

Catalogue no. 11-522-X
ISSN 1709-8211

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

Statistical Disclosure Control Analysis for Small Area Estimation

by Cissy Tang

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

Statistical Disclosure Control Analysis for Small Area Estimation

Cissy Tang¹

Abstract

Currently, Statistics Canada has no official guidance on confidentiality rules for releasing small area estimate. In recent years, there has been increasing demand from Research Data Centre (RDC) researchers for comprehensive confidentiality guidelines such that they can publish small area estimates in their research. This confidentiality analysis applies to area-level small area estimation.

A simulation study is conducted in R to create simulated populations from which samples are selected. The simulated populations contain an auxiliary variable, a variable of interest, and domain information. The strength of the relationship between the auxiliary variable and the variable of interest is controlled through an “error” variable with a random component. Stratified random samples are drawn, and area-level small area estimates are calculated using the “sae” R package (Molina and Marhuenda 2015). The simulation is run for various sampling rates and various levels of association between auxiliary and response variables to observe the impact on disclosure risk and to identify potential areas of disclosure risk. The risk of disclosure of the small area estimate is compared against the direct Horvitz-Thompson estimate to demonstrate that small area estimates are inherently less risky than direct estimates, especially when sampling rates are extremely low. The results are then analyzed and finally, confidentiality guidelines for the release of area-level small area estimates are proposed. The paper will outline the simulation process and discuss justifications for the proposed confidentiality guidelines.

Key Words: Statistical disclosure control; Confidentiality; Small area estimation.

1. Introduction

1.1 Background

Currently, Statistics Canada has no official guidance on confidentiality rules for releasing small area estimates and prior to this, no official study has yet been conducted on the subject. In recent years, The Centre for Confidentiality and Data Access has received increasing demands from Research Data Centre (RDC) researchers for comprehensive confidentiality guidelines such that they can publish small area estimates in their research.

For survey and administrative data, the typical direct estimation minimum counts used are 5 and 10, respectively. These minimum counts refer to the minimum number of respondents or records needed for a weighted frequency (or unweighted frequency in the case of administrative data), and statistics such as mean, variance, proportion, and percentile, to be released. In small area estimation (SAE), the estimate is a linear combination of a direct estimate and a model estimate constructed from auxiliary data. Therefore, there is more noise in the resulting estimate compared to direct estimates, which suggests that the minimum counts for SAE won't need to be as strict as they are for direct estimates. On the other hand, domains used in SAE can potentially be so small that there are very few or no respondents in some domains. However, this does not necessarily imply a heightened risk of identity, attribute, or inferential disclosure as would be the case when the number of respondents is low in direct estimation. In SAE, when the number of respondents in a domain is extremely small, the variance of the direct estimate will be high, resulting in a SAE relying more heavily on the model estimate than the direct estimate. Thus, the possibility of removing the minimum count requirement entirely for releasing SAE is explored in this simulation study. The proposed removal of minimum count requirements for SAE will allow researchers to examine rare population subdomains where direct sample size may be lacking. The goal of this simulation study is to create disclosure control guidelines for SAE, both for use in Statistics Canada's Research Data Centres, as well as for official data releases.

¹Cissy Tang, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A0T6
(cissy.tang@statcan.gc.ca)

1.2 Types of Statistical Disclosure

There are three key types of statistical disclosure that will be used as a framework to assess disclosure control risk for SAE. The first type is identity disclosure. Identity disclosure happens when an individual in the population can be identified from a table (Hundepool, 2024, p.4). This typically occurs when there is a small cell with 1 or 2 respondents with a unique characteristic. In direct estimation, implementing a minimum count, usually 5 or 10, prevents the possibility of identity disclosure. The second type is attribute disclosure. Attribute disclosure occurs when traits are discovered about a sub-population, even if identities of members of the subpopulation are unknown (Hundepool, 2024, p.214). Attribute disclosure typically happens when zero cells or full cells are released, leading to intruders finding out that either everybody or nobody in a subpopulation has a certain trait. To prevent attribute disclosure, noise can be added so that zero or full cells are never released in frequency tables. The third type is inferential disclosure. Inferential disclosure happens when information previously unknown can be inferred with a high degree of certainty from released data, sometimes in combination with previously released data, or data available from other sources (Hundepool, 2024, p.218).

2. Simulation Study

2.1 Scope

The simulation study will cover area-level (Fay-Herriot) small area estimation for a probability survey. The Empirical Best Linear Unbiased Predictor (EBLUP) of the area-level (Fay-Herriot) small area estimate consists of a ratio of the direct estimate $\hat{\theta}$ and a synthetic estimate $z'_i\hat{\beta}$. The coefficient is determined by the variance and error of the direct estimate and the synthetic estimate, respectively. This is the estimator that will be used to assess the disclosure risk of releasing area-level SAE.

$$\hat{\theta}_i^{EBLUP} = \hat{\gamma}_i\hat{\theta}_i + (1 - \hat{\gamma}_i)z'_i\hat{\beta}, \quad \hat{\gamma}_i = \frac{b_i^2\hat{\sigma}_v^2}{b_i^2\hat{\sigma}_v^2 + \hat{\psi}_i},$$

where $\hat{\theta}_i$ is the direct estimate, $z'_i\hat{\beta}$ is the synthetic estimate, $\hat{\sigma}_v^2$ is the model variance, b_i are fixed constants used to account for heteroscedasticity, and $\hat{\psi}_i$ is the predictor of the smoothed design variance.

2.2 Simulation Methodology

A microdata file containing 50 mutually exclusive population groups (the domains) is simulated in R. Stratified random samples are then drawn, stratified by the domains. The variable of interest is a flag which is equal to 1 when the person has the characteristic, and equal to 0 otherwise. To assess its impact on disclosure risk, different sampling fractions are used. Assuming 100% response rate, the Horvitz-Thompson estimates and stratum variances are calculated using the design weight of each sampled unit. Then, the stratum variances are smoothed using a log-linear ordinary function, which improves the quality of the small area estimates (You and Hidioglou, 2024).

SAE relies on the presence of auxiliary information for all domains. This information will be used to build a model for synthetic estimates. The auxiliary information is important especially when the direct estimate is of poor quality, in which case the small area estimate will rely more on the synthetic estimate than the direct estimate. The auxiliary variable will be used to generate the synthetic estimate in the form of a model. To ensure that all aspects of the linear regression model can be controlled, including the intercept, the slope, and the error, the auxiliary variable will be generated first using arbitrary parameters. The auxiliary variable is then used to generate the “true” rate of the characteristic of interest in the population, which will be used to simulate the population matching those rates. Using this method, the degree of correlation can be varied between the “true” population rate and the auxiliary variable by specifying the intercept, slope, and error to mirror a variety of situations to investigate. In other words, the error term can be increased to simulate an auxiliary variable with very poor predictive ability or can be decreased to simulate an auxiliary variable with strong predictive ability.

From the direct estimate and the synthetic estimate mentioned above, the SAE is calculated for use in the disclosure risk assessment. The “sae” package in R (Molina and Marhuenda 2015) was used to calculate the SAE. The function used to calculate the Fay-Herriot estimator is eblupFH(). The estimation method used to estimate the model error in eblupFH() is the Restricted Maximum Likelihood Method (REML), which yields unbiased estimators of variances. Table 2.2-1 summarizes the simulation parameters that will be analysed and discussed in section 3.

Table 2.2-1
Simulation Parameters and Scenarios

Variable	Simulation Parameter(s)	
	Scenario A	Scenario B
Number of Domains	50	
Population Size	Randomly selected within the range [25000, 75000]	
Auxiliary Variable Rate (x)	Record-level binary variable in each domain generated to match the domain rate (uniform distribution where $\min=0.03$ and $\max=0.1$)	
Slope (b_1)	1.01	
Intercept (b_0)	0.005	
Error	Uniform distribution with ...	
	... $\min=0.005$ and $\max=0.05$... $\min=0$ and $\max=0$
Population Characteristic Rate	$= b_1 x + b_0 + error$	
Sampling Fractions	0.1, 0.01, 0.001	

3. Simulation Results and Analysis

3.1 Simulation Results and Analysis

The simulation is first run using ‘Scenario A’ parameters in Table 1, using sampling fractions of 0.1, 0.01, and 0.001. There is a randomly generated error term for the synthetic estimate in the following four data visualizations.

As the sampling fraction increases, the Horvitz-Thompson (direct) estimate becomes closer to the true population value. Figure 3.1-1 is a histogram of frequencies of the absolute difference between the Horvitz-Thompson(direct) estimates and the true population values used in the simulation.

Figure 3.1-1
Histogram of frequencies of the absolute difference between the Horvitz-Thompson(direct) estimates and the true population values used in the simulation. (Scenario A)

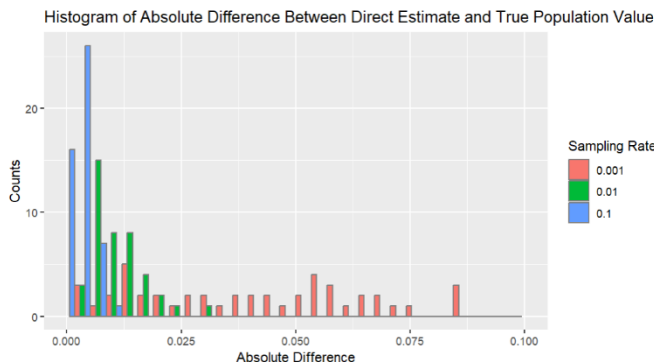


Figure 3.1-2
Histogram of frequencies of the absolute difference between the Fay-Herriot EBLUP (small area) estimates and the true population values used in the simulation. (Scenario A)

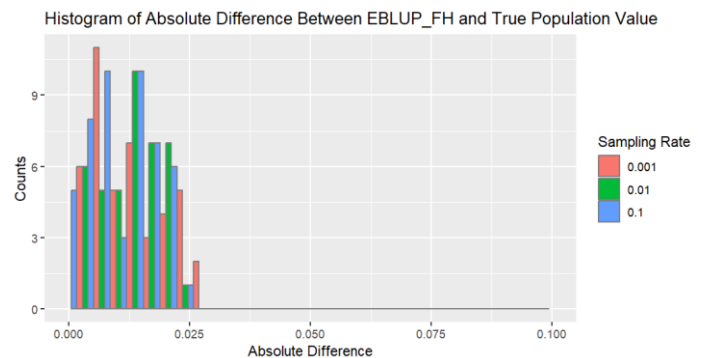


Figure 3.1-2 demonstrates that in this simulation scenario, SAE is more effective at estimating the true population parameter compared to the direct estimate. The data visualization in Figure 3.1-2 is a histogram of frequencies of the absolute difference between the Fay-Herriot EBLUP (small area) estimates and the true population values. The X-axis is the same for both histograms. The absolute difference when Sampling Rate = 0.001 is much smaller when SAE is used compared to direct estimation.

As the sampling rate decreases, the SAE is much less likely to be the same or close to the direct estimate. Figure 3.1-3 shows the absolute difference between the Fay-Herriot EBLUP (small area) estimate and the direct estimate. Even when the sample size is extremely small (under 5 respondents) for a given domain, releasing the SAE will likely not reveal the value of the direct estimate. While it can't be guaranteed that in every simulation, the SAE will be different than the direct estimate, the simulation demonstrates that as sampling rate decreases, the absolute differences between the SAE and the direct estimates increases, adding noise to the direct estimate.

Figure 3.1-3
Histogram of frequencies of the absolute difference between the Fay-Herriot EBLUP (small area) estimates and Horvitz-Thompson(direct) estimates. (Scenario A)

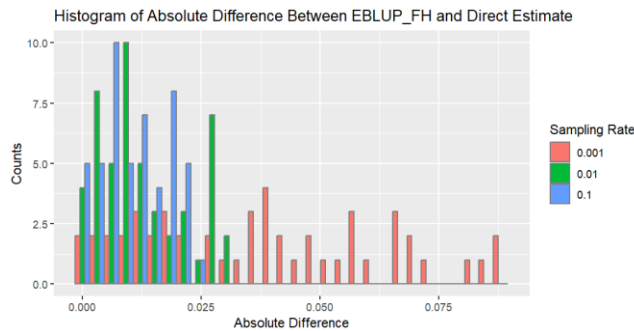
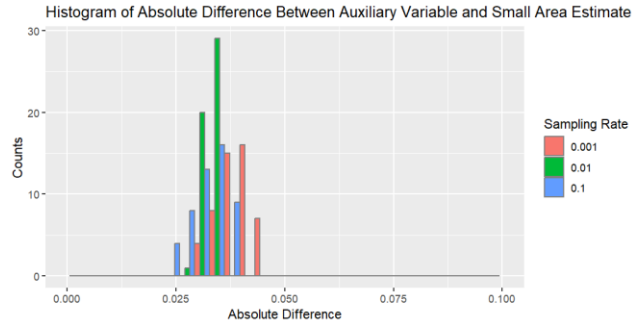


Figure 3.1-4
Histogram of frequencies of the absolute difference between the auxiliary variable and the Fay-Herriot EBLUP (small area) estimates (Scenario A)



The next step is to analyze what impact the size of the error term in the synthetic model has on disclosure risk. As the sampling fraction decreases, the SAE becomes closer to the synthetic estimate. Therefore, if the synthetic estimate is extremely close or identical to the auxiliary variable, then when the sampling fraction is small, releasing the SAE may present a disclosure risk for the auxiliary variable.

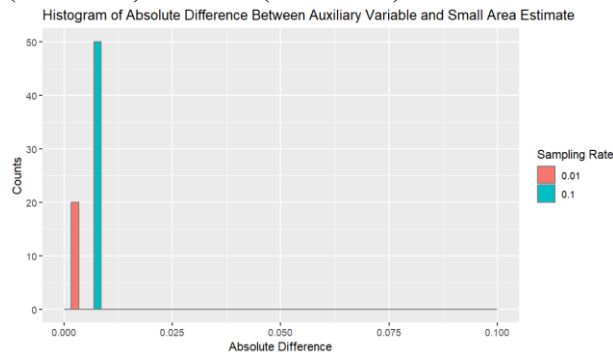
Figure 3.1-4 shows the absolute difference between the auxiliary variable and the SAE when there is a random error component to the synthetic estimate, for Scenario A. Even though the absolute difference decreases as the sampling rate increases, there is a large amount of variation and noise such that it would be difficult to re-derive the auxiliary variable from the SAE.

However, if the random error component present in the synthetic estimate is removed, the small area estimate becomes much closer to the auxiliary variable. Now consider a similar simulation (Scenario B, Table 2.2-1) where the randomly generated error term is removed (error = 0 for every simulated record).

In Figure 3.1-5, the histograms for Sampling Rate = 0.001 was not produced. In the process of a random simulation, there were one or more domains that had zero respondents with the targeted characteristic. This was not an issue for scenario A because since the error term is positive for all domains, the rates of the characteristic of interest in scenario A are strictly greater than the rates in scenario B. Thus, scenario B had a higher chance of drawing samples that results in a direct estimate value of 0 for each domain.

The `eblup_FH()` function in the “sae” R package does not work when there is a zero estimate for the variance in any of the domains. When the sampling rate is so small such that zero-estimates occur, the practical work-around would be to set the variance of the domain so large such that the `eblup_FH()` function is manipulated into outputting the synthetic estimate as the small area estimate for that domain. In Figure 3.1-5, even though histograms for Sampling Rate = 0.001 has not been produced, the absolute difference between the auxiliary variable and small area estimate is already very close to 0, even when the sampling rates are relatively large at 0.01 and 0.1. Figure 3.1-5 demonstrates that when the auxiliary variable has an extremely high predictive power for the variable of interest, releasing the small area estimate presents a disclosure risk for the auxiliary variable.

Figure 3.1-5
Histogram of frequencies of the absolute difference between the auxiliary variable and the Fay-Herriot EBLUP (small area) estimates (Scenario B)



3.2.1 Disclosure Risk from Direct Estimate

The simulation supports the hypothesis that when there are very few respondents (less than 5) sampled in a domain, the SAE of that domain is unlikely to reveal information that can be attributed to any real respondents. Thus, the disclosure risk of the small area estimate in this situation is low and a minimum count does not need to be imposed when releasing SAE.

A special case to consider would be a situation where a direct estimate is 0 or does not exist for a domain. In practice, this would be a domain in which the number of respondents is so few that the direct estimate is 0, in which no respondents have the characteristic of interest. In traditional direct estimation, a zero cell would not be released, as this may cause attribute disclosure. However, in SAE, when the direct estimate variance is 0, the `eblupFH()` function does not work as intended and needs to be manipulated by setting the variance to an extremely high number to produce the desired small area estimate, which is equal to the synthetic estimate. Therefore, in cases where there are zero respondents with the characteristic of interest in the target domain in the direct sample, the small area estimate is unlikely to cause an attribute disclosure through its direct estimate component. Thus, for cases like this, it is only necessary to evaluate the potential disclosure risk from the synthetic estimate.

It would also be prudent to not release the direct estimate and the unweighted sample counts at the same time as releasing the small area estimate. This will add an extra layer of uncertainty and protection.

3.2.2 Disclosure Risk from Indirect Estimate

The simulation shows that in the case where the direct estimate is composed of very few respondents, the small area estimate will be identical or extremely close to the synthetic estimate. Thus, it's important to ensure that the synthetic estimate itself is not disclosive.

A common disclosure concern in linear regression modeling is a saturated model. A saturated model is one where there is a parameter for each unique combination of the covariates. A saturated model perfectly predicts the auxiliary variables. In statistical disclosure control terminology, saturated models are called "table-equivalent", since the model itself implies descriptive tables of the auxiliary variables used in the model. If the model used to predict the synthetic estimate is saturated, then releasing a small area estimate for a domain implies the release of the auxiliary information as a table. Under this scenario, assuming a minimum count is not imposed for the number of respondents required for SAE to be published, the publication of data on a domain where there are zero or only a few respondents with the characteristic of interest can potentially cause an identity or attribute disclosure of the auxiliary variables for respondents in the domain. This would be a disclosure of the auxiliary variable, which must also be kept confidential under the Statistics Act. Therefore, it is recommended to not release SAE with synthetic estimates derived from a saturated model.

Another type of model to be wary of are regression models with high predictive power. To measure the strength of the auxiliary variable z_i for the prediction of θ_i , Hidiroglou, Beaumont, and Yung derived a formula to calculate the

R^2 coefficient in their paper “Development of a small area estimation system at Statistics Canada” (Hidiroglou, Beaumont, and Yung 2019). $0 < R^2 < 1$, where a value close to 0 implies a weak model and a value close to 1 represents a strong model. As Miller (2024) states, when the R^2 value is very close to 1, the relationship between the auxiliary variable and predictor is close to being deterministic. This type of model may cause inferential disclosure. If $\hat{\theta}_i$ is published for a domain with very few respondents, such that the SAE relies almost entirely on the model estimate, then the auxiliary variable z_i could be inferred for the population in that domain with high accuracy. This is an issue if the auxiliary variable is considered confidential, such as tax or health data. To reduce disclosure risk from this scenario in practice, the release of small area estimates can be restricted where the model used to calculate the synthetic estimates has an R^2 coefficient of greater than 0.95 (Miller 2024). The threshold of 0.95 is selected to be consistent with modeling confidentiality vetting rules in Research Data Centres.

4. Proposed Confidentiality Guidelines and Conclusion

4.1 Proposed Confidentiality Guidelines for SAE

Considering the simulation results, here are the proposed confidentiality guidelines for area-level (Fay-Herriot) small area estimation:

- Restrict the release of small area estimates where the synthetic estimate is derived from a saturated model.
- Restrict the release of small area estimates where the synthetic estimate is derived from a model where the R^2 coefficient is greater than 0.95.
- No restrictions on minimum count. That is, there is no minimum number of units required in the direct estimate for a small area estimate to be published.
- Recommend against releasing the direct estimate, sampling information, and unweighted descriptive data along with small area estimates.

4.2 Conclusion

In conclusion, the simulation study has demonstrated that small area estimates are safe to release even when the sample size is extremely small, except in cases where the synthetic estimate is based on a saturated model or where the synthetic estimate has extremely high predictive power. To officially implement the SAE confidentiality rules in Statistics Canada Research Data Centres (RDC), consultation with and approval from the RDC vetting committee is needed. Furthermore, consultation with users of SAE both within Statistics Canada and the external academic community is needed to ensure that the rules are sufficient, clear, and comprehensive before official implementation.

References

Hidiroglou, M.A., Beaumont, J.-F. and Yung, W. (2019), “Development of a Small Area Estimation System at Statistics Canada”. *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, 45(1). Paper available at [Development of a small area estimation system at Statistics Canada | Semantic Scholar](#)

Hidiroglou, M. A. and You, Y. (2024), ‘Application of Sampling Variance Smoothing Methods for Small Area Proportion Estimation’, *Proceedings of Statistics Canada Symposium* 2022.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E. S., Seri, G., Wolf, P.-P. D., Tent, R., Młodak, A., Gussenbauer, J., & Wilak, K. (2024), *Handbook on Statistical Disclosure control*, Centre of Excellence on SDC.

Miller, J. (2024) ‘Disclosure Risk of Parametric Regression Output’, Unpublished Statistics Canada Internal Paper.

Molina, I and Marhuenda, Y (2015), ‘sae: An R Package for Small Area Estimation’, *The R Journal*, 2015 -007.