BIOST 555 Final Project

Benjamin Stan
March 16, 2022

Introduction

Human immunodeficiency virus (HIV) is a virus that attacks the immune system by targeting CD4 T cells. It is spread through sexual intercourse, contact with infected blood, sharing needles, and from mother to child through birth or breastfeeding.[1] The infection progresses through three stages, with the most severe stage being Acquired Immunodeficiency Syndrome (AIDS). In this condition, patients can suffer from increasingly severe opportunistic infections that result from a depleted immune system. In the absence of treatment, patients with AIDS survive approximately three years. With modern treatments, however, patients can live for decades with HIV. The life expectancy for a 20 year-old person with HIV has increased from 39 in 1996 to 70 in 2011.[2]

HIV is a health concern in all areas of the world but is particularly prevalent in Sub-Saharan Africa, which has 67% of the world's infections.[3] Among the countries in this region is Cameroon, which has a surface area of 475,000 km^2 and a population of 27 million.[4] Roughly half the population of Cameroon now lives in urban areas.[3] As of 2020, there are an estimated 500,000 people living with HIV in the country, and the disease has a prevalence of 3.0% among adults aged 15-49.[5] This prevalence varies by gender, with women having higher rates (4.0%) than men (1.9%). The disease prevalence has fluctuated greatly in recent decades, reaching a peak of 11% in the early 2000s.[3] Approximately 56% of those infected are aware of their HIV status, making testing and information dissemination critical areas of investment for public health entities.[6] The goal of this analysis was to better characterize spatial trends in the prevalence of HIV in Cameroon through small area estimation methods.

Data Description

This analysis used data from a Demographic Health Survey (DHS) conducted in Cameroon in 2018. This was the fifth health survey conducted by the organization in the country since 1991 and covered topics such as family planning methods, maternal and child health, domestic violence, malaria, and HIV/AIDS.[7] The full data can be found on the DHS website (https://dhsprogram.com/data/dataset/Cameroon_Standard-DHS_2018.cfm?flag=1). Fieldwork was conducted from June to December 2018. The sampling mechanism used by the DHS aims to generate representation at the national, residence, and regional level. The stratification is performed by region and urban/rural designation.[8] It uses a two-stage cluster design. The first stage is a sampling of enumeration areas (EA) from the local census with probability proportional to size. The second stage is a sampling of households within the EA with equal probability systematic sampling.[9] The data contains weights for each individual, which are derived from the household selection probability, household response rate in the stratum, and the individual response rate in the stratum. The survey interviewed a sample of 13,527 women

aged 15-49 in all selected households and 6,978 men aged 15-64 in half of selected households. Because of safety concerns, areas of the country's south and west were not surveyed. The DHS recommends against generating estimates from these regions or comparing to other, more represented areas.

The analysis combined data across three sources coming from the DHS. The primary data source was the HIV dataset, which contained testing information for all individuals administered HIV tests. These tests were conducted in an anonymous, informed and voluntary manner. Blood samples were taken via finger prick and taken to a lab for testing. Full individual-level data was also used for the purpose of calculating covariate values and for associating each HIV test result with a corresponding strata, cluster, and region. Each of the clusters had an associated location found in the DHS GPS dataset. These locations were randomly displaced for the purpose of confidentiality. Urban cluster locations were relocated a maximum of two kilometers from original location, and rural clusters were relocated a maximum of ten kilometers. The random displacement kept each cluster within its second administrative region, a detail that our analysis utilized. Maps for first and second administrative levels were obtained from the Database of Global Administrative Areas (GADM; https://gadm.org/maps/CMR.html).

Methods

Each individual in the HIV data set was associated with a cluster. The DHS GPS dataset was joined with the GADM second administrative level dataset in order to associate each cluster with its corresponding second administrative region. With this combined information, we were able to proceed with area-level methods. The primary outcome was prevalence of HIV, with each subject being identified as either HIV positive or HIV negative. Records with an unknown HIV test result were dropped from the analysis.

Several methods were used to estimate the prevalence of HIV in each of the second administrative regions. The first two methods assumed that the prevalence was independent of spatial relationships and calculated the simple (unweighted) proportions and direct (weighted) proportions. The simple method considered the binomial model below:

$$\begin{aligned} Y_i|p_i &\sim_{iid} Binomial(n_i, p_i) \\ \hat{p_i} &= y_i/n_i \\ \hat{se} &= \sqrt{\hat{p_i}(1-\hat{p_i})/n_i} \end{aligned}$$

The direct method considered a modified proportion based on weights:

$$\begin{aligned} \hat{p_i}^w &= \frac{\sum_{k \epsilon S_i} w_k y_k}{\sum_{k \epsilon S_i} w_k} \\ \hat{se} &= \sqrt{\hat{p_i}^w(1-\hat{p_i}^w)/n_i} \end{aligned}$$

Next we approached the problem using spatial smoothing with a Fay-Herriot model. Using a logit transform on the weighted estimates, this discrete spatial model was fit with an ICAR relationship between neighboring regions. The framework is below:

$$\hat{\theta}_i = log\left[\frac{\hat{p}_i^w}{(1-\hat{p}_i^w)}\right]$$
$$\hat{\theta}_i|\theta_i \sim_{iid} N(\theta_i, \hat{V}_i)$$
$$\theta_i = \alpha + b_i$$

In this model, $\hat{V}_i$ correspond to the estimated design-based variances of $\hat{\theta}_i$, α is the intercept, and b_i are BYM2 random effects. This model was further expanded to include a covariate. The covariate of interest was the survey response to the question "Ever used anything or tried to delay or avoid getting pregnant?". The answers were treated as a binary variable and aggregated in both an unweighted and weighted manner at the second administrative level. When incorporating this information into the model, the spatial and weighting framework remained the same, but a covariate was included in the estimate as follows:

$$\theta_i = \alpha + x_i^T\beta + b_i$$

The corresponding estimates and standard errors for all methods were then compared.

Results

The initial dataset contained records for 14,611 individuals surveyed and tested for HIV. The removal of those with an unknown HIV test result reduced this number by 27. Preprocessing involved taking the geospatial data provided by DHS, which is stored at the cluster level as a latitude and longitude, and mapping it onto the GADM second administrative level.

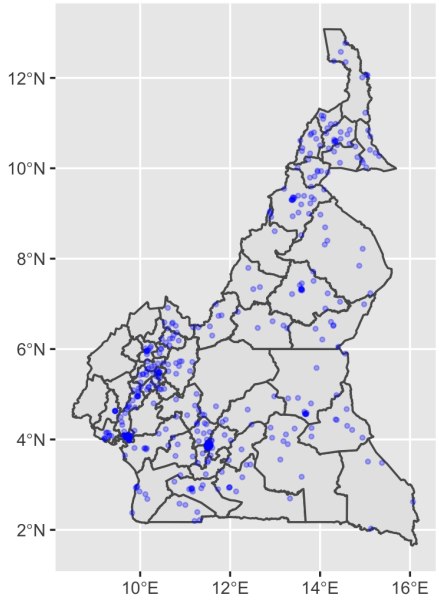

Figure 1. Mapping of clusters onto second administrative regions.

There are 430 clusters and 58 second administrative regions. There are no clusters located in seven regions in the southwest area of the country. This is due to the aforementioned security

concerns. Due to the grouped nature of this missing data, no attempts will be made to model the prevalence in these seven regions. The number of clusters in the remaining 51 regions ranged from 1 to 44. The full distribution of the number of clusters per region can be found in the Supplementary Materials. The regions with the greatest number of clusters correspond to the cities of Douala and Yaounde.

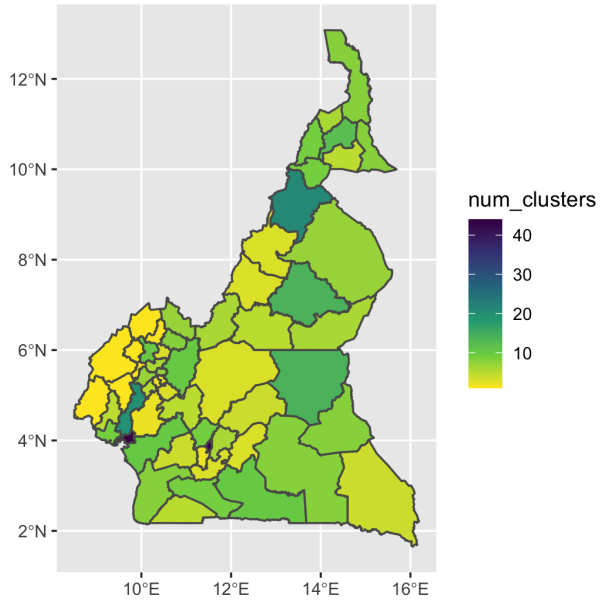

Figure 2. Number of clusters by second administrative region.

The spatially independent estimates were generated and mapped along with their corresponding standard errors. The unweighted values map can be found in the Supplementary Materials, and the weighted values are shown below:

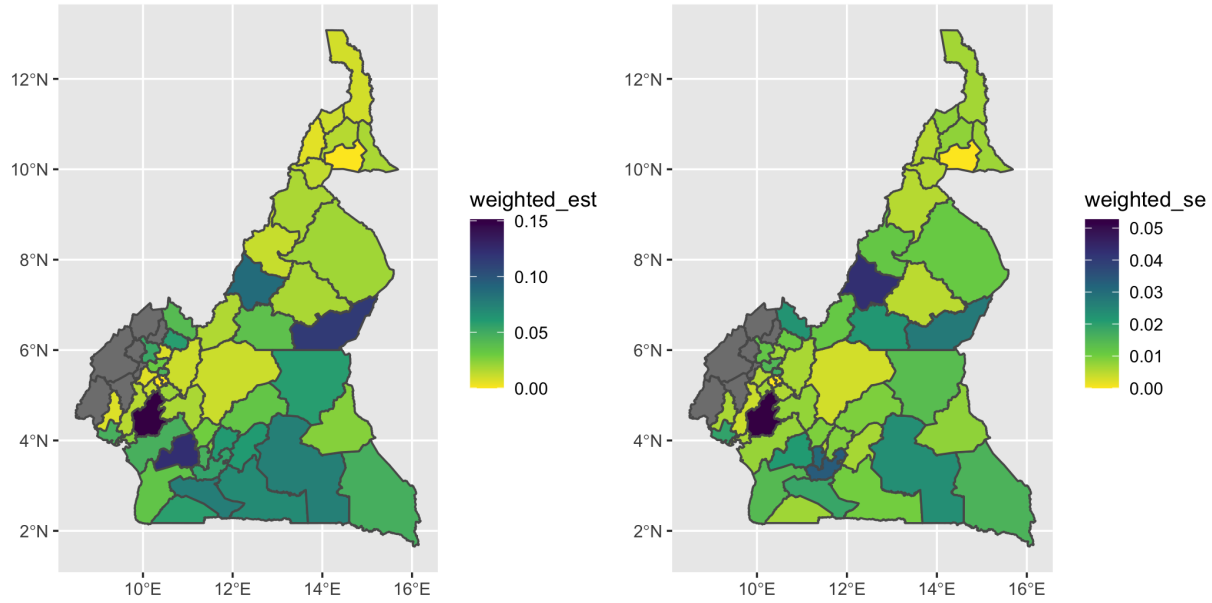

Figure 3. Weighted estimates (left) and corresponding standard errors (right).

The simple (unweighted) and direct (weighted) estimates were then compared to one another.

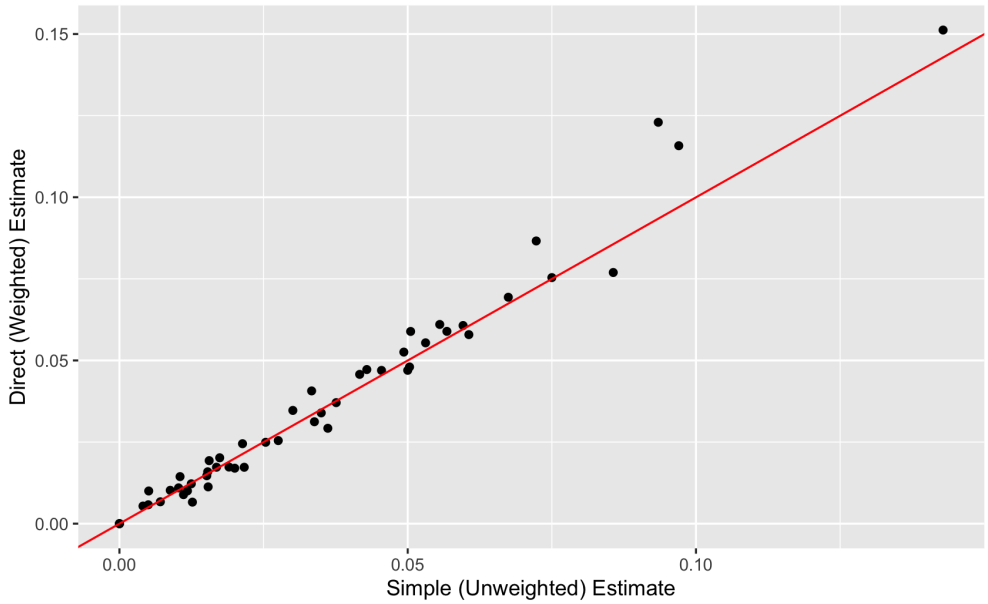

Figure 4. Comparison of weighted and unweighted estimates for HIV prevalence.

The next method to incorporate was the smoothed weighted estimates of the Fay-Herriot model. The posterior median estimates and corresponding posterior standard deviations are below:

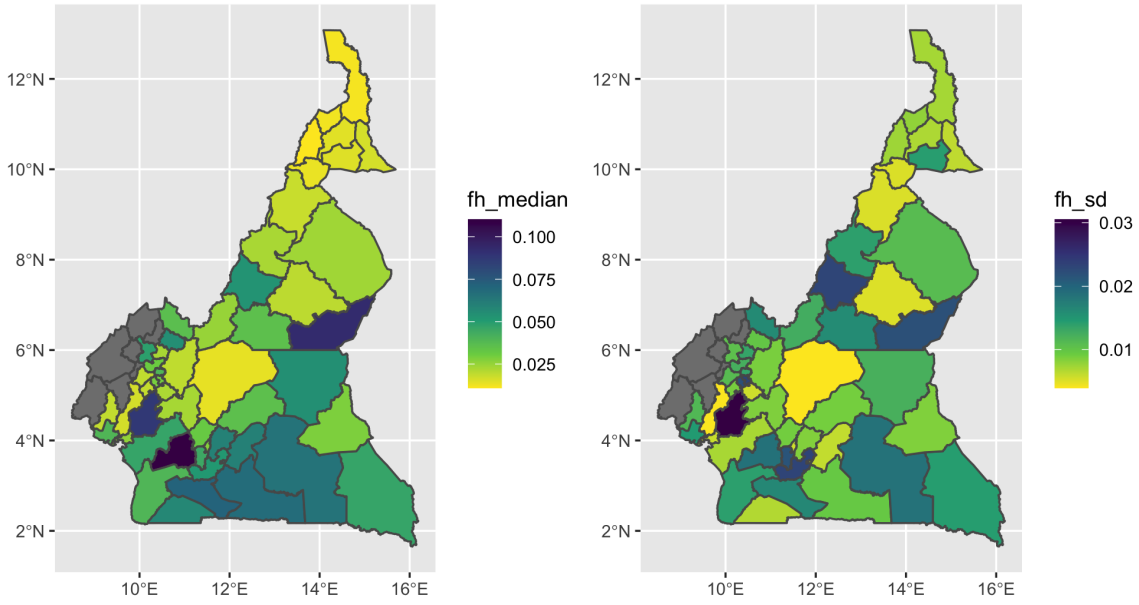

Figure 5. Posterior medians (left) and posterior standard deviations (right) from the FH model.

These FH estimates were compared to the spatially-independent weighted estimates to determine the impact of the applied spatial smoothing. A similar comparison was made for the standard errors of the two methods.

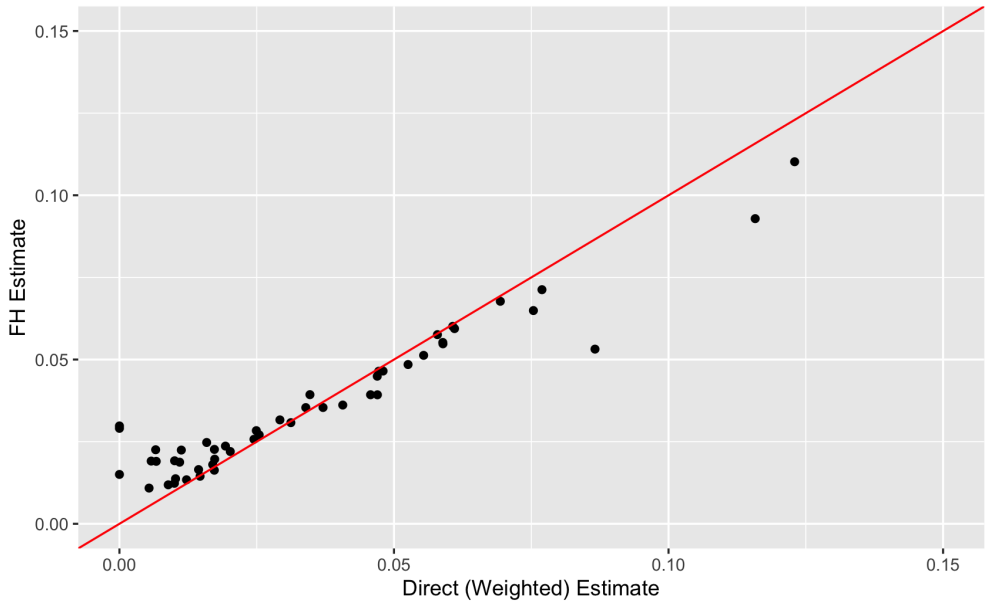

Figure 6. Comparison of Fay-Herriot posterior median estimates to weighted estimates.

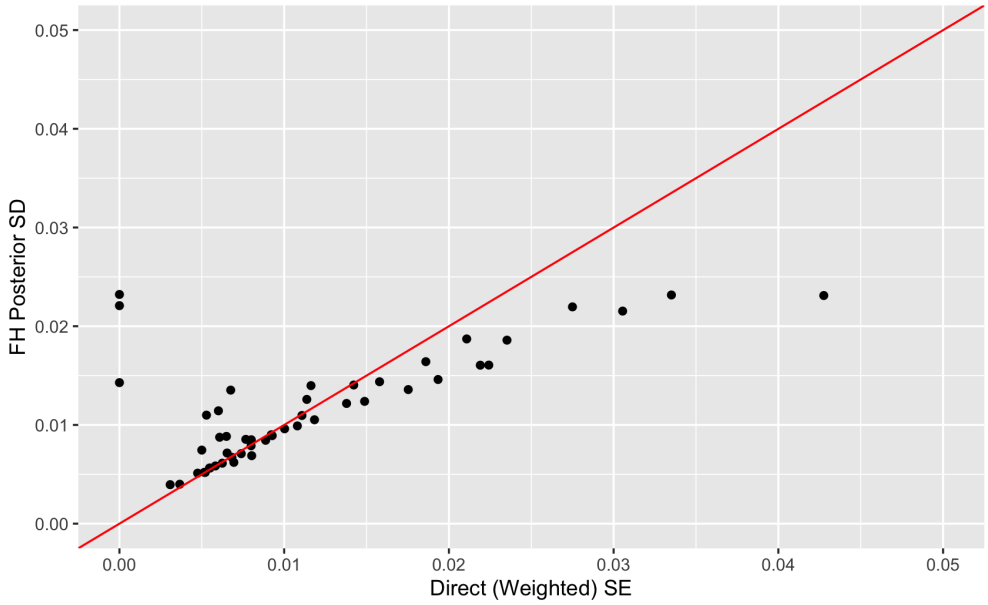

Figure 7. Comparison of Fay-Herriot posterior standard deviations to weighted standard errors.

The Fay-Herriot method was then expanded to include a covariate. This covariate was the proportion from each region that answered affirmatively to the survey question "Ever used anything or tried to delay or avoid getting pregnant?". The map of the covariate values can be found in the Supplementary Materials. The plot of the weighted estimates against the covariate values along with a smoother are shown in the graph below:

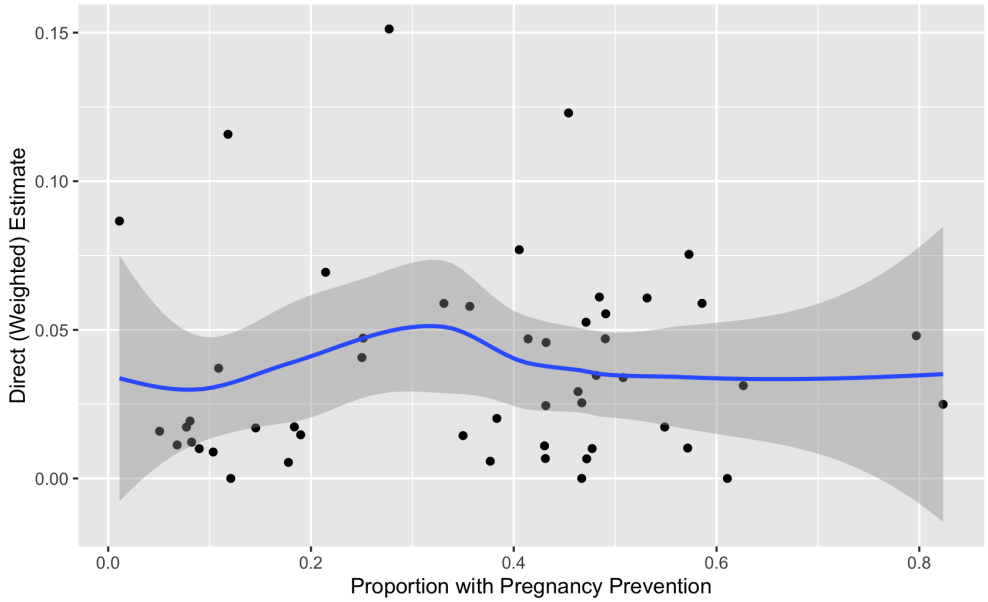

Figure 8. Weighted estimates versus proportion of respondents that answered affirmatively to the survey question "Ever used anything or tried to delay or avoid getting pregnant?" by region.

The outputs of the Fay-Herriot models with and without a covariate were then compared. One measure to compare is ϕ, the proportion of the total variance attributed to spatial random effects. In the model without a covariate included, the posterior median value for this parameter is 0.305, indicating that 30.5% of the total variance is attributed to the spatial random effect. In the FH model with a covariate, the corresponding value is 0.412, thus the spatial random effect accounts for a larger proportion of the variance in the model with a covariate. The comparison of the posterior median values is below:

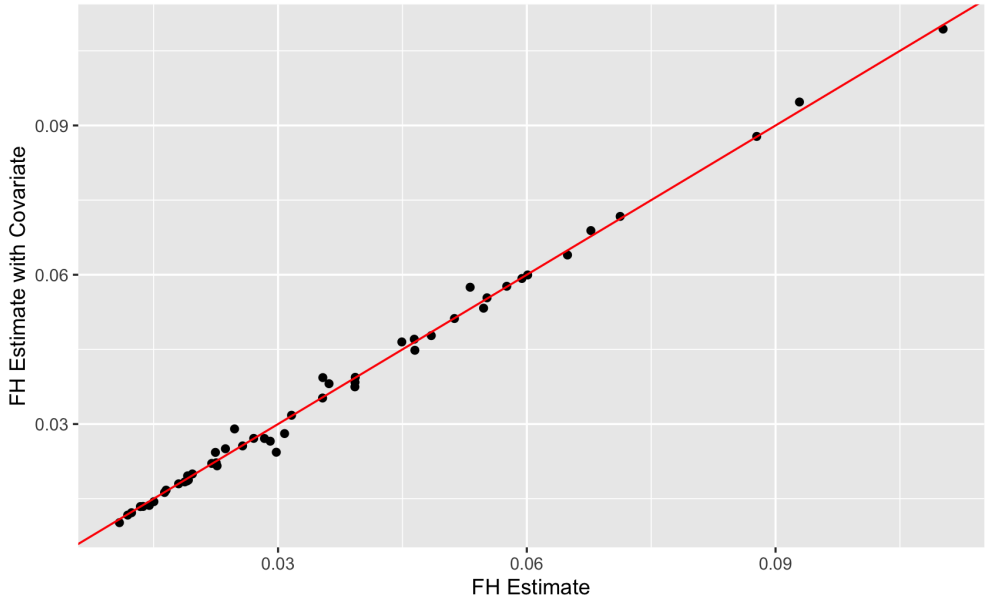

Figure 9. Comparison of Fay-Herriot estimates with covariate to those without a covariate.

Discussion

The methods applied in this analysis attempt to characterize the prevalence of HIV in the 51 second administrative regions that contain clusters. The weighted estimates and corresponding standard errors in Figure 3 show great variability between regions. There are three regions, Hauts Plateaux, Koung Khi, and Mayo Kani, which have a weighted prevalence of zero. The region of Nkam has the highest prevalence at 15.1%. The areas with higher prevalence tend to be concentrated in the southern portion of the country. The standard errors of the estimates range from 0 to 0.0527, and this range is expected given the relationship between prevalence and standard error. The comparison of unweighted and weighted estimates in Figure 4 shows little difference between the two methods. One noteworthy finding is that the three regions with greatest unweighted prevalence had even higher estimates in the weighted metric.

The Fay-Herriot estimated values and corresponding posterior standard deviations were less extreme relative to the weighted estimates, as shown in Figure 5. The range of posterior median estimates was 1.09% to 11.0%, and the maximum posterior standard deviation was 0.0306. The comparison of weighted and Fay-Herriot estimates in Figure 6 shows the effect of spatial smoothing. Low weighted estimates are increased in the Fay-Herriot model, while higher ones are decreased. In Figure 7, it can be seen that the regions with largest weighted standard error have lower posterior standard deviations in the FH model. The model was then extended to incorporate a covariate based on the proportion of individuals in a region that answered affirmatively to the question "Ever used anything or tried to delay or avoid getting pregnant?". Figure 8 shows that the association between this covariate and the weighted prevalence metric is not strong, and this is confirmed by the minimal difference in estimates contained in Figure 9. Despite intuition that greater utilization of pregnancy prevention methods would lead to lower HIV rates, this variable appears to provide little value to our estimation.

There are several areas in which our analysis is limited. The mapping from cluster to second administrative region is essential to the methods applied in this analysis. The details of this can be seen in Figure 1 and Figure 2. The assumption of this method is that the clusters are representative of the region in which they lie. There are seven regions with only one cluster, so one possible limitation of this analysis is that these clusters do not represent the entirety of their region. Another assumption was made regarding the second administrative level covariate. The responses used to obtain covariate values were based on all survey respondents in the region, not just those that were tested for HIV. If there was a discrepancy in covariate values between these two populations, it could have led to differing performance in the model fit with a covariate. In spite of these limitations, the methods generated insight and greater understanding of HIV prevalence in Cameroon. These estimates can be used by public health entities for the purpose of targeting preventative intervention.

References

1. *About HIV/AIDS*. (2021, June 1). Centers for Disease Control and Prevention. Retrieved March 15, 2022, from https://www.cdc.gov/hiv/basics/whatishiv.html#:~:text=What%20is%20HIV%3F-,HIV%20(human%20immunodeficiency%20virus)%20is%20a%20virus%20that%20attacks%20the,care%2C%20HIV%20can%20be%20controlled.

2. *Facts About HIV: Life Expectancy and Long-Term Outlook*. (2018, April 27). Healthline. Retrieved March 15, 2022, from https://www.healthline.com/health/hiv-aids/life-expectancy#:~:text=In%201996%2C%20the%20total%20life.

3. Mbanya, D., Sama, M., & Tchounwou, P. (2008). Current Status of HIV/AIDS in Cameroon: How Effective are Control Strategies? *International Journal of Environmental Research and Public Health*, 5(5), 378–383. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3699997.

4. *Cameroon Population (2019) - Worldometers*. (2019). Worldometers.info. Retrieved March 15, 2022, from https://www.worldometers.info/world-population/cameroon-population.

5. *Cameroon*. (2016). Unaids.org. Retrieved March 15, 2022, from https://www.unaids.org/en/regionscountries/countries/cameroon.

6. *National HIV Survey in Cameroon Shows Advances in Treatment and Remaining Challenges to Achieving Epidemic Control*. (2021, April 29). PHIA Project. Retrieved March 15, 2022, from https://phia.icap.columbia.edu/national-hiv-survey-in-cameroon-shows-advances-in-treatment-and-remaining-challenges-to-achieving-epidemic-control/

7. *2018 Demographic and Health Survey Summary Report Cameroon.* (2019). Retrieved March 15, 2022, from https://dhsprogram.com/pubs/pdf/SR266/SR266.pdf.

8. *Analyzing DHS Data*. (n.d.). Dhsprogram.com. Retrieved March 15, 2022, from https://dhsprogram.com/data/Guide-to-DHS-Statistics/Analyzing_DHS_Data.htm.

9. *The DHS Program - DHS Methodology*. (n.d.). Dhsprogram.com. Retrieved March 15, 2022, from https://dhsprogram.com/Methodology/Survey-Types/DHS-Methodology.cfm.