USAID
FROM THE AMERICAN PEOPLE

DHS WORKING PAPERS

Design-based Small Area Estimation: An Application to the DHS Surveys

Ruilin Ren

2021 No. 180

DEMOGRAPHIC AND HEALTH SURVEYS

September 2021

DHS Working Papers No. 180

Design-based Small Area Estimation: An Application to the DHS Surveys

Ruilin Ren[1]

ICF
Rockville, Maryland, USA

September 2021

[1] ICF

Corresponding author: Ruilin Ren, International Health and Development, ICF, 530 Gaither Road, Suite 500, Rockville, MD 20850, USA; phone: +1 301-407-6500; fax: +1 301-407-6501; email: Ruilin.ren@icf.com.

Acknowledgments: The author is grateful for helpful reviews by Shireen Assaf and Thomas Pullum.

Editor: Diane Stoy
Document Production: Natalie Shattuck

This study was conducted with support from the United States Agency for International Development (USAID) through The DHS Program (#720-OAA-18C-00083). The views expressed are those of the authors and do not necessarily reflect the views of USAID or the United States Government.

Recommended citation: Ren, Ruilin. 2021. *Design-based Small Area Estimation: An Application to the DHS Surveys.* DHS Working Papers No. 180. Rockville, Maryland, USA: ICF.

CONTENTS

TABLES

FIGURES

ABSTRACT

This study explores design-based small area estimation methods using Demographic and Health Survey (DHS) data collected by The DHS Program, an international program funded by United States Agency for International Development. The DHS surveys are household-based, two-stage cluster surveys that provide key survey indicators for a country's first-level administrative unit, or region. The DHS Program has received increasing requests from host countries for subregional indicator estimates that can be used for policymaking and development planning. Increasing sample size is usually not feasible for meeting this need. One solution is using small area estimation techniques to produce reliable estimation of subregions. This study explores a method for creating a survey domain that covers a small area by pooling clusters or sample units close to the small area from one single target survey or similar surveys conducted in recent years. 'Close' can mean geographically, in time and space, or in other demographic, social, religious, cultural, or economic measures. A survey domain created in this way is easy to analyze with design-based domain analysis tools such as parameter estimation, variance estimation, and confidence intervals for small areas. This study uses data from the 2010 and 2014-15 Rwanda DHS surveys and the proposed methods to produce district-level total fertility rates and childhood mortality rates, which were not provided in the DHS survey reports due to insufficient sample sizes at the district level. The methods described here can be used to produce estimates at the district or subregional level for other surveys and other indicators.

Keywords: small area estimation, design-based, survey domain, total fertility rate, childhood mortality rates, DHS surveys

1 INTRODUCTION

Small area estimation (SAE) techniques have received increased attention from numerous requests for subregional-level data that can be used for policymaking and development planning. The Demographic and Health Survey (DHS) Program has also received increasing requests from host countries for subregional-level indicator estimates. Since 1984, The DHS Program, an international program funded by United States Agency for International Development (USAID), has collected, analyzed, and disseminated high-quality data on population, health, HIV, malaria, nutrition, and health care services through approximately 400 surveys in 90 countries. The DHS surveys are household-based, two-stage cluster surveys that provide key survey indicators for a country's first-level administrative unit, or region. Increasing sample size to produce direct subregional estimates, especially for the total fertility rate (TFR) and childhood mortality rates (CMR) which require large sample sizes, is not usually feasible because of potential concerns about data quality and cost. One solution is using SAE techniques to produce reliable estimates for key subregional indicators. There are many ways to produce SAE estimates by borrowing "strength" from other data sources such as census and administrative data. The SAE techniques typically use auxiliary information to first construct models (Kott 1989; Ghosh and Rao, 1994), which are used to either predict the small area characteristics, produce model-assisted estimates (Särndal et al. 1992; Tikkiwal et al. 2013) for small areas, or improve the precision of direct estimates of small areas simply through a sampling weight calibration procedure (Chambers and Chandra, 2008).

This research explores a design-based SAE methodology that uses data collected within a single target survey or from similar surveys conducted in recent years. We use data collected by the DHS surveys to test the proposed methods. The DHS surveys are household-based, two-stage cluster surveys conducted in a 5-year cycle in low- and middle-income countries. Key survey indicators are reported for the first-level administrative unit or region. In this research, we have attempted to produce reliable estimates of the TFR and CMR for the country's second-level administrative units or districts, which are usually not reported in the final survey report. The goal was to create a survey domain that covered the small area by pooling clusters close to the small area either geographically, in time and space, or by using other demographic, social, religious, cultural, and economic measures within one survey or from similar recent surveys. A survey domain created this way is easy to analyze with design-based domain analysis tools such as parameter estimation, variance estimation, and construction of confidence intervals. This approach assumes that individuals who live geographically close or who are close in other related measures may have similar demographic characteristics, even when they live in different regions or districts. The DHS surveys illustrate that most DHS key indicators change slowly in time, especially the TFR and CMR. Therefore, combining data from two or more similar surveys conducted by The DHS Program in the same country may enhance the power of analysis and the production of reliable small area estimations. Data from DHS surveys are typically more reliable and often more timely than data from other sources such as census or administrative records. The SAE estimates are usually not consistent in that they cannot be aggregated to align with survey estimates at a higher or regional level. An adjustment procedure, or more generally, a calibration procedure, can be applied to small area estimates to achieve desirable consistency.

2 DESIGN-BASED SMALL AREA ESTIMATION

In practice, most large-scale sample surveys have complex designs, including multistage and multiphase probability proportional to size (PPS) sampling procedures with stratification and clustering. Sampling weight, an expansion weight, is calculated as the inverse of the overall inclusion probability with possible adjustments for nonresponse and other calibration factors. Let S be a sample selected with a complex design, let $Y_i, i \in S$ be the sample observations of the variable of interest Y, let $w_i, i \in S$ be a set of expansion weight, and let $\hat{T}_w$ and $\hat{M}_w$ be the population total and mean estimates using the expansion weight:

$$\hat{T}_w = \sum_{i \in S} w_i Y_i, \quad \hat{M}_w = \sum_{i \in S} w_i Y_i \,/\, \sum_{i \in S} w_i \tag{1}$$

The variance and the variance estimation cannot be calculated without bias, but by approximations, for example, by Taylor Linearization (Woodruff 1971) approximation. The Taylor Linearization method is widely used in commercialized statistical software such as SAS, SPSS, and STATA. The Jackknife Repeated Replication Method (Efron & Tibshirani 1993) can also be used for variance estimation for more complex statistics such as the TFR and CMR.

Suppose that the total population can be subdivided into a large number of small areas or domains $U = \bigcup_1^A U_a$ with unknown small area totals $T_a = \sum_{i \in U_a} Y_i$, $1 \le a \le A$; and the sample S can be subdivided into corresponding small area subsamples $S = \bigcup_1^A S_a$ with small subsample size. The aim is to efficiently estimate the area total, or its mean based on S_a for each small area. The most intuitive estimate of T_a or M_a is the direct estimate based on the subsample S_a (non-missed small area):

$$\hat{T}_a = \sum_{i \in S_a} w_i Y_i \,, \quad \hat{M}_a = \sum_{i \in S_a} w_i Y_i \,/\, \sum_{i \in S_a} w_i \tag{2}$$

The variance and variance estimation of the direct estimates are straightforward with domain estimation tools.

Direct estimate based only on a small area sample is usually inefficient because of the small sample size. There are many ways to construct small area estimates by borrowing 'strength' based on spatial or structural properties of the small area, including design-based, model-assisted, and model-based methods. Design-based SAE techniques use auxiliary information from data outside of the survey to improve the reliability of the direct estimates, including the ratio estimator, regression estimator, or more generally, calibration estimators. In the following subsections, we present some basics of the design-based SAE techniques and our proposed nearest neighbor methods. The aim of this study is not to compare the different SAE methods, but to introduce a different approach and a different concept for SAE. Borrowing "strength" is not restricted to other sources outside of the target survey, and we can borrow "strength" from within the targeted survey.

2.1 Traditional Design-based SAE Techniques

The traditional design-based SAE techniques use auxiliary information available outside of the survey data to improve the reliability of the direct estimate. Suppose auxiliary information is available from reliable

sources with known area total for each small area. Let $T_{xa} = \sum_{i \in U_a} X_i$ be the known area total for an auxiliary variable X in each domain. $\hat{T}_a$ and $\hat{T}_{xa}$ are the direct estimates based on the area samples:

$$\hat{T}_a = \sum_{i \in S_a} wY_i \,, \quad \hat{T}_{xa} = \sum_{i \in S_a} w_i X_i \tag{3}$$

Usually $\hat{T}_{xa} \neq T_{xa}$; if we can calibrate the sampling weights $w_i, i \in S_a$ to $w_i^c, i \in S_a$ such that the direct estimate of the known area total can be determined without error using the calibrated weights:

$$\hat{T}_{xa} = \sum_{i \in S_a} w_i^c X_i = T_{xa} \tag{4}$$

then we have good reason to believe that the area total estimate of the variable of interest using the calibrated weight

$$\hat{T}_a = \sum_{i \in S_a} w_i^c Y_i \tag{5}$$

should be a better estimate for the small area total T_a if X and Y are well correlated. When the auxiliary variable X defines a classification of the total population, i.e.,

$$X_{ij} = 1 \text{ if Unit i} \in C_j, i = 1,2, \dots, j = 1,2, \dots$$

then a straightforward calibrated estimator is the post-stratified estimator:

$$\hat{T}_a^P = \sum_j N_{aj} \hat{\bar{Y}}_{aj} \text{ or } \hat{T}_a^P = \sum_j \sum_{i \in S_a \cap C_j} w_i^c Y_i \tag{6}$$

where $\hat{\bar{Y}}_{aj} = \frac{\sum_{i \in S_a \cap C_j} w_i Y_i}{\sum_{i \in S_a \cap C_j} w_i}$ is the area mean estimation with post-stratification calibrated weights $w_i^c = \frac{N_{aj} w_i}{\sum_{i \in S_a \cap C_j} w_i}$. N_{aj} is the total number of units in small area a which is in class j. This estimator requires that all the classes are present in the small area sample. This can be a problem when the number of classes or the number of small areas is large, and the total sample size is small.

When auxiliary variable X is a continuous variable, a general regression estimator, which is a model-assisted estimator, is given by:

$$\hat{T}_a^{Reg} = \sum_{i \in S_a} w_i^c \, Y_i \tag{7}$$

where $w_i^c, i \in S_a$ is a set of regression weights:

$$w_i^c = w_i \left\{ 1 + \left[T_{S_a}^{-1} X_i / v^2(X_i) \right]^t \left(T_{xa} - \hat{T}_{xa} \right) \right\} \tag{8}$$

with $T_{S_a} = \sum_{i \in S_a} w_i \, X_i X_i^t / v^2(X_i)$, $v^2(\mathrm{x})$ is the variance function of the regression model. Another way to express the general repression estimator is:

$$\hat{T}_a^{Reg} = \sum_{i \in S_a} w_i \, Y_i + \hat{\beta}_a \left(T_{xa} - \hat{T}_{xa} \right) \tag{9}$$

with $\hat{\beta}_a$ is the estimated regression coefficient $\hat{\beta}_a = \sum_{i \in S_a} w_i Y_i \left[T_{S_a}^{-1} X_i / v^2(X_i)\right]^t$ over the small area. This estimator satisfies $\sum_{i \in S_a} w_i^c X_i = T_{xa}$.

When the auxiliary variable X is a continuous single variable, the regression estimator becomes a ratio estimator if the model variance function is a linear function $v^2(x) = x$, the regression coefficient $\hat{\beta}_a$ is becoming a simple ratio $\hat{\beta}_a = \frac{\sum_{i \in S_a} w_i Y_i}{\sum_{i \in S_a} w_i X_i} = \hat{R}_a$, and the estimator is a ratio estimator.

$$\hat{T}_a^R = \frac{\sum_{i \in S_a} w_i Y_i}{\sum_{i \in S_a} w_i X_i} T_{ax} = \hat{R}_a T_{ax} \tag{10}$$

When the sample size in the small area is small, the estimator $\hat{\beta}_a$ may not be stable. One option is to use $\hat{\beta}_S$, with the regression coefficient based on the full sample, in the place of $\hat{\beta}_a$:

$$\hat{T}_{asyn}^{Reg} = \sum_{i \in S_a} w_i Y_i + \hat{\beta}_S \left(T_{xa} - \hat{T}_{xa}\right) \tag{11}$$

This is a synthetic estimator that assumes the same regression model to be valid across the small areas.

The advantage of the design-based SAE is that variance estimation, and hence the confidence interval, are straightforward using the linearization variance estimation method. The disadvantage of design-based SAE is that it does not extend to missed areas—that is, areas for which there are no sample observations. Some of the methods proposed in this paper, however, will work even for missed areas.

2.2 Proposed Nearest Neighbor SAE Method

In this section, we describe an SAE method that uses the nearest neighbor technique. Sampling units located geographically close or close to other related measures correlated with the variable of interest may have similar characteristics to the study variables. The survey may also have collected other information that can be used as distance measures, such as GPS coordinates of the sample points, demographic, social, cultural and economic measures, and time and space data from previous surveys of the same kind in the near past. We pool the sampled sampling units "close" to the small area together with the sampled sampling units from the small area to form a group, a nearest neighborhood, and then treat it as a survey domain. A domain created in this way could be a true survey domain or a pseudo-domain, depending on the definition of the distance measure. If the distance measure defines a fixed subpopulation that does not depend on any sample selection results, then the domain is a true survey domain. For example, all sampling units located within a fixed distance to a fixed geographical point within a small area form a true survey domain. All sampling units located within a fixed distance to one or a group of sampled sampling units can form a pseudo-domain because it depends on the random selection results, which define a random subpopulation. This is acceptable because we are not seeking to estimate the population characteristics of the domain. Instead, we are inferring the population characteristics of the small area. By treating them as a survey domain, we can use all the known statistical inference techniques of survey domain analysis to estimate the population characteristics of the small area, such as small area total, its variance, and variance estimation. Let S_a^+ be the enlarged sample that includes the small area sample, plus the borrowed sampling units from nearest neighbor areas. The small area total can be estimated by

$$\hat{T}_a^* = N_a \times \frac{\sum_{i \in S_a^+} w_i Y_i}{\sum_{i \in S_a^+} w_i} \quad \text{or} \quad \hat{T}_a^* = \hat{N}_a \times \frac{\sum_{i \in S_a^+} w_i Y_i}{\sum_{i \in S_a^+} w_i} \tag{12}$$

depending on whether the small area population size N_a is known or unknown, where $\hat{N}_a = \sum_{i \in S_a} w_i$ is the estimate of the area population size when it is unknown, where the weights $w_i, i \in S_a^+$ is a set of expansion weights associated to the sampling units. This is a ratio or ratio type estimate that uses the domain analysis tools and the linearization method. Variance and confidence interval estimations are straightforward. If the sample size is very small in the small area, $\hat{N}_a$ may have low reliability, but the small area mean estimation $\hat{M}_a^* = \frac{\sum_{i \in S_a^+} w_i Y_i}{\sum_{i \in S_a^+} w_i}$ can be reliable if the nearest neighbor areas have adequate sample size.

When similar surveys had been conducted in the same area in recent years, and the characteristics to be estimated are relatively stable over time, we can then combine two surveys to increase the sample size for small areas, which uses the time-space nearest neighbor. For example, The DHS Program's Senegal Continuous Survey always combines data collected in 2 consecutive years to produce regional-level TFR and CMR estimates. Let $S_a^{(1)}$ and $S_a^{(2)}$ be the small area samples from the previous survey and current (target) survey, respectively, and $\hat{T}_a^{(1)}$and $\hat{T}_a^{(2)}$ be the corresponding direct estimates of the area total over small area a. Then

$$\hat{T}_a^* = \alpha \hat{T}_a^{(1)} + (1-\alpha)\hat{T}_a^{(2)} = \alpha \sum_{i \in S_a^{(1)}} w_i Y_i + (1-\alpha) \sum_{i \in S_a^{(2)}} w_i Y_i \tag{13}$$

is an estimate of the current area total with a proper weighting factor α $(0 < \alpha < 1)$. The two direct estimates can be weighted equally or weighted with an importance weight. If weighted equally, the estimate represents the small area total at a time point between the two surveys. The estimates can also be weighted by using their variance to achieve minimum variance for the combined estimate. When assuming that the two surveys are independent, all analysis is simple and directly based on standard survey data analysis tools and techniques. To simplify notations in the formula, the values of Y_i in the two different terms represent the sample values of the variable of interest observed at different occasions. To make the combined estimate close to the target survey or to the previous survey, we can use an adjustment procedure, or a calibration procedure, to calibrate the sampling weights together with the weighting factor to achieve certain known constraints. For example, assuming an auxiliary variable X with sample observations in the previous survey and current survey, and known area totals $T_{xa}^{(1)}$ and $T_{xa}^{(2)}$, then the following estimator will be a ratio estimator for the area total for the current survey:

$$\hat{T}_a^* = T_{xa}^{(2)} \frac{\hat{T}_a^*}{\hat{T}_{xa}^*} \tag{14}$$

where $\hat{T}_{xa}^*$ is the sample estimate (13) with the Y_i replaced by X_i. Replacing $T_{xa}^{(2)}$ by $T_{xa}^{(1)}$, the area total of the auxiliary variable at the time of the previous survey, (14) will be a ratio estimator of the area total for the previous survey. The variance estimation and confidence interval for estimator (14) are straightforward.

It is desirable that small area estimates produced by different methods be consistent with reliable higher-level estimates. For example, the SAE at the district level should be consistent with regional-level estimations. That is, the SAE produced at district level should be able to aggregate to regional-level estimates

and match the regional-level estimates, which are considered reliable because of larger sample sizes. There are different ways to adjust the SAE for consistency. Let B be a broad area containing a number of small areas, with the simplest adjustment as a ratio-type adjustment:

$$\hat{T}_a^{\Delta} = \frac{\hat{T}_B}{\sum_{a \in B} \delta_a \hat{T}_a^*} \times \hat{T}_a^* \tag{15}$$

where $\hat{T}_B$ is the broad area or higher-level estimate based on the full sample S_B from the broad area B, and δ_a is the relative size of small area a within the broad area B:

$$\hat{T}_B = \sum_{i \in S_B} w_i \, Y_i \, , \qquad \delta_a = \frac{\sum_{i \in S_a} w_i}{\sum_{i \in S_B} w_i} \tag{16}$$

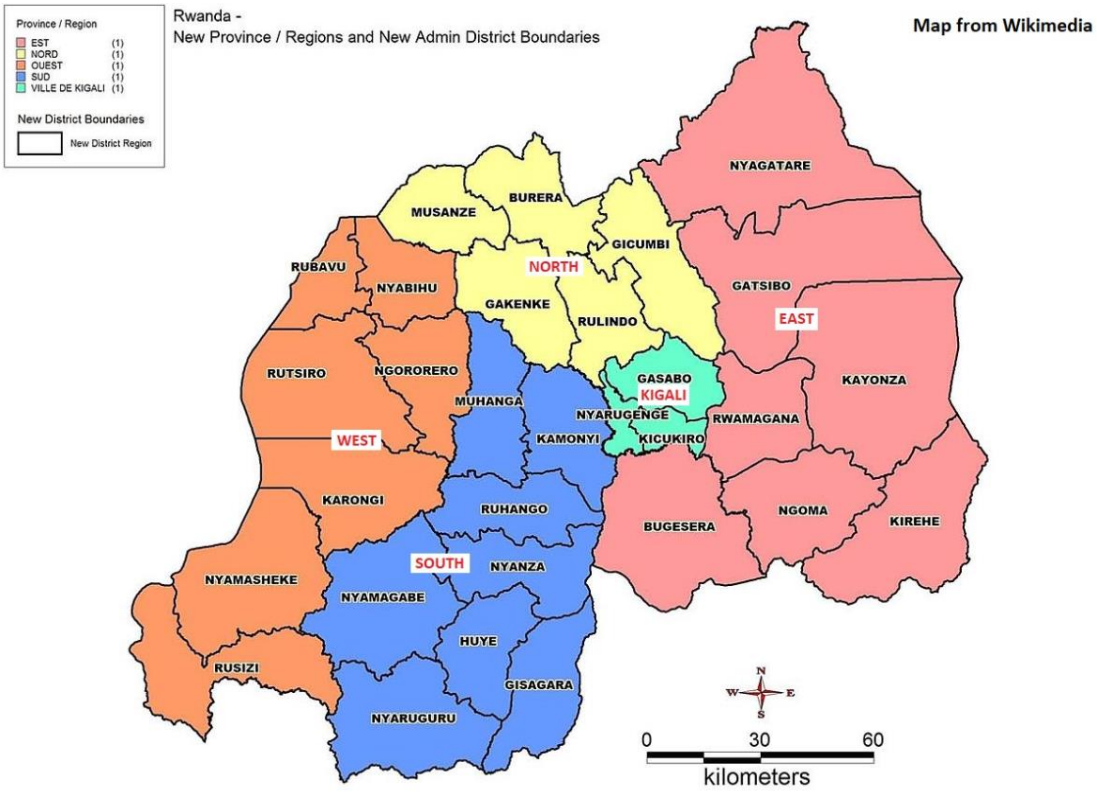
$\hat{T}_a^{\Delta}$ is consistent in that it can be aggregated to the broad area estimate $\hat{T}_B = \sum_{a \in B} \delta_a \hat{T}_a^{\Delta}$. This adjustment is simply a parallel transformation. A more general adjustment can use a "reverse calibration" procedure by treating $\hat{T}_B$ as the target total and the $\hat{T}_a^*$s as "weights," especially when the number of small areas from the broad area is large. The variance of the consistency-adjusted estimator can be estimated by the Jackknife method.

The proposed methods in this study are different from the Broad Area Ratio Estimator (BARE) (Asian Development Bank 2020), which pools all neighboring small areas together from a broad area, and where a homogeneous assumption is made that all small areas have the same mean as the broad area. It is also different from the reweighting method of Schirm and Zalansky et al. (1997), which uses the full sample including the small area with adjusted sampling weights, and where weights of the full sample are adjusted to catch the small area population size or other known population characteristics of the small area.

3 DATA USED IN THIS STUDY

In this study, we use the Rwanda DHS 2014-15 as the target survey and Rwanda DHS 2010 as the auxiliary survey. Rwanda has 5 provinces, each of which is subdivided into a number of districts, for a total number of 30 districts. The smallest province is Kigali City Province, which has only 3 districts. The largest province is the East Province with 7 districts. The second largest province is the South Province, with 8 districts. The districts in Rwanda are quite homogeneous in population size. Figure 1 is a map of Rwanda with provinces and districts delineated. The Rwanda Demographic and Health Survey 2014-15 was the fifth DHS in Rwanda that followed surveys in 1992, 2000, 2005, and 2010. Rwanda's administrative units had been reformed in 2006, and this reduced the number of provinces from 11 to 5. According to the reformed administrative units, Rwanda is divided into provinces; each province is subdivided into districts; each district into sectors, each sector into cells, and each cell into villages. Rwanda DHS 2010 is the first DHS that used the new province and district specifications, followed by the DHS 2014-15.

Figure 1 Map of Rwanda's provinces and districts



Rwanda DHS 2014-15 and DHS 2010 have exactly the same design; both are household-based, two-stage cluster surveys, with a designed sample size of 492 clusters, 12,792 households, and 26 households per cluster. The sample allocation adopted was an equal size allocation with 16 clusters and 416 households per district, except for the 3 districts in Kigali City Province where 20 clusters and 520 households per district were allocated. Table 1 presents the detailed sample allocation of the number of clusters and households, and number of women age 15-49 interviewed. Women of reproductive age 15-49 and children under age 5 are the two main targeted populations of the DHS surveys. As with all DHS surveys, the surveys collected

data on basic demographic, reproductive health, women and children's basic health, and family planning, as well as a full birth history for all interviewed women, which is the main data source for the TFR and CMR. The number of women age 15-49 interviewed per district varies from 390 to 665 in the 2010 survey, and 375 to 653 in the 2014-15 survey. The district-level sample size is adequate for many indicators, but is too small for a direct estimate of the TFR and CMR. What we call "small area" here is relative to the specific variables where a reliable estimate requires a much larger sample size. The TFR and CMR estimates require at least double the sample size to produce reliable estimations at the district level. The DHS surveys control the 95% confidence interval for the TFR at the survey domain level to be shorter than one child; and control the CMR estimation precision at the survey domain level with a coefficient of variation less than 20%. The DHS experiences show that we need to interview at least 800 to 1,000 women age 15-49 per survey domain to produce reliable estimations of the TFR and CMR in high-fertility-level countries such as most African countries. In the Rwandan settings, we must double the sample size at the district level to meet the minimum sample size needed for the TFR and CMR estimations.

Table 1 Sample allocation of clusters and households and number of women age 15-49 interviewed for Rwanda DHS 2010 and DHS 2014-15

		Rwanda DHS 2010			Rwanda DHS 2014-15		
Province	District	Number of clusters selected	Number of households selected	Number of women interviewed	Number of clusters selected	Number of households selected	Number of women interviewed
Kigali City	Nyarugenge	20	520	617	20	520	637
Kigali City	Gasabo	20	520	608	20	520	586
Kigali City	Kicukiro	20	520	665	20	520	653
South	Nyanza	16	416	390	16	416	385
South	Gisagara	16	416	428	16	416	427
South	Nyaruguru	16	416	433	16	416	424
South	Huye	16	416	424	16	416	439
South	Nyamagabe	16	416	423	16	416	453
South	Ruhango	16	416	420	16	416	403
South	Muhanga	16	416	395	16	416	447
South	Kamonyi	16	416	427	16	416	457
West	Karongi	16	416	417	16	416	428
West	Rutsiro	16	416	451	16	416	411
West	Rubavu	16	416	442	16	416	434
West	Nyabihu	16	416	455	16	416	418
West	Ngororero	16	416	460	16	416	426
West	Rusizi	16	416	442	16	416	512
West	Nyamasheke	16	416	471	16	416	431
North	Rulindo	16	416	466	16	416	414
North	Gakenke	16	416	429	16	416	427
North	Musanze	16	416	464	16	416	450
North	Burera	16	416	413	16	416	450
North	Gicumbi	16	416	427	16	416	429
East	Rwamagana	16	416	456	16	416	454
East	Nyagatare	16	416	442	16	416	405
East	Gatsibo	16	416	467	16	416	435
East	Kayonza	16	416	445	16	416	433
East	Kirehe	16	416	426	16	416	375
East	Ngoma	16	416	398	16	416	457
East	Bugesera	16	416	470	16	416	397
	Rwanda	**492**	**12,792**	**13,671**	**492**	**12,792**	**13,497**

Table 2 presents the district profile for selected indicators based on data from the target survey Rwanda DHS 2014-15. These indicators are closely correlated with the TFR and CMR. The district profile is used for the construction of nearest neighborhood in one of the proposed methods and can be referenced when interpreting the SAE estimates of the TFR and CMR. We calculated the percentage of interviewed women living in urban areas (urban), living in and below the second wealth quintile (poor), living in and above the fourth quintile (rich), the percentage of literacy, having no education, having secondary or higher education (secondary-higher), the percentage of never married, currently married (married), currently pregnant (pregnant), currently use a modern contraceptive method (contraceptive use), and the number of births given in last 3 years (birth3), number of children ever born, and number of living children. For these indicators, the sample size at district level is large enough to produce reliable estimations based only on the survey data from the district. The average coefficient of variation for all indicators in this table is around 10%, implying good precision for domain-level estimates.

Table 2 District profile for select variables, Rwanda DHS 2014-15

District	Urban	Poor	Rich	Literacy	No education	Secondary-higher	Never married	Married	Pregnant	Contraceptive use	Birth3	Children ever born	Living children
Nyarugenge	0.78	0.08	0.88	0.90	0.04	0.41	0.41	0.47	0.08	0.51	0.35	1.92	1.73
Gasabo	0.73	0.11	0.83	0.92	0.04	0.40	0.41	0.51	0.08	0.51	0.36	1.81	1.68
Kicukiro	0.90	0.04	0.94	0.94	0.04	0.52	0.50	0.39	0.04	0.46	0.26	1.49	1.40
Nyanza	0.09	0.55	0.29	0.77	0.14	0.14	0.30	0.54	0.08	0.43	0.39	2.40	2.11
Gisagara	0.02	0.72	0.15	0.69	0.16	0.10	0.35	0.51	0.07	0.50	0.38	2.45	2.07
Nyaruguru	0.02	0.52	0.20	0.73	0.18	0.22	0.37	0.55	0.10	0.34	0.38	2.61	2.27
Huye	0.15	0.39	0.45	0.83	0.08	0.29	0.43	0.46	0.06	0.47	0.34	1.95	1.73
Nyamagabe	0.08	0.47	0.29	0.76	0.15	0.20	0.45	0.46	0.05	0.56	0.29	2.15	1.93
Ruhango	0.12	0.46	0.34	0.85	0.10	0.21	0.38	0.50	0.06	0.49	0.34	2.24	1.96
Muhanga	0.16	0.25	0.57	0.87	0.08	0.26	0.40	0.51	0.06	0.53	0.31	2.04	1.87
Kamonyi	0.15	0.29	0.47	0.86	0.06	0.21	0.41	0.49	0.08	0.51	0.33	2.09	1.86
Karongi	0.07	0.37	0.38	0.82	0.09	0.30	0.44	0.48	0.09	0.40	0.33	2.02	1.81
Rutsiro	0.03	0.59	0.18	0.73	0.19	0.14	0.33	0.59	0.09	0.42	0.44	2.48	2.22
Rubavu	0.41	0.45	0.43	0.76	0.18	0.25	0.37	0.52	0.06	0.44	0.45	2.46	2.16
Nyabihu	0.13	0.74	0.11	0.76	0.18	0.17	0.36	0.53	0.07	0.47	0.35	2.40	2.07
Ngororero	0.05	0.47	0.32	0.75	0.22	0.18	0.35	0.55	0.07	0.45	0.38	2.39	2.04
Rusizi	0.16	0.42	0.36	0.80	0.12	0.25	0.45	0.47	0.06	0.37	0.37	2.36	2.10
Nyamasheke	0.02	0.51	0.27	0.81	0.09	0.18	0.38	0.54	0.10	0.34	0.47	2.29	2.13
Rulindo	0.03	0.42	0.38	0.83	0.09	0.21	0.38	0.53	0.09	0.51	0.37	2.11	1.85
Gakenke	0.07	0.32	0.41	0.79	0.08	0.21	0.41	0.52	0.05	0.58	0.27	2.08	1.80
Musanze	0.30	0.38	0.43	0.83	0.12	0.27	0.40	0.49	0.05	0.67	0.30	2.13	1.87
Burera	0.02	0.54	0.26	0.74	0.15	0.16	0.39	0.52	0.07	0.44	0.33	2.31	2.10
Gicumbi	0.07	0.38	0.35	0.80	0.12	0.22	0.40	0.51	0.05	0.54	0.32	2.44	2.16
Rwamagana	0.09	0.29	0.50	0.84	0.10	0.23	0.36	0.51	0.07	0.47	0.38	2.47	2.09
Nyagatare	0.09	0.33	0.42	0.72	0.21	0.18	0.31	0.58	0.10	0.48	0.44	2.68	2.23
Gatsibo	0.06	0.42	0.30	0.74	0.17	0.16	0.31	0.55	0.09	0.45	0.42	2.76	2.26
Kayonza	0.09	0.32	0.39	0.79	0.14	0.19	0.35	0.54	0.11	0.47	0.39	2.47	2.11
Kirehe	0.04	0.42	0.37	0.73	0.17	0.13	0.27	0.62	0.06	0.50	0.42	2.69	2.23
Ngoma	0.04	0.45	0.34	0.73	0.13	0.19	0.32	0.56	0.07	0.47	0.41	2.48	2.10
Bugesera	0.10	0.37	0.38	0.83	0.15	0.18	0.28	0.59	0.08	0.41	0.45	2.56	2.23
Rwanda	**0.19**	**0.38**	**0.42**	**0.80**	**0.12**	**0.23**	**0.38**	**0.52**	**0.07**	**0.47**	**0.37**	**2.28**	**1.99**

It is worth noting that the sampling weights in the DHS data are normalized weights that are relative weights, and the normalizations are survey specific. The normalization factor is a constant which is the estimated sampling fraction at the national level. This requires that pooling data together from different surveys, the weights must be adjusted or denormalized after pooling. In this study, for the time-space nearest neighbor method in which we combine data from the 2010 DHS and 2014-15 DHS, the sampling weights were denormalized by dividing the weights by the estimated sampling fraction for each survey, although the sampling fractions of the two surveys were very close.

4 METHODS AND RESULTS

In this section, we present the results of the study using the Rwanda 2010 and 2014-15 DHS data. Our target indicators are the TFR and CMR at district level, which were not reported in the surveys' final report because of insufficient sample size. We explored five methods of creating the nearest neighborhood of the small area.

1. The first method uses a time-space nearest neighbor by simply combining the two surveys 2010 and 2014-15, so that the sample size is doubled for each district.
2. The second method uses the geographical nearest neighbor method based on just the target survey DHS 2014-15. This involved selecting "donor" clusters from the neighboring districts within the same province or from other provinces based on a geographical distance measure calculated with the GPS information collected at each cluster center. A list of donor clusters is identified based on their distance to the targeted district center. We call this method "district center nearest neighbor".
3. The third method is similar to the second. For each cluster from the targeted district, a list of donor clusters from neighboring districts is identified based on their distances from the target cluster center. We call this method "cluster center nearest neighbor".
4. The fourth method is also similar to the second method. It uses a more complex measure of the distance from the district center to the nearest neighbor by creating a profile for each district and cluster using GPS coordinates, women's demographic characteristics, and the wealth quintile. The wealth quintile reflects the living conditions and economic status of the households, and is strongly correlated with the TFR and CMR. We call this method "district center nearest neighbor with a composite distance measure".
5. The fifth method is a hybrid that uses data on nearest neighborhoods constructed with the other four methods to form hybrid nearest neighborhoods.

The GPS coordinates collected by the DHS surveys are subject to a random displacement for data confidentiality concerns. The scale of the displacement is usually small and does not cross the boundaries of the country's second-level administrative units. In Rwanda, the displacement is within districts. Therefore, the displacement may only have a small impact on the distance measures.

4.1 Time-space Nearest Neighbor

This method uses time-space nearest neighbor by simply combining data from the 2010 and 2014-15 surveys, since all sampled clusters in the 2010 survey in a district are the nearest neighbors geographically and in time for the clusters in the same district for the 2014-15 survey. This doubles the sample size for each of the 30 districts and meets the minimum sample size requirement for the TFR and CMR estimations at the domain level. In the pooled data, all districts have 32 clusters and 984 households, except the three districts in Kigali City Province where each has 40 clusters and 1,040 households. The TFR (for the 3 years before the survey) and CMR (for the 10 years before the survey) are calculated with the standard procedure based on the combined sample as if they were from a single survey. However, an importance weight can be used to reflect the user's subjective judgment as indicated in equation (13). For example, a larger weight can be assigned to the target survey and a smaller weight to the auxiliary survey. The weighting factor can be area/district-specific. In this study, we tested equal weights, province-level TFR and CMR variance weights,

and an importance weight. The results produced by different weights were very close. The results reported here used the equal weight option. The calculated TFR and CMR without adjustment represent a reference period between the two surveys, roughly from 2010 to 2012 for the TFR, and from 2003 to 2012 for the CMR. A consistency adjustment with the 2014-15 provincial-level TFR and CMR estimates made the estimates lean toward the 2014-15 survey. A consistency adjustment can also be made for the age-specific fertility rates or the TFR. The results reported here are adjusted with the provincial-level TFR. We report here only the results for infant mortality rate (IMR), which is one of the five CMRs.

4.2 District Center Nearest Neighbor

This method uses the small area center point as a reference point, calculates the distance of the other clusters from other areas, and takes a number of clusters closest to the small area central point as a nearest neighborhood, which creates a true survey domain. The small area central point is usually easy to obtain information. When the sample size from the small area is not too small, such as in the Rwanda DHS 2014-15, a district central point, calculated based on the GPS coordinates of the sample clusters, should be very close to the district central point, which is the central point of habited areas that is better and more meaningful than the actual geographical center. Suppose a group of such clusters are identified, noted as S_a^D, and the population characteristic estimation has the same formula as given in equation (12). In this study, we calculated the district center based on the sample points, and then calculated the distance to the district center for each of the clusters that are not from the target district, and took the first 20 clusters closest to the targeted district center. Table 3 shows the construction of the neighborhoods for each of the 30 districts, with the number of donor clusters from other districts within the same province, and number of donor clusters from districts in other provinces. Table 3 has four panels, with the first and second panels showing the province name and district name, the third and fourth panels the number of donor clusters by district codes 1, 2, …, 8 within province, and then by province codes 1,2, …, 5 from other provinces. The table shows that some districts did not borrow any clusters from other provinces, but some districts borrowed many clusters from other provinces. The largest number is 17 clusters borrowed by district Bugesera in the East Province.

Table 3 Construction of nearest neighborhood with distance to district center

		From other districts within province								From other provinces				
Province	**District**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**
Kigali City	Nyarugenge		7	11							2			
Kigali City	Gasabo	8		12										
Kigali City	Kicukiro	13	7											
South	Nyanza		3		6		11							
South	Gisagara	2		2	16									
South	Nyaruguru		4		12	4								
South	Huye	4	9	4		3								
South	Nyamagabe	4		3	7		1					5		
South	Ruhango	11						6	3					
South	Muhanga						3		8			7	2	
South	Kamonyi						3	2		14			1	
West	Karongi		6			3		2			9			
West	Rutsiro	4		7		9								

(continued…)

Table 3—*Continued*

		From other districts within province								From other provinces				
Province	**District**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**
West	Rubavu		8		11								1	
West	Nyabihu			6		4							10	
West	Ngororero		7		5						8			
West	Rusizi							16			4			
West	Nyamasheke	4					13				3			
North	Rulindo		5			8				7				
North	Gakenke	6		2	3						4	5		
North	Musanze		2		7							11		
North	Burera	2	6	10		2								
North	Gicumbi	11			2									7
East	Rwamagana			2	7		4			7				
East	Nyagatare			10									10	
East	Gatsibo	3	5		2					2			8	
East	Kayonza	10		2		1	7							
East	Kirehe	1			6		13							
East	Ngoma	7			5	7		1						
East	Bugesera	1					2			14	3			

The district center nearest neighbor method also works for missed areas if the area center is known. In this case, the SAE for a missed area will be based only on donor clusters from neighboring areas.

4.3 Cluster Center Nearest Neighbor

This method uses a cluster-level nearest neighborhood. Distances to a target cluster for each of the donor clusters from neighboring districts from the same province or from a neighboring province are calculated. A list of donor clusters closest to a target cluster is identified for each target cluster from the target district. Since one donor cluster can be the nearest neighbor for several target clusters, we kept the distinct donor clusters. To control the neighborhood size, 5 closest donor clusters were identified for each target cluster with a distance cutoff. The distance cutoff is district-specific. The same distance cutoff was used for each of the target clusters within same district, which varies from 4 km to 25 km by district, with a target of 16 distinct donor clusters identified for each district to reach the smallest sample size required for reliable TFR and CMR estimation. Table 4 provides the construction of the neighborhood for each of the 30 districts, with detailed distribution of donor clusters. The number of donor clusters ranges from 8 to 19 per district. Some districts borrowed clusters only from the neighboring districts within the same province, while some districts borrowed clusters from districts in other provinces. The largest number of clusters borrowed from other provinces is 14, borrowed by district Bugesera in the East Province.

Table 4 Construction of nearest neighborhood with distance to each cluster

		From other districts within same province								From other provinces				
Province	**District**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**
Kigali City	Nyarugenge		5	6							1		1	
Kigali City	Gasabo	7		7										
Kigali City	Kicukiro	7	8											1
South	Nyanza		4		2		9							1
South	Gisagara	3		5	10									
South	Nyaruguru		3		7	4								
South	Huye	6	8	3		2								
South	Nyamagabe	3		2	6		1					5		
South	Ruhango	8						2	6			2		1
South	Muhanga						3		5			7	4	
South	Kamonyi						3	3		8			3	1
West	Karongi		5			1		2			9			
West	Rutsiro	5		5		8								
West	Rubavu		5		7								1	
West	Nyabihu			5		3							7	
West	Ngororero	2	5		6						4			
West	Rusizi							8						
West	Nyamasheke	5					10				3			
North	Rulindo		1		2	6				5	3			
North	Gakenke	4		4	3						4	2		
North	Musanze		3		6							6		
North	Burera	2	4	7		3								
North	Gicumbi	8			2					4				4
East	Rwamagana			2	8		3			5				
East	Nyagatare			8									5	
East	Gatsibo	4	4		2								4	
East	Kayonza	6		3		2	6							
East	Kirehe				4		10							
East	Ngoma	4			5	5		3						
East	Bugesera	1					1			9	5			

4.4 Nearest Neighbor with Other Distance Measures

In Sections 4.2 and 4.3, we used geographical distance to create the nearest neighborhood. In this section, we use a composite distance measure by creating a profile for each of the 30 districts and each of the 495 clusters. As presented in Table 2 for the district, the profile considers the wealth index, women's individual education level, marital status (ever married or never married), current pregnancy status, use of modern contraception, number of births in the past 3 years, number of children ever born, and number of children that survive. The profile uses the mean of each variable. The distance of a cluster to a district for a specific variable is the absolute value of the difference between the district and the cluster means. A composite distance measure that measures the distance of a cluster to a district is the weighted sum of distance on each variable plus the geographical distance

$$d = \sum \alpha_i d_i \tag{17}$$

Because of the scale difference and the correlations of the variables with the TFR and the CMR, the α-coefficients play an important role in the composite distance measure. It is difficult to determine what is the best coefficient for each variable. The determination of the coefficients in equation (17) can be based on the user's personal judgment, or through complex analysis such as factor analysis or logistic regression. The distance function can be uniform for all districts, by province, or even by district. After some simulations, we found that a much simpler distance measure works well, which includes only the geographical distance based on GPS (*gps*), the wealth index (*widx*), women's individual education level (*edu*), number of children ever born (*ceb*), and number of living children (*clv*):

$$d = \alpha_1 d_{gps} + \alpha_2 d_{widx} + \alpha_3 d_{edu} + \alpha_4 d_{ceb} + \alpha_5 d_{clv} \quad (18)$$

In this study, the α-coefficients were determined based on the author's personal judgement and simulations. We used a uniform distance measure for all districts, but separately for the TFR and CMR because of their different nature. Table 5 presents the details of the construction of the nearest neighborhoods for TFR estimation, with the α-coefficients uniformly set to (0.1, 2, 5, 5, 0) for all districts. Table 6 presents the details of the construction of the nearest neighborhoods for CMR estimation, with the α-coefficients uniformly set to (0.1, 5, 2, 2, 5) for all districts. The construction of the nearest neighborhoods for the TFR and CMR are different, although they are quite similar based on the visual patterns shown in the two tables.

Table 5 District-center-based nearest neighborhood with a fixed number of 20 clusters borrowed using the composite distance measure for the TFR

		From other districts within same province								From other provinces				
Province	**District**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**
Kigali City	Nyarugenge		7	8							2		2	1
Kigali City	Gasabo	10		7							1		2	
Kigali City	Kicukiro	9		8							1		1	1
South	Nyanza		3		4	1	9	1	1					1
South	Gisagara	6		4	7	2	1							
South	Nyaruguru	4	4		7	5								
South	Huye	1	3	4		6	3	1				2		
South	Nyamagabe	5	1	4	1		4					5		
South	Ruhango	9	1		1			4	5					
South	Muhanga						3		7	3		5	2	
South	Kamonyi						4	7		4		1	4	
West	Karongi		2			2		4			12			
West	Rutsiro	6		3	2	4					3		2	
West	Rubavu	2	6			4	1				1		6	
West	Nyabihu		3	6		3							8	
West	Ngororero	3	3	1	1						6		6	
West	Rusizi	1						14			5			
West	Nyamasheke	5					11				4			
North	Rulindo		4		4	4				3	5			
North	Gakenke	6		2	5						6	1		
North	Musanze	1	9		7							3		
North	Burera	4	6		7		3							
North	Gicumbi	7	3		4					1				5

(continued…)

Table 5—*Continued*

		From other districts within same province								From other provinces				
Province	**District**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**
East	Rwamagana			3	5		2	4		4			2	
East	Nyagatare			7									13	
East	Gatsibo	3	3		4					2			8	
East	Kayonza	7		4		4	3	1		1				
East	Kirehe	3			6		8	3						
East	Ngoma	4			6	5		4		1				
East	Bugesera	4					4			4	8			

Table 6 District-center-based nearest neighborhood with a fixed number of 20 clusters borrowed using composite distance measure for the CMR

		From other districts within same province								From other provinces				
Province	**District**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**1**	**2**	**3**	**4**	**5**
Kigali City	Nyarugenge		7	8							2		2	1
Kigali City	Gasabo	9		9									2	
Kigali City	Kicukiro	9	9								1			1
South	Nyanza		3		4		9	1	2					1
South	Gisagara	7		4	7		2							
South	Nyaruguru	2	4		8	6								
South	Huye	2	3	3		7	4					1		
South	Nyamagabe	5		3	4		4					4		
South	Ruhango	8						4	7			1		
South	Muhanga						4		7	3		4	2	
South	Kamonyi						2	9		5			3	1
West	Karongi		2			2		3			13			
West	Rutsiro	7		3	2	6					2			
West	Rubavu	1	7		4	1					1		6	
West	Nyabihu		2	6		3							9	
West	Ngororero	3	3	1	1						7		5	
West	Rusizi	2						13			5			
West	Nyamasheke	6					11				3			
North	Rulindo		5		2	5				4	4			
North	Gakenke	5		2	6						7			
North	Musanze	1	8		7							4		
North	Burera	3	6	8		3								
North	Gicumbi	8	4		3					1				4
East	Rwamagana			3	7	2	4			2			2	
East	Nyagatare			10									10	
East	Gatsibo	3	2		5					2			8	
East	Kayonza	8		4		5	3							
East	Kirehe	1			8		9	2						
East	Ngoma	4				7	4	4		1				
East	Bugesera	4				1	1			6	8			

4.5 Hybrid SAE Method

The method used for creating the nearest neighborhoods does not need to be the same for all districts, or for all districts in the same province. The method for creating the nearest neighborhood for each district can be district-specific, as described in the previous sections. This can be a time-consuming task. Instead of creating an independent nearest neighborhood with a different method for each district, we can select a specific nearest neighborhood created in the previous sections, for a specific district, to demonstrate a hybrid SAE estimation. The selection depends on the author's personal judgment of best fit based on knowledge about the district. For simplicity, in this study, we selected nearest neighborhood at the provincial level. This means that all the districts in the same province have the nearest neighborhood created by the same method. For the results presented in this paper for the TFR, we used the nearest neighborhoods created in Section 4.2, the district center nearest neighbor, for the districts in the first two provinces, and we used the nearest neighborhoods created in Section 4.4, nearest neighbor with other distance measures, for the last three provinces. For the CMR, we used the nearest neighborhoods created in Section 4.4, nearest neighbor with other distance measures, for the district in the province of Kigali City; the nearest neighborhoods created in Section 4.2, the district center nearest neighbor, for the districts in the South and North provinces; the nearest neighborhoods created in Section 4.1, the time-space nearest neighbor, for the districts in the West province; and the nearest neighborhood created in Section 4.3, the cluster center nearest neighbor, for the districts in the East Province. The selection of the nearest neighborhoods was not guided by any numerical measures but was based on the author's personal judgments and evaluations.

4.6 Results

In this section, we present the results using the Rwanda DHS 2010 and Rwanda DHS 2014-15 data and the SAE methods to estimate the district-level TFR and CMR. The TFR is the average number of children a woman would give birth to during her whole childbearing age 15-49. It is calculated as the sum of the 7 standard 5-year age-specific fertility rates (ASFR), and multiplying by 5, for a specific reference period. The standard reference period is the 3 years before the survey. The ASFR is the ratio of live births to women-years of exposure, for a specific age group. The 7 ASFRs are $ASFR_{15\text{-}19}$, $ASFR_{20\text{-}24}$, $ASFR_{25\text{-}29}$, … $ASFR_{45\text{-}49}$. The TFR is a complex ratio. Childhood mortality rates (CMR) refer to the probability of dying between birth and a specific timepoint in the child's life, under age 5, expressed per 1,000 live births. The CMR encompasses five different rates: neonatal mortality rate (the probability of dying within the first month of life), post-neonatal mortality rate (the probability of dying between the first month of life and the first birthday), infant mortality rate (IMR) (the probability of dying between birth and the first birthday), child mortality rate (the conditional probability of dying between the first and the fifth birthday, for children who reached the first birthday), and under-5 mortality rate (the probability of dying between birth and the fifth birthday). These are complex rates, calculated with smaller segments of age. Here we focus on the numerical results for IMR because it is not just a measure of the risk of infant death, but is used more broadly to evaluate community health status, poverty and socioeconomic status, and the availability of quality primary health care services. We calculated the direct estimates and the five SAE estimates for each district together with their 95% confidence intervals for each estimate with the Jackknife variance estimation method. The program used for the variance calculation is a SAS program developed by ICF, which is also the standard program used for sampling error calculations for all DHS surveys.

DHS reports the TFR for the past 3 years and controls survey precision with a 95% confidence interval less than one child wide at the domain level. The confidence interval for direct estimates for all districts has a mean width of 1.43 children, which is beyond our precision control for domain-level estimation. The average length of the confidence intervals for the five SAE estimates is 1.01 for the time-space nearest neighbor estimate, 0.93 for the district center nearest neighbor method, 0.96 for the cluster center nearest neighbor method, 0.87 for the composite distance measure, and 0.88 for the hybrid estimates, which are all under our controlled precision as domain-level TFR estimation. Although some of the SAE estimates do not have confidence interval widths less than one child, a few have a confidence interval only slightly wider than one child. These can be improved by using a district-specific SAE method in practice. The SAE estimates were adjusted for consistency with the provincial TFR based on formula (15) and (16), by using the provincial TFR and district SAE TFR in the place of the totals, with the weights being the total weighted number of women years of exposure by district. Table 7 shows the results of the direct estimate and the five consistency-adjusted SAE estimates by district, together with the length of confidence interval (LenCI). The confidence intervals for the consistency-adjusted SAE estimates were calculated based on their Jackknife variance estimations. We did not calculate a full Jackknife variance, but used a somewhat "conditional" Jackknife with the nearest neighborhood fixed and a partially Jackknifed adjustment factor in (15). The Jackknife is performed only on the TFR estimation of the target district, and partially on the adjustment factor through the Jackknifed TFR of the target district. A full Jackknife would require an automated program that is not currently available. We will continue this study and develop an automated program for future use.

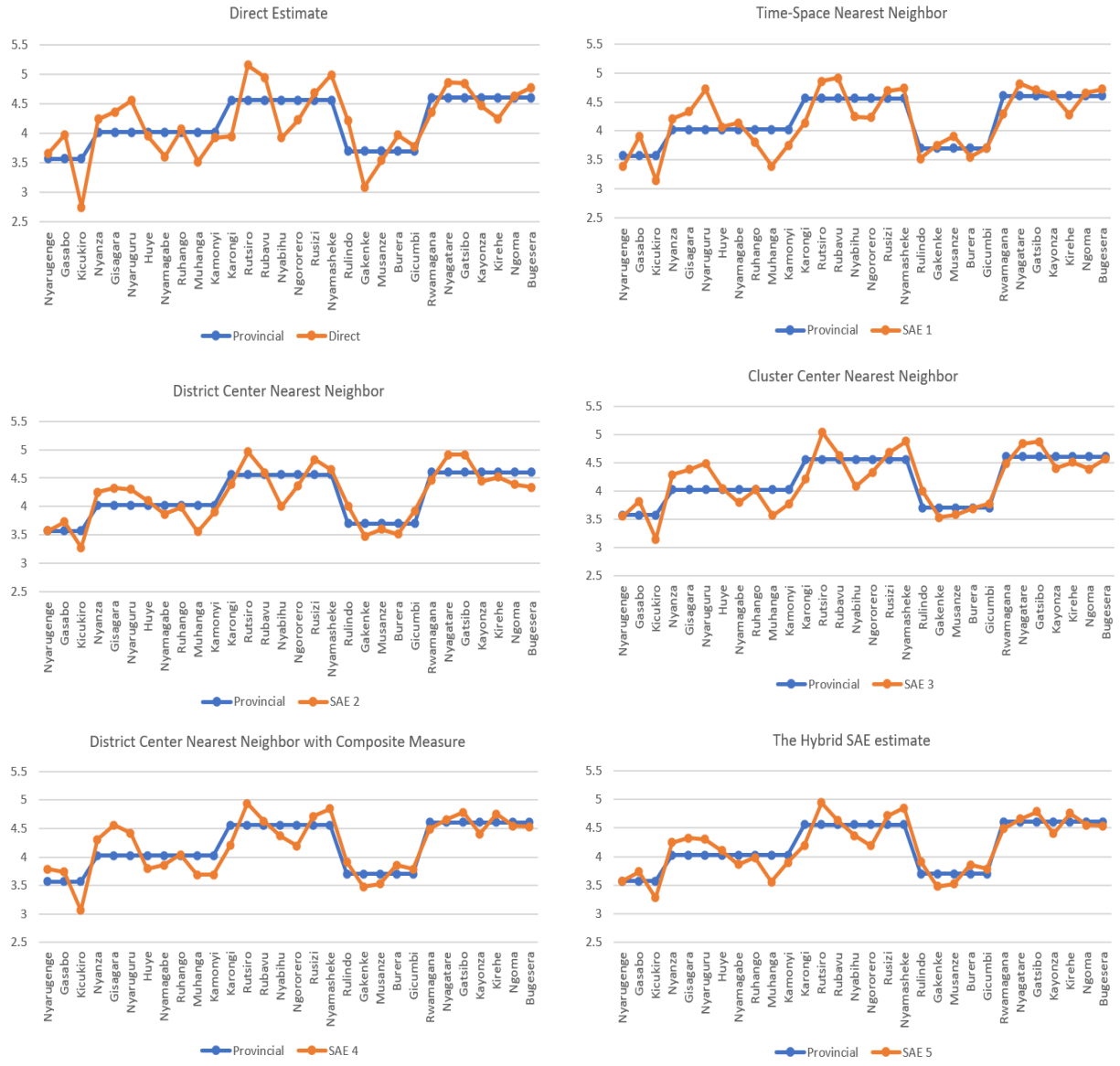
Figure 2 below is an illustrative presentation of the numerical results, with the various estimates of the TFR plotted against the provincial TFR estimate at the same scale. We can see that the curves of the SAE 2, SAE 3, and SAE 4 are all close to the provincial estimates with moderate variations, while the direct estimate and the SAE 1 have larger variations. The SAE 5, the hybrid estimate, seems to have the best fit. We believe that district-level estimates with moderate variations compared to the provincial estimate may more accurately reflect the true situation. For all the estimates, district Kicukiro of Kigali City Province has the lowest TFR. This district is mostly urban (90%), with 94% of the interviewed women living in the fourth wealth quintile (rich) or above, the highest education level (52% of them have secondary or higher education), the lowest rate of currently married (39%), the lowest rate of currently pregnant (4%), and a high rate of current use of modern contraception (46%). While the district Rutsiro of West Province has the highest TFR in all the estimates, this district is among the mostly rural (3% urban) districts, with 59% of the interviewed women living in the second wealth quintile (poor) or lower, among the districts with lowest education level (only 14% of them have second or higher education), and with a high rate of currently married women (59%).

Table 7 Direct estimate and the consistency-adjusted SAE estimates and their length of confidence interval for the TFR (the past 3 years) by district

District	Direct estimate		Time-space nearest neighbor		District center nearest neighbor		Single cluster nearest neighbor		District center composite measure		Hybrid SAE estimates	
	Direct	LenCI	SAE 1	LenCI	SAE 2	LenCI	SAE 3	LenCI	SAE 4	LenCI	SAE 5	LenCI
Nyarugenge	3.66	1.47	3.38	0.79	3.57	0.82	3.55	0.87	3.78	0.70	3.57	0.82
Gasabo	3.98	1.31	3.91	0.53	3.73	0.57	3.82	0.55	3.74	0.47	3.73	0.57
Kicukiro	2.75	1.14	3.14	0.70	3.28	0.80	3.14	0.74	3.07	0.76	3.28	0.80
Nyanza	4.24	1.29	4.21	0.74	4.25	0.71	4.28	0.81	4.31	0.69	4.25	0.71
Gisagara	4.36	1.29	4.33	0.79	4.32	0.67	4.38	0.74	4.56	0.63	4.32	0.67
Nyaruguru	4.56	1.68	4.72	0.91	4.30	0.82	4.48	0.99	4.42	0.85	4.30	0.82
Huye	3.95	1.29	4.06	0.75	4.10	0.75	4.04	0.74	3.79	0.75	4.10	0.75
Nyamagabe	3.60	1.21	4.14	0.82	3.87	0.66	3.79	0.64	3.86	0.68	3.87	0.66
Ruhango	4.07	1.27	3.80	0.71	3.98	0.66	4.02	0.75	4.04	0.69	3.98	0.66
Muhanga	3.52	1.10	3.39	0.71	3.56	0.60	3.57	0.65	3.68	0.59	3.56	0.60
Kamonyi	3.93	1.22	3.75	0.70	3.90	0.76	3.77	0.70	3.69	0.69	3.90	0.76
Karongi	3.95	1.58	4.14	0.81	4.39	0.89	4.21	0.91	4.20	0.86	4.20	0.86
Rutsiro	5.15	3.47	4.85	0.97	4.96	0.99	5.04	0.97	4.94	0.92	4.94	0.92
Rubavu	4.95	1.65	4.91	0.92	4.59	1.00	4.63	1.08	4.63	0.92	4.63	0.92
Nyabihu	3.93	1.36	4.25	0.95	4.00	0.87	4.08	0.85	4.37	0.77	4.37	0.77
Ngororero	4.22	1.84	4.23	0.86	4.37	0.89	4.33	0.96	4.19	0.82	4.19	0.82
Rusizi	4.68	1.40	4.69	0.95	4.83	0.63	4.68	0.80	4.71	0.73	4.71	0.73
Nyamasheke	4.99	1.00	4.73	0.65	4.65	0.73	4.88	0.62	4.85	0.75	4.85	0.75
Rulindo	4.22	1.46	3.52	0.70	4.01	0.85	3.99	0.76	3.92	0.73	3.92	0.73
Gakenke	3.09	1.32	3.75	0.89	3.47	0.71	3.53	0.78	3.48	0.68	3.48	0.68
Musanze	3.55	1.19	3.91	0.90	3.60	0.66	3.58	0.67	3.52	0.66	3.52	0.66
Burera	3.97	1.22	3.55	0.65	3.51	0.72	3.68	0.80	3.86	0.66	3.86	0.66
Gicumbi	3.77	1.14	3.71	0.71	3.92	0.63	3.77	0.66	3.78	0.71	3.78	0.71
Rwamagana	4.35	1.09	4.29	0.63	4.47	0.84	4.48	0.83	4.49	0.68	4.49	0.68
Nyagatare	4.86	1.63	4.81	0.87	4.92	0.91	4.84	0.99	4.66	0.88	4.66	0.88
Gatsibo	4.85	1.75	4.70	0.86	4.91	0.89	4.87	0.94	4.79	0.84	4.79	0.84
Kayonza	4.47	1.20	4.62	0.93	4.45	0.72	4.40	0.72	4.40	0.69	4.40	0.69
Kirehe	4.25	1.75	4.28	0.89	4.52	0.99	4.51	1.08	4.76	0.89	4.76	0.89
Ngoma	4.63	1.61	4.66	0.91	4.39	0.87	4.39	0.87	4.55	0.85	4.55	0.85
Bugesera	4.78	1.01	4.72	0.83	4.34	0.91	4.57	0.79	4.53	0.61	4.53	0.61
Average		1.43		0.80		0.78		0.81		0.74		0.75

Note: SAE 1=Time-space nearest neighbor; SAE 2=District center nearest neighbor; SAE 3=Cluster center nearest neighbor; SAE 4=District center nearest neighbor with composite distance measure; SAE 5=Hybrid SAE estimate.

Figure 2 **Direct estimates and the five SAE estimates plotted against the provincial estimates for the TFR in the past 3 years**



Note: SAE 1=Time-space nearest neighbor; SAE 2=District center nearest neighbor; SAE 3=Cluster center nearest neighbor; SAE 4=District center nearest neighbor with composite distance measure; SAE 5=Hybrid SAE estimate.

For the CMR estimation, the DHS program reports the CMR for the past 10 years at the domain level and controls the survey precision with a coefficient of variation under 15% for domain-level estimations. Due to the large number of tables, we only report the results for the IMR, the infant mortality rate. The direct estimates for all districts have an average coefficient of variation (CV) of 24%, which is far above our controlled precision. The time-space nearest neighbor has an average CV of 15.4%, which is slightly over 15%. The district center nearest neighbor has an average CV of 16.8%; the cluster center nearest neighbor has an average CV of 17.7%; the district center nearest neighbor with composite distance measure has an average CV of 16%; and the hybrid SAE estimate an average CV of 16.2%. Although all estimates have a CV slightly over 15%, some have a CV near 20%. Doubling the sample size at district level is not enough

for CMR estimation. This can be solved by increasing the number of clusters borrowed or increasing the nearest neighborhood size. We did not resort to district-specific remedies here because the aim of this study is not to obtain the best estimate for any specific district. Instead, our aim was to present the ideas and basic findings. The SAE estimates were adjusted for consistency with provincial-level IMR based on formula (15) and (16), by using the provincial-level IMR and the district-level SAE IMR in the place of the totals, with the weights being the weighted number of children under age 5 exposed to risk of death in the past 10 years by district. Table 8 shows the results of the direct estimate and the five consistency-adjusted SAE estimates by district, together with their CV. As we did for the TFR estimation, we did not calculate a full Jackknife variance estimation for the consistency-adjusted SAE estimates because we do not yet have an automated program for this task.

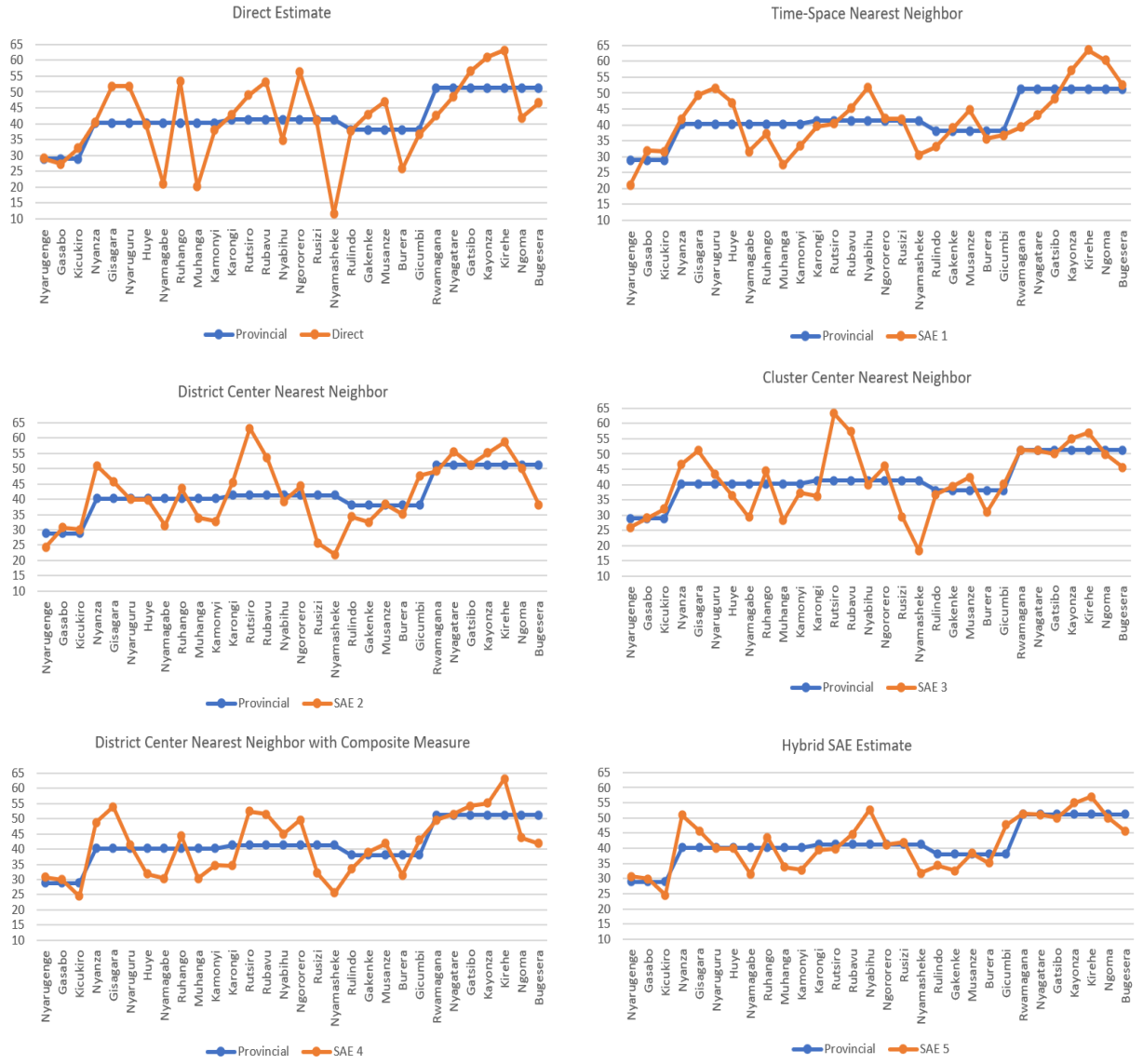
Figure 3 is an illustrative presentation of the results, with the various estimated IMRs plotted against the provincial IMR estimate on the same scale. The curve for the direct estimate and SAE 1, SAE 2 and SAE 3 have larger variations compared to the provincial IMR estimates, while SAE 4, the nearest neighbor with composite measure, has relatively small variation. The best fit is the hybrid estimate SAE 5. We believe that district-level estimates with moderate variations compared to the provincial estimate may more accurately reflect the true situation. For example, from the direct estimate, district Nyamasheke of West Province has the lowest IMR, 11.5 per 1,000 live births, which must be largely underestimated because this province is mostly rural (only 2% urban), has the lowest secondary education (only 14% of the interviewed women have secondary or higher education), 59% of the interviewed women living in the bottom two wealth quintiles, and only 18% in the top two wealth quintiles. For the hybrid SAE, district Kicukiro of Kigali City Province has the lowest IMR, 24.5 per 1,000 live births. This district is the most developed district in Rwanda, and has the lowest TFR too, as described above. The district Kirehe in the East Province has the highest IMR, 57 per 1,000 live births, and is among the most underdeveloped districts. It is mostly rural (4% urban), with 42% of the interviewed women in the bottom two wealth quintiles, the lowest education level (only 13% have secondary or higher education), the highest rate of currently married (62%), and a high TFR (4.76).

Table 8 Direct estimate and the five consistency-adjusted SAE estimates and their length of confidence interval for IMR (the past 10 years) by district

District	Direct estimate		Time-space nearest neighbor		District center nearest neighbor		Single cluster nearest neighbor		District center composite measure		Hybrid SAE estimates	
	Direct	CV	SAE 1	CV	SAE 2	CV	SAE 3	CV	SAE 4	CV	SAE 5	CV
Nyarugenge	29.04	0.22	20.57	0.11	24.35	0.14	25.99	0.14	30.72	0.11	30.72	0.11
Gasabo	27.31	0.24	32.14	0.07	30.76	0.08	29.02	0.09	29.94	0.07	29.94	0.07
Kicukiro	32.28	0.25	31.51	0.13	30.11	0.13	32.00	0.14	24.50	0.15	24.50	0.15
Nyanza	40.49	0.29	42.01	0.15	50.95	0.14	46.54	0.17	48.80	0.13	50.95	0.14
Gisagara	51.95	0.25	49.28	0.12	45.70	0.15	51.20	0.15	53.94	0.12	45.70	0.15
Nyaruguru	51.88	0.18	51.54	0.11	39.99	0.12	43.37	0.14	41.27	0.11	39.99	0.12
Huye	39.63	0.16	47.26	0.14	39.90	0.16	36.26	0.15	31.87	0.14	39.90	0.16
Nyamagabe	20.98	0.37	32.20	0.16	31.48	0.16	29.27	0.19	30.30	0.15	31.48	0.16
Ruhango	53.57	0.27	36.18	0.16	43.52	0.15	44.42	0.16	44.35	0.14	43.52	0.15
Muhanga	20.12	0.37	27.85	0.17	33.83	0.15	28.21	0.20	30.41	0.17	33.83	0.15
Kamonyi	38.00	0.17	33.14	0.11	32.79	0.12	37.20	0.11	34.75	0.13	32.79	0.12
Karongi	42.88	0.23	39.35	0.14	45.51	0.17	36.19	0.18	34.49	0.17	39.35	0.14
Rutsiro	49.04	0.23	39.83	0.14	63.20	0.11	63.34	0.11	52.39	0.12	39.83	0.14
Rubavu	53.18	0.16	44.69	0.11	53.58	0.09	57.31	0.09	51.47	0.09	44.69	0.11
Nyabihu	34.91	0.25	52.66	0.09	39.35	0.15	40.01	0.18	44.99	0.14	52.66	0.09
Ngororero	56.28	0.28	41.19	0.15	44.36	0.17	45.96	0.18	49.52	0.15	41.19	0.15
Rusizi	40.97	0.30	42.01	0.13	25.73	0.23	29.47	0.27	32.13	0.19	42.01	0.13
Nyamasheke	11.52	0.39	31.72	0.20	21.95	0.25	18.39	0.25	25.60	0.22	31.72	0.20
Rulindo	37.83	0.22	32.95	0.11	34.27	0.14	36.66	0.12	33.49	0.13	34.27	0.14
Gakenke	42.99	0.28	38.69	0.20	32.50	0.17	39.36	0.15	39.05	0.15	32.50	0.17
Musanze	47.02	0.15	44.64	0.10	38.27	0.10	42.39	0.10	41.80	0.11	38.27	0.10
Burera	25.85	0.36	36.10	0.15	35.05	0.14	30.90	0.17	31.44	0.16	35.05	0.14
Gicumbi	36.58	0.22	36.76	0.11	47.71	0.11	40.15	0.12	42.99	0.12	47.71	0.11
Rwamagana	42.44	0.16	39.08	0.12	49.13	0.12	51.32	0.13	49.64	0.11	51.32	0.13
Nyagatare	48.53	0.17	42.64	0.11	55.50	0.11	51.13	0.13	51.54	0.11	51.13	0.13
Gatsibo	56.49	0.23	47.62	0.14	51.17	0.12	50.01	0.14	54.18	0.11	50.01	0.14
Kayonza	60.95	0.21	56.92	0.13	55.20	0.12	55.03	0.13	55.15	0.11	55.03	0.13
Kirehe	63.02	0.15	63.50	0.10	58.61	0.12	57.00	0.13	63.14	0.11	57.00	0.13
Ngoma	41.80	0.27	61.63	0.11	49.99	0.13	49.94	0.14	43.79	0.15	49.94	0.14
Bugesera	46.51	0.26	52.77	0.12	38.22	0.18	45.63	0.17	41.90	0.15	45.63	0.17
Average		0.24		0.13		0.14		0.15		0.13		0.14

Note: SAE 1=Time-space nearest neighbor; SAE 2=District center nearest neighbor; SAE 3=Cluster center nearest neighbor; SAE 4=District center nearest neighbor with composite distance measure; SAE 5=Hybrid SAE estimate.

Figure 3 **Direct estimates and the five consistency-adjusted SAE estimates plotted against the provincial estimates for IMR in the past 10 years**



Note: SAE 1=Time-space nearest neighbor; SAE 2=District center nearest neighbor; SAE 3=Cluster center nearest neighbor; SAE 4=District center nearest neighbor with composite distance measure; SAE 5=Hybrid SAE estimate.

5 CONCLUSION

In this paper, we proposed SAE methods that use data from a single target survey or similar surveys conducted in recent years in the same area. This approach assumes that sampling units close to each other, in geographical distance or in other distance measures, tend to be similar. The methods create a survey domain that covers the small area by pooling sampling units/clusters from neighboring areas. We use data collected from the Rwanda DHS 2010 and DHS 2014-15 to illustrate the methods. The Rwanda DHS 2014-15 was the target survey. We produced SAE estimates for the TFR and CMR for each of 30 districts, a level of disaggregation not included in the final survey report.

We first generated the nearest neighborhoods with different methods that included the *time-space nearest neighbo*r, combining the 2010 DHS data with the 2014-15 survey data and doubling the sample size at district level. The *district center nearest neighbor* method pooled clusters from other districts within the same province or in other neighboring provinces that are geographically close to the target district center, and used the GPS information from each cluster in the 2014-15 survey to identify the 20 closest donor clusters for each district. The *cluster center nearest neighbor* method pooled clusters from other districts that are geographically close to a target cluster in the target district, and identified a different number of closest donor clusters, with a target of about 16 donor clusters for each district. A *composite distance measure* method integrates women's individual demographic characteristics with geographical distance and wealth index into a district profile, and then applies the *district center nearest neighbor* method using a composite distance measure. A *hybrid method* takes nearest neighborhoods created by different methods at the provincial level, by applying the provincial level best-fitting method for all districts in the same province. We described the hybrid method but did not explicitly provide the nearest neighbor construction table because it is included in other tables.

We produced direct estimates and the various SAE estimates with their variance estimates and confidence intervals for the TFR and IMR from each of the 30 districts for the target survey, Rwanda DHS 2014-15. The SAE estimates reported here are the consistency-adjusted estimates. The estimates are plotted against the provincial level survey estimates to provide a visual representation of each method's performance. For both TFR and IMR estimations, all SAE estimates performed better than the direct estimates; among the four basic SAE estimates, the district center nearest neighbor with composite distance measure performs better. The hybrid method performed the best among all the five SAE. For simplicity, the SAE estimates reported here are the most basic estimates that used a uniform method at the provincial level, and not a district-specific method. Results can be improved when we construct district-specific nearest neighborhoods with a district-specific method.

REFERENCES

Asian Development Bank. 2020. *Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices*.
http://dx.doi.org/10.22617/TIM200160-2.

Chambers, R. and H. Chandra. 2008. "Improved Direct Estimators for Small Areas." University of Wollongong, Working Paper 03-08. Wollongong, New South Wales, Australia: University of Wollongong Centre for Statistical and Survey Methodology.
http://ro.uow.edu.au/cssmwp/3/.

Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York, USA: Chapman & Hall/CRC.
https://doi.org/10.1007/978-1-4899-4541-9.

Ghosh, M. and J. N. K. Rao. 1994. "Small Area Estimation: An Appraisal." *Statistical Science* 9 (1): 55-76.
https://doi.org/10/1214/ss/1177010647.

Lahiri, P. and J. N. K. Rao. 1995. "Robust Estimation of Mean Squared Error of Small Area Estimators." *Journal of the American Statistical Association* 90 (430): 758-766.
https://doi.org/10.2307/2291089/.

Marker, D. A. 1999. "Organization of Small Area Estimators Using a Generalized Linear Regression Framework." *Journal of Official Statistics* 15 (1): 1-24.

National Institute of Statistics of Rwanda (NISR), Ministry of Health (MOH), Rwanda, and ICF International. 2012. *Rwanda DHS 2010 Final Report*, Calverton, Maryland, USA: NISR, MOH, ICF International.
http://dhsprogram.com/pubs/pdf/FR259/FR/259.pdf.

National Institute of Statistics of Rwanda (NISR), Ministry of Finance and Economic Planning (MFEP), Ministry of Health (MOH), Rwanda, and ICF International. 2016. *Rwanda DHS 2014-15 Final Repor*t. Rockville, Maryland, USA: NIS, MFEP, MoH of Rwanda, and ICF International.
http://dhsprogram.com/pubs/pdf/FR316/FR/316.pdf.

Platek, R., J. N. K. Rao, C. E. Särndal, and M. P. Singh. 1987. *Small Area Statistics*. New York, USA: John Wiley & Sons.
https://pt.booksc.org/book/683662/39046b.

Prasad, N. G. N. and J. N. K. Rao. 1990. "The Estimation of the Mean Squared Error of Small Area Estimators." *Journal of the American Statistical Association* 85 (409): 163-171.
https://doi.org/10.2307/2289539.

Särndal, C. E. 1984. “Design-consistent versus Model-dependent Estimation for Small Domains.” *Journal of the American Statistical Association* 79 (387): 624-631. https://doi.org/10.2307/2288409.

Särndal, C. E. and M. A. Hidiroglou. 1989. “Small Domain Estimation: A Conditional Analysis.” *Journal of the American Statistical Association* 84 (405): 266-275. https://doi.org/10.2307/2289873.

Särndal, C. E. 1984. “Design-consistent versus Model-dependent Estimation for Small Domains.” *Journal of the American Statistical Association* 79 (387): 624-631. https://doi.org/10.2307/2288409.

Schirm A. L., A. M. Zaslavsky, and L. Czajka. 1997. *Large Numbers of Estimates for Small Areas*. Working Paper. Washington, DC, USA: Mathematica Policy Research. https://econpapers.repec.org/RePEc:mpr:mprres:5a8722c29a9b49ee88462504763ebcc8.

Wolter, K. M. 1985. *Introduction to Variance Estimation*. New York, USA: Springer-Verlag. https://link.springer.com/content/pdf/10.1007%2F978-0-387-35099-8.pdf.

Woodruff, R. S. 1971. “A Simple Method for Approximating the Variance of a Complicated Estimate.” *Journal of the American Statistical Association* 66 (334): 411-414. https://doi.org/10.2307/2283947.