Krista Steuben
March 14, 2018
555 – Statistical Methods
for Spatial Epidemiology

Final Project:

Spatial Variation of HIV Prevalence in Zimbabwe in 2015

Introduction:

The HIV epidemic is one of the most pressing concerns in public health. According to The Global Burden of Disease project, HIV contributes 2.4% of global DALYs (disability-adjusted life years) (http://ihmeuw.org/4d1w). Previous studies have shown that there can be significant variation in the prevalence of HIV within a country and that identifying areas where burden of HIV infection is concentrated may help identify at-risk populations and ultimately allow optimal resource allocation to occur within countries.

Certain HIV interventions may affect various age groups differently. Assume resources for a certain intervention are allocated based on total population prevalence trends. If the prevalence trends for the intended age group for that intervention differ from the total population trends, then the resources may not be optimally allocated. It may be worthwhile to look into the spatial variation for different age groups when allocating resources.

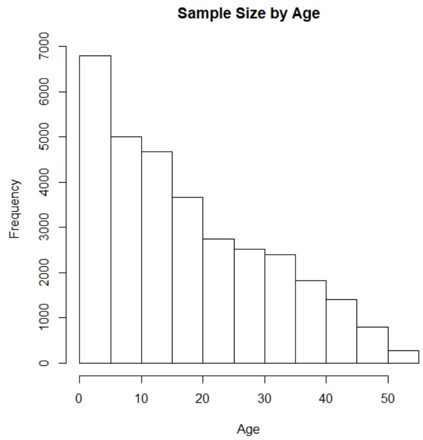
The aim of this project is to map HIV prevalence in Zimbabwe in 2015 across different age groups using small area estimation methods. An analysis of these maps will look into the spatial variation of the prevalence of HIV for various age groups (0-14, 15-24, 25-34, and 34-49 years old) and compare that to the spatial variation of HIV prevalence for the total population (0-54 years old).

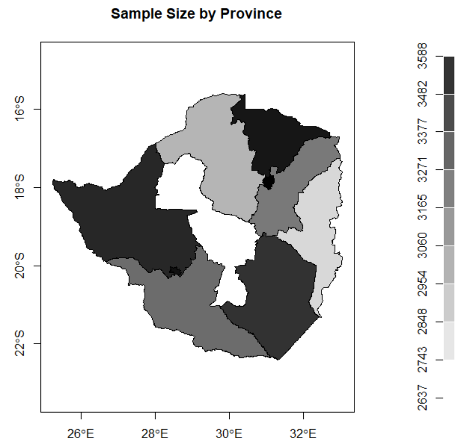
Data Description:

The data for this analysis comes from the Zimbabwe Demographic and Health Survey from 2015. Access to the microdata of this survey is found at the following website: https://dhsprogram.com/data/dataset/Zimbabwe_Standard-DHS_2015.cfm?flag=0.

In order to prep the data for analysis, the microdata from the household members recode was merged onto the HIV recode according to a unique identifier. Variables not needed for the analysis were dropped leaving only survey design, weights, age, sex, and HIV test result variables. Rows in the dataset where the HIV test result was indeterminate, inconclusive or missing were dropped. As prompted by DHS, the weights were set to the household weight for individuals under the age of 15 and to the HIV weight for individuals 15 years or older.

The resulting data set contained 32,112 observations over 10 provinces.





Note that the sample sizes monotonically decrease with age and vary by about 30% between the most and least represented provinces.

Methods:

Modeling was done through small area estimation techniques with the HIV test result (hiv03) as the binary outcome of interest and the 10 provinces of Zimbabwe as the small areas of interest. Two models were considered in this analysis: Model A – weighted and unsmoothed, and Model B – weighted and spatially smoothed.

Only weighted models were considered in order to capture the sampling design of the DHS survey.

Model A:

Model A is a weighted and unsmoothed model. The small area estimator of HIV prevalence in province i is $\widehat{p}_i$ (the survey-weighted estimated of the prevalence in area i) with a standard error of $\widehat{SE}(\widehat{p_i})$. The estimator and the standard errors are both calculated with the svydesign and svyby functions of the survey package in R.

The maps of the estimator of prevalence and standard errors by province were created using the mapvariable function from the SpatialEpi package in R.

Model B:

Model B is the weighted and spatially smoothed Bayesian design-based hierarchical model. The weighted estimator of prevalence is

$$\widehat{\theta_i} = logit\ \widehat{p}_i \sim N(\theta_i, V_i)$$

where $\widehat{V_i} = \frac{var(\widehat{p_i})}{\widehat{p_i}^2(1-\widehat{p_i})^2}$ is assumed known and θ_i is estimated through the prior model.

The prior model is $\theta_i = \beta + \epsilon_i + S_i$ where $\epsilon_i \sim_{iid} N(0, \sigma_\epsilon^2)$ are the random effects, $S_i|S_j, j \in ne(i) \sim N(\overline{S_j}, \frac{\sigma_s^2}{m_i})$ are the spatial effects, and $\beta, \sigma_\epsilon^2, and\ \sigma_s^2$ are given the default priors in INLA.

This model was implemented using the INLA package for R. The .graph file indicating the neighbors of each province to be used for the spatial effects was created using the nb2INLA function. Maps were created using the mapvariable function from the SpatialEpi package in R.

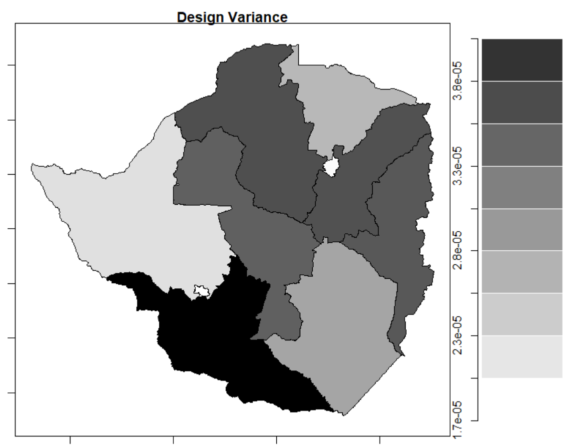
Age Breakdowns:

Maps for each age group was created by subsetting the dataset to only people within a given age range and then implementing Model A and Model B as above.

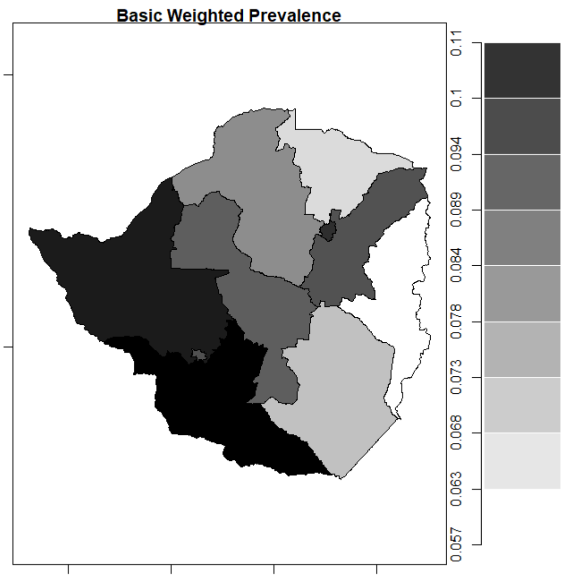
Results:

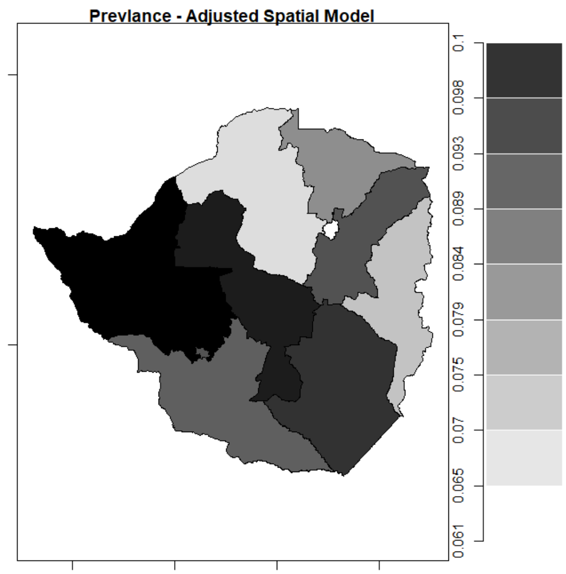
Model Selection:

While it is clear that including the proper weights and designs is necessary to adjust for the sampling design, it is unclear if spatial smoothing is necessary. The map below of the design variances from Model A show that while there is some difference in design variance by province, the variance is small for each province in Zimbabwe. This suggests that the non-smoothed model may be adequate.



When compared to an unsmoothed model, spatial smoothing models generally reduce variability but may introduce bias. A comparison of the prevalence maps of the models can aid in determining if smoothing is beneficial. There are two key differences between the maps to consider.



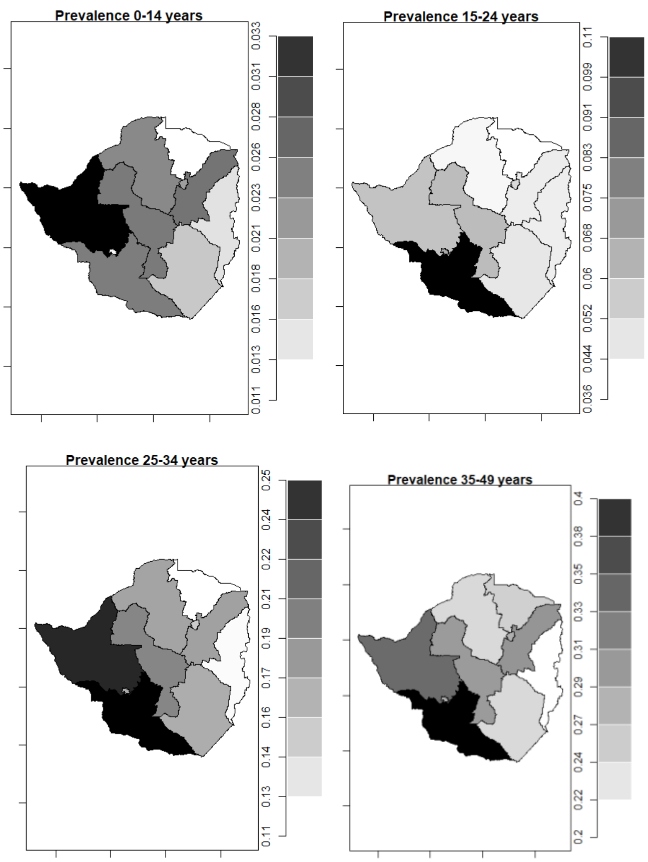


One key difference is that compared to the surrounding areas, the cities Bulawayo and Harare have higher prevalence in Model A than in Model B. According to antenatal care clinic statistics, HIV prevalence tends to be higher in urban compared to rural areas of Sub-Saharan Africa (https://www.sida.se/contentassets/6eb2fb5838cf4325b8d41ff3e58ab50a/hiv-and-aids-in-urban-settings-in-sub-saharan-africa_1049.pdf). Since HIV is spread by certain person-to-person interaction, areas where people have more interaction will tend to have higher prevalence. This provides some evidence that Model A may be a more appropriate model for this data.

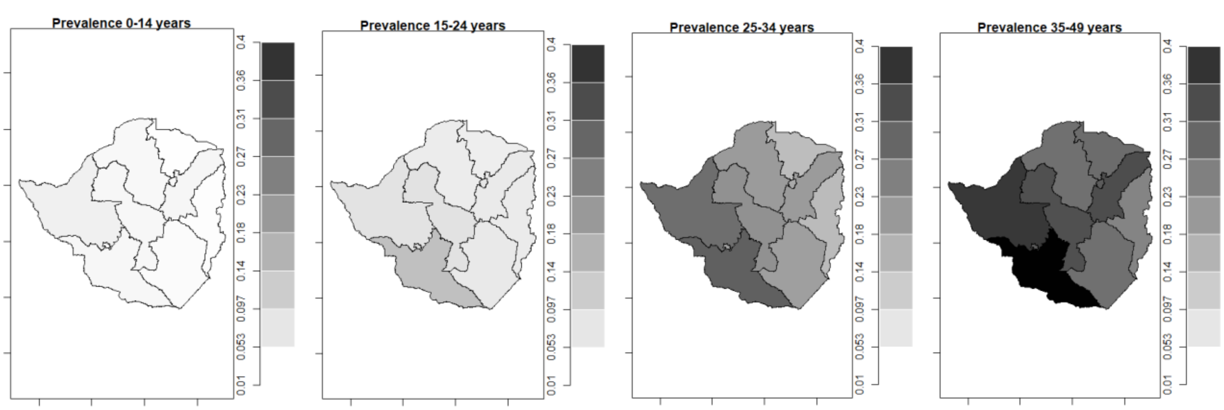
The two models also differ in their spatial trends. High prevalence is more concentrated in the south and west of the country in Model A. By contrast, in Model B, the higher prevalence is pulled towards the center of the country. Based on the Global Burden of Disease Compare visualization (http://ihmeuw.org/4d1q), Botswana and South Africa (the countries bordering the south and west of Zimbabwe) had higher prevalence than Mozambique and Zambia (the countries bordering the north and east) in 2015. Assuming that there is enough migration across the borders of these countries to influence HIV prevalence, one would expect the provinces along the southern and western borders to have relatively high prevalence rates. In the spatial smoothing model, cases along the edge of the country will automatically smooth in towards the center of the country. This suggests that the basic weighted model, Model A, is likely a better representation of the true HIV prevalence trends within the country.

A comparison of the prevalence maps from the two models against prior knowledge of the HIV epidemic suggests that the basic weighted prevalence model is stronger than the spatial smoothing model. As such, Model A will be used for further analysis.

Age Breakdown:



The maps of the age breakdowns shows that while there are some commonalities of spatial variation among age groupings, there are also some notable differences. In general, the highest prevalence tends to be in the south-west and the lowest prevalence in the north east for all age groupings. The highest prevalence province is different for 0-14 year olds from the other age groups. In addition, the 15-24 age grouping map has the most striking difference between the province with the highest prevalence and the remaining provinces.



When the maps of the prevalence of different age groupings are mapped with the same color scale, the trend of higher prevalence with higher age becomes apparent. This is not a surprise though, since HIV is not curable.

Discussion:

There are some limitations to this analysis. First, performing the analysis at a finer level than provinces would be more informative of the spatial variation of HIV prevalence. The spatial smoothing models may be improved by this refinement, especially since many of the pairs of provinces share a border. Second, the decision to choose the weighted model was infuenced by general knowledge of the HIV epidemic that may not be as applicable in Zimbabwe. There was also an assumption that the prevalence of neighboring countries could inform the prevalence in areas near the borders. This assumption could be validated or invalidated by doing a wider analysis involving multiple countries. Finally, there are inherent issues with HIV testing data. The stigma around HIV may lead people to refuse to participate in the study. There are also errors collecting and processing the blood samples. Also, the expense of performing these blood tests leads to small sample sizes. Especially if performing this analysis at a smaller scale, having larger sample sizes can lead to less biased and varied results.

To conclude there appears there was spatial variation of HIV prevalence within Zimbabwe in 2015. The south west of the country has the highest prevalence, and this trend is present in all of the age breakdowns studied. HIV prevalence is higher for older age groups, but the spatial trend is fairly similar among the different age groups studied. The age group that appears to have the most unique spatial trend is the 15-24 year old age group. If an intervention was to be implemented that focuses on young adults, it would be worthwhile to consider the spatial trends of HIV prevalence 15-24 year olds rather than just the spatial trends of the total population when allocating resources.

An interesting next step would be to to perform an allocation analysis in order to quantify the improvements of considering differences in spatial trends among different age groups. It would also be interesting to investigate the spatial variation of HIV prevalence among different age groups for other heavily burdened countries to see if similar differences appear.