Jarl Mooyaart, Akira Soto-Nishimura, Mathias Czaika and Helga de Valk

QuantMig Migration Data Inventory: Documentation

Deliverable 4.1

QuantMig has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299.

History of changes

Version	Date	Changes
1.0	29 Oct 2020	Issued for Consortium Review
1.1	30 Oct 2020	First version submitted as official deliverable to the EC
1.2	19 April 2020	Version for documentation website inventory

Suggested citation

Mooyaart J, Soto-Nishimura A, Czaika M and de Valk H (2020) QuantMig Migration Data Inventory: Documentation. QuantMig Project Deliverable D4.1. The Hague and Krems: Netherlands Interdisciplinary Institute (NIDI) and Danube University Krems (DUK)

Dissemination level

PU Public

Acknowledgments

This document reflects the authors' view and the Research Executive Agency of the European Commission are not responsible for any use that may be made of the information it contains. We are very grateful for useful comments and advice from Jakub Bijak, Sarah Nurse and Albert Kraler.

Cover photo: iStockphoto.com/Guenter Guni

Table of Contents

List of Tables

List of Acronyms

Organizations

CORDIS - The Community Research and Development Information Service - https://cordis.europa.eu/

EASO – European Asylum Support Office - https://www.easo.europa.eu/

ESS – European Social Survey - https://www.europeansocialsurvey.org/

EU LFS – European Union Labour Force Survey - https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey

Eurostat - Eurostat is the statistical office of the European Union - https://ec.europa.eu/eurostat

Frontex – European coast guard and border agency - https://frontex.europa.eu/

GESIS – Data Catalogue, Databestandskatalog (DBK), Leibniz Institute for the Social Sciences in Mannheim, Germany - https://www.gesis.org/

IOM – International Organization for Migration - https://www.iom.int/

IOM GMDAC – IOM's Global Migration Data Analysis Centre - https://gmdac.iom.int/

ILO – International Labour organization – www.ilo.org

IMISCOE- International migration, integration and social cohesion - https://www.imiscoe.org/

JPI Urban Europe – Joint Program Initiative Urban Europe - https://jpi-urbaneurope.eu

KCMD – Knowledge Centre on Migration and Demography - https://bluehub.jrc.ec.europa.eu/catalogues/data/

NORFACE - New Opportunities for Research Funding Agency Cooperation in Europe - https://www.norface.net/

OECD - The Organisation for Economic Co-operation and Development - https://www.oecd.org/

UN – United Nations - https://www.un.org/

UN DESA – United Nations Department of Economic and Social Affairs - https://www.un.org/en/desa

UNHCR – The UN refugee Agency – the United Nations high commissioner for refugees - https://www.unhcr.org/

World Bank - The World Bank is an international financial institution that provides loans and grants to the governments of low and middle income countries for the purpose of pursuing capital projects - https://www.worldbank.org/

Research projects

CrossMigration - Current European and Cross-National Comparative Research and Research Actions on Migration - https://migrationresearch.com/

ETHMIGSURVEYDATA - The International Ethnic and Immigrant Minorities' Survey Data Network - https://ethmigsurveydatahub.eu/

EMM Survey Registry – Ethnic and Migrant Minorities Survey Registry -

https://registry.ethmigsurveydatahub.eu/surveys

IMEM – Integrated Modelling of European Migration - http://www.imem.cpc.ac.uk/

MIMOSA – Migration Modelling for Statistical Analysis - http://mimosa.cytise.be/

Prominstat - Promoting Comparative Quantitative Research in the Field of Migration and Integration in Europe - http://www.prominstat.eu/drupal/?q=node/64

QuantMig - Quantifying Migration Scenarios for Better Policy

REMINDER - Role of European Mobility and its Impacts in Narratives, Debates and EU Reforms - https://www.reminder-project.eu/

TEMPER – Temporary versus permanent migration research project - http://www.temperproject.eu/ - http://www.temperproject.eu/wp-content/uploads/2019/12/Working-Paper-14-TEMPER.pdf

THESIM - Towards Harmonised European Statistics on International Migration - http://research.icmpd.org/projects/migration-statistics/thesim/

1. Introduction

To address migration challenges with solid evidence has sparked an increasing demand of researchers and policymakers alike to inform the public not only on current migration situations, but also for improving the evidence base of migration foresights. An important input for predictive migration modelling is quantitative migration data on migration flows and migrant stocks. Migration flow data register the number of immigrants and emigrants in a certain geographical unit (e.g. country) within a certain period of time, whereas migrant stock data give information on the number of people of migrant origin residing in a geographical unit (e.g. country) at a certain point in time. While there are many different sources providing such data, several challenges in assessing the reliability, comparability, and comprehensiveness of data remain. Not only do different datasets focus on different types or categories of migrants, they also use different definitions for what constitutes migration and who counts as a migrant. Some data are based on counting the event of migration moves (event-based) while others infer flow data from the change in the resident population in two points in time (transition based; see Rees and Willekens 1981). Moreover, migration stocks and flow data are hampered by the often-mentioned problems of coverage and consistency as each of potentially several individual moves of an individual must be identified, counted, and reported by respective data collection agencies (see Fassmann 2009; Raymer et al. 2013).

In the last two decades, there have been several attempts and significant improvements in harmonizing different data sources into larger cross-country migration datasets covering all EU and often non-EU countries. Concerning European migration flow and stock data, primarily collected by the national statistical offices of the EU member states, the European Commission has adopted regulations in order to improve the cross-country comparability of data in measuring immigration and emigration. These data are collected in the EUs' statistical office Eurostat database[1]. In addition, multiple research groups have made efforts to create harmonized cross-comparative datasets and using sophisticated methods to estimate and impute missing migration data. Examples of such projects that aimed to combine different sources of migration data and to estimate and fill in data gaps are the MIMOSA (Kupiszewska et al. 2009; De Beer et al. 2009) and IMEM (Raymer et al. 2012; Wiśniowski et al. 2016) projects.

Yet, while these efforts to combine data information sources are useful, there is still a whole range of other data on migration flows and stocks that are provided in different formats and by various data providers. There is a clear need for a more inclusive and more comprehensive overview of the migration data landscape, which gives orientation not only on the scope and comprehensiveness of existing migration flow and stocks datasets, but also on the level of detail and quality of these datasets. It is therefore important to provide an inventory of what unique data exist, and what exact type of data each of these datasets contains, supporting users of migration data to quickly find relevant information on European migration. Researchers, policy makers, journalists and other stakeholders may want to know what type of migrants enter and leave the EU and how migrants move within the EU. Users may be interested in data on specific entry categories such as

[1] Commission Regulation (EC) No 1178/2008 of 28 November 2008 amending Council Regulation (EC) No 1165/98 concerning short-term statistics and Commission Regulations (EC) No 1503/2006 and (EC) No 657/2007 as regards adaptations following the revision of statistical classifications NACE and CPA (Text with EEA relevance)

for asylum, study, work, or family reunification, but also in migrants' demographic characteristics including the age structure, gender composition, or educational backgrounds.

This is not the first time an overview of migration data is compiled. There have been several attempts before cataloguing or visualizing migration stocks and flows data. This meta-database is constructed to offer users a structured overview and access to the data sources that can help understand, explain and predict migration to, from and within the EU. This meta-database is explicitly taking a user's perspective in compiling and assessing the different data sources. Its purpose is not only to catalogue existing datasets on migration to, from and within the EU, but also to provide key information on each dataset that may be relevant for the user. We aim to not overload this inventory with information on each of the datasets, but rather strike a balance between relevance, comprehensiveness, and user-friendliness. This inventory also includes information on different quality criteria, which can help users to identify the appropriate datasets.

This data inventory forms the basis of two work packages within the QuantMig project, namely on internal EU mobility and migration to and from the EU (this background document is deliverable D4.1 which goes together with the data inventory overview which is deliverable D5.1). In addition to this background document (D4.1), all information on the data sources is summarized in a browsable database (D5.1). This documentation provides background information to the Excel-based data inventory. As migration to the EU and within the EU are both conceptually and empirically interlinked, we do not explicitly separate datasets on intra-EU migration from datasets with wider geographical coverage. However, we provide explicit information on datasets that contain information on intra-EU migration and mobility, so that the two deliverables can be identified. This data inventory on migration flows and stocks forms the basis for follow-up analyses of the QuantMig project on internal EU mobility (WP 4) as well as on evaluating of EU migration patterns since 1990 with a specific focus on the role of migration policies (WP 5).

This data inventory also aims to shed light on datasets that capture aspects of European migration that have received relatively little attention. For instance, a large and increasing group of residents in the EU comprises of third country nationals, many of whom are also mobile within the EU by crossing intra-EU borders. In the next section, we briefly review prior data overviews and indicate how this inventory builds upon them. Thereafter, we outline the design and structure of this data inventory, discussing its key elements including the data search strategy and selection criteria. In the last section, we introduce the data quality assessment and how it builds on previous approaches.

2. Previous inventories of migration data

Over the past decades, several data overviews have been compiled for various purposes. These include for example the migration databases constructed by the IMEM, MIMOSA and THESIM projects (see list of acronyms for abbreviations and links to the relevant information websites if available). The THESIM project created an overview of the data availability of international migration data and relevant metadata in EU countries in 2006, providing recommendations for harmonizing data collection, whereas MIMOSA and IMEM provided a one-off data harmonization effort that helped improve data coverage by producing migration estimates between 2002 and 2008 based on official migration statistics (mostly administrative data) as provided by national statistical offices and Eurostat. These estimates can help to fill in missing data and thus to make more accurate migration predictions. While these projects have provided

substantive information and insight on European migration patterns, they only included a part of all migration data available. Moreover, these projects, by and large, focused on general migration patterns within and to the EU, but not on specific migrant groups or categories, such as refugees or labour migrants. Also, the projects used data the first decade of the 2000s and have not been updated since. For realizing the aims of the QuantMig project, a comprehensive and up-to-date overview of available datasets is essential for analyzing recent patterns and future scenarios of European migration.

Several migration databases and meta-compilations have been compiled over the past decade, although not always specifically for the purpose of gathering (migrant) stock and (migration) flow data to improve migration predictions. Examples of such data inventories include the data catalogue by the European Commission's Knowledge Centre on Migration and Demography (KCMD) or the Migration Research Hub of the European network of migration scholars working together in the International Migration, Integration and Social Cohesion network (IMISCOE) (see list of acronyms for links to websites). These overviews help users in searching for relevant data, but in order to get a clearer idea of what is covered by the data, the related documentation needs to be studied first.

Moreover, there is a range of data overviews that have the purpose of visualizing migration stocks or flows. The largest data overview of this kind is the Migration Data Portal, provided by IOM GMDAC, which relies on and combines information from different databases, like the one by the United Nationals Department of Economic and Social Affairs (UN DESA). The IOM GMDAC database allows users to select specific countries and types of migration, such as family and forced migration. Another example is the flow monitoring and displacement tracking matrix of the International Organisation for Migration (IOM) that visualizes the number of displaced persons on a world map and provides a flow monitoring of irregular migration through the Mediterranean and Western Balkan routes. While these websites are very informative for users interested in the broader migration picture, they are not directly aimed for users interested in further analyzing contemporary or predicting future migration, as the combined data (data containing multiple years and origin/destination countries) is not easily accessible. Both data inventories and respective websites visualizing migration data mainly rely on global datasets, such as census/administrative data (e.g. Eurostat, UN DESA data) and large-scale survey data (e.g. the European Labour Force Survey (EU-LFS), also available through Eurostat.

Another branch of migration databases have been constructed as part of EU funded research projects. Example of such overviews include the Prominstat database, the REMINDER data inventory and EMM Survey Registry (see list of acronyms for links to websites). The data overviews of these academic projects are not only listing different migration data but are also providing substantive information on how migration is conceptualized and measured, and what types or forms of migration are covered. The Prominstat database as the oldest of the three has not been updated since 2008. The REMINDER data inventory is the outcome of a recently finished Horizon 2020 project, with a focus on mobility within the EU only. While Prominstat and REMINDER do contain some smaller scale surveys, they primarily cover official statistics and large-scale cross-country surveys. The EMM Survey Registry, produced by the ETHMIGSURVEYDATA network, has an exclusive focus on national surveys on migration. It provides detailed meta-information on each of the datasets and is a still ongoing project in 2020.

3. The design of the QuantMig data inventory

The aim of the QuantMig migration data inventory is to provide a comprehensive overview of all types of datasets, whether large administrative datasets or relatively small-scale surveys, that capture migration flows and/or migration stocks, capturing both migration *to*, *from* and *within* Europe. Following and expanding on approaches as also taken in earlier projects, like the REMINDER and Prominstat databases, we classify and characterize datasets. We do so by including information on the definition of migration or migrants and detailing indicators that are relevant in estimating migration flows and stocks, i.e. migration characteristics (citizenship/nationality, country of birth, country of destination and previous/next residence), demographic characteristics (age, gender, education, occupation). In doing so, we aim to lay the foundation for a framework or infrastructure in which future datasets on European migration stocks and flows can easily be added and included. In addition to listing and characterizing the different datasets, we also include a quality assessment based on objective criteria that address some fundamental aspects including coverage, timeliness, or completeness of the different datasets included in this inventory. The quality assessment builds on recent work of this kind and selects indicators that proved to be relevant in that context.

In summary, the objectives of the QuantMig data inventory are:

1. Creating an overview of all available stock and flow data covering migration to, from and within the EU
2. Providing a balance between comprehensiveness and relevance of the data overview with the aim of providing a useful and user-friendly interface
3. Including a quality assessment of the datasets

Before we describe the methodology of the data inventory, we first outline how we searched for datasets and what criteria the data should fulfill in order to be included in this inventory.

3.1 Search strategy for data included in the database

Given the extensive data overviews that have been compiled before, the first step entailed going through these existing databases. Our search started with the following meta-databases: Prominstat, KCMD's Migration Data Catalogue, IOM GMDAC's Global Migration Data Portal, REMINDER database, EMM Survey Registry, Migration Research Hub (CrossMigration), and the preparation of survey tools for merged dataset from the TEMPER project (an ongoing project that is focused on creating a repository of migration surveys and origin surveys, see list of acronyms for links to website and repository). Linked to this search, we queried the online statistical databases of international organisations providing migration data, i.e., United Nations Higher Commissioner for Refugees (UNHCR), the World Bank, the Organization for Economic Co-operation and Development (OECD), the International Labour Organization (ILO) and the International Organization for Migration (IOM). In addition, we extensively used data from statistical office of the EU, Eurostat.

While these sources cover most of the data that is currently available, additional searches have been carried out including:

- EU funded projects in the CORDIS database[2]
- Large funding agencies that cover migration topics in the EU, such as NORFACE and JPI Urban Europe
- GESIS data archive[3]
- Finally, we contacted a selection of migration data experts to ask for their assessment of completeness of our work and if needed identify datasets that were missed in our search so far.

3.2 Data selection criteria

Before outlining the structure of the data inventory, we specify the criteria for selecting data to be included in our data inventory. We basically aimed to include datasets that may help predicting and understanding migration towards, from, and within the EU. With that in mind, we defined the following requirements for selecting datasets:

1) Stock and flow data

In principle, we include any dataset (e.g. administrative/register, census, survey) that covers migration flows or stocks. Datasets contain one or more of the following characteristics of a migrant: 1) country of citizenship 2) country of birth 3) country of previous or next residence. Furthermore, we require that data can be analyzed using quantitative methods given the aim and scope of the QuantMig project. We thus exclude qualitative data that may provide additional insight but are too limited to offer a broader, cross-country picture of migration volumes and dynamics. We do not incorporate datasets that use 'digital trace' data such as social media or cellphone tracking data to capture flows of migrants. While this type of data is interesting and promising for many reasons, the usability of such data for the purposes of this database is limited. Most digital migration data stems from platforms such as Facebook, Twitter, Yahoo Email, etc. and therefore capture a selective type and group of migrants (namely those registered on such platforms). Additionally, datasets created with this approach are often not open access or included in data repositories, as the data is owned by the respective social media companies.

2) Country level

Datasets to be included in the QuantMig data inventory should thus have at least one variable that indicates the country of origin of the migrant, that is the country where the immigrant originates from, which we may refer to the country of birth, country of nationality or the previous country of residence, depending on the definition used by respective datasets. This also applies to destinations (countries of next residence) of emigrants. This does not necessarily mean that we only include data

[2] The CORDIS database lists and provides information on all EU-supported R&D activities, including programs (H2020, FP7 and older) projects - https://cordis.europa.eu/, which

[3] https://www.gesis.org/en/institute/departments/data-archive-for-the-social-sciences, which is a data archive for social sciences

at the country level. If data is also disaggregated by specific sub-national regions within countries, this dataset will also be included, as long as all regions within a country are covered. This implies that datasets that include only migrants from specific ethnicities or minorities (e.g. Roma) or specific regions or cities are not included in the QuantMig database. Finally, we only include data that cover two or more different destination countries, except for datasets that do not identify destination countries such as the Missing Migrants dataset[4]. Therefore, data from the national statistical offices from each EU country are excluded. While the national statistical offices may in some cases contain more detailed migration data for their country, we decided to not include these separately in our database as in principle relevant information should be available via Eurostat. Furthermore, in another EU funded project, extensive overviews of available national register data of European migration and stocks are provided. Thus, rather than replicating their work, we decided to refer to their findings (see the REMINDER project).

3) EU focus

The focus of this data inventory is on the EU countries; this implies that the data should cover immigration/emigration flows into, out of, or between EU countries or refer to the absolute number (stock) of immigrants in EU countries. We require a dataset to cover at least two EU countries as destination countries for it to be included. When it comes to emigration from EU countries, we only include datasets that cover emigration that are reported by EU institutions or EU based surveys. Even though EU emigration could be captured from an immigration perspective of the destination country, for instance the number of EU emigrants (EU citizens and other) who migrated to the United States in a particular year according to US immigration statistics, such detailed immigration data are not available for the vast majority of non-EU countries, and so, this information –even if available in some cases - is not captured.

4) Recent data

This data inventory is mainly constructed to help researchers analyze current migration with the aim to also analyze future migration flows. Hence, the focus is on data that are more recent. The inventory limits to datasets that include information on migration after 1 January 2000, even though we do note the complete time range for which the respective data is available. Hereby, migration pre- and post-2004 enlargement of the EU and related mobility within Europe is covered for, as well as different other types of migration within and to Europe.

5) Accessibility

The QuantMig data inventory aims to cover primarily those data that are openly accessible. At the same time, we do also include datasets for which we were able to locate an "access point" to the data (even though the data as such are not directly accessible). That is either a website from where the user can apply or an email address when it comes to providing access to data upon users' request.

[4] As stated on website: "Missing Migrants Project tracks deaths of migrants, including refugees and asylum-seekers, who have died or gone missing in the process of migration towards an international destination. Please note that these data represent minimum estimates, as many deaths during migration go unrecorded." - https://missingmigrants.iom.int/downloads

4. Key indicators of the data inventory

This data inventory aims to provide an overview of existing migration flow and migrant stock data for informing subsequent work within the QuantMig project, but also to assist other researchers and policy makers to identify and obtain different kinds of data useful for addressing their research questions or as input for their policy analyses. In constructing the inventory, we formulate key guiding questions that users may (usually) have when searching for migration data. The design and construction of the data inventory follows these questions:

What is the source of the data?

There are many different sources of migration data. Data are collected by national statistical offices and registries, European and international migration institutions (such as UNHCR, Frontex, IOM, EASO) and as part of scientific projects. While different actors collect data, much of the migration data are compiled into larger datasets. We make a distinction therefore between a data source and a data provider. The data source refers to the institution or actor that was responsible for the data collection, whereas the data provider is the institution or actor that combines the data and provides access to the data to users. For example, Eurostat is a data provider that combines secondary data from national statistical offices, making the national statistical offices the data source. In some cases, the data provider is the same as the data source, for instance in the case when data is collected and provided access to by a scientific project.

Different data sources collect different types of data. The most common types of migration data are register data, (population) census and nationally representative surveys. The way that data is collected has implications for the structure and design of the data. Register data provide numbers on migration flows and stocks and their related characteristics of the persons and households involved, whereas surveys collect microdata in which migration information is linked to an individual person or household, and they provide additional information on other (demographic) characteristics of a migrant (household). Furthermore, while most data are collected cross-sectionally, some data follow migrants on multiple time-points, i.e. they have a longitudinal design. In our database we therefore distinguish:

1. Type of data
2. Design of data
3. Whether the data represents a migration stock and/or flow
4. Level of aggregation of data: micro or macro.

Migration data can take very different forms and are not always presented in the same way. Register data is often presented in the form of tables that can be downloaded online by the user. Large international data providers such as Eurostat, OECD, UN etc., are examples of sources where users can download different types of data tables, and can (often) adjust these tables, as long as information is available and cross tabulations are allowed. Here we make the choice to generally treat each data table as a separate dataset, unless there are reasons to believe that two or multiple datasets represent the same data. The decision to treat separate data tables as one dataset, and therefore as one entry in our inventory, was most relevant to data from Eurostat and the ILO. We consider data tables from the same data provider to come from the same dataset if:

1. Two data tables contain the same information with the only difference being a more or less detailed operationalization of one variable. For example, if two datasets contain the same information, except for one data table having 5-year age categories, while the other having

15-year age categories, we only report the data with the more detailed information.

2. A data table is clearly a sub-dataset extracted from another data table. For example, Eurostat provides separate tables for youths (young adults), which are clearly extracted from the data tables that cover all age groups. In this case the data table that refers to the largest data population is chosen and included in the list.

What is the coverage of the data in terms of time range, destination and origin countries and types of migrants?

In our data inventory we cover individual EU destination countries and origin countries as distinguished by the original data. This database also includes EU emigration data, defined as migration *from* an EU country towards another EU or non-EU country. Although the focus for this data inventory is at the country-level, we also include information at the sub-national level, if entire countries are covered. Another aspect of coverage is related to time periods. We include information for what years data has been collected.

Furthermore, migrants have different reasons to migrate from one country to another, whether it is for work or for protection. Based on such reasons and often related entry permits, different migrant categories are distinguished, such as labour migrants or refugees. Categorization of migrants is often blurred as categorical boundaries are not clear-cut and migrants may 'belong' to several categories. For instance, a migrant may be categorized as a family migrant but may just as well work after migration and could thus also be classified as a labour migrant (cf. Bijak and Czaika 2020). This limitation applies especially to data on entry or residence permits as someone can only have one type of permit at a time. Here, we report whether the data distinguishes between different migrant categories or if it focuses on specific types of migrants, as defined in a particular dataset. With respect to data that focuses on migration by type we refer to the definitions in the respective data documentation reports and display this in our categorization. In cases where the dataset contains self-reported reason(s) we indicate that data includes reasons for migration, whereas in other cases we simply list the different types of migrants as defined by the dataset.

How is migration measured?

We apply the following definitions of a migrant's origin: first, the country of origin of a migrant is typically defined by (i) the country of birth, (ii) the country of citizenship or nationality, or (iii) the country of previous or next residence. In our dataset we therefore indicate whether information is available on these three migrant characteristics. Secondly, we consider someone being a migrant when residing in a certain country (de facto) or where someone is lawfully expected to live (de jure). Usually migrant status is determined by how long and with what purpose a migrant is residing in a country. Countries use different criteria regarding the time period as to when someone is considered a permanent or temporary migrant in their country, and varying definitions of time periods when someone is emigrated from a country of origin. Different definitions make comparisons between countries difficult, that is why there is an increasing demand for using similar definitions. For decades there have been efforts from the UN to harmonize international migration statistics, resulting in multiple recommendations for a shared definition on what constitutes a migrant. Currently the most widely accepted definition is the one formulated by the UN in 1998, which defines a migrant as *a person who moves from their country of usual residence for a period of at least 12 months*. Also, the European commission has taken steps to increase comparability between migrant statistics collected by the national statistical offices of EU member states. In 2006, the THESIM project catalogued the different migrant definitions used by national statistical offices

of the EU countries and provided recommendations for harmonizing migration data collection within the EU. This laid the foundation for several European regulations (such as Regulation (EC) No 862/2007), providing a legal framework which facilitated the construction of cross-national migration datasets, which continues to be updated with its latest amendment in 2020[5].

Yet, while these recommendations and regulations have improved comparability, differences remain across different data sources and data providers. In this data inventory, we therefore incorporate the specific migration definition as reported in the documentation, and the respective time period after which someone is considered and counted as a migrant, if this information is available in the documentation. Note that these harmonization efforts are mainly focused on general migration flow and stocks statistics, other datasets/surveys on specific migrant groups (such as students, temporary workers, asylum seekers) often use their own specific definitions, measures and concepts.

How recent are the data?

While data may be recently published, this does not necessarily mean that the data is very recent. There is always a time lag between data collection and data publication. Particularly for short-term forecasting (nowcasting) it is important to have the most recent data available. As mentioned above, we indicate all the years for which data is available. Apart from that, we also indicate whether the data is being updated or not.

How complete is the data?

While data may cover many origin and destination countries, there may be missing data entries for certain years or certain dyads (i.e. origin-destination combinations). This particularly applies to register data that aims to cover all migration moves and migrants in contrast to surveys which often apply to smaller samples. Yet for cross-national surveys there can be missing data in the form of some countries not participating in a certain year. Linking to previous points on how attempts are made to fill data based on different sources, still not all data points (year-dyad combinations) can always be filled. One can use existing data and respective estimates for imputing missing data, as done in the IMEM and MIMOSA projects. In some cases, national statistical offices and others like the World Bank (Artuc et al. 2014) also impute missing data on migration based on what is the most likely information; overall this is however done with reluctance and only when it is clear that this information can be trusted (a person e.g. is no longer living on a certain address).

We therefore indicate the level of completeness (proportion of missing data) in the dataset. Missing data are shown as a percentage for the entire dataset and for the least two years (time points) of the dataset. In datasets that are bilateral (i.e., country of destination and country of origin), we consider an observation to be missing when a destination-origin combination is missing. For example, if the dataset has Austria as a destination and Germany as an origin and Spain as destination and Portugal as an origin but does not have Spain as a destination and Germany as an origin, we consider this to be missing. Therefore, we consider completeness along the three dimensions destination, origin, and year. We do not consider countries that never appear in the dataset. We do also not consider completeness on other combinations such as sex, education, age,

[5] Regulation (EU) 2020/851 of the European Parliament and of the Council of 18 June 2020 amending Regulation (EC) No 862/2007 on Community statistics on migration and international protection

etc. This means that if in a dataset a country has separate information on sex (male and female migrants) and another country only provides information for total migration (without sex breakdown), we do not consider this as missing. For surveys we consider completeness based on participation. For example, if Austria participated in a survey in 2018 but not in 2019, we consider this to be a missing country-year.

What other characteristics of migrants are available in the dataset?

While all sorts of information on migrants could help predicting migration flows, we focus on characteristics that are most often used, which come down to basic demographic dimensions. These include (i) sex, (ii) age, (iii) educational level, and (iv) occupational status of the migrant.

What is the quality of the dataset? A quality assessment framework

There have been multiple ways in which the quality of migration data has been assessed. In the EC Regulation of 2008[i] there is a section on quality assessments. In this regulation, data from national statistical offices is assessed along the criteria of relevance, accuracy, timeliness, coherence, accessibility and comparability. In the IMEM project, migration data from different EU countries is assessed on their migration definition, accuracy and their coverage (Van der Erf 2010). In a recent work of Nurse and Bijak (2019) data covering Syrian migration flows are assessed on several objective and semi-objective criteria and ranked according to a traffic-light color scheme. In their quality assessment they incorporate the following dimensions: purpose (relevance), timeliness, trustworthiness, disaggregation, target populations, transparency, completeness and sample design.

In our data inventory, the purpose of the quality assessment is, however, slightly different. Rather than providing an overall score, we want to have an assessment that may help users in their decisions of selecting data. For some users, the overall completeness of a dataset may not matter as much when the dataset is excelling at a certain aspect, they are interested in. In other words, we leave the overall interpretation on the data quality more to the user but want to highlight the strengths of specific data in order to come to a balanced decision on usage. Furthermore, judging the datasets on their accuracy and reliability is beyond the scope of this inventory, for which we refer to the IMEM, THESIM and Prominstat scientific projects. Finally, as we include a very broad range of data and various types of data, we do not aim to provide an overall score for data quality. For instance, register data aiming to capture all migration flows for a given year will always provide a more accurate estimate of migration flows compared to a survey. However, surveys may contain more information on migrant characteristics and/or certain migration drivers, which may in a different way help predict future migration.

Consequently, we have assessed the data quality here along the following five dimensions:

- Geographical coverage, i.e. in terms of origin and destination countries
- Migration characteristics, i.e. citizenship, country of birth, previous or next country of residence, the more characteristics the better
- Completeness of data: (i) to what extent are cells of origin countries filled, with focus on most recent data. (ii) For survey data: to what extent are destination countries included for all country-years of repeated cross-sectional or longitudinal surveys
- Target type (migration type disaggregation). Does data distinguish different types of migrants (e.g. labour, family, student, refugee etc.), is it focused on one particular type of

migrant or is this unspecified

- Timeliness: The time lag between data collection and data publication

Finally, we consider each data table as presented by data providers as a separate dataset. This means that the quality assessment applies to the specific data table rather than the possible overarching dataset from which this data table is retrieved. For instance, a data provider may present separate data tables of which one contains information only on country of birth, while the other table contains information only on country of citizenship. Even if these data tables are retrieved from the same larger dataset (often not visible to the user) we apply the quality assessment to the data tables themselves. That is, the data tables as entered in the inventory form the unit of assessment. This means that in terms of the quality assessment of different migration background characteristics both data tables are judged to have only one migration background characteristic.

5. Overview of data inventory

The following overview lists the categories (columns) in the browsable data sheet and the definitions of these categories:

Name provider and source

*Dataset name- N*ame of dataset as described by the data provider.

*Data provider- W*hat institution provides the data?

Data source- If the providers and the source are not the same, where the data providers get the raw data from.

Basic characteristics of migration data

Data source type - The data inventory predominantly contains register, census and survey data. Some datasets contain information from multiple data types. In this case, we list all data types that are included in the dataset. For instance, some datasets (e.g. Eurostat) combine data from country register data with data from surveys when for some countries data from their registers are not available.

Data design - Whether data is cross-sectional (measured at one time point), or repeated cross-sectional (measurement at different time points, but different populations/samples) or longitudinal (same persons observed at multiple time points) and finally some surveys include retrospective migration histories (migration history as indicated by respondent). The combination of Data source type and Data design reveal also whether there is information on year of arrival and duration. For instance, country register data on number of immigrants arriving in a given year, will provide you information on how many immigrants from a certain origin arrived at that year, but will not reveal information on duration of stay, whereas retrospective histories will give you full detail on year of arrival and duration of stay, but only for a sample of individuals. Census data often also have some (although limited) information on duration of stay.

Stock or flow - Stock data provides information on how many individuals with a migration background are living in a certain EU country. Flow data indicates migration movements within a certain time period. This can be yearly statistics from registers, but also surveys in which respondents indicate retrospectively what their time of entering the country of destination was.

Micro data- In microdata, individuals have their own id identifier.

Migrant characteristics and definitions

We indicate (yes or no) whether the data themselves or the publicly available information on the data (e.g. documentation and reports) show or report information on:

Citizenship/nationality

Country of birth

Previous/next residence- (Residence before arriving in / after leaving the current residence).

Migrant type - We indicate if a dataset is focused on a particular type of migrant, for example: refugees, asylum seekers, students, labour migrants etc. In case there is no specific focus on a particular group of migrants, we indicate that migrant type is "not specified". However, if there is a dataset with no focus on a particular migrant group, but contains a variable with the reason of migration, we indicate this, stating that the data includes "reasons for migration".

Migration definition detail - If the information is available, the explicit definition as applied by the data providers on what constitutes a migrant is provided.

Migration definition short - A shorter simplified definition, i.e. country of birth, citizenship/nationality or previous/next residence.

Migration time definition - If applicable and information is available, minimum time period after entry when a person is counted as a migrant.

Coverage

Origin countries- Which countries are covered as origin countries

Destination countries- Which countries are covered as destination countries

Intra-EU- Whether (yes/ no) the data covers data on migration/mobility within the EU.

Aggregation level - Indicates on which level origin and destination are defined. Here, we distinguish national (country level) and subnational (regional or lower level). Subnational data will have country level aggregates, as else the dataset would not have been included in this data inventory.

Years - The temporal coverage in terms of for which years data has been collected.

Update status - Whether the data is still updated or not with data of more recent years.

Demographic characteristics

Sex - indicate whether this information is available in the data (yes or no)

Age - indicates whether there is information on is either measured continuously or in age categories

Labor – indicates whether there is information on employment status or the occupational status. For occupational status it is specified if the International Standard Classification of Occupations (ISCO) is applied or some other classification.

Education – indicates whether the International Standard Classification of Education (ISCED) is applied or whether some other categorization is used.

Quality assessment

Table 1 summarizes the criteria and thresholds applied in the quality assessment of the datasets. The quality assessment criteria are assessed and recorded as follows.

Q1 geographical coverage - If 100 or more origin and/or destination countries, we indicate 'high' coverage, and we indicate 'low' coverage if 10 or fewer origin and/or destination countries and medium coverage for those in between.

Q2 migrant characteristics - 'High' if all three dimensions are present (citizenship, country of birth, previous/next residence) – 'medium' if two out of three – 'low' if only one out of three.

Q3 Completeness - In terms of completeness we consider all available data, but also the data availability in the last two years. We give both characteristics equal weights (50%). For instance, if a dataset has 20% missing information overall and 10% in the two most recent years, we consider this data having 15% missing data. We provide a 'high' score for datasets that have a missing rate lower than 10%, a 'medium' score for a rate under 33.3% (lower than 1/3 missing data) and datasets with a higher percentage of missing data receive a 'low' score.

Q4 Target type - Here there is no high-medium-low categorization, but rather we distinguish
1) Multiple - data that distinguishes between multiple types of migrants or has a variable indicating the reason for migration.
2) Specific - if the data is on specific group of migrants (e.g. labour, refugee, student etc.) Also, if the data distinguishes different types of "displaced persons" (e.g. internally displaced persons, refugees, asylum seekers etc.) it is labelled specific.
3) Not specified - if the data makes no specification on different types of migrants.

Q5 Timeliness - Here we consider the time between data collection and data publication. We consider two factors. How recent the data is and the time lag between data collection and data publication. We consider 'high' timeliness if data collected in year T is published in T. It has 'medium' timeliness if the last data stems from T–1 and 'low' timeliness if the latest data is available for T–2 or older.

Table 1 Overview Quality Assessment Indicators

Quality assessment indicator	**High score**	**Medium score**	**Low score**
Q1 Geographical coverage	>=100 origin/destination countries	>10 & <100 origin/destination countries	<=10 origin/destination countries
Q2 Migrant characteristics	3 indicators	2 indicators	1 indicator
Q3 Completeness	<10% missing	<33.3% missing	>33.3% missing
Q4 Target type	Multiple migration types	One specific migrant type	Not specified
Q5 Timeliness	*T*	*T*–1	*T*–2 or earlier

Note: For surveys with only one round and datasets where missing rates were not possible to assign such as in the Missing Migrants dataset or where there are not countries of destination, we indicate not applicable, i.e. "NA". For cases in which we do not have access to the data, we indicate that an assessment for a particular criterion is "unknown".

Additional information

Missings - percentage of missingness in dataset as described above (last paragraph page 11).

Accessibility - the accessibility of the data (as in whether it 1) open access 2) can apply for access 3) limited aggregated data, apply for more detailed data 4) restricted 5) unclear on how to get access.

Link data – link to website where data can be accessed or applied for

Link report – link to reports written based on data from respective dataset

Link documentation – link to documentation/methodology of dataset

Notes – additional notes added by creators

Browsable database

The data inventory (QuantMig - Migration Data Inventory (soton.ac.uk)) is a browsable inventory with search filters and column selections. For more information on how to use the data inventory, see the instruction video

References

Artuç, E., Docquier, F., Özden, Ç., & Parsons, C. (2014). *A global assessment of human capital mobility: the role of non-OECD destinations*. The World Bank.

Bijak, J. and M. Czaika (2020). Assessing Uncertain Migration Futures: A Typology of the Unknown, QuantMig background paper D1.1, University of Southampton, UK

De Beer, J., R. van der Erf and J. Raymer (2009). Estimates of OD matrix by broad group of citizenship, sex and age, 2002-2007. Report for MIMOSA-project.

Fassmann, H. (2009). European migration: Historical overview and statistical problems. *Statistics and reality. Concepts and measurements of migration in Europe*, 21-44.

Kupiszewska, D., A. Wiśniowski, P. Ekamper, J. Bijak and J. Raymer (2009). Estimation of population stocks by broad group of citizenship, sex and age for 1st January 2002–2008. Report for MIMOSA-project.

Nurse, S., & Bijak, J. (2019). Data and Knowledge for Modelling Asylum Migration [Workshop paper]. Uncertainty and Complexity of Migration 2019, London, England.

Raymer, J., Forster, J. J., Smith, P. W., Bijak, J., & Wisniowski, A. (2012). Integrated modelling of European migration: background, specification and results (No. 2012004). Norface Research Programme on Migration, Department of Economics, University College London.

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W., & Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association*, *108*(503), 801-819.

Rees, P.H. & Willekens, F. (1981). Data Bases and accounting Frameworks for IIASA's Comparative Migration and Settlement Study. IIASA Collaborative Paper. IIASA, Laxenburg, Austria: CP-81-039

United Nations, Department of Economic and Social Affairs (1998). Recommendations on Statistics on International Migration, Revision 1. Sales No. E.98.XVII.14

Van der Erf, R. (2010). Initial Assessment of the Quality of International Migration Data. Discussion paper project" Integrated Modelling of European migration"

Wiśniowski, A., Forster, J. J., Smith, P. W., Bijak, J., & Raymer, J. (2016). Integrated modelling of age and sex patterns of European migration. *Journal of the Royal Statistical Society. Series A, (Statistics in Society), 179(4)*, 1007.

■

---

[i] EU Regulation quality assessment details:
REGULATION (EC) No 763/2008 article 6, Quality assessment

1. For the purpose of this Regulation, the following quality assessment dimensions shall apply to the data to be transmitted:

— 'relevance' shall refer to the degree to which statistics meet the current and potential needs of users,

— 'accuracy' shall refer to the closeness of estimates to the unknown true values,

— 'timeliness' and 'punctuality' shall refer to the delay between the reference period and the availability of results,

— 'accessibility' and 'clarity' shall refer to the conditions under and modalities by which users can obtain, use and interpret data,

— 'comparability' shall refer to the measurement of the impact of differences in applied statistical concepts and measurement tools and procedures when statistics are compared between geographical areas, sectoral domains, or over time, and

— 'coherence' shall refer to the adequacy of the data to be reliably combined in different ways and for various uses.