United Nations ECE/CES/2023/6

Economic and Social Council

Distr.: General
11 April 2023

Original: English

Economic Commission for Europe

Conference of European Statisticians

Seventy-first plenary session
Geneva, 22–23 June 2023
Item 4 (b) of the provisional agenda
Coordination of international statistical work in the
United Nations Economic Commission for Europe region:
Hard-to-reach groups in administrative sources

Hard-to-reach groups in administrative sources

Prepared by Denmark with contributions from Canada, Italy, New Zealand and United States of America

Summary

This in-depth review deals with the concepts and ways of capturing hard-to reach groups in administrative sources. The document summarizes the experience of national statistical offices in accessing hard-to-reach groups and describes problems and challenges. It also proposes further possible work to improve the access to hard-to-reach groups using administrative data. The last section summarizes the discussion by the Bureau of the Conference of European Statisticians at its meeting in February 2023.

As an outcome of the review, the CES Bureau supported further work in this area and decided to establish a new task force, as recommended in the paper. Denmark will chair the task force.

The Conference is invited to endorse the outcomes of the in-depth review, discuss the main findings and recommendations identified in this paper, and provide further input to the work of the new task force.

I. Executive summary

1. This in-depth review was mandated by the Bureau of the Conference of European Statisticians (CES) and deals with the concepts and ways of capturing *hard-to reach* groups in administrative sources. This note summarizes the experience of national statistical offices in accessing *hard-to-reach* groups and describes problems and challenges. It also proposes further possible work to improve the access to *hard-to-reach* groups using administrative data.

2. The drive to 'leave no-one behind', enshrined in the 2030 Agenda for Sustainable Development, has come to provide the backdrop for many efforts at improving the quality and coverage of statistics. Calls abound for multiple disaggregation of data, to permit analysis of the impacts of intersecting vulnerabilities. Yet many of the dimensions along which users would like data to be disaggregated concern population groups that are considered *hard-to-reach*: that is groups that tend to be under-represented either because they are numerically very small; because they are hard to identify, e.g. due to a lack of standardized definitions or because of a lack of data collection on the relevant variables; because they choose not to be identified, e.g. due to stigma associated with group membership; because they are systematically excluded from standard collection techniques and sampling frames, e.g. people living in institutions; because they are physically hard to reach, e.g. those living in remote areas or without a fixed abode; or because they are hard to enumerate even once identified and sampled, e.g. people living with dementia, people who do not speak the national language, and young children.

3. The generalized shift towards use of administrative sources for censuses and surveys heightens some of the challenges encountered when trying to safeguard and improve the availability of data and statistics on *hard-to-reach* groups. Some examples include: ensuring inclusion of the primary and secondary homeless and undocumented migrants in statistics derived from administrative registers; representing marginalized ethnic, religious and gender minorities and people with disabilities in statistics if administrative sources are not able to or do not routinely capture these characteristics; and producing sex-disaggregated statistics to permit gender analysis of crucial topics, when the administrative sources used to provide the information are gathered at the level of the household rather than the individual.

4. Five countries (Canada, Italy, New Zealand, USA and Denmark) have contributed to the work on this paper. Even though the statistical systems in those countries are different, they all face similar challenges in identifying *hard-to-reach* groups in their administrative registers.

5. *Hard-to-reach* populations in administrative data can be interpreted in different ways. One of the interpretations is incompleteness of administrative registers, making some groups, such as children or elders *hard-to-reach* and hence describe with data. This approach is described by the cases of Canada and the US. Another interpretation is selecting some groups, such as homeless, illegal immigrants or indigenous people and then trying to capture them in administrative records. This approach is described by Denmark, Italy, New Zealand and the US.

6. The current landscape of using administrative registers in order to access *hard-to-reach* groups shows few practices that allow to identify specific *hard-to-reach* groups in administrative registers but their application for statistics varies from country to country. Data on *hard-to-reach* groups can in some instances be retrieved from administrative records, mainly supported by surveys in the particular field, but currently there is not much evidence that such data is being used in production of regular, law bound statistics. Furthermore, the examples discussed show that identification of hard-to-reach populations is depending on a country context but is often supported by surveys or censuses that try to capture attributes, such as ID number, address, date of birth helping in identification of hard-to-reach populations in administrative registers.

7. It is recommended to establish a task team to investigate whether it is, at the current stage, possible to implement work aiming at improving the access to *hard-to-reach* groups in administrative registers.

II. Introduction

8. Across all areas of social statistics as well as various areas of economic and business statistics there is a widespread and rapid trend towards the use of administrative sources either to complement traditional census and survey sources, or to replace them. Some countries are already using administrative sources to conduct a full census. There has been a stronger focus on using administrative sources for statistics since the adoption of the 2030 Agenda, which again has been accelerated by the Covid-19 pandemic. There are many advantages to this overall trend, including reduced respondent burden, faster production of statistics, and the concomitant reduced costs. There are also well-known drawbacks, such as a dependence on the content of administrative sources and the challenges in gathering data on subjective characteristics or other variables that are not recorded in administrative sources.

9. During the pandemic, the NSOs have had to seek out administrative sources to make up for the absence or shortcomings of data collected by traditional means. Simultaneously, the emerging data demands and the 2030 Agenda call for a better disaggregation of data both to have a better insight into different segments of population and also to ensure compliance with ‘Leave no one behind’, the inherent principle of the 2030 Agenda. The provision of information on various population segments is challenging from statistical point of view, as it can be difficult to get ‘survey’ information about those segments, both in administrative and physical/geographical, sense. This, taken together with a fact of constantly falling survey response rates makes it difficult to provide sufficient evidence for groups that are *hard-to-reach.*

10. This paper focuses on describing current initiatives aiming to provide a better information on *hard-to-reach* groups from administrative data. It provides information on what groups are currently considered as *hard-to-reach*, describes a selection of ongoing practices and proposes possible next steps. It is not intended as an all-encompassing review of approaches to representation of *hard-to-reach* groups in statistics, nor of the importance of disaggregation for the Sustainable Development Goals (SDG) but it focuses on administrative sources, with an emphasis on presenting possible approaches.

11. Finally, an important distinction on *hard-to-reach* groups in the context of administrative data has to be made. *Hard-to-reach* populations can be interpreted in two ways:

(a) Groups that are *hard-to-reach* in any statistical context, i.e. in survey or administrative data, such as homeless, illegal immigrants, etc.

(b) Groups that are *hard-to-reach* in administrative registers due to underreporting in consequence of a time lag in reporting, such as children, highly mobile population segments (youth) or elders.

This paper describes both cases.

III. Scope of the statistical area covered

12. The concept of *hard-to-reach* populations from a statistical perspective is typically due to the fact that many standard survey sampling techniques are difficult, or often fail, since target populations cannot be accessed through frames based on traditional data sources, such as a list of dwellings. For example, if members of a target population are rare or stigmatized in the larger population, it may be expensive and/or difficult to contact them using traditional probabilistic approaches. As an alternative, administrative data offers the potential to improve frame coverage for some target populations, but may also lead to other *hard-to-reach* or “hidden” populations for various groups of interest.

13. However, a cross-cutting identification of *hard-to-reach* groups and their coverage by administrative sources is not a straightforward task, as the groups can either be unreachable by administrative sources or the access can vary between countries. The definition of *hard-to-reach* groups can also vary between countries in spite of international attempts to provide a uniform definition. The reasons why members of a population group are *hard-to-reach* can vary according to the context of each national, geographic, or social environment.

14. In the context of administrative sources, the *hard-to-reach* groups are usually not covered or easily identifiable and need to be complemented by some form of surveys in order to get a broader background for statistics/analysis. Linking the two types of data (survey and administrative) can be challenging as there could be a need for an ‘identifier’ to make it possible. *Hard-to-reach* groups are in fact *hard-to-reach*.

15. Mapping of identification of *hard-to-reach* groups and their coverage by administrative sources is a challenging task. When it comes to concrete data, there is a little if any overlap between data administrative from sources and data describing *hard-to-reach* groups. Most overlaps are created by survey data that capture a unique identifier of a person in question, hereby allowing for linking it to administrative registers. There is though a wide array of possibilities where data cannot be used in a convincing way. An example here could be a person that has been assigned a housing but due to, say, mental problems is living as a homeless. In such example, the person will not be captured by administrative data but maybe by a specific survey mapping the homelessness.

IV. Country practices

16. The information presented below is a summary of the responses provided by National Statistical Offices from Canada, Italy, New Zealand, USA, and Denmark on the subject. Identifying countries for this paper built on a subjective knowledge of country application of administrative registers in the production of statistics.

17. The main ‘nerve’ of data from administrative registers implies that they cover a total of a very big subset of population. However, if particular groups are not captured in administrative registers they cannot be described by data from this source. On the other hand, if captured, individual records can open up for a broader statistical analysis of the group.

18. The paragraphs below describe experiences from selected countries on how to identify *hard-to-reach* groups in administrative registers

A. Canada

19. There are some individuals and population groups considered “hard-to-reach” in a traditional census enumeration design. For the discussion below, we present examples noted during research into the use of administrative data in the Canadian Census of population. Canada does not currently have a population register, so our research on the increased use of administrative data for the census relies on linked administrative records. Generally, administrative data often begins with individual-level information, so persons of interest could be sampled from a more targeted person-level frame, as opposed to identifying individuals through dwelling-based surveys.

20. In the Canadian context, we could add First-nations people living on-reserve, people living in remote or Northern geographies, and individuals living in collective dwellings (i.e. nursing homes) as *hard-to-reach*. In cases where members of these population groups use public services or interact with organizations that collect information, administrative data offers the potential to help enumerate some individuals and population groups that are hard-to-reach using traditional census enumeration.

21. Collection disruptions due to a natural disaster or a pandemic can also create unplanned *hard-to-reach* populations. For the 2021 Census, Statistics Canada developed a statistical contingency plan based on the secure, responsible, and appropriate use of administrative data to support its collection in the case of such a disruption. While the 2021 Census enumeration was a success, some collection units (a census collection geography) had lower-than-expected response rates due to the COVID-19 pandemic, among other challenges. There are approximately 49,000 collection units in Canada. From those 1,045 of those collection units had low response rates, along with good administrative data available, thereby making them eligible for imputation using administrative data. For non-responding dwellings in those collection units, age, sex at birth, and number of usual residents were

imputed. This corresponds to a small proportion of the Canadian population, but Statistics Canada is planning an increased use of administrative data for the 2026 Census.[1]

22. The increased use of administrative data for census and survey programs might require a reconceptualization of the concept of *hard-to-reach* individuals and populations. Unlike in a traditional census, *hard-to-reach* populations in administrative data are not always conceptualized according to shared characteristics of the members of that population, such as their ethnic, geographic, or socioeconomic characteristics. Instead, new types of *hard-to-reach* populations emerge from the perspective of administrative data coverage, which might or might not coincide with *hard-to-reach* populations in a traditional census. We note below some examples of new types of *hard-to-reach* populations from the Canadian context.

23. First, in Canadian context, an administrative data source might be an incomplete sample of the overall eligible population, as the data was derived for administrative purposes as opposed to statistical usage. This raises the question of who appears or does not appear in administrative data sources at a given point in time. For example, research using administrative data in the 2021 Canadian Census showed delays in receiving vital statistics information close to Census day of May 11, 2021, which led to an over-representation of individuals aged 80 years and older, along with an under-representation of children less than one year old.

24. Second, conceptual differences in the types of information collected in census/survey programs versus the types of information available in administrative data might create new *hard-to-reach* groups. In Canada, as in other countries without a population register, differences in concepts between administrative information and survey information often arise. For instance, while spousal and dependent information is well-established in many administrative sources, other relationships such as common-law may be under-represented in administrative data since tax and traditional Census concepts may differ. We also note less representation of single-parent families in administrative data compared to the traditional Census. Any such attribute of an individual or a population that is not collected nor well-defined by an administrative record makes this a *hard-to-reach* population in administrative data.

25. Third, the concept of *hard-to-reach* in administrative data can include obfuscation of an individual's information due to multiple conflicting records from different linked administrative data sources. In Canada, probabilistic record linkage is used to link multiple administrative records together, followed by statistical models to identify single individuals for Census enumeration. While not necessarily related to the statistical aspects of *hard-to-reach* populations, in countries without a population register, it becomes challenging to assign an address or a geography to some individuals within the administrative data sources. For this situation, the real-life individual is not necessarily *hard-to-reach*, but de-duplicating and identifying single individuals across linked records might be difficult. Conflicting administrative records could refer to: (a) two different people; (b) a single person at different points in time; or (c) a single person at a single point in time who has multiple addresses and phone numbers? In the Canadian context, research into statistical integration methods is progressing, using small area estimation techniques and hierarchical linear models which combine survey and administrative data to mitigate these difficulties.

26. <u>*Hard-to-reach* age groups</u>: 0-1 years, 18-24 years, and 80+ years: Three distinct age groups emerge as *hard-to-reach* in administrative data for different reasons. Infants age 0-1 are under-covered due to a lag in receiving and integrating birth administrative data in a timely manner. Individuals aged 18-24 years are a highly mobile population, thus their administrative sources of geographic information are often conflicting, and assignment of a usual place of residence for these individuals is challenging. For individuals over 80 years of age two issues are of note: first, an overcoverage due to delays in receiving death records for certain individuals, and secondly the incorrect enumeration within administrative data since older adults age 80+ may reside in long-term care homes, but be enumerated as usual residents by a partner still residing at home. Moreover, the spouses, accountants, or adult

[1] More information on the statistical contingency plan can be found at the following link: https://www12.statcan.gc.ca/census-recensement/2021/ref/98-304/2021001/app-ann1-7-eng.cfm.

children of these individuals often assist with filing taxes and other government administrative duties, and therefore these older adults are sometimes erroneously situated at the location of the tax filer (e.g. adult child), and not at their usual place of residence (e.g. collective dwelling).

27. *Hard-to-reach* emigrants: Canada does not require citizens or residents to register when they leave the country. The vast majority of Canadians who leave the country return within a few days or weeks, but a small proportion will reside abroad long-term, yet retain fiscal ties to Canada. Emigrants from Canada are therefore difficult to detect in administrative data, because of the high number of false-positive cases of Canadians who are out of the country during the census, but who reside in Canada and return (or intend to return) during the year. This is also an issue for traditional census collection, and follow-up of emigrants using administrative data is considered for coverage studies.

B. Denmark

28. Denmark has an extensive system of administrative data that constitutes the backbone of the statistical system. The system builds on three main registers: population register, dwelling register and business register. The registers get frequently updated via automated data transfer system so the data show the latest status available. Via unique identifier, the system allows for linking the data making producing statistical time series across different domains possible. Furthermore, Denmark is one of the countries where a clear majority of social statistics is produced by the use of administrative data. Every citizen has attached a personal number for administrative purposes. The information in the personal number covers many dimensions, such as sex, age, and address. The personal number is being used as a unique identifier in linking the data to other statistical domains, such as education, health, income, employment etc. This usage gives a detailed information on various population groups in an anonymized way. Furthermore, Denmark is conducting census against the background of administrative data.

29. When the picture comes to *hard-to-reach* groups they are however not so easily identifiable in the administrative registers as there can be a myriad of situations where the data does not provide information on the recent status of a person in question. For example, there are situations where homeless people can have an address or a person does not have an address but is not being considered homeless.

30. Some subgroups of homeless people can be captured by shelter and care home statistics, a statistical survey conducted by Statistics Denmark. Here, the persons in need of a sleepover in shelters and care homes are, in majority of cases, asked to provide their personal number. This information can then serve as an input to statistics. The challenging part is here that it is not known is what share of homeless people is looking for a sleepover in shelters and care homes. This, consequently, makes it difficult to provide any form of a mark-up for general statistics. The Danish Centre for Social Science Research conducts a biannual mapping of the homeless people in Denmark and Statistics Denmark is, via the shelter and care home statistics contributing to this mapping. However, no further statistics on homeless people are being produced with this mapping as an input source. Another complicating factor is that voluntary organizations providing assistance to the homeless people often do it on a condition that it is not obligatory to collect personal number of the persons in question.

31. Illegal immigrants in their nature cannot be captured by administrative registers. In some cases, they may be given a temporary identification number mainly for administrative purposes. This number does not contain other information than age, sex and country of origin and cannot serve for statistical purposes as it is usually given once and if the same immigrant somehow again gets in contact with authorities, he/she will be assigned another temporary identification number, without a link to the previous one. Only in very sporadic cases, when an illegal immigrant is recognized with certainty, he/she will be assigned the same identification number as previously, which theoretically could give a background for a short series of information on the immigrant. However, this is not a sufficient base to produce statistics, not even on experimental basis.

32. Administrative registers open up for different possibilities of identification of *hard-to-reach* groups, however the possibilities require an extensive effort. As production of statistics on *hard-to-reach* populations is not legally required neither in Denmark or European Union, the full range of possibilities of producing statistics on *hard-to-reach* populations, such as linking information from surveys to administrative registers is yet to be explored.

C. Italy

33. To replace the decennial census, in 2018 the Italian National Institute of Statistics (Istat) launched the Permanent Population and Housing Census (PPHC), according to Istat modernization program, which places the integrated system of statistical registers at the core of statistical production. At the core of the PPHC is the Population Register (hereafter, using the Italian acronym, RBI), whose main sources are the local population registers of Italian municipalities, while two sample surveys (Areal survey and List survey) are conducted annually to support registers, in the broad sense of assessing their quality and to add information that is missing, incomplete or of insufficient quality. This allows the yearly availability of census statistics.

34. As to the population count, it was first obtained by applying correction coefficients for undercoverage and overcoverage errors to individuals classified as usual residents in RBI (the capture-recapture model was adopted for direct estimates of the coverage errors of RBI, with the population register representing the ‘first capture’ and field data being the ‘second capture’). In 2020, following the cancellation of the surveys due to the Covid-19 pandemic, a different methodology was used, entirely based on the use of ‘administrative signs of life’ (Sol) and the application of classification criteria to statistical registers. More precisely, usual residents (in RBI) with no SoL according to other sources were considered RBI overcoverage, while individuals not recorded in RBI as usual residents but with SoL in other sources were identified as the RBI undercoverage. This correction was then applied at the micro level, operating through the reclassification of individual records in RBI.

35. This obliged push towards a larger use of administrative data has called for a rethinking of the statistical framework for the quality assessment of the estimation processes of the PPHC and, more generally, of the PPHC design, with survey data used for the quality measurement of a, from now on, fully register-based population count estimation. An audit survey is planned to be conducted every 2-4 years to provide quality measures of the register based population size estimation, while a small-scale areal survey could be performed in order to evaluate the undercoverage of administrative sources which, by definition, don’t include populations such as undocumented migrants. At the same time, Istat is working to improve the use of SoL in the new cycle of the PPHC (post-2021) and to take into account also the misplacement error of the population register, which has not been evaluated so far. Furthermore, the acquisition of new sources (e.g. utilities archives such as energy consumption/smart meters data that will most likely provide objective assessment elements with regard to the actual place of usual residence) should represent a turning point.

36. Within this general design, some population subgroups have been distinguished, whose peculiarities require the adoption of a dedicated approach and which have therefore been excluded from the scope of the above mentioned census surveys. These *hard-to-reach* groups are: homeless and people without a fixed abode who, even in conventional censuses, represent a typical "hard to count" population due to the high risk of undercoverage; people living in institutions and people living in formal/informal/unauthorised settlements.

37. Therefore, in order to estimate the size of these population subgroups, a separate survey on *hard-to-reach* groups’ addresses is being conducted yearly on the local population registers. More precisely, the addresses corresponding to institutions, homeless[2] and

[2] Indeed, according to Italian legislation, homeless and people without a fixed abode have the right to be registered in the local population registers, and to this purpose Municipalities have to use fictitious addresses. In some cases addresses of NGOs providing assistance to homeless, migrants and people in need are also used.

formal/informal settlements are identified within RBI and submitted to the validation of municipalities' census offices through an online platform. For each confirmed address (including the newly added), the number of individuals by sex and citizenship is requested to be filled in. RBI is then updated based on such information, i.e. individuals belonging to the three groups are identified in RBI through their addresses and flagged so that basic data on these groups can be derived from the population register.

38. This process allows to obtain a more precise count of the *hard-to-reach* groups within the register-based population count. Individuals belonging to the registered segments of these groups were identified in RBI and included in the population count as such i.e. were not subject to the application of correction coefficients (2018-2019 count) or to reclassification based on SoL (2020 and on census counts).

39. Nevertheless, especially for what concerns the homeless, there is obviously only a partial overlap between the registered population and the actual one. Many primary homeless (especially undocumented migrants) are not registered in any municipality, while others are not registered as such (as they are still officially registered as members of their former households); at the same time, some of the people registered at the fictitious addresses are nomads or people who don't have a place of usual residence because of their job.

40. Given the known under-representation in administrative data of primary and secondary homeless and, more generally, the issues related to their identification, and based on the need to collect further data besides the basic ones that can be derived from the variables available in the register (i.e. the localisation and the sex, age and citizenship distribution), two *ad hoc* field surveys on homeless and people without a fixed abode are currently being planned. A survey on users of canteens, dormitories and other basic services provided to homeless and people living in extreme poverty conditions will be conducted in a selected number of municipalities across the country (the ones with the highest incidence of homeless population); while a point-in-time survey will be carried out in the 14 biggest municipalities. The latter will be aimed at producing a quantitative estimate of street homeless, and possibly at reconciling the data collected on the field with the figures related to the registered homeless, while the former will collect more qualitative data on the socio-demographic profile and living conditions of people in extreme poverty.

41. Furthermore, due to some quality issues, the next wave of the administrative survey on *hard-to-reach* groups will be conducted at the household code level for homeless and individuals living in formal/informal settlements. Indeed, if for people living in institutions, the address information is sufficient to identify with a high level of accuracy the target population in the register, it is not the same for the other two aggregates. Therefore the household codes (instead of the addresses) will be submitted to validation through the online platform, so that it will be easier to identify the relevant population in RBI in cases when the aggregate data declared by the municipalities do not coincide with the ones derived from RBI. Finally, the risk of undercoverage for people living in institutions (e.g. older adults residing in long-term care homes) has to be mentioned, as in many cases (for fiscal or other reasons) they are still registered at the address of their former homes. This issue will be addressed within the wider problem of misplacement mentioned above.

D. New Zealand

42. Stats NZ's Integrated Data Infrastructure (IDI) provides access to de-identified linked microdata for researchers. The Integrated Data Infrastructure includes data from a range of government agencies, Stats NZ surveys and the Census of Population and Dwellings. The data includes birth and death registrations, and international border movements, and for topics including education, work, income, benefit payments, justice, and health, gathered from a range of government agencies.

43. There is no national population identifier in New Zealand, and no national population register designed for administrative purposes. To facilitate data integration in the Integrated Data Infrastructure, a central list of population identities (the Integrated Data Infrastructure spine) is constructed by combining three high quality data sources: birth registrations, tax registrations and visa applications. The data sources are linked through a pair-wise

probabilistic linking. The Integrated Data Infrastructure spine is designed to include as far as possible all those who have ever lived in New Zealand, including those on work or study visas, and in 2021 included around 10 million individuals (the current population of New Zealand is around 5 million people). An encrypted identifier is assigned for each identity in the Integrated Data Infrastructure that is common across all datasets. All other data sources are then linked to the Integrated Data Infrastructure spine.

44. The Integrated Data Infrastructure processes derive basic demographic characteristics - age, ‘sex or gender’, ethnicity, and usual residence address - as far as possible, for everyone on the Integrated Data Infrastructure spine. These variables are derived by combining information from a range of data sources.

45. An admin resident population at a given reference date can be constructed from the Integrated Data Infrastructure spine. It includes those individuals in the Integrated Data Infrastructure spine who have activity in selected administrative data sources over a two-year period up to the reference date. Those who have died before the reference date are identified by a link to death registrations data and are excluded. International border movements data is used to exclude anyone who was not a New Zealand resident on the reference date, for example a resident who migrates to live overseas, or a short-term visitor to New Zealand

46. Young adults. Almost everyone in the Integrated Data Infrastructure spine has an age, the variable has a very high quality and young adults can be well-identified in the Integrated Data Infrastructure. With respect to the resident population, we have found that the discrepancy between the admin-derived resident population and the official figures is fairly uniform across all ages and young adults do not stand out. However, it is more difficult to provide an accurate address for young adults using only administrative sources. Comparing the admin-derived usual residence address with census, consistency is lowest for young adults, in particular the 20 to 24 years age group. A similar pattern is seen at all geographic levels, though it is less pronounced for the larger geographies.

47. Indigenous people. Māori are the indigenous people of New Zealand. Māori descent and Māori ethnicity are two main ways of identifying who is Māori in New Zealand.

48. Māori descent is collected on birth registrations for the child and their parents and is available since 1995. The 2013 Census provides good coverage for those born in New Zealand before 1995, and for migrants who had arrived in New Zealand by 2013. Used in this way, the 2013 Census can be considered as a one-off administrative data source which is valuable in providing historical information before birth registrations were available. The quality of Māori descent data from these two sources is very good for people who we do have a value for, but 14 percent were missing a value for Māori descent in 2018. The electoral roll also collects Māori descent, and the combination of birth registrations and electoral roll data (those 18 years and older who enroll to vote) would help to fill the missing-data gap and provide information for much of the population on an ongoing basis. The recent passing of the Data and Statistics Act 2022 allows Stats NZ to access electoral data, and research will be a priority once data is available.

49. In New Zealand, ethnicity is a measure of cultural identity and is a key social factor used to describe the population. Around 17 percent of the New Zealand population are estimated to be of Māori ethnicity.

50. Ethnicity is collected by a number of government agencies on a regular basis. While agencies aim to collect ethnicity according to the statistical standard for ethnicity, various constraints on collection and processing mean that the quality of ethnicity information varies. There are also legitimate reasons why responses may differ since people’s view of the ethnic group they belong to can change over time, and people do not always give the same answers about their ethnicity in different situations. To harmonise the information from different sources, the Integrated Data Infrastructure applies a method that ranks data sources by the quality of their ethnicity data and selects the highest quality source available.

51. Using only administrative sources, an ethnicity is available for nearly everyone. While there is some concern about the collection of Māori ethnicity by some agencies, the quality ranking method generally performs well. The age distribution of the administrative

population census closely follows the patterns of the official data but slightly lower, and administrative population census counts by age are similar to or higher than the 2018 Census.

52. Homeless. Although homeless people are identified by some government departments and non-government organizations who provide services for them, administrative data about those who are homeless in New Zealand is limited and fragmented.

53. Stats NZ developed a definition of homelessness with four categories: ‘living without shelter’ and ‘living in temporary accommodation’, ‘sharing accommodation’ and ‘uninhabitable housing’. There is no comprehensive or standard way on identifying or counting the homeless population.

54. Most estimates of homelessness rely on the five-yearly census. The census data is supplemented by administrative data for organizations including Night shelters, Women’s refuge, and other accommodation targeted at people who lack access to minimally adequate housing. The problems with the 2018 Census will have had a significant effect on the quality of information about severe housing deprivation. This underlines the importance of developing other sources of data to monitor severe housing deprivation – providing more regular data and allowing comparison with the findings from census.

55. Illegal migrants. New Zealand has very good administrative data to identify migrants who stay in New Zealand illegally.

56. New Zealand is separated from its nearest neighbours by thousands of kilometers of open sea. This geographic isolation means that there is very little opportunity for asylum seekers, refugees, or any other migrants to cross New Zealand’s borders without going through formal border controls. Stats NZ expects that very few people enter New Zealand illegally, and that the bigger issue is likely to be those who enter the country on valid visas but overstay the terms of their visa.

57. Government departments collect and maintain data on international travellers and migrants. Customs New Zealand holds passport data for virtually every arrival and departure across the New Zealand border. The Ministry of Business, Innovation and Employment holds visa applications, visa approvals and border movements data (with limited data on New Zealand and Australian citizens who have free rights of entry). Immigration New Zealand monitors compliance with visa conditions and provides information to the Continuous Reporting System on Migration that include data on deportations and voluntary departures. Those admitted to New Zealand as refugees and asylum seekers can also be identified from Immigration data in the Integrated Data Infrastructure and are the subject of a recent study.

58. Mental health and addiction. Health data within the Integrated Data Infrastructure includes information about mental health and addiction related to those using health services. Service use observed in administrative microdata is one starting point for understanding mental health and addiction conditions. To assist researchers, five key sources of mental health and addictions data have been consolidated in the Integrated Data Infrastructure into a single table of events.

59. Two tiers of the health service are captured in the Integrated Data Infrastructure and provide clearly coded mental health and addiction events. These are specialist services as recorded by the Project for the Integration of Mental Health Data, and hospital admissions as recorded by National Minimum Dataset. The third tier, community health services, are not available in the Integrated Data Infrastructure.

60. In addition, there are three supporting services that cut across different service tiers and types and provide information about mental health and addiction conditions. These are laboratory tests, dispensed medications, and medical certificates used for welfare support.

61. Addiction events in the consolidated table are limited to alcohol and other drug abuse (also referred to as substance use disorder).

62. While the focus of the administrative data in the Integrated Data Infrastructure is on uptake in the use of health services, identification of the prevalence of mental health and addiction need has been primarily done via surveys.

63. Disabled people. The main information about disabled people comes from the census, and the Disability survey conducted after the census.

64. Administrative data on disability is scarce. Some organizations do report on their service delivery to the disabled community (for example through health, education or government benefits). There has been ongoing encouragement from the disability sector for programs and services to collect data related to disability which in turn contributes to the availability of disability data. The Office for Disability Services has compiled guidance for organizations on how to collect disability data in an appropriate and effective way and examples and guidance on how to assess the quality of administrative data and how to use it.

E. United States

65. United States has a decentralized statistical system. At this point in time, the United States Census Bureau does not have consolidated information about the US population members based on past census and administrative record information from birth, death, tax, health, and other sources. Because of this, *hard-to-reach* populations in relation to administrative records can be different for the United States Census Bureau than other countries.

66. The United States does have particular populations that are harder-to-reach than others. One group that is *hard-to-reach* are children. The United States has had undercount of children especially young children for our past three decennial censuses. It has also been found that under coverage in administrative records for young children was more as compared to other age groups as well. This can have impacts on statistical content and quality since these census results have been used as a base in our population estimates program to allocate federal funds, as population controls for demographic surveys, and as the denominators for calculating vital rates. One solution was to use national-level estimates from our 2020 Demographic Analysis program that utilizes the birth, death, immigration, and emigration data to form a blended base for population estimates program.

67. Another *hard-to-reach* group are people with disabilities. The Census Bureau has surveys like the American Community Survey and the Survey of Income and Program Participation that provide information about people with disabilities. One analysis used one-year estimates from the American Community Survey were able to examine disability rates for children and the monetary and nonmonetary costs associated with their care. The Census Bureau continue to identify and obtain administrative data for this population. The Social Security Administration administers the Supplemental Security Income program that is a means tested cash assistance program. The Survey of Income and Program Participation uses this information in the editing and imputation. This program recipients are also eligible for health insurance coverage in most states. By obtaining the Supplemental Security Information and medical insurance coverage administrative data, the Census Bureau has seen one potential solution of combining these administrative data with Census Bureau surveys to produce more in-depth statistical analyses.

68. As one solution, the Census Bureau is undertaking a modernization effort that combine data-science with traditional survey methods to diversify our data products and place data at the center of our approach. One aspect of this is the creation of our Frames program. The Frames program is a growing variety of linked datasets. While many of these datasets already exist as standalone entities at the Census Bureau, the Frames approach will collocate these and any number of curated datasets and provide an easy and efficient way to link them for purposes both familiar (e.g., providing a tailored survey frame) and unanticipated (e.g., answering a new question about jobs and COVID-19 vaccination rates). These linked, augmented and continuously updated datasets will provide a more comprehensive means for maintaining and updating the inventory of our nation’s addresses, jobs, businesses, people, and other linked data. Centralization and ability to link to other records will increase efficiency, reduce duplicative efforts to maintain and manage data and greatly expand our capacity to answer critical questions about the nation’s population and economy at multiple geographic scales. As part of this potential solution, the Census Bureau

will continue to identify and pursue administrative data sources to fill in the gaps for children, disabilities, and other *hard-to-reach* populations.

V. Main findings

69. *Hard-to-reach* populations in administrative data can be interpreted in different ways and their definition is dependent on country circumstances. One of the interpretations is incompleteness of administrative registers or linked administrative databases, making some groups, such as children or elders *hard-to-reach* and hence describe with data. This approach is described by the cases of Canada and the US. Another interpretation is selecting some groups, such as homeless, illegal immigrants or indigenous people and then trying to capture them in administrative registers to get a more complete information. People with disabilities are another group often mentioned in this context, however the data situation for people with disabilities does not make wider analysis possible. This approach is described by Denmark, Italy, New Zealand, and the US.

70. There has not been found evidence for well-established mechanisms in identifying *hard-to-reach* populations, such as homeless people or illegal immigrants in administrative registers. The examples discussed show that identification of *hard-to-reach* populations is depending on a country context but is often supported by surveys that try to capture attributes, such as ID number, address, date of birth helping in identification of *hard-to-reach* populations in administrative registers.

71. Examples provided show a practice of complementing administrative data with statistical surveys or census data to get a little broader information on *hard-to-reach* groups. This can enrich the existing information and also allow for linking data with across different data sets.

72. The described countries have different individual initiative in order to improve access to *hard-to-reach* populations. Provided there is interest for improving the access to *hard-to-reach* populations in administrative registers in a more cross-cutting manner, there is need for further cooperation in the field.

VI. Recommendations

73. There are many ongoing initiatives in individual countries aiming at capturing *hard-to-reach* populations in administrative data and the understanding of *hard-to-reach* populations can vary from country to country. Some countries interpret it as specific segments of the population, such as homeless, other as specific population segments that are not fully covered by administrative data.

74. The diverging understanding of *hard-to-reach* populations could point to the need of some cross-cutting development of the work on *hard-to-reach* populations, so that countries could benefit from a common framework of concepts and some general guidelines in the field.

75. One of the steps in developing work with *hard-to-reach* populations in administrative registers could be identification of cross-cutting issues faced by countries and then delimiting the broad application of *hard-to-*reach populations to few specific groups, such as homeless, illegal immigrants or other groups, depending on the findings.

76. The delimitation of *hard-to-reach* populations could serve as a point of departure for further investigation of whether there is a common ground for an analysis of how to better identify those groups in administrative registers.

77. If such a common ground is lacking, it could be considered to outline a list of best practices in getting access to *hard-to-reach* populations in administrative registers in different countries.

78. One of the factors that could promote a wider work on the use of data on *hard-to-reach* groups could be a clearly defined demand, such as legislative framework or international recommendations.

79. Against the background of the above it is recommended to establish a Task Force in order to outline the need for future cooperation on *hard-to-reach* populations in administrative data.

VII. Discussion by the Bureau of the Conference of European Statisticians

80. The CES Bureau made an in-depth review of hard-to-reach groups in administrative sources at its February 2023 meeting. The following comments were made in the discussion:

(a) The paper gives a very good overview of the issues related to this very important topic;

(b) A conceptual framework could help to identify what we know and what we do not know about hard-to-reach groups. For some groups we may have very limited information such as the total number of persons belonging to the group, or have no information at all;

(c) Work could be useful on some basic principles, and focusing on certain policy relevant groups, such as people with a disability, migrants, ethnic minorities, homeless, children, and older persons. The policy perspective is important. Who are these groups? What are their characteristics? How are they integrated?

(d) Administrative sources do not capture some people because they do not meet the conditions for inclusion. We need to be mindful of the design of the sources, and that they may be imperfect. The experience of countries with register-based statistical systems will be particularly important. All sources should be considered, not just administrative sources. Data integration is key;

(e) Some people do not want to be included and may try hard not to be identified. Special methods should be developed to identify those people, and multiple sources should be used.

81. In conclusion, the CES Bureau supported further work in this area and agreed with the establishment of a new task force, as recommended in the paper. Denmark will chair the new task force. In addition to the countries that already contributed to the paper (Canada, Italy, New Zealand and United States), Ireland, Mexico, United Kingdom, Eurostat, OECD and UNSD expressed interest in joining the task force. The Secretariat will prepare the terms of reference for the new task force, for review by the Bureau at the October 2023 meeting.

VIII. Acknowledgements

82. Statistics Denmark would like to extend special thanks to Sylvain Delisle, Karelyn Davis, James Falconer, from Statistics Canada; Donatella Zindato from Istituto Nazionale di Statistica (ISTAT); Christine Bycroft, Pubudu Senanayake, Nathaniel Matheson-Dunning, Hannes Diener from Statistics New Zealand; and Vincent T Mule Jr and Andrew D Keller from U.S. Census Bureau.

IX. References

Bowlby G., Morel J.: Use of Administrative Data in the Canadian Census, UNECE Twenty-fourth Meeting of the Group of Experts on Population and Housing Censuses, September 21-23 (2022):

https://unece.org/sites/default/files/2022-07/ECE_CES_GE.41_2022_7-2210829E.pdf

Statistics Canada (2021). Guide to the Census of Population. Appendix 1.7 – Use of administrative data to impute non-responding households in areas with low response rates,

https://www12.statcan.gc.ca/census-recensement/2021/ref/98-304/2021001/app-ann1-7-eng.cfm

Bernardini A., Chieppa A., Cibella N., Solari F, Zindato D.: Evolution of the Italian Permanent Population Census. Lessons learnt from the first cycle and the design of the Permanent Census beyond 2021, UNECE Twenty-fourth Meeting of the Group of Experts on Population and Housing Censuses, September 21-23 (2022): https://unece.org/statistics/documents/2022/07/working-documents/evolution-italian-permanent-population-census

Falorsi, S.: The Italian experience on the Population and Housing Census: the Master Sample, UNECE Nineteenth Meeting of the Group of Experts on Population and Housing Censuses, October 4-6 (2017): https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2017/Meeting-Geneva-Oct/Day2_1130_Italy_falorsi_presentation.ppt__1_.pdf

Gallo, G., Zindato, D.: Annex H. Italy case study, in UNECE (2018), Guidelines on the Use of Registers and Administrative Data for Population and Housing Censuses, https://unece.org/guidelines-use-registers-and-administrative-data-population-and-housing-censuses-0.

Gallo, G., Zindato, D.: Italy: The combined use of survey and register data for the Italian Permanent Population Census count in UNECE (2021), Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses, https://unece.org/statistics/publications/CensusAdminQuality

Stats NZ (2022a). Experimental administrative population census: Data sources, methods, and quality (second iteration). Retrieved from www.stats.govt.nz.

Milne, B. J., D’Souza, S., Andersen, S. H., & Richmond-Rakerd, L. S. (2022). Use of population-level administrative data in developmental science. Annual Review of Developmental Psychology, 4. https://doi.org/10.1146/annurev-devpsych-120920-023709

Stats NZ, 2022b Information for Māori about the experimental administrative population census

Amore K. (2016). Severe housing deprivation in Aotearoa/New Zealand: 2001-2013. He Kainga Oranga/Housing & Health Research Programme, University of Otago, Wellington.

Amore, K. Viggers H, Howden-Chapman P (2021). Severe housing deprivation in Aotearoa New Zealand, 2018: June 2021 Update. He Kāinga Oranga / Housing & Health Research Programme Department of Public Health University of Otago, Wellington. Retrieved from www.hud.govt.nz.

Arezoo Malihi, Annie Chiang, Jay Marlowe, Barry Milne, Dan Exeter (2022) Navigating data in the IDI to unlock refugee settlement trajectories link is to a presentation – may have paper forthcoming.

SWA’s guide to using Integrated Data to understand mental health and addiction conditions.

Fernandez, L., Shattuck, R., and Noon, J. (2018), "The Use of Administrative Records and the American Community Survey to Study the Characteristics of Undercounted Young Children in the 2010 Census"

Center for Administrative Records Research and Application Working Paper Series #2018 -05

https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/carra-wp-2018-05.pdf

Khubba, S., Heim, K. and Hong, J (2022), "National Census Coverage Estimates for People in the United States by Demographic Characteristics" 2020 Post-Enumeration Survey Estimation Report, PES20-G-01

https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/national-census-coverage-estimates-by-demographic-characteristics.pdf

Velkoff, V., & Hartley, C. (2022). Moving Forward With the U.S. Census Bureau’s Annual Population Estimates Post-2020. Harvard Data Science Review, 4(4). https://doi.org/10.1162/99608f92.4ba61ca4

Young, N. (2021) , “Childhood Disability in the United States: 2019,” ACSBR-006, American Community Survey Briefs, U.S. Census Bureau, Washington, DC, 2021

https://www.census.gov/library/publications/2021/acs/acsbr-006.html

Thieme, M. (2022), "Technology Transformation at the Census Bureau: Building a Modern, Data-Centric Ecosystem"

https://www.census.gov/newsroom/blogs/research-matters/2022/10/technology-transformation.html