United Nations

ECE/CES/2018/2

Economic and Social Council

Distr.: General
27 March 2018

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-sixth plenary session
Geneva, 18–20 June 2018
Item 4 (b) of the provisional agenda
Data integration for measuring migration

Guidance on data integration for measuring migration

Note by the Task Force on data integration for measuring migration

Summary

The document presents a short version of the *Guidance on data integration for measuring migration*. The Guidance was prepared by a Task Force composed of Spain (Chair), Austria, Italy, New Zealand, Switzerland, United Kingdom, United States, Eurostat and UNECE.

The current short version of the Guidance is prepared for translation purposes and it excludes the country case studies. The Guidance includes: a discussion of the definition of data integration, with reference in particular to international migration statistics (Chapter II); a discussion of the results of the survey on data integration practices (Chapter III); summary results of the thirteen national case studies (Chapter IV); a discussion and list of metadata (Chapter V), the conclusions (Chapter VI) and a number of general recommendations for countries (Chapter VII).

The full text of the Guidance (including the country case studies) has been sent to all members of the Conference of European Statisticians for electronic consultation. It is available at: http://www.unece.org/index.php?id=47411. Subject to a positive outcome of the consultation, the CES plenary session will be invited to endorse the Guidance.

Contents

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. The UNECE High-level Group on Modernisation of Official Statistics (HLG-MOS) had a project on data integration in 2016–2017. The project included experiments with data integration in different areas and led to developing a practical online guide to data integration for official statistics.[1] Eurostat has supported research on methods to improve data integration, such as the ESSnet project in the area of Integration of Survey and Administrative Data.[2] 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 international 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 that may concern overall migration stock and flow data, and enhance the richness of migration data by adding socio-demographic or economic dimensions to existing data.

3. At the European level, the importance of data integration to improve migration statistics was reaffirmed on the occasion of the 2017 annual conference of NSI Directors General, which included a statistical session on ‘Population Movements and Integration Issues - Migration Statistics’.[3] The ‘Budapest Memorandum’ discussed at the conference and approved by the European Statistical System Committee includes among the action points: *“To support the identification, assessment and adoption of new methods and data sources, particularly the increased use for statistical purposes of administrative data sources of appropriate quality ensured through ongoing quality assessment – either single registers, linked data from several administrative sources or combined with survey sources, and the opportunities offered by new data sources (e.g. Big Data)”*.

4. 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). ‘Big data’ from utility and phone companies are also being considered in some countries as supplementary information. 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.

5. 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

[1] https://statswiki.unece.org/display/DI/Data+Integration+Home

[2] https://ec.europa.eu/eurostat/cros/content/data-integration_en

[3] 103rd DGINS Conference (Budapest, 20–21 September 2017). Papers and presentations are available at: http://www.ksh.hu/dgins2017/presentations.html

production of metadata to facilitate comparisons between migration estimates produced by different countries.

6. 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.

7. Based on these discussions, the Conference of European Statisticians (CES)[4] 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 work of the Task Force.

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 II. Chapter III provides an analysis of the results of the survey of migration data providers. Practical examples of data integration in migration are shown through in-depth case studies from a number of countries. The summary result of the case studies are presented in Chapter IV (the case studies are available in the full version of this document, in English only). A proposal on metadata to be considered with regard to international migration statistics is presented in Chapter V. Finally, the conclusions and general recommendations based on the survey and the case studies are provided in Chapter VI.

II. Definition of 'data integration'

A. General features

11. 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

[4] 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.

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).

12. 'Data integration' is often mentioned as one of most effective approaches for enriching and improving the statistical information. Various activities and projects over the last years dealt with data integration, and adopted different definitions. For instance, according to the Generic Statistical Business Process Model (GSBPM), data integration is an activity in the statistical business process when data from one or more sources are integrated. In the HLG-MOS project mentioned above, data integration is defined as "the activity when at least two different sources of data are combined into a dataset".[5] 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.

13. 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. In the case of migration, datasets 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.

14. 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.

15. 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.

16. 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.

17. 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

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

privacy and national security that countries may oppose any request of international migration microdata exchange. Examples of this latter approach do exist in the Nordic countries, where international microdata exchange is used to improve the quality of the population registers.

18. 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.

19. The reduction of response burden is sometimes mentioned as one of the opportunities offered by data integration. To date, most use of alternate data sources for migration statistics 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.

20. Inputs from alternatives data sources may be used also for the sake of validating the statistical output of a single, official data source. In this case, it can be considered as a matter of *comparison* rather than *integration*. This may also 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.

21. Another case is when each data source covers a specific population (migrant) sub-group. The pooling of these data to produce an overall statistical output could be labelled as 'compilation' rather than pure 'integration', as there is no actual merging at the level of the single record (either individuals or aggregated entities).

22. 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.

23. 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

24. 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"*.

25. 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.

26. For the purposes of this document, 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"*.

27. Data integration can be processed at two main levels of aggregation: micro- and macro-data integration[6], defined as follows:

(a) 'Micro-data integration': the integration of data based on record linkage/statistical matching of individual level records using key identifying variables; and

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

28. 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.

29. *"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.[7]

30. 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.

31. 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.

[6] In some contexts the terms 'micro-macro' is used when micro and macro level data are combined. This approach can be considered here as a case of macro-data integration.

[7] 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.

III. Survey on data integration practices

32. 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.

33. The questionnaire was forwarded in September 2016 to National Statistical Offices of UNECE member countries[8]. Fifty-six countries provided responses to parts of the survey that had relevance to their data collection practice.

34. The results of the survey were analysed between the end of 2016 and the first quarter of 2017. The main results are summarized in the following paragraphs.

A. Overview of main results of survey

35. Some form of data integration is essential for enabling statistical production of international migration flows. Where possible, data integration is done at the individual record level but this is mainly achievable for countries with complete population registers that account for exits of nationals and entries of foreign-nationals.

36. For many countries the actual migration event is not registered by a single source. Statistics of migration flows can only be derived by merging contributing components that account for changes in the country's resident population due to migration.

37. Statistics on international migration flows using a single source are produced when it is possible to maintain an administrative collection of the migration event (actual border crossings) or when a total population register based system is available. However, quality assurance and further disaggregation of these statistics may require integration with supporting data sources. This may just be in the form of an adjustment of migration flows following a comparison with available statistics from other countries (mirror statistics). Grouping of individuals in a single source may be required for longitudinal observations of border movements as a way of classifying migrant status.

38. The survey results highlighted an extensive variety in the uses of data integration. In some countries population registers of nationals may not be linkable at the unit record level to a separate administrative collection on foreign nationals. In such cases the population register may serve as a source for cross-checking and imputation of missing data.

39. A population census or a population register would be the common sources for producing statistics on the foreign-born population. The population census may serve as a base population and subsequent integration with an administrative collection enables the updating of the foreign-born population statistics.

40. When multiple sources are informing the migration statistics some countries have reported on the need for a prioritization assessment of the sources in terms of timeliness, completeness, reliability, coherence and other quality dimensions.

[8] 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.

B. Recommendations

41. As countries advance and standardize registers of social and population statistical datasets, data integration is more likely to take place at the unit record level across multiple sources. A greater attention to a quality assessment of the integrated data will necessarily become an integral part of the production of international migration statistics.

42. It is increasingly important for statistical offices to provide a transparent presentation of the rules and methods used for the data integration processes that produce the statistics of international migration flows and other types of international migration statistics. This opens up opportunities to develop methodological standards of data integration practices across the national statistical offices in the future. Individual countries will always have their own collection and data integration practice but there are likely to be more open technical forums for discussion of rules and methods used in the combining of sources for the production of migration statistics.

C. Summary of results

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

44. 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.

45. 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.

46. 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.

47. 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 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.

IV. Case studies

48. Taking into account the information collected in the survey on data integration practices, the task force identified a number of countries with significant data integration experiences. These countries were invited to provide a short text describing the primary migration data sources, the reasons why data integration is considered necessary, the methods used and the benefits of integration.

49. The texts provided by the countries are presented in the full version of the present document (available in English only) as a collection of significant practical examples of data integration on migration.

A. Summary of the case studies

50. The case studies show a broad range of situations that differ widely with regard to the sources available for migration statistics, the limitations of those sources, and the data integration approaches adopted. The examples provided by the case studies should not be considered as fully representative sample of the data integration approaches developed in the various countries. Moreover, when evaluating the experiences of the various countries, it should be kept in mind that national contexts may differ significantly with regard to a number of important aspects, and that a data integration approach adopted successfully in one country is not necessarily the best solution for other countries. Notwithstanding these limitations, the analysis of the case studies presented allows drawing some key points that could be useful for other countries.

51. With regard to *the countries with a population register*, in several cases (including Austria, Italy, Latvia, and Spain*) this register is the main source of migration data*. In these countries data integration is carried out mainly to improve the data quality. Data from other registers can be used, among other purposes, to adjust for missing de-registrations of emigrants or to identify deceased persons. In several countries data integration is also used to provide additional information on variables not included in the primary sources.

52. Several countries adopt practices based on the *'presence signals'* (in some countries called 'signs of life') of individuals in the various registers. In *Austria*, for instance, the analysis of the presence signals in other registers is carried out to produce an estimate of those who are in the population register but do not live anymore in the country. Data from the Austrian Social Security are used to identify deceased persons, and from various other administrative sources to adjust for missing de-registrations of emigrants.

53. In *Italy*, where only persons who are legally resident in the country can be included in the population register, the analysis of presence signals in the various sources allows improving the international migration estimates by identifying persons who are in the population register but have probably emigrated, and those who immigrated but are not in the population registers.

54. In *Latvia*, the analysis of the presence signals is carried out using various registers, including those on revenue, social insurance, education, health, and employment. Emigration estimates are produced also using mirror statistics from destination countries. Moreover, the quality of migration statistics is evaluated using data from household surveys, including LFS and EU-SILC.

55. In *Spain* the national population register ('padrón') is managed by the national statistical institute INE, and all persons living in the country – irrespective of their legal situation – have to be registered. In connection with the 2011 census, the registered population was linked with many other administrative registers in order to identify persons whose presence in the country was doubtful and those who were dead. For doubtful records, data integration techniques based on administrative data and a sample survey

carried out in connection with the census allowed estimating the number of persons who were in the register but were not actually living in the country.

56. In some countries (including Hungary, Israel, Netherlands and Switzerland), migration statistics are produced using *systems where the population register is one of several data sources*, not necessarily the most important one. In *Hungary*, a relatively complex system of registers is used, where the various sources are complementary to each other. Data on immigration and emigration are produced by integrating data among the various sources, in particular the Office for Immigration and Asylum and the Central Population Register managed by the Ministry of the Interior. Data integration allows producing high quality migration statistics and provides additional information to that available from the primary sources.

57. In *Israel*, data on migration flows and stocks derive mainly from border control registrations, and are provided by the Population and Immigration Authority (which maintains also the population register) to the Central Bureau of Statistics. The data are combined with other administrative files (including tax files, education files, etc.) and with population census and other statistical surveys to provide estimates of grater quality and add variables from other sources related to the same individual or events.

58. In the *Netherlands*, migration statistics (like all social and demographic statistics) are based on the System of Social Statistical Databases, resulting from data integration among over fifty administrative registers. It is considered that virtually all the population living in the country is registered, with few exceptions. Concerning emigration, about one third of those leaving the country do not de-register. When the authorities find that someone is missing an investigation is conducted and, if not found, the person can be registered as emigrated to an unknown country.

59. In *Switzerland*, three main registers are relevant for migration data: 1) the local population registers maintained by the municipalities; 2) a federal register for legally resident foreigners holding a permit of stay; and 3) a special federal register for foreign diplomats, members of consulates, international organizations, their family members and similar categories. A modest form of data integration is carried out by matching data collected at the federal level and those collected at local level. Moreover, some variables available only in the register for legally resident foreigners are added to the local population register records pertaining to the same observations.

60. With regard to the *countries without a population register*, the main sources of data on migration include passenger cards filled at borders, data on passports, visas and stay permits. As for the countries with population registers, data integration is carried out using various sources to improve the data quality and obtain more comprehensive information. In particular, data integration may help improving the information for immigrants on the actual place and date of settlement, which often differ from those stated at the time of arrival.

61. In *Australia*, data on arrivals and departures are compiled by the Australian Bureau of Statistics based on information derived from various processing systems including passenger cards, passport and visa information. Data are linked using a personal identifier which allows ABS to also generate travellers’ individual histories. Migrant data sets are also linked with census data and tax information.

62. In *Canada*, Statistics Canada receives administrative data on immigrants and non-permanent residents. Since the intended province and date of settlement not always corresponds with the actual place and date of settlement, data integration is carried out using other sources such as tax files. Record linkage is also carried out using data from other sources (such as the Canadian population census, the National Household Survey, the US Office of Migration Statistics) to provide estimates of great quality and obtain more comprehensive information on socio-economic characteristics of migrants.

63. In *New Zealand*, imaged passenger cards are linked to electronic border movements and passenger records to include additional information and provide validation. Moreover, for travellers who have indicated their intentions to stay in the country or leave it on a long-term basis, travel and passenger data are linked for the 16 months prior to the reference month, in order to confirm the status of New Zealand resident at the time of travel.

64. In the *United Kingdom* there are many sources of official statistics that provide information on different aspects of international migration flows and stocks. Concerning flows, the main source is the International Passenger Survey (IPS), a multi-purpose sample survey that collects information from passengers as they enter or leave the country. As the IPS does not take into account the changing intentions of passengers, in order to produce long-term international migration estimates data integration is carried out using data from IPS, LFS, Home Office immigration administrative data, and estimates of the flows at the land border of Northern Ireland (not covered by the IPS).

65. In the *United States* the Census Bureau produces annual estimates of international migration flows by adding various subcomponents derived from different data sources. For some subcomponents a limited integration (at the aggregate level) is carried out using some foreign and administrative data, including the American Community Survey, foreign censuses, and military movements from the Department of Defense. Integration of international migration data at the micro-level is not done at the Census Bureau, although some experiments were carried out.

66. To summarize, the case studies show that virtually all countries covered use data integration to improve the quality of migration statistics, and to provide additional information to that available from the primary sources.

67. In countries with a population register, data integration – also through the analysis of the ‘presence signals’ – may allow identifying: 1) persons who emigrated but are still in the population register; 2) persons who immigrated but are not in the population register, and 3) deceased persons.

68. In countries without a population register, data integration may help improving the information for immigrants on the actual place and date of settlement, which often differ from those stated at the time of arrival.

69. The case studies showed many examples of data integration conducted at micro or macro level. 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. On the other hand, an 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.

70. As expected, many different types of sources are used for data integration, including various administrative or statistical registers, other administrative data (including border control data), and statistical surveys (including censuses). Few countries use ‘mirror statistics’ from destination countries to improve their emigration estimates. This is done mainly at the aggregate level, although there are examples of international exchange of microdata, namely among the Nordic countries (not included in the case studies), or restricted to neighbour countries.

V. Metadata

71. When migration estimates produced by different countries are compared, it is very important to have access to comprehensive metadata, in order to assess in detail the comparability of the data.

72. 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.

73. 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:

(a) Name;

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

(c) Owner (including statistical authority of another country);

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

(e) Population of reference;

(f) Date/period of reference of the data;

(g) Frequency of updating;

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

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

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

(k) Description of any method applied to transform the original dataset into an input suitable for data integration;

(l) Dimensions of the dataset (n x p);

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

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

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

(p) 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.

74. 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:

(a) Step number;

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

(c) Frequency of the integration procedure;

(d) Variables used for integration (if applicable);

(e) Description of the methodology for integrating the datasets;

(f) Share of data overlapping;

(g) 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;

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

75. 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.

76. 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:

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

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

(c) Difference in the timeliness between the final integrated dataset and the less timely single dataset: t* – max(t);

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

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

(f) Difference in the timeliness between the final integrated dataset and the most timely single dataset: t* – min(t);

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

77. 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.

VI. Conclusion

78. 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. Moreover, there are increasingly different forms of migration, and there are is a variety of sources that provide partial measurement of migration.

79. The challenge for statistical offices is to combine data from the different sources into reliable and integrated information on migration. The 13 case studies showed the different ways and data sources that countries use to improve their migration statistics.

80. The Task Force conducted a survey to assess the potential of many different sources for producing migration statistics. According to the replies, most countries (31 of the 56

responses) use *more than one source for generating migration statistics*. Approaches differ because the circumstances and data sources vary greatly across countries. The overall tendency to involve more administrative sources in measuring migration applies both to countries that base their population statistics on classical methods (surveys, exhaustive censuses) as well as to countries that base their statistics on population registers.

81. The choice of sources for migration data often depends on *geography*. Countries that are relatively isolated or with fewer points of entry and exit (e.g. New Zealand, Malta) can produce migration statistics based on very different sources than others with many access points or without checkpoints, like countries in the Schengen Area in Europe.

82. 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 concepts (e.g. other time periods) are adopted. In addition, different administrative sources within the same country can offer information based on different concepts.

83. In the European Union countries, migration statistics are intended to be *fully consistent with changes in migrant stock*. 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.

84. 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.

85. 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.

86. By combining sources, *'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 in the combination of different administrative records in cases such as Latvia or Austria where logistic regressions are used for estimating resident population from evidence in different administrative registers. The analysis of the 'presence signals' may also help identifying persons who immigrated but are not in the population register (in Italy), and deceased persons (in Austria and Spain).

87. As for statistics on migration flows, there are some examples of integration in the strict sense, i.e., the combination of sources to obtain a single database. For example, in Hungary and Canada different sources are micro-linked. This micro-integration is carried out by linking information on foreigners across different administrative sources and with the population register, to avoid overlapping and to provide additional variables.

88. An initiative that is beginning to be demanded from users and would be very promising is *the exchange of individual level 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. Nordic countries provide a good example of such exchange.

VII. Recommendations

89. Based on the work carried out by the Task Force and the content of this document, a number of recommendations to national statistical offices are presented in this section. These recommendations are considered relevant to all countries, regardless of their current situation with the availability and use of administrative data for migration statistics.

A. Improve access to administrative data for national statistical offices

90. In countries where administrative records are not or only at some extent available to the statistical offices to produce official statistics, a priority is to take measures to facilitate the access to administrative records and generalizing their use by the statistical office. This may require support at the political level and, in some cases, changes in the legislation.

B. Use administrative data for migration statistics

91. Many administrative sources contain information that can be used in migration statistics. These include registers and databases on population, visas, passports, border control, residence and work permits, and asylum seekers and refugees. It is recommended to examine the potential of all these administrative sources for migration statistics and use them for the purpose of migration statistics to the extent they are fit for this purpose. This requires cooperation with relevant ministries, agencies and administrations responsible for the data.

92. It is advisable to produce several 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. However, their publication 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. The simple publication of data from various sources is a first mechanism for integrating information on migrants.

C. Combine data from different sources using ‘presence signals’

93. In those countries where administrative data are available for statistics, the combination of sources using ‘presence signals’ methods is very promising. The point in common to those different approaches is the usage of several sources to estimate a probability of an individual’s residence in the country based on their presence in these sources. Combining data from different sources in such way allows to improve estimates of current population and stock of foreigners.

D. Pay attention to the quality of integrated data

94. As countries advance and standardize registers of social and population statistical datasets, data integration is more likely to take place at the unit record level across multiple sources. A greater attention to a quality assessment of the integrated data will necessarily become an integral part of the production of international migration statistics.

E. Be transparent about data integration methods used and develop standards

95. It is increasingly important for statistical offices to provide a transparent presentation of the rules and methods used for the data integration processes that produce the statistics of international migration flows and other types of international migration statistics. This opens up opportunities to develop methodological standards of data integration practices across the national statistical offices in the future. Individual countries will always have their own collection and data integration practice but there are likely to be more open technical forums for discussion of rules and methods used in the combining of sources for the production of migration statistics.

F. Promote international comparison and exchange of migration data

96. Another source of improvement can be found in the international comparison and in the exchange of data on migration, foreign born population and foreigners. Administrative sources typically underestimate emigration due to the low motivation of emigrants to register the event of moving out of the country. However, immigration data of countries where the emigrants go could cover these events. This can be checked using the so-called ‘mirror statistics’ that show the migration flow from country A to country B through the lenses of emigration statistics of country A and immigration statistics of country B. Furthermore, data from destination countries can provide information on the number and characteristics of emigrants and citizens of a country of origin. Exchanging data with other countries can thus be very useful for understanding migration processes and for improving their systems of measuring migration.

97. International data sources are very helpful for this. In the European case, Eurostat publishes quite detailed data of inflows and outflows that allow this comparison between countries of the European Union. In its Clearing House (database) of Migration Statistics, UNECE publishes such data from countries of Eastern Europe, Caucasus and Central Asia.

98. A breakthrough could be achieved if the databases were shared between countries to link at the individual level. There are already real cases like those of the Nordic countries. Today it is unlikely that exchanges of individual level data could be generalized, but there might be some initiatives between countries due to the intensity of their exchanges or others circumstances. The success of these exchanges could encourage others to follow the same path.

VIII. Further work

99. At the international level, further work could address the potential of using big data for migration statistics and the ways how the analysis of longitudinal data can be incorporated into annual migration statistics and complement the available cross-sectional measurements.

A. The potential of big data

100. The new and non-conventional data sources, such as mobile phones, credit cards and social networks — generally known as big data — could potentially be useful for migration statistics. Currently, statistical offices have great difficulty in accessing these data, which are in most cases privately owned. In the coming years, steps should be taken to gain access to these sources and find ways to use them for the purpose of official statistics. This information alone would probably be insufficient and biased, but when integrated with conventional sources it could improve migration estimates. It would be important to share

the emerging practices internationally, to support countries’ first steps towards using the potential of such data for migration statistics.

B. Longitudinal data

101. Probably the future of data integration is not only in the improvement of quality in the measurement of migratory flows. The integration of data may allow new data on the migrant population to be available. The possibility of connecting population databases (or surveys) with files for employment, social security, taxes, health, education systems, among many others, will be able to provide more and better information in the future on living conditions or the social integration of the migrant population.

102. A clear example is found in longitudinal studies information. Official statistics focus almost exclusively on providing cross-sectional information on migrants. For example, there is barely any information about the duration of the migratory episodes (does a migrant live in the host country for a long time?) Or about very relevant time intervals when studying migratory flows: the time elapsed since arrival in the country of destination until obtaining a job, or between losing a job and emigrating. Longitudinal data are needed for addressing the question of whether migrant socio-economic outcomes improve or worsen over time. They are also needed to understand family migration and emerging patterns such as circular migration.

103. Sharing of good practices in creating datasets from the longitudinal perspective and using them in regular migration statistics would support NSIs in their work. Developing recommendations on how to harmonize such statistics internationally could be the subsequent step.

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.