UN Expert Group on Migration Statistics

Task Force 3: Data Integration for Disaggregated Statistics on International Migration

Final Report

2025 (Unedited Version)

Contents

Chapter 1. Introduction 3

A. Data Integration in the context of the new conceptual framework on international migration statistics 3

B. Definitions of Data Integration 5

C. How data integration can improve migration statistics 7

D. Use of outputs derived from integrated data in official statistics 11

Chapter 2. Macro-Data Integration Methods 13

A. Overview 13

B. Statistical adjustment methods 17

C. Statistical modelling methods 20

D. New and hybrid methods 25

E. Challenges for macro-data integration 27

Chapter 3. Micro-data Integration Methods 30

A. Overview 30

B. Creating/Enabling the legal framework 31

D. Micro-data integration data linkage methodology 39

E. Challenges for micro-data integration 45

Chapter 4. Assessing and Communicating Results 47

A. Overview 47

B. Estimate assessment/validation 48

C. Communication with key stakeholders and dissemination of integrated data 50

D. Use of outputs derived from integrated data in official statistics 53

Chapter 5. Conclusions and Future work 55

References 61

Appendix I: Definitions of Key Concepts (A Glossary of Terms) 64
Appendix II: Task Force Membership..... 66
Appendix III (country case studies)..... 68

Chapter 1. Introduction

A. Data Integration in the context of the new conceptual framework on international migration statistics

Accurate and timely data are crucial to measure international migration patterns as well as to track migrant populations and their changes over time. Several global initiatives have stressed the need to collect, use, and improve migration data to develop evidence-based migration policies and mainstream migrants into national development planning.[1]

The new conceptual framework[2] on international migration and temporary mobility, developed by the UN Expert Group on Migration Statistics and endorsed by the UN Statistical Commission in its 52nd Session, recommend that migration measures and indicators could be improved by integrating multiple data sources and through strengthening the alignment between statistics on migration stocks and flows, and other components of change. Additionally, the framework stresses that it is important to distinguish between change in country of residence flows and temporary population flows, as statistics about temporary populations (such as seasonal workers, international students) need to be systematized and included in the broader picture of international migration and mobility. The need for information on the magnitude of migrant populations and their changes over time also imply the requirement of rich datasets covering demographic processes such as births, deaths, immigration, emigration, citizenship acquisition, and country of previous or next residence, sufficiently disaggregated by variables like age, sex, birthplace, and citizenship status. To meet these requirements, it could be necessary to combine

[1] The 2030 Agenda for Sustainable Development (A/70/1); the Global Compact for Safe, Orderly, and Regular Migration (A/73/195); International Migration Statistics, Report of the Secretary General, Statistical Commission, Fiftieth session, 2019.

[2] Developed by the UN Expert Group on Migration Statistics, Task Force 2, https://unstats.un.org/unsd/demographic-social/migration-expert-group/task-forces/TF2-ConceptualFramework-Final.pdf

data from multiple sources, such as census data with other types of data, such as registers on births, deaths, and migration / border crossings.

Data integration, which can be simply defined as the process of combining data from two or more sources to produce statistical outputs,[3] represents a useful strategy for improving the quality and availability of migration statistics. In many circumstances, the combination of multiple data sources can provide more timely, accurate, and granular data than relying on a single data source. Compared to the alternative strategy, which is new data collection, data integration also incurs significantly lower operational cost and respondent burden.

By leveraging existing data sources, data integration may lead to the production of migration statistics on past and current migration patterns which would otherwise be missing from typical data sources. Previous guidelines on data integration[4] and development of longitudinal data for migration statistics[5] have made recommendations to better utilize existing administrative data sources, such as tax filings, border control and visa records data, in combination with more traditional data sources, namely census, civil registration data, and household surveys.

A previous report by the UNECE (2019) found that many countries are interested in data integration as a strategy to produce better migration statistics; however, the practices and definitions employed by countries are extremely heterogenous. In the framework of the Expert Group on Migration Statistics, Task Force 3 on Data Integration seeks to support countries to produce sufficiently disaggregated data for the measurement of international migration by means of integration of micro- and macro-data techniques.

Towards that goal, this technical report aims to provide insights on a variety of methodologies used to integrate migration data at the micro and macro levels, drawing from both

[3] SDMX 2009
[4] UNECE-HLG MOS 2017
[5] UNECE 2020

academic and statistical literature, as well as country-specific examples (case studies) provided by Task Force members from National Statistical Offices (NSOs). The report will also consider the extent to which specific legal, policy, and technical contexts impact various data integration initiatives. Taken together, the report will provide examples so that countries can anticipate the major challenges and opportunities in mainstreaming the use of integrated data in official migration statistics.

B. Definitions of Data Integration

There is no internationally agreed-upon definition of data integration in the statistical community; however, previous initiatives such as the UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS)[6] have arrived at a general working definition of data integration as "an activity when data from one or more sources are integrated."

Specific to the context of data integration for migration statistics, a more recent report by the UNECE (2019) clarifies that data integration is "a statistical activity on two or more datasets resulting in a single enlarged and/or higher quality dataset." An "enlarged" dataset means more/better coverage of population, such as through combining data on various subgroups. "Higher quality" could be understood in some different ways, for example, it may mean that the content of the dataset is enriched (with more variables) or that the data becomes more accurate (with errors removed) or the estimates can be compiled in a more timely manner. Data integration may sometimes reduce coverage of the population and/or variables of interest, such as by removing duplication errors or by excluding unmatched observations. Thus, it is crucial to recognize that the goals for data integration projects can be quite heterogenous.

Data integration may be further defined by the levels of data that it utilizes, including micro- and macro-data integration. Micro-data integration refers to the integration of data at the

[6] https://statswiki.unece.org/pages/viewpage.action?pageId=169018059

record (individual) level, whereas macro-data integration combines data that are already aggregated to produce new statistics. The two broad subsets of data integration have different feasibility conditions. Micro-data integration is only feasible when record-level data from more than two sources are available *and* linkable with key identifying variables. In contrast, macro-data integration may use statistical outputs from multiple existing data sources, which are unlinkable, such as not having access to or missing personal identification information (PII), or less constrained by legal restrictions on data confidentiality. In other words, micro-data integration creates new combined datasets which can produce statistics on international migration, while macro-data integration includes methods to produce international migration statistics via the integration of aggregated data from multiple sources.

While data integration may seem like a new concept, some countries have employed integration methods to produce population statistics, including statistics on migration, for decades. Population registers are typical examples of integrated data at the micro level. As defined by the United Nations,[7] the population register represents "a mechanism for the continuous recording of selected information pertaining to each member of the resident population of a country or area, making it possible to determine up-to-date information about the size and characteristics of the population at selected points in time." Within a population register, information about an individual from various administrative registers such as birth, death, marriage, and migration are linked by personal identification variables, though in some cases the same information is included on several registers.

In many Nordic countries, for example, records are linked by a single personal identification number that is unique to every individual. When the complete register information about immigration and emigration are integrated, it is then possible to tally statistics about

7 https://unstats.un.org/unsd/demographic/sources/popreg/popregmethods.htm#:~:text=The%20term%20%E2%80%9Cpopulation%20register%E2%80%9D%20was,coordinated%20linkage%2C%20of%20selected%20information

migration stocks and flows from the population register. Information on age, sex, and socio-economic characteristics may be used to produce disaggregated migration statistics and to track longitudinal changes in the socio-economic conditions of immigrants.

At the macro level, demographic accounting is a common high-level strategy for producing population statistics by combining aggregated data on birth, death, and migration from different sources. In the United Kingdom, for example, the decennial census is a major source for producing statistics on the country's population count, but this only occurs every ten years. The country does not have a population register, which means that an annual count of the population cannot be derived directly. Instead, demographic accounting is used to produce annual statistics (estimates) during intercensal years, by subtracting deaths and emigrations and adding births and immigrations to the previous year's population count, which come from a variety of data sources.[8] Other more complex methods of macro-data integration exist, as will be discussed later in this report.

Beyond the simple examples above, data integration for migration statistics is a diverse and growing field, matching the growing need for migration data worldwide and the proliferation of new data and integration methods. This technical report seeks to provide an overview of integration methodologies, paying close attention to existing data contexts and practical strategies to assess the resulting integrated data and generated statistics. Drawing examples from both academic research and national practice, this report will enable countries to envision and improve on data integration projects that best suit their goals, data contexts, and technical capacity.

C. How data integration can improve migration statistics

Measuring migration is particularly difficult because migration is a process of change that rarely gets fully captured in a single data source. Few countries have complete population registers that

8 https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/populationestimatesfortheukmid2020methodsguide

allow them to track changes in the usual residence of individuals over time. Even with population registers, some shortcomings may affect the coverage and accuracy of migration statistics.[9] For example, a population register may only include its citizens or legal permanent residents, thus missing information about foreign residents, temporary non-resident movers, or unauthorized migrants. In cases where separate registers of foreigners are available, such as in Switzerland, integration efforts are directed towards linking the two registers. Countries also differ in requirements about updates in individual's status, for instance, failure to de-register is a common source of error that affects estimates of emigration, which might be alleviated by information (e.g "signs of life) from other linked sources.

The integration of multiple data sources at both the macro and micro level helps fill in gaps: an "enlarged" dataset could improve coverage of the migrant population, while "enriched" data typically increase the available information about migrants, as well as improving the timeliness, accuracy and level of detail (e.g. subnational estimates) for producing quality migration statistics. For example, one of the main impetuses for the United States to develop its Integrated Database on International Migration (IDIM) is to better estimate migrants for smaller levels of geography. In other words, use of one data source (e.g. administrative) to supplement inherent weaknesses in other types of data sources (e.g. survey-based estimates).

With appropriate statistical methods, data integration may also help fill the void for some missing migration measures, for example, to generate estimates of the undocumented migrant population in the United States,[10] or to create estimates of country-to-country migration flows from migration stock data and auxiliary information.[11]

[9] Poulain, Herm & Depledge 2013
[10] Van Hook et al 2016
[11] Abel 2017, Azose & Raftery 2019

Data integration can also help produce migration statistics for migration-related events which regular data cannot adequately measure due to measurement lag or newly emerging concepts of migration. This is often in response to new user or policy needs, perhaps due to changes in migration patterns resulting from natural disasters or pandemics, humanitarian crises, or extreme changes in national migration policy.

Per the United Nations' Principles Governing International Statistical Activities, official statistics are defined as statistics produced by government agencies, which can inform debate and decision making both by governments and the wider community. It may also be helpful to consider that the production of official migration statistics is a reiterative process where the goal is not only to produce one single set of official statistics, but also to assess the quality of published statistics and periodically revise and improve both data collection and processing to better meet key stakeholder needs.[12] Data integration can contribute both directly and indirectly to this goal. Data integration may also involve adding new data sources to help refine an existing data estimation procedure, for example in cases where external shocks (hurricanes, global pandemics) disrupt the usual migration flow estimates, such as between the United States and Puerto Rico after Hurricane Maria in 2017.

Further, an important "side effect" of data integration is the triangulation of measurements with data from multiple sources, which allows statisticians to evaluate and improve original data sources. In fact, some countries consider this to be an important reason for initiating data integration projects. For example, the Unified Migration Data Analytical System (UMAS) in Georgia was introduced in 2016, and operational since 2019, with data quality improvement as its major goal. The data integration process in Georgia has highlighted inconsistencies and errors in

[12] Raymer et al 2015

the original data sources and led to various recommendations for agencies to improve their own data collection and data processing procedures.

Data integration at the macro level is primarily focused on creating new statistics from outputs of existing datasets. In many cases, macro-data integration is the only possible strategy because microdata are not available or of insufficient quality, or not accessible due to legal constraints in data confidentiality and privacy. Macro-data integration methods are as diverse as the types of input data that it utilizes and can be motivated by many different factors. Simple projects may involve compiling various components of migrant stocks or flows from different data sources. More complex projects, such as those trying to capture specific hard-to-measure migrant groups, like undocumented migrants, or indicators, like migration rates, need to leverage a large range of statistical methods. Further, macro-data integration may introduce unconventional data sources such as flight or social media data to improve migration estimates when existing sources are inadequate.

At the micro level, data integration typically provides richer information about migrant populations. Different types of data can be combined at the record level, including for example administrative-only registers (e.g. tax records, social benefits, national health insurance, etc.), population registers (which already include multiple administrative registers), surveys, and censuses —as long as they include some PII to allow linkage, and at least one dataset includes a migrant identifier. Integration projects may produce longitudinal datasets which can be updated when new data become available, and this helps to understand long-term changes in the migrants' life and migrant population. Additionally, even simply combining data from cross-sectoral sources can help address important macro-level policy questions, such as migration integration indicators (language acquisition, naturalization), economic impacts of migration (education, tax filings, skills brought by return migrants), and welfare and health system impacts (usage of health care and social benefits).

D. Use of outputs derived from integrated data in official statistics

Integrated migration data can help fill many important gaps of missing data in official statistics. Centralized population registers are good examples of how integrated microdata are used to produce important statistics on migrant population and migration flows. In many cases, foreign residents and temporary population are not included in general population registers, and therefore some integration activities are needed to combine information on those subgroups to general registers. In the case of Switzerland, both foreign residents and temporary population are included in general (local) population registers, as well as in specific (federal) foreigner registers, thus both are combined to produced information on the general population.

In the absence of population registers, micro-data integration projects can be very resource intensive due to the technical expertise and infrastructure required both to integrate data and to continuously update the new integrated database when new information becomes available. The Longitudinal Immigration Database (IMDB) in Canada is a good example of such resource investment. The longitudinal database provides extensive data on the migrant population, as well as long-term changes in migrant's social-economic status after landing in Canada over a time span of greater than 35 years. Once established, the maintenance and subsequent incorporation of more data sources into the database are considered more cost effective than new data collection efforts.

Some countries have few available microdata sources to work with, thus have used macro-data integration methods to develop official statistics for certain forms of migration for a long period of time. For example, emigration statistics are typically incomplete, thus countries such as Canada develop estimates of emigration and net international migration flows based on aggregated outputs from multiple data sources.

Some challenges remain to be considered for migration data integration. First, the statistics produced from integrated data may be incompatible with statistics compiled from non-integrated

sources, demanding both producers and users of statistics to be mindful about their differences. Second, some countries may be resistant to use estimates from integrated data as official statistics. This resistance could come from different sources (NSOs, policy makers, political lobbyists/think tanks, etc.) and there could be many different reasons for this, including concerns about data quality, reconciliation of different estimates from different sources, political implications, etc. Switzerland, for instance, does not use any statistical estimates in official statistics, as they rely on exhaustive and high-quality register data. Third, as previous reports have found, countries may lack the required technical expertise to complete integration projects.[13] The international community can provide some help to initiate data integration projects, for example, one of the first attempts to integrate migration data in Moldova in 2014-2019 was supported by the United Nations Population Fund (UNFPA) and the Swiss Agency for Development and Cooperation (SDS).

[13] UNECE 2019, UNECE 2020

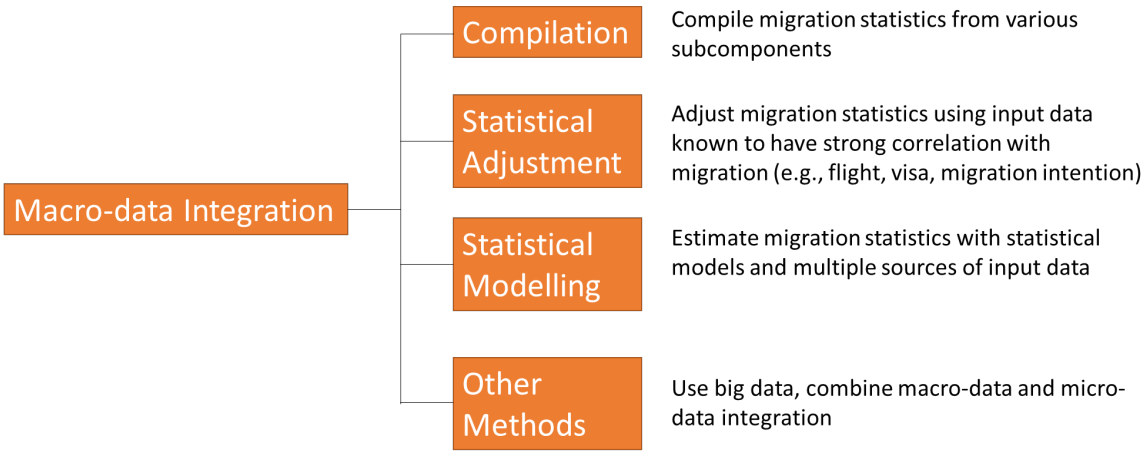
Chapter 2. Macro-Data Integration Methods

A. Overview

Macro-data integration refers to the process of combining data which are aggregated from individual-level records. In other words, it integrates outputs of various data sources to generate new migration statistics based on other existing outputs. Emigration statistics constitute one such example, as many countries do not keep track of departing individuals with sufficient information (such as duration of time away from country, country of destination, etc.). Additionally, certain subpopulations such as children, temporary non-resident population, and undocumented migrants are difficult to measure in most data sources. Macro-data integration is a strategy for combining data from various sources to better meet the needs of data users, both by making missing migration statistics available or by providing better coverage and more robust estimates of the target population.

Macro-data integration is frequently used when microdata are unavailable, either because they are not collected or inaccessible due to inadequate PII for linkages, limited resources, legal restrictions regarding data confidentiality and privacy, lack of coordination between agencies, or for other reasons. As noted earlier, some countries might have used macro-data integration techniques extensively without calling them "integration." Many examples come from countries with sparse population data, for example, when most population statistics are drawn from censuses which have an extended time lag (sometimes over 10 years) in between. Macro-data integration methods are thus diverse as they may rely on multiple types of available input data, including administrative data, survey data, as well as aggregated data from the private sector (e.g., flight records).

Figure 1. Categories of macro-data integration



In this report, we divide macro-data integration methods into three broad categories: (1) additive or "compilation" methods, (2) statistical adjustment methods, and (3) statistical modelling methods according to Figure 1 above. A fourth category is "other methods," like the use of big data or a combination of macro- and micro-data integration, but few practical examples of these methods exist in the production of official migration statistics.

Compilation methods refer to integration techniques that combine known subcomponents to produce migration statistics. For example, combining information on foreign work permit arrivals, with information on foreign student arrivals, with information on asylum seekers arriving during a particular period of time, to come up with a single immigration estimate. In contrast, statistical adjustment methods are used when the migration measure is available but lacks accuracy or timeliness. Auxiliary data sources with strong correlation to the measure of interest may be used to adjust the statistics. For instance, flight data may be used to correct estimates of migration flows in cases where flight records and migration flows are found to be strongly correlated.

Finally, statistical modelling methods refer to the broad range of techniques used when both the migration statistic in question and its subcomponents are completely missing from one or more datasets. In these cases, information from one dataset is used to supplement/enhance data

missing or of low quality in a second dataset, or shared characteristics in both datasets can be used to estimate characteristics missing from one of the datasets. Migration flow data represent one such example. When migration flow data are missing, it is possible to develop models to borrow strengths from available data sources, such as migration stocks, country-to-country economic relationships, and cultural proximity[14] to estimate the missing statistics.

In practice, a macro-data integration project may combine two or three categories of methods described above. For instance, emigration flow statistics may first be compiled using various known components and then statistically adjusted with auxiliary data sources. Statistical modelling may sometimes be added to provide detailed measure of uncertainty (e.g., the confidence interval for an estimate) or to estimate characteristics of missing populations (e.g. emigrants). While relatively rare, some projects may combine both macro- and micro-data integration techniques.

The obvious advantage of macro-data integration methods is that most aggregated data are already available, sometimes even for public access. This means that estimates based on integrated macrodata may be produced relatively quickly to provide more timely insights on migration patterns than what can typically be used (e.g., annual household surveys or decennial censuses).

The key challenge for macro-data integration projects is that input data from various sources typically come already aggregated by data providers. Statisticians thus have little control over differences in concepts and operational definitions, as well as potentially diverse coverage and time frames used in each source. Therefore, many integration projects require extensive assessments of data sources to confirm they are suitable for the project goals and chosen integration methodology. Alternatively, data users need to accept (and document) conceptual differences

[14] Raymer et al. 2019, for instance, use country-pair characteristics such as whether two countries share a common language, common colonial history, and land borders (contiguity) as inputs for estimating migration flows. For another example of cultural proximity, see Lanati and Venturin (2021).

between data sources, and simply use new figures or make adjustments to one dataset based on trends seen between data sources, knowing this is the best estimate that can be made given the data user has no control over data inputs.

Immigration statistics from another country may constitute the necessary subcomponents to compile emigration estimates, also known as "mirror statistics." This practice is common amongst neighboring countries (such as some countries in the Latin America region[15]) or groups of countries with strong collaborative agreements (such as countries in the European Union). For instance, in Latvia, emigration statistics are compiled based on immigration statistics from other countries, namely Denmark, Finland, Sweden, Norway, Spain, the Netherlands, Austria, Iceland, and Germany.[16] Another example is Chile, who estimated Chileans living abroad by asking other NSOs about the total number of Chileans in their country according to census and consulate surveys[17]. A notable limitation of this practice is that sometimes it is not possible to obtain information about country-specific origins and destinations due to missing or inaccessible data, while the timeliness of censuses, definitions used to measure migration, and availability of information is not the same for each country.

Compilation methods are also used to improve the accuracy of migrant population data, by adding unaccounted groups and subtracting populations who do not meet the definition of a migrant (e.g. non-residents). For instance, international students are often considered temporary visitors although their length of stay can exceed one year. In terms of subtraction, migrants who have obtained citizenship need to be removed from the foreign citizen-migrant stock population count.

[15] See https://celade.cepal.org/bdcelade/imila/
[16] UNECE 2019
[17] https://chilesomostodos.gob.cl/chilesomostodos/documentos/segundo-registro-de-chilenos-en-el-exterior-dicoex-ine

The U.K.'s Office of National Statistics started a new data integration project in 2019 to produce Administrative-based Migration Estimates (ABMEs) with an initial plan to integrate data representing various subcomponents, including (1) long-term migrants with a National insurance number who have resided in the U.K. for at least 12 months, (2) estimates of international students enrolled in first-year courses who might not yet have a National insurance number, (3) estimates of the proportion of migrants who obtained citizenship, and (4) estimates of long-term migrant (including international student) emigration. The in-progress project has reconciled various input data sources to account for both migrant status and history of migrants' activity (whether they remained in the U.K. for over 12 months and whether they changed status). Statistical modelling is used to fill in some gaps, such as in the case of naturalizations.

Similarly, in Mexico, the annual number of foreign residents with regular migratory status is compiled using data from three governmental agencies, namely the Migration Policy, Registration, and Identity of Person Bureaus (UPMRIP), the Secretariat of Foreign Relations (SRE), and the National Institute of Geography and Statistics (INEGI). Based on four subcomponents, the results helped to estimate the number of irregular migrants in Mexico, namely the population not possessing immigration documents authorizing regular residence in Mexico. When aligned with the total number of migrants derived from the 2020 Population and Housing Census, the results suggest that estimates for the population of migrants in irregular migratory status was too low from both Census and administrative derived sources. This exercise presents an example of the use of complementary data sources to estimate populations for which direct data are not available.

B. Statistical adjustment methods

Statistical methods are typically used to produce new migration statistics by using estimates from one or more other data sources to adjust an existing estimate. The input data sources usually have a strong statistical or theoretical relationship with the migration statistics in question. The methods

are sometimes called "correlation methods" to highlight the strong correlation across the measurements used.

Unexpected events such as natural disasters (e.g., hurricanes), global pandemics, or extreme policy changes can cause migration patterns to fluctuate abruptly. These events contribute to large and sudden flows of migration which cannot be immediately captured in surveys. In the U.S., annual estimates of international migration are drawn primarily from the American Community Survey (ACS) and the Puerto Rico Community Survey (PRCS). Household sample surveys do not measure migration events in real-time, rather, they capture migration events once the migrant is included in the sample. When there are large annual fluctuations in the magnitude of migratory movements, migration events will not be fully picked up until later (usually in the following survey year).

The U.S. Census Bureau has developed several data integration initiatives to provide up-to-date and accurate estimates of migration during the 2017-2021 period. Auxiliary data sources such as monthly Airline Passenger Traffic data were combined with the ACS and PRCS to adjust for migration patterns between Puerto Rico and the United States after Hurricane Maria. To adjust for the impact of the global COVID pandemic, a number of different administrative data sources were integrated with survey data at the macro level, including monthly visa issuance and refugee data from the U.S. State Department and U.S. Citizenship and Immigration Services.[18] The auxiliary data were chosen based on empirical observations that they have strong correlation with observed migration patterns using the ACS. The integration method generated adjustment factors were then applied to initial survey estimates to generate more accurate and timely migration measures.

[18]https://www.census.gov/library/stories/2021/12/net-international-migration-at-lowest-levels-in-decades.html

Another example of a statistical adjustment method is how Canada adjusted its usual methods to take into consideration the impact of the COVID-19 pandemic on international migration flows. Preliminary estimates of emigration and return migration are released approximately three months after the end of the reference period. These estimates are typically based on Canada Child Benefit program data from previous years under the assumption that recent trends continue. This assumption was less appropriate given the abrupt decline in migration stemming from the pandemic. As a result, starting in March 2020, monthly data on American visas issued by U.S. consulates in Canada were used to adjust the estimation method for emigration, while monthly Primary Inspection Kiosks airport data and data from the Global Affairs Canada's Registration of Canadians Abroad (ROCA) service were used to adjust the number of return migrants. Similar to the U.S. example, these sources were integrated by computing monthly ratios between current data and estimates using the previous method, and then applied to monthly visa, Primary Inspection Kiosks and ROCA data to calculate timely numbers of emigrants and return migrants, taking into consideration the impact of the COVID-19 pandemic.[19]

The Long-Term International Migration (LTIM) database in the U.K. is another good example of a statistical adjustment method. The International Passenger Survey (IPS) which captures travelers' migration intentions is used to estimate the number of long-term migrants who intend to immigrate to or emigrate from the U.K. for a period longer than 12 months. The LTIM adjusts these estimates based on the proportion of cases where actual migration behaviors do not match intentions (also known as the "switchers."). Starting in 2004, new questions were introduced to the IPS to collect data on respondents whose completed trips were different from their intention. Final LTIM data are compiled from IPS figures (including intended trips and switchers), Northern Ireland migration flow data from the Northern Ireland Statistics and Research Agency (NISRA),

[19] https://www150.statcan.gc.ca/n1/pub/91f0015m/91f0015m2020002-eng.htm

asylum seeker flow data from the Home Office, and preliminary adjustments data from the Home Office, the Department for Work and Pensions, Higher Education Statistics Agency, and the Census.[20] Further, refugee and Northern Ireland migration flows are incorporated into LTIM estimates, accounting for the fact that the IPS does not capture refugee and Northern Ireland migrant population. To date, it remains debatable whether migration intentions should be used as a proxy for actual migration behaviors, with subsequent estimates showing that past migration estimates based on the IPS were inaccurate. The IPS was suspended in March 2020 due to the COVID-19 pandemic, which also coincided with the U.K.'s new initiative to produce migration statistics from administrative data (i.e., ABMEs estimates). This marks a methodological shift towards compilation methods and potentially micro-data integration with multiple sources of administrative data.

C. Statistical modelling methods

Statistical modelling methods improve the quality and availability of migration statistics by borrowing strengths from other data sources that offset weaknesses in the primary data source (including conceptual coherence, timeliness, and spatial-temporal coverage) to improve final estimates. They help to improve migration estimates by accounting for: (1) missing data (e.g., certain characteristics are not asked in household surveys or census questionnaires), (2) improved balance between data accuracy and timeliness (3) inaccurate estimates for small populations and small areas/geographies, and (4) inconsistent measurement operationalizations over time and space. This category of methods may include familiar technical names such as imputation or indirect estimation.

[20] https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/longterminternationalmigrationestimatesmethodology#data-sources-used-to-compile-final-estimates-of-ltim

For example, the American Community Survey (ACS), which is the primary data source for estimating the total immigrant population in the U.S., does not include legal status variables needed to identify refugees and undocumented migrants, though these migrants are included in its sampling frame. As such, it is impossible to directly estimate the number of U.S. undocumented migrants and refugees from the ACS. However, other data sources containing demographic and social-economic characteristics of undocumented migrants and refugees can be used to generate patterns which help identify similar individuals in ACS data.

In the case of refugees, researchers from the Migration Policy Institute (MPI) rely on various U.S. administrative data sources including the Bureau of Population, Refugees, and Migration (PRM), the Worldwide Refugee Processing System (WRAPS), the Office of Refugee Resettlement (ORR), Department of Homeland Security (DHS), and archival reports from Immigration and Naturalization Services (INS). Such data are used as the basis for imputing refugee status for respondents in ACS data using two variables: country of birth and year of arrival in the United States. Specifically, refugee status was assigned to every country/year combination in which refugee admissions in the DHS and WRAPS data exceeded 40 percent of the estimated foreign-born population identified in the ACS data. After that, other record-level characteristics in the ACS such as age, English proficiency, educational attainment, employment status, median household income, poverty levels, health insurance coverage, cash welfare receipt, and food stamp receipt are used to confirm that refugee assignments effectively capture the aggregated characteristics of the U.S. refugee population.[21]

A more complex imputation procedure was used to generate estimates for the undocumented immigrant population using ACS data. To impute this missing characteristic, researchers from MPI collaborated with academic researchers Jennifer Van Hook and James

[21] Capps et al 2015

Bachmeier, borrowing information from the Survey of Income and Program Participation (SIPP), which is the only nationally representative Census Bureau survey that asks non-citizens to report whether they have legal permanent resident status. The researchers analyzed SIPP and ACS datasets and used a statistical process known as "multiple imputation" to assign immigration status in the ACS. Multiple imputation procedures mapped noncitizens' characteristics such as country of birth, year of U.S. entry, age, gender, and educational attainment across the two surveys. After this imputation process, federal government administrative data were used to identify recent refugees and asylees. Those who were initially assigned to be unauthorized immigrants, but who matched the characteristics of refugees were reassigned to be legal immigrants.[22] Although the methods used in the cases of refugee and undocumented immigrant imputation above do not integrate or link microdata, they do require access to record-level data to perform imputations and to evaluate the plausibility of the resulting aggregated estimates.

For small populations and smaller areas/geographies where migration statistics may be unavailable or inaccurate, integration projects may use data from multiple sources to generate disaggregated statistics and benchmark them.[23] For example, the very small size of the Indigenous population in Australia is one such case. Academic researchers James Raymer and colleagues used log-linear models to identify the key structures and patterns of migration flows from the three most recent Australian quinary censuses (2001, 2006, and 2011). A projection model with specific specifications and smoothing algorithms to overcome the small-number-cell issue associated with the very small Indigenous population size was then developed to project migration flows by origin, destination, age, and sex forward in 5-year increments to 2031.

[22] Van Hook et al 2016
[23] Raymer et al 2020, Wilson 2017

Another case is the inability of border control data to provide estimates for geographies below state or territory level in Australia.[24] Here, Raymer and colleagues developed a methodology for estimating detailed immigration and emigration flows for Australian sub-state regions by birthplace, age and sex covering a 35-year period from 1981 to 2016. The methods combine census data, publicly available international migration statistics, and international passenger movement data. Quinquennial Census data are then used to validate and adjust the final estimates.

It is noteworthy that missing data might involve a country or world region. In the UN International Migration Stock database, several countries do not share any migration statistics. In this case, another country or group of countries is used as a model to impute missing values. The "model" countries are selected on the basis of various characteristics, including the use of the same criterion for enumerating international migrants, geographical proximity, and migration experience. For instance, estimates for the Democratic People's Republic of Korea are imputed based on data from the Eastern Asia region.[25]

Statistical modelling methods can integrate multiple non-migration data sources to generate migration estimates. One example is the generation/distribution model, a two-step model that first estimates the possible number of migrants that can be "generated" within a migration system, and then "distributes" the migrants into bilateral migration flows based on covariate information.[26] The model relies primarily on covariate data, which are measures theoretically known to have a relationship with migration flows, such as population size, GDP, contiguity, and bilateral trade.

Cross-country comparisons of international migration patterns are difficult and confusing due to differences in definitions and measurement methods. Data integration may help to overcome

[24] Wilson 2017; Raymer et al. 2020
[25] https://www.un.org/en/development/desa/population/migration/data/estimates2/docs/MigrationStockDocumentation_2019.pdf
[26] Willekens & Baydar 1986

the conceptual coherence of migration measures, including different definitions of a migrant and different migration duration criteria.

The Integrated Modelling of European Migration (IMEM) project[27] represents one such initiative to harmonize estimates for migration flow data among 31 countries in the European Union and European Free Trade Association from 2002 to 2008, using data collated by Eurostat. A Bayesian model was used to correct inadequacies in available data and for estimating completely missing flows. The project integrates Eurostat data with various data sources containing covariate information (such as contiguity, bilateral trade flows) and experts' opinions on the effects of undercount, measurement, and accuracy of data collection systems. The result is a synthetic database of international migration flows with measures of uncertainty for international migration flows and other model parameters.

Migration flow data, which captures the number of migrants entering and leaving a country over the course of a specific period (such as one year), are dynamic statistics that may lead to a better understanding of past patterns and prediction of future trends. Compared to static migration stock or migrant population data, migration flow data are much less readily available.

Several data integration efforts in recent years have been developed to estimate migration flows from migration stock data, which is also known as the "flows from stocks" estimation method. Academic researcher Guy Abel developed a method to indirectly derive the number of global bilateral (country-to-country) migrant flows by integrating information on bilateral migrant stocks, births, and deaths in a demographic accounting system. The major challenge is that there are many possible combinations of migration events that can take place over the period to match the observed changes in bilateral migrant stocks. Abel used a log-linear model to estimate the minimum number of migrant flows that must have happened to match observed changes, taking

[27] Raymer et al. 2013; Wiśniowski et al. 2016

into account births and deaths. The results constitute a database of global bilateral migration flows in 5- and 10-year intervals covering the period from 1960 to 2015, with breakdowns by sex.[28] Subsequent studies further consider the plausibility of the "flows from stocks" method and add measures of uncertainty using Bayesian methods to the flow estimates.[29]

D. New and hybrid methods

Some examples do not fit neatly into the definition of macro-data integration, including cases which use (1) both micro- and macro-data integration methods, or (2) new data sources, such as big data.

The case study of Chile provides an example of combined micro- and macro-data integration methods. While the bulk of data integration is done at the record level (micro), the final estimates of migration flows are compiled by adding the aggregated number of foreign citizens who entered the country after the Census date and subtracting the aggregated count for those who were outside of the country or who have died during the estimation period. In the U.K., the newly developed ABMEs are awaiting potential linkage at the record level. When micro-data integration is ready, record linkages will help to evaluate and refine existing integration methods at the aggregated level. Another example is the United States' development of their own administrative database (the Integrated Database for International Migration (IDIM)) to measure international migration, which links tax and social security information to identify migrants. However, these administrative sources lack information on most of the foreign student population, thus they must use aggregate national, state, and county estimates of foreign students from the ACS or other sources to supplement this missing population.

[28] Abel 2017
[29] Azose and Raftery 2019

Big data, especially social media data, represent a new opportunity for migration data integration. In a recent study, Facebook Advertising data is used to generate estimates of migrant populations (or migrant stocks). The data contain aggregated counts of Facebook users who could be cross classified by various dimensions, such as place of birth (hometown), current and previous locations, age, and gender. The assumption is that migrants can be identified as those being flagged as "Expats" by Facebook's algorithm. Estimates for the United States are subsequently integrated with data from the ACS to validate results, with the key finding being that age and country of origin are the two main sources for systemic bias. While a key limitation of Facebook data is that the variables are not necessarily measured and documented according to the standards for scientific research, statistical models and particularly Bayesian hierarchical models can be used to harmonize differences.[30] Estimates derived from social media data represent unique new sources for migration data which can arguably be timelier and provide better coverage for hard-to-reach populations (e.g. certain race or ethnic groups, recent or undocumented migrants, refugees, etc.), compared to traditional data sources, particularly after certain events (e.g. "migration shocks") which cause rapid changes in migration patterns.[31]

Cell phone data is another potential source of information on migrants, though most studies have been regulated to analysis of commuting and internal mobility patterns.[32] For example, it might be possible to look at the destination of phone calls originating from specific places in a country using mobile positioning data to gauge the geographic distribution of specific-migrant groups, including irregular migrants, in those countries.[33] In addition, it could be possible to use cell-phone data to validate the stock of migrants measured in other data sources (e.g. census), if it contains information to identify the populations of interest.

[30] Zagheni, Weber & Gummadi 2017
31 Gendronneau et al 2019
[32] UNESCAP (2021)
33 Luca et al. 2021

However, there are still many limitations to use of "big data" in the production of official statistics, including access to data, data privacy and legal concerns, coverage of the total population of interest, self-selection bias, data quality and measurement issues (e.g. limited ability to identify migrants by nativity, resident or legal status, or reason for move), multiple and fake/inactive accounts, changes to methods used to collect data, etc. For more information about new and emerging data sources, one can refer to the forthcoming UN Task Force report on Operationalization of Conceptual Frameworks and Sources of Data on International Migration and Temporary Mobility, produced within the framework of the Expert Group on Migration Statistics, which addresses this topic in more depth.[34]

E. Challenges for macro-data integration

Macro-data integration represents resourceful initiatives to generate migration statistics in the face of an already challenging situation of missing, inconsistent, and incomplete data on migration and migrant populations. Two specific challenges remain to be considered: internal consistency and estimate validation.

First, differences in input data sources constitute a major challenge for macro-data integration efforts.[35] Aggregated data from various sources are susceptible to variations in definitions, measurement universes, as well as spatial and temporal coverage. For instance, the estimation year for migration statistics in the United States covers the period from June of the previous year to July of the current year, which differs from the calendar year or the tax year. Therefore, careful data cleaning must be done to reconcile such differences. In cases where input data sources are different from one another, some statistical adjustment methods must be applied at the data-cleaning and production stages. The notion of a 'true measure' is key for alignment;

34 See Expert Group on Migration Statistics section, Task Force 4 sub-section of the UNSD migration statistics website: https://unstats.un.org/unsd/demographic-social/sconcerns/migration/index.cshtml
35 See for example, https://www150.statcan.gc.ca/n1/pub/91f0015m/91f0015m2018001-eng.htm

each data source needs to be assessed in relation to the measure of interest prior to integration. For example, it is often difficult to compare and reconcile absolute migration numbers coming from different data sources (e.g. administrative data vs survey estimates), thus rather than making adjustments based on total values, one must use historical trends seen between data sources to make adjustments. This has been the case of the United States' recent macro-data integration efforts, which have leveraged historical relationships between flight data, administrative data, and survey estimates to make adjustments to take into account the impact of natural disasters (e.g. Hurricane Maria, COVID-19 pandemic) on migration patterns.

Collaborative agreements across different data providers may also help to resolve issues of data inconsistency, and stronger collaborative efforts among different entities of the national statistical system should be encouraged. While organizational structures and data release policies might pose challenges that hinder collaboration, specific collaborative agreements could generate synergies in terms of how initial data are tabulated, e.g., based on a universal conceptual framework, definition, or time frame. Similarly, international collaborative agreements may help to enhance the comparability of "mirror statistics," for instance, country-specific migration flows may be disclosed to a partnered statistical office for data integration purposes.

Validation is another major challenge for macro-data integration projects. In situations where migration data are missing, incomplete, or inaccurate, it can be challenging to find a true value to validate the resulting estimates. Sometimes, the true value is simply missing. Governmental statistics offices and researchers have come up with multiple methods to assess the results' plausibility, such as by consulting historical trends, and comparing estimates from different sources and methods. Thorough estimate assessment with multiple sources and clear communications about the size and the causes of error are good practices to enhance data user's perceptions about the estimate's validity and usability. It should be noted that measures of uncertainly can be difficult to generate using macro-data integration methods, as datasets which

would normally be used to produce error indicators, are combined to produce the final estimate. As they are, it is often beneficial to consider the resulting measure of uncertainty as a range rather than a fixed number.[36]

It can be difficult communicating measures of uncertainty to the general public, but there is benefit for more advanced data users to have this information, or at least make more quality assessments publicly available. However, as macro-data integration incorporates a wide range of available data sources, it limits external or "independent" sources to validate the model, which could be a significant concern, given official statistics need to be accepted by the public.

[36] The report elaborates on the topic of validation in Chapter 4. For more details about estimate assessment, readers may refer to Raymer et al. 2015.

Chapter 3. Micro-data Integration Methods

A. Overview

Micro-data integration refers to the process of integrating multiple data sources at the record - level. This is done through linking individual records across several different data sources, with an emphasis on ensuring that information from different sources is from the same individual. Among other things, integrated microdata may be used to identify migrants through their activities (e.g., border crossings (entries/exits), address registration, etc.), to measure duration of residence since immigration, and to examine migrants' changes in socio-economic characteristics over time. Many different types of data can be linked at the individual level, including multiple administrative data sources, censuses, and household surveys. These datasets can be compiled for a number of different applications, including cross-sectional and longitudinal analysis to inform topics like migrant integration, change of immigration status, and the social and economic impact of migration over time. Micro-data integration thus generates new, more robust, data with no additional respondent burden. Compared to new data collection, micro-data integration tends to generate lower operation costs and produce quality migration estimates in a timelier manner. Further benefits of micro-data integration include improved geographic coverage vis a vis sample survey data.

Because micro-data integration involves access to personal identifying information of individuals, it requires a legal framework that allows statisticians in governmental agencies to combine data based on information that can uniquely identify people, while also assuring data confidentiality. The enabling legal framework is even more crucial when data sources are managed by different governmental agencies. Additionally, for countries which are newly embarking on micro-data integration, the initial manpower and financial and technical resources needed to set up a data infrastructure, potentially consisting of millions of records, could be daunting. It is not accidental that the forerunners in the development and utilization of population registers were

countries with small populations, but this is currently of lesser concern as computing power has increased considerably.

Once the legal and technical frameworks are in place, the actual integration methodologies are relatively simpler than those employed in macro-data integration. Key methodological concerns are centered upon accuracy of the matched (or linked) records. When an actual identifying variable (such as a national identification number) is available, exact matching can be used. In contrast, probabilistic matching methods based on names, date of birth, addresses or other common characteristics must be used when unique identifiers do not exist.

B. Creating/Enabling the legal framework

The legal framework for data sharing and cooperation between governmental agencies is central to all micro-data integration projects and is based on bilateral relations and agreements. The legal framework is necessary for all countries and integration of data for statistical purposes should be performed by NSOs. However, legal frameworks vary sharply across different countries and are especially relevant for countries with less centralized national statistical systems or countries where data integration is performed outside the NSO. Centralized population registers represent the most facilitating framework for micro-data integration. Commonly found in European countries, and most advanced amongst Nordic countries, population registers are already integrated data systems, and continuously update selected information pertaining to each member of the resident population of a country or area. To update a register means to process "identifiable information with the purpose of establishing, bringing up to date, correcting or extending the register in such a way that it can be maintained as a continuous set of records."[37]

[37] Poulain, Herm & Depledge 2013

Legal and regulatory frameworks are crucial to interoperability, especially when it comes to the sharing and integration of data assets between organizations and across national borders.[38] While the legal framework in countries with population registers allow for the transfer of "identifiable information" across relevant government agencies for statistical purposes, their primary purpose is still administrative. In Norway, the principle is that each register is responsible for a certain type of information, and data should then be exchanged between agencies that need them. [39] Not only does data exchange minimize the need for governmental agencies to develop their own parallel data (saves resources), but it also reduces burden for citizens, in that they have a "single point of contact" to provide information, while also reducing the amount of contact needed to provide this information. Administrative registers are adapted according to each agency's responsibilities and roles, while additional procedures are taken to ensure personal data protection in the centralized system. The overarching legal framework thus facilitates micro-data integration for all necessary purposes, including to generate migration statistics.

In countries without population registers, which lack the established agreements found in register-based countries, different government agencies need to establish legal agreements with one another for sharing data, which is often coordinated by the NSO. In some cases, countries may already have data sharing systems in place, or established practices for digital public administration, in which case data can be shared with NSOs or other statistical agencies for statistical purposes within those general governmental agreements. Concerns about personal data protection, and the potential use of the shared data, are key elements in these data sharing agreements. These agreements typically dictate how data sources are linked, how the linkages are maintained in subsequent updates, and how the data will be used, stored, and for how long a period.

[38] Data Interoperability: A Practitioner's Guide to Joining Up Data in the Development Sector (https://unstats.un.org/wiki/pages/viewpage.action?pageId=36144005)
[39] See for example the Norwegian Act of National Registration of 2016, https://lovdata.no/dokument/NL/lov/2016-12-09-88

For example, the Longitudinal Immigration Database (IMDB) in Canada contains integrated microdata contributed by different governmental departments: the department for Immigration, Refugees, and Citizenship Canada (IRCC) provides annual data on immigrant admissions, the Canadian Revenue Agency provides tax files and child tax benefit files, and Statistics Canada integrates the various data sources together and updates new information into the IMDB database. The specific agreements with regards to the legal process for data integration, as well as the terms for data access and exchange between departments, are revised every five years. When challenges between departments arise during the revision process, they are addressed through inter-departmental meetings and approval processes. Additionally, any change to the set of variables or coverage of the population in the agreement must be approved at the ministerial level. To ensure data confidentiality, information from the IMDB is only published as annual aggregated summary tables by Statistics Canada, per their confidentiality guidelines. The micro-level data files are only available to Statistics Canada's researchers and employees. External researchers can access data when deemed employees of Statistics Canada, as well as members of Research Data Centers located throughout the country, after having obtained specific security clearance to use the data for research purposes.

Different interpretations in terms of what constitute confidential personal data may impact collaboration across governmental departments, including how data integration is conducted. The Republic of Moldova's 2019 project to integrate border crossing records with the State Population Register to identify migrants by their duration of stay provides one such example. Under the scope of this project, two data sources were integrated: (1) records of international border crossings from the General Inspectorate of Border Police (GIBP) and (2) data on local residence from the State Register of Population, which come from Moldova's Public Service Agency. The National Bureau of Statistics (NBS) serves as the central statistical authority to integrate the two data sources and produce official statistics according to the project's stated goal. According to Moldova's national

legal framework, specifically the law on official statistics, the NBS believes that it is authorized to access and process individual data, including personal identifiers, in order to link records across different data sources. The GIBP, however, did not consider personal identifiers to be individual data. This disagreement between the two agencies became a major limitation to the integration project. To overcome this disagreement, the NBS needed to modify the legal framework to clarify the use of personal identification for statistical data integration purposes, which added considerable time to implement the project. This is a good example of the need to modify legal frameworks when conducting data integration projects, especially in countries who are new to micro-data integration.

Once in place, an enabling legal framework may introduce transformative changes to the way international migration statistics are produced. In the United Kingdom, for example, the Office of National Statistics (ONS) long recognized that the existing methodology for estimating international migration, which was based primarily on the International Passenger Survey (IPS), was inadequate for the evolving needs to better understand international migrant patterns and to produce more accurate estimates at the subnational level. The 2017 Digital Economy Act[40] was the necessary change that allowed ONS to collaborate across the Government Statistical Service to better share and maximize the value of existing administrative data through integration at the record level.

Legal frameworks for data sharing may also exist between countries. Currently, agreements for international exchange of population register information are only available amongst Nordic countries (Sweden, Denmark, Finland, Norway, Iceland, Greenland, and Faroe Islands).[41] For example, in the Nordic system, persons have a unique identification number in each system which can be used and stored as a variable across country registers. Therefore, if a person moves to

[40] https://www.legislation.gov.uk/ukpga/2017/30/contents/enacted
[41] Poulain, Herm & Depledge 2013

Sweden from Norway, Sweden's population register would inform Norway's population register of the move (and vice versa), and Norway then "emigrates" this person from Norway.

Like mirror statistics at the macro level, exchange of micro-level data between NSOs can improve the accuracy of migration statistics, particularly for measurement of emigration, which is often not recorded when a person leaves a country. It should be noted that micro-data exchange between countries is generally not a widely accepted practice and is often met with resistance from statistical agencies due to data confidentiality and other concerns. For example, in 2021 Eurostat proposed a new regulatory framework on population statistics (European Statistics on Population) that among other things would allow NSOs to exchange individual level data.[42] However, this proposal was met by resistance from member countries (for many reasons) and is currently no longer being considered.

C. Creating/Enabling the technical framework

The technical framework is another key dimension for micro-data integration projects. As micro-data integration typically involves individual-level records with observations repeated over time, the data itself can quickly become overwhelming both in size and complexity of system architecture. Further, regular production of statistics based on integrated data requires repetition of integration activity once new data become available, especially when creating a longitudinal database. Similar to the process of "updating" population registers, the systematic repetition of integration activity requires clearly defined technical rules and a well-trained IT and statistical staff to ensure that the continuous integration process can be carried out smoothly. This leads to potential issues that need to be considered.

[42] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12958-Data-collection-European-statistics-on-population-ESOP-_en

First, the technical design of the integration project must align with its statistical goals, namely the type of statistics the project seeks to produce and its structural constraints, i.e., the quality and availability of current data sources. These have strong implications for the selection of input data files, including the "spine" and additional data sources.

The "spine" refers to the key dataset to which additional information will be linked to. In principle, spine selection should optimize coverage of the targeted populations, so this will vary across projects. For instance, in the United States, the Numident database, which combines Social Security Number records with death records, was selected as the spine to the IDIM because it provides relatively good coverage for legal U.S. permanent residents. This helps to achieve a key project goal which is to produce immigration estimates for the foreign-born population (including both citizens and non-citizens). In contrast, the Longitudinal Immigration Database in Canada was designed to better understand immigration processes and outcomes, thus it was built from the country's Integrated Permanent and Non-permanent Resident File, which covers all immigrants (i.e., non-citizens) who have landed in Canada since 1952 and all individuals with non-permanent resident permits since 1980.

Once the spine has been selected, additional data sources are added to expand the integrated database in two aspects: (1) to improve coverage of populations, and (2) to add more details, such as migration events or socio-demographic characteristics. The selection of additional data sources is also guided by project goals, and as such, potential data sources may be eliminated if they do not have adequate or redundant information on either the targeted populations or the measurements of interest, or have insufficient personal identifiable information (PII) to make quality linkages.

It is also necessary to consider other factors involved in the operationalization of migration in integrated data sources. Paramount to the identification of migrants is information on country of birth, while country of citizenship is an important variable as well. Knowing when a migrant

arrived in a country is critical, as entry into a new data system is not always the same as when they arrived in a country (though it often serves as a proxy measure of this). Previous country of residence is helpful information for determining country specific flows, in lieu of using country of birth or nationality as a proxy for this. Current address information is important for looking at the subnational distribution of migrants. When the same information (e.g. year of naturalization) for an individual on multiple datasets is different, care must be taken to decide which data source is most accurate. Additional consideration should be taken on how long it takes a migrant to enter data sources used as the "spine" for data integration, since this can lead to undercoverage of specific groups of migrants, especially those newly arrived in the country.

Similarly, significant effort must be invested in assessment/validation and pre-processing of input data sources before integration can happen. Due to the longitudinal nature of immigration data, both the data sources and the individual records may involve changes over time that hinder data integration. Changes in administrative processes or data structures may induce change in the coverage of populations (i.e., who get counted and who do not) as well as definitions for key variables (e.g., the duration of stay required to be counted as an immigrant) in the data sources. At the record level, individuals may change variables like names (e.g., after marriage), citizenship status, and in some cases, identification numbers (e.g., one person may be given different visa numbers). While some changes can be reconciled rather simply by realigning records (e.g., removing duplicate records caused by name changes) or by adding data on sub-populations captured in a different source, irreconcilable changes may result in the either missing data, inclusion of duplicate records, or mismatched records, which should be accounted for. If possible, it is important to include digital data sharing and database governance management systems, which can function as quality controls, help standardize variables, and anonymize Personal Identification Numbers (change to a synthetic number), in order to better use the data for statistical purposes.

Missing data is even a problem for Nordic countries.[43] In Nordic centralized population registers, tracking of unique identification numbers are limited to citizens and people born in each of the five Nordic countries. Identification numbers for foreigners are kept in a separate database which suffer from various data quality issues. Coverage of foreigners can be inaccurate as some individuals are not immediately given identification numbers when they arrive. The lag between arrival and identification can be up to several years.[44] Return migration is an additional source for inaccuracy, such that returnees may not have de-registered themselves when they originally left the country.[45]

There are also potential issues related to data cleaning, and the accuracy and coding of variables that are not originally designed to measure international migration. For example, in the case of "country of birth" codes on the United States' Numident file, a two-letter code is used to denote "place of birth," for both countries and states. As a result, a code of "CA" could mean either "California" or "Canada," necessitating cross-validation with "city of birth" information. Further, upon evaluation of data, a number of persons with a country code of "CH" (China) were born in the city of Santiago, suggesting a miscode of many Chilean (code "CL") born immigrants in the database. Even for databases that use numeric codes similar coding errors can exist and need to be addressed, even in the case of well-established integrated systems in countries like Norway.

Finally, as significant IT and statistical skills are required for the development and maintenance of data integration projects, the sustainability of such projects over time must be considered from the start. Additionally, data sharing agreements are often established for a finite

[43] Maret-Ouda J, Tao W, Wahlin K, Lagergren J. (2017). Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. Scandinavian Journal of Public Health. 2017;45(17_suppl):14-19.
[44] Careja, R., Bevelander, P. (2018) Using population registers for migration and integration research: examples from Denmark and Sweden. CMS 6, 19.
[45] Gauffin K. (2022) The illusion of universality: The use of Nordic population registers in studies of migration, employment and health. Scand J Public Health 50(2):269-271.

period thus need to be renewed regularly, while even when in effect, data might not be delivered in a timely manner. To overcome these potential obstacles, consultations with key stakeholders and data users often help to establish clear visions for technical demands. One good example is Georgia's integrated data platform, the Unified Migration Data Analytical System (UMAS), which was initiated in 2016 and has been operational since 2019. The project was not designed to generate specific migration statistics, but rather to pool data from different state agencies to facilitate cross-institutional data analyses and reports on migration issues. This data-driven approach in Georgia comes with high and growing financial costs for IT systems and specialists. However, it also allows the integrated database to meet multiple needs, including cross-verification of reports across state agencies and to provide recommendations for improving data quality of input sources. It is also important to note that after initial startup cost, resource consumption may be reduced over time. In the case of Canada, administrative data integration is found to be an effective and cost-effective tool. While significant investment and effort is required at the start and during major data augmentations, once everything has been established, cost and maintenance is lower than other (new) data collection processes.[46]

D. Micro-data integration data linkage methodology

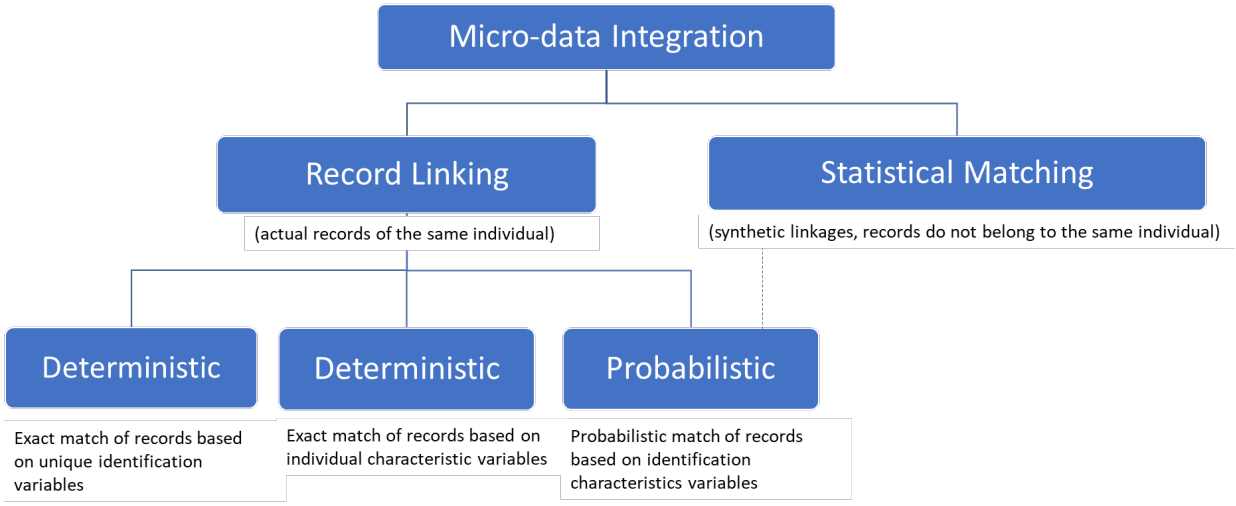
Micro-data integration produces linked records where information about one individual is drawn from several different data sources (e.g., nationality from one data source and age from another source). In general, the primary methodological concern is to ensure that information from various sources are correctly attributed to the same individual. This process includes not only data linkages, but also post-linkage methods to deal with validation, duplication, conflict resolution, and editing and imputation of incomplete or missing values.[47] As such, the focus of this section

[46] https://www.statcan.gc.ca/en/about/policy/admin_data

47 For more detailed information about post-linkage validation methods, see UNECE (2020),"Guidance on the Use of Longitudinal Data for Migration Statistics," or UNSD (2022), "Handbook on Registers-Based Population and Housing Censuses," Chapter IV, "Transforming administrative data into census data."

is on data linkage methods. As illustrated in Figure 2, there are four methodological approaches to link data at the record level, two using deterministic, and two using probabilistic methods: (1) deterministic match of records based on unique identification variables included in multiple datasets; (2) deterministic match of records based on individual characteristic variables; (3) probabilistic match of records based on identification characteristic variables; and (4) statistical matching using probabilistic models to generate synthetic linkages (and data) from records which do not belong to the same individuals.

Figure 2. Methodological approaches to linking record level data



Deterministic (or "exact") matching is done when a person can be linked across datasets using a pre-existing unique identifier (direct matching), or by linking through shared "true" identification variables that exist across datasets (name, sex, date of birth, residence, etc.). A "true" identification variable refers to a unique attribute that is given and maintained by a specific administrative office to ensure that the attribute does not change over time and that it is unique to

an individual, i.e., the same personal identification number will not be given to more than one individual and one person cannot have more than one valid identification number at a given time.

An example of the first method of deterministic matching using a unique identifier is the Personal Identification Number used in the Norwegian central population register, given to all persons born in Norway (regardless of citizenship) and to all persons immigrating to Norway (even including some non-residents). Statistics Norway monitors the uniqueness of the Personal Identification Number, for instance, to reconcile errors when two numbers are given to the same person. Using the Personal Identification Number as the unique identifier, information collected by different state agencies (multiple datasets) and over time can be linked to the same record. Additional variables that do not change over time (time invariant), such as date of birth, may be used to identify and correct possible errors. Further, variables used for identification (e.g. date of birth, sex, etc.) must be changed if any of the information elements change across datasets, meaning the updated status of the variables must be the same in all datasets before an exact match can occur.

In the absence of unique identifiers, it is also possible to make direct deterministic linkages based on shared individual characteristic variables like age, sex, date of birth, and address. In this type of deterministic matching, a person is often determined to be the same person if several (more than one) of these variables are shared across datasets. This is more typical in small populations where different types of names are common, and combinations of date of birth, name, sex, and address are unique. This method differs from probabilistic matching using these same variables, where a unique identifier is created using the likelihood that individuals across datasets are the same person.

In contrast to deterministic matching, the most common probabilistic matching method involves matching records based on the likelihood that a person is the same across two or more

datasets. This method generates synthetic identification variables to determine this likelihood, which are based on a combination of variables that may uniquely identify an individual. This method often generates a unique identifier based on these probabilities, which is used to link new data with existing records when new data becomes available. This strategy is adopted when a unique identifier variable is not available, such as in the case of the Longitudinal Immigration Database in Canada, the IDIM in the United States, or Georgia's UMAS. In Georgia's case, probabilistic matching is used since different identification variables (national identification number and passport number) are used by different governmental agencies. Relatively time invariant variables, such as date of birth, place of birth, sex, and name are used in combination to identify the likelihood a person is the same individual across data sources.

A probabilistic matching procedure typically compares several fields of values between two records and assigns a weight that indicates a possible match between them. The comparison method and the weight threshold for records to be considered a match differ across different matching algorithms.[48] Probabilistic matching methods recognize that identifiers for identical individuals can diverge in different datasets for many reasons, such as data entry errors, surname changes, or address changes which cannot always be updated immediately.

As the literature on probabilistic matching continues to evolve, specific matching algorithms are determined by each statistical office and relevant stakeholders. Linkage information captured over time may also be used to facilitate current and future efforts, such as in the case of Canada where changes in linkage fields are maintained in the Social Data Linkage Environment. It is also important to note that pre-processing to standardize formats of identifying variables such as date of birth, names in different languages, or name changes are crucial to the quality and success of subsequent matching procedures.

[48] For a detailed overview of matching algorithms, please refer to UNECE-HLG MOS (2017).

In many cases, unique identifier ("true identification") numbers only cover a specific population (e.g., citizens) while omitting others (e.g., foreign citizens). A hybrid approach which combines true identification variables and probabilistic matching is useful to expand the coverage population while also improving the quality of matching. In Chile, the national identification number is a true unique identifier assigned to its citizens and legally resident foreigners. However, when foreigners initially enter the country without formal residence status they enter on a passport but are not assigned a national identification number (not picked up as an in-migrant). If they later become a resident they are assigned a national identification number, thus when they leave the country they are only counted as an emigrant. To rectify this, a matching algorithm uses national identification number and other variables including surname, first name, middle name, country of nationality, and date of birth to match records for foreigners who initially entered the country without residence. This hybrid approach allows them to ascertain that the level of matching in their 2020 migration estimates was close to 90%.

A hybrid approach may also expand linkable data sources, such as in the case of the United States, where the Social Security Number (SSN) maintained by the Social Security Administration uniquely identifies all individuals, including both citizens and non-citizens who intend to legally work in the country. The SSN is thus a unique identifier variable which can link records to tax filling information, but it cannot be used to link other data sources at the Census Bureau, as they do not collect SSNs. A hybrid approach was developed, first to use data with SSNs to assess the quality of the chosen probabilistic matching algorithm, and then to match data without SSNs using name, sex, age, and address information. A synthetic unique identification number called a Personal Identification Key is created and provides an anonymized unique identifier for subsequent data linkage across data sources owned by the Census Bureau.[49]

[49] https://www.census.gov/about/adrm/linkage/working-papers.2014.html

In addition to record linking though deterministic or probabilistic matching, statistical matching is a probability model-based imputation method based on similar units. Employing both parametric and non-parametric methods, statistical matching may be used to impute important characteristics, such as whether an individual is a migrant in datasets where such information is unavailable. From a micro-data integration perspective, a new "synthetic" dataset is created from multiple datasets (which do not contain the same units), where data on all variables is available for every unit. These "synthetic' records are based on an informative set of common variables across the original datasets. However, statistical matching methods have not been widely used to produce migration statistics. Currently, one project is exploring the possibility of integrating the German Central Register of Foreigners (an administrative data source) and the Micro-Census (a labor force survey) with weighted random hot deck imputation.[50]

Beyond data linking, there are also methodological considerations pertaining to the operationalization of migration measures, e.g., to measure emigration or immigration events. The rules of operationalization are usually developed based on statistical offices' expert knowledge to logically infer connections across data sources. The "signs of life" principle, for example, is typically used to ascertain the presence of an individual in a country through crucial activities such as tax filling, work, or education. In the United States' IDIM, Social Security records are linked with annual tax files to help remove individuals who apply for a Social Security Number but never actually migrate to the United States (i.e., no sign of life). In Moldova, the national statistics office applies logical rules to identify errors in border-crossing events recorded in administrative data. An "illogical itinerary" is flagged when an individual has two movements in the same direction (e.g., two consecutive exits or entries).

[50] UNHCR (2018) International Recommendations on Refugee Statistics

Regarding undercoverage of specific groups, family association is a method used to capture dependent family members and children who may not have their own individual records in the main data sources. In the United States' IDIM, tax filling information was used to determine the number of dependents recorded in foreign-born tax returns, thus helping to complete the count of the foreign-born population, as family dependents are often not eligible to receive social security numbers. Canada is another example, as integrated child benefit data provide information about immigrant children connected to specific immigrant parents and families.

E. Challenges for micro-data integration

Integrated microdata is a powerful tool to enhance migration statistics, and linking records is widely used and accepted in register-based statistical systems. However, data access can still be a major challenge for integration efforts in other country contexts. Even when microdata are available, the legal framework on data privacy and confidentiality determines whether and how data can be used for statistical purposes. Thus, national statistical offices must develop collaborations with relevant governmental agencies to create enabling rules for their data integration projects.

There also data privacy risks that must be assured, especially when integrating data and measuring vulnerable populations like refugees or irregular migrants. This can complicate the development of legal data sharing arrangements between agencies, as privacy and confidentiality concerns are particularly salient when linking records using probabilistic methods. Particular care must be taken to ensure confidentiality of individuals, through the anonymization or pseudonymization of individual records, both used in the matching process, as well as when data are disseminated or accessible by the data user community. While outside the scope of this Task Force report, methods of disclosure avoidance can range from simpler methods of data masking, like data suppression or data swapping, to more complex methods of data perturbation, including incorporation of complex differential privacy models or the creation of synthetic datasets.

Unequal coverage and quality issues in data sources further pose challenges for micro-data integration projects. Some specific migrant groups (e.g. those in irregular status) might not be included in datasets, while for migrants who are included in one dataset, unmatched cases to another dataset must be dropped, resulting in less reliable results. Additionally, significant time and effort may be required to evaluate and correct errors in data sources. Correction of errors may be a significant challenge for countries with lower levels of technical capacity or low data quality. In the pre-processing stage, the specific goals of an integration project may be altered to align with actual conditions of data quality, as well as budget and time constraints.

Finally, a time lag is usually involved when measuring migration through integrated microdata, which is especially pertinent when measuring migration flows. Some lags are inherent in definitional measurement. For example, if an emigrant is defined as someone staying away for at least 12 months, it would take one year to identify an emigration event. In other words, the measurement of migration is lagged one year from the actual event, though this issue is not specific to integrated data, which might enable finer measurement of the emigration process through linkages to other sources. The complex process needed to integrate data may contribute towards additional time lags (in terms of data sources used), especially in new projects where significant pre-processing work is required. Legal and administrative barriers on data sharing (and when data are actually delivered) in specific countries may exacerbate the timeliness issue. With this being said, integration of existing data sources may still potentially generate more timely estimates than new data collection.

Chapter 4. Assessing and Communicating Results

A. Overview

Integrated data can be used to produce official statistics, thus necessitating proper evaluation of results and communication to the public. This is particularly pertinent to macro-data integration, though less so when micro-data integration is well established and routinely used to produce register-based statistics. Both macro- and micro-data integration combine multiple data sources, and in the process, they also combine multiple sources of errors inherent in those data sources, in addition to potential errors associated with the integration methodology.[51] This is linked to the perception that integrated data are less authentic than primary data sources. A survey conducted in 2017 by the Data Integration Project[52] found that public acceptance and trust issues caused significant barriers to data integration, even more so than methodological issues, technical skills, and budget needs. A thorough data assessment plan and clear communication with stakeholders—including data providers, data users, and the general public—are key to the success and sustainability of data integration projects. Key assessment strategies include comparisons between alternative data sources and demographic measures, as well as determination and reporting of measurement errors, which can be included in model-based error estimates. In general, assessments should provide data users with a good sense of bias in the integrated data, i.e., whether the data have higher or lower values compared to the true values.[53] Comparisons with traditional estimation methods may further help data users recognize the extent of improvement (e.g., in terms of data quality, accuracy, or coverage), while demographic accounting can be used as a validation tool to ensure migration stocks are aligned with migration flows.

Communication with stakeholders is key, as they can provide feedback to validate newly produced estimates, as well as help determine specific data user needs. Dissemination of official

[51] For a detailed discussion, see the "Error Framework" section in UNECE 2020.
[52] UNECE 2019
[53] Raymer et al 2015

statistics, progress reports, methodological notes, and usage of estimates in public discourses play a key role in improving acceptance and support for data integration. Additionally, academic and policy researchers may contribute to data dissemination by producing research reports when given access to integrated data.

B. Estimate assessment/validation

The validity of integrated data may be assessed using two types of comparisons: (1) with known ("true") values or (2) with alternative data sources. As demographic processes (birth, death, and migration) are ultimately linked to changes in population size and composition, the population count (such as from a population census) is frequently used as a benchmark to validate migration estimates derived from integrated data. Obviously, the assumption here is that the population count, as well as birth and death registrations, are accurate. Known values also play an important role in micro-data integration, such that matches may be used to assess and improve probabilistic matching algorithms. In some cases, however, even the population count may be problematic due to the lengthy time gap between censuses or due to significant under-coverage of undocumented migrant populations, such as in the cases of the United States and Mexico. When known values are absent, estimate assessments must rely on alternative data sources of information on migration stocks and flows, such as residence permits, flight data, border control and/or visa data. However, use of "benchmark data" to evaluate integrated results can be complicated by the fact that these same data sources are often those which are used for integration, meaning there is not true independence between the integrated-based estimate and estimates from benchmark data. Finally, expert opinion may also serve as an alternative source for assessment.

Assessment on the size of error helps users understand the range of plausible values for a specific migration measure. Typically, statisticians report whether the resulting estimates are higher or lower than the true value, often with an associated level of confidence for the estimate. For example, evaluations have indicated that estimates of the number of emigrants in Canada,

which is based on macro-data integration, are underestimations. Because the true value is unknown, improvements in this estimate is recognized by the inclusion of better and more complete sources of data, such as through the incorporation of United States immigration statistics, which is a major destination for emigrants originating from Canada.

In the case of an unknown true value, model-based error estimations may be used to provide insight about the range of uncertainty. In the United Kingdom, for instance, migration estimates derived from the International Passenger Survey (IPS) are provided with uncertainty measures. Multiple assessments of errors (e.g., various sources of errors and opinions from multiple experts) may also be incorporated through a Bayesian model, such as in the case of the Integrated Modelling of European Migration (IMEM) project, which harmonizes estimates for migration flow data among 31 countries in the European Union and European Free Trade Association.

Knowing the sources of errors helps data users understand data limitations. Missing data and hard-to-reach populations are common causes for bias in migration estimates, and one of the reasons integrated data methods are used in the first place. When the cause of error is linked to specific data sources, data users can further envision how data collection and processing methodology can be improved. In Chile, for example, part of the reason previous migration estimates lacked accuracy was because they were partially based on population censuses that were conducted 15 years apart, while there were also problems with administrative records used in the previous estimates (e.g. duplicates were not removed and limited coverage of residences). Chile's new estimation method is based on integrated macro- and micro-data thus helps improve estimates by using more recent administrative data sources which are available more frequently.

Assessments of data quality further help evaluate the validity of estimation methods over time and especially in the event of shocks and extreme policy changes. In the United States, for instance, existing survey-based methods for estimating annual migration flows to and from Puerto

Rico and the United States were shown to be inaccurate, as they did not capture the surge in migration to the United States from Puerto Rico after Hurricane Maria in late 2017. Similarly, these survey-based methods did not measure the sudden influx of return migration flows of US residents back to the United States due to fear of extreme travel policy restrictions at the beginning of the COVID-19 pandemic, nor did they measure the sudden reduction in international migration due to border and visa office closures. New estimation methods therefore integrated additional data sources, including flight data (between the US and Puerto Rico and the US and the rest of the world) and administrative data on visas issued abroad, to improve the accuracy of the resulting estimates due to these migration-impacting events.

When combining different data sources, the integration process often necessitates synthesizing definitions of migration measures, target populations, geographic units, and time frames. An indirect effect of this practice is that it enhances comparability of statistics based on integrated data sources to earlier/other migration statistics releases. In the long run, this allows for multiple assessments based on different sources of published statistics, while it also increases the usability of estimates based on integrated data.

C. Communication with key stakeholders and dissemination of integrated data

Communicating results is an integral part of validation, as comments and feedback from outside the data integration team (both internally and externally) help to recognize blind spots and refine integration methods. In addition to validation, communication with stakeholders helps to maintain the trust and relevancy of the integrated data and resulting statistics. Key stakeholders for migration data integration projects, and data integration projects in general, typically include data providers, external and internal data users, policy makers, and the general public.

Data providers play an important role in the integration process, as data access is an important precondition for integration, particularly for micro-data integration. In country contexts

where data integration is relatively new, communication is key for cementing collaborations between statistical offices and relevant governmental agencies which provide input data. In the case of Chile, for instance, the recent integration project was founded upon the collaboration of five governmental agencies. The Migration Service of the Ministry of the Interior was in charge of integrating administrative microdata, while the national statistical office collaborated in the development of the integration methodology and macro-data integration with census data. In Chile, the integration methodology and interim results were communicated not only to the five data providers, but also to relevant public servants representing over 20 public institutions. This helped participating agencies legally commit to future collaboration, as well as establishing an official methodology for processing and integrating migration data.

In countries where data integration has a longer history, a specific program may be established to facilitate data sharing and data integration, in addition to evaluating results and engaging stakeholders. In the United States, the Federal-State Cooperative for Population Estimates[54] (FSCPE) works in cooperation with the Census Bureau to produce population estimates. FSCPE agencies supply some state-specific vital statistics and information about group quarters, like college dorms or prisons, though they do not provide any international migration data. The Census Bureau produces and sends population estimates based on integrated data to FSCPE agencies for review and comment prior to release. A similar organization exists in Canada, the Federal-Provincial-Territorial (FPT) Committee on Demography, where any significant changes in data sources or methodology are presented for approval. Like the US case, some data sharing agreements exist with provinces and territories, but not for international migration. Committees like the FPT and FSPCE are critical from a statistical user standpoint given the use of migration and population data in legislation and funding allocation.

[54] https://www.census.gov/programs-surveys/popest/about/fscpe.html

Given the broad range of data users, the methods used and resulting estimates from integrated data are usually disseminated via multiple channels. Many national statistical offices disseminate outputs (including data tables, graphs, metadata, written reports, and method statements) on websites that are available for public access. Both public and academic conferences are also good venues for statistical offices to provide more detailed explanations of integration methodology and results. The international statistical community is also a good venue to share and learn about new techniques and good practices for data integration. An example of these types of activities can be seen in Chile, where the NSO's press and communications team made infographics, tweets, and posted on both social media and their institutional website to disseminate information about their new integrated methodology and results of their international migration estimates. Chile also participated in various regional webinars organized by the Economic Commission for Latin America and the Caribbean's(ECLAC) population division to present their collaborative work, including the methodology and estimates results.

Beyond outputs, specific data users may be given access to integrated databases to conduct their own analyses. This enhances the range of outputs or estimates that can be generated with integrated data. At the same time, data users may contribute their own unique insights about data quality or integration methods, which help to improve the integration process in the long run. Access to integrated data (especially for microdata) are typically granted upon strict conditions of data confidentiality and specific rules about how the data can be used or interpreted. Some examples of this data access model include the Federal Statistical Research Data Center[55] in the United States and the Canadian Research Data Center Network.[56] However, there are often restrictions on which data can be accessed, especially if seeking data from another agency (e.g. tax data), thus the ability to link microdata could be limited for external users.

[55] https://www.census.gov/about/adrm/fsrdc/locations.html
[56] https://crdcn.org/

Communication and dissemination of results to the general public is crucial for maintaining public support and securing future legitimization and funding for statistical integration projects. Websites are a common venue for releasing public information, though the use of social media platforms to communicate results and announce new reports is becoming more commonplace. Statistical offices also use news releases and public forums to explain methodologies and results in more details. Progress reports are helpful for integration projects that are relatively new. For example, for the new micro-data integration project in the United Kingdom, the Office of National Statistics (ONS) provide many updates as they develop the integration methodology[57] and produce new estimates[58] along the way.

D. Use of outputs derived from integrated data in official statistics

There are a wide range of practices surrounding the use of integrated data in official statistics. Integrated data from population registers, where individuals are directly matched with a personal identification number, is the least controversial and widely used to produce migration statistics in register-based countries. This has been the case in register-based countries, like Nordic countries, for decades. Other countries such as the United Kingdom have a long-standing tradition of producing not only migration but population statistics in general based on integrated estimates. In such context, there is stronger acceptance of integrated estimates as official statistics.

Nevertheless, estimates derived from integrated data represent an important first step to fill the void of missing migration indicators and help improve the quality of migration data. In countries where migration measures are traditionally derived from censuses, new integration

57 https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/internationalmigrationdevelopingourapproachforproducingadminbasedmigrationestimates/2021-04-16#transformation-of-migration-statistics

58 https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/methodsformeasuringinternationalmigrationusingrapidadministrativedata#overview-of-rapid

methods with alternative data sources, such as administrative data, further enhance the timeliness and relevancy of migration measures.

Estimates drawn from integrated data will soon prove to be useful in the context of the new guidelines and conceptual framework on international migration and mobility, where more policy-relevant migration indicators are required at a minimum.[59] The guidelines also recommend using integrated data sources to strengthen the alignment of migration stocks and flows, and to enhance available information on demographic processes disaggregated by detailed characteristics. The increasing presence and usefulness of data integration projects may help gain more public acceptance and trust moving forward, particularly from a macro-date integration perspective.

[59] See Task Force 1.

Chapter 5. Conclusions and Future work

The new conceptual framework on international migration and temporary mobility calls for a number of migration measures and indicators which require improved and more timely migration data, which could benefit from the integration of multiple data sources. Data integration has the potential to create richer datasets which could help operationalize the conceptual framework, improving consistency between migration stock and flow statistics by drawing from the strengths of different data sources.

The UN Expert Group on Migration Statistics has also developed a series of indicators to improve the measurement of international migration and mobility on important policy areas, such as migration stocks and flows, irregular migration, access to basic services, and social integration. Several of these indicators could be computed using data enriched through integration methods, particularly when further disaggregation of indicators is required. For example, in addition to basic indicators to measure the stock and flow of immigrant and temporary populations, micro-data integration could be particularly relevant for obtaining disaggregated information on age, sex, country of birth or citizenship, duration of stay, educational attainment, and location of residence within a country. From a macro-data integration perspective, for example, these methods could be particularly useful for developing indicators to estimate irregular migration, a group which is often missing or under covered on both administrative and household survey data sources.

This report also contributes to other work of the UN Expert Group on Migration Statistics, such as the Task Force on Operationalization of Conceptual Frameworks and Sources of Data on International Migration and Temporary Mobility. This group discusses potential uses of traditional data sources (population census, household surveys and administrative registers) for the production of migration statistics, while also examining the use of new and emerging data sources to improve measurement of international migration and temporary mobility. Data integration is especially pertinent when applied to alternative data sources like social media and mobile

positioning, as these “big data” sources often require integration with more traditional data sources to produce reliable and valid estimates.

In summary, this report provides an overview of data integration methodologies in multiple contexts, covering both practical statistical work and academic literature with the goal to support countries to produce sufficiently disaggregated data for the measurement of international migration by means of data integration, including both macro- and micro-data techniques. There is not a one-size-fit-all approach to describe all data integration methodologies. The examples clearly show how data integration methodologies are developed and refined to fit specific project goals, which are further limited by conditions of project needs, data access, data availability, and technical readiness in each individual context. It is also common for project goals to be revised to align with existing conditions. On this note, pre-processing steps to evaluate and prepare input data sources are crucial for setting and re-setting project goals.

Macro-data integration techniques are often chosen when microdata are unavailable, either due to microdata not being collected or inaccessible due to various reasons. A wide range of aggregated outputs from various data sources including administrative data and auxiliary data such as flight data may be used for macro-data integration. Methodologies may range from simple data compilation to complex modelling, as determined by project goals, available data, and technical capacity.

Micro-data integration techniques must be supported by an enabling legal and technical framework that allow statisticians to process and match multiple data sources at the record level. While micro-data integration methodologies may seem less diverse than those used in macro-data integration, most micro-data integration projects still involve rigorous and complex pre- and post-processing steps to ensure that input data are suitable, and that integration can be done with minimal errors while still meeting project goals. Specific to micro-data integration, sometimes the

goal to enrich data with more variables may come at the expense of reduced coverage (i.e., unmatched cases will be dropped).

Within each project, it is important to note that integration methodologies are developed and refined over time. Estimates based on integrated data are typically assessed (or validated) for the magnitude and causes of error. Some countries provide estimates of uncertainty, such as a range for measurement errors, while some do not, either due to computational difficulty or preferences to release single numbers. Documenting causes for error is important because they help data users understand why estimates must be produced with uncertainty and how future estimates can address such error. Methodological adjustments are often made to address causes of bias and improve the accuracy and timeliness of estimates.

Countries are also distinct in their tradition regarding the use of estimates based on integrated data as official statistics. Some countries have a long history of producing official statistics from integrated data, while other countries do not include estimates derived from integrated (particularly macro) data in official statistics. While country-specific preference may be resistant to change, the increased presence and usefulness of estimates drawn from integrated data will likely help enhance public trust and eventual acceptance for these estimates. Effectively communicating results plays an important part in this. Communication with key stakeholders, including data providers, data users, and the general public can happen via multiple platforms, such as websites, news releases, published reports, social media, and interactive workshops. It is important to communicate with these stakeholders on a regular basis to ensure the relevancy of migration statistics and understand emerging data needs given migration patterns can evolve rapidly. In some cases, researchers are given conditional access to integrated microdata, which further expands the production and use of estimates based on these data sources.

We conclude this report with some suggestions for future work to expand the use and utility of data integration to improve international migration statistics.

- **Improve access to administrative data sources to produce migration statistics**

Administrative data sources such as border control data, visa records, and tax filings, are good candidates for data integration to measure international migration. Data access, however, is a barrier for many countries to enable integration of administrative data. Collaborative agreements between national statistics offices and relevant governmental agencies which collect and process the respective administrative data can help reduce barriers to access, especially for microdata. Notably, collaborative agreements should consider the broader legal framework within a country regarding data confidentiality and privacy to avoid possible misunderstanding regarding what an agency can process (e.g., access the identifying variables to match records). As such, it is critical that any agreements include as much personal identification information(PII) as possible to allow for quality linkages between data sources.

For macro-data integration, collaborative agreements may generate synergies across administrative data sources, such that a common definition, data processing method, coverage, or time frame will be used across different agencies. This can facilitate the integration of aggregated macro data from various sources and can be used to get earlier access to aggregate tabulations, as opposed to waiting for the official public release.

- **Develop international data exchanges and collaborations**

Data exchange at both the micro- and macro-levels can greatly improve the quality of migration statistics. The Nordic Agreement on Population Registration is one example of data exchange between population registration authorities at the micro level, which indirectly allows countries to better measure emigration of residents, once information is processed by NSOs. More international collaborative agreements, especially amongst groups of countries that have strong migration ties,

should be developed. An example of this is the recent Memorandum of Understanding between the NSOs of Canada, Mexico, and the United States, with the extent purpose of improving international migration statistics through the exchange of methodological information and statistical data. While there are often limitations on sharing microdata between countries, macrodata exchange, such as in the form of special tabulations, can enhance the ability of countries to measure difficult migration phenomena (e.g. use of country-specific immigration flow data from destination countries to measure outmigration flows from origin countries).

- **Expand and promote methods of communication and dissemination with key stakeholders**

Communication and dissemination of results is critical for developing and maintaining data integration projects. Expansion of existing and new methods of communication and dissemination is recommended. These can take the form of promoting webinars, conferences, and other venues where participants can discuss data integration methodologies, exchange documents, and learn more about data and metadata. Social media and other emerging platforms, if not already used, should be considered as a method to release this public information. Regular communication with stakeholders is important to stay abreast of current and emerging data user needs in the field of international migration and temporary mobility. If possible, researchers should be provided with opportunities to have conditional access to integrated data, which broadens the utility and functionality of these data sources.

- **Combine micro and macro integration methodologies**

When data are available, macro- and micro-data integration techniques can be combined to further improve integration methodologies. This could help address inherent weakness in each type of integration approach. For example, if micro-data integration of administrative sources results in significant undercoverage of a migrant population, then it could be possible to use macro-data tabulations (e.g. from a census or household survey) to supplement or enhance estimates produced

by the linked administrative data. In the example of Chile, micro-data integration provides rich information, yet still has undercoverage of the entire migrant population. An additional step of compilation to incorporate aggregate data about missing populations helps provide better migration estimates. On the flipside, estimates derived from micro-data integration can serve as an additional benchmark to validate estimates generated from macro-data integration techniques and to help finetune integration methodologies.

- **Provide better understanding of potential new data sources, uses, and limitations**

In the context of the new guidelines and conceptual framework on international migration and mobility, more migration indicators, hence detailed disaggregated data on migration, are a requirement. There is also a need to respond to and identify emerging user needs, especially for use in evidence-based policy making. To fully measure complex migration patterns in a timely manner, it is important to explore new data sources as well as potential uses and limitations. For example, the emerging usage of big data, such as social network data for migration estimates, is a promising venue for future work. Big data provide unique richness and potentially good coverage of hard-to-reach populations yet come with their own set of limitations to produce official statistics. Methods on how to leverage big data for either micro-data integration (linking individuals from big datasets to individuals) or macro-data integration (use of other data sources to adjust estimates generated from big data) still need to be developed, which is especially true for their use to produce international migration statistics. This recommendation ties into the previous mentioned work of the UN Expert Group on Migration Statistics on the use of new and emerging data sources to measure migration and temporary mobility.

References

Abel, G.J. (2017). "Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2015." Int Migr Rev. https://doi.org/10.1111/imre.12327

Azose, J.J., Raftery, A.E. (2019). "Estimation of emigration, return migration, and transit migration between all pairs of countries." Proc Natl Acad Sci116(1):116-122. https://doi.org/10.1073/pnas.1722334116

Capps, R., Newland, K., Fratzke, S. et al. (2015). "The integration outcomes of U.S. refugees: Successes and challenges." Washington D.C.: Migration Policy Insitute. https://www.migrationpolicy.org/sites/default/files/publications/UsRefugeeOutcomes-FINALWEB.pdf

Careja, R., Bevelander, P. (2018). Using population registers for migration and integration research: examples from Denmark and Sweden. CMS 6, 19.

de Beer, J., Raymer, J., van der Erf, R., & van Wissen, L. (2010). "Overcoming the Problems of Inconsistent International Migration data: A New Method Applied to Flows in Europe." European Journal of Population 26 (4), 459–481. http://www.jstor.org/stable/40928479

Eurostat (2013). "Statistical matching: a model based approach for data integration." Luxembourg. https://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF.pdf/477dd541-92ee-4259-95d4-1c42fcf2ef34?t=1414780333000

Gauffin K. (2022). "The illusion of universality: The use of Nordic population registers in studies of migration, employment and health." Scandinavian Journal of Public Health 50(2):269-271.

Gendronneau, C., Wiśniowski, A., Yildiz, D. et al. (2019). "Measuring Labour Mobility and Migration Using Big Data: Exploring the potential of social-media data for measuring EU mobility flows and stocks of EU movers." European Commission, Directorate-General for Employment Social Affairs and Inclusion. https://www.rand.org/content/dam/rand/pubs/external_publications/EP60000/EP68037/RAND_EP68038.pdf

Lanati, M & Venturin, A. (2021). "Cultural change and the migration choice." Rev World Econ 157, 799–852. https://doi.org/10.1007/s10290-021-00418-1

Luca, M., Barlacchi G., Oliver, N. & Lepri (2021). "Leveraging Mobile Phone Data for Migration Flows." https://www.researchgate.net/publication/352016918_Leveraging_Mobile_Phone_Data_for_Migration_Flows

Maret-Ouda J, Tao W, Wahlin K, Lagergren J. (2017). Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. Scandinavian Journal of Public Health. 45(17_suppl):14-19.

Poulain, M., Herm, A. & Depledge, R. (2013). "Central Population Registers as a Source of Demographic Statistics in Europe." Population, 68, 183-212. https://doi.org/10.3917/popu.1302.0215

Raymer, J., Rees, P., Blake, A. (2015). "Frameworks for Guiding the Development and Improvement of Population Statistics in the United Kingdom." Journal of Official Statistics 31(4): 699-722. http://dx.doi.org/10.1515/jos-2015-0041

Raymer, J., Biddle, N. & Campbell, P. (2017). "Analysing and Projecting Indigenous Migration in Australia." Appl. Spatial Analysis 10, 211–232. https://doi.org/10.1007/s12061-015-9179-6

Raymer, J., Guan, Q., Ha, J.T. (2019). "Overcoming data limitations to obtain migration flows for ASEAN countries. Asian and Pacific Migration Journal." 2019;28(4):385-414. https://doi.org/10.1177/0117196819892344

Raymer, J., Bai, X., Liu, N. et al. (2020). "Estimating a Consistent and Detailed Time Series of Immigration and Emigration for Sub-state Regions of Australia." Appl. Spatial Analysis 13, 411–439. https://doi.org/10.1007/s12061-019-09310-w

Raymer J, Bai X, Liu N and Wilson T (2020) Estimating a Consistent and Detailed Time Series of Immigration and Emigration for Sub-state Regions of Australia. Applied Spatial Analysis and Policy, 13, 411-439.

SDMX (2009). "SDMX Content-Oriented Guidelines – Annex 4: Metadata Common Vocabulary". Available at http://www.sdmx.org.

UNECE (2019). "Guidance on data integration for measuring migration." New York and Geneva: United Nations.

UNECE (2020). "Guidance on the Use of Longitudinal Data for Migration Statistics." New York and Geneva: United Nations.

UNECE-HLG MOS High-Level Group for the Modernisation of Official Statistics (2017). "In-depth review of data integration". Document ECE/CES/2017/8 for the Conference of European Statisticians meeting of 19-21 June 2017.

UNESCAP (2021). Stats Brief. Issue 29. "Big Data for Population and Social Statistics." April 2021.

UNSD (2022). "Handbook on Registers-based Population and Housing Censuses." New York: United Nations. https://unstats.un.org/unsd/demographic-social/publication/handbook-registers-phc.pdf

Van Hook, J., Bachmeier, J.D., Coffman, D., & Harel, O. (2016). "Can We Spin Straw into Gold? An Evaluation of Immigrant Legal Status Imputation Approaches," Demography 52 (1): 329-54. www.ncbi.nlm.nih.gov/pmc/articles/PMC4318768/

Willekens, F., & Baydar, N. (1986). “Forecasting place-to-place migration with generalized linear models.” In: Woods, R, Rees, P (eds) Population Structures and Models: Developments in Spatial Demography, London: Allen & Unwin, pp. 203–244.

Wilson T (2017) Methods for estimating sub-state international migration: the case of Australia. Spatial Demography, 5, 171–192.

Wiśniowski, A., Forster, J.J., Smith, P.W.F., 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): 1007-1024.

Zagheni, E., Weber, I. and Gummadi, K. (2017). “Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants.” Population and Development Review, 43: 721-734. https://doi.org/10.1111/padr.12102

Appendix I: Definitions of Key Concepts (A Glossary of Terms)

Big data: data sources that are categorized by their high volume, velocity and variety of data.

Compilation: Macro-data integration method which combines known subcomponents of migration to produce migration statistics.

Data integration: the process of combining data from two or more sources to produce statistical outputs.

Deterministic (unique identifier) matching: Data linking method to match records based on unique identification variables include on multiple datasets.

Deterministic (exact) matching: Data linking method to match records based on individual characteristics variables.

Emigration (flow) includes all persons leaving the country to become a part of another country`s resident population within a given year, including persons with national or foreign citizenships or stateless persons.

Foreign-born population (stock) includes all persons who reside in the country at a particular time who were born in another country.

Foreign citizen population (stock) includes all persons who reside in the country at a particular time who do not hold national citizenship, including those without citizenship (stateless).

Immigration (flow) includes all persons entering the country and becoming part of the resident population within a given year, including persons with national or foreign citizenships or stateless persons.

International mobility includes all movements that cross international borders within a given year.

International migration includes all movements resulting in a change in the country of residence (a subset of international mobility) within a given year.

Longitudinal database: A dataset that tracks the same individual information over time.

Macro-data integration: methods to produce international migration statistics via the integration of aggregated data from multiple sources.

Micro-data integration combines two or more datasets to create a new combined datasets which can produce international migration statistics.

Mirror statistics: immigration statistics to a (destination) country from another (origin) country may be used to partially compile emigration estimates from the origin country.

Native-born population (stock) includes all persons who reside in the country at a particular time who were born in the same country.

Official statistics: statistics produced by government agencies, which can inform debate and decision making both by governments and the wider community.

Personal Identifiable Information (PII): information that can be used to determine an individual's identity. This can be either alone, or in combination with other information.

Population register: a mechanism for the continuous recording of selected information pertaining to each member of the resident population of a country, making it possible to know the size and characteristics of the population at a given time.

Probabilistic Matching: Data linking method to match records based on the likelihood that a person is the same across two or more datasets, based on shared identification characteristics.

Spine: This micro-data integration term refers to the key (primary) dataset to which additional information (datasets) will be linked to.

Statistical adjustment: macro-data integration method used to produce new migration statistics by using estimates from one or more other data sources to adjust an existing estimate. Auxiliary data sources with strong correlation to the measure of interest may be used to adjust the statistics.

Statistical Matching: Data creation method which uses probabilistic models to develop synthetic data for individuals from records on two or more datasets that do not belong to the same individuals.

Statistical modelling: Macro-data integration method used when information is missing from one or more datasets. In these cases, information from one dataset is used to supplement/enhance data missing or of low quality in a second dataset, or shared characteristics in both datasets can be used to estimate characteristics missing from one of the datasets.

Synthetic data are data that are modeled from combining two or more real datasets, which has its roots in imputation methods. The goal is to reproduce the statistical properties and patterns of the existing dataset(s) by modelling its probability distribution and sampling it out to a new dataset.

Appendix II: Task Force Membership

The Task Force is comprised of representatives from countries, academia, and relevant international organizations. The United States and UNSD jointly coordinate the work of the Task Force.

Jason Schachter, United States (co-chair)

Haoyi Chen, ISWGHS (co-chair)

Julien Bérard-Chagnon , Canada

Tristan Cayn, Canada,

Jing Shen, Canada

Julibeth Rodriguez Leon, Chile

Miguel Ojeda Labourdette, Chile

Andres Felipe Copete ,Colombia

George Jashi, Georgia

Nikoloz Nikuradze, Georgia

Graciela Martinez Caballero, Mexico

Nicéforo Delgadillo, Mexico

Karima Belhaj , Morocco

Kare Vassenden, Norway

Aurelia Spataru, Republic of Moldova

Marcel Heiniger, Switzerland

Johanna Probst, Switzerland

Dominic Weber, United Kingdom

Becca Briggs, United Kingdom

Jo Zumpe, United Kingdom

Alison Beck, United Kingdom

Lauren Medina, United States

James Raymer, Australian National University

Jasmine Trang Ha, Western University

Gimapaolo Lanzieri, Eurostat

Elisa Benes, ILO

Rifat Hossain, WHO

Maria Isabel Cobos (Secretariat), UNSD

Meryem Demirci (Secretariat), UNSD

Appendix III (country case studies)

(see attached document)