Distr.: General
20 October 2017

English

United Nations Economic Commission for Europe

Conference of European Statisticians

Work Session on Migration Statistics
Geneva, Switzerland
30-31 October 2017
Item 3 of the provisional agenda
Data integration and administrative data

Draft report on 'Data Integration for Measuring Migration'

Note by the UNECE Task Force on Data Integration for Measuring Migration

Summary

This document presents the draft report on 'Data Integration for Measuring Migration', prepared by the UNECE Task Force on 'Data Integration for Measuring Migration'.

This document includes a discussion of the definition of data integration, with reference in particular to migration statistics. A number of case studies dedicated to different countries are presented, describing the primary migration data sources, the reasons why data integration is necessary, the methods used and the benefits of integration.

The document also includes a discussion and a tentative list of relevant metadata, and a number of general recommendations for the countries.

I. Introduction

1. Data integration has been a topic of great interest in recent years, especially as data needs have grown under constrained fiscal climates, combined with increased respondent burden and privacy concerns. Eurostat has supported research on methods to improve data integration, such as the ESSnet project in the area of Integration of Survey and Administrative Data1. This project was an initial attempt to create a common methodological basis for integrating different data sources.

2. However, many countries have an interest in data integration specifically for the purposes of migration measurement. Given the challenges in collecting migration statistics, it is often useful to collect data from several different sources. Data integration can reduce coverage or accuracy problems for overall migration stock and flow are concerned and also enhance the richness of migration data by adding socio-demographic or economic dimensions to existing data.

3. Methods to integrate different administrative data sources within a country (e.g. population registers, health, work, taxation, or education records) to supplement information missing from various sources (e.g. to measure outmigration for those who fail to deregister) have proved successful in some countries. In other cases non-administrative data sources can provide information missing from administrative sources (e.g. use of household surveys to collect information not included in registers). Other uses of data integration could be in reconciling different migration figures as derived from different sources, particularly in estimates for "hard-to-count" migration populations, such as irregular migrants or emigrants.

4. Noting the importance and timeliness of this topic, in 2014, the UNECE-Eurostat Work Session on Migration Statistics discussed the utilization of different administrative data sources to measure international migration. The discussion focused on the opportunities and challenges faced when using administrative data, particularly given the differing levels of development of administrative data systems across the region. Ways to improve cooperation between national migration services, statistical agencies, agencies in charge of registers, and other producers of administrative data were considered as were methods of integrating various administrative data sources to improve measurement of migration. Participants also emphasised the need for practical advice and guidance on the production of metadata to facilitate comparisons between migration estimates produced by different countries.

5. The Work Session concluded by recommending further methodological work on the topic of integration of multiple data sources for measuring international migration, including data sources within a country and between different countries, and good practices in communication between national statistical offices and producers of administrative data.

[1] https://ec.europa.eu/eurostat/cros/content/data-integration-finished_en

6. Based on these discussions, the Conference of European Statisticians (CES)2 Bureau established the Task Force on Data Integration for Measuring Migration in October 2015, to prepare a set of guidelines and description of good practices for integrating different data sources to improve the measurement of immigration, emigration and net migration. This publication presents the results of the Task Force.

7. This publication defines data integration in measuring migration as *any statistical activity aiming to enlarge the content and/or to improve the quality of statistics in migration by linking and/or comparing two or more datasets, either on national or on international level.*

8. The publication provides a general overview of this subject as well as an overview of the types of data integration that are already in use in various countries – whether disseminated through regularly published data or as part of pilot studies. Principles of best practices based on this overview are also provided to serve as guidelines for improving data integration for measuring migration in different countries. Country experiences are documented in this publication on the basis of a survey of migration data providers in nearly 50 countries as well as on more detailed case studies for several countries.

9. Data integration is addressed in this publication by examining both *macro-data integration* – the comparison/statistical modelling based on data which are aggregates (statistics) of individual level records – as well as *micro-data integration* – the integration of data based on linkage/matching of individual level records. Differing levels of overlaps of variables and/or individuals between different sources are also analysed.

10. Further details on the above concepts are provided as part of the conceptual background and definition of data integration given in Chapter 2. Chapter 3 provides an analysis of the results of the survey of migration data providers. Practical examples of data integration in migration are shown in Chapter 4 through in-depth case studies from Italy, New Zealand, Spain, Switzerland and the United Kingdom, and shorter summaries from several other countries. Conclusions and recommendations based on the survey and these case studies are provided in Chapter 5 with areas of future work on this topic addressed in Chapter 6.

[2] The Conference of European Statisticians is composed of national statistical organizations in the UNECE region (for UNECE member countries, see www.unece.org/oes/nutshell/member_states_representatives.html) and includes in addition Australia, Brazil, Chile, China, Colombia, Japan, Mexico, Mongolia, New Zealand and Republic of Korea. The major international organizations active in statistics in the UNECE region also participate in the work, such as the statistical office of the European Commission (Eurostat), the Organization for Economic Cooperation and Development (OECD), the Interstate Statistical Committee of the Commonwealth of the Independent States (CIS-STAT), the International Labour Organization (ILO), the International Monetary Fund (IMF), the World Trade Organization (WTO), and the World Bank.

II. Definition of 'data integration'

A. General features

12. The measurement of migration has since long proved to be a very challenging task. Although in recent time considerable efforts have been made in both the academic and official statistics domains to improve the quality of migration statistics, there are still relevant margins of uncertainty as for their accuracy. More recently, in a wider statistical context, the attention has been drawn on the opportunities offered by 'data integration' as potential additional action for the improvement of statistics (e.g., Eurostat, 2009, 2013; ESSnet, 2008, 2011, 2014; UNECE-HLG MOS 2017).

13. 'Data integration' is often mentioned as one of most effective approaches for enriching and improving the statistical information. Yet, to the best of our knowledge, there is not an internationally agreed definition of data integration in the statistical community. There are various features which may help to characterize data integration, such as number and typology of data sources, methodology, timeliness, response burden, and not least its purpose.

14. Intuitively, integrating requires the availability of at least two inputs. These inputs are the datasets, i.e. organized information derived from selected data sources by means of a statistical operation, which in the case of migration can be classified in four categories: (datasets derived from) statistical surveys (including exhaustive surveys, i.e. censuses), administrative registers, big data and geographical information. These typologies can be integrated in all possible combinations and the process can be repeated, generating datasets of mixed nature which can be in turn integrated with other datasets.

15. The way datasets are integrated mainly depends upon whether the information they contain is on micro or macro level and, for individual records, upon the availability of a common identifier. For the integration at micro level, this latter feature makes the difference in the choice between statistical matching and record linking, the latter usually applied when an identifier is available in all datasets being integrated. In the integration at macro level, i.e. between data aggregated from single records, the process should take place in two steps: first, cleaning from differences in concepts and operational definitions; second, reconciling (‘balancing’) the data, using various statistical techniques or mere expert opinion.

16. In the case of migration data, the most common individual identifier is the Personal Identification Number (PIN). In fact, micro-level integration is preferred for the setup of population registers, of which migration statistics are often a by-product. It should be noted that a PIN is not the only possible common identifier, as different identification coding can be used to link data related to sub-national aggregations, administrative entities, file records, and so on.

17. The data sources used as input to a population register can be many and in some countries additional data sources have been integrated over time, enriching the statistical information on the population and, indirectly, on migration. This can be the case also for registers of the foreign population / aliens registers, although usually to a slighter extent. Data integration can thus be seen as an expanding process, where

to the first integrated datasets, others are added following lessons learnt, refinement of integration methodologies, and access to new data sources.

18. Another distinctive feature of migration measurement is the fact that, given the nature of its phenomenon shared by two countries (either receiving and sending countries, or hosting country and country of birth, or whatever the variable for identifying migrants), data sources other than national can be used. The exploitation of this intrinsic international feature is mostly limited to the exchange of aggregated data, given the high concerns about privacy and national security that countries may oppose any request of international migration microdata exchange. Examples of this latter approach do exist in the Northern European countries, where it is used to improve the quality of the population registers. However, this is about the owner (country) of the data source rather than about specific features of data integration.

19. Adding new data sources may lead to changes in the timeliness of the statistical output, i.e. changes in the delay between the reference period and the availability of results. Whilst the general rule would be that the timeliness of a specific output from integrated data is given by the data source with the larger delay, statistical modelling may actually reverse this situation by giving the opportunity to release more timely estimates. It is likely that a worsening of the timeliness would only be accepted when the data integration brings a relevant improvement of the coverage or accuracy of the statistical measurement.

20. The reduction of response burden is sometimes mentioned as one of the opportunities offered by data integration. Usually, however, such reduction is achieved by replacing a data source by another data source with a lower burden on the respondents/data providers or by improving the statistical operation applied on the data source. From a conceptual perspective, this may be seen rather as an enhancement of the efficacy than an enrichment of the data availability, and therefore not exactly peculiar/pertinent to data integration.

21. Inputs from alternatives data sources (including from another country) may be used also for the sake of validating the statistical output of a single, official data source. In this case, it is matter of *comparison* rather than *integration*. However, this may be the first step in a process of progressive integration of these data sources, especially when the additional data sources do not comply with the requirements of official statistics.

22. Another peculiar case is when each data source covers a specific population (migrant) sub-group. Here also it would be more appropriated to label the pooling of these data to produce an overall statistical output as 'compilation' rather than 'integration', as there is no actual merging at the level of the single record (either individuals or aggregated entities).

23. From the considerations above, it derives that the qualifying feature of data integration is certainly the use of multiple datasets, but conditionally to the way these are used jointly. In fact, there are statistical operations which are somehow border cases with data integration, such as data compilation and data comparison. As for the other features, the level of detail of the data matters for the methodology to be applied, rather than for the identification of a data integration activity. Timeliness cannot be used either as identification criterion, given that data integration may in principle lead to an improvement as well as a worsening of the delay of the results,

and that timeliness improvements can be gained as well by applying statistical operations other than integration. Likewise, any generic reference to quality improvement of the statistical output (including reduction of response burden) would not help, given that any statistical operation should in principle aim to such an effect. Obviously, this does not imply that data integration does not have such positive effects, but only that those references are not useful to identify distinctive features of data integration.

24. The outcome of a data integration activity should be an 'enriched' or 'higher quality' dataset. Thinking in terms of a generic data matrix $n \times p$, with *n* records and *p* variables, such enrichment could take place in both directions, improving the coverage (i.e., increasing *n*) and/or the information on the same records (i.e., adding new variables to the *p* set). As for the higher quality, this would not be reflected in changes of the size of the resulting dataset, but rather in its content (e.g., in the weights of a sample survey).

B. Working definition

25. An available definition of data integration is from the IT environment (SDMX, 2009) and it is the following: "*The process of combining data from two or more sources to produce statistical* outputs." In the same reference, it is also clarified that "*Data integration can be at the micro-level, where it is often referred to as matching, or at the macro-level*".

26. The definition above is based on inputs ("*two or more sources*") and purpose ("*to produce statistical outputs*"). The latter is rather generic, because a statistical output can be seen as the end of any statistical process and it does not necessarily require data integration.

27. For the purposes of this report, the following working definition is hereby proposed: "*Data integration is a statistical activity on two or more datasets resulting in a single enlarged and/or higher quality dataset*."

28. Data integration can be processed at two main levels of aggregation: micro- and macro-data integration, defined as follows:

i. 'Micro-data integration': the integration of data based on record linkage/statistical matching of individual level records using key identifying variables.

ii. 'Macro-data integration': the combination of data based on aggregates (statistics) of individual level records.

29. The reference to a generic 'statistical activity' leaves it open the use of any methodology, either on macro or micro level, including expert opinion. It covers as well the work on conceptual differences, a paramount step in any statistical process. The explicit reference to statistical matching and/or record linking, although peculiar to data integration, would have excluded the macro-level integration.

30. "*Two or more datasets*" highlights the multidimensional nature of data integration. It also indicates that, whilst integration ends in a single outcome, it is an activity that needs to be repeated each time such outcome is targeted. In other words, for the regular production of statistics from more than one dataset, integration is not

an occasional, once for all, statistical activity. The 'integrated' result is generated each time a fresher data input is available[3].

31. The use of the word 'datasets' instead of 'data sources' points to the fact that subject of integration are actual data, and not the data sources from which they are derived. For instance, merging two existing sample surveys in a single encompassing survey by modifying the questionnaires, the sampling design, etc., is in this context not considered "data" integration. It also supports the view expressed just above that the integration does not transform the input data sources in new, mixed data sources: it is the outputs of those data sources that are the core of the activity. The integration operation is thus likely to be as systematic as the production of statistics from the data on which it is applied.

32. The outcome of data integration is first and foremost a "*single*" dataset: a new set of data where all the information from the input datasets is re-organized in a harmonized fashion. This is not simply the union of multiple datasets, as redundancies, conceptual differences, and any other factor of bias are supposedly properly treated. This is an "*enlarged*" dataset, meaning a structured new set of data whose dimensions cannot be smaller than the largest corresponding dimensions of the individual input datasets, and/or an "*higher quality*" dataset, whose elements have been changed due to data integration with a (possibly measurable) gain in the statistical quality.

C. Survey on data integration practices

33. In order to collect information needed to carry out its work, the task force carried out a survey on data integration practices. A specific questionnaire was designed to collect information on current practices in the use of combined or integrated data sources for the measurement of immigration and emigration flows, and for outputting statistics of the migrant population and the foreign-born population.

34. The questionnaire was forwarded in September 2016 to National Statistical Offices of UNECE member countries[4]. Fifty-six countries provided responses to parts of the survey that had relevance to their data collection practice.

35. The results of the survey were analysed between the end of 2016 and the first quester of 2017. The main results are summarized in the folllowing paragraphs.

D. Summary of results

36. Nearly all respondent countries produce statistics on international migration flows and more than half of these countries would use a combination of data sources.

[3] Identifying conceptual differences in a context of data integration is possibly an exception to the 'repetition rule', because such activity would need to be carried out only once, unless the conceptual framework of the input data changes over time.

[4] In addition to the 56 countries that are members of UNECE, the questionnaire was also addressed to other countries that participate regularly in the activities of the Conference of European Statisticians, including Australia, Chile Colombia, Japan, Mexico and New Zealand.

Some countries would use a central register of all registrations and de-registrations of residents, and integration of the source at the individual record level with administrative collections of emigration and immigration flows, and deaths, supported a refined production of migration flows. Other countries would integrate population registers with survey information or other administrative collection of non-nationals at a macro-level.

37. Overall, about a third of the respondent countries (total of 35 responses) integrated the data sources at the macro-level using statistical modelling or other method to combine aggregated data, and a slightly less proportion of countries would apply integration methods at the unit record level. However, it should be noted that a few countries used other sources for cross-checking and complimentary information only. About half of respondent countries noted observation units from the integrated data sources were partially overlapping; a smaller proportion had identical observation units when combining sources; and a few reported the use of mutually exclusive observation units.

38. Less than half the countries prepare statistics on international migration flows from a single source. Statistics on migration flows from a single source can frequently be produced when it is possible to maintain an administrative collection of all border crossings or when migration flows can be derived from a population register. Another less frequent option, is the use of the population census to indirectly estimate migration flows from population stock measures.

39. Around 90 percent of respondent countries produced statistics or collected data on the foreign born population. Just over a half of these countries would use more than one data source and usually by data integration at the unit record or macro level. The main types and features of the sources used, either single or integrated, were a 5-yearly population census, population register, survey, or administrative collection. The main purpose of integration was reported to be for production of statistics on the foreign born population.

40. Around 60 percent of respondent countries reported they did not produce other statistics on migrants or the foreign born population. So a few countries reported that from integration of data sources allowed the preparation of statistics such as asylum seekers and grants, residence permits, work permit holders, irregular foreign workers, socio-economic characteristics of migrants, and reason for settlement.

III. Case studies

A. Italy[5]

1. Primary source for the measurement of migration statistics: the population registers ("Anagrafi").

41. Since the early nineties, the Italian National Institute of Statistics (Istat) has been trying to make the best use of the available administrative sources to produce international migration statistics, overall the population registers ("Anagrafi"). This has allowed for a large statistical documentation to adequately catch the main socio-demographic features of the migration phenomenon.

42. During the last years, many changes have occurred under administrative point of view as well as under the statistical data management and collection procedures. The population registers digitalization and law changes have improved the administrative system productivity and decreased the citizens' administrative burden. Such changes have strongly affected the statistical production and allowed to improve the data quality in terms of timeliness, coverage and consistency with other administrative sources.

43. In Italy official statistics on international migration flows are based on local population registers under the responsibility of the Ministry of Interior which is in charge of supervising local administrations and national security. Changes of residence data source provide the number of international migration flows, the information on both the origin and destination of a single movement as well as the main socio-demographic characteristics of migrants. Istat yearly collects individual data from more than 8'000 municipalities. The data collected are not usable as they stand because data quality problems are present due to errors during the data entry, missing information or other invalid data. The validated data are cross checked against the data by aggregated form (demographic balance) and the inconsistencies are removed in order to produce a unique, more accurate and consistent statistics as required by the Regulation EC No 862/2007 of the European Parliament and of the Council.

2. Limitations of the original source. Why is integration necessary?

44. Even though the quality has considerably improved over the last decade, accuracy of this register still represents a critical issue mainly for the difficulty in recording international emigration but also due to the discrepancies between the administrative concepts and the international definitions. These lead to three main limitations of the original source:

(a) Over-coverage: In Italy, nationals living abroad and foreigners who have left the country permanently or long-term should be removed

[5] Prepared by the Directorate for Social Statistics and Population Census, Italian National Institute of Statistics (Istat)

from the population register. However, emigrants don't see real motivations to notify the authorities of their departure and likely they remain years in the registers before being removed as deregistration ex-officio.

(b) Under-coverage: statistics are referred to a "de jure" legally resident population, and not to the "de facto" one: the foreign population usually resident in Italy but without official residence are excluded from the statistics. Although coverage of the usually resident foreign population in Italy is monitored, there are foreign citizens who do not want to be recorded in the population register, especially those nationals from the European Union member States. Moreover, those who have lost the requirements for staying in Italy and immigrants who entered without authorization cannot be recorded in the population register. The latter, who do not appear in any administrative sources, represent a hard-to-count sub-population and they require a suitable survey to be counted.

(c) Time criteria: no time criterion is applied.

3. Methods used for data integration

45. A solution to face the coverage of population registers could be the use of an integrated system of registers: official population registers would be linked together with other subject-specific administrative sources (related to labor, education, taxation, earnings, etc.). This system could be used to identify groups that correspond to the national or international definition of "usually resident population". Under-coverage in population registers could be detected by using individuals' signals of presence on Italian territory coming from other registers; whereas the absence of signals for people in the population register could be evidence of over-coverage.

46. ISTAT arranged a Working Group with the aim of testing the use of administrative information for Census purposes and for the production of population counts. The results of a trial carried out by ISTAT's researchers will be briefly presented in this paragraph.

47. Data integration is necessary to define a preliminary workflow on how to process available administrative data for deriving usually-resident population counts. ISTAT has built an integrated system of the administrative sources to manage the increasing number of administrative data sets acquired for statistical uses and to maximize the benefit deriving from the huge amount of information available. To this aim it was necessary to centralize some functions for acquisition, storage, integration and administrative data quality evaluation in the system called SIM (Integrated System of Microdata). This system manages social and economic administrative data for: individuals and household characteristics (demographic aspects, employment status, level of education; places, in terms of residence, labor or education). The integration step in the SIM system is the process of linkage and physical integration of microdata coming from different sources: depending on the linking variables available, a suitable integration strategy and a set of algorithms are applied.

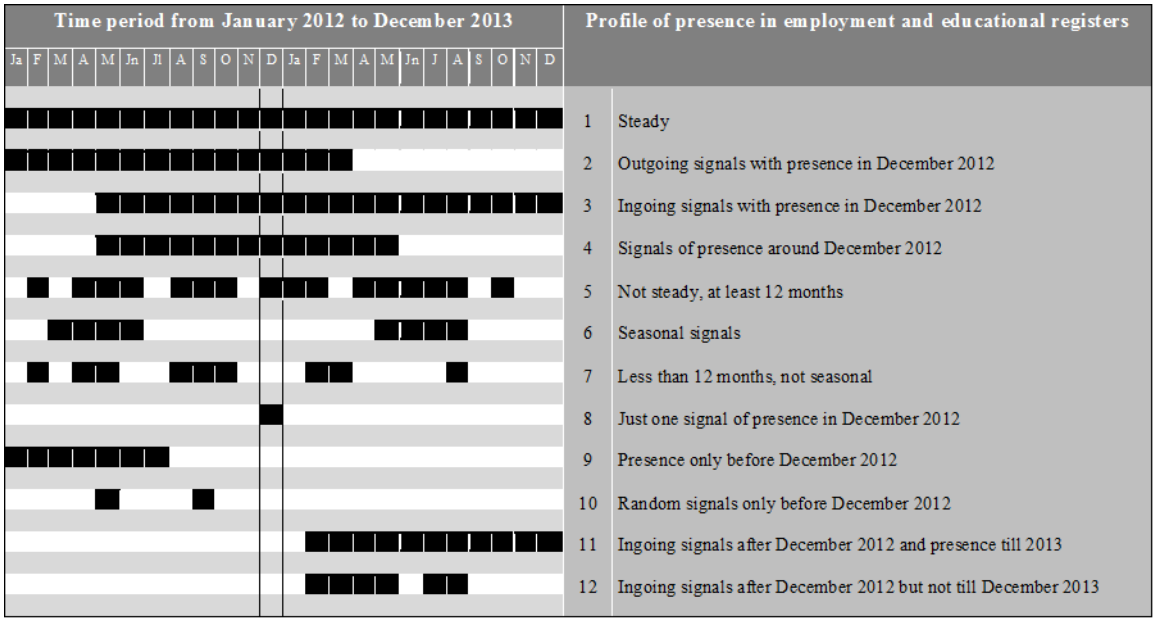
48. Integration means: (i) identifying each object (individual, economic unit) in the administrative data sources with a unique and stable (over time) ID number; (ii)

defining, for each object, the logical and physical relationships among sets of data coming from different sources.

49. SIM has been created with the aim of supporting the statistical production process: the use of the SIM IDs enables the building of data structures that could constitute the starting point for the statistical processes based on administrative data. To evaluate how administrative data could improve population counts, a thematic database of 'administrative signals' was created in order to support an initial experiment. Data used in the trial came from specific administrative sources already stored and integrated in the SIM system: Population registers, Permits to Stay, Employees and Self-employed, Compulsory Education, University Students, Retired People, Non-Pension Benefits, Income and Taxation.

50. By using a standardized time reference for the duration of an activity, it was possible to derive a monthly presence scheme. This feature consists of a sequence with a length of 24 digits, each of them representing whether a person was present for a specific activity and in which geographical area of the country. Clustering on these sequences, evident in the majority of the activity signals, showed the existence of patterns of continuity (Figure 1).

Figure 1: Monthly presence scheme of continuity's patterns in job and study activities



51. The evaluation of consistency between the registers examined determines a classification of all the cases recorded in the experimental database, into groups that are useful when producing population counts. The process to derive these groups was organized in three steps (Table 1). The first step involved linkage between the experimental Population Register (PR), the activity signals from Labor and Education registers (LE) and the Permits to Stay Register (PS); in the second step, which could explain the absence of previous activity signals, retired people or people with other benefit signals emerged from all the individuals without signals in the LE; finally, the Tax Returns Register (TR) provided indirect signals on people who could in some way justify their presence in Italy.

52. It is possible to organize the above mentioned groups according to evaluation of usual residence. So, the individuals belonging to Group A, Group E and Group G could be considered "eligible" as the usually-resident population using the usual residence concept with 12 months' time criteria. Instead, Group C, Group D and Group H require a more thorough analysis.

53. At the end of 2012, about 3 million people were recorded in Group H which represents the sub-population recorded in the population register (PR) without signals from other sources. Almost 75% of these people had Italian citizenship.

Table 1: Proceeding scheme and counts of population groups (in thousands) according to their eligibility or uncertainty to be usual resident in Italy

Step I					**Step II**			**Step III**		
PR	**LE**	**PS**	Group	Count	**RR-NPR**	Group	Count	**TR**	Group	Count
61,068	37,704	3,378			20,764			26,649		
Signals					Signals			Signals		
Yes	Yes	-	**A**	**36,618**						
Yes	No	-	B	24,450	Yes	**E**	**14,485**			
					No	F	9,965	Yes	**G**	**6,939**
								No	H	3,026
No	Yes	-	C	1,086						
No	No	Yes	D	351						

Table legend:

Eligible as usually residents

Uncertain residents

Administrative sources legend:

PR	Population Register
LE	Labor and Education Registers
PS	Permits to stay Register
RR-NPR	Retired and Non-Pension Benefits Registers
TR	Tax Returns Register

4. Benefit of integration with statistics from other sources

54. The integration of Italian administrative data sources proved to be very effective in exploiting the wealth of information already available. However, as the signals can also represent a temporary or occasional presence, it is necessary to carry out a characterization process by constructing derived variables which then make it

possible to identify cases of permanent presence that correspond to the usual residence international definition. In the trial, ISTAT defined a preliminary workflow to integrate the use of administrative sources and the official population registers in order to calculate the usually-resident population. Using this workflow, it is possible to define a group of individuals eligible to be included in the usually-resident population of Italy at a given reference date. This group of possible usual residents totaled 62.6 million individuals in 2012 (on December 31). From this group of individuals, it is possible to identify three main sub-groups:

(a) The subpopulation present in the population register, without signals from other sources (3.0 million individuals in 2012) that could be emigrated without notifying the departure.

(b) The subpopulation present in the population register that showed signals in other sources (58.1 million individuals in 2012);

(c) The subpopulation not present in the population register but with signals from other sources (1.5 million individuals in 2012) that could be immigrated without notifying the arrival.

55. The analysis of signal strength based on its continuity over time is only the starting point in the use of longitudinal data that can be processed with administrative sources. The main objective of the future trials should be, therefore, the study of longitudinal models, over several years, to produce subpopulation estimates which are more stable in relation to the fluctuations linked to the labor market, from which signals are derived.

References

Tucci E., Marsili M. and Terra Abrami V. "Improving quality of international migration outcomes by incorporating the micro-approach in managing current demographic accounting (MIDEA) and statistical population registers (ANVIS), UNECE/Eurostat Work Session and Workshop on Migration Statistics 10-12 September 2014, Chisinau (Republic Of Moldova)

Prati, Gallo, Chieppa, Tomeo et al. Feasibility study of using usual residence concept with 12 month's criteria for all the breakdowns of population, births and deaths requested in the regulation (art. 8) - Note prepared By the National Institute of Statistics of Italy for the REGULATION EU 1260/2013 on European Demographic Statistics

B. New Zealand

56. As the principal agency responsible for the processing of international movements in to and out of NZ, Statistics NZ provides timely and accurate statistics on three passenger types: overseas visitor, New Zealand-resident travelers, and permanent and long-term migrants. Having an advantageous geographic situation with border entry by sea ports and airports only, travelers' self-completed passenger cards have been the main data collection.

57. The content of the cards have evolved over time but in the longer term there are expectations to reduce further the information passengers provide on the paper forms. There has also been a strong reliance on integrated electronic files of international passenger movements as a source of validation. Statistics NZ is investigating the use of alternative sources and statistical methods for obtaining the departure information with a view to reducing or remove the reliance on the departure card.

1. Data collections

(a) Sources for measuring migration flows

58. Releases of NZ international migration statistics are based on the electronic capture of all passenger arrivals and departures supplied to Statistics NZ by the NZ Customs Service. The records include travel identities such as carrier details, travel mode and port, as well as key passport information such as date of birth, sex and country of citizenship. Records include the processing of passengers who require a visa to enter NZ. Visa types are categorized according to the policy and immigration decision making requirements of Immigration NZ. All border movements and passenger data captured by the NZ Customs Service are regularly (daily) transferred to Statistics NZ.

59. The NZ Customs Service also supplies Statistics NZ with the physical passenger cards which were completed by travelers at the time of travel. Statistical information captured by the passenger card include additional traveler information such as country of last/ next residence, travel purpose, and intended length of stay in NZ or absence from NZ. It also enables the classification of passenger categories (long-term, overseas visitor, resident).

60. Statistics NZ undertakes the scanning of the passenger cards using image recognition technology for automatic coding of text responses. This is followed by a manual completion process of records not clearly recognized and coded by the imaging software (about 1 in 10 cards). Some response fields are required from each passenger card for the purpose of linking to the electronic record supplied by NZ Customs Service. Other response fields of the passenger card are only required for sampling purposes and vary by passenger type and direction of movement. Table 2 summarizes the data content of the migration data collections combined for the production of monthly statistics of international passenger movements.

Table 2: International migration data collections and processing

Data collections		
Movement records	**Passenger records**	**Imaged passenger cards**
Actual date Schedule date Prime carrier/ ship name Co-share code Direction NZ Port Travel mode	Customs port Passport number Citizenship Date of birth First names (text) Last names (text) Sex Visa type code Visa date Country of birth Route port code Co-share code	Travel date Passport number Carrier/ ship name Direction Birth date Intended length of stay/ absence Travel purpose Country of main destination Country of last/next permanent residence Overseas state Occupation NZ locality (city/ district)
Processing for statistical production		
Loading of movements and passenger data Validation Editing/ correction Linking of imaged passenger card data Coding Linking of previous passenger classification history Derived variables: - Age - Length of stay/ absence (days) - Passenger type (overseas visitor, NZ resident traveler, permanent- and long-term migrant) Imputation Sampling and weighting Production of monthly statistics of passenger movements		
Series of linked travel histories for final definition of migrant status		
Passenger identifier Identifier linkable to other movement and passenger records (Identifier linkable to Stats NZ Integrated Data Infrastructure) Direction of movement Date-time of movement Date of birth Derived long-term migrant status by ‘12/16 month’ rule		

61. A passenger may change their intentions of stay/absence after their arrival or departure, which may result in the recorded passenger type at time of travel becoming incorrect. Statistics NZ does not revise published statistics to adjust for such changes. There is a need for an alternative method for classifying passengers after having observed their actual patterns of stay in NZ. A time series of linked travel sequences for individual travelers enables a final classification of their migrant status.

(b) Sources for measuring migration stocks

62. The five-yearly population census provides measures of the overseas born population by capturing the country of birth and the year of arrival in NZ for the usual resident population. At the 2013 Census, the overseas born resident population count was around 1 million, representing nearly one quarter of the total NZ usual resident population (Figure 2). In 1986, the overseas born population of NZ was less than 15 percent and there have been marked increases in the number and proportion of the resident overseas born population until 2006 (figures 2 and 3).

63. A post-enumeration survey (PES) is undertaken shortly after the census to evaluate the completeness of census coverage. Estimates of under-enumeration and over-enumeration as measured by the PES for population groupings defined by age, sex, ethnic groups and geographic area. The five-yearly historic PES (1996-2013) have not provided estimates of the net census undercount of the overseas born population (Statistics NZ, 2014). However, for the 2018 PES, Statistics NZ has made plans to include the country of birth question. This is acknowledging the need to improve accuracy and extend statistics of the overseas born population.

Figure 2

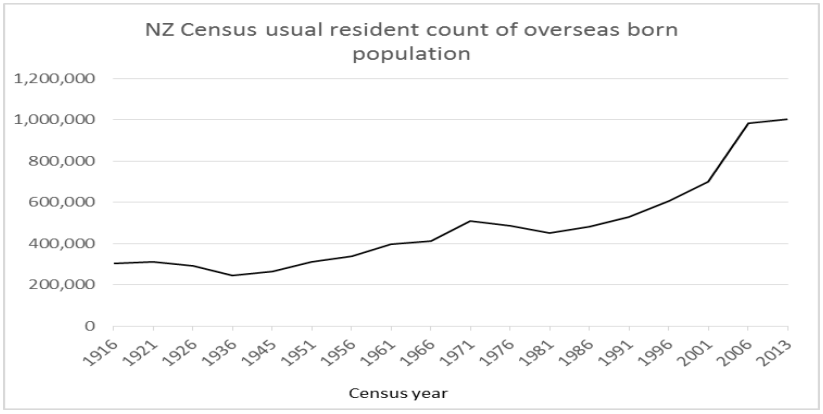
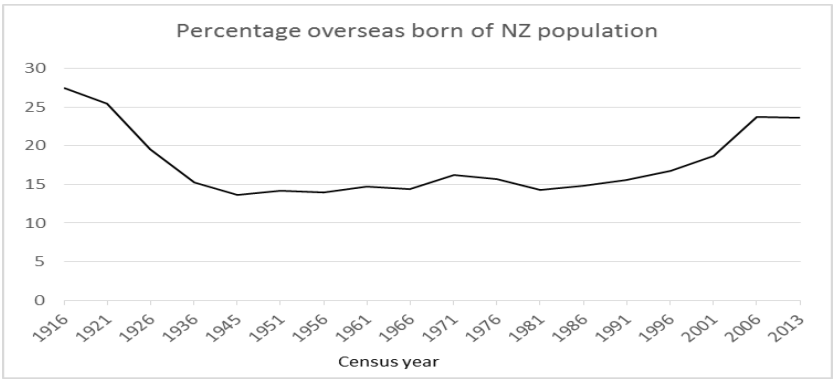


Figure 3



2. Purpose of data integration

(a) Migration flows

64. The linking of imaged passenger card records to electronic border movements and passenger records is primarily undertaken to include additional passport information as well as carrier details available in the electronic capturing of border movements. The integration of border movements with passenger records also serves as a validation of the actual passengers with border movements at a given time. For some passengers there will be no filled-in cards but electronic records will be available, or there may be filled-in cards but no electronic records. A manual linking process is included for optimizing the coverage of international passenger movements.

65. Linked travel and passenger classification sequences for the 16 months prior to the reference month are created for travelers who have indicated their intentions to stay in NZ or leave NZ long-term. The purpose is to confirm the NZ resident status at the time of travel using the information on the card about whether they have been living in NZ for 12 months or more. Given the integrated data sources as well as the classification history for the passenger the evaluation serves as a data quality assurance check on initial counts of long-term arrivals and departures.

66. A passenger may change their intention about stay/absence after arriving in NZ or leaving NZ. An additional measure of migrant status takes account of when an international passenger is included in or excluded from the resident population after observing times of stay in NZ or absence from NZ. The measure uses linked travel histories over a follow-up period of 16 months and applies operational rules for determining a change in resident status. The '12/16-month rule' is independent of the individual's legal residence status and of the information stated on the passenger card.

67. By observing individuals' travel sequences over time, analysis of actual durations of stay or absence may produce results quite different from the information initially recorded on the passenger cards. A person may have transitioned from being counted as an overseas visitor arrival to being a NZ resident departure at next point of overseas travel by going through stages of onshore temporary visa approvals to permanent residency. On-shore transitions leading to a change in resident status can result in over- or under-reporting of initial estimates of migrant arrivals and departures.

68. The use of travel histories also introduces added opportunities for extending the measures of international migration statistics. For example, there may be future interests in measures of short-term migration. Further, passengers' classification histories represent a longitudinal data source enabling statistical measures of return and circular migration. Integration with the Statistics NZ Integrated Data Infrastructure (Statistics NZ, 2016) also facilitates other analysis of migration trajectories and outcomes.

3. Data integration methodology

69. Files of border movements and passenger data are automatically linkable through the record filing system provided by NZ Customs Service. The deterministic integration of these electronic files with imaged passenger cards includes an automatic step which links on passport number, name, and date of birth, carrier / ship name, and direction of travel. At least two of passport number, name and date of birth must have full agreement for the record to be a match. A manual step is added to correct for non-linked records by looking at images of names and incorrect response to the linking variables.

70. A small number of electronically recorded passenger movements have no associated filled-in passenger cards. Imputation for missing responses on the card is carried out using past travel history and passenger characteristics. Conversely, the passenger card may be the only source of border movement capture for some passengers. They mainly represent passengers arriving or departing by ship, and electronic records are created accordingly for these passengers.

71. The integration of these records with past 16-month passenger classification sequences is undertaken by deterministic linking on passport numbers, name fields and date of birth. Records are considered matched if there is full agreement on two of the three linking variables.

72. The data integration of border movements to create the series of unique passenger IDs is a probabilistic linking process using the IBM QualityStage software. It applies the Fellegi-Sunter methodology (Statistics NZ, 2015) over a three-pass matching process; the first pass is deterministic followed by two probabilistic matching steps. The blocking and linking variables used for this process, including the chosen cut-off value, are given in Table 3.

Table 3

Pass	Blocking variables	Linking variables	Cut-off value
1	Movement record identifier	NA	0
2	Date of birth Sex	Name Nationality	21.25
3	Date of birth Sex of females Nationality First and middle names	First and middle names Name	21.26

73. The movement record identifier is derived from the passport identifiers at the point of travel whereas the data integration methodology assigns a unique passenger ID. This unique traveler identifier takes care of passengers changing passport nationality or using more than one passport (second linking pass). The last linking pass accounts for females changing their surname. For more information on the data integration methodology and the application of the ‘12/16-month’ rule refer to the reference (Statistics NZ, 2017).

4. Use of data integration with other sources

74. Historic datasets of border movements and passenger classifications collected by the processes described above are stored within the Statistics NZ Integrated Data Infrastructure (IDI) (Statistics NZ, 2016). Travel and migration data in the IDI also include datasets containing information on the immigration approval decisions and criteria, as well as the temporary visa categories (e.g. student, work) granted on entry to NZ. The IDI research database contains integrated person-level data collected from a range of government agencies, Statistics NZ surveys including the 2013 Census, and non-government organizations.

75. Integration of travel and migration data with other sources such as the tax data allows longitudinal studies of migrants' contribution to the labour market. In particular, labour market outcomes may be compared across specific groups of migrants and for different geographic settlements.

References

Statistics NZ, International Travel and Migration, Data Collection Methodology, http://datainfoplus.stats.govt.nz/Item/nz.govt.stats/f705ca38-ea6e-453f-b1d9-a95dc0fcaf59#/nz.govt.stats/d83affa2-c0e1-4629-9aff-cc8452eb41c6#

Statistics NZ, Defining migrants using travel histories and the '12/16-month' rule http://www.stats.govt.nz/browse_for_stats/population/Migration/define-migrants-travel-rule.aspx

Statistics NZ, Integrated Data Infrastructure http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx (2016)

Statistics NZ, Data integration manual http://www.stats.govt.nz/methods/data-integration/data-integration-manual-2edn.aspx (2015)

Statistics NZ, Coverage in the 2013 Census based on the NZ 2013 post-enumeration Survey http://www.stats.govt.nz/browse_for_stats/population/census_counts/report-on-2013-post-enumeration-survey.aspx (2014)

C. Spain[6]

1. Introduction

76. INE-Spain does not currently apply data integration techniques to produce migration statistics. The only source from which these statistics are obtained is the population register, although statistical treatments are performed on the raw data included in the population register to improve the aggregate figures on flows.

77. However, the 2011 population census in Spain was a clear example of integration of different sources to improve the measurement of population stocks and especially the stock of foreign residents in Spain. The migratory flows for period 2008-2011 were consequently revised. The method used for data integration in the 2011 census is described below.

2. Primary source for the measurement of migration statistics: the population register ("Padrón")

78. The main source in Spain as regards both population stocks and migration statistics is the population register, named Padrón in Spanish. Padrón is the official list of residents in each of the 8,100 municipalities in Spain. Padrón, as a list of neighbors of every municipality, comes from a very old tradition in Spain, dating from the middle ages.

79. There are as many registers as municipalities in Spain. But there is a law, in force since 1996, integrating these municipal lists into a single national database. There are also legal procedures to keep this database and the municipal files interconnected and updated on a monthly basis.

80. The law requires municipalities to exchange these records. This is made through the statistical office of Spain, INE. So, unlike other countries where the police or other administrative bodies are in charge of population registers, in the case of Spain, INE is the national institution that coordinates this single national population register. INE receives every month all changes produced in every municipality, performs validations and forwards these results to all the municipalities, to avoid duplications and also to include deaths, births or acquisition of Spanish citizenship that INE receives on a monthly basis (or more frequently) from the Civil Register.

81. All consular offices of Spain throughout the world (around 250) are also connected to Padrón like the municipalities. Any Spanish national leaving a municipality within Spain to live abroad should declare his/her new residence at the closest consular office. Consular offices send this information to INE and INE forwards these data to the municipalities into the monthly coordination files.

82. According to the law, there is no restriction for registering in Spain in terms of legal situation. All people living or willing to live in Spain, regardless of their

[6] Prepared by the Directorate for Socio-Demographic Statistics, INE.

legal situation have the right to be registered (this is actually an obligation) and they normally are, since being registered brings many advantages and no drawbacks. The benefits of being registered may comprise free access to public health system, public schools, services and rights granted by the municipalities, as well as cheaper transport and many others.

83. For each person, the register contains gender, date of birth, place of birth (country, in the case of foreigners), nationality, educational attainment and the national identity number. Foreigners in legal situation have an ID number. For people in irregular situation the passport number is stored.

84. Padrón is also a longitudinal database: all previous places of residence (within Spain) are also stored, thus allowing longitudinal analysis and a very close and precise monitoring of internal migration.

3. Limitations of the original source. Why is integration necessary?

85. As noted, Padrón is the source for population figures and migration statistics. It does not mean that the population count directly corresponds to the number of people registered at a given moment. Some statistical treatments are applied to the raw data included in the population register to provide the population figures.

86. The main drawback of a system based on a population register like this one is the difficulty it faces while measuring emigration. With regards to immigration it seems that registered inflows are likely closer to the true values. People may find many reasons to register but not many for de-registering.

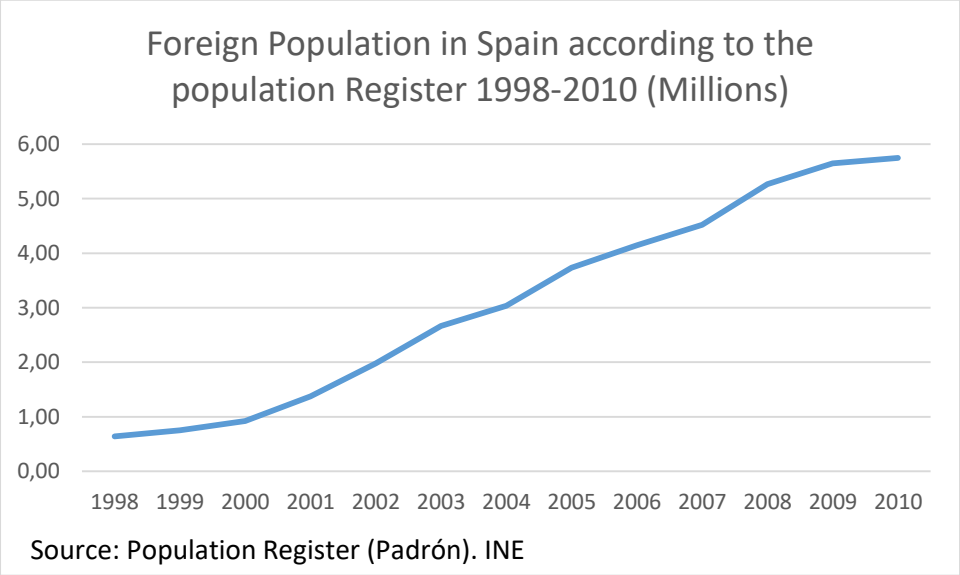
87. When the Population Register was set up, in 1996, the number of foreign nationals was 540,000 (1.4% of the population of Spain - 39.7 million at that time). Some years back, in the Census 1991, it was almost the same (470,000). Spain had a very small net international migration in the first half of the nineties.

88. But the second half of the 1990s, and especially the first decade of 21st century were very different. Because of the growth of the economy, the number of foreign residents increased dramatically. In only ten years (2000 to 2010) the number of foreign residents according to the population register grew from 1 million to near 6 million (from 1.4% to 12.2% of the population in 2010 in relative terms).

89. The register had to address an important challenge for which it had not been designed: how to count foreign people, and more specifically how to measure emigration. No process had been designed for de-registering foreign people when they did not declare their departure.

90. A new legal procedure was approved in 2006. According to their legal status, foreign residents can be divided into two groups: EU citizens or non-EU but having a permanent residence permit (let us call them “EUP” foreigners) and the rest, i.e., non-EU citizens that hold a non-permanent residence permit or do not hold a residence permit (“non-EUP” foreigners).

Figure 4



91. Since 2006, all non-EUP foreigners (roughly 50% of total foreigners at that time) are required to renew their inscriptions in Padrón every two years. Every city council must send notifications to all non-EUP foreigners whose inscriptions are about to expire. If they do not renew their registration, after an administrative process that may take some months, they are de-registered from the municipal records.

92. A similar process was set up for EUP citizens but it was built later, in 2009, and was not fully applied until 2016.

93. At the moment when the decisions on the population census had to be taken, around 2008-2009, there was an important uncertainty on the real number of foreign residents in Spain. How many of the nearly 6 million foreigners registered since 1996 had left the country without declaring so to public administration? How many Spaniards had left the country without declaring a new residence abroad to consular offices?

94. Eventually, in 2009, the methodology for the census 2011 was outlined. It was conceived as a combination between registers (linking Padrón with others) and a survey addressed to 10% of population, mainly used to collect all demographic variables not included in registers.

95. The main concern when designing the census methodology was how to combine registers to improve the measurement of population stocks, both for Spaniards and foreigners, considering that there was little evidence about outflows in the Population Register, notably for the foreign population. In the case of Spaniards, at least there is a source of information on outflows, the declaration at consular offices.

4. Methods used for data integration

96. The 2011 population census included a complex procedure of data integration to improve the population count. This procedure was used to count the so-called

"doubtful" population, i.e. registered residents in Padrón without evidence of residence in Spain in other administrative registers (tax collection agency and social security databases among others).

97. During the 2011 Census, in order to count the population stock, the registered population (both Spaniards and foreign residents) was linked with many different administrative registers to provide to each person a likelihood of residence, a number called "count factor". As a result of this process a "pre-census" file was built. The main principle is simple. If a person is correctly registered and if he/she is receiving a salary or a pension, or paying taxes during 2011, then this person is very likely to be residing in Spain. These verified records are named "sure" population. But people not found in other records are considered "doubtful" population.

98. The pre-census file was the result of this process, with roughly 47.4 million records.

99. Around 15 different criteria by which a record could be considered doubtful were defined. After this linkage, the results were:

i. 97.7% of the pre-census file records (46,372,000) were considered "sure" population (a count factor equal to 1 was assigned to each person in this group).

ii. Around 40,000 records (0.1 % of population) were removed from the pre-census file after a probabilistic record linkage with death records. They were considered errors in the population register and were assigned a count factor equal to 0.

iii. For the remaining 2.2% of records the evidence was inconclusive, so they were considered doubtful records. There were 1,046,000 such records, and 87% of them were foreigners. Therefore, their "count factor" was unknown.

100. The first finding of this process was that doubtful population were mainly migrants, but not economic migrants. During the decade before the 2011 census Spain had received large flows of migrants from Ecuador, Morocco or Romania, but there were also important and continuous inflows of Germans or Britons that came to Spain to live after their retirement. These "good weather migrants", often referred to as sun-seekers, are more difficult to find in registers: they can be registered in Padrón but they do not receive a pension from Spanish government nor do they pay taxes in Spain.

101. The "count factor" for doubtful records was calculated using the "census survey" (as mentioned, the 2011 census in Spain comprised a survey addressed to 10% of population).

Table 4: Doubtful records in 2011 Pre-census file. Selected nationalities

Nationality	Total Population in pre-census file with CF ≠ 0	Doubtful records	% doubtful
Total	**47,418,916**	**1,046,433**	**2.2%**
Spain	41,589,484	136,452	0.3%
Romania	886,631	120,216	13.6%
United Kingdom	394,998	112,100	28.4%
Morocco	806,012	61,459	7.6%
Germany	195,667	58,690	30.0%
Bolivia	197,113	37,492	19.0%
Paraguay	90,788	32,107	35.4%
Bulgaria	173,061	31,269	18.1%
Brazil	105,216	28,352	26.9%
France	120,780	27,482	22.8%

102. For this purpose, doubtful records were grouped into population classes defined by sociodemographic characteristics: age groups, nationality and province of residence (there are 52 provinces in Spain). The pre-census file was partitioned into these classes. Each class contained sure and doubtful records. If a class did not contain any doubtful record, then no estimation was needed, and the population for that class was equal to that from the pre-census file.

103. All classes containing doubtful records were configured to include at least 1,000 such records. Classes with less than 1,000 doubtful records were grouped in order to reach that threshold. The grouping strategy was to first select wider age groups, then wider territory areas and lastly grouping nationalities. 724 clusters or classes were created, with at least 1,000 doubtful records in each.

104. In order to calculate these count factors we needed to estimate the proportion of sure population in the survey. In a given class (i) of the pre-census file, the total number of records, Ti was split into the number of records of sure population Si and the doubtful ones Di,

$$T_i = S_i + D_i$$

105. To improve the estimation of the total population, the count factor was applied to doubtful records in class i as shown below to create a new estimation of total population for this class, Ti*:

$$T_i^* = S_i + CF_i * D_i$$

106. The population from the survey was grouped into the same 724 classes according to the same characteristics (age, territory, nationality). After that, the sample and the pre-census file records were linked at the individual level to determine which records from the sample had been previously classified as doubtful in the pre-census file.

107. With those count factors incorporated to every doubtful record, the pre-census file became the Weighted Census File (WCF) and allowed for the derivation of

census population figures. The population of a given geographic area, T_i^*, was obtained as the sum of the count factors of the WCF records in that area (given that CF=1 for all sure records).

108. This procedure provided statistical figures for the census, but of course, it cannot be used for administrative purposes to update the population register, since it is not possible to determine which individual citizen should be counted and which one should not.

109. The most noteworthy class, in terms of the proportion of doubtful population was that of German nationals aged 80 and over in the province of Alicante (a place in the Mediterranean coast), comprising nearly 3,000 persons, 70% of whom were considered doubtful. The largest class in terms of absolute number of doubtful persons is that of citizens of the United Kingdom, also living in the province of Alicante, aged 60 to 65 years old. Comprising 5,963 doubtful records, out of 19,317 UK Citizens included in this group in the pre-census file.

110. Some selected cases are shown in table 5.

Table 5: Doubtful population in the pre-census file group by Nationality-province-age

Group (Nationality, province, age)	Population in pre-census file (T_i)	Sure records (S_i)	Doubtful records (D_i)	% doubtful records	Count factor (CF_i)	Estimated population $T_i^*=S_i+(D_i*CF_i)$
United Kingdom, Alicante, 60 to 64 years	19,317	13,354	5,963	30.9%	0.347	15,426.1
United Kingdom, Alicante, 65 to 69 years	21,180	15,287	5,893	27.8%	0.222	16,598.1
Germany, Alicante, 70 to 74 years	6,703	2,633	4,070	60.7%	0.269	3,729.1
United Kingdom, Alicante, 55 to 59 years	11,732	7,663	4,069	34.7%	0.254	8,695.3
Romania, Madrid, 25 to 29 years	32,909	29,867	3,042	9.2%	0.257	30,649.8
Pakistan, Spain, 25 to 29 years	4,384	3,329	1,055	24.1%	1.370	4,774.5
Morocco, Madrid, 0-9 years	17,162	16,158	1,004	5.9%	0.115	17,046.6

5. Benefit of integration with statistics from other sources

111. As stated before, the number of doubtful records were 1,046,000 in round numbers. The average count factor for all classes was 0.424, which means that these doubtful records were counted as a population of approximately 440,000 people.

112. The CF was higher than 1 in only 22 classes, meaning that these populations were considered to be sub-registered. This was the case of nationals of Pakistan aged between 25 and 29, with a count factor of 1.37 and affecting 1,055 doubtful records (subsequently counted as a population of 1,445).

113. As a result of these calculations, the population of Spain reached 46,815,916 inhabitants as of November 1st 2011, some 450,000 inhabitants below the one provided by the Population Register at the same date. While the number of Spaniards hardly changed, the stock of foreigners according to the census decreased by 480,000

compared to the population register figures to reach a total number of 5,252,473 (11.2 % of the population of Spain, 8% fewer foreigners than those considered in the population register).

114. The most prominent case of decrease in population figures was that of UK nationals. Though according to the Register there were 391,194 UK nationals, the number according to the census was only 312,098.

115. In conclusion, the use of data integration techniques, combining administrative data and survey-based data improved the population figures, mainly on foreign nationals. Moreover, since 2011, the population register has been greatly improved. Procedures have been put in place in town councils to verify the effective residence of foreign population, especially "EUP" foreigners and to improve the population figures provided by the population register.

References

http://www.ine.es/en/censos2011_datos/cen11_datos_metodologia_en.htm

Argüeso, Antonio and Vega, Jorge L. (2013) "A population census based on registers and a "10% survey" methodological challenges and conclusions". Statistical Journal of the IAOS - Volume 30.

D. Switzerland

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

117. 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. The 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.

118. 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.

119. The register survey is the 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 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).

120. 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.

1. Primary data sources for the measurement of migration stocks and flows: the population registers

121. The main data sources for official annual migration statistics (demographic stocks and flows) are the administrative registers at the federal, cantonal and municipal level.

122. There are three different (types) of registers:

(a) 2,500 local population registers (maintained by the municipalities or in some cases the cantons7) for the national and non-national population;

[7] 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. It receives data extracts containing pre-defined variables in a pre-defined standardized data format from each register at regular intervals.

(b) 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);

(c) 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).

123. 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 the two federal registers). Since there is no single data source which provides 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 the 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 one.

124. 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 the sources.

125. The following decisions have been established for the production of annual migration stock and flow data:

i. 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.

ii. 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.

iii. The register of reference for foreign nationals who have been issued legitimation cards by the Federal Department of Foreign Affairs is the Ordipro information system.

126. Consequently, "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.

127. The local and the 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 the 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 the 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.

128. Moreover, the characteristics relating to an individual can differ from one register to another. Depending on the task or the position of the register in the administrative process, from a statistical point of view it can be assumed that the quality of some of the characteristics or variables (such as sex, age, marital status, citizenship, country of last/previous residence etc.) is better or more coherent in some registers than in others. Therefore it may be necessary to prioritize the sources for the modalities of certain characteristics. For the production of annual migration statistics it is again the "register of reference" which determines which modality is taken into account.

129. However, the Central Migration Information System provides some exclusive variables that are missing at the local level. They are added to the 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 the different administrative steps of the asylum procedure.

130. This data "enhancement" procedure has been implemented into the 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 unambiguous linking of records of one individual in different sources.

131. 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.

2. Other data sources for the measurement of migration stocks and flows not integrated into the production of annual migration statistics

132. There are additional sources of migration data in Switzerland, the two most important ones being the Swiss Labor Force Survey (LFS) and the Structural Survey (an annual sample survey of 200,000 people which is part of the new census system and collects variables which are not currently available in the registers, e.g. language, religion, educational attainment).

133. A key aspect of the new Swiss census system is the possibility of linking of data from population registers with data from an annual sample survey by means of the social security number. The information contents of both sources can thus be combined and supplement each other. However, at present no such matching or combining of data is being made within the context of annual migration statistics.

134. Migration-relevant information from sample surveys are presented separately and provide complementary insights into stocks and flows of migrants. Moreover, survey data and register data do not only differ in geographic scope and data collection methods but also in the population universe. Both the LFS and the Structural Survey cover only persons aged 15 and over living in private households.

E. The United Kingdom

135. The Office for National Statistics (ONS) produces statistics on international migration to and from the UK which contribute to our understanding of the make-up of society and the changing shape of the population. Our key migration statistics are:

i. Long-Term International Migration estimates (flows)

ii. Short-Term International Migration estimates (flows)

iii. Population of the United Kingdom by Country of Birth and Nationality (stocks)

136. This case study focuses on the ONS Long-Term International Migration (LTIM) estimates.

1. Data sources for measurement of migration statistics

137. The ONS Long-term International Migration (LTIM) flows data are based on the International Passenger Survey (IPS) with several adjustments made based on other survey and administrative data sources.

138. The International Passenger Survey (IPS) is a multi-purpose survey that collects information from passengers as they enter or leave the UK. It is mainly used to provide data about international migration, travel expenditure and tourism.

139. The IPS is a sample survey which is conducted at all the main entry and exit points from the UK including airports, sea ports and at the Channel Tunnel.

140. To collect the data, a specific entry or exit point is sampled on a particular day. A strict method of counting is used which ensures each person has the same opportunity of inclusion in the survey. Passengers are systematically selected for interview at fixed intervals (for example, every 1 in 20th person) from a random start.

141. Between 700,000 and 800,000 people are interviewed on the IPS each year. Of those interviewed, approximately 4,000 people each year are identified as long-term migrants. These are non-UK residents who state that they intend to stay in the UK for at least 12 months (immigrants) or UK residents who state that they intend to reside outside of the UK for at least 12 months (emigrants). The remainder are identified as either short-term migrants, visitors to and from the UK, or UK residents returning after or leaving for a short stay overseas.

142. Further information on the IPS can be found in the International Passenger Survey Quality Information in Relation to Migration Flows.

2. Purpose of data integration

143. Although the IPS is the prime source of long-term migration, there is no single, all-inclusive system in place to measure all movements of migrants into and out of the UK. The IPS has some limitations with respect to measuring immigration and emigration, as it:

i. Is a sample survey and so the estimates are subject to a degree of uncertainty

ii. Captures very few asylum seekers who may be entering or leaving the UK, or non-asylum enforced removals from the UK

iii. Does not take into account the changing intentions of passengers; these are passengers who intended to remain in or out of the UK for 12 months, but actually spent less than a year (migrant switchers) and those who believed they would be staying or leaving for less than a year but actually spent longer (visitor switchers)

iv. Does not capture those who are crossing the land border between the UK (Northern Ireland) and the Republic of Ireland.

144. Therefore, it is necessary to use a combination of data from different sources that have different characteristics and attributes in order to produce estimates of LTIM.

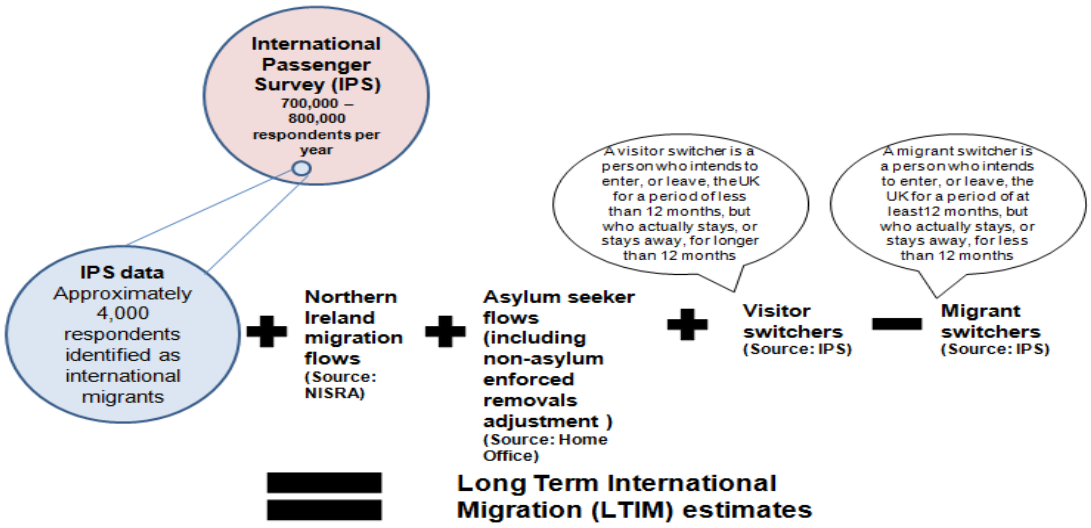
145. Further information on the IPS can be found in the International Passenger Survey Quality Information in Relation to Migration Flows.

3. Data integration methodology

146. The methodology outlined here was first applied in 2009 for the calculation of the 2008 estimates and revisions were made to earlier years as appropriate.

147. Estimates of LTIM are about 90% based on data from the IPS. A more comprehensive estimate of long-term international migration is produced by combining the IPS data with the information provided by additional data sources. Adjustments are also made for migrants who change their intentions known as visitor and migrant switchers. This more inclusive estimate is referred to as Long-Term International Migration (LTIM) (see Figure 5).

Figure 5: Calculation of Long-Term International Migration



Source: Office for National Statistics

148. None of the data sources used, including the IPS, are specifically designed to capture information solely on long-term migration. Estimates of LTIM are produced using the following main data sources:

i. International Passenger Survey (IPS)

ii. Labour Force Survey (LFS) - provide a geographical distribution of migrants for the calibration of IPS

iii. Home Office immigration administrative systems; which provide data on asylum seekers and their dependants for 1991 onwards, from 2013 data on non-asylum enforced removals are also used and from 2015, adjustments for people resettled in the UK under various resettlement schemes

iv. Forecasted long-term international migration estimates based on previous GP registrations from the Northern Ireland Statistics and Research Agency (NISRA) for estimating long-term international migration to and from Northern Ireland and the rest of the world, from 2008 onwards (forecasted data is replaced with final data for LTIM final annual estimates). This captures land routes that are not surveyed by the IPS.

149. The published LTIM figures are broken down to show estimates by variables such as citizenship and age and sex. To produce estimates for each of these variables, data from the sources that contribute to LTIM also need to be broken down by the same variables.

150. Migrant data from the IPS is available broken down by each variable. Data on Northern Ireland flows and asylum seeker data are not, and need to be derived using a series of assumptions. In addition, the IPS data used to calculate the visitor switcher adjustments are based on a relatively small sample size each year, but still need to be broken down in the same way.

151. As data received from the supplementary sources are not as detailed as the data collected by the IPS, the level of analysis that can be performed on the LTIM data is limited. While LTIM tables usually contain just 1 variable, cross-tabulations of IPS data are available separately.

152. LTIM estimates are shown with the confidence intervals for the IPS component of the estimate in order to give you an indication of the accuracy of the estimate. The uncertainty associated with the IPS component of the estimate is used to calculate statistically significant changes in the LTIM estimates. However, when interpreting these confidence intervals and statistically significant changes, you should be aware that there is no method for quantifying the error associated with the non-survey components of LTIM, which are unlikely to be random.

153. Further information on the methodology including the assumptions used to produce IPS and LTIM estimates ("Methodology to estimate LTIM"), and changes to the methodology over time, is available on the ONS website. This also includes copies of the questionnaires used in the IPS.

4. Use of data integration with other sources

154. There are many sources of official statistics that measure the number and characteristics of international migration into and out of the UK (flows) as well as the migrants who have settled in the UK (stocks). Taken together they provide a rich picture of migration in the UK. It is important to understand that these sources measure different things: some measure flows, some measure stocks, some measure workers, some students and some only measure the characteristics of those migrating from outside the EU. Each source is valuable in its own right in measuring particular aspects of international migration.

155. As noted, none of the data sources including the IPS, while offering the best data currently available, are specifically designed to capture information solely on LTIM and alone cannot quantify all population inflows and outflows. Data integration is essential to ensure the best possible estimate of international migration is made, trusted and used confidently by the UK government and the wider community.

156. ONS is committed to providing the right information to inform public and policy debate and international migration is a topic of considerable debate and there is a high demand for trusted data and analysis. Regular methodological changes are made to ensure estimates are as inclusive and robust as possible, particular where new data sources become available. Details of changes can be found in the Methodology to estimate LTIM. In addition in February 2017 ONS published a note on plans for the development of international migration statistics with a particular focus on making better use of current administrative data sources (see International migration data and analysis: Improving the evidence).

F. Canada

1. Original/primary source for the measurement of migration statistics

(a) For Population Estimates purposes

157. Immigration, Refugees and Citizenship Canada (IRCC) collects and processes immigrants' and non-permanent residents (NPRs) administrative files. It then provides Statistics Canada (STC) with the information.

158. For immigration, the files are used to estimate the number and characteristics of people granted permanent resident status by the federal government for a given period.

159. For NPRs, the files required to produce their estimates include visitor's permits, work permits, study permits, Minister's permits, refugee status claims, landings, applications for landing and deportations. The information is used to estimate the number and characteristics of people granted non-permanent resident status by the federal government.

160. Emigration estimates require a distinction between persons establishing a permanent residence in another country (i.e., emigrants), persons living temporarily abroad while not maintaining a usual place of residence in Canada ; other temporarily leave Canada and then return (i.e., net temporary emigrants), and finally the portion of emigrants who have returned to Canada (i.e., returning emigrants).

161. Emigrants are estimated from administrative sources in terms of the gross flow of migrants out of Canada. The Office of Immigration Statistics of U.S. Department of Homeland Security provides data on Canadians who acquire permanent immigrant status in the U.S. This data source is used in estimating emigration to the United States. In order to estimate emigration to other countries, information on notification of departure from the Canada Child Tax Benefit (CCTB) program and tax data from Canada Revenue Agency (CRA) is used. Various adjustments are made to the data, for example to correct for incomplete coverage.

162. Net temporary emigrants are estimated with few sources. Data from the Reverse Record Check, the census coverage study used to measure undercoverage, are used to estimate the number of persons leaving the country temporarily; while data from the National Household Survey (NHS), combined with Demography Division's estimates of returning emigrant, are used to estimate the number of temporary emigrants returning.

163. Returning emigrants are Canadian citizens or immigrants having previously emigrated from Canada and subsequently returned to Canada to re-establish a permanent residence. Again, data from the CCTB program and tax data are used in estimating returning emigrants.

(b) For analytical characteristics

164. The permanent and non-permanent resident data provided by IRCC offers relevant information for analytical research and policy needs. This data source is a census of the immigrant population and, as such, offers the capacity to study small populations in detail. Data source relating to permanent residents contains socio-demographic information of immigrants who have landed since 1980. Characteristics related to knowledge of languages, destination in Canada, country of origin, intended occupation or level of education are obtained at the time of admission.

165. For NPRs, pre-admission files provide detailed geographic information and the main activity of the individual such as work or study during specific periods after 1980.

166. The Census of Population provides a statistical portrait of the country every five years. The 2016 Census identifies permanent and non-permanent residents living in Canada. It provides many socio-economic characteristics of immigrants such as family composition, knowledge of languages, labour force status and income.

2. Limitations of the original or main source. Why is integration necessary?

(a) For Population Estimates purposes

167. Immigration data are straightforward administrative data. However, for provincial and territorial estimates, the file obtained from IRCC indicates the province or territory of intended destination on arrival, rather than the province or territory in which the immigrant actually settles. In a small number of cases, information on the province of destination is missing. These cases are distributed proportionally between the provinces and territories according to the observed distribution of immigrants for whom this information is available.

168. For NPRs data, there is the same limitation for the intended province of settlement as for immigration data. There is also a lack of information on the departure dates of NPRs as some of them leave the country before the end of the validity of their permits. Since it is not mandatory to inform the federal government before leaving the country, STC does not have any additional sources to obtain this information.

169. Moreover, both immigrants and NPRs are assigned a Client Identification Number (CID) by IRCC. This identification number is a key variable as it allows identifying each person only once. The CID is used in the production of population estimates because it allows STC to create a biography of each individual, even if the person holds more than one permit at the same period.

170. Unlike immigration, there is no legal provision in Canada to maintain records for persons leaving the country either on a temporary or permanent basis. Therefore, estimates of the number of emigrants and persons living temporarily abroad and their

characteristics must be derived through secondary sources such as Canadian administrative files or immigration statistics of the United States. Also, few data sources allow for a measure of emigration and these sources often use different definitions of emigration. Given its nature, temporary emigration is notably hard to measure. As a result, various evaluations suggest that Canada underestimates the levels of emigration and temporary emigration. In that context, integration is necessary to make the best possible use of the strengths of the various datasets. For instance, as the United States is the country of destination of most emigrants from Canada, using American immigration data is thought to improve our measure of Canadians moving to the US.

(b) For analytical characteristics

171. The administrative data from IRCC does not provide information on any outcomes after admission. Information related to socio-economic outcomes, changes in the family structure, and changes in residence may only be obtained from linkages to external data sources such as annual tax files.

172. The Census of Population does provide up-to-date outcomes on immigrants after landing but it does not include any information on admission conditions or selection criteria.

3. Methods used for data integration

(a) For Population estimates purposes

173. In the case of immigration and NPRs, information is integrated at the individual record. All the files contain the same variables and are linked using a unique CID. Measuring the number of immigrants entering Canada in a given period is straightforward, and adjustments to the data are not required. Information is available for each person entering Canada under immigrant status from IRCC's administrative file. Every month, IRCC makes a data file containing the records of immigrants for the previous month available to STC, as well as any additions or updates to data already received.

174. For the NPR population, there are two major subgroups that are administratively different (permit holders (PH) and refugee status claimants (RSC)); their estimates must, therefore, be produced separately. For more information regarding the methodology of the RNPs, please refer to http://wwwstaging.statcan.gc.ca/pub/91-528-x/2015001/ch/ch5-eng.htm#n1

175. As for emigration (emigration, net temporary emigrants and returning emigrants), please refer to the methodological document: http://wwwstaging.statcan.gc.ca/pub/91-528-x/2015001/ch/ch6-eng.htm

(b) For analytical characteristics

176. To circumvent the limitations of the administrative immigration data, Statistics Canada uses exact matching record linkage techniques to combine the information with other data sources such as tax files. The Longitudinal Immigration Database (IMDB) is updated every year and includes all immigrants who have landed in Canada since 1980 with outcomes from annual tax files since 1982. For processing year 2014, 89% of immigrants were linked to the tax files. As such, the database is an administrative census with a longitudinal design, therefore no sampling is done.

177. In addition, the 2011 National Household Survey was linked with immigration administrative data to connect admission conditions with the socio-economic outcomes available on the survey. For the 2016 Census of Population, record linkage is being used to add variables related to admission category to the census database.

4. Benefit of integration with statistics from other sources

(a) For Population Estimates purposes

178. The main benefit of integrating different sources of data is to provide estimates of great quality that will be useful not only for STC, but also for its many external users. Finally, STC is currently working on a Statistical Population Register. It would then be possible to link immigration and emigration data through a unique personal identifier.

(b) For analytical characteristics

179. The record linkage between Census survey and the immigration administrative file provides additional analytical capacity while reducing the burden on respondents. This approach is a good alternative to obtain details on immigration selection category that respondents may neglect, overlook, or cease to remember over time.

180. Being mostly transactional in nature, IRCC immigration administrative files are somewhat restricted with regard to performance and research indicators. The IMDB is created from the record linkage between immigration and tax data files to provide a more comprehensive source of data on socio-economic outcomes of immigrant tax filers. This data source contains detailed information on labour market behavior and covers a period long enough to assess the impact of characteristics at admission, such as education or knowledge of French or English. Elements such as pre-admission experience of immigrants (i.e. study experience in Canada), mobility within Canada, or changes in the family composition take advantage of the longitudinal aspect possible with each new tax report.

G. Israel

1. Original/primary source for the measurement of migration statistics

(a) For Population Estimates purposes

181. The Israeli Population and Immigration Authority (PIA) is responsible for the border control registration (BCR) and for the registration at the population registry (PR). PIA provides the Central Bureau of Statistics (CBS) separate files for Israelis and for foreigners. These administrative data make it possible to distinguish between immigrants and non-permanent residents (NPR) and allows for measuring the flows and the stock of emigrants.

182. For NPRs, the files required to produce their estimates include Visa at admission time. Although the files include different types of visas we only refer to visas for foreign workers. PIA provides CBS with a special registration for non-permanent residents who entered Israel illegally, not through an organized border.

183. For Permanent residents, CBS uses the new registration at the PR with distinction between new citizens (mostly Jewish or of Jewish extraction who immigrate to Israel under the Law of Return) and non-citizens (mostly family reunification).

184. For emigration estimation the CBS uses the BCR to examine the length of stay abroad.

185. An out emigrant is considered one who has gone abroad and stayed abroad for at least one year. A Return emigrant is considered someone who has returned from a stay abroad of at least one year and who has resided in Israel for at least three consecutive months after returning from abroad.

(b) For analytical characteristics

186. The permanent and non-permanent residents' data provided by PIA offers primary source of administrative data for analytical research and policy needs. This data source is a census of the immigrant population and, as such, offers the capacity to study small populations in detail. Data source related to permanent residents contains socio-demographic information of immigrants who have landed in Israel since 1970. Characteristics related to destination in Israel, country of origin, intended occupation or level of education are obtained at the time of admission. For additional administrative information we link permanent residents with other administrative data (education, employment, income…).

187. The Census of Population and all household surveys include a module of questions that allow us to know the year of immigration and the country of birth of permanent residents. This module enables CBS to get information on all topics of the surveys.

2. Limitations of the original or main source. Why is integration necessary?

(a) For Population Estimates purposes

188. The information available to foreigners (non-permanent residents) is the information received at the admission day. This information lacks geographical breakdown to the place of residence. The information identifying the individual in these files does not allow the linkage of records with other sources (administrative or surveys). These limitations restrict CBS in preparing the data beyond the data existing in the administrative files. There are also some cases, where no linkage is possible between arrival and departure records. This may affect our stock estimation. This limitation exists especially for records with tourist visa and limits us in the estimation of over stayers (unreported migration).

189. Many Israelis have citizenship of another country. In most cases, the BCR knows how to identify dual citizenship and registers their border crossings even if the person has crossed the border with a non-Israeli passport. In some cases the BCR misses the information of dual citizenship. This limitation affects the estimation of emigration.

3. Methods used for data integration

(a) For Population Estimates purposes

190. In the case of Permanent resident immigration, information is integrated at the level of the individual record. All the files contain the same variables and are linked using a unique ID (personal identifier of PR). Measuring the number of immigrants entering Israel in a given period of time is straightforward, and adjustments to the data are not required. Information is available for each person entering Israel under immigrant status from PIA administrative file. Every month, PIA prepares a data file containing the records of immigrants of the previous month and makes it available to CBS, as well as any additions or updates to data already received.

191. For the NPR population, there are three major subgroups that are administratively different: permit foreign workers, irregular foreign workers and immigrants who cross the border illegally; their estimates must, therefore, be produced separately. In the case of permit foreign workers and illegally foreigners we get a list from PIA and we use it as is. For more information about the methodology In the case of irregular workers, please refer to http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2016/mtg2_WS/14_Israel_Sheps_ENG.pdf.

192. As for emigration, please refer to the methodological document: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2006/wp.10.e.pdf

(b) For analytical characteristics

193. In order to complete information or carry out longitudinal studies, the CBS uses exact-matching record linkage techniques to combine the information of immigrants with other administrative files (as tax files. Education files, etc.), and with population censuses and surveys.

194. The matching is also used to add variables from data sources that are related to admission day.

4. Benefit of integration with statistics from other sources

(a) For Population Estimates purposes

195. The use of unique ID in all administrative files and the collecting of ID in most of Israeli surveys and in the population census allow CBS to combine all these sources together. The main benefit of this operation is to provide estimates of greater quality that will be useful on the national level.

196. In addition, the integration of information enables the CBS to reduce the response burden and reduce costs of data collection.

(b) For analytical characteristics

197. The CBS allows its employees and external users to conduct data processing and research using the combined data. Researchers receive access to data in research rooms located in the CBS buildings. There is no standard data file on immigrants that is accessible to researchers. Usually every researcher defines his or her data according to the needs of the research. The data allow users to research the absorption and integration of immigrants in the Israeli society and in the economy, enabling the monitoring of their population movements in the population. The family relations in the PR allow users to get information about the second generation as well.

198. In the case of emigration, studies were performed on the characteristics of the emigrants before departure and afterward. Linked data of emigrants from various sources allows the study of the brain drain phenomenon, for example.

H. Netherlands

1. Primary sources for the measurement of migration statistics

199. The Dutch demographic statistics are entirely based on the Dutch population registers. As such, describing immigration statistics in the Netherlands basically boils down to describing the definitions and practices used in the population registers. However, some limitations are present. By integrating administrative data from many sources into the System of Social Statistical Datasets (SSD), Statistics Netherlands faces these limitations.

200. Everyone who enters the Netherlands is registered as a resident (immigrant) provided

i. His/her stay is legal according to the Immigration Act (on people who do not have the Dutch nationality);

ii. The intended stay is at least two thirds of the forthcoming six months;

iii. The person is properly identified. The latter means that a valid passport or other official document is shown for identification.

201. Every child born in the Netherlands whose mother is registered as a resident is also registered as a resident. Children who are born abroad to a mother who herself is registered as a resident of the Netherlands are registered (as immigrants), provided the children will live in the country.

202. Emigration relates to persons who leave the Netherlands and intend to stay abroad for at least two thirds of the forthcoming twelve months and who inform the municipal authorities of their departures. They are included in the Dutch emigration statistics.

2. Limitations of the original or main source. Why is integration necessary?

203. There are two exceptions to the rules for registration of immigrants, who make the registrations not entirely complete.

204. The first applies to so-called privileged persons, including foreigners working on Dutch soil as diplomatic, consular or military officials, or in an international organisation. Since they benefit of a special 'privileged' status and are not considered foreigners under the auspices of the Immigration Act, they are given the choice whether or not to be entered in the population register.

205. Asylum seekers set another exception. Their registration takes place only six months after their arrival in the Netherlands, irrespective of their expected duration of stay, unless they are granted a residence permit within six months. In that case they are registered when the residence permit is granted. However, children born to asylum seekers who are not yet registered are registered directly after birth. This leads to the somewhat odd situation that a new-born baby is registered whereas the parents and siblings are not (yet).

206. When it comes to emigration, about one in three persons who leave the country does not notify the municipal authorities of his or her departure. When the authorities find out that someone is 'missing', the law stipulates that they must investigate his or her whereabouts. If the investigations lead to the conclusion that the persons remains missing, he or she is administratively registered as emigrated to a country qualified as 'unknown'. They are included in the statistics of administrative removals. If they return to the Netherlands they are included in the statistics of administrative entrances.

3. Description of the methods used for data integration

207. All production processes within Statistics Netherlands concerning social or spatial statistics are based on the System of Social Statistical Databases (SSD). The SSD contains a wealth of information on persons, households, jobs, benefits, pensions, education, hospitalizations, crime reports, dwellings, vehicles and more. Before the SSD, various statistical registers were scattered within Statistics Netherlands and were not standardized. By the process of micro-integration, the SSD resulted in a library of standardized and linked statistical registers, as well as an organization which has been put in place to control various aspects of the system.

208. See for more information on the SSD: *Bakker, B., Van Rooijen, J., & Van Toor, L. (2014). The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. Journal of the International Association for Official Statistics, 30: 411–424.*

209. When new registers come to Statistics Netherlands, they also converge into the SSD, using the same system of standardization, linkage and micro-integration. Recently, Statistics Netherlands added data from the Central Agency for the Reception of Asylum Seekers (COA) to the SSD. This will make solve the problem as described above, as soon asylum seekers will be registered from the first day entering The Netherlands. Hopefully other sources, such as PROBAS (a database containing privileged persons) of RNI (a municipal registration of non-residents) will be added to SSD soon.

4. Benefits of integration with statistics from other sources

210. The benefits of integration with statistics from other sources are evident. First of all, integral data is available, reducing the need for surveys. This reduces the costs of gathering data and increases the number of respondents. The benefits of standardisation and micro-integration of data are that information from different sources can be linked to the same persons. This result in a data library containing more than fifty administrative registers. The integrated data of Statistics Netherlands therefore offer a rich source of data, which allows researchers not only to measure migration, but also gives the opportunity to study social-economic integration with a broad range of variables.

I. Hungary

1. Original/primary source for the measurement of migration statistics

211. At the Hungarian Central Statistical Office (HCSO) the Population and Social Protection Statistics Department produces the annual population number, the annual population figures are based both on the registers of the vital statistics and on the migration data.

212. The number of foreigners having a residence or a settlement document as well as those with a refugee status and persons under subsidiary protection, who have a registered address in Hungary are calculated from the administrative registrations.

213. The migration data are produced using different types of registers, these data sources are under scope of various authorities, and these authorities are the followings:

214. Immigration and Asylum Office (IAO). The Office of Immigration and Asylum is responsible for two different types of registers. A memorandum of cooperation exists between the HCSO and IAO since the year 2009:

i. EEA-register, the register contains data of EU and EEA citizens and their accompanying family members.

ii. IDTV-register, this register contains data of third-country nationals. Both of the registers are established for administrative purposes, the statistical data collection is a "by-product" of the registers.

215. According to the registers of IAO two types of foreign migrants can be distinguished:

i. Non-national immigrants are foreign citizens who applied for, and were granted residence permit at the Immigration and Asylum Office, and who hold a registration certificate (residence card in the case of EEA citizens, or residence permit in the case of third-country nationals).

ii. Non-national emigrants are persons who used to own a document which entitled them for a legal stay at the territory of Hungary, but this document expired or their residence permit was invalidated. From 2012, their number contains estimations based on the Census of 2011.

216. Ministry of Interior. The Ministry of Interior runs the central population register (CPR), the CPR contains data of persons immigrating in Hungary and data of persons emigrating from Hungary who are obliged to register an address in the CPR (both of Hungarians and foreigners, according to the current Hungarian law about two third of foreign citizens legally present in Hungary).

217. In the CPR data of persons whom international protection is granted and data of persons naturalized in Hungary are stored. A person naturalized in Hungary is someone who became Hungarian citizen by naturalization (he/she was born as a foreign citizen) or by renaturalization (his/her former citizenship was abolished). The rules of naturalization in Hungary were modified by the Act XLIV of 2010. The act introduced the simplified naturalization procedure from 1 January 2011 onwards,

and made it possible to obtain Hungarian citizenship without a residence in Hungary (for persons who can prove having Hungarian ancestors).

218. The data published by the HCSO refer only to those new Hungarian citizens who have an address in Hungary.

219. Ministry of Human Capacities: National Health Insurance Fund of Hungary (NHIF). The NHIF is also one of the main data source of migrants who are insured in Hungary which is the waste majority of the population. Regarding the tasks of calculating the migrant population the data of NHIF are used for establishing the number of immigration and emigration of Hungarian citizens (according to the law in power the Hungarian citizens shall deregister in the NHIF if he/she leaves the country to another EEA MS and vice versa shall register back if he/she returns to Hungary from a stay in an EEA country. According to this migrants with Hungarian nationality can be divided into the following categories:

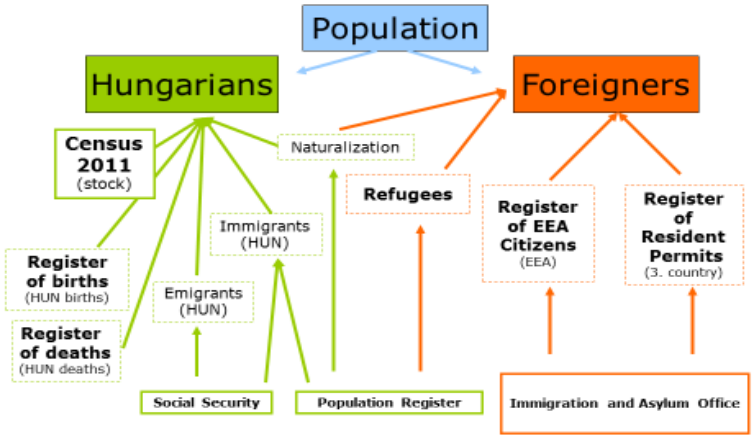
(a) National emigrants: persons who left Hungary and reported it to the National Health Insurance Fund.

(b) National immigrants belong to one of these two subgroups:

i. Hungarian citizens who returned from an emigration period and reported it to the National Health Insurance Fund.

ii. Hungarian citizens who established Hungarian address (and registered in the Population Register) after having been granted Hungarian citizenship while living abroad (without having a Hungarian residence previously).

Figure 6: Data sources of annual Hungarian population calculation



220. The Census 2011 is highlighted on the chart because the administrative data of 31 December 2011 was harmonized with the Census 2011 results, and the data of following years are based on this harmonization.

2. Limitations of the original or main source. Why is integration necessary?

221. The above described characteristics of the Hungarian migration statistics system stipulates a very complex system because the relevant data on migration are stored in various registers and in some cases even if the data contains migration related data it is stored in various sub-registers, e.g. data of foreigners.

222. In order to produce the number of immigrants and emigrants (both of Hungarian and foreigners) it is inevitable to integrate the data on migration from the different data sources.

223. In case of the Hungarian migration data production it has to be emphasized that it is not necessary to highlight a one register as main data source because in our case the data sources are almost equally relevant for the migration related data productions, the data sources are complementary to each other.

224. Production of migration data (Hungarian citizens and non-nationals). For the production of flow and stock data of migrants in Hungary the HCSO utilizes the data of registers of Immigration and Asylum Office (EEA-register and IDTV-register). During the procedure these two databases which are provided by the Immigration and Asylum Office are merged first, in order to clear the data and remove double registered records, the data are linked to CPR as well – it is essential to outline that CPR contains data of foreigners who have address registration [third-country nationals with residence permits (short-term stay) are not obliged to register in the CPR], second the registers of OIA are merged to CPR because the data of persons taken under international protection are registered in CPR.

225. The clearing of the primary data sources is an important step in the procedure because the primary data source contains numerous double registered data.

226. In order to get the migration data of Hungarian citizens HCSO currently does not use any data integration procedure, the data source for flow of Hungarian citizens is purely the NHIF.

3. Methods used for data integration

227. The data of registers at OIA, the data of CPR are provided twice a year to HCSO (at the beginning of the year and at the end of the year). The data provision is based on record level and the records are linked together upon a developed key-variable using SAS program. The linking method is a deterministic linking via unique identifiers which enables high quality integrated data sets, of course this integration method imposes high demands on the administrative data infrastructure. Mismatches occur from misspelling and name changes.

228. First the data of OIA are cleared from double registered data separately and afterwards the cleared dataset is interlinked with data of CPR.

4. Benefit of integration with statistics from other sources

229. On the one hand data integration of migration data is inevitable for HCSO because even in one authority the relevant data are stored in different registers in Hungary, on the other hand the data integration enables of production of high quality. Data integration is necessary for production of migration related data and it is also important for the annual population calculation.

230. Data integration creates additional information and the motivation for data integration emerges from a particular shortcoming of administrative data.

J. Latvia

1. Original/primary source for the measurement of migration statistics

231. Population number is estimated on the basis of data files provided by Office of Citizenship and Migration Affairs (hereinafter – OCMA) from the Common Migration Information System sub-system Population Register – individual data, incl. information on gender, birth date, birth country, citizenship, legal marital status, code of administrative territory of declared place of residence in compliance with Classification of Territories and Territorial Units (CATTU), residence permit.

232. Since January 2013, the CSB has been also receiving information on registration of civil status (marriage, birth or death) from OCMA Common Migration Information System sub-system Common System for Registration of Civil Status (CSRCS) as annual data file at individual level (previously information was received from county civil registry offices in paper form). With the help of the above data information is produced on the annual number of births and deaths.

233. Analysis of Population and Housing Census 2011 data resulted in acquiring information that helped CSB specifying population number and composition thereof in the country. On 1 January 2011, number of Latvia population accounted for 2 074.6 thousand. As compared to the information published prior (according to the Register), population number was 155 thousand or 7% smaller. The data confirmed that part of country population does not fulfil requirements of the Population Register Law and information included in the register on population migration is incomplete.

2. Limitations of the original or main source. Why is integration necessary?

(a) Problems identified within Population Register

234. Non or late deregistration – non-registered migration – even if according to the Population Register Law people have to inform the register about the changes – de jure marital status, children, usual place of residence etc., people not always do it (according to the 2011 Census results, 7% of population of Latvia had been moved abroad, but had not changed usual place of residence in the register (according to the register, they resided in Latvia). Also, analysing information from Population Register few years after Census, it is possible to see that there are quite a lot of people who correct the information in the register dilatory;

235. Very old persons – persons, who are in the register, however their usual place of residence is not Latvia, had died abroad - the Population Register didn’t receive the information, they stay in the register as alive. The worst situation – a person had left the country without deregistration, the usual residence in the register is still Latvia, however he or she had died abroad.

3. Methods used for data integration

236. In 2012 the CSB worked out a new method for estimating the number of population. The method is based on the Population Register and on individual data from other administrative registers (all data sources have ID codes which are used to merge data).

237. To evaluate the residence status of each individual the Logistic regression model had been developed. With the help of administrative register data, on each person registered within the Population Register there are more than 200 characteristics variables developed. The aim of the model is to predict the probability (number falling within the interval from zero to one) of being a resident for each individual, to estimate the status of the actual place of residence at the beginning of the year for every resident registered (on individual level) in Latvia. Necessary probability to be included in the resident population differs depending on age and gender.

238. The model had been developed using data from 2011 Census on actual place of residence and data from administrative data sources on 2010, 1 January 2011 or 1 March 2011.

239. In the result database on individual level on 1st January for certain year is set up. The database allows preparing necessary data tables for national and international data users.

240. Apart from OCMA data, population statistics is produced by using also individual data from other administrative registers available to the CSB and meeting corresponding time period (starting from 2010). CSB has access to data in following administrative registers:

i. State Revenue Service (SRS);
ii. State Social Insurance Agency (SSIA);
iii. Ministry of Education and Science (MES);
iv. Agricultural Data Centre (ADC);
v. Rural Support Service (LRSS);
vi. National Health Service (NHS);
vii. State Employment Agency (SEA);
viii. 32 from 58 higher education institutions (89% of total number of students);
ix. Road Traffic Safety Directorate data;
x. Prison Administration.

241. CSB has information also from Social Security Administration Information System (SSAIS) on persons that have received social benefits from local governments. However, this information is available only starting from 2012. Thus currently mentioned data are used to adjust number of population in age group 18 - 30 and to assess model quality.

242. Starting from 2016, population number is specified by also using SSIA data on persons who have been paid benefits or pension and who are located in social care institutions, as well as additional analysis is being carried out regarding the usual place of residence of a mother if the usual place of residence of the child (under 16 years of age) and father is Latvia according to estimates.

243. International long-term emigration from Latvia to another country theoretically corresponds with international long-term immigration data from Latvia to the respective country. This relation is called mirror statistics, and it is used to estimate international long-term emigration from Latvia.

244. When estimating international long-term emigration, information regarding immigration from Latvia received from other countries, e.g. Denmark, Finland, Sweden, Norway, Spain, the Netherlands, Austria, Iceland, Germany, is used. Not all countries prepare data on immigrants from Latvia, as Article 3 of the Regulation of the European Parliament and of the Council on Community statistics on migration and international protection stipulates that countries may provide information on immigrants in breakdown by group of previous usual place of residence: EU Member States; European Free Trade Association countries; candidate countries; other non-member countries.

245. To estimate emigration of the Latvian population to the United Kingdom and Ireland, the following information was used: the amount of UK National Insurance Numbers granted for the first time and the amount of Ireland Personal Public Service Numbers granted for the first time. It should be noted that these data are used only to evaluate general trends, because the respective numbers are given also to the Latvian residents staying in the UK or Ireland for less than one year.

246. To evaluate the quality of estimation household survey data are used (data on individual level from Labour Force Survey (LFS), Survey "EU Statistics on Income and Living Conditions" (EU-SILC), European Health and Social Integration Survey (EHSIS); European Health Interview Survey (EHIS). Evaluation of the results of the method was one of 2015 Micro Census's tasks.

247. Detail information about the methodology is available on CSB web page (http://www.csb.gov.lv/sites/default/files/dati/demstat_metodologija_eng.pdf).

4. Benefit of integration with statistics from other sources

248. The main benefit of integrating different administrative data sources is to provide more precise estimates of population that will be useful not only for statistical office, but also for national and international data users. The method developed will be used for Census 2021 which will be register based.

249. The method had been presented and positive comments received from the Latvian Statistical Association, experts from the Central Bank of Latvia, demographers. The main conclusion was – if the information on usual place of residence in the Population Register is of so low quality, the developed method is the best solution.

K. Australia

1. Sources of migration statistics

251. Administrative information on persons arriving in, or departing from, Australia is collected via various processing systems, passport documents, visa information, and incoming and outgoing passenger cards (see Appendix 1 of reference below). Incoming persons provide information in visa applications except those travelling as Australian or New Zealand citizens. These administrative data are collected by the Australian Government Department of Immigration and Border Protection (DIBP) under the authority of the Migration Regulations (Migration Act, 1958).

252. The Australian Bureau of Statistics (ABS) statistics on overseas arrivals and departures (OAD) are compiled using information from DIBP sources. Overseas movements are collected and matched (where possible) by DIBP and then stored with movement records on the Travel and Immigration Processing System (TRIPS). Each month the matched OAD movement records are supplied to the ABS and then processed. A unique personal identifier is the mechanism used for linking all the data sets at DIBP and then used by the Australian Bureau of Statistics to generate traveller histories of each individual traveller.

253. Quarterly net overseas migration (NOM) estimates are sourced from this processed monthly OAD matched data and then combined with monthly extracts of unmatched OAD records. Unmatched OAD records are those where an inward/ outward movement has been recorded by DIBP within the TRIPS system, but the data has not been able to be matched with an equivalent passenger card. Statistics on overseas migration exclude: multiple movements; the movements of operational air and ships' crew; transit passengers who pass through Australia but are not cleared for entry; passengers on pleasure cruises commencing and finishing in Australia and undocumented arrivals. From 1 July 2006 onwards, foreign diplomatic personnel and their families are also excluded.

254. For a reference to the methods on how Australia measures international migration see
http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/3412.0Technical%20Note12014-15?opendocument&tabname=Notes&prodno=3412.0&issue=2014-15&num=&view=

2. Sources for measuring migration stock

255. As mentioned above for international migrants the source for country of birth is data provided from Australia's DIBP which is a combination of passport and visa information. For deaths, the source for country of birth is information provided by the Birth & Death registrars in each State and Territory of Australia. All births are recorded as Australian born (obviously). In addition, the Australian Census is used once every five years to re-base the Australian Population including by Country of Birth.

256. For further information see Explanatory Note 5 in http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/3412.0Explanatory%20Notes12014-15?opendocument&tabname=Notes&prodno=3412.0&issue=2014-15&num=&view=

3. Use of data integration with other sources

257. One of the most useful variables has been a unique personal identifier (PID) for each individual who crosses the Australian border, in or out. This allows a lot of additional uses for the data including linking an individual over time and within various systems.

258. The Australian Bureau of Statistics in collaboration with the Australian Department of Immigration and Border Protection link migrant data sets with Census data. In addition, recent work has also linked this same data with Tax information - see relevant document below.

259. See the following releases by most recent first:

http://www.abs.gov.au/AUSSTATS/abs@.nsf/productsbyCatalogue/70AA0E84BE9D586ACA2575400017B0F3?OpenDocument

http://www.abs.gov.au/AUSSTATS/abs@.nsf/productsbyCatalogue/3B2787D4377D2D84CA2573F7000DDE5B?OpenDocument

http://www.abs.gov.au/AUSSTATS/abs@.nsf/productsbyCatalogue/706907E56F9F5128CA257BEA00111584?OpenDocument

http://www.abs.gov.au/AUSSTATS/abs@.nsf/productsbyCatalogue/C916B16440BF9B60CA257C7E000FC851?OpenDocument

http://www.abs.gov.au/AUSSTATS/abs@.nsf/productsbyCatalogue/208F6F1B4DED24ECCA257EB50011C711?OpenDocument

260. For a Guide to all the other Migrant Statistical Sources see:

http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/3414.0Main%20Features42011%20(Edition%202)?opendocument&tabname=Summary&prodno=3414.0&issue=2011%20(Edition%202)&num=&view=

L. Austria

261. Since 2002, Austrian migration statistics are based upon data from the Central Register of Residence ("Zentrales Melderegister (ZMR)). Residents who establish their home in a private or institutional household have the legal obligation to register and de-register their residence within 3 days of moving. All residents, independent of their nationality[8] or length of stay must register if their stay exceeds three days[9]. Statistics Austria receives and processes all residence registrations and de-registrations on a quarterly basis.

262. Data from the Central Register of Residence are enhanced by integrating data in two separate processes. First, information on deaths from the Organisation of Austrian Social Security ("Hauptverband der Sozialversicherungsträger" (HV)) is linked to the residence register at the unit record level adding information of deceased persons and their missing de-registrations. Second, in order to adjust for missing de-registrations of emigrants, yearly estimations identify potential nominal members employing information from several administrative data sources. While both steps link data on the micro-level in the final stage, the underlying processes differ substantially.

1. Adjusting for uncounted deaths in the population register

263. Data from the Organisation of the Austrian Social Security serves as a supplementary data source for means of improving the data quality of Austrian population and migration statistics. The Organisation of the Austrian Social Security is the umbrella organisation of all social security funds in Austria. It collects information on all persons insured in Austria and their dependents. Their data base covers information contingent on the insurance status, but not necessarily on the residence registration status[10]. Therefore, the residence register and the social security data base have partly overlapping populations but also cover categories of persons which are either only represented in one or the other data base.

264. Registration of deaths is compulsory for all deaths occurring on national territory as well as deaths of Austrian nationals occurring abroad in the Central Register of Civil status (ZPR). Succeeding the death registration, the deceased is de-registered in the residence register. However, deaths of Austrian residents of foreign nationality, occurring on foreign territory are not subject to registration. The notice of death may reach the Austrian authorities only with delay.

265. The social security data base thus supplies complementary information on deaths and is therefore employed to refine the total population and migration count. These data are matched to the migration flows from the residence register via bPK,

[8] Including asylum seekers and refugees

[9] Foreign diplomatic personnel are exempt from registration.

[10] The social security data base covers i.e. commuters from abroad without residence in Austria but does not supply any information on persons who have never been insured, received pension payments, child support or any other social service.

an anonymised 27-digit key that allows connecting information from different administrative data sources.

2. Estimation of potential nominal members in the population register:

266. At the reference date of register-based censuses (the last being on 31st October 2011) people occurring only in the Central Register of Residence, but in no other administrative register are identified as suspicious cases necessitating further investigation. They are addressed to confirm their presence in Austria. Those not having responded are seen as nominal members and therefore excluded statistically from the population. Municipalities also get knowledge of the results and are asked to clean up their administrative records.

267. For the inter-censal period a similar exercise is undertaken annually. In this case, people are not addressed directly to confirm their presence, but rather a share of those identified as unique cases in the Central Register of Residence are statistically excluded from the total population. The likelihood of being a nominal member is determined by applying a logistic regression model[11].

268. The results of this exercise are integrated into quarterly population statistics once a year (upon publication of the final results). Deviations between quarterly population statistics and those identified as nominal members by the so-called mini register-based census are integrated in two ways:

i. by relocation abroad of non-recognised registered residents and

ii. by relocation from abroad of previously non-recognised residents.

269. The population adjustment for nominal members thus has an impact on the migration flows in a direct way. Potential nominal members not recognised in population stocks are assumed to have relocated abroad (and thus statistically counted as emigration). In turn, previously non-recognised residents can re-join the population stock if they show life-signs in other administrative records than the residence register in the consecutive year. In this case, a statistical record for re-entry to Austria (immigration) is created.

270. In contrast to directly linking information from a single other data source at the micro level (as in step 1), a different methodological approach is used: first, several data sources are consulted to identify registered residents who are potential nominal members, and second, statistical estimation techniques finally select the individuals who will no longer be recognised residents. Here, data linkage on the unit record level creating statistically produced emigrations and immigrations is only a downstream process in the estimation of nominal members from various administrative records.

[11] Detailed documentation in German:
http://www.statistik.at/wcm/idc/idcplg?IdcService=GET_PDF_FILE&RevisionSelectionMethod=LatestReleased&dDocName=073537

M. United States

1. Adjusting for uncounted deaths in the population register

271. The US Census Bureau produces annual estimates of net international migration flows (NIM) to and from the United States, which are used as inputs for national and subnational population estimates. Integration of migration data is limited, though some foreign and administrative data are used to calculate a few subcomponents of NIM. At the same time, there has been a major effort at the Census Bureau to better leverage preexisting administrative data to reduce costs associated with Census data collection, though this has yet to be applied to international migration data. Some future work in this area with regards to migration data has been proposed, but is still at a nascent stage.

272. Data integration only occurs at the additive level, in that estimates of NIM subcomponents are derived from different data sources and then combined to produce a final estimate. The NIM estimate produced by the US Census Bureau is made up of several subcomponents which are calculated separately: 1) Foreign-born immigration, 2) foreign-born emigration, 3) net migration to/from Puerto Rico, 4) net native-born international migration, and 5) net military movement. The foreign born and Puerto Rico components utilize the American Community Survey (ACS) and its Puerto Rico equivalent. These subcomponents make up the bulk of the NIM estimates, but we also add a constant net native subcomponent that is derived from analysis of US-born or US citizens measured from foreign data sources (primarily censuses). The final subcomponent is net military movement to and from the US, which comes from administrative data supplied by the Department of Defense.

2. Movement of the Armed Forces Population to and from Overseas

273. The net military movement component uses the estimate of the net overseas movement of the Armed Forces population from data collected by the Department of Defense, Defense Manpower Data Center (DMDC). DMDC provides monthly tabulations of military personnel stationed or deployed outside the United States by age, sex, race, Hispanic origin, and individual branches of service within the Department of Defense. We assume that changes in the overseas military population, excluding deaths, indicate movement of personnel into and out of the United States. To derive estimates of net international movement of the armed forces at the state and county-level, we primarily use DMDC data by age, sex, race, Hispanic origin, and county. To improve the geographic distribution of military movement around certain domestic military installations, we use county grouping information derived from the most recent ACS five-year file.

274. This figure, at both the national and subnational level, is simply added to the other NIM subcomponents to produce a final NIM estimates. This is the only type of integration of international migration data we currently conduct.

3. Data integration at the micro-level

275. Integration of international migration data at the micro-level is currently not done at the Census Bureau, though we have experimented with looking at country of birth data from the Social Security Administration file (NUMIDENT) and merging it with Internal Revenue Service (IRS) tax data, to analyze domestic migration patterns of the foreign born. We have also discussed the potential of linking administrative data (from the Department of State) on refugees to census or ACS data in order to help determine refugee status through probabilistic methods. Other administrative micro-level data exist at the Department of Homeland Security and their Office of Immigration Statistics (OIS). OIS has undertaken their own efforts to link various administrative datasets across Federal agencies, but at the US Census Bureau, we have only begun to think about possible ways to integrate migration data at these levels.

IV. Conclusions and recommendations

277. Migration statistics are probably the most difficult element in the field of social statistics, not only from an operational point of view, but also from a conceptual point of view. While it is true that there are international definitions, it is not always easy to measure the strict concept of a migrant. There are also a variety of sources that provide partial measurements of migrations.

278. The traditional method for measuring migration is indirect, based on the comparison of population figures obtained in two consecutive censuses. But obviously this method is very insufficient today, when the migratory movements happen to have a capital importance. In addition, more and more sources are available and can provide relevant, though sometimes partial, information on migratory movements (surveys, administrative records of all kinds). The challenge for statistical offices is therefore to be able to combine these pieces into reliable and integrated information on migrations.

279. The existence of many different potential sources for migration data has been assessed by the Task Force. According to the replies to the survey, most of the countries (31 of the 56 responses) use more than one source for generating migration statistics.

280. Trying to establish a taxonomy of methods or techniques of data integration for measuring migration in the international context is almost impossible. When analyzing the methods for producing migration statistics there are many different approaches because circumstances and sources are different in every country.

281. The concepts that are measured and the possible methods of integration that can be used are constrained by the main sources from which migration statistics are obtained. But there is a tendency to incorporate more administrative sources to measure migration in countries that base their population statistics on classical methods (surveys, exhaustive censuses) as well as in those that base their statistics on population registers.

282. In addition, the choice of sources for migration data depends on other elements. For example, geography greatly influences the methods for producing these statistics. Countries that are relatively isolated or with fewer points of entry and exit (New Zealand, Malta) can produce migration statistics based on very different sources than others with many access points, even without checkpoints, like countries in the Schengen Area in Europe.

283. Another important limitation when studying migration statistics is that there are different definitions of the migrant population in different administrative sources. In many cases, the standard time frame of 12 months of actual or planned stay is not the one used for the preparation of migration statistics but other similar concepts (e.g. other time periods). In addition, different administrative sources within the same country can offer information based on different concepts.

284. In some countries, migration statistics are intended to be fully consistent with stock changes (e.g. European Union countries). In many other cases, migration statistics do not pursue this full consistency between flows and stocks. Moreover, many countries produce separate data on long-term migrants, asylum seekers, non-permanent foreign population, short-term migrants, etc.

285. We can consider that a first step in data integration refers to the selection and use of different administrative sources to quantify different flows or stocks of different subpopulations. There are statistical offices that do not publish a single product "Migration statistics" but different statistics on the migration phenomenon under different methodological frameworks and concepts. In some cases these offices even compile these different data into “migration data reports” highlighting the importance and the richness of having different data on the phenomenon. The simple publication of data from various sources is a first mechanism for integrating information on migrants.

286. A first recommendation can be made in this sense. It is advisable to produce a number of statistical outputs that provide information on migration stocks, as well as inflows and outflows, covering different types of migration flows. In some cases different statistics may provide not fully consistent data but it can help to give a global picture and to increase awareness that a single product does not provide all the necessary information on international migration.

287. In terms of data integration, there are some relevant examples of combinations of data both at macro and micro levels. As regards to micro-integration, that is, a combination of sources at the individual record level, there are several examples that can illustrate the trend for the future. The combining of records may generate overlaps and these are addressed through adjustment and calibrations processes. This integration is taking place to improve the measurement of stocks and also to produce more complete statistics of the flows.

288. In general, the main weakness of the population registers is their difficulty in detecting emigration. If a person does not inform the authorities of their exit from the country, the administrative register has to be updated by administrative procedures that are not always easy or immediate, so the population registers may contain persons no longer residing in the country.

289. By combining sources, "life signals" or "presence signals" can be detected. Thus, a population register can be linked with tax or social security files to be more certain about the presence of people in the country. It is of particular importance the combination of different administrative records in cases such as Estonia, Latvia or Austria where logistic regressions are used for accounting resident population from evidences in different administrative registers.

290. There are also examples of countries working on the integration of different sources for the construction of a statistical population register. In some cases (e.g. Netherlands), already consolidated records are available, while in others, such as Italy, it is still in a design phase.

291. Another example of combining data to improve the accuracy of population register figures is the 2011 Population Census in Spain. Records were combined between the population register and a survey sampling 10% of population. This survey was also used to improve the figures on stocks of (national and) foreign people assigning likelihood of residence to doubtful records in the population register.

292. As for statistics on migration flows there are also some examples of integration in the strict sense, i.e., the combination of sources to obtain a single database. For example in cases like Hungary or Canada different sources are micro-

linked. This micro-integration is carried out by linking administrative sources on foreign people among them and with the population register to avoid overlapping and to provide additional variables.

293. Another interesting example of macro-integration is found in the United Kingdom, where a border passenger survey is combined, by calibrating, with another source (Labour Force Survey) to allow territorial breakdown of incoming flows. A pre-requisite for the use of mirror statistics will be the compatibility of definitions of the different types of migration flows.

294. Mirror statistics in the field of migration can be considered another example of integration. The comparison of national aggregate data on migration flows with the corresponding inverse flow data from other countries' national sources (databases of foreign population, migration figures) at least help to shed some light on the quality of information on migration in different countries. A further step is to use data from other countries to improve national figures. There are some examples of this practice (Romania, Poland).

295. An initiative that is beginning to be demanded from users and would be very promising is the exchange of individual data between countries for statistical purposes. The main implication of measuring migration flows is that they affect one country of origin and another of destination. So it is logical to think that the sharing of the statistical data will improve these statistics. The exchange of individual data for statistical purposes is likely to develop further, although it is necessary to overcome barriers, not just legal nor technical. But there are already examples of exchange between Nordic countries.

296. This picture of different sources and methodologies, on the one hand, shows how difficult it is to make particular recommendations on data integration; but on the other hand, showing that many different initiatives are taking place around the world, may encourage other countries to improve migration figures using sources at hand. The general recommendations seem clear: publishing different data may not be seen by users as a weakness of the statistical system but, on the contrary, as a proof of its richness.

297. Statistical offices should focus their efforts on collecting data from different sources to study the migratory phenomena, trying to know and exploit them thoroughly and to combine them into statistical information that is informative and relevant to the society. This is not an easy task and inconsistencies may show up, but the benefit for the society makes it worth the effort.

A. Metadata

298. The metadata can be ideally split in three parts: the first is specific to each data source, describing their main features and how their datasets are transformed before they are subject to integration; the second part looks at the way the datasets generated from the listed data sources are integrated; the third part would be based on some measures of quality, basically to prove its increased informative content due to integration. This information should be provided for each relevant migration statistics, i.e. for migrants stocks, migration flows, and any of their sub-groups of interest.

299. The first information that should be provided is the list of data sources used to produce the integrated data. In such a list, the following metadata could accompany each data source:

i. Name;

ii. Typology (administrative register, sample survey, census, big data, etc.);

iii. owner (including statistical authority of another country);

iv. Rules of access / data provision (if the owner is other than national NSI)

v. Population of reference;

vi. Date/period of reference of the data;

vii. Frequency of updating;

viii. Timeliness (i.e., the period of time between the date of availability of the data and their date/period of reference);

ix. Level of detail of the original data (micro, macro);

x. Level of detail of the dataset used for integration (micro, macro);

xi. Description of any method applied to transform the original dataset into an input suitable for data integration;

xii. Dimensions of the dataset (n x p);

xiii. The variables contained in the data source (or broad description if they are many);

xiv. The variables contained in the data sources that are retained for integration (possibly subset of the previous item);

xv. The variables used to integrate this data source with the other(s) (possibly subset of the previous item);

xvi. Availability of the PIN (this is in principle included in the previous item, but given its relevance it can be useful to dedicate a specific space to it) and its main features whether relevant.

300. Assuming that the integration process is implemented sequentially, for each step, i.e., for each pairs of datasets being integrated, the following information could be provided:

i. Step number;

ii. Datasets involved (using names or number from the list of individual data sources);

iii. Frequency of the integration procedure;

iv. Variables used for integration (if applicable);

v. Description of the methodology for integrating the datasets;

vi. Share of data overlapping;

vii. Main issues / difficulties in the integration of these datasets, such as reporting of non-linkage, methods and results from estimation of false-positive link rates;

viii. Dimensions (n x p) of the resulting integrated dataset.

301. In some cases, and particularly when migration statistics are derived from population registers, the number of data sources may be relatively high, and thus the reporting of the metadata can become burdensome. In such cases, it may be suitable to prepare just once a general report on the functioning of the system. However, it should also be taken into account that it may be important for the users to get a clear understanding on how the final data are produced through a possibly complex migration processing system. Complexity should not prevent transparency.

302. The third part of the metadata, which requires much further work, can be developed around the following initial set of quantitative measures of quality, aiming to highlight the positive effect of data integration:

i. Difference in the number of records / size of the covered population between the final integrated dataset and the smallest single dataset: n* - min(n);

ii. Difference in the number of variables between the final integrated dataset and the smallest single dataset: p* - min(p);

iii. Difference in the timeliness between the final integrated dataset and the less timely single dataset: t* - max(t);

iv. Difference in the number of records / size of the covered population between the final integrated dataset and the largest single dataset: n* - max(n);

v. Difference in the number of variables between the final integrated dataset and the largest single dataset: p* - max(p);

vi. Difference in the timeliness between the final integrated dataset and the most timely single dataset: t* - min(t).

vii. Gain in the quality of estimates, such as percentage decrease of the variance of the estimates, or any quantitative measure of the improved accuracy.

303. The metadata can be enriched by a qualitative assessment of the impact of integration, such as feedbacks given for the improvement of the original data sources and reduction of response burden, possibly supported by specific references / examples, and by an overall conclusion about the integration process.

V. Future work

305. The metadata structure provided in the previous paragraphs is only a preliminary input to a better worked version of the metadata which could benefit of further reflections and test applications as well.

References

ESSnet (2008): Project on Integration of Survey and Administrative Data. Material available at: https://ec.europa.eu/eurostat/cros/content/isad-0_en

ESSnet (2011): Project on Data Integration. Material available at: https://ec.europa.eu/eurostat/cros/content/data-integration_en

ESSnet (2014): Project on Handbook on Methodology of Modern Business Statistics ('Memobust handbook'), Module on Macro Integration. Material available at: https://ec.europa.eu/eurostat/cros/content/macro-integration_en

Eurostat (2009): "Insights on Data Integration Methodologies". Proceedings of the ESSnet-ISAD workshop, Vienna, 29-30 May 2008. Eurostat Methodologies and Working Papers. Available at: http://ec.europa.eu/eurostat/en/web/products-statistical-working-papers/-/KS-RA-09-005

Eurostat (2013): "Statistical matching: a model based approach for data integration". Eurostat Methodologies and Working Papers. Available at: http://ec.europa.eu/eurostat/en/web/products-statistical-working-papers/-/KS-RA-13-020

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

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