misleading conclusions and recommendations in the area of international migration. Therefore, a more comprehensive metadata on international migration is a crucial component in projects devoted to harmonizing and modelling migration flows such as IMEM (Raymer et al. 2013) and similar projects (Del Fava at al. 2019).

Our work is performed as a part of the Human Migration Database project aimed to provide consistent data on migration with the highest possible quality. The first stage of this sub-project is focused on a systematic analysis of existing metadata and includes efforts to improve their completeness and quality by collecting additional information from national data sources. In particular, these efforts are directed to fill in the information gaps related to time-specific contexts, focusing on changes in definitions and data collection procedures, as well as on potential data quality issues.

The second stage of the project is to construct non-arbitrary classification of the accuracy **(i)**, undercounting **(ii)**, and coverage **(iii)** of data based on the detailed assessments concerning different methodological issues as well as their time of occurrence and scale.

Specifically, accuracy **(i)** is classified according to **a)** primary sources of data on international migration flows, **b)** any discontinuity over time in the set of data sources used (e.g., switching between different sources or adding new sources, merging register with census data, etc.), **c)** the scale of destinations/origins misreporting.

Undercounting **(ii)** criteria are based on collecting the following information and data: **a)** are there any incentives or regulations requiring potential migrants to register/de-register when immigrating/emigrating? In particular, is de-registration or registration mandatory? Is de-registration done automatically by authorities? What are the time limits in each of these cases? Are there any benefits of registration for migrants? **b)** Is migration always reported in the year of the actual migration or it could be reported long after? **c)** Do migration statistics face a high rate of non-response for data on duration of stay abroad, so that returns after a short-term migration cannot always be properly separated from immigration?

Coverage **(iii)** is defined according to the following criteria: **a)** are all population groups covered equally by registration and de-registration? Are asylum seekers included in migration flows? Under which conditions are asylum seekers taken into account? In particular, is permanent/temporal residence of asylum seekers necessary? **b)** Are there any differences in the migration regulations for foreigners and nationals? Does data quality differ between nationals and foreigners?

**References**

Del Fava E., D. A. Jdanov, A. Jasilioniene, D. Jasilionis, and E. Zagheni. 2019. Integrated Modeling of International Migration Flows Using Multiple Data Sources. SocArXiv cma5h. DOI: 10.31219/osf.io/cma5h.

Raymer J., A. Wiśniowski, J. Forster, P. Smith, and J. Bijak. 2013. Integrated Modeling of European Migration. *Journal of the American Statistical Association, 108*(503), 801-819.

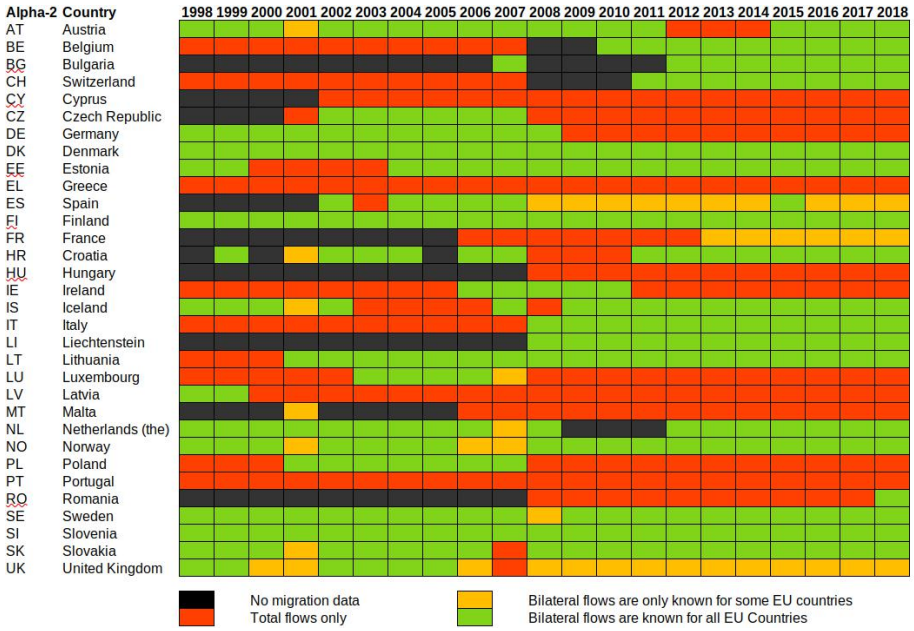


**Figure 1.** Emigration data quality chart for different countries (rows) and years (columns). Black marks years-country combinations with missing data, red - only totals flows are available, yellow – some EU countries are not listed as next residence but total flows are given, green – all EU countries are on the list of next possible residence.

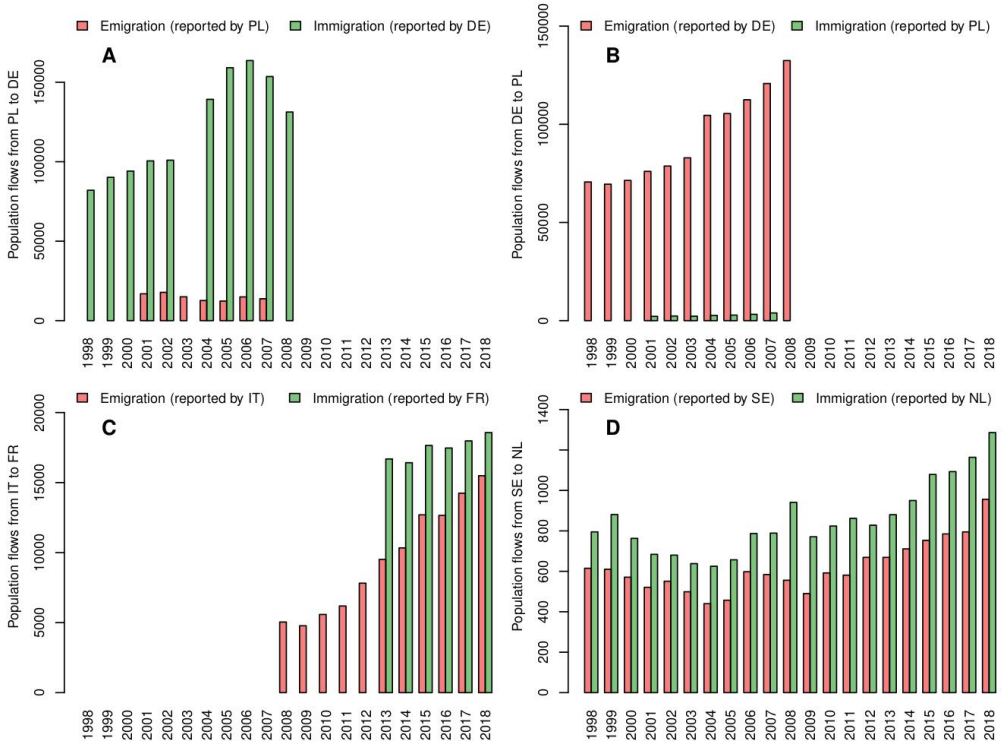


**Figure 2.** The origin-destination flows for 4 pairs of countries, broken down by immigration and emigration data sources. Bilateral migration flows are only available for a limited number of countries and years in the Eurostat database. **A.** Migration from PL to DE. On the one hand PL emigration seems to be highly under-counted probably due to problems of de-registration and permanent definition of duration of stay in PL, but on the other hand DE data may be overcounted due to criteria not based on length of stay for the definition of a migrant in DE for 1998-2008. **B**. Migration from DE to PL. Returning migration of Poles seems to be highly undercounted due to missing re-registrations or DE data is overcounted. **C**. Migration from IT to FR. Even though IT and FR have high quality migration data (register data and census data respectively), IT data seems to experience a de-registration problem. Moreover, emigration and immigration show different trends. **D**. Migration from SE to NL. There seems to be a problem with de-registration of people migrating from SE to NL.