Appendix III: Country case studies on data integration practices

These country case studies represent the experiences of Task Force members. The case studies summarize information from relevant countries' experiences, much of which may not (yet) be available for public readership. They provide greater detail than the examples referenced in the report.

Examples of data integration are organized around six main themes: (1) Data needs, (2) Collaborative framework for data integration, (3) Data sources, (4) Integration methodology, (5) Data evaluation strategy, and (6) Dissemination/ Communication of results.

Table of Contents

Canada

Overview

Data integration for international migration statistics has been used at Statistics Canada for many decades for the purpose of calculating estimates for international migrant flows.

Statistics Canada conducts both micro- and macro-data integration to produce migration statistics using administrative, survey and census data. These projects support many activities related to migration such as the calculation of demographic estimates, replacing the collection of migration characteristics in censuses and the creation of an administrative database for the purposes of studying longitudinal data.

Three data integration examples are included:

1. Estimating the number of emigrants
2. Probabilistic linking of administrative sources in the Longitudinal Immigration Database (IMDB)
3. Data quality assurance project for the 2016 Census

Example 1: Estimate of the number of emigrants (macro-data integration)

1. Data needs

The need for timely statistics on emigration that cannot be fully met by existing data for various reasons, including lack of timeliness, conceptual differences, incomplete coverage, biases, etc.

2. Collaborative framework for data integration

The collaborative framework involves Statistics Canada, the Canada Revenue Agency, and the US Department of Homeland Security (DHS).

3. Data sources

DHS data provide the number of Canadian children and adults migrating to the United States (main country of Canadian emigration).

From the Canada Revenue Agency, Child Benefit (CCB) data provide the number of children emigrating from Canada to a country other than the United States. CCB data is adjusted for coverage, bias and delay using tax data and population estimates. The number of adults emigrating to a country other than the United States is estimated using CCB and tax data. Tax data are also used to disaggregate the number of emigrants by province/territory, age and sex.

4. Integration methodology

After the relevant adjustments are made to tax and CCB data, the number of emigrants from DHS and CCB are added together.

Key challenges of this integration method include getting annual US data, communicating with Canadian data providers when data and/or trends are changing, and evaluating results by comparing them with those of other sources (Canadian and international).

5. Data evaluation strategy

The results are estimates of the number of emigrants by province/territory, age and sex, computed every quarter and 3 months after the end of the previous quarter. They are released as tables on Statistics Canada's website. These estimates are in turn integrated with those of other components of demographic growth to compute population estimates. A notable limitation of these estimates is that they underestimate emigration.

The more sources that are integrated, the higher is the risk that one source will have issues. Mitigating these risks requires communication with data providers and evaluation of each source. Moreover, as sources available to measure Canadian emigration are few and flawed, it is critical to do regular coherence analysis between these sources and to understand their strengths and weaknesses.

6. Dissemination/Communication of results

Methods are discussed with provincial/territorial partners at official meetings. They are also described in a methodological guide published on StatCan's website. Ad-hoc methodological documents are released from time to time.

Future plans include studying the use of emigration signals (e.g. sign of life approach) from various tax databases (short-term plan) and developing new models with border data when these data become available (long-term plan).

Example 2: Probabilistic linking of administrative sources in the Longitudinal Immigration Database (IMDB) (micro-data integration)

1. Data needs

The IMDB provides detailed and reliable information on socioeconomic outcomes of immigrants after their admission, such as employment income and mobility. It connects short- and long-term outcomes with characteristics at admission, such as immigrant admission category, source country and knowledge of official languages.

The database also provides information on pre-admission experience in Canada, such as issuance of work or study permits. The content of the IMDB has continuously been expanding. It now includes citizenship acquisition since 2005, as well as express entry data and details on the admission of economic immigrants. In addition, information on wages and salaries, as well as children's family outcomes, and information on settlement services provided to immigrants since 2013 has also become available.

2. Collaborative framework for data integration

The database brings together immigration information from Immigration, Refugees and Citizenship Canada (federal department) and taxation data from the Canada Revenue Agency (federal department). Provincial immigration departments also provide some immigration data, as well as a financial support to the initiative.

3. Data sources

The IMDB combines multiple administrative immigration data sources (citizenship, temporary and permanent residency, and settlement activity records) with annual tax files (T1 Family Files -T1FF) and employment records (T4 supplementary file) through probabilistic record linkages, and is updated with death information from the Canadian Mortality Database.

Agreements regarding the legal process for data integration, as well as data access/exchange between departments, are revised every 5 years. Challenges evolve regularly between departments involved and include a thorough approval process. All data integration files published by Statistics Canada are released on their website. Any change to the set of variables or coverage of the population must be approved at the ministerial level and information is updated on Statistics Canada’s website accordingly.

4. Integration methodology

The integration method uses name matching processes developed by methodologists at Statistics Canada. The immigration files are validated for accuracy using code sets and comparisons with the previous year's files.

No coefficients of variation are produced for estimates as the IMDB is considered a census of all immigrants within each of the reference years.

Of immigrants admitted between 1980 and 2020, 85.3% were linked to at least one T1FF record. Details on data accuracy for information from the T1FF may be consulted under the T1FF entry. Of immigrants admitted between 1980 and 2021, 70.3% were linked to at least one T4 record.

5. Data evaluation strategy

The data integration project was initiated in 1997. Since then, data sources have been updated annually and extensive documentation covers the validation and the analytical capacity of the product, as well as data dictionaries and descriptive articles that inform the official availability of data. While annual release is a recurring target, new data developments and changes in file configuration happen almost every year, bringing new challenges to the team.

6. Dissemination/Communication of results

Data tables and articles are prerequisites to officially release data products at Statistics Canada (for both integrated or survey data). Multi-dimensional tables and at least one official descriptive article is released to the public when updated data are released. As mentioned earlier, the data integration activity is published on Statistics Canada's website.

It is generally agreed that over time administrative data integration is an efficient and cost-effective tool. However, investment and effort are required for its creation and expansion. Once data acquisition and processing has been completed, and the approvals and training of employees are in place, the cost and maintenance of integrated data is lower than other data collection processes. The future plan is to augment the administrative data portfolio by reaching out to additional data providers (such as provinces and territories) and attempt to add additional federal data sources (such as health and education administrative databases).

Example 3: Data quality assurance project for the 2016 Census (macro-data integration based on micro-data)

1. Data needs

Citizenship and immigration census topic have the benefit of a wealth of available administrative data that can be used for certification. Moreover, IRCC's administrative immigrant landing file (ILF) was linked to the 2016 Census for the purpose of improving processing and the addition of variables, meaning it could be used to directly confirm the accuracy of certain immigration data, such as year of immigration and place of birth, as well as help certify variables like immigrant status via linkage rates.

2. Collaborative framework for data integration

The Census 2016 certification project brought together immigration information from Immigration, Refugees and Citizenship Canada (federal department) and taxation data from the Canada Revenue Agency (federal department).

3. Data sources

The Immigration Landing File, or ILF, includes all immigrants who have landed (i.e., acquired Permanent Resident status) in Canada from 1980 to 2016. It is a registry of all incoming immigrants and their characteristics at the time of becoming a permanent resident of Canada. The file is very useful in determining exactly how many individuals came to Canada as immigrants over any period of time between 1980 and 2016, as well as any characteristic available on the ILF (e.g., place of birth).

The landing file, however, cannot be used to determine the current population of immigrants in Canada by the same characteristics, since it does not track individuals over time, or get updated when they leave the country or die. However, by providing the exact number of incoming immigrants, it produces an upper bound for current immigrant population estimates.

4. Integration methodology

Calculating measures like match rates (how often a reported/imputed value matched the linked value from the ILF) helped gauge the accuracy of imputation and overall estimation with high detail for place of birth and year of immigration.

This was an additional tool in the certification process of the Census. This new process increased the quality of census estimates as it informed on the weighting strategy and external data was used in the imputation process.

5. Data evaluation strategy

As an additional tool for census certification, such data integration adds to the precision of census coverage and the quality of estimates. The 2021 census directly used administrative data to provide certain information on respondents and reduce response burden. The 2021 census replacement project used the result of the certification success of the 2016 census.

6. Dissemination/Communication of results

The project was not created for publication and no specific product was disseminated (aside from internally at Statistics Canada).

Additional resources

Published guides about the census 2016 certification process

Published guides about the census 2016 certification process

Fact sheet on the replacement project objectives for the census 2021

Place of Birth, Generation Status, Citizenship and Immigration Reference Guide, Census of Population, 2021

Chile

Overview

Previous efforts to integrate administrative records with the census have been at the macro level. The first effort was undertaken in 2014, because the 2012 Census was deemed to be invalid by a National Commission, and post-census migratory flows showed a significant increase in immigration to the country. The Ministry of Interior therefore estimated the number of foreigners living in Chile based on data from the 2002 Census and permanent residence permits (PD) issued between 2002 and 2014.

After conducting a census in 2017 immigration flows to Chile increased rapidly, so in 2018 the National Statistics Institute (INE) and the Department of Foreign Affairs and Migration (now known as the National Migration service) initiated a new methodology for estimating the foreign population by integrating census data with administrative records related to migratory flows.

The latest approach combines macro- and micro-data integration methodology, the result of collaborative work between INE and four other public institutions. Technical work included the analysis and cleaning of administrative data from administrative sources (micro-level), which were combined with data from the last population census (macro-level). This process provided an updated estimate of the total number of foreign residents in the country at different geographic levels, and by sex, age, and country of origin.

Example: Estimate of the foreign population living in Chile 2018-2019-2020

1. Data needs

Migration flows to and from Chile have intensified over the past decade, creating a growing need to update the country's migration policy, for which official statistics only partially cover the phenomenon. Since the measurement of the foreign population in the census of April 2017, there has been an enormous increase in applications for residence permits, which quickly made the census estimate obsolete.

Upon looking at administrative data to update these estimates, several difficulties were discovered. These included the fact that the register of border crossings only covered entrances to and exits from Chile by legal points of entry, and that the system for issuing residence permits covered only the applications of foreigners within national territory, while it excluded applications from abroad.

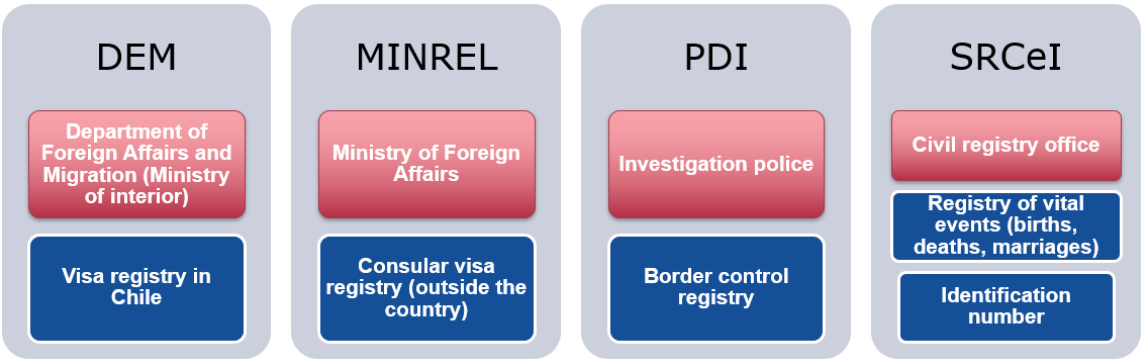
In addition, there were no household surveys specifically focused on migration, and there was an interval of fifteen years between the latest (2017) and previous population and housing censuses (2002). Population censuses in Chile used to be the main source of information for international migration, but this was during a period when it accounted for a small share of the population (1% or less), which was not the case after

2017. Therefore, a new methodology was needed to determine the number and characteristics of the foreign population in Chile, as well as to measure recent migration flows. This methodology was envisioned as an integration of administrative data from public services related to migration with census data on the population.

2. Collaborative framework for data integration

The INE collaborated with four institutions: (1) The Department of Foreign Affairs and Migration (DEM), (2) the Ministry of Foreign Affairs (MINREL), (3) the Investigation Police (PDI), and (4) the Civil Registration and Identification Office (SRCeI). All four partners provided data, as summarized in Figure 1 below. Additionally, INE and DEM worked together to design the methodology and both institutions were responsible for the final estimates, as well as dissemination and publication of data.

Figure 1: Administrative data for measuring regular immigration flows



Source: National Statistics Institute (INE) – Department of Foreign Affairs and Migration (DEM)

One of the main challenges has been that, for the last three set of estimates, there was no regulatory framework that explicitly permitted the exchange of information between different institutions for statistical purposes. This represented a significant challenge because the estimates were tied to specific cooperation agreements that were often dependent upon political will of the moment. Because data exchange takes place between counterpart institutions, it would be beneficial for INE to have access to nominated data for future estimates. This has recently been advanced with the signing of a collaboration agreement between INE and DEM, which will permit INE to access this nominated data, though access to data from other institutions still needs to be clarified.

3. Data sources

The Department of Foreign Affairs and Migration (DEM) of the Ministry of Interior and Public Security is the entity in charge of immigration policy, including the granting of residence permits and nationalization (which can be requested while inside national territory, with no need to leave the country). DEM contributes by indicating the "number of people who declared their intention to reside in Chile by applying for a

residence permit." This variable is used for the period after the date of the census (April 19, 2017), since the 2017 Census already provided data on the number of immigrants up to the day of the census.

The Ministry of Foreign Affairs (MINREL) contributes to the formulation of Foreign Policy and granting of residence permits requested outside national territory. It contributes by indicating the "number of people who declared their intention to reside in Chile by applying for a visa" (likewise covering the period after the 2017 Census).

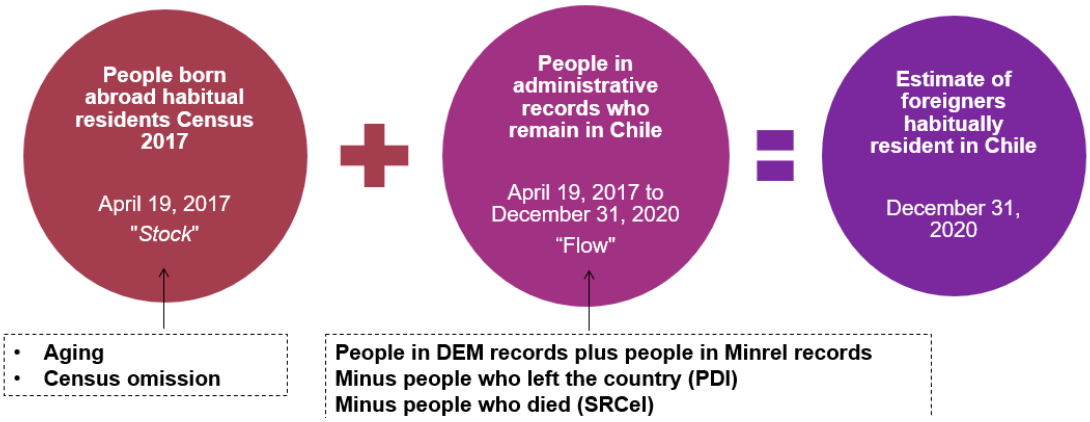
The Investigation Police (PDI), the entity that controls Chilean borders, contributes by indicating the number of people who entered the country after the 2017 Census, as well as the number of people from this group that later left the country.

The Civil Registration and Identification Office (SRCeI) is the entity in charge of granting identification documents and keeping records of births and deaths that occur in the country (including other variables such as marriages, divorces, and other vital statistics). The SRCeI contributes by reporting the number of deaths of immigrants, which is subtracted from the estimated number of foreign residents.

4. Integration methodology

The population census forms the basis of the estimate, but it is adjusted for census omissions and for mortality of the population. To the census figure were added information contained in administrative records on foreigners who entered the country after the census date, while subtracting those who were outside the country or who have died during the period up to the date of the estimate (December 31 of 2018, 2019, and 2020). The integration methodology is summarized in Figure 2 below.

Figure 2: Methodology for estimating foreign residents in Chile



At the macro-level, Chile's methodology includes combining information from the 2017 Census with information from administrative records created after the census, according to variables of sex, age group, country, and place of residence.

To link records at the micro-level, we searched the DEM database for relevant cases, eliminating duplicates and using information from initial residency applications. For information on applicants for a residence permit after the date of the census, we added MINREL data from records of residence permits (visas) requested at the consular level and matched them with information from SRCeI, through their Chilean identification number (RUN), in order to identify those who died and subtract them from the estimate.

On the other hand, to identify people who have left the country, as well as those who have entered the country after applying abroad for a consular visa, data was matched with the records from border control, which are administered by the Investigation Police (PDI). This process was carried out by the PDI, who receive personal information from the DEM database. The PDI then generates an identification algorithm that includes the RUN, and adds variables of surnames, first and middle names, and date of birth. When the person does not have a RUN, their surname, first and middle names, and date of birth are the main identification variables. For the 2020 estimate, the level of matching was close to 90%.

Regarding micro-integration, a match percentage is calculated between the base information (from the Immigration Department and the Ministry of Foreign Affairs) and information from the control of borders (PDI) and deaths (Civil Registry). This percentage has been increasing over the years, reaching 90% by 2020.

5. Data evaluation strategy

To evaluate the methodology, several discussion panels with experts from ECLAC and academic specialists in migration were held, as well as in working groups with other State institutions. Regarding a calculation of a possible coverage error rate compared to other sources, this was not possible since estimates were not comparable to other sources. For example, immigration flows are recent, while survey frames are based on outdated census address listings and do not consider the diversity of migrant locations in their methodological design. Despite this, comparisons have been made in terms of structure by age, sex, and country, with two large surveys: the labor force survey (ENE) and the National Socioeconomic Characterization Survey (CASEN), obtaining similar results for main groups of migrants.

Some verification actions are carried out by Immigration Department, which is the institution that compiles and processes information from administrative records. These actions consist of comparing datasets sent by the PDI, given that information accumulates progressively year by year, and differences or problems could be found in dataset between different years, which allows for clarifying and improving information.

6. Dissemination/Communication of results

INE recently published the third version of their estimates, which measure the foreign population residing in Chile in three different years (December 31 of 2018, 2019, and 2020).

In addition to estimates, an important outcome of this project is a recently signed legal commitment to work collaboratively among five public institutions.

This commitment resulted in the consolidation of an official methodology, shared by the five public institutions, for establishing the official number of immigrants residing in the country, as well as for measuring their characteristics and distribution at the national, regional, and communal levels.

Furthermore, the commitment includes the provision of intercensal and up-to-date estimates of total immigrants by sex, age, country of nationality, and place of residence, for use in the creation of public policies that more accurately reflect the nature of Chilean society.

Related links:

Estimates of foreigners habitually resident in Chile as of December 31, 2018-2020 (A technical report for entire country, in Spanish)

Estimates of foreigners habitually resident in Chile as of December 31, 2018-2020. (A technical report on regional and communal disaggregation, in Spanish)

Estimates of foreigners habitually resident in Chile as of December 31, 2018-2020 (Tabulated data, in Spanish)

Statistical tables, publications, infographics, and methodology on demographic statistics (in Spanish)

Georgia

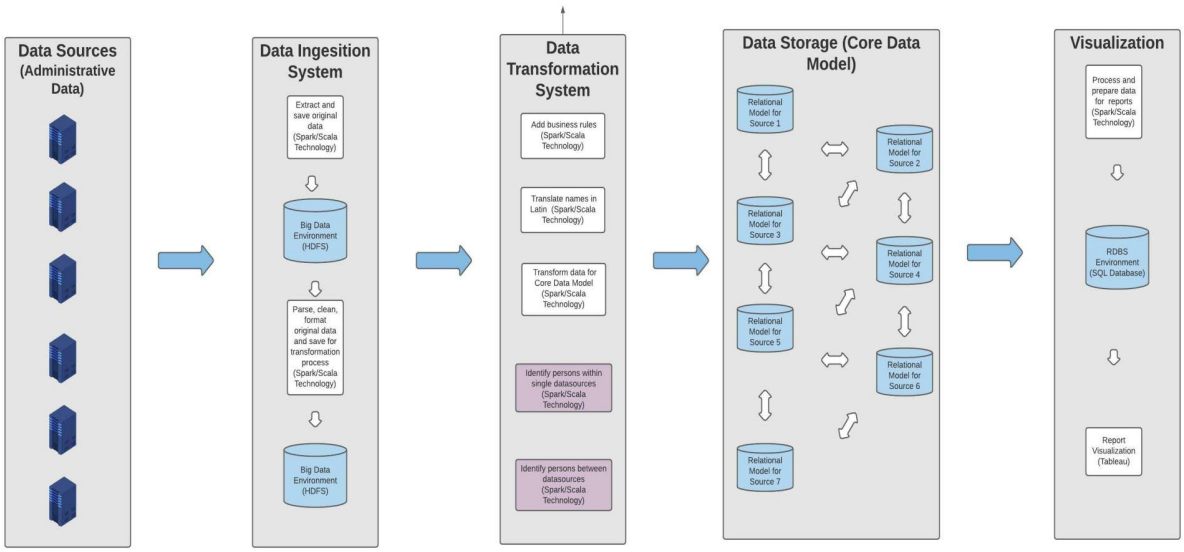
Example: The Unified Migration Data Analytical System (UMAS)

1. Data needs

Sound migration management policy greatly depends on decisions made on the basis of migration data analysis. Logically, data related to migration is dispersed in different state agencies. These agencies, according to their mandates, conduct analysis of their own data; however, they need access to analytical reports produced by other state agencies, as well as cross-institutional reports (capturing and processing data from several state agencies).

In 2016, work started by the Public Service Development Agency (PSDA), under the Ministry of Justice of Georgia, aimed to process data accumulated by different state agencies in one space, for analytical purposes. This newly created analytical system is called the Unified Migration Data Analytical System, or UMAS. It ensures that records extracted from databases of different state agencies are integrated into one database (micro-integration). The System was developed based on a Big Data platform, as summarized in Figure 1 below.

Figure 1: UMAS Data flow architecture



2. Collaborative framework for data integration

Data integration is conducted within the PSDA by in-house human resources and with the help of invited experts. In order to ensure effective communication with data provider state agencies, an inter-agency Working Group (i.e., UMAS WG) functions within the State Commission on Migration Issues (SCMI), which is the collaborative body designed to manage migration on the principle of the "Whole-of-Government-Approach."

3. Data sources

Data from the Public Service Development Agency, the National Agency of Public Registry, the Ministry of Internal Affairs, and the Ministry of Education are stored and processed in the UMAS. Data descriptions are provided in Table 1 below.

Table 1. Description of data sources

Source	Data Description
Public Service Development Agency	Residence permits issued to foreigners
Public Service Development Agency	Georgian citizenship (granted / revoked)
National Agency of Public Registry	Business registry data (data on foreigners doing business in Georgia)
National Agency of Public Registry	Real estate registry - information about foreigners who own real estate in Georgia
Ministry of Internal Affairs	Border crossings
Ministry of Foreign Affairs	Visas issued by MFA's consular department, as well as Georgian diplomatic representations and consular services abroad (immigration visas)
Ministry of Education (Education Management Information System)	Foreign students

4. Integration methodology

The main challenge related to data integration is linking records received from different agencies, which is done through personal identification fields such as first name, last name, date of birth, citizenship, gender, and personal ID or passport number (under strict control of the Law on Personal Data Protection). A deterministic Record Linkage Algorithm is used for this purpose, with Jaro Winkler distance.

Two main approaches are used when working within the analytical systems: goal-driven and data-driven. In the first case, demands towards the system are clearly defined; while in the second case, demands are vague and analysts are tasked to define (within identified data sources) fields, which might be interesting for the purpose of the system. In the case of UMAS the second approach was chosen, which created more data than needed.

Currently, two main challenges for data integration are: (1) Possible financial shortages related to the growing need for IT specialists, and (2) Difficulties improving the quality of data sources.

5. Data evaluation strategy

The quality of analytical reports should be a priority, meaning that reports extracted from UMAS should be identical to the results of internal reports produced by state

agencies. The second important task is improving data quality within data sources, based on recommendations, in order to ensure better data integration.

6. Dissemination/Communication of results

The main purpose of UMAS is generating analytical reports. Hence, a platform for providing requested reports is built into the System. When requesting new reports, a team responsible for UMAS administration creates these reports and ensures access for the requesting agency.

Information on actual challenges regarding efficient UMAS operation, such as improving data quality, refining methodology of producing analytical reports and optimization of data related business processes are shared within the format of the interagency WG on UMAS. It should also be noted that providing recommendations to improve data quality is one of the purposes of UMAS.

Mexico

Example: Compiled information about the foreign population

1. Data needs

In the Mexican Housing and Population Census 2020, the main questionnaire, which was applied to the entire population, collects information on people's country of birth. On the other hand, the expanded questionnaire, which was applied to a sample, includes a question about nationality. As the sample design is not focused to estimate foreign population, whose presence is small in the country, some results have low statistical precision. Data integration is thus needed to compile information about the foreign-citizen population (i.e., people whose are not Mexican by nationality). These statistics are vital for the design of public policies, programs, and services such as the right to identity and nationality, health, education, and employment needs in Mexico. For that, it is important to include a question on people's nationality in the main questionnaire.

2. Collaborative framework for data integration

The Migration Policy, Registration, and Identity of Persons Bureau (UPMRIP, by its Spanish abbreviation) oversees data integration about the foreign population. Data sources come from UPMRIP and other governmental agencies, namely:

- Migration statistics on permanent residents and temporary migrants collected by UPMRIP
- The Population and Housing Census 2020 which was collected by the National Institute of Statistics and Geography (INEGI)
- The number of naturalized foreigners which was issued by the Secretariat of Foreign Relation (SRE)
- Deaths of foreigners registered and occurring in Mexico which was collected from INEGI's Vital Statistics database

3. Data sources

The expanded (sample) questionnaire of the Population and Housing Census 2020 was used as the primary source for estimating the total population of foreigners (regardless of migratory status) in Mexico. This questionnaire contained a question on whether the person has Mexican nationality. The expanded questionnaire was applied to approximately 4 million households selected through a stratified probabilistic sample clustered in a single stage. A known constraint for estimates drawn from the expanded questionnaire of the Census is that it only represents population residing in private homes. The main questionnaire also collects information about migrants living in shelters.

The number of foreigners with regular migratory status was compiled from various components: existing and new permanent residents, existing and new temporary migrants, naturalized citizens, and deaths of foreigners in Mexico. This was done by integrating migration statistics from UPMRIP, naturalization records from SRE, and mortality data from INEGI.

The number of foreigners with irregular migratory status was calculated by subtracting the number of foreigners with regular status from the total foreigner population estimated using the Census.

Regular status in Mexico is determined from the authorization of stay or residence permit issued by the immigration authority to a person of nationality other than Mexican. This is not included in the population census, so this information can only be obtained from residence permits, which are identified as "Migration statistics on permanent residents and temporary migrants collected by UPMRIP."

Because the calculation is based on country of nationality, "The number of naturalized foreigners which were issued by the Secretariat of Foreign Relation (SRE)," this source registers people born in another country who obtain Mexican nationality and cease to be part of the foreign immigrant universe. Deaths of registered foreigners occurring in Mexico was collected from INEGI's Mortality database.

Finally, the population census considers residents born in another country, who do not have Mexican nationality and who do not have a blood relationship with a Mexican, as foreigners. Because the census does not ask about immigration status, the foreign-citizen estimate includes both residents with regular or legal status and those who do not have a residence permit, either because the document expired or because they entered Mexico irregularly and decided to stay and reside without the corresponding immigration authorization.

4. Integration methodology

Macro-data integration methodology was chosen based on a theoretical understanding of the relationship between the subcomponents of desired migration statistics. A major challenge for the project is that aggregate data obtained from various sources contain different assumptions and temporality.

Several adjustments were made to data sources prior to integration. First, a temporal cut-off date (e.g., date of registration, event, procedure) was selected for all data sources. The date of the Census (March 15, 2020) was selected for this cut off. Second, it was assumed that foreigners who obtained permanent residence status in Mexico would continue to stay in the country, except for very old individuals. Estimates of permanent residents thus include those who have an open file with the migration authority and exclude those aged 99 or older. Third, estimates of temporary residents were adjusted by the date of their documents' issuance or renewal to ensure validity beyond the cut-off date.

A compensating equation, similar to the basic equation used in demographic analysis, adds and subtracts the components of a population (deaths, births, emigration, and immigration). Specifically, the equation is defined as:

$$P_{rrt} = P_{rp0} - N_{et} + NR_{tpr} - D_{et}$$

Where:

P_{rrt} = Population with regular stay at time *t*.
P_{rp0} = Population with permanent residence at the beginning of the period.
N_{et} = Naturalizations of foreigners at time *t*.
NR_{tpt} = New temporary and permanent residents of year *t*.
D_{et} = Deaths of foreigners at time *t*.

Data substitution:
P_{rrt} = (431,895) - (75,516) + (12,373) - (33,213)
P_{rrt} = 381,931

The estimate indicates that as of March 15, 2020, there were 381,981 foreign people with regular residence in Mexico.

A second equation was then used to estimate the population of foreigners in an irregular migratory situation in Mexico:

$$PERI_t = PEC_t - PRR_t$$

Where:

$PERI_t$ = Foreign population with irregular residence in Mexico at time *t*
PEC_t = Census foreign population (estimated) at time *t*
PRR_t = Foreign population with regular residence at time *t*

Data substitution:
$PERI_t$ = (390,065) - (381,931)

The final estimate indicates that as of March 15, 2020 there were 8,134 people with irregular residence in Mexico, that is, without migration documentation proving their residence. This estimate of the resident foreign population without legal residence authorization is low and suggests a need to use a longer time series for other sources of information to improve estimates of this specific population.

5. Data evaluation strategy

The estimated result may provide multiple interpretations and the estimation process can still be improved. For those who know the dynamics of migration flows that cross through Mexican territory to reach the United States, the obtained result will seem illogical. However, it should be considered that the estimates are only made on the basis of residents in private homes and omit the population in shelters or that do not habitually

reside in Mexico. This is why to accurately calculate this population it is necessary to carry out another type of estimation, one not comparable with data from censuses or surveys carried out in homes.

Furthermore, mortality time series data are only available from 2012. The project team continues to seek alternative sources of administrative registrations that strengthen information on deaths.

6. Dissemination/Communication of results

Some of this research has been released publicly[1], while the estimates of migrants with irregular residence are currently only used internally. Full results are expected to be published in 2024.

1 http://www.politicamigratoria.gob.mx/work/models/PoliticaMigratoria/CEM/Estadistica/anuario/ANUARIO_2020.pdf

Moldova

Overview

The National Bureau of Statistics (NBS) is relatively new to data integration, having only started its first project in 2019 to integrate data for the period 2014 (year of the latest census) to present. In the first project, which is described in more details below, they disseminated estimates of the usual resident population (URP) at the national level for the period 2014-2019 and produced two analytical reports on population and migration.[2] The project provided technical training and assistance to develop competencies within the NBS to produce annual estimates of international migration (on the basis of individual border crossing data, instead of legally endorsed aggregated migration data) and the URP. Within a new project, which aimed to estimate international migration and URP at the subnational level, international migration was estimated at the level of municipalities, for the period 2014 - 2020, and URP, at the same level of detail, for the period 2014-2022.

Example: Revision of the usual resident population (URP) estimates at the national and subnational levels for 2014-2022.

1. Data needs

The demographic situation of the Republic of Moldova is one of the most publicly debated topics. A general concern about the steady decline in the country's population over the last 30 years, and unfavorable changes in its structural characteristics, creates an increasing need for careful and in-depth study of migration patterns.

Past population estimates in the Republic of Moldova have been inaccurate resulting in a large over-estimation of population size, explained by use of the "legal residence" concept for annual estimates, which differs from the internationally recommended "usual residence" concept used in the latest census.[3] The Population and Housing Census of 2014 (adjusted for undercount) provided accurate estimates of the usual resident population (URP), but annual updates to census-based estimates in subsequent years were impossible without parallel accurate estimates of annual net migration. The country's State Population Register, maintained by the Public Service Agency, includes some migration data, but this is likely incomplete because only legally registered migrations (i.e., persons who officially declared they were changing country of residence) are recorded. Some additional data on migration are produced by the Bureau

[2] Both reports available here: https://statistica.gov.md/en/statistical-methodologies-104.html

[3] See for example: Producing reliable mortality estimates in the context of distorted population statistics: the case of Moldova, Olga Penina, Dmitri Jdanov, Pavel Grigoriev, 2015; Adapted Global Assessment of the National Statistical System of Moldova, Vera Herrmann, Claudia Junker, Bronislava Kaminskiene, Günter Kopsch, Jason Schachter 2013

for Migration and Asylum, but this is also an incomplete data source because annual numbers of return migrants are not recorded.

2. Collaborative framework for data integration

Understanding the critical importance of this topic, the NBS, with the assistance of UNFPA and SDC who supported the operation, devoted significant efforts to identify proper data sources and develop a data integration methodology. The concern is not only to derive reliable estimates, but also to establish a process for updating intercensal population estimates every year in a timely manner. The data integration project thus involves collaboration with the General Inspectorate of Border Police from the Ministry of Interior (GIBP), the Public Service Agency (PSA), and the Electronic Governance Agency. The Collaborative framework is enhanced by a joint governmental technological platform for data exchange, MConnect,[4] where personal and other data held in public registers and administrative data sources can be uploaded by public authorities and institutions who collect them. Once uploaded, the data can then be used, according to requests endorsed by the Electronic Governance Agency, by other public institutions to fulfill their relevant functions.

A key challenge for the Collaborative framework is that the various government offices have not agreed upon a common definition of individual and personal data. In this context, individual data were perceived to include non-confidential information about individuals, whereas personal data contain confidential information that enables record-linking. For the 2014-2018 period, up to present, NBS and GIBP have agreed with PSA to perform record-linking for datasets held by each agency and provide pseudonymized[5] linked data to NBS to produce population and migration estimates. In December 2022, amendments to the 2017 national Law of official statistics[6] were approved by Moldova's Parliament, through which NBS aimed to introduce the necessary clarifications to the legal framework, to be able to access and process personal data directly, as part of the individual data concept.

3. Data sources

There are a few possible data sources in Moldova that can be used to produce international migration estimates. Potential data sources are evaluated both for reliability and their potential for timely updates in subsequent years. Two data sources were found to satisfy these requirements: (1) the registration of cross-border international movements conducted routinely by GIBP, and (2) data on internal migration from the State Population Register (SPR) on the locality of legal residence inside the country

[4] https://mconnect.gov.md
[5] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN
[6] https://www.legis.md/cautare/getResults?doc_id=134938&lang=ro

4. Integration methodology

The integration methodology involves linking of individual records from the State Population Register provided by PSA, with data on border-crossing records provided by the GIBP. For current projects, linking was done for Moldovan citizens only, in two stages. Record-linking at the first stage (by name, surname and birth date) was undertaken by GIBP within the border-crossing dataset to identify Moldovan citizens with dual citizenship (crossed a border with documents other than a Moldovan passport). At the second stage, the dataset with Moldovan citizens who have crossed the border are linked (by Identification Number of Person - IDNP) with data on their legal place of residence from the State Population Register. The resulted datasets on border crossings by Moldovan and foreign citizens, after pseudonymization of IDNP by DIBP, was sent to NBS to perform the estimation procedure.

Besides working with pseudonymized data, the key challenge for these data integration projects is to measure international migration based on a large number of border-crossing movements recorded each year, most of which were not migration related. On average, there were about 21 million movements per year in the period 2013-2018 (compared to the 2.7 million usually resident population in 2018). Additionally, migration estimates for any single year require assessing movements for three consecutive years (the year before, the year of movement, and the year after).

Specifically, a measure of duration abroad was calculated for each individual. This includes consecutive movements, and the cumulative addition of these intervals, for anyone crossing the border during a period of one full year before and one full year after any specific movement of entrance or exit to/from the country. Then, an immigration status is applied if an individual spent more than 275 days abroad during the year before the movement and then spent more than 275 days in the Republic of Moldova. An emigration status is applied if an individual spent more than 275 days in the Republic of Moldova in the year preceding the movement and then spent more than 275 days abroad during the year of movement. Individuals with no movements indicating a migration status will be categorized as non-migrants in the calculation year.

An additional methodological step was taken to identify and remove “illogical itineraries.” Illogical itineraries are cases when a specific individual is registered with two consecutive movements in the same direction. Any entry to the country can be followed (or not) by an exit and it would be illogical to have two consecutive entries (or exits) without an exit (entry). These “illogical itineraries” may be the result of errors in the assignation of movements to a specific individual (the middle missing movement was not properly assigned to the individual) or that the individual crossed the border without being registered by the system (like in the case an individual left or entered the country through the Transnistrian border). In practice, it was decided to “impute” a new movement in cases where the illogical itinerary created an interval of less than 183 days.

The imputed movement was inserted in the middle of the illogical interval allowing the incorporation of this individual into the migration estimates calculations. There were annually less than 2% of cases that have been imputed, and the average length of the imputed intervals was less than 30 days. About 19% of individuals had at least one imputed interval but in most of cases, imputed intervals were of a few days (the average imputed interval was of 15 days). In extreme cases, when an individual has illogical intervals of 183 days or more, this individual was excluded from the calculations (less than 1% of individuals crossing the border in any specific year).

5. Data evaluation strategy

Exhaustive checks have been conducted to verify as far as possible that imputations and exclusions for the illogical itineraries do not significantly affect the calculation of migration estimates. The resulting migration estimates have been further checked for completeness and consistency over the four years and reliable international migration estimates for the Republic of Moldova have been calculated for the first time. These estimates allowed for the calculation of revised population estimates by usual residence for 2014-2017 and onwards, which have been published by NBS on an annual basis.

6. Dissemination/Communication of results

The NBS collaborated with UNFPA to organize a public event in July 2019 to release revised population estimates of the Republic of Moldova for the period 2014-2018, and again in July 2022, to release estimates at the subnational level for the period 2014-2022. Prior to this, NBS organized several meetings with the expert community, including central authorities (as main data users) and academia and researchers to present the methodology used to estimate international migration and the usual resident population.

The result was an integrated data set of border crossings and data on the locality of residence at the individual level, but still pseudonymized. Two reports are available from the first sub-project: (a) Revision of the population estimates 2014 – 2019 of the Republic of Moldova in alignment with UN recommendations[7], and (b) International Migration Estimates for the Republic of Moldova 2014-17. [8]

[7] https://statistica.gov.md/files/files/Metadate/alte/Revision_pop_estimates_2014-2019_R_M_in_alignment_with_UN_recommendations.pdf
[8] https://statistica.gov.md/files/files/Metadate/alte/International_Migration_Estimates_R_M_2014-17.pdf

Switzerland

Overview

The Swiss Federal Statistical office (FSO) does not currently apply any data integration techniques in the strictest sense[9] to produce migration statistics. The only source for such statistics are the population registers.

The production of annual migration data is an integral part of the Population and Households Statistics (STATPOP). It is based on one of the various surveys conducted within the framework of the federal population census. These statistics provide information regarding population size and composition at the end of a given year, as well as population change and its demographic components during the same year.

In 2010 the traditional decennial census was replaced by an integrated statistical system which provides annual data. The new census system combines the use of administrative registers with sample surveys. It consists of four different annual surveys, among them – and of particular importance to migration statistics – the register survey.

Example: Integrated data of foreign nationals to calculate migration stock and flow statistics

1. Data needs

Whereas Swiss citizens are registered at the local level only, foreign nationals can appear in several registers (in one or more local population register, as well as in one of two federal registers). Since there is no single data source which provide a comprehensive picture of migration of non-nationals it is good practice to use multiple sources. While in some instances these sources will lead to the same findings, in other instances details and trends may differ. However, this does not necessarily mean that one source is 'right' and another is 'wrong' or that one source is 'better' than another.

Therefore, for all categories of foreign nationals (e.g., permanent residents, asylum-seekers, etc.) it is necessary to select the source on which the number of migrants (stocks) or migratory flows to be considered in official statistics is to be based. Timeliness, completeness, reliability, and other factors are critical to prioritize sources.

2. Collaborative framework for data integration

The register survey is the primary source for STATPOP. It exploits existing and harmonized administrative data from both centralized and local population registers and is therefore particularly well suited for producing data on migration flows: Registers are

[9] production of international migration statistics outputs (including both statistical reports and data) by (1) matching data between multiple data sources at the unit record level ("micro level") and/or (2) combining data sources at the "macro level" using statistical modeling.

dependent on the timely registration of the population and keep track of all changes of residence, thus ensuring a continuous update of individual records. Administrative sources are frequently able to do so despite some of their well-known drawbacks (e.g. statistical data collection is not a priority, definitions and coverage depend on legislation and administrative rules).

The nationwide social security number which uniquely identifies a person (universal PIN) plays an important role both in the maintenance of administrative registers and in their statistical use, particularly in linking information from register and survey data.

3. Data sources

Data from three types of registers are integrated:

- Local population registers: there are 2,500 local registers maintained by the municipalities or in some cases the cantons[10]
- Central Migration Information System (an aliens register maintained by the State Secretariat for Migration at the federal level) for the legally resident non-national population (i.e., people in possession of an official permit of stay);
- Ordipro information system (an aliens register maintained by the Federal Department of Foreign Affairs) for non-nationals that are entitled to privileges and immunities and not subject to Swiss immigration laws, i.e., staff of diplomatic missions, consular posts, permanent missions and intergovernmental organizations (including spouses, partners, unmarried children as well as private household employees).

4. Integration methodology

Characteristics relating to an individual can differ from one register to another. The Central Migration Information System provides some exclusive variables that are missing at the local level. They are added to local population register records pertaining to the same observations (individuals or migratory events). These variables include "nationality of spouse/partner," "purpose of stay" or "migration motive," "date of first issue of permit," "expiry date of permit" and for asylum-seekers detailed information about different administrative steps of the asylum procedure.

This data "enhancement" procedure has been implemented into annual migration data production and can be considered a modest form of data integration. Data collected at the federal level are matched with data collected at the cantonal or municipal level. The availability of the social security number in each of the involved registers allows the direct linking of records of one individual across different sources.

[10] There is no central population register covering the entire (national and non-national) population of Switzerland. In addition, FSO does not maintain a statistical population register, but rather collects statistics on the total population (STATPOP). It receives data extracts containing pre-defined variables in a pre-defined standardized data format from each register at regular intervals.

Beyond record-linking, "data integration for measuring migration" in Switzerland refers to the deliberate choice of one administrative register rather than another for specific sub-groups of the non-national population.

The following decisions have been established to produce annual migration stock and flow data:

- The register of reference for the "permanent resident foreign population" (long-term migrants, i.e. stocks of persons who have resided in Switzerland for at least 12 months and flows of persons who arrive and leave Switzerland and have been issued legal settlement or residence permits that are valid for at least 12 months) are the local population registers.
- The register of reference for the "non-permanent resident foreign population" (short-term migrants, i.e., stocks of persons who have resided in Switzerland for less than 12 months and flows of persons who arrive and leave Switzerland and are in the possession of legal permits that are valid for less than 12 months) is the Central Migration Information System.
- The register of reference for foreign nationals who have been issued legitimization cards by the Federal Department of Foreign Affairs is the Ordipro information system.

5. Data evaluation strategy

Local and federal data sources are not mutually exclusive. They have partially overlapping observation units as well as partially overlapping variables. This overlap is due to the fact that local population registers are supposed to include all resident foreign nationals, who in addition should be included in one of the two federal registers as well.

However, no attempt is made to systematically compare registers in order to identify individuals (stocks) and migratory events (flows) that are recorded in multiple registers or in a single register only (and therefore missing elsewhere). The count of migrants and migration inflows and outflows is neither augmented by observations found in an administrative database that is not considered to be the "register of reference" for a specific population sub-group, nor are observations in the "register of reference" disregarded or deleted if they are not found in another register.

6. Dissemination/Communication of results

In the field of migration statistics FSO makes a clear distinction between statistics and estimates, the former broadly representing the product of an (as far as possible) exhaustive compilation of records from primary data source(s), the latter the outcome of probabilistic/statistical models which might involve combining information from various sources. Official Swiss migration statistics do not include any estimations – neither at the individual nor at the aggregated level.

United Kingdom

Overview

The Office for National Statistics (ONS) are responsible for producing official statistics on international migration to and from the United Kingdom (UK).

Migration statistics are vitally important. They feed into population estimates, which underpin a huge range of policy decisions and inform public debate. The users—including central government policymakers, local public sector staff, and members of the public—are interested in how migration impacts society and the economy, especially since the coronavirus (COVID-19) pandemic.

Example: Administrative-based Migration Estimates (ABMEs) to improve measurements for all forms of migration into and out of the UK

1. Data needs

The new data integration project began in 2019 and is still in progress. The key goal is to place administrative data at the core of evidence on international migration. For many years, migration flow estimates in the UK were underpinned by the International Passenger Survey (IPS), a random face-to-face survey at airports and other ports of entry and exit. It has been long acknowledged that the IPS was being stretched beyond its original purpose. For example, the survey captures traveler intentions to stay longer than 12 months, as the basis for proxying long-term migration. Subsequent analysis has highlighted that often intentions didn't match behaviors, and that headline estimates of migration based on the IPS were inaccurate. These findings stimulated the current data integration efforts, which have since accelerated following the suspension of the IPS in March 2020 due to the coronavirus pandemic.

The integrated administrative data project will exploit the larger sample sizes offered to enable more precise migration estimates, at more granular levels (e.g., smaller geographies and groups of people.) There are also plans to produce much more detailed insights on particular topics relating to migration, such as migrant workers in the labour market. These research strands will play to the particular strengths of these administrative data and are not possible using the IPS.

Data integration will enable more timely and accurate production of migration statistics which can feed into the overall population estimates. The current project is part of a wider project to transform how population estimates are produced. The UK's existing system for creating annual population estimates is heavily reliant on the decennial census. While this provides granular data at the lowest levels of geography every 10 years, it delivers less detail throughout the interim years. Additionally, the quality of population estimates declines as one moves further away from the census year. ONS is moving towards using integrated administrative data for its population system rather than relying so heavily on the census, and ABMEs will be part of that.

2. Collaborative framework for data integration

The focus is to produce Administrative-based Migration Estimates (ABMEs) by bringing together information from two government administrative datasets: the Registration and Population Interaction Database (RAPID), which includes data from the Department for Work and Pensions (DWP) and Her Majesty's Revenue and Customs (HMRC); and Home Office border data. Other data partners include the National Health Service (NHS) and the Higher Education Statistics Authority (HESA).

3. Data sources

The project combines both marco-level and micro-level data integration. In the first and current phase, macro-data integration takes aggregates from two sources:

- RAPID - to measure migration of EU nationals
- Exit Checks to produce ABMEs - to measure migration on non-EU nationals.

First, the Registration and Population Interaction Database (RAPID), developed by the Department for Work and Pensions (DWP) provides a single coherent view of interactions across the breadth of systems in DWP, HM Revenue and Customs (HMRC) and local authorities via Housing Benefits. These interactions include benefits, employment, self-employment, pensions, and in-work benefits. RAPID contains a record of everyone who has a National Insurance number (NINo), which are used to administer social security payments and taxation. Every migrant who wishes to work in the UK is allocated a unique NINo. Within RAPID, it is possible to identify overseas nationals who have registered for a NINo. This source is used to measure the migration of EU nationals, who historically where able to travel to the UK without a visa.

Second, Exit Checks data are collected at the UK border by the Home Office. This record level data source has the potential to provide a direct measure of the movement of people in and out of the UK. Data on visas issued and people's actual travel movement across the border are used in the data integration project.

Some adjustments (detailed below) are made to RAPID data, and those involve two additional data sources: data on international students from the Higher Education Statistics Agency (HESA) and estimates of foreigners who obtained UK citizenships from the Home Office's Migrant Journey data.

In future stages of the project, the plan is to carry out micro-data integration, specifically by linking RAPID and Exit Checks at the record level. For some hard-to-measure groups, such as children and students, linking of additional data sources may be beneficials, and include: (1) School censuses, (2) data from the Higher Education Statistics Agency (HESA), and (3) Personal Demographic Service, which is a database of all National Health Service patients.

4. Integration methodology

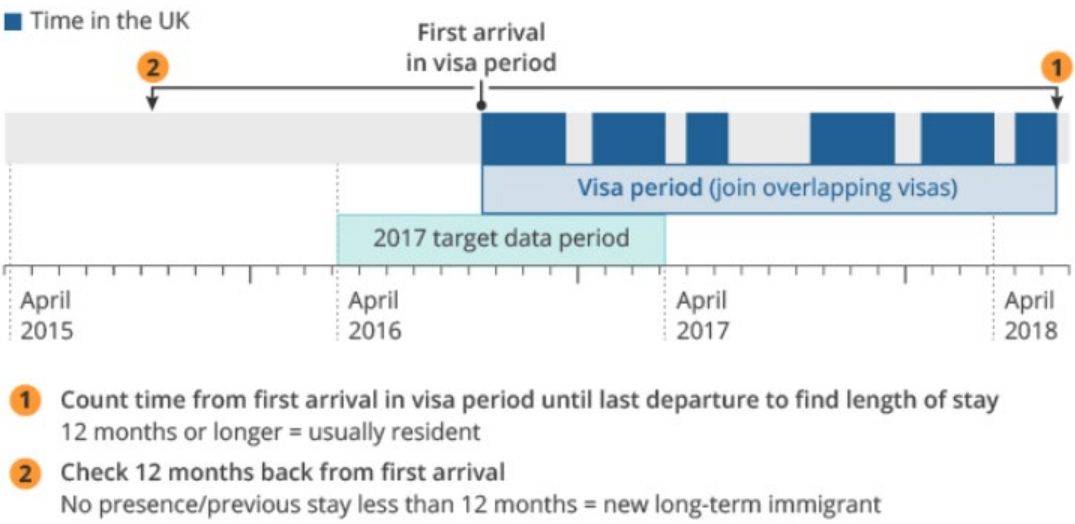
The macro-integration phase involves multiple steps. First, several adjustments are made to RAPID data to address coverage gaps. Second, a separate method is applied to estimate long-term immigration to the UK based on Exit Checks data.

Two important adjustments are made to the RAPID data source. First, foreign students who do not work alongside their studies will not be covered in the RAPID database. Similarly, students who left are also not captured in the data. Aggregated data of first-year foreign students and those graduating from the Higher Education Statistics Agency (HESA) and the Graduate Outcomes data published by the Department for Education are used to add and remove students from RAPID estimates. Second, to better distinguished UK and non-UK nationals, the Home Office's Migrant Journey data was used to generate an estimate of proportion of migrants who would gain UK citizenship within 10 years.

The ONS also developed a method for using Exit Checks to produce estimates of long-term immigration to the UK for non-EU nationals. The method uses individual level data, taking first arrival and last departure within a visa period as an approximation for length of stay in the UK. Visa periods are constructed by linking together any consecutive or concurrent visas held. If there is a gap between visas, then a new visa period is started. Visits from non-visa nationals and those on long term visit visas are excluded.

The process of estimating long-term international immigration for any given 12-month reference period is illustrated in Figure 2 and uses a three-step process. The first step is identifying those people who have a visa period with a first arrival date within the reference period. The second step is to use the time between the first arrival and last departure within a visa period to identify whether they have been resident in the country for 12 months or more (that is, whether they meet the usual residence threshold applied in the UN definition).

Figure 2: Illustrative example of identifying long-term international immigration using border data



Lastly, the third step is to look at any previous visa period to determine if this is a new long-term immigrant or one who has previously been in the country. If no presence is identified in the country during the 12 months preceding first arrival on a given visa, or the previous visa period had a length of stay of less than 12 months, then this pattern of travel will be considered as identifying a new long-term immigrant.

5. Data evaluation strategy

The Home Office has provided updated annual datasets to support the development and testing of this new methodology. Currently the data used cover the time period 8 April 2015 to 7 April 2019, which allows the production of indicative estimates of long-term international immigration for three time points: 2016 to 2017, 2017 to 2018, and provisional data for 2018 to 2019.

The Home Office also publishes annual reports on the quality of border data collected under the exit checks program, improvements to data quality, and some statistical findings from the data. Quality indicators on the attributes of data have been developed and are regularly reported on within these reports.

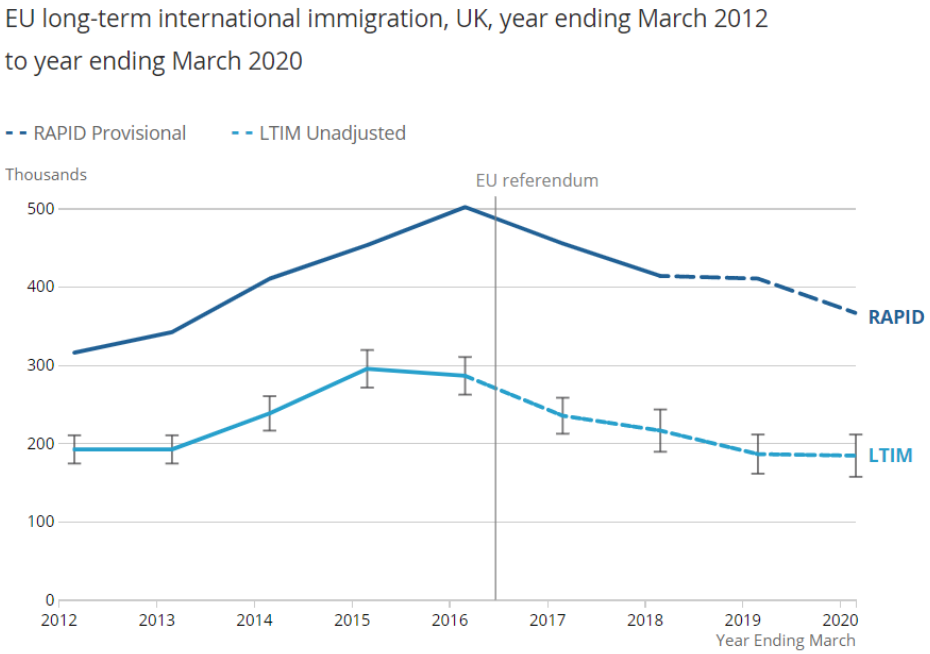
The ONS has also produced an error framework for longitudinal administrative sources using border data and we have used this approach to help identify the areas of the methodology for border data-based immigration estimates where further quality assurance or development is required.

6. Dissemination/Communication of results

In April 2021, the ONS produced an interim progress report with illustrative migration estimates for 2010 to 2019, based on macro-level integration of RAPID and Exit Checks. The findings highlight discrepancies between our ABME estimates, and our

official Long-Term International Migration (LTIM) estimates which we previously produced using the IPS survey data.

Figure 1: RAPID data show consistently higher estimates of long-term EU arrivals compared with LTIM data



Source: Office for National Statistics – LTIM; Department for Work and Pensions – RAPID, with adjustments applied

As shown in Figure 1 above, for EU nationals, both estimates from RAPID and LTIM data show a similar trend in the number of arrivals into the UK. However, for the year ending March 2012 onwards, estimates from RAPID show consistently higher numbers of long-term arrivals, with RAPID estimates being nearly double those from LTIM. The most likely driver for this difference is the uncertainty of these migrants in their intentions to move to the UK when responding to the IPS.

ONS is transparent about their move towards data integration and strive to keep users informed. Several progress reports and research articles have been published on the Office's website since 2019. Central to this is the article 'Population and migration statistics system transformation – overview', which is continually updated. The progress is also published via blogs, including Ensuring our population statistics meet the needs of everyone and Giving the best picture of international migration in times of a pandemic

ONS convenes regular meetings with senior stakeholders across the UK Government Statistical Service and manages an expert group comprised of top academics and researchers from the population and migration statistical world. These two groups provide continuous input and feedback about the Office's methodological plans.

Additional resources:

‘International migration: developing our approach for producing admin-based migration estimates’.

‘International migration: Exploring international migration concepts and definitions with Home Office administrative data.

United States

Overview

The US Census Bureau has made a concerted effort in recent years to integrate migration data to improve its estimates of net international migration (NIM). These efforts have included both macro- and micro-integration data initiatives and have been motivated by a need to improve the timeliness of migration estimates which have typically relied on large household survey data collected by the Census Bureau.

The current methodology works well when migration patterns are consistent overtime, but tend to suffer when migration patterns quickly change, due to natural disasters (e.g., hurricanes, global pandemics) or extreme policy changes (e.g., Executive Orders[11]). To provide more up-to-date and accurate estimates of migration, the Census Bureau has worked to integrate, at the aggregated level, administrative data produced by other agencies with data collected from its surveys. Further, the Census Bureau has made micro-data integration efforts by linking administrative data to improve measurement of international migration, particularly at the subnational level.

Three examples are included in this appendix:

1. Macro-data integration to measure net migration to/from Puerto Rico
2. Macro-data integration to measure migration to/from the United States during the COVID-19 Pandemic
3. Micro-data Integration to create the Integrated Database on International Migration (IDIM)

Example 1: Macro-data integration to measure net migration to/from Puerto Rico

1. Data needs

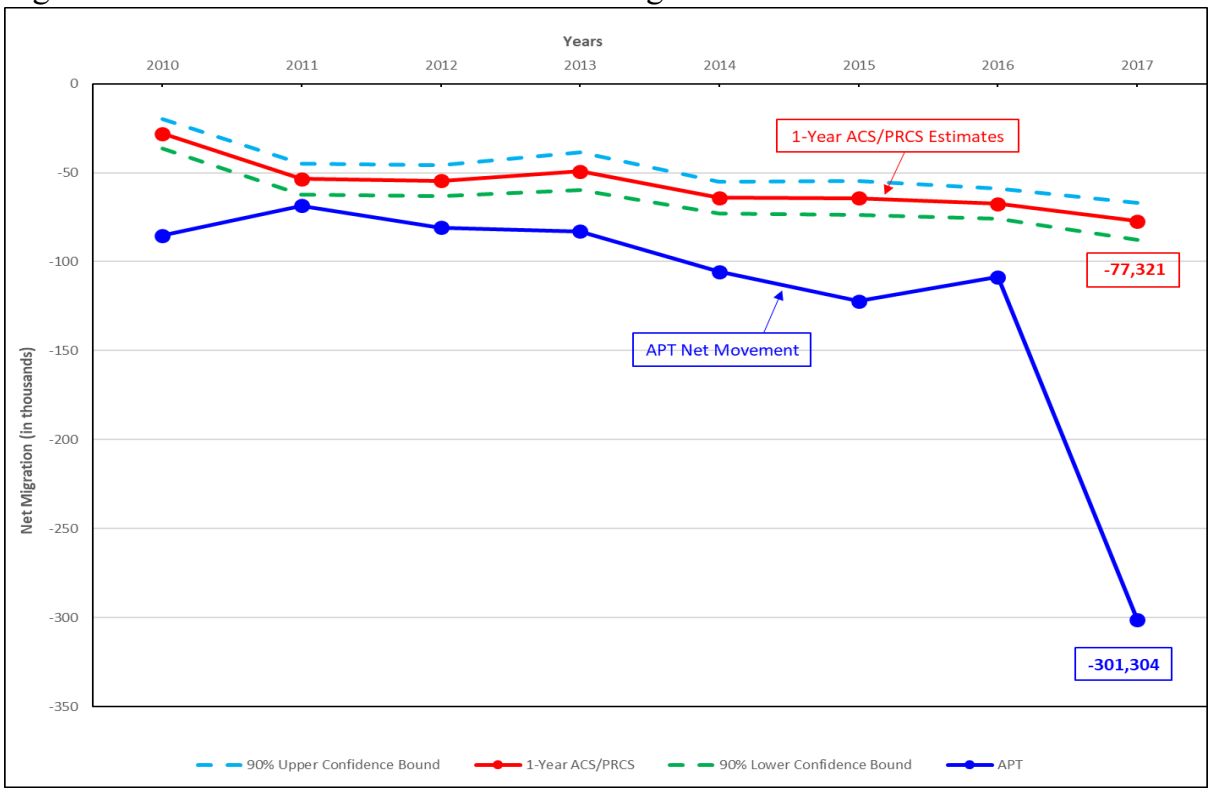
The Commonwealth of Puerto Rico is an unincorporated territory of the United States, with a population of over 3 million persons and a long history of migration to the U.S. mainland. In September 2017, Category 5 Hurricane Maria made landfall on the island of Puerto Rico, resulting in extensive damage, loss of human life, and out-migration to the mainland United States.

Household sample surveys like the ACS are not designed to pick up sudden mass movements of persons, since retrospective survey-based migration data tend to "lag" actual migration events. The former methodology used to estimate migration between Puerto Rico and the United States utilized the ACS and its counterpart, the Puerto Rico Community Survey (PRCS). As a result of the hurricane, 2017 ACS and PRCS

[11] For example, Executive Proclamation 10014, of April 22, 2020. Suspension of Entry of Immigrants Who Present a Risk to the United States Labor Market During the Economic Recovery Following the 2019 Novel Coronavirus Outbreak.

migration flow estimates yielded inadequate measures of net migration for the 2018 estimates year (July 2017-June 2018). However, monthly airline data did show a mass exodus of persons from Puerto Rico to the United States during this time period.

Figure 1. Net Puerto Rico-to-United States Migration: 2010 to 2017



Sources: U.S. Census Bureau, American Community Survey and Puerto Rican Community Survey; U.S. Bureau of Transportation Statistics Form 41, T100 (International) Segment All Carriers

2. Collaborative framework for data integration

Given the constraints of ACS/PRCS data and evidence of outmigration using airline data, staff explored further methodological improvements involving this data source. Most notably, a "flight-based methodology" which relies on publicly available T-100 Passenger Airline. Flight passenger information released by the Bureau of Transportation Statistics (BTS) is released 6 months after flights occur. This time gap was inadequate for producing our estimates, so an annual collaborative agreement was developed between BTS and the Census Bureau to enable access to preliminary data with a 3-month time lag, which allows us to us these data for our annual estimates.

3. Data sources

In our new methodology, ACS/PRCS data were integrated with monthly Airline Passenger Traffic data (APT) from the Bureau of Transportation Statistics (BTS) to improve previous estimates. Historically, APT data have consistently shown higher net out-migration from Puerto Rico to the United States than ACS/PRCS estimates. To account for this inherent difference between data sources, the revised method "blends"

ACS/PRCS and APT data, which provided an estimate that better reflected the impact of Hurricane Maria.

4. Integration methodology

The "blended" methodology used for our 2018 estimates was based on the assumption that APT data were a better reflection of the true impact of Hurricane Maria on migration patterns than the ACS/PRCS, and that ACS and APT data followed similar historical patterns, thus this relationship could be leveraged to create a better estimate for the time period. As seen in Figure 1, ACS and APT data tended to follow similar patterns prior to 2017, with APT data consistently showing more net out-movement than ACS net out-migration.

Since our previous methodology had been based on the ACS/PRCS, efforts were made to make the two data sources as comparable as possible, compiling monthly flight data for the 2017 calendar year to coincide with the ACS/PRCS estimation period. [12] We also limited flight information to domestic flights between the United States and Puerto Rico, excluding international flights.

The method applied a simple ratio, using the ratio of ACS/PRCS-to-APT net migration results over a two-year period: 2015 to 2016. The calculated ratio was applied to the APT Puerto Rico-United States net migration figure measured for calendar year 2017, to remain methodologically consistent with previous ACS/PRCS-based estimates. The calculated ratio was applied to the APT Puerto Rico-United States net migration figure measured for calendar year 2017, to remain methodologically consistent with previous ACS/PRCS-based estimates.

Since the Census Bureau's 2018 population estimates represent the population on July 1, 2018, we also considered return migration to Puerto Rico in early 2018. In order to account for January return migration, we used the 12-month APT time period from February 2017 to January 2018 prior to applying our adjustment ratio to make the time period as ACS-equivalent as possible, while still taking into consideration post-Hurricane Maria return migration to Puerto Rico. Shifting our time period one month helped account for return migration, yet also kept most months (11 of 12) within the ACS/PRCS-equivalent 2017 calendar year. This modification resulted in an APT-Puerto Rico-United States net migration figure of -215,166, which when adjusted by the APT-ACS ratio yielded a final figure of -123,399 net migration between Puerto Rico and the United States.

A macro-data integration method was again applied for both the 2019, to account for return migration post-hurricane Maria, and 2020 estimates years, to account for the impact of COVID-19.

[12] This differs from the 2018 estimates year, which covers July 1, 2017 to June 30, 2018.

A 2020 COVID-19 factor for Puerto Rico was calculated using APT data, but since this was a partial year adjustment (March-June 2020), the blended method, based on a full calendar year of APT and ACS/PRCS data, could not be used. As such, monthly flight data to/from Puerto Rico and the United States was used to make an adjustment to Puerto Rico net migration for the March-June 2020 period.

To account for the impact of COVID in 2020 we used a partial year adjustment factor. Using APT data, the net passenger total was calculated for the months of March to June 2020. Next, a ratio was calculated using the historical average net passenger movement for March to June 2010-2019, excluding 2018 (Hurricane Maria), and the March to June 2020 net passenger total. We then reduced the ACS/PRCS seasonal total by multiplying one-third of the 2019 ACS/PRCS annual total by the net passenger movement ratio. For the final estimate, two-thirds of the 2019 ACS/PRCS estimates were applied to the ACS/PRCS reduced seasonal total. The overall impact of the COVID-19 adjustment was reduced net outmigration from Puerto Rico.

5. Data evaluation strategy

Data quality and validity are observed through analysis of annual and historical trends. This is done using ACS/PRCS household surveys and APT data. Airline passenger flight patterns to and from Puerto Rico typically follow consistent seasonal patterns. Normally, March through May are net passenger outflow months, while June, the start of the tourist season, is a passenger inflow month. Annual estimates from the ACS/PRCS also remain reasonably consistent throughout the decade. Any deviation in data trends can be seen as an indicator of change. In addition to our annual data evaluation and estimates, we consult with the Puerto Rico Institute of Statistics, as well as the Bureau of Transportation Statistics, to ensure reliability in our data production. Finally, we also consulted academic literature on post-Hurricane Maria migration patterns to help validate our results.

6. Dissemination/Communication of results

Along with our regular annual estimates and methodology release, the Census Bureau presents its findings and methodological changes at professional conferences (e.g. FSCPE winter and spring meetings, PAA conference, etc.), international task forces, and international meetings on migration statistics.[13] Results are further released via publications.[14] In addition, several blogs were released, such as "Revisiting Methods to Better Reflect the Impact of Disaster" and "New Population Estimates Rely on Monthly Flight Data to Capture Puerto Rico's Migration Flows", which can be found via Census.gov under the webpage "America Counts: Stories Behind the Numbers." Census Bureau staff are also available via email or by telephone for further communication throughout the year.

[13] https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2019/mtg2/3.8_Integration_USA.pdf
[14] https://unece.org/statistics/publications/use-new-data-sources-measuring-international-migration; https://www.census.gov/library/stories/2020/08/estimating-puerto-rico-population-after-hurricane-maria.html

Example 2: Macro-data integration to measure migration to/from the United States during the COVID-19 Pandemic

1. Data needs

The COVID-19 pandemic greatly impacted movement to and from the United States starting in March 2020. As a result of several executive policy decisions, land borders were closed between the United States, Canada, and Mexico, except for essential travel, and flight restrictions were implemented for non-US citizens and permanent residents traveling from China, the European Union, and other countries. These restrictions lead to a 98% reduction in international flight traffic in April 2020. Further, US consulates abroad were closed and stopped issuing visas, which greatly reduced the number of approved visas from this time forward. Several Executive Orders reduced movement from other countries, including a ban of non-residents from certain countries and a "pause" of legal immigration, including employment-based visas. Finally, in March 2020, the US State Department encouraged the return of US citizens and legal residents living abroad, for fear of border closures.

All these factors impacted international migration to and from the United States for the later part of the 2020 estimates period, which covered the July 1, 2019 to June 30, 2020 time period. We anticipated that the 2019 ACS (which did not cover the pandemic period) would not be an adequate measure of EY 2020 international migration patterns, since we knew that international migration was very low (near zero) for the months of April-June 2020, thus adjustments would be necessary. To reflect the impact of the COVID-19 pandemic, we adjusted preliminary ACS-based NIM estimates to reflect lower migration levels for the last quarter of the 2020 estimates period.

2. Collaborative framework for data integration

Data were collected independently by migration staff at the Census Bureau. Currently, there is no formal framework for receiving migration-related data from other United States federal agencies. Therefore, staff searched for public statistics and reports on international mobility released by various federal agencies. Staff compiled data primarily on in-flows to the United States based on visa issuances, recent international students, and humanitarian migrants, as well as APT flight data from BTS. Special tabulations of I-94 data were received from the Office of Immigration Statistics. Statistics Canada shared special tabulations on aggregate numbers on persons entering Canada from the United States to help establish a base trendline of out-migration during 2020-21. These numbers included permanent and temporary migrants from Canada's IRCC and passenger arrivals recorded from Primary Inspection Kiosks at major Canadian airports.

3. Data sources

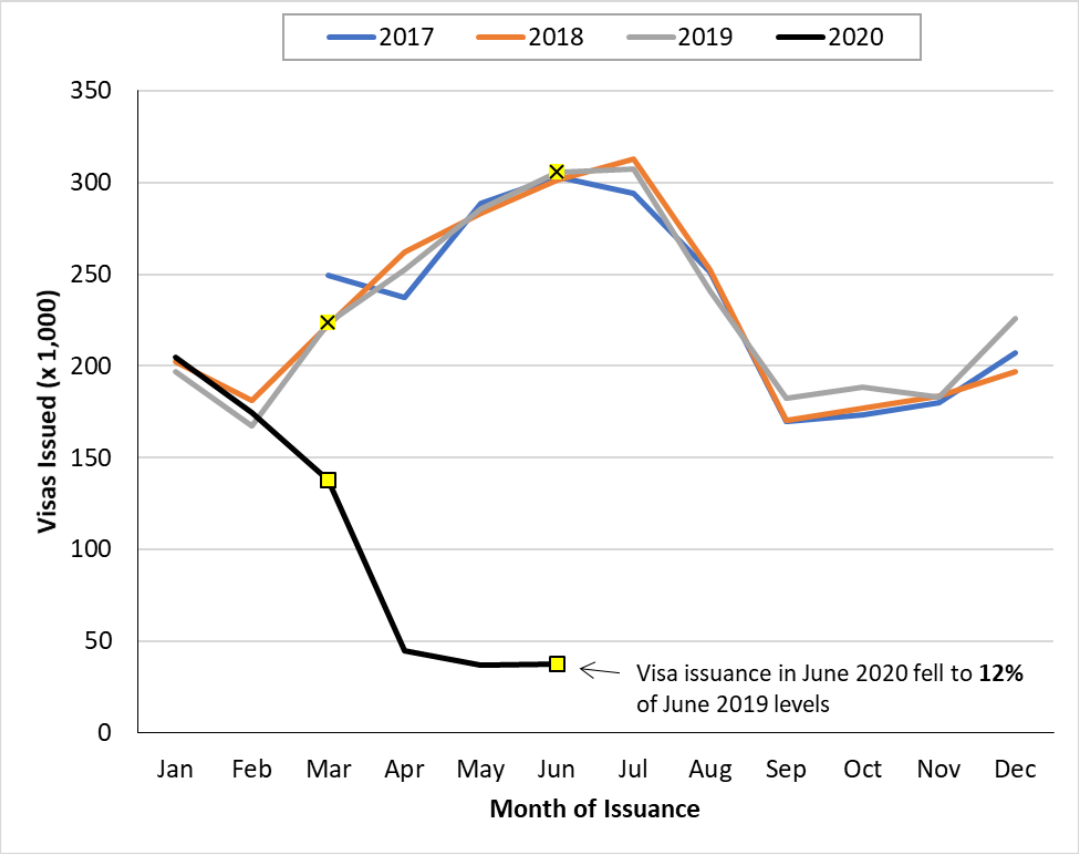
We did this by integrating 2019 ACS and 2020 monthly time series on international mobility from auxiliary data sources. Auxiliary data sources included Airline Passenger Traffic (APT) data from the Bureau of Transportation Statistics (BTS), visas issued overseas from the Bureau of Consular Affairs, I-94 arrival data from the Department of Homeland Security (DHS), and internal-use flight entry and visa data from Statistics Canada and border crossing data with Mexico.

4. Integration methodology

For each NIM component, we estimated alternative migration levels for the last quarter of 2020 based on different assumptions from the auxiliary data. For all components, we assume migration that occurred before the pandemic followed similar migration levels from the 2019 ACS input data, thus individual COVID adjustment factors were only applied to the March-June period. The final national NIM total, which was about 25% lower than our original estimate, was an average of the alternative estimates.

A specific example of the method to adjust the foreign-born immigration NIM component is as follows. First, we established there was a historical relationship in trends, if not actual levels, between visas issued overseas and our NIM foreign-born immigration estimates. Next, based on monthly administrative data, visas issued abroad and non-immigrant flight arrivals, we knew that in-migration levels were low for the April-June 2020 time period (see Figure 2). Using this information, we applied an adjustment factor (10%) to one-quarter of the 2019 ACS foreign-born immigration total, to calculate an adjusted national foreign-born immigration total.

Figure 2. Visas Issued Overseas by month: 2017-2020

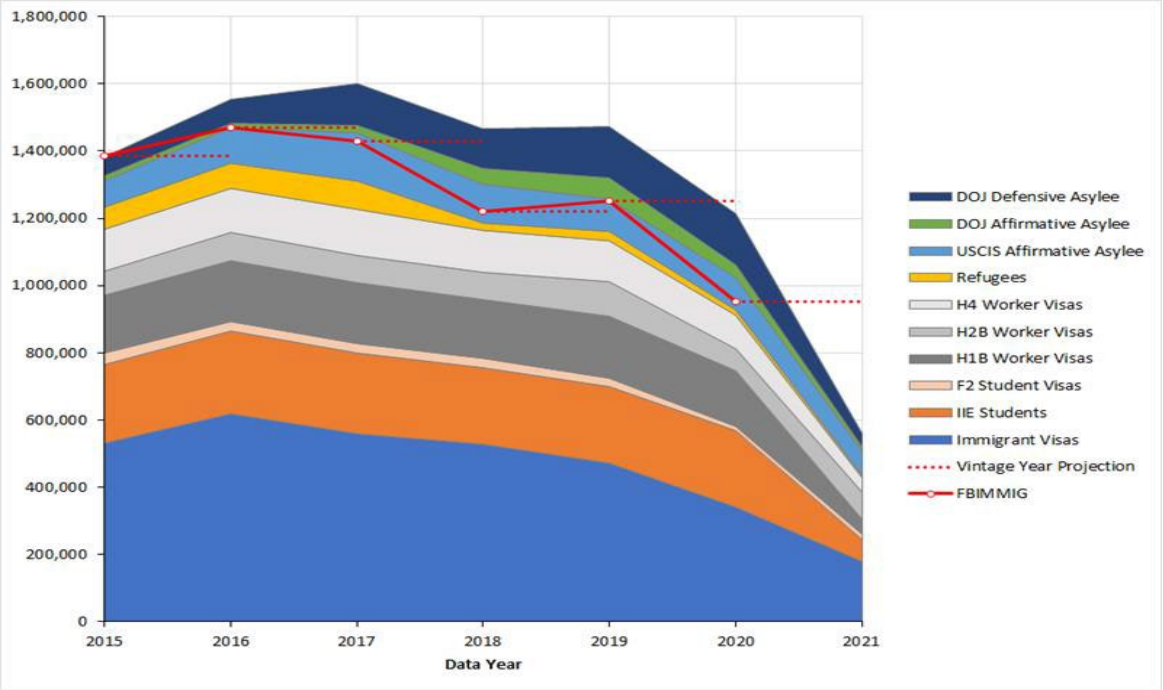


Source: US Department of State, Bureau of Consular Affairs Visa Statistics

For estimates in 2021, we decided to not use the 2020 ACS because the pandemic had a large impact on data collection. The ACS sample was greatly reduced due to limitations immediately after the COVID epidemic (1/3 of the sample lost), which disproportionately impacted certain groups, leading to high non-response bias. Due to this bias, the Census Bureau declared the 2020 ACS an experimental data set, and discouraged direct comparisons be made to earlier years.

Adjustments were made based on 2019 ACS data and trends seen in administrative data between 2019 and 2021. Administrative sources used included data from the US Department of Justice, the Institute of International Education, United States Citizenship and Immigration Services, and the US State Department Bureau of Consular Affairs and Refugee Processing Center.

Figure 3. Foreign-Born Immigration Measured by Administrative Sources: 2015-2021



Source: Department of Justice, Institute of International Education, United States Citizenship and Immigration Services, US Census Bureau V2020 estimates, US State Department Bureau of Consular Affairs and Refugee Processing Center.

5. Data evaluation strategy

ACS migration data are the foundation of NIM estimates. We evaluate administrative data based on how closely they track with the annual ACS going back to 2010. Administrative data are not directly comparable to survey-based estimates due to definitional and conceptual differences. Therefore, our analyses focus primarily on similarities in annual trends (instead of annual levels) between survey and administrative data. Due to substantial changes in the 2020-21 ACS survey operations, we are unable to evaluate how well our COVID adjustments performed compared to the observed survey data. Instead, we assume the close relationship between the ACS and administrative data before 2020 would have continued during the pandemic.

We consult with our colleagues at the Department of Homeland Security, Office of Immigration Statistics. They provide feedback on the quality of administrative sources, reasonableness of the results, and recommendations for other data sources.

6. Dissemination/Communication of results

The Census Bureau presented methodologies and results at professional conferences and regular stakeholder meetings with state-level governments throughout the pandemic. A blog entry about the COVID adjustment was published on the Census.gov

America Counts website in December 2021.[15] An updated entry on the post-COVID recovery in migration is planned for December 2022. The Census Bureau releases a methodology statement as part of the official estimates on the Census.gov website every December.

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

Example 3: Micro-Data Integration to create the Integrated Database on International Migration (IDIM)

1. Data needs

The Census Bureau currently produces annual estimates of international migration using the ACS as a primary data source. Sample surveys come with limitations, such as annual fluctuations due to sampling error, particularly at the subnational level, and lagged measurement of migration events. To mitigate these issues, we are researching the use of administrative records to supplement the ACS to create and distribute international migration estimates and characteristics.

2. Collaborative framework for data integration

The U.S. Census Bureau has longstanding data sharing agreements with the Social Security Administration (SSA) and the Internal Revenue Service (IRS) to access micro-data for estimates production and research. SSA data include the Numident file, which is a record of applications for social security cards. IRS tax records are currently used to calculate domestic migration rates, and there are agreements to utilize these data for international migration research.

3. Data sources

At present, IDIM combines SSA and Internal Revenue Service (IRS) data, but additional data sets will be added in the future. While there are potentially many data sources which could be integrated into IDIM, we have not been able to access micro-level data from DHS, thus are focusing on administrative data currently available at the Census Bureau, namely the Social Security Numident and IRS tax filing information. Variables like age, sex, and foreign-born status can be gleamed from the Numident, while geographic detail is available from IRS files. These, and other, data sources can be linked via Personal Identification Keys (PIKs), which are anonymized unique identifiers for individuals. PIKs are most easily created using directly matched encrypted Social Security Numbers (SSNs), but are also created by probabilistically matching name, sex, age, and address information.

4. Integration methodology

The Numident is our primary administrative data source (aka “spine”) for estimating national-level foreign-born immigration, and it is used in the first phase of processing. The Numident is a micro-record dataset that combines SSN records with Census Bureau death records. It includes data on demographic characteristics, country of birth, and citizenship status. It does not include geographic locator information (e.g., addresses).

First, the foreign born are identified using citizenship variables from the Numident (this includes non-citizens authorized to work and naturalized citizens). We then use record creation year as a proxy for year of entry into the US. Lastly, we remove individuals

who died the same year they migrated. This step results in an estimate of foreign-born immigrants by year with demographic characteristics, albeit a clear overestimation. We expect an overestimation at this point, as this universe includes SSN applicants who never actually migrated to the United States.

In the second phase of processing, we use a "signs of life" method, matching Numident records to IRS tax form 1040 filings to confirm entry into the United States. The Numident contains all applications for SSNs, including individuals who never actually migrated to the United States. To remove this group from our estimates, we match IRS data to restrict the universe to authorized migrants who worked and filed taxes in the United States, as well as both working and non-working naturalized citizens. This step also assigns MAFID geocodes, giving us a baseline estimate of foreign-born immigration with demographic characteristics at the county level. We expect an underestimation at this point, as we are missing migrants who fail to file taxes, as well as authorized migrants who did not work.

In the third phase of processing, we use exemptions on the IRS 1040 tax filings to account for non-working spouses and dependents. Since only non-citizens authorized to work are eligible for SSNs, a significant number of dependent migrants in the IRS tax filings do not match with the Numident. By using the number of claimed exemptions, we can determine how many dependents recorded in a tax return should be included in the migration universe. Current tax law prohibits claiming non-resident dependents, with the noted exception of dependents living in Mexico or Canada (https://www.irs.gov/publications/p501).

In the planned fourth phase of processing, we will estimate the migrant student population. Most foreign students and exchange visitors do not qualify to receive an SSN, and they are not entered into either the Numident or the IRS tax filing data. Because of this, we need to obtain data on this population from other sources. We are currently looking at using extracts from the Student Exchange and Visitor Programs (SEVIS) and the ACS, or other datasets as they become available.

The end result should include the following immigrant populations: Naturalized citizens, non-citizens authorized to work and who filed taxes (as well as their non-working resident dependents), and foreign-born students and exchange visitors. Missing populations would be: Citizens born abroad of American parents, unauthorized migrants, working migrants who did not file a tax return, and non-dependent non-working household members. The Estimates would be at the national, state, and county levels and would have demographic detail (age, sex, race, and Hispanic origin).

5. Data evaluation strategy

Evaluation of the IDIM is an on-going process. IDIM data are primarily evaluated through comparison to survey-based estimates produced by the ACS, as well as other administrative data sources. In addition to macro-comparisons, we have linked the ACS to IDIM to evaluate those matching both files, as well as those included on the ACS

who are not on the IDIM. This evaluation shows that the current version of IDIM is missing specific populations as detailed below. Results have been presented at several professional meetings.[16]

The IDIM is currently missing a few key populations: students and exchange visitors, and the unauthorized migrant population. However, overall levels of foreign-born immigration derived from the IDIM match trends seen in both our Vintage 2020 estimates and other administrative sources. IDIM's age/sex distribution show less variation than ACS inputs, and are generally on parity with vintage estimates. State-level geographic distributions are plausible, and in some cases appear to be outperforming our vintage estimates. The most current year of IDIM estimates is much more recent than the ACS or vintage estimates, with virtually no delay at the expected time of official annual production.

Future work will look at ways to measure foreign student and exchange visitors, develop methods to improve measurement of race/Hispanic origin characteristics, improve coverage of children, and investigate methods to estimate foreign-born emigration leveraging the longitudinal nature of IRS tax filings. Future data sets will be integrated as they become available. While still under development, IDIM, in conjunction with the ACS, has the potential to improve international migration estimates produced by the US Census Bureau, and serves as a good example of micro-data integration.

6. Dissemination/Communication of results

IDIM research is on-going and has yet to be officially incorporated into our official NIM estimates. Research is presented at professional conferences and regular stakeholder meetings with state-level governments.

[16] See for example, https://unece.org/sites/default/files/2022-10/A3_Presentation_Miller_ENG.pdf