Joint modelling spatial patterns of disease risk for data from multiple sources: An application on HIV prevalence data from antenatal sentinel and demographic and health surveys in Namibia.

Dismas Ntirampeba [1,3] Isaak Neema [2] Lawrence Kazembe [3]

[1]Namibia University of Science and Technology

[2]Namibia Statistics Agency

[3]University of Namibia

Outline

Abstract

Background

- In disease mapping field, scholars often encounter data from multiple sources.

But such data are fraught with challenges including:

- lack of representative sample;
- incomplete;
- and may be spatially misaligned.

Methods

- Bivariate spatial model is used in order to deal with the sampling bias and misalignment
- Data used to estimated HIV prevalence: 2014 National HIV Sentinel survey (NHSS) among pregnant women aged 15-49 years attending antenatal care (ANC) and the 2013 Namibia Demographic and Health Surveys (NDHS).

Abstract

Results

- Districts and constituencies in the northern part of Namibia were found to be highly associated with HIV infection.

Factors significantly associated with HIV:

- place of residence, gender, gravida, marital status, number of kids died
- wealth index, education, and condom use

Abstract

Conclusions

- The study had unveiled determinants of HIV infection in Namibia and had shown areas at high risk through HIV prevalence maps
- The study has shown that the power of predicting HIV prevalence using the DHS data source can be enhanced by jointly modelling other HIV data such as NHSS data.

Keywords

HIV prevalence, joint bivariate analysis, spatial analysis, Namibia

Introduction

- A downwards change in the trajectory of AIDS epidemic has been achieved worldwide (WHO and UNAIDS, 2015), but by the end of 2014:
- 36.9 million people were estimated to live with HIV (UNAIDS, 2015), of which about 70 % (25.8 million) are found in sub-Sahara Africa.
- 2 million of people were newly infected with HIV, a large portion (1.4 million) of which is said to be in sub-Sahara Africa (UNAIDS, 2015).
- In Namibia, the HIV prevalence among adults and children was estimated to be around 26000, of which 11000 were newly infected (MoHSS, 2014).

Introduction

- The NHSS among pregnant women attending antenatal care (ANC) and DHS are the commonly used tools to monitor the HIV prevalence trend in a country.
- However, each one of the two data sources has its own weaknesses that may lead to inaccurate estimation of HIV prevalence.
- NHSS limitations: accessibility of ANC sites and exclusion of some categories of the population (e.g. men and non-pregnant women) (Manda et al., 2015).
- DHS limitation: a significant non-response drawback (Manda et al., 2015).

Introduction

In face of these limitations, a joint analysis of data from different sources is proven to be useful (Manda et al., 2012):

- Avoids multiple testing on same data
- Helps dealing with identifiability in random effect parameters estimation
- Increases precision and efficiency of parameter estimates
- Captures disease specific covariates and as well as to carry pairwise and cross-covariances inferences between different sources

Introduction

Some of approaches of multivariate techniques are:

- multivariate normal distribution and iterative generalised least squares (IGLS) method, but underestimate the variation associated with sources (Manda et al., 2012)
- Multivariate conditional autoregressive (MCAR) is commonly used in mapping of multiple diseases as it enables to account for within and /or between areal associations
- Bellier et al. (2013) had jointly analysed multiple data sources by including an observability parameter
- Guo and Carlin (2004) had used a full Bayesian approach to link longitudinal and survival data.

Introduction

- Shared-component model pioneered by (Knorr-Held and Best, 2001):
- This model splits the disease profile into two components, namely the disease-specific component representing spatially varying factors, and the shared component which is a proxy of unobserved spatially varying factors that are common to both or all diseases
- It had been extensively used in joint analysis of multiple health outcomes (e.g (Downing et al., 2008; Onicescu et al., 2010)).

Introduction

- There is a rich literature on analyses of determinants of HIV and its geographical spread, but most of the analyses used were based on univariate methods for different data sources
- One notable study by Manda et al. (2015) has used a shared-component modelling approach to jointly analyse data from DHS and ANC surveys.
- District level HIV prevalence rates were used and also two circumstantial covariates were considered as determinants of HIV.
- Data were first aggregated at district level and then a spatial bivariate modelling approach was applied on aggregated rates. In this situation, a misalignment in data sources is avoided

introduction

- This has some limitations as, for instance, many covariates available from ANC or DHS would not be used in the joint analysis

Ways to include most of ANC and/or DHS covariates would be:

- Compute averages at district level and use them in joint modelling.
- Alternatively, a model that allows different neighbourhood structures may be useful as it would permit to model data available at different block levels.
- Therefore, the primary objective of this study is to develop a **joint spatial model for NHSS and DHS data**, which enables the estimation at any location of the constituency or district level **while dealing with misalignment in data.**

Data

2013 DHS data:
Sampling methodology:

- Stage 1: 554 enumeration areas (EAs) were selected using EA size proportional stratified probability
- Stage 2: 20 households were selected from each EA using equal probability systematic sampling approach

Instruments:
Three questionnaires (Household questionnaire, woman's questionnaire, and the man's questionnaire) to address questions on household characteristics and assessed women's and men's knowledge of HIV.

Data

Sample size:
A total 9176 women and 3950 men formed part of 2013 DHS interviews.
Note: Survey included HIV testing women and men aged between 15 and 64 years selected throughout the country.

2014 NHSS data:

- Objective: Determine HIV prevalence among pregnant women age 15-49 years attending antenatal care (ANC) clinics at public health facilities in Namibia.
- NHSS started in 1992 and has expanded from 8 sites to 35 district sites supplemented by 98 satellite facilities.

Data

- Sampling techniques, sample size and data collection methods were based on the World Health Organization (WHO) guidelines for conducting HIV surveys among pregnant women and other groups.
- Sample size: 7 920 women enrolled for 2014 NHSS, majority of them are multi-gravida.

Data

Variables:

Table: Summary of variables used in this study by source

	DHS	NHSS
Variables	HIV status Place of residence Gender Age of the respondent Head of household Marital status Number of kids dead Education Wealth	HIV status Age of the respondent Gravidity

Data

Table: variables(cont)

	DHS	NHSS
	Stayed away of home Sexual activity (in last 4 months) Age at first sex condom use Had STI in last 12 months	

Separate modelling for data source

Univariate models:

$$y_{ij} = \begin{cases} 1 & \text{if disease incident is observed at location } i \text{ for dataset } j \\ 0 & \text{otherwise} \end{cases}$$

Then $y_{ij} \sim Bernoulli(p_{ij})$, where p_{ij} is the probability of recorded incident at location i from dataset j.
Thus, the independent model fitted to dataset $(j = 1, 2)$ is given by

$$logit(p_{ij}) = \beta_0 + \sum_{k}^{r} \beta_k X_{ijk} + f_j(g_i) + z_j(s_i), \qquad (1)$$

Separate modelling for data source

Where β_0 is the intercept, x_{ijk} is the k^{th} linear covariate of dataset j in a given health district facility i or constituency i, $f_j(\cdot)$ is a function of a non-linear covariate, g_i is a vector of ages, and $z_j(s_i)$ is a Gaussian random field.

- Eq. 1 can be split into two separate (univariate) models as follows (first stage of Bayesian hierarchy):

$$logit(p_{i1}) = \beta_{01} + \sum_{k}^{r} \beta_k X_{i1k} + f_1(g_i) + z_1(s_i), \quad (2)$$

$$logit(p_{i2}) = \beta_{02} + \sum_{k}^{r} \beta_k X_{i2k} + f_2(g_i) + z_2(s_i), \quad (3)$$

- For the Gausssian random field, a multivariate Gaussian distribution was assumed:

$$z(s) \sim N(0, \Sigma), \tag{4}$$

- where Σ is the covariance matrix
- Elements of the covariance matrix Σ are specified as a function of the marginal variance of the process σ_z and the Matérn correlation function Cor_M as follows:

$$\Sigma_{ij} = \sigma_z Cor_M(z(s_i), z(s_j)), \tag{5}$$

Separate modelling for data source

where The Matérn correlation function is given by;

$$Cor_M(z(s_i), z(s_j)) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\kappa \parallel s_i - s_j \parallel)^{\nu}\kappa_{\nu}(\kappa \parallel s_i - s_j \parallel), \qquad (6)$$

where $\parallel . \parallel$ denotes the Euclidean distance, $\kappa_{\nu}(.)$ is the modified Bessel function of second order, k and ν are scale parameter and smoothness parameter, respectively.

Separate modelling for data source

Specification of prior distributions(second stage of hierarchy:)

- Inverse Gamma prior distributions were assigned to k, ν, and σ_z
- $\beta \sim \ N(0, \tau_\beta^{-1} I)$,weakly informative Gaussian priors
- $\Delta_{g_i} \sim \ N(0, \tau_{g^{-1}} I)$,first order random walk process(penalised regression spline approach using,) for details see (Okango et al., 2015; Ngesa et al., 2014)

Joint modelling of HIV prevalence from DHS and NHSS data sources

Joint modelling of HIV prevalence from DHS and NHSS data sources:

Considering the bivariate model, which pools the two datasets, let y_{ij} be a binary indicator of HIV incidence at location i from dataset ($j = 1, 2$). Then y_{ij} follows a $Bernoulli(p_{ij})$, p_{ij} is the probability of recorded HIV incident pertaining to the j^{th} dataset. The bivariate model is then given by;

$$logit(p_{i1}) = \beta_0 + \sum_{k}^{r} \beta_k X_{i1k} + f_1(g_{i1}) + z_1(s_i), \tag{7}$$

$$logit(p_{i2}) = \beta_0 + \sum_{k}^{r} \beta_k X_{i2k} + f_1(g_{i2}) + z_2(s_i) + \gamma z_1(s_i), \tag{8}$$

with $j = 1$ for NHSS and $j = 2$ for DHS

Joint modelling of HIV prevalence from DHS and NHSS data sources

where,

- X is vector of linear covariates with corresponding regression parameters β;
- g_{ij} is the vector of ages which are assumed to follow a random walk of order 1;
- $z_1(s_i)$ is a Gaussian random field shared between both responses, the interaction parameter γ describes how much of the structure captured in $z_1(s_i)$ is also inherent the $logit(p_{i2})$.
- Similar prior distributions to those specified for univariate models were assigned for parameters and hyperparameters of the joint model.

Estimation of parameters and model diagnostics

Estimation of parameters and model diagnostics:

- Posterior distribution

Posterior distribution: obtained by taking the product of likelihood function together with the prior and hyper distributions. In this study, posterior distribution is given by;

$$p(\theta|y_{ij}) \propto \prod_{i=1}^{n} L(y_{ij}, p_{ij}) \prod_{g=1}^{2} [p(\Delta g_i|\tau_g^{-1})p(\tau_g^{-1})] \prod_{k=1}^{r} p(\beta_k)p(\tau_{\beta_k}^{-1}) \prod_{j=1}^{2} p(z_j|k_j, \nu_j, \sigma_z \quad (9)$$

- stochastic partial differential equation (SPDE) approach with INLA was employed to estimate posterior marginal distributions and any other posterior inferences
- The best model was identified using the deviance information criterion (DIC)

Table: Nested models to be fitted in this study

Model	GRF	Shared component	covar
M_{U1}: Uni.model for NDHS data	$\surd$	-	-
M_{U2}: Uni.model for NHSS data	$\surd$	-	-
M_{U12}: Uni.model for NDHS data +covar	$\surd$	-	$\surd$
M_{U22}: Uni.model for NHSS data+covar	$\surd$	-	$\surd$
M_{J1}: Biv.model for NDHS & NHSS data	$\surd$	$\surd$	-
M_{J2}: Biv.model for NDHS & NHSS data+covar	$\surd$	$\surd$	$\surd$

Descriptive statstics

Spatial distribution of observed HIV prevalence from DHS and NHSS:

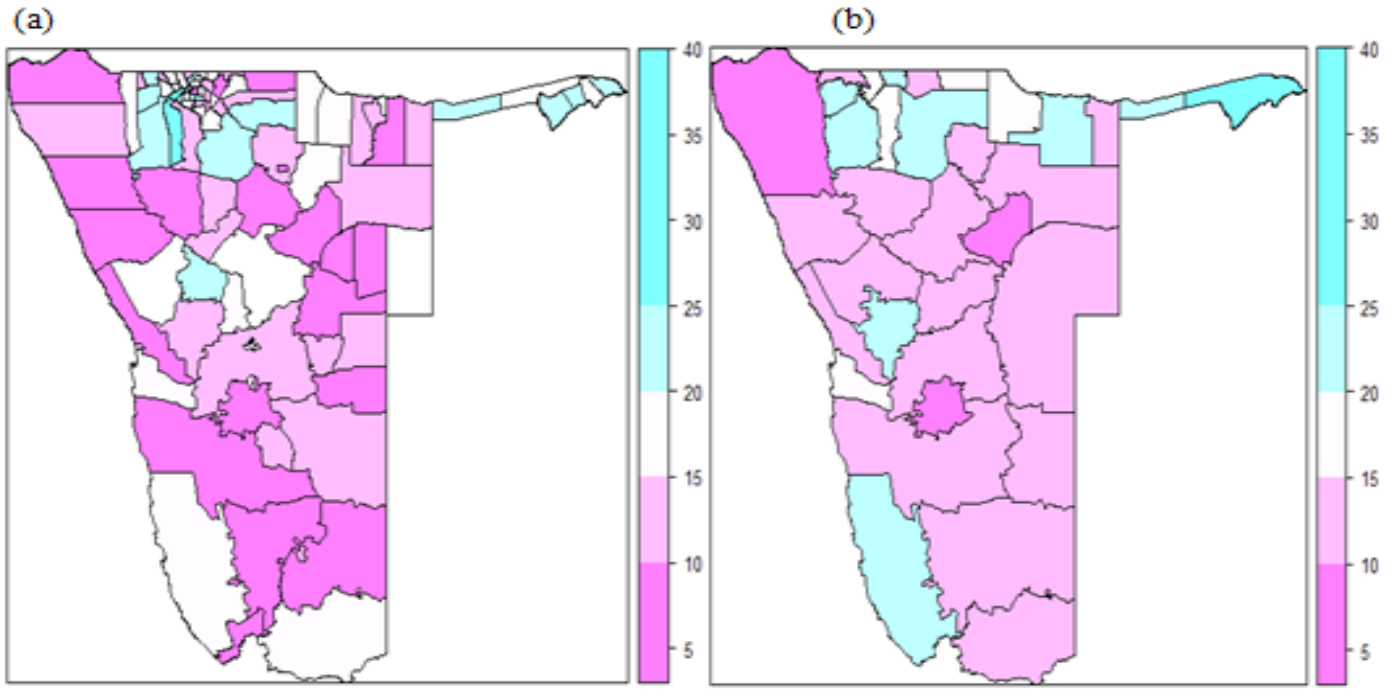

Figure: Crude HIV prevalence: (a) Constituencies HIV prevalence (2013 NDHS data); (b)Health districts prevalence (2014 NHSS data)

Comparison of fitted models

Table: DIC values for fitted models

Model	DIC		
M_{U1}	7011.89		
M_{U2}: Uni. model for NHSS data	6872.59		
M_{U12}: Uni.model for NDHS data +covar	6344.00		
M_{U22}: Uni.model for NHSS data+covar	6388.33		
M_{J1}: Biv. model for NDHS & NHSS data	NDHS-DIC	NHSS-DIC	Total
	7003.498	6870.218	1387
M_{J2}: Biv. model for NDHS & NHSS data+covar	NDHS-DIC	NHSS-DIC	Total
	5998.11	6355.98	1235

Estimated HIV prevalence

Estimated HIV prevalence:

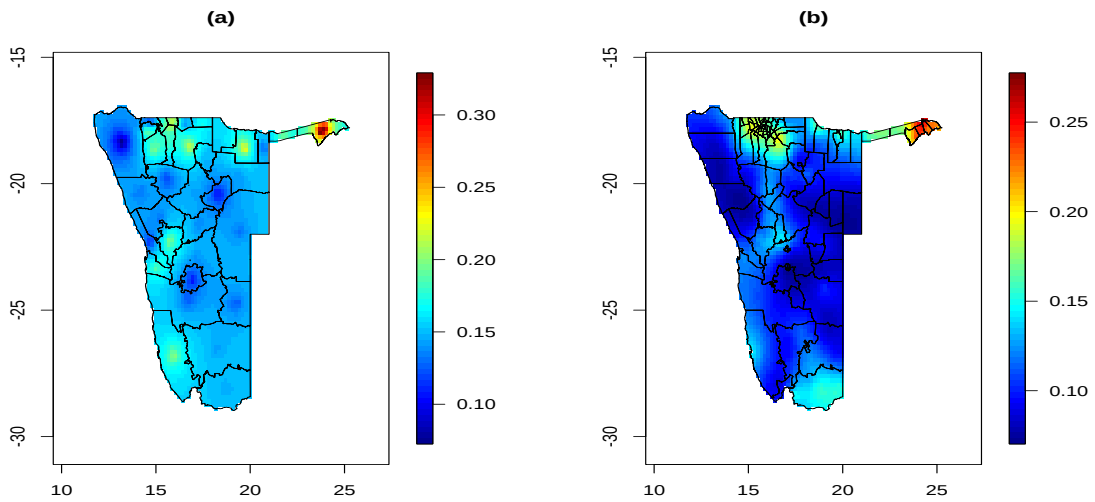

Figure: Estimated HIV prevalence using separate models: (a) HIV prevalence estimates from 2014 NHSS data; (b)HIV prevalence estimates from 2013 DHS data

Estimated HIV prevalence

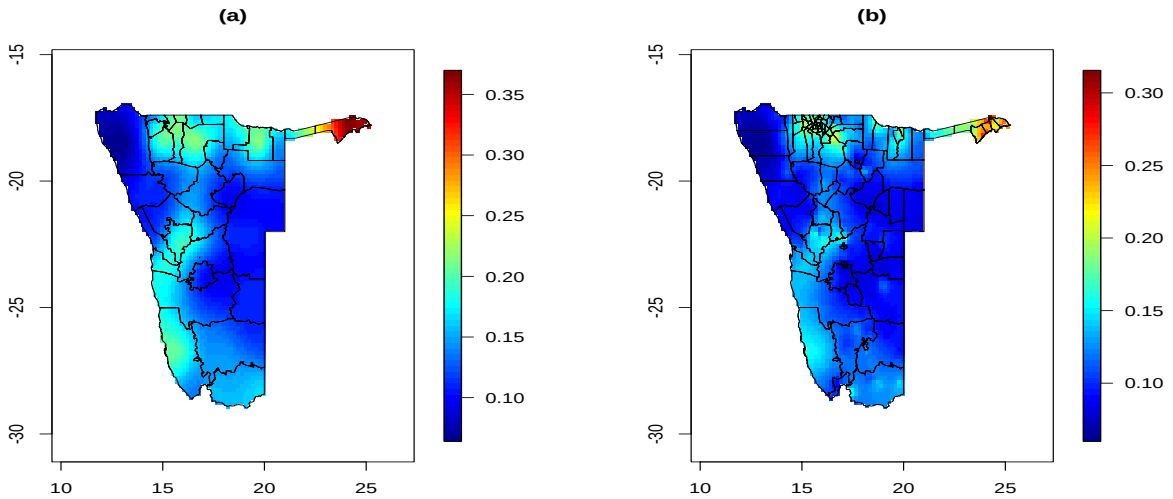

Figure: Estimated HIV prevalence using the bivariate model: (a) HIV prevalence estimates for 2014 NHSS data; (b)HIV prevalence estimates for 2013 NDHS data

Nonlinear effects

Non linear trajectory of HIV random effects:

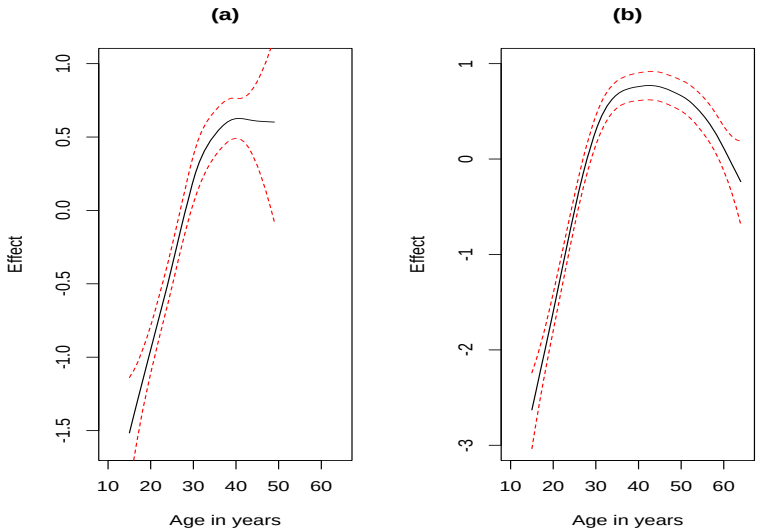

Figure: Estimated nonlinear effects of age on HIV infection and corresponding confidence intervals: (a) NHSS data; (b) DHS data

Table: Estimated covariate effects and their 95 % credible intervals (CI)

Joint(bivariate) model		
Covariate	OR	95 % CI
β_{01}	0.12	(0.07, 0.23)
Gravida		
Prima-gravida (ref)	1.00	
Multi-gravida	1.88	(1.52, 2.32)
β_{02}	0.08	(0.04, 0.18)
Place Residence		
Rural (Ref)	1	
Urban	1.53	(1.27, 1.84)
Gender		
Female	1	
Male	0.68	(0.58, 0.79)

Table: Estimated covariate effects and their 95 % credible intervals (CI)

Joint(bivariate) model		
Head of household		
Male (Ref)	1	
Female	1.14	(0.97, 1.33)
Marital status		
Never in union (Ref)	1	
Married	0.72	(0.58, 0.89)
Living with a partner	1.41	(1.16, 1.73)
Widowed	1.46	(1.06, 2.02)
Divorced	1.07	(0.66, 1.75)
Separated	1.41	(1.04, 1.91)

Table: Estimated covariate effects and their 95 % credible intervals (CI)

Number of Kids dead		
No child died (Ref)	1	
One child died	1.84	(1.48, 2.29)
More than one child died	2.69	(1.84, 3.91)
Education		
No education (Ref)	1	
Primary	1.09	(0.87, 1.37)
Secondary	0.84	(0.66, 1.06)
Higher	0.63	(0.41, 0.96)
Wealth index		
Poorest (Ref)	1	
Poorer	0.93	(0.79, 1.09)
Middle	1.10	(0.89, 1.35)
Richer	0.78	(0.61, 0.99)
Richest	0.33	(0.24, 0.46)

Table: Estimated covariate effects and their 95 % credible intervals (CI) (continued)

Stayed away from home		
Did not move away (Ref)	1	
Moved away	0.93	(0.79, 1.09)
Sexual activity		
Never had sex (Ref)	1	
Not active	0.98	(0.90, 1.07)
Active	1.15	(1.06, 1.26)
Age at first sex (in years)		
Never had sex (Ref)	1	
≤ 11	1.29	(0.87, 1.91)
12 to 14	1.08	(0.67, 1.73)
15 to 17	1.47	(0.99, 2.17)
18 and above	1.26	(0.85, 1.87)

Table: Estimated covariate effects and their 95 % credible intervals (CI)

Joint(bivariate) model		
Covariate	OR	95 % CI
Condom used		
No(Ref)	1	
Yes	1.78	(1.53,2.07)
Had STI in last 12 months		
No (Ref)	1	
Yes	1.05	(0.96, 1.16)

Discussion

- A full Bayesian framework through SPDE approach with INLA to model jointly the two data sources available at two different block levels
- The bivariate model, which used a spatial shared component that acts as a surrogate of HIV risky behaviours among pregnant women in order to improve the estimation of HIV prevalence using DHS source, was found to be more appropriate in estimating HIV prevalence
- The interaction parameter γ(0.76,CI:0.50 to 1.3) was found to be significant

Discussion

- The joint analysis of DHS and ANC sources has enhanced the estimation of HIV prevalence using the demographic and health survey (DHS). This finding agrees with the study Manda et al (Manda et al., 2015).
- Relationship between age and its effects on HIV infection followed an inverted U shape. This finding agrees with other studies (Okango et al., 2015; Ngesa et al., 2014).
- Urban areas had higher risk of getting infected compared to rural areas. This finding has been reported in many other studies (Manda et al., 2015; Okango et al., 2015; Ngesa et al., 2014; Amornkul et al., 2009).

Discussion

- This finding could be used to design focused public campaigns against HIV/AIDS
- Poverty levels were inversely associated with the likelihood of HIV infection. (Chege et al., 2012) showed that unwanted or forced sex was related to lack of resources and the ability to obtain resources.
- HIV infection was found to be significantly related with head of a household. It has been shown that the male-headship is a proxy of a better socio-economic status (Musenge et al., 2013). Mufune et al. (2014),Chege et al. (2012) advanced the complex of inferiority of women and the struggle to obtain leadership positions and power to make decisions as possible explanations.

Discussion

- The gender and HIV infection relationship was confirmed in many studies Manda et al. (2015); Amornkul et al. (2009); Chege et al. (2012); Barankanira et al. (2016), including our study.
- Education was found to be negatively associated with HIV infection. this finding was confirmed by other similar studies (e.g. Okango et al. (2015); Ngesa et al. (2014); Chege et al. (2012))
- Possible reasons: limited sexual education in Namibia schools, no proper training provided to teachers in this matter, Life -skills subject not taken seriously as it is not examinable (Mufune et al., 2014)

Discussion

- At contrary of the intuition, the condom use covariate was found to be positive related to HIV infections, also reported in the work of Ngesa et al. (2014)
- One of the justifications provided: men use condoms in the earlier stage of relationship with their partners and later on give up on using condoms.
- Number kids died had positive significant effects of HIV infection. This might imply that kids could have been infected by their mothers.

- Sexual behaviour or biological characteristics such as sexual activity, age at first sex and STI are associated with HIV infection. These finding could be used to identify groups with high risk where greater efforts should be directed.
- Based on findings of this study, we proposed that great efforts in terms of primary and secondary HIV interventions should be concentrated on constituencies in the northern part of Namibia.
- Unlike other study that a same underlying spatial process for different sources, the bivariate model developed it is possible to specify different spatial processes (e.g. a Poisson and Bernouilli processes) through the link function.
- Weakness of the study: This study did not take into account the bias that might be induced by random displacements of the positions of HIV (DHS data source) cases

conclusions and recommendations

- The study had unveiled determinants of HIV infection in Namibia and had shown areas at high risk through HIV prevalence maps
- The study had used a bivariate modelling approach that helped dealing with spatially misaligned data
- The study has shown that the power of predicting HIV prevalence using the DHS data source can be enhanced by jointly modelling other HIV data such as NHSS data
- Findings and prevalence maps could be used by the Ministry of health and social services and any health policy makers to **identify groups of people in need of HIV support** and **where they live** in order to **efficiently allocate resources that are increasingly becoming scarce**

References

Amornkul, P. N., Vandenhoudt, H., Nasokho, P., Odhiambo, F., Mwaengo, D., Hightower, A., Buvé, A., Misore, A., Vulule, J., Vitek, C., et al. (2009). HIV prevalence and associated risk factors among individuals aged 13-34 years in Rural Western Kenya. *PloS one*, 4(7):e6470.

Barankanira, E., Molinari, N., Niyongabo, T., and Laurent, C. (2016). Spatial analysis of HIV infection and associated individual characteristics in Burundi: indications for effective prevention. *BMC public health*, 16(1):1.

Bellier, E., Neubauer, P., Monestiez, P., Letourneur, Y., Ledireach, L., Bonhomme, P., and Bachet, F. (2013). Marine reserve spillover: Modelling from multiple data sources. *Ecological informatics*, 18:188–193.

Chege, W., Pals, S. L., McLellan-Lemal, E., Shinde, S., Nyambura, M., Otieno, F. O., Gust, D. A., Chen, R. T., and Thomas, T. (2012).