Technical Documentation: Preventing & Reducing Homelessness Integrated Data Project

Table of Contents

PROJECT: Preventing & Reducing Homelessness Integrated Data Project

PREPARED BY:
Data Science Partnerships Program
Digital Platforms and Data Division
Office of the Chief Information Officer
Ministry of Citizens' Services

DATE: 2022-04-05

Purpose

This document summarizes the analytical methods used to update phase one of the *Preventing and Reducing Homelessness Integrated Data* project with data from the year 2020 as well as updated estimates for 2019. This update of phase one remains focused on implementing a cross-agency analytic definition of homelessness based on administrative data. To support the presentation of this estimated population a background table on shelter occupancy is also included in this package. In addition, COVID-19 cases were identified for the estimated homeless population using COVID testing data.

Project Overview

The Ministry of Attorney General and Minister Responsible for Housing, in partnership with BC Housing (BCH), Ministry of Citizens' Services and the Ministry of Social Development & Poverty Reduction (SDPR), have undertaken an integrated data project with a focus on homelessness in British Columbia.

Phase one of this project estimated the population who experienced homelessness in B.C. by building an analytical definition of homelessness defined by SDPR's *Employment and Assistance Program* usage, and integrating BCH's definition of homelessness, defined by *Emergency Shelter Program* usage. Several attribute variables were defined that address the demographic and geographic characteristics of this population. This is a novel approach and represents a new method, using administrative data, of estimating homelessness for British Columbia. A full technical summary of the 2019 estimate can be found at this url: https://www2.gov.bc.ca/assets/gov/housing-and-tenancy/social-housing/supportive-housing/tech_doc_preventing_and_reducing_homelessness_integrated_data_project_province_of_british_columbia_2021.pdf

The project is enabled by the Data Innovation Program, a data integration and analytics program for the B.C. government. While every B.C. ministry collects and manages its own data, the Data Innovation Program can securely link and de-identify data from multiple ministries, giving government analysts a broader understanding of complex issues. The Data Innovation Program is based on world-leading best practices for managing safe access to confidential or sensitive data, following the Five Safes model to protect data and reduce the risk of sensitive data being accessed or used inappropriately.

The data science methods and implementation are being conducted by the Data Science Partnership program—a service provided through the Digital Platform and Data Division of the Office of the Chief Information Officer, Ministry of Citizens' Services—which aims to accelerate adoption of data-driven digital government and supporting evidence-based decision making by providing cross-government data science support.

Data

Data available through the Data Innovation Program is always de-identified to protect privacy. This means identifiers such as names, driver's licence numbers and personal health numbers are removed and replaced by a project-specific studyid used to link individuals across data sets in the secure analytics environment.

Data sets used in Phase one of this project were:

B.C. Employment and Assistance (SDPR, Data Innovation Program)

- Involvement data
- No Fixed Address data

Emergency Shelter Program (BCH, Data Innovation Program)

- Shelter Stays data
- Shelter Attributes data
- Clients data

Health data (Data Innovation Program)

- COVID Test Results
- B.C. Central Demographics data

Statistics Canada Table 17-10-0005-01 (open-licensed data)

- Population Estimates on July 1st, by age and sex

Statistics Canada Table 17-10-0139-01 (open-licensed data)

- Population estimates, July 1, by census division, 2016 boundaries

Statistics Canada Table 17-10-0137-01 (open-licensed data)

- Population estimates, July 1, by economic region, 2016 boundaries

Data Integration & Provisioning

Data sets are linked and provisioned into a secure analytics environment by Population Data B.C., an internationally-recognized academic organization that has facilitated population-based research for over 20 years.

Population Data B.C. uses a *population directory* as the primary key for data linkage. It includes all the individuals about whom Population Data B.C. has information. The directory has been built with Medical Services Plan (MSP) registration & premium billing data going back to 1985. Since MSP covers most of the B.C. population and issues a unique lifetime Personal Health Number, this directory provides the largest representative coverage of residents in B.C. to support high quality linkage across data sets.

The data linkage rate, defined here as the percentage of individuals linked to the population directory for a given data set, is dependent on the number and kind of identifiers contained in the data set. For data used to estimate the homeless population, SDPR data contains Personal Health Numbers, while BC Housing's data contains only provided name, date of birth and gender. As a result, the linkage rate of the BC Housing data was lower than the SDPR data.

While the administrative data sets available through the Data Innovation Program are comprehensive, administrative data is still a sample of the actual population and all derived estimates have some associated uncertainty. Data set linkage rates offer quantification of some of this uncertainty across linkages in *all* data sets in the Data Innovation Program. While "unlinked" studyids within a data set represent studyids not linkable to *any* data in the Data Innovation Program, this is distinct from individuals present in *some* (but not all) data sets. For example, some individuals were present in both the SDPR and BCH data while others were present in only one or the other. For the definition of homelessness, this enables inclusions of individuals using either service. However, when linking to additional data sets, for example the B.C. Central Demographic file, it means that not all individuals in the SDPR and BCH data are linked or found in the B.C. Central Demographic data. Further complicating linkages, while the *population directory* developed by Population Data B.C. is continuously updated, the data provisioned to a given project may have been linked based on older versions of the directory, resulting in "linked" individuals who are not present in any other data sets. Fortunately, this number remains low and does not presently affect the population level analysis described here.

Data Preparation

A number of BC Housing clients matched multiple studyids in other administrative data sets. Population Data B.C. would typically collapse these studyids into a single id and restate all domain data sets accordingly before provisioning the data into the secure analytics environment. In this case, to expedite data release, that work was passed along to the project team by providing an PopData Extra ID table that defines the BC Housing client's relationship to other data sets. The step-by-step process to subsequently collapse those studyids is as follows:

- Find all unique combinations of studyid and clientid in the PopData Extra ID file
- Find all unique combinations of studyid and clientid in the BC Housing Clients file
- Find all rows where clientid is equal in both PopData Extra ID and BC Housing Clients files

This process, recommended by Population Data B.C., results in a crosswalk table that allows for identifying duplicated ids outside of the BC Housing data and restating data sets by associating duplicated ids with the canonical studyid from the BC Housing data.

Statistical Disclosure

This project meets or exceeds the statistical disclosure guidelines set out by the Data Innovation Program. To maintain future flexibility most data outputs have been aggregated well above the cell size thresholds to reduce the risk of future re-identification if more fine-grained results are exported.

In the table hl_age_and_gender.csv there are three cells that do not meet the rules of thumb. These three cells all share status as "Unknown" age and gender. These are individuals for which this demographic information is not found in any of the shelter, income assistance or health demographics data. The information in these cells *does not* represent non-binary individuals in the data, rather the data simply does not exist. The risk of re-identification for these individuals is very low. Future data updates may result in exports that no longer have "Unknown" ages and gender. In such as case, it could be assumed that the data counts were dispersed into the categories with known age and gender. However, this small increase would not be disclosive given the broader nature of these categories. Moreover, because linkage rates tend to vary with data updates, it would be impossible to know if changes in numbers reflect better linkage rates or updated demographic information.

Background Materials

Shelter Occupancy in British Columbia

The shelter occupancy was determined from the Shelter Stays data which provides the start and end date of an individual's shelter stay. Using those dates, we were able to establish how many people were registered into a single shelter per day. That table was joined (using the shelter_org_id field) to the Shelter Attributes data to label each shelter into either Emergency Response Centre, Temporary Shelter or Emergency Shelter Program (Year Round Shelters). The resulting table was further aggregated to present the number of people per date using a shelter type. Shelter data was subset to include only shelter visits after March 28, 2020. This was a simple method to meet statistical disclosure requirements.

Study Population: Estimated 2019 & 2020 Homelessness Population

The estimated population experiencing homelessness was defined using service use patterns from both the BCH Shelter Stays data and SDPR B.C. Employment and Assistance data. The definition period was chosen because it represents a common temporal period in both data sets and because a calendar year facilitates year-over-year comparisons as new data becomes available. The study population was defined at the monthly time scale, primarily because B.C. Employment and Assistance is delivered on a monthly basis. An individual was defined as experiencing homelessness in a given month:

a) if they had three consecutive months of income assistance (including the current month) with no fixed address or;

b) they spent any time in a shelter in the given month

Using this criteria, the population estimate was generated for each month. For the months of January and February 2018 income assistance data was required to determine if an individual met the definition. For an annual estimate, unique studyid's from the monthly estimate were compiled to form the annual population. These two temporal resolutions defined whether a person experienced homelessness during a given month or whether a person experienced homelessness at any point during a given year.

Study Population Attributes

Several demographic and geographic population attributes were added to the above population on both monthly and annual time scales. For attributes where this makes sense, the annual and monthly attributes are differentiated below.

Data Source

The means of entry into the population was recorded to determine whether an individual entered the study population via service usage definitions from the SDPR data, BCH data or both. At the monthly resolution, this measure directly results from whether a person entered into the population by their usage of income assistance, shelters or both. At the annual resolution, this measure expresses whether a person ever used either or both services sufficiently to meet the homeless definition.

Homeless Category

A chronic homelessness sub-population, distinct from the non-chronic homelessness population was also defined. This category was defined using an individuals past 12 months of service usage for a given month. For some months, this includes reaching into the previous year's shelter and income assistance data. We implemented the following criteria to define homeless categories:

- Calculate the cumulative number of nights spent in a shelter over the past 12 months
- Calculate the cumulative number of shelter visits separated by 30 days (unique visits) over the past 12 months
- Calculate the longest period of consecutive monthly income assistance for those individuals that reported no fixed address over the past 12 months
- Apply the following definition for a given month:
  - Non-Chronic Homelessness: **3 to 5 months** of consecutive income assistance reporting no fixed address OR **180 or fewer** days in a shelter OR **1 or 2 unique visits** to a shelter (separated by 30 days)
  - Chronic Homelessness: **6 or more months** of reporting no fixed address OR **greater than 180 days** in a shelter OR **3 or more unique visits** to a shelter (separated by 30 days)
- In instances where the above criteria resulted in differing non-chronic homelessness and chronic homelessness outcomes between the services, a studyid was associated with chronic homelessness

For the annual estimate, if a person is assigned chronic status during any month (based on their previous 12 month of service usage) in a given year, they were assigned chronic status for all of that year. A small portion of the BCH shelter data contained individuals with overlapping time intervals at different locations. That is, some people were registered in different shelters at the same time. Those overlapping intervals were merged to create one continuous interval. This was possible when the overlapping intervals were in the same city. For instances where an individual was present in two census subdivisions, the intervals were not merged. This is a very small number of shelter visits and therefore represents an acceptable loss of data accuracy.

Demography

This project had three sources of age and gender data for the study population: BCH client data, SDPR Income Assistance data and the B.C. Central Demographic data. This section outlines the hierarchy of those data sources and the process to determine which to use.

Date of Birth

The primary data source for date of birth is the B.C. Central Demographic file . When no date of birth was available in that file, date of birth from the SDPR Income Assistance data was used. These two data sources are the most reliable for date of birth because they require some level of verification for a service to be provided. If an individual still had not been assigned a date of birth, the BCH clients data was used which is a self-reported measure. Age was calculated as the age (as an integer) as of December 31 which is the end of the definition period. To meet conservative statistical disclosure goals, ages were grouped into three groups: 24 & under, 25 to 55, and 55 & over. To compare age distributions of the homelessness population relative to the general population, population estimates by age from Statistics Canada were used.

Gender

We have integrated gender identity data from various sources and specified a hierarchy for data quality. Each of the Central Demographic files, SDPR Income Assistance data and BCH clients data have gender indicators of varying quality.

The hierarchy of data quality used for gender identity data for this integrated project is as follows:

1. BCH client data is collected using open-ended responses and therefore is the priority gender indicator

2. The SDPR income assistance data, which collects binary genders, is the most recent data and is the next priority gender indicator

3. B.C. Central Demographic data, which collects binary genders, is the most dated data source and is therefore the last choice for gender identity. Note that while B.C. Central Demographic file is the least used data source for gender identity in this circumstance, it is still a reliable data source

The number of individuals identifying in the shelter data as non-binary was very small. This low number presented a possible statistical disclosure risk for that population of individuals. Therefore, individuals identifying as non-binary in the shelter data were recoded to their binary gender data present in the SDPR or MSP data.

Geography

To estimate the location of individuals experiencing homelessness, the location where a person was accessing services was used as a proxy. The hierarchy for this geographic proxy was:

1. Location of a shelter where a person visited
2. Location of the Service B.C./SDPR issuing office where a person accessed income assistance

The BCH shelter data provides location information for individual shelter visits while the SDPR income assistance data only provides monthly location information. For both of these data sets, the location was provided at either the municipal or the census subdivision level. Adopting a conservative approach to statistical disclosure, the export aggregated these data to the census division level. Further, the census divisions of 'Mt. Waddington' and 'Central Coast' were merged into one as were 'Northern Rockies' and 'Peace River'. The resulting census divisions and merged census divisions provided geographies to evaluate where services were used and by proxy where individuals were experiencing homelessness. The population was further aggregated to economic region to provide an additional geography. To calculate rates of homelessness relative to the general population, population estimates of census division from Statistics Canada were used.

Movers and Non-Movers

To resolve individuals that moved only within and between census division (or economic region) a geographic flag was added to each studyid indicating either the geography that a person spent all of their time in or a 'multiple cd' (or 'multiple er') flag if they moved between or among geographies. This flag was applied at both the annual and monthly time scales.

COVID-19 Testing Data

For COVID-19 tests which were able to be linked to an individual, a flag was added to the COVID-19 test dataset to indicate if the person was part of the population experiencing homelessness or part of the reference population (i.e. the rest of British Columbia). Only test results of positive or negative were retained, with a small number of inconclusive results being discarded.

Using the location marker discussed above, the spatial distribution of COVID-19 cases by economic region for the homeless population was estimated. For individuals residing in multiple locations, the most common location was chosen. For the very small number of individuals who resided in more than 1 common location, a location was randomly chosen from those possibilities. Additionally, an automatic error was implemented into the code to prevent this occurring for greater than 2% of individuals.

Appendix

This appendix is provided as specific documentation of the data analysis conducted on provisioned Data Innovation Program data in the secure analytics environment.

Software

This analysis is implemented in the R programming language (R Core Team 2021). The code used to generate this analysis was reviewed by three data scientists. Key tools used to complete this work include the Apache Arrow project (Richardson et al. 2021), the tidyverse (Wickham et al. 2019), cansim (von Bergmann and Shkolnik 2021), dipr (Albers and Hazlitt 2020) and the R package targets (Landau 2021) for project organization. All code is stored under the git version control system and shared inside the secure environment.

Raw Data

All data was converted from compressed fixed width files into parquet files for ease of analysis. Significant testing against Population Data B.C. data provisioning metrics occurred to ensure that conversions were done accurately.

Social Development and Poverty Reduction Data

Involvement data

- filenames: "idosdpr2018-2019.bceainvolvement.A.dat.gz", "idosdpr2020.bceainvolvement.A.dat.gz"
- columns used: ym, fileid, deprltncd, birthdt_yymm

Nfa data

- filenames: "idosdpr2018-2019.bceanfa.B.dat.gz", "idosdpr2020.bceanfa.B.dat.gz"
- columns used: ym, fileid, nfa, csdname
- subsetting: only rows where nfa == 1

BC Housing Data

Client Data

- filename: "bchousing_hifis2017-2020.hifis_clients.A.dat.gz"
- columns used: gender, dobyyyy, dobmm

Shelter Stays Data

- filename: "bchousing_hifis2017-2020.hifis_clnts_shlt_stays.B.dat.gz"
- columns used: shelter_stay_start_date, shelter_stay_end_date, shelter_org_id, shelter_census_sub_division, clientid

Shelter Attributes

- filename: Shelter attributes_HIFIS.xlsx
- columns used: org_id, shelter_type

Extra ID data - filename: "bchousing_hifis2017-2020.hifis_extra_clntid_popid_xlk.A.dat.gz" - columns used: clientid

Demographic data

- filename: "demographics1986-2020.B.dat.gz"
- columns used: sex, dobyyyy, dobmm

Health data

COVID Test Results V01

- filename: covidtest2020-20210719.A.dat.gz
- columns used: covid_is_valid_moh_clnt_identifier, covid_covid19_result

Data Processing

Restating Method

The extra id data represent linkages to multiple studyids in other data sets that have not been collapsed to a single id. Popdata refers to this term as "restating" and we have adopted the same terminology here. Restating typically happens with popdata. In this case, however, that work has been passed along to the research team. The step by step process is as follows:

- Find all unique combinations of studyid and clientid in the bchousing_hifis2017-2020.hifis_extra_clntid_popid_xlk.A.dat.gz file
- Find all unique combinations of studyid and clientid in the bchousing_hifis2017-2020.hifis_clients.A.dat.gz file
- Find all rows where clientid is equal in both extra_id and clients data
- Remove rows where extra_id studyids and clients ids are already the same

This process results in a crosswalk table that allows us to identify duplicated ids outside of the BC Housing data and associated them with the canonical studyid from the BC Housing data. This process is informed by popdata and is their recommend approach. The restating process occurred with any data outside of the BC Housing data.

Geographic Data

BCH client data was provisioned with the geographic variable shelter_stay_city. SDPR data was provisioned with the geographic variable csd_name. These variables can both be considered census subdivision (CSD) because of the commonality of municipality and CSD. However, because of slight differences in the names of geographic locations between SDPR, BCH and the BCStats population estimates, a crosswalk table was created. This approach a) ensures that population estimates were assigned to the correct CSD and b) resolves slight naming differences between the SDPR and BCH geographic data sets.

Manipulations: A movement flag was derived for a given geography. This binary flag specified when a person (at monthly or annual time scales) moved outside of a given geographic area. For example, if a person moved between or among census divisions during a year, they would receive a flag of "multiple cd".

SDPR data

Definition

- Homeless: At least 3 months of consecutive income assistance with the nfa == 1 for a given month during a 12 month period.
- Attributes (hl_category)
- Non-chronic homelessness: 3 to 5 months of consecutive income assistance with the nfa == 1 during the past 12 months for any given month in a given year.

Chronic homelessness: 6 or months of consecutive income assistance with the nfa == 1 during the past 12 months for any given month in a given year.

Manipulations: The nfa data was linked with the involvement data using clientid as a primary key to add studyid to the nfa data. The merged nfa-involvement data was subset for only rows where nfa == 1. Each row represents a monthly income assistance payment. We calculated the number of months of consecutive income assistance with an nfa == 1 for a given year for a given month. This allowed for monthly assignment of the hl_category attribute. To assign an annual attribute of hl_category, the occurrence of any chronic homelessness status during any month resulted in an annual status of chronic homelessness.

BC Housing data

Definition

- All individuals who spent any time in a shelter
- Attributes (hl_category)
- Non-chronic homelessness: all individuals who spent 180 or fewer days in a shelter or who had 1 or 2 shelter visits separated by 30 days during the past 12 months for any given month in a given year.
- Chronic homelessness: all individuals who spent more than 180 days in a shelter or who had 3 or more shelter visits separated by 30 days during the past 12 months for any given month in a given year.

Manipulations: The shelter attribute was calculated on a monthly basis. For a given month, shelter visits were subset into that month and the preceding 11 months to assign the hl_category attribute. Visits separated by 30 days and the cumulative number of days spent in a shelter were calculated and the BC Housing definition was applied to classify individuals as experiencing non-chronic homelessness or chronic homelessness. In instances where cumulative number of days and number of shelter visits resulted in different classifications, chronic homelessness was taken as the classification. To assign an annual attribute of hl_category, the occurrence of any chronic homelessness status during any month resulted in an annual status of chronic homelessness. A small portion of the BCH shelter data contained individuals with overlapping time intervals at different locations. That is, some people are registered in different shelters at the same time. Those overlapping intervals were merged to create one continuous interval. This was possible when the overlapping intervals were in the same city. For instances where an individual was present in two cities, the intervals were not merged. This is a very small number of shelter visits and therefore represents an acceptable loss of data accuracy.

Demography data

Age data

Age data was derived from three ranked sources. Age for an individual was chosen in the following order:

- B.C. Central Demographic file: most authoritative source for accurate birth data.
- SDPR data: largely also derived from registry data but also providing more values due to slightly more complete data
- BCH data: least accurate age data as it is entirely self reported

Gender data

We established a hierarchy of our data source:

- BCH data collects gender using open ended response and is therefore the priority gender indicator.
- SDPR income assistance data, which collects binary gender, and is the most recent data is therefore the next priority.
- B.C. Central Demographic file, which also collects binary gender, is the least current data source and therefore, while still very reliable, is the lowest priority.

Manipulations: Studyids with two birthdays were removed from the demographic data. To address concerns of statistical disclosure, the category of "Non-Binary" was ultimately removed from the dataset. All instances of this category of gender were in the BCH data and therefore replaced by either SDPR and demographics data. Age was calculated as the age as of December 31, of a given year which is the end of the accrual window and then rounded down to the nearest integer.

Merging BC Housing and SDPR data

- Data was merged in the primary studyid key keeping track of whether an individual was present in the BC Housing data, SDPR data or both.
- In instances where the BC Housing and SDPR resulted in different homeless classifications, chronically homeless was chosen.
- To define the annual data_source variable, any service usage (from either SDPR or BCH data) resulted in a data_source flag for that individual.

References

Albers, Sam, and Stephanie Hazlitt. 2020. *Dipr: Provide Functions to Efficiently Import SRE Data*.

Landau, William Michael. 2021. "The Targets r Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing." *Journal of Open Source Software* 6 (57): 2959. https://doi.org/10.21105/joss.02959.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Jonathan Keane, Romain François, Jeroen Ooms, and Apache Arrow. 2021. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

von Bergmann, Jens, and Dmitry Shkolnik. 2021. *Cansim: Accessing Statistics Canada Data Table and Vectors*. https://CRAN.R-project.org/package=cansim.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.