scientific data

OPEN

DATA DESCRIPTOR

A synthetic vulnerable population dataset for fine scale geographical equity analysis and urban planning

Jérémy Gelb[1]✉, Philippe Apparicio[2] & Hamzeh Alizadeh[1]

Assessing the social and economic vulnerability of populations within a given area is essential for conducting environmental equity evaluations and devising effective public policies to mitigate disparities. However, prevailing indicators used to measure socio-economic vulnerability exhibit several shortcomings. Primarily relying on factor analysis, these indicators face challenges in terms of comparability over time, lack of standardized scales, and inherent limitations associated with composite indicators. To address these shortcomings, we propose a novel approach that estimates the number of potentially vulnerable individuals by constructing a synthetic population. Our methodology, developed using open tools and datasets, offers a scalable solution applicable to the entire Canadian context. The resulting percentage of potentially vulnerable populations demonstrates strong correlations with traditional vulnerability indicators commonly used in Canada, while overcoming their inherent limitations. The generated dataset holds significant potential and serves as a valuable resource for both researchers and governmental organizations. It provides a robust foundation for conducting equity analyses, assessments, and policy evaluations, thereby facilitating evidence-based decision-making processes aimed at promoting social and economic inclusivity.

Background & Summary

Introduction. In this article, we describe a new methodology for quantifying the socio-economic vulnerability in urban areas. Specifically, we propose to directly measure the number (or percentage) of people in a potential situation of socio-economic vulnerability. This indicator can be used to study the impacts of urban projects and policies in terms of environmental equity in a much more intuitive way when compared to the classical vulnerability indices. Moreover, the proposed method addresses several common issues related to the classical vulnerability indices, namely the absence of an interpretable measurement scale, the challenge of temporal comparability, the addition and substitution effects, and the arbitrary weighting of contributing factors[1,2]. These challenges are extensively discussed in the following sections. To attain this objective, we create a synthetic vulnerable population that focuses on simulating individuals or groups characterized by a heightened potential for vulnerability in Canada. This vulnerability may be associated with various factors such as economic disadvantage, ethnic minority status, or other social determinants[3,4]. This synthetic population serves as a valuable tool for researchers, policymakers, and planners to analyze and address issues related to social equity, public health, urban development and transportation without compromising the privacy and confidentiality of actual vulnerable individuals.

Synthetic populations enable researchers and practitioners to conduct analyses, simulations, and assessments without directly accessing individual-level data, thereby addressing privacy concerns while still providing valuable insights into population dynamics and behaviours. A synthetic population refers to a simulated representation of a real population created through statistical methods and modelling techniques[5]. It is designed to mimic the demographic, socioeconomic, and other relevant characteristics of an actual population while preserving privacy and confidentiality[6,7]. A second objective of the study is to validate the ability of the proposed measure to adequately represent the phenomena of socio-economic vulnerability. To this end, we evaluate the quality of the synthetic population obtained (internal validation) and assess the association of the estimated proportion of potentially vulnerable people with two well-known indicators, the Canadien index of multiple deprivation[8] and the index of material and social deprivation[9] (external validation).

[1]Autorité Régionale de Transport Métropolitain – équipe Recherche et Valorisation des Données, 1001 Boulevard Robert-Bourassa, Montréal, QC, H3B 4L4, Canada. [2]Université de Sherbrooke - département de géomatique, 2500 Boulevard de l'université, Sherbrooke, QC, J1K 2R1, Canada. ✉e-mail: jgelb@artm.quebec

We also describe the resulting database for the four largest metropolitan areas in Canada. This database is particularly useful for research on social vulnerability, as well as for government planning exercises aimed at environmental equity.

The data produced for this study is available for download on Zenodo[10] along with all the necessary scripts for replicating the results or generating a synthetic vulnerable population for a different region.

Vulnerable populations. Social vulnerability is a concept with blurred boundaries. Many terms have been used to describe vulnerable populations, such as *disadvantaged, underprivileged, medically undeserved, poverty stricken, socially marginalized, distressed population and underclasses*[11]. From the outset, it is important to distinguish between an individual concept of vulnerability (vulnerable individuals) and a population concept (vulnerable populations)[12]. This article specifically focuses on the examination of the population concept of vulnerability.

The genesis of the concept of population vulnerability can be traced back to the health sector, where the designation of "vulnerable populations" serves to identify individuals potentially more impacted by specific exposures, such as children, pregnant women, asthmatics, and the elderly. In this context, vulnerability is only perceived through a physiological lens, encompassing aspects of susceptibility stemming from inherent physical conditions. Subsequently, the analytical scope has expanded to incorporate dimensions of social vulnerability, with a heightened emphasis on marginalized segments of society, including individuals grappling with economic hardship, homelessness, ethnic minorities, and indigenous populations. Early work on these populations mainly focused on the consequences of this vulnerability at an individual level, emphasizing downstream factors, particularly from a health perspective. As the discourse evolved, the research paradigm has expanded to encompass a more comprehensive examination of vulnerable populations and the intricate societal mechanisms contributing to the genesis of this vulnerability. This shift in focus is notably characterized by a consideration of upstream factors, which encapsulate environmental characteristics, disparities, and educational discrimination, among other determinants[13]. This turning point in the conceptualization of vulnerability can be attributed to Aday's theoretical framework of vulnerability[3], associating individual physical dimensions, such as age, gender, ethnic origin, and disabilities, with economic-social capital factors, including income, education and housing, along with community capital aspects like relationships, family structure and social networks. Moreover, this framework underscores the significance of resource availability in delineating both physical and mental health states. This nuanced theoretical foundation marks a pivotal departure from earlier perspectives by illuminating the multifaceted nature of vulnerability and emphasizing the intricate interplay of individual and societal factors in shaping health outcomes.

The accumulation of different stressors and the lack of resources to cope with them leads to a situation of vulnerability, i.e. "a stressful social disorganization as a normative reality of life"[14]. More recently still, postmodern critical analysis has deconstructed the concept of vulnerability to propose a non-dichotomous definition based on a gradient, linking resilience and vulnerability[15] and implying that each human has his or her own level of vulnerability in his or her social context. However, this type of relativistic definition offers rather limited perspectives in practice since it does not guide decision-making and intervention prioritization[4]. The difficulty of clearly defining the vulnerability of populations has led to the adoption of an approach that Wrigley and Dawson[4] named population listing. Many studies adopt a rudimentary approach by delineating vulnerability through the enumeration of population groups, such as the youth, elderly, visible minorities, and low-income individuals, etc. While this method serves to illustrate vulnerability through exemplification, it falls short of offering a direct and explicit definition of the concept. Therefore, resulting lists exhibit a notable tendency for overlapping categories, rendering them inherently incomplete and offering merely a heuristic portrayal of the vulnerability construct. Furthermore, the observed tendency for these delineated groups to intersect, as highlighted by Shi and Stevens[11], prompts considerations of cumulative vulnerability and the intricate dynamics of intersectionality.

What these vulnerable populations have in common is that they have greater needs than the rest of the population for various resources, or at least potentially suffer more from a limited access to these resources[16]. In short, a relatively broad definition of the social vulnerability of populations could therefore be a fusion of social, cultural, economic, political, and institutional processes. These intricate elements collectively shape socioeconomic differentials, influencing both the exposure to and the subsequent recovery from various hazards[17].

Nowadays, there are many areas of public intervention that consider the dimension of vulnerable populations. In addition to health, these include transport, and more specifically, mobility[18,19] and accessibility[20,21]. Add to this the issue of housing[22], green spaces[23], exposure to various nuisances[24,25], or food deserts[26].

The cross-disciplinary use of vulnerability and its application in diverse contexts underscores the necessity for a more nuanced and comprehensive approach to elucidate its multifaceted nature. It also highlights the relevance of having tools to identify vulnerable populations and measure their geographical distribution. This need is crucial both for the academic discourse, notably to enrich the growing literature on environmental equity, and for the practical sphere to aid decision-making and policy implementation.

Measuring vulnerability. The considerable diversity and heterogeneity of the definitions of vulnerability within the scientific literature are mirrored by an equally diverse landscape in its definition and measurement methods within the realm of public policy. While not exhaustive, we endeavour to provide a succinct overview of population vulnerability indicators, drawing from both Canadian and international contexts, used both in the academic and practical spheres. In fact, there are several articles presenting indicators of socio-economic vulnerability, but their use in urban planning is rather limited. For a more in-depth review of the literature on this topic, we refer the reader to the recent review of Mah *et al.*[27]. The five indicators presented below are very similar to those examined in this literature review. They include the main identified domains (at risk populations, education, micro-level socioeconomic status, older population, household composition, employment, housing, etc.). The authors also conclude that the indicators reviewed are very similar in terms of domains included and methodology.

	Canadien Index of Multiple Defavorisation (Canada)	Index of material and social deprivation (Quebec)	French deprivation index (France)	Social Vulnerability Index (USA)	Equity Index of Living environments (Montreal)
age	Proportion of 65 years or older			Proportion of 65 years or older	
	proportion of 14 years or younger			Proportion of 17 years or younger	
ethnocultural background	Proportion of the population belonging to a visible minority				Proportion of the population belonging to a visible minority
	Proportion of the population belonging to Indigenous people			Proportion of the population belonging to a visible minority	Proportion of the population belonging to Indigenous people
	Proportion of immigrants				Proportion of recent immigrants
Household composition	Proportion of single-person households	Proportion of 15 year or older living alone		Single-parent household with children under 18	Proportion of single-person households
	Proportion of the adults being married or living in common law couples	Proportion of 15 year or older separated, divorced or widowed			
		Proportion of single parent households			
Income and occupation	Proportion of the active population	Proportion of 15 year or older employed	median income of households	Persons below 150% poverty	Proportion of low income based on Market Basket Measure poverty thresholds
	Ratio between jobs and population	Median income for 15 year or older	Proportion of labourer in the active population between 15 and 64 years	Civilian (age 16+) unemployed	
	Proportion of the population receiving social assistance from the government		Proportion of unemployed people in active population between 15 and 64 years	Housing cost burden	
Education	Proportion of 25 years or older without a high school diploma	Proportion of 15 years or older without a high school diploma	Proportion of the 15 years or older with a with a high school diploma	Proportion of adults without a high school diploma	Proportion of adults without diploma
Housing	Proportion of homeowners			multi-unit structures	Core housing need
	Proportion of the population which moved in the last 5 years			mobiles homes	
	Proportion of dwellings needing major repairs			crowding (At household level	

Table 1 . Main dimensions in five widely used social vulnerability indices.

In Canada, Statistics Canada produced an indicator based on the 2016 census called the Canadian Index of Multiple Deprivation (CIMD)[8,28], which combines four dimensions: residential instability, economic dependence, situational vulnerability and ethnocultural composition (see Table 1). According to Statistics Canada, it is a geographic index of disadvantage and marginalization. Its purpose is to provide a quantitative measure of these phenomena for policy planning and evaluation, or resource allocation. Methodologically, the index is derived through the summation of scores, ranging from 1 to 5, corresponding to the four delineated dimensions, derived through a principal component analysis (PCA). It is an extension of the foundational groundwork laid by the 2006 Canadian Marginalization Index.

Similarly, the National Institute of Public Health in Quebec publishes since the late 1990s a material and social deprivation index[9]. It comprises six indicators relating to health status, material deprivation and social network fragility (see Table 1). A PCA is used to reduce the original variables to two distinct dimensions (social and material), which are individually evaluated based on a 1 to 5 scale (quintiles of the first two components of PCA). These two dimensions can then be analyzed individually or in combination to create categories of bivariate vulnerability.

In France, INSERM[29,30] offers an index of social disadvantage based on a PCA on four socio-demographic variables (namely, unemployment rate, proportion of blue-collar workers in labour force, proportion of adults with a high school diploma, median income of households). However, only the first factor from the PCA is retained.

In the United States, the Agency for Toxic Substances and Disease Registry employs the Social Vulnerability Index[31]. It combines 16 variables encompassing socio-economic and ethnic dimensions, household characteristics, as well as housing and transportation. These variables are converted into percentiles before being added together to obtain a final vulnerability score. This index holds the same name as the indicator proposed by Cutter *et al.*[32], which is based on 42 variables combined in 11 factors by a PCA. The latter has had a major impact in the scientific literature, but also in public policy in the United States[17].

More recently, the City of Montreal[33] has also published an equity index of living environments. This multidimensional indicator is designed to guide public intervention in neighbourhoods where socio-environmental vulnerability is prevalent. It combines 23 variables, grouped into six dimensions: social vulnerability (4 variables), economic vulnerability (2), environment (5), accessibility to urban resources (6), accessibility to cultural, sports and

leisure resources (3), urban safety (3). The variables of each dimension are introduced into a PCA, in which only the first factor is retained. These synthetic variables were then converted into quintiles, and the final score for each sector of analysis is obtained by counting the number of dimensions for which this sector belongs to the last quintile.

These index construction approaches have several advantages like their ability to reflect the multidimensional aspect of vulnerability and to measure it along a gradient. In fact, they assume that vulnerability is a latent variable and therefore not directly observable. It must therefore be reconstructed using proxies. On the other hand, these composite indicators have a number of important limitations[1]:

- The absence of an interpretable measurement scale: These scores do not allow for natural interpretation and cannot be compared through mathematical operators. Consequently, asserting that a statistical sector is «X times» more vulnerable than another according to these indicators lacks substantive significance.
- The challenge of temporal comparability: This emanates from the reliance on factor analysis and various stages of data standardization in deriving these scores. Whenever new data is considered, these steps affect the new indicator scale, making it difficult to directly compare with previous results.
- Addition and substitution effects: Since these scores aggregate various dimensions, they suggest that a high score in one dimension can offset a low score in another dimension. This unintended consequence is undesirable as it may obscure disparities or deficiencies in specific dimensions, compromising the precision and comprehensiveness of the assessment.
- Arbitrary weighting: Given the aggregation of multiple variables in these scores, a weighting mechanism is imperative. The selection of such weights (or the absence of weights) may hinge on statistical, empirical, or simply guided by judgment, but remains debatable in all cases.

These limitations have a direct impact on the quality of composite vulnerability indicators as planning tools. While their primary strength lies in pinpointing vulnerable sectors within a specific area at a particular moment, their efficacy diminishes when employed to track temporal progress or establish project or policy targets. In other words, the multiple limitations of these indicators hinder their applications in real-world usages. We can, however, observe a form of consensus in the identification of potentially vulnerable populations for these indicators. This is consistent with the "listing" strategy identified by Wrigley and Dawson[4] and mentioned in the previous section. The Table 1 lists the most common population groups used to construct these indicators.

Article contribution. Considering the limitations outlined in the previous sections, we propose a different approach to the spatial measurement of social vulnerability. Instead of the conventional practice of formulating vulnerability scores, we identify and quantify vulnerable populations. We therefore seek to determine for a geographical entity the number of people in a potential situation of socio-economic vulnerability, rather than calculating a socio-economic vulnerability score. This approach clearly tackles the limitations outlined above.

- Firstly, looking at the number or share of people facing a situation of socio-economic vulnerability is a clear way to understand and interpret the situation.
- Secondly, it is easy to compare these indicators over time to see how things are changing.
- Thirdly, they can be used to effectively measure the impact of projects or policies and set achievable goals for improvement.

In Canada, we can work with census microdata, which provides an exhaustive, fine-scaled data set. However, the access to the complete microdata is limited, due to privacy rules that protect the confidentiality and anonymity of people who respond to the census. Instead, Statistics Canada releases only a representative sample of this detailed data as open data. As an alternative, we suggest utilizing this open data to create a synthetic vulnerable population.

In the following sections, we elaborate on the approach employed to build a synthetic vulnerable population for four major Canadian metropolitan areas, Montreal, Vancouver, Toronto and Calgary. We limited our analysis to four regions to facilitate the presentation of the results. The regions of Montreal, Vancouver, Toronto and Calgary were chosen because they are the four biggest Canadian metropolitan areas and because they face different realities in terms of demographics, immigration and socioeconomic vulnerability[34,35].

This construction relies on the utilization of the Public Use Microdata File (PUMF), and the Census Profiles, which constitutes a conventional compilation of census data at the dissemination area level. We then evaluate the results obtained by performing an internal validation (adjustment quality of the synthetic population) and an external validation (in comparison with the existing vulnerability indices). Finally, we conclude by illustrating the advantages and limitations of using a vulnerable synthetic population in the formulation, evaluation and monitoring of public policies.

Methods

Data sources. We leverage two main data sources for the generation of a vulnerable synthetic population:

- First, we use census profile data from Statistics Canada[36,37] at the Dissemination Area (DA) level. These geographic units, with a population ranging between 400 and 700, represent the finest granularity at which census data are aggregated and disseminated (Census profiles).
- Second, the Public Use Microdata File[38,39] (PUMF), which contains individual data for a sample representing 2.7% of the Canadian population. The individuals are grouped by provinces and metropolitan areas.

Dimension	Census profile data	transformation	PUMF data	transformation
Visible minority	Total visible minority population in private household	None	VISMIN – Visible minority	Dichotomized: 0 if not in a visible minority, 1 otherwise
Indigenous people	Total of North American Indigenous (Indigenous identity)	None	ABOID - Indigenous: Indigenous identity	Dichotomized: 0 if not in a North American Indigenous, 1 otherwise
low education level	No certificate, diploma or degree Secondary (high) school diploma or equivalency certificate, in private households for the population aged 15 years and over	sum of the two totals	HDGREE – Education: Highest certificate, diploma or degree	Dichotomized: 1 if No certificate, diploma or degree OR Secondary (high) school diploma or equivalency certificate, 0 otherwise
low income level	In low income based on the Low-income cut-offs, after tax (LICO-AT) for the population in private households to whom low-income concepts are applicable	None	LICO_AT – Income: Low-income status based on LICO-AT	Dichotomized: 1 if Member of a low income economic family or low income person aged 15 years and over not in an economic family, 0 otherwise
People living alone	Private households by household size - 100% data, 1 person	None	HHSIZE – Household size	Dichotomized: 1 if 1 person, 0 otherwise
Single parent household (2016)*	Lone-parent census families in private households - 100% data 1 child, 2 children, 3 or more children	The number of children was approximated up to six children by adjusting a decreasing power function to each Dissemination area minimizing the total absolute difference. The estimated number of households with 3, 4, 5, or 6 children was then integerised to fit the known number of households with 3 or more children. The estimated number of children in lone-parent census families was then added to the known number of parents to obtain the number of persons in lone-parent census families	HHTYPE – Household type	Dichotomized: 1 if One-census-family household with or without additional persons: Lone parent family, 0 otherwise
Single parent household (2021)*	Parents in one-parent families & Children In a one-parent family	sum of the two totals		
Unemployed adults	Population aged 15 years and over, Unemployed	None	LFACT – Labour: Labour force status	Dichotomized: 1 if unemployed, 0 otherwise or not applicable
recent immigrants	Immigrant status and period of immigration for the population in private households, 2011 to 2016	None	YRIMM – Immigration: Year of immigration	Dichotomized: 1 if within 2011 and 2016, 0 otherwise
age	Age groups 0 to 14 years, 15 to 19 years, 20 to 24 years, 65 years and over	None	AGEGRP – Age	Dichotomized in 4 groups accordingly

Table 2. Variables used in the generation of the synthetic potentially vulnerable population.

Data from the 2016 and 2021 censuses are used for both data sources to generate a DA level synthetic populations corresponding to these two respective years. As stated earlier and illustrated in Table 1, there is a relative consensus on the characteristics of socio-economically vulnerable populations. We have retained all the dimensions present in Table 1 (age, ethnocultural background, household composition, income and occupation, education) except housing. Indeed, this dimension stands out as a material measure, whereas the other dimensions are individual or household characteristics. However, it would be possible to extend the methodology by using, for example, the variable of "core housing need". Table 2 describes these dimensions and the variable associations between census profile data and PUMF microdata variables.

It should be noted from the outset that the synthetic population we are seeking to produce does not need to be hierarchical. In other words, since we want to estimate the number of people per dissemination area in a potential situation of socio-economic vulnerability, we need to create data at the individual level without assigning each individual to a household. This simplifies the process of creating the synthetic population, eliminating the need for employing the Hierarchical PUMF file.

Data cleaning and imputation. The data from the census profiles has been pre-processed, mainly in two ways. Firstly, we ensured that neither individual categories nor cumulative sums surpassed the total population count within each DA. It should be noted that specific variables within the census profiles are derived from estimates based on the responses of 25% of participants who completed the comprehensive questionnaire. Also, the overall figures in the census data are randomly rounded to a multiple of 5. Therefore, there is a potential occurrence where the value of certain variables surpasses the estimated total population within a DA.

To mitigate this discrepancy for these variables, we converted the absolute number of individuals into percentages using their respective totals. Then, we recalculated the absolute number by applying these percentages to the total population figure for each DA. The resulting absolute numbers were then integerized to the nearest multiple of 5.

Secondly, we address the issue of completing missing data. Due to privacy concerns and the need for comprehensive data, certain values may be missing from the DAs. Instances where all variables were missing were removed from the dataset. For the remaining observations, we used an imputation method based on the 10 nearest neighbours. More specifically, each variable was converted into a ratio by relating it to the total number of people in the DA, then they were centred and reduced. The total number of people per DA was also centred and reduced. Finally, for each observation with missing data, the 10 nearest neighbours were identified through Euclidean distance calculation based on the retained characteristics. The medians of their values were computed

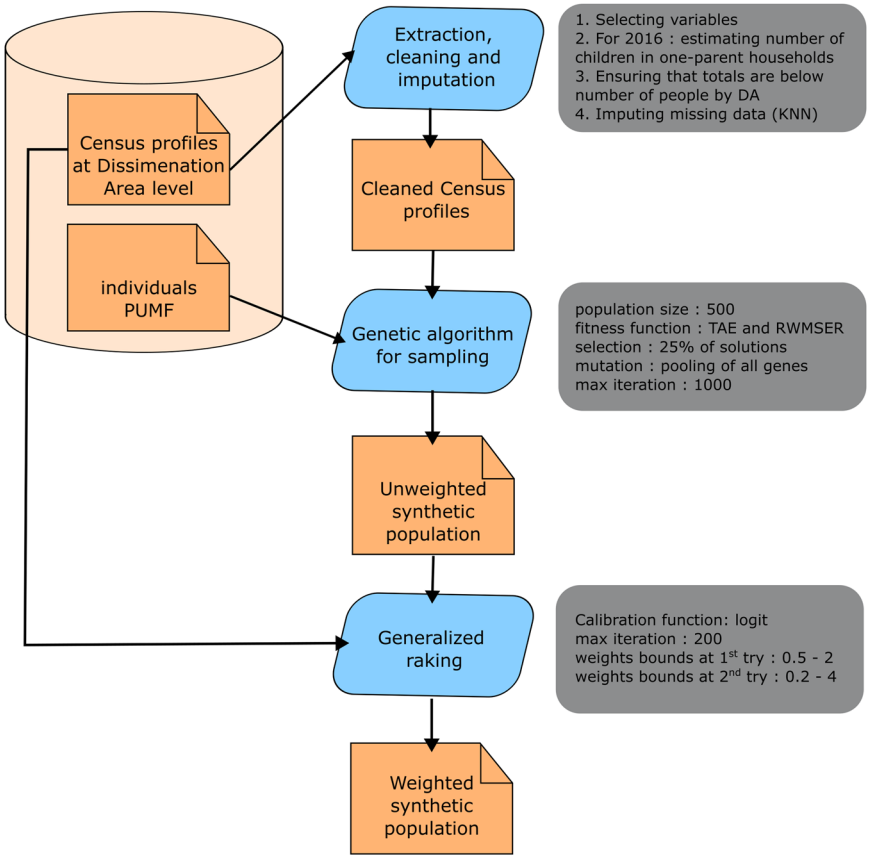

Fig. 1 Process of synthetic population generation.

and used as imputed values. These imputed values were then reverted into absolute ratios and numbers by applying the inverse transformation functions. Overall, this process impacted a very limited number of observations: 149 in Toronto (1.9%), 49 in Montréal (0.7%), 34 in Calgary (1.9%) and 49 in Vancouver (1.4%).

Generating a synthetic population. A synthetic population can be defined as a database representing agents for whom all the characteristics under study are known. They are constructed from other sources of aggregated or incomplete data and aim to reproduce a real population.

Three main families of methods are used to generate a synthetic population[5,40]: 1. Synthetic reconstruction, 2. Combinatorial Optimization and 3. Statistical Learning.

In our case, we have access to a representative subset of individuals through the PUMF data, coupled with aggregate totals derived from census profiles. Notably, our synthetic population does not need to be hierarchical. We therefore opted for a hybrid approach combining Combinatorial Optimization and Synthetic reconstruction, as recommended by Yaméogo *et al.*[40], and taking advantage of the benefits of both methods[41]. More precisely, the proposed methodology comprises two main steps:

1. Sampling by DA seeking to reproduce as closely as possible the margins of the census profile.
2. Weighting of individuals by DA to reduce any discrepancies with the targets.

The proposed methodology is summarized in Fig. 1 and elaborated below.

Combinatorial optimization methods include various optimization algorithms such as simulated annealing, hill climbing, genetic algorithms and greedy heuristics. Genetic algorithms have so far been little used for this type of exercise, despite their ability to find very efficient solutions to problems with many dimensions and to avoid getting stuck in local minima[5]. This is probably due to the need for the analyst to choose a large number of parameters for this type of model, such as the fitting function, the agent selection function, the genetic mixing function, the number of iterations, the number of agents and the learning speed. We decided to opt for this method after obtaining less satisfactory results with the simulated annealing and hill climbing algorithms.

A genetic algorithm is a heuristic optimization function directly inspired by the biological model of evolution[42,43]. It is made up of five central components:

- **The initialization step**
  It is the starting state of the algorithm where all the parameters are selected randomly or accordingly to specified priors.

- **Fitness function**
  It is responsible for evaluating the performance of the solutions obtained. It assigns a measure of how well each solution addresses the optimization goal. Solutions with higher fitness values are favoured in subsequent stages.
- **Selection**
  It is responsible for selecting a subset of solutions for each generation. Solutions are chosen based on their fitness, with higher fitness solutions being more likely to be selected, mimicking the natural selection process.
- **Crossover**
  Also known as recombination, this stage involves the exchange of genetic information between pairs of selected solutions emulating the genetic crossover mechanism.
- **Mutation**
  It introduces random changes to the genetic information of some individuals in the population, simulating genetic mutations to inject diversity. It introduces new values for the parameters at each generation and prevents the algorithm from converging prematurely to suboptimal solutions.

These components collectively form an iterative process, wherein the population undergoes cycles of evaluation, selection, crossover, and mutation. The initial population of solutions has randomly drawn parameters, so their adjustment is weak. At each iteration, the selection function chooses a set of solutions with the highest fitness scores. The parameters of these solutions are then crossed to generate a new generation of solutions, based on the best previous solutions. At each generation, the mutation function adds new values for the parameters, drawn at random, to explore the space of solutions in greater depth. After a certain number of iterations, or if the algorithm reaches a convergence criterion, the best solutions from the last generation are retained as potential solutions.

Below, we describe the key components of the genetic algorithm we have implemented to perform DA-level sampling from PUMF data.

Initialization. We initiate the process by generating 500 solutions. Each solution consists of a sample of the PUMF data, achieved with random draws. In this context, each sampled individual is a gene within its respective solution. It is important to highlight that if a particular category's margin value is zero, individuals in the PUMF data falling into this category are removed from the microdata. For example, if in a specific DA, no low-income individuals are present, all individuals in the microdata belonging to this group are removed systematically prior to the sampling.

Fitness function. We use two fitness functions to evaluate the performance of our synthetic population. The first function is the Total Absolute Error (TAE), calculated by summing the absolute deviations of each population category count (see Table 2) in the local synthetic population from the expected margins for the DA. This aggregated error is then rounded up to a threshold corresponding to 1% of the population of each DA. For example, if a DA has a total population of 500 people, then the TAE for this DA is rounded to the nearest multiple of 5. In this way, solutions with a difference in TAE less than this threshold are considered equivalent. To distinguish between these equivalent solutions and to prioritize those that yield more balanced outcomes for each variable, a second fitness function, the square root of the weighted mean of the relative squared deviations (RWMSER), comes into play.

In essence, the TAE offers a measure of overall accuracy, capturing the sum of absolute errors across variables, while the RWMSER steps in to refine the assessment, providing a nuanced evaluation that prioritizes balanced performance across variables when distinguishing between comparable solutions. This dual-fitness approach enhances the algorithm's ability to select and refine solutions, contributing to the robustness of our synthetic population generation process.

$$TAE_i = \sum_{j=1}^{k} |O_i^j - \quad M_i^j|$$

$$\mathrm{RWMSER}_i = \sqrt{\frac{\sum_{j=1}^{k} \left(\frac{|O_i^j - \quad M_i^j|}{M_i^j} \right)^2 \cdot M_i^j}{\sum_{j=1}^{k} M_i^j}}$$

Where:

- i is a specific population DA
- j is a specific category
- O_i is a vector with the counts for each category in the obtained sample for DA i
- M_i is a vector with the real counts for each category (margins) for DA i

	2016				2021			
Calgary	**TAE**	**TAE/N**	**MRE**	**RWMSRE**	**TAE**	**TAE/N**	**MRE**	**RWMSRE**
1%	11	0.016	0.018	0.015	10	0.018	0.015	0.015
25%	23	0.039	0.033	0.031	22	0.038	0.032	0.030
50%	29	0.047	0.042	0.038	28	0.046	0.040	0.037
75%	36	0.058	0.053	0.047	37	0.057	0.050	0.045
99%	95	0.086	0.099	0.081	87	0.083	0.089	0.074
Montreal	**TAE**	**TAE/N**	**MRE**	**RWMSRE**	**TAE**	**TAE/N**	**MRE**	**RWMSRE**
1%	11	0.019	0.017	0.017	10	0.019	0.016	0.017
25%	21	0.038	0.031	0.031	20	0.036	0.031	0.030
50%	26	0.046	0.040	0.037	26	0.044	0.038	0.036
75%	32	0.055	0.051	0.045	31	0.052	0.049	0.043
99%	62	0.078	0.097	0.076	64	0.075	0.100	0.071
Toronto	**TAE**	**TAE/N**	**MRE**	**RWMSRE**	**TAE**	**TAE/N**	**MRE**	**RWMSRE**
1%	11	0.018	0.015	0.013	10	0.017	0.014	0.013
25%	22	0.038	0.031	0.028	21	0.038	0.029	0.027
50%	27	0.046	0.041	0.036	27	0.045	0.038	0.035
75%	36	0.055	0.052	0.044	36	0.054	0.049	0.043
99%	105	0.097	0.103	0.076	114	0.091	0.098	0.071
Vancouver	**TAE**	**TAE/N**	**MRE**	**RWMSRE**	**TAE**	**TAE/N**	**MRE**	**RWMSRE**
1%	13	0.021	0.017	0.015	12	0.021	0.016	0.015
25%	24	0.041	0.032	0.029	24	0.039	0.031	0.028
50%	29	0.048	0.041	0.036	29	0.046	0.039	0.034
75%	36	0.057	0.052	0.044	37	0.053	0.048	0.042
99%	80	0.090	0.088	0.072	82	0.087	0.085	0.068

Table 3. Adjustment of the synthetic populations before weighting.

Selection. Our selection function orders the solutions for each generation according to the rounded TAE and the RWMSER, then retains the first quarter of the best solutions.

Crossover and mutation. Following the selection process, we calculate the overall score for each solution. This score is derived by taking the reciprocal of the solution's rank, as determined in the selection phase, and subsequently dividing it by the maximum among the reciprocals of the ranks. The resulting values, denoted by w, are then assigned to the respective sampled individuals. The individuals present in the different solutions are then grouped together in a single set denoted by E. To augment this set, E is completed by a random selection of individuals from the PUMF data (mutations), representing 35% of the total number of individuals present in E. These individuals added to the data are assigned a value of w equal to the mean of w in E. Finally, E is used to generate a new generation of 500 solutions by randomly drawing individuals from E, with a sampling probability proportional to w. This approach also known as gene pool recombination allows for the evolution of the population of solutions as a whole rather than its individual members[44].

Convergence criterion. To limit the number of iterations, we have determined a convergence criterion. This involved assessing, over the preceding five iterations, whether both the TAE and the RWMSER failed to exhibit improvement beyond their respective tolerance thresholds. Specifically, the tolerance thresholds were set at 1% of the DA population for the TAE and 0.001 for the RWMSER. Additionally, the maximum allowable number of iterations was limited to 1000.

Calibrating the obtained synthetic population. One of the most common methods for adjusting a sample to align with the known characteristics and margins of a population is the Iterative Proportional Fitting Procedure[45]. This method is often used in multi-agent modelling exercises, during which the weights are converted to integers, and individuals are duplicated according to these integer values to construct a synthetic population. The classic IPFP method has been the subject of several proposed enhancements. In particular, modifications have been suggested to incorporate a hierarchical element into its fitting process. The iterative proportional updating method[46], and the hierarchical iterative proportional fitting[47] are examples of such adaptations. Additionally, a family of approaches known as Generalized Raking[48], offers an alternative perspective. It formulates the problem of assigning weights as a constrained optimization exercise, penalizing deviations from a weight of 1. Consequently, solutions obtained avoid a recurring problem with classical IPFP, which may assign disproportionately large weights to specific observations.

In our context, we opted for the Generalized Raking method since our synthetic population does not need to be hierarchical. In addition, this procedure is applied after the preselection of observations conducted by our genetic algorithm. It is therefore essential to penalize weights that are far from 1, given that the pre-selected individuals closely align with the target margins for each DA.

	2016				2021			
Calgary	**TAE**	**TAE/N**	**MRE**	**RWMSRE**	**TAE**	**TAE/N**	**MRE**	**RWMSRE**
95%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
99%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
99.90%	5	0.021	0.023	0.020	2	0.005	0.002	0.002
Montreal	**TAE**	**TAE/N**	**MRE**	**RWMSRE**	**TAE**	**TAE/N**	**MRE**	**RWMSRE**
95%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
99%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
99.90%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
Toronto	**TAE**	**TAE/N**	**MRE**	**RWMSRE**	**TAE**	**TAE/N**	**MRE**	**RWMSRE**
95%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
99%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
99.90%	11	0.011	0.004	0.005	8	0.011	0.003	0.005
Vancouver	**TAE**	**TAE/N**	**MRE**	**RWMSRE**	**TAE**	**TAE/N**	**MRE**	**RWMSRE**
95%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
99%	0	0.000	0.000	0.000	0	0.000	0.000	0.000
99.90%	6	0.011	0.014	0.015	6	0.012	0.003	0.005

Table 4. Adjustment of the synthetic populations after weighting.

	Calgary		Montreal		Toronto		Vancouver	
	2016	**2021**	**2016**	**2021**	**2016**	**2021**	**2016**	**2021**
Definition 1	69.8%	73.7%	66.9%	68.6%	79.6%	82.3%	79.9%	82.3%
Definition 2	33.0%	34.3%	33.8%	33.9%	43.6%	44.9%	43.7%	44.3%
Definition 3	10.5%	11.1%	13.2%	11.9%	17.2%	16.6%	17.0%	15.8%
Definition 4	33.9%	34.9%	34.7%	34.4%	44.3%	45.3%	44.7%	44.7%
Definition 5	14.3%	13.5%	17.5%	14.5%	21.5%	18.9%	22.4%	18.8%

Table 5. Proportion of potentially vulnerable population for each proposed definition.

In this study, the reweighting process is initiated for each DA, employing initial weight limits of 0.5 and 2. If convergence is not achieved for a specific DA, the procedure is repeated using weight limits of 0.2 and 5. It should be noted that the weights obtained are retained even if the algorithm fails to converge. The 10 best samples selected by the genetic algorithm are reweighted and the weighted sample with the best RWMSER is selected as the final solution.

Data Records

The potentially vulnerable synthetic population generated for Montreal, Calgary, Vancouver and Toronto CMAs is available on Zenodo[10]. The data is published with the code required to reproduce the results.

The data is provided at the individual level (one line per individual) as *parquet* files (binary data readable with several softwares like R[49]) to reduce the size of repository. At the DA level, the data is provided as a *gpkg* file and can be open with any GIS software. The data is available for 2016 and 2021.

The file *pop_vuln_AD.gpkg* contains the estimates of the number of potentially vulnerable people at the DA level. It contains 8 layers: *Calgary_2016, Montreal_2016, Toronto_2016, Vancouver_2016* and the same four layers of 2021. They contain a set of geographic identifier columns (PRIDU, DRIDU, SDRIDU, ADIDU, …) that can be used to join this dataset with other StatCan data. The columns *vuln1, vuln2, vuln3, vuln4*, and *vuln5* are the estimated number of potentially vulnerable people. Similarly, the columns *vuln1_prt* to *vuln5_prt* are the estimated proportions of potentially vulnerable people. The numbers in the columns' names refer to the five definitions tested in the next section to determine whether a person should be considered as potentially vulnerable. We recommend using the definition 5 if the user does not want to use their own definition.

The individual data are stored in *.parquet* files like *synth_pop_weighted_2016_Calgary*. The name and the year are always specified in the same way. The columns available are:

- *ADIDU*, the geographical identifier linking each individual to its DA
- *w*, the weight given to the individual for calibration;
- *age014, age15_24, age25_64, age65p*, a set of binary variables indicating the age group of the individual;
- *minoritevis*, a binary variable indicating if the individual is a member of a visible minority;*aboriginal*, a binary variable indicating if the individual declared itself as 'First Nations people, Métis or Inuit';
- *immig_recent*, a binary variable indicating if the individual has immigrated in Canada during the last five years;
- *loweduc*, a binary variable indicating if the individual has a level of education equivalent or below secondary;
- *unemployed*, a binary variable indicating if the individual is unemployed;
- *lone_parent*, a binary variable indicating if the individual is a member of a household with only one parent;

	% of the population potentially vulnerable (2016)	Canadian Index of Multiple Deprivation 2016				
		Economic dependency	Residential instability	Situational vulnerability	Ethnocultural composition	Global score
Calgary	definiton 1	0.131	0.413	0.46	0.757	0.733
	definiton 2	0.126	0.483	0.563	0.76	0.813
	definiton 3	0.148	0.495	0.583	0.709	0.814
	definiton 4	0.125	0.495	0.565	0.753	0.82
	definiton 5	0.102	0.586	0.576	0.658	0.83
Montreal	definiton 1	0.453	0.662	0.546	0.685	0.842
	definiton 2	0.445	0.731	0.596	0.722	0.864
	definiton 3	0.387	0.741	0.585	0.724	0.806
	definiton 4	0.435	0.75	0.6	0.716	0.871
	definiton 5	0.314	0.83	0.59	0.674	0.815
Toronto	definiton 1	0.276	0.276	0.443	0.813	0.631
	definiton 2	0.368	0.419	0.573	0.796	0.762
	definiton 3	0.362	0.507	0.621	0.725	0.772
	definiton 4	0.367	0.436	0.576	0.79	0.771
	definiton 5	0.322	0.601	0.61	0.688	0.794
Vancouver	definiton 1	0.297	0.232	0.428	0.796	0.71
	definiton 2	0.365	0.317	0.491	0.781	0.77
	definiton 3	0.377	0.374	0.484	0.744	0.767
	definiton 4	0.365	0.331	0.49	0.776	0.775
	definiton 5	0.322	0.481	0.44	0.696	0.758
		Canadian Index of Multiple Deprivation 2021				
	% of the population potentially vulnerable (2021)	Economic dependency	Residential instability	Situational vulnerability	Ethnocultural composition	Global score
Calgary	definiton 1	0.404	0.779	0.199	0.560	0.762
	definiton 2	0.495	0.754	0.205	0.660	0.844
	definiton 3	0.500	0.696	0.212	0.677	0.830
	definiton 4	0.498	0.749	0.206	0.657	0.844
	definiton 5	0.557	0.659	0.192	0.661	0.832
Montreal	definiton 1	0.604	0.741	0.411	0.522	0.831
	definiton 2	0.674	0.739	0.437	0.575	0.857
	definiton 3	0.681	0.718	0.385	0.567	0.799
	definiton 4	0.685	0.737	0.431	0.571	0.859
	definiton 5	0.759	0.688	0.316	0.537	0.795
Toronto	definiton 1	0.231	0.846	0.281	0.503	0.640
	definiton 2	0.382	0.826	0.382	0.611	0.772
	definiton 3	0.456	0.750	0.404	0.633	0.778
	definiton 4	0.388	0.824	0.383	0.609	0.775
	definiton 5	0.502	0.733	0.386	0.604	0.782
Vancouver	definiton 1	0.169	0.830	0.336	0.372	0.678
	definiton 2	0.254	0.808	0.403	0.447	0.758
	definiton 3	0.309	0.742	0.423	0.431	0.756
	definiton 4	0.260	0.806	0.403	0.441	0.759
	definiton 5	0.361	0.723	0.391	0.366	0.740

Table 6. Correlation at the DA level between each definition and the CIMD.

- *lone_people*, a binary variable indicating if the individual is a member of a household with only one person;
- *low_income*, a binary variable indicating if the individual is a member of a household with a low income, based on the measure LICO After Tax;

Technical Validation

The second objective of the article is to validate the proposed methodology by ensuring the quality of the synthetic population obtained (internal validation) and the association between the estimated proportion of potentially vulnerable people and two well-known indicators: the Canadien Index of Multiple Defavorisation and the Index of material and social deprivation (external validation). Note that we used the existing ICDM and IMSD indicators to validate the construction of our measure because they have been widely used indicators in Canada

	% of the population potentially vulnerable	INSPQ index of social and material deprivation	
		grouping 1 R^2	grouping 2 R^2
Calgary 2016	Definition 1	0.350	0.359
	Definition 2	0.407	0.487
	Definition 3	0.417	0.533
	Definition 4	0.409	0.496
	Definition 5	0.415	0.557
Montreal 2016	Definition 1	0.440	0.516
	Definition 2	0.505	0.616
	Definition 3	0.510	0.626
	Definition 4	0.510	0.623
	Definition 5	0.524	0.634
Toronto 2016	Definition 1	0.412	0.299
	Definition 2	0.438	0.449
	Definition 3	0.420	0.508
	Definition 4	0.436	0.458
	Definition 5	0.405	0.526
Vancouver 2016	Definition 1	0.389	0.307
	Definition 2	0.417	0.394
	Definition 3	0.408	0.420
	Definition 4	0.412	0.396
	Definition 5	0.365	0.413
Montreal 2021	Definition 1	0.301	0.368
	Definition 2	0.332	0.433
	Definition 3	0.298	0.415
	Definition 4	0.331	0.433
	Definition 5	0.289	0.407

Table 7. R^2 between the five definitions and the IMSD.

since their creation. Despite their limitations documented earlier, they are solid indicators of socio-economic vulnerability.

Table 3 shows the different percentiles of our goodness-of-fit indicators at DA level for the synthetic population obtained before weighting. We also report the scaled TAE (the error divided by the population in a DA: TAE/N) and the Mean Relative Error (the mean of the relative errors: MRE). The MRE consists of the mean absolute error of each category divided by the total size of the category.

We can see that for each region, the genetic algorithm selects samples that reconstruct efficiently the targeted margins. For 99% of the DAs, the scaled TAE represents less than 9.7% of their total population and the RWMSRE is less than 0.1. The genetic algorithm reached the convergence criterion for all the DAs (median of the number of iterations = 60; 95^{th} percentile = 110). Our observations indicate that, on average, an increase in the population size of a DA of 1000 people requires 17 additional iterations to reach convergence, although this relationship is not strictly linear. The Pearson correlation coefficient between the DA population and the number of iterations required to reach convergence is 0.54 ($p < 0.001$). The average computation time for a 600-person DA was 9 seconds and increased by 20 seconds for each 1000 inhabitants (using a 2.44 GHz processor). The calculation speed can be an issue in genetic algorithms, but the proposed algorithm produces results in a time-efficient manner and is easy to parallelize, enhancing computational speed and resource utilization.

As indicated in Table 4, when the weighting is applied, a notable reduction in errors within the samples generated by the genetic algorithm is observed. Remarkably, for 99.9% of the DAs, the scaled TAE values are less than 2% of their total population. The fit is very satisfactory and indicates that the proposed method can accurately reconstruct the target margins at DA level. The distributions of weights obtained are almost identical for all metropolitan areas in 2016 and 2021, with the 1st percentile of all weights being 0.79 and the 99th percentile being 1.20. These low weights underline the quality of the sample produced by the genetic algorithm and the marginal adjustment role of the weighting.

The spatial autocorrelation of the errors, measured with Moran's I statistic with a Queen contiguity matrix, is less than 0.01 for all four regions for the scaled TAE, MRE and RWMSRE, indicating that the errors are randomly distributed in space in both 2016 and 2021. The TAE was not tested considering that it is a counting variable.

Using the synthetic populations obtained, we calculated the total number of people in a potential situation of vulnerability for each DA. We tested five definitions:

1) Individuals falling into at least one of the specified categories, encompassing those aged 65 and above, low-income households, single-person households, single-parent households, unemployed adults, recent immigrants, visible minorities, aboriginal, low level of education.

2) Individuals falling into at least two of the aforementioned categories.
3) Individuals falling into at least three of the aforementioned categories.
4) Individuals living in a low-income household or in at least two of the aforementioned categories.
5) Individuals falling into a low-income household or into at least three of the aforementioned categories.

The first definition appears to reflect an overly broad scope of socio-economic vulnerability. Indeed, it is quite possible to belong to one of these groups without being in a situation of potential vulnerability. Rather, socio-economic vulnerability arises from the nuanced interplay of various factors, a concept encapsulated by the notion of intersectionality. The second and third definitions acknowledge this complexity by stipulating the necessity of possessing two or three distinct characteristics. The last two definitions assign a higher importance to the low-income parameter acknowledging that belonging to a low-income household is sufficient to place its members in a situation of vulnerability.

Table 5 shows the percentages of the total population considered as potentially vulnerable, according to these five definitions, for the four metropolitan areas in 2016 and 2021.

Definitions 1, 2 and 4 lead to a potentially vulnerable population that may encompass an excessively large proportion of individuals. To determine the best fitting definition, we calculated the Pearson correlation coefficient between these percentages and results obtained by the 2016 and 2021 CIMD (Canadian Index of Multiple Deprivation) indicator (see Table 6). This indicator is built from four distinct dimensions that can be aggregated into a global indicator by calculating the average of the quantiles of the 4 sub-dimensions.

With the exception of the first definition, the correlation values with the overall indicator are very strong (between 0.76 and 0.87). This underlines that the proposed method measures a concept equivalent to that of the CIMD and is valid for both censuses. The strongest correlation values are obtained for Montreal and Calgary, and are slightly weaker in Toronto and Vancouver. This result could be explained by their very different ethnocultural compositions or by the contribution of other variables to the socio-economic vulnerability. It highlights the need to adapt the definition of vulnerable population by metropolitan area. Also, the correlations vary greatly among the sub dimensions of the CIMD. In 2016, ethnocultural composition has the highest correlation with our measure. In 2021, we observe a significant increase in the correlations with economic dependency and residential instability, while the correlation with ethnocultural composition decreases. Situational vulnerability tends to be the dimension with the lowest correlation with our measure of the potentially vulnerable population. This result highlights the complementary role of our measure as a quantitative tool for assessing the number of potentially vulnerable people. Existing indicators provide more detail on the sub-components of socio-economic vulnerability.

We then compared the percentages of the vulnerable population obtained by DA with the Institut National de Santé Publique du Québec's (INSPQ) Index of Material and Social Deprivation (IMSD), published in 2016. This indicator comprises two dimensions, each categorized into five quintiles. These two dimensions are combined to produce a classification of DAs. Two different groupings of five classes are proposed by the INSPQ. The first distinguishes DAs with high social and material deprivation, DAs with high social or material deprivation, privileged DAs and DAs in between. The second can be understood as a gradient going from the very privileged DAs to the DAs summing both material and social deprivation.

Table 7 shows the R-squared obtained by crossing our percentage of vulnerable populations and the categories of the two groupings. For each metropolitan area, a Wilcoxon test established that the means of the percentages of vulnerable populations were different between the groupings of the INSPQ at the 0.01 threshold. Moreover, the DAs in the most disadvantaged categories were well associated with the highest rates of vulnerable populations, indicating a great correlation between the two methods to measure social vulnerability.

It is worth noting that definitions 5 and 3 systematically have higher R-squared values. Similarly, the R-squared values are always higher when we compare the percentage of potentially vulnerable populations with the second grouping. This can probably be explained by the fact that the second grouping is of a more quantitative nature than the first grouping and is therefore closer to our own assessment of the concentration of vulnerable populations. For information purposes, we were also able to calculate the R-squared for the Montreal CMA in 2021 because the IMSD data is already available for Quebec. However, the indicator was not standardized for the same geographical scale in 2016 (Canadian regions) and in 2021 (health and social regions), which limits the direct comparison. Excluding 2021, the R-squared obtained for the definition 5 and grouping 2 range from 0.63 in Montreal to 0.41 in Vancouver. These values can be considered as a strong indicator of association between our measure of potentially vulnerable populations and the categories of the IMSD indicator. Indeed, it is not straightforward to propose a correlation measure between a qualitative variable and a percentage variable. The R-squared is used here as a crude measure of association, since the categories of the IMSD do not have a direct directionality like the CIMD.

Once again, the measured association is weaker for Toronto and Vancouver, confirming the conclusions drawn from the results in Table 6 and the need to adapt the definition of vulnerability locally.

Results imply that the final definition of the population in a potential situation of vulnerability can benefit from a local adjustment. Depending on the specific subject and objectives of the study, some definitions may prove more adequate than others.

In this study, the fifth definition offers very satisfactory results when compared with both the CIMD and the IMSD, suggesting potential as a reference point. Figure 2 shows the percentages of the population in a situation of potential vulnerability obtained in 2016 and 2021 for the four regions according to definition five.

We also tested the difference in the composition of populations identified as potentially vulnerable between the PUMF data and the weighted synthetic populations. This was carried out by comparing two *beta* generalized linear models. We first modelled the share of each possible profile of potentially vulnerable population as a function of the different binary vulnerability variables. The second model incorporated an interaction term between

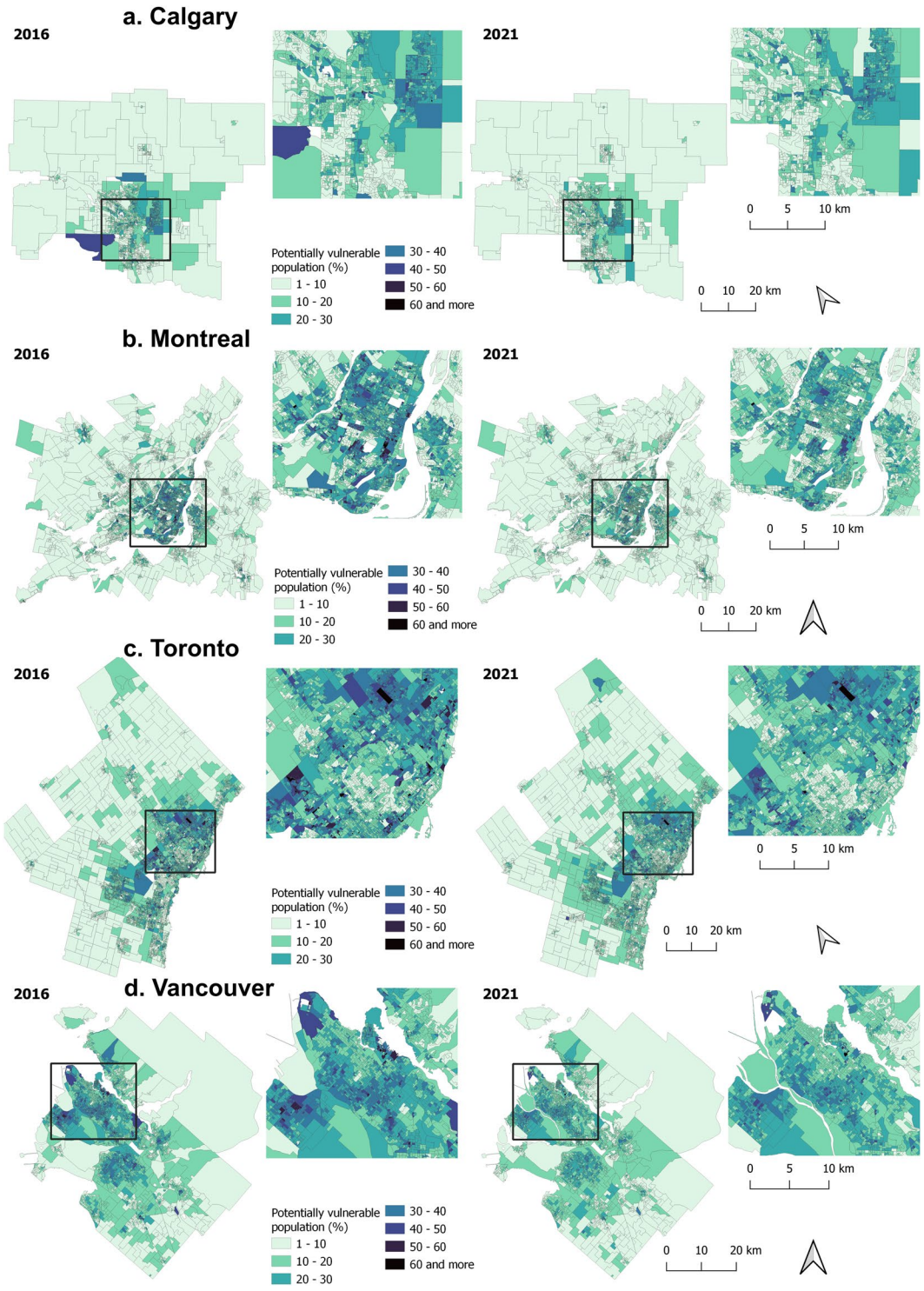

Fig. 2 Maps of potentially vulnerable population.

each binary variable and a separate binary variable distinguishing between PUMF data and data from synthetic populations. The two models were then compared using a likelihood ratio test. The null hypothesis, implying no disparity in the proportion of potentially vulnerable population profiles, was retained if the difference between the models was not statistically significant at the 0.01 threshold. In other terms, the distinction of coefficients between PUMF and synthetic data did not improve the model in any way. Table 8 shows the p-values for these various tests. All the tests are non-significant, even when the p-values are not adjusted for multiple testing and thus higher than what they should be. In other words, the composition of the synthetic populations does not differ significantly from the composition of the original PUMF data.

	p-value 2016	p-value 2021
Calgary	0.687	0.997
Montreal	0.791	0.955
Toronto	0.809	0.999
Vancouver	0.947	0.999

Table 8. P-values of likelihood ratio tests comparing the composition of the original and synthetic population.

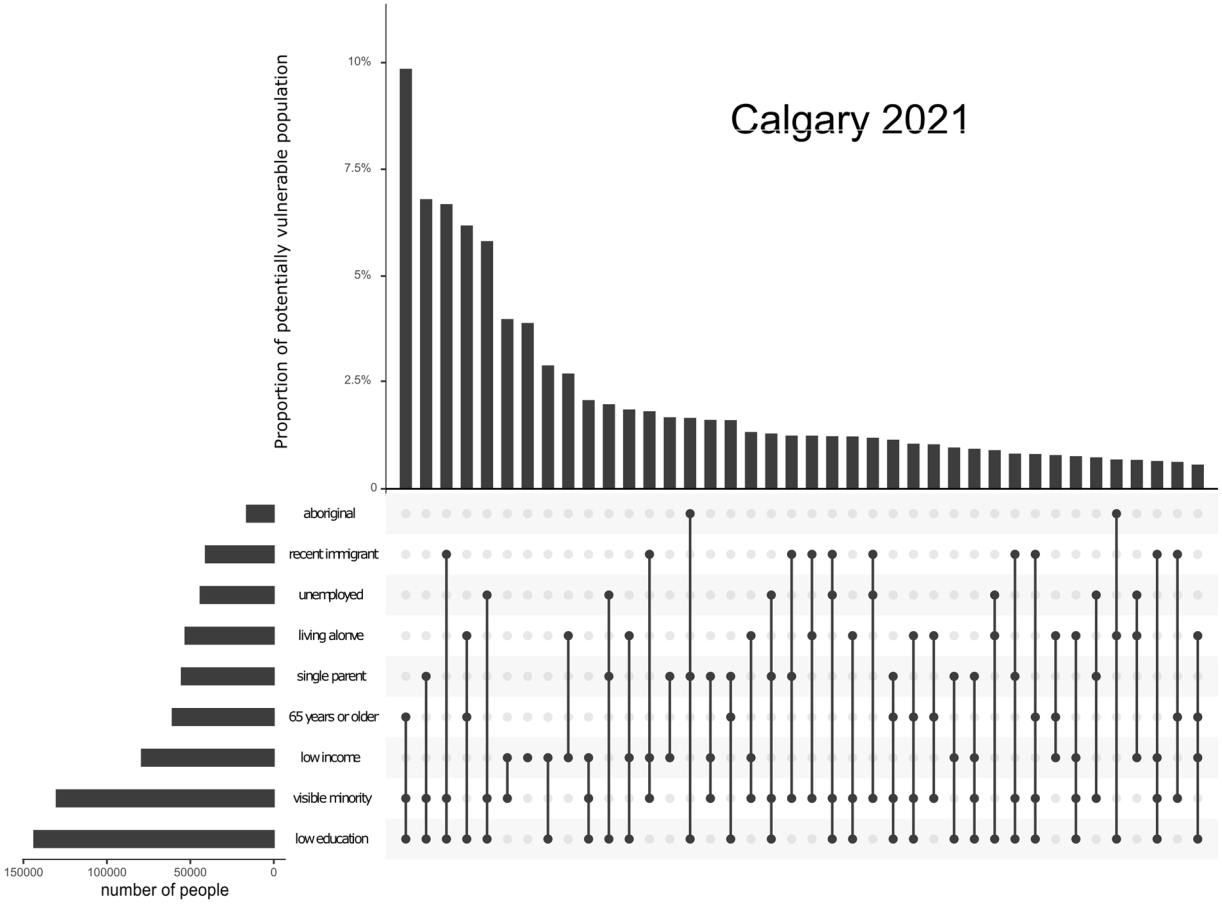

Fig. 3 Composition of Calgary's potentially vulnerable population.

Finally, we look at the composition of the population determined as potentially vulnerable according to the criteria outlined in definition 5. For brevity, we present only the results for 2021.

In 2021, in Calgary (Fig. 3), the most common profile group in our potentially vulnerable population includes people who are 65 or older, have a lower level of education and belong to a visible minority. This group accounts for 10% of the potentially vulnerable population. The top five group, covering 37% of the potentially vulnerable population, all share the characteristics of lower education levels and mostly belong to visible minorities. Interestingly, it's only in the sixth group that the variable of low income becomes noticeable.

In the Montreal context in 2021 (Fig. 4), the demographic group most prominently featured within the potentially vulnerable population comprises individuals aged 65 and above, residing in solitary conditions, and possessing a limited educational background. This specific demographic segment constitutes 14% of the overall potentially vulnerable population. Notably, the top five demographic groups collectively contribute to 37% of the entire vulnerable population. Furthermore, these groups exhibit a greater degree of heterogeneity (i.e. a more diverse set of characteristics) compared to their counterparts in the city of Calgary.

In the demographic context of Toronto in 2021 (Fig. 5), similar to the situation observed in Calgary, the predominant groups within the potentially vulnerable population are found to be associated with specific demographic characteristics. Notably, individuals characterized by a low level of education, those identifying as a visible minority, and those aged over 65 or belonging to a single-parent household constitute the most predominant groups of the potentially vulnerable population. Specifically, these two groups contribute 14% and 11%, respectively, to the overall potentially vulnerable population. Among the identified variables, the attributes of being visible minorities and possessing a low level of education are the most influential. They play a pivotal role

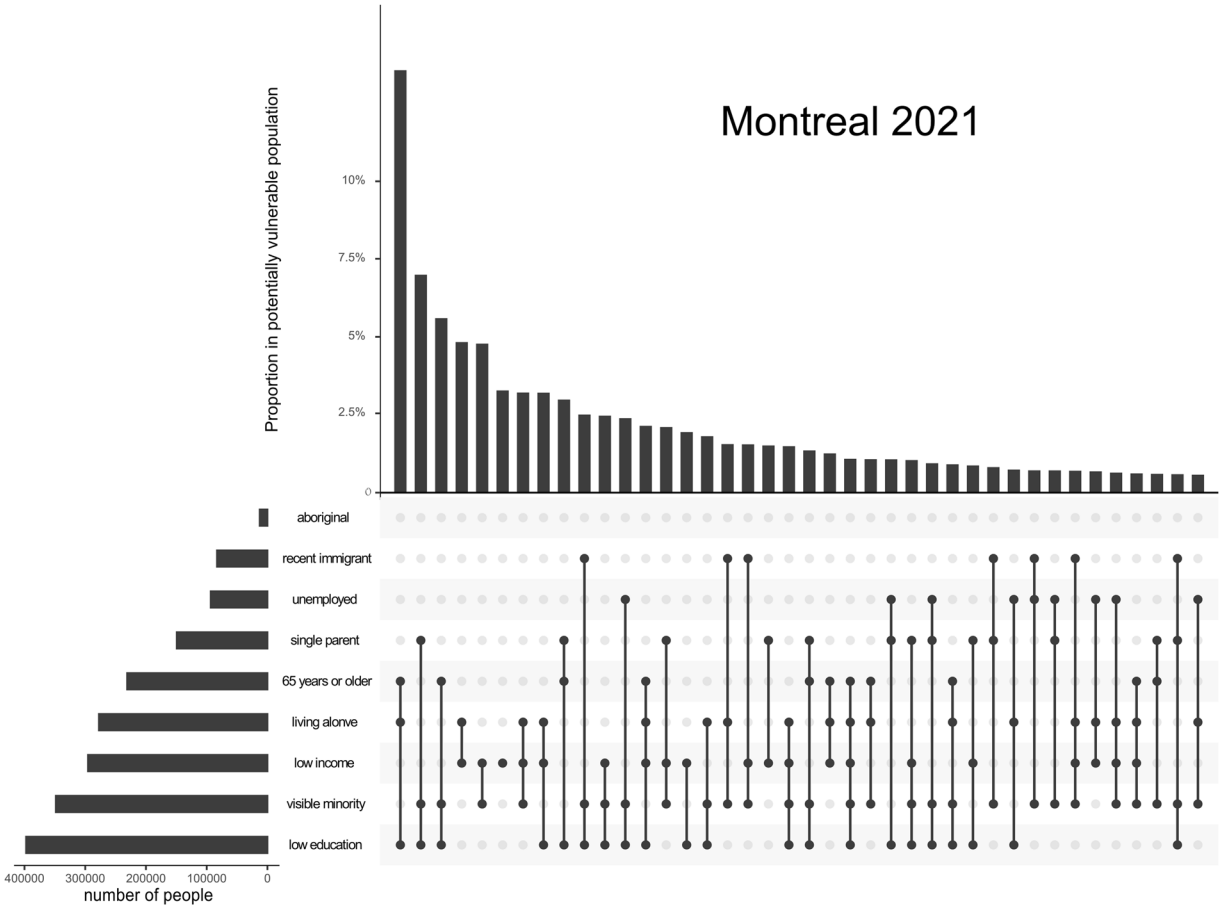

Fig. 4 Composition of Montreal's potentially vulnerable population.

in defining four out of the five initial groups, which, in turn, collectively represent a substantial 43% of the entire potentially vulnerable population.

In Vancouver, in 2021 (Fig. 6), the first 2 groups are identical to those found in Calgary and Toronto, constituting 17% and 8% of the total vulnerable population, respectively. It is interesting to note a certain similarity in the composition of the potentially vulnerable populations for these four metropolitan areas. The defining characteristics of the primary groups predominantly revolve around variables such as visible minority status, lower educational attainment, and seniors. It should be noted that the weak presence of groups representing indigenous populations in the graphs may be attributed to their minimal representation within the overall vulnerable population. This absence is due solely to the method of representation (upset plot) in this section. A study focusing on this specific group of people could extract the relevant part of the data and map their presence and analyze their characteristics. However, we would not advocate analyzing a tiny subset of data because the uncertainty is higher when considering very specific and rare profiles.

For every Metropolitan Area, the two characteristics that are the most present within the vulnerable populations are the low level of education and the belonging to a visible minority (as shown by the left histograms). They are then followed by the low income and age (65 years or older). Montreal is slightly different than the other Metropolitan Areas, indeed the gap between the two main groups and the following is much lower for Montreal. This indicates a more diverse vulnerable population in the Montreal Metropolitan Area. The analysis of these charts suggests that the proposed method in this article could be extended by creating groups of similar profiles within the potentially vulnerable people. Such segmentation could help to identify different dimensions of socio-economic vulnerability and reveal their geography.

Usage Notes

The synthetic population methodology introduced in this article primarily aims to estimate the total count of potentially vulnerable individuals per DA. These estimates are provided at the individual level, offering users the flexibility to adjust the definition of vulnerability based on specific requirements. It is important to note that the absence of a hierarchical household structure in the data may constrain their applicability in certain contexts, such as multi-agent models or multi-level statistical analyses. Nevertheless, the individual-level data hold significant value for examining the demographic composition of the potentially vulnerable population and may also facilitate longitudinal analyses.

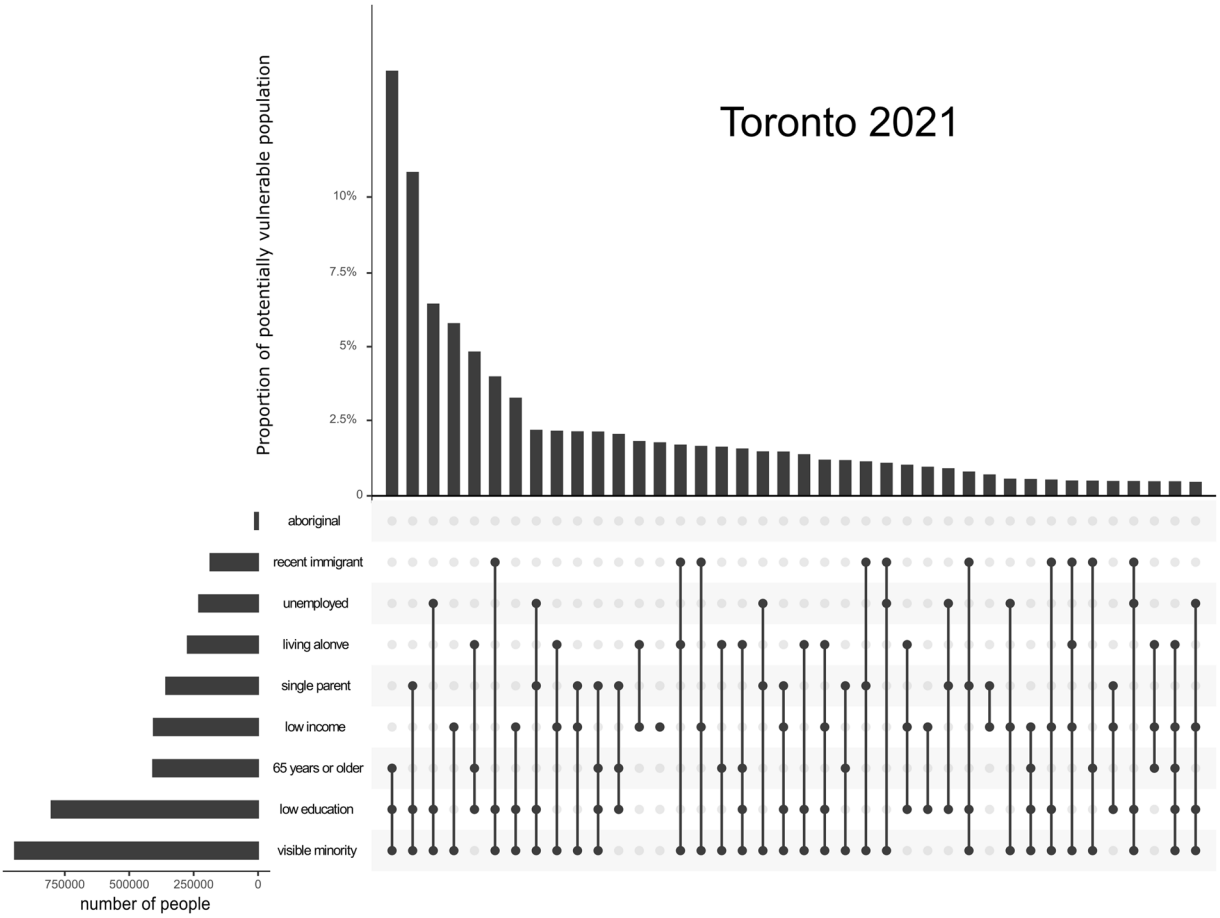

Fig. 5 Composition of Toronto's potentially vulnerable population.

In this study, we showcased findings for four Canadian metropolitan areas. Our internal and external validation analysis demonstrated the high quality of the synthetic population adjustment and the strong association of the estimated proportions of potentially vulnerable people and well-known indicators of socio-economic vulnerability.

With the open access code and freely accessible data provided with this publication, there is the opportunity to broaden the application of our methodology to encompass all DAs across Canada. However, if the method is applied in DAs outside a Census metropolitan area or a Census agglomeration covered by the PUMF data, then the results might be less accurate. In such cases, it will be necessary to use individuals located at the provincial level. Although encouraging, the results suggest that the proposed methodology should be adapted locally (weaker correlations were observed for Vancouver). However, the proposed methodology is flexible enough to allow adjustment of the variables used to define the potentially vulnerable population. The proposed definition can also be used as a benchmark for studies analyzing multiple cities.

The choice of a genetic algorithm in our method implies that the obtained results are stochastic. In this study, we present a single version of the produced synthetic population, although we ran the analysis several times with different random starts and obtained similar results. Nonetheless, the supplementary material (the supporting code and data) enables users to generate multiple iterations of the presented results, facilitating sensitivity and uncertainty analyses.

This measure should not be seen as a replacement for indicators such as CIMD or IMSD, which offer an understanding of vulnerability from a multivariate perspective. Rather, we propose it as a complementary quantitative indicator. The inclusion of an indicator that quantifies the number of potentially vulnerable individuals within a DA holds significant utility for land-use planning and assessing the impact of public policies.

Our proposal departs from conventional methods that rely on composite indices. It offers a new perspective on how to construct meaningful measures of socio-economic vulnerability that focus on the people's intersectional situations. We have outlined its advantages several times, but in this article, we have only scratched the surface of its possibilities. By intersecting this measure with other variables measuring urban amenities or nuisances, it is possible to easily quantify a potential situation of inequity. For instance, one could compare the proportion of individuals exposed to air pollution levels exceeding recommended thresholds with the proportion of vulnerable populations facing the same exposure. Similarly, the average exposure levels in DAs, weighted by both total population and vulnerable population, could be calculated. The ratio between these two values could serve as an indicator of equity in exposure. Also, one could use segmentation techniques to identify the main profiles of vulnerable people and analyze their geographic distribution. In terms of urban planning, public transit agencies could use such data to establish planning targets like reducing the proportion of the vulnerable population with

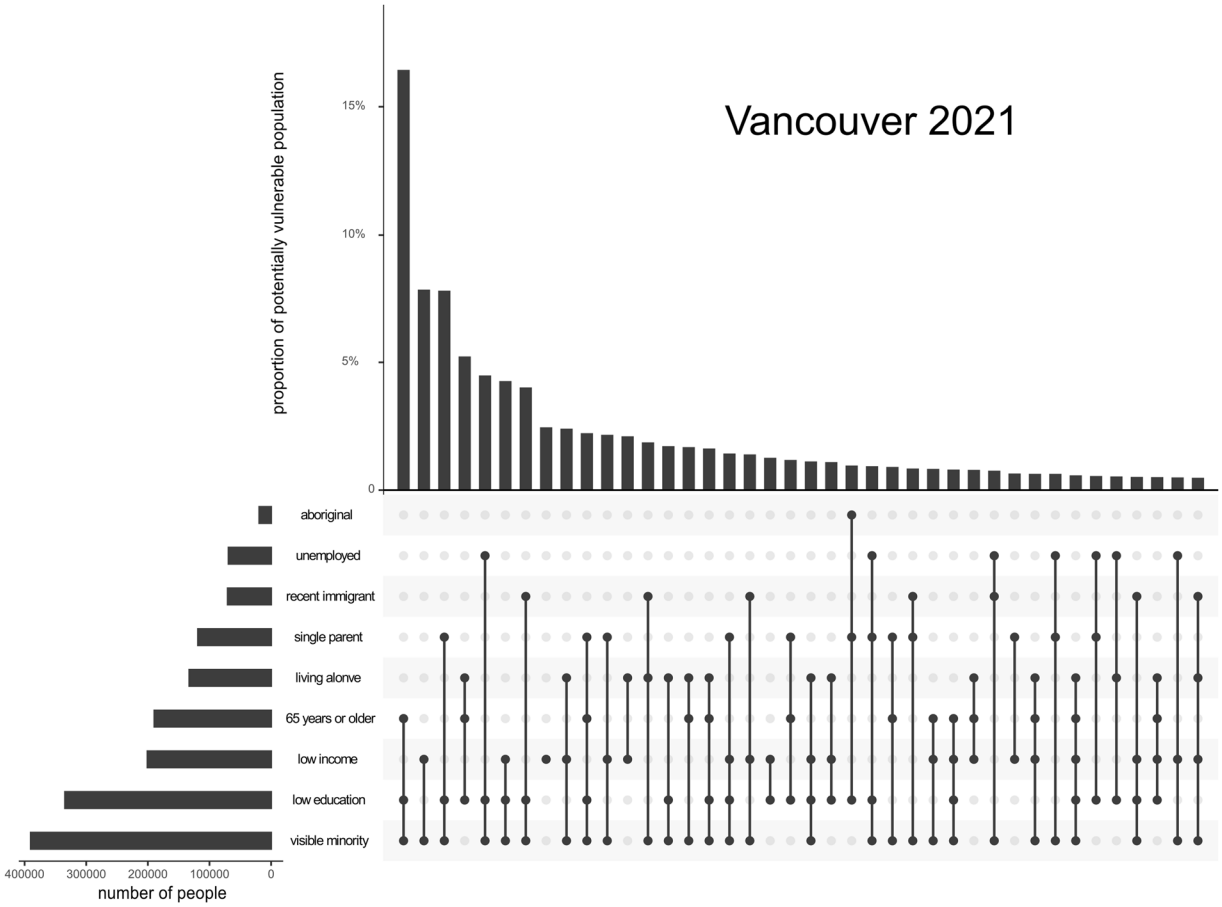

Fig. 6 Composition of Vancouver's potentially vulnerable population.

low levels of accessibility to public transit. Such analyses could be readily conducted using any variables measuring urban nuisances or resources. Similarly, the impact of a public policy or urban project could be directly expressed in terms of total population and vulnerable population affected. Thus, the proposed indicator holds significant potential and serves as a robust foundation for conducting equity analyses, assessments, and policy evaluations.

Code availability

The R scripts (4.2.1) developed for this project are freely available in the same repository on Zenodo[10], accompanied by the generated synthetic data. A comprehensive list of the required packages is provided in the *requirement.txt* file, all of which can be downloaded directly from CRAN, the main package depository for R.

Received: 15 April 2024; Accepted: 14 August 2024;
Published online: 31 August 2024

References

1. OECD, Union, E. & Commission, J. R. C.-E. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. isbn:978-92-64-04346-6 (OECD Publishing, 2008).
2. Greco, S., Ishizaka, A., Tasiou, M. & Torrisi, G. On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. *Soc Indic Res* **141**, 61–94 (2019).
3. Aday, L. A. *At Risk in America: The Health and Health Care Needs of Vulnerable Populations in the United States*. (John Wiley & Sons, 2002). isbn:978-0-7879-5932-6.
4. Wrigley, A. & Dawson, A. *Vulnerability and Marginalized Populations. in Public Health Ethics: Cases Spanning the Globe* (eds. H. Barrett, D. *et al.*) isbn:978-3-319-23846-3 (Springer, 2016).
5. Chapuis, K., Taillandier, P. & Drogoul, A. Generation of Synthetic Populations in Social Simulations: A Review of Methods and Practices. *JASSS* **25**, 6 (2022).
6. Nicolaie, M. A., Füssenich, K., Ameling, C. & Boshuizen, H. C. Constructing synthetic populations in the age of big data. *Popul Health Metrics* **21**, 19 (2023).
7. Prédhumeau, M. & Manley, E. A synthetic population for agent-based modelling in Canada. *Sci Data* **10**, 148 (2023).
8. Statistics Canada. The Canadian Index of Multiple Deprivation, Statistics Canada Catalogue no. 45-20-0001, https://www150.statcan.gc.ca/n1/pub/45-20-0001/452000012019002-eng.htm (2019).
9. Pampalon, R. *et al.* Un indice régional de défavorisation matérielle et sociale pour la santé publique au Québec et au Canada. *Can J Public Health* **103**, S17–S22 (2012).
10. Gelb, J., Apparicio, P. & Alizadeh, H. Development of a synthetic vulnerable population for four major metropolitan regions in Canada. *Zenodo* https://doi.org/10.5281/zenodo.10819314 (2024).
11. Shi, L. & Stevens, G. D. *Vulnerable Populations in the United States*. (John Wiley & Sons, 2021). isbn:978-1-119-62767-8.

12. Chesnay, M. de & Anderson, B. *Caring for the Vulnerable*. (Jones & Bartlett Learning, 2019). isbn:978-1-284-14681-3.
13. Mechanic, D. & Tanner, J. Vulnerable People, Groups, And Populations: Societal View. *Health Affairs* **26**, 1220–1230 (2007).
14. Pearlin, L. I. The Stress Process Revisited. in *Handbook of the Sociology of Mental Health* (eds. Aneshensel, C. S. & Phelan, J. C.) 395–415 (Springer US, 1999). isbn:978-0-387-36223-6.
15. Havrilla, E. Defining vulnerability. *Madridge Journal of Nursing* **2**, 63–68 (2017).
16. Bullard, R. D. & Wright, B. H. Environmental Justice for all: Community Perspectives on Health and Research. *Toxicol Ind Health* **9**, 821–841 (1993).
17. Spielman, S. E. *et al*. Evaluating social vulnerability indicators: criteria and their application to the Social Vulnerability Index. *Nat Hazards* **100**, 417–436 (2020).
18. Kar, A., Carrel, A. L., Miller, H. J. & Le, H. T. K. Public transit cuts during COVID-19 compound social vulnerability in 22 US cities. *Transportation Research Part D: Transport and Environment* **110**, 103435 (2022).
19. Morency, C., Paez, A., Roorda, M. J., Mercado, R. & Farber, S. Distance traveled in three Canadian cities: Spatial analysis from the perspective of vulnerable population segments. *Journal of Transport Geography* **19**, 39–50 (2011).
20. Deboosere, R. & El-Geneidy, A. Evaluating equity and accessibility to jobs by public transport across Canada. *Journal of Transport Geography* **73**, 54–63 (2018).
21. El-Geneidy, A. *et al*. The cost of equity: Assessing transit accessibility and social disparity using total travel cost. *Transportation Research Part A: Policy and Practice* **91**, 302–316 (2016).
22. Pendall, R., Theodos, B. & Franks, K. Vulnerable people, precarious housing, and regional resilience: an exploratory analysis. *Housing Policy Debate* **22**, 271–296 (2012).
23. Enssle, F. & Kabisch, N. Urban green spaces for the social interaction, health and well-being of older people— An integrated view of urban ecosystem services and socio-environmental justice. *Environmental Science & Policy* **109**, 36–44 (2020).
24. Carrier, M., Apparicio, P. & Séguin, A.-M. Road traffic noise in Montreal and environmental equity: What is the situation for the most vulnerable population groups? *Journal of Transport Geography* **51**, 1–8 (2016).
25. Schweitzer, L. & Zhou, J. Neighborhood Air Quality, Respiratory Health, and Vulnerable Populations in Compact and Sprawled Regions. *Journal of the American Planning Association* **76**, 363–371 (2010).
26. Larsen, K. & Gilliland, J. Mapping the evolution of 'food deserts' in a Canadian city: Supermarket accessibility in London, Ontario, 1961–2005. *International Journal of Health Geographics* **7**, 16 (2008).
27. Mah, J. C., Penwarden, J. L., Pott, H., Theou, O. & Andrew, M. K. Social vulnerability indices: a scoping review. *BMC Public Health* **23**, 1253 (2023).
28. Matheson, F. I., Dunn, J. R., Smith, K. L. W., Moineddin, R. & Glazier, R. H. É. laboration de l'indice de marginalisation canadien: un nouvel outil d'étude des inégalités. *Can J Public Health* **103**, S12–S16 (2012).
29. Rey, G., Jougla, E., Fouillet, A. & Hémon, D. Ecological association between a deprivation index and mortality in France over the period 1997 – 2001: variations with spatial scale, degree of urbanicity, age, gender and cause of death. *BMC Public Health* **9**, 33 (2009).
30. Rey, G., Rican, S. & Jougla, E. Mesure des inégalités de mortalité par cause de décès. Approche écologique à l'aide d'un indice de désavantage social. *Bull Epidémiol Hebdomadaire* **8**, 87–90 (2011).
31. Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L. & Lewis, B. A social vulnerability index for disaster management. *Journal of homeland security and emergency management* **8** (2011).
32. Cutter, S. L., Boruff, B. J. & Shirley, W. L. Social vulnerability to environmental hazards. in *Hazards vulnerability and environmental justice* 143–160 (Routledge, 2012). isbn:978-1-84977-154-2.
33. Ville de M. *Indice d'équité des milieux de vie, rapport méthodologique*. https://donnees.montreal.ca/dataset/1c1a9b72-efaa-4484-9b19-2ab9a5561cfa/resource/7754ac6b-2e97-4aa5-97e9-8ef7a38ddb03/download/rapport-methodologique.pdf (2023).
34. Heisz, A. *Canada's global cities: Socio-economic conditions in Montreal, Toronto and Vancouver*. vol. 010 (Statistics Canada Ottawa, 2006). isbn:0-662-43672-5.
35. Breau, S., Shin, M. & Burkhart, N. Pulling apart: new perspectives on the spatial dimensions of neighbourhood income disparities in Canadian cities. *J Geogr Syst* **20**, 1–25 (2018).
36. Statistics Canada. Census Profile, 2016 Census - Statistics Canada Catalogue number 98-316-X2016001. (2016), Statistics Canada Catalogue number 98-316-X2016001, https://www150.statcan.gc.ca/n1/en/catalogue/98-316-X2016001.
37. Statistics Canada. Census Profile, 2021 Census - Statistics Canada Catalogue number 98-316-X2021001. (2021), Statistics Canada Catalogue number 98-316-X2021001, https://www150.statcan.gc.ca/n1/en/catalogue/98-316-X2021001.
38. Statistics Canada. Individuals File, 2016 Census of Population (Public Use Microdata Files). (2019), Statistics Canada Catalogue number 98M0001X, https://www150.statcan.gc.ca/n1/en/catalogue/98M0002X.
39. Statistics Canada. Individuals File, 2021 Census of Population (Public Use Microdata Files). (2023), Statistics Canada Catalogue number 98M0001X, https://www150.statcan.gc.ca/n1/en/catalogue/98M0002X.
40. Yaméogo, B. F., Gastineau, P., Hankach, P. & Vandanjon, P.-O. Comparing Methods for Generating a Two-Layered Synthetic Population. *Transportation Research Record* **2675**, 136–147 (2021).
41. Ilahi, A. & Axhausen, K. W. Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta. *Regional Studies, Regional Science* **6**, 623–636 (2019).
42. Forrest, S. Genetic algorithms. *ACM computing surveys (CSUR)* **28**, 77–80 (1996).
43. Kramer, O. Genetic Algorithms. in *Genetic Algorithm Essentials* (ed. Kramer, O.) 11–19 (Springer International Publishing, 2017). isbn:978-3-319-52156-5.
44. Mühlenbein, H. & Voigt, H.-M. Gene Pool Recombination in Genetic Algorithms. in *Meta-Heuristics* (eds. Osman, I. H. & Kelly, J. P.) 53–62 (Springer US, 1996). isbn:978-1-4612-8587-8.
45. Deming, W. E. & Stephan, F. F. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics* **11**, 427–444 (1940).
46. Ye, X., Konduri, K., Pendyala, R. M., Sana, B. & Waddell, P. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. in *88th Annual Meeting of the transportation research Board, Washington, DC* (2009).
47. Müller, K. & Axhausen, K. W. Multi-level fitting algorithms for population synthesis. *Arbeitsberichte Verkehrs-und Raumplanung* **821**, (2012).
48. Deville, J.-C., Särndal, C.-E. & Sautory, O. Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association* **88**, 1013–1020 (1993).
49. R Core Team.*R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2021).

Author contributions

Jérémy Gelb. main researcher, data structuration, curation, analysis, coding and article redaction. Philippe Apparicio. validation of the methodology and article reviewing. Hamzeh Alizadeh. project supervisor, validation of the methodology and article redaction.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.