Public Health Ontario | Santé publique Ontario

Small Area Analysis:

A primer for Public Health Units

Primer

March 2018

Public Health Ontario

Public Health Ontario is a Crown corporation dedicated to protecting and promoting the health of all Ontarians and reducing inequities in health. Public Health Ontario links public health practitioners, frontline health workers and researchers to the best scientific intelligence and knowledge from around the world.

Public Health Ontario provides expert scientific and technical support to government, local public health units and health care providers relating to the following:

- communicable and infectious diseases
- infection prevention and control
- environmental and occupational health
- emergency preparedness
- health promotion, chronic disease and injury prevention
- public health laboratory services

Public Health Ontario's work also includes surveillance, epidemiology, research, professional development and knowledge services. For more information, visit publichealthontario.ca.

How to cite this document:
Ontario Agency for Health Protection and Promotion (Public Health Ontario). Small area analysis: a primer for public health units. Toronto, ON: Queen's Printer for Ontario; 2018.

ISBN: 978-1-4868-1605-7

Public Health Ontario acknowledges the financial support of the Ontario Government.

Authors

Analytic Services, Informatics
Public Health Ontario

Acknowledgements

The authors wish to express their sincere appreciation for their guidance, input, and review of this document by the advisory committee members from Halton Region Health Department, Ottawa Public Health, Oxford County Public Health and Emergency Services, Toronto Public Health, and Toronto Central LHIN. We also thank our colleagues at Toronto Public Health and Windsor-Essex County Health Unit, for taking the time to test the document from a usability perspective.

Disclaimer

This document was developed by Public Health Ontario (PHO). PHO provides scientific and technical advice to Ontario's government, public health organizations and health care providers. PHO's work is guided by the current best available evidence at the time of publication.

The application and use of this document is the responsibility of the user. PHO assumes no liability resulting from any such application or use.

Contents

Background

Neighbourhood level information is essential for Public Health Units (PHUs) to effectively assess, plan, and evaluate health services and programs in order to decrease health inequities and address the needs of priority populations at the local level.

One way to analyze this information is by utilizing small area analysis. Small area analysis (SAA) describes statistical methods or techniques used to produce adequate estimates when there is insufficient information or small sample size. It is also referred to as small area estimation or small domain analysis. Although the term 'area' is used, the concept can also refer to groupings other than geographical districts, such as socio-demographic groups or other structural characteristics that partition the population.

Survey data such as the Canadian Community Health Survey (CCHS) and the Rapid Risk Factor Surveillance System (RRFSS) are available for estimating population health practices and behaviours. However, the sampling design and weighting scheme is structured such that health indicator estimates are often only recommended for calculation at the PHU, Local Health Integration Network (LHIN) and provincial levels.

PHUs face a number of challenges with trying to obtain high quality estimates at the neighbourhood level. Resources are often unavailable to collect additional samples to accurately represent the neighbourhood, or the sample size in existing data sources may be too small to provide reliable health indicator estimates. While administrative data do not pose the same challenges as survey data, many of these data sources do not contain variables for neighbourhoods and would need to be added by custom geospatial methods.

Rationale for resource document

We created this document in response to the results of a survey completed by all 36 PHUs in 2017. The majority of PHUs stated being asked by community partners or LHINs to present or analyze data for small areas, and almost all PHUs routinely or occasionally examine information below the health unit level. The majority of health units report data by municipality and information was most often used for the purposes of program planning and population health assessment.

Although more than half of the PHUs have applied small area analysis methods to complex survey data such as CCHS, less than 20% of health units are moderately or very comfortable/confident with small area methods. Survey respondents also reported an interest in having SAA resources, such as a how-to guide that would show how to prepare the data, SAA method options, and how to calculate estimates based on Ontario specific examples.

The submission of two proposals related to the topic of SAA from two separate PHUs during the 2016 Locally Driven Collaborative Projects (LDCP) funding cycle also shows there is an interest in SAA from the field. With *Patients First*[1] formalizing the connection between LHINs and Boards of Health in December 2016, along with the modernization of the *Ontario Public Health Standards* (OPHS 2018), we anticipate a growing need for small area population health data.

Purpose of document

The purpose of this document is to provide PHUs with a decision making process when addressing questions at the small area level, including examples of small area analysis methods and when they may be applicable. This document is not intended to be a step-by-step manual. As SAA encompasses a variety of methodologies, we do not provide an exhaustive list of techniques. Instead, we discuss methods that PHUs could find practical to use in their routine work, based on information we gathered from the literature, discussions with other organizations doing similar analysis, and results from a survey of all PHUs in Ontario.

It is expected that users of this document will have experience in using common data sources available for population health assessment and surveillance, reporting indicators using complex survey data, and statistical model building techniques. When choosing a small area method, knowledge and expertise about the local context, availability of resources and understanding of the situation and methodology should be included as part of the decision making process.

Considerations for SAA

When presented with a request asking for small area information, a number of issues should be discussed with the requestor prior to beginning the analysis. As considerable staff time and resources are often required to develop, check, and release information, there should be justified demand to do this work.

It is important that everyone involved in the decision to undertake a SAA exercise have a clear understanding of the decisions, trade-offs, and assumptions that need to be made in order to provide the end user with useful/meaningful output to answer the original question.

Is the large area estimate sufficient?

Depending on the purpose of the request, the nature of the problem you are trying to solve, and assumptions you are willing to make, a PHU, LHIN or other large area estimate may be adequate to answer the question posed. Below are some questions to help initiate your discussion with the requestor when considering if a large area estimate may be sufficient.

- Are the characteristics of the small area similar to the area with an appropriate estimate?
- How will the information be used in the decision making process?
    - Do you have a specific purpose in mind for the information (e.g., to support allocation of large funds for public health program and service planning, creating targeted interventions, to fulfil requirements in the OPHS?
    - Will the information be used to support action? Is the information required before public health action can be taken?
    - Are there activities that may arise from the small area analysis? Interventions that target large populations (e.g., water fluoridation) may not be supported by a small area analysis and may be best supported by PHU-level analysis.

- Do you anticipate having to produce this information on a regular basis? Or is the information required for a one time purpose?
- What is the tolerance for imprecision? The number of assumptions or approximations made may limit the usefulness of results from SAA for some questions.

What domain size is required for a meaningful analysis?

There is a relationship between the domain size for which you are producing an estimate, and the degree of accuracy/confidence in the output produced. As the size of the domain becomes smaller (e.g., sample size becomes smaller), the less confident or accurate the estimate becomes. It is important to determine if the minimum domain size to produce accurate enough results that are meaningful to the decision maker aligns with the small area request.

When assessing a request asking for domains of questionable size (i.e., is the domain too small?), a general sample size calculation can be used to get an approximate sense of what domain size is needed to produce a meaningful estimate. For example, would it be more appropriate to calculate estimates at the level of Census Subdivision (CSD) or Dissemination Area (DA)? Knowing the prevalence of factor (x) with confidence interval (y), the required domain size (z) can be calculated using a direct estimation method. This calculation applies to both survey and administrative data.

Is the data representative of the small area?

As it is impractical to survey the entire population, weights are used to adjust for survey respondents having unequal probabilities of being included in the survey sample. Applying weights in the analysis allows for the results to be generalized to the appropriate sampling frame (e.g., PHU or LHIN[*] for CCHS data). When the decision is made to undertake an analysis of survey data that is not based on the intended sampling design frame (i.e., analyzing CCHS data at the CSD level instead of a PHU or LHIN level), one should be aware of the potential for data quality issues that may arise. Depending on the SAA request, you will have to determine what elements/indicators (e.g., sex, age structure, income) are important for evaluating the appropriateness of the small area sample. Although two examples for assessing CCHS data representativeness are described below, you may need to take a different approach depending on the goal of the analysis.

EXAMPLE ONE: ESTIMATE FOR LHIN SUB-REGION

In the first example, the Health Analytics Branch of the Ontario Ministry of Health and Long-Term Care,[2] used three qualities to determine the suitability of a combined CCHS survey sample (2007/08 and 2009/10) for producing estimates at the level of LHIN sub-region. Data were assessed for sample coverage, sample representativeness, and age structure of the sample (Table 1). Each LHIN sub-region was scored and categorized into 'unrestricted use' (overall score of 0 to 3 points), 'use with caution' (overall score of 4 to 5 points) and 'do not release' (overall score of 6). LHIN sub-regions were also suppressed if serious issues with data quality were identified that were not captured by the scoring system (e.g., if an area received a score of 1 or more in all three flags).

[*] Oversampling to create reliable estimates at the LHIN level ended in the 2015 CCHS cycle

Table 1: Assessment criteria for producing small area samples and the quality of estimates derived from them

Flagging scheme	Score
Sample coverage • Number of sampled respondents (un-weighted) who contribute to resulting calculation	Sample ≥30 then score = 0 Sample 10-29 then score = 1 Sample <10 then score = 2
Sample representativeness • % difference = (Census LHIN sub-region population – CCHS weighted LHIN sub-region population) / CCHS weighted LHIN sub-region population	% diff <5 then score =0 % diff 5-49 then score =1 % diff ≥50 then score = 2
Age structure of the sample • Absolute difference = Σ(CCHS weighted sub-region [% of age group*] – Census LHIN sub-region [% of age group])	abs diff <5 then score =0 abs diff 5-9 then score =1 abs diff ≥10 then score =2

*age groups = 12-19yr, 20-29yr, 30-44yr, 45-64yr, 65+yr

EXAMPLE TWO: ESTIMATE FOR SERVICE DELIVERY AREA

In a second example, Toronto Public Health (TPH) has used the following internal general guideline to help determine the appropriateness of analyzing CCHS data at the level of Service Delivery Area (SDA) for one of their programs. Data are first assessed against four criteria – age and sex coverage, age and sex distribution (absolute difference), age and sex distribution (relative difference), low income distribution (Table 2). Depending on the outcome of each criterion, the decision to produce estimates at the SDA level is made. If the data are deemed appropriate for SDA analysis, a subsequent check is then applied to assess if the information can be reported with caution and a footnote (coefficient of variation (CV) = 16.6 to 25), or should be supressed (CV >25[†])).

Table 2: Assessing CCHS data for use at the level of Service Delivery Area

Criteria	Threshold	Decision
Three age groups and two sex groups* sufficiently covered within the sample (n)	Okay: n ≥ 50 Caution: n is 30 to 50 Problem: n ≤ 30	No analysis at SDA level if one or more of the three age groups or two sex groups has n<30

[†] TPH uses a threshold lower than what is recommended by Statistics Canada.

Criteria	Threshold	Decision
Absolute weighted (N) distribution for SDA in three age groups and two sex groups* similar to expected Absolute difference = weighted CCHS (group %) – Census (group %)	Okay: <5% points Caution: 5% to <10% points Problem: >10% points	**Flag if one or more of the three age groups or two sex groups exceeds 10% points.
Relative weighted (N) distribution for SDA in three age groups and two sex groups* similar to expected Relative difference = [(CCHS group % - Census group %) / Census group %] x 100	Okay: <15% Caution: 15% to <30% Problem: ≥30%	**Flag if one or more of the three age groups or two sex groups exceeds 30%
Absolute weighted (N) % of low income measure (LIM) for SDA similar to expected Absolute difference = CCHS (% living below the LIM) – T1FF (% living below the LIM)	Okay: <5% points Caution: 5% to <10% points Problem: ≥10% points	**Flag if one or more of the three age groups or two sex groups exceeds 10 percentage points

*Three age groups (20-39, 40-64, 65+), two sex groups (male, female)
**No analysis at SDA level, if a flag is raised for at least two of the three criteria

Should modelling be considered to create an estimate?

If data are not adequately representative of a small area, modelling might be considered. Prior to undertaking a modelling exercise, you should consider if developing a model is an appropriate option. As written by British statistician, George Box, 'all models are wrong, some are useful'.[3] Sometimes it is more important to have partial answers than to have the most complete answer before taking action. Models may be valuable in situations where an area has unique characteristics that relationships from larger areas cannot easily explain (e.g., environmental or service based factors) or where adjusting for confounding is required.

Although more complex SAA methods (i.e., creating statistical models) should theoretically improve the accuracy/confidence of an estimate, there is no guarantee that any increase you gain by conducting this analysis outweighs the time, effort and resources required. Some issues to think through if you are considering a modelling exercise include:

- Your confidence in the predictors being able to explain your variable of interest. Are you confident that the predictor variables you've included in the model are valid and most relevant to the outcome variable of interest?
- The robustness of the model: the more assumptions you make, the less robust the model becomes.
- How informative is the model output? Depending on the complexity of the model, the output may be more difficult to interpret and communicate to a lay audience.

- Does the method need to be transparent? Complex methods, such as model building, may be more difficult to replicate by others and often are not as transparent as simpler methods.
- Consider the importance of the decision being made. Information that may be used to drive further enquiry or have low resource implications may be more tolerant of estimate imprecision/inaccuracy than high resource decisions.
- How often do the data change? Updates to model inputs will impact the results of the model output.
- Availability of resources and expertise to undertake a modeling exercise. Model building can be complex and may require a higher level of statistical expertise to implement.

Defining geographic areas

Boundary relationships and alignment

Generally, the geographic area selected for the small area will depend on the type of data you have and the question of interest. A number of geographic areas can be created using various levels of geography in Ontario. Common geographies include:

- Statistics Canada's Standard Geographical Classification (SGC)
- Public Health Units (PHUs)
- Local Health Integration Network (LHIN)/ LHIN sub-region
- Postal codes

Each of these areas were established for various purposes and as a result, their boundaries are interrelated and may overlap with one another. For example, SGC geographic areas were initially developed for disseminating statistics from the population census and LHINs were established as a way to plan, fund, and manage health services.

Statistics Canada's Standard Geographical Classification (SGC)

Many geographies make up the SGC, ranging in size from the provincial level to the block level. Unfortunately, not all of these geographies line up well with each other, making it difficult to aggregate numerator or denominator data between geographies. Statistics Canada has created the 'Hierarchy of standard geographic units' for the Census years as a quick reference guide to determine how (and if) each of the geographies align.[4]

It is strongly recommended to follow the above hierarchy and accompanying reference documentation when trying to aggregate between census geographies. Since not all census geographies line up, as visualized within the hierarchy document, caution needs to be used when aggregating. For example, if one wanted to represent data at the census subdivision (CSD) level and they had dissemination area (DA) attributes, it would be a relatively simple aggregation. However, if one only had forward sortation areas (FSAs), aggregation to CSD would not be recommended, unless one is willing to accept the corresponding degree of error.

PHUs and LHINs/LHIN sub-regions

PHUs are largely made up of municipalities, townships, districts, counties and/or cities, as outlined in Regulation 553 – Areas Comprising Health Units of the *Health Protection and Promotion Act*.[5] LHINs are not as neatly-structured in their make-up as PHUs, often splitting cities and other administrative boundaries.

Each LHIN is further divided into LHIN sub-regions. Since the LHIN sub-regions are relatively new, alignment to other boundaries (e.g., SGC, PHUs) is still being explored. Initial investigations, however, show that the LHIN sub-regions only align to LHIN boundaries.

PHU and LHIN boundaries (and LHIN sub-regions), in many parts of the province do not line up, making any type of data aggregation next to impossible in many areas (Appendix A).

Postal Codes

Postal codes were designed specifically for the purposes of efficient postal delivery and therefore, do not take into account administrative boundaries (including any health related boundaries). Postal codes are made up of forward sortation areas (FSA) (the first three characters) and local delivery units (LDUs) (the last three characters). The first character of the postal code essentially designates the code to a province/territory, with characters 'K', 'L', 'M', 'N' and 'P' falling within Ontario. The more characters available from a postal code, the more likely one can assign an area covered by that postal code. The areas of postal code coverage do not align well with administrative boundaries, and therefore aggregating to other geographies using postal codes generates additional concerns. Particularly in rural areas, postal codes or FSAs may include disparate areas that don't align well with the question being asked.

Full postal codes are often represented by points within a geographic information system (GIS) instead of a polygon, which represents an area. This method of representation further complicates trying to assign postal codes to administrative levels of geography. Statistics Canada has created a product known as the Postal Code Conversion File (PCCF) to aid in the assignment of a postal code to a census DA. The PCCF is a data file which provides a single link called the single link indicator (SLI) which assigns a postal code to the best single match census geography.[4,6]

A companion product (PCCF+) uses a population weighted method to assign postal codes to all census geographies covered. The PCCF+ is only available in SAS. Representing postal codes using points may convey an inaccurate message of the distribution of the events being attributed to or within the small area.

Custom boundaries

The initial small area analysis question should be reviewed to determine if existing, predefined boundaries can be used. If not, either custom geographic boundaries may need to be developed, or the question may need to be modified in order to use predefined boundaries. An example of custom boundaries are health unit neighbourhoods. The process of creating custom boundaries will depend on many factors (e.g. physical and demographic aspects of neighbourhoods)[7] and is beyond the scope of this document. To help you decide whether or not custom boundaries should be created as part of your SAA request, we have provided a list of questions and issues to consider:

- Could the question be answered using existing geographies?
    - Due to the complexity and number of factors involved, the process of creating custom geographies should be avoided wherever possible.
- What type of geographic information is available in the data that I am working with?
    - Do these data support the planned creation of custom geographies (i.e., the tabular data are already aggregated to such geographies)?
    - Is auxiliary data readily available (e.g., denominator data)? Custom geographies may require additional work to align data to match the new customized aggregated boundaries. If custom geographies are made up of existing census geographies (e.g., combinations of Census Tracts and Dissemination Areas), additional work would be required to combine the data to match the new customized aggregated boundaries.
- What am I trying to compare across the geographic areas?
    - Are there characteristics of the geography (e.g., per cent of low income housing, population density, diversity of community resources) or issues with existing geographies that should be avoided that may influence the outcome based on the aggregation of the data?
- How will the results be communicated? Depending on the target audience, geographic areas should be based on familiar or relevant frameworks (e.g., municipalities, PHUs).
- What are the time and resources available? Custom boundaries often require additional time and resources compared to using predefined boundaries.
    - Custom geographic boundaries require someone with GIS experience and knowledge to ensure the boundaries are generated correctly and can ensure the created boundaries are in the proper format for statistical, analytical and visualization purposes.
- Will the data be used for a one time purpose, or for ongoing analysis? If you need to be able to compare the geographic area over multiple years or on an ongoing basis, you should account for changes or updates to geographic boundaries. Some of the determinants that directed the original neighbourhood boundaries may be changed by built environment processes (e.g. new roadways) or changes in the underpinning neighbourhood characteristics due to effects like gentrification.

Types of small area methods

Methods in small area analysis can be classified in many ways, for example, direct estimation versus indirect estimation, model based versus design based, and Bayesian versus the frequentist approach.[8-12] Small area methods can range in complexity from simpler techniques such as pooling or aggregating datasets to increase sample size[13] to complex model building.[12] Although simpler methods can be easily interpreted, and are reproducible and transparent, modelling methods can be used to produce more precise estimates, but often require more expertise and time to implement. There is no one-size-fits-all solution; the SAA approach depends on the nature of the data and the specific question being answered. The choice of small area methods often depends on context, client needs, and availability of both data and staff resources. Four common small area methods are described and compared below, in addition to a decision aid for to assist users select a SAA method (Table 3).

Pooling data

Pooling data consists of combining data over time or space to increase sample size for the small area.[14-17] Before pooling, the data should be assessed for comparability of variables in measuring the same quantities (e.g., sample design, survey questions, modes of interviewing) and similarities in the target group characteristics (i.e., comparability of samples).[18,19] Think about what the target population of the pooled sample represents, as an artificial population is created when multiple surveys are combined (e.g., different populations surveyed at different points in time). This may be of particular concern if the phenomenon being measured is changing over time (e.g., smoking rates) or the underlying population is changing (e.g., a growing population in a region of new development).

Two common approaches are often discussed when combining data from multiple surveys: separate and pooled. In the separate approach, estimates are calculated for each survey separately and then combined. For the pooled approach, data are combined to create a single data file where survey weights are adjusted and/or scaled, and the data are treated as one large sample. Generally the resulting estimate will vary based on the method selected. Depending on the goal of the analysis, one method may be preferred over the other.[13]

ASSUMPTIONS/REQUIREMENTS

- The variables being combined measure the same thing and were measured the same way.
- The concept/characteristic being measured is comparable across the datasets/cycle.
- The populations being targeted/sampled by the different sources are similar/comparable.
- The estimate should be similar between the cycles to produce a meaningful combined estimate.
- Geography boundaries have not changed between cycles/surveys.
- Requires independence between samples for variance calculation.

ADVANTAGES

- Simple to conduct, relatively straight forward.
- Easy to interpret.

LIMITATIONS

- Changes in the indicator rate over time are obscured if pooling data over multiple time periods.
- Artificial population is created when multiple surveys are combined.

Synthetic estimation – Indirect

Synthetic estimation is a term used to describe an indirect method where a reliable direct estimator is obtained for a large area that spans several small areas and then used to derive an indirect estimator for a small area. Generating estimates using synthetic estimation involves applying estimates from the large area for specific population factors (e.g., age group, sex, social class) to the small area population composition. Synthetic estimation assumes that the estimates for each subgroup apply uniformly in the small area as the large area and the estimate for the large area is unbiased and valid.[11,20] Often the weighted average of the mean values for characteristics of the subgroups in the large area are

calculated with weights that are proportional to the distribution of the subgroups in the small area population.

ASSUMPTIONS/REQUIREMENTS

- Large area needs to be of sufficient size so that a direct estimate is reliable.
- Information correlated with the variable of interest is available at the small area level to derive an estimate that adjusts for compositional differences in small areas.
- Assumes the estimate of proportion for a given large area applies uniformly to each and every small area. This means that it assumes the differences in the outcome of interest are solely due to differences in the socio-demographic composition.
- Assumes that the same deterministic relationship (i.e., exact relationship) between the variable of interest (response) and the auxiliary variables (predictors) holds across a range of small areas.

ADVANTAGES

- Simple to conduct, relatively straight forward.
- Can allow for changes in the proportion of outcome by demographic characteristics of the small area.
- Easy to interpret.

LIMITATIONS

- Difficult to verify if small areas are homogenous within the large area.
- Does not take into account differences between the small areas that are not explained by the predictors.

Synthetic estimation – Regression model

Synthetic estimates can be improved by using a model for the construction of the estimate, as this allows for the effect of variables (e.g., age and sex) and the interactions between these factors, on outcomes of interest (e.g., presence of disease or risk factor) to be estimated. A regression based synthetic model typically involves using standard generalized linear modelling techniques to select a model to predict the outcome for each small area. The regression model estimates are then applied to the auxiliary data at the small area (e.g., Census) using the same explanatory variables.

Depending on the outcome of interest, this could include linear models for continuous data or generalized linear models (e.g., Poisson or logistic). As most public health data are area level data with count (discrete) data, a Poisson model can be used. Area level models are also chosen when person level data are not available, or when reliable auxiliary data are only available at the area level.

ASSUMPTIONS/REQUIREMENTS

- Requires the availability of high quality auxiliary data to area-level and/or unit-level that are potentially correlated (both theoretically and observed) with the variable of interest.
- Requires access to statistical software.

- For the Poisson distribution, the variance is equal to the mean. Where the variance is greater than the mean (over dispersion), adjustments to the Poisson model is required to account for this variance, either by using a quasi-Poisson regression, or a negative binomial regression.

ADVANTAGES

- Able to produce a higher level of accuracy compared to synthetic estimation – indirect.
- Area level auxiliary data requires area-based data (i.e., grouped data vs. individual data).
- Able to predict estimates for areas with sparse or no samples.

LIMITATIONS

- Requires a higher level of statistical expertise to implement and interpret results.
- Works best when all relevant auxiliary variables that help predict the response variable are available, accurate, and can be included in the model.
- Model building and validation can be resource intensive.

Synthetic estimation is appealing because of its simplicity. The demographic characteristics of the small area are often available from the Census, and the direct estimates by these demographic classes are easily obtainable from national/provincial surveys.[21] Because this method is based on the assumption that rates for each subgroup apply uniformly across all areas,[22] it assumes that the differences in health behaviour measures between areas are due solely to differences in their demographic composition. If two areas had the same composition with respect to the demographic variables used, they would have the same expected prevalence rates.[23] Thus, synthetic models assume that the outcome of interest is fully accounted for in the auxiliary data/model predictors, and that this relationship applies across all small areas.

Small area random effects models

A number of statistical techniques exist for deriving estimates using more sophisticated modeling approaches. Generally referred to in the literature as small area estimation (SAE), these techniques may include an additional error component to account for between-area variation, or variation between individuals within an area, that cannot be explained using only the auxiliary information. These techniques result in the production of less biased estimates with more confidence.

The trade-off of using these methods is they are more complicated to derive and operationalize within the context of population health assessment activities. Often small area models require a higher level of statistical skill and some familiarity with specialized statistical software. More complex models may not necessarily provide more accurate estimates, and results from simpler models or methods may be similar to those from complex ones. The benefit from using more complex methods needs to be considered against quality, cost, time, and effort required.

Model-based methods can take two approaches, based on data at either the unit (person) level or at the area level. A unit level model incorporates individual level data as inputs (e.g., an individual response to the CCHS) and links it with individual level auxiliary data, while an area level model takes area level measures as inputs (e.g., census data for a particular age group) and uses auxiliary data observed at the

area level. Area level summary measures can be constructed by aggregating individual level data, enabling one to apply an area level model to unit level data.

Both unit and area level models are a form of generalized linear mixed models with random effects on area level. Random effects models are extensions of traditional regression models. In addition to the fixed effects variables that are found in traditional regression models, a random effects model includes a "random effects" term for the data level of interest. Random effects are included in these models to account for variations between small areas or individuals within an area that cannot be explained only using auxiliary information. These random effects terms do not have coefficients associated with them; however, distribution assumptions must be satisfied for random effects to work. For example, one can add random effects μ_i on each area i in the dataset and assume they are independently normally distributed with mean 0 and variance σ_μ^2. The model will estimate all the μ_i and σ_μ^2 in addition to the coefficient estimations with the data.

Both unit level models (sometimes referred to as nested models) and area level models can borrow strength from related auxiliary data (e.g., census and administrative records) by including them as covariates. This further improves the estimation precision/reliability, which is one of the reasons that they are popular in the small area analysis literature. When the outcome is a continuous measure, a widely used model for creating small area estimates is the Fay-Herriot model. The Fay-Herriot model (FH model) is a random effects model that can be defined as:

$$\mathbf{Y_i = X_i\beta + e_i + \mu_i}$$

- Y_i is the direct estimator for area i
- X_i is a vector of auxiliary variables for area i
- e_i are the independently normally distributed error terms with variance σ^2 as in traditional linear regression accounting for sampling or measurement errors
- μ_i are the area level random effects

Once model parameters have been estimated using unit or area level methods, additional techniques can be applied to produce unbiased estimates of measures with greater confidence. One technique is the Best Linear Unbiased Predictor (BLUP) method. The BLUP method produces estimates as a weighted average of direct estimators and model based estimators, with the latter obtained from unit or area models mentioned above. Since BLUP relies on the assumption of normality of the outcome of the data, BLUP is applicable only for linear mixed models (e.g., outcome considered as continuous) and not for logistic or Poisson models (e.g., outcomes considered to be binary or count). When w_i are estimated by applying estimated variances σ_μ^2 and σ^2, the estimator is referred as Empirical BLUP (i.e., the EBLUP), which is the estimate of BLUP using the data.

The BLUP is constructed as a weighted average of direct estimator and regression based estimator using the auxiliary variables:

$$\hat{\mathbf{Y}}\mathbf{= w_i Y_i + (1- w_i) X_i\hat{\boldsymbol{\beta}}}$$

- Y_i is the direct estimator for area i
- X_i is a vector of auxiliary variables for area i

- w_i is the amount of total unexplained variation between areas, defined as the weight $(w_i) = \sigma_\mu^2 / (\sigma_\mu^2 + \sigma^2)$

This simple random effect model can be extended in many ways to model more complex data with additional assumptions. For example, unit level mixed effect models deal with individual level data by assuming e_{ij} are normally distributed. Spatial FH models borrow information from neighbours by assuming random effects are correlated to a neighborhood pattern. Empirical Bayesian (EB) and Hierarchical Bayesian (HB) can be used to produce BLUP when outcomes are normally distributed.

The EB method and the HB/full Bayesian method can be used to calculate estimates/predictors for models of binary and count based outcomes data. The EB method estimates the posterior density of the measure of interest by using empirical estimates of model parameters through their marginal distribution, while the HB method is a full Bayesian approach estimating all parameters through their posterior distributions.

Other techniques are able to:

- Incorporate survey weights (e.g., pseudo EBLUP)
- Borrow strength through:
    - Temporal smoothing (e.g., using multiple cycles of CCHS data and account for temporal correlation)
    - Spatial or spatiotemporal smoothing (e.g., spatial Fay-Herriot, Bayesian techniques to incorporate spatial correlations)
    - Auxiliary variables (e.g., more general model building approaches)

More recent research is focused on addressing the challenges of utilizing multiple small area estimation techniques simultaneously (e.g., Bayesian spatial modelling with survey weights) in order to produce less biased/unbiased estimation with more precision, and with more efficient computations.

ASSUMPTIONS/REQUIREMENTS

- All assumptions/requirements in regression based methods in the previous section.
- Outcomes are continuous measures (i.e., cannot be binary or count measures).
- Area level random effects does not depend on sampling or measurement errors (i.e., e_i and μ_i are independent).

ADVANTAGES

- Can account for areas that have unique characteristics that are not adequately explained by the auxiliary variables.
- Incorporates random effects which allow certain small areas to have different characteristics than that predicted by the auxiliary variables.
- Better able to predict differences between small areas compared to regression based synthetic models.

- Incorporating a random effect variable into the model can take into account between area variation that is not accounted for by the auxiliary variable.

LIMITATIONS

- More complex than synthetic models.
- Estimating the random effects terms is generally complex and may require Bayesian estimation techniques depending on the assumed structure and distribution of the random effects.

Decision aid for selecting a SAA method

The decision aid (Table 3) is intended to help users understand the benefits and limitations of the SAA methods presented in this document. It enables users to identify appropriate options for their situation and/or question, rather than simply providing them with an answer. Decision aids can be used when there is no single 'best' choice and the end user is recognized as the expert for judging values, consequences, and trade-offs to select an option most appropriate for the question they are trying to answer. As this document does not provide an exhaustive list of SAA techniques, users may identify SAA methodologies not mentioned in this decision aid that could better address their small area needs.

Table 3: Decision aid framework for assessing small area methodologies

Question	Direct	Combining surveys	Synthetic indirect	Synthetic model	Model (Fay Herriot)
What data is needed for this analysis?	Estimates are derived from a single survey file	Two or more survey files that are comparable (e.g., question asked, sampling frame, survey collection method)	Survey file and auxiliary data*	Survey file and high quality auxiliary data*	Survey file and high quality auxiliary data*
Time and resource(s) required to produce estimate?	Low	Low to Moderate	Moderate	High	High
When might this method be used?	Producing routine/ongoing estimates	Producing routine/ongoing estimates, research or specific policy questions	Producing routine/ongoing estimates, research or specific policy questions	Producing estimates for research or specific policy questions	Producing estimates for research or specific policy questions

Question	Direct	Combining surveys	Synthetic indirect	Synthetic model	Model (Fay Herriot)
Will I need special software to do this analysis?	Can be done using a standard statistical package	Can be done using a standard statistical package	Can be done using a standard statistical package	Can be done using a standard statistical package	R offers a free SAE package add-on Statistics Canada offers a G-EST macro in SAS
What level of SAA knowledge is required?	Low – estimates are produced using appropriate survey weights	Moderate – estimates are produced using a separate or pooled approach	Moderate – estimates are produced then projected to population estimates/ projections	High – requires knowledge of regression. Wide variety of models to choose from to model particular types of data	High – requires knowledge of random effects regression models. Wide variety of models to choose from to model particular types of data. Additional BLUP/EBLUP analysis can be applied to improve estimate.
How easy are the findings to interpret and communicate?	Easy to understand and communicate to lay audience	Easy to understand and communicate to lay audience	Fairly easy to understand and communicate to lay audience	Understanding and communicating model based results are more complex	Understanding and communicating model based results are more complex
Is the estimation method transparent so others can replicate?	Simple to replicate Methods are transparent	Simple to replicate Methods are transparent	Simple to replicate Methods are transparent	Requires knowledge of modelling methods	Requires knowledge of advanced statistical methods

*the Australian Bureau of Statistics[22] describes auxiliary data as one or more variables obtained from either administrative or census that are included in the model as explanatory variables. The auxiliary data should:

- Comprehensively cover the entire population scope for which small area estimates are required. If an auxiliary data item is not available for the unselected part of the population then small area predictions cannot be made and affected data items cannot be included in the model
- Include reliable geographic information so that all units belonging to a small area can be accurately identified
- Be contemporaneous with the target variable and other auxiliary data used in the model

Case studies

Examples of small area methods are described below based on data from the 2007/08 CCHS (unless otherwise stated) using *current smoker* as the output of interest. For additional information on the CCHS variables used refer to Appendix B.

Pooled approach

Data from the 2007/08 and the 2009/10 CCHS cycles were appended using the pooled approach.[13] For this example, *current smoker* (the output of interest) was reviewed for consistency (i.e., wording, skip pattern) across the two CCHS cycles. To adjust for increased sample size using the pooled approach, weights were scaled by a constant factor, by multiplying it by the inverse of the number of cycles combined.

$$\alpha \text{ (constant factor)} = \frac{1}{k} \quad \text{where k is the number of pooled cycles}$$

For this example, the sampling weights were multiplied by the constant factor ½ because two cycles were combined.[24] For statistics such as ratios, proportions and means, using the original weights or the weights that have been adjusted using a common constant factor will give the same result.[13]

Synthetic Indirect

Using current smoker as the output of interest from the CCHS, we created estimates for age and sex subgroups, as this information was readily available in the dataset (Table 4). Although we created subgroups by age group and sex in this example, other groupings could be used to adjust for compositional differences in the small areas. Tables were created to get a sense of how the CCHS age and sex distribution compared to Census age and sex distributions (Appendix C). Although we grouped age into groups (20-44, 45-64, and 65 +) for illustrative purposes, categories should be defined and selected so they will improve the estimator and to provide sufficient sample to ensure stability. Additional analysis should be undertaken to ensure that this holds true.

The synthetic estimate can be calculated using:

$$\text{Estimate of current smokers in small area} = \sum \left(\frac{\text{No. of current smokers in the large area for each group}}{\text{Population count in large area for each group}} \times \text{Population count in small area for each group} \right)$$

Using PHU A as the large area, estimates for current smokers are calculated by age/gender (Table 4).

Table 4: 2007/08 CCHS estimates of current smokers by age/sex groups for PHU A

Age/Sex Group	Current smokers – males (PHU A)	Current smokers – females (PHU A)
20-44yrs	0.23	0.32
45-64yrs	0.14	0.18
65+yrs	0.04	0.07

To estimate the number of current smokers in CSD 1 (a CSD within PHU A), we would apply the age/sex group estimates obtained for PHU A (Table 4) to CSD 1 sub-group populations (third column in Table 5). Note, as CCHS data are used in this example, CV results should be reviewed according to Statistics Canada guidelines. If CVs are large, one may want to consider using a different reference such as Province (Appendix D) or larger subgroup aggregates (e.g., adults 20-64yrs).

Table 5: Population and estimated number of current smokers by age/sex group in CSD 1

Age/Sex Group	Current smoker estimate (PHU A)	CSD 1 population (from 2006 Census)	Estimated number of current smokers in CSD 1
Male 20-44	0.23	26685	6142
Male 45-64	0.14	21255	3062
Male 65+	0.04	8545	373
Female 20-44	0.32	29190	9310
Female 45-64	0.18	22605	4145
Female 65+	0.07	10825	760
Total		119105	23795

The overall estimated proportion of current smokers in CSD 1, based on this method would be 20.0%:

$$\text{Estimated number of current smokers (CSD 1)} = (0.23 \times 26685) + (0.14 \times 21255) + \cdots (0.07 \times 10825)$$
$$= 23795 \text{ current smokers}$$

$$\text{Predicted estimate} = \frac{\text{Number of current smokers in CSD 1}}{\text{CSD 1 population over 20yrs}}$$

$$= \frac{23795}{119105} = 0.20$$

Regression synthetic model – Poisson

The data for the model was prepared as follows (Figure 6a):

- The number of current smokers by CSD was obtained from the 2007/08 CCHS
- Estimates for current smokers were calculated for each CSD
- Auxiliary information from the 2006 Census were obtained for variables of interest (proportion of people in small area (Pop2006over20), proportion of people per age group and sex (M2, M3, M4, F2, F3, F4), proportion of people with low income (LowInc), proportion of people with high school education (Edu1), proportion of people unemployed (Unemploy)) for each CSD of interest
- The 2007/08 CCHS and 2006 Census datasets are then joined by the CSD variable

Figure 6a: Sample SAS input dataset of current smokers from 2007/08 CCHS and auxiliary information from 2006 Census

geodcsd	n1	Estimate	M2	M3	M4	F2	F3	F4	Edu1	LowInc	Unemploy	Pop2006over20
3524001	54	0.1540961609	0.3322749346	0.264661935	0.1064001992	0.3422039859	0.2650058617	0.126905041	0.0565353276	7.7	5.3	119105
3524002	80	0.1958272698	0.3397910731	0.262804685	0.1373219373	0.3329431047	0.2659798642	0.1698665418	0.072973124	7.1	4.6	124105
3524009	21	0.2450735443	0.4154131848	0.2304549675	0.0748375116	0.414029243	0.2341291875	0.091615769	0.090664557	3.7	3.6	39390
3524015	37	0.310840938	0.3487057966	0.2582938389	0.0911410864	0.3586566092	0.2498204023	0.1120689655	0.1034201954	3.6	4.1	39210
3525005	304	0.2507502804	0.3436577871	0.2649423065	0.1310771485	0.3356639305	0.2625205215	0.1668372767	0.1570981211	14	6.5	379780
3526003	23	0.359166058	0.2980769231	0.2864010989	0.1624313187	0.2944046845	0.2888744307	0.1945348081	0.1512578616	7.5	6.7	22830
3526011	29	0.5643822309	0.2974018795	0.2913211719	0.1857379768	0.2782426778	0.2798117155	0.2384937238	0.1613742842	8.8	7.4	14620
3526014	.	0	0.2857142857	0.3166421208	0.1354933726	0.3057722309	0.3166926677	0.1404056162	0.1565934066	4.4	4.6	4950
3526021	9	0.2885463894	0.3298429319	0.2580403889	0.095736724	0.3335889571	0.2507668712	0.1050613497	0.195193008	3.1	4.8	9065
3526028	11	0.1788045721	0.2668344871	0.3152926369	0.1560730019	0.2706812652	0.3169099757	0.1763990268	0.0679554774	3.2	5.6	12145
3526032	43	0.2577967119	0.3285538714	0.2763211798	0.1493240475	0.3204089506	0.2758487654	0.1857638889	0.157622739	10.2	7.3	38680

Using SAS, the dependent variable (Smokcases_over20 calculated by multiplying the Estimate*Pop2006over20) and independent variables (M2, M3, M4, F2, F3, F4, LowInc, Edu1, Unemploy) was inputted into a Poisson model to produce a current smoking estimate (Predicted Value) for each CSD domain of interest. A copy of the SAS code to produce the results can be found in Appendix E.

Figure 6b: Output from SAS Poisson model for current smokers

	geodcsd	n1	Estimate	Pop2006over20	Smkcases_over20	Predicted Value	Lower Confidence Limit	Upper Confidence Limit
117	3524001	54	0.1540961609	119105	18353	21521.005334	19747.354409	23453.960515
118	3524002	80	0.1958272698	124105	24303	26043.461436	24095.937699	28148.39132
119	3524009	21	0.2450735443	39390	9653	9091.0801416	7784.3057528	10617.226605
120	3524015	37	0.310840938	39210	12188	8653.2264423	7966.1952103	9399.5095381
121	3525005	304	0.2507502804	379780	95229	80077.101995	75475.213154	84959.578064
122	3526003	23	0.359166058	22830	8199	5490.3265647	5163.4630127	5837.8816142
123	3526011	29	0.5643822309	14620	8251	4148.6335204	3779.0588919	4554.3508526
124	3526014	.	0	4950	0	1121.4936946	1014.2286413	1240.1031245
125	3526021	9	0.2885463894	9065	2615	2052.058321	1853.2178196	2272.2333599
126	3526028	11	0.1788045721	12145	2171	2576.2711506	2322.2211487	2858.1141141
127	3526032	43	0.2577967119	38680	9971	9547.7936252	9139.9435409	9973.8431317
128	3526037	10	0.3964271174	13770	5458	3423.5020741	3220.8581144	3638.8956094
129	3526043	58	0.3343040671	62810	20997	14963.198542	14263.317272	15697.422019
130	3526047	7	0.1156463127	11640	1346	2414.6691049	2059.1221968	2831.6080005

The overall estimated number of current smokers in CSD 1 (GEODCSD 3524001), based on this SAS Poisson model would be 21,521 (Figure 6b).

Random effects regression model – Fay Herriot

The input data was prepared as follows (Figure 7a):

- The number of current smokers by CSD was obtained from the 2007/08 CCHS
- Weighted estimate counts (Estimate) and variance (bs_var) was produced
- Auxiliary information from the 2006 Census were obtained for variables of interest (number of people in small area (totalpopover20), number of people per age group and sex (M20to44, M45to64,M65plus, F20to44, F45to64, F65plus), number of people with low income (lowinc), education status (no highschool)) for each CSD of interest
- The 2007/08 CCHS and 2006 Census datasets are then joined by the CSD variable

Figure 7a: Sample R dataset of current smokers from 2007/08 CCHS and auxiliary information from 2006 Census

geodcsd	samplesize	Estimate	bs_var	bs_sd	bs_cv	totalpopover20	M20to44	M45to64	M65plus	F20to44	F45to64	F65plus	nohighschool	lowinc
3524001	54	20721.32	10538722.21	3246.34	15.67	119105	26685	21255	8545	29190	22605	10825	5065	12752
3524002	80	26583.64	13784633.06	3712.77	13.97	124105	26835	20755	10845	28440	22720	14510	6530	11673
3524009	21	7083.02	10274873.57	3205.44	45.26	39390	11185	6205	2015	11185	6325	2475	2865	1996
3524015	37	15903.07	18244320.8	4271.34	26.86	39210	9565	7085	2500	9985	6955	3120	3175	1990
3525005	304	96874.59	37513120.8	6124.8	6.32	379780	84435	65095	32205	86895	67960	43190	42140	70638
3526003	23	6609.46	5859411	2420.62	36.62	22830	4340	4170	2365	4525	4440	2990	2405	2244
3526011	29	13281.37	12975468.36	3602.15	27.12	14620	2690	2635	1680	2660	2675	2280	1550	1637
3526014	0	0	0	0	0	4950	970	1075	460	980	1015	450	570	290
3526021	9	1899.12	458725	677.29	35.66	9065	2205	1725	640	2175	1635	685	1340	408

SAE PACKAGE IN R

Currently R offers a list of packages that perform small area estimations. One of the conventional packages is named “sae” (Small Area Estimation). This package includes functions to estimate the EBLUPs of both area-level SAE models (i.e., Fay-Herriot model) and individual level models. Documentation is available describing how to run the sae package in R, and Fay-Herriot model (FH model) options.

For areas with a zero sample size, two options are available to create a FH model estimate. The variance estimate can be assigned a non-zero positive value or a value of zero.

If a non-zero variance value is assigned, one will need to determine what value should be assigned. As the variance value becomes larger, the estimate becomes more dispersed, meaning the outcome will be less homogenous with more extreme values. If one decides to assign a zero value for the variance estimate, the estimate will be based solely on the synthetic estimate, instead of being a composite estimate of the synthetic and direct estimate.

In this example, we will use the “mseFH” function to fit an area level model and assigned zero value to the variance estimate. Using the sae R package, the dependent variable (Estimate) and independent variables (M20to44, M45to64, M65plus, F20to44, F45to64, F 65plus, nohighschool, lowinc) were applied to a FH model. A copy of the R code detailing how the model is run and how results are constructed is available in Appendix F. The FH model produces count estimates (FH_count) for the number of current smokers, and the CV for the count and the Confidence Interval (CI) around the proportion can be

calculated based on the Mean Squared Error (MSE) variance estimate of the EBLUP, since EBLUP is an unbiased estimator.

Figure7b: Sample R output dataset of current smokers

geodcsd	samplesize	Estimate	totalpopover20	FH_count	MSE	FH_CV	SE	FH_proportion	UL	LL
3524001	54	20721.32	119105	23183.69402	1327138.506	4.96907428	0.009672264	0.194649209	0.213606846	0.175691572
3524002	80	26583.64	124105	26661.28549	1492047.373	4.581526689	0.009842423	0.214828456	0.234119605	0.195537307
3524009	21	7083.02	39390	7410.703984	552875.4661	10.03354016	0.0188767	0.188136684	0.225135153	0.151138215
3524015	37	15903.07	39210	8292.536177	397423.3742	7.602200655	0.01607792	0.211490339	0.243003062	0.179977616
3525005	304	96874.59	379780	87529.56597	6044173.277	2.808753834	0.006473458	0.230474396	0.243162374	0.217786417
3526003	23	6609.46	22830	4973.965637	206322.8545	9.13210483	0.019896091	0.217869717	0.256866055	0.178873379
3526011	29	13281.37	14620	3812.543464	225022.9995	12.44224221	0.032446367	0.260775887	0.324370767	0.197181007
3526014	0	0	4950	1318.75668	127.8558052	0.857423837	0.00228431	0.266415491	0.270892738	0.261938243
3526021	9	1899.12	9065	1742.122715	142532.7607	21.67098268	0.041647558	0.192181215	0.273810429	0.110552002
3526028	11	2343.79	12145	2177.277085	175394.2778	19.23508051	0.034483409	0.179273535	0.246861017	0.111686054
3526032	43	10044.59	38680	8885.654912	218727.0071	5.263344086	0.01209107	0.229722206	0.253420703	0.206023708
3526037	10	4635.01	13770	2942.13009	192296.1321	14.90470562	0.031845739	0.213662316	0.276079965	0.151244667
3526043	58	20652.64	62810	14967.43374	341891.6986	3.906581902	0.009309267	0.238296987	0.25654315	0.220050824

The overall estimated number of current smokers in CSD 1 (GEODCSD 3524001), based on this SAE R package would be 23,183 (Figure 7b).

G-EST SOFTWARE IN SAS

Upon request, Statistics Canada can provide a SAS package titled G-EST[‡] (Generalized Estimation System) free to the requester after a license agreement has been signed. Statistics Canada developed this software to produce small area estimates under the Fay-Herriot area level linear regression model with EBLUP estimation. The program is run through a SAS macro designed to run under SAS 9.3 and requires the IML (Interactive Matrix Language) add-on. Extensive documentation has been created by Statistics Canada that includes information on the parameters for the macro and inputs, as well as how to run the program in the SAS environment.

Following the macro call described in Statistics Canada's G-EST Small Area Estimation documentation, the program will produce an output dataset that includes estimates for each small area of interest. For situations where an area has no samples (either no data was collected for that area or the n value of the outcome of interest is zero) the G-EST macro will create a synthetic estimate instead of a composite estimate (which is a combination of the synthetic and direct estimate).

In this example the "FH" method was selected. A copy of the SAS macro and details of how the data was prepared are available in Appendix G. The FH model produces count estimates. The FH MSE can be used to calculate a CI, around a FH proportion.

[‡] http://www.statcan.gc.ca/pub/12-206-x/2016000/activity-activite-eng.htm

Figure 7c: Sample SAS output dataset for current smokers

GEODCSD	SAMPLE_ESTIMATE	SAMPLE_VARIANCE	FH_GVF_SMOOTHED_VARIANCE	FH_ESTIMATE_TYPE	FH_ESTIMATE	FH_MSE	FH_PREDICTED
3524001	20721.32	10538722.21	7936402.76	COMPOSITE	23675.74	2481560.33	24599.77
3524002	26583.64	13784633.06	58596935.06	COMPOSITE	30329.98	3822904.18	30841.06
3524009	7083.02	10274873.57	233612.84	COMPOSITE	7274.35	214407.79	9296.16
3524015	15903.07	18244320.8	512169.91	COMPOSITE	14950.51	425137.74	10801.33
3525005	96874.59	37513120.8	611543888.5	COMPOSITE	88821.79	8246784.29	92351.86
3526003	6609.46	5859411	198228.95	COMPOSITE	6528.19	182797.72	5772.93
3526011	13281.37	12975468.36	125872.38	COMPOSITE	12844.62	119491.43	5205.2
3526014	0	.	.	SYNTHETIC	1156.45	2289038.93	.
3526021	1899.12	458725	75517.84	COMPOSITE	1899.94	73143.31	2348.54
3526028	2343.79	984312.38	100187.53	COMPOSITE	2349.27	96071.09	2800.08
3526032	10044.59	5343594.74	311681.23	COMPOSITE	10026.46	275242.71	10018.25
3526037	4635.01	3708895.07	91214.58	COMPOSITE	4587.35	87765.75	3752.74
3526043	20652.64	18700712.04	1617243.73	COMPOSITE	18904.29	987662.44	16495.7

The overall estimated number of current smokers in CSD 1 (GEODCSD 3524001), based on this SAS macro would be 23,675 (Figure 7c).

Comparison of results from case studies

A summary of results for smoking status estimates for PHU A from the various estimate produced from the case studies above is presented in Table 8.

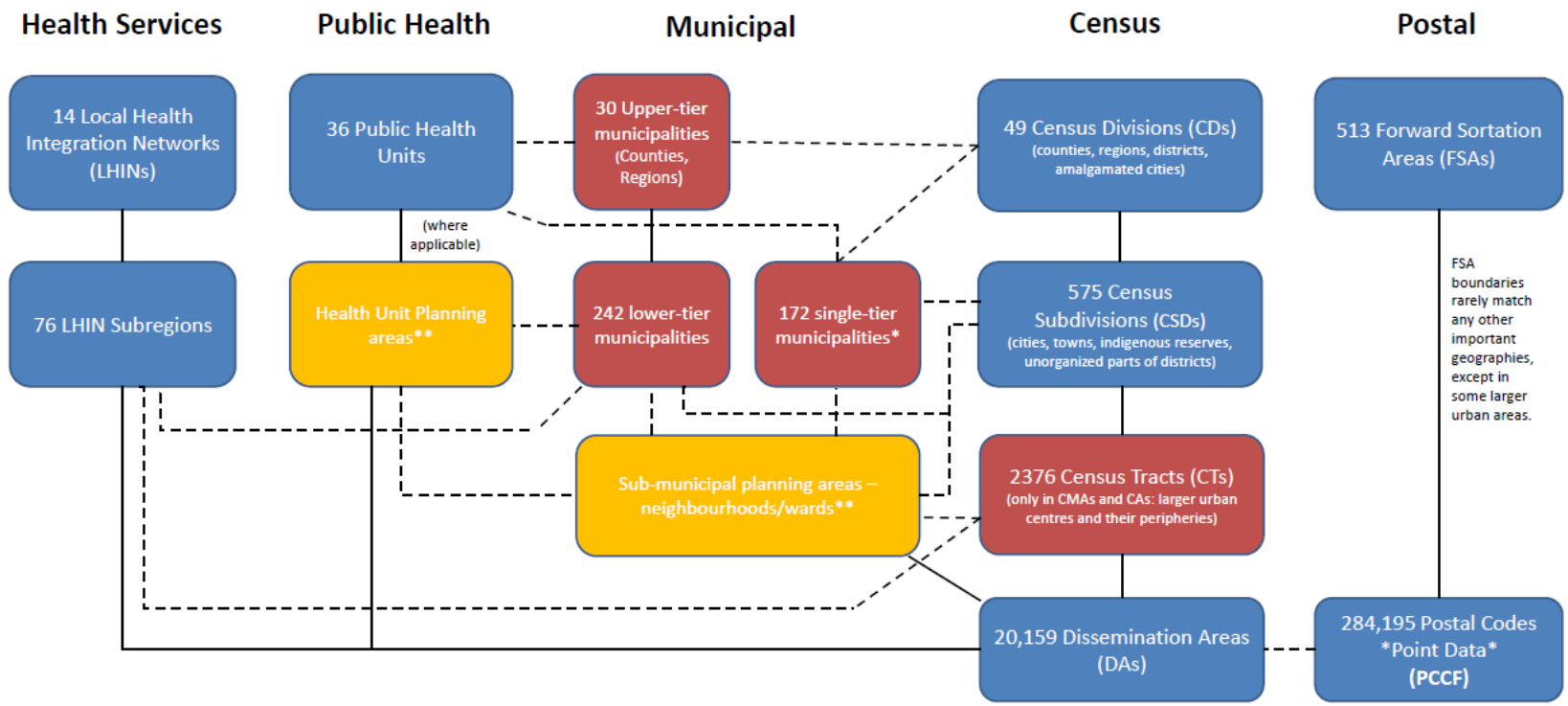
Table 8: Comparison of estimates using different SAA methods

Area	Population (20yrs+)	Direct Estimate	Pooled	Synthetic indirect	Synthetic model	FH SAE R package	FH G-EST SAS macro
CSD 1	119,105	15.4%	13.3%	20.0%	18.1%	19.5%	19.9%
CSD 2	124,105	19.6%	21.6%	19.3%	21.0%	21.5%	24.4%
CSD 3	39,390	24.5%	22.8%	21.5%	23.1%	18.8%	18.5%
CSD 4	39,210	31.1%	24.8%	20.4%	22.1%	21.1%	38.1%

- The results from the pooled estimates are similar to that of the direct estimate. Differences may be due to combining results from different time periods.
- Variation from the synthetic indirect compared to the direct estimate may be due to differences in age-sex composition. The assumption that the indicator of interest only varies by age and sex and does not capture all the variation between small areas and allow for other effects. This method still makes a strong assumption that the association between age and sex are similar between the small and large area.
- Synthetic models might lead to improved estimates that are more accurate or precise than direct estimates when there is high quality auxiliary data. Synthetic indirect estimates exhibit the least variation across CSDs of all methods presented. This is likely to be indicative of shrinkage, the reduction in distributional spread of small area estimates being more tightly centred around the overall mean than would be expected of the true small area values.
- Estimates produced by applying random effects models (i.e., FH model in the R package and SAS macro), can account for between-area variations other than those explained by auxiliary variables. The accuracy of these methods are dependent on both the accuracy of the model as well as the auxiliary data.

The estimates are quite variable depending on the method used. Selecting a method that provides the most reliable and defensible small area estimate can be a challenging decision. It will depend on a combination of factors such as the purpose of the estimate, assumptions you are willing to make, sound knowledge of social or economic dynamics behind the data, and good judgement.

Appendix A: Alignment of Ontario geographies



With few exceptions, LHINs have no regard to PHU or municipal boundaries. However, some LHIN Subregions were designed to correspond with municipal and/or PHU boundaries.

In Southern Ontario, PHUs match the boundaries of one or multiple upper-tier or single-tier municipalities. In 8 northern PHUs (and including Renfrew), these boundaries do not entirely match district boundaries.

Unorganized townships and remote areas in Northern Ontario (parts of 8 PHUs) have no municipal government.
Municipalities do not include indigenous reserves.

As of the 2016 Census

Postal Code Conversion File (PCCF) can link postal code points with any polygon (such as those to the left), and combine with other data, such as the Census.

Legend

Full coverage of province

Partial coverage of province

Boundaries, where used, are developed by local agencies., differs across province

Full geographic concordance

Partial/limited geographic concordance

*Single-tier municipalities may be amalgamated cities, like Toronto, Ottawa, or Chatham-Kent; separate from the county they are located in (such as cities of London, Windsor, Guelph) or in northern districts (Timmins, Wawa, Kenora).

**Used by City of Toronto, other larger PHUs/municipalities

Appendix B: Data for case studies

Data from the following sources are used to illustrate SAA methods.

Canadian Community Health Survey (CCHS) 2007/08

The Canadian Community Health Survey (CCHS) is cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population. It is designed to provide reliable estimates at the health region level every 2 years.

GEODHR4

The variable 'health region of residence of respondent' is used to define the large area

GEODDA06

The 2006 census dissemination area variable is used to define the small area

GEODCSD

The variable census sub-division is used to define the small area

DHH_AGE

The age variable is used to create post-strata age groups. The age groups varied depending on the outcome of interest.

DHH_SEX

The sex variable is used to create post-strata sex groups.

SMKDSTY

The variable 'type of smoker' was used to define the proportion of people who are current smokers. This outcome variable of interest was defined as SMKDSTY = DAILY SMOKER (1) + OCCASIONAL SMOKER (FORMER DAILY SMOKER) (2) + ALWAYS AN OCCASIONAL SMOKER (3)

2006 Census

The Census of Population (census) enumerates the entire Canadian population. It is designed to provide information about people and housing units in Canada by their demographic, social and economic characteristics. It is a reliable basis for the estimation of the population of the provinces, territories and local municipal areas. The census also provides information about the characteristics of the population and its housing within small geographic areas and for small population groups. This supports planning, administration, policy development and evaluation activities of governments at all levels, as well as data users in the private sector.

Appendix C: Comparability of survey to population

Basic sample coverage checks were conducted to assess whether or not the CCHS survey respondents were representative of the small area, based on census data.

Age group and sex comparison

Table 9: Weighted age-sex group proportions from the CCHS (overall) compared to the age-sex group proportions from the Census

Area	Age Sex Group	CCHS %	Census%
PHU A	F20 to 44	47.6%	46.8%
	F 45 to 64	35.0%	34.8%
	F 65+	17.5%	18.4%
	M20 to 44	49.4%	48.4%
	M45 to 64	35.7%	36.0%
	M 65+	14.9%	15.6%
CSD 1	F20 to 44	52.8%	56.0%
	F 45 to 64	32.9%	31.6%
	F 65+	14.3%	12.4%
	M20 to 44	41.3%	57.6%
	M45 to 64	38.3%	32.0%
	M 65+	20.4%	10.4%
CSD 2	F20 to 44	43.2%	43.3%
	F 45 to 64	36.5%	34.6%
	F 65+	20.4%	22.1%
	M20 to 44	49.4%	45.9%
	M45 to 64	29.8%	35.5%
	M 65+	20.8%	18.6%
CSD 3	F20 to 44	52.8%	56.0%
	F 45 to 64	32.9%	31.6%
	F 65+	14.3%	12.4%

Area	Age Sex Group	CCHS %	Census%
	M20 to 44	41.3%	57.6%
	M45 to 64	38.3%	32.0%
	M 65+	20.4%	10.4%
CSD 4	F20 to 44	53.9%	49.8%
	F 45 to 64	33.5%	34.7%
	F 65+	12.5%	15.6%
	M20 to 44	49.1%	49.9%
	M45 to 64	42.9%	37.0%
	M 65+	8.0%	13.1%

Number of respondents

General CCHS release guidelines from Statistics Canada dictate that users examine the unweighted number of sampled respondents. If this unweighted number is less than 10, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate.

Table 10: Number of sampled respondents and current smokers based on CCHS 2007/08 by area

Area	Total Sample	# Current Smokers
PHU A	1132	192
CSD 1	391	54
CSD 2	458	80
CSD 3	118	21
CSD 4	165	37

Appendix D: Estimates, CV, CI for reference population for synthetic indirect method

The quality of estimates produced with CCHS data is measured with the coefficient of variation (CV), produced using bootstrap weights. The CV magnitude will depend on the domain of interest and the prevalence of the characteristic. Disaggregating estimates further by age group or sex will increase the coefficient of variation.

Table 11: PHU A estimates from CCHS 2007/08 for current smokers

AgeSex grp	Sample size	Estimate	CV	LCI	UCI
Males 20-44	52	0.23	15.72	0.16	0.30
Males 45-64	25	0.14	24.04	0.08	0.21
Males 65+	7	0.04	53.34	0	0.09
Females 20-44	60	0.32	13.07	0.24	0.40
Females 45-64	35	0.18	21.81	0.10	0.26
Females 65+	13	0.07	29.97	0.03	0.11

Table 12: Ontario estimates from CCHS 2007/08 for current smokers

AgeSex grp	Sample size	Estimate	CV	LCI	UCI
Males 20-44	2071	0.32	2.89	0.30	0.34
Males 45-64	1630	0.24	3.77	0.22	0.26
Males 65+	437	0.10	7.72	0.08	0.12
Females 20-44	1782	0.21	3.08	0.20	0.23
Females 45-64	1593	0.19	4.23	0.17	0.20
Females 65+	582	0.08	6.23	0.07	0.09

Appendix E: SAS code for Poisson model

```
proc genmod data=SmkModel; /*source dataset*/

/*Dependent variable    Independent variables age-sex categories, level of education, income level, employment status*/

model smkcases_over20yrs = Male2 Male3 Male4 Female2 Female3 Female4 Education Income Employed/

dist =  Poisson /*choose Poisson distribution */

link =  log  /*a requirement to run the Poisson model */

offset = ln_population2006 /* Scale by population */

scale =  deviance; /*over dispersion adjustment  */

output out = Synthetic_indirect_method p=pred lower=lcl upper=ucl ; /*output predictions and confidence intervals*/

run;
```

Appendix F: Sample R code for FH model

Users can install the package by typing “install.packages(sae)” in the R console. Once the library is installed, it has to be loaded with “library(sae)” in order to access the functions come with the package.

There are two functions that can be used for Fay-Herriot model (FH model), “eblupFH” and “mseFH”. Selecting the function “eblupFH” will produce EBLUPs for the FH model. If the “mseFH” function is selected, the output will produce EBLUPs for the FH model as well as a Mean Squared Errors (MSE) of the EBLUP.

The MSE of an EBLUP could be used as the variance estimate of the EBLUP since EBLUP is an unbiased estimator. One can type “?” with the function name to see the help file on the usage of a particular function, e.g., “?mseFH”.

The “mseFH” function is used to fit an area level model. Once the data is read into R, the call function “result = mseFH(Estimate ~M20to44 + M45to64+ M65plus + F20to44+ F45to64+ F65plus+ nohighschool+ lowinc, vardir = bs_var, method = "FH", data = theData)” is applied. The functional call is very similar to any other regression call, with the only addition being “vardir = bs_var” to specify the sampling variance of the direct estimators.

****

```
require(sae)

theD = read.csv("data_for_sae_updated.csv")

theD[theD$bs_var==0,] #check to see if areas that have 0 variance

##FH estimate, this will not give estimate when estimated variance is 0, i.e. bs_var = 0

result = mseFH(Estimate ~M20to44 + m45to64+ M65plus + F20to44+ F45to64+ F65plus+
nohighschool+ lowinc, bs_var, method = "FH", data = theD)

##Eblup estimates when bs_var = 0 is 0

result$est$eblup[theD$bs_var==0]

####option 1, when bs_var is 0, make it a large variane, for example lets set it to maximum variance in the dataset

theD$bs_var[theD$bs_var==0] = max(theD$bs_var)

result = mseFH(Estimate ~M20to44 + m45to64+ M65plus + F20to44+ F45to64+ F65plus+
nohighschool+ lowinc,  bs_var, method = "FH",data = theD)
```

```
theD$FH_count = result$est$eblup[,1] #assign eblup back to dataset

theD$MSE = result$mse #assign standard error estimate back to dataset

theD$FH_CV = sqrt(theD$MSE)/theD$FH_count *100 #calculate CV for count

theD$SE = sqrt(theD$MSE)/theD$totalpopover20 #caclulate SE of proportion

theD$FH_proportion = theD$FH_count/theD$totalpopover20 #calculate estimated proportion

theD$UL = theD$FH_proportion + 1.96*theD$SE #calculate upper bound of proportion

theD$LL = theD$FH_proportion - 1.96*theD$SE #calculate lower bound of proportion

theD$LL[theD$LL < 0 ] = 0 #if lower limit is below 0 set to 0

theD$LL[theD$UL > 1 ] = 1 #if upper limit is above 1 set to 1



d1 = theD #assign a copy for comparision with option 2


####Option 2, this is what G-est from StatsCan does, for replace 0 estimates with synthetic estimators

theD = read.csv("data_for_sae_updated.csv")

result = mseFH(Estimate ~M20to44 + m45to64+  M65plus + F20to44+  F45to64+  F65plus+
nohighschool+  lowinc,

        bs_var, method = "FH",data = theD)


result_synthetic = predict(lm(Estimate ~M20to44 + m45to64+  M65plus + F20to44+  F45to64+  F65plus+
nohighschool+  lowinc,  data = theD), data = theD, se.fit = T)



theD$FH_count = result$est$eblup #assign eblup back to dataset

theD$FH_count[theD$bs_var==0] = result_synthetic$fit[theD$bs_var==0] #for 0 estimates, replace
synthetic estimators
```

```
theD$MSE = result$mse #assign standard error estimate back to dataset

theD$MSE[theD$bs_var==0] = result_synthetic$se.fit[theD$bs_var==0] #for 0 estimates, replace synthetic estimators

theD$FH_CV = sqrt(theD$MSE)/theD$FH_count *100 #calculate CV for count

theD$SE = sqrt(theD$MSE)/theD$totalpopover20 #caclulate SE of proportion

theD$FH_proportion = theD$FH_count/theD$totalpopover20 #calculate estimated proportion

theD$UL = theD$FH_proportion + 1.96*theD$SE #calculate upper bound of proportion

theD$LL = theD$FH_proportion - 1.96*theD$SE #calculate lower bound of proportion

theD$LL[theD$LL < 0 ] = 0 #if lower limit is below 0 set to 0

theD$LL[theD$UL > 1 ] = 1 #if upper limit is above 1 set to 1



###comparison of option 1 and 2

plot(d1$FH_count[d1$bs_cv=="."], theD$FH_count[theD$bs_cv=="."], xlab = "Option 1", ylab = "Option 2")
```

Appendix G: Sample SAS macro for G-EST

```
%Gest_SaeAreaFH(
/* specify Fay-Herriot method */
Method = FH,
/* specify to include an intercept in the model */
IncludeIntercept = 1,
/* specify that direct estimates for CSDs have been provided */
SampleEstimateType = DIRECT,
/* specify the smoothing method of generalize variance function */
SmoothingMethod = GVF,
/* specify not to run a mean model since we are modelling totals */
ApplyMeanModel = 0,
/* specify that results should be scaled up to the ON direct estimate */
Benchmark = 1,
/* ensure that the model only returns positive counts */
FullerMaxRule = 1,
/* specify the number of iterations used to solve the model equation (100000 is the max) */
MaxIter = 100000,
/* set to default */
Epsilon = 1e-10,
/* specify the small area variable name */
AreaGroup = geodcsd,
/* specify the dataset containing direct estimates and variance by small area */
DirectEstimateFile = sample_estimates,
/* specify the direct estimate variable name */
EstimateVariable = estimate,
/* specify the variance of direct estimate variable name */
VarianceVariable = bs_var,
/* not required as variance of direct estimates has been provided */
StandardErrorVariable = ,
CoefficientOfVariationVariable = ,
/* specify the dataset containing auxiliary data by small area */
AreaAuxiliaryFile = auxiliary_data,
/* specify the auxiliary variable names to be used in the model */
AuxiliaryVariables = m20to44 m45to64 m65plus f20to44 f45to64 f65plus popnodegree poplowinc,
/* specify the auxiliary variable names to be used for smoothing */
GVFVariables = m20to44 m45to64 m65plus f20to44 f45to64 f65plus popnodegree poplowinc,
/* specify the dataset containing the indicator name(s) and benchmark value(s) */
ParameterFile = parms,
/* specify the benchmark variable name in parameter file */
PopulationBenchmarkVariable = ONtotal,
/* specify the dataset containing auxiliary data for by small area */
AreaCountsFile = auxiliary_data,
/* specify the total population variable name for the small areas */
PopulationSizeVariable = Pop2006over20,
/* specify the sample size variable name (in auxilary dataset) */
SampleSizeVariable = n1,
/* not required */
ModelCoefficientFile = ,
ErrorCoefficientVariable = ,
/* specify the name of the output dataset containing the small area estimates and variance */
SaeEstimateFile = saa.sae_estimates,
DiagnosticPlots = 1 2 3 4, /* specify what model diagnostic plots to run */
/* not required */
DiagnosticCatalog = ,
/* set macro to generate small area estimates */
MacroExecutionMode = 3,
/* set to default */
GestOptions = TIME HEADER FILEINFO
);
```

List of Acronyms

BLUP	Best Linear Unbiased Predictor
CCHS	Canadian Community Health Survey
CI	Confidence Interval
CSD	Census Subdivision
CV	Coefficient of Variation
DA	Dissemination Area
EB	Empirical Bayesian
EBLUP	Empirical Best Linear Unbiased Predictor
FH model	Fay-Herriot model
FSA	Forward Sortation Areas
G-EST	Generalized Estimation System
HB	Hierarchical Bayesian
IML	Interactive Matrix Language
LDCP	Locally Driven Collaborative Projects
LDU	Local Delivery Unit
LIM	Low Income Measure
MSE	Mean Squared Error
mseFH	Mean squared error estimator of the EBLUP under a Fay-Herriot model
PCCF	Postal Code Conversion File
RRFSS	Rapid Risk Factor Surveillance System
SAA	Small Area Analysis
SAE	Small Area Estimation
SDA	Service Delivery Area
SGC	Standard Geographical Classification
SLI	Single Link Indicator
T1FF	T1 Family File

References

1. Bill 41, *Patients First Act, 2016*, SO 2016, c 30. Available from: https://www.ontario.ca/laws/statute/S16030

2. Ward M, Bennett K, Bains N. Evaluating dataset validity for small area estimation using combined cycles of the Canadian Community Health Survey. Presented at: Producing Reliable Estimates from Imperfect Frames: Statistics Canada 2013 International Methodology Symposium. 2013 Oct 16; Ottawa, ON.

3. Box GEP. Robustness in the strategy of scientific model building. Madison, WI: University of Wisconsin; 1979. Robustness in statistics; p. 201-36.

4. Statistics Canada. Postal code conversion file (PCCF), reference guide. Ottawa, ON: Minister of Industry; 2016. Appendix B: Hierarchy of standard geographic units for dissemination, 2011 census. Available from: http://www.statcan.gc.ca/pub/92-154-g/2016001/ap-an/ap-anb-eng.htm

5. *Areas Comprising Health Units*, RRO 1990, Reg 553 . Available from: https://www.ontario.ca/laws/regulation/900553

6. Wilkins R (Statistic Canada, Health Statistics Division). PCCF+ version 5F user's guide: automated geographic coding based on the Statistics Canada postal code conversion files including postal codes though July 2009 [Internet]. Ottawa, ON: Her Majesty the Queen in Right of Canada; 2010 [cited 2018 Feb 12]. Available from: http://odesi2.scholarsportal.info/documentation/PCCF+/V5F/MSWORD.PCCF5F.pdf

7. Parenteau M-P, Sawada M, Kristjansson EA, Calhoun M, Leclair S, Labonté R, et al. Development of neighborhoods to measure spatial indicators of health. J Urban Reg Inf Syst Assoc. 2008;20:43-55. Available from: http://create.usc.edu/sites/default/files/publications/levelingtheplayingfield-enablingcommunity-basedorganizatio_0.pdf#page=43

8. Pfeffermann D. New important developments in small area estimation. Stat Sci. 2013;28(1):40-68. Available from: https://projecteuclid.org/euclid.ss/1359468408

9. Goodman MS. Comparison of small-area analysis techniques for estimating prevalence by race. Prev Chronic Dis. 2010;7(2):A33. Available from: https://www.cdc.gov/pcd/issues/2010/mar/09_0026.htm

10. Jia H, Muennig P, Borawski E. Comparison of small-area analysis techniques for estimating county-level outcomes. Am J Prev Med. 2004;26(5):453-60.

11. Ghosh M, Rao JNK. Small area estimation: an appraisal. Stat Sci. 1994;9(1):55-76. Available from: https://projecteuclid.org/euclid.ss/1177010647

12. Rao JNK, Molina I, editors. Small area estimation. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2015.

13. Thomas S, Wannell B. Combining cycles of the Canadian Community Health Survey. Health Rep. 2009;20(1):53-8. Available from: https://www.statcan.gc.ca/pub/82-003-x/2009001/article/10795-eng.htm

14. Terashima M, Rainham DG, Levy AR. A small-area analysis of inequalities in chronic disease prevalence across urban and non-urban communities in the Province of Nova Scotia, Canada, 2007-2011. BMJ Open. 2014;4(5):e004459. Available from: http://bmjopen.bmj.com/content/4/5/e004459.full

15. Penney TL, Rainham DG, Dummer TJ, Kirk SF. A spatial analysis of community level overweight and obesity. J Hum Nutr Diet. 2014;27 Suppl 2:65-74.

16. Seliske L, Norwood TA, McLaughlin JR, Wang S, Palleschi C, Holowaty E. Estimating micro area behavioural risk factor prevalence from large population-based surveys: a full Bayesian approach. BMC Public Health. 2016;16:478. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-016-3144-4

17. Knutson K, Zhang W, Tabnak F. Applying the small-area estimation method to estimate a population eligible for breast cancer detection services. Prev Chronic Dis. 2008;5(1):A10. Available from: https://www.cdc.gov/pcd/issues/2008/jan/06_0144.htm

18. Wendt M (Statistics Canada, Social and Aboriginal Statistics Division). Considerations before pooling data from two different cycles of the general social survey [Internet]. Ottawa, ON: Her Majesty the Queen in Right of Canada; 2007 [cited 2018 Feb 12]. Available from: http://www23.statcan.gc.ca/imdb-bmdi/document/8011_D1_T9_V1-eng.pdf

19. Roberts G, Binder D. Analyses based on combining similar information from multiple surveys section on survey research methods Paper presented at: Statistics: from Evidence to Policy: JSM 2009. 2009 Aug 3; Washington, DC. Available from: https://pdfs.semanticscholar.org/70ef/9ec615640c12f029fe0d026b0166ca88e2a7.pdf

20. Singh VK, Seth SK. An efficient family of synthetic estimators for small areas and its applications. J Stat Appl Pro Lett. 2015;2(1): 59-69. Available from: http://www.naturalspublishing.com/files/published/qm50x6rx728no0.pdf

21. Rahman A. Estimating small area health-related characteristics of populations: a methodological review. Geospat Health. 2017;12(1):495. Available from: http://www.geospatialhealth.net/index.php/gh/article/view/495/544

22. Australian Bureau of Statistics . A guide to small area estimation. Version 1.1 [Internet]. Canberra, AU: Australian Bureau of Statistics; 2006 [cited 2018 Feb 12]. Available from: http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/3a60738d0abdf98cca2571ab00242664/$FILE/May%2006.pdf

23. Bajekal M, Scholes S, Pickering K, Purdon S. Synthetic estimation of healthy lifestyles indicators: Stage 1 report [Internet]. Prepared for the Department of Health. London, UK: National Centre for Social Research; 2004 [cited 2018 Feb 12]. Available from: http://old.iph.ie/files/file/Synthetic_Estimation_Stage_1_Report.pdf

24. Statistics Canada. The Research and Data Centres information and technical bulletin. Ottawa, ON: Minister of Industry; 2014;6(1). Available from: http://www.statcan.gc.ca/pub/12-002-x/2014001/article/11901-eng.pdf

Public Health Ontario
480 University Avenue, Suite 300
Toronto, Ontario
M5G 1V2

647.260.7100
communications@oahpp.ca
www.publichealthontario.ca