Small-Area Estimation (SAE) in the application of Prevalence of Malaria among children aged 6-59 months in Kenya at the Region Levels

Jennifer Ci
Stat 544 Final Project

Abstract: The aim of this analysis is to provide an overview of malaria prevalence in Kenya using small area estimation (SAE) techniques in comparing direct estimators and indirect estimators. We will first review the current status of malaria control in Kenya and the role of SAE in improving our understanding of the prevalence and distribution of the disease. Next, we will discuss the challenges and limitations of using SAE to estimate malaria prevalence in Kenya and consider the implications of these estimates for effective malaria control and prevention strategies. Finally, we will identify areas for future research that can help improve our understanding of the factors contributing to malaria transmission and inform more targeted interventions.

1 Introduction

Malaria is a significant public health concern in Kenya, with the country ranking among the high-burden malaria countries in sub-Saharan Africa. Despite considerable efforts to control and prevent malaria transmission in the country, it remains a leading cause of morbidity and mortality, especially among children under five years old and pregnant women. Accurate estimates of malaria prevalence are crucial for planning and implementing effective control and prevention strategies. However, obtaining reliable estimates at a small area level can be challenging due to limitations in data availability and quality.

According to the 2020 World Malaria Report by the World Health Organization, there were an estimated 5.6 million cases of malaria in Kenya in 2019, resulting in 3,634 deaths attributed to the disease. Kenya accounted for approximately 3% of global malaria cases in 2019, with children under five years of age being the most vulnerable group. The report also noted that while malaria prevalence in Kenya has been declining over the years, progress has been slower in recent years, with a slight increase in malaria cases reported in 2018 (WHO, 2020).

The direct estimates become problematic when the sample sizes in the subgroups of interest are small. In this situation, model-dependent methods are increasingly being used to produce what are termed as "small area" or "indirect" estimates. Small area estimation (SAE) is a statistical technique that is increasingly being used to estimate disease prevalence in areas where health data is limited or unavailable. SAE utilizes statistical models to combine data from multiple sources, including household surveys, health facility data, and remote sensing data, to produce estimates of disease prevalence at a small area level.

In the context of malaria, SAE can provide a more accurate understanding of the distribution of the disease and its burden, which can inform targeted interventions and resource allocation in Kenya.

Several studies have employed SAE methods to estimate disease prevalence in various settings. For example, a study conducted in Brazil used SAE to estimate smoking prevalence at the district level (Bernal, 2020), while another study in Kenya, Tanzania, and Mozambique estimated HIV prevalence at the first sub-national level (Saldarriaga, 2021). These studies have demonstrated the usefulness of SAE in providing accurate estimates of disease prevalence at a small area level.

In summary, the application of SAE methods to estimate the prevalence of malaria among children aged 6-59 months in Kenya at a small area level is crucial for identifying areas that require targeted interventions to reduce the disease burden. This analysis will provide valuable insights into the distribution of malaria in Kenya and inform policy and decision-making related to malaria control efforts.

2 Data description

In this study, we used the Kenya MIS 2020 from DHS (DHS) and geospatial data(GADM). MIS collects data on all of the internationally recognized malaria indicators including diagnostic blood testing of children under five with fever (DHS). Kenya has 7 regions, each of which is subdivided into a number of counties, for a total number of 47 counties. Figure 1 shows a map with divided county boundaries for Kenya. The stratification is based on urban and rural areas in each region for the DHS program. For Kenya 2020, the stratification was county (47) and urban/rural (2), Nairobi and Mombasa are entirely urban, so there are 92 strata in total. The 2020 Kenya MIS is a two-stage stratified cluster sample design.

For the first step of the sampling process, 301 clusters were chosen at random from the NASSEP V master sample frame. These clusters were composed of 134 urban and 167 rural areas. An equal probability selection method was used, and each sampling stratum was selected independently. In the second stage, systematic random sampling was utilized to randomly select 30 households per cluster from a list of households in the sampled clusters. Field teams also updated maps and recorded geographic coordinates during the listing process. However, due to a lack of security, two clusters were not listed, one cluster was not listed due to pastoralist migration, and eight clusters were not listed because they contained less than 30 households, which end with 290 clusters. Figure 2 shows the Cluster locations for Kenya, with the majority being the West part of Kenya (Province: Nyanza and Western).

The weight was calculated by MIS as household weight (hv005) for a particular household is the inverse of its household selection probability multiplied by the inverse of the household response rate in the stratum.

Figure 1 Map of Kenya

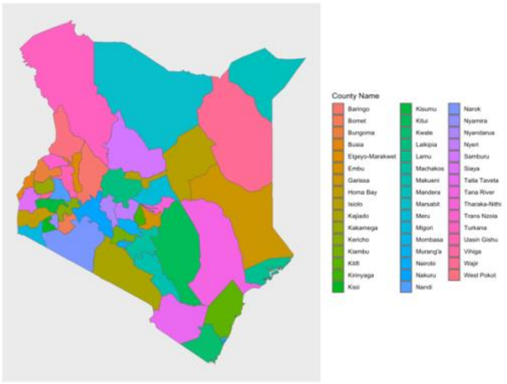
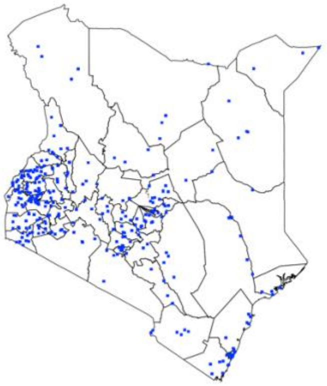


Figure 2: Cluster locations in Kenya MIS, with county boundaries



3 Methods

A naïve estivation ignoring the survey design with a binomial model was performed as the first step since it is the simplest method. Direct Estimation (Horvitz-Thompson) will then perform account for the different design weights associated with each sample. Given that the MIS survey collects multistage cluster design samples, the spatial Bayes Fay Harriot models accounts BYM2 spatial effects (an iid normal random effect and an intrinsic CAR (ICAR) random) with smoothing over space will be performed lastly. In order to perform the smoothing, we needed to construct a spatial adjacency matrix of the regions. We performed the following analysis using R, the BYM2 model was implemented in the R package SUMMER (Martin, 2018).

3.1 Naïve Estimation

Let yi and ni denote the number of children testing positive for malaria using either microscopy or RDT and the total number of children in areas respectively. Ignoring the survey design, the naive estimate for the prevalence of malaria is $\widehat{p_i} = \frac{y_i}{n_i}$ with associated variance being$\frac{p_i(1-p_i)}{n_i}$. This naive estimate can be easily calculated by tabulating the data.

Yi|pi ~iid Binomial(ni, pi)

3.2 Direct Estimation - Horvitz and Thompson Estimator

A direct estimate of a quantity in a specific area only uses data on the variable of interest from that area (Horvitz and Thompson, 1952). To account for the different design weights associated with each sample, we can estimate the survey weighted estimates (the Horvitz-Thompson estimates) using the survey package. The weighted estimator of the log-likelihood for our binary malaria data is

$$\sum_{i=1}^{n}\sum_{k=1}^{m_i} w_{ik}\{y_{ik}\log P_i + (1-y_{ik})\log(1-P_i)\}$$

$$\text{logit}(\hat{p}_i) \sim N(\theta, V)$$

where $y_{ik}$is the binary malaria outcome on child k in area i, with associated weight w_{ik}.

3.3 Indirect Estimation - Fay – Herriot Area Level Model

The direct smoothed Fay – Herrio method is an example of an area-level model. It treats the direct (weighted) estimates as data, and then fits a spatial smoothing model. To estimate the spatially smoothed estimator, we fitted the data to the following model (Fay,1979):

$\widehat{\theta_i} \sim N(\theta_i, \widehat{V_i})$

With $\widehat{\theta_i} = \log\left(\frac{\hat{p}_i}{(1-\hat{p}_i)}\right)$ where $\hat{p}_i$ being the direct (design-based weighted) estimate and $\widehat{V_i}$ is the variance of this estimate for each region and $\theta_i = \log\left(\frac{p_i}{(1-p_i)}\right) + Si + \epsilon_i$,

$$S_i \sim ICAR(\sigma_s^2)$$

$$\epsilon_i \sim_{iid} N(0, \sigma_\epsilon^2)$$

Si is the spatial smoothing, ϵi is the unstructured random effects.

4 Results

The original dataset for Kenya had a sample size of 30651 observations. After deleting the observation containing missing data in the malaria information only (no missing values in the geographic data were found), the final sample size was 11357.

Table 1 shows the distribution of the respondents included in the study by selected background characteristics. Of the total of 11357 children aged 6–60 months who were tested for malaria, 50.6% were female. A majority of participants were aged 49-60 months (74.6). Most participants lived in rural areas (66.4%). Western and Nyanza were the top two provinces with the most Malaria cases among children for the reported population (49.8%, 38.4%).

Table 1 Descriptive statistics with the pooled sample of MIS data in Kenya

	Malaria Negative	Malaria Positive	Overall	P-value
	(N=9966)	(N=1391)	(N=11357)	
Age (month)				
0-12	421 (4.2%)	20 (1.4%)	441 (3.9%)	<0.001
13-24	762 (7.6%)	62 (4.5%)	824 (7.3%)	
25-36	721 (7.2%)	60 (4.3%)	781 (6.9%)	

	Malaria Negative	Malaria Positive	Overall	P-value
37-48	767 (7.7%)	77 (5.5%)	844 (7.4%)	
49-60	7295 (73.2%)	1172 (84.3%)	8467 (74.6%)	
Sex				
Female	5053 (50.7%)	697 (50.1%)	5750 (50.6%)	0.917
Male	4913 (49.3%)	694 (49.9%)	5607 (49.4%)	
Area				
Rural	6399 (64.2%)	1142 (82.1%)	7541 (66.4%)	<0.001
Urban	3567 (35.8%)	249 (17.9%)	3816 (33.6%)	
Province				
Central	517 (5.2%)	0 (0%)	517 (4.6%)	<0.001
Coast	1263 (12.7%)	81 (5.8%)	1344 (11.8%)	
Eastern	1188 (11.9%)	13 (0.9%)	1201 (10.6%)	
Nairobi	54 (0.5%)	1 (0.1%)	55 (0.5%)	
North Eastern	557 (5.6%)	0 (0%)	557 (4.9%)	
Nyanza	2230 (22.4%)	534 (38.4%)	2764 (24.3%)	
Rift Valley	2351 (23.6%)	69 (5.0%)	2420 (21.3%)	
Western	1806 (18.1%)	693 (49.8%)	2499 (22.0%)	

Figure 3 shows the results of different approaches to SAE. Based on the weighted smoothed estimation, the estimated prevalence for all regions of Kenya ranged from 0.2% to 51.4%. Our analysis allowed us to identify the prevalence of malaria in each small region in Kenya.

In general, malaria prevalence is highest in more Western parts of Kenya than in Eastern parts of Kenya, with the highest prevalence being in Western Busia(51.4%), Nyanza Siaya(42.3%), and Western Bungoma (32.5) according to the smoothed weighted model (Table 2). We can see that the regions with the highest malaria prevalence tend to be gathered on the East side of Lake Victoria and the West side of Lake Turkana. The Supplemental Material contains tables for detailed estimated prevalence and uncertainty estimates for each region for all three models.

According to Table 2, both naïve estimation and direct estimation gave zero as the estimator for the region that lacks cases collected.

Figure 3: Mapping of the estimates and uncertainty estimates for the three estimation. Top: Estimates (posterior median for the Weighed Smoothed Estimation) of Malaria prevalence. Bottom: Uncertainty estimates (posterior standard deviations for the Weighed Smoothed Estimation, standard errors for remainder) of Malaria prevalence.

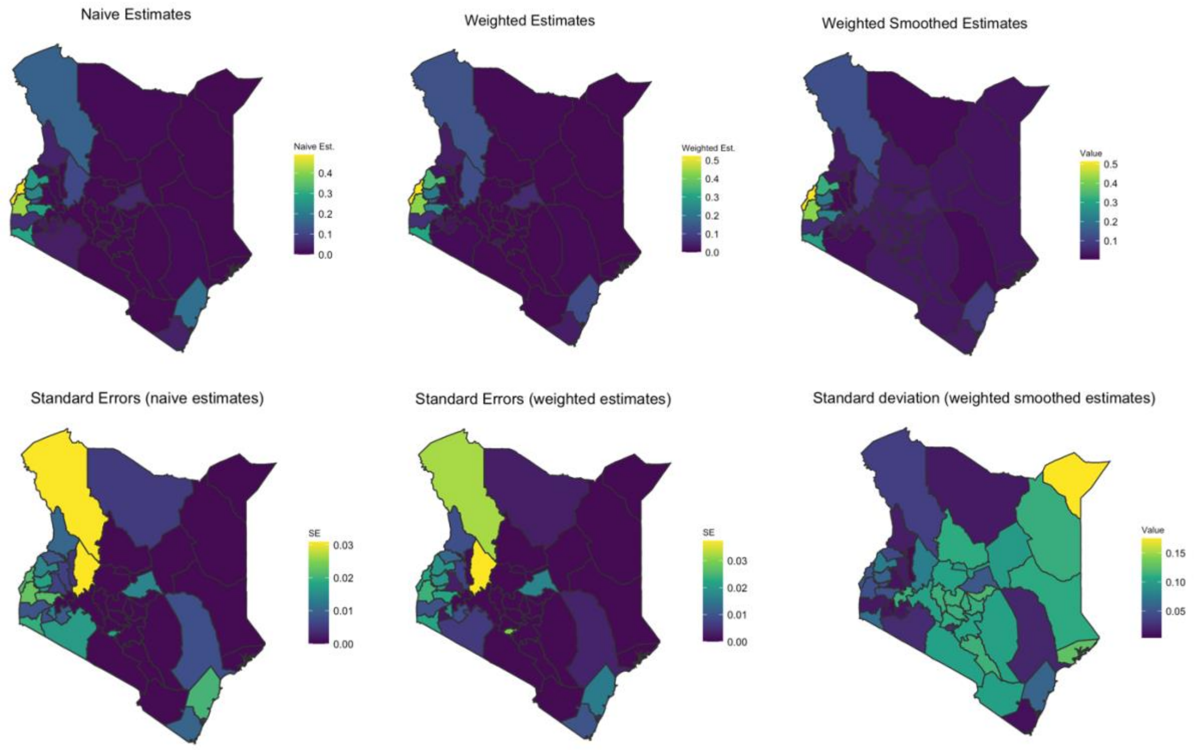


Table 2 for Malaria Prevalence in each estimation

Province	Region	Naïve	Direct	Smoothed Direct
Central	Kirinyaga	0.0	0.0	3.3
Central	Kiambu	0.0	0.0	3.4
Central	Murang'a	0.0	0.0	3.4
Central	Nyeri	0.0	0.0	3.4
Central	Nyandarua	0.0	0.0	3.5

Province	Region	Naïve	Direct	Smoothed Direct
Coast	Mombasa	0.6	0.4	0.8
Coast	Tana River	0.8	0.3	1.1
Coast	Lamu	0.0	0.0	2.6
Coast	Taita Taveta	0.0	0.0	3.2
Coast	Kwale	3.9	3.6	3.5
Coast	Kilifi	17.9	11.9	8.9
Eastern	Marsabit	0.5	0.3	0.8
Eastern	Isiolo	0.0	0.0	2.8
Eastern	Kitui	0.0	0.0	3.1
Eastern	Embu	0.0	0.0	3.3
Eastern	Machakos	0.0	0.0	3.3
Eastern	Makueni	0.0	0.0	3.3
Eastern	Meru	5.0	5.8	4.5
Nairobi	Nairobi	1.8	3.1	3.1
North Eastern	Garissa	0.0	0.0	2.5
North Eastern	Wajir	0.0	0.0	2.5
North Eastern	Mandera	0.0	0.0	2.6
Nyanza	Kisii	1.3	0.3	0.9
Nyanza	Nyamira	3.0	3.2	3.2
Nyanza	Homa Bay	6.5	6.9	6.9
Nyanza	Kisumu	27.4	23.7	22.7
Nyanza	Migori	28.4	30.9	28.4
Nyanza	Siaya	41.6	43.2	42.3
Rift Valley	Elgeyo-Marakwet	0.5	0.0	0.2

Province	Region	Naïve	Direct	Smoothed Direct
Rift Valley	Nandi	0.6	0.3	0.9
Rift Valley	Uasin Gishu	0.6	0.7	1.4
Rift Valley	Narok	3.4	0.9	1.7
Rift Valley	Bomet	2.7	2.2	2.4
Rift Valley	Trans Nzoia	1.7	2.5	3.1
Rift Valley	Kajiado	0.0	0.0	3.3
Rift Valley	Laikipia	0.0	0.0	3.4
Rift Valley	Samburu	0.0	0.0	3.4
Rift Valley	West Pokot	4.5	3.4	3.4
Rift Valley	Nakuru	0.0	0.0	3.5
Rift Valley	Kericho	0.0	0.0	4.6
Rift Valley	Baringo	10.3	12.2	7.1
Rift Valley	Turkana	15.3	13.4	12.5
Western	Vihiga	10.9	13.2	13.0
Western	Kakamega	22.2	21.4	20.3
Western	Bungoma	26.2	35.0	32.5
Western	Busia	49.0	52.3	51.4

5 Discussion

5.1 Conclusions

Small area estimation techniques have shown great potential in estimating the prevalence of malaria at a local level. These techniques allow for more accurate and precise estimates than traditional survey methods by utilizing both survey and auxiliary data. It is helpful to extend small area estimation to produce the needed reliable estimates for all areas including those small local areas with a small sample size. SAE can help identify areas with the highest malaria prevalence, which can be useful for targeting interventions, such as bed net distribution, insecticide spraying, and education campaigns. By focusing on areas with the highest prevalence, interventions can be more effective and efficient, leading to better outcomes. Furthermore, it can help monitor the impact of interventions and evaluate the effectiveness of malaria prevention and control programs. By using SAE to estimate malaria prevalence over time, policymakers can track progress, identify areas where interventions are working, and adjust strategies as needed.

The naive estimate described before fails to account for the survey design and is thus subject to bias and incorrect variance estimation. A naive estimator is a simple approach where we estimate the parameter of interest for a small area by using the sample mean of the variable of interest for the entire population. This approach assumes that the small area has the same characteristics as the entire population, which may not be true. As a result, the naive estimator may have high bias and low precision, leading to inaccurate estimates for the small area.

Direct estimator, on the other hand, is a more sophisticated approach that accounts for the survey design. The direct estimator is calculated by taking a weighted average of the sample data, where the weights are based on the sampling design. This approach takes into account the variability of the population within the small area and the differences between the small area and the entire population. The direct estimator typically has lower bias and higher precision than the naive estimator, resulting in more accurate estimates for the small area.

In SAE (Small Area Estimation), the direct estimator can be further improved by smoothing the estimates as an area-based model. The Fay – Herriot smoothed model is a method that incorporates available auxiliary information, which can include demographic and socioeconomic characteristics of the small area such as the spatial structure. By incorporating this auxiliary information, the direct smoothed estimator can produce more accurate estimates of the parameter of interest. The smoothed direct estimates is very reliable for examining the subnational variation, but the direct estimates are often unreliable for that estimation.

As we mentioned previously, the naïve estimation and direct estimation failed to estimate the regions with no cases recorded. It is important to note that the absence of reported cases within a particular geographic region does not necessarily indicate a complete lack of infection within that area. Therefore, in order to estimate the prevalence of malaria infection more accurately, it is essential to incorporate information from neighboring areas through surrogate estimation methods. The indirect estimation's smoothing process has a significant benefit, as it shrinks the estimates, thus providing an estimate for those unrecorded areas.

5.2 Limitations

For our dataset, there was more than half of the missing observation on malaria, but we did not perform other analyses dealing with the missing variables since the spatial effects were not clear in this content. Given a large number of missed observations, a missing-at-random analysis for further model improvement. Although socioeconomic and demographic variables influence the prevalence of malaria and these variables are likely spatial correlated, our model did not include covariates. The model also could be improved by adjusting for potential confounding variables including the source of water, the trip to obtain water, the toilet facility, and the total number of rooms (Ayele DG, 2012).

Despite its potential benefits, there are challenges and limitations to using SAE for estimating malaria prevalence in Kenya. For example, the quality and availability of data can vary between regions, and the accuracy of SAE estimates can depend on the appropriateness of the statistical models used. Moreover, the implementation of effective malaria control programs depends on understanding the local context and factors contributing to malaria transmission, which may not be captured in SAE estimates alone.

References

World Health Organization. World malaria report 2020: 20 years of global progress and challenges. Geneva, Switzerland: World Health Organization, 2020.

Bernal, R.T.I., de Carvalho, Q.H., Pell, J.P. et al. A methodology for small area prevalence estimation based on survey data. Int J Equity Health 19, 124 (2020). https://doi.org/10.1186/s12939-020-01220-5

Saldarriaga EM. HIV-Prevalence Mapping Using Small Area Estimation in Kenya, Tanzania, and Mozambique at the First Sub-National Level. Ann Glob Health. 2021 Sep 27;87(1):93. doi: 10.5334/aogh.3345. PMID: 34692427; PMCID: PMC8485867.

The DHS Program – Available Datasets. Funded by USAID. Accessed March 13, 2023. https://dhsprogram.com/data/available-datasets.cfm.

GADM - Accessed March 13, 2023. https://gadm.org/data.html

The DHS Program – MIS Overview. Funded by USAID. Accessed March 13, 2023. https://dhsprogram.com/Methodology/Survey-Types/MIS.cfm

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663–685.

Ayele DG, Zewotir TT, Mwambi HG. Prevalence and risk factors of malaria in Ethiopia. Malar J. 2012 Jun 12;11:195. doi: 10.1186/1475-2875-11-195. PMID: 22691364; PMCID: PMC3473321.

Gething, P.W., Patil, A.P., Smith, D.L. et al. A new world malaria map: Plasmodium falciparum endemicity in 2010. Malar J 10, 378 (2011). https://doi.org/10.1186/1475-2875-10-378

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of James–Stein procedure to census data. Journal of the American Statistical Association, 74, 269–277.

Martin, B. D., Li, Z. R., Hsiao, Y., Godwin, J., Wakefield, J., and Clark, S. J. (2018). SUMMER: Spatio-Temporal Under-Five Mortality Methods for Estimation in R. R package version 0.2.1.