

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Advancing Equitable Data Collection: Insights from Statistics Canada's Statistical Integration Methods Division Disaggregated Data Action Plan Research Project

by Andrew Pearce, Kenza Sallier and Christiane Laperrière

Release date: September 8, 2025



Statistics
Canada Statistique
Canada

Canada

Advancing Equitable Data Collection: Insights from Statistics Canada’s Statistical Integration Methods Division Disaggregated Data Action Plan Research Project

Andrew Pearce, Kenza Sallier and Christiane Laperrière¹

Abstract

As part of answering the call to action for the United Nations’ (UN) 17 Sustainable Development Goals, as well as addressing social, economic, and equity challenges within Canada, Statistics Canada’s five-year development phase for the Disaggregated Data Action Plan (DDAP) was funded in 2021 to support data driven decision around these challenges. In turn, the document “Guiding Principles: Leveraging the 2021 Census of Populations Data for DDAP Groups of Interest” were created. The guiding principles document explains the organizational framework of the DDAP in the Agency, describes existing data sources, addresses ethical and privacy concerns, and centralizes sampling methods tailored for DDAP initiatives while accounting for characteristics which can complicate sampling and data collection procedures.

Key Words: Data disaggregation; Sampling; Rare populations; 2021 Census.

1. Introduction

1.1 Organizational Context

As part of the UN’s 2030 Agenda, there was a global call to action to address their 17 Sustainable Development Goals, among which are reduced inequalities, gender equality, and no poverty (United Nations, 2024). Similarly, in Canada, social movements for Indigenous rights, racial justice, and economic equity are at the forefront of society. Whether it be international or Canadian challenges, both require the use of official statistics to guide the decision-making processes that surround them. In response to these needs, the Government of Canada funded Statistics Canada’s five-year development phase for the Disaggregated Data Action Plan (DDAP) in 2021.

In terms of concrete actions, a disaggregated data strategy such as Statistics Canada’s DDAP refers to a systematic approach of collecting, analyzing, and releasing data on a particular topic by breaking down the data into smaller, more granular categories. As a National Statistical Organization, Statistics Canada governs the implementation of its disaggregated data strategy through the usage of rigorous methods, standards, and quality control procedures, each ensuring the accuracy and reliability of the produced statistics while also accounting for privacy, confidentiality, and ethical considerations. As a result, Statistics Canada’s DDAP covers all aspects of the data life cycle, or the Generic Statistical Business Process Model (UNECE, 2019).

1.2 The 2022/23 Census Operations Section Research Project

As previously mentioned, the DDAP is currently being implemented in many Statistics Canada programs, from the sampling design to the collection stage. For social programs and surveys, the design stage of the project involves determining the population size of the in-scope disaggregated groups and their associated guidelines to choose

¹ Andrew Pearce, Statistics Canada, 150 Tunney’s Pasture Driveway Ottawa, Canada, K1A 0T6 (andrew.pearce@statcan.gc.ca); Kenza Sallier, Statistics Canada, 150 Tunney’s Pasture Driveway Ottawa, Canada, K1A 0T6 (kenza.sallier@statcan.gc.ca); Christiane Laperrière, Statistics Canada, 150 Tunney’s Pasture Driveway Ottawa, Canada, K1A 0T6 (christiane.laperriere@statcan.gc.ca)

appropriate methods required to meet the disaggregated data targets. As this paper will highlight, traditional sampling methods are not always sufficient to allow for robust estimation given the methodological challenges that arise from the specific characteristics of subpopulations targeted by the DDAP. The census operations section of the Statistical Integration Methods Division (SIMD) took the initiative of leading a research project, which outlined how to leverage the 2021 Census of population data, which, at the time of writing, is the source of the most detailed socio-demographic data of the Canadian society to support these procedures. In particular, the census of population data can identify the four employment equity groups (women, indigenous peoples, persons with disabilities, and members of visible minorities), as well as breakdowns of these, and other groups at varying levels of geography.

Ultimately, the main objective of this research project was to develop an internal document recommending guiding principles based on population size and characteristics to determine whether direct estimates are possible (e.g., via oversampling), or whether other indirect methods and modeling techniques are required. Specifically, two main deliverables, presented within this paper, were identified to achieve these objectives. The first deliverable consists of an internal document of guiding principles for employees at any level in the Agency (e.g., survey managers) interested in producing estimates of parameters for specific DDAP subgroups of interest. The guiding principles document also serves as a central location for information pertaining to data sources, ethical and practical considerations, response burden, and includes a literature review of appropriate sampling/data collection procedures in the context of the DDAP. The second deliverable are tables of counts based on the 2021 Census of population data that were produced to support practical applications of these guiding principles by allowing subpopulation prevalence of DDAP groups of interest to be assessed by integrating intersectionality, which, based on the Public Health Agency of Canada (2022), “refers to how sources of discrimination overlap and reinforce each other.”

2. Data, Ethical, and Practical Considerations

2.1 Data Ethics and Response Burden

Statistics Canada has an internal ethical review process in place for the acquisition of survey and non-survey data, aiming to evaluate the ethical justification for both the acquisition and proper use of the data. Ethical reviews are a necessary condition to ensure the Agency can communicate why it needs to acquire the data and how it will safeguard its proper use, helping to ensure a wider social acceptance and maintain public trust for the use of data (Deblois & Côté 2021). Ethical reviews are integrated in many parts of the data life cycle, the same data life cycles covered by the DDAP. As such, DDAP initiatives can raise specific ethical challenges outside those common in typical surveys. If challenges such as complex historical heritages/sensitivities, socioeconomic barriers, and general distrust of the government are not managed alongside the need for information, long lasting negative effects may occur, hindering future initiatives centered around these subpopulations.

Alongside data ethics, another important aspect to consider when selecting a data collection approach in the context of the DDAP is response burden. Response burden is the effort required by respondents to provide personal information through a questionnaire, an interview, or via direct measurement (common in health-related surveys). When not managed properly, higher response burden can lead to higher total and item non-response and, depending on the topic, could raise concerns from the public. Furthermore, the effects of response burden are compounded when interested in DDAP groups as these groups are often overrepresented in the samples to produce reliable estimates. For the same reasons as above (sensitivities, barriers, and distrust) and difficulty in collecting data, when burden is left unchecked it can make an already challenging population to sample, much more difficult to sample in the future.

Currently, there are processes in place at Statistics Canada to control the various aspects of response burden, from consultations with experts to ensure the questionnaire is properly designed, to following the Necessity and Proportionality Framework to ensure only the relevant information is asked, proportional to the sensitivity and social acceptability of the topic (Statistics Canada, 2024b). One way to monitor response burden is by using sample selection history files, which document the past selections of a given household in previous surveys. Selection history files ensure a better coordination between surveys, both inside and outside the same division, while helping to minimize the chance of re-selecting the same dwellings in a short time frame, in turn, reducing item and total non-response rates.

3. Main Deliverables

3.1 Guiding Principles Document

The main deliverable of the research project is a document internal to Statistics Canada named “Guiding Principles: Using the 2021 Census of Population Data to Produce Statistics on DDAP Groups of Interest”. The DDAP subgroups of interest include the previously mentioned employment equity groups (women, indigenous peoples, persons with disabilities, and members of visible minorities), as well as immigrants to Canada, people with a low-income status, seniors, and veterans. The creation of this document was motivated by the methodological challenges encountered when collecting statistical information on subgroups of interests for DDAP initiatives. Due to characteristics commonly found in subgroups targeted by the DDAP, traditional sampling processes may not be suitable and other survey designs and estimation methods should be considered. Consequently, it was identified that Statistics Canada would benefit from a comprehensive guide serving as a central location for both methodologically traditional and non-traditional approaches for sampling and collecting information on subpopulation matching DDAP characteristics. For each sampling method and data collection procedure outlined in the guide, the corresponding theoretical factors, important considerations, as well as both advantages and disadvantages are provided.

3.1.1 Characteristics of DDAP Subpopulations

As previously mentioned, the size and characteristics of a DDAP subgroup of interest can pose methodological challenges at the sampling, collection, and estimation stages. As a result, traditional sampling processes or data collection procedures may not be suitable, and other non-traditional survey designs and estimation methods should be considered.

Introduced by Kish (1987) is a general classification system for subgroups in a population based on their size in said population. This domain classification is shown in Table 3.1.1-1 and includes an example of each classification taken from the tables of population counts.

Table 3.1.1-1
Domain classification system as proposed by Kish (1987)

Domain Classification	Prevalence in Population	Example based on 2021 Census of Population
Major	≥10%	Individuals who immigrated to Canada
Minor	1% - 10%	Veterans of the Canadian Military
Mini	0.1% - 1%	Transgender Persons in Canada
Rare	< 0.1%	Canadians who are First Nations and Métis

Although domain size gives insight into the sampling feasibility of different subpopulations, it is not the only consideration when choosing an appropriate sampling method in the context of DDAP subgroups of interest. Another challenging characteristic of a subpopulation is their hiddenness, which Smith (2014) describes as “subgroups that are socially concealed” and that “in many cases, the groups are actively hiding because of the attributes that makes them the focus of research”. A common example associated with hidden populations is the topic of drug usage. Being considered a sensitive topic, it is unrealistic to assume that reliable information of drug user’s habits can be obtained via a traditional questionnaire since respondents may lie about their membership.

3.1.2 Oversampling Methods

One challenge when estimating parameters of interest for a minor population or smaller, is if the subpopulation is not specifically targeted in the sampling design, it is likely that there will be too few respondents in the subpopulation of interest to produce reliable estimates. Oversampling methods can be used to target the subpopulation(s) of interest, leading to the subpopulation(s) being overrepresented in the sample with respect to their observed prevalence in the general population. Oversampling methods, such as disproportionate stratification wherein the sample size of one, or many of the strata are not proportional to the size of the strata in the population (Daniel, 2011), are common within the Agency. Seeing wide usage in social surveys, but also in DDAP initiatives to ensure enough members of the target

population are sampled. Recently, as part of the 2024 Census of population test (Statistics Canada, 2024a), three DDAP groups of interest (people with military experience, indigenous peoples, people experiencing hidden homelessness) were specifically sampled disproportionately more relative to their size in the Canadian population. This allowed for a more accurate assessment of how the proposed changes for the next census cycle would affect them. Although not always an oversampling method, two-phase or multiphase sampling procedures are often used at Statistics Canada. In fact, post-censal surveys, such as the Canadian Survey on Disability, use the census of population as the first phase, then uses the obtained domain identifiers to sample their subpopulation(s) of interest in the following phase(s). Lastly, the use of multiple sampling frames can be useful when sampling small DDAP groups of interest as one frame may undercover the subpopulation of interest, while another may properly cover it while uncovering the rest of the population.

Within the Agency, there is closely related idea to oversampling, an oversample. Both result in respondents of the targeted subpopulation being overrepresented, however, unlike oversampling a subpopulation, which is planned for in the sampling design of the survey, an oversample is a supplemental, secondary sample to the survey's main sample. The oversample is drawn to meet data goals outside the baseline requirements of the main survey. As an example, an oversample was used in the 2021/22 Canadian Community Health Survey to better understand health status, health care utilization, and health determinants in the four employment equity groups. Here, an oversample was used as it was too late in the survey design stage to modify the sampling design.

3.1.3 Methods for Hidden Populations

When targeting hidden populations, the stigmatization or involvement in illicit activities of the population can affect collection procedures (Heckathorn, 1997). Specifically, such populations may purposely disclose incorrect information on domain membership to the survey teams over fears of being judged or prosecuted. Thus, due to the hidden nature of some populations, full or partial identification is rarely available and, even when it is, concerns over its accuracy can arise. However, hidden populations often have other characteristics which can in turn be leveraged to sample and collect data from them. For example, some hidden populations are known to frequent various known locations allowing for time-location based sampling, while others may have an underlying social structure allowing for chain referral-based methods.

Time-Location Sampling (TLS) can be used to sample members of subpopulations which are known to frequent various locations, such as subpopulations without a fixed address and those that are mobile (commonly moving from place to place). In short, TLS works by screening individuals by performing the sampling at locations they are known to visit (e.g., food banks and shelters for the homeless population). A frame of known locations and times can be constructed for sampling, yielding probabilistic results. However, undercoverage may occur if individuals of the population visit other locations at other times that are unknown to the survey team, or if locations are excluded from the frame (e.g. if they deemed these areas to be dangerous). Conversely, by sampling at different locations, at different times, it is possible that the same member of the population is selected into the sample multiple times leading to issue of multiplicity (Kalton, 2009).

When sampling populations that are socially connected, chain referral methods can be appropriate. Chain referral methods are a family of sampling methods, which depend on an initial set of respondents (the seeds). Upon their participation in the survey, they are asked to refer other people they know within the population into the survey in exchange for some incentive. The sequential referral of others to the survey creates referral chains, which progress and grow until a desired sample size or an estimate's precision is attained. Discussed within the guide are two of these methods: snowball sampling and Respondent Driven Sampling (RDS). Snowball sampling can yield probabilistic results if the initial seeds are selected randomly from a sampling frame, however this is not feasible for hidden populations in the context of DDAP. RDS, which can be thought of as an extension of snowball sampling, can provide probabilistic results independently of the initially selected seeds (Heckathorn 1997) through a set of rigorous assumptions which are hard to meet in practice.

3.1.4 Practical Considerations

Another main objective of the guiding principles document was, through a set of practical considerations and questions, to guide the reader to a suitable survey method or data collection strategy for their own specific set of needs.

The set of practical considerations and questions begins by asking the reader “Do you have a strong justification for collecting data on a specific DDAP subgroup?”. This question highlights the fact that while DDAP aims to make data disaggregation a standard practice within the Agency, the justification for collecting data must extend beyond simply filling a current data gap; especially if sensitive questions are asked. Readers of the guide are recommended to follow the Necessity and Proportionality Framework to better assess their justification for collecting this data. Response burden should also be considered before finding appropriate sampling or data collection methods. Readers can refer to the section in the guiding principles document on response burden which outlines available tools referenced within the document to assess the burden their initiative would place on the possibly small, DDAP subgroup of interest.

After the reader evaluates the burden and ethical implications of their initiative, they are then asked, “What are the characteristics of your DDAP subgroup of interest?”, aiming to assess subpopulation size, hiddenness, and identification level. Identification level, the proportion of the subpopulation for which domain identifiers are available, can vary for many reasons. Of course, the subpopulation may be hidden, and domain identifiers would be missing for this reason. Certain identification questions may only be asked on the long-form census questionnaire and are therefore only available for roughly 25% of the population. Furthermore, it may be the case that identifying questions were not asked in the census, but instead on other social surveys. In these scenarios, readers are directed towards using internal database searching tools to assess the current identification level by determining what data sources exist for the given subpopulation. Lastly, when identifying questions that are not asked on the census nor on social surveys, proxy domain identifiers can be used instead, using available variables that are correlated with the population of interest, to identify those likely to be in the subpopulation of interest.

Once these three characteristics are assessed, the reader will be guided to an appropriate section of the document outlining important consideration for their situation, appropriate sampling methods for their needs, as well as pros and cons of these methods. The reader is guided by a set of questions such as “is your population of interest hidden?”, or “given that your population of interest is not hidden, what level of identification do you have for it?”. For example, if the reader’s subpopulation of interest is very small (e.g., mini, or rare as in Table 3.1.1-1), they will be directed to information pertaining to small area estimation or combining cross-sectional samples. If their subpopulation completely lacks, or has very little identifying information prior to collection, the reader would instead be presented with information on screening approaches.

3.2 Tables of Population Counts

To help meet the methodological challenges related to the sample size of subpopulations of interest, as well as make more concrete recommendations, another outcome of the research project was the internal production of tables of counts for many DDAP subgroups of interest based on the 2021 Census. However, before producing these tables, the source of the data had to be decided between three possibilities available at Statistics Canada: the 2021 Census Response Data Base (RDB) that includes the raw responses provided by respondents, the 2021 Census Edit & Imputation (E&I) database that includes cleaned-up, aggregated, imputed, and derived variables based on the respondent’s answers, and the Dissemination database, which includes variables that are scheduled for release to the public and have successfully passed confidentiality rules. It was decided to use the 2021 E&I for a few reasons, namely, due to the cleanliness of the data, which is particularly useful when there are variables already derived by other teams with more expertise (e.g., producing counts for the non-binary population). Secondly, the dissemination database may not have all the necessary information due to the applied confidentiality and suppression rules for these counts to be cleared for public use (e.g., non-binary counts are restricted).

Using the E&I database, person-level population counts of DDAP subgroups, broken down by intersectionality factors, were produced to highlight their sampling feasibility and determine which sampling strategies to consider based on their estimated population size. Initially, through discussions in a DDAP working group within the Agency and through standardized definitions aligned with the 2021 Census data, the following subpopulations were identified as the groups of interests: gender minorities (transgender, non-binary), Indigenous peoples (First Nations, Métis, Inuit) and Indigenous identity, visible minority groups, and active members or veterans of the Canadian military. However, after consulting survey managers in the Agency to better understand their survey needs when implementing DDAP initiatives, the considered groups were expanded to include persons living with an activity limitation, immigrants to Canada, and various religious groups. For all subpopulations of interest, distributions across gender (cisgender male, cisgender female, transgender male, transgender female, gender diverse), province/territory, and age group were

produced. For some subgroups of interest, distributions by income status and population center size were produced, and in the case of Indigenous peoples, counts were further disaggregated by on and off a reserve.

In terms of producing the tables of counts, there were some other considerations besides the data source, the subgroups of interest and intersectionality variables. For some population groups, the information used to identify them was sourced exclusively from the long-form census (e.g., military status) which is only sent to roughly 25% of dwellings. As a result, long-form weights had to be used to produce tables of weighted counts for these populations.

Similarly, imputed values were also accounted for. Given that one of the purposes of the DDAP is to sample specific subgroups of interest, accurate information is required to ensure the validity of a person's domain membership. Since the imputed value indicating their membership in the subpopulation may be different from their true value, all records where their membership variable was imputed were excluded from the counts of said subpopulation. These tables are only accessible internally in the Agency, wherein access is provided exclusively to social survey methodologists who request access and give proper justification of their use.

4. Conclusion

Motivated by the United Nations' 17 Development Goals and sociodemographic issues in Canada, the census operations section successfully carried out an innovative research project centered around using the 2021 Census of population data to compile and document important considerations when conducting DDAP initiatives. Synthesized and centralized within the guiding principles document are discussions and resources for the organizational context of the DDAP, data ethics, response burden, tables of counts for subpopulations broken down by intersectionality factors, methodological challenges (population size, hiddenness, identification level), and the appropriate sampling methods and data collection procedures that can overcome these challenges.

References

- Daniel, J. (2012), "Choosing the type of probability sampling", *Sampling essentials: Practical guidelines for making sampling choices*, pp. 125-174.
- DeBlois, S., and D. Côté (2021), "Perceptions of data sensitivity: Study with Questionnaire Testing Participants", unpublished report, Ottawa, Canada, Statistics Canada.
- Heckathorn, D. D. (1997), "Respondent-driven sampling: a new approach to the study of hidden populations", *Social problems*, 44.2, pp. 174-199.
- Kish, L. (1987), *Statistical Design for Research*. New York: Wiley & Sons, Inc.
- Kalton, G. (2009), "Methods for oversampling rare subpopulations in social surveys", *Survey Methodology*, pp. 125-141.
- Statistics Canada (2024a), "2024 Census Test" [Online] Available at: [2024 Census Test](#). Accessed: 2025-01-09.
- Statistics Canada (2024b), "Principles of Necessity and Proportionality", an article published on the Statistics Canada website: [Principles of Necessity and Proportionality](#).
- Public Health Agency of Canada (2022), "How to integrate intersectionality theory in quantitative health equity analysis? A rapid review and checklist of promising practices", an article published on the Government of Canada website. [How to integrate Intersectionality Theory in Health Equity analysis - Canada.ca](#).
- Smith, T. W. (2014), "Choosing the type of probability sampling", *Sampling essentials: Practical guidelines for making sampling choices*, pp. 125-174.

United Nations Economic Commission for Europe Statistics Division (2019), “Generic Statistical Business Process Model”, an article published on the UNECE website. [GSBPM v5.1 | UNECE](#).

United Nations Department of Economic and Social Affairs (2024), “Sustainable Development” [Online] Available at: <https://sdgs.un.org/goals> . Accessed: 2024-12-13.