**Preliminary Form to Add Social Determinants to CSDUL**

**Request date (2025-07-13):**

|  |  |
| --- | --- |
| **Researcher (name and affiliation):** | Sean Leipe, McMaster University  Jackie Villanueva, McMaster University |
| **Node Lead (name and affiliation):** | Marisa Young, McMaster University |
| **Indicator or Model Name:** | Family, Recreation, & Safety Resources (FRSR) |

**Purpose of the document**

This document includes several questions that must be answered by the researcher interested in adding indicators or models into CSDUL. These questions pretend to briefly explain the mathematical and theoretical framework of the indicator or model being incorporated. The researcher must be able to fill out every question clearly and concisely, supporting their explanation with respectable academic sources.

The document will be added to the model or indicator documentation in CSDUL-OUT and CSDUL-RDC. It must serve as a quick and straightforward introduction to the indicator or model for anyone interested and give relevant references to guide the learning process to other researchers.

**To be completed by the responsible analyst.**

**If there are questions that cannot be answered because of the nature of the indicator/model, write N/A.**

**You can support your completion using the example document located in this link:** [**Documents - Add inputs to CSDUL - 02 - Example.docx - Google Docs**](https://docs.google.com/document/d/1t4_Bh5pRtHzd8GQ3ifJWY2zjjjch8DFf/edit)

1. **Will you share the inputs through CSDUL-RDC, CSDUL-OUT, or both? CSDUL-OUT**

1. **Explanation of the indicator/model.** 
   1. **In simple words, explain what the indicator/model to be added consists of.**

The indicators for Family, Recreation & Safety Resources (FRSR) represent community amenities and resources that support families and safety. *Education resources* include institutions like schools, colleges, and universities, which are identified using codes specific to educational facilities. *Food stores* cover various food distribution locations, such as grocery stores and food depots, where residents can purchase food. *Recreation and Entertainment* represent places for sports, entertainment, and outdoor activities, such as parks, swimming pools, and amusement parks. *Police and Security Services* include locations providing police and security support, such as police stations, while *Fire Protection Services* encompass facilities like fire stations that offer fire-related emergency assistance. These indicators collectively capture key resources and amenities that contribute to family well-being, community support and safety. These data come from the Digital Map Technologies Inc. (DMTI), which is a Canadian company specializing in geospatial data and mapping solutions. It provides high-precision location-based datasets widely used in business, government, and academic research. The raw data are accessible to any researcher at a university that holds a Scholars GeoPortal license.

* 1. **Are there assumptions associated with the indicator/model? If there are, please briefly describe them.**

Yes, the indicators associated with the FRSR do involve several assumptions:

*Proximity and Accessibility:* A key assumption is that the presence of these resources within a certain geographic area (e.g., census divisions) means they are accessible to residents. This assumes that closer proximity directly translates to access and usability for families, which may not account for individual mobility limitations or other barriers.

*Classification, Comparison, and Accuracy:* The FRSR indicators are based on Standard Industrial Classification (SIC) codes to categorize resources, assuming these codes are consistently accurate in representing each resource type. SIC codes are a system for categorizing establishments based on their primary economic activity, historically used for economic censuses, business directories, and resource mapping, and still common in datasets like those used here. They provide a standardized way to identify and group organizations such as schools, grocery stores, or clinics.

SIC codes vary compared to other forms of classification. For example, the Canadian Classification of Functions of Government (CCOFOG) organizes government expenditures by their purpose (e.g., education, health, housing) rather than by the type of establishment. While SIC focuses on industries and producers, CCOFOG focuses on how public funds are allocated across functional areas of government activity.[[1]](#footnote-1)

Despite the benefits of SIC codes, note that there can be classification errors or limitations in how specific types of facilities are coded, as outlined in Young et al. (2023). Please see Section 2.9 for additional information on these potential biases.

*Community as a Source of Social Infrastructure:* The indicators assume that the community resources also serve as “social infrastructure,” fostering social bonds, safety, and collective efficacy. This assumption is based on Klinenberg's (2018) model, which posits that community spaces help residents form networks, enhancing mutual support and reducing individual stress.

These assumptions underlie the FRSR framework's approach to understanding how community resources contribute to enhancing health and well-being for residents.

* 1. **How is the indicator/model derived? Support your explanation with formulas when possible.**

The FRSR indicators are derived through a structured approach involving data collection, classification and counting across census tracts.

*Data Collection and Classification*: FRSR indicators are based on data from DMTI Spatial, which provides information on various community resources across Canada. Resources are categorized using Standard Industrial Classification (SIC) codes, which identify specific resource types like education facilities, food stores, recreation facilities, and police and fire stations. This classification system allows each relevant point of interest to be categorized according to the FRSR resource types.

More information on DMTI Spatial and SIC codes can be found [here](https://drive.google.com/file/d/1emOEWLaWxJeFzedKeXp3pVlLs7z3XUmz/view).

*Counting Resources by Census Tract:*  For each census tract, the number of resources in each category (e.g., education facilities, food stores) is tallied. These counts provide the baseline measure for each type of family and / or safety resource within the census tract.

We include the following recommendations to future researchers using these indicators:

*Standardizing Counts by Population:* To enable comparisons across census tracts with different population sizes, each resource count should be standardized per capita. This creates a measure of resources available per population unit within each tract, making the data comparable across areas with varying population densities.

*Z-Score Transformation:* For those wishing to use these data for descriptive and simple comparison analysis, we recommend that the standardized counts for each resource category be transformed into z-scores, which normalize the data across census tracts. This transformation ensures that each resource type is centred around the same scale, allowing for direct comparisons within and across census tracts.

* 1. **What is the unit of analysis of the indicator/model? (e.g. households, persons, cities)**

The unit of analysis is the census tract.

* 1. **How can the indicator be integrated with other datasets?**

Using the census tract identifier, counts can be aggregated to different levels of census geography through a tabular join. Any data that has been collected at (or aggregated to) the census tract level or coarser can be joined to this dataset by CTUID. We would not be able to get FRSR counts down to a finer level of geography from census tract, but could extract using finer boundaries (e.g. dissemination area)

* 1. **What are the boundaries of the indicator/model?**

Values can range from zero into the hundreds. However, we did not detect any census tract with

more than 100 or any given resource type. Value ranges may vary if researchers choose to standardize counts by population.

* 1. **If you want to add a model to CSDUL, is this associated with a hypothesis? If yes, please describe their:**
     1. **Null hypothesis**
     2. **Alternative hypothesis**
     3. **The implications of rejecting the null hypothesis**
  2. **What is the interpretation of the values of the indicator/model?**

Higher values equal greater resources of that given classification.

* 1. **Based on the literature and your experience working with this indicator/model, is it possible to identify weaknesses in its calculations or assumptions? To facilitate your answer, you can focus on:**

1. **Potential biases**
2. **Overestimation**
3. **Underestimation**
4. **Omitted variables**
5. **Endogeneity**
6. **Datasets’ problems**

There are limitations to our indicators related to validity and reliability. To assess these

weaknesses, we conducted spot checks for each SIC code within and across each resource type.

Due to limited RA hours, we were only able to spot check two census tracts across each

year of data per SIC code. Based on these spotchecks, we provide a series of [queries and recommendations](https://docs.google.com/document/d/1lycc3JaAVn4PNslXmE6TZal_w3pEEPt2/edit?usp=drive_link&ouid=115043681076556168159&rtpof=true&sd=true) (see attached doc). Note, we have included datasets with indicators in counts in their original form, as well as filtered datasets that incorporate these recommendations.

Below, we summarize the weaknesses we found from these checks in accordance with your listed criteria. We also provide other limitations based on our assumptions.

*Potential Biases.* Broad or heterogeneous SIC groupings (8211, 8221, 8299, 7999) and inclusion of Quebec‑only 8222 introduce systematic classification and regional biases—e.g., gambling or lottery outlets counted as “recreation,” campus sub‑buildings counted as separate institutions, and Quebec’s junior colleges inflating provincial education counts.

*Overestimation.* Education resources are overstated by duplicated primary/secondary listings, fragmented postsecondary campuses (many buildings coded as separate colleges/universities), misclassified therapeutic or niche training sites (8299), gambling/lottery outlets inflating recreation (7999), and stacked or duplicate points when not collapsed.

*Omitted Variables/Information.* The indicators omit qualitative and distal aspects (capacity, quality, suitability, proximity).

1. **Does the indicator/model have other mathematical or computational versions (not syntax) to build it? (provide references)**

Yes, the Python programming language is used to build the indicator. Shapefiles containing point records with associated x/y coordinates and categorical information are first downloaded from Scholar’s GeoPortal, as provided to the library by DMTI Spatial. Given a list of SIC codes corresponding to different types of FRSRs and an additional shapefile of census geography boundaries in (multi)polygon format, points corresponding to each code are selected and counted within each polygon of the boundary shapefile. The scripts provide a CSV file containing the amount of points corresponding to each FRSR category within each census geography polygon.

We have attached the [program](https://docs.google.com/document/d/12sF2ND85ZLBVjetXMWRiwyUYXk6R7HWrRMnWOeGivTI/edit?usp=drive_link) for future researchers’ reference.

* 1. **Why are you building the indicator/model as you propose? Are there advantages compared to other versions?**

There are several decisions made apriori in the construction of these indicators that we justify below:

1. *The use of DMTI data.* These are the most comprehensive data for researcher access across Canada. These data are mapped using DMTI’s CanMap platform, ensuring accuracy and comprehensive coverage. It provides high-precision location-based datasets widely used in business, government, and academic research. The CanMap platform is refreshed quarterly for universities (via the Scholars GeoPortal licence) and monthly for enterprise users, giving rooftop-accurate layers that plug straight into ArcGIS, QGIS or Stata/GeoDa workflows. The upside for us and CSDUL is then quite apparent: instead of relying on coarse census proxies, we can model “real” community infrastructure at metre resolution.
2. *The use of SIC codes versus other industry classification codes*. In our prior study, (Young et al. 2024) we found that EPOIs contained far fewer missing primary SIC codes than NAICS codes. We also found that many food and education locations appearing in the dataset, for example, lacked a NAICS code yet still had a valid SIC code, demonstrating that a NAICS‑only pull would have underrepresented actual on‑the‑ground resources.
3. *The distribution of raw counts of resources.* We provide recommendations above of how to alternately standardize the FRSR counts for comparison across the selected geographical boundaries. However, we leave those decisions to the researcher using the data and provide recommendations, only with the data in raw form.
4. **Do you see potential improvements for the indicator/model? This could involve using other datasets, refining calculations, or modifying assumptions, among others.**

We note the following improvements could be undertaken to ensure greater validity of our FRSR indicator.

1. Additional spotchecking would be helpful as we add other SIC codes as indicators.
2. Secondary validations of proximity of resources to residents based on alternate data sources could be undertaken (i.e., case studies of certain census tracts). Following from these secondary sets of analyses, an average “error term” could be incorporated into the FRSR indicators to correct for distance / proximity of residents to a given resource.
3. When modelling these indicators with individual-level resident data, incorporate FRSR from contiguous (adjacent) neighbourhoods, since residents’ effective proximity to resources often extends beyond the boundaries of their own census tract and may depend on amenities located just across a tract line.
4. The 2014 data are missing from DMTI Spatial Inc. After inquiring, it appears they simply didn’t collect data this year. It might make sense to average the values from 2013 and 2015 to get counts for 2014 across the FRSR.
5. Duplicate entries have been removed, with deduplication based solely on address remaining under consideration as some sites may host multiple distinct resources, found in overcounts for SIC 8221 (postsecondary institutions).
6. As well, keyword-based filters have been introduced for SIC 7999, with points containing “lottery” or “lotto” excluded to prevent the conflation of gambling-related services with socially beneficial recreation.
7. Further refinement is to be explored for SIC 8299, where misclassified therapeutic clinics (e.g., massage or RMT providers) may inflate these counts. Manual review may be needed to ensure valid count retainment if key-word filters are to be introduced.
8. As well, unique to the province of Quebec, SIC 8222 has been retained as having educational resources when aggregated according to the FSFR paper’s classifications. Due to this regional specificity, flagging or excluding this SIC from national composite indices may be of interest.

**What inputs are to be added to CSDUL? Write “X”**

|  |  |
| --- | --- |
| **X** | Raw or intermediate datasets required to create the indicator/model. |
| **X** | Codes that create the indicator/model (be sure that your code is clear enough to be replicated in the future for yourself or any other researcher). |
| **X** | Documentation that explains step by step the entire process that builds the indicator or model. |
| **X** | Results, which consist of the list of variables, indicators, or model results. |
| **X** | Support files. They can be papers, chapter books, codes, etc. |

**References**

Klinenberg, E. (2018). *Palaces for the People*. Crown.

Young, Marisa, Sean Leipe, and Diana Singh. 2023. “Best Practices for Measuring Community Characteristics Across Canada.” *Canadian Geographer, 68*(1): 115-28.

Young, Marisa and Diana Singh. 2024. “The Canadian Family-Friendly Community Resources for Better Balance, Health and Well-Being Study.” *Community, Work & Family, 27*(4): 472-98.

1. COFOG has a hierarchical coding structure with three levels:

   Division (2-digit code) – Broad functions of government (e.g., *01 General public services*, *07 Health*, *09 Education*).

   Group (3-digit code) – Sub-functions within each division (e.g., *071 Medical products, appliances, and equipment* under Health).

   Class (4-digit code) – More detailed categories of activities (e.g., *0721 General hospital services*). [↑](#footnote-ref-1)