The State of Migration and Mobility Data in Europe:

A systematic quality assessment of Eurostat migration data

M. J. Dańko[1,2], J. E. Mooyaart[3], E. Del Fava[1], D. Jasilionis[1], D. A. Jdanov[1,4], and E. Zagheni[1]
[1] Max Planck Institute for Demographic Research, Rostock, Germany; [2] Department of Public Health, University of Copenhagen, Copenhagen, Denmark; [3] Netherlands Interdisciplinary Demographic Institute (NIDI)-KNAW, University of Groningen, Groningen, Netherlands; [4] National Research University Higher School of Economics, Moscow, Russia
Email: danko@demogr.mpg.de

Abstract
The paper assesses, in a systematic way, the availability and quality of migration and mobility data in Europe, with a focus on bilateral migration flows based on the official statistics reported by the Eurostat. First, using real examples from specific countries, the paper discusses general data availability and comparability issues, including definitions, undercounting, coverage, and accuracy. Second, a thorough analysis of available metadata reported by national and international data repositories is performed in order to understand specifics of migration data collected by national statistical bodies. The study concludes that available metadata is often incomplete, imprecise, or even missing. The observed inaccuracies in migration statistics and lacking metadata may lead to wrong evidence for policymaking and should be taken into account in specific statistical models that estimate actual migration flows. The study lays the foundations for the Human Migration Database project, which aims at producing internationally comparable migration estimates on the basis of rigorous data quality assessment and advanced statistical framework.

Introduction
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, such as providing only migration stocks or total migration counts and not containing information on bilateral flows. Furthermore, many of these databases, at least partly, use data from the same sources. For European countries, the most complete resource including data and metadata on international migration flows is the Eurostat Database, containing annual data from national statistical institutes.

Since 2007, new EU regulations (Reg (EC) 862/2007) have aimed to increase the comparability of migration data across EU countries by ordering national statistical offices to collect and categorize data according to one unified standard. Although progress has been made over time, still, the methodology of data collection, and thus data quality, varies considerably by country. The main goal of this paper is to assess the quality of available migration data for European countries with the focus on the information collected and reported by Eurostat. Thus, the geographical scope of the current report is concentrated on the EU and EFTA 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
The data sources for the Eurostat database are the national official statistics 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 timeliness and quality of 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. The official statistics may come from one or more data sources, using different approaches and procedures for gathering the data. Different data types can be combined, and different estimation methods can be applied. For example, censuses are sometimes merged with register data to improve the migration estimates (e.g., Lithuania (LT) merged the register data with the Population and Housing Census 2011; Statistics Lithuania, 2017). Importantly, approaches towards data collection may change over time leading to improvements or deteriorations in the quality of data. Most migration data is based on information on country of citizenship or country of birth of the migrant. Unfortunately, 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).

There are two main issues related to international migration data. The first issue concerns data availability, i.e., when not all information on bilateral flows is recorded. The second issue is related to data comparability, i.e., to what extent data on international migration across countries are collected in the same way. The comparability issues of the official statistics can be classified into the four categories: **(i)** definition of duration of stay, **(ii)** undercounting, **(iii)** coverage, and **(iv)** accuracy.

First, migrants are defined based on a minimum duration of stay, which is the minimum length of time the migrant must reside inside the country of destination to be officially classified as an immigrant by that country. This definition can vary not only among different countries, but also within a country across different population groups. For example, during the period 1998-2007 Denmark used different definitions for duration of stay for migrants coming from the Nordic, European, and other countries. The definition of duration of stay can also change with time. The most important change in this definition occurred around 2008, when the EU introduced the 12-month minimal duration of stay criterion (Reg (EC) 862/2007). However, some EU countries did not change the duration definition; instead, *ad hoc* methodologies were implemented to abide by the new Eurostat requirements (e.g., in Austria (AT)). Such re-estimation can potentially lead to a bias. It also occurred that, after the EU regulation, some countries no longer shared information on bilateral flows with Eurostat, only providing the institute with the information on the total flows (see Germany (DE), Poland (PL), Czechia (CZ), and Luxembourg (LU) in Figure 1).

Second, the undercounting bias occurs when individuals do not report their arrival (immigration) or, more likely, their departure (emigration) to the registries or other administrative bodies. A remarkable example of under-reporting is Poland (PL, Figure 2A), as the emigration from Poland to Germany is highly underestimated in the official emigration statistics. This problem also applies to other Eastern European countries. For example, Slovakia (SK) appears to have a similar scale of undercounting as Poland. In Lithuania, the undercounting of emigration has declined following administrative measures introduced in 2010, which consisted of a requirement for all (including those *de facto* living abroad) permanent residents to make regular compulsory health insurance contributions (Klüsener et al., 2015). The exceptionally high undercounting rates for PL in PL-DE migration data are likely the result of not only lack of reporting emigration, but also differences in definition of migration: PL considers migrants to be those with the intention of moving away permanently, whereas there is no length of stay criteria in DE (1998-2008). See Figure 2 for more examples of potential counting bias.

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 reporting for emigration of foreigners, but unreliable reporting of emigration 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. Nonetheless, Eurostat metadata may 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 specifics of the data sources originally used for collecting the official migration information. For the registers, data accuracy refers to the chance of making random mistakes in the registration or de-registration process. On the other hand, 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, coverage, 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 is 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.

What is also unclear is whether the metadata are based on the information provided by the national statistical offices or rely on the external assessment performed by Eurostat. For instance, when the data for a particular country are described as "reliable", it is unclear on what grounds the data are considered reliable and who performed data quality checking. In broader terms, the metadata appear to lack transparency on how data from national statistical offices is assessed.

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 regarding international migration.

Our work is performed as a part of the Human Migration Database project, whose goal is to provide consistent data on migration of the highest 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. 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 consistent classifications of duration of stay **(i)** undercounting **(ii)**, coverage **(iii)**, and accuracy **(iv)** of data based on a detailed assessment of the different methodological issues as well as their time of occurrence and scale. Such assessment relies on the answers to a series of questions related to each of these four aspects. The questions regarding the **(i)** duration of stay are the following ones: **a)** What is the time definition of immigration and emigration? **b)** Are there any differences among the distinct sub-populations in this respect? **c)** Is migration always reported by date of occurrence or only by date of registration, including the delays? 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 or emigrating? **b)** Are de-registration or registration mandatory or are there any sanctions or benefits of registration/de-registration for migrants? **c)** Are there register-linkage procedures based on "signs of life approach" to identify the permanent place of residence of individuals with unclear migration status (e.g., persons who never de-registered living *de facto* abroad)? 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 considered? Is permanent/temporal residence of asylum seekers necessary? **b)** Are there any differences in the migration regulations for foreigners and nationals? Does undercounting differ between nationals and foreigners? Finally, accuracy **(ii)** is classified according to **a)** the primary data sources on international migration flows, **b)** the detection of the discontinuities 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.).

Together, the systematic assessment of Eurostat metadata, the gathering of additional data from the national statistical offices (especially when not disseminated by Eurostat), and the new systematization of the country-specific official statistics in terms of duration of stay, undercounting, coverage, represent a crucial component in the development of the next generation of models that, along the line of the IMEM project (Raymer et al. 2013), aim at harmonizing the available migration data to provide a more accurate picture of the international movement in Europe (e.g., Del Fava at al. 2019).

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.

Klüsener S., Stankuniene V., Grigoriev P., Jasilionis, D. 2015, The Mass emigration context of Lithuania: patterns and policy options. *International Migration*, 53(5), 179-193.

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.

Statistics Lithuania. 2017. Gyventojų tarptautinės migracijos statistinio tyrimo metodika [International migration survey methodology]. Vilnius: Statistics Lithuania. [in Lithuanian]. https://osp.stat.gov.lt/documents/10180/576507/Tarp_migr_metod.pdf

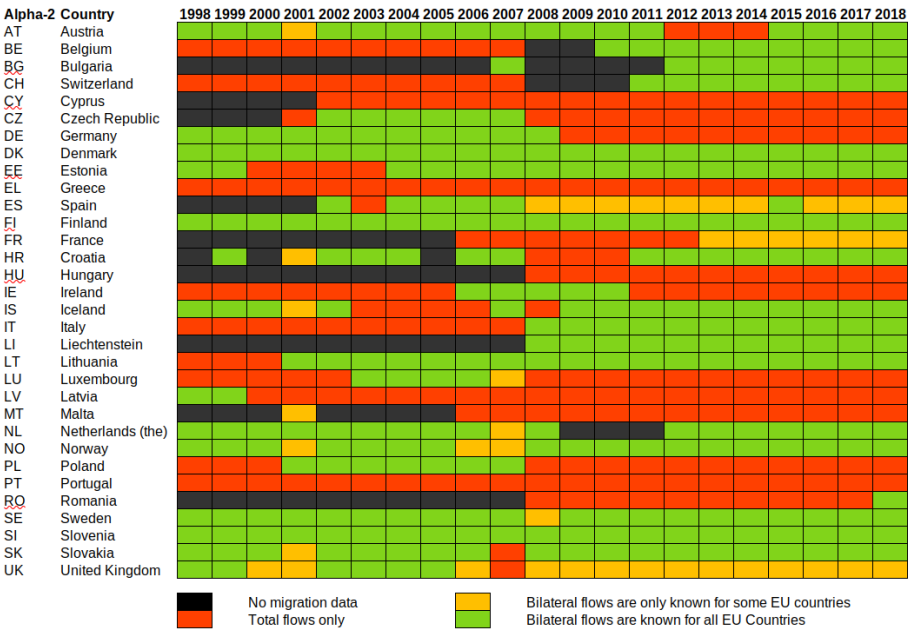

Figure 1. Emigration data quality chart for different countries (rows) and years (columns). Black marks years-country combinations with missing data; red - only totals flows are available; yellow – some EU countries are not listed as next residence, but total flows are given; green – all EU countries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

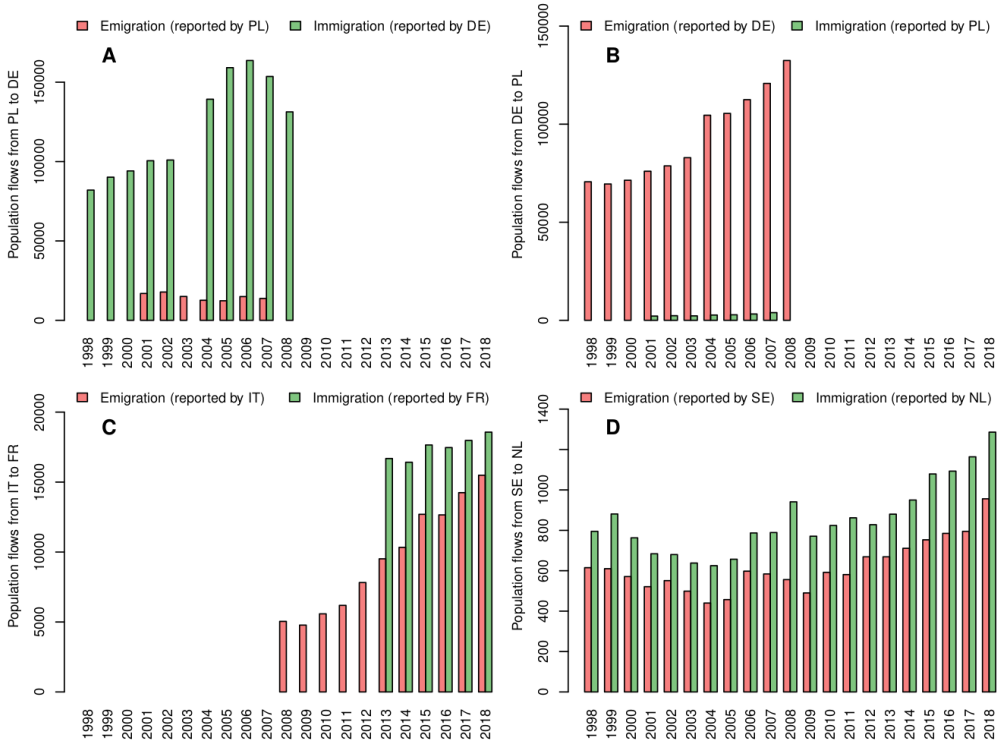

Figure 2. The origin-destination flows for four pairs of countries, broken down by immigration and emigration data sources. Bilateral migration flows are only available for a limited number of countries and years in the Eurostat database. **A.** Migration from PL to DE. On the one hand PL emigration 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.