A Introductory Short Course on:

Small Area Estimation of Health Indicators

Samuel Manda
Department of Statistics, University of Pretoria
South Africa

Presented at:

Strengthening Biostatistics and Data Science Education and Research in Uganda Meeting IBS Uganda Research Meeting

Kampala, Uganda

12/05/25-15/05/25

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

This presentation

- Introduction to Small Area Estimation (SAE) methods
  - ❑ Example Applications
    - ➢ Estimation of voluntary medical male circumcision (VMMC) coverage for each one of the 52 districts in South Africa:
      - ❖Funded by Centre for Disease Control (USA)
      - ❖Survey: 5th South African National HIV Prevalence, Incidence, Behaviour and Communication Survey (SABSSM)
        - ➢ Estimation of high impact child health interventions for each of the 32 districts in Malawi before and after the Integrated Health Systems Strengthening (IHSS) program.
      - ❖Funded by UNICEF
      - ❖Surveys: 2004 and 2015 Malawi Demographic and Health Surveys

- Multivariate Small Area Estimation Methods
  - Example application to UNAIDS HIV "95-95-95" targets in South Africa

Small Area Estimation

- Demand for local level estimates of the many health indicators are needed for assessing international development frameworks such as the Substanable Development Goals (SDGs).
- The SDGs calls for disaggregated data based on income class, gender, ethnicity, geographic location, migration status, disability status
  - Reliable small area statistics but sample sizes are too small to provide direct (or area specific) estimators with acceptable accuracy.
- Thus could achived by Censuses and administrative records, but have limited scope.
- Health surveys often powered to provide enough samples of the population at the region or country levels
  - Not at lower administrative level needed for decision making.
  - Estimates could be imprecise and unstable (large variances) due the small numbers accrued at these levels.
  - Reliable estimates have been shown to be highly associated with number of observations falling in these lower levels
- Examples of small areas: county, municipality, three digit occupation counts within a province, health regions, even a state by age-sex-race groups.
- Domain or subpopulation is called a small area if the domain-specific sample size is small.
- Borrow strength from related areas through linking models based on auxiliary data such as recent census and administrative records. This leads to indirect estimators.
- Parameters of interest: Area means, totals, proportions and quantiles. Complex measures: Poverty indicators

Possible solutions

- Design surveys to be representative at the desired geographic level.
  - Very costly as it will require large sample size -> need to have sufficient sample size for each small area to be estimated
- Systematic evaluation and validation of several spatial models for generating robust estimates at small area level (UNAIDS HIV work)

- Several statistical techniques that are concerned with the estimation of efficient estimates for small areas with small sample sizes.
  - These statistical procedures are collectively termed Small Area Estimation (SAE) methods
  - Leverage available survey data to provide local estimates of health indicators
    - Survey data are complemented with auxiliary data (e.g. census, administrative records) to produce small area estimates.

Small Area Estimation (SAE)

- SAE is any of several statistical techniques involving the estimation of parameters for small sub-populations or areas.
- The term "small area" refers to a small geographical area, for example, a district in our cases.
- SAE methods used when the sample size within any district is small to generate accurate health estimates.
- Additional data from census records, administrative and routine information, and other national surveys that exist for these districts were used to obtain estimates.
  - These data include district-level age, education level obtained, religion, literacy levels, geography (provincial and district level), ethnicity and deprivation (poverty) levels.

Survey set up

- Suppose the interest is to estimate a finite population parameter ϑ_i for small area i $(i = 1, 2, \ldots, N)$ where N is the number of small areas each of size N_i.

- Also let n_i be the sample size in area i.

- Let R_{ij} is binary (1/0) for subject i being sampled or not

- Assume we are conducting a prevalence survey, then ϑ_i is the small area proportion/prevalence/coverage
  - So outcome y_{ij} is binary (1/0); e.g. HIV yes or no; VMMC yes or no

Direct Survey Estimator

- Aim of survey: To estimate the small area population event count, $y_i = \sum_j^{N_i} y_{ij}$ or the small area proportion $\vartheta_i = {}^{y_i}/_{N_i}$ in area area i $(i = 1, 2, \dots, N)$

- Using the sample data, we calculate ϑ_i^{DIR} as a direct design-unbiased estimator
  - For example. The area prevalence/coverage estimate given by the Horvitz-Thompson (Horvitz and Thompson, 1952; Sarndal et al., 1992):

$$\vartheta_i^{DIR} = \frac{\sum_{j=1}^{N_i} R_{ij}\, w_{ij} y_{ij}}{\sum_{j=1}^{N_i} R_{ij}\, w_{ij}} = \frac{1}{n_i} \sum_{j=1}^{N_i} R_{ij}\, w_{ij} y_{ij}$$

where w_{ij} are the normalised weights such that $\sum_{j=1}^{N_i} R_{ij}\, w_{ij} = n_i$. The variance of ϑ_i^{DIR} is given by

$$\widehat{var}\left(\vartheta_i^{DIR}\right) = \frac{1}{n_i(n_i - 1)} \left(\frac{N_i - n_i}{N_i}\right) \sum_{j=1}^{N_i} R_{ij} w_{ij}^2 \, (y_{ij} - \vartheta_i^{DIR})^2$$

- Under simple random sampling (SRS), the direct estimator is $\vartheta_i^{DIR(s)} = {}^{y_{is}}/_{n_i}$ where $y_{is} = \sum_j^{n_i} y_{ij}$ with $variance\left(\vartheta_i^{DIR(s)}\right) = {}^{\vartheta_i^{DIR(s)}(1 - \vartheta_i^{DIR(s)})}/_{n_i}$.

Small Estimation

- Under the Fay–Herriot (FH) model, we have three linking models, namely

1) Sampling model (usually on a transformation of θ_i^{DIR})

$$f_i\left(\vartheta_i^{DIR}\right) = f_i(\vartheta_i) + \varepsilon_i \; i = 1,2,\dots,N$$

where θ_i and ε_i are unknown area parameter to be estimated and the sampling error of the direct estimator, θ_i^{DIR}, respectively. It is assumed that $\varepsilon_i \sim N(0, \sigma_{ei}^2)$ and σ_{ei}^2 is known.

2) Model linking area parameter to area-level covariates (predictor variables) or auxiliary variables or additional data X_i

$$f_i(\vartheta_i) = X_i^T\beta + \upsilon_i; i = 1,2,\dots,N$$

where β and υ_i are unknown regression parameters to be estimated and random area-specific effect to account for between area variation or unexplained variability between the areas, respectively. It is assumed that $\upsilon_i \sim N(0, \sigma_\upsilon^2)$ and σ_υ^2 is unknown and to be estimated from the data.

3) Linking linear model combining 1) and 2)

$$\theta_i^{DIR} = X_i^T\beta + \upsilon_i + \varepsilon_i; i = 1,2,\dots,N$$

FH method (2)

- Using the Best Linear Unbiased Predictor (BLUP) estimation, the Fay-Herriot model estimator for SAE is

$$\theta_i^{FH} = \gamma_i \theta_i^{DIR} + (1 - \gamma_i) X_i^T \hat{\beta}; i = 1,2, \dots, N$$

$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} \theta, V = \text{Diag}(\sigma_\mu^2 + \sigma_{\varepsilon 1}^2 +, \dots, \sigma_\mu^2 + \sigma_{\varepsilon N}^2)$ where the mixing weight γ_i are given by:

$$\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_{\varepsilon i}^2}$$

which is the proportion of total variance due to between area variation.

- The BLUP is also the best unbiased predictor under assumed normality of θ_i^{DIR} and θ

FH method (3)

- Thus, the FH estimator of a small area is just a weighted average of a direct estimator θ_i^{DIR} and regression estimator given by $X_i^T \hat{\beta} + \hat{v}_i$,
- It moves towards the direct estimate from the survey when the sample sizes are large (or when sampling variance in area i is small; and leans more to the regression synthetic estimator as the between area variance increases.
- The random effect v_i is sometimes modelled with spatial error model or conditional autoregressive (CAR) models to account for spatial dependency.
- In practice, we need to estimate model parameters: Fay and Herriot (1979) method of moments (MM), maximum likelihood (ML) or restricted ML (REML
- Estimation of random effect variance is complex and challenging, as could lead to negative estimates which are truncated to zero.
- Active research area

Method based on estimating counts

- Aim of survey: To estimate the small area population count, $y_i = \sum_j^{N_i} y_{ij}$ or the small area proportion $\vartheta_i = {}^{y_i}/_{N_i}$ in area area i $(i = 1, 2, \dots, N)$
- Ignoring the sampling design, the sample count y_{is} in area i can be assumed to follow a $Binomial(\pi_i, n_i)$. The non-sample count $y_{i(ns)}$is also $Binomial(\pi_i, N_i - n_i)$. The counts are assumed to be independent binomial variables with π_i being a common "positive" probability.
- The model linking the probability π_i with the covariates X_i is the logistic linear mixed model of form $logit(\pi_i) = X_i^T \beta + \upsilon_i$
- A plug-in empirical predictor (EP) of the population vent count y_i is obtained as

$$\hat{y}_i^{EP} = y_{is} + E(y_{i(ns)}) = y_{is} + (N_i - n_i)\left(\exp(X_i^T \beta + \upsilon_i)(1 + \exp(X_i^T \beta + \upsilon_i))^{-1}\right)$$

This, the proportion estimate in area i is given by $\hat{p}_i^{EP} = \hat{y}_i^{EP} / N_i$

South Africa District level estimation of Voluntary Medical Male Circumcision (VMMC)

Samuel Manda, Tarylee Reddy and SAMRC), Khangelani Zuma and HSRC

Background

- South Africa continues to have the highest burden of HIV in the world, with about 7.9 million people living with HIV (PLHIV)
- The HIV transmission is largely heterosexual
- Voluntary medical male circumcision (VMMC) is one of the proven biomedical interventions for the prevention of heterosexual HIV transmission
- Consequently, voluntary medical male circumcision (VMMC) has become an additional HIV prevention strategy in the fight against HIV in the country
- South Africa is one of the 14 priority countries identified by World Health Organization (WHO) and Joint United Nations Programme on HIV/AIDS (UNAIDS) for VMMC campaigns
  - with a set national coverage goal of 80% for VMMC for men ages 15-49
- Hence, there has been mass scale-up of VMMC in the country
- However, there is no there is no comprehensive understanding of local spatial patterns of VMMC coverage.

Rationale for modelling VMMC coverage

- There is growing evidence that male circumcision prevalence may play a role in driving the local variations in the HIV epidemic
- It is therefore important to have a comprehensive understanding of local variation in coverage of VMMC
- Modelling methods that use location linked data might better describe local patterns of VMMC coverage
  - Providing a good understanding of the gaps in male circumcision coverage at local level
- This will help inform the South Africa NSP current focus on doing the "right things, in the right place and at the right time"
  - To maximize impact and efficiency in an era of declining funding for HIV programs.
- Hence the purpose entails integrating and modelling publicly available surveillance data, population data and routine VMMC data for improving district-level estimation of VMMC coverage in South Africa

Aims and objectives

- The aim was to estimate district level estimate of VMMC coverage
  - By applying spatial modelling using the small area estimation (SAE) method developed previously for improving precision of district level estimates of HIV prevalence
  - By projecting VMMC coverage to 2019 which is essential for COP 2020 planning in order to informs impactful allocation of limited VMMC program resources in South Africa
- Specific objectives

Methodology: Data and Sources

Spatial estimation of VMMC coverage at district level

- The modelling used the district, provincial and national geocoded boundary data 2011 census
  - From Statistics South Africa
- 2017 South African National HIV Prevalence, Incidence and Behaviour Survey
  - From Human Sciences Research Council (HSRC), South Africa
- Routine data on VMMC coverage
  - From the South African District Health Information System (DHIS)
- 2017 South Africa Mid-year population estimates and 2011 census
  - From Statistics South Africa
- District-level demographic data from the mid-year population estimates
  - From Statistics South Africa

Modelling was prioritized by age and geographical location

- Age 10-14 years, 15-35 years, 35 years and older
- 52 district including 27 PEFPAR focus districts and 25 additional non-PEPFAR districts

Implemetation

Direct district level estimates were transformed from the original scale to the logit scale to align the models to the normal linear mixed model.

For prevalence of 1 or 0, these were adjusted to 0.5%, 0.25%. 0.1%.

Variance on the logit scale was computed using a delta-method approximation.

Several models, with and without spatial covariance structures, and with different combinations of the covariates were considered.

AIC or Bayesian Information Criteria (BIC), Likelihood Ratio Test (LRT) and Deviance were used to select the best fitting model

Methodology: Model parameters (1)

Variable	Defined as	Time /year	Source	Unit
Population data				
Gridded population data	SA gridded population at 100m grid	2017	World pop data www.worldpop.org	Raster -100m, 1x1 km grid
Male population			2017 mid-year population estimates	
Male population			2011 census data	

Methodology: Model parameters (2)

Variable	Defined as	Time /year	Source	Unit
Covariates : Auxiliary predictors (out of surveys)				
Medical male circumcision	Male circumcision coverage/prevalence	2017/18	District health barometer and Demographic health survey	District
Male circumcision	Percent MC coverage at district level Imprecise estimates at district level	2017/18	Demographic health survey	District
HIV prevalence	District level prevalence	2017/18	Fifth South African national HIV, incidence and behaviour survey, Demographic health survey and Antenatal survey	District

Methodology: Model parameters (3)

Variable	Defined as	Time /year	Source	Unit
Religion	Roman Catholic, Protestant/other Christian denominations and other religious groups	2016/17	Statistics SA, Community Survey	District
Language spoken at home (proxy for ethnicity)	Afrikaans, English, Isindebele, Isixhosa, Isizulu, Sepedi, Sesotho, Setswana, Siswati, Tshivenda, Xitsonga and Other	2016/17	Statistics SA, Community Survey	District
Education	highest educational level- • percent of individuals with secondary or higher education • no education, primary, secondary or higher	2016/17	Statistics SA, Community Survey	District

Methodology: Model parameters (4)

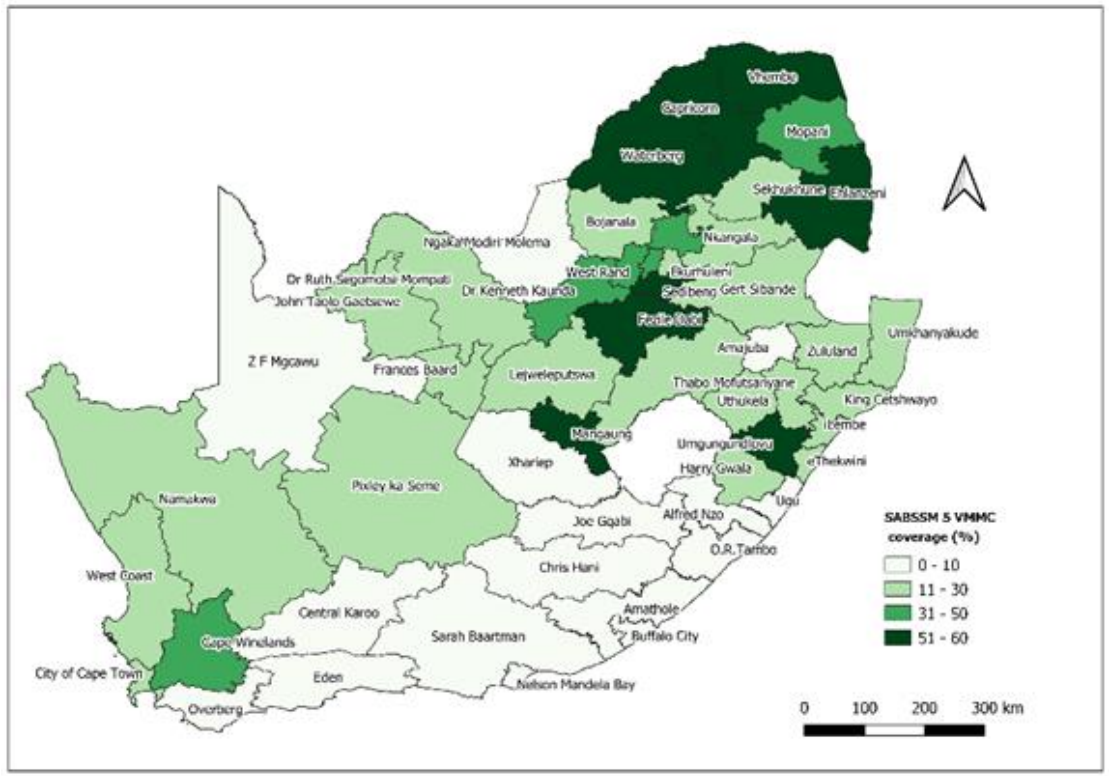
Variable	Defined as	Time /year	Source	Unit
Residence	urban or rural or farm	2016/17	Statistics SA, Community Survey	District
Wealth status/index	Ordinal variable that describes standard of living as determined by material possessions. The resulting asset scores were used to define wealth quintiles index: • Poorest, poorer, middle, richer, richest. • Use index to calculate a poverty variable as the percent of poorest and poorer people	2014	SAMPI SA Multidimensional poverty index (2014)	
Marital status	percentage married vs not married	2016/17	Statistics SA, Community Survey	District

Summary statistics of VMMC coverage (%) by estimation method

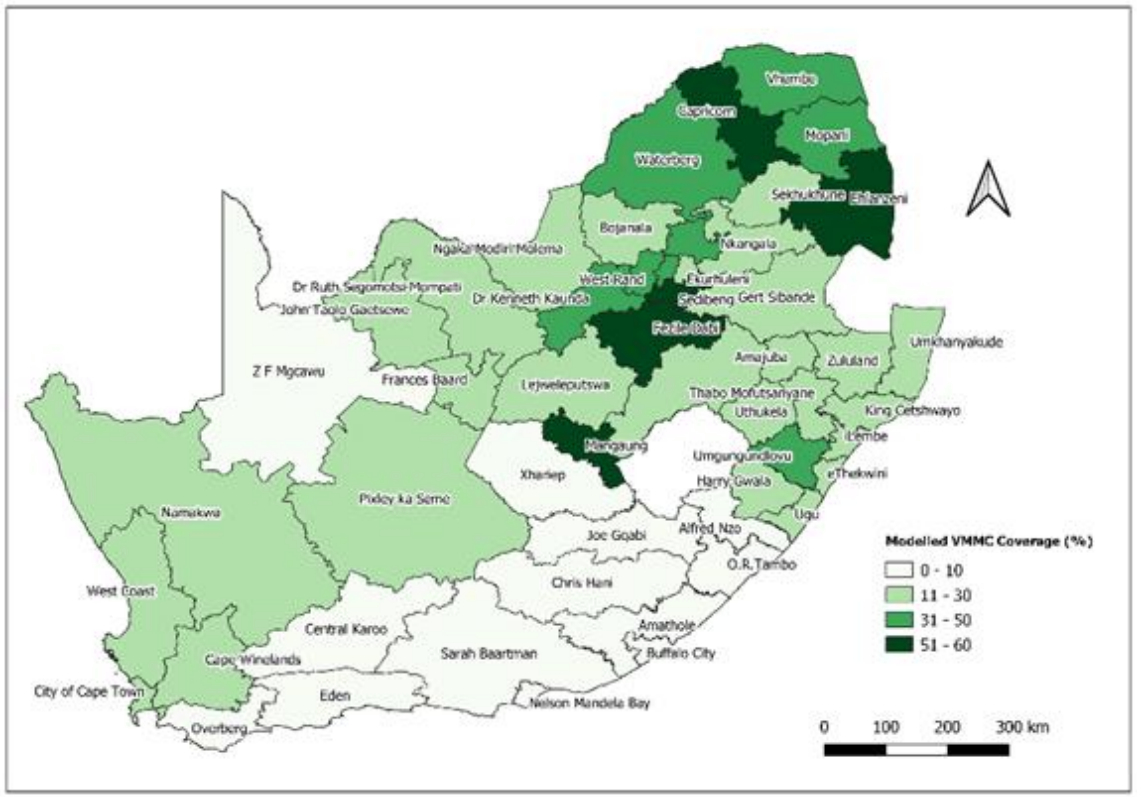
	VMMC coverage					
	10-14 years			≥15 years		
	Mean	Median	IQR	Mean	Median	IQR
Direct survey estimate	21.99%	18.60%	6.70-31.18	25.72%	27.40%	16.70-32.6
Improved estimates	21.34%	19.00%	9.92-28.90	25.29%	26.41%	16.40-34.19
Projected estimates 2019	25.04%	22.98%	12.06-33.21	29.64%	32.67%	17.69-40.04

VMMC coverage 10-14 years:2017

Direct Estimate

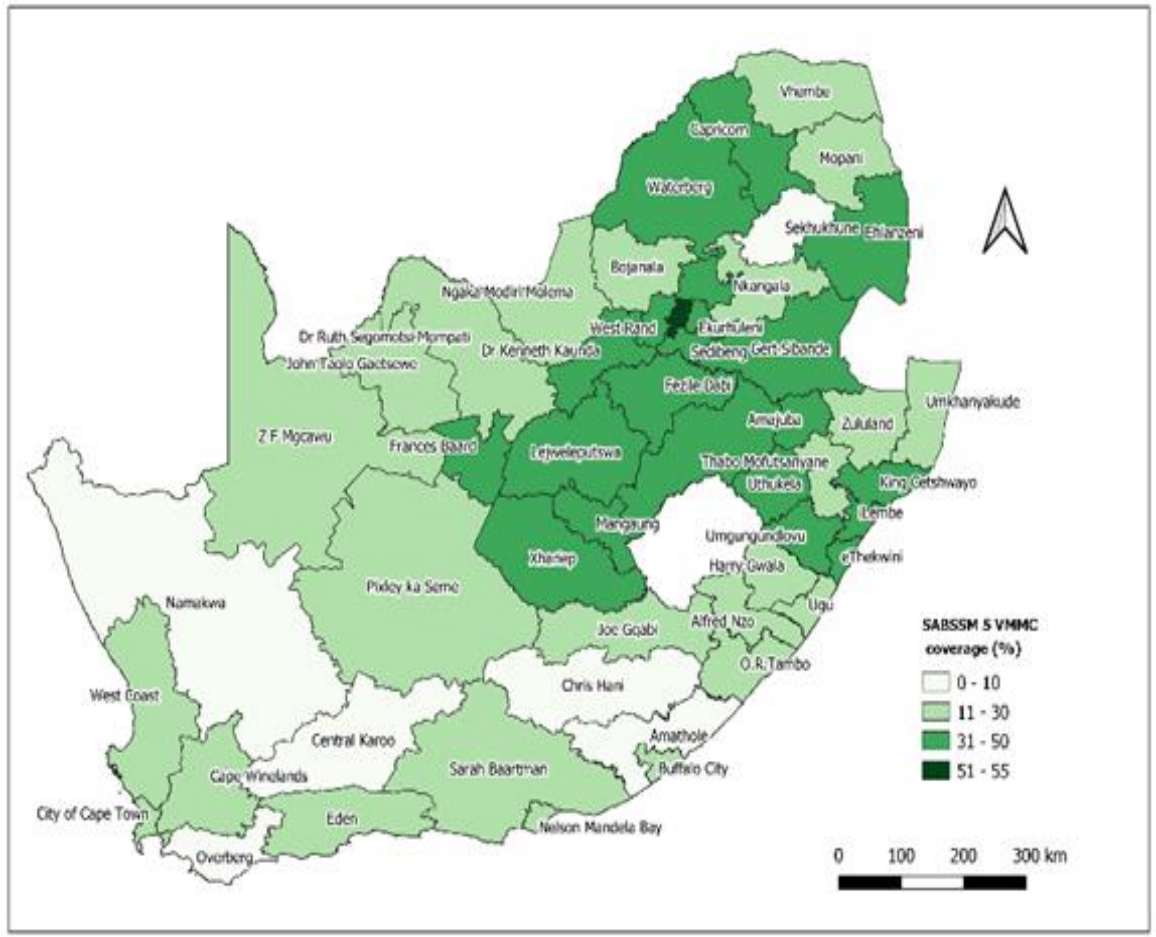


F-H Estimate

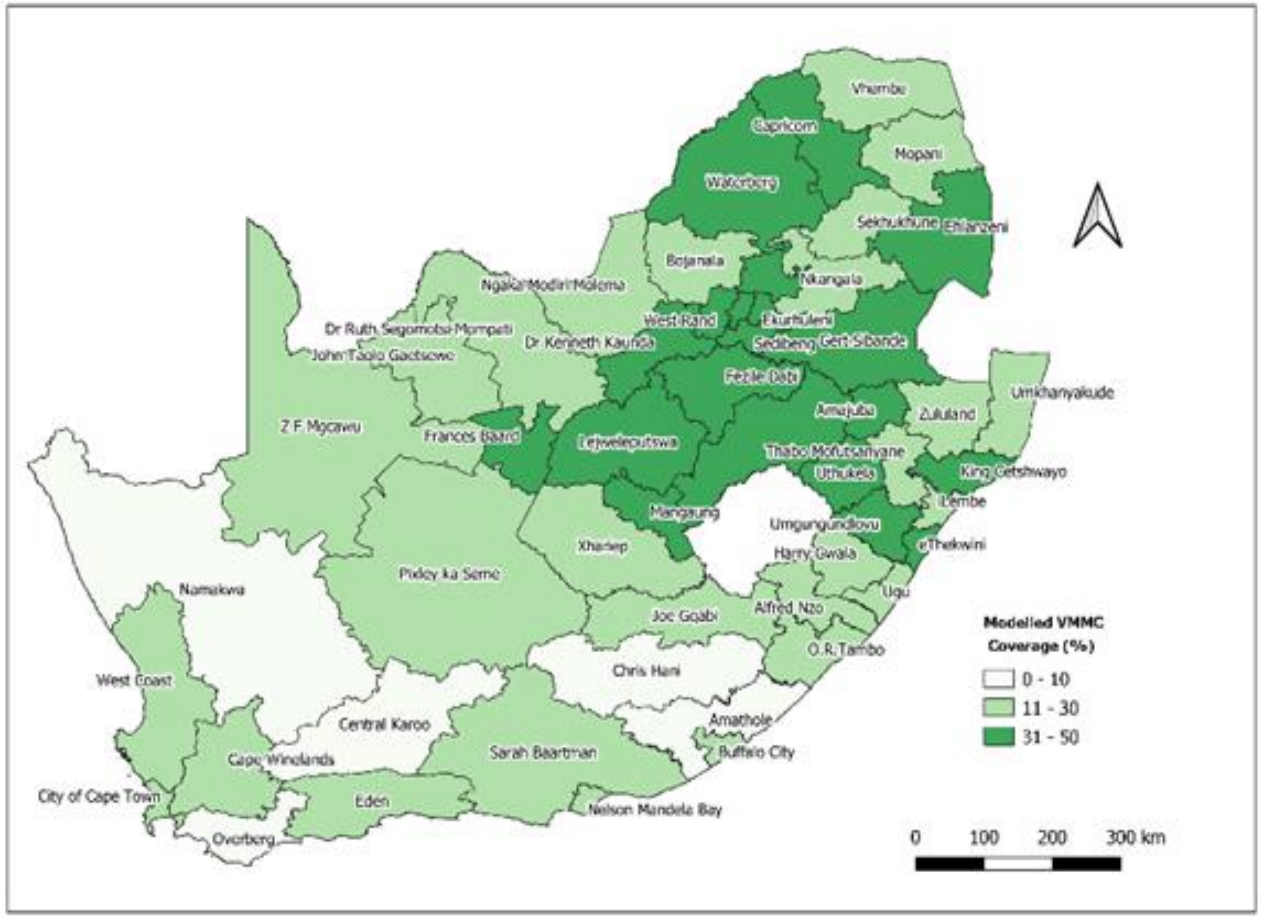


VMMC coverage >=15 years:2017

Direct Estimate



F-H Estimate



Discussion

- Male circumcision is one of the most cost effective and once-off intervention to reduce the spread of HIV from men to women and vice versa.
- The South African Government has put in place plans to achieve a target of at least 80% and 90% voluntary Medical Male Circumcision (VMMC) among men aged 15 years and older and men aged 10-14 years, respectively.
- Obtained reliable estimates of VMMC coverage at the district level to aid the local and national level planning for VMMC interventions
- VMMC coverage rates in 2017 show a high degree of variability across districts in South Africa, in all the age groups considered.
- None of the high VMMC districts exceeded the target of 90% and 80 coverage in 10-14 or 15 and older, respectively.
- These findings highlight the need for targeted VMMC interventions, particularly in the Western Cape, Eastern Cape and Northern Cape.
- VMMC estimates were based on self-report data, which may be subject to bias.

Hierarchical Bayesian FH

- $\theta_i^{DIR}|\theta_i \sim ind\ N(\theta_i, \sigma_{\varepsilon i}^2), \mathrm{i} = 1, \ldots, \mathrm{N}$
- $\theta_i|\beta, \sigma_v^2 \sim ind\ N(X_i^T\beta, \sigma_v^2), \mathrm{i} = 1, \ldots, \mathrm{N}$
- Priors: $f(\beta) \propto 1, f(\sigma_v^2) \sim IG(v_0, \omega_0)$

Obtaining full conditional distribution for the Gibbs sample. For example, the area estimate:

$[\theta_i \left| \theta_i^{DIR}, \beta, \sigma_u^2 \right] \sim N\left(\gamma_i y_i + (1-\gamma_i)X_I'\beta, \gamma_I \sigma_I^2\right)$, where $\gamma_I = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_{\varepsilon i}^2}, i = 1, \ldots, N$;

$$\beta \left| \theta_i^{DIR}, \theta, \sigma_v^2 \right] \sim N_p \begin{pmatrix} (\sum_{i=1} X_i X_i')^{-1} (\sum_{i=1}^N x_i \theta_i \\ \sigma_u^2 (\sum_{i=1} X_i X_i')^{-1} \end{pmatrix},$$

$$[\sigma_V^2 \left| \theta_i^{DIR}, \theta, \beta \right] \sim \mathrm{IG} \begin{pmatrix} v_0 + \frac{N}{2}, \omega_0 \\ + \frac{(\sum_{i=1}^N (\theta_i - X_i'^2 \beta)^2 +}{2} \end{pmatrix}$$

where $i = 1, \ldots, N$;

Multivariate Fay-Herriot model

- Suppose we want to estimate characteristics of R variables in N areas,
- $Let\ y_i = (y_{1_i}, \dots, y_{R_i})^T$, be a direct estimator of $\theta_i = (\theta_{1i}, \cdots, \theta_{Ri})'$. The sampling model

$y_i = \theta_i + e_i, e_i \sim MVN(0, V_{\varepsilon i}\ )$, i= $1, \dots N$

Second component has:

- $\theta_i = X_i\beta_i + u_i, u_i \sim MVN(0, V_u\ )$, i= $1, \dots N$

- Combining the two forms a multivariate linear mixed effect model:

$$y_i = X_i\beta_i + u_i + e_i, \text{ i= } 1, \dots N$$

where u_i and e_i are independent.

- Several options for the random variance matrix
  - Univariate specification
  - Diagonal matrices
  - Autoregressive multivariate (AR(1))
  - Heteroskedastic autoregressive (HAR(1))

Coverage of High Impact Child Health Interventions in Malawi. Integrated Health Systems Strengthening (IHSS) program

Samuel Manda, Tanya Doherty +SAMRC

IHSS program (UNICEF)

- The Catalytic Initiative to Save a Million Lives (CI) was “an international partnership with the goal of strengthening health systems to accelerate progress on the health-related Millennium Development Goals (MDGs).
- The CI sought to “strengthen health systems by delivering life-saving health and nutritional services to disadvantaged children and pregnant women to dramatically reduce child and maternal mortality” in Africa and Asia.
- Department of Foreign Affairs, Trade and Development Canada (DFATD) supports the UNICEF Integrated Health Systems Strengthening (IHSS)
- IHSS) was implemented in six African countries: Ethiopia, Ghana, Malawi, Mali, Mozambique and Niger between 2007 and 2013, was aimed at increasing coverage indicators for high impact child health interventions.
  1) Measles vaccine
  2) Antimalarial treatment
  3) Oral rehydration salts (ORS) for diarrhoea
- Routine childhood vaccination is among the most cost-effective, successful public health interventions available.
- Substantial investments to expand vaccine delivery throughout Africa and strengthen administrative reporting systems
  - Most countries still require robust measures of local routine vaccine coverage and changes in geographical inequalities over time

Aim of this study.

- Estimate of the three child health interventions at subnational level in Malawi before and after the IHSS programme.
- Examine changes in the coverage rates at subnational level in Malawi before and after the IHSS programme.
- Survey data from the 2004 and 2016 Demographic and Health Surveys
- We used multivariate Bayesian F-H model to estimate coverage of the child interventions

Malawi Demographic and Health Survey: Sampled Enumeration Areas pre (2004) and post (2016) IHSS implementation

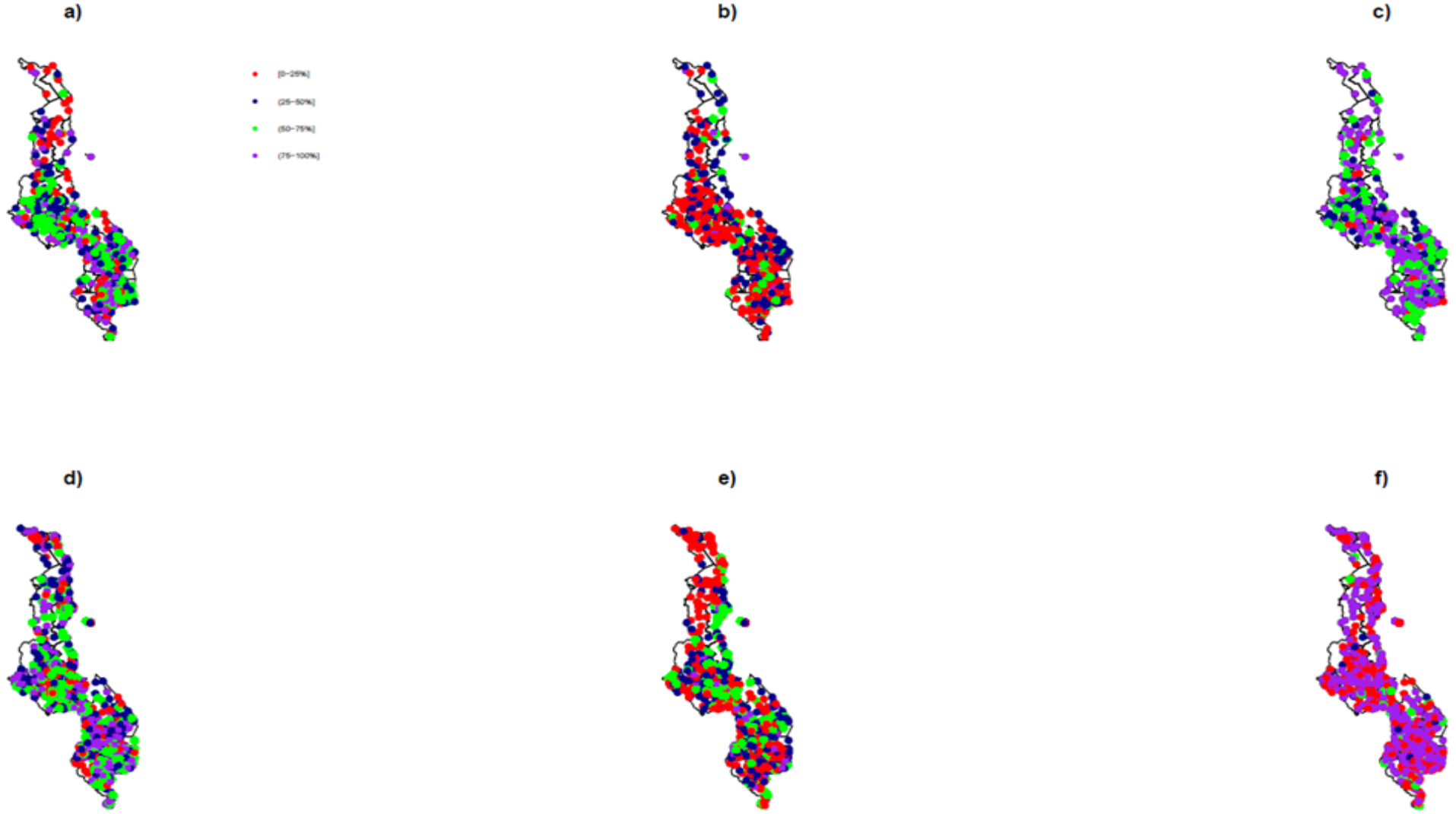
2004

2016

Improved (FH) estimates

Health indicators and coverage	2004		2016	
	Proportion	95% CI	Proportion	95% CI
Prevalence				
Diarrhoea	0.226	0.213;0.240	0.219	0.210;0.229
Malaria	0.378	0.362;0.393	0.292	0.281;0.303
Coverage				
Diarrhoea treatment (ORS)	0.612	0.586;0.636	0.648	0.632;0.663
Antimalarial treatment	0.284	0.264;0.305	0.376	0.354;0.397
Measles Vaccination	0.787	0.764;0.808	0.910	0.894;0.923

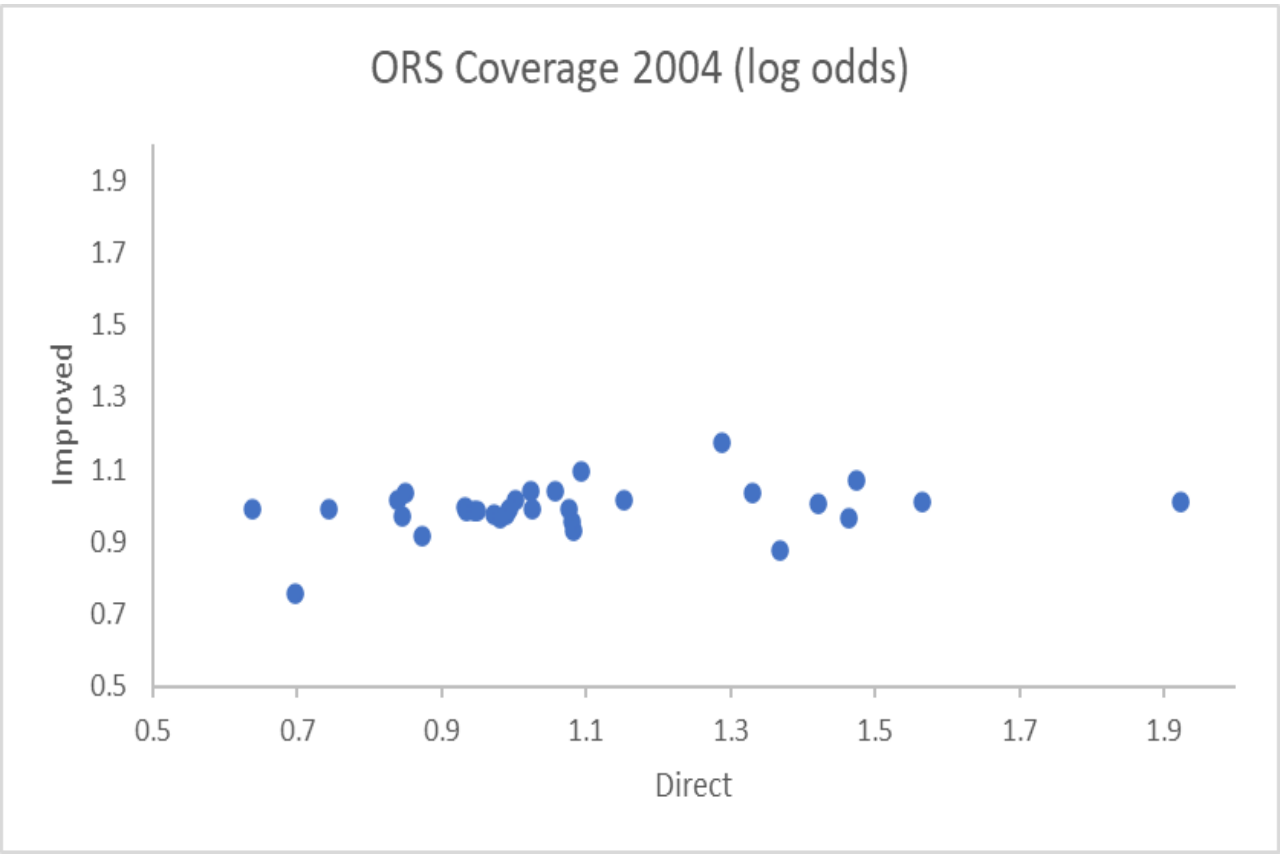
ORS treatment, Malaria treatment, and Measles treatment coverage pre-and post IHSS at the EA level

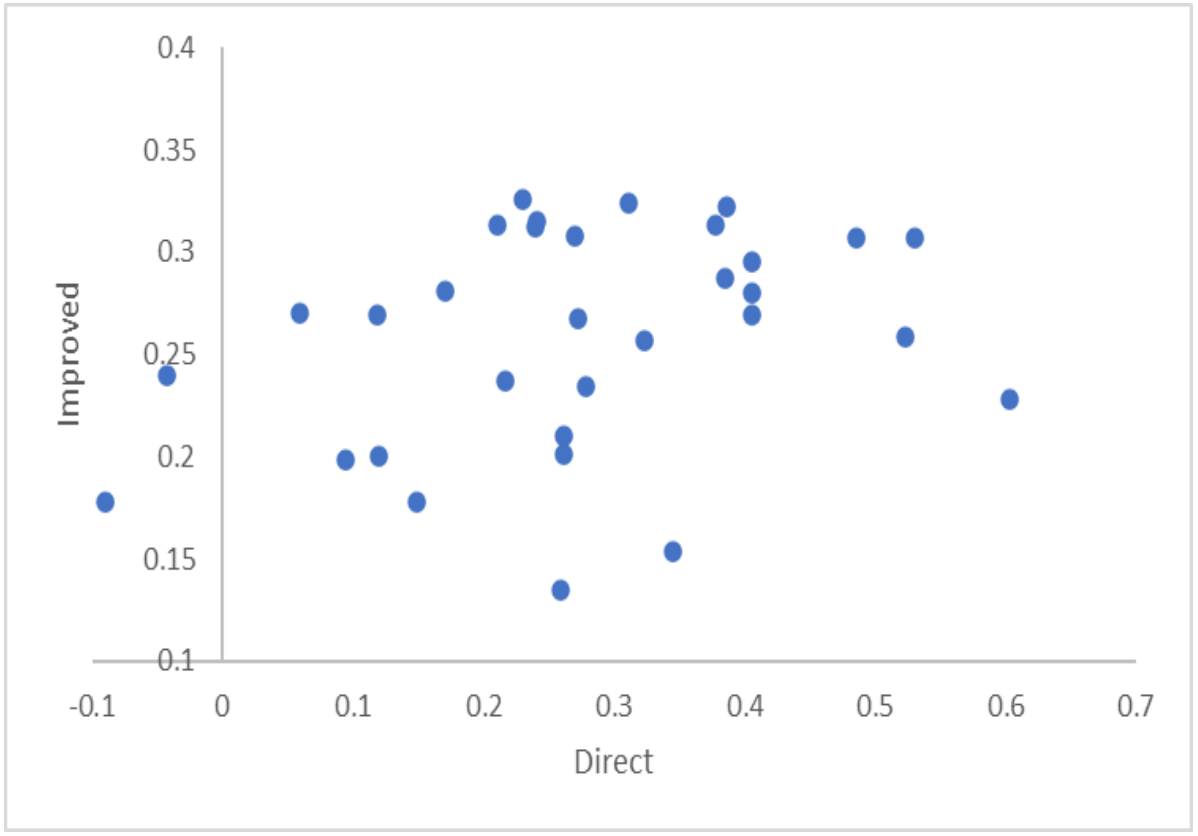


Complementary data

- Integrated Household Survey (IHS2) 2004/05
- The 2018 Malawi Population and Housing Census
  - Population density
  - Source of Cooking Energy
  - Economic Active Women Unemployed
  - Sex ratio
  - Literacy level of women

Direct vs FH estimates: ORS coverage





South Africa AIDS Targets

- South Africa continues to have the highest burden of HIV in the world, with about 7.9 million people living with HIV (PLHIV)
- The HIV transmission is largely heterosexual
- HIV testing and knowledge of HIV serostatus are effective in HIV prevention.
- Effective promotion of knowledge of HIV status ensures timely access to HIV prevention options or linkage to HIV treatment and support services needed to stay healthy.
- UNAIDS "95-95-95" targets, which aim for 95% of all people who are living with HIV to know their HIV status, 95% of all HIV positive people to be on antiretroviral treatment, and 95% of all people receiving antiretroviral therapy to have viral suppression by 2025 (Heath et al, AIDS, 2021).
- To access the virtual roundtable series recordings, please click below:
- This study considers:
    1) Recent HIV testing coverage (12 month window)
    2) HIV Prevalence among 15-49 year olds
    3) ARV uptake Coverage
    4) HIV viral load suppression

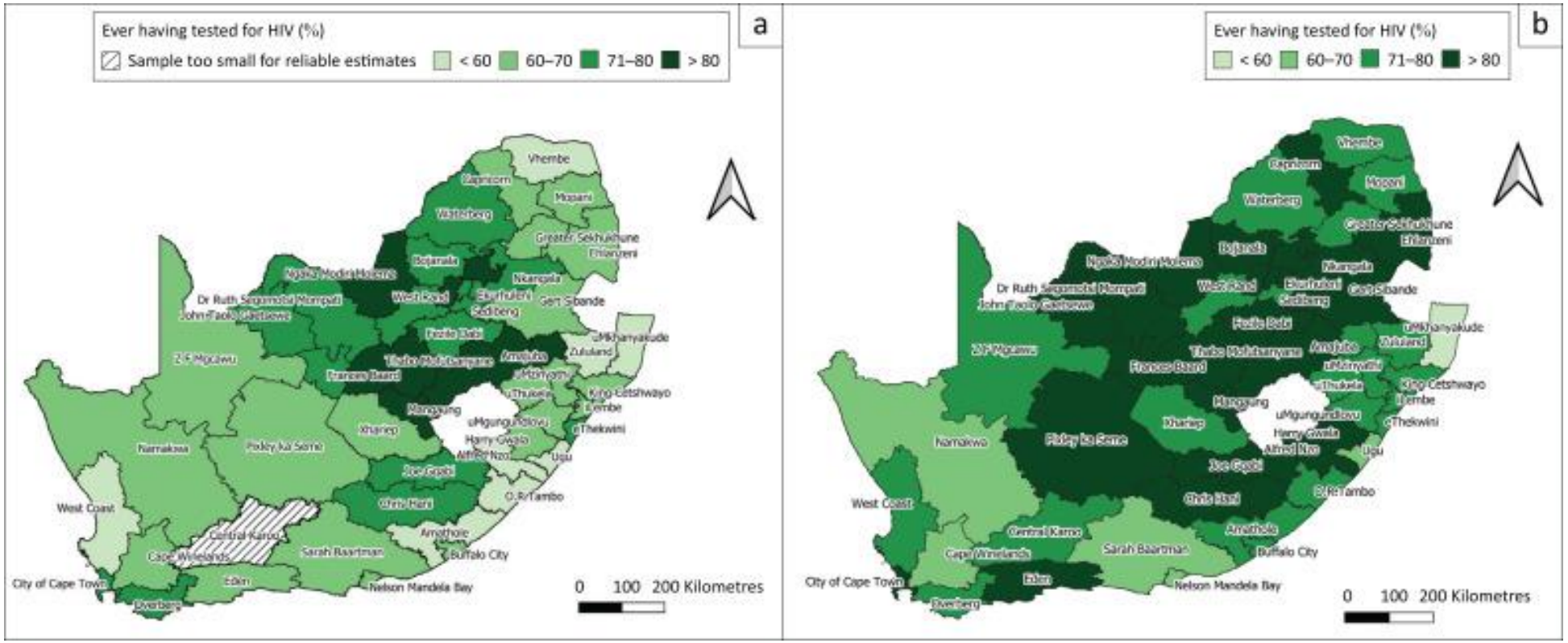
Direct Estimates:

- The Fifth South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, (SABSSM V), a population-based cross-sectional survey of households in South Africa, was designed to assess the prevalence and trends of key HIV–related indicators.
- The survey was conducted between January and December 2017 by the Human Sciences Research Council (HSRC) and provides information on national and sub-national progress toward HIV epidemic control in the country.
- Emplyed a complex multistage-stratified cluster sampling design was applied
- The sample was stratified by province and geo-type where geo-type was categorised as urban, rural formal (farms), and rural informal (traditional tribal areas and rural villages)
- A stratified sample of 1000 small area layers (SAL) was randomly sampled with probability proportional to size, where the number of households, or visiting points (VPs), within SAL was used as a measure of size from the national sampling frame

Summary stats per district

Variable	Obs	Mean	Std. dev.	Min	Max
hiv_status	52	21.33288	8.265195	4.87	37.83
arv	52	59.25038	17.72428	2.34	100
vl_suppre	52	60.30808	15.59321	22.62	100
test12	52	51.25269	8.173033	33.28	65.09
hiv_pstive~e	52	84.11538	104.6489	2	427

Geographical coverage: proportion of people who have ever been tested for HIV amongst (a) male and (b) female participants aged 15 years and older in the 52 districts in South Africa (Jooste et al, 2021)



Indirect Estimate Regressors:

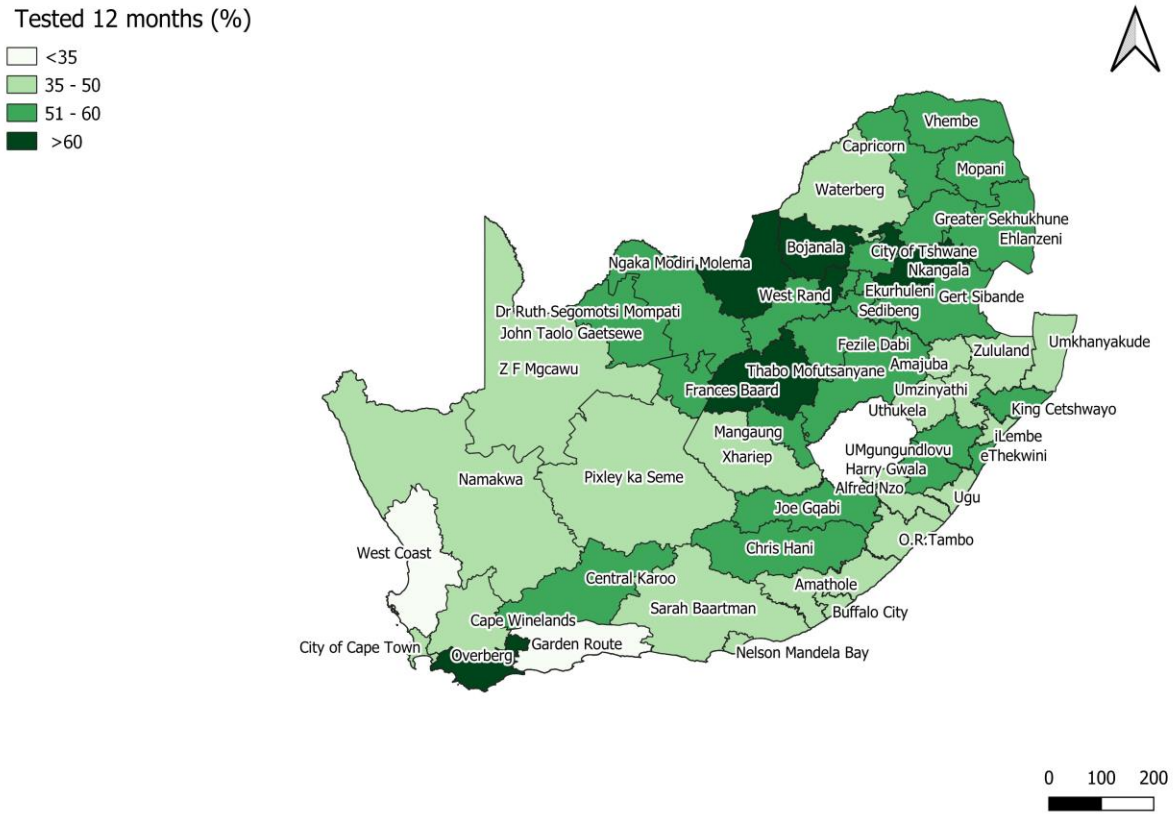
Auxilary variables

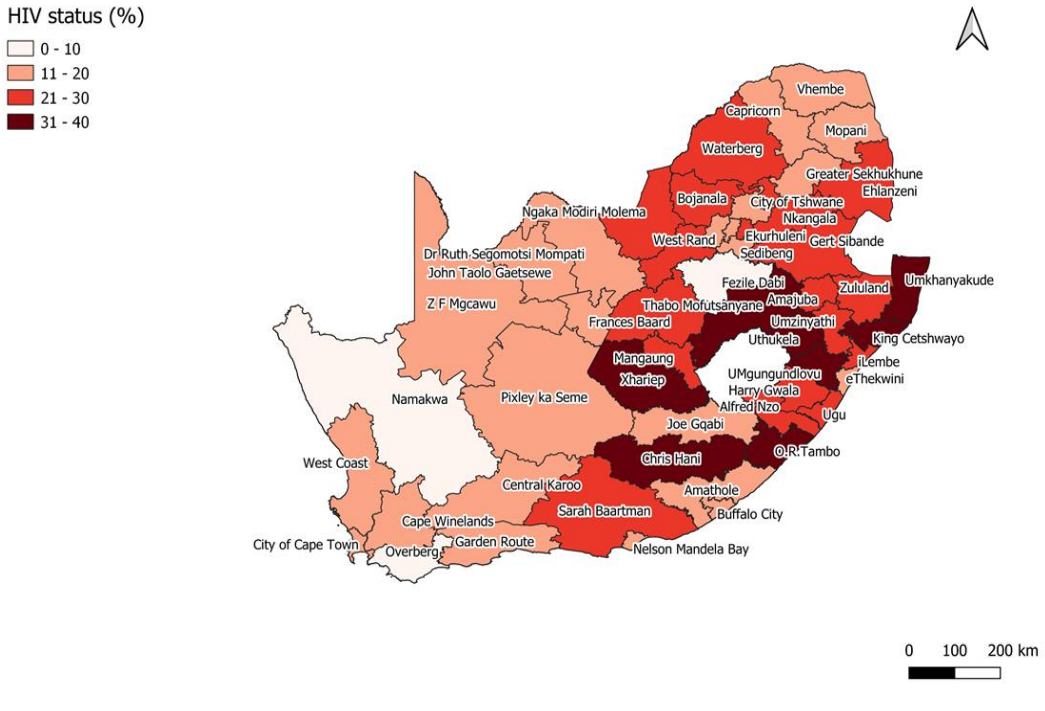
Variable	Categories	Source	Year	Unit	Extracted
Religion	Christian, Islim, Hindu, Traditional African, Judaism, None, Other	Statistics South Africa (StatsSA) Census	2022	District	Yes
Education	Highest level of Education (Secondary/Higher Edu)	StatsSA General Household Survey / Census	2022	District	Yes
Race	Black African, Coloured, Indian/Asian, White	StatsSA Census	2022	District	Yes
Marital Status	Married, Not Married	StatsSA Census	2022	District	Yes
Place of Residence	Formal dwelling, informal dwelling, traditional dwelling, Other	StatsSA Census	2022	District	Yes
Language	Afrikaans, English, IsiNdebele, IsiXhosa, IsiZulu, Sepedi, Sesotho, Setswana, SiSwati, Tshivenda, Xitsonga, Other	StatsSA Census	2022	District	Yes
Variable	**Categories**	**Source**	**Year**	**Unit**	**Extracted**
ANC HIV prevalence	Continuous variable	NICD	2022	District	Yes
VMMC coverage_sabssm5	Continuous variable	SABSSM	2017	District	Yes
VMMC coverage_dhs	Continuous variable	DHS	2016	District	Yes
Any MC	Continuous variable	DHB	2017	District	Yes
HIV_Prevalence_sabssm5	Continuous variable	SABSSM5	2017	District	Yes
MPI	Continuous variable	SA Community survey 2016	2001–16	District	Yes

Some correlations between the synthetic predictors

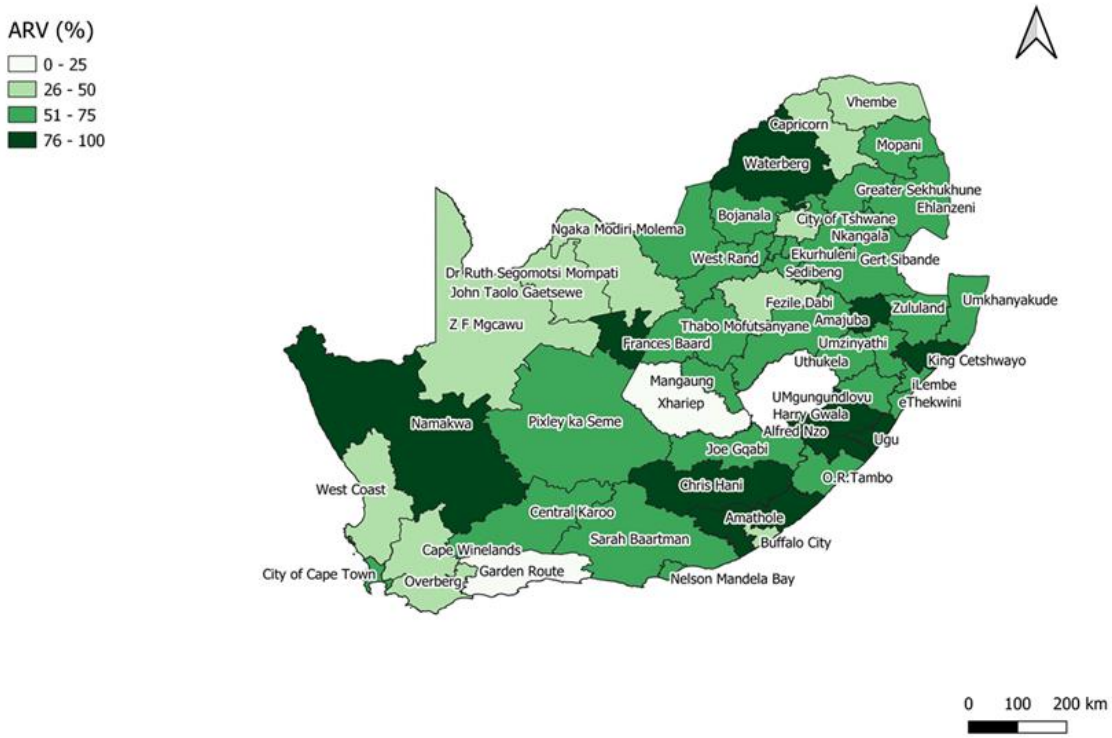
	p_HHS	p_vmmcdhs17	p_dhb2017	p_MCnat2019	p_anc2017	p_secondplus	p_christ	p_tribalfarm	p_unmarried	se_qtile	p_afrikaans	p_English	p_Isindebele	p_Isixhosa	p_Isizulu	p_Sepedi	p_Sesotho	p_Setsw
	VMMC (SABSSM)	VMMC (DHS)	VMMC (DHB)	MC (Nature,2019)	ANC HIV	Secondary/ higher education	Religion	Tribal/ farm locations	marital staus (unmarried)	SAMPI poverty index	Afrikaans speaking	English speaking	Isindebele speaking	sixhosa speaking	Isizulu speaking	Sepedi	Sesotho speaking	Setsw speaking
p_HHS	1																	
p_vmmcdhs17	0.6617*	1																
p_dhb2017	0.5726*	0.7333*	1															
p_MCnat2019		-0.3170*		1														
p_anc2017	0.3379*	0.5062*	0.5456*		1													
p_secondplus	0.4399*					1												
p_christ			-0.4383*	-0.4348*	-0.5959*	-0.3780*	1											
p_tribalfarm			0.3090*			-0.4540*	-0.3711*	1										
p_unmarried		0.3593*	0.4090*		0.5220*	-0.4906*	-0.4263*	0.8023*	1									
se_qtile					0.4704*	-0.4777*	-0.4247*	0.8003*	0.8743*	1								
p_afrikaans		-0.3958*	-0.5548*	-0.2815*	-0.6422*		0.7277*	-0.7313*	-0.7575*	-0.7750*	1							
p_English	0.2845*				0.3194*	0.7095*		-0.5567*	-0.3566*	-0.3704*		1						
p_Isindebele		-0.3161*		0.6288*		0.3443*			-0.2929*				1					
p_Isixhosa	-0.3749*	-0.5305*	-0.6845*					-0.3987*	-0.3230*		0.3265*			1				
p_Isizulu	0.5656*	0.7420*	0.7720*		0.7292*	0.3123*	-0.4973*		0.3679*		-0.5396*	0.3108*		-0.4023*	1			
p_Sepedi	0.5334*			0.3550*		0.3086*							0.5413*	-0.4426*		1		
p_Sesotho	0.5579*							-0.3327*	-0.3577*							0.5230*	1	
p_Setsw	0.3484*				-0.4827*		0.4842*	-0.3994*	-0.5214*	-0.4427*	0.5196*		0.3337*			0.6731*	0.5929*	1

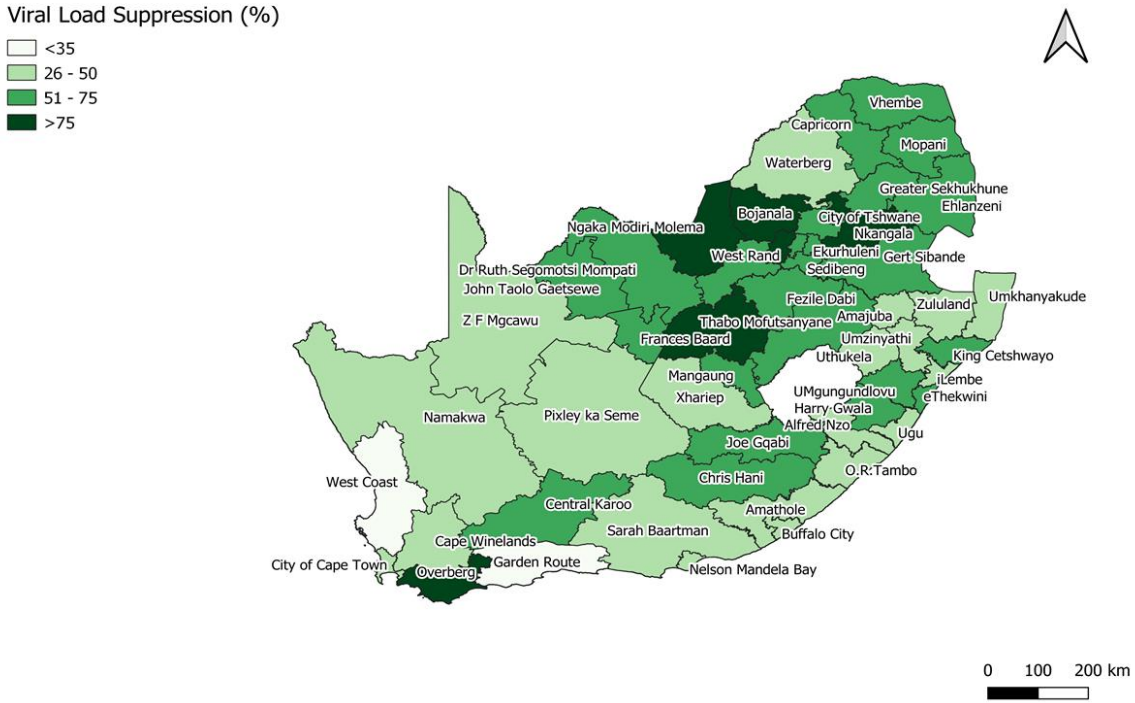
FH Estimates of SA 95-95-95 AIDS Targets





FH Estimates of SA 95-95-95 AIDS Targets: II





Discussions

SAE methods should be used for local level estimates of health especially samples sizes are not adequate

Problems with quality and availability of auxiliary data (e.g. time misalignment between survey and auxiliary data)