Modelling HIV/AIDS Epidemic in Nigeria

Jude Ikechukwu Eze

A Dissertation Submitted to the

University of Glasgow

for the degree of

Doctor of Philosophy

Department of Statistics

March 2009

Abstract

Nigeria is one of the countries most affected by the HIV/AIDS pandemic, third only to India and South Africa. With about 10% of the global HIV/AIDS cases estimated to be in the country, the public health and socio-economic implications are enormous.

This thesis has two broad aims: the first is to develop statistical models which adequately describe the spatial distribution of the Nigerian HIV/AIDS epidemic and its associated ecological risk factors; the second, to develop models that could reconstruct the HIV incidence curve, obtain an estimate of the hidden HIV/AIDS population and a short term projection for AIDS incidence and a measure of precision of the estimates.

To achieve these objectives, we first examined data from various sources and selected three sets of data based on national coverage and minimal reporting delay. The data sets are the outcome of the National HIV/AIDS Sentinel Surveillance Survey conducted in 1999, 2001, 2003 and 2005 by the Federal Ministry of Health; the outcome of the survey of 1057 health and laboratory facilities conducted by the Nigerian Institute of Medical Research in 2000; and case by case HIV screening data collected from an HIV/AIDS centre of excellence.

A thorough review of methods used by WHO/UNAIDS to produce estimates

of the Nigerian HIV/AIDS scenario was carried out. The Estimation and Projection Package (EPP) currently being used for modelling the epidemic partitions the population into at-risk, not-at-risk and infected sub-populations. It also requires some parameter input representing the force of infection and behaviour or high risk adjustment parameter. It may be difficult to precisely ascertain the size of these population groups and parameters in countries as large and diverse as Nigeria. Also, the accuracy of vital rates used in the EPP and Spectrum program is doubtful. Literature on ordinary back-calculation, nonparametric back-calculation, and modified back-calculation methods was reviewed in detail. Also, an indepth review of disease mapping techniques including multilevel models and geostatistical methods was conducted.

The existence of spatial clusters was investigated using cluster analysis and some measure of spatial autocorrelation (Moran I and Geary c coefficients, semivariogram and kriging) applied to the National HIV/AIDS Surveillance data. Results revealed the existence of spatial clusters with significant positive spatial autocorrelation coefficients that tended to get stronger as the epidemic developed through time. GAM and local regression fit on the data revealed spatial trends on the north-south and east - west axis.

Analysis of hierarchical, spatial and ecological factor effects on the geographical variation of HIV prevalence using variance component and spatial multilevel models was performed using restricted maximum likelihood implemented in $R^{©}$ and empirical and full Bayesian methods in $WinBUGS^{©}$. Results confirmed significant spatial effects and some ecological factors were significant in explaining the variation. Also, variation due to various levels of aggregation was prominent.

Estimates of cumulative HIV infection in Nigeria were obtained from both

parametric and nonparametric back-calculation methods. Step and spline functions were assumed for the HIV infection curve in the parametric case. Parameter estimates obtained using 3-step and 4-step models were similar but the standard errors of these parameters were higher in the 4-step model. Estimates obtained using linear, quadratic, cubic and natural splines differed and also depended on the number and positions of the knots. Cumulative HIV infection estimates obtained using the step function models were comparable with those obtained using nonparametric back-calculation methods. Estimates from nonparametric back-calculation were obtained using the EMS algorithm. The modified nonparametric back-calculation method makes use of HIV data instead of the AIDS incidence data that are used in parametric and ordinary nonparametric back-calculation methods. In this approach, the hazard of undergoing HIV test is different for routine and symptom-related tests. The constant hazard of routine testing and the proportionality coefficient of symptom-related tests were estimated from the data and incorporated into the HIV induction distribution function. Estimates of HIV prevalence differ widely (about three times higher) from those obtained using parametric and ordinary nonparametric back-calculation methods. Nonparametric bootstrap procedure was used to obtain point-wise confidence interval and the uncertainty in estimating or predicting precisely the most recent incidence of AIDS or HIV infection was noticeable in the models but greater when AIDS data was used in the back-projection model.

Analysis of case by case HIV screening data indicate that of 33349 patients who attended the HIV laboratory of a centre of excellence for the treatment of HIV/AIDS between October 2000 and August 2006, 7646 (23%) were HIV positive with females constituting about 61% of the positive cases. The bulk of infection was found in patients aged 15-49 years, about 86 percent of infected

females and 78 percent of males were in this age group. Attendance at the laboratory and the proportion of HIV positive tests witnessed a remarkable increase when screening became free of charge. Logistic regression analysis indicated a 3-way interaction between time period, age and sex. Removing the effect of time by stratifying by time period left 2-way interactions between age and sex. A Correction factor for underreporting was ascertained by studying attendance at the laboratory facility over two time periods defined by the cost of HIV screening. Estimates of HIV prevalence obtained from corrected data using the modified nonparametric back-calculation are comparable with UN estimates obtained by a different method.

The Nigerian HIV/AIDS pandemic is made up of multiple epidemics spatially located in different parts of the country with most of them having the potential of being sustained into the future given information on some risk factors. It is hoped that the findings of this research will be a ready tool in the hands of policy makers in the formulation of policy and design of programs to combat the epidemic in the country. Access to data on HIV/AIDS are highly restricted in the country and this hampers more in-depth modelling of the epidemic. Subject to data availability, we recommend that further work be done on the construction of stratification models based on sex, age and the geopolitical zones in order to estimate the infection intensity in each of the population groups. Uncertainties surrounding assumptions of infection intensity and incubation distribution can be minimized using Bayesian methods in back-projection.

Acknowledgement

To God be the glory great things He has done! When I consider the depth from where I started and the height I have attained, I cannot but appreciate the blessings of God upon my life. For I know that promotion comes not from the east, nor from the west, nor from the south but from God. All I have, I have received from God, not because of my hardwork or personal qualities, for it is not he that willeth nor he that runneth but of God that showeth mercy. It is not by power and might but by the Spirit of God. Every good and perfect gift comes from God, the father of light. You selected me out of the multitude that applied for the Commonwealth scholarship, you made my admission in Glasgow University possible, you started this journey with me and you have been faithful to the end. Father, what you have done and what you are doing is good, I appreciate you Lord and I ascribe this great success to You alone. Be thou glorified in my life forever in Jesus name, Amen.

My gratitude goes to the Commonwealth Scholarship Commission for giving me this rare opportunity to be one of her prestigious scholars. I feel overwhelmed by this privilege to be counted among the select few from the hundreds of thousands that applied. I thank the Executive Secretary, Dr Kirkland, my former and current award administrators, Sabina Eboll and Selina Hannaford and the British Council team at Manchester, you all were there all the time I needed your

advice and assistance.

I thank my supervisor Prof. John McColl who has been a mentor and a great source of encouragement. His advise and wealth of experience made this project a great success. I am also most grateful to his wife, Isabel, and daughter, Ruth for their support and prayers in the last three years and their wishes for me for the future. May God bless you all abundantly.

I am grateful to my wife, Edith, the love of my life and the mother of my children. She has been a blessing and great pillar of support and encouragement. Thanks very much for your love and understanding and making that house a home. And to my bundles of joy, my three stars: - Victor, Vincent and Chibuikem, I bless you all for being there with your welcoming smiles each time I walk into the house.

To my father, Late Mr. Patrick Animaoke Eze, am grateful. How I wish you are living to see this day when your lifelong prayers are answered. The Lord called you home on June 25th 2006, barely eight months into my research. I remember your tears when I could not be registered in the secondary school due to lack of funds. I remember how you went about to everyone you know trying to raise money so I could go to school. And today, those tears are replaced with joy. I know you will be proud and happy where you are now. I am also grateful to my mother, Mrs. Virginia Uzoamaka Eze, who never ceased to pray and wished that one day I will study abroad and be a doctor. This day has the Lord answered those age long prayers. Thanks also to my brothers for their prayers.

I will not fail to thank my colleague, Boikanyo Makubate, and his wife for accommodating me for the first few days on my arrival to Glasgow and also for being a good friend through the years. May God bless and reward you immensely.

I thank all the academic and administrative staff of the Department who have been very kind and well-disposed towards me always. Thanks to Claire F, Ludger, Harper, Ben, Stephen, Mike, Agostino, Beverley and Kathleen for all your assistance.

I am very grateful to Dr. Gavin Shaddick, my external examiner, Prof. Adrian Bowman, my internal examiner and Prof. Marian Scot, my convener for a thorough examination of my work and for creating a cordial atmosphere throughout the viva.

To all the postgraduate students in the department, I say a big thanks, especially to my officemates Caroline, David, Firdaus and Maria who have been very helpful in various ways.

I am grateful also to members of the Victory Family and African and Caribbean Fellowships for their prayers and especially to Dr/Mrs. Komolafe for their care and support. Also to Rev. Fr. Joe Boyle and Pastor/Mrs T. Abu and other Pastors who prayed for this success. May God reward you all.

To all my friends too numerous to mention both in Nigeria and here in the UK, I say a big thanks for your good wishes and prayers. Most especially I am grateful to Mr/Mrs M.E. Obiegbu who have been a great source of encouragement for so many years now. May God bless you all.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

The first case of AIDS was diagnosed in Nigeria in 1985 in a young female teenager aged 13 years but was reported in 1986. This case was diagnosed in Lagos, the former capital city and the most populous city in Nigeria (16)(152). The Nigerian public received the news of the presence of AIDS in the country with doubt and disbelief. AIDS was perceived as the disease of American homosexuals - a disease of a distant land which had no place in Nigerian society. People, especially youths, were very sceptical of the presence of HIV in their environment. They saw the whole story as a hoax and a ploy by the Americans to discourage sex. This sceptism was entrenched in the many acronyms developed for AIDS one of which was *"**A**merican **I**dea for **D**iscouraging **S**ex"*. The government was then more interested in a debate about the origin of the disease, even denying that the disease was a threat to the Nigerian nation. This stance of the government and the general public was based on the fact that the first HIV positive individual identified in the country was a sex worker from one of the West African countries.

This led to an erroneous belief that the disease was foreign and incapable of affecting Nigerians (120). Consequently, the Nigerian public under-reacted to the news of AIDS and the government did virtually nothing to curb the spread of the disease (146), (152).

Due to the perception of the public about the disease together with the religious and cultural belief of most Nigerians, that death is pre-ordained and must come when it is due, there was little or no behavioural change in matters of sex and sexual practices (120) (162). Consequently, the AIDS virus spread silently and unnoticed through the sexual networks in all social classes, professions, age groups, genders, regions, zones, states, towns and villages in the country. The HIV/AIDS epidemic in Nigeria grew rapidly into a generalized epidemic when it was largely thought to be concentrated only within a few sub-populations (278). Twenty-two years after the first case was reported, the disease has become a massive epidemic which has become not only a health burden but also a socio-economic problem. It has affected every facet of Nigerian society and has eaten deep into the Nigerian nation.

Since 1991, the Nigerian government has conducted National Sentinel Surveys in order to monitor the trend and extent of the epidemic. The prevalence rate was 1.8% in 1991, rose to 5.8% in 2001 and declined slightly to 5.0% in 2003 and 4.4% in 2005 (93),(92), (90), (89). It appears from these estimates that the epidemic has peaked and is on the decline.

In 2001, the UNAIDS/WHO introduced a new version of its epidemiological model for all developing countries. The model known as Estimation and Projection Package (EPP) makes use of the surveillance data from each HIV sentinel site in estimating HIV/AIDS prevalence rates. Also the Spectrum Program utilizes the information on the birth rate, death rate and the output of EPP for the

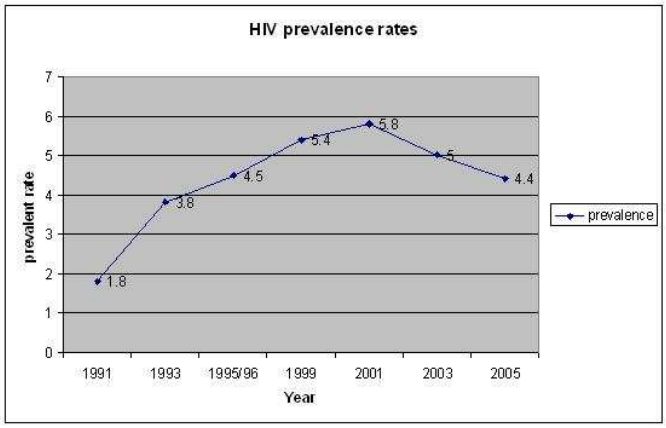

Figure 1.1. *National HIV prevalence rates estimated from sentinel survey data. Source: Federal Ministry of Health*

country in calculating the estimated number of people living with HIV, number of new infections, number of AIDS cases, number of AIDS deaths, number of orphans, etc(280).

These methods of estimating the national adult HIV/AIDS prevalence in Nigeria were based upon the outcomes of the Sentinel Surveys of pregnant women attending 85 selected antenatal clinics (ANC). The extent to which this data source is representative of the entire adult population of Nigeria is doubtful. This is because not all pregnant women attended antenatal clinics and not all women of adult age were pregnant at the time of the surveys. Also the selection procedure of the survey sites systematically excludes private clinics where many births occur. Therefore, national estimates based on these surveys rely on a fraction of women who attended the selected antenatal clinics. Also the Spectrum Program of UNAIDS and WHO makes use of vital rates which were obtained from a poor vital registration system. Studies (98), (167), (292) in some countries indicate that ANC estimates tend to overestimate the population based prevalence rates of females. Unfortunately, there have been no such population-based survey in Nigeria.

Given the observed limitations in the current estimation procedures in Nigeria, it is the intention of this research to examine other modelling approaches for HIV/AIDS epidemic that could be more suitable and give more accurate estimates for the Nigeria epidemic scenario.

1.2 The Country's Profile

Nigeria is a country in the west of sub-Saharan Africa with a population of about 140 million people (2), (192) and occupying a land area of 923,768 square kilometers. There are 373 ethnic groups each with its own language. Among these languages, Ibo, Yoruba and Hausa/Fulani are the major languages spoken by about 40 percent of the Nigeria population. English is the official language. The three main religions are Christianity, Islam and Traditional.

The country is divided into 36 states and a Federal Capital Territory (FTC). These states are then subdivided into 774 local government areas. The country's system of government is a three-tier- structure presidential system: the Federal Government, the State Governments and the Local Governments. The states are further grouped into six geo-political zones as follows:

Zone	States
North-East	Borno, Yobe, Bauchi, Gombe, Taraba and Adamawa
North-West	Sokoto, Kebbi, Zamfara, Kastina, Kano, Jigawa, and Kaduna
North-Central	Plateau, Nassarawa, Niger, Kogi, Benue, Kwara, and FTC
South-East	Anambra, Enugu, Ebonyi, Abia, and Imo
South-West	Ogun, Osun, Ekiti, Ondo, Oyo, and Lagos
South-South	Edo, Delta, Bayelsa, Rivers, Akwa Ibom and Cross Rivers

Table 1.1. *The grouping of the Nigerian States into Zones*

Nigeria is the world's most populous black nation with an annual population growth rate of 3.2% and a total fertility rate of 5.7. The infant and under five

mortality rates are 100 and 201 per 1000 live births respectively (193). The crude birth and death rates are 41 births and 17 deaths per 1000 population per annum respectively. The gross national income is $640 per capita and the gross domestic product per capita growth rate is -3.1% (287).

Nigeria, formerly a British colony, got her political independence on 1st October 1960 and was subjected to military rule for about 30 years. However, since 1999, Nigeria has been under democratic rule. The country has abundant human and natural resources. Despite this wealth of resources, most Nigerians live in penury. The 2007/2008 Human Development Report (HDR) ranked Nigeria as 158th out of 177 countries with Human Development Index score of 0.47, life expectancy index score of 0.359 and GDP index score of 0.404. Nigeria is about the 25th poorest nation in the world (284).

1.3 The HIV/AIDS Situation

The first case of AIDS was reported in Nigeria in June 1986. By the end of 1986, only 2(two) cases were officially reported. In 1987, another 2(two) cases were reported. 1988 saw 33 cases and ten years later, in 1998 alone 18,490 cases were reported, followed by 16,188 cases in 1999. 9715 and 3661 cases were reported in 2000 and 2001 respectively (278).

Notwithstanding the fact that the reported cases of AIDS were beset with incompleteness due to under diagnosis and underreporting, the data gives an idea of the inception and trend of the epidemic in Nigeria especially from the mid 1980's till 1999. Data for 2000 and 2001 appear unrealistic and may not represent the true situation of the disease at that time. The excitement of the

Location	Prevalence (%)	Group	Time Period	Source
Ibadan	21.3	Pregnant women attending ANC in the inner city	May-Nov 2001	a
Ibadan	34.3	Commercial Sex Workers	2002	a
Ondo State	12.8	Pregnant women attending ANC in 9 towns	2001-2003	a
Jos	8.9	Pregnant women	Oct. 2001-Jan.2003	a
Jos	39.2	Commercial sex Workers	1993-2002	a
Jos	11.7	Blood donors (males)	Jan.-Mar. 2004	b
Jos	14.1	Pregnant women attending ANC	Jan.- Mar. 2004	b
Zamfara	35.4	All tests	May 2003- Jan 2004	c
Zamfara	24.4	All tests	Jan- Dec 2003	d
Zamfara	35.4	All tests	May 2003 - Jan 2004	c

Table 1.2. *HIV Prevalence Rates estimated from various sources. a: Various Contributors to Regional and Global Conferences on HIV/AIDS: Compiled by the Nigeria Institute of Medical Research and NACA. Sept. 2003; b: Jos: Our Lady Apostolic Catholic Hospital; c: Zamfara: Duala Hospital, Gusau; d: Zamfara: Royal Medical Laboratories. Source:* ***UNDP (2004)***

discovery and the reporting of the new disease seem to wane with time. Data from the sentinel surveillance survey suggest higher figures for 2000 and 2001.

Estimates from the National HIV/Syphilis Sentinel Surveys give a better idea of the extent of the epidemic in the country over the period. Figure 1.1 shows the national prevalence rates from 1991 to 2005.

These estimates are the median of all the observed rates in the zones. This means that the explosive rates in some zones are masked by these national median rates. For instance in 2003, a site in Calabar had a prevalence of 12.7% but the national prevalence for that year was 5.0%. Overall, these surveys indicate that the trend of the epidemic peaked in 2001 and is on the decline in recent years. Table 1.2 shows estimates of prevalence in some towns, indicating that the official estimates from the sentinel surveys may not be very precise.

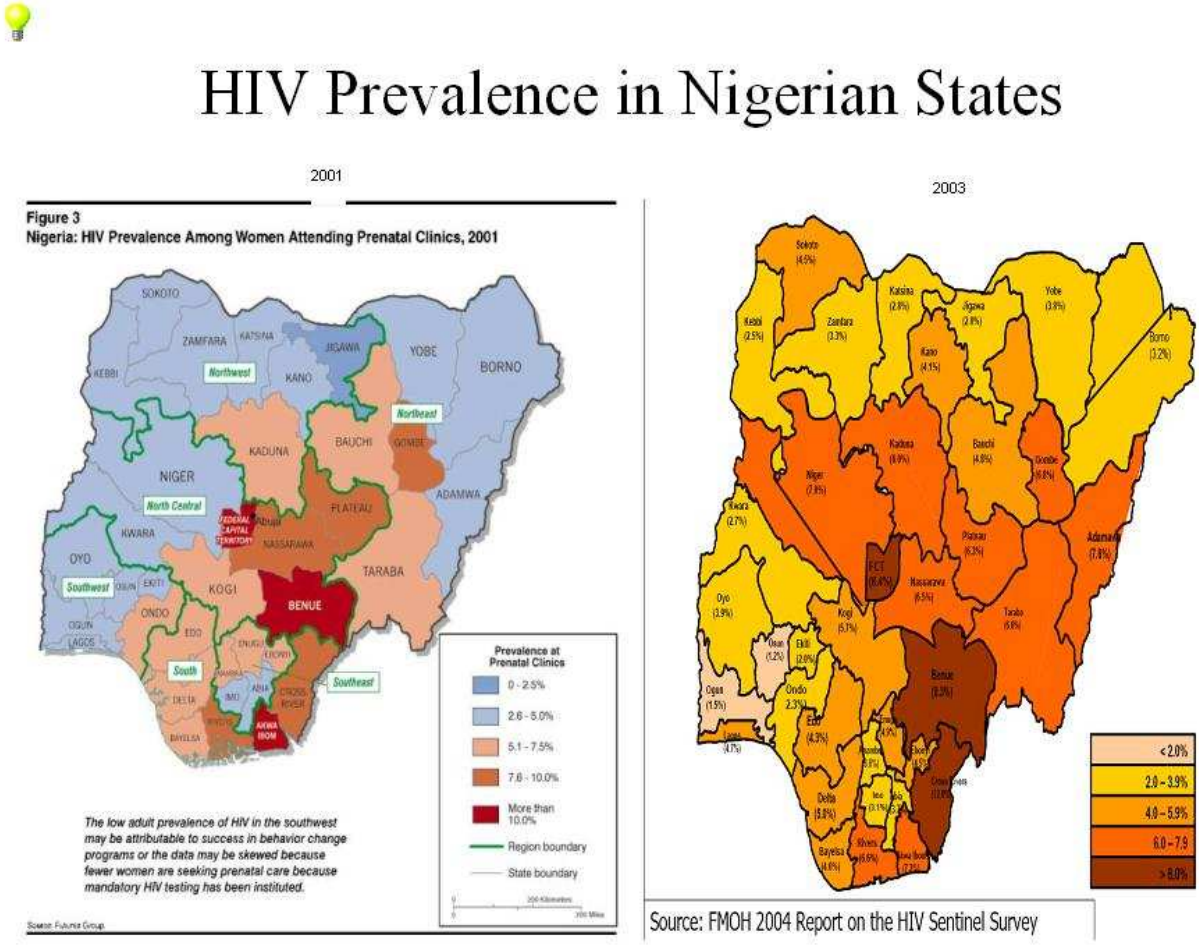

Figure 1.2. *Map of HIV prevalence in Nigeria by States (2001 and 2003). Federal Ministry of Health*

The spatial trend of the epidemic is more visible from the maps showing the outcome of the national HIV/AIDS sentinel survey. The spatio-temporal distribution of the disease indicates that at the earlier stage of the epidemic, more states in the southern part of the country were affected than those in the northern part. However, over the years, the epidemic appears to have concentrated in the middle belt (north central zone) and the south-south zone of the country. see Figure 1.2

Rather than decline as suggested by the data on reported AIDS cases and the sentinel prevalence rates, the UNAIDS/WHO estimates and projections based on

the surveillance data show that the number of persons infected is on the increase. The estimates are represented in Figure 1.4 with the two scenarios representing assumed levels infection intensity. Recent research evidence based on some factors that fuel the epidemic suggests that the epidemic is still emerging and the worst is yet to come.

According to USAID estimates in 2005, Nigeria ranks third amongst countries with the highest number of people infected with HIV and the number of AIDS deaths (see Figure 1.3). Currently, Nigeria is responsible for 20 per cent of Africa's total AIDS figure and 10 per cent of the world's, with an estimated 3.86 million people living with HIV/AIDS and 221,000 AIDS related deaths and 370,000 new HIV infections annually. About 540,000 patients are estimated to require Antiretroviral treatment and by 2006 only about 81,000 patients were estimated to be receiving treatment (283),(191),(279) . The average life expectancy declined from 53.8 years for women and 52.6 for men in 1991 to 46 and 47 years for women and men respectively in 2007 (297),(287).

60 per cent of the total infected population are young people below 25 years of age and about 61.5 per cent of all adult infections are women (279). About 1.3 million children are estimated to be living with HIV and AIDS which they contracted from their mother through breastfeeding or during birth. National survey indicated that by age 15, 25 per cent of young Nigerians had initiated sex and by age 18, 50 per cent of them have had sex (283). Given the large proportion of these youths (44 per cent of the total population), the low condom use and other forms of high risk sex, the propensity or the potential for the epidemic to grow higher than its present level is almost certainty unless honest efforts are directed towards the curbing of the epidemic.

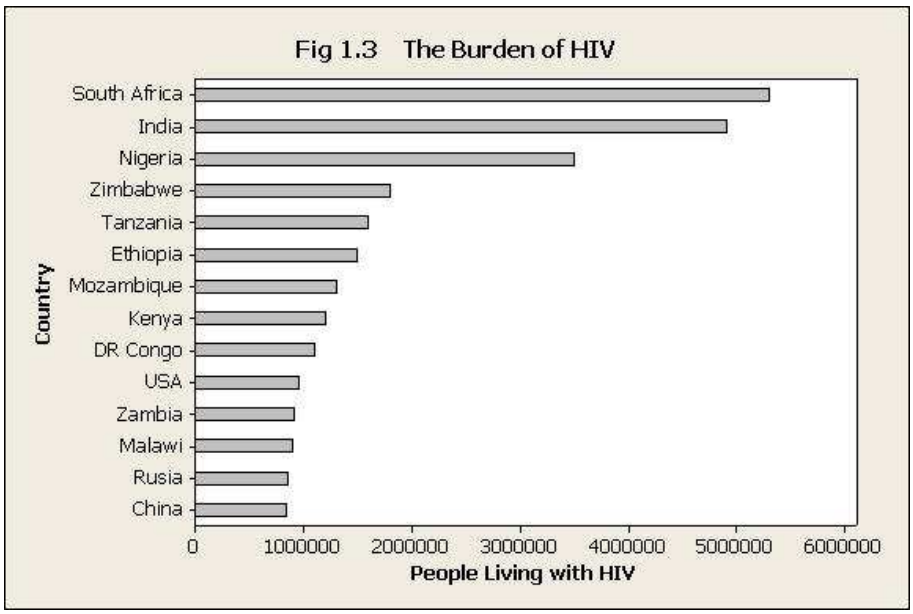

Figure 1.3. *The world's most affected countries. Source: USAID, 2005*

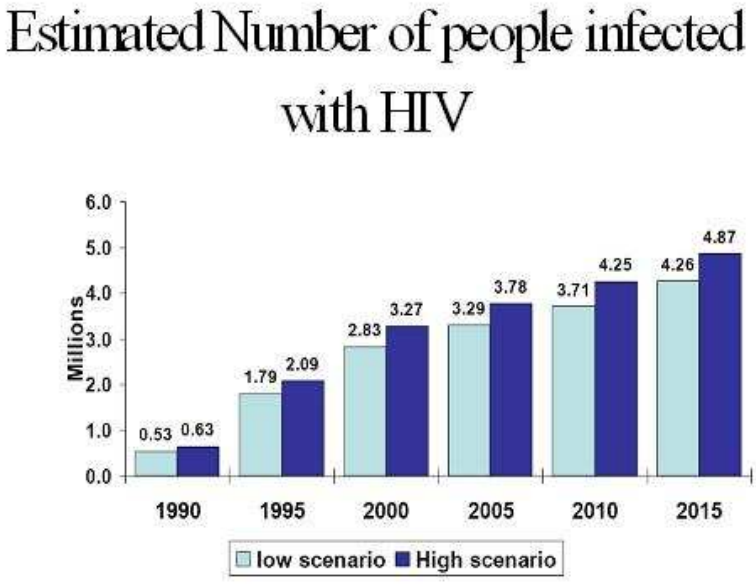

Figure 1.4. *Predictions of HIV infection in Nigeria. Source: UNAIDS 2004*

1.4 Modes of HIV Transmission in Nigeria

About 80 percent of HIV transmissions is via heterosexual transmission with blood transfusion accounting for 5 per cent of infections. Mother-to-child transmission, needle sharing, scarification, etc share the remaining 15 percent (283). However, recent findings reported that blood transfusion accounts for up to 10 per cent of new HIV infections in Nigeria (65) (29). This is due to the fact that not all hospitals have the technology to effectively screen blood (3) and the implementation of the blood policy guideline is restricted to only major Federal and State institutions and a few select private hospitals. It is a well known fact that a greater proportion of Nigerians patronize these private health facilities (152). There is a high demand for blood because of blood loss from surgery, childbirth, road traffic accident, anaemia and malaria in the country. Studies (76) (119) which investigated blood transfusion related HIV/AIDS in children screened for HIV between 1996 and 2001 at the University Teaching Hospital at Ile-Ife, and between 1989 and 1996 in the University Teaching Hospital at Enugu, found that about 66.7% and 68% respectively, of all children who were HIV positive were infected through blood transfusion. Most of the transfusions were done in the private hospitals and blood collected from private laboratories.

The predominance of heterosexual transmission in Nigeria was confirmed by several studies that focused on some high-risk and vulnerable groups. A survey conducted by the Federal Ministry of Health in 2003 and 2004 gave the information in Table 1.3.

Group	Prevalence(%)
Sex Workers	35 - 66%
Long distance truck drivers	20 - 25%
Sexually Transmitted Disease Patients	11.5 -13%
TB Patients	17%
Injection drug users	8.9%

Table 1.3. *Prevalence in Vulnerable Groups*

1.5 Driving force of the epidemic

The major driving force of the epidemic is basically the behavioural attitude of Nigerians in matters of sex. Some of the observed factors inducing the spread of the pandemic are: rampantly high risk sex, low risk perception, high poverty levels, harmful traditional practices, high stigma and discrimination of patients, Low levels of education, high levels of sexually transmitted infections, mother to child transmission and blood transfusion.

Of all these factors, poverty is the core causative factor. The booming commercial sex trade, international human trafficking (especially of young girls and women), the massive rural-urban migration with its social consequences and some cultural practices like wife inheritance are fueled by poverty. According to UNDP (283), poverty manifests itself in Nigeria in various ways, namely:- human poverty, physiological deprivation, income poverty, poor macroeconomic performance, negative impact of public expenditure and social exclusion. All of these, combined with pandemic corruption and fiscal indiscipline, have compelled certain risky behaviour on the citizenry as they struggle to survive. Hence, the risk of HIV assumes a lower priority as people are more concerned about immediate consequences of survival than the chances of contracting HIV where the effects are not immediate. Therefore, poverty increases the vulnerability to HIV infection and the speed and scale of the epidemic. Given the deteriorating health,

education and social services, the curbing of the epidemic is impracticable. Malnutrition and compromised immune systems due to exposure to other diseases may make people more susceptible to HIV infection. Therefore, the relationship between HIV/AIDS and poverty is "bi-directional" (134), a relation which Whiteside (22) aptly described as a poverty/epidemic circle: poverty increases the spread of AIDS and AIDS increases poverty.

Poverty in Nigeria is worrisome and paradoxical. Nigeria has all the resources to be one of the richest and most advanced, but sadly, it is among the poorest 25. It is the 6th largest producer of oil and yet it is the poorest OPEC country (276). It is estimated that about two-third of Nigerians live below one dollar per day and about 85 per cent of Nigerians are vulnerable to poverty. Also, there appears to be a link between inequality and AIDS (22), (215).

The National HIV/AIDS and Reproductive Health survey (NARHS)(91) of 2003 indicate that about 9% of women aged 15-49 years and 18.4% of men aged 15-64 years engage in extra and premarital sexual activity. The survey also revealed that youths are at greater risk as about 14% of female and 25% of male youth respectively engage in non-marital sex. Also, only about 32% of women and 50% men use condoms during risky sex.

Another factor that may fuel the epidemic in the country is the wide range of traditional practices such as wife hospitality, spouse sharing (180) , polygamy, wife inheritance and concubinage. Studies show that HIV prevalence is high in communities (example Benue, Kogi, and Nsukka) where these practices are common. Other cultural practices that may be privy to HIV transmission are female genital cutting (25) and some traditional Healers' practices (78)

1.6 A description of available data on HIV/AIDS in Nigeria

Below is a brief description of some available sources of HIV/AIDS data in Nigeria. The list is however not exhaustive, the sources described here are the major ones which have a national outlook except for some few localized surveys on sexual networking.

1.6.1 National Sentinel Surveillance Data.

The most comprehensive national data on HIV/AIDS in Nigeria is the Sentinel Surveillance data already mentioned above. This data is generated from a population-based survey of the prevalence of HIV/AIDS in all the zones, states, major towns and some rural areas in the country. The target population of the study is pregnant women aged between 15-49 years attending antenatal clinics (ANC) in selected health facilities in all the states in the country.

At least two sentinel sites are selected in each state and the selection of these sites depends on their meeting some set criteria which include inter-alia, availability of functional antenatal clinic services with qualified and willing staff. The number of sites selected per state is proportional to the population of the state.

The first HIV Sentinel Surveillance in Nigeria was conducted in 1991 in 9 states and covered 44 sites. The second was in 1993 which covered 17 states and 64 sites. 1995 and 1999 surveillances were in 21 states and 84 sites and 18 states and 74 sites respectively. 2001, 2003 and 2005 sentinel surveys covered the entire 36 states with a total of 85 sites.

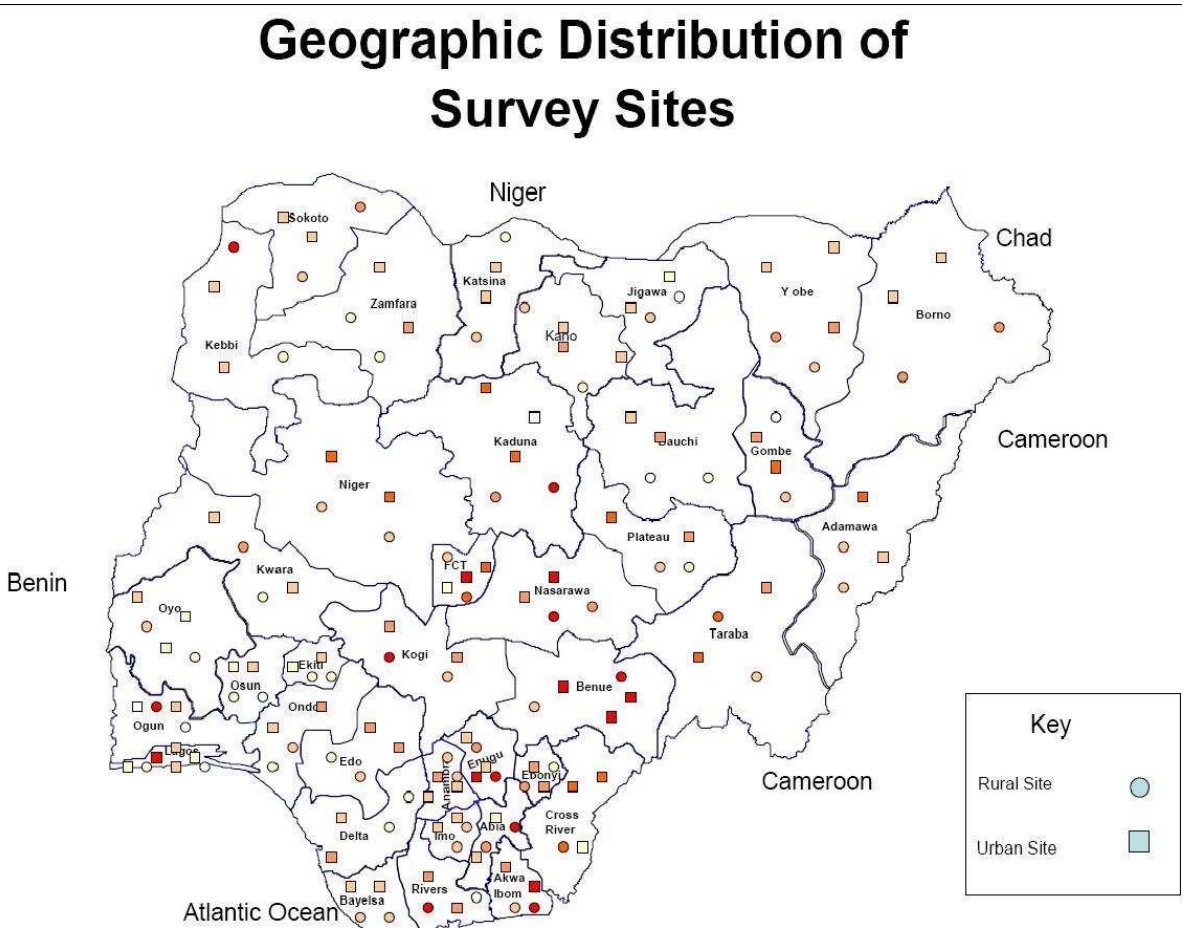

Figure 1.5. *Distribution of HIV Sentinel Survey Sites. Source: Federal Ministry of Health*

The targeted sample size per site in the recent surveys was 300 and the sampling procedure used was consecutive sampling whereby clients who met the eligibility criteria were selected one after the other until the desired sample size is achieved. A woman is said to meet the eligibility criteria if she is aged 15-49 years, pregnant and is presenting herself for the first time for booking in the antenatal clinic during the survey period.

The unlinked anonymous method of blood sample collection is usually adopted, using syphilis screening as the entry point. Information on zone, state, site , age and marital status of the patients are also collected.

Earlier surveys collected information on the HIV status of some population subgroups like the commercial sex workers (CSW), STD clinic attendees, and tuberculosis (TB) patients. These population subgroups were removed from the sentinel exercise since 1999.

The data is tabulated by zone, state, site, rural/urban status, and age. Since 1991, the national prevalence rate is estimated from the information obtained from these surveys. In a recent study of the quality of sero-surveillance in low and middle-income countries (240), Nigeria's surveillance system was adjudged as fully functioning based on scores allocated to certain information on the survey from 2001 through 2007.

Although the sentinel surveillance provides the bulk of the data upon which national estimates are based, it is not without limitations. The extent to which the sentinel population is representative of the entire population is questionable. This is because not all women aged 15-49 were pregnant at the time of the survey and not all of them who were pregnant attended antenatal clinics. The surveyed sites were only public clinics and given the poor facility situation in these clinics,

many who can afford it prefer private clinics, others may opt for the traditional health practitioners.

1.6.2 Reported cases of AIDS

The Federal Ministry of Health publish data on the reported cases of AIDS. This data is collected from various hospitals and published centrally. The data is also published annually in the UNAIDS/WHO Epidemiological fact sheet. Surprisingly, data from the two sources differ slightly.

AIDS case reporting in Nigeria is fraught with low reporting as not all cases are diagnosed because of poor health facilities and many patients may not visit the public health clinics throughout the time of their illness, either due to poverty or preference for traditional or religious health outfits. Reporting delays is another set back of the data. It takes a very long time for the Federal Ministry of Health in Abuja to receive data on HIV/AIDS diagnosis from the States ministries.

A more comprehensive data on reported cases of HIV/AIDS in Nigeria is the data published by the Nigerian Institute of Medical Research (NIMR), Federal Ministry of Science and Technology (83). The institute surveyed 659 health facilities comprising of 289 public and 370 private hospitals and 398 laboratory facilities made up of 181 public and 217 private laboratories in the six geopolitical zones of the country. The survey aimed at constructing a national HIV/AIDS database by collating and articulating reliable qualitative and quantitative national data on HIV/AIDS by pooling together all existing epidemiological, clinical, socio-economic data and scientific publications on various aspects of the epidemic in the country. Data was therefore retrieved from the records of the 1057 health and laboratory facilities on all diagnosed HIV infections, AIDS cases

and AIDS-related deaths between 1989 and 1999.

1.6.3 Data from HIV/AIDS centres of excellence

Some University Teaching Hospitals are designated centres of excellence for research and treatment of HIV/AIDS. They publish data mostly on the number screened and the number testing positive for HIV. The data is extracted from the registers kept in the HIV laboratories. The register contains some demographic information of the patients like name, sex, age, place of residence and occupation which are usually not published. Other personalized information on the patients can be obtained from the records of the patients in the HIV/AIDS clinic.

The data may not be very representative of the general population. The number screened and the cases detected may only be a small proportion of the individuals and total cases within the catchment area.

1.6.4 Other surveys in collaboration with some international agencies

Some international organizations collaborate with governmental and non-governmental structures in Nigeria in conducting surveys in some specific aspects of HIV/AIDS. Some of such surveys are the National Demographic and Health Survey (NDHS)(194), (193), the National HIV/AIDS and Reproductive Health Survey (NARHS) (91) and the Behavioural Surveillance Survey (BSS)(88).

The NDHS was conducted in 1980, 1990 and 1999 and 2003. The survey was not specifically designed to monitor sexual behavioural changes in HIV/AIDS scenario but was designed to solicit information on the socioeconomic and health

conditions of selected respondents throughout the nation. The BSS arose as a result of the need to monitor changes in HIV risk behaviours among Nigerians given the 5.4% prevalent rate recorded among pregnant women attending ante-natal clinics in 1999 (47). BSS which was conducted in 2000, focused on high risk and vulnerable groups like female sex workers, male truck drivers, male in-school youth and female in-school youth (aged 18-19 years). However, similar information were collected in the two surveys some of which were age at first intercourse, knowledge about AIDS, condom use, STI, etc.

The National HIV/AIDS and Reproductive Health Survey (NARHS) is a nationally representative survey conducted in 2003. 10,090 respondents were surveyed of which 5,128 were women aged 15-49 years and 4,962 were men aged between 15-64 years. The objectives of the survey was to provide information on levels of HIV preventive knowledge and behaviour, sexually transmitted diseases, stigma and discrimination against persons living with HIV/AIDS, etc. Data from this survey, like those from NDHS and BSS, may assist in predicting behaviour changes and provide ecological data for model construction.

1.6.5 Surveys on Sexual Networking

Some surveys on sexual networking were conducted in Lagos State (38), Ekiti district (121), south-western Nigeria (67), Calabar (263) and within the Nigeria Police Force (82) between 1990 and 1995. Data from these surveys may be useful in estimating the level of concurrent partnership in the population and in constructing network models. The problem with these data is that they were conducted at different places at different times, hence, it may not be easy to combine them for meaningful statistical modelling.

1.7 The Scope of the Research

The research reported in this thesis focuses on developing models that could describe or predict the HIV/AIDS epidemic in Nigeria. We shall however, first review current methods adopted by the WHO/UNAIDS in estimating HIV/AIDS epidemic in Nigeria.

Several attempts were made at obtaining more detailed data from various sources but the establishments in charge of the data were unable to grant the request.

The scope of the research depends largely on the nature of data available on HIV/AIDS in Nigeria. Also, the nature of the data depends on the source from which the data were collected. For this study, three sets of data are proposed:

- Reported cases of AIDS and HIV
- HIV screening results
- HIV Sentinel surveillance data

Each data set shall be used for the construction of suitable models that will describe or predict the HIV/AIDS epidemic in the country.

We shall explore modelling procedures developed for other countries and apply it or a modified version to the Nigeria scenario, taking into account the peculiarities of the country's epidemic.

Data on reported cases of AIDS and HIV is published centrally by the National Bureau of Statistics (NBS) and the Federal Ministry of Health (FMOH). Also, the Nigerian Institute of Medical Research (NIMR) published HIV/AIDS data

collected from 1057 health and laboratory facilities on all cases of HIV/AIDS diagnosed between 1989 and 1999 from selected public and private facilities. We adjudge the NIMR data to be more comprehensive than that published centrally by FMOH because most private health and laboratory facilities do not report cases to the FMOH since they have no legal obligation to do so. The FMOH data is not only beset with underreporting but also reporting-delays due to red tape and bureaucratical bottle-necks associated with collating and publishing official statistics in Nigeria. For instance, as of August 2008, reported cases of HIV/AIDS for 2003 to 2007 were yet to be published and those published for 2001 and 2002 were not only doubtful but misleading. Therefore, models constructed using such data should be amenable to adjustment or correction for these reporting errors. Another limitation is that the data is not classified by age and sex and is published as annual totals. The use of the NIMR data will, to some extent, solve the problem of reporting delays since the data were collected retrospectively with the date of diagnosis. Specifically, we shall used the NIMR data for the construction of back-projection models. We shall consider two aspects of back-projection methods: the parametric and the nonparametric back-projection. In doing this, we shall apply the data just as it is without correcting for underreporting. The next step will be to examine appropriate ways or methods for correcting for underreporting. In order to achieve this, we shall study attendance to, and HIV screening test results from, the HIV laboratory of a Centre of Excellence for the treatment of HIV/AIDS in the country over two periods of time. Period one is the time when patients were required to pay for testing and treatment and period two is the time when testing and treatment were free. This will be serve as a surrogate measure of underreporting due to poverty only.

At the asymptomatic stage of HIV infection, the infected person may be unaware of the infection and may spread the virus unknowingly within the population. These individuals who are not yet diagnosed could be described as "hidden". The determination of the size of this hidden population is of utmost importance as it explains the quantum of HIV infection in the population within a given period. We shall endeavour to estimate the size of the hidden population of HIV positive individuals in Nigeria from estimates from the back-projection models.

Data on results of HIV screening is individual level data which contain information on age, sex, date and result of the test. The age/sex distribution of patients will form the basis for stratification and each stratum will be studied and modelled separately. We shall employ the tool of logistic regression to compare the distribution of the burden of HIV infection among the the various sub-groups and also to study the effect of the interaction of some covariates on the spread of the disease. Information obtained from this data will assist us in determining the correcting factor for underreporting in the national data.

We shall study the spread of the epidemic among the six geo-political zones in Nigeria and attempt to fit spatial models for comparison. In particular, we shall study the tendency of the disease to cluster in geographic space. To achieve this, we shall apply the tool of spatial autocorrelation analysis, variograms, and kriging. The data proposed for this analysis are the outcome of the sentinel surveillance conducted biannually from 1991 up to 2005. The data were collected from pregnant women attending antenatal clinics in various sites in the 36 States of the Federation and the Federal Capital Territory (FTC), Abuja. The States are grouped into six geopolitical zones. Thus, the ANC data are hierarchically structured. We shall explore this structure in the study of the variations in the

intensity of the epidemic by employing the tools of multilevel analysis and spatial multilevel models. In estimating the multilevel models, we shall seek to utilize the restricted maximum likelihood, empirical and fully Bayesian approaches.

Chapter 2

Literature Review

A review of available literature reveals that not much work have been done on building models for HIV/AIDS epidemic in Nigeria. However, the UNAIDS/WHO over the years have developed some general models which are applied to various countries depending on the nature of the epidemic.

2.1 UNAIDS/WHO methods

This section reviews all the methods used by UNAIDS/WHO (280), (282) in obtaining and projecting estimates of HIV/AIDS prevalence in Nigeria.

HIV prevalence estimation

The HIV prevalence rate in Nigeria is estimated by applying a *general* formula on data from the HIV Sentinel Surveys (HSS) of women aged 15-49 years who attended antenatal clinics (ANC). The sample estimate of the prevalence rate is obtained by dividing the number of positive cases found in the sample

by the sample size. To estimate the HIV prevalence in the adult population (15-49 years), the prevalence rate for sampled women is multiplied by the total adult (15-49 years) population. This method was also used by UNAIDS/WHO in estimating the seroprevalence rate in high-risk groups or any specific group of interest. The method is based on the assumption that the HIV prevalence obtained from the antenatal clinic attendees, with adjustment for the male to female ratio, is a surrogate for HIV prevalence in the total population of aged 15-49years old. Recent studies (292), (98),(167) have confirmed that the ANC estimates generally overestimate the population-based survey prevalence, especially for younger women and men. Gouws et al(98) found that the ANC overestimates the population-based survey prevalence by about 20% and recommended that HIV prevalence derived from the ANC surveillance data be multiplied by about 0.8 to adjust for overestimation in countries where population-based HIV surveys have not been conducted.

Projection methods

The following methods were used by UNAIDS/WHO to project estimates of HIV/AIDS in Nigeria.

Delphic survey method This method was used in early stage of the epidemic in the late 1980s to obtain projected estimates of the HIV/AIDS prevalence rate in Nigeria. Opinions were obtained from knowledgeable experts in iterative fashion. The average and range of their guesses were used as projections. While this method has the advantage of speed of information gathering and low cost, it is highly subjective and may produce estimates that are widely ranging. Also, it may be difficult to find experts in quantitative epidemiology who are very familiar with the demographics of the country. However, this method was useful

at the time when data on HIV/AIDS were scant in the country.

Mathematical and computer /simulation methods Used in the early 1990s for short and long term projections of HIV prevalence. The size of the risk population, the number of their current partners, their partner exchange rate and the rate of mixture between the risk population and the not-at-risk population were used as input parameters. The general uncertainty surrounding accuracy of these input parameters made estimation and projection of HIV/AIDS incidence and prevalence by this approach very unreliable.

Scenario/modelling This approach was used for the short term estimation and projection of AIDS cases and AIDS deaths. It was developed by the Surveillance, Forecasting and Impact Assessment unit of the former WHO Global Program on AIDS (GPA). A scenario is defined as an outline of any series of events, real or imagined. The HIV/AIDS scenario can be constructed with or without models to fit the observed HIV/AIDS data and trends. The procedure as outlined in the WHO document (296) is as follows:

- Available HIV seroprevalence data are assembled and analyzed and used to estimate the most recent pattern(s), prevalence and trends of HIV infection for a specific population.
- Different HIV patterns and prevalence levels can be constructed with some confidence based on these data and other epidemiological observations.
- Using an AIDS model, we can derive the annual and cummulative estimates and projection of AIDS cases/deaths and other HIV-related conditions, using the general HIV scenario(s) constructed above.

Epimodel This model was used to estimate past and current prevalence and

make short-term predictions of AIDS cases and AIDS deaths. It was developed particularly for countries where AIDS case reporting was incomplete and unreliable. It uses estimates of HIV prevalence at a selected point in time and distributes this prevalence by annual HIV-infected cohorts back to the estimated start of the epidemic along a selected epidemic curve. It then applies the annual progression rates from HIV infection to the development of AIDS to each of the annual HIV cohorts to calculate the annual number of adult AIDS cases and deaths. The use of a single point prevalence and time may yield biased AIDS estimates. High HIV prevalence estimates will produce high estimates of AIDS cases. The stage of the HIV epidemic will also affect the HIV prevalence used. Estimates of HIV prevalence at the increasing phase of the epidemic will be higher than that at the declining stage of the epidemic.

Estimation and Projection Package (EPP) This method is currently being used by UNAIDS/WHO (280), (296) (221) for the estimation and short-term projection of HIV/AIDS estimates. It is designed for a generalized or concentrated epidemic where more data are available and uses yearly HIV prevalence for at least five years for all population groups and some curve fitting parameters as inputs. Developed in 2001, it tries to find the curve that best describes the trend of national adult HIV prevalence over time. In a generalized heterosexual epidemic like Nigeria, the model is fitted to urban and rural HIV prevalence data of women attending ANC separately. The estimates are then combined to produce a national estimate. Four major parameter inputs are

- t_0 : the start year of the HIV/AIDS epidemic
- r : the force of infection. It is the summary of sexual contact and transmission probability.

- f_o : the initial proportion of the adult population that is exposed to the risk of infection
- ϕ : the behaviour or high risk adjustment parameter which determines the extent to which susceptible people who die of AIDS are replaced by people who were not at risk.

The population groups are

- X = not-at -risk population
- Z = at-risk population
- Y = infected population
- N = X+Y+Z = Total population

The change in these population group is given by the following differential equations

$$\frac{dZ}{dt} = f(\frac{X}{N})E_t - (\mu + \frac{rY}{N} + \iota)Z \tag{2.1}$$

$$\frac{dX}{dt} = 1 - f(\frac{X}{N})E_t - \mu X \tag{2.2}$$

$$\frac{dY}{dt} = (\frac{rY}{N} + \iota)Z - \int_0^t (\frac{rY_x}{N_x} + \iota_x)Z_x g(t-x)dx \tag{2.3}$$

where $f(\frac{X}{N})$ is the fraction of those entering the adult population (E_t) who enter

the at risk group Z and is defined as

$$f(\frac{X}{N}) = \frac{\exp\left[\phi(\frac{X}{N} - (1 - f_0))\right]}{\exp\left[\phi(\frac{X}{N} - (1 - f_0))\right] + \frac{1}{f_0} - 1} \tag{2.4}$$

$\iota = 1$ for the first year of the epidemic and 0 for another years f = proportion of those entering the adult population who enter the At-Risk group. If $\phi > 0$, the proportion of people entering the at-risk group is increased. g = density function describing the progression to AIDS death since HIV infection and is given as

$$g(x) = (\frac{\mu + \alpha x^{\alpha-1}}{\beta})exp[-\mu x - (\frac{x}{\beta})^{\alpha}] \tag{2.5}$$

where α is the shape parameter of the Weibull distribution fitted to the HIV survival times and β is the position parameter defined in terms of the median survival time, m, as

$$\beta = \frac{m}{[ln(2)^{1/\alpha}]} \tag{2.6}$$

It is recommended that α be predefined based on available empirical data and that three values of m be used each corresponding to slow, medium and rapid progression respectively. The fixed parameters of the model are

- crude adult (15+) death rate μ
- number entering the adult population at time t, E_t
- force of mortality due to AIDS, x years after infection

For a heterosexual epidemic like Nigeria, E_t can be defined in terms of HIV negative children

$$E_t = B^-_{t-15} l \tag{2.7}$$

$$B^-_{t-15} = b[X_{t-15} + Z_{t-15} + (1-\nu)\varepsilon Y_{t-15}] \tag{2.8}$$

l is the cohort survival proportion to age 15, b is the birth rate and ν is the probability of vertical transmission. ε is fertility reduction due to HIV infection. It is assumed that HIV positive births do not survive to adulthood. However, the projected number HIV positive births is given as

$$B^+_t = \nu\varepsilon Y_t \tag{2.9}$$

In the implementation of the EPP model , it is assumed that the parameters are fixed. EPP searches for the best values of T_o, f_o, and r that best fit the observed surveillance data. The best fit is obtained by minimizing the sum of squared errors between the model curve and the surveillance estimates.

Spectrum This is a programme that uses the prevalence projection produced by EPP to calculate other estimates like numbers of people infected, new infections, AIDS cases and AIDS deaths. These calculations are based on population estimates provided by the UN population Division and model patterns.

The vital registration system in Nigeria is not yet efficient. Most vital events

are not recorded. Therefore estates of birth rate and death rate used in this model may be misleading. The estimates of other parameters such as ϕ and the size of the at-risk population may not be reliable. The data used is the surveillance data which may not be very representative of HIV prevalence in the entire adult population.

Since 2001, the EPP and Spectrum has undergone several metamorphosis (271) all aimed at overcoming some observed limitations. The most recent version, EPP2007, incorporates uncertainty estimation(164) for generalized epidemic using the technique of Bayesian Melding. Prior distribution for the model parameters are specified using expert knowledge. Other major improvement on the program include changes in the urban-rural population ratio, calibration of HIV prevalence measured at ANC in countries with generalized epidemics, based on a comparison of HIV prevalence from ANC to national population-based surveys, longer survival of HIV patients due to antiretroviral therapy(ART) and the quantification of number of people eligible for ART (220),(282), (281).

2.2 Review of other methods

Given the limitations of the methods presently used for HIV/AIDS prevalence estimates, we were compelled to review methods applied in other countries with a view to checking their suitability to the Nigeria scenario. The following models were reviewed.

2.2.1 Extrapolation models

Under this method, AIDS incidence is modelled as a function of time. Time may assume any mathematical form; linear, polynomial, log-linear function, etc and extrapolated into the future (172), (173), (236). The problem with this model is that the extrapolated estimates are greatly influenced by the mathematical function chosen for time. Also, it is not possible to obtain projected estimates for HIV prevalence or incidence. This method also assume that the trends will remain unchanged which is unlikely.

2.2.2 Epidemic theory models

This model generally makes use of a partitioned population:–the susceptible, infective and not-at-risk groups (125), (126). The deterministic approach could be used to describe an epidemic in large populations with large number of people infected. The problem with this method is that it is not easy to determine precisely the proportions of individuals in each category, the extent of mixing between the susceptible and the infective, the partner exchange rate, concurrent partners, behaviour change parameters, etc. Also, some of these models are deterministic, they assume that once the initial conditions and parameters are specified, the prevalence function and infection curves are uniquely determined and can be found by recursive methods (230). The stochastic version of these models may be useful in describing epidemic in limited geographic areas or subgroups. The models are usually complicated with so many unknown parameters and involve assumptions that are often not verifiable. A detailed account of the procedures for deterministic and stochastic modelling of AIDS/HIV epidemiology can be found in Wai-yuan (304)

2.3 Backcalculation models

First proposed by Brookmeyer and Gail (227), the backcalculation method is used to reconstruct the historical infection rates that may have occurred and have generated the observed pattern of AIDS diagnosis. That is, the historical pattern of HIV infection curves or rates is reconstructed using information on observed AIDS incidence and knowledge of the incubation period distribution. Incubation period is defined as the duration between HIV infection and AIDS diagnosis. The reconstruction of the infection curve and the estimation of the time since infection, in turn, makes it possible to predict the future number of AIDS cases (227), (229), (257).The back-projection equation is formulated based on the fact that for an individual to be diagnosed with AIDS at time t, he must have been infected with HIV at some time s in the past and hence the incubation period is the time between s and t, that is $(t - s)$. This technique is widely accepted because of its efficient application and requires fewer assumptions and parameter inputs when compared with other modelling approaches like the epidemic theory models.

Generally, the back-projection model is given as

$$\mu_t = \sum_{s=1}^{t} \lambda_s f_{t-s,s} \tag{2.10}$$

μ_t is the mean AIDS incidence at time t, λ_s is the mean HIV incidence at time s and $f_{t-s,s}$ is the probability density function for someone infected at time s and diagnosed at time t. μ_t is known from the AIDS diagnosis and $f_{t-s,s}$ is known from other epidemiological studies (231), (139). λ_s is then estimated using a deconvolution process. However, the nature of each of these components

of the back-projection model has a far reaching effect on the outcome of the back-calculation procedure. We discuss below some of the effects for each component.

The infection curve (λ) The choice of the infection curve for λ may have some consequences on projected estimates (69), (97), (288), (257). For instance, DeGruttola and Lagakos (288) fit four different shapes of parametric infection curve to the US AIDS incidence data for the period 1981-1987 and found that the curves differ dramatically in the most recent years but give similar AIDS incidence in the distance past. The reason for this is not far fetched; the use of a strong parametric model, such as the exponential model, as the infection curve will produce estimates that fit the distant past of the AIDS data very well but will portray an exponential growth in the most recent portion even when the AIDS incidence data suggest otherwise. To overcome this, Brookmeyer and Gail (227) (229) Rosenberg and Gail (257) suggest the use of flexible models such as step functions with about four or five steps. The justification of this is that these models are sufficiently flexible so that later portions of the infection curve can vary independently of early values of the curve. Rosenberg et al (258) investigated the performance of step function models and found that they yielded the smallest percentage root mean square error and bias in the short-term projection of AIDS incidence and estimating cumulative HIV infections. The estimates of the cumulative HIV infections and the projection of the AIDS incidence are based on the integrals of the estimated infection curve, and given the discontinuities in step function models, the integral over the infection curve appear not to be feasible. Consequently, Rosenberg and Gail (257) suggested the use of spline functions for the infection curve

The incubation density $(f(t))$ Earlier work on HIV/AIDS backcalculation (227), (229), (257) assumed that the incubation distribution is stationary

over the time period. Also, the incubation distribution adopted for any model has an influence on the HIV incidence estimation (207), (6), (97), (256),(261). To investigate this Baachetti et al (207) used four different incubation period distributions to estimate HIV incidence through 1990 using AIDS diagnosis data and found that the estimates vary substantially in the most recent past (1987-90). According to Brookmeyer and Gail (103)(230), assuming a shorter incubation distribution leads to lower estimates of cumulative number of infections and a slow incubation distribution requires large values of infection curve to fit the AIDS incidence series while fast incubation distribution require small values of the infection curve. However, most authors (226), (229), (97) agree that use of backcalculation for short-term projection of AIDS incidence and estimates of AIDS incidence at the earlier part of the epidemic is very reliable and is relatively insensitive to the choice of incubation distribution. This is because in the first few years following infection, the estimated incubation distribution has increasing hazard and flexible models of infection curve can adapt to a particular incubation distribution to fit the AIDS incidence.

AIDS data series Since the AIDS incidence is an important component of the backcalculation model, it is required to be reliable, up to date and accurate. Unfortunately, the accuracy of the AIDS incidence series is affected by reporting delay and underreporting. Reporting delay is measured as the time between AIDS diagnosis and time when report of such diagnosis is received by the AIDS statistics coordinating agency. Often, this delay may range from few a months to a year or more. In Nigeria, it takes quite a long time for the Federal Ministry of Health to gather information on diagnosed cases of AIDS from the 36 states and the Federal Capital Territory (FTC), Abuja. For instance, as at July 2008, reported cases of AIDS for 2003 to 2007 were yet to be published partly because

the data were not yet collected from the states. Therefore, the use of such data in backcalculation calls for some form of adjustment on the data if backacalculated estimates and projections will be reliable. Some authors (73), (74), (49), (226), (232),(254), (138) have suggested various adjustment procedures. These adjustments are performed on the data before backcalculation is applied. Brookmeyer and Damiano (226) suggested that data delayed for up to 7 to 9 months be increased by 22% and those delayed for 16-24 months by 6%. The data proposed for backcalculation in this research is free from the limitations of reporting delay because they were extracted retrospectively from the records of health and laboratory facilities in the country and grouped based on the dates of diagnosis. What is of great concern in the data is underreporting. In the United States, it is estimated that about 15% of AIDS cases are never reported and consequently, the delay-adjusted AIDS data is further inflated by $\frac{1}{0.85}$ (50). The proportion of unreported cases in Nigeria is far more than that of the US due to the reasons listed in the first and sixth chapter . We are not aware of any study that has established the extent of underreporting in Nigeria. Therefore, we shall attempt to obtain an approximate underreporting rate using data collected from one of the centers of excellence for treatment of HIV/AIDS.

INCUBATION DISTRIBUTION.

The incubation distribution is a very important component of the backprojection model. The uncertainty surrounding the choice of this distribution has attracted the interest of very many statisticians. Various attempts have been made to derive an appropriate distribution that could reflect the true incubation distribution. The first of such attempts was made by Lui et al (139) who studied 100 transfusion-associated AIDS cases reported to the Centers for Disease Control

(CDC) as at April 1, 1985. The dates of infection were assumed to be the dates of transfusion with infected blood and they defined incubation period as the period from transfusion to the diagnosis of the first opportunistic disease associated with AIDS. Using a simple average, 2.6 years was estimated as the average incubation period for transfusion-related AIDS cases. In order to correct for length-biased sampling resulting from the fact that the sample does not include those exposed persons who have long incubation periods and have not yet been diagnosed, the authors assumed a family of probability densities to describe the incubation distribution and obtained a maximum likelihood estimate of mean incubation period as 54 months with 90% confidence bound of (2.6 years, 14.2 years). This study has the limitation that the dates of infection were determined retrospectively and it is difficult to estimate the probability that a member of an infected cohort would develop AIDS in a specified time period. Also, those with long incubation period were selectively excluded from the study. These limitations in the work of Lui et al (139) led to the need to follow up a cohort of individuals. Goedert et al (135) studied a cohort of individuals who were already infected with HIV but whose dates of seroconversion are not known and estimated that the cumulative probability of developing AIDS within 3 years of follow up is 0.36 . This type of study is widely known as a prevalent cohort study. Brookmeyer and Gail (228) considered the bias inherent in such study and show that failure to adequately adjust for duration of infection will under some conditions, bias relative risk estimates toward 1.0. Other improved studies that led to interval censoring in the estimation of infection times are the hemophiliacs cohort who were regularly seen at a treatment centre in Hershey, Pennsylvania and their serum samples stored since the mid 1970s (71), (136) and the San Francisco City Clinic Cohort study involving homosexual men enrolled in hepatitis B vaccine trial (187), (9). Most of the derived estimates of the incubation distribution were obtained using the

data from the three studies above.

The incubation period distribution $F(t)$ is defined as the probability that an HIV infected person develops AIDS within t years after infection. That is, if I is the random variable representing the incubation period, then,

$$F(t) = P(I \leq t)$$

and its probability density function is $\frac{dF}{dt} = F'(t) = f(t)$ and survival function $S(t) = 1 - F(t)$. Its hazard function $\frac{f(t)}{S(t)}$ is the risk of developing AIDS at time t after infection conditional on not having AIDS just before t. The hazard of progression to AIDS increases with time. Below are some parametric distributions that have been used to model the incubation distribution.

The Weibull Distribution Given the Weibull density function $f(t) = \alpha\lambda(\lambda t)^{\alpha-1}exp{-}(\lambda t)^{\alpha}$, where $\lambda = \frac{1}{\sigma} > 0, \alpha > 0, t > 0$, the incubation period distribution is given as

$$F(t) = 1 - e^{-\lambda t^{\alpha}} \tag{2.11}$$

and the hazard function is

$$\lambda(t) = \lambda\alpha t^{\alpha-1}$$

which is monotonically increasing if $\alpha > 1$ and decreasing otherwise. This model was used in deriving estimates of incubation periods among haemophiliacs (231), (45), among homosexuals (140), and blood transfusion recipients (139), (61),(87). See also Blythe and Anderso (209), Wilkie (63) and Boldson et al (85). The limitation of the model is that it assumes that the hazard function increases

indefinitely and is proportional to the power of time since infection which may not be true in the long run.

The Gamma Distribution

The *pdf* is given as

$$f(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{\Gamma(k)}$$

where k is the stages which infected individual must pass through before developing AIDS and $\lambda > 0$ is a constant representing the hazard function of transiting from one stage to the next. Many authors (85) (209) (212)(87) have applied this model in estimating the incubation period.

Log-Logistic Distribution

Lui et al (140) and Lawless and Sun (138) considered the log-logistic distribution for the incubation distribution. The simple form of the distribution is given as

$$F(t) = 1 - (1 + (\lambda t)^{\beta})^{-1} \tag{2.12}$$

$\lambda > 0, \beta > 0$. If $\beta > 1$, the hazard function increases initially until it reaches maximum and then decreases as $t \to \infty$. This distribution has the same behaviour as the Lognormal distribution in that it assumes that the logarithm of the incubation period follow a normal distribution. Rees (177) and Boldson et al (85) assumed a Lognormal distribution for the incubation distribution.

The Gompertz Distribution Wilkie (63) and Lui (139) modelled the incubation period distribution as Gompertz whose cumulative density function

is given as

$$F(t) = 1 - e^{-\alpha(e^{\beta(t-s)}-1)} \; for \; t \geq \; s \tag{2.13}$$

The Staging Models The progression from HIV infection to AIDS can be divided into stages that are random and unequal in duration. The hazard functions of transition may differ from one stage to the other this distinguish the stage model from the gamma model where the transition rates are assumed equal and constant. Brookmeyer and Liao (232) used the model as

$$F(t) = \int_0^t \lambda_1 \alpha s^{\alpha-1} e^{-\lambda_1 s^\alpha} \{1 - e^{-\lambda_2(t-s)}\} ds \tag{2.14}$$

The equation models the depletion of CD4+ cell. The time from infection to depletion of CD4+ to less than 200 CD4+ T cell is assumed to follow the Weibull distribution with hazard function $\lambda_1 \alpha t^{\alpha-1}$ and the time from less than 200 CD4+ to AIDS follow an exponential distribution with hazard function λ_2, then the incubation distribution is given as a convolution of the Weibull and exponential distributions given above.

Mixture models This procedure partitions the infected individuals into two groups and assign them different incubation distribution: α the proportion with incubation distribution $F_1(t)$ and $(1-\alpha)$ with incubation distribution $F_2(t)$, then the incubation distribution for the entire population of infected individuals

as

$$F(t) = \alpha F_1(t) + (1 - \alpha)F_2(t) \tag{2.15}$$

Lui et al (140) used the mixture model to study incubation period among homosexual men also, Auger et al (115) used a mixture of Weibull distributions in a study of the incubation period among maternally infected newborns.

2.3.1 Criticism of Backcalculation

Most criticisms of backcalculation are based on the limitations or uncertainties surrounding the components and assumptions of the model some of which we have discussed in the previous section.

The incubation distribution is the most criticized. For instance, most estimates of the incubation period were estimated from data obtained from the study of a few cohorts who may not be representative of the general population in which the distribution is applied. The transfusion-associated AIDS cases study (139), the Multicenter AIDS Cohorts Study (MACS) (15), (169), the San Francisco Men's Health study (208) which provided data for the estimation of the incubation period may not adequately represent the entire population. The infection times of the transfusion-associated AIDS cases were ascertained retrospectively after AIDS had developed, thus leading to right truncation. Also, the periodic testing of a cohort of hemophiliacs (136) and homosexual men in San Francisco (9) until they test positive and the ascertainment of the date of seroconversion from the interval defined by the latest screening test that was negative and the

earliest screening that was positive, give rise to doubly censored data. This censoring has been handled by imputing the seroconversion time as the midpoint of the interval (140) (9), (131). It has been shown (290), (214) that the midpoint approach will only be optimal if the seroconversion density is uniform over the censoring interval, but the intervals are often sufficiently long to make this assumption void. Other authors have used parametric (231) and nonparametric (290)(208) modelling of both the seroconversion and incubation distribution.

Initial formulation of backcalculation assumes that the incubation distribution is stationary through the period and that the probability of progression to AIDS is the same for all infected individuals. Studies have shown that progression to AIDS depends on age. Infants and the elderly have shorter incubation periods than other infected individuals. Also the treatment of HIV infected individual may lengthen the incubation period therefore faulting the stationarity assumption. Recent developments have modified the incubation distribution to accommodate changes in incubation period due to therapy.

The accuracy of AIDS incidence data is affected by underreporting and reporting delay. Delays in reporting affect the most recent AIDS data series. These two sources of errors may undermine estimates obtained using backcalculation. To overcome this limitations, some underreporting and reporting delay adjustments have been proposed (50), (75), (138) (226).

The AIDS incidence data contain little information about the most recent infections because of the long incubation period. The numbers of individuals infected in the last one or two years are not reliably reflected in the AIDS incidence data. This makes the model imprecise in estimating the recent infection rates .

Backcalculation models do not give information about future rates and assumes that there is no migration in the population.

2.3.2 Developments in Backcalculation Method

Advancements in backcalculation arise mainly in the attempt to overcome or rectify the limitations or uncertainties in backcalculation methods as reviewed in the previous section. The first issue of concern was the problem of data truncation caused by reporting delays of AIDS incidence data. In an effort to compensate for the effects of these delays, several authors(75) (254) (289)(206) (232) have suggested various ways of correcting such truncation and incorporating reporting delay distribution in backcalculation methods. Also the concern of whether the information contained in the AIDS incidence data is sufficient to reconstruct the HIV incidence curve led researchers to seek other sources of information to augment or substitute the AIDS data. Some authors have used a combination of AIDS data and information on date of AIDS diagnosis (273), date of HIV diagnosis (199)(250). Mariotto and Verdecchia (35) substituted AIDS mortality data for AIDS registered cases and Deuffic-Burban and Costagliola (44) considered the use of pre-AIDS mortality data instead of AIDS data. Chau and colleagues (102) (101) used data HIV positive test only in place of AIDS data. While Bellocco and Marschner (224) analyzed jointly the HIV and AIDS surveillance data.

The introduction of effective therapy for the treatment of AIDS resulted in the elongation of the AIDS incubation period distribution. This led to the adjustment of the incubation distribution to accommodate changes in the distribution due to treatment (259), (104), (260), (233). Therefore, rather than use a single incubation distribution, a family of distributions indexed by the calendar year

of infection such that $F(t)$ becomes $F(t|s)$. This gives the probability that an individual infected with HIV at calendar time s develops AIDS within t year of infection. Hence, incubation distribution can be obtained for different cohorts.

The recognition that the incubation period distribution may vary with duration of infection led to the development of the stage model which assumes that HIV infected individuals pass through stages before developing AIDS. There could be two stages (225), (232) or multiple stages defined by clinical criteria (171) or by multiple CD4 levels (155). Assumptions are made to incorporate treatment into the model which has the effect to reduce the hazard of transition from one stage to the other.

The stage models do not incorporate the effect of changes in AIDS surveillance definition in the hazard model. The time-since-infection (TSI) models (259) incorporate both treatment and the effect of redefinition of AIDS through the hazard model. Unlike the stage models, the TSI models do not assume reduced hazard due to treatment effect until some time has elapsed since infection. Thus, the efficacy function is assumed to vary with time since infection.

In a study to investigate the incubation period of AIDS in patients infected via blood transfusion, Medley et al (87) found that mean incubation time for children is much shorter than that of adult patients and the incubation period for patients older than 59 years is less than that of younger adults. Also Goedert and his colleagues (136) found that hemophiliacs over the age of 30 at infection are at higher risk of progression to AIDS than individuals 19 -30 years old at the time of infection. These findings led to the development of backcalculation models whose incubation distributions depend on age (255). Others (196), (46), (37) incorporated age as a covariate in backcalculation model because it carries information about incubation period of the individual.

Due to the lack of identifiability of HIV incidence curve Becker et al (197) (196) adopted a nonparametric approach to backcalculation using the maximum likelihood estimation procedure implemented in EM algorithm that incorporates smoothing at each step of iteration. The advantage of this method is that it gives the data greater power to determine the shape of the estimated infection intensity and avoids the assumption of parametric distribution of infection curve. Several authors (250), (102), (101) used this methods in conjunction with HIV data in backcalculation method.

2.3.3 Parametric back-projection

We use the term parametric back-projection to represent all back-calculation approaches where a particular functional form is assumed for the HIV incidence curve or the AIDS incidence curve. This includes the use of step functions, splines or any other form of parametric function. In particular, we shall review the works of Brookmeyer and Gail (227)(229)and Rosenberg and Gail (257).

- Let $T_0, T_1, T_2, \ldots, T_J$ be the time points (for example month, quarter, year) at which counts of AIDS cases are available.
- T_0 is the start of the epidemic and T_J is the latest time at which AIDS data are available.
- Partition the calender time into $J+1$ intervals as $(T_0, T_1], (T_1, T_2], \ldots, (T_{j-1}, T_j], (T_J, T_\infty]$. Since there is no available data for the last interval $(T_J, T_\infty]$, the total number of individuals infected is not known.
- Let Y_j be the number of AIDS cases diagnosed in the jth interval $(T_j - T_{j-1}), j = 1, 2, 3, \ldots, J$ and $Y_\cdot = \sum_{j=1}^{J} Y_j$

- Let Y_{J+1} be the unobserved counts for the last interval $(T_J, T_\infty]$
- The total number of individual infected with HIV as at T_J is $N = Y_\cdot + Y_{J+1}$

Let $\nu(s)$ be the infection curve which specifies the expected number of individuals infected with HIV in the time interval $(s, s+\delta s)$. And let the general families of all infection curves defined by a basis function $G = \{g_1(s), g_2(s), \ldots, g_I(s)\}$ be a basis set. Then $\nu(s) \in G$ if

$$\nu(s) = \sum_{i=1}^{I} g_i(s)\beta_i \tag{2.16}$$

The total number of persons infected in an interval $(t_{j-1}, t_j] \subset (T_0, T_J]$ is

$$\int_{t_{j-1}}^{t_j} \nu(s)ds = \sum_{i=1}^{I} \{G_i(t_j) - G_i(t_{j-1})\}\beta_i \tag{2.17}$$

where

$$G_i(t) = \int_0^t g_i(s)ds$$

Hence the total number of individual infected before T_J is computed as

$$N = \int_0^{T_J} \nu(s)ds = \sum_{i=1}^{I} \{G_i(T_J)\}\beta_i \tag{2.18}$$

Given the incubation density $f(t)$ and the incubation distribution $F(t)$, we can express the expected number of persons infected in the interval (T_{j-1}, T_j) as

$$E(Y_j) = \int_0^{T_j} \nu(s) f(t) dt$$

$$= \int_0^{T_j} \nu(s)\{F(T_j - s) - F(T_{j-1} - s)\} ds$$

$$= \int_0^{T_j} \sum_{i=1}^{I} g_i(s)\beta_i\{F(T_j - s) - F(T_{j-1} - s)\} ds$$

$$= \sum_{i=1}^{I} \beta_i \int_0^{T_j} g_i(s)\beta_i\{F(T_j - s) - F(T_{j-1} - s)\} ds$$

$$= \sum_{i=1}^{I} x_{ji}\beta_i \tag{2.19}$$

$$= X\beta \tag{2.20}$$

where

$$x_{ji} = \int_0^{T_j} g_i(s)\{F(T_j - s) - F(T_{j-1} - s)\} ds \tag{2.21}$$

$F(t)$ is the incubation distribution and for $t < 0$, $F(t) = 0$. This distribution is assumed known from other studies. Also the general family of infection curve G, is usually assumed to be either a step function (227) or spline function (257). Having assumed the form of F(t) and G, x_{ji} is computed and β_i is estimated via regression methods. In particular, the multinomial maximum likelihood , the quasi-likelihood and the Poisson likelihood approach will be applied in the

estimation of β.

Estimation when G is a basis of indicator function

Here the general family of infection curve G is defined by a set of basis functions $g_i(s) = I\{s \in (t_{i-1}, t_i]\}$, where I(A) is unity in the event of A and zero elswhere. Hence $\nu(s)$ as defined in equation (2.10) is a step function.

Let $\mu_j = E(Y_j)$ and $X = [x_{ji}]$ be the $J \times I$ design matrix. $\mu' = (\mu_1, \mu_2, \ldots, \mu_J)$ such that $\mu = X\beta$. Let p_j be the probability that an individual infected before T_J is diagnosed in the jth interval and it is given as $p_j = \frac{\mu_j}{N}$ and $Np = \mu$

To accommodate those who were infected before T_J but not yet diagnosed, we construct a $(J+1) \times I$ augmented design matrix X^* whose $(J+1)^{th}$ row is the vector $x'_{J+1} = (x_{J+1,1}, x_{J+1,2}, \ldots, x_{J+1,I})$. Hence $\mu^* = X^*\beta$ and $p_j^* = \mu_j^*/N$

The Multinomial Maximum Likelihood Estimate

We assume that the infection times of the N individuals are independently and identically distributed and that the number of diagnosis within the intervals follow a multinomial distribution such that

$$Y^* = (y_1, y_2, \ldots, y_j, y_{j+1}) \sim Mult(N, p_1, p_2, \ldots, p_j, p_{j+1})$$

and the likelihood function is given as

$$L(N, \theta, Y) = \prod_{i=1}^{J+1} \frac{N!}{y_1! y_2! \ldots, y_j!(N - y_.)!} p_1^{y_1} p_2^{y_2} \ldots p_J^{y_J} p_{J+1}^{y_{J+1}}$$

The log-likelihood is given as

$$\begin{aligned}\ln L(N,\theta,Y) &= \log N! - \sum_{j=1}^{J} \log y_j! - \log(N-y_.)! + \sum_{j=1}^{J+1} y_j^* \log p_j^* \\ &= \log N! - \sum_{j=1}^{J} \log y_j! - \log(N-y_.)! + \sum_{j=1}^{J+1} y_j^* \log(\frac{\mu_j^*}{N}) \\ = logN! - \sum_{j=1}^{J} \log y_j! - \log(N-y_.)! &+ \sum_{j=1}^{J+1} y_j^* \log \mu_j^* - (\sum_{j=1}^{J+1} y_j^*) \log N \\ = logN! - \sum_{j=1}^{J} \log y_j! - log(N-y_.)! &+ \sum_{j=1}^{J+1} y_j^* \log \mu_j^* - N \log N\end{aligned}$$

$$\approx logN! - log(N-y_.)! - N\log N + \sum_{j=1}^{J+1} y_j^* \log \mu_j^* \tag{2.22}$$

We note that N is implicitly a linear function of $\beta, N = \sum_{i=1}^{I} G_i((T_J)\beta_i$. Differentiating the log-likelihood above with respect to β, Rosenberg and Gail (257) obtained the following results;

$$\frac{\partial l_{MULT}}{\partial \beta} = (\ln \mu_{J+1} - \ln N + \psi(N) - \psi(N-y_.) - 1)\Delta + (N-y_.)\frac{x_{J+1}}{\mu_{J+1}} + X'\{diag(\mu)\}^{-1}y \tag{2.23}$$

Also, they derived the information matrix with respect to β as

$$I_{MULT}(\beta) = \{\frac{1}{N} - \psi'(N) + \psi'(\mu_{J+1})\}\Delta\Delta' + X^{*\prime}\{diag(\mu^*)\}^{-1}X^* \tag{2.24}$$

Where $\partial \ln N!/\partial N = \psi(N)$ is the digamma function and $\partial^2 \ln N!/\partial N^2 = \psi'(N)$

is the trigamma function (168). Δ is given as

$$\Delta_i = \sum_{j=1}^{J+1} x_{ji}$$

Using the score and the information given above, Rosenberg and Gail (257) obtained the multinomial maximum likelihood estimate using the updating scheme of the Fisher's scoring algorithm;

$$\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + I_{MULT}(\hat{\beta}_{(n)})^{-1} \frac{\partial l_{MULT}}{\partial \beta}|_{\hat{\beta}_{(n)}} \tag{2.25}$$

Quasi-likelihood Estimate

Using the iteratively reweighted least squares, the Quasi-likelihood is maximized by;

$$\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}(y - \hat{\mu}) \tag{2.26}$$

Where

$$\hat{\Sigma} = diag(\hat{\mu}) - \frac{\hat{\mu}\hat{\mu}'}{N}$$

Poisson regression

Here the observed AIDS diagnosis are assumed to be independent Poisson

variates with means $\mu_j = x_{ji}\beta_i$. The Poisson likelihood function is given as

$$\ln L = \sum_{j=1}^{J} y_j \ln \sum_{i=1}^{I} x_{ji}\beta_i - \sum_{j=1}^{J}\sum_{i=1}^{I} x_{ji}\beta_i - \sum_{j=1}^{J} \ln y_j!$$

and the Poisson regression parameters are updated by

$$\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + (X'\{diag(\hat{\mu})\}^{-1}X)^{-1}(X'\{diag(\hat{\mu})\}^{-1}X) \tag{2.27}$$

The incubation distribution

The distribution of the waiting time between HIV infection and AIDS diagnosis is assumed to follow the Weibull distribution (139) (231)(45) (140). That is

$$F(t) = 1 - e^{(-\lambda t^{\gamma})} \tag{2.28}$$

Estimation when G is a spline function

The basis set is assumed to be a quadratic spline. According to Rosenberg and Gail (257), the flexibility of the step function can greatly be enhanced by using a spline function $g_i(s)$ with knots at $t_l, l = 1, 2, \ldots, L$ as a basis for G with the requirements that $\nu(s)$ be continuous at the knots or that $\nu(s)$ and its derivative $\nu'(s)$ be continuous. Using the '+' function notation, they defined $\nu(s)$ as

$$\nu(s) = \sum_{j=0}^{n} \beta_{0j} s^j + \sum_{l=1}^{L} \beta_{ln}(s - t_l)_+^n. \tag{2.29}$$

Applying this to the US AIDS data, Rosenberg and Gail (257) assumed a single knot in January 1982 and letting $n = 2$, then

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2 + \beta_{12}(s - t_1)_+^2 \tag{2.30}$$

Hence, $g_1(s) = 1,\ g_2(s) = s,\ g_3(s) = s^2$ and

$$g_4(s) = \begin{cases} (s - t_l) & t_l \geq T_L \\ 0 & otherwise \end{cases}$$

Hence,

$$x_{ji} = \int_0^{T_j} g_i(s)\{F(T_j - s) - F(T_{j-1} - s)\}ds,\ i = 1, 2, 3 \tag{2.31}$$

integrating within the intervals of $g_i(s)$ and

$$x_{j4} = \int_{t_l}^{T_j} g_4(s)\{F(T_j - s) - F(T_{j-1} - s)\}ds \tag{2.32}$$

The estimate of the HIV population in the time interval i is given as

$$N_i = \int_{g_i(s)} \nu(s)ds$$

and the estimate of the HIV population through the years in which AIDS diagnosis data is available is

$$N = \int_0^{\tau_J} \nu(s)ds$$

2.3.4 Non-parametric Back-projection

The ordinary back-calculation defines the yearly, quarterly, or monthly incidence of AIDS as an expression in terms of the average incidence of HIV in the corresponding time period. This is as given in equation 2.10

$$\mu_t = \sum_{s=1}^{t} \lambda_s f_{t-s,s}$$

where μ_t is the mean AIDS incidence at time t, λ_s is the mean of HIV incidence at time s and $f_{t-s,s}$ is the probability density function for someone infected at time s and diagnosed at time t.

A major disadvantage of the parametric back-projection is the problem of identifying the functional form of the HIV incidence curve because different forms may be consistent with the observed AIDS incidence. In order to avoid this limitation, Becker et al (197) proposed an imposition of a smoothness restriction on a non-parametric form and estimates are obtained using the non-parametric maximum likelihood approach implemented in the EM algorithm. This approach has the following advantages:

- It avoids the assumptions of parametric distribution of the infection curve.
- It gives the data greater power to determine the shape of the estimated intensity curve.
- It overcomes the problem of identifying the form of the infection (HIV incidence) curve.
- It ensures that all estimated values of HIV incidence is non-negative.

- The estimator of the HIV incidence has an explicit formular.

The EM algorithm

If we let A_t be the number of AIDS cases diagnosed at time t and a_t be the observed diagnoses at time t such that $E(A_t) = \mu_t$. Also if N_{st} is the number of individuals infected with HIV at time s and diagnosed at time t and n_{st} is its realized but unknown value. Hence a_t is generated by n_{st} as

$$\sum_{s=1}^{t} n_{st} = a_t$$

We assume that each infected person has an incubation duration independent of the incubation duration of others and $f_{d,t}$ is the probability that the incubation period is d given that the person was diagnosed at time t. Also the HIV incidence $N_1, N_2, \ldots, N_\tau$ are assumed to be independent Poisson variates and $A_1, A_2, \ldots, A_\tau$ be distributed independently as Poisson.
Also, the conditional distribution of N_{st} given $A_t = a_t$ is binomial such that

$$E(N_{st}/A_t = a_t) = \frac{a_t \lambda_s f_{t-s}}{\mu_t}$$

The EM algorithm makes use of a more complete data set in maximum likelihood estimation. Here the HIV incidence data N_t is considered more complete than the AIDS incidence data since the AIDS data set depend on the HIV data.

Hence the maximum likelihood for the more complete set is given as

$$L(\lambda; n) = \prod_{t=1}^{\tau} \frac{(\lambda_t f_{d,t})^{n_{st}} e^{-\lambda_t f_{d,t}}}{n_{st}!}$$

$$\ln L(\lambda; n) = \sum_{t=1}^{\tau} \sum_{s=1}^{t} \{n_{st} \ln \lambda_t f_{d,t} - \lambda_t f_{d,t}\} - \sum_{t=1}^{\tau} \ln n_{st}!$$

$$\approx \sum_{t=1}^{\tau} \sum_{s=1}^{t} \{n_{st} \ln \lambda_t f_{d,t} - \lambda_t f_{d,t}\}$$

The expectation stage, the E-step, is given as

$$\sum_{t=1}^{\tau} \sum_{s=1}^{t} \left\{ \frac{a_t \lambda_t f_{d,t}}{\mu_t} \ln \lambda_t f_{d,t} - \lambda_t f_{d,t} \right\}$$

$$= \sum_{t=1}^{\tau} \sum_{s=1}^{t} \left\{ \frac{a_t \lambda_t f_{d,t}}{\sum_{s=1}^{t} \lambda_t f_{d,t}} \ln \lambda_t f_{d,t} - \lambda_t f_{d,t} \right\}$$

Becker et al (197) obtained the M-step (the maximization stage) as

$$\lambda'^{[L+1]} = \frac{\lambda_t^{[L]}}{F_{\tau-t,t}} \sum_{d=0}^{\tau-t} \frac{a_{t+d} f_{d,t}}{\sum_{i=1}^{t+d} \lambda_t^{[L]} f_{t+d-i,i}} \tag{2.33}$$

Where $d = t - s$, is the time since infection, $(t = 1, 2, 3, \ldots, \tau)$. a_t is the observed AIDS incidence at time t. At time $t < 1$, it is assumed that the disease has not yet emerged in the population and τ is the last time when reliable data

can be obtained and

$$F_{\tau-t,t} = \sum_{d=0}^{\tau-t} f_{d,t}$$

In order to obtain a smooth curve, smoothing step is attached to the EM algorithm above to obtain $\lambda^{[L+1]}$ as

$$\lambda^{[L+1]} = \sum_{i=0}^{k} w_i \lambda'^{[L+1]}_{t+i-k/2} \tag{2.34}$$

w_i is a symmetrical moving average weight and k is the window width. k is an even integer and for our analysis we choose $k = 2$.

Convergence

It is expected that the likelihood will increase at each iteration. In order to establish the maximum likelihood estimates, we incorporated a convergence criterion in the EMS iteration. The following criteria which are based on the parameter λ were used at various times.

$$\sum_{t=1}^{\tau} |\lambda_t^{[L+1]} - \lambda_t^{[L]}| < \epsilon \tag{2.35}$$

$$\sum_{t=1}^{\tau} \frac{|\lambda_t^{[L+1]} - \lambda_t^{[L]}|}{\lambda_t^{[L]}} < \epsilon \tag{2.36}$$

$$\frac{|\sum_{t=1}^{\tau} \lambda_t^{[L+1]} - \sum_{t=1}^{\tau} \lambda_t^{[L]}|}{\sum_{t=1}^{\tau} \lambda_t^{[L]}} < \epsilon \tag{2.37}$$

The last two criteria were used in searching for the convergence in the EM algorithm, while the first criterion was used in the EMS algorithm. In some cases, because of the imprecision of the estimates near the most recent time τ, we used $\tau' = \tau - 1$ as the upper limit in the convergence criteria above.

2.3.5 The modification of the non-parametric back-projection

The ordinary back-calculation method, as considered in the previous section, makes use of diagnosed AIDS data in reconstructing the HIV infection curve. This approach has the limitation of not predicting precisely the HIV incidence in the recent past due to the long incubation period between HIV infection and AIDS diagnosis. In order to overcome this limitation, Cui and Becker (250) and Chau et al (102) suggested the use of HIV data in back-calculation for estimating HIV incidence curve.

According to Chau et al (102), HIV data has the following advantages:

- It contains more information than the AIDS data set because not all HIV infected individuals will develop AIDS by the time of analysis but some of them may undergo an HIV test
- It is not affected by the redefinition of AIDS
- It is not affected by treatment effects as it is unlikely that individuals receive treatment before HIV diagnosis
- The induction period between HIV infection and HIV diagnosis is shorter than the incubation period which is the time between HIV infection and

AIDS diagnosis. This short period implies that HIV data set contains more information than the AIDS data set.

- HIV data are more available than the AIDS data.

The back-projection method as given in the last section is modified using HIV diagnosis data instead of AIDS incidence data (250)(102). Hence the modified back-projection is given as

$$\mu'_t = \sum_{s=1}^{t} \lambda_s f'_{t-s,s} \tag{2.38}$$

where μ'_t is the mean number of HIV positive diagnosis at time t, λ_s is the mean HIV incidence at time s and $f'_{t-s,s}$ is the probability density function of the induction period for someone infected with HIV at time s and diagnosed with HIV at time t. The induction period (D') is defined as the period between HIV infection and HIV diagnosis.

The induction probability density function can be derived from HIV diagnosis hazard function which according to Cui and Becker (250), can be expressed as an additive model through the natural hazard function for AIDS.

The natural hazard function associated with the waiting time between HIV infection and AIDS diagnosis (incubation period) is known and is given as

$$P(x/u) = \frac{f(x/u)}{1 - F(x/u)} \tag{2.39}$$

The hazard of an infected person taking a positive HIV test is assumed to arise from two sources:

- Routine HIV testing: This refers to tests performed for reasons other than illness. We assume that the hazard for this test is a constant ν.
- Symptom-related testing: Are tests performed because of illness and its hazard is assumed to be proportional to AIDS natural hazard function.

Hence the hazard function for HIV diagnosisis is an additive hazard model given by

$$p'(x/u) = \begin{cases} \nu + \gamma p(x/u) & if\ x + u \geq \tau_0 \\ 0 & if\ x + u < \tau_0 \end{cases} \tag{2.40}$$

where τ_0 is the time when HIV diagnosis data became available and the unknown parameter γ is the coefficient of proportionality.

In order to estimate the two parameters ν and γ, we need to obtain information on HIV diagnosis. If we define D' and D as the induction period for HIV

and incubation period for AIDS respectively, then the probability that an HIV infected individual may have HIV test before developing AIDS is given as

$$Pr(D' < D/u) = \int_0^\infty f'(x/u)[1 - F(x/u)]dx = P$$

where

$$f'(x) = \{\nu + \gamma p(x)\}e^{-\int_0^x \{\nu+\gamma p(w)\}dw} \tag{2.41}$$

and

$$f(x) = p(x)e^{-\int_0^x p(w)dw}$$

Simplifying, we have that

$$Pr(D' < D/u) = \int_0^\infty [\nu + \gamma p(x/u)]e^{-\int_0^x [\nu+(\gamma+1)p(w/u)dw}]dx = P \tag{2.42}$$

Also let R be the waiting time from infection to HIV diagnosis if only routine HIV tests were available and S is the waiting time from infection to HIV diagnosis if only test were conducted when symptoms occur, then

$$f_R(x) = \nu e^{-\nu x}$$

and

$$f_S(x) = \gamma p(x) e^{-\int_0^x \gamma p(w) dw}$$

are the probability density function for the routine (R) and the Symptom (S) testing respectively, and

$$p(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}$$

is the incubation period density.

Hence, the proportion of HIV positive tests from routine testing is given as

$$Pr(R < S) = \int_0^\infty f_R(x)[1 - F_S(x)]dx$$

$$= \nu \int_0^\infty e^{-\int_0^x (\nu+\gamma p(w))dw} dx \tag{2.43}$$

If we assume that HIV routine tests and HIV-related symptoms tests contribute proportionately (say by P^*) to make individuals go for HIV diagnostic test, then,

$$Pr(R < S/D' < D) = \frac{Pr(R < S \bigcap D' < D)}{Pr(D' < D)} = P^*$$

$$Pr(R < S \bigcap D' < D) = Pr(R < S/D' < D) Pr(D' < D)$$

$$= P^* P$$

Also

$$Pr(R < D) = \int_0^\infty \nu e^{-\nu x}[1 - F(x)]$$

$$= \nu \int_0^\infty e^{-\int_0^x [\nu + p(w/u)]dw} dx \approx PP^* \tag{2.44}$$

Therefore in estimating the two parameters ν and γ, we may solve equations 2.42 and 2.44 simultaneously (102) (250)

Since the induction period distribution is not affected by drug therapy, the different drug regimes as defined for the AIDS incubation distribution function does not apply here. Therefore, the induction distribution function is given as

$$F'(x/u) = \begin{cases} 1 - e^{-\int_0^{x+0.5} p'(w/u)dw} & if\ x + u \geq \tau_0, \\ 0 & otherwise \end{cases} \tag{2.45}$$

Hence the HIV incidence is updated as

$$\lambda'^{[L+1]} = \frac{\lambda_t^{[L]}}{F'_{\tau-t,t}} \sum_{d=0}^{\tau-t} \frac{h_{t+d} f'_{d,t}}{\sum_{i=1}^{t+d} \lambda_t^{[L]} f'_{t+d-i,i}} \tag{2.46}$$

where h_t is the number of HIV diagnosis at time t and f' is as defined above

2.4 Spatial Analysis

The main concern of spatial statistics is to account for observation correlational effects arising from geographic configuration of data (60). The geographical configuration of HIV prevalence rates is assessed with a view to investigating the presence of spatial autocorrelation in the distribution of the data. Griffith and Layne (7) assert that observations are correlated strictly due to their relative locational positions resulting in spillover of information from one location to another. Hence, spatial autocorrelation is defined as the relationship among a single quantitative variable that results from the geographical patterning of the areas in which the values occur. It is a measure of similarity of objects within an area, the degree to which a spatial phenomenon is related to itself in space(57). Therefore, the spatial distribution of HIV prevalence rates in Nigeria is determined by the arrangement of the site prevalence rates in space and the geographic relationships among them.

Spatial autocorrelation exist in two forms - positive or negative spatial autocorrelation. If it is positive, the HIV prevalence rate at a given site tends to be similar to the prevalence rate of a nearby site. Conversely, negative autocorrelation among the site prevalence rates indicates that dissimilar rates are in nearby or adjacent locations. We shall investigate whether there is this systematic spatial variation in the distribution of HIV prevalence rates. To do this, we employed the tools discussed below.

2.4.1 Cluster Analysis

According to Jain et al (178), cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Such that the within cluster variation is less than between cluster variation. Sites or States with similar prevalence rates were grouped into clusters in order to investigate the spatial structure of HIV prevalence rates in Nigeria. To achieve this, we used the model-based hierarchical agglomerative clustering in $S-plus^{©}$ using the ellipsiodal distribution S^{*} (86) as the clustering criterion because it easily adapts and is robust to the choice of the size and shape of the cluster.

The clustering procedure entails the calculation of the intervariable similarities or dissimilarities and the grouping of the variables according to their mutual dissimilarities (20).An exaple of a measure of similarity between variables i and j is given by $\sqrt{1-\tau_{ij}^2}$ where τ is the Pearson correlation coefficient. If d is the dissimilarity coefficient, then d is symmetric if $d(i,j) = d(j,i)$, non-negative and $d(i,i)$ is zero. d is metric if, $d(i,k) \leq d(i,j) + d(j,k)$ or ultrametric if $d(i,k) \leq max(d(i,k), d(j,k))$ (189). Hierarchical clustering methods can be described as approximating a dissimilarity by an ultametric dissimilarity

The distance measures used in computation of similaries or dissimilarities include Euclidean and the Mahalanobis distances. The Euclidean distance is the most popular and is commonly used to evaluate the the proximity of values in two or three-dimensional space. It however has the disadvantage of being influenced by large-scale differences between variables. Also linear correlation among variables can affect the distance measure. The Mahalanobis distance measure offers a solution to this problems by assigning different weights to different variables

based on their variances and pairwise linear correlations.

The choice of the number of clusters can be obtained by a visual inspection of the dendrogram for natural clusters in the data. This method is subjective and may not yield the optimal number of clusters. A rule of the thumb for the choice of an appropriate number of clusters k is given as

$$k \approx (\frac{n}{2})^{1/2} \tag{2.47}$$

where n is the number of objects(159)

The plot of the agglomeration coefficient against number of clusters can be used to choose the appropriate number of clusters. The appropriate number of clusters is found at the elbow of the graph. Also, incremental changes in the coefficient may be used as an indicator of the number of clusters. A large increase means that dissimilar clusters have been merged, hence, the number of clusters prior to the merger is the most appropriate (154). A measure of within-cluster homogeneity relative to between-cluster heterogeneity known as the cubic clustering criterion is a good indicator of the appropriate number of clusters. The number of clusters is indicated at the peak of the cubic cluster criterion.

Other methods for determining the appropriate number of clusters are the Akaike Information Criteria and the Bayesian Information Criteria when it is possible to obtain a likelihood function of the clustering model, for instance using the k-means model.

2.4.2 Moran's Index and Geary's Ratio Statistics

The Moran's I (217) and Geary's c (235) are the most commonly used global measures of spatial autocorrelation. They provide an indication of the nature and extent of spatial autocorrelation present in the HIV prevalence data. Both approaches require a measure of connectivity for all pairs of HIV sites or states. In this research, we adopted the binary adjacency weights such that $w_{ij} = 1$ if sites i and j are neighbours and zero otherwise. Other choices of weights can however be adopted (62).

The Moran's I statistic is given as

$$I = \frac{N \sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i \sum_j w_{ij} \sum_i (x_i - \bar{x})^2} \tag{2.48}$$

Under the null hypothesis, its expected value is given in equation 2.49 as

$$E(I) = \frac{-1}{N-1} \tag{2.49}$$

The test of significance is done using analytical expectation and variances based on the neighbourhood structure assumed in the spatial weighting and are asymptotically normally distributed. Hence, the test statistic, Z is

$$Z = \frac{I - E(I)}{S_{E(I)}}$$

The Moran's I is a spatial univariate extension of the Pearson correlation coefficient. It measures covariance between data pairs. Its value varies between -1 and 1, where values between 0 and 1 indicate a positive spatial correlation between variables, values between -1 and 0 indicate negative association, and values equal

to 0 means there are no spatial autocorrelation. Perfect correlations are indicated by values equal to 1 or -1.

A Moran scatter plot gives a visualization of the type and strength of spatial autocorrelation in a data series. It regresses a spatially lagged variable on the original standardized variable (165). The steps for the construction of the scatter plot is outlined in Griffith and Layne (7) as follows: (a) Center the variable X, $(X - 1^T X 1/n)$. (b) Compute $W(X - 1^T X 1/n)$, where W is the adjacency matrix of binary weight and 1 is a vector of 1's. (c) plot (b) on the y-axis against (a) in the x-axis. This gives the Moran scatter plot. (d) Regressing (b) on (a) using a non-intercept regression model gives slope of the regression line for the scatter plot. The Moran coefficient can be expressed a ratio of regression coefficients as

$$I = \frac{\hat{\beta}_{XCX}}{\hat{\beta}_{1C1}}$$

where $\hat{\beta}_{XCX}$ is the regression coefficient obtained in step (d) above and $\hat{\beta}_{1C1}$ is the regression coefficient obtained by regressing $W1$ on 1 using a non-intercept model.

The Geary c statistic given in equation 2.50 is always positive and has values ranging between 0 and 2. It emphasizes differences between pairs of observations, rather than covariation between them as in Moran's I. It is therefore more sensitive to differences in small neighbourhoods. Squared differences between one value and an outlier value will have a disproportionate effect on the coefficient. The coefficient tends to over-emphasize areas with large number of neighbours and underemphasize those with fewer number of neighbours. If the HIV sites are spatially unrelated with one another,the expected value of the Geary coefficient

will be 1. Values less than 1 entail positive autocorrelation, while values greater than 1 indicate negative autocorrelation

$$c = \frac{(N-1)\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2\sum_i \sum_j w_{ij} \sum_i (x_i - \bar{x})^2} \tag{2.50}$$

2.4.3 Variogram

A Variogram provides a measure of spatial correlation by describing how sample data are related to distance and direction. The estimator of the semivariogram function, denoted by $\gamma(h)$, was originally defined by Matheron (96) as

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (Z(s_i) - Z(s_j))^2 \tag{2.51}$$

where N(h) is the set of all pairwise Euclidean distances $s_i - s_j = h$, that is, the lag (spatial distance) between site s_i and s_j, $|N(h)|$ is the number of distinct pairs in $N(h)$ and $Z(s_i)$, $Z(s_j)$ are data values at spatial locations s_i and s_j, respectively. In our case, $Z(s_i)$ is the observed number of positive HIV cases (or its transformed form)in site (or state) s_i. Note that $2\gamma(h)$ is the variogram.

For the semivariogram or the variogram to be a valid parameter of the stochastic process, $Z(s_i)$ should be intrinsically stationary. If the spatial process is intrinsically and second-order stationary, then the covariance function is given as

$$C(h) = Cov[Z(s), Z(s+h)]$$

such that

$$C(0) = Cov[Z(s), Z(s+0)] = Var[Z(s)]$$

and $E(Z(s)) = \mu$

Under this condition, it is possible to write the variogram in terms of the covariance function as

Var[Z(s)- Z(s+h)] = Var(Z(s)) +Var(Z(s+h)) - 2Cov(Z(s), Z(s+h))
= 2[Var(Z(s) - C(h)]

$$= 2[C(0) - C(h)] = 2\gamma(h) \tag{2.52}$$

where $C(h)$ is estimated as

$$\hat{C}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (Z(s_i) - \bar{Z})(Z(s_j) - \bar{Z}) \tag{2.53}$$

and $\bar{Z} = \sum_{i}^{n} Z(s_i)/n$

If the process is intrinsic and not second-order stationary, the covariance function does not exist and therefore is not a parameter of the process. When this happens, the semivariogram is adopted (204). A plot of $C(h)$ against h is known as covariogram. The autocorrelation function is given as

$$\rho(h) = \frac{C(h)}{C(0)} = 1 - \frac{\gamma(h)}{C(0)}$$

A plot of $\rho(h)$ against lag h is called the correlogram. Both covariogram and correlogram are used as indicators of spatial autocorrelation at different lags.

The parameters of an empirical variogram are the sill, range and Nugget effect. The *nugget* is the value of the variogram at $h = 0$. That is, $\lim_{h\to 0}\gamma(h) \to c_0 > 0$, then c_0 is the nugget effect. It is the micro-scale variation causing discontinuity at the origin (96). However Cressie (185) insists that this is not mathematically feasible if we require that $\lim_{||h||\to 0} E(Y(s+h) - Y(s))^2 \to 0$. Hence, if continuity is expected at the micro-scale, the only possible reason for $c_0 > 0$ is measurement error, if number of measurements is fairly large. There is a problem of determining c_0 from data whose separations are too large to give accurate micro-scale information. In practice, it is determined by extrapolating variogram estimates from lags closest to zero.

Sites close to each other in geographic space are expected to have more similar values than sites located farther apart, therefore, the variation between sites increases as the lag h increases until it gets to a distance h where the correlation between the sites is almost zero. At this point, the variogram levels-off or becomes flat. This point is known as the *sill* or *amplitude* of the variogram. From equation (2.52), if $C(h) \to 0$ as $h \to \infty$, then $2\gamma(h) = 2C(0)$, the quantity $C(0)$ is the *sill* of the semivariogram. The *partial sill* is defined as the difference between the sill and the nugget effect, $(C(0) - c_0)$. The *range* is the lag at which the variogram reaches the sill. Beyond the point of the range, autocorrelation is zero.

If the covariance function or the semivariogram function depend solely on the absolute distance between points h and do not depend on direction, the function is termed *isotropic*. However, if it depends on both distance and direction, the function is said to *anisotropic*.

Robust Variogram Estimation

The variogram estimation as given in equation (2.51) has the disadvantage of being negatively influenced by outliers. In order to alleviate this negative impact of outliers, Cressie and Hawkins (186) suggested eliminating the squared differences from the Matheron estimator (equation 2.51) and replacing it with the fourth power of the square root of absolute differences and correcting for the resulting bias. It is given as

$$\hat{\gamma}(h) = \frac{\frac{1}{2}\{\frac{1}{|N(h)|}\sum_{N(h)}|Z(s_i) - Z(s_j)|^{1/2}\}^4}{0.457 + \frac{0.494}{|N(h)|}} \tag{2.54}$$

The Variogram Clouds This is the distribution of the variance between all pairs of points at all possible distance h. There are main variance functions commonly used in the construction of variogram clouds. They are; the squared difference cloud given as $(Z(s+h)-Z(s))^2/2$ and the square root-difference cloud expressed as $\sqrt{(}|Z(s+h)-Z(s)|)/2$. The variogram cloud is obtained by plotting any of these functions against the spatial lag h. When plotted in conjunction with box plots, it is a good diagnostic tool for detecting outliers, spatial trends and variability with increasing distance (237).

2.4.4 Fitting the theoretical Semivariogram

Various functions can be fitted to a semivariogram plot in order to explicitly specify the spatial similarity present the data. Models commonly used for this purpose include linear, exponential, spherical, Gaussian, power, circular, Bessel, rational quadratic, De Wijsian, and wave/hole (7). Most of these functions are

bounded but the linear and power functions increase without bounds. The choice of the model to be fitted depends on the shape of the empirical semivariogram and the researcher's belief as to the nature of the operating processes. Initial values of the semivariogram parameters are obtained by inspecting the the semivariogram plot. In this thesis, we adopted the spherical function for the semivariogram model and it is given as

$$\gamma(h) = \begin{cases} 0 & if\ |h| = 0 \\ C_0 + C_1[1.5(\frac{h}{r}) - 0.5(\frac{h}{r})^3] & if\ 0 < |h| < r \\ C_0 + C_1 & if\ |h| \geq r \end{cases} \tag{2.55}$$

where C_0 is the nugget, r is the range and C_1 is the sill. These parameters are estimated using either the maximum likelihood (ML), ordinary least squares (OLS), generalized least squares (GLS) or the weighted least squares (WLS) approach. The maximum likelihood estimate relies on normality assumptions and is affected by small sample sizes. The GLS has the advantage of making no assumptions about the distribution of the data and is more robust than the ML whenever the distribution of the attribute variable is misspecified (234). However, parameter estimation in the semivariogram requires the variance-covariance matrix which is not easy in GLS (185). The weighted least squares (especially the nonlinear) approach usually performs better than other methods (185), (64).

2.4.5 Kriging

The geostatistical method of kriging is used in this thesis for the purpose of spatial estimation of HIV prevalence estimates in Nigeria and to interpolate these

estimates onto a continuous surface of infection grid. Using the framework of spatial linear models (185)(201), we propose to obtain smoothed HIV prevalence estimates that can be represented on a map. The use of kriging is informed by the presence of spatial correlation in the data as established in the previous sections. The method is known to be useful in expressing properly the spatial correlation and overcomes areal bias problems, thus giving rise to better disease maps (201) (95) (262).

Different types of kriging methods are reviewed in Cressie (185), but the most common types are simple, ordinary and universal kriging. The simple kriging assumes that the spatial mean of the random field $\mu(s)$ is known, while ordinary kriging assumes that $\mu(s)$ is unknown but constant across locations (204). Universal kringing is an adaptation of ordinary kriging that accommodates trend. It can be used to obtain local estimates in the presence of trend and to estimate the underlying trend (237). Hence, universal kriging requires knowledge of both the trend model and a semivariogram or covariance function for the data. Usually, a polynomial trend surface is used. A high polynomial degree (large number of regression coefficients) is needed in order for the prediction to be locally adequate everywhere and capture simple spatial variation.

However, the choice of the variogram or covariance function and the presence or absence of a nugget effect have an impact on kriging. When the semivariogram is modelled without a nugget effect, then kriging leads to direct interpolation at the sampling sites and prediction equals the observation at the sampled sites. Also, predicted residuals are equal to model residuals and predictions at any other site have the tendency to shrink towards the value of the estimated trend surface at that place. When semivariogram with a nugget is used,a smoother prediction surface is obtained as predictions tend to be closer to the mean surface with less

residuals (201).

Ordinary kriging makes use of models of spatial correlation to calculate a weighted linear combination of observed data resulting in estimates of the observed location or specified unobserved location. Weights are choosen so that the average error for the model is zero and the modelled error variance is minimized(110).

Given the spatial data $Z(s)$ such that

$$Z(s) = \mu(s) + e(s)$$

$$e(s) \sim (0, \Sigma)$$

$$E[Z(s)] = \mu(s),\ Var[Z(s)] = \ \Sigma$$

Cressie (185) obtained the kriging weights for ordinary kriging in terms of the semivariogram as

$$\lambda' = \left(\gamma(s_0) + \mathbf{1}\frac{1 - \mathbf{1'}\Gamma^{-1}\gamma(s_0)}{\mathbf{1'}\Gamma^{-1}\mathbf{1}}\right)' \Gamma^{-1} \tag{2.56}$$

and the optimal linear predictor of $Z(s_o)$ was obtained as

$$p_{ok}(Z; s_0) = \lambda' Z(s) \tag{2.57}$$

with variance

$$\sigma^2_{ok}(s_0) = 2\lambda'\gamma(s_0) - \lambda'\Gamma\lambda$$

where $\Gamma = \gamma(h)$ and $\gamma(s_0) = \gamma(s_0 - s_i),\ i = 1, 2, \ldots, n$

2.5 Disease Mapping

2.5.1 Introduction

The interest here is in investing the spatial and non-spatial variability of the risk of HIV infection in Nigeria and to produce smoothed maps of prevalence rates. The true underlying distribution of HIV prevalence rates will be estimated and the spatial patterns displayed on the map. Thus, areas (sites and states) with unusually high risk can stand out distinctively and this may prompt risk assessment and resource allocation by the health policy makers (24). We shall also seek to determine the ecological association between zone-level covariates and risk of HIV infection using a model-based approach (48), (56),(130). Hence, it may be possible to identify the socio-economic and cultural effects that contribute to the variation in infection rates and also to improve the stability of estimates of the risk of infection for each site and state by taking into account their geographical locations in relation to other areas and the rate of infection in those areas (113). The study of geographical correlation aims at exploring geographic variations in exposure to some life-style or behavioural factors in order to understand the HIV/AIDS aetiology in the country.

Several epidemiological measures can be displayed on a map - the crude rate, standardized rate, statistical significance of local deviations of risk from the overall rates and other smooth version of the standardized rate derived using some model approach. The Standardized Morbidity (or Mortality) Ratio (SMR) is the common choice of epidemiologists.

2.5.2 The Standardized Morbidity Ratio (SMR)

Let O_i denote the observed number of positive cases of HIV in the ith site or state, E_i the expected number of HIV positive cases in the ith site or state which is known and the relative risk of HIV infection in the ith site or state be denoted by θ_i . The observed number of positive cases O_i is assumed to be distributed as Poisson with mean $E_i\theta_i$. Hence, the likelihood is given as

$$L(\theta) = \prod_{i=1}^{n} \frac{\exp(-E_i\theta_i)}{O_i}(E_i\theta_i)^{O_i} \propto \prod_{i=1}^{n} \theta_i^{O_i}\exp(-\sum_{i=1}^{n} E_i\theta_i)$$

And

$$lnL(\theta) = \sum_{i=1}^{n} O_i ln\theta_i - \sum_{i=1}^{n} E_i\theta_i$$

Differentiating with respect to θ and solving, we have

$$\hat{\theta}_i = \frac{O_i}{E_i}$$

where,

$$E_i = N_i \frac{\sum_{i=1}^{n} O_i}{\sum_{i=1}^{n} N_i}$$

Here, N_i is the number of women tested in each site or state within the period of the survey.

Therefore, standardized morbidity ratio is defined as the ratio of observed to expected counts of HIV positive cases in each site or state. The estimator

is unbiased but because it is based on a sample of size one, it is a saturated model estimate, as it leaves no degree of freedom for goodness of fit test. The display of SMR on a map can be accomplished using the chloropleth method (79) by classifying the SMR in to class intervals and assigning specific colours to each class. The determination of the class size is often subjective. One way of achieving this is to divide the SMR into equal range of classes. This method of classification is good when the distribution of SMR is not skewed but may be problematic when the data is highly skewed as most of the SMR may belong to one or few classes. Hence such display on the map may distort the spatial variation of the disease inherent in the population, as the majority of the areas may look similar when they are not. The SMR can also be classified using its percentiles but this method has the risk of classifying similar SMRs into different classes such that some areas will appear very heterogeneous on the map when indeed they are not. The search of a natural divisions in the distribution of SMRs is another method of obtaining class intervals of SMRs. These natural groupings or clusters do not assign SMRs arbitrarily into classes but may be determined through the use of statistical methods that minimize within-class variation. See Muechrcke and Muechrcke (219) for more details on the choice of class intervals.

The use of SMR as an epidemiological measure may be informative but has some limitations. Clayton and Kaldor (54) and Lawson et al (24) outlined some limitations of the use of SMR for disease mapping as follows:

- The variance of SMR is large in areas with small population and small in areas with large population with few cases. These may form the extremes of the map and may dominate its pattern because no account is taken of the varying population size over the map.

- It does not differentiate between regions where no event is observed.
- It does not reveal the underlying structure in the data and it is not parsimonious because of its saturated form.

The variance of the estimator of the relative risk is proportional to E_i, hence areas with small population have very large relative risk as the number of expected cases (E_i) in the denominator is small. Conversely, areas with large population have high values of E_i and hence small relative risk if the number of cases is small. The display of significance levels in place of the SMR on the map in order to avoid the shortcomings of SMR tends to give more advantage to areas with large populations since they are more likely to attain significance, even if excess risk is small (54) (13) ,(48).

To overcome these limitations, various models have been proposed by different authors. For instance Kelsall and Diggle (149), (150) proposed a non-parametric approach for the estimation of spatial variation in relative risk and Marshall (**?**) extended the linear Bayes methodology to disease mapping. Kafadar (158) considered smoothing separately the numerator and denominator of the relative risk taking into cognisance the assumed Poisson variation in the numerator and the variation due to counting and recording errors of the population at risk in the denominator. However it has been established by other authors, (150), that a common smoothing constant is preferred over separate estimation of different smoothing constants for the numerator and denominator. Downer (243) proposed the smoothing of estimates of disease rates by penalizing the Poisson maximum likelihood estimates using the inter-site distance penalty. MacNab and Dean (308) proposed spatiotemporal models that use autoregressive local smoothing across the spatial dimension and B-spline smoothing over the temporal dimension

with the object of identifying temporal trends and the production a series of smoothed maps. Recent advances in disease mapping have availed the Bayesian methods in arriving at smooth disease rates.

2.5.3 Empirical Bayesian Approach

The need to identify extreme rates for areas with small population or rare diseases led to the development of empirical Bayesian estimation in disease mapping. According to Marshal (244), Efron and Morris (30),(31), (32) appear to be the first to suggest this approach for regional estimates of disease rates. Several authors (246), (54), (130) and (211) have proposed modified models in which observations at site level are mutually independent but are conditional on an underlying spatial process. Tsutakawa et al (246) in particular, derived improved estimates of cancer mortality rates using empirical Bayesian (EB) approach that treats the true rates as samples from an unknown prior distribution that needs estimation.

The Bayesian approaches combines two types of information: the information provided by the observed number of HIV cases in each site or State described by the Poisson likelihood $L(\theta/O)$ and prior information on the relative risks specifying their variability in the overall map, summarized by their prior distribution (13). The empirical Bayesian estimator assumes that the relative risks follow some distribution $f(\theta)$ and Bayesian inference about the unknown relative risk θ is based on the marginal posterior distribution

$$g(\theta/O,\gamma) = \int_i g(O/\theta) f(\theta/\gamma) d\theta$$

where O is the observed HIV counts and γ are the unknown hyper-parameters

and the probability density function of the relative risk is given by $f(\theta_i)$. The likelihood function of the relative risks for the observed HIV counts is the product of n independent Poisson distributions given as

$$g(O/\theta) = \prod_{i}^{n} g(O_i/\theta_i)$$

The prior distribution reflects prior beliefs about the relative risks and is parameterized by the hyper-parameter γ give as $f(\theta/\gamma)$. The empirical Bayes approach employs the estimates of the hyper-parameters which are obtained by maximizing the marginal likelihood of O/γ

$$g(O/\gamma) = \int_{i} g(O/\theta) f(\theta/\gamma) d\theta$$

That is, the estimates of the hyper-parameters are obtained from the data. If the areas are independent, the marginal posterior distribution is also independent and is given as

$$g(\theta_i/O_i, \gamma) \propto g(O_i/\theta_i) f(\theta_i/\gamma)$$

Hence, empirical Bayes estimate of the relative risk is the posterior mean or median evaluated at the maximum likelihood estimate of the hyper-parameter. The maximization of the likelihood can be achieved using EM-algorithm (27). The posterior mode or the maximum a posteriori (MAP) obtained using penalized likelihood maximization has been used as a measure of location in disease

mapping (246),(166).

The penalized quasi-likelihood (PQL) method has been used in empirical Bayesian disease mapping (184),(41) ,(307). The point estimates obtained through this procedure are consistent and nearly unbiased for the relative risks but the variability in these estimates is often underestimated because no allowance is made for the uncertainty in the hyper-parameter. Hence the confidence intervals for the relative risk based on the estimated variance of the posterior are very narrow. Another disadvantage of empirical Bayesian estimation is that it may not be able to provide an adequate description of the true dispersion in the rates (13), (148). To overcome these limitations Devine and Louis (202) and Devine et al (203) have suggested a constrained empirical Bayesian estimator, and Macnab et al(310) proposed an EB bootstrap methodology using type III parametric bootstrapping (198) and a sample reuse method (43). Other authors have suggested the complementary use of empirical Bayesian and fully Bayesian inferential techniques (306) and (8) proposed the use of product partition models (PPM) approach.

2.5.4 Fully Bayesian Approach

The fully Bayesian (FB) approach assigns prior distribution to all parameters and the parameters of these prior distribution are assigned hyperprior distributions to cope with their possible variability. Thus, it incorporates variability in the hyper-parameters when specifying the hyperprior distribution. Inference on the relative risks is based on estimated posterior distributions where uncertainties associated with the estimates are taken care of by specifying vague hyperpriors. The joint posterior distribution of the relative risks and the hyper-parameters

given the data is

$$g(\theta, \gamma/O) \propto g(O/\theta) f(\theta/\gamma) f(\gamma)$$

And the marginal posterior distribution for the relative risk given the data is

$$g(\theta/O) = \int g(\theta, \gamma/O) d\gamma$$

Inference about the relative risk is based on this marginal posterior distribution. Working directly with this distribution requires analytic approximation procedures to the integral such as Laplace method or numerical evaluation of the integral (246). Estimation can also be done using the Markov Chain Monte Carlo (MCMC) simulation methods which allow the samples to be drawn from the joint posterior distribution and the marginal posteriors $g(\theta/O)$ and $g(\gamma/O)$ (13). Also see Gilks et al (301) and Brooks (269) for details of the MCMC computational algorithm.

Lawson et al (24) gave a review of the fully Bayesian disease models and Bernadinelli and Montomoli (166) compared comprehensively the empirical and fully Bayesian methods. The authors judge the FB as the preferred method because it considers the uncertainty of the parameters of the model whereas the EB conditions the estimation on the point estimates of the parameters. As a result, the EB is less accurate. FB procedure is more computer intensive, EB is useful for initial risk assessment and can serve as exploratory analysis, and if strong variation in risk rates is established in the exploratory analysis, FB estimation can be employed to detect clusters of disease. Best et al (183) give a summary of the hierarchical Bayesian models that are used for disease mapping using fully Bayesian estimation. They described a three-level hierarchical model as a natural

model for disease mapping based on aggregation of underlying individual level risks as $O_i \sim Poisson(\theta_i E_i), i = 1, 2, 3, \ldots, n$

$\beta \sim p(./\lambda)$

$\lambda \sim \pi()$

Where O_i is the observed HIV positive count in site i, E_i is the expected number of cases in site i, β_i is the log relative risks, $p(./\lambda)$ is an appropriate second stage prior distribution and λ is the hyper-parameter of this second stage model with hyperprior distribution.

2.5.5 The Gamma-Poisson Model

Clayton and Kaldor (54) were the first to adopt a random-effect (or mixture) model that assumes a parametric probability density function for the distribution of relative risks between areas. They suggested that the independently and identically distributed relative risk θ_i follow a gamma distribution with scale parameter α and shape parameter ν - the hyper-parameters, that is with mean $\frac{\nu}{\alpha}$ and variance $\frac{\nu}{\alpha^2}$. Hence,

$$f(\theta_i/\gamma) = \frac{\alpha^\nu \theta_i^{\nu-1} \exp -\alpha\theta_i}{\Gamma(\nu)}$$

Conditioned on θ_i ,O_i the observed HIV counts, are Poisson variates with mean $\theta_i E_i$. It then implies that the marginal density of O_i has a closed form negative binomial distribution with unconditional expectations

$$E(O_i) = E_i \frac{\nu}{\alpha}$$
$$var(O_i) = E_i \frac{\nu}{\alpha} + E_i^2 \frac{\nu}{\alpha^2}$$

The scale and shape parameters of the negative binomial model are estimated by maximizing the marginal likelihood. The empirical Bayesian estimate of the posterior expectation is

$$E(\hat{\theta}_i/O_i, \alpha, \nu) = \frac{O_i + \hat{\nu}}{E_i + \hat{\alpha}}$$

This is a compromise between the observed relative risk $\frac{O_i}{E_i}$ and the mean $\frac{\nu}{\alpha}$ of the distribution of the relative risk.

The Gamma-Poisson model has the advantage of estimating the full posterior which can be used to give confidence interval and hypothesis tests and its estimate of mean and variance via maximum likelihood is superior to that obtained through the methods of moments as in linear Bayesian methods (245).

This model can account for covariates by making the prior mean a function of the covariates (161). Marshal (245) extended the gamma-Poisson model by proposing a non-iterative distribution-free approach using weighted moments to estimate a prior mean and variance and pointed out the difficulties arising in iterative methods of estimation. Tsutakawa (247) developed the Gamma-Poisson model further by letting the gamma scale parameter depend on a geographic effect with an inverse gamma distribution. He then used the Poisson likelihood and Gamma framework to estimate relative risks for geographic regions with additional random effects component. See Lawson (23) for detailed definition of random effects in disease mapping. Area-level or ecological covariates can be included by modelling the logarithm of the relative risk as a linear function of the covariates (54), (118), (112) as:

$$E[log(\theta_i)] = x_i^T \beta$$

And

$$E(\theta_i) = \frac{\nu}{\alpha_i} = \exp(x_i^T \beta)$$

This model assumes that there is no extra-Poisson variation. e_i^{β} is the relative risk due to risk factor i .

2.6 Multi-level Models

2.6.1 Variance Component models

In this section, we examine simultaneously the effects of individual-level and group-level factors on risk of HIV infection. The data defines a multilevel structure:- sites are within states and the states are grouped into zones. Information on HIV positivity is collected from individuals in the sites. Hence, it is possible to have a three-level (site, state and zone) model or a two-level (site/state or state/zone) model as the data is also aggregated by state. Langford et al(112) extended the multilevel models developed by Goldstein (105), (107) to disease mapping. The simplest Poisson multilevel model is the one that incorporates a measure of the extra-Poisson variation in the model as a high level variable. Given the covariates x_i, the logarithm of the mean of the relative risks is

$$log(\mu_i) = log(E_i) + \alpha + x_i^T \beta + u_i$$

where u_i are the heterogeneity effects or extra-Poisson variation caused by

variation among underlying populations at risk in the areas considered. $log(E_i)$ is the offset which accounts for the population at risk and α is the intercept.

$$u_i \sim N(0, \sigma_u)$$

and

$$log(\theta_i) \sim N(\mu_i, \sigma_u)$$

where

$$\mu_i = x_i^T \beta$$

This model can be extended to a model with more than one higher level of geographical aggregation (107) (118), (112). For a model consisting of two levels, site i nested in the state j , then the observed HIV counts in site i becomes

$$O_{ij} \sim Poisson(\theta_{ij} E_{ij})$$

And the log-linear model becomes

$$log(\mu_{ij}) = log(E_{ij}) + \alpha + x_{ij}^T \beta + u_{ij} + v_j$$

$$u_{ij} \sim N(0, \sigma_u), v_j \sim N(0, \sigma_v)$$

And a three-level model comprising of site i nested in state j nested within zone k we have that

$$O_{ijk} \sim Poisson(\theta_{ijk} E_{ijk})$$

and

$$log(\mu_{ijk}) = log(E_{ijk}) + \alpha + x_{ijk}^T \beta + u_{ijk} + v_{jk} + y_k$$

$$u_{ijk} \sim N(0, \sigma_u), v_{jk} \sim N(0, \sigma_v), y_k \sim N(0, \sigma_y)$$

where u_{ijk} are the random effects for the sites, v_{jk} are the random effects for the states and y_k are the random effects for the zones.

Hence $log(\theta) \sim MVN(\mu, \Sigma)$. where Σ is a block diagonal comprising of the variance of the three random effects due to site, state and zone respectively.

The models we have considered so far are the variance component models. The effects of the spatial distribution of the sites and the states were not taken into cognisance.

The variance component models can be fitted using the quasi-likelihood, iterative generalized least squares (IGLS), Fisher scoring algorithm or the restricted iterative generalized least squares (RIGLS). The detailed account of the algorithm for the estimation procedure of the multilevel model is given in Goldstein (105).

Iterative Generalized Least Squares

This method is based on generalized least squares which gives the maximum likelihood estimates for hierarchically structured data (105). A simple model of fixed and random effects (107),(184) is given as,

$$Y = X\beta + Z\theta \tag{2.58}$$

where $X\beta$ is the fixed part and $Z\theta$ is the random part. Y is the observed vector of events being modelled by predictor variables X and the fixed parameters β, and the predictor variable Z with random coefficients θ. The design matrices X and Z need not be the same. Z may represent variables random at any level in the model.

The procedure of the IGLS is a two-stage process for estimating the fixed parameters and the variances and covariances of the random parameters in successive iterations. The first stage is to estimate the fixed parameters using the ordinary least squares regression and taking the higher level variances to be zero. The vector of residuals from this initial model is then used to construct the initial values for the dispersion matrix V. The dispersion matrix is then used in the estimation of the fixed parameters using the generalized least squares estimation procedure as

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \tag{2.59}$$

again the vector of residuals are computed as

$$\tilde{Y} = Y - X\hat{\beta} \tag{2.60}$$

and we obtain the cross-product matrix of the residuals, $Y^* = \tilde{Y}\tilde{Y}^T$ such that

$$V = E(Y^*) = E(\tilde{Y}\tilde{Y}^T) \tag{2.61}$$

We then stack the columns of the cross-product matrix into a vector,

$$Y^{**} = vec(\tilde{Y}\tilde{Y}^T) = vec[(Y - X\hat{\beta})(Y - X\hat{\beta})^T] \tag{2.62}$$

and Y^{**}will then be used as the response variable in a regression equation to estimate the random parameters. The covariance of the random coefficients θ is estimated as

$$cov(\hat{\theta}) = (Z^T V^{*-1} Z)^{-1} Z^T V^{*-1} Y^{**} \tag{2.63}$$

where V^* is the Kronecker product of V, that is $V^* = V \bigotimes V$. Assuming multivariate normality, the estimated covariance matrix for the fixed parameters is

$$cov(\beta) = (X^T V^{-1} X)^{-1} \tag{2.64}$$

Goldstein and Rabash (108) gave the estimate of the random parameters as

$$cov(\hat{\theta}) = 2(Z^T V^{*-1} Z)^{-1} \tag{2.65}$$

Hence, the iterative procedure continues alternating between estimation of the fixed and random parameter vectors until convergence is achieved.

Fisher Scoring algorithm

This iterative technique is used to obtain the maximum likelihood estimate of the hyperparameters γ which are updated using estimates from the pth iteration as

$$\hat{\gamma} = \gamma^{(p)} + i^{(p)-1} U^{(p)}$$

where $i^{(p)}$ is the Fisher's information matrix and $U^{(p)}$ is the score statistic and both of them are evaluated at $\gamma^{(p)}$. See Breslow and Clayton (184)for more details in this estimation procedure.

Penalized Quasi-likelihood (PQL)

Given that the observed number of cases (O_i) in each site follow the Poisson distribution with mean μ_i and that

$$log(\mu_i) = log(E_i) + \alpha + x_i\beta + u_i \tag{2.66}$$

this equation implies a nonlinear (logarithmic) relationship between the observed number of cases O_i and the predictor part of the model. Hence, the normal distribution approximation does not directly apply here. In order to estimate the random parameters $\hat{u}_i$ from the model we use the penalized quasi-likelihood estimation procedure which involves the application of an approximate linearizing technique at each iteration using a first and second order Taylor series approximation. If

$$\mu_i = f(H)$$

where $H = \alpha + x_i\beta + u_i$ and if H_t is the value of the linear predictor H at iteration t, then $f(H_{t+1})$ is expressed as a function of H_t through a second-order Taylor expansion about the current fixed and random part estimates as

$$f(H_{t+1}) = f(H_t) + x_i(\beta_{t+1} - \hat{\beta}_t)f'(H_t) + (u_{t+1,i} - \hat{u}_{t,i})f'(H_t)$$

$$+(u_{t+1,i} - \hat{u}_{t,i})^2 f''(H_t)/2 \tag{2.67}$$

The first two terms on the right-hand side provide the updating function for the fixed part of the model and the last two terms are for the estimation of the random part. See Breslow and Clayton (184), Goldstein (107) and Goldstein and Rasbash (109) for a full description of the linearizing procedure. For the Poisson distribution

$$f(H) = f'(H) = f''(H) = exp(X_i\hat{\beta}_{ti} + \hat{u}_i) \tag{2.68}$$

Langford et al. (1999) gave the extension of the PQL procedure to spatial models as follows:

$$log(\mu_i) = log(E_i) + \alpha + x_i\beta + u_i + v_i \tag{2.69}$$

The random parameters $\hat{u}_i$ and $\hat{v}_i$ are estimated from the model using the procedure outlined above as

$$\mu_i = f(H)$$

where $H = \alpha + x_i\beta + u_i + v_i$ and

$$f(H_{t+1}) = f(H_t) + (\alpha_{t+1} - \hat{\alpha}_t) + x_i(\beta_{t+1} - \hat{\beta}_t)f'(H_t) + (u_{t+1,i} - \hat{u}_{t,i})f'(H_t)$$

$$+(v_{t+1,i} - \hat{v}_{t,i})f'(H_t) + (u_{t+1,i} - \hat{u}_{t,i})^2 f''(H_t)/2 + (v_{t+1,i} - \hat{v}_{t,i})^2 f''(H_t)/2 \tag{2.70}$$

The first three terms on the right-hand side provide the updating function for the fixed part of the model and the last four terms is for the estimation of the random part.

Marginal Quasi-likelihood (MQL)

The linearizing procedure given equations (2.67) and (2.70) above can lead to convergence problems or the model may fail if residuals are very large. To

overcome this limitation the MQL procedure (184),(107) can be adopted whereby the second-order terms in the equations (2.67) and (2.70 are omitted. In extreme cases, estimates can be based only on the fixed part of the model such that

$$H_t = X_i\hat{\beta}_t$$

This procedure has the disadvantage of producing biased estimates when the sample size is small. However, it can be corrected using bootstrap procedures (107). Generally, the PQL procedure gives better estimates than the MQL (107).

Restricted Iterative Generalized Least Squares

This is an extension of the IGLS. Like the maximum likelihood estimates, the IGLS estimates are biased. Goldstein (106) shows that a slight modification of the IGLS by restricting the model to take account of the sampling variations in the parameters can lead to unbiased estimates of the fixed and random parameters. Given the general model,

$$Y = X\beta + Z\theta$$

such that $E[(Z\theta)(Z\theta)^T] = V$ and $cov(\hat{\beta}) = (X^TV^{-1}X)^{-1}$ Then the

$$E[(Y - X\hat{\beta})(Y - X\hat{\beta})^T] = V - Xcov(\hat{\beta})X^T = V - X(X^TV^{-1}X)^{-1}X^T$$

where X is the design matrix for the fixed effects in the model with full rank. V is then updated at each iteration using its current value $\hat{V}$ as

$$V = (Y - X\hat{\beta})(Y - X\hat{\beta})^T + X(X^T\hat{V}^{-1}X)^{-1}X^T$$

The last term $X(X^T\hat{V}^{-1}X)^{-1}X^T$ can be regarded as a bias correction term. Under the assumption of multivariate normality this procedure is equivalent to restricted maximum likelihood.

2.6.2 Spatial Multilevel Models

The idea behind this section is that areas close to one another in geographical space share the same environmental, socio-economic, cultural and demographic factors that influence disease rates and are more likely to share similar relative risks. Ignoring this dependence, where it exists, will result in the standard errors of the ecological regression coefficients being too small if the dependence is positive or too large if the dependence is negative. Thus, we need to reflect this prior knowledge in the model by incorporating a spatial component into the multilevel models considered in the previous section. This can be achieved using the nearest neighbour Markov random field models (127), (303), (13)

The introduction of the geographical structure of the relative risks into the model results in a more complex prior model as it imposes conditional independence structure on the relative risks such that each relative risk is conditionally independent of all other relative risks, given a small set of geographically adjacent areas. Clayton and Kaldor (54) proposed the multivariate lognormal prior, which have the capability of accommodating the spatial dependence of the relative risks. Supposing that the relative risks are correlated where the correlation is dependent on geographical proximity and that the relative risk can be considered to be Gaussian, (127), (160), (94), (54) estimated the log relative risks using conditional autoregressive (CAR) method given as

$$E(\beta_i/\beta_j, j \neq i) = \mu_i + \rho \sum_j W_{ij}(\beta_j - \mu_j)$$

$$var(\beta_i/\beta_j, j \neq i) = \sigma^2$$

Where W is the adjacency matrix of the map defined as $W_{ij} = 1$ if i and j are adjacent sites or states and 0 otherwise. Also

$$E(\beta) = \mu$$

$$cov(\beta) = \Sigma = \sigma(I - \rho W)^{-1}$$

This model was modified slightly in Yasui et al (303)as

$$E(\beta_i/\beta_j, j \neq i) = \alpha + \frac{\rho \sum\limits_{j \in \delta_i} w_{ij}(\beta_j - \alpha)}{\sum\limits_{j \in \delta_i} w_{ij}}$$

$$var(\beta_i/\beta_j, j \neq i) = \frac{\sigma^2}{\sum\limits_{j \in \delta_i} w_{ij}}$$

$E(\beta_i) = \alpha$ and σ^2 is the scale parameter. ρ is the spatial dependent parameter, δ_i is the set of neighbourhood sites for site i and w_{ij} are weights indicating the proximity of each area to its neighbourhood areas. When $\rho = 1$, this model is called Gaussian intrinsic autoregression model (238).

$$cov(\beta) = \Sigma = \sigma(I - \rho MW)^{-1}M$$

Where I is an identity matrix, M is a diagonal matrix with

$$M_{ii} = \frac{1}{\sum_{j \in \delta_i} w_{ij}}$$

Besag et al (130) developed a simple spatial model for the distribution of relative risks of a disease by considering β to be a sum of a Gaussian intrinsic autoregression prior. This model is an extension of the convolution Gaussian prior proposed by Besag (128) and Besag and Mollié (129) given as

$$O_i \sim Poisson(\mu_i)$$

$$log(\mu_i) = log(E_i) + \alpha + x_i\beta + u_i + v_i$$

where u_i are residuals with no spatial structure; they are the heterogeneity effects between sites or states which are independent of one another. Its variance σ_u^2 is the non-spatial extra-Poisson variation which is assumed to be a Gaussian homoscedastic white noise. v_i are residuals with spatial structure. They are spatially dependent random effects and may have any one of a number of structures describing adjacency or nearness in geographic space. They are spatially structured and represent the spatial contribution of neighbouring areas. σ_v^2 is the spatial extra-Poisson variation. If the variation in β is spatially structured, σ_v^2 will dominate the total variation, if the variation of β is spatially unstructured, σ_u^2 will dominate the total variance.

From Clayton and Bernadinelli (55), the logarithm of the relative risk can be

expressed as

$$log(\theta_i) = u_i + v_i$$

where u_i are the unstructured heterogeneous effects and

$$u_i \sim N(0, \sigma_u^2)$$

v_i are the spatially structured effects through an intrinsic Gaussian autoregression

$$(v_i|v_j, j \neq i) \sim N(\bar{v_i}, \sigma_{v_i}^2)$$

where $\bar{v_i}$ is the mean of all the neighbours of site i

$$\bar{v_i} = \frac{\sum_j w_{ij} v_j}{\sum_j w_{ij}}$$

$$\sigma_{v_i}^2 = \frac{\sigma_v^2}{\sum_j w_{ij}}$$

$w_{ij} = 1$ if i and j are contiguous and 0 otherwise. The structured and unstructured effects are independent of one another. Also,

$$var[log(\theta_i)/log(\theta_j), j \in \delta_i, \sigma_v, \sigma_u] = \frac{\sigma_v^2}{\sum_j w_{ij}} + \sigma_u^2$$

is the conditional variance of the log relative risk θ_i given all other $\theta_j's$. If

σ_v^2/σ_u^2 is small, then unstructured heterogeneity dominates, whereas if σ_v^2/σ_u^2 is large, the spatially structured variation dominates.

The multivariate representation of the model with spatial dependence is given as

$$log(\theta) = \beta \sim MVN(\mu, \Sigma)$$

where

$$\beta = \{\beta_1, \beta_2, \ldots, \beta_n\} = \{\log(\theta_1), \log(\theta_2), \ldots, \log(\theta_n)\}$$

$$\Sigma = \sigma^2 \Omega$$

and Ω_{ij} is the correlation between β_i and β_j. If $\Omega_{ij} = 0$, then β_i and β_j are marginally independent. Ω is usually specified as a parametric function of distance, d_{ij}, between the centroids of each pair of areas. The function chosen must ensure that Σ be positive definite. To achieve this, exponential decay function is adopted (36), (274) such that

$$\Omega_{ij} = f(d_{ij}, \phi) = \exp(-\phi d_{ij})$$

where $\phi > 0$ controls the rate of decrease of correlation with distance, with large values representing rapid decay. The disc model can also be used to determine the correlation between two points which is defined as proportional to the intersection area of two discs of common radius ϕ centered on the points (252).

For the independent normal or multivariate normal priors of the relative risks, the hyperpriors for the inverse variances (σ_v^2 and σ_u^2) are the conjugate gamma distributions with specified parameters. When there is lack of information on the strength of u and v, vague Gamma priors with means

$$\frac{2}{var[log(\theta_i)]}$$

for σ_u^{-2} and

$$\frac{2}{\bar{w}var[log(\theta_i)]}$$

for σ_v^{-2} are assumed (166),(12), where $\bar{w} = \frac{\sum_{ij} w_{ij}}{n}$, n is the number of sites.

Another form of Gaussian Markov Random Field (GMRF) prior specification for the multivariate normal adopted by many authors (41), (307),(305) for modeling random spatial effects is

$$v \sim N\{0, \Sigma(\sigma^2, \rho)\}$$

and

$$\Sigma(\sigma^2, \rho) = \sigma^2 D^{-1}$$

where

$$D = \rho R + (1 - \rho) I_J$$

σ represents the spatial dispersion parameter and ρ,$(0 \leq \rho \leq 1)$ is the spatial autocorrelation parameter, R is a $J \times J$ neighbourhood matrix whose jth diagonal element is the number of neighbours of the corresponding site and the off diagonal elements in each row is -1 if the corresponding areas are neighbours and 0 otherwise. I_J is an identity matrix of order J. Hence the model depends on the neighbourhood structure and not on distance between sites. The prior specification for v above also represents a conditional autoregressive (CAR) model with conditional distribution of v_j given as (305)

$$v_j/v_{k\neq j} \sim N\{\frac{\rho}{1-\rho+\rho n_j}\sum_{j\sim k} b_k, \frac{\sigma^2}{1-\rho+\rho n_j}\}, j = 1, \ldots, J,$$

where n_j is the number of neighbours for site j and $j \sim k$ means that site j is a neighbour of site k. Note that when $\rho = 0$, the sites are independent and when $\rho = 1$, we have Gaussian intrinsic autoregressive prior.

2.6.3 Multiple Membership Multiple Classification Models

This is an extension of the multilevel models which is applied on data where the lowest level unit is a member of more than one higher classification unit. The standard model fits mainly two classifications; the area (site) classification that captures the non-spatial variation and the multiple membership neighbour classification that adjusts for effects due to neighbouring areas. This model was first applied by Hill and Goldstein (295) and was further developed by Rabash and Browne (141). Browne et al (294) extended this to disease modelling by

applying it to the analysis of Scottish lip cancer data set as follows:

$$O_i \sim Poisson(\lambda_i),$$

$$log(\lambda_i) = log(E_i) + \beta_0 + x_i\beta_i + u^{(2)}_{site(i)} + \sum_{j \in Neighbour(i)} w^{(3)}_{ij} u^{(3)}_j$$

$$u^{(2)}_{site(i)} \sim N(0, \sigma^2_{u(2)}), u_j \sim N(0, \sigma^2_{u(3)})$$

where $w_{ij} = 1/n_i$ and n_i is the number of neighbours for site i. $u^{(2)}_i$ *and* $u^{(3)}_j$ are the vectors of residuals for the random effects for site classification (2) and neighbours classification (3) respectively. This model is similar to the conditional autoregressive (CAR) model considered in the previous section. They differ slightly in the manner in which spatial correlation is estimated. The CAR model estimate the spatial correlation through the variance structure rather than through the multiple membership relationship.

2.7 Monitoring Convergence

To effectively monitor convergence in any iterative simulation method like the Gibbs sampler or Metropolis algorithm, Gelman and Rubin (4) recommended the use of several independent sequences with starting points sampled from an over-dispersed distribution. Earlier works (5) have confirmed that the use of a single series from the Gibbs sampler provides a false sense of security. The use of multiple sequence makes it possible to obtain a distributional estimate for each estimand at each iterative simulation and an estimate of how much sharper the distributional estimate might be if the simulations were continued indefinitely. This investigation is accomplished using the components of variance from the

multiple sequences. The detailed procedure could be found in Gelman and Rubin (4).

The process entails running $m \geq 2$ independent sequences, each of length $2n$, with starting values drawn from an over-dispersed distribution. The first set of n iterations is discarded to minimize the effect of the starting distribution. Diagnosis is therefore based on the last set of n iterations. For each parameter of interest, we obtain the following: The *between* sequence mean variance B given as

$$B = \frac{1}{n} \sum_{i=1}^{m} \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{m-1}$$

The average of the *within*-sequence variance

$$W = \frac{1}{m} \left\{ \frac{\sum_{ij}^{mn} (x_{ij} - \bar{x}_{i.})^2}{n-1} \right\}$$

The estimate of the target variance obtained as weighted average of W and B

$$\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{B}{n}$$

Define the variance of the Student's t distribution of the estimand x, as

$$V = \hat{\sigma^2} + \frac{B}{mn} \tag{2.71}$$

$$= \frac{1}{n} \left\{ (n-1)W + \frac{m+1}{m} B \right\} \tag{2.72}$$

Convergence is monitored by estimating the factor by which the scale of current distribution for x might be reduced if the iteration is continued in the limit as $n \to \infty$. This reduction factor is the ratio R of the current variance estimate, $\hat{v}$, to the *within*-sequence variance, W, with a factor to account for the extra variance of the Student's t distribution. R is estimated as

$$\hat{R} = \frac{\hat{V}}{W}\left(\frac{df}{df-2}\right) \tag{2.73}$$

$$= \frac{1}{n}\left\{(n-1) + \frac{m+1}{m}\frac{B}{W}\right\}\frac{df}{df-2} \tag{2.74}$$

The scale reduction is estimated as $\sqrt{\hat{R}}$. If the potential scale reduction is high, further iterations are needed to improve inference about the targeted distribution. If $\hat{R}$ approaches 1, then convergence is reached.

Note as $n \to \infty$, $\hat{R} \propto \frac{\hat{V}}{W}$

Chapter 3

Spatial Analysis

3.1 Justification for Spatial Analysis

Nigeria is a multi-cultural, multi-ethnic and multi-lingual society which lies between longitudes 3 and 14 and latitudes 4 and 14. At its widest, Nigeria measures about 1200km from east to west and about 1050km from north to south. About 374 pure ethnic groups make up the entity called Nigeria and trado-cultural practices vary between these ethnic groups. However, some of the ethnic groups share some cultural and religious affinity. Political boundaries were designed to cluster communities based on these affinities and currently, there are 774 Local Government Areas (LGAs). These LGAs are in turn grouped under 36 States and a Federal Capital Territory (FTC), Abuja. The states are grouped into six geopolitical zones namely, south-south, southeast, southwest, north central, northeast and northwest. See Table 1.1

The December 2006 National Census gave the population count for the country as 140,003,542 people, obtained from 662,000 Enumeration Areas (EAs) (2).

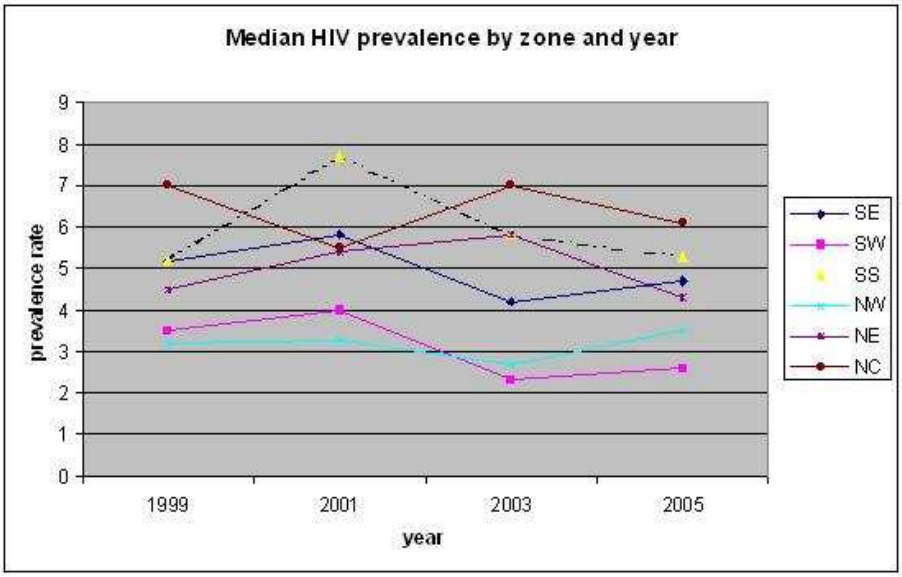

Figure 3.1. *Time series plot of prevalence rates by zone*

These individuals form human clusters that are distinguishable by their languages, traditions, socio-cultural practices, and religion. The human clusters make up the hamlets, villages, and towns/communities spread across the land of Nigeria.

Several studies (194), (193),(1), (88), (33) indicate that socio-cultural practices and behaviours that encourage the spread of HIV/AIDS vary widely among the states and zones. Also, results from sentinel surveys indicate that the median HIV prevalence rate varies significantly across states, zones and by rural and urban locations (93),(92), (90), (89). The graph in Figure 3.1 shows the estimates of the median prevalence rates for each of the six geopolitical zones in the years 1999, 2001, 2003 and 2005. Estimates were based on the results of the HIV Sentinel Surveillance for these years and calculated as weighted averages of the urban and rural prevalence rates using the rural and urban population as weights. Variation between zones over the years is distinct in the graph with the North-central and South-south zones in the lead.

Independent studies on sexual networking in some parts of the country indicate that there is a large network of premarital and extramarital sexual behaviour within the surveyed subpopulations (121), (122), (263), (200) (277), (195). The size and nature of the observed network vary from one subpopulation to the other. The patterns of the networks appear to depend on some social and economic determinants and on the cultural context within which the sexual behaviour takes place ((33), (277),(195), (267)). One's sex ideology seems to be a function of geophysical, socio-cultural and economic environment and personal risk orientation. Most of these studies identified age of sexual debut, age of first marriage, rate of partner exchange, rate of contact with sex workers, concurrent partnership and level of education as factors affecting premarital and extramarital sex. These factors are known to vary by ethnic and religious groups in Nigeria and elsewhere. For example, in a study of unmarried women in the US, it was found that having multiple sexual partners is linked to age at first sex and birthplace (265).

Adherence to religious tenets in one's behaviour may have an effect on health and disease transmission (51). Religions place some constraints on sexuality and studies have shown that religiosity and religious affiliation are negatively correlated with sexually transmitted diseases (188), (218). The core Muslim states in Nigeria are located in the northeast, northwest and north-central geopolitical zones. According to Malamba et al (270), Muslims have lower risk for HIV infection than non-Muslims, possibly due to the protective effect offered by male circumcision. Specifically, Gray(218) tested the hypothesis that Islamic religious affiliation negatively associates with HIV seropositivity and concluded that the percentage of Muslims in the population negatively and significantly predicted the prevalence of HIV among sub-Saharan African countries. Six out of seven studies enabling within-population comparisons revealed lower HIV prevalence

rates among Muslims. According to him, Muslims have lower alcohol consumption as well as a higher rate of circumcision compared with non-Muslims. This may not be the case with Nigeria. HIV prevalence is high in some core Muslim states especially those in the north central and northeast zones. Although prevalence is low in the northwest zone, it is not the lowest among the zones and it does appear from figure 3.2 to be on the increase. Also male circumcision is very common among other religious and ethnic groups. While religious inclination and belief may be a factor in explaining the spread of HIV/AIDS in Nigeria (as it may impose restriction on certain behaviour), it does seem that other factors may be more associated with the epidemic. Hence, we need to consider religion alongside other factors.

The association between alcohol consumption and increased high-risk sexual behaviour is documented in various studies ((145), (223), (153), (80) (81)). Association between a history of alcohol consumption and being HIV sero-positive has been established by some of these authors. Alcohol diminishes personal control, increases risk-taking, and reduces the ability to make informed choices around safer sex (266). Alcohol impairs various aspects of the immune system and increases the susceptibility to HIV infection (66). According to BBC News (2003), a team of researchers led by Prof Bagby of Louisiana State University, found that rhesus monkeys that were given regular doses of alcohol before being exposed to SIV virus (a monkey equivalent of HIV in humans) had 64-fold increase of SIV virus in their blood (one week after the exposure) than the control who were given doses of sugar solution before the exposure. Hence alcohol consumption may increase susceptibility to infection upon exposure to HIV.

The Muslims in Nigeria are located mainly in the north and partly in the west of the country. Recently, the observed HIV prevalence rates in some core

Muslim states are low. According to a study (28) conducted to determine risk factors for HIV among pregnant women attending antenatal clinics in the city of Jos located in the North-central zone, women of other faith were found to be more likely to be infected with HIV than Muslim women. Feyisetan and Pebley (33) confirmed that Muslims are less likely than other religious groups in Nigeria to have had premarital sex due to the more conservative nature of Islamic culture and the more protective attitude of Islam towards women. Other explanations are the closed sexual networks - since Muslim men are legally allowed to have up to four wives, the pressure for casual sex partners is appreciably reduced (266) - and the ritual washing which could increase penile hygiene and reduce the risk of sexually transmitted diseases. The sale of alcohol is prohibited in all core Muslim communities in Nigeria thereby making access to alcohol very difficult for the population.

The North Central geopolitical zone of Nigeria houses communities and towns whose culture is mixed. The religion is mainly Islam, with their emirate administered from the far north, but they have socio-cultural affinity with the south. Several studies have shown that the Okun tribe found mainly in Kogi state and some part of Kwara and Ekiti states still practice the culture of spouse sharing. In this culture, men do have and maintain sexual relationship with their kin's wife without any conflict. This is because family and clan members view themselves as one and consider what belongs to a kin as belonging to every member of the clan including their wives (181). One study (179) reported that 65 per cent of 1029 sexually active respondents were involved in this practice. Also, some communities in Benue state were known for their culture of men offering their wives, daughters or sisters as the highest gift of honour ("kola") to their visiting special friends. The woman is expected to sleep with the visitor throughout the

visit or some days thereof. According to Mbiti (251), in some societies brothers have sexual rights to the wives of their brothers (here, the word brother has wider meaning such that a person has hundreds of brothers). Where the age group system is taken seriously, as among the Maasai, members of one age group who were initiated in the same batch are entitled to have sexual relationship with wives of fellow members.

In the South-East geopolitical zone, cultural practices that encourage extramarital and premarital sex are common. If a man dies without having children (male children) or if the man is impotent, the wife is encouraged to have sexual relationships with men (often from the kin's men) in order to raise children for the husband. Also, a couple that do not have male children may encourage one of their daughters to remain unmarried and raise male children for the perpetuation of the family lineage (195). Temporary infertility can induce a woman to have extramarital sexual relationships. A woman who fails to achieve pregnancy in the first few years of marriage often considers her position as a wife threatened, and in order to save her marriage and be free from societal pressure, especially from her husband and his relatives, gets involved in extramarital sex. The majority of the sex partners of this category of women are often those they approach for solutions to their childlessness; medical practitioners, spiritualists and traditional healers (267). In some parts of this zone, tradition permits an elderly woman to marry a "wife". In this practice, if a woman is unable to achieve pregnancy throughout her childbearing age, she may marry a wife and choose a man of her choice (often her husband if still alive and they are in a good relationship) to raise children in her name. Sometimes the wife is allowed to make her choice of men. Procreation is the main reason for extramarital sex in traditional Igbo society. High premium is placed on a male child and women will do anything to have one.

In recent times, however, modernization, materialism and socioeconomic gains lure more women into extramarital and premarital sex.

The Southwest geopolitical zone of Nigeria is not without some practices that encourage extramarital and premarital sex. Several studies (121), (122); (200) confirm the existence of sexual networks in some cities and towns in the zone. Some authors have commented on the loose marriages among the Muslim women in Ibadan- they frequently alternated between wifehood and prostitution.

In the South-South zone, widespread prostitution was known as far back as the colonial era especially in the upper Cross River Basin. The practice of prostitution as a profession in Nigeria began with the advent of British rule. The Britons brought Cross River under their control between 1888 and 1909; by the 1920s, prostitution had become a substantially developed trade in the area. By the 1930s it was reported that about 12 percent of the female population in the Nta clan, 15 percent of the Uhumunu females and about 33 percent adult females in the Nnam group had joined prostitution and migrated with the Europeans to different parts of Nigeria and the west coast of Africa, Cameroon and Equatorial Guinea (39). Many of these women abandoned their husband and children for prostitution. The Calabar and Ogoja provinces (now Cross River and Akwa Ibom states) were so bedevilled with prostitution that, up till today, prostitution is known in many parts of Nigeria as "Akunakuna" named after one of the village groups in the then Obubura division. The carefree sexual behaviour of women from this part of Nigeria is well known. The frequent divorce, remarriage and loose parental supervision of adolescents, and the consequent early sexual debut in these societies, were captured in a study by (263). Prostitution in this zone has taken a more advanced form. International prostitution, which began in the second half of the 1980s following the introduction of the structural adjustment

program and its consequent economic crunch, is very popular among the Edos in this zone. According to Aghatise (68), about 80 percent of Nigerian women and girls trafficked into Europe came from the midsouth region and belong to the Edo ethnic group and most of them are from polygamous homes. Aghatise (68) also noted that most polygamous Edo men abdicate the caring for their children and abandon the task to the women. According to him, most of the Nigerian women trafficked to Italy in the 1980s were either married or separated.

Given these socio-cultural and behavioural scenarios, which may have direct links with HIV transmission and spread within the communities, it is expected that there should be variations in the rate and pattern of spread in the different communities (or social and cultural clusters) in Nigeria. The two graphs in figure 3.2 show some variations in observed HIV prevalence among pregnant women attending antenatal clinics in the surveyed sites (community or town) in the six zones for 2001 and 2003. Starting from the far left (Umuahia), the first group is the Southeast zone, followed by the South-south, Southwest, North-central, Northeast and Northwest zone. The variation within and between zones is very prominent and the patterns of spread appear similar and consistent when the two years are compared, with slight decreases in some sites in 2003.

The spread of HIV/AIDS within and between these clusters depends on the level of interactions between individuals with varying characteristics and behaviours. According to Schinazi (242), it is suspected that diseases like HIV can only spread in populations where people are grouped in clusters in which individuals have repeated and sustained sexual contact. Hagenaars et al (275) insist that the persistence of an infectious disease within a population depends on both the disease's transmission characteristics and the pattern of mixing between hosts. This mixing pattern is affected by spatial heterogeneity. Population

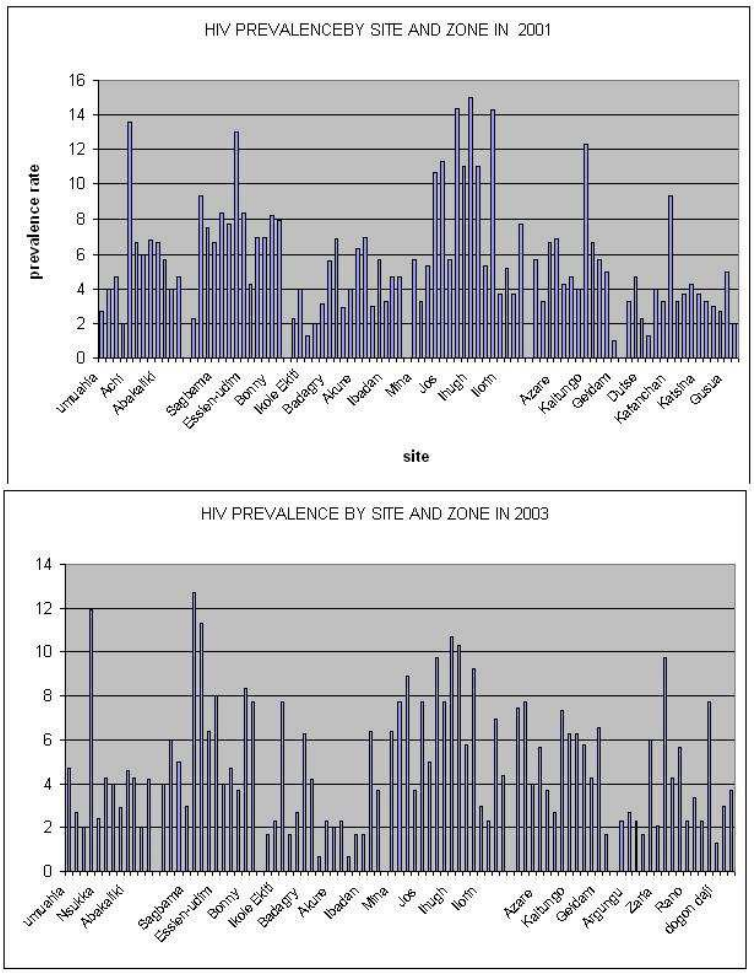

Figure 3.2. *Plot of site prevalence rates in women attending antenatal clinics grouped by zones*

heterogeneity can be achieved by dividing the population into homogenous subgroups based on spatial location, sex, behaviour, genetics, or other factors (300). In spatially heterogeneous sub-populations, an individual is more likely to contact other individuals within the same spatially defined subpopulation than those outside. Persistence of an epidemic between sub-populations is enhanced by rescue effects defined as transmissions between sub-populations that act to re-infect sub-populations where the disease has gone extinct. Also rescue effects are most effective if the coupling between sub-populations is sufficiently strong to generate frequent between-subpopulation transmissions but not so strong that spatial heterogeneity is lost (34) (239). However, if the infection rate outside the subpopulation (or social cluster) is low, an epidemic is possible if the social cluster and the within cluster infection rate are large enough (242).

All the papers on sexual networking, sexual behaviour and premarital and extramarital sex in Nigeria reviewed in this chapter, reveal that these practices are sufficient to sustain the spread of HIV/AIDS. Repeated and sustained sexual contacts were established in most of the studies. In a study of sexual behaviour, HIV related knowledge and condom use among commercial bus drivers and motor park attendants in Lagos, Nigeria, (77), it was found that about 74.3 percent of men had multiple sex partners and a strongly woven network of sexual relationships which include their wives, regular sex partners, commercial sex workers, young female hawkers, schoolgirls, and market women within and outside the park. There was consistent and regular condom use at a rate of 11.6 percent and knowledge of risk factors for STDs was poor. The spread of HIV in this type of network, where high-risk sex and low condom use is predominant, is a certainty. Over the years, the HIV/AIDS epidemic in Nigeria has been sustained

as a result of these risky behaviours. Interactions between social clusters, represented by the individual communities, are mirrored by the sexual networks which transcend geophysical community boundaries. Hence, the infection rates of HIV are expected to vary by communities but communities close to each other may show some similarities in the rates according to the persistence and rescue effect theory. This is confirmed by the graphs in Figure 3.2.

3.2 HIV clusters in Nigeria

In this section, we examine whether there is any tendency of HIV cases to cluster systematically in a particular area or whether the spread is random. That is, the similarity or dissimilarity of the spatial distribution of HIV prevalence among the Nigerian States. We start with a descriptive analysis of data pertaining to the 36 states of the Federation and the Federal Capital Territory. Thereafter, we shall consider observed values from individual sites. In doing this, we shall make use of the data generated from the National HIV Sero-prevalence sentinel Surveys conducted in 1999, 2001, 2003 and 2005 (93),(92), (90), (89). Data are available for all the HIV sentinel surveillance sites in 2001 and 2003. We first examined the data for symmetry and applied some transformation techniques on the data. We compared four different methods namely; natural log, logit, arcsine, Freedman- Turkey transformation methods. The natural log transform of the prevalence rates seems to be more appropriate.

The graphs in Figure 3.3 show perspective plots of HIV prevalence rates at the state level for 1999, 2001, 2003 and 2005 respectively. A close look at the shape of the plots suggests some heaping of HIV positive cases in some locations in the country. It is important to note that the change in the heaping over time is

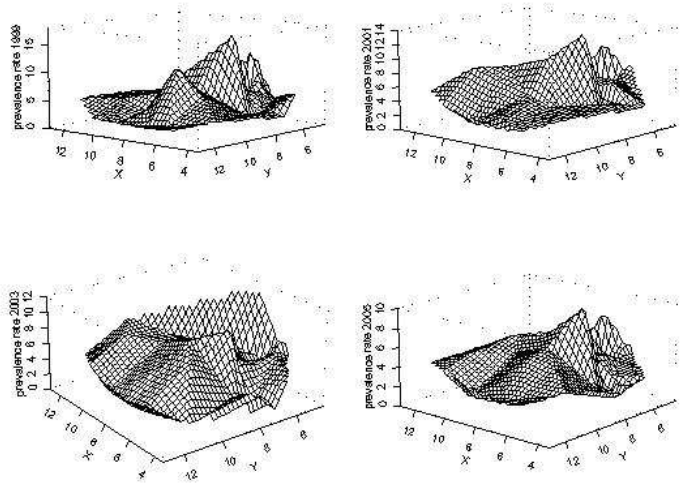

Figure 3.3. *Perspective plot of the natural log transform of HIV prevalence rates*

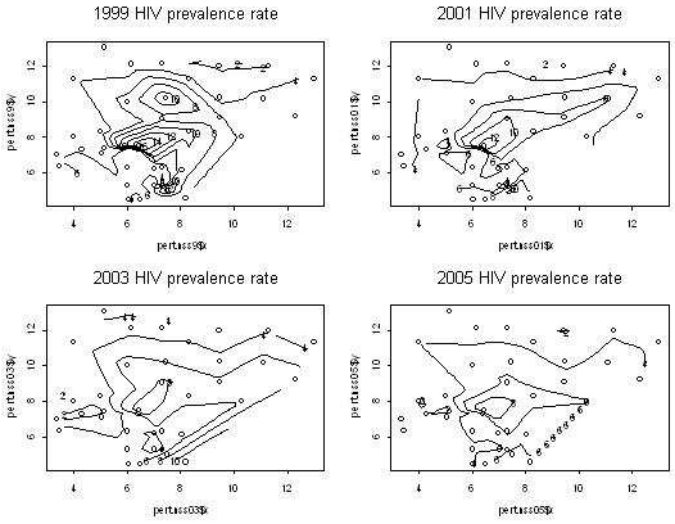

Figure 3.4. *contour plot of the natural log transform of HIV prevalence rates*

not appreciable. It does appear as though the states in the North central and the South-south consistently bear heavier burdens of the HIV epidemic. This burden is indicated by the heaping of cases in the plots in the locations of the two zones. In order to see the concentration of HIV cases more clearly, the prevalence rates were represented in contour plots.

The contour plots in Figure 3.4 indicate uneven rate of spread of HIV in the country. This unevenness is implicated by the "hills" and "valleys" shown in the contours, depicted by the degree of concentration of the lines.

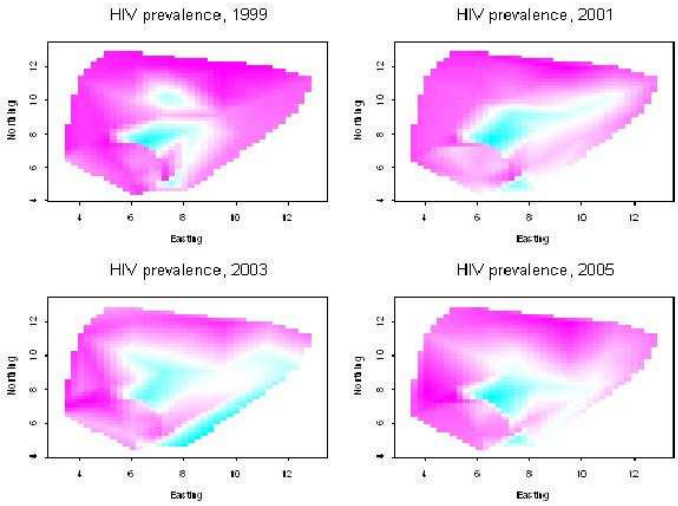

Figure 3.5. *contour image of the natural log transform of HIV prevalence rates*

A clearer picture of the concentration of the epidemic in Nigeria is shown in figure 3.5. These are colour images of the contours given in Figure 3.4. The blue and whitish patches indicate server or heavy concentration of cases, with the blue patches being the most server. The epicenters of the epidemic over the years has consistently been the North central and some part of the South-south zone of the country. It appears as though the spread of the epidemic is facing the direction of the Northeast and Southeast as it spreads from the North central and South-south. The spread over time can be discerned from the plots. In 1999, the epidemic was concentrated around the Cross River/Akwa Ibom axis, North central and Kaduna/Kano areas. By 2003 and 2005, the virus had spread to the Eastern half of the country and gradually advancing into the Northwest and Southwest.

To further investigate the clustering of HIV cases in the country, we employed the tools of cluster analysis

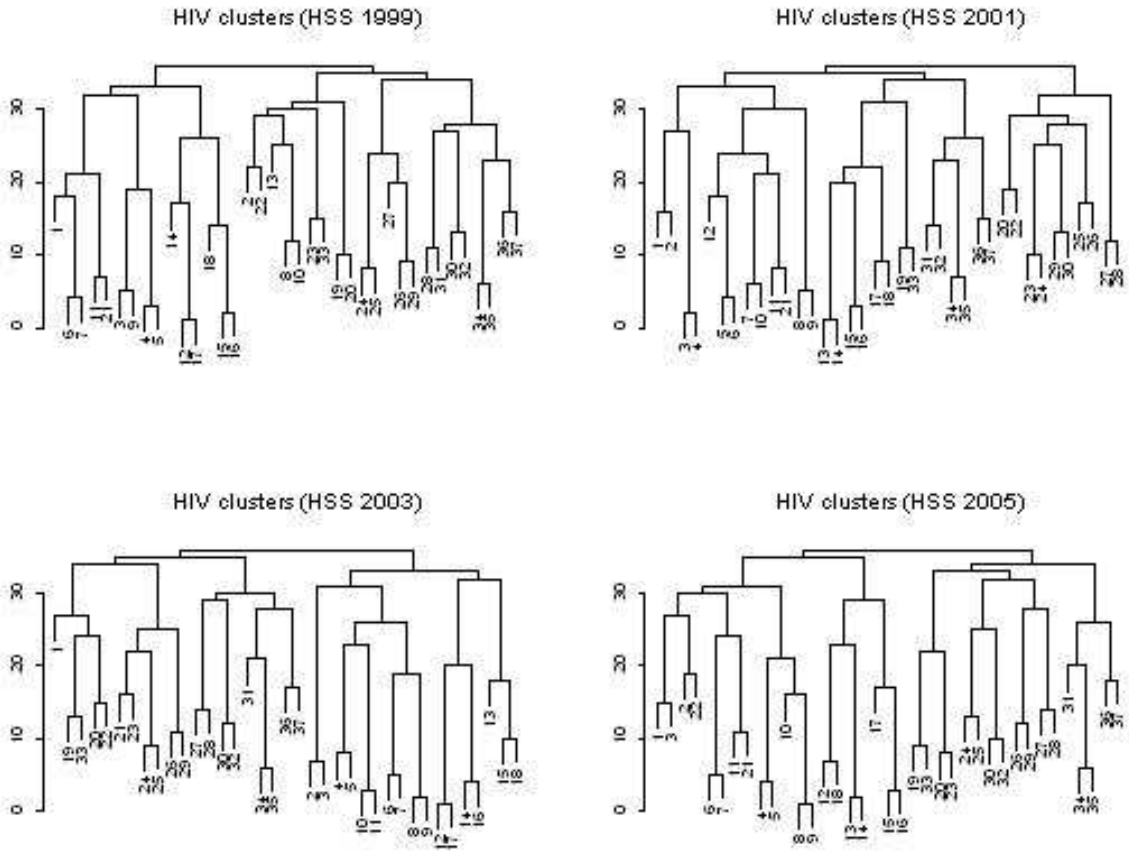

Figure 3.6. *State clustering using the natural log transform of HIV prevalence rates*

3.2.1 Cluster Analysis

From the analysis above, it does appear that there are some natural groupings of the States based on the level of HIV prevalence. This clustering tendency among the 36 states and the Federal Capital Territory as observed in the previous section is investigated in this section using the hierarchical agglomerative clustering method. This analysis was performed on data obtained from the National HIV Sentinel Survey conducted in 1999, 2001, 2003 and 2005. These data were obtained by unlinked screening of blood samples collected from pregnant women attending selected prenatal clinics within the period of the survey. See section 1.4.1 for detailed description of the data.

The results of the analysis support the findings of the contour analysis. The clusters show the groupings of nearby States with similar HIV prevalence rates. Broadly, it seems that each geopolitical zone is a major cluster. The graph in Figure 3.7 displays the spatial distribution of five major clusters obtained from the cluster analysis for each year. The analysis for 1999 clearly distinguishes the states in North central zone as a distinct cluster. But in 2001, this cluster merged and became one cluster with the Northeastern states, only to appear again as a separate cluster in 2003 and 2005. The behaviour of these clusters in these years corresponds to what we observed in the image map (Figure 3.5). For better appreciation of Figures 3.6 and 3.7, the numbers used in the clustering corresponding to serial number assigned to each state is given in Table 3.1

SS	SE	SW	NC	NE	NW
Cross Rivers 1	Anambra 7	Ondo 12	Kwara 18	Taraba 25	Jigawa 31
Akwa Ibom 2	Imo 8	Lagos 13	Niger 19	Adamawa 26	Kano 32
Rivers 3	Abia 9	Ogun 14	FTC 20	Borno 27	Kaduna 33
Bayelsa 4,	Ebonyi 10	Oyo 15	Kogi 21	Yobe 28	Kastina 34
Delta 5	Enugu 11	Osun 16	Benue 22	Gombe 29	Zamfara 35
Edo 6		Ekiti 17	Nassarawa 23	Bauchi 30	Kebbi 36
			Plateau 24		Sokoto 37

Table 3.1. *The grouping of the Nigeria States into Zones*

The plots in Figure 3.7 show the five major HIV clusters in the country over the years. The plots are offshoots of the dendograms (Figure 3.6). The numbers in the plots indicate the number assigned to the state positioned at the beginning of each of the major clusters. For instance in 1999, 12 is Ondo state in the southwest zone of the country, this number is used to represent all states in the same cluster with Ondo state. And for 2005, 19 is Niger state in the north-central zone.

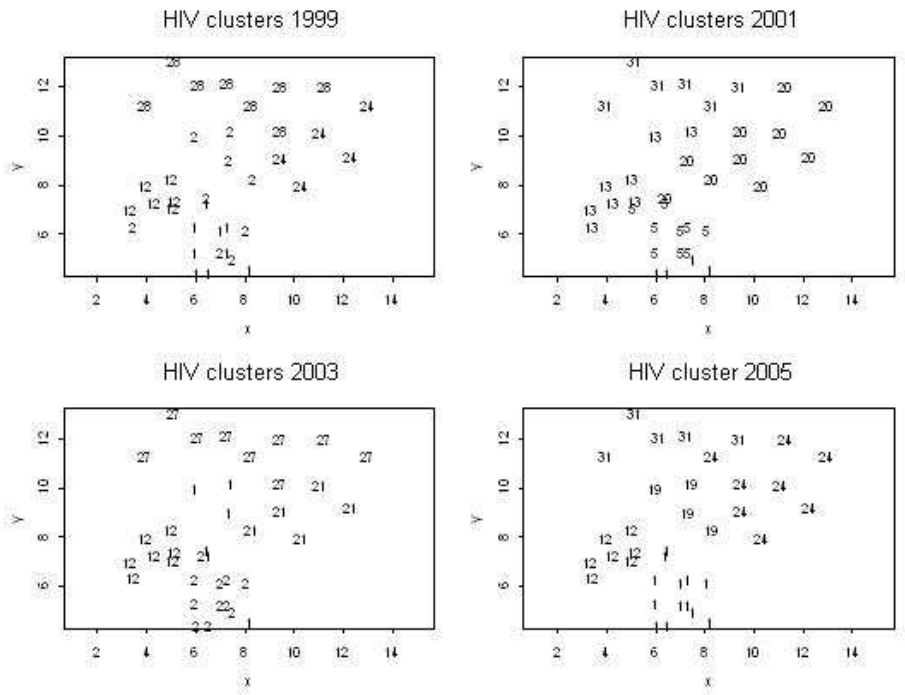

Figure 3.7. *Major HIV clusters in Nigeria*

3.2.2 The Regression Tree

The search for possible structural patterns of the prevalence of HIV among the neighbourhood groupings of the States led to the use of regression trees. The resulting pruned regression trees and their corresponding partition plots obtained using S–plus are shown in Figure 3.8. The tree groups local neighbourhoods where estimated prevalence rates are similar. The resulting partitions show the estimates of the average HIV prevalence rate for the group of States in each leave or partition. The partitions follow the observed trend of the epidemic observed in previous sections. However, the estimate of average prevalence rate for the Northwest in 1999 and 2001 is a bit of a surprise as most of the states in that partition are observed to have low prevalence rate. The trees for 2003 and 2005 appear to have more efficient partitions and estimates as expected given the observed data. Their partitions are consistent with those of the contour plots where the far North and the Southwest were distinguished for their low

prevalence. In all the plots, the North-central zone stands out as the hot point of HIV. The north-south or vertical clustering of the states in the early years of the epidemic is visible from the vertical strips in the plots. The 2003 tree attempted to group some states in the South-south with very high prevalence rates with their counterparts in the North-central. This phenomenon was observed in the hierarchical clustering where Akwa Ibom State in the South-south was clustered with Benue State in the North-central. The vertical partitioning seem to give way to horizontal strips in the 2005, this may suggest homogeneity of average prevalence rates among immediate neighbouring states in the recent years.

3.3 The Search for Spatial Trends

In the last section, we examined the tendency of HIV cases to cluster in some local neighbourhoods. In this section, we explore the spatial trend of the disease through the States as we travel from the west to the east and from the north to the south. To achieve this, we fit smooth functions of the prevalence rates on the latitude and the longitude using GAM function in S+SpatialStats. The local regression (loess) was used as the smooth function through the northing and easting and the State prevalence rates for 1999, 2001, 2003 and 2005 were used as the separate responses. The plots in Figure 3.9 show the results of the two-way GAM fit- easting and northing. Each plot indicates a steep rise toward the centre and sloping downward towards the west and far north. The north-south plots (labeled as Northing in the x-axis) for 2001, 2003 and 2005 show high prevalence of HIV in the south-south Nigeria. This is indicated by the rise in the lower part of the graph. The east-west plots (labeled Easting) depict also that prevalence is higher in the east than the west.

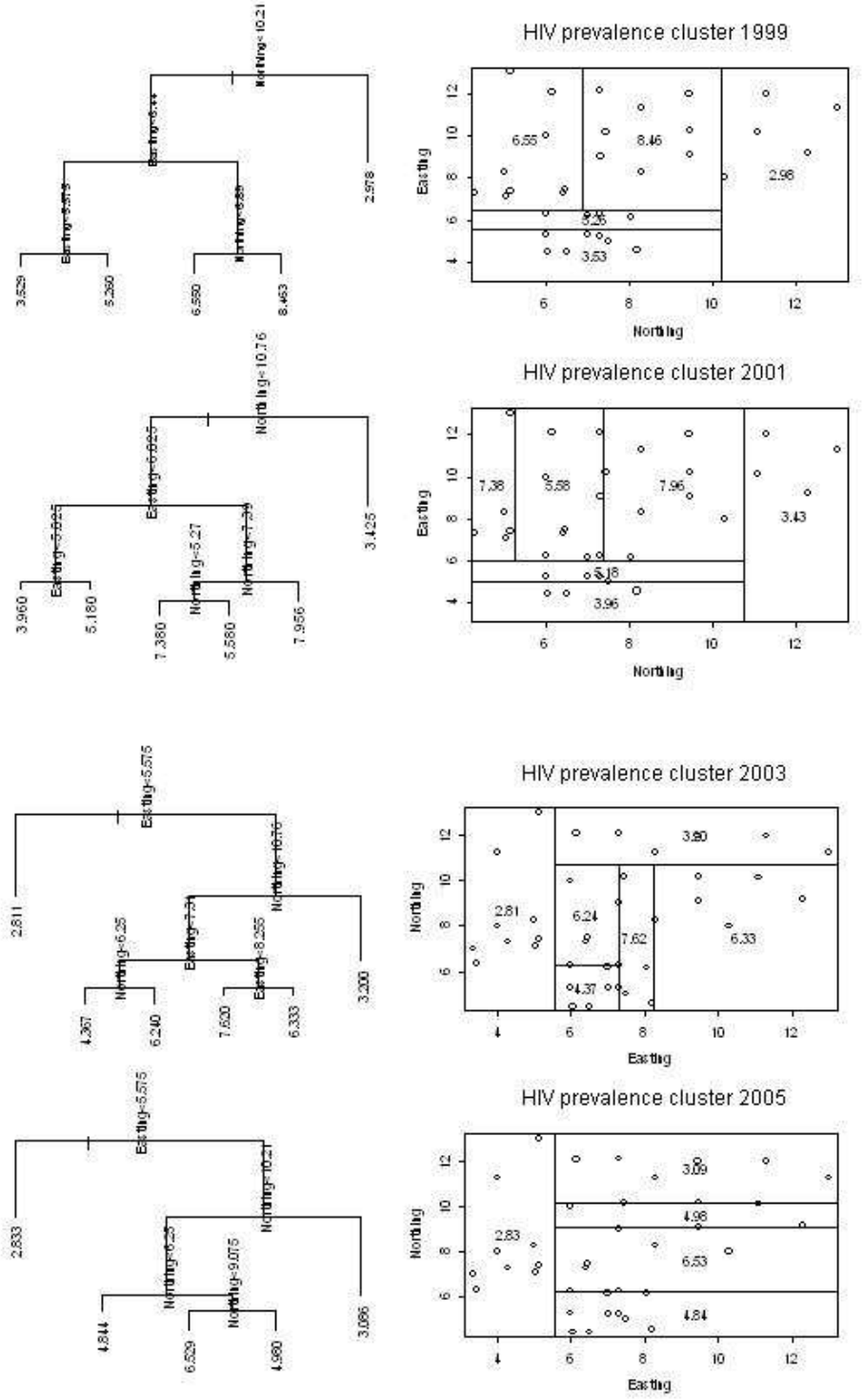

Figure 3.8. *Regression tree of HIV clusters in Nigeria*

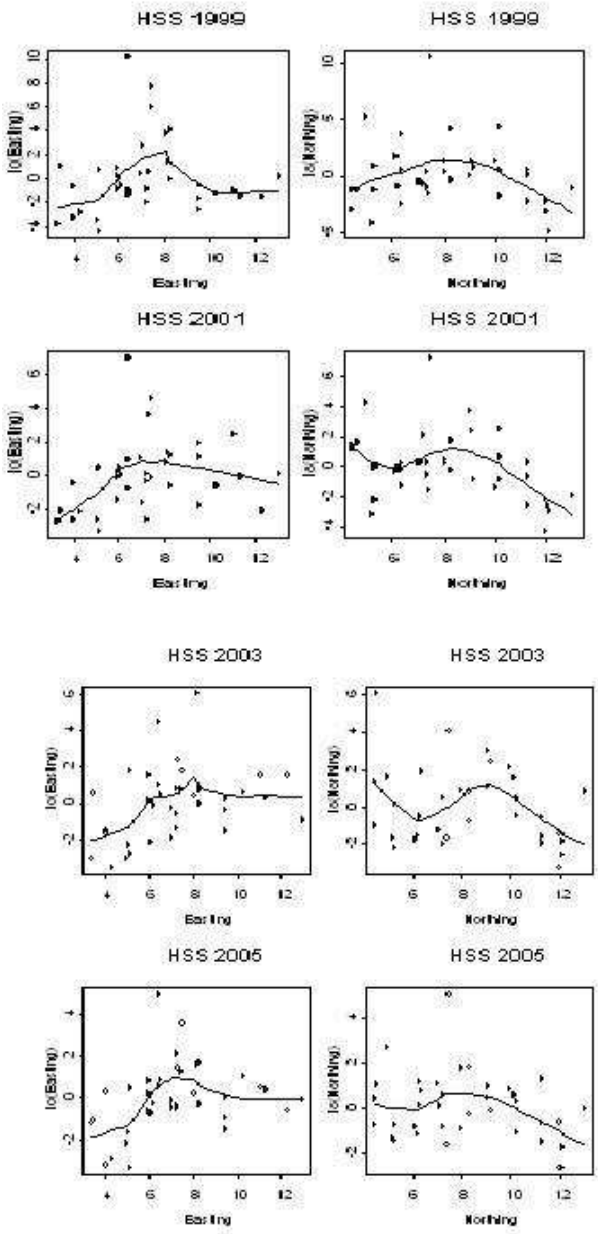

Figure 3.9. *Plots of spatial trend of the log prevalence obtained using a two-way GAM*

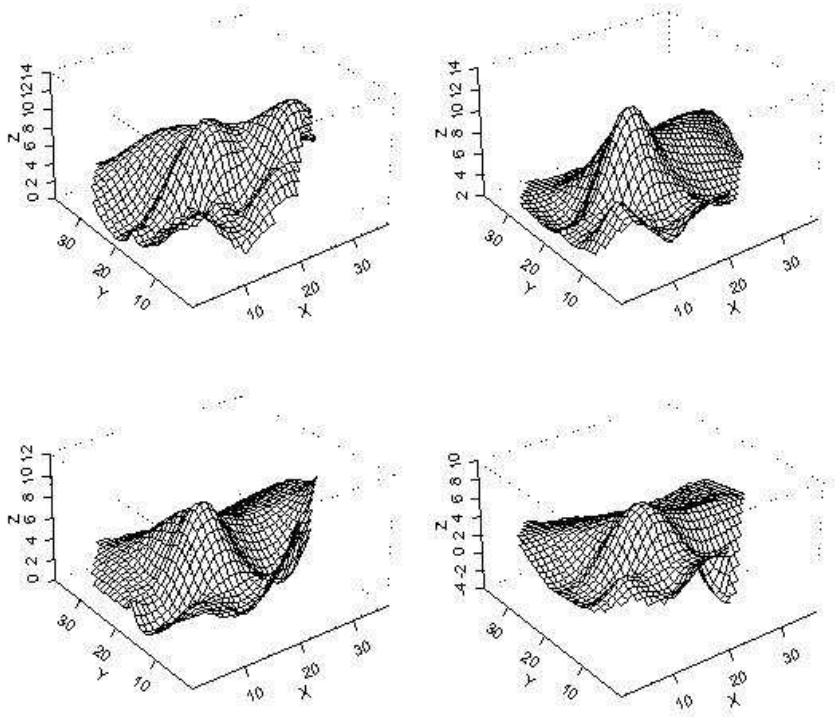

Figure 3.10. *Plots of surface trend of the natural log transform of HIV prevalence rates obtained using loess function*

It is important to note that these plots allow us to view the trend only in two directions because of its additive nature. In order to examine the entire trend surface, we fit a loess function to the data, making prevalence a function of the product of the two coordinates. A better view of the trend surface is shown in the surface plot of the predictions. The plots are given Figure 3.10. The individual plots represent predicted values of the levels of HIV prevalence in the Nigerian States for the four years under consideration. For all the years, the heavy concentration of HIV cases in the North central is very distinct. The elevated points in the South-South correspond to the estimates for Cross Rivers and Akwa Ibom States.

The protruding is more prominent when a nonparametric smoothing method (293) is applied to the data as shown in the plots in Figure 3.11. The consistency of high prevalence in the North central and South-south zones over the years is

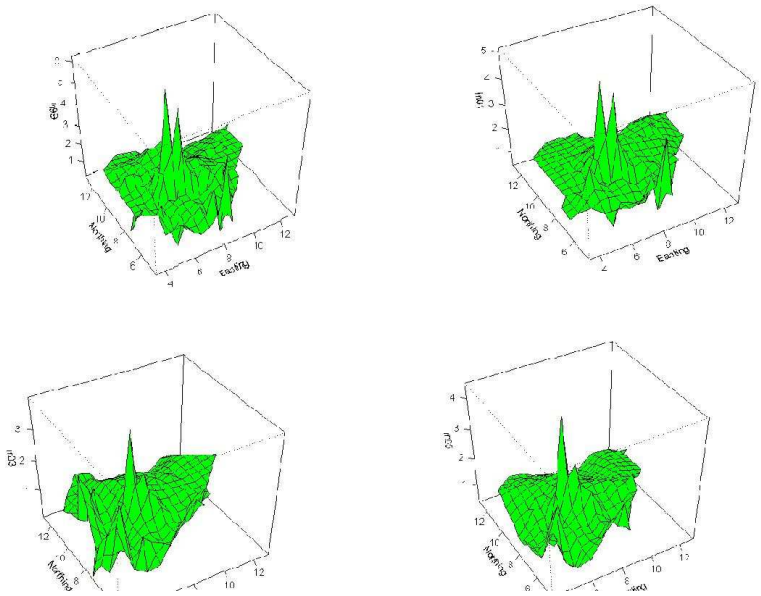

Figure 3.11. *Plots of surface trend obtained using sm regression*

clearly depicted in the plots.

3.4 Measures of Spatial Correlation

In this section, we examine how prevalence rates are related to each other within a defined distance and direction. It is expected that sites or States that are close neighbours will have similar rates than States or sites far apart given the trado-cultural and socio-economic behaviours that may transcend political boundaries.

3.4.1 Correlogram

The plots shown in Figure 3.12 are correlogram plots. Each point in the graph is a measure of the relationship (represented by ρ) between the natural log prevalence rates in states separated by a common Euclidean distance. The

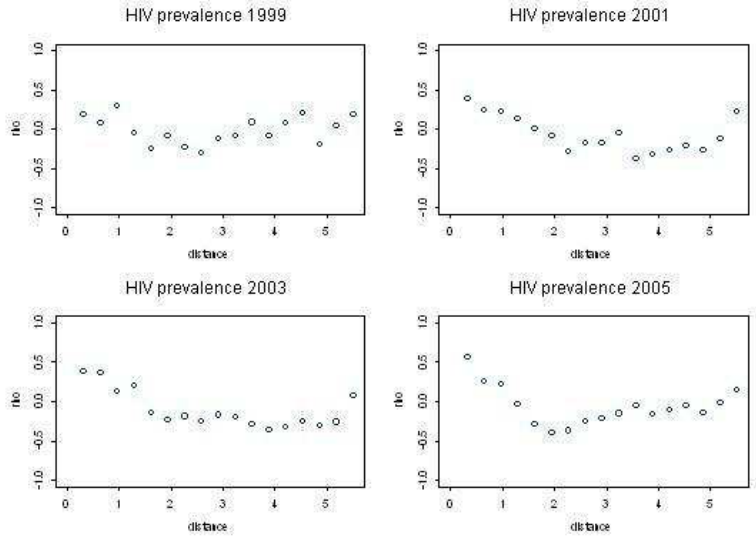

Figure 3.12. *Correlogram plots of the natural log transform of HIV prevalence rates*

distance between any two State capitals is of interest here. It is clear from the plots that relationship tends to diminish as distance increases. Note the high positive correlation between states whose Euclidean distance is less than one unit (about 100 miles). This may be interpreted to mean that HIV epidemic is similar among states that are close neighbours.

The variation in time of the spatial autocorrelation is visible in the plots in Figure 3.12. The correlation between close neighbours appears to increase with years. If we perceive correlogram as a measure of similarity of HIV prevalence among close neighbours, it may be right to say that the prevalence of HIV infection becomes more and more similar among States that are close neighbours as the years go by. The absolute value of the slope of the plots seems to increase with time. The covariogram for this data also show the same behaviour. We note here that the distance between the centroids of the state capitals were calculated from the data locations measured by their longitude and latitude. One unit in x-axis is approximately 100 miles.

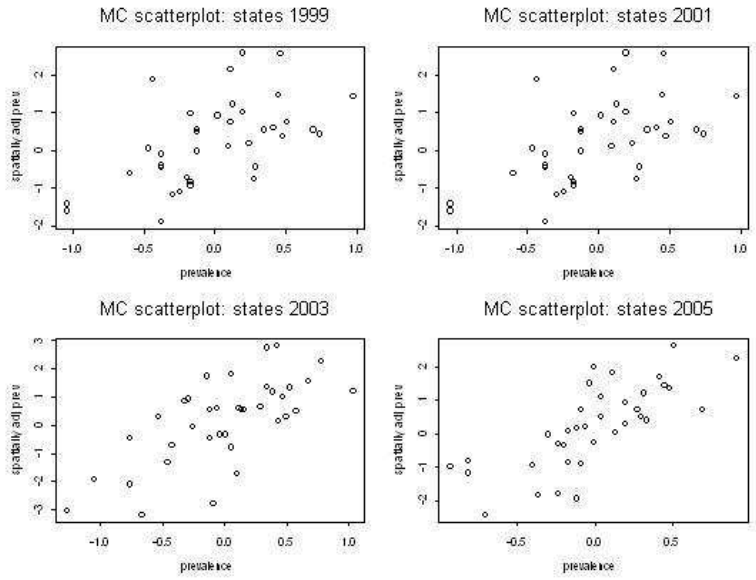

Figure 3.13. *MC scatter plots of the natural log transform of HIV prevalence rates*

3.4.2 The Moran I and Geary c Statistics

The Moran Coefficient (MC) is a covariance measure of spatial autocorrelation. Its scatter plots shown in Figure 3.13 were obtained by plotting centred value of the log prevalence rates weighted by the neighbourhood matrix against the centred log prevalence rates (7). The graph can be described as plotting the spatially adjusted rates against the rates corrected for mean. The plots indicate positive autocorrelation that tends to increase with time. This suggests that the number of States with similar prevalence rate clustering within a defined geographic space increases with time.

MC is known to be the most powerful test for spatial autocorrelation (7),(111) (62), (57). Table 3.2 shows the computed statistics for Moran (217) and Geary(235) measures of spatial autocorrelation. The Moran coefficients were estimated using the regression ratio approach and the Moran index, both methods gave identical estimates. All the coefficients are significant at 5% level of significance considering

	Moran		Geary	
Year	coef	p-value	coef	p-value
1999	0.294	7.18E-4	0.706	0.021
2001	0.293	7.27E-4	0.611	0.0023
2003	0.423	2.09E-6	0.539	2.96E-4
2005	0.476	1.16E-7	0.485	5.31E-5

Table 3.2. *Estimates of spatial autocorrelation using Moran and Geary statistics*

the normal p-values and the permutation p-values of the estimates. Therefore, there is a strong evidence against the null hypothesis of no spatial correlation in the data sets. As seen in the scatter plots, these coefficients confirm positive autocorrelation among the States over the four-time period. Hence, these tests confirm our earlier finding that there exist different clusters of the epidemic in the country. That is, there appear to be different forms of the epidemic in the country. Nearby States on average, have similar prevalence rates and the number of state forming these clusters appear to increase with time as depicted by the increase in Moran coefficient over time.

The Geary coefficient (GC) is never negative and it is sensitive to particularly large differences between the prevalence rates (62). The Geary Coefficients as shown in Table 3.2 seems to suggest that the differences between the prevalence rates declined over the years. This may explain the decreasing trend exhibited by its coefficient over time. However, it is noteworthy that the difference between its values and 1 (the mean of the null model) increased with time. This increasing difference signify increasing strength of positive spatial correlation in the data over the years. The further away the Geary coefficients are from 1, the stronger the spatial correlation. Thus, it appears from the result that HIV prevalence rates among the States become more and more similar with time and this is especially so among States that are geographically close to each other. This finding is in line with the result obtained using Moran I statistic.

3.4.3 The Semivariogram Clouds

We employ the tool of variogram cloud in search of possible spatial structures and potential spatial outliers inherent in the HIV prevalence data. Also, the distribution of the variance between all possible pairs of States at some selected Euclidean distances will be examined. We consider two types of variogram clouds- the squared-differences and the square-root differences cloud.

The plots in Figure 3.14 are the squared-difference cloud for the four-time periods - 1999, 2001, 2003, and 2005. The presence of spatial outliers is evident in the four plots. The unusual values noticed in three of the four-time period are due to the state numbered 22. This state is Benue state located in the North Central geopolitical zone of Nigeria. The possible spatial outlier in 2003 is the state numbered 1. This is Cross Rivers state located in the South-South geopolitical zone of Nigeria. A closer look at the plots show that the state numbered 2 closely follows state number 22 in 1999, 2001 and 2005. State number 2 is Akwa Ibom state in the South-South geopolitical zone of Nigeria. These three States are known for their high HIV prevalence rates. It does appear from the graphs that the number of points that constitute outliers tend to decrease with time, leading us to suspect that other states are closing the gap with time. The spatial structure of the HIV prevalence among the states appears to be approximately the same in the four-time points.

The spatial variation with increasing distance is clearer when viewed from the boxplots of the variogram cloud given Figure 3.15 below. The boxplots seem to define a general pattern of low variation in the first 200 miles (southwest zone) followed by high variance in the central and low variation in the far north. This observed pattern is in line with the results obtained from the loess estimate. The

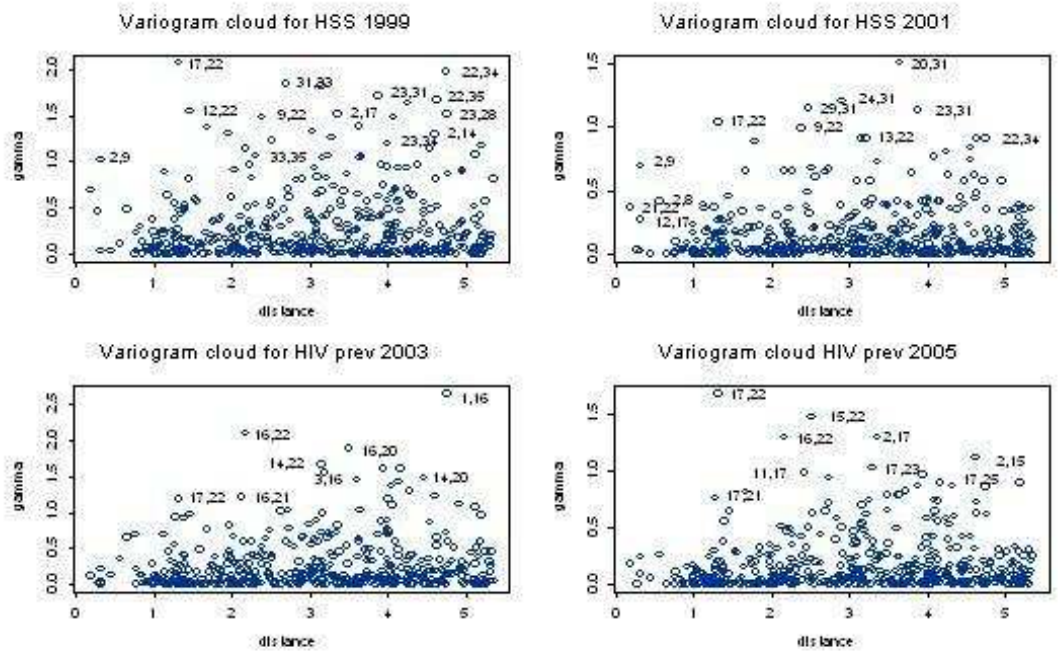

Figure 3.14. *Variogram cloud plots of the natural log transform of HIV prevalence rates*

distance is estimated from the south-south to the far north using the longitudes and the latitudes of the state capitals. The boxplot at the near zero distance depict states with abnormal prevalence rate in the south-south zone (Cross Rivers and Akwa Ibom) as identified in the preceding paragraph and in the North-South GAM plot (Figure 3.9). The outliers are more distinct. However, the distribution of the variogram cloud is skewed (185) (237), hence care should be taken in labelling any observation as unusual or outlier based on the variogram clouds.

The plots in Figure 3.16 are intended to investigate what happens to the spatial structure and the seeming outliers as the distance of the variogram cloud is reduced from approximately 5 to 3units using the 2005 HIV prevalence.

This reduction in distance shows some form of spatial structure with the variance fluctuating with increasing distance. The elimination of the State number 22 (which was noted as an outlier) from the reduced distance semivariogram cloud (second row of Figure 3.16) only slightly improved the variability structure as

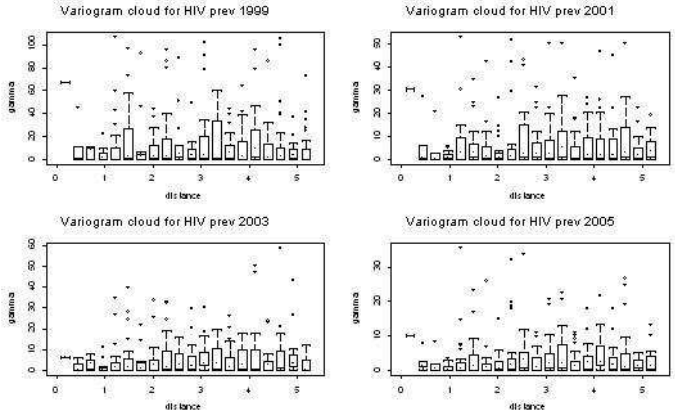

Figure 3.15. *Variogram cloud box plots of the natural log transform of HIV prevalence rates*

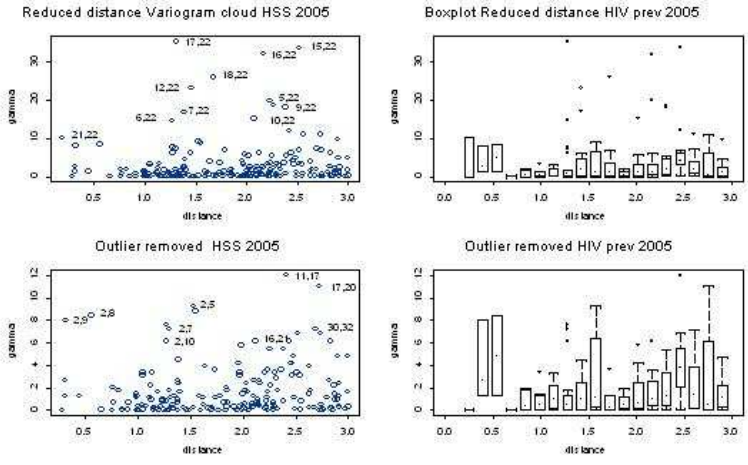

Figure 3.16. *Reduced distance Variogran cloud plots of the natural log transform of HIV prevalence rates*

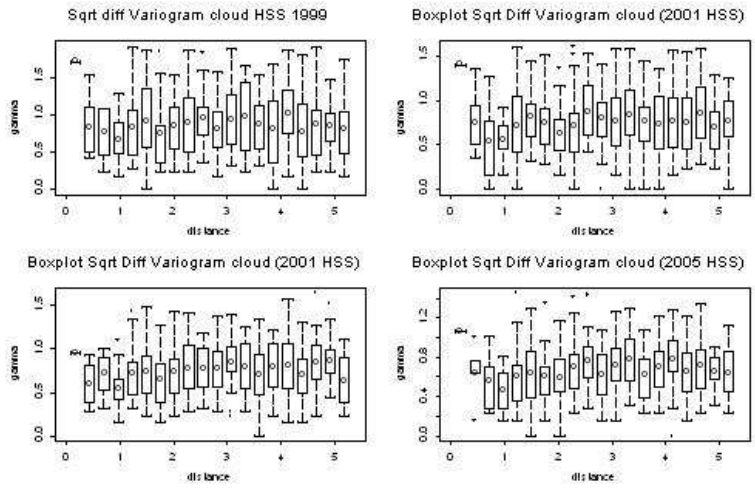

Figure 3.17. *Square root semivariogram cloud plots of the natural log transform of HIV prevalence rates*

depicted by the boxplots but some outlying points are still visible. State number 2 (Akwa Ibom state in the south-south zone with a prevalence rate of 8%) is a possible spatial outlier at very low distance. State number 17 (Ekiti state in the southwest zone) with the least prevalence rate of 1.6% may be a possible outlier at a distance of about 300miles. This goes to confirm that squared differences semivariogram cloud may not be the best tool to investigate outlying points in spatial analysis

The square-root differences cloud is considered in Figure 3.17. The symmetry of the boxplots is improved and very few outliers are observed. The abnormal boxplot at the near zero distance in all the four year period is outstanding. A line joining the means (o's in the box) or the medians (dots in the box) gives a fair idea of the pattern or trend of the variation in the HIV prevalence among the states in Nigeria.

3.4.4 The Semivariogram

In order to investigate further the spatial correlation observed in the correlogram plots above, we use the semivariogram. The semivariogram will throw more light on the nature, degree and extent of spatial correlation. The directional variogram describes not only how data are related with distance but also with direction. In order to obtain a better view of the nature of the spatial correlation, we need to remove the spatial trend earlier observed in the data when we applied the generalized additive model.

Omni directional semivariogram

The directional semivariograms for 1999 and 2005 (Figure 3.18) were obtained using the residuals from the local regression models. It does appear that the positive trend observed in the various directions has been removed and the shape of the semivariogram appears to be the same for all directions. Therefore, it may be right to say that spatial dependency of HIV prevalence is approximately the same in all directions. This implies isotropy or absence of anisotropy.

Using the residuals obtained from the loess estimates, the isotropic empirical spherical semivariograms in Figure 3.19 were obtained and the lines fitted using the weighted nonlinear least squares. The close fitting of the empirical semivariograms to the model suggests a good choice of model for the spatial dependence structure of the model. There is a marked increase in the range of the semivariograms in 2001 indicating that the number of States in close geographical space with similar prevalence could be more in 2001 than other years. The range and the nugget effect are least in 2005 implying that spatial correlation is higher among States separated by shorter distances in geographic space. The shallow

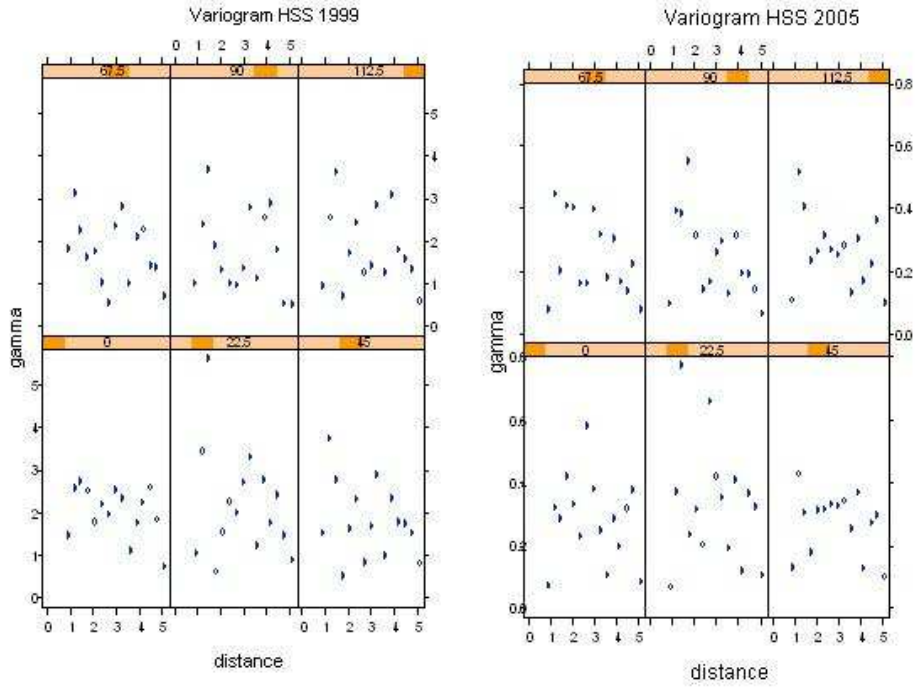

Figure 3.18. *Omnidirection semivariogram plots of the natural log transform of HIV prevalence rates*

slope of the semivariograms suggests the presence of positive spatial correlation. It indicates less variability in the prevalence rates with increasing distance. Also, since the sill is a measure of the overall spatial variability, the decrease over time can be explained to mean that the variation in prevalence rates of HIV in the States within the distance defined by the spatial lags decreases with time. This suggests that as the years go by, the prevalence of HIV in States close to each other becomes increasingly similar.

3.4.5 Kriging

Kriging estimates were obtained using universal kriging approach by fitting a cubic polynomial to the residuals obtained from the loess estimates. The dependence structure of the spatial distribution was modelled by the empirical spherical semivariogram which was established in the previous section as being appropriate

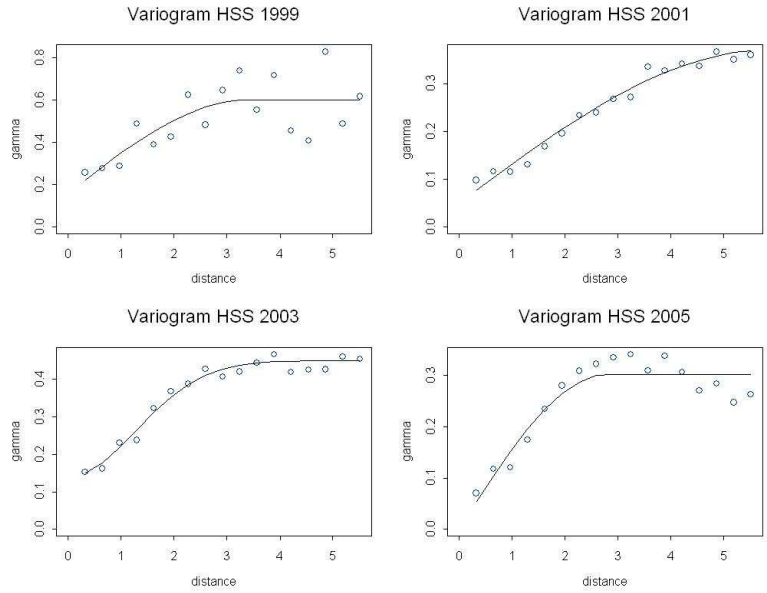

Figure 3.19. *Spherical Variogram plots of the natural log transform of HIV prevalence rates*

for the model. Our choice of the universal approach is informed by the need to simultaneously model the spatial correlation and trend. The graphs in Figure 3.20 show the wireframe plots of the kriged estimates. The spatial patterning of HIV prevalence rates is vivid in the plots. Going from west to east, the low prevalence rates of the west and far north east and the outstanding high prevalence of north central and the moderately high prevalence of the south-south is distinct. This spatial structure defines the spatial trend adequately.

All the analyses in the sections above were done using HIV sentinel survey data for the 36 States and the Federal capital Territory. Similar analyses were also conducted using the data at site level. Results and interpretations obtained are identical to those presented above.

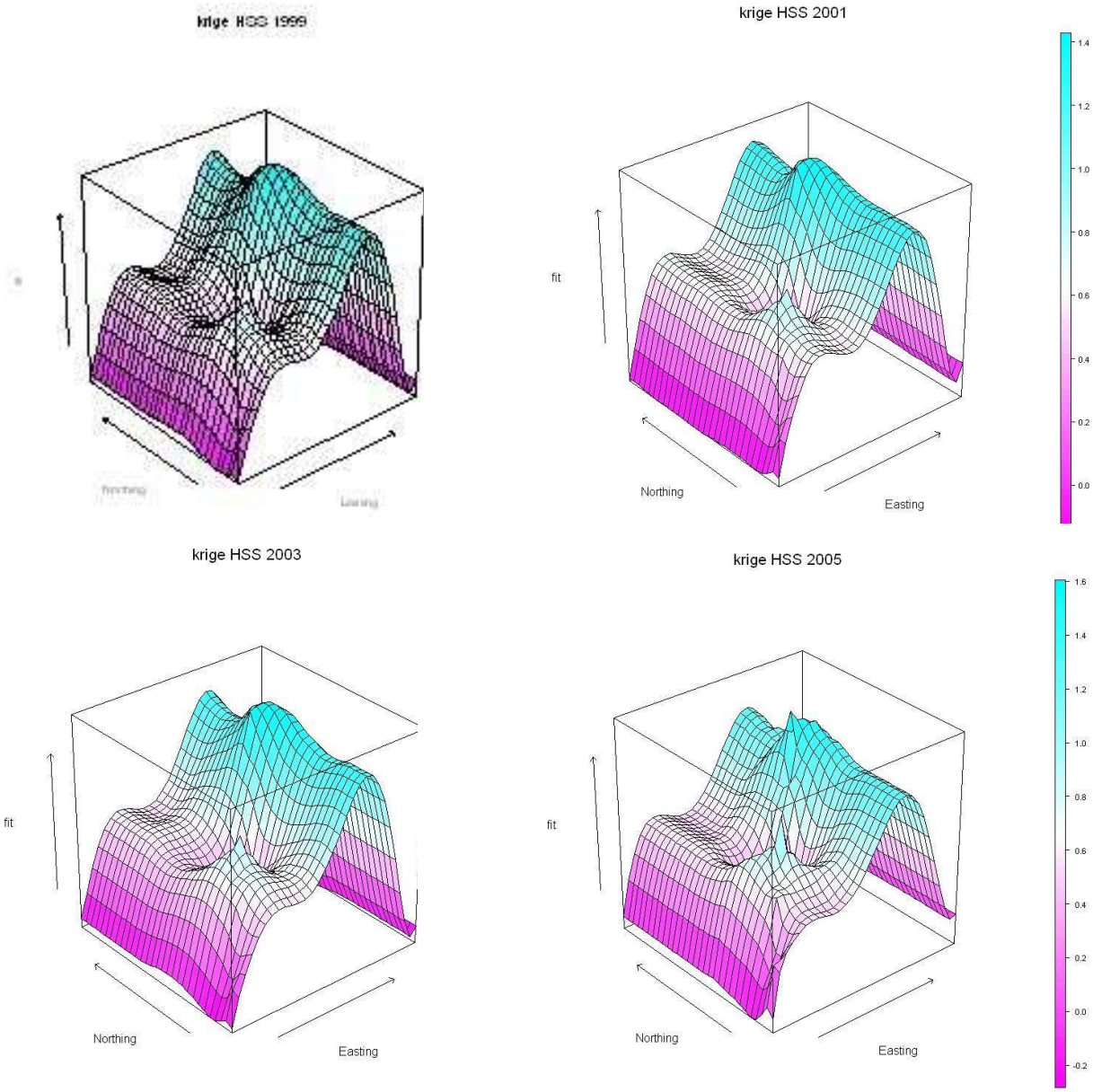

Figure 3.20. *Plot of krige estimates of the natural log transform of HIV prevalence rates*

Chapter 4

The multi-level models

4.1 The Variance Component model

Multilevel analysis allows the extra Poisson variation inherent in the HIV positive counts to be modelled. Thus, it is possible to estimate the dependence of HIV positivity on some ecological covariates simultaneously with a measure of variation at the site, state and zone group levels. We shall estimate the variance components in various models starting with the null model given in equation 4.1 as

$$O_i \sim Poisson(\mu_i)$$

$$log(\mu_i) = log(E_i) + \alpha + u_i \tag{4.1}$$

where u_i is the heterogeneity effect at the site level and $u_i \sim N(0, \sigma_u)$. It is assumed here that all the variation is at a single level that of sites, the lowest level of aggregation. An extension of this model comprises of other possible sources of variation as given in equation 4.2 where site i is nested in state j and zone k:

$$log(\mu_{ijk}) = log(E_i) + \alpha + u_{ijk} + v_{jk} + y_k \tag{4.2}$$

where

$$u_{ijk} \sim N(0, \sigma_u),\ v_{jk} \sim N(0, \sigma_v),\ y_k \sim N(0, \sigma_y)$$

are respectively variance components at the levels of sites, states and geopolitical zones. $Log(E_i)$ is the logarithm of the expected number of cases which is an offset that accounts for the different populations at risk of infection. α is the constant term which is required if $log(E_i)$ are centered, if the sum of the observed number of cases is not equal to the sum of the expected number of cases or if the covariates x_i are not centered (113). In our case, the last condition holds.

Some zone-level variables such as condom use within a group, availability and cost of treatment, norms, number of partners, and number of sexual contacts may influence the risk of HIV infection in an individual within the zone. Therefore, these zone-level factors have direct and indirect effects (72),(291),(157) on the individual's risk of infection. In this analysis, we have used some ecological factors known to affect the variation in the distribution of number of HIV cases. Eight risk factors at zonal level are considered: the median age at first sex (in years), proportion of women engaging in risky sex, proportion in polygamous marriage,

proportion using condoms, percentage literacy level, proportion who had sex at least once in the last week preceding the survey, proportion using condoms in risky sex and proportion who have a history of other sexually transmitted infections (STIs). These covariates were obtained from the National Demographic and Health Survey (NDHS)(193) conducted in 2003. These data pertain only to women aged between 15-49 years and are only available at the zone or regional level. We also note that the HIV data used in conjunction with these covariates also pertain to women of the same age group. The HIV data were collected from the National HIV Sentinel Surveillance (92) which surveyed pregnant women aged 15-49 years attending ante-natal clinics in the 85 selected survey sites in 2003. See Chapter 1 for detailed discussion of data sources.

Hence a model with single (site) level and the covariates is given as

$$log(\mu_i) = log(E_i) + \alpha + X\beta + u_i \tag{4.3}$$

where X is the design matrix containing values of the ecological factors

$u_i \sim N(0, \sigma_u)$

A two-level (site and state) model is

$$log(\mu_{ij}) = log(E_i) + \alpha + X\beta + u_{ij} + v_j \tag{4.4}$$

where site i is nested in state j and

$$u_{ij} \sim N(0, \sigma_u),\ v_j \sim N(0, \sigma_v)$$

For three levels (site, state and zone), the equivalent model is given as

$$log(\mu_{ijk}) = log(E_i) + \alpha + X\beta + u_{ijk} + v_{jk} + y_k \tag{4.5}$$

where site i is nested in state j and zone k and

$$u_{ijk} \sim N(0, \sigma_u),\ v_{jk} \sim N(0, \sigma_v),\ y_k \sim N(0, \sigma_y)$$

are variations due to sites, states and zone differences respectively.

The variance component models can be fitted using the quasi-likelihood, iterative generalized least squares (IGLS), Fisher scoring algorithm or the restricted iterative generalized least squares (RIGLS). The detailed account of the algorithm for the estimation procedure of the multilevel model is given in Goldstein (105).

The estimates of the fixed and random components of the model were obtained using the restricted maximum likelihood (REML) method in $R^{©}$ and the iterative generalized least squares (IGLS) in $MLWin^{©}$. The full model using the eight

covariates described above is given as

$$log(\mu_{ijk}) = log(E_i) + \alpha + \sum_{i=1}^{8} x_i\beta_i + u_{ijk} + v_{jk} + y_k \tag{4.6}$$

In the course of the iterative procedure in the estimation of parameters, it was observed that some of the covariates were collinear as estimates obtained using $MLwiN^{©}$ in some cases were misleading or even meaningless and, in $R^{©}$, the X^TX matrix was not positive definite. To overcome this problem, we adopted the forward-backward stepwise regression approach in order to select the best subset of regressors that will be used in 4.6. The final estimates of the model given in Table 4.1 and 4.2 were obtained in $R^{©}$ using the Laplace method.

Predictor	Model 1 Estimate (s.e)	Model 2 Estimate (s.e)	Model 3 Estimate (s.e)
Fixed Part			
α	-0.1265(0.063)	-0.1379(0.078)	-0.1325(0.134)
Random Part			
$\sigma^2_{u(\text{site})}$	0.259(0.509)	0.138(0.372)	0.142(0.377)
$\sigma^2_{v\ (\text{state})}$		0.127(0.356)	0.041(0.202)
$\sigma^2_{y(\text{zone})}$			0.0844(0.290)
Extra-Poisson	0.999	0.998	0.990

Table 4.1. *Estimates from the null model and the Variance Component models*

Table 4.1 shows the results for the null and the variance component models. All the three models adequately accounted for the extra-Poisson variation inherent in the HIV prevalence rates as indicated by the last row of Table 4.1. Model 1 assumes that the variation in HIV infection in the country is due solely

to differences between the sites. Comparing this with Model 3, the contribution of the various hierarchies of aggregation of the population to the variation in HIV infection is made more visible. The table shows that about 32% of the variation in infection is due to the differences among the geopolitical zones. Differences among the states accounted for about 15% of the total variation. However, there is large variability among the sites. This may suggest the presence of spatial clustering of the sites.

Table 4.2 shows the models with the selected regressors. Of the eight covariates considered, three were found to be significant in explaining the variation in HIV positivity especially among the zones. The covariates are; regular use of condom, polygamy and involvement in risky sexual behaviours. The inclusion of these social and behavioural variables in the model reduced the random variation by almost 34%. It is worthy of note that variation at the zone level is completely explained by these zone-level covariates, as shown under Model 3 in Table 4.2. Also, there is a larger variability within the states (measured by the site variance of 0.14525) than between the states ($\sigma_v^2 = 0.02575$). This suggests a need to further investigate the spatial effects of the site locations.

As can be seen, polygamy and risky sexual behaviours are positively associated with the risk of HIV infection in the country, while condom use is shown to be negatively related to HIV positivity. Hence, any intervention programme targeted towards reduction in the prevalence of polygamy and risk behaviours and increase in the use of condoms could expected to be very effective in fighting the spread and the scourge of HIV/AIDS epidemic in Nigeria. We need to state that the parameter estimates in each of the models are log relative risks of being HIV positive. Therefore, the relative risk for condom use as obtained in Model 3 is $exp(-0.02518) = 0.975135$. It then means that for every 1 percent point increase

Predictor	Model 1 Estimate (s.e)	Model 2 Estimate (s.e)	Model 3 Estimate (s.e)
Fixed Effects			
α	0.0155(0.504)	-0.000425(0.538)	0.01437(0.538)
β_1 (CondomUse)	-2.4676(0.777)	-2.5019(0.832)	-2.5180(0.832)
β_2 (Polygamy)	3.1494(1.053)	3.234(1.125)	3.2005(1.125)
β_3 (Riskysex)	1.8706(0.521)	1.895(0.556)	1.89114(0.556)
Random Effects			
$\sigma^2_{u(\text{site})}$	0.1715(0.414)	0.14537(0.381)	0.14525(0.381)
$\sigma^2_{v\ (\text{state})}$		0.0256(0.16)	0.02575(0.16)
$\sigma^2_{y(\text{zone})}$			$5.00e-10$(2.2e-05)
Extra-Poisson	0.9866	0.9874	0.9874

Table 4.2. *Estimates from the Variance component models after stepwise selection of covariates*

in the use of condom at the zone level will reduce HIV infection in the zones by 2.49%. Similarly, 1 percent increase in polygamous and risky sexual practices is associated with increase in the risk of HIV infection by an estimated 3.25 and 1.91% respectively.

Due to the multicolinearity experienced in the estimation of the model in $R^{©}$, we explore, in the next section, the use of empirical Bayes procedure in the derivation of estimates for the multilevel models.

4.1.1 Empirical Bayes Estimation

.

The estimates obtained using the empirical Bayes procedure implemented in $WinBUGS14^{©}$ are shown in Tables 4.3, 4.4 and 4.5. Table 4.3 show estimates from the full model while Table 4.5 show the estimates obtained from the best regressors. Due to the problem of multicollinearity, some estimates from the

full model are misleading. The sign of some of the estimates is contrary to what is expected in real life situation. For instance, literacy and condom use in risky sex are estimated as being positively associated with the risk of HIV infection and risky sexual practices appear to be negatively associated with HIV infection. Although these estimates are not statistically significant in explaining variations in HIV infection, we expect that, at least, they should be meaningful in real world situations. To obtain a better model, we adopted the backward stepwise regression and Table 4.4 show more meaningful estimates when some of the covariates were removed from the model. Table 4.5 contains the final model.

More variables were found to be significant in explaining variations in the risk of HIV infection than were obtained using REML implemented in $R^{©}$. Surprisingly, age at sex debut appear to be positively associated with risk of HIV infection. The reason for this may be attributed to differences in culture among the zones. Age at first sex is lower in the core Muslim states in the far North of the country where HIV prevalence is also low and higher in the southern states where HIV prevalence is relatively high. Also, this calls to mind the issue of ecological bias where the group effect is at variance with (or not equal to) what is expected at the individual level. That is, the group effect parameter is not equal to individual-level parameter (144) (253). Polygamy and frequency of sex are positively associated with the risk of HIV infection. Results from Table 4.5 indicate that a percentage increase in polygamous practices may likely trigger the risk of HIV infection by an estimated 12.9% across the zones and a percentage point increase in exposure to sexual contacts is estimated to increase the risk of HIV infection by about 7.0%. However, the same unit increase in the use of condom in each sexual contact is estimated to reduce the risk of infection by 3.4% in the zones. Also, it is interesting to note from Table 4.4 that increase in

Parameters	Estimate	s.e	95%CI
Fixed Effects			
α	-9.269	2.794	(-14.34, -3.309)
Sexage	0.282	0.171	(-0.075, 0.556)
Risksex	-3.518	6.076	(-15.71, 7.393)
Polygamy	10.8	4.519	(2.254, 19.92)
Literacy	3.224	3.817	(-3.459, 10.68)
Usecdmrs	1.262	1.728	(-2.338, 4.575)
Freqsex	3.376	1.421	(0.605, 6.253)
CondomUse	-3.436	1.382	(-6.034 , -0.668)
STI	15.33	14.9	(-11.68, 44.07)
Random Effects			
$\sigma^2_{u(\text{site})}$	0.211	0.0492	(0.131, 0.320)
$\sigma^2_{v(\text{state})}$	0.021	0.027	(6.5e-4, 0.095)
$\sigma^2_{y(\text{zone})}$	0.065	0.210	(6.7e-4, 0.450)

Table 4.3. *Estimates from the full model of the Variance component model using empirical Bayes method*

literacy level is associated with a reduction in the risk of HIV infection by 1.3%. Although the estimate of relative risk for literacy is not statistically significant, it worthy of note that investment in education aimed at increasing literacy levels across the zones has some positive contribution in the fight against HIV/AIDS. Also a percentage point increase in risky sexual behaviours can enhance the risk of HIV infection by as much as 3.06%.

Relating this result to the information from the covariates as shown in Table 4.6, polygamy is highest in the Northeast region and lowest in the Southeast region of the country. It then means that a population where polygamous practice is at the same proportion with Northeast zone has a risk of HIV infection that is 12.9% higher than a society where polygamy is at the same proportion with the south east zone. Likewise, frequency of exposure to sex is lowest in the Southwest region and highest in the Northwest region. Therefore, a society where exposure to sex is as frequent as in the Northwest zone has the risk of HIV infection that is 6.94% higher than a population where exposure is the same as the Southwest

Parameters	Estimate	s.e	95%CI
Fixed Effects			
α	-10.91	4.956	(-24.29,-4.313)
Sexage	0.5172	0.2339	(0.1573,1.115)
Risksex	3.012	2.34	(-1.573,7.41)
Polygamy	7.236	3.608	(0.7879,14.66)
Literacy	-1.274	1.967	(-5.169,2.241)
Freqsex	3.469	1.943	(0.2626, 8.378)
CondomUse	-3.247	1.469	(-6.055 , -0.4358)
Random Effects			
$\sigma^2_{u(\text{site})}$	0.2093	0.0482	(0.1308, 0.3155)
$\sigma^2_{v(\text{state})}$	0.02	0.0269	(0.01022, 0.0923)
$\sigma^2_{y(\text{zone})}$	0.06418	0.138	(0.0199, 0.3885)

Table 4.4. *Estimates from the Variance component model using empirical Bayes method*

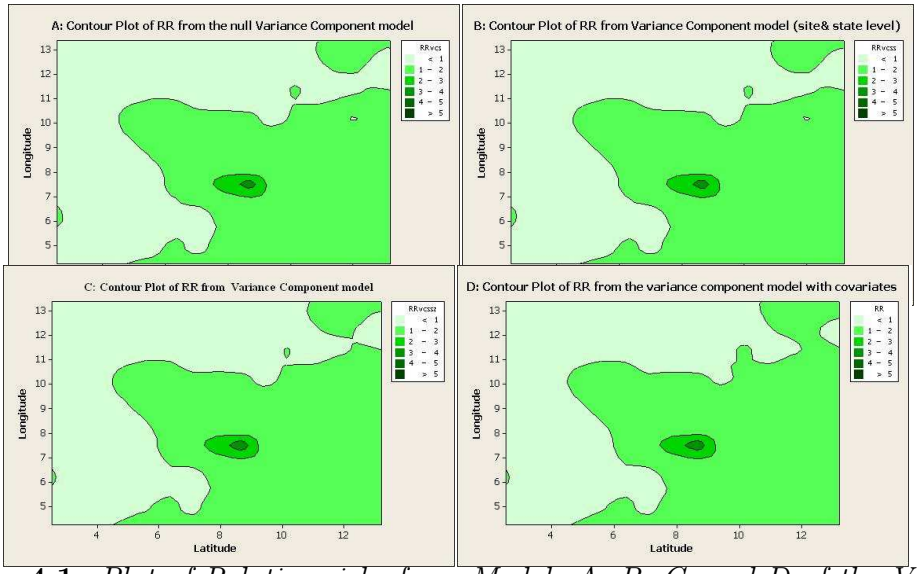

Figure 4.1. *Plot of Relative risks from Models A, B, C, and D of the Variance Component models*

Parameters	Model A Est(s.e) 95%CI	Model B Est(s.e) 95%CI	Model C Est(s.e) 95%CI	Model D Est(s.e) 95%CI
Fixed Effects				
α	-0.137(0.069)	-0.136(0.074)	-0.148(0.182)	-21.38(6.289)
	(-0.274, -0.005)	(-0.283, 0.008)	(-0.555,0.216)	(-32.7,-11.43)
Sexage				0.9765(0.2755)
				(0.555,1.471)
Polygamy				12.1(3.373)
				(5.505,18.35)
Freqsex				6.71(2.149)
				(2.981,10.79)
CondomUse				-3.474(0.868)
				(-5.172,-1.762)
Random Effects				
$\sigma^2_{u(\text{site})}$	0.317(0.063)	0.280(0.067)	0.2406(0.0537)	0.2072(0.047)
	(0.211, 0.459)	(0.166, 0.428)	(0.153,0.3621)	(0.1295,0.3134)
$\sigma^2_{v(\text{state})}$		0.0458(0.05)	0.0188(0.02592)	0.0184(0.02341)
		(0.0009, 0.178)	(0.0006,0.09234)	(0.0006,0.0838)
$\sigma^2_{y(\text{zone})}$			0.1504(0.2254)	0.0353(0.1079)
			(0.00979,0.6314)	(0.0006,0.2110)

Table 4.5. *Final models for the Variance component model using empirical Bayes method*

zone. The data suggest that condom use is highest in the Southwest zone and lowest in the Northeast and North-central zones. Hence, a population where condom use is as regular as in the Southwest zone has a risk of infection that is 3.4% lower than a population whose condom use is at the same level as the North central or North east zone.

Figure 4.1 shows the contour plots of the relative risk estimates from models A, B, C and D. The plot for the models A and B are identical. The introduction of the third hierarchy (the zone) in model C and the covariates in model D evened out some areas in the far Northeast that were isolated in models A and B. The low risk of infection in the Southwest and far North and the high risk of infection in the North-central is evident in the plots. The plots suggest the need to study the spatial clustering of HIV infection in Nigeria.

The large variability among the sites is very conspicuous from Table 4.5. Using model C, differences among the sites accounted for about 58.7% of the total variation. Differences among the zones make up about 36.7% and that of the states accounted for about 4.6% of the total variation. This calls for a check for the existence of spatial variation among the sites which may explain part of the site variation. Interestingly, the introduction of the covariates into the model reduced the total variation by 36%. Breaking down this explained proportion of variation across the hierarchies, it shows that the covariates explained 13.9% of the site differences, 2.2% of the state differences and 76.6% of the differences among the zones. Hence, due to the large variation at the site level, we are compelled to go a step further to investigate the spatial patterning of the risk of HIV infection by applying the spatial multilevel analysis. Also, the empirical Bayesian method has the limitation of not accounting for the variation in the prior parameters. To overcome this limitation, the full Bayesian method is used

Parameters	Lowest	Highest	Relative Risk
Sexage	Northwest	Southwest	1.0098
Polygamy	Southeast	Northeast	1.1286
Freqsex	Southwest	Northwest	1.0694
CondomUse	Northcentral	Southwest	0.9659

Table 4.6. *Estimates of the effects of the covariates*

in obtaining estimates for the spatial multilevel models.

4.2 Spatial Multilevel models

The spatial variation of HIV prevalence rates in Nigeria as established in the previous chapter may be more distinct if the multilevel structure of the data is incorporated. Also the large variation at the site level may be indicative of the spatial structuring of HIV prevalence in the country.

We shall seek to break down the influences on the distribution of HIV infection into three separate categories: within area effects, hierarchical effects and neighbourhood effects (112),(10). The geopolitical boundaries imposed on the states and zones of Nigeria are artificial, individuals in sites close to each other tend to share common socio-cultural, religious and behavioural factors that influence the spread of HIV. Therefore spatial smoothing of the HIV relative risk distribution might remove any variation imposed on the data as a result of the geopolitical groupings. Employing the techniques of multilevel modelling also makes it possible to account for the interclass correlation (116),(272) effects between the neighbourhood groupings.

The model incorporating the spatial effects is given as

$$Log(\mu_i) = log(E_i) + \alpha + X\beta + u_i + v_i \tag{4.7}$$

where u_i are the heterogeneity effects measuring differences between sites, and

$$v_i = \sum_{i \neq j} z_{ij} v_j^*$$

are the spatial effects which are weighted sums of a set of independent random effects v_j^*. v_j^* are independent residuals which are the effect of area j on other areas, moderated by a measure of the proximity z_{ij} of each pair of areas (112).

$$z_{ij} = \frac{w_{ij}}{w_{i.}} \tag{4.8}$$

Here, $w_{ii} = 0$. If interest is to ensure that the variance contribution is the same for all areas, $w_{i.}$ is chosen to be $(\sum_{j \neq i} w_{ij})^{0.5}$ or $w_{i.} = \sum_{j \neq i} w_{ij}$ if interest is to ensure that the variance of an area decreases as the number of neighbours increases (112). Here, we adopt the last criterion and use the adjacency matrix such that $w_{ij} = 1$ if site i and site j are neighbours and 0 otherwise.

Equation 4.7 can be written as

$$log(\mu_i) = log(E_i) + \alpha + X\beta + Z_u\theta_u + Z_v^*\theta_v^* \tag{4.9}$$

which in matrix notation is given as

$$log(\mu_i) = \{log(E_i)\, 1\, X\} \begin{pmatrix} 1 \\ \alpha \\ \beta \end{pmatrix} + (Z_u\ Z_v^*) \begin{pmatrix} \theta_u \\ \theta_v^* \end{pmatrix} \tag{4.10}$$

Z_u is the identity matrix and $Z_v^* = \{z_{ij}\}$ is the matrix of weights.

The variance covariance matrix of the random part is given as

$$\Sigma_\theta = var\left\{ \begin{pmatrix} \theta_u \\ \theta_v^* \end{pmatrix} \right\} = \begin{pmatrix} \sigma_u^2 I & \sigma_{uv} I \\ \sigma_{uv} I & \sigma_v^2 I \end{pmatrix} \tag{4.11}$$

Hence

$$var\left\{ \begin{pmatrix} u \\ v^* \end{pmatrix} \right\} = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix} \tag{4.12}$$

The variance of the relative risk conditional on the fixed parameters is then,

$$var(log(\mu)/X\beta) = Z\Sigma_\theta Z \tag{4.13}$$

Given the spatial model in equation 4.10,

$$var(log(\mu)/X\beta) = \sigma_u^2 Z_u Z_u^T + \sigma_{uv}(Z_u Z_v^{*T} + Z_v Z_u^T) + \sigma_v^2 Z_v Z_v^{*T} \tag{4.14}$$

4.2.1 Fully Bayesian estimation

In estimation of the spatial multilevel model we adopted the fully Bayesian approach. Under this method, the distribution of the hyper-prior is specified instead of assigning values to it as we did in the empirical Bayes method. This allows the variability of the hyperprior parameters among the sites to be taken care of. Since sites close to one another share common culture and behavioural practices that may affect the risk of HIV infection, it is expected that the prevalence of HIV infection may be similar among these sites. Hence, there may exist a local spatially structured variation in the prevalence rates. The nearest neighbour Markov random field (MRF) models or the conditional intrinsic Gaussian autoregressive (CIGAR) model are usually used to express this prior knowledge (182) (13). Under this prior model, the conditional distribution of the relative risk in site i, given the relative risks in all other areas $j \neq i$, depends only on the relative risks in the neighbouring areas of area i. The relative risks have a locally dependent prior probability structure, the variance of this spatial structure is often of interest.

The ordinary form of the conditional Gaussian autoregression (127) as used by Clayton and Kaldor (54), assumes that the conditional variance is constant, this is not very appropriate for irregular mapping where the number of neighbour varies. Intrisic Gaussian autoregression (130) is more suitable for irregular maps as the conditional variance of the log relative risk for site i given all other sites j is inversely proportional to the number of neighbouring areas of area i defined as $w_{i.}$ in equation 4.8. We adopted this later approach.

Hyperpriors for the Fixed and Random Effects

The fixed parameters may assume any value on the real line. Therefore a flat prior is assigned to them. This prior is the uniform prior defined within the open interval $(-\infty, \infty)$. For the random effect parameters, the multivariate Wishart prior is adopted. The inverse variance matrix is distributed as

$$\Sigma^{-1} \sim Wishart(2\hat{\Sigma}, 2)$$

which is a Wishart distribution with precision matrix $2\hat{\Sigma}$ and 2 degrees of freedom. The degree of freedom is the order of the matrix. Initial values of the precision matrix were assumed and the model was fitted using $WinBUGS14^{©}$.

Fitting the Spatial Models

Given that the observed number of HIV positive cases is Poisson with mean μ_i, that is

$$O_i \sim Poisson(\mu_i),$$

and that the generalized mixed effect model for μ is given as

$$log(\mu_i) = log(E_i) + \alpha + \sum_{k=1}^{8} x_{ki}\beta_k + u_{site(i)} + \sum_{j \in Neighbour(i)} w_{ij}v_j \qquad (4.15)$$

$$u^{(2)}_{site(i)} \sim N(0, \sigma_u^2), v_j \sim N(0, \sigma_v^2)$$

We have thus assumed that variation is due to differences among the sites and differences due to their neighbourhood patterning, hence it is possible to estimate

	Model E			Model F		
Parameters	Est.	s.e	95%CI	Est.	s.e	95%CI
Fixed Effects						
α	-0.128	0.095	(-0.31,0.062)	-22.33	10.93	(-45.27,-8.03)
Sexage				1.003	0.48	(0.37,2.01)
Polygamy				13.21	4.75	(6.02,23.10)
Freqsex				6.812	3.71	(1.73, 14.42)
CondomUse				-2.938	0.85	(-4.62 , -1.30)
Random Effects						
$\sigma^2_{u(\text{site})}$	0.220	0.06	(0.08,0.35)	0.11	0.063	(0.01, 0.23)
σ_{uv}	-0.137	0.05	(-0.26,-0.05)	-0.09	0.046	(-0.20, -0.02)
$\sigma^2_{v(\text{spatial})}$	0.135	0.18	(0.02,0.69)	0.25	0.34	(0.01, 1.12)

Table 4.7. *Estimates from the Spatial model*

these two sources of variation with a measure of similarity of the prevalence rates between sites which are geographically close to one another.

There is an appreciable increase in the parameter estimates and their standard errors when a combination of the fully Bayes and spatial model is used. Age at first sex, polygamy and frequent exposure to sex are positively associated with the risk of HIV infection. While condom use is found to be negatively associated with the risk of HIV risk.

The estimate for polygamy gave a relative risk of 1.1412, indicating that a one percent increase in polygamous practices in the zones has the potential of increasing the risk of HIV infection by at least 14.12% in the zones. Also, frequent exposure to heterosexual practices is associated with an increase in the risk of being infected. Age at first sex appear to have a significant positive association with the risk of contracting HIV. The good news is that regular use of condom in each sexual contact has the advantage of reducing the risk of HIV infection. The practice of polygamy as a form of marriage is more rampart in the North East zone, analysis suggest that a society that practices polygamy at the same level with the North East zone has 14.12% higher risk of contracting HIV than a society

Parameters	Lowest	Highest	Relative Risk
Sexage	Northwest	Southwest	1.010
Polygamy	Southeast	Northeast	1.14
Freqsex	Southwest	Northwest	1.07
CondomUse	Northcentral	Southwest	0.97

Table 4.8. *Relative Risks between areas of highest and lowest levels of the risk factors*

whose practice of polygamy is at the level of South East zone. A population where exposure to sex of its members is as frequent as that of the North west zone is 1.0705 times more at risk of HIV infection than a population whose frequency of exposure is the same with that of South west zone. Any society whose rate of condom use is at the same level as the South west has about 3.25% lower risk of HIV infection than a population whose rate of condom use is at the same level as North central or North east zone.

Worthy of note is the random part of the model. Spatial effects are highly significant in explaining variability in the risk of HIV infection. In Model F, the spatial effect dominates the heterogeneity effect since $\frac{\sigma_v^2}{\sigma_u^2} > 1$ (approximately 2). Also, there is a negative autocorrelation among the sites. This suggests that models that did not take this into account underestimated the variation in the parameters. This is evident when we compare the estimates of the standard error of the parameters in the spatial model with that of the previous model (see Table 4.5).

The plots of the estimates of the relative risks against the latitude are shown in Figure 4.3. We have used the latitude instead of the longitude for better positioning of the estimates along the horizontal axis, giving a picture of the trend observed in section 3.4 of chapter three. Two things can be noted from the plots; sites with relative risk greater than one could be identified, and the spatial clustering of the relative risks among close neighbours. Sites in the same state or

zone (represented with the same colour) share similar estimates of relative risk. The South West (SW) zone has generally very low risk of HIV infection. This is not surprising as condom use is highest and frequency of exposure to sex is lowest in the zone. Also, the age at first sex is highest in the south west zone. The sites in the North Central (NC) zone are hot spots of HIV. Many of them have relative risks greater than one. A site in Benue state has an estimated relative risk of 5.5. This means that individuals within this catchment area are about 6 times more at risk of HIV infection than their counterpart in the South west. Also, the contour plots and the map of estimates of relative risks in Figures 4.3, 4.2 and 4.4 respectively, give credence to this spatial structuring of relative risks of HIV infection.

The models considered in this section do not take into account other higher levels of classification. The site are nested in the States and the States are in the Zones. In the next section, we examine the random effects on the variation in the relative risks due to States and Zones.

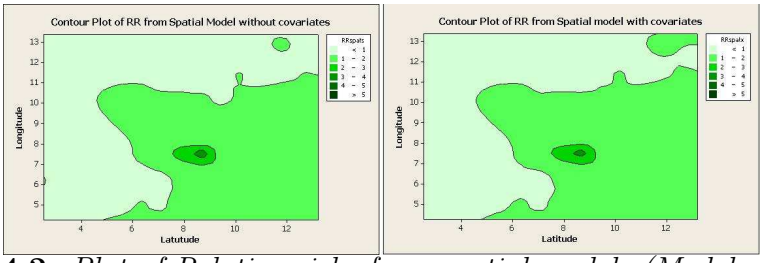

Figure 4.2. *Plot of Relative risks from spatial models (Models E and F)*

4.2.2 Incorporating Higher Levels into the Spatial model

It is possible to add higher hierarchical geographic levels to the spatial multi-level model considered above. The HIV sentinel data, as was earlier stated, can be described in three levels - site, state and zone. We have only considered the sites and the effects of their neighbours on variation in HIV prevalence rates. In

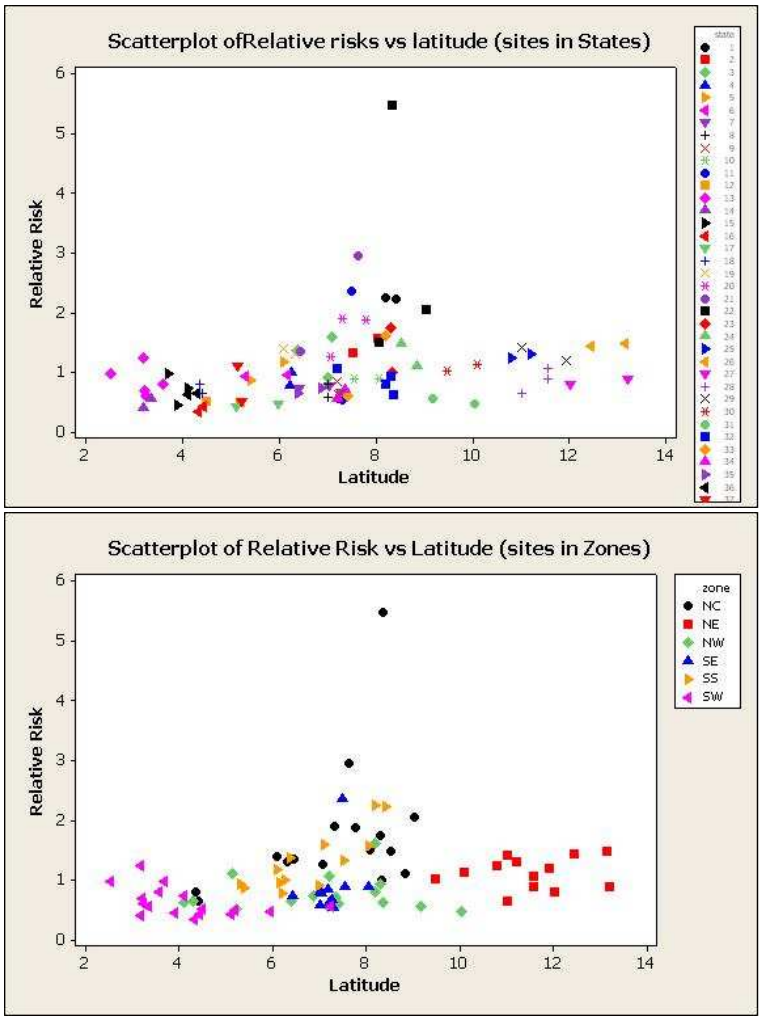

Figure 4.3. *Plot of Relative risks from spatial model against latitude - States and Zones identified*

this section, we introduce estimates of variation in prevalence rates at the hierarchy of the state and zone level into the model. Therefore, similarities of rates within the same state or the same zone will be accounted for. Hence an extension of model 4.15 is given as

$$log(\mu_i) = log(E_i) + \alpha + \sum_{k=1}^{8} x_{ki}\beta_k + u_{site(i)} + \sum_{j \in Neighbour(i)} w_{ij}v_j + s_{kl} + z_l \quad (4.16)$$

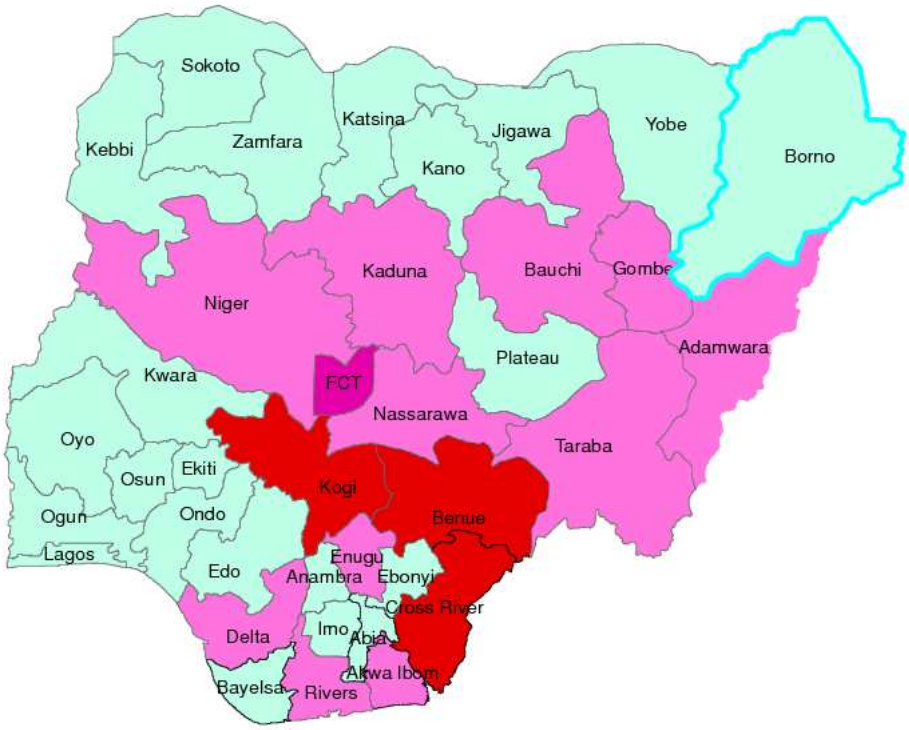

Figure 4.4. *Map of Relative risks from spatial models (Models F). Light green (0.4-1.0), Pink (1.0-1.5), dark pink (1.5-2.0), Red (2.0-3.0)*

$$u^{(2)}_{site(i)} \sim N(0, \sigma_u^2), v_j \sim N(0, \sigma_v^2),\ s_{kl} \sim N(0, \sigma_s^2),\ z_l \sim N(0, \sigma_z^2)$$

where the additional random effects are that of the state s_{kl} and zone z_l which are assumed to be distributed as normal with mean zero and variance σ_s^2 and σ_z^2 respectively.

We define the hyper-prior for the inverse variance terms Σ_s^{-1} and Σ_z^{-1} as Gamma distributed. And using the fully Bayes procedure the results in table 4.9 were obtained.

The model estimates are shown in Table 4.9. Note that literacy and risky sexual practices are not statistically significant. We adopted this model in preference to the model with only significant covariates based on a comparison of their DIC. The difference between the DICs is less than 2. The estimates of the

fixed parameters and their standard errors are similar to those obtained using the model excluding the higher hierarchies. However, the inclusion of the two higher levels has resulted to an increase in the estimate of variation due to spatial dependence by more than 140% (comparing models E and G). The spatial effect is made more distinct. Consequently, the spatial autocorrelation coefficient which was hitherto statistically significant has become insignificant. The spatial autocorrelation appear to have been completely explained by the spatial variation. It does seem from this result that some of the sites within the same zone and state are spatially dependent. The heterogeneity effect represented by σ^2_{site} declined by more than 42 per cent due to the incorporation of the higher levels into the model.

Comparing models G and H, the inclusion of the covariates reduced the total variation by about 39% and variation due to differences among the zones was reduced by almost 60% and that due to spatial effects by more than 42%. Therefore, it does mean that the covariates accounted for about 39% of the variation in the risk of HIV infection.

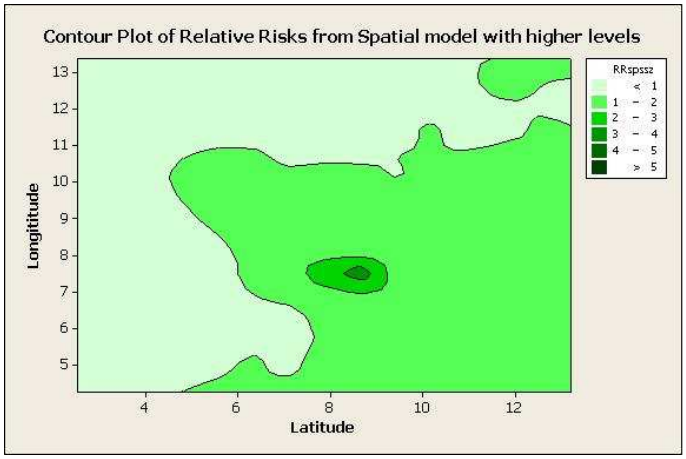

Figure 4.5. *Plot of Relative risks from spatial models incorporating higher levels*

Parameters	Model G Estimate	s.e	95%	Model H Estimate	s.e	95%
Fixed Effects						
α	-0.127	0.178	(-0.48,0.23)	-19.9	10.72	(-40.61, -3.82)
Sexage				0.914	0.496	(0.16, 1.87)
Riskysex				2.019	2.625	(-3.41, 6.93)
Polygamy				10.950	4.959	(2.22, 20.99)
literacy				-0.985	2.067	(-4.85, 3.28)
Freqsex				6.335	3.521	(0.66, 13.17)
Usecdm				-3.015	0.917	(-4.85, -1.23)
Random Effects						
$\sigma^2_{z(\text{zone})}$	0.142	0.237	(0.006, 0.61)	0.058	0.256	(0.0007, 0.37)
$\sigma^2_{s(\text{state})}$	0.013	0.0217	($5.7E{-}4$, 0.07)	0.017	0.030	($6.0E{-}4$, 0.079)
$\sigma^2_{u(\text{site})}$	0.127	0.083	(0.003, 0.27)	0.111	0.059	(0.005, 0.22)
σ_{uv}	-0.081	0.0536	(-0.197, 0.02)	-0.071	0.045	(-0.17, 0.006)
$\sigma^2_{v(\text{spatial})}$	0.323	0.424	(0.006, 1.35)	0.186	0.289	(0.004, 1.01)

Table 4.9. *Estimates from Spatial model incorporating higher levels*

4.3 Multiple Membership Multiple Classification Model

Parameter estimates obtained using the multiple membership multiple classification approach are fairly similar to that of the spatial model that incorporate estimates of autocorrelation which we considered in the previous section. The estimates of the standard errors are also quite similar indicating that the model adequately capture the variation in the parameters. However, they differ in the estimates of the random effects. Variation due to differences between sites or the heterogeneity effect is larger than the effect due to spatial dependence.

The interpretation of the estimates (shown in Table 4.10) is similar to those already considered. Age at first sex, polygamy and Frequency of exposure to heterosexual contact are positively associated with HIV prevalence rate while condom use is inversely associated with the prevalence of HIV infection. Highly

Parameters	Estimate	s.e	95%
Fixed Effects			
α	-26.52	12.37	(-52.750,-4.754)
Sexage	1.197	0.5462	(0.236,2.341)
Polygamy	14.57	5.319	(5.032,25.570)
Freqsex	8.306	4.268	(0.695, 17.240)
CondomUse	-3.457	1.02	(-5.475 , -1.493)
Random Effects			
$\sigma^2_{u(\text{site})}$	0.197	0.0496	(0.114, 0.307)
$\sigma^2_{v(\text{spatial})}$	0.1394	0.1608	(0.00107, 0.562)

Table 4.10. *Estimates from the MMMC model*

Parameters	Lowest	Highest	Relative Risk
Sexage	Northwest	Southwest	1.012
Polygamy	Southeast	Northeast	1.16
Freqsex	Southwest	Northwest	1.081
CondomUse	Northcentral	Southwest	0.966

Table 4.11. *Estimates of covariate effect (MMMC model)*

polygamous society like communities in the Northeast zone, has risk of HIV infection that is about 16% higher than communities that are mainly monogamous-like those in the southeast zone. Also, the more frequent a population is exposed to heterosexual intercourse, the more the risk of HIV infection. This risk is about 8.07% higher in such society than a society with less frequent exposure. Populations prone to regular use of condom as in the Southwest zone has a risk of HIV infection that is 3.4% lower than that of populations where condom use is as low as in the North central zone (see Table 4.11)

The plot of the estimates of the relative risks against the latitude is shown figure 4.6. The clustering of sites by zones is discernible. Most sites in the North central and South-south have significantly high risk of HIV infection. The spatial pattern is also evident in Figure 4.7

In this chapter we have investigated the contribution of the various hierarchical levels and some ecological factors to variations in the distribution of HIV

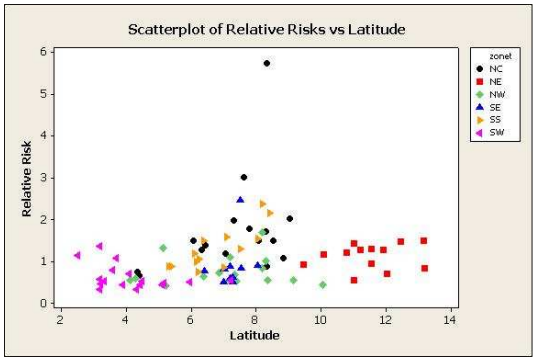

Figure 4.6. *Plot of Relative risks from Multiple Membership Multiple Classification(MMMC) model against latitude*

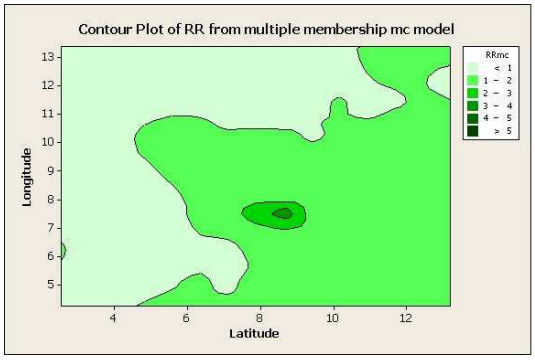

Figure 4.7. *Contour plot of Relative risks from Multiple Membership Multiple Classification model*

prevalence in Nigeria. The use of the variance component models indicated large variability among the sites followed by the zones and the least variation among the states. However, the variance component model ignores the fact that sites geographically close to one another might have similar prevalence rates due to common socio-cultural, religious and behavioural factor shared by individuals within the same neighbourhood which may influence the spread of HIV. To account for this possible clustering of prevalence rates, spatial models were applied to the data. Estimates from this model show a significant negative autocorrelation among the sites. Improved parameter estimates were obtained with slightly larger standard error estimates. Indicating that the variance component models underestimated variability associated with the parameters. Also, spatial effects

were significant.

The standard spatial model which estimate the heterogeneity and spatial effects and a measure of a measure of similarity of prevalence rates tend to ignore the fact that the sites are nested within other higher levels - states and zones. We extended the spatial model by incorporating the two higher hierarchies into the model. Estimates obtained from this model are an improvement on the standard spatial models. The spatial effect is also more prominent.

The multiple membership multiple classification model assumes that random variation in the prevalence of HIV can be explained only by the site heterogeneity effect and the neighbourhood patterning. It therefore neglects the effects that may be attributed to the hierarchies in which the lowest level might be nested.

4.4 Monitoring Convergence

To establish convergence when fitting, we used three different criteria; the history trace plots, Gelman-Rubin diagnostics (4) and the Monte Carlo error as a percentage of the posterior standard deviation. To achieve this, we ran two parallel chains using different starting values with the aim of obtaining an equilibrium distribution of the Markov chain (301). From this point of equilibrium, the joint distribution of the sample values is expected to converge to joint posterior distribution. Further iteration from this stationary point produces dependent sample assumed to have come from the posterior distribution. The period from the first iteration till convergence to the posterior distribution is called the *burn-in period*. This burn-in period is usually discarded and further iterations done in order to

obtain samples from the joint posterior distribution for posterior inference. Monitoring the convergence of every parameter in a multi-parameter model is not practical, therefore we need to make a decision on the relevant parameters to monitor.

Using the trace or time series plots to monitor convergence, the patterns produced by the parallel chains were observed until they overlap and remain so as the number of iterations increases. The stabilization of this overlap indicates convergence.

The chain trace plot of some fixed and random terms in the variance component model is shown in Figure 4.8. Two parallel chains (the red and the blue lines) were run simultaneously for 900,000 iterations from different starting points. For the fixed part, the beta parameters differ significantly in the convergence behaviour. While $\beta[5]$ reached convergence at an early iterative stage, $\beta[4]$ is yet to reach convergence even after 600,000 iterations. This problem of convergence of the parameters could be overcome by centering the parameters (142). Convergence in the random part of the model was easier to achieve than that of the fixed part. As can be see from the plots, the site, state and zone variances converged at the early stage of the iteration and remained stable to the end.

We also monitored the convergence of the iterative sampling using the Gelman-Rubin convergence test. The time series plot of the components of the test is shown in Figure 4.8. The green line is the width of the central 80% interval of the *pooled* runs. The blue line is the average width of the 80% *within* the individual runs and the red line is the ratio (R) of the green and the blue (ratio of the pooled and the within). If the starting values are suitably over-dispersed, R would generally be greater than 1 (142),(11) and is expected to decline to 1 as $n \rightarrow \infty$. Hence at convergence, $R \rightarrow 1$ and both the green and the blue

Parameters	sd.	90,000 iterations MCE	MCE % sd	sd.	175,000 iterations MCE	MCE % sd
Fixed Effects						
α	8.677	0.3547	4.09	10.93	0.3765	3.44
Sexage	0.383	.0157	4.09	0.4802	0.01655	3.45
Polygamy	3.822	0.1524	3.99	4.749	0.1606	3.38
Freqsex	2.977	0.1207	4.05	3.712	0.1272	3.43
CondomUse	0.817	0.027	3.30	0.846	.0.0227	2.68
Random Effects						
$\sigma_u^2(site)$	0.0606	0.00204	3.37	0.0627	0.00178	2.85
σ_{uv}	0.0449	0.00115	2.55	0.04606	0.00091	1.98
$\sigma_v^2(spatial)$	0.319	0.01178	3.69	0.3345	0.01022	3.06

Table 4.12. *Convergence test using Monte Carlo error*

line should consistently overlap and stabilize, possibly merging with the red line. Figure 4.9 shows that most of the parameters reached convergence at the 5,000 iterations. However, 10,000 iterations was used as the burn-in period for this model.

After convergence, we ran further iterations in order to improve the inference on the posterior estimates. The length of this further iteration is determined by monitoring the Monte Carlo error and the sample standard deviation. Using the rule of the thumb as suggested by Spiegelhalter et al.(142), the iteration is stopped when the Monte Carlo error of each parameter of interest is less than 5% of the sample standard deviation. An example is shown in Table 4.12 for the estimates of the spatial model at 90,000 and 175,000 iterations after a burn-in of 10,000. The longer the iteration, the better the convergence and hence better improved posterior inference.

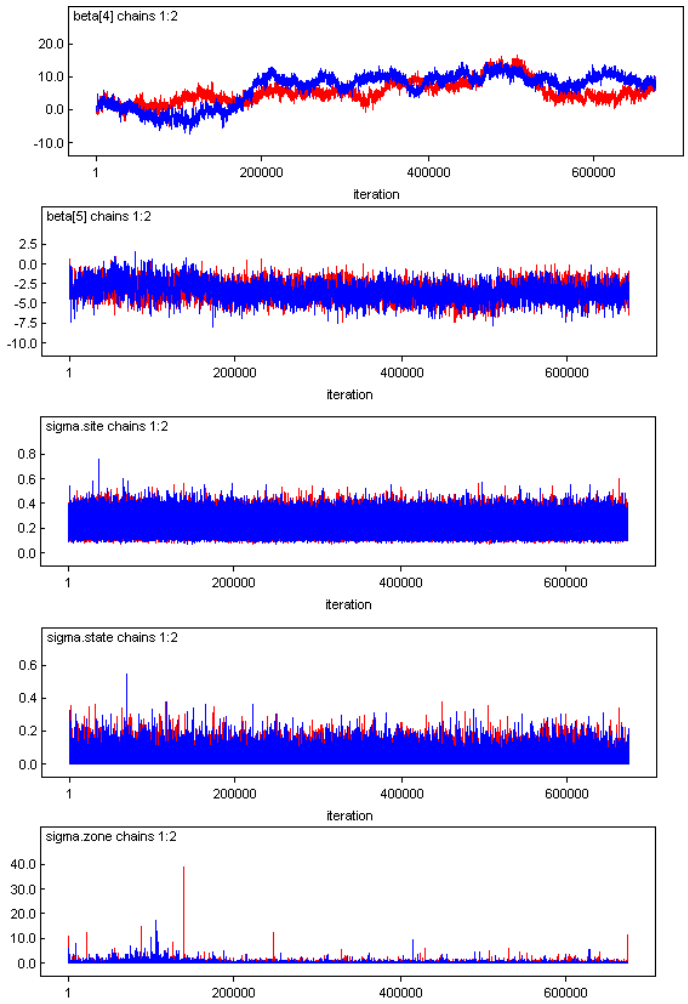

Figure 4.8. *Trace plots of some fixed and random terms in the variance component model. The red and blue lines represent two parallel chains*

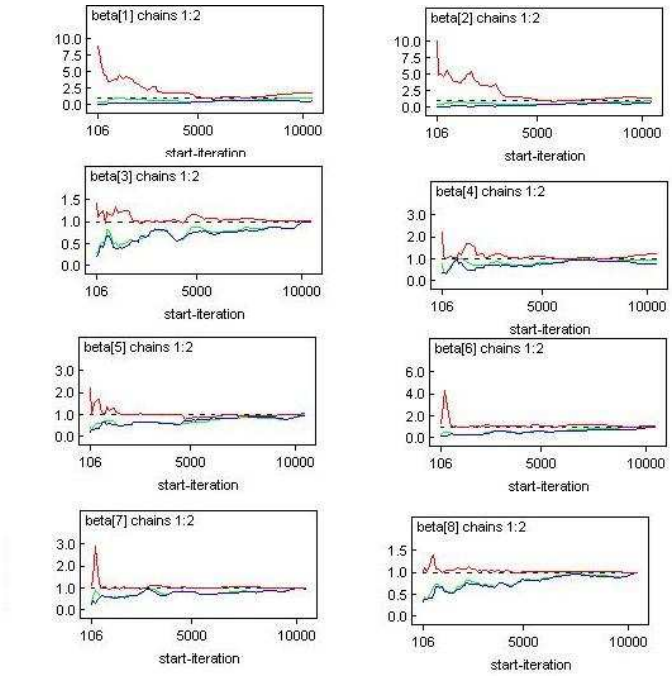

Figure 4.9. *Gelman-Rubin convergence plot of some fixed terms in the spatial model*

Chapter 5

Back-projection Models

5.1 Introduction

In this chapter, we shall consider two aspects of back-projection: Parametric and nonparametric back -projection methods. Generally, as discussed in chapter 2, the back-projection model is given as

$$\mu_t = \sum_{s=1}^{t} \lambda_s f_{t-s,s} \tag{5.1}$$

where μ_t is the mean AIDS incidence at time t, λ_s is the mean HIV incidence at time s and $f_{t-s,s}$ is the probability density function for someone infected at time s and diagnosed at time t. μ_t is known from the observed AIDS diagnosis. So assuming $f_{t-s,s}$ is known from other studies, λ_s is then estimated.

5.2 Parametric back-projection

We use the term parametric back-projection to represent all back-calculation approaches where a particular functional form is assumed for the HIV incidence curve or the AIDS incidence curve. In particular, we shall review the works of Brookmeyer and Gail (227)(229) and Rosenberg and Gail (257). We shall then reproduce the results of the later and apply the method to Nigeria AIDS data. We shall also seek to apply this approach to other countries where different methods were applied in order to compare the results.

5.3 Estimation when G is a basis of indicator functions

5.3.1 Application to American AIDS diagnosis data

We first reproduce the results of Rosenberg and Gail (257) for American AIDS diagnosis data adjusted for reporting delays. The incubation distribution is assumed to be Weibull and the parameters of this distribution were estimated by Brookmeyer and Goedert (231)) using data from the National Cancer Institute's multicentre haemophilia cohort. Parametric regression techniques were used to obtain these parameters. The cumulative distribution is given as

$$F(t) = \begin{cases} 1 - e^{(-0.0021t^{2.516})} & t > 0 \\ 0 & otherwise \end{cases} \tag{5.2}$$

with a median incubation period of 10 years.

The data is quarterly data and spans the period January 1977 and March 1988. Applying the step function model such that $g_i(s)$ is an indicator function, this period is partitioned into four intervals over which $g_i(s)$ is constant. The four intervals are; January 1977–January 1981, January 1981–January 1983, January 1983–January 1985, January 1985–April 1988. T_0 is regarded as 1st January 1977 and its value is 0 because it is assumed that there were no infections before that time. T_1 correspond to 1st January 1982.

We define $g_i(s)$ as a step function thus;

$$g_i(s) = \begin{cases} 1 & t_{i-1} < s \leq t_i \\ 0 & otherwise \end{cases} \tag{5.3}$$

For the four steps given above, $g_1(s) \in [0, 4), g_2(s) \in [4, 6), g_3(s) \in [6, 8)$& $g_4(s) \in (8, 11.25]$

Recall that x_{ji} is computed as

$$x_{ji} = \int_0^{T_j} g_i(s)\{F(T_j - s) - F(T_{j-1} - s)\}ds$$

X^* is the X matrix augmented by an additional row vector (X_{J+1}) containing information on the unknown infections that may be diagnosed after time J (J = 26 corresponding to first quarter of 1988). X_{J+1} is computed as

$$x_{J+1,i} = \int_0^{T_{J+1}} g_i(s)\{F(T_{J+1} - s) - F(T_J - s)\}ds$$

$$= \int_0^{\infty} g_i(s)\{F(\infty - s) - F(11.25 - s)\}ds$$

$$= \int_0^{\infty} g_i(s)\{1 - F(11.25 - s)\}ds$$

$$= \int_0^{\infty} g_i(s)\{e^{-0.0021(11.25-s)^{2.516}}\}ds$$

For instance,

$$x_{J+1,1} = \int_0^{4} e^{-0.0021(11.25-s)^{2.516}ds} = 2.2682$$

Hence

$$X_{J+1} = (\ 2.2682\ 1.61566\ 1.84169\ 3.21278)$$

when this is added as the 27th row of matrix X we obtain the augumented matrix X^*

$$\Delta_i = \sum_{j=1}^{J+1} x_{ji}$$

Hence,

$$\Delta = (\ 3.998861\ 2.000004\ 2.00004\ 3.250016)$$

Applying the methods of multinomial maximum likelihood, Poisson likelihood and the Quasi-likelihood we obtained the following estimates shown in Table 5.1. These exactly replicate the findings of Rosenberg and Gail (257) .

We note that the estimates of the number of AIDS diagnosis in each quarter were the same irrespective of the method of estimation used. That is, the multinomial likelihood, Quasi-likelihood and the Poisson likelihood gave the same quarterly estimate of the AIDS incidence as shown in Table 5.1. Thus, each of the methods gave a residual variation (measured by the χ^2 in the last column of Table 5.1) of 131.88.

The parameter estimates obtained using the three methods differ slightly. The Quasi-likelihood and the Poisson likelihood gave estimates slightly different from that obtained by the multinomial likelihood maximization as shown in Table 5.2.

The estimates of the number of individuals ($\hat{N}$) previously infected with HIV that gave rise to the observed number of AIDS cases are shown in Table 5.3. The Multinomial estimate is slightly higher than that of Poisson and Quasi-likelihood approach. In all, the models estimated a rising epidemic between 1977 and 1988.

A visual impression of the relationship between the observed and expected AIDS count is shown in the graph in Figure 5.1. The model-based estimates are in very good agreement with the observed data.

Quarter	*Observed (y)*	*Expected*($\hat{\mu}$)	$\frac{y-\hat{\mu}}{\sqrt{\hat{\mu}}}$	$\frac{(y-\hat{\mu})^2}{\hat{\mu}}$
1977:1-1981:4	374	413.49	-1.94	3.77
1982:1	185	141.03	3.70	13.71
1982:2	200	199.23	0.05	0.003
1982:3	293	273.7	1.17	1.36
1982:4	374	365.86	0.43	0.18
1983:1	554	476.89	3.53	12.47
1983:2	713	608.56	4.23	17.92
1983:3	763	762.33	0.02	0.0006
1983:4	857	936.66	-2.6	6.77
1984:1	1147	1141.42	0.17	0.03
1984:2	1369	1368.25	0.02	0.0004
1984:3	1563	1621.18	-1.4	2.1
1984:4	1726	1900.45	-4.0	16.0
1985:1	2142	2206.96	-1.4	1.9
1985:2	2525	2544	-0.4	0.14
1985:3	2951	2914.64	0.67	0.45
1985:4	3160	3321.37	-2.8	7.8
1986:1	3819	3765.48	0.9	0.8
1986:2	4321	4247.61	1.12	1.26
1986:3	4863	4769.38	1.36	1.83
1986:4	5192	5330.24	-1.9	3.59
1987:1	6155	5930.96	2.9	8.46
1987:2	6816	6571.42	3.0	9.1
1987:3	7491	7251.05	2.82	7.94
1987:4	7726	7969.68	-2.73	7.45
1988:1	8483	8726.59	-2.61	6.8
TOTAL	75762	75758.43	4.31	131.88

Table 5.1. *Observed and expected AIDS counts for USA HIV/AIDS epidemic. Source: Rosenberg and Gail (1991)*

Time(Step)	Multinomial	QL	Poison regression
Jan 1977–Jan 1981	2104.82	2105.61	2105.61
Jan 1981–Jan 1983	111579.89	111557.88	111557.88
Jan 1983–Jan 1985	131923.96	131982.41	131982.41
Jan 1985–April 1988	224202.42	224159.53	224159.53

Table 5.2. *Parameter estimates*($\hat{\beta}$) *for American HIV/AIDS epidemic*

Time (Step)	Multinomial	QL/Poision
Jan 1977–Jan 1981	8419	8422
Jan 1981–Jan 1983	223160	223116
Jan 1983–Jan 1985	263848	263965
Jan 1985–April1988	728658	728518
Jan 1977–April 1988	1224083	1224021

Table 5.3. *Estimates of the Numbers previously infected with HIV*

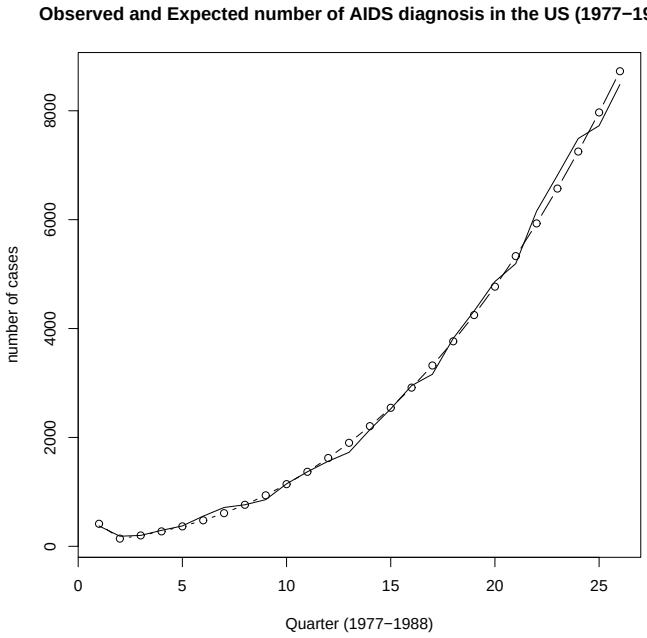

Figure 5.1. *Observed and Expected number of AIDS diagnosis in the US (1977-88). Infection curves are assumed to be step functions*

5.3.2 Application to the Nigerian AIDS data

The procedure as defined above was applied to the Nigerian AIDS data as published by the Nigerian Institute of Medical Research. The data is the number of AIDS diagnosis between January 1989 and December 1999. To determine the number of steps in a step function model, we considered the fact that the first cases of AIDS were reported in the country in 1986. It can be argued that the HIV/AIDS epidemic started in the country sometime before 1986 since it takes about 6 to 10 years before AIDS is diagnosed after HIV infection. In order to cover this period, we extended the years for our estimation to 1980 with assumption that the AIDS cases diagnosed in 1989 might have been infected with HIV within the last 10 years.

The decision on the number steps was informed by the behaviour of the epidemic as depicted by the data. Based on this, three steps were identified. The steps are 1st January 1980 – 1st January 1989, 1st January 1989 – 1st January 1993, 1st January 1993 – 31st December 1999. Hence $T_0 = 1980$ but data as published started from 1st January 1989, therefore, $T_1 = 1989$.

However, we noticed a slight curve in the graph of the AIDS incidence data between 1st January 1993 and 1st January 1995. We decided to view this period as a separate step. Hence, we now have four steps instead of three by partitioning the last step in the paragraph above accordingly. We shall compare results from the two step function models.

Using the approach described above, we obtained the design matrix for the 3-step function as shown in Table 5.4:

To obtain the augmented matrix X^*, we compute the X_{J+1} row vector given as

G_1	G_2	G_3
1.384214	0	0
0.497843	0.0006	0
0.577008	0.0062119	0
0.642123	0.021344	0
0.6875	0.048472	0
0.71015	0.088228	0.0006
0.709168	0.136257	0.0062119
0.685644	0.186993	0.021344
0.642409	0.235845	0.048472
0.583618	0.278629	0.088825
0.514226	0.31177	0.142469

Table 5.4. *Design matrix X for the Nigerian AIDS incidence curve obtained by assuming three steps for the infection curve*

$$x_{J+1,i} = \int_0^{T_{J+1}} g_i(s)\{F(T_{J+1} - s) - F(T_J - s)\}ds$$

Hence, $X_{J+1} = (1.867\ 2.686\ 5.692)$

Adding X_{J+1} as the $12th$ row of the matrix in Table 5.4, we have the augmented matrix X^*. The sum of the columns gives an estimate of time Δ_i in each step as given below.

$\Delta = G_i(T_J) = (10\ 3.9809\ 5.999948)$

And using the multinomial, Quasi-likelihood and the Poisson likelihood estimation methods, we obtained the same estimates of AIDS incidence as shown in Table 5.5.

The methods estimated precisely the cumulative number of individuals diagnosed with AIDS between January 1980 and December 1999. As shown in Table 5.5, a total of 12316 cases of AIDS were diagnosed within the period and

Year	*Observed (y)*	*Expected*($\hat{\mu}$)	$\frac{y-\hat{\mu}}{\sqrt{\hat{\mu}}}$	$\frac{(y-\hat{\mu})^2}{\hat{\mu}}$
1980-1989	8	49	-5.8	34.0
1990	14	20	-1.3	1.7
1991	160	45	17.3	298.6
1992	102	106	-0.4	0.16
1993	198	216	-1.08	1.16
1994	327	383	-2.9	8.3
1995	625	697	-2.7	7.47
1996	1381	1234	4.18	17.47
1997	1940	2032	-2.04	4.18
1998	3219	3103	2.09	4.37
1999	4342	4444	-1.38	1.89
Total	12316	12316	5.9	379.34

Table 5.5. *Observed and estimated AIDS cases for Nigeria*

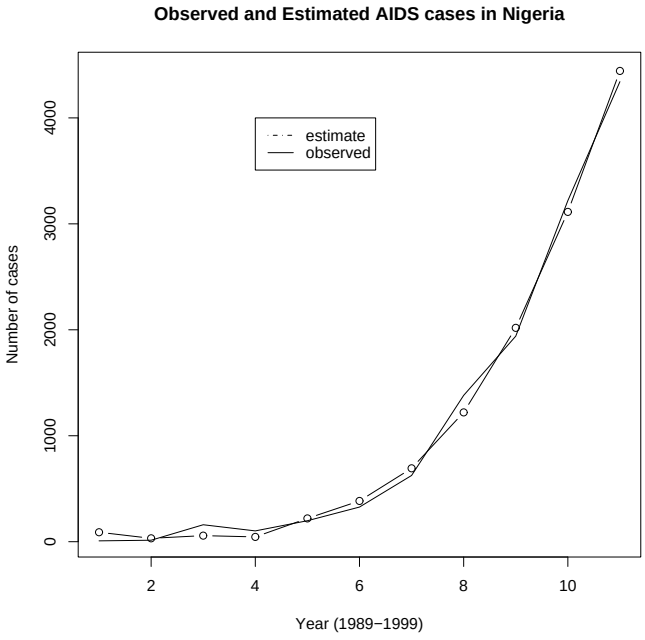

Figure 5.2. *Observed and estimated AIDS cases for Nigeria obtained using three steps for the infection curve*

Time(Step)	Multinomial	QL	Poison regression
Jan 1980–Jan 1989	35.1929(3.51)	35.1919(4.07)	35.1919(4.07)
Jan 1989–Jan 1993	3910.48(123.33)	3910.56(122.65)	3910.56(122.65)
Jan 1993–Dec 1999	22431.35(517.48)	22430.94(515.44)	22430.94(515.44)

Table 5.6. *Parameter estimates($\hat{\beta}$) and their (standard error) for the Nigerian HIV/AIDS epidemic obtained from the three steps model*

the estimate predicted this accurately. However, when the annual estimates are compared with the observed values for each year, some gaps are noticed. A measure of this deviation of the estimates from the observed counts is shown in the last column of Table 5.5. The total residual variation is large at 379.34. One explanation for the large residual variation, when compared with that obtained using the US data, may be the structure of the Nigeria data. A closer scrutiny of the residual variations indicates that 1991 contributed about 79 percent of the total variation (298.6 out of 379.34). The models were unable to match the sharp increase in the number of AIDS cases diagnosed in 1991 which increased from just 14 cases in 1990 to 160 cases in 1991. Several attempts were made to obtain better estimate of the number of AIDS diagnosis for 1991 by adjusting the positions of the steps but all to no avail. It seems to suggest that a steeper curve should be assumed for the infection intensity in order to capture the steep rise in 1991.

As shown in Table 5.6 infection intensity shows an increasing trend from one step (time interval) to the other. The three methods gave equivalent parameter estimates. The standard errors indicate a greater uncertainty in estimating the HIV positive population in the more recent years.

Estimates of the number of persons previously infected with HIV, part of which were later diagnosed with AIDS, is shown in the Table 5.7. The table

Time (Step)	Multinomial	QL/Poision
Jan 1980–Jan 1989	352	352
Jan 1989–Jan 1993	15017	15018
Jan 1993–Dec 1999	134588	134586
Total	150582	150580

Table 5.7. *Estimates of the Numbers previously infected with HIV in Nigeria obtained using the three steps model*

gives an idea of what the population of people living with HIV was in Nigeria in these time intervals. It suggests that, as at December 1999, 150,582 persons were infected with the virus.

Data collected from the 1057 health and laboratory facilities show that about 63,387 HIV positive cases and 12,316 AIDS cases were diagnosed in Nigeria between January 1989 and December 1999. This means that a total of 75,703 HIV/AIDS cases were observed. The difference between this total and the estimated cumulative HIV infection gives an idea of how large the hidden or undiagnosed cases could have been. It does seem from the estimate that about half of the HIV population were not diagnosed.

The Four Steps Model

We also examined what the estimates would be if the number of steps were extended to four. Doing this, we obtained the following:

$$X_{J+1} = (1.867\ 2.686\ 1.769\ 3.923) \text{ and } \Delta = (9.500903\ 4.000347\ 2.000294\ 3.999625)$$

Comparing Tables 5.6 and 5.9, the parameter estimates obtained from the four-step model are not very different from the those obtained using three steps. However, the standard error for each parameter estimate in the four-step model is higher than that of the estimate in its corresponding position in the three-step model. Consequently, it appears that the addition of the fourth step has not

G_1	G_2	G_3	G_4
1.384214	0	0	0
0.497843	0.0006	0	0
0.577008	0.0062119	0	0
0.642123	0.021344	0	0
0.6875	0.048472	0	0
0.71015	0.088228	0.0006	0
0.709168	0.136257	0.0062119	0
0.685644	0.186993	0.020747	0.0005969
0.642409	0.235845	0.04226	0.006211936
0.583618	0.278629	0.067481	0.021344
0.514226	0.31177	0.093997	0.048472

Table 5.8. *Design matrix obtained using four steps*

Time(Step)	Multinomial	QL	Poison regression
Jan 1980–Jan 1989	37.96(4.23)	37.9679(4.24)	37.9679(4.24)
Jan 1989–Jan 1993	3695.20(151.08)	3694.898(152.53)	3694.898(152.92)
Jan 1993 - Jan 1995	25623.13(1561.71)	25628.48(1621.51)	25628.48(1623.27)
Jan 1995–Dec 1999	16221.65(2980.64)	16209.74(3147.72)	16209.74(3148.08)

Table 5.9. *Parameter estimates* $(\hat{\beta})$ *and their standard error (four-step model)*

significantly improved the estimates. Table 5.10 shows the estimates of the AIDS incidence and their residual variation.

There is also no significant improvement in the standard error of the estimates. 1991 still contributed the most standard error. Comparing the total residual variations obtained in the two models as measured by the chi square, the slight decrease in the four-step model seem not enough to warrant the additional step.

Tables 5.3, 5.7 and 5.11 seem to suggest that the multinomial maximum likelihood estimate consistently have higher estimate for the most recent infection than the Poisson and Quasi-likelihood estimates.

Year	*Observed (y)*	*Expected*($\hat{\mu}$)	$\frac{y-\hat{\mu}}{\sqrt{\hat{\mu}}}$	$\frac{(y-\hat{\mu})^2}{\hat{\mu}}$
1980-1989	8	53	-6.15	37.78
1990	14	21	-1.55	2.39
1991	160	45	17.19	295.52
1992	102	103	-0.12	0.015
1993	198	205	-0.5	0.25
1994	327	368	-2.15	4.62
1995	625	690	2.46	6.05
1996	1381	1258	3.46	11.96
1997	1940	2080	-3.06	9.37
1998	3219	3127	1.64	2.7
1999	4342	4366	-0.37	0.13
Total	12316	12315.99	5.938	370.8

Table 5.10. *Observed and estimated AIDS cases in Nigeria and their (standard errors) obtained using four-step model*

Time (Step)	Multinomial	QL/Poision
Jan 1980–Jan 1989	380	380
Jan 1989–Jan 1993	14781	14780
Jan 1993–Jan 1995	51246	51257
Jan 1995- Dec 1999	64887	64839
Total	131293	131256

Table 5.11. *Estimates of the Numbers previously infected with HIV obtained using four-step model*

5.4 Estimation when G is a spline function

Here the basis set is assumed to be a spline. According to Rosenberg and Gail (257), the flexibility of the step function can greatly be enhanced by using a spline function $g_i(s)$ with knots at $t_l, l = 1, 2, \ldots, L$ as a basis for G with the requirements that $\nu(s)$ be continuous at the knots or that $\nu(s)$ and its derivative $\nu'(s)$ be continuous. Using the '+' function notation, they defined $\nu(s)$ as

$$\nu(s) = \sum_{j=0}^{n} \beta_{0j} s^j + \sum_{l=1}^{L} \beta_{ln}(s - t_l)_+^n. \tag{5.4}$$

Applying this to the US AIDS data, Rosenberg and Gail (257) assumed a single knot in January 1982 and letting $n = 2$, then

$$\nu(s) = \beta_{00} + \beta_{01} + \beta_{02} s + \beta_{12}(s - t_1)_+^2 \tag{5.5}$$

Hence, $g_1(s) = 1, g_2(s) = s, g_3(s) = s^2$ and

$$g_4(s) = \begin{cases} (s - t_l) & t_l \geq T_L \\ 0 & otherwise \end{cases}$$

.

Hence,

$$x_{ji} = \int_0^{T_j} g_i(s)\{F(T_j - s) - F(T_{j-1} - s)\}ds, \ i = 1, 2, 3 \tag{5.6}$$

integrating within the intervals of $g_i(s)$ and

$$x_{j4} = \int_{t_l}^{T_j} g_4(s)\{F(T_j - s) - F(T_{j-1} - s)\}ds \tag{5.7}$$

The estimate of the HIV population in the time interval i is given as

$$N_i = \int_{g_i(s)} \nu(s)ds$$

and the estimate of the HIV population through the years in which AIDS diagnosis data is available is

$$N = \int_0^{\tau_J} \nu(s)ds$$

5.4.1 Application to American AIDS diagnosis data

Using equations 5.6 and 5.7 we obtained the design matrix X for the US data and using the quasi-likelihood method, we obtained the estimates shown in Table 5.12.

This result is exactly the same as obtained by Rosenberg and Gail (257). There is an improvement in the estimates as a result of the assumed spline infection intensity curve. The residual variance as obtained here is lower than that obtained when step function was assumed for the infection curve.

The estimates of the number of persons previously infected with HIV in the time intervals is given in the Table 5.13.

This estimate of the HIV population obtained using the quadratic spline is

Quarter	*Observed (y)*	*Expected*($\hat{\mu}$)	$\frac{y-\hat{\mu}}{\sqrt{\hat{\mu}}}$	$\frac{(y-\hat{\mu})^2}{\hat{\mu}}$
1977:1-1981:4	374	376.21	-0.11	0.01
1982:1	185	171.00	1.07	1.46
1982:2	200	228.00	-1.9	3.63
1982:3	293	300.00	-0.42	0.18
1982:4	374	387.47	-0.68	0.47
1983:1	554	492.12	2.79	7.78
1983:2	713	616.27	3.9	15.18
1983:3	763	761.86	0.04	0.002
1983:4	857	929.08	-2.36	5.59
1984:1	1147	1124.95	0.66	0.43
1984:2	1369	1346.09	0.62	0.39
1984:3	1563	1595.88	-0.82	0.68
1984:4	1726	1875.93	-3.46	11.98
1985:1	2142	2187.66	-0.98	0.95
1985:2	2525	2532.4	-0.15	0.02
1985:3	2951	2911.31	0.74	0.54
1985:4	3160	3325.33	-2.87	8.22
1986:1	3819	3776.14	0.70	0.49
1986:2	4321	4261.94	0.90	0.82
1986:3	4863	4785.49	1.12	1.26
1986:4	5192	5346.23	-2.11	4.45
1987:1	6155	5944.14	2.73	7.48
1987:2	6816	6578.96	2.92	8.54
1987:3	7491	7250.21	2.83	8.0
1987:4	7726	7957.25	-2.59	6.72
1988:1	8483	8698.93	-2.32	5.36
TOTAL	75762	75762.04	-0.04	100.32

Table 5.12. *Estimates of AIDS cases in the USA obtained by assuming a spline infection intensity and using the Quasi likelihood method*

Time	Spline estimates
Jan 1977–Jan 1981	41267
Jan 1981–Jan 1983	172867
Jan 1983–Jan 1985	316045
Jan 1985–April 1988	652311
Jan 1977–April 1988	1182490

Table 5.13. *number of persons previously infected with HIV in the US*

slightly less than that obtained when the step function was used. The step function estimated that there 1,224,083 (*multinomial*) or 1,224,021 (*quasi-likelihood and Poison regression*) HIV positive individuals in the United States between January 1977 and April 1988. However, it is noted from the estimates that the quadratic spline and the step function gave the same pattern of infection intensity within the time intervals considered above.

5.5 Application to Nigerian data

Applying the same technique to the Nigeria AIDS data, we applied various forms of spline function. We obtained estimates using the linear spline, quadratic spline, cubic spline and the natural cubic spline. Specifically, we considered the following splines where $t_1, t_2, \ and \ t_3$ are knot positions:

The linear splines

$$\nu(s) = \beta_{00} + \beta_{01}s$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{11}(s - t_1)_+$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{11}(s - t_1)_+ + \beta_{12}(s - t_2)_+$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{11}(s - t_1)_+ + \beta_{12}(s - t_2)_+ + \beta_{13}(s - t_3)_+$$

The quadratic splines

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2 + \beta_{11}(s - t_1)_+^2$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2 + \beta_{11}(s - t_1)_+^2 + \beta_{12}(s - t_2)_+^2$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2 + \beta_{11}(s - t_1)_+^2 + \beta_{12}(s - t_2)_+^2 + \beta_{13}(s - t_3)_+^2$$

The cubic splines

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2 + \beta_{03}s^3$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2 + \beta_{03}s^3 + \beta_{1l}(s - t_l)_+^3$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{02}s^2 + \beta_{03}s^3 + \beta_{11}(s - t_1)_+^3 + \beta_{12}(s - t_2)_+^3$$

The natural cubic splines

$$\nu(s) = \beta_{00} + \beta_{01}s$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{11}(s - t_1)_+^3$$

$$\nu(s) = \beta_{00} + \beta_{01}s + \beta_{11}(s - t_1)_+^3 + \beta_{12}(s - t_2)_+^3$$

.

The selection of the positions of the knots was informed by the pattern of the AIDS epidemic in Nigeria as depicted by the data used in this analysis. We selected 1989, 1992 and 1995 for the first, second and third knot respectively.

We progressively increased the number of knots in each category of spline in order to investigate the effects of the number of knots on the estimates. The determination of the optimum number of knots was based on the precision of the estimates as measured by the residual variance and ability of the estimates to produce a positive and feasible value of the HIV population. We note that our estimates agree with the findings of Stone (143) that the number of knots has some effect on the estimates. However, we used single knot in different positions and found that the estimates obtained differ. Analysis suggest that the precision of a single knot spline depends on the order of the spline and the distance of the knot from the origin of the data.

The further away the single knot is from the origin, the less precise the back-calculation method is in estimating the HIV population. However, the cubic spline appear to give better estimates of the AIDS diagnosis as the single knot is moved further away from the origin but the estimate of the HIV population

β	Number of knots					
	1(1989)	1(1992)	1(1995)	2(1989,1992)	2(1989,1995)	3(89,92,95)
	(a)	(b)	(c)	(d)	(e)	(f)
β_0	-105.85	-321.06	-564.0	-171.52	-129.04	-209.64
β_1	54.90	155.36	270.37	85.09	65.37	103.04
β_2	2771.11	9432.48	65016.81	1636.05	2479.17	719.54
β_3				4255.28	12116.68	10680.77
β_4						-33469.25
$\hat{N}$	146454	253038	558221	198156	230239	44514

Table 5.14. *Parameter Estimates from the linear spline. Models (a), (b) and (c) are single-knot splines with positions at 1989, 1992 and 1995. Models (d) and (e) are two-knot splines with positions at the years indicated and model (f) is a three-knot spline positioned in the years 89, 92 and 95 respective*

progressively decreased as the distance from the origin increases. We assessed the precision using the χ^2 given as $\sum \frac{(y-\hat{\mu})^2}{\hat{\mu}}$ where y is the observed value and $\hat{\mu}$ is the expected value obtained from the spline models.

The linear spline

The simple linear spline (with no knots) was unable to converge to a specific solution. In fact, different number of iterations gave different estimates and even numbers of iteration yielded negative estimates. The shifting of single knot to different points improved the estimates with that nearest to the origin having the least residual variance (see models (a), (b), and (c) in Tables 5.14, 5.15 and 5.16). Estimates obtained using two knots (models (d) and (e))were better than those obtained using single knot judging by the residual variance. It seems that linear splines with three or more knots over-fitted the model. While their residual variances are smaller than that of two knots spline, their estimates of the HIV population are unrealistic.

Observed AIDS cases	Number of knots 1(1989) (a)	1(1992) (b)	1(1995) (c)	2(1989,1992) (d)	2(1989,1995) (e)	3(89,92,95) (f)
8	11	5.97	4.25	8.32	9.58	7.8
14	28.48	69.78	119.05	40.40	32.44	47.87
160	54.26	119.26	204.71	72.38	60.62	82.96
102	113.11	183.13	315.43	128.77	119.88	134.85
198	235.18	262.80	451.49	227.97	237.24	212.87
327	455.49	381.25	612.06	401.11	443.97	352.43
625	810.69	608.04	795.32	701.30	772.73	642.38
1381	1335.84	1044.24	1007.30	1195.0	1256.37	1202.67
1940	2061.51	1807.99	1405.8	1953.77	1952.0	2093.21
3219	3011.17	3023.45	2477.53	3047.66	2971.74	3029.53
4342	4199.27	4810.09	4923.05	4539.30	4459.43	4329.43
12316	12316	12316	12316	12316	12316	12316

Table 5.15. *Estimate of the diagnosed AIDS cases obtained using the linear splines*

It is worthy to note the behaviour of the parameter estimates as we shift the single knot some distance away from the origin. The large variations in the parameter estimates due to these shifts are also reflected in the $\hat{N}$ since they are estimated directly from these estimates. Even when two knots were used in the model, the effect of the distance between the two knots is also felt both on the parameter estimates and the estimate of the number of persons living with HIV/AIDS.

It appears from the Table 5.16 and the graphs in Figure 5.3 that the linear spline with three knots is a better model for predicting number of AIDS cases diagnosed each year. However, our interest is not only on the ability of the model to predict precisely the AIDS cases but also its ability to predict a feasible HIV population – that is, the number of persons previously infected with HIV some of which were later diagnosed as AIDS cases. Table 5.16 gives a summarized

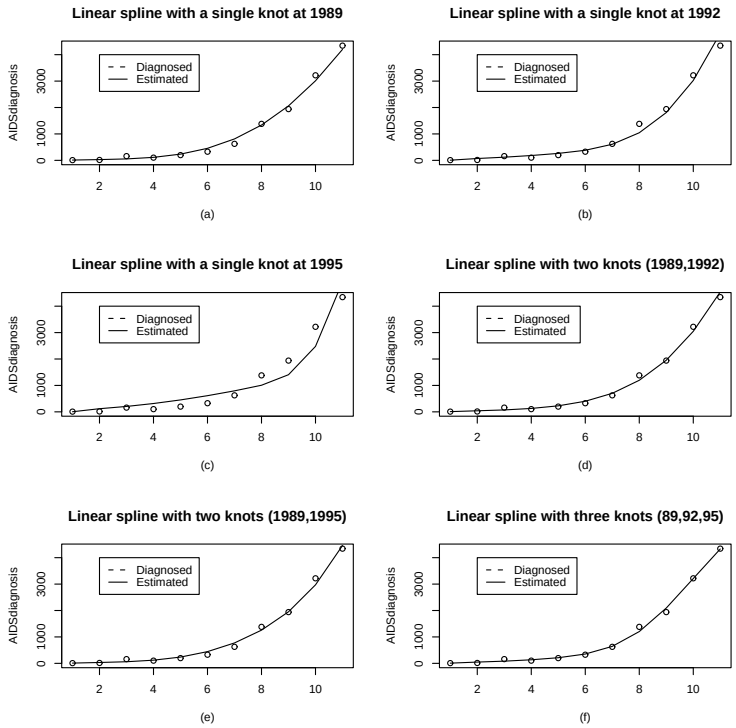

Figure 5.3. *Plot of observed (*dots*) and Backcalculated AIDS estimates (*line*) when infection curve is linear spline function*

information of the residual variance and estimate of the cumulative number of HIV/AIDS infection.

A total of 75,703 cases of HIV and AIDS were recorded in the 1057 health and laboratory facilities surveyed in the country between 1989 and 1999. Therefore, we expect our model to give estimates of number of cases of HIV and AIDS ($\hat{N}$) greater than this observed cases. Using this criterion, we eliminate the model with three knots and the model with the least residual variation may be selected. Thus, model (d) is selected.

	Number of knots					
	1(1989) (a)	1(1992) (b)	1(1995) (c)	2(1989,1992) (d)	2(1989,1995) (e)	3(89,92,95) (f)
	0.82	0.69	3.44	0.01	0.26	0.005
	7.36	44.59	92.70	17.25	10.48	23.96
	206.16	13.92	9.76	106.03	162.92	71.54
	1.09	35.94	144.41	5.57	2.67	8.0
	5.88	15.97	142.31	3.94	6.49	1.04
	36.24	7.72	132.76	13.69	30.82	1.83
	42.53	0.47	36.47	8.30	28.24	0.47
	1.53	108.61	138.64	28.95	12.36	26.44
	7.16	9.64	202.99	0.09	0.07	11.21
	14.34	12.65	221.91	9.63	20.57	0.03
	4.85	45.55	68.58	8.58	3.09	0.04
χ^2	327.87	295.75	1193.98	202.05	277.98	144.58
$\hat{N}$	146454	253038	558220	198157	230239	44514

Table 5.16. *Estimates of residual variance and total number of persons infected with HIV (linear spline)*

5.5.1 The quadratic spline

The ordinary quadratic model (without knots) could not converge to a feasible solution. Negative estimates were obtained in very large number of iterations. The graphs in Figure 5.4 show the behavior of the quadratic spline model for our data as the position of the single knot is varied. The estimates become less precise as the distance between the origin and the knot increases. Hence graph (a) appear better than (b) and (c). The graph number (d) and (e) are quadratic splines with two and three knots respectively and they seem a better fit than (a).

The repositioning of the single knot changes the value of the beta estimates drastically and consequently, the estimate of the cumulative number of those infected with HIV, $\hat{N}$,varied with the changing beta values. Table 5.17 indicate that the farther away the single knot is from the origin, the estimated number of persons living with HIV/AIDS increases. The results suggest that there is no need for models with more than two knots if we base our judgement on the value

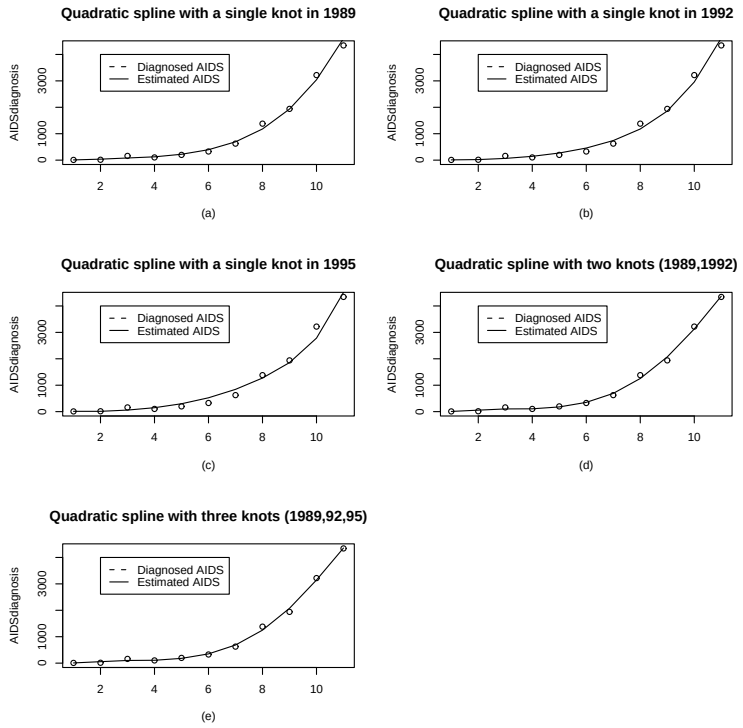

Figure 5.4. Plot of observed (*dots*) and Backcalculated AIDS estimates (*line*) when infection curve is a quadratic spline function

β	Number of knots				
	1(1989) (a)	1(1992) (b)	1(1995) (c)	2(1989,1992) (d)	3(89,92,95) (e)
β_0	-53.87	424.27	650.44	-588.25	-561.69
β_1	-17.67	-423.03	-607.32	435.38	412.29
β_2	13.86	66.85	89.90	-45.26	-42.15
β_3	579.72	1610.09	14604.10	1333.73	1273.09
β_4				-2533.59	-2144.17
β_5					-2626.35
$\hat{N}$	220713	268625	419823	118827	90562

Table 5.17. *Parameter estimates obtained using the quadratic spline*

Observed AIDS cases	Number of knots 1(1989) (a)	1(1992) (b)	1(1995) (c)	2(1989,1992) (d)	3(89,92,95) (e)
8	7.95	9.03	11.22	7.58	7.58
14	37.19	21.16	11.42	55.73	55.07
160	81.28	65.25	58.51	102.43	101.41
102	124.79	143.92	148.55	106.73	108.14
198	221.66	269.64	298.15	178.35	180.20
327	394.58	457.77	523.77	350.42	349.93
625	692.5	740.24	842.68	690.67	685.45
1381	1179.30	1178.38	1271.99	1254.74	1248.31
1940	1932	1857.04	1855.94	2070.88	2075.29
3219	3041.45	2956.76	2782.64	3128.45	3147.02
4342	4602.68	4616.83	4511.14	4370.02	4357.60
12316	12316				

Table 5.18. *Estimates of the diagnosed AIDS cases from the quadratic spline model*

of $\hat{N}$ alone. Tables 5.18 and 5.19 compares the closeness of each model to the observed data.

The estimates (a), (d) and (e) in the Table 5.18, are closer to the observed values than estimates (b) and (c). To choose the best model of the three, we need to consider the residual variances and parsimony.

Going by the principle of parsimony, model (d) is better than (e) since the difference between their residual variances is negligible (see Table 5.19). Model (a) has fewer parameters than model (d) and also higher residual variance than model (d). Hence, model (d) is adopted.

	Number of knots				
	1(1989) (a)	1(1992) (b)	1(1995) (c)	2(1989,1992) (d)	3(89,92,95) (e)
	0.0	0.12	0.92	0.02	0.02
	14.46	2.42	0.58	31.25	30.65
	76.24	137.58	176.04	32.36	33.85
	4.16	12.21	14.58	0.21	0.35
	2.53	19.03	33.64	2.16	1.76
	11.57	37.36	73.92	1.57	1.50
	6.58	17.94	56.23	6.24	5.33
	34.50	34.84	9.34	12.31	14.10
	0.03	3.71	3.81	8.27	8.82
	10.36	23.26	68.43	2.62	1.65
	14.76	16.36	6.34	0.18	0.06
χ^2	175.2	304.83	443.85	97.59	98.07
$\hat{N}$	220713	268625	419823	118827	90562

Table 5.19. *Estimates of residual variance and total number of persons infected with HIV (quadratic spline)*

5.5.2 The cubic spline

Unlike the linear and the quadratic models, the cubic model converged to a feasible solution. Also, as the distance between the single knot and the origin increases, the cubic spline tend to produce better estimates (see the graphs in Figure 5.5).

β	Number of knots				
	0 (a)	1(1989) (b)	1(1992) (c)	1(1995) (d)	2(1989,1992) (e)
β_0	-1198.58	-3069.14	-2388.68	-1911.79	2327.41
β_1	1378.58	3542.57	2693.84	2152.96	-2982.57
β_2	-338.90	-851.68	-633.73	-508.65	771.14
β_3	23.59	55.33	40.67	33.16	-50.37
β_4		-110.94	-358.60	-5559.14	519.89
β_5					-1558.10
$\hat{N}$	219696	159220	101845	-86082	-7664

Table 5.20. *Parameter estimates obtained using the cubic splines*

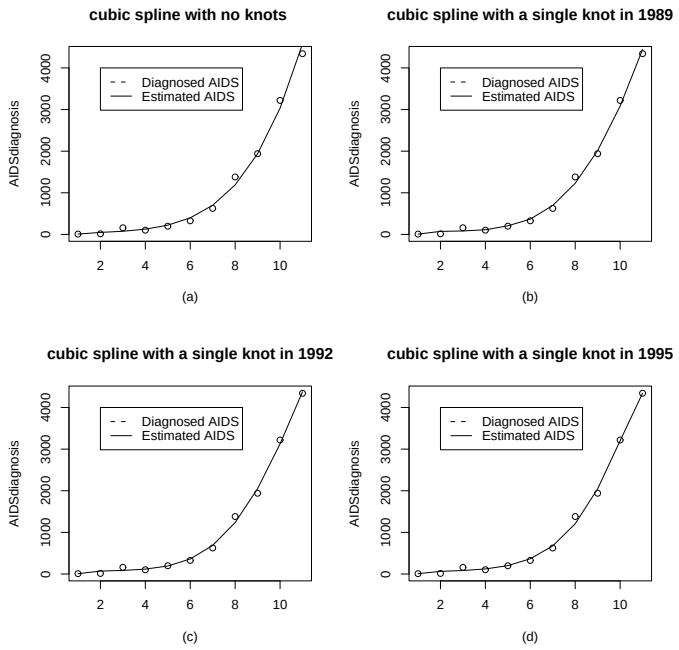

Figure 5.5. *Plot of observed (*dots*) and Backcalculated AIDS estimates (*line*) when infection curve is a cubic spline function*

Table 5.20 shows that the increase in the number of parameters (knots) has diminishing effect on the estimate of the total number of persons previously infected with HIV ($\hat{N}$). If it is possible to restrict the parameters to assume only positive values, then it may be possible to obtain increasing values of $\hat{N}$ but this may also affect the estimates of AIDS diagnosis incidence in Table 5.21.

The estimates became better with the inclusion of a knot and as the knot moves farther away from the origin. However, Table 5.22 is more revealing. The residual variance, represented by the χ^2, and the back-projected estimates of the HIV population $\hat{N}$, indicate that models (a), (b) and (c) are better fits than models (d) and (e). Considering the residual variances of these three, model (c) may be chosen.

Observed AIDS cases	Number of knots				
	0 (a)	1(1989) (b)	1(1992) (c)	1(1995) (d)	2(1989,1992) (e)
8	8.0	7.79	7.74	7.78	7.86
14	48.75	73.38	70.60	63.41	31.67
160	75.85	83.28	86.22	83.75	90.03
102	126.79	112.68	117.84	121.93	157.78
198	222.30	206.02	192.07	201.13	169.86
327	398.37	367.77	361.28	366.03	381.31
625	700.85	695.98	687.89	674.93	665.59
1381	1189.70	1229.13	1236.79	1206.10	1205.36
1940	1938.40	2014.23	2051.56	2046.78	2069.32
3219	3033.50	3082.83	3126.70	3200.81	3203.0
4342	4573.50	4442.91	4377.33	4343.35	4334.22
12316	12316	12316	12316	12316	12316

Table 5.21. *Estimates of diagnosed AIDS cases; cubic spline models*

	Number of knots				
	0 (a)	1(1989) (b)	1(1992) (c)	1(1995) (d)	2(1989,1992)) (e)
	0	0.005	0.009	0.006	0.003
	24.79	48.05	45.37	38.50	9.86
	98.36	70.67	63.13	69.42	54.38
	4.85	1.01	2.13	3.25	19.72
	2.66	0.31	0.18	0.05	4.66
	12.79	4.52	3.25	4.16	7.74
	8.21	7.24	5.75	3.69	2.47
	30.76	18.76	16.81	25.36	25.59
	0.001	2.74	6.07	5.57	8.08
	11.34	6.01	2.72	0.10	0.08
	11.72	2.29	0.29	0.00	0.014
χ^2	200.45	161.62	145.72	150.13	132.60
$\hat{N}$	219696	159220	101845	-86082	-7663.66

Table 5.22. *Estimates of residual variance and total number of persons infected with HIV (cubic spline)*

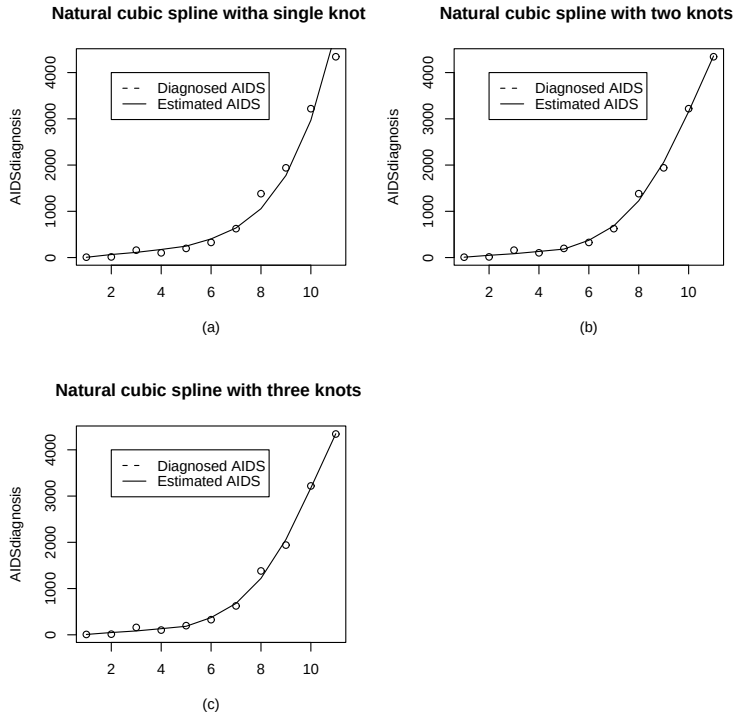

Figure 5.6. Plot of observed (*dots*) and Backcalculated AIDS estimates (*line*) when infection curve is a natural spline function

5.5.3 The Natural spline

The natural spline gave good estimates of the AIDS incidence diagnosis and these estimates improved as the number of knots increased. However, they performed poorly in the back-projection estimates of the HIV population.

Due to the inability of all the natural spline models to give feasible estimates of the number of persons previously infected with HIV and AIDS, we are unable to select any of them for further consideration. Each of the models underestimated the observed number of persons diagnosed for HIV or AIDS.

We now bring together the selected models from all the spline models in

β	Number of knots		
	1(1989) (a)	2(1989,1992) (b)	3(89,92,95) (c)
β_0	-300.05	-214.30	-216.20
β_1	145.42	105.32	106.21
β_2	118.47	221.71	216.06
β_3		-849.40	-761.09
β_4			-1368.55
$\hat{N}$	26471	59367	10785

Table 5.23. *Parameter estimates from the Natural spline models*

Observed AIDS cases	Number of knots		
	1(1989) (a)	2(1989,1992) (b)	3(89,92,95) (c)
8	6.12	7.96	7.91
14	65.52	48.83	49.20
160	112.11	83.11	83.74
102	174.81	132.07	132.92
198	246.97	183.09	184.45
327	402.67	373.69	372.45
625	640.67	684.13	679.02
1381	1056.35	1228.06	1220.36
1940	1774.77	2059.77	2060.66
3219	2968.66	3155.89	3176.70
4342	4868.02	4359.39	4348.59
12316	12316		

Table 5.24. *Estimates of AIDS diagnosis from Natural spline*

	Number of knots		
	1(1989)	2(1989,1992)	3(89,92,95)
	(a)	(b)	(c)
	0.52	0.00	0.001
	40.49	24.84	25.18
	20.46	71.14	69.45
	30.33	6.85	7.19
	9.71	1.21	1.0
	4.22	5.83	5.55
	0.35	5.11	4.30
	99.78	19.05	21.15
	15.38	6.96	7.07
	21.11	1.26	0.56
	56.84	0.07	0.01
χ^2	309.19	142.33	141.45
$\hat{N}$	26471	59367	10785

Table 5.25. *Estimates of residual variance and total number of persons infected with HIV (Natural spline)*

order to make a closer comparison and possibly select the most outstanding of the models. Table 5.26 gives the model from each of the splines functions. It may be inferred from the table that the estimate of the cumulative number of HIV infected patients in Nigeria range from 101845 to 198157 depending on the form of spline assumed for the infection curve. We recall that the estimate for this population obtained from the step function model was 150582 and 150580 for the multinomial and Poisson regression respectively. These step function estimates are about the mid-point of the range of spline function estimates.

Given the size of the residual variance in each model in Table 5.26, the quadratic spline model seems to be the better model. Thus, selecting this model, the estimated cumulative number of HIV/AIDS cases is 118,827.

Observed AIDS cases	Linear 2(1989,1992) (a)	Quadratic 2(1989,1992) (b)	Cubic 3(89,92,95) (c)
8	8.32	7.58	7.74
14	40.40	55.73	70.60
160	72.38	102.43	86.22
102	128.77	106.73	117.84
198	227.97	178.35	192.07
327	401.11	350.42	361.28
625	701.30	690.67	681.89
1381	1195.00	1254.74	1236.79
1940	1953.77	2070.88	2051.56
3219	3047.66	3128.45	3126.70
4342	4539.30	4370.02	4377.33
χ^2	202.05	97.69	145.72
$\hat{N}$	198157	118827	101845

Table 5.26. *Estimates of HIV/AIDS from selected spline models*

5.6 Non-parametric Back-projection

A major disadvantage of the parametric back-projection is the problem of identifying the functional form of the HIV incidence curve as different forms may be consistent with the observed AIDS incidence. In order to avoid this limitation, Becker et al (197) proposed an imposition of a smoothness restriction on a non-parametric form and estimates are obtained using the non-parametric maximum likelihood approach implemented in the EM algorithm.

5.7 Application to Hong Kong data

Becker et al(197) and Chau et al (102) applied the above method to Australia's and Hong Kong's AIDS incidence data respectively. We reproduce the results of Chau et al (102) for Hong Kong and extend the same technique to Nigerian AIDS data.

Applying this technique to Hong Kong data, Chau et al (102) assumed that the incubation period distribution is Weibull distribution as suggested by Brookmeyer and Goedert (231). The scale parameter β for this distrbution was estimated based on a median incubation period of 10 years (133). For someone infected with HIV at time u, the natural hazard function for AIDS diagnosis at time x is given as

$$p(x/u) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \tag{5.8}$$

Hence,

$$p(x/u) = \frac{2.516}{12.147}\left(\frac{x}{12.147}\right)^{1.516}$$

Recognizing the effect of different regimes of treatment on incubation distribution, the approach of Muñoz and Hoover (14) was adopted and three different

time (or treatment) regimes were recognized.

$$F(x/u) = \begin{cases} 1 - e^{-\int_0^{x+0.5} p(w/u)dw} & for\ 1979 \leq u \leq 1987 \\ 1 - e^{-\int_0^{x+0.5} \{1-0.125(u-9)\}p(w/u)dw} & for\ 1988 \leq u \leq 1991 \\ 1 - e^{-\int_0^{x+0.5} 0.5p(w/u)dw} & for\ 1992 \leq u \leq 2000 \end{cases} \tag{5.9}$$

and,

$$f(x/u) = \begin{cases} F(x/u) & for\ x = 0 \\ F(x/u) - F(x-1/u) & for\ x = 1, 2, \ldots \end{cases}$$

The discrete time is adjusted by adding 0.5 units to the year x in deriving the cumulative distribution.

Using the above information and equations 5.8 and 2.35, we obtained the result in Figure 5.7 for Hong Kong

The graph in Figure 5.7 suggests that the HIV infection Hong Kong reached its peak in 1991 and subsequently declined. However, it took some time (about nine years), for the observed AIDS incidence to reach its peak in 1999 after which the number of AIDS cases diagnosed diminished. In order to have a measure of the extent of uncertainty in the estimates, we apply the nonparametric bootstrap procedure to obtain point-wise confidence interval for the estimates of HIV incidence.

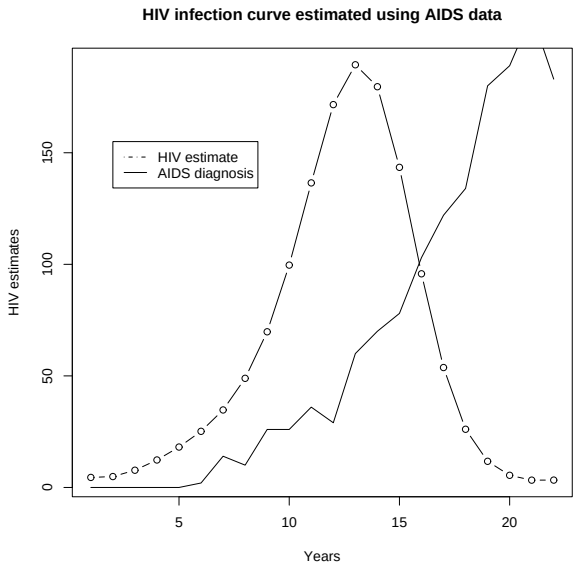

Figure 5.7. *Observed AIDS cases(solid line), Estimated HIV infection(dotted line)*

5.7.1 Precision of the estimates

The bootstrap estimates of precision was used to obtain a 95 per cent confidence interval for the estimates. The following steps were undergone

- Using the observed data, we obtained *unsmooth* estimate of the infection curve by performing the EM algorithm
- We then substituted these unsmooth estimates into equation 5.1 to obtain the mean AIDS diagnosis
- Using the mean estimates of AIDS diagnosis obtained above as the mean of the Poisson process that brought about the observed AIDS incidence, we generated for each estimate, 1000 Poisson random variables
- We then applied each of the 1000 rows (as values of observed AIDS) to the EMS algorithm to obtain smooth estimates of the HIV infection curve

Year	Lower bound	HIV estimates	upper bound
1979	1	5	8
1980	1	5	9
1981	3	8	12
1982	8	12	25
1983	12	18	25
1984	17	25	35
1985	24	35	48
1986	36	49	65
1987	52	70	86
1988	69	100	114
1989	90	137	144
1990	106	172	174
1991	115	189	200
1992	117	180	220
1993	107	143	231
1994	81	96	243
1995	54	54	265
1996	21	26	327
1997	11	12	445
1998	6	6	633
1999	3	4	818
2000	2	4	822
Total	936	1345	4949

Table 5.27. *Bootstrap confidence interval for HIV incidence estimates for Hong Kong*

- Each column resulting from the previous step is arranged in ascending order of size.

- The upper bound is the 975th (or 950th) row and the lower bound is the 25th (or 50th) row. This gives the 95% (or 90%) confidence interval of the estimates– which are point-wise confidence interval of the estimates

Doing this, we obtained the results displayed in Table 5.27

In Figure 5.8 solid lines represent the point-wise confidence intervals, the cross line represent the estimates of the HIV infection curve and the triangular points represent the observed HIV diagnosis. The imprecision of the back-projection

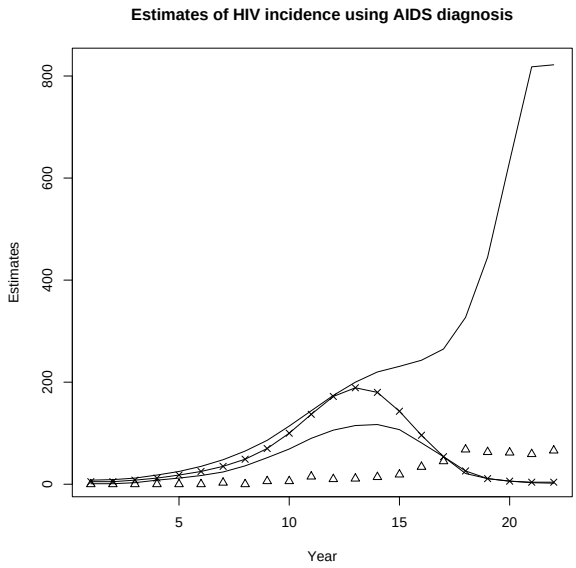

Figure 5.8. *Observed AIDS cases(triangular points), Estimated HIV infection(crossed line), 95% CI(solid lines)*

model in estimating accurately the recent incidence of HIV is very clear in the graph in Figure 5.8. As can be seen from the graph, the estimates obtained using AIDS data underestimate the HIV incidence in the recent past as depicted by the widths of the confidence interval at the most recent past.

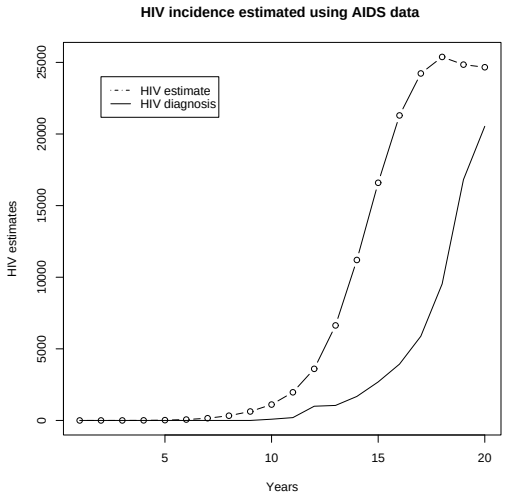

Figure 5.9. *Observed HIV cases(solid line), Estimated HIV infection curve(dotted line)*

5.8 Application to Nigeria

Applying this same approach to Nigerian AIDS data as published by the Nigeria Institute of Medical Research (83), we observe that antiretroviral therapy (ART) was introduced in the country in mid 2001 and was available only for 10,000 patients out of an estimated 600,000 who needed the drug in the country as at that time. Therefore, there were no effect of treatment therapy on incubation distribution of AIDS in Nigeria as at the period being considered in this analysis. Hence, we shall assume that the cumulative distribution of the incubation period for the country is

$$F(x/u) = \begin{cases} 1 - e^{-\int_0^{x+0.5} p(w/u)dw} & for\ 1980 \leq u \leq 1999 \end{cases}$$

and $p(x/u)$ is as defined in equation 5.8. Using the above incubation distribution function, estimates obtained is represented in Figure 5.9,

Figure 5.9 shows the observed HIV diagnosis and the estimated HIV incidence

Year	Diagnosed HIV	HIV estimates
1980	0	0
1981	0	0
1982	0	1
1983	0	6
1984	0	21
1985	0	62
1986	0	155
1987	0	331
1988	0	625
1989	87	1110
1990	198	1966
1991	993	3604
1992	1049	6635
1993	1676	11204
1994	2690	16595
1995	3932	21296
1996	5878	24227
1997	9531	25381
1998	16816	24844
1999	20537	24661
Total	63387	162724

Table 5.28. *Diagnosed and Estimated HIV cases in Nigeria obtained using non-parametric back-projection*

obtained using AIDS diagnosis data. The distance between the two lines, at any time point, gives a measure of the hidden HIV population. That is, those infected with HIV in the population but not yet diagnosed. The graph also indicates that the HIV incidence is on the increase and is yet to peak. The seeming decrease at the top of the graph could be the effect of the back-calculation procedure as it imprecisely estimate the most recent infections. Table 5.28 gives the observed and expected HIV incidence.

The last column of the Table 5.28 shows the likely number of persons infected with HIV in the country each year. It shows for instance that in 1983, about 6 persons already had the virus in the population none of which were diagnosed. Also in 1989, it is estimated that about 1,110 persons were infected but only 87

of them were diagnosed. Cumulatively, it is estimated that between 1980 and 1999, about 162,724 Nigerians were infected with the virus but only 75,703 cases were diagnosed. The estimates in this table are point estimates, a measure of the level of imprecision of these estimates are obtained using the nonparametric bootstrap procedure in the next section.

5.8.1 Precision of the Estimates

The unsmooth EM algorithm could not converge for the Nigeria AIDS and HIV data. This made it difficult obtaining the bootstrap confidence interval for the estimates. A careful look at the estimates obtained at the end of each iteration shows that as the number of iteration increase, the estimates for the last three years rapidly tend towards zero. The entire estimates vanish to infinity whenever the number of iterations goes beyond a certain point. Several attempts were made at correcting this, including the elimination of the last three or two values of the estimates when setting the criterion for convergence, but no meaningful results were obtained. That is, the EM algorithm could not converge even with this criterion. There were no problems of convergence with the EMS- that is, EM algorithm with a smoothing step.

We tried to investigate the cause of this by using the Nigerian data with the Hong Kong model and the Hong Kong data with the Nigerian model. We found that the Hong Kong data converged in the Nigerian model but the Nigeria data could not converge in the Hong Kong model. We then found that the algorithm converged if we divided the Nigerian data consistently by 100. This suggests that there is a relationship between the magnitude of the individual data and

Year	Lower bound	HIV estimates	upper bound
1980	0	0	20
1981	0	0	22
1982	0	1	40
1983	0	6	78
1984	0	21	150
1985	1	62	250
1986	1	155	442
1987	30	331	713
1988	142	625	1171
1989	401	1110	2016
1990	917	1966	3604
1991	1924	3604	6070
1992	3521	6635	9862
1993	6383	11204	15134
1994	10312	16595	21880
1995	11293	21296	31324
1996	8811	24227	45533
1997	5111	25381	68003
1998	2512	24844	98124
1999	2420	24661	101360
Total	53779	162724	405796

Table 5.29. *Estimates of HIV incidence in Nigeria and Bootstrap confidence interval*

the convergence of the EM algorithm that has not been documented by other researchers using this back-projection methods. We were unable to determine the reason for this behaviour which was a continuing concern, but we proceeded to carry out the analysis using scaled data. Dividing the AIDS data by 100 in order to run the algorithm, then multiplying up by 100 the end points of the derived interval estimates, we obtained the 95 percent bootstrap confidence interval as shown in Tables 5.29 and Figure 5.10

Figure 5.10 shows the HIV incidence estimates (*cross points*), their corresponding 95 per cent *point-wise bootstrap confidence intervals (solid lines)* and the observed HIV positive cases (*triangular points*). It is estimated that the cumulative number of infected individuals in the country from 1980 to 1999 lie

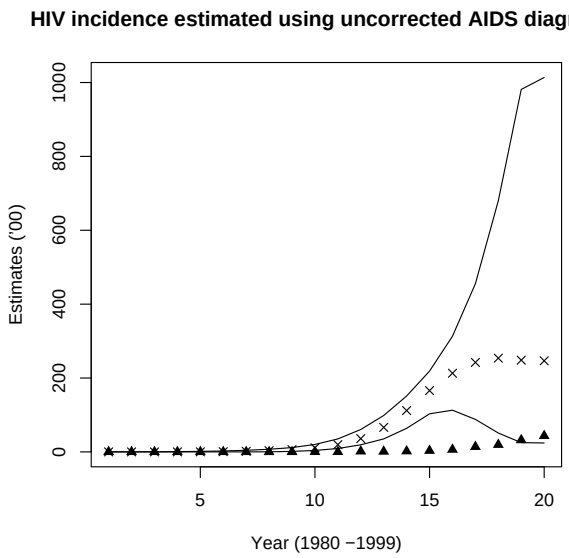

Figure 5.10. *Bootstrap 95%CI (solid lines), Estimate of HIV incidence (cross points), and Observed HIV cases (triangular points)*

somewhere between 53,779 and 405,796 with the best estimate as 162,724.

The increased uncertainty in estimating the recent incidence of HIV infection is depicted by the wide interval at the right end of the graph.

5.8.2 Projection

In order to project the estimated AIDS incidence to a near future (say 2001), we took the following steps

- We extended the dimension of the incubation distribution matrix from 20x20 to 22x22
- Due to the imprecision of estimates of the most recent HIV infections, the most recent diagnosis were not used in the projection. To correct for this limitation, it is usually assumed that the infection intensity is constant for some period of time. For instance, Chau and his colleagues (102) assumed

that the incidence of HIV for the years after 1997 are at the same level as in the year 1997. Also, average of the three most recent years can be used as a constant infection intensity for projections.

- The observed data were extended by two years (up to 2001) using the assumption that rate of diagnosis is the same as in 1999

- Using the expanded incubation distribution and the extended data, the EM algorithm was performed to obtain estimates of HIV infection up to the forecast years.

- Result from the last step was substituted into equation (5.1) to obtain the estimated mean diagnosed AIDS cases. This gives the projected estimates of AIDS diagnosis.

- In order to obtain the point-wise confidence interval for the projected estimates, the mean AIDS diagnosis obtained in the step above is used as the mean of a Poisson process assumed to have generated the observed diagnosis. With these means, 1000 Poisson random samples were generated.

- The 1000 Poisson random samples were then arranged in ascending order and the 25th and 975th percentiles were obtained and plotted on the same graph with the estimates of AIDS diagnosis. See Figure 5.11

In 2000 and 2001, it is predicted that 4404 and 4992 cases of AIDS respectively, were diagnosed in Nigeria. Between 1989 and 2001, it is estimated that between 13,400 and 29,800 cases of AIDS were diagnosed with best estimate being 21,002 cases, of which only 12,316 cases were diagnosed.

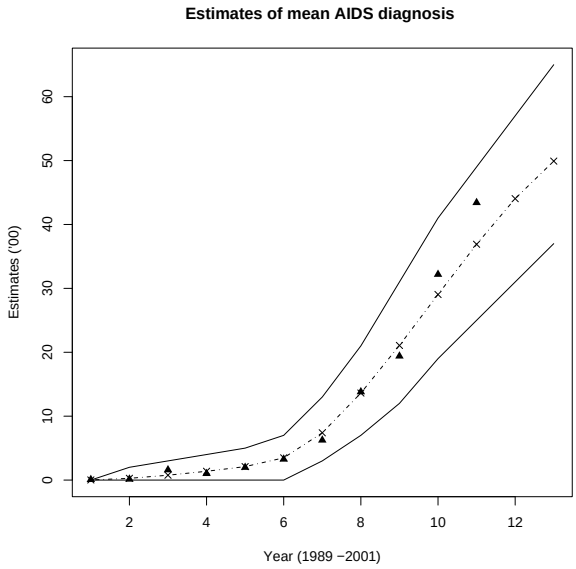

Figure 5.11. Estimates and projected number of AIDS cases in Nigeria (*dotted line and cross*), observed cases of AIDS (*triangular points*)and the 95 per cent confidence interval (*thick lines*)

Year	Lower bound	Observed AIDS cases	Estimated cases	Upper bound
1989	0	8	2	0
1990	0	14	28	200
1991	0	160	76	300
1992	0	102	138	400
1993	0	198	211	500
1994	0	327	348	700
1995	300	625	740	1300
1996	700	1381	1362	2100
1997	1200	1940	2107	3100
1998	1900	3219	2904	4100
1999	2500	4342	3690	4900
subtotal	6600	12316	11606	17600
2000	3100	-	4404	5700
2001	3700	-	4992	6500
Total	13400	-	21002	29800

Table 5.30. *Observed and estimated diagnosed number of AIDS cases in Nigeria and the bootstrap CI*

5.9 The modification of the non-parametric back-projection

The ordinary back-calculation method, as considered in the previous sections, makes use of diagnosed AIDS data in reconstructing the HIV infection curve. The approach has the limitation of not predicting precisely the HIV incidence in the recent past due to the long incubation period between HIV infection and AIDS diagnosis. In order to overcome this limitation, Cui and Becker (250) and Chau et al (102) suggested the use of HIV data set in back-calculation for estimating HIV incidence curve.

5.10 Application to Hong Kong data

Chau et al (102) applied the above techniques on the HIV/AIDS data for Hong Kong. They identified the proportion of routine and symptom-related tests as 0.20 and 0.26, equating these values to equations (2.41) and (2.42) respectively, they obtained the parameters $\nu = 0.0306$ (the constant hazard for routine testing) and $\gamma = 2.385$ (the proportional coefficient of synptom related HIV testing).

Using the natural hazard function for the diagnosis of AIDS at time x given HIV infection at time u as suggested by Brookmeyer and Goedert (231), we reproduce the result obtained in Chau et al (102) as shown in Table 5.31

Year	HIV cases	HIV estimate
1979	0	3
1980	0	3
1981	0	7
1982	0	21
1983	0	39
1984	2	54
1985	14	60
1986	10	63
1987	26	70
1988	26	85
1989	367	113
1990	29	151
1991	60	194
1992	70	231
1993	78	260
1994	103	275
1995	112	268
1996	134	242
1997	180	195
1998	189	139
1999	212	94
2000	183	85
Total	1394	2653

Table 5.31. *Observed and estimated diagnosed number of HIV positive cases in Hong Kong*

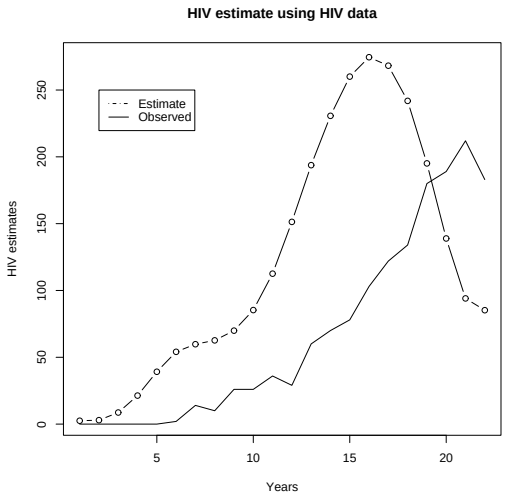

Figure 5.12. *Observed (solid line) and estimated (dotted line) HIV diagnosis in Honk Kong*

Year	$\nu = 0.01$ $\gamma = 2.385$	$\nu = 0.10$ $\gamma = 2.385$	$\nu = 0.0306$ $\gamma = 0.5$	$\nu = 0.0306$ $\gamma = 1.5$	$\nu = 0.0306$ $\gamma = 3.0$
1979	7	0.2	4	3	2
1980	8	0.2	4	4	2
1981	18	1	11	10	8
1982	32	6	27	24	20
1983	47	18	56	44	37
1984	55	37	94	65	50
1985	60	49	121	76	54
1986	66	52	136	81	56
1987	80	56	149	90	63
1988	104	61	163	105	78
1989	139	76	192	133	104
1990	181	104	240	173	141
1991	222	137	296	218	181
1992	254	165	342	256	217
1993	271	196	386	288	246
1994	268	230	422	306	261
1995	242	258	433	302	257
1996	197	275	412	276	232
1997	144	260	341	223	189
1998	95	209	233	155	137
1999	62	152	143	100	95
2000	58	137	126	89	87
Total	2610	2479	4331	3021	2517

Table 5.32. Sensitivity Analysis

5.10.1 Sensitivity analysis

Table 5.32 shows the sensitivity analysis for different values of the two parameters γ and ν. As can be seen, the estimates are insensitive to changes in the parameters except for the third and fourth scenario where the value of γ is 0.5 and 1.5 respectively. It appears the estimates are more sensitive to changes in γ (symptom related proportionality coefficient) than in ν. However, variation in the parameters result in a shift in the estimated peak of the epidemic and different values for the median incubation period.

Year	Lower bound	Observed HIV cases	Estimated HIV cases	Upper bound
1979	1	0	3	5
1980	2	0	3	5
1981	5	0	9	13
1982	15	0	21	27
1983	30	0	39	46
1984	43	2	54	61
1985	48	14	60	70
1986	51	10	63	76
1987	58	26	70	86
1988	73	26	85	105
1989	98	36	113	133
1990	131	29	151	174
1991	167	60	194	218
1992	201	70	231	256
1993	223	78	260	284
1994	231	103	275	299
1995	221	122	268	299
1996	189	134	242	299
1997	142	180	195	275
1998	98	189	139	273
1999	66	212	94	287
2000	62	183	85	290

Table 5.33. Observed and estimated HIV positive cases and the bootstrap C I

5.10.2 Bootstrap Confidence Interval

The imprecision of the back-projection model in estimating accurately the recent incidence of HIV is very clear in the two graphs in Figures 5.13 and 5.14. The solid lines represent the point-wise confidence intervals, the cross points represent the estimates of the HIV infection curve and the triangular points represent the observed HIV diagnosis. As can be seen from the graphs, the estimates obtained using HIV diagnosis data also underestimate the HIV incidence in the recent past. This is worse when AIDS diagnosis data is used as seen in the previous section.

The drop in the shape of the observed HIV incidence has an influence on

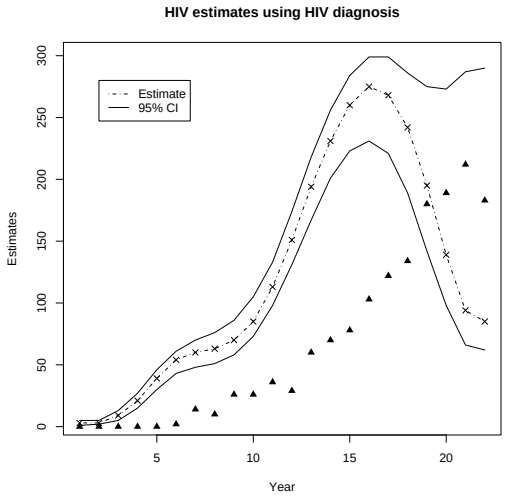

Figure 5.13. *95% CI (solid line), Estimated (crossed dotted line) and Observed(triangular) HIV diagnosis*

the outcome of back-projection. In order to eliminate this effect, the data was cut in the year where the drop is noticed (year 2000). The estimates plotted in the graph in Figure 5.14 with the corresponding confidence interval was obtained using the data without year 2000. It can be seen that the drop in the estimated infection curve has been eliminated and instead the curve now levels up after the peak. This may suggest that the HIV incidence is still on the increase beyond 1999.

5.10.3 Projection

In projecting future HIV incidence, the effect of the slight drop in 2000 is taken into cognisance. Table 5.34 and Figure 5.15 show the fitted and projected value of HIV incidence and the 95% point-wise confidence interval using data up to and including year 2000. While Table 5.35 and Figure 5.16 show that obtained by dropping year 2000 data. While both approach projected an increasing epidemic for years 2001 and 2002, the elimination of year 2000 from the data series resulted

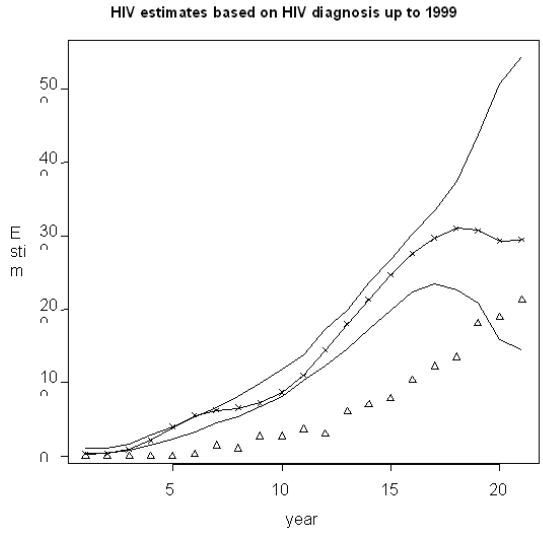

Figure 5.14. *95% CI (solid line), Estimated (crossed dotted line) and Observed (triangular) HIV diagnosis. Using data up to 1999*

in moderately higher estimate for the projected years and also the estimate of cumulative HIV infection up to 2000 is higher than the observed cumulative HIV diagnosis. The graph in Figure 5.15 (obtained using the whole data) appears to have peaked and almost stable at the peak while that of Figure 5.16 (obtained using data up to 1999) seems to be on the increase.

Year	Lower bound	Observed HIV cases	Estimated HIV cases	Upper bound
1984	0	2	2	5
1985	2	14	6	11
1986	5	10	11	18
1987	10	26	17	26
1988	15	26	25	36
1989	23	36	34	46
1990	32	29	44	57
1991	42	60	56	71
1992	54	70	70	87
1993	68	78	87	105
1994	86	103	106	126
1995	105	122	128	147
1996	123	134	146	168
1997	140	180	165	191
1998	157	189	181	206
1999	167	212	194	220
2000	175	183	202	227
subtotal	1204	1474	1474	1747
2001	177		206	234
2002	180		207	236
Total	1561		1887	2217

Table 5.34. *Fitted and projected HIV positive cases and the bootstrap C I using data up to year 2000*

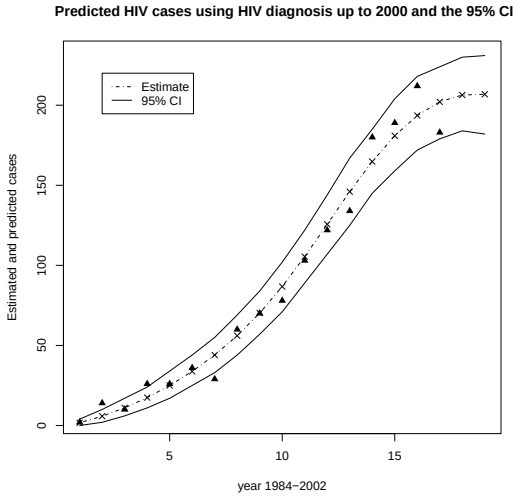

Figure 5.15. *95% CI (solid line), fitted and projected estimates(crossed dotted line) and observed (triangular) HIV diagnosis. Using data up to 2000*

Year	Lower bound	Observed HIV cases	Estimated HIV cases	Upper bound
1984	0	2	2	4
1985	2	14	6	11
1986	5	10	11	18
1987	10	26	18	26
1988	16	26	25	36
1989	23	36	34	46
1990	32	29	44	58
1991	41	60	56	70
1992	53	70	70	87
1993	68	78	85	104
1994	84	103	104	124
1995	104	122	124	147
1996	123	134	146	171
1997	141	180	167	194
1998	161	189	187	213
1999	178	212	205	236
2000	191	183	219	249
subtotal	1232	1474	1503	1794
2001	202		230	261
2002	206		237	267
Total	1640		1970	2322

Table 5.35. *Fitted and projected HIV positive cases and the bootstrap CI using data up to year 1999*

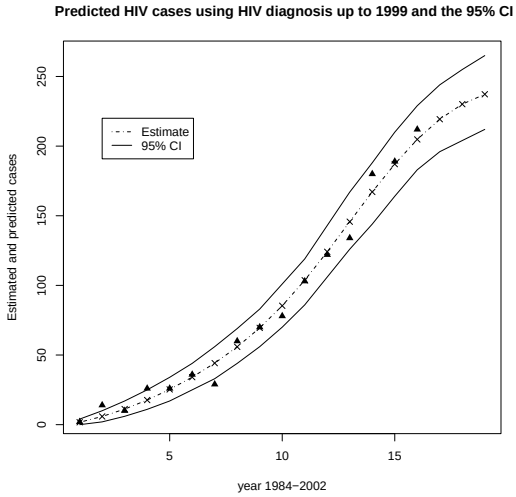

Figure 5.16. *95% CI (solid line), Fitted and projected estimates(crossed dotted line) and observed (triangular) HIV diagnosis. Using data up to 1999*

5.11 Application to Nigerian data

This model approach may be said to be ideal for modeling the Nigerian HIV/AIDS epidemic. This is because, the HIV test data is more readily available than the AIDS diagnosis data. Underreporting is much higher with AIDS data than HIV data. People undergo HIV tests for one reason or the other (routine, mandatory or symptom related tests), once confirmed positive, many of them are lost to traditional healers, prayer houses and the many self-acclaimed HIV/AIDS doctors spread all over the country. Only a very small proportion of those who tested positive to HIV make use of the orthodox medicine (hospital) when AIDS symptoms develop. the majority of the persons recorded as AIDS cases were individuals who were not aware of their HIV status until the symptoms of AIDS manifest.

Therefore any model based on HIV data will provide better insight into the epidemic situation in the country provided the data is fairly representative of the entire population.

To obtain estimates of the HIV incidence curve for Nigeria, HIV data set collected from 659 testing sites and published by the Nigeria Institute of Medical Research (83) is used. The data indicate that of all the HIV positive cases diagnosed between 1989 and 1999, 40.6 per cent were clinical AIDS diagnosis, 29.6 per cent were routine tests, and 29.8 per cent were blood donors and pretrans. Using this information, we obtained the parameter estimates of the routine and simultaneous reporting as $\nu = 0.0349$ and $\gamma = 0.851$ respectively. Hence, the hazard for HIV positive test in year x for someone infected in year u is

$$p'(x/u) = 0.0349 + 0.851\left(\frac{2.516}{12.147}\right)\left(\frac{x}{12.147}\right)^{1.516}$$

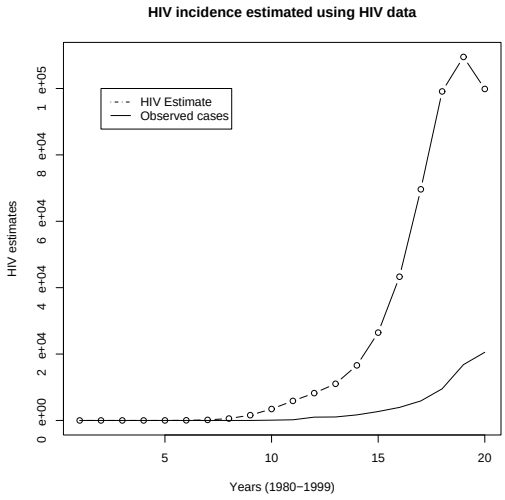

Figure 5.17. *Estimates of HIV diagnosis in Nigeria*

Using the induction distribution in equation 2.45, the shape of the induction density is given in Figure **??**

The median induction period is computed to be 8.9 years. This indicates that most of the patients underwent HIV tests at a later stage of the infection, may be as a result of one sickness or the other. This information justifies our pushing back the estimation to start from 1980.

Figure 5.17 shows the graph of the estimates of incidence of HIV positive cases (broken lines) and the observed number of HIV positive tests (solid lines) obtained using the modified back-projection (HIV data only). Estimates show a rising trend of the epidemic. The seeming drop in the HIV incidence of the most recent year (1999) can be attributed to the inability of the back-projection method to accurately estimate the most recent past incidence of HIV infection. The observed data show a continual increase in incidence over the years. Table 5.38 gives the observed and estimated number of HIV positive cases. The difference between the two columns give the estimated number of infections yet unobserved.

The model estimated the cumulative number of HIV infections between 1980

Year	HIV cases	HIV estimate
1980	0	0
1981	0	0
1982	0	0
1983	0	1
1984	0	8
1985	0	40
1986	0	167
1987	0	573
1988	0	1585
1989	87	3444
1990	6198	5889
1991	993	8227
1992	1049	11043
1993	1676	16598
1994	2690	26462
1995	3932	43297
1996	5878	69628
1997	9531	99112
1998	16816	109508
1999	20537	99861
Total	63387	495439

Table 5.36. *Observed and estimated number of HIV infection in Nigeria*

and 1999 to be approximately 495,439 as against 75,703 cases diagnosed between 1989 and 1999. Comparing this estimate with that obtained using AIDS data (see Table 5.28), an appreciable increase in the estimate is observed. The estimate obtained using HIV data is more than 3 times that obtained using AIDS incidence data. As earlier stated, the HIV data contain more information about HIV infection than the AIDS data. Hence, it is expected that estimates obtained using this approach should be more reliable than that obtained using AIDS data. A measure of precision of this estimate is obtained by the bootstrap procedure.

5.11.1 Bootstrap Estimates of precision

As earlier pointed out, the EM algorithm was unable to converge for the Nigeria AIDS and HIV data. Consequently, the HIV data was divided by 1000. The bootstrap confidence interval was then obtained using the steps outlined in the previous section.

It is estimated that the number of individuals infected with HIV in Nigeria between 1989 and 1999 lie somewhere between 347,792 and 621,494 with the best estimate being 495,439. Observe that the width of the interval gets larger as we approach the recent years. However, when compared with that obtained using the AIDS data, the precision of the estimates using this method seem better. The graph in Figure 5.18 gives a pictorial view of the point-wise spread.

Year	Lower bound	HIV cases	HIV estimate	Upper bound
1980	0	0	0	79
1981	0	0	0	82
1982	0	0	0	103
1983	0	0	1	155
1984	4	0	8	288
1985	22	0	40	536
1986	91	0	167	935
1987	304	0	573	1606
1988	794	0	1585	2752
1989	1775	87	3444	4533
1990	3419	198	5889	6992
1991	5779	993	8227	10799
1992	9279	1049	11043	16104
1993	14873	61676	16598	25098
1994	23267	2690	26462	37614
1995	36372	3932	43297	54322
1996	53504	5878	69628	77691
1997	66794	9531	99112	104184
1998	67599	1686	109508	133680
1999	63916	20537	99861	143941
Total	347792	63387	495439	621494

Table 5.37. *Observed and estimates of number of HIV infection in Nigeria with the 95% bootstrap CI*

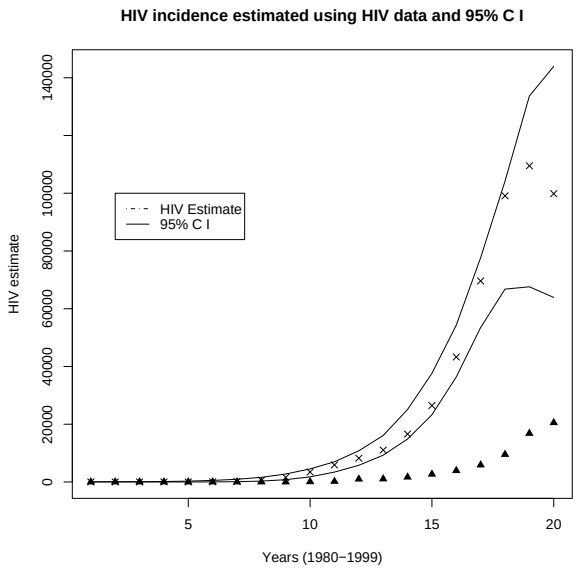

Figure 5.18. *The 95% point-wise bootstrap confidence interval (solid lines) for estimates of HIV incidence (cross points) and observed HIV positive cases (triangular points)*

Year	Lower bound	HIV cases	HIV estimate	Upper bound
1989	0	87	67	200
1990	0	198	278	600
1991	300	993	774	1400
1992	600	1049	1207	1900
1993	1000	1676	1733	2600
1994	1700	2690	2639	3700
1995	2900	3932	4126	5500
1996	4100	5878	5655	7100
1997	8200	9531	10140	12100
1998	13500	16816	15877	18200
1999	15900	20537	18607	21500
subtotal	48200	63387	61103	74800
2000	17800		20680	23500
2001	19800		22678	25900
Total	85800		104461	124200

Table 5.38. *Observed, estimates and bootstrap CI for HIV positive cases*

5.11.2 Projection

In order to obtain the estimates of the future number of HIV infections (say for 2000 and 2001), we followed the steps outlined in the previous section. Based on the observed data, we obtained the fitted and projected number of HIV positive cases and their 95 per cent point-wise confidence interval. It is projected that there were 20,680 and 22,678 diagnosed HIV positive cases in 2000 and 2001 respectively.

Figure 5.19 gives an interesting picture of stages of the epidemic in Nigeria. It seems to define three obvious stages approximately coinciding with the three positions of the steps and knots used in the previous section. The shape of the graph depicts a rising trend of the epidemic indicating that in the near future, the number of infected individual will be on the increase if no pragmatic program is adopted to curb the spread of the virus.

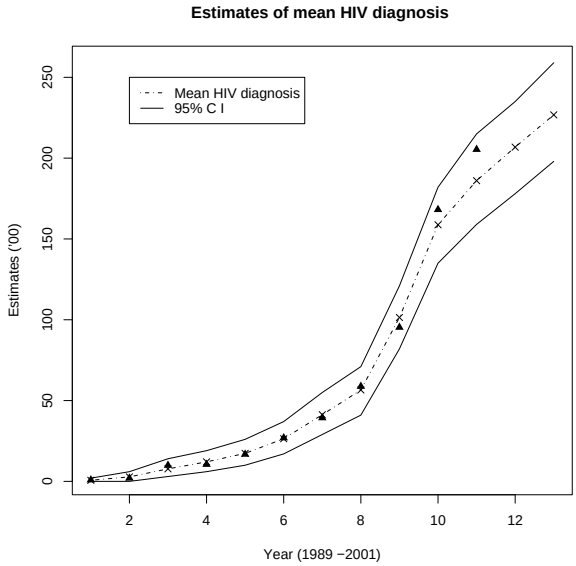

Figure 5.19. *95% CI (solid line), fitted and projected estimates(crossed dotted line) and Observed (triangular) HIV diagnosis in Nigeria*

5.12 Comparison of the parametric and non-parametric Back-projection

We refer to the parametric back-projection as all back-projection method that assumes a form of parametric distribution for the infection intensity. Non-parametric back-projection is the converse, no parametric distribution is assumed for the infection curve. We have assumed two parametric forms for the infection curve; the step function and the spline function and have obtained back-projected estimates for different scenarios of HIV/AIDS epidemic using data from Nigeria, America and Hong Kong. Estimates were also obtained from the same sets of data using non-parametric back-projection.

In this section, we compare results obtained from the parametric back-projection and that obtained from the non-parametric back-projection with a view of establishing the suitability of the methods in the different epidemic scenarios.

Experience in this research shows that the parametric methods are flexible and can yield infinitely many solutions depending on the number of steps (or knots) and the positions of the steps (or knots). Results in Tables 5.6 and 5.9 give credence to this. The extension of the number of steps from three to four gave different estimates of the infection intensity for HIV/AIDS epidemic in Nigeria. Also it is possible to obtain negative estimates of the infection curve in some time intervals which is unrealistic in real world situation. See Tables 5.14, 5.17, 5.20 and 5.23. However, this problem can be overcome by placing restrictions on the parameters. The estimates shown in these tables indicate the effect of the number and positioning of the knots on estimates of HIV infection intensity, the repositioning of the single knot and the increase in the number of knots resulted in large differences in the estimates of the cumulative HIV infection. It is important also to point out that the different parametric forms assumed for the HIV infection curve yielded different estimates of the cumulative HIV infection. Table 5.40 gives a summary of estimates of cumulative HIV infection obtained using the step function and spline for America and Nigeria. It seems from the table that estimates obtained using the spline is slightly lower than that obtained using the step functions. This depends however, on the choice of the position and number of steps or knots. Another factor that may affect the estimate of the HIV population is the choice of the type of spline function. Our analysis show that different spline function gave different estimates of the HIV infection curve. See Table 5.26.

Surprisingly, all these models (different splines and step functions) gave good fit of the AIDS incidence data. Though slight variation is noticed in the value of the AIDS incidence per time period, the cumulative number of AIDS diagnosis is the same. See the third columns of Tables 5.5 and 5.10 and Tables 5.15, 5.18

Year/quarter	Observed	Parametric	Nonparametric
1977.1-81.4	374	413.49	249
1982.1	185	141.03	184
1982.2	200	199.23	256
1982.3	293	273.7	340
1982.4	374	365.86	434
1983.1	554	476.89	540
1983.2	713	608.56	658
1983.3	763	762.33	791
1983.4	857	936.66	941
1984.1	1147	1141.42	1114
1984.2	1369	1368.25	1313
1984.3	1563	1621.18	1543
1984.4	1726	1900.45	1809
1985.1	2142	2206.96	2116
1985.2	2525	2544	2466
1985.3	2951	2914.64	2861
1985.4	3160	3321.37	3300
1986.1	3819	3765.48	3782
1986.2	4321	4247.61	4303
1986.3	463	4769.38	4862
1986.4	5792	5330.24	5451
1987.1	6155	5930.96	6065
1987.2	6816	6571.42	6691
1987.3	7491	7251.05	7310
1987.4	7726	7969.68	7906
1988.1	8483	8726.59	8484
TOTAL	76362	75758.43	75770

Table 5.39. *Estimates of AIDS cases in the US obtained using parametric and non-parametric back-projection*

Country	Multinomial	QL/Poission	Spline
Nigeria	150582	150580	118827
America	1224083	1224021	1182490

Table 5.40. *Estimates of the Numbers previously infected with HIV in Nigeria and America*

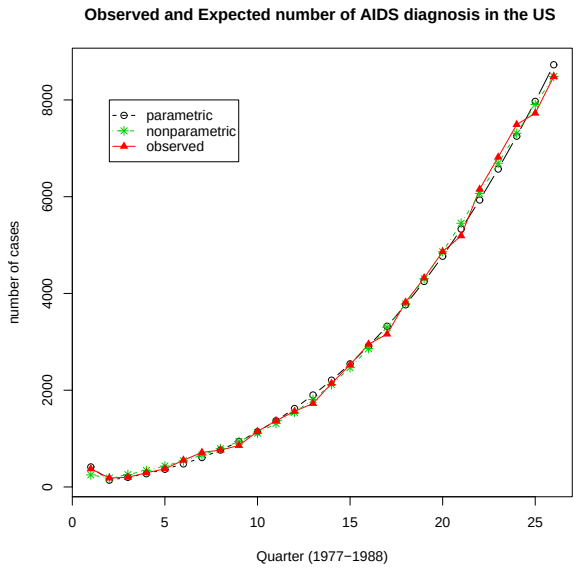

Figure 5.20. *Parametric and nonparametric estimates of AIDS incidence in America*

and 5.21 which show estimates of AIDS diagnosis incidence obtained from the 3 and 4-step function models and the various spline models.

Nonparametric back-projection has an explicit formulae for the generation of estimates of the infection intensity and this formulae is easily implemented using the EM algorithm. Under this method, the data is given greater power to determine the shape of the infection intensity. Hence, assumptions about the nature of HIV curve is not required. The method is a bit rigid but could be affected by the choice of the smoothing parameters. Also, convergence could be difficult for large values of AIDS or HIV counts.

Generally, estimates obtained using the parametric and nonparametric back-projection methods differ. This difference is larger in the estimates of HIV population. There are also differences in the estimates of the AIDS incidence as shown in Tables 5.39, 5.41 and 5.42

It seems the two methods adequately predict the AIDS prevalence as indicated

Year	Observed	Parametric	Nonparametric
1984	0	1	0
1985	3	2	1
1986	0	3	3
1987	6	4	5
1988	6	6	7
1989	15	8	9
1990	10	11	11
1991	11	13	14
1992	14	18	19
1993	19	23	26
1994	34	31	34
1995	45	43	41
1996	68	55	49
1997	63	68	57
1998	62	71	63
1999	59	57	69
2000	66	66	73
TOTAL	480	480	484

Table 5.41. *Estimates of AIDS cases in Hong Kong obtained using parametric and non-parametric back-projection*

by the totals in the three columns of Table 5.39 although little variations are noticed in the estimate of the incidence of AIDS cases diagnosis. The identical nature of the estimates obtained from the two methods are more prominent in the US and Nigeria epidemic situations. The graphs of the the observed and estimates of the AIDS incidence obtained using the two approaches indicate that the two methods produced similar estimates especially for the US and Nigeria. See Figures 5.12 and 5.22. However it appears that the parametric approach produced better estimates of AIDS incidence for Honk Kong as depicted in Figure 5.21.

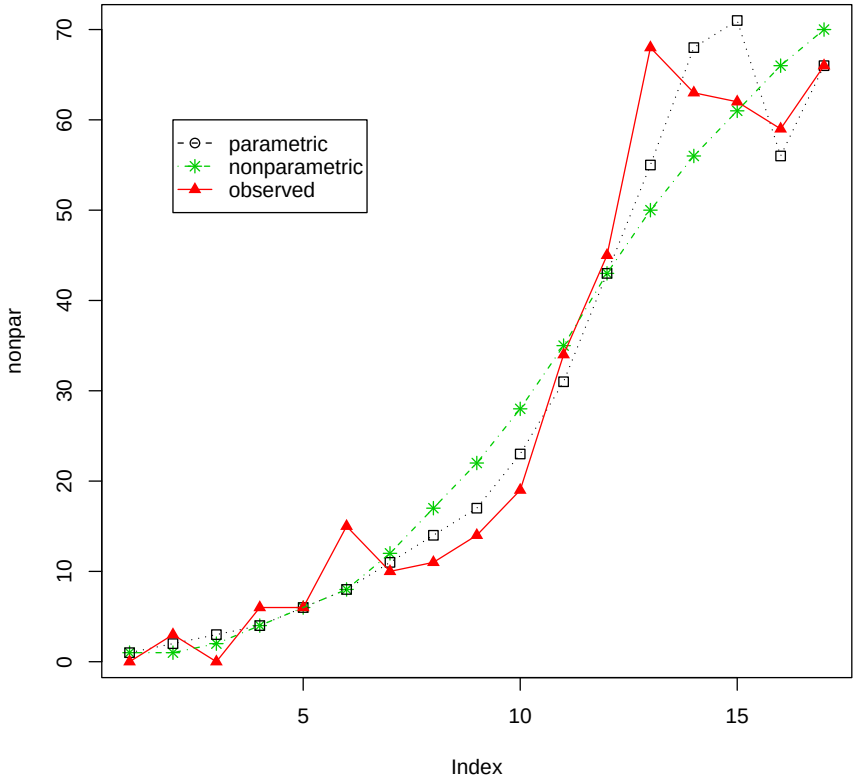

Figure 5.21. *Parametric and nonparametric estimates of AIDS diagnosis in Hong Kong*

Year	Observed	Parametric	Nonparametric
1989	8	89	20
1990	14	32	29
1991	160	57	62
1992	102	45	123
1993	198	220	236
1994	327	385	435
1995	625	693	768
1996	1381	1220	1286
1997	1940	2018	2034
1998	3219	3113	3047
1999	4342	4443	4336
TOTAL	12316	12315	12376
$\hat{N}$		152608	162724

Table 5.42. *Estimates of AIDS cases in Nigeria obtained using parametric and non-parametric back-projection*

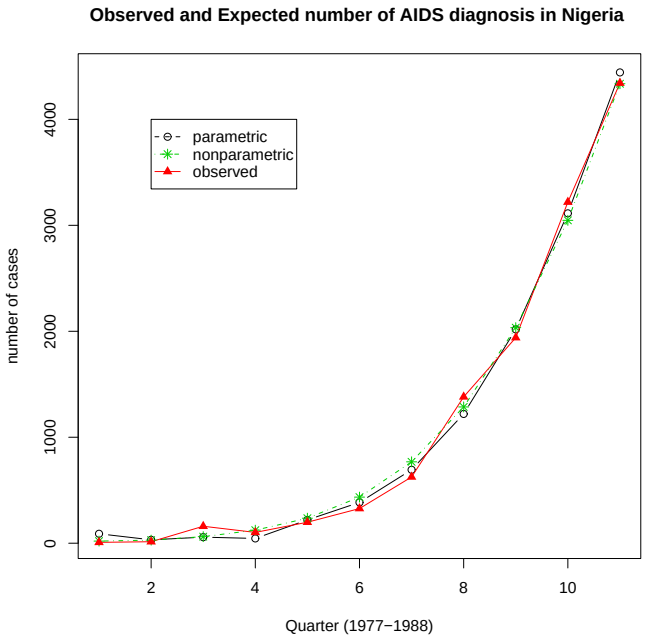

Figure 5.22. *Parametric and nonparametric estimates of AIDS diagnosis in Nigeria*

Chapter 6

Analysis of the HIV Screening Data

6.1 Descriptive Analysis

The data used in this chapter is HIV screening data collected from one of the Nigerian centres of excellence for HIV/AIDS research and treatment. It is a individual level data set comprising of all persons who presented themselves for HIV screening between October 2000 and August 2006. Information collected includes Date of screening, Result of the test(HIV positive or negative), Age, and Sex of the patient. In all, 33349 patients attended the HIV laboratory within the period under review. Of these, 7646 (about 23%) were confirmed positive for HIV.

A look at Table 6.1 and Figure 6.1 indicates that there is a marked increase in the number of patients screened from August 2005. The effect of policy change on HIV/AIDS screening and treatment is very evident in the data. Prior to August

	MALE			FEMALE		
Age	Oct 2000 July 2005	Aug 2005 Aug 2006	%change	Oct 2000 July 2005	Aug 2005 Aug 2006	%change
Below 2yrs	3.0	30.2	893.7	1.9	21.0	1007.7
2-14yrs	8.8	35.8	309.3	5.8	39.5	575.2
15-24yrs	14.4	83.8	483.8	28.4	189.5	567.1
25-34yrs	31.0	166.8	438.6	40.1	303.8	658.3
35-49yrs	22.8	138.8	510.1	20.6	169.7	722.9
50yrs& above	12.4	57.2	362.3	8.6	78.2	808.4
Total	5350	6664		6114	10423	

Table 6.1. *Average number of persons screened per month*

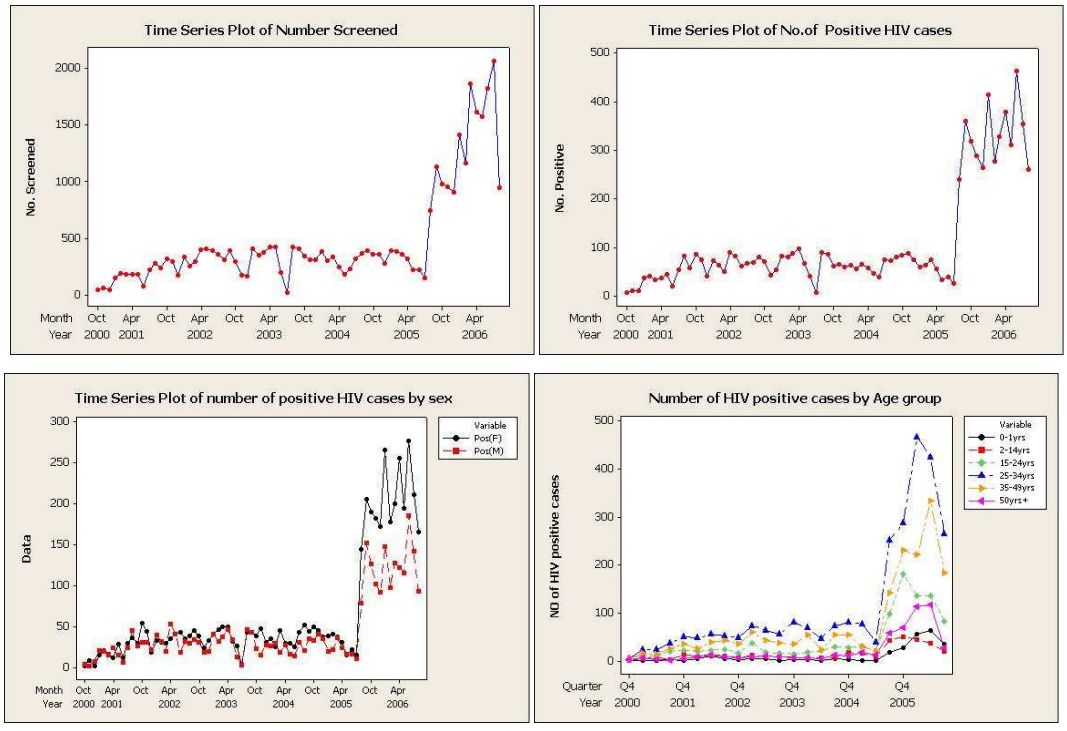

Figure 6.1. *Number of patients screened and number who tested positive for HIV by sex and age*

2005, patients paid a minimum fee of N10,000.00 (about £40) for a comprehensive HIV test (26). There is no doubt that, given the level of poverty in Nigeria, many patients would have been discouraged from presenting themselves for screening because of their inability to pay the bill. Also, testing kits available then were outdated and in most cases very slow such that, for some tests like CD4 count, only about three or four counts could be done per day. However in 2003, an international organization came to the aid of the Nigerian government, screening and treatment of HIV/AIDS became free of charge and state of the art screening equipment was installed in some selected centres of excellence. The surge in the number of patients seeking HIV/AIDS services is an indication of the extent to which poverty had hindered patients from utilizing hospital facilities. The free Voluntary Counselling, Testing and Treatment (VCTT) as adopted by these internationally sponsored centres, encouraged even the most poor and vulnerable to volunteer for HIV tests.

As can be seen from the plots, the difference between the total numbers of males and females screened prior to the intervention of the international organization is minimal. However, there is an outstanding difference between the two sexes in the current period of free test with the females clearly outnumbering the males. The burden of poverty falls more on females, it is therefore not surprising that more females than males attended when services became free.

When the data is partitioned by time and by sex and age (Table 6.1), a clearer picture of the effects of poverty on the sexes is presented. When the percentage change in attendance for the two time periods is considered by sex and age, the increase in average attendance per month for the males is between 300 to 900 percent, but for the females it ranges between 500 to 1007 percent. For all the age groups, the percentage change in monthly average attendance for the females

		MALE			FEMALE			All		
		No.		%	No.		%	No.		%
Time	Age(yrs)	+ve	Total	+ve	+ve	Total	+ve	+ve	Total	+ve
1	< 2	29	176	16.5	18	110	16.4	47	286	16.4
1	2-14	82	508	16.1	79	339	2.3	161	847	19.0
1	15-24	74	833	8.9	309	1648	18.7	383	2481	15.4
1	25-34	376	1796	20.9	635	2324	27.3	1011	4120	24.5
1	35-49	388	1320	29.4	259	1196	21.7	647	2516	25.7
1	≥ 50	76	717	10.6	78	499	15.6	154	1216	12.7
2	< 2	99	392	25.3	98	273	35.9	197	665	29.6
2	2-14	107	466	23.0	84	513	16.4	191	979	19.5
2	15-24	165	1090	15.1	482	2464	19.6	647	3554	18.2
2	25-34	521	2168	24.0	1167	3950	29.5	1688	6118	27.6
2	35-49	514	1805	28.5	600	2206	27.2	1114	4011	27.8
2	≥ 50	170	743	22.9	208	1016	20.5	378	1759	21.5

Table 6.2. *Age/Sex distribution of HIV positive cases. Time period 1 = Oct 2000- July 2005, time period 2= Aug 2005-Aug 2006*

is higher than for the males. Due to their higher financial status, men had been more capable of paying the screening and prescription bills than the women. The free screening and drugs made more women who hitherto could not afford the cost of these services present themselves for screening. The same trend was also noticed for the children aged below 2 years. In this part of the country, the male child is highly valued and parents would give all for the health of the child. This might explain why the proportion of the male children was higher than that of the females at time 1 (when treatment was not free).

It is worthy of note also that the increase was not only in the number of laboratory attendees. Table 6.2 shows that there is a noticeable increase in the number and proportion of HIV positive test results in time 2 when compared with time 1. This increase is greatest in the youngest and oldest age groups. Also the average number of positive diagnoses per month was 59 cases for time 1 and 327 cases for time 2, thus giving a monthly average rate of increase of 5.6 times the number of cases recorded per month in time period one. This gives an idea of

the quantum of underreporting of HIV/AIDS cases in the country before 2005.

Overall, the proportion of female patients who were HIV positive was higher than the males. For females, patients aged between 25-34 years were more affected by the virus. For the males, the proportion of patients aged between 35-49 years was slightly higher than those aged 25-34 years. This combined age group accounted for about 66 and 70 percent of all HIV positive results for the females and males respectively.

It also seems from the data that young adolescent females are more affected than their male counterparts. That is, the data suggest that more females were infected at younger ages than the males. For instance, about 20 percent of infected females were in age group 15-24 years but for the males, only 9 per cent of them were in that age group. This is also true for age group 25-34 years, the proportion of females infected is higher than the proportion of males infected. However, at older ages (35- 49, 50 and above), the proportion of males infected is higher. In all, the bulk of infection was found in patients aged between 15-49 years. For the females, about 86 per cent of the infected were in this age range and for the males; about 78 per cent of those infected belong to this age range. This age distribution of HIV positive patients has some policy implications in the design of outreach programs to fight the spread of the virus. The focus or target groups can be easily determined.

6.2 Formal Analysis - The Logistic Regression

In order to estimate the dependence of test outcome on the explanatory variables sex, age and time period of the test, logistic regression was applied to the

coefficients	Estimate	Std Error	z-value	p-value
intercept	-1.63	0.258	-6.33	2.45e-10
Time2	1.052	0.287	3.67	0.000248
Age2	0.44	0.288	1.529	0.126
Age3	0.165	0.265	0.622	0.534
Age4	0.65	0.262	2.49	0.0126
Age5	0.346	0.267	1.29	0.196
Age6	-0.055	0.286	-0.191	0.849
SexM	0.008	0.328	0.025	0.98
Time2*Age2	-1.49	0.336	-4.43	9.26e-06
Time2*Age3	-0.999	0.298	-3.351	0.000805
Time2*Age4	-0.943	0.293	-3.22	0.00128
Time2*Age5	-0.75	0.299	-2.507	0.0122
Time2*Age6	-0.723	0.322	-2.245	0.0247
Time2*SexM	-0.513	0.370	-1.387	0.166
Age2*SexM	-0.465	0.373	-1.248	0.212
Age3*SexM	-0.87	0.356	-2.446	0.01446
Age4*SexM	-0.3588	0.3365	-1.066	0.286
Age5*SexM	0.4013	0.341	1.177	0.239
Age6*SexM	-0.455	0.371	-1.226	0.2204
Time2*age2*sexM	1.390	0.441	3.152	0.00162
Time2*age3*sexM	1.065	0.407	2.617	0.00887
Time2*age4*sexM	0.583	0.383	1.522	0.128
Time2*age5*sexM	0.168	0.388	0.432	0.666
Time2*age6*sexM	1.102	0.425	2.592	0.0095

Table 6.3. *Parameter estimates for logistic regression model- all data*

data using $R^{©}$. The full model is given in equation 6.1 where x_i $(i = 1, 2, 3)$ are age, sex and time period respectively. The output shown in Table 6.3 was obtained. The analysis show that the model is significant with a change in deviance of 555 on 23 degrees of freedom and $p < 0.001$.

$$\log(\frac{p_i}{1-p_i}) = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_1x_2+\beta_5x_1x_3+\beta_6x_2x_3+\beta_7x_1x_2x_3 \quad (6.1)$$

Table 6.3 show that 3-way interactions are statistically significant, indicating that differences in the proportion of HIV positive cases between age and sex is

coefficients	Estimate	Std Error	z-value	p-value
intercept	-1.63	0.258	-6.33	2.45e-10
Age2	0.44	0.288	1.529	0.126
Age3	0.165	0.265	0.622	0.534
Age4	0.65	0.262	2.49	0.0126
Age5	0.346	0.267	1.29	0.196
Age6	-0.055	0.286	-0.191	0.849
SexM	0.008	0.328	0.025	0.98
Age2*SexM	-0.465	0.373	-1.248	0.212
Age3*SexM	-0.87	0.356	-2.446	0.01446
Age4*SexM	-0.3588	0.3365	-1.066	0.286
Age5*SexM	0.4013	0.341	1.177	0.239
Age6*SexM	-0.455	0.371	-1.226	0.2204

Table 6.4. *Parameter estimates for logistic regression model- Oct 2000 till July 2005 only*

different in the two time periods. For parsimony, we tested the significance of the 3-way interactions alone by fitting a model without a 3-way interactions and comparing it with the full model. Result gave a change in deviance of 40.90 on 5 degrees of freedom with $p < 0.0001$, indicating that the model with 3-way interaction is better than that with just 2-way interactions.

To further investigate the interaction between age and sex without the influence of time, we partitioned the data by time period. Table 6.4 and 6.5 show results when the data was partitioned into the two time periods, representing the period when screening was paid for (Time 1) and the period when it is free (Time 2). The result of the analysis is shown in Table 6.4. So the age/sex interaction is significant even in the time when patients were expected to foot the medical bills. Carrying out the equivalent test using data for Time period 2, we obtain the result in Table 6.5.

The age/sex interaction model is a better fit to the data than the simple model in the two time periods. It seems therefore, that the variations in the proportion of patients diagnosed as being HIV positive is greatly influenced by the sex and

coefficients	Estimate	Std Error	z-value	p-value
intercept	-0.58	0.126	-4.596	4.31e-06
Age2	-1.0508	0.174	-6.051	1.44e-09
Age3	-0.834	0.136	-6.133	8.63e-10
Age4	-0.29	0.131	-2.213	0.0269
Age5	0.405	0.135	-3.0	0.0027
Age6	-0.777	0.148	-5.244	1.57e-07
SexM	-0.505	0.172	-2.945	0.00323
Age2*SexM	0.925	0.236	3.917	8.95e-05
Age3*SexM	0.195	0.198	0.987	0.324
Age4*SexM	0.224	0.182	1.228	0.219
Age5*SexM	0.59	0.186	3.065	0.00218
Age6*SexM	0.6472	0.2076	3.117	0.00183

Table 6.5. *Parameter estimates for logistic regression model- Aug 2005 till Aug 2006 only*

age of the patients. Hence, differences between the proportion of HIV positive males and that of females is different for each age group.

Conclusion

There is a marked increase in the number of persons seeking HIV advice and services as a result of the free counseling, test and treatment introduced in the hospital in 2005. This increase was more for women and children. Also, there was a noticeable increase in the number and proportion of diagnosed HIV positive cases. Some age groups had an increase of up to 10 times what it was at time period 1. The effect of poverty on the diagnosis and treatment of HIV/AIDS in the country is very clear in this data as depicted by the sharp increase in the number of patients when treatment became free. It may not be out of place to argue that any decision or estimation based on the information on HIV/AIDS obtained prior to 2005 may be misleading. This is because the data on which such decision or estimation was based is an under representation of what was obtainable. Under diagnosis and underreporting of cases were very prominent at

that time.

The data suggest that the proportion of females who tested positive for the virus is higher than the males. The females constitute about 61 percent of the total number of positive cases, the remaining 39 percent were males. Also the females were infected more at younger ages than males. The bulk of infection was found in patients aged 15-49 years. This age group accounted for about 86 percent of all cases in the females and about 78 percent of the males. For the females, patients aged 25-34 years were more affected by the virus. While males aged 35-49 years were slightly more affected than males aged 25-34. This seems to imply that older males infect younger females.

A formal analysis of the data using the logistic regression model indicates that there is a 3way interaction between time, age and sex. A search for a simpler, less complex model was done but in all, it seems that the 3way interaction model is the best. The data was further partitioned by time to remove the effect of time and test the 2way interaction between sex and age for each time period. This analysis confirmed that the 2way interaction in both time periods 1 and 2 was significant. Hence, the probability of testing positive to HIV infection depends on the sex and age of the patient.

6.3 Correction for underreporting

It became clear from the review of available data on HIV/AIDS in Nigeria that there were serious problems of underreporting of cases (especially AIDS cases). Some of the factors identified as contributing to underreporting include:

- Many patients may prefer alternative medicine. Most patients see the disease as art of witchcraft, poison from an enemy, or work of evil spirits. Hence they opt for traditional or spiritual healers
- Health practitioners may not record cases so as to protect patients from stigma
- It is not mandatory for private laboratories/hospitals to report cases. Most Nigerians prefer private health care for reasons ranging from efficiency to prompt attention.
- No effective central statistical coordination of available data
- Lack of Education
- Poverty which makes it difficult or impossible for patients to seek medical help
- Patients may die of other diseases before they are diagnosed of HIV/AIDS

Due to lack of information on the reporting delay and underreporting distribution of HIV/AIDS cases in Nigeria, it becomes impossible to apply the methods proposed by (254) and (232) where stationary and non-stationary probabilities of reporting delay were used to correct the data.

We note that while most developed countries contend with the issue of reporting delay, the developing countries like Nigeria, are faced with the crises of huge underreporting and underdiagnosis of cases, where a high proportion of HIV/AIDS cases were never reported due to the reasons mentioned above.

In the previous chapter,we analyzed the data without correcting for underreporting. Estimates of the infection curve were obtained using the AIDS and HIV diagnosis data as published by the Nigeria Institute of Medical Research. Based on the knowledge gained from the cases-by-case data analysis above, it becomes expedient that the data be adjusted for underreporting. In order to correct the published data for underreporting, we need further information on the extent of underreporting within each time period.

Analysis in the previous section gave us an insight into the extent of underreporting in Nigeria. This perceived rate of underreporting could be argued to be solely due to poverty, as there was a drastic increase in the number of attendance to the HIV laboratory when diagnosis and treatment became free in August 2005. We are also aware that some individuals may not utilize the free services offered by the hospital due to lack of information of its availability, lack of education, lack of transport fare (as many have to travel long distances to access the facility) and fear of being stigmatized. We are also aware that the surge in the number of patients seeking HIV services in this laboratory could mean that patients outside its catchment areas may be attracted by its free services. However, there are little evidence to support this.

The ratio of the mean monthly diagnosis of HIV positive cases between the two time periods was adopted as the correction factor based on assumptions that the observed rate of underreporting is constant over the time periods and that this rate of underreporting is the same for all geopolitical regions in the country.

It may be right to argue that this correction factor only took care of under-reporting due to poverty and not encompassing other factors which are known to affect HIV/AIDS data reporting. Since it is not easy to quantify the extent of the effects of these other factors on HIV/AIDS underreporting in Nigeria, it becomes pertinent to rely on the only available information. We also note that the information we have on underreporting in our data is on HIV diagnosis and not purely on AIDS incidence, our first attraction was to apply this correction factor only to HIV test data. However, since a large proportion of the tests were symptom induced, we decided also to extend the correction fraction on AIDS data. We believe that the underreporting of AIDS incidence is higher than that of HIV in Nigeria but we shall use this correction factor as a proxy for the rate of AIDS cases underreporting.

The mean number of HIV positive diagnosis for time periods 1 and 2 were estimated to be 58.6 and 326.7 respectively, the average rate of increase per month therefore is 5.575. Correcting the data by this rate, the new estimates of HIV incidence were obtained using the original (when AIDS data is used) and the modified (HIV data is used) back-projection models as shown in Tables 6.6 and 6.7 respectively.

With this correction, it is estimated that a cumulative number of about 907208 persons were infected with HIV in Nigeria as at December 1999. This point estimate of the population number of HIV infected persons is expected to lie somewhere between 261157 and 2566703. The width of this confidence interval is a signal to the possible imprecision of predicting HIV prevalence using AIDS diagnosis data in back-projection. Comparing these results with the uncorrected estimates in Table **??**, there is an appreciable difference in the two estimates.

Year	Lower	Estimate	Upper
1980	0	1	119
1981	0	2	134
1982	0	7	272
1983	0	32	539
1984	0	117	983
1985	3	348	1660
1986	20	867	2747
1987	119	1845	4484
1988	535	3487	6968
1989	1693	6186	11911
1990	4125	10958	19316
1991	9707	20091	33025
1992	19778	36989	56440
1993	32854	62462	86458
1994	49448	92520	124642
1995	56908	118732	174084
1996	40893	135072	263132
1997	23302	141501	437921
1998	11092	138505	654779
1999	10680	137486	687089
Total	261157	907208	2566703

Table 6.6. *HIV incidence estimated original back-projection model (using AIDS diagnosis) and the 90% confidence interval*

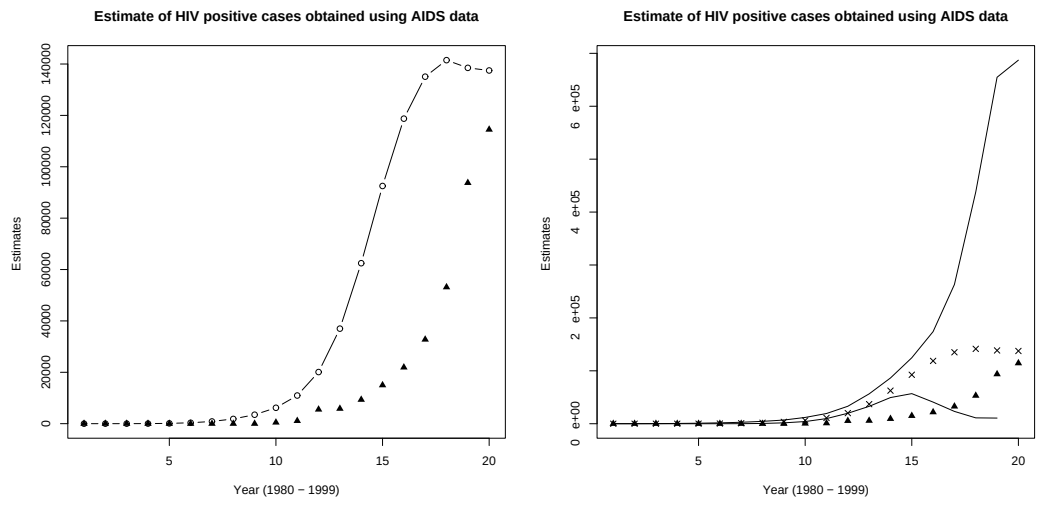

Figure 6.2. *HIV incidence estimated using AIDS data*

Year	Lower	Estimate	Upper
1980	0	0	92
1981	0	0	100
1982	0	1	244
1983	1	7	547
1984	8	43	1071
1985	52	221	2332
1986	276	928	4444
1987	1159	3192	8209
1988	3607	8837	14789
1989	8859	19198	25057
1990	17806	32831	41072
1991	31525	45863	62427
1992	50923	61564	94646
1993	80104	92532	142871
1994	127636	147525	212195
1995	197167	241386	306162
1996	293920	388182	441838
1997	366072	552564	596577
1998	362088	610522	762214
1999	337052	556738	810986
Total	187255	2762134	3527873

Table 6.7. *HIV incidence estimated from the modified back-projection model (using HIV diagnosis) and the 90% confidence interval*

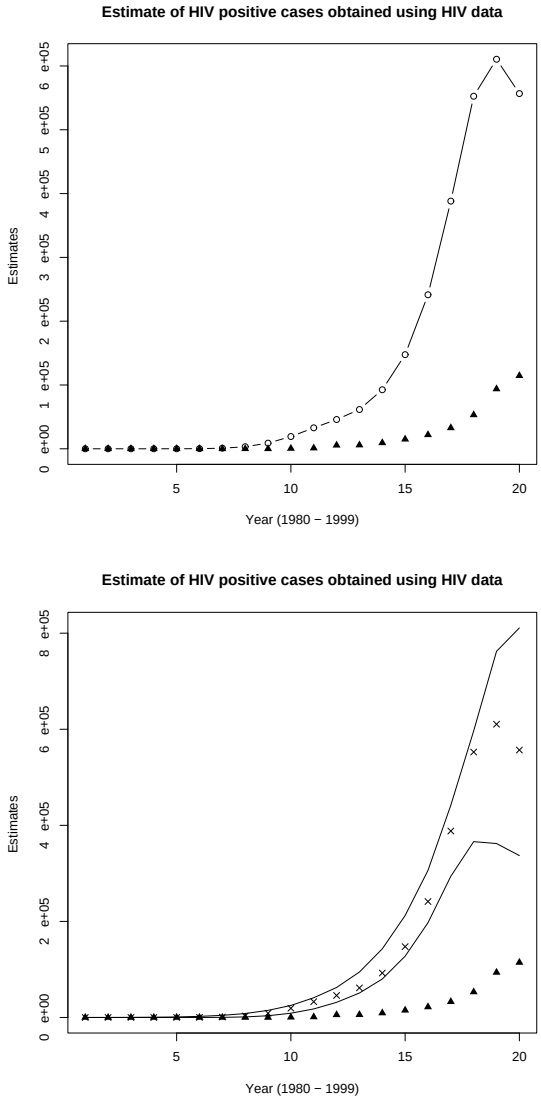

Figure 6.3. *HIV incidence estimated using HIV diagnosis*

The HIV prevalence estimate obtained using HIV incidence data is far higher than that obtained using AIDS data. Here, it is estimated that about 2762134 persons were living with HIV/AIDS. It is noteworthy that the 90 per cent confidence interval obtained using the HIV data also encompass that obtained using AIDS data.

This estimate closely approximate that obtained for Nigeria by the United Nations Joint Action Committee on AIDS (UNAIDS) in 1999. UNAIDS estimated

β	3 Steps	4 Steps
β_0	196.50(9.61)	211.99(10.01)
β_1	21798.09(289.62)	20594.66(361.08)
β_2	125063.18(1217.08)	142906.92(3832.86)
β_3		90345.04(7433.16)

Table 6.8. *Parameter estimates and their standard error obtained using the parametric (step function) back-projection*

that as at December 1999, about 2700000 persons were living with HIV/AIDS in Nigeria (Time magazine, Feb. 12, 2001). The method and data used by the UN were quite different from the one used here.

Applying the corrected data in the parametric back-projection under the assumption that the infection curves follow a step fuction (see section 5.3.2), it still appears that the three-step model is more efficient than the four-step model given the standard error of the estimates of the parameters as shown in Table 6.8. The two models gave different estimates of the HIV/AIDS population. With the three-step model, an estimate of 839536 was obtained and the four-step model gave an estimate of 731692. See Table 6.10

Year	Observed	3 Steps	4 Steps
1989	45	272	293
1990	78	111	118
1991	892	249	250
1992	569	591	576
1993	1104	1192	1144
1994	1823	2137	2053
1995	3484	3886	3844
1996	7699	6880	7015
1997	10816	11329	11593
1998	17946	17297	17434
1999	24207	24718	24341
Total	68663	68663	68663

Table 6.9. *AIDS incidence estimates obtained using the parametric (step function) back-projection*

Time	3 Steps	Time	4 Steps
Jan 1980-Jan 1989	1965	Jan 1980-Jan 1989	2120
Jan 1989-Jan 1993	87192	Jan 1989-Jan 1993	82379
Jan 1993-Dec 1999	750379	Jan 1993-Jan 1995	285814
		Jan 1995-Dec 1999	361380
Total	839536	Total	731692

Table 6.10. *Estimates of number of persons living with HIV/AIDS obtained using the parametric (step function) back-projection*

Chapter 7

Discussion and Conclusion

7.1 Summary

The aim of this thesis was to develop epidemic models that could describe and predict the HIV/AIDS epidemic in Nigeria. To achieve this, we focused on two broad approaches, namely, spatial epidemiology and backcalculation methods. The choice of these methods was informed by the nature of available data. After a careful review of all available sources of data on HIV/AIDS, two sets of data, collected from two different sources were adjudged better than others based on the criteria of national coverage and minimum reporting delay. The data adopted for this research were the outcome of the survey of 1057 health and laboratory facilities (public and private) conducted by the Nigerian Institute of Medical Research (NIMR) in 2000 and the outcome of the National HIV/AIDS Sentinel Surveillance Survey conducted biannually by the Federal Ministry of Health biannually between 1991 and 2005.

A review of the literature reveals that there exist wide spread networks of

premarital and extramarital and other risky sexual practices capable of sustaining the HIV/AIDS epidemic in Nigeria. The nature and extent of these practices vary across the six geopolitical zones of the country. Some communities have socio-cultural and religious affinity which tend to have direct influence on their sexual behaviours and, consequently, on the prevalence of HIV/AIDS in these communities. Cluster analysis, using data from the HIV Sentinel Surveillance which surveyed 85 selected sites (communities), reveal clustering or spatial patterning of HIV prevalence rates among the sites and States. The nature of the patterning seems to change with time which may reflect distinct phases of the epidemic and a tendency for the prevalence rates to become increasingly similar or dissimilar with time. Analysis suggests that the sites are broadly clustered within their geopolitical zones such that each geopolitical zone appears to be a distinct cluster.

The fit of two-way GAM and sm-regression models to the data identified spatial trends in the east-west and north-south directions. The trends indicate low prevalence in the west and far north, moderately high prevalence in the east and south-south and very high prevalence in the centre (North-central). Some of the towns in the North-central zone are known for the practice of spouse sharing and the giving of wife or daughter to distinguished guests as a welcome gift. Condom use is also lowest in this zone.

The correlograms depict the highest positive spatial autocorrelation among sites and States geographically close (less than 100miles) to each other, suggesting that the epidemic is similar among States or sites that are in close proximity. This correlation also tends to increase with time. Thus, prevalence of HIV infection gets more and more similar among States and sites that are close neighbours as the years of the epidemic advanced. The Moran scatter plots, Moran and Geary

coefficients corroborated these findings of significant positive autocorrelation that tends to get stronger with time. Therefore, this confirms that nearby states or sites, on average, have similar prevalence rates and the number of States or sites forming these clusters appear to increase with time. The analysis of outliers indicates that the number of outliers in the semivariogram cloud decreased with time signifying increased similarity in the HIV prevalence levels among States and sites. The semivariogram analysis supports this decline in the overall spatial variation over time as indicated by the reduction in the size of the sill as the epidemic advanced in years. The neighbourhood effects, among the States, varied from 340miles in 1999, 580miles in 2001 and 279miles in 2005. Kriging estimates clearly show these spatial patterning and trend with some local focal points found in the south-south and north central zones.

A literature search implicated some ecological factors as being of importance in explaining variation in HIV prevalence in Nigeria. Data on some of these risk factors were obtained. However, these data were only available at the zone level while the prevalence data were at the site level. The effect of the factors on the geographical variation in the risk of HIV infection was established by fitting multilevel variance component models and examining the fixed parameter estimates. Of the eight ecological factors for which data were available, four of them were consistently found to be highly significant in explaining variation in the relative risk of HIV infection. The factors are age at sex debut, polygamy, frequency of sex and condom use. Polygamy, which may be seen as a proxy measure of concurrent multiple sex partnership, has the greatest effect on the risk of exposure to HIV infection. The next in line is frequency of exposure to heterosexual contact. Condom use is negatively associated with the risk of HIV infection. The spatial effect on the distribution of the risk of HIV infection was studied by fitting spatial

multilevel models and examining the random parameter estimates. Variations in the risk of HIV infection within States and within zones was prominent in the models. Accounting for the ecological and spatial effects, as previously discussed, significantly reduced the random variability in HIV prevalence across the zones. Thus, relative risk estimates obtained using these models are relatively precise and are expected to give an accurate map of the distribution of HIV prevalence in Nigeria.

In order to investigate the temporal development of the epidemic, various forms of back-projection methods were applied to the Nigerian HIV/AIDS historical data. Broadly, the methods used can be grouped as parametric and nonparametric. In order to apply the parametric methods, particular functional forms had to be assumed for the HIV intensity curve. Specifically, we assumed that the family of infection curves could be defined by a basis function which could either be an indicator function or a spline function. Estimates were obtained under each assumed intensity curve using three methods, namely multinomial likelihood, Poisson regression and quasi-likelihood methods. Step function model estimates of HIV infection intensity obtained using the Poisson regression and quasi-likelihood methods are almost identical and differ only slightly from those obtained using the multinomial likelihood method. Estimates of the AIDS incidence obtained using different spline function models appear to be better fits than those obtained using the step function models judging by their lower residual variances measured by the chi squares. However, the estimates of the cumulative HIV infection deffer substantially depending on the number and positions of the steps and knots.

In order to avoid the problem of specifying the parametric form of the infection curve, we adopted nonparametric back-projection methods. These allow the data

more power to determine the shape of the estimated intensity curve. Two approaches were undertaken: a nonparametric method where AIDS incidence data was used to reconstruct the HIV incidence curve and a modified nonparametric form that made use of the HIV diagnosis data to reconstruct the HIV incidence curve. The use of the latter method is informed by the fact that HIV incidence data contains more information about the HIV incidence curve, has a shorter incubation period and is not affected by the treatment regimes. Estimates of the cumulative number of HIV infection differ widely in the two methods. This estimate is about three times higher when HIV diagnosis data is used than when AIDS incidence data is used. In both cases, a nonparametric bootstrap pointwise confidence interval was constructed to quantify the precision of the estimates. The imprecision of the back-projection methods for estimating or predicting the most recent incidence of AIDS or HIV is clear in the two methods but greater when AIDS data was used.

While data used for the back-projection models may be assumed to be free from reporting delays, given the fact that they were collected retrospectively, it is not free from underreporting. In order to study the extent of under reporting, we collected case-by-case data from the HIV laboratory of a centre of excellence for the treatment of HIV/AIDS. The data covered two time periods, a period when HIV screening services were paid for by the patients followed by a period when services were free. Analysis indicated sharp increases in the number of patients seeking the laboratory services, and consequently in the number of diagnosed HIV positive cases, once the services became free. More females than males turned up for the free HIV services. Also, increase in attendance varied by age group. This suggests that the most poor and vulnerable who hitherto could not afford the bills, were given a life line by the free services. On average, the number of HIV

positive diagnoses in the free-test period was more than five times greater than it was when screening was paid for. Formal analysis of data using logistic regression indicated that the probability of being diagnosed as HIV positive depends on main effects and interactions between the patient's age, sex and time period. Based on some simplistic assumptions, the historical data used in the back-projection models was corrected for underreporting using a correction factor estimated by comparing the mean HIV diagnosis in the two time periods. The modified non-parametric Back-projected estimate of cumulative number of HIV infections (N = 2,762,134) obtained using the corrected data are comparable with that obtained by the UNAIDS for Nigeria in 1999 (N= 2,700,000) using other approaches.

7.2 Limitations of the research

Restricted availability of data on HIV/AIDS in Nigeria is a serious limitation on this research. The nature of available data, to a great extent, tailored the direction of this thesis. Data used in the construction of back-projection models do not cover the period of the inception of the epidemic and the recent years. The data were collected on all HIV/AIDS diagnosed between 1989 and 1999, thus omitting the first reported cases in 1986 and data up to 1988. Also, there was no information about the epidemic in the recent years (2000 - 2008). The aggregation of the data at national level made it impossible to conduct more detailed analysis. The study and modelling of the trend of the epidemic across the various demographic strata of the Nigerian society was hindered by the non-stratification of the data. It is strongly recommended that data be placed on the public domain after striping all patient identities and made more accessible to researchers. Aggregation of data should be avoided as much as possible, at

the least, data should be published by sex and age for each of the 36 states and Abuja.

The ecological covariates used in the construction of the multilevel models were also aggregated at the zone level. This high level aggregation results not only in loss of information but also, conclusions based on the data run the risk of being affected by the ecological fallacy. It is important to note here that our conclusions under the multilevel models relate only to populations at the zone level. We strongly recommend that data arising from surveys should be published at the level of Enumeration Areas or towns in which they were collected. At best, they could be left at the individual level.

The back-projection models adopted an incubation period distribution estimated from other studies. It is uncertain to what extent this distribution represents the Nigerian epidemic scenario. The estimates from the parametric back-projection models depend on the positions and number of steps and knots. Therefore, these methods are capable of generating infinitely many different estimates. The methods can even yield negative estimates unless restrictions are placed on the parameters. The choice of the smoothing constants in the non-parametric back-projection also may affect the estimates. Due to the long incubation period, the method lends itself to only short-term projection and predictions of even the most recent infections is highly imprecise.

7.3 Further work

Subject to data availability, it may be worthwhile to construct and fit stratified models describing the behaviour of the epidemic by gender and age. The reason

behind this thinking is that the risk of HIV infection or the infection intensity vary by sex and age as established in chapter six. Analysis suggests also that HIV prevalence differs by State and geopolitical zone, therefore, models representing the infection curve for each of the States or geopolitical zones can be revealing.

Ecological analysis may yield better results if information on the ecological covariates were available at the lower levels of aggregation. This is because data at individual level contain much more information than zone aggregated data. Further research may seek to evaluate the model using data at lower hierarchical levels. Also, it my be possible to study the changes in the trend of the epidemic over time by fitting a spatiotemporal model that estimates the time effect on HIV prevalence rates in different parts of the country.

The uncertainty surrounding the incubation period distribution, the infection curve, smoothing constants and data sources suggests the need to explore methods that allow for additional information known about the epidemic to be entered into the model. Bayesian approaches (58)(176)(216) might be helpful in this regard as they allow for formal treatment of uncertainty and inclusion of extra information.

Sexual networks models (59) (213) (170) (174) may be developed using some survey outcomes for some towns in the country. To do this effectively, further information on the size of each category of the infective, at-risk and not-at-risk populations; the extent of mixing between the susceptible and the infective, the partner exchange rate, concurrent partners, probability of infection per sexual contact and other behaviour change parameters.

Bibliography

[1] Federal Office of Statistics. Nigeria Demographic and Health Survey. 1990.

[2] National Population Commission (NPC). National Report on 2006 National population Census. 2006.

[3] NgEX! Ministry of health alerts Nigerians to the transfusion of unsafe blood in hospitals. *http://www.ngex.com/*, 2008.

[4] Gelman A and Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511, 1992.

[5] Gelman A and Rubin DB. *A single series from the Gibbs sampler provides a false sense of security.* In Bayesian Statistics : Bernado JM, Berger JO, Dawid AP and Smith, eds. Oxford University Press, 1992.

[6] Gigli A and Verdecchia A. Uncertainty of AIDS incubation time and its eefects on back-calculation estimates. *Statistics in Medicine*, 19:175–189, 2000.

[7] Griffith D. A. and Layne L. J. *A casebook for spatial statistical data analysis. A compilation of analysis of different thematic data sets.* Oxford University Press, New York, 1999.

[8] Hegarty A and Barry D. Bayesian disease mapping using product partition models. *Statistics in Medicine*, 27:3868–3893, 2008.

[9] Hessol N A, Lifson, O'Malley P M, and et al. Prevalence, incidence and progression of human immunodeficiency virus infectionin homosexual and bisexual men in hepatitis B vaccine trials 1978-1988. *American Journal of Epidemiology*, 130:1167–1175, 1989.

[10] Lawson A, Bohning D, Biggeri A, Lesaffre, and Veil JF. *Disease Mapping and Its Uses. In: Disease Mapping and Risk Assessment for Public Health.* John Willey and Sons Ltd, 1999.

[11] Lawson A, Browne W, and Vidal Rodeiro C L. *Disease Mapping with WinBUGS and MLwiN.* West Sussex: John Wiley and Sons Ltd, 2003.

[12] Mollié A. *Bayesian Mapping of Disease. In: W Gilks, N Richardson and D J Spiegelhalter, editors. Markov Chain Monte Carlo in Practice* . Chapman and Hall, London, 1996.

[13] Mollié A. *Bayesian and Empirical Bayes Approaches to Disease Mapping. In: Lawson A, Editor. Disease Mapping and Risk Assessment for Public Health* . John Willey and Sons Ltd, 1999.

[14] Munoz A. and Hoover D. R. The role of cohort studies in evaluating AIDS therapies. *In AIDS Clinical Trials (eds Finkelstein D. M. and Schenfeld D. A.). New York: Wiley*, pages 423–446, 1995.

[15] Munoz A., Wang M. C.and Bass S, Taylor J M G, Kingsley L.A., Chmiel J. S, and Polk B. F. Acquired immunodeficiency syndrome (AIDS)-free time after Human Immunodeficiency Virus Type 1 (HIV-1) seroconversion in homosexual men . *Journal of American Epidemiology*, 130:530–539, 1989.

[16] Nasidi A, Harry T O, Ajose-Coker O O, and et al. Evidence of LAV/HTLV III infection and AIDS related complex in Lagos. Nigeria. *II international Conference on AIDS, Paris France, June 23-25*, FR86–3, 1986.

[17] Nasidi A and Harry TO. The Epidemiology of HIV/AIDS in Nigeria: In AIDS in Nigeria: A nation in the Threshold. *Harvard Centre for population and Development Studies*, 2006.

[18] Osho A and Olanyinka BA. Sexual practices conducive to HIV transmission in Southwest Nigeria . *The continuing African HIV/AIDS epidemic*, pages 85–91, 1999.

[19] Sabin C A, Lee C A, and Philips A N. The use of Backcalculation to estimate the prevalence of severe immunodeficiency induced by HIV in England . *Royal Statistical Society*, 157(1), 1994.

[20] Thielemans A, Hopke P. K., De Quint P, Depoorter A. M., Thiers G, and Massart D. Investigation of the Geographical Distribution of Female Cancer Patterns in Belgiun using Pattern Recognition Techniques. *International Journal of Geography*, 17(4):724–731, 1988.

[21] Velayati A, Bakayev V, Bahadori M, Tabatabaei S, Alaei A, Farahbood A, and Masjedi M. Religious and Cultural Traits in HIV/AIDS Epidemics in Sub-Saharan Africa. *Archive of Iranian Medicine*, 10(4):486–497, 2007.

[22] Whiteside A. Poverty and HIV/AIDS in Africa. *Third World Quarterly*, 23(2):313–332, 2002.

[23] Lawson AB. Disease map reconstruction. *Statistics in Medicine*, 20(14):2183–2204, 2001.

[24] Lawson AB, Briggeri AB, Boehning D, Lesaffre E, Veil JF, Clark A, and et al. Disease mapping models: an empirical evaluation. *Statistics in Medicine*, 19(17-18):2217–2241, 2000.

[25] Etokidem AJ. HIV/AIDS transmission through Female Genital Cutting: a case report. *International Conference on AIDS*, 15:D10677, 2004.

[26] Adeneye AK, Adewole AZ, Musa D, Onwujekwe NN, Odunukw ID, Araoyinbo TA, Gbajabiamila PM, Ezeobi EO, and Idigbe EO. Limitations to access and use of Antiretroviral Therapy (ART) among HIV positive persons in Lagos, Nigeria . *World Health and Population*, pages 1–11, 2006.

[27] Dempster AP, Laird NM, and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Series B*, 39:1–38, 1977.

[28] Sagay AS, Kapiga SH, Imade GE, Sankale JL, Idoko J, and Kanki P. HIV infection among pregnant women in Nigeria. *International Journal of Gyncology and Obsterics*, 90:61–67, 2005.

[29] AVERT. HIV/AIDS in Nigeria. *http://www.avert.org/aids-nigeria.htm*, 2008.

[30] Efron B and Morris C. Stein's estimation rules and its competitors - an empirical Bayes approach . *Journal of American Statistical Association*, 68(65):117–130, 1973.

[31] Efron B and Morris C. Data analysis using Stein's estimation and its generalization . *Journal of American Statistical Association*, 70:311–319, 1973.

[32] Efron B and Morris C. Stein's paradox in Statistics . *Scient. Am.*, 236:119–127, 1976.

[33] Feyisetan B and Pebley AR. Premarital Sexuality in Urban Nigeria. *Studies in Family Planning*, 20(6):343–354, 1989.

[34] Grenfell B and harwood J. Metapopulation dynamics of infectious diseases. *Trend in Ecological Evolution*, 12:395–399, 1997.

[35] Mariotto A. B. and Verdecchia A. Using AIDS data to reconstruct HIV/AIDS epidemics. *Statistics in Medicine*, 19:161–174, 2000.

[36] Ripley B. *Spatial Statistics*. Wiley, New York, 1981.

[37] Williams B, Gouws E, Wilkinson D, and Karim S.K. Estimating HIV incidence rates from age prevalence data in epidemic situation. *Statistics in Medicine*, 20:2003–2016, 2001.

[38] Oloko BA and Omoboye AO. Sexual networking among some Lagos state adolescent Yoruba students. *Health Transition Review*, 3(2), 1993.

[39] Naanen BBB. Itinerant Gold Mines: Prostitution in the Cross River Basin of Nigeria, 1930- 1950. *African Studies Review*, 34:57–79, 1991.

[40] BBC. Alcohol increases HIV risk. *http://newsvote.bbc.co.uk/mpapps*, 2003.

[41] Leroux BG, Lei X, and Breslow N. *Estimation of Disease rates in small areas: A new mixed model for spatial dependence. In Statistical Models in Epidemiology, the Environment and Clinical Trials.* Halloran ME, Berry D. eds. Springer New York, 1999.

[42] Williams BG, Lloyd-Smith JO, Gouws E, Hankins C, and et al. The Potential Impact of Male Circuncision on HIV in Sub-Saharan. *PLoS Medicine*, 3:1032–1040, 2006.

[43] Carlin BP and Gelfand AE. A sample reuse method for acurate parametric Bayes confidence intervals. *Journal of the Royal Statistical Society, Series B*, 53:189–200, 1991.

[44] Deuffic burban S and Costagliola D. Including pre-AIDS mortality in back-calculation model to estimate HIV prevalence in France, 2000. *European Journal of Epidemiology*, 21:389–396, 2006.

[45] Darby S C, Doll R, Thakrar B, and et al. Time from infection with HIV to onset of AIDS in patients with hemophilia in the UK . *Statistics in Medicine*, 9:681–689, 1990.

[46] Marschner I. C. and Bosch R. J. Flexible assessment of trend in age-specific HIV incidence using two-dimensional penalized likelihood . *Statistics in Medicine*, 17:1017–1031, 1998.

[47] Panchaud C, Woog V, Singh S, Darroch JE, and Bankole A. Issues in Measuring HIV Prevalence: The case of Nigeria . *African Journal of Reproductive Health*, 6:11–29, 2002.

[48] Pascutto C, Wakefield JC, Best NG, Richardson S, Bernardinelli L, Staines A, and et al. Statistical issues in the analysis of disease mapping data. *Statistics in Medicine*, 19(17-18):2493–2519, 2000.

[49] Zeger S L; See L C; and Diggle PJ. Statistical Methods for monitoring the AIDS epidemic. *Statistics in Medicine*, 8:3–21, 1989.

[50] Centers for Disease Control. Estimates of HIV prevalence and projected AIDS cases: Summary of a workshop Oct 31 - Nov 1, 1989. *Morbidity and Mortality Weekly Report*, 39:110–119, 1990.

[51] Ellison CG and Levin JS. The religion-health connection: evidence for future directions. *Health Education and Behaviour*, 25(6), 1998.

[52] Uneke CJ, Ogbu O, Alo M, , and Ariom T. Syphilis serology in HIV-positive and HIV-negative Nigerians: The public health significance. *Online Journal of Health and Allied Sciences*, 5(2), 2006.

[53] Okunna CS and Dunu IV. Religious constraints on reporting HIV/AIDS in Nigeria. *The World Association of Christian Communication*, pages 1–12, 2007.

[54] Clayton D and Kaldor J. Empirical bayes estimates of age-standardized relative risks for use in disease mapping.

[55] Clayton D and Bernardinelli L. *Bayesian Methods for mapping disease risk. In P. Elliot, J. Cuzick, D. English and R. Stern, editors, Geographical and Environmental Epidemiology: methods for Small-Area Studies.* . 1992.

[56] Clayton D, Bernardinelli L, and Montomoli C. Spatial correlation and Ecological analysis . *International Journal of Epidemiology*, 22:1193–1201, 1993.

[57] Cliff A. D. and Ord J. K. *Spatial Processes*. Pion, London, 1981.

[58] De Angelis D, Gilks W R, and Day N E. Bayesian projection of the acquired immune deficiency syndrome epidemic. *Applied Statistics*, 47(4), 1998.

[59] Eames K T D and Keeling M J. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *PNAS*, 99(20), 2002.

[60] Griffith A. D. A comparison of six analytical disease mapping techniques as applied to West Nile Virus in the coterminous United States . *International Journal of Health Geography*, 4(18), 2005.

[61] Kalbfleisch J D and Lawless J F. Inference based on retrospective ascertainment: An analysis of data on transfusion related AIDS. *Journal of American Statistical Association*, 84:360–372, 1989.

[62] Walter S. D. Assessing spatial patterns in disease rates. *Statistics in Medicine*, 12(12):1885–1894, 1993.

[63] Wilkie A D. Population projections for AIDS using an Actuarial model. *Phil. Trans. Royal Soc., London B*, 625:61–74, 1989.

[64] Zimmerman D. and Zimmerman M. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, 33:77–91, 1991.

[65] Daily Trust Newspapers. Infected blood causes 10 per cent of HIV/AIDS. *http://www.dailytrust.com/*, 2006.

[66] Meyerhoff DJ. Effects of Alcohol and HIV infection on the Central Nervous System. *Alcohol research and Health*, 25, 2001.

[67] Owuamanam DO. Sexual networking among youth in south-western Nigeria. *Health Transition Review*, 5, 1995.

[68] Aghatise E. Trafficking for Prostitution in Italy. *Violence Against Women*, 10:1004, 2004.

[69] Day N E, Gore S M, McGee M A, and South M. Prediction of the AIDS epidemic in the UK: the use of the back projection methods. *Philosophical Transition Royal Society, London B*, 325:123–134, 1989.

[70] Esu-Williams E. Sexually transmitted diseases and condom interventions among prostitutes and their clients in Cross River State . *Health Transition Review*, 5:223–228, 1995.

[71] Eyster M. E., Gail M. H, Ballard J. O, and et al. Natural history of human immodeficiency virus in hemophiliacs: effects of T-cell subsets, platelet counts and age. . *Annals of Internal Medicine*, 107:1–6, 1987.

[72] Halloran M E and Struchiner C J. Study designs for dependent happenings. *Epidemiology*, 2:331–338, 91.

[73] Haris J E. Delay in reporting acquired immunodeficiency syndrome(AIDS). *Working Paper No. 2278. National Bureau of Economic Research*, 1987.

[74] Haris J E. Reporting delays and the incidence of AIDS. *Journal of American Statistical Association*, 85:915–924, 1990.

[75] Harris J. E. Reprting delays and the incidence of AIDS . *Journal of American Statistical Association*, 85:915–924, 1990.

[76] Adejuyigbe EA, Durosinmi MA, Onyia FN, and Adeodu OO. Blood transfusion related paediatric HIV/AIDS in Ile-Ife, Nigeria. *AIDS CARE*, 15(3):329–335, 2003.

[77] Ekanem EE, Afolabi BM, Nuga AO, and Adebajo SB. Sexual Behaviour, HIV-Related knowledge and condom use by intra-city commercial bus drivers and motor park attendants in Lagos, Nigeria. *African Journal of Reproductive Health*, 9(1):78–87, 2005.

[78] Peters EJ, Immananagha KK, Essien OE, and Ekott JU. Traditional healers' practices and the spread of HIV/AIDS in south eastern Nigeria. *Tropical Doctor*, 34(2):79–82, 2004.

[79] Cromley EK and McLafferty SL. *GIS and Public Health.* The Guilford press, New York, first edition, 2002.

[80] Adalaf EM and Smart RG. Risk taking and drug use behaviour: an examination. *AIDS Education prevention*, 11:287, 1983.

[81] Zukerman EM. Sensation-seeking beyound the Optimal level of arousal . *Hillsdale, NJ: Erlbaum* , page 449, 1979.

[82] Akinnawo EO. Sexual networking, STDs and HIV/AIDS transmission among Nigerian Police Officers. *Health Transition Review*, 5, 1995.

[83] Idigbe EO, Ibrahim MM, Ubuane TA, Onwujekwe DI, Otoh I, and Adedoyin JA. HIV/AIDS in Nigeria: Survey of Health and laboratory facilities, 1989-1999. Technical Report. *Nigerian Institute of Medical Research, Lagos*, 2000.

[84] Renne EP. Condom use and the popular press in Nigeria. *Health Transition Review*, 1:41–56, 1993.

[85] Boldson J L et al. On incubation time distribution and the Danish AIDS data . *Journal of Royal Statistical Society A*, 151:42–43, 1988.

[86] Banfied J. F. and Raftery A. E. Model-Based Gaussian and non-Gaussian Clustering. *Biometrics*, 49:803–821, 1993.

[87] Medley G F, Anderson R M, Cox D R, and Billard L. Incubation period of AIDS in patients infected via blood transfusion. *Nature*, 328:719–721, 1987.

[88] Family Health International (FHI). HIV/AIDS Behavioural Surveillance Survey: Nigeria 2000. Report and Data Sheet . *http://www.fhi.org/en/hivaids/pub/survreports/bssnigeria2000.htm*, 2000.

[89] Federal Ministry of Health; Abuja. Technical Report, 1999 National HIV Sero-prevalence Sentinel Survey. 2000.

[90] Federal Ministry of Health; Abuja. Technical Report, The 2001 National HIV/Syphilis Sentinel Survey among pregnant Women attaending Ante natal Clinics in Nigeria. 2001.

[91] Federal Ministry of Health; Abuja. Technical Report, National HIV/AIDS and Reproductive Health Survey. 2003.

[92] Federal Ministry of Health; Abuja. Technical Report, 2003 National HIV Sero-prevalence Sentinel Survey. 2004.

[93] Federal Ministry of Health, Abuja. Technical Report, 2005 National HIV/Syphilis Sero-prevalence Sentinel Survey among pregnant Women attaending Ante natal Clinics in Nigeria. 2006.

[94] Cook D G and Pocock S J. Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics*, 39:361–371, 1983.

[95] Law D. C. G., Serre M. L.and Christakos G, Leone P. A., and Miller W. C. Spatial analysis and mapping of sexually transmitted disease to optimize intervention and prevention strategies. *Sexual Transmitted Infections*, 80:294–299, 2004.

[96] Matheron G. Principles of Geostatistics. *Economic Geology*, 58:1246–1266, 1963.

[97] Taylor J M G. Models for HIV infection and AIDS Epidemic in the United States. *Statistics in Medicine*, 8:45–58, 1989.

[98] Fowler TB Gouws E, Mishra V. Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalized epidemics: implications for calibrating surveillance data. *Sexually Transimitted Infections*, 84(1):i17–i23, 2008.

[99] Schmid GP, Buve A, Mugyenyi, Garnett GP, and et al. Transmission of HIV-1 infection in sub-Saharan Africa and effect of elimination of unsafe injections. *The Lancet*, 363, 2004.

[100] Gupta GR and Weiss E. Women's lives and sex: Implications for AIDS prevention. *Culture, Medicine and Psychiatry*, 17:399–412, 1993.

[101] Chau P. H and Yip P. S. F. Non-parametric back-projection of HIV positive tests using multinomial and poisson settings. *Applied Statistics*, 31(5), 2004.

[102] Chau P. H, Yip P. S.F., , and Cui J. S. Reconstructing the incidence of Human immunodeficiency virus (HIV) in Hong Kong by using data from positive tests and diagnoses of acquired immune deficiency syndrome. *Applied Statistics*, 52(2), 2003.

[103] Gail M H and Rosenberg P S. Perspectives on using backcalculation to estimate HIV prevalence and project AIDS incidence. *In Jewell N P, Dietz K, and Farewell V T eds., AIDS Epidemiology: Methodological Issues, Birkhauser, Boston*, 1992.

[104] Gail M H, Rosenberg P S, and Goedert J.J. Therapy may explain recent deficits in AIDS incidence . *Journal of AIDS*, 3:296–306, 1990.

[105] Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73(1):43–56, 1986.

[106] Goldstein H. Restricted Unbiased Iterative Generalized Least Squares Estimation . *Biometrika*, 76:622–623, 1989.

[107] Goldstein H. *Multilevel Statistical Models*. Edward Arnold, London, 1995.

[108] Goldstein H and Rasbash J. Efficient computational procedure for the estimation of parameters in multilevel models based on iterative generalized least squres. *Computational Statistics and Data Analysis*, 13:63–71, 1992.

[109] Goldstein H and Rasbash J. Improved Approximations for Multilevel Models with Binary Responses . *Journal of the Royal Statistical Society, Series A*, 159(3):505–513, 1996.

[110] Isaaks E. H. and Srivastava R. M. *Introduction to Applied Geostatistics*. Oxford University Press, New York, 1989.

[111] Jacqmin-Gada H, Commenges D, Nejjari C, and Dartigues J. Test of geographical correlation with adjustment for explanatory variables: an application to dyspnoea in the elderly. *Statistics in Medicine*, 16:1283–1297, 1997.

[112] Langford I H, Leyland AH, Rasbash J, and Goldstein H. Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistical Society Series C Applied Statistics*, 48:253–268, 1999.

[113] Leyland A H, Langford IH, Rabash J, and Goldstein H. Multivariate spatial models for event data. *Statistics in Medicine*, 19, 2000.

[114] Human Rights Watch. Accss to condoms and HIV/AIDS information: A global health and human rights concern. 2004.

[115] Auger I. and Thomas P.and De Gruttola V. Incubation Period for Pediatric AIDS Patients . *Nature*, 336:575–577, 1988.

[116] Graubaud B I and Korn E L. Regression analysis with clustered data. *Statistics in Medicine*, 13:509–522, 1994.

[117] Langford IH, Leyland AH, Rasbash J, and Goldstein H. Multilevel modelling of the geographical distributions of diseases. *Applied Statistics*, 48(2):253–268, 1999.

[118] Langford IH, Bentham G, and McDonald A L. Multilevel modelling of geographically aggregated health data: a case study on malignant melanoma mortality and uv exposure in the european community. *Statistics in Medicine*, 17:41–57, 1998.

[119] Emodi IJ and Okafor GO. Clinical manifestations of HIV infection in children at Enugu, NIgeria . *Journal of tropical pediatrics*, 44(2):73–76, 1998.

[120] Orubuloye IO and Oguntimehin F. Death is pre-ordained, it will come when it is due: attitude of men to death in the presence of AIDS in Nigeria. *Resistances to Behavioural change to Reduce HIV/AIDS infection*, pages 101–111, 1999.

[121] Orubuloye IO, Caldwell J, and Caldwell P. Sexual networking in Ekiti District of Nigeria. *Studies in Family Planning*, 22(2):61–73, 1991.

[122] Orubuloye IO, Caldwell JC, and Caldwell P. The cultural, social and attitudinal context of male sexual behaviour in urban south-west Nigeria. *Health Transition Review*, 5:207–222, 1995.

[123] Orubuloye IO, Cadwell P, and Cadwell JC. The role of High-Risk Occupation in the spread of AIDS: Truck drivers and Itinerant Market Women

in Nigeria. *International Family Planning Perspectives*, 19(2):43–48+71, 1993.

[124] Orubuloye IO, Omoniyi O P, and Shokunbi W A. Sexual networking, STDs and HIV/AIDS in four urban gaols in Nigeria. *Health Transition Review*, 5, 1995.

[125] Bailey N T J. *The mathematical theory of infectioud diseases*. New York: Hafner Press, 1975.

[126] Bailey N T J. Simplified Modelling of the Population Dynamics of HIV/AIDS . *Journal of Royal Statistical Society, Series A*, 151(1):31–34, 1988.

[127] Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[128] Besag J. Towards bayesian image analysis. *Journal of Applied Statistics*, 16:259–302, 1989.

[129] Besag J and Mollié A. Bayesian mapping of mortality rates. *Bulletin of International Statistical Institute*, 53:127–128, 1989.

[130] Besag J, York J, and Mollie A. Bayesian Image Restoration with Two Applications in spatial Statistics . *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

[131] Biggar J. AIDS incubation in 1890 HIV seroconverters from different exposure group. *AIDS*, 4:1059–1066, 1990.

[132] Cadwell J. Reasons for limited sexual behaviour change in the sub-Saharan

African AIDS epidemic, and possible future intervention strategies . *Resistance to behaviour change to reduce HIV/AIDS infection*, 15(3):329–335, 1999.

[133] Chin J. Estimation and projection of HIV infection AIDS cases in Hong Kong. *Hong Kong Government*, 1994.

[134] Collin J and Rau B. Africa: HIV/AIDS and poverty. *University of Pennsylvania - African Studies center*, 2000.

[135] Goedert J J, Biggar R J, Weiss S H, and et al. Three year incidence of AIDS in five cohorts of HTLV-III-infected risk group members . *Science*, 231:992–995, 1986.

[136] Goedert J J, Kessler C M, Aledort L M, and et al. A perspective study of human immunodeficiency virus type 1 infection and the development of AIDS in subjects with hemophilia. *New Englan Journal of Medicine*, 321:1141–1148, 1989.

[137] Healey M J and Tillett. Short-term extrapolation of AIDS epidemic . *Royal Statistical Society, A*, 151(1):50–65, 1988.

[138] Lawless J. and Sun J. A comprehensive Back-calculation framework for the estimation and prediction of cases. In: AIDS Epidemiology: Methodological issues, eds., Jewel N. P., Dietz K., and Farewell V. T. . *Birkhauser Basel*, pages 81–104, 1992.

[139] Lui K J, Lawrence DN, Morgan WM, Peterman T A, Haverkos H W, and Bregman D J. Model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome . *Proc. atl. Acad. Sci. USA: Applied Mathematics*, 83:3051–3055, 1986.

[140] Lui K J, Darrow W W, and Rutherford G W. Model-based estimate of the mean incubation period for AIDS in homosexual men. *Science*, 240:1333–1335, 1988.

[141] Rabash J and Browne WJ. *Non-Hierarchical Multilevel models. In: Leyland A, Goldstein H, eds. Multilevel modelling of Health statistics.* Chichester: Willey, 2001.

[142] Spiegelhalter D J, Thomas A, Best N, and Lunn D. *WinBUGS User Manual, version 2.0* . MRC Biostatistics Unit, Cambridge, 2004.

[143] Stone C. J. Comment: Generalized additive models . *Royal Statistical Society, Series B*, 1:312–314, 1986.

[144] Wakefield J. and Shaddick G. Health exposure modelling and the ecological fallacy. *Biostatistics*, 7:438–455, 2006.

[145] Baldwin JA, Maxwell CJ, Fenaughty AM, Trotter RT, and Stevens SJ. Alcohol as a factor for HIV transmission among American Indian and Alaska Native drug users . *American Indian Alaska Native Ment. Health Res.*, 9(1):1–16, 2000.

[146] Caldwell JC, Orubuloye IO, and Caldwell P. Underreaction to AIDS in sub-Saharan Africa. *Social Science Med.*, 34:1169–1182, 1992.

[147] Caldwell JC, Orubuloye IO, and Caldwell P. Obstacles to behavioural change to lesson the risk of HIV infection in the African AIDS epidemic: Nigerian research. *Resistance to Behavioural Change to reduce HIV/AIDS Infection*, pages 113–124, 1999.

[148] Wakefield JC, Best NG, and Waller LA. *Bayesian approaches to disease mapping. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, editors. Spatial*

Epidemiology: Methods and Applications , volume 59. Oxford: Oxford University Press, 2000.

[149] Kesall JE and Diggle PJ. Non-paramtric Estimation of Spatial Variation in Relative Risk . *Statistics in Medicine*, 14:2335–2342, 1995.

[150] Kesall JE and Diggle PJ. Spatial Variation in Risk of Disease: Non-paramtric binary regression approach. *Journal of Royal Statistical Society, Series C*, 47(4):559–573, 1998.

[151] Oyefara JL. Food insecurity, HIV/AIDS pandemic and sexual behaviour of female commercial sex workers in Lagos metropolis, Nigeria. *Social Science Med. Social Aspects of HIV/AIDS*, 4:626–635, 2007.

[152] Mafeni JO and Fajemisin OA. HIV/AIDS in Nigeria: situation, response and Prospects, Key Issues . *Policy Project, Nigeria*, 2003.

[153] Paul JP, Stall R, and Davis F. Sexual risk of HIV transmission among gay/bisexual men in substance abuse treatment. *AIDS Education prevention*, 5(1):11–24, 1993.

[154] Ketchen D. J. Jr and Shook C. L. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17:441–458, 1996.

[155] Longini IM. Jr. Modelling the decline of CD4+ T-lymphocyte counts in HIV infected individuals. . *Journal of Acquired Immune Deficiency Syndromes*, 3:930–931, 1990.

[156] Longini IM. Jr., Clark W. S., Gardner L. I., and Brundage J. The dynamics of CD4+ T-lymphocyte decline in HIV infected individuals: Markov Chain

modelling approach . *Journal of Acquired Immunodeficiency Syndromes*, 4:1141–1147, 1991.

[157] Jone K and Duncan C. Individuals in thier ecologies: Analysing the geography of chronic illness with a multilevel modelling framework. *Health and Place*, l(1), 1995.

[158] Kafadar K. Smoothing Geographical data, Particularly rates of Disease . *Statistics in Medicine*, 15:2539–2560, 1996.

[159] Mardia V. K, Kent J. T., and Bibby J. *Multivariate Analysis*. Academic Press, 1979.

[160] Ord K. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126, 1975.

[161] Manton KG, stallard E, Woodbury MA, Riggan WB, Creason JP, and Mason TJ. Statistically adjusted estimates of geographical mortality. *Journal of the National Cancer Institute*, 78:805–815, 1987.

[162] Adeokun L. Social and Cultural factors affecting the HIV epidemic: In AIDS in Nigeria: A nation in the Threshold. *Harvard Centre for population and Development Studies*, 2006.

[163] Adeokun L, Okonkwo P, and Ladipo OA. The Epidemiology of HIV/AIDS in Nigeria: In AIDS in Nigeria: A nation in the Threshold. *Harvard Centre for population and Development Studies*, 2006.

[164] Alkema L, Raftery AE, and Brown T. Bayesian melding for estimating uncertainty in national HIV prevalence estimates. *Sexually Transimitted Infections*, 84(1):i11–i16, 2008.

[165] Anselin L. The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. *GISDATA Specialist Meeting on GIS and Spatial Analysis, Amsterdam*, 1993.

[166] Bernadinelli L and Montomoli C. Emperical bayesian versus fully bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 11:983–1007, 1992.

[167] Montana LS, Mishra V, and Hong R. Comparison of HIV prevalence estimates from antenatal care surveillance and population-based surveys in sub-Saharan Africa. *Sexually Transimitted Infections*, 84(1):i78–i84, 2008.

[168] Abramowitz M and Stegun I A. Handbook of Mathematical functions with Formulars, Graphs and Mathematical Tables. *Washington DC: National Bureau of Standards*, 1964.

[169] Kou J. M., Taylor J M G, and Detels R. Estimating the AIDS incubation period from a prevalence cohort. *Ameriacn Journal of Epidemiology*, 133:1050–1057, 1991.

[170] Kretzschmar M and Wiessing L G. Modelling the spread of HIV in social networks of injecting drug users . *AIDS*, 12:801–811, 1998.

[171] Longini I. M., Clark W. S., Byers R. H., and Ward J. W.and et al. Statistical analysis of the stages of HIV infection using a Markov model. . *Statistics in Medicine*, 8:831–843, 1989.

[172] McEvoy M and Tillet H E. Some problems in the prediction of future numbers of cases of the acquired immunodeficiency syndrome in the uk. *Lancet*, ii:541–542, 1985.

[173] Morgan W M and Curan J W. Acquired immunodeficiency syndrome: current and future trends. *Public Health Report*, 101:459–465, 1986.

[174] Morris M, Levine R O, and Weaver M. Sexual networks and HIV program design . *www.Synergyaids.com/documents/SexualNetworksHIVprogramdesign.*

[175] Raab G M, Allardice G, Goldberg D J, and McMenamin J. Modelling Human Immunodeficiency Virus Infection and Acquired Immune Deficiency Syndrome cases in Scotland: Data sources, prior information and Bayesian estimation . *Royal Statistical Society, A*, 161(3):367–384, 1988.

[176] Raab G M, Gore S M, Goldberg D J, and Donnelly C A. Bayesian Forecasting of the Human Immunodeficiency Virus Epidemic in Scotland . *Journal of Royal Statistical Society, A*, l57(1), 1994.

[177] Rees M. The sombre view of AIDS. *Nature*, 326:343–345, 1987.

[178] Jain A. K.; Murty M.N. and Flynn P. J. Data Clustering: A Review. *ACM computing surveys*, 31(3):264–307, 1999.

[179] Osagbemi MO and Adepetu AA. Gender differences in the reason for participation in spouse sharing among okun in Nigeria. *African Journal of Reproductive Health*, 5(2):36–55, 2001.

[180] Osagbemi MO and Jegede AS. Spouse-sharing practice and reproductive health promotion among Okun people of Nigeria. *African Population Studies*, 16(2):91–116, 2001.

[181] Osagbemi MO, Joseph B, Adepetu AA, Nyong AO, and Jegede AS. Culture and HIV/AIDS in Africa: Promoting Reproductive Health in Light of

Spouse-Sharing Practice among Okun People, Nigeria. *World Health and Population*, 2007.

[182] Best N, Waller L A, Thomson A, Conlon E M, and Arnold R A. *Bayesian models for spatially correlated diseases and exposure data (with discussions).* In Bayesian Statistics: eds. Bernardo J M, Berger J O, Dawin A P, Smith A F M. Oxford University Press, 1999.

[183] Best N, Richardson S, and Thomson A. A comparison of bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14(1):35–59, 2005.

[184] Breslow N and Clayton D. Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88:9–25, 1993.

[185] Cressie N. *Statistics for Spatial Data.* John Wiley and Sons, Inc, New York, 1991.

[186] Cressie N and Hawkins D. M. Robust estimation of the variogram. *Journal of International Association of Mathematical Geology*, 12:115–125, 1980.

[187] Hessol N, Rutherford G W, O'Malley P M, Doll L S, Darrow W W, and Jaffe H W. The natural history of human immunodeficiency virus infectionin cohorts of homosexual and bisexual men. *Third international Conference on AIDS, Washington DC, Department of Health and Human Srevices and the World Health Organisation*, 1987.

[188] Seidman S N, Monsher M D, and Aral S O. Women with multiple sexual partners: United States, 1988 . *American Journal of Public Health* , 82(10):1388–1394, 1992.

[189] Venables W. N. and Ripley B. D. *Modern Applied Statistics with R* . Springer, USA, 2002.

[190] National Agency for the Control of AIDS, Abuja. Nigeria UNGASS Report. 2005.

[191] National Agency for the Control of AIDS, Abuja. Nigeria UNGASS Report. 2007.

[192] National Buteau of Statistics (NBS). Federal Republic of Nigeria 2006 Population Census. Official Gazatte . *www.nigeriastat.gov.ng*, 2007.

[193] National Population Commission, Abuja. Nigeria Demographic and Health Survey, year = 2003,.

[194] National Population Commission, Abuja. Nigeria Demographic and Health Survey. 1999.

[195] Ezumah NE. Gender Issues in the Prevention and Control of STIs and HIV/AIDS; Lessons from Awka and Agulu, Anambra State, Nigeria. *African Journal of Reproductive Health*, 7:89–99, 2003.

[196] Becker N.G. and Marschner I. C. A method for estimating the age-specific relative risk of HIV infection from AIDS incident data. *Biometrika*, 80:165–178, 1993.

[197] Becker N.G., Watson L. F., and Carlin J. B. The method of nonparametric back-projection and its application to AIDS data . *Statistics in Medicine*, 10:1527–1542, 1991.

[198] Laird NM and Louis TA. Empirical bayes confidence intervals based on bootstrap samples. *Journal of American Statistical Association*, 82(25):739–750, 1987.

[199] Aalen O. O, Farewell V T, Angelis D D, and Day N E. The use of human Immunodeficiency Virus diagnosis in formation in Monitoring AIDS epidemic. *Journal of Royal Statistical Society, Series A*, 157(1):3–16, 1994.

[200] Adegbola O and Babatola O. Premarital and extramarital sex in Lagos, Nigeria. *The Continuing African HIV/AIDS Epidemic*, pages 19–44, 1999.

[201] Berke O. Exploratory disease mapping: kriging the risk function fron regional count data. *International Journal of Health Geographics*, 3(18), 2004.

[202] Devine O and Louis T. A constrained empirical bayes estimator for spatially correlated incidence rates. *Statistics in Medicine*, 13(4):1119–1133, 1994.

[203] Devine O, Louis T, and Halloran ME. Empirical bayes method for stabilizing incidence rates before mapping. *Epidemiology*, 5(4):622–630, 1994.

[204] Schabenberger O. and Gotway C. A. *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC, New York, 2005.

[205] J. C. Okeibunor. Threats of AIDS and condom use in a Nigerian urban community: Implications for fertility regulation in Nigeria. *Union for African Population Studies*, (36), 1999.

[206] Baachetti P. The impact of lengthening AIDS reporting delays and uncertainty about underreporting on incidence trends and projections. *Journal of AIDS*, 7:860–865, 1994.

[207] Baachetti P, Segal M R, and Jewell N P. Backcalculation of HIV Infection Rates. *Statistical Science*, 8(2):82–101, 1993.

[208] Bacchetti P. and Jewell N. P. Nonparametric estimation of the incubation period of AIDS based on a prevalence cohort with unknown infection times. *Biometics.*, 47:947–960, 1991.

[209] Blythe S P and Anderson R M. Distributed incubation and infection periods in models of the transmission dynamics of human immunodeficiency virus (HIV) . *IMA Journal of Mathematics Applied in Medicine and Biology*, 5:1–19, 1988.

[210] Cadwell P. Prostitution and the risk of STDs and AIDS in Nigeria and Thailand. *Health Transition Review*, 5:167–172, 1995.

[211] Diggle P, Moyeed R, and Tawn J. Model based Geostatistics. *Journal of Royal Statistical Society, Series C*, pages 299–350, 1998.

[212] Freund H. P and Book D. L. Determination of the spread of HIV from the AIDS incidence history. . *Math. Biosciences*, 98:227–241, 1990.

[213] Garnett G P. An introduction to mathematical models in sexually transmitted disease epidemiology. *Sexually Transmitted Infections*, 78:7–12, 2002.

[214] Jewel N. P. Some statistical issues in studies of epidemiology of AIDS. *Statistics in Medicine*, 9:1387–1416, 1990.

[215] Poit P, Greener R, and Russel S. Squaring the circle: Poverty, and Human Development. *PLoS Medicine*, 10:1571–1575, 2007.

[216] Wild P, Commenges D, and Etcheverry B. A hierarchical Bayesian approach to the back-calculation of numbers of HIV-infected subjects . *The Statistician*, 42:405–414, 1993.

[217] Moran PA. The interpretation of Statistical maps . *Journal of Royal Statistical Society, Series B*, 10:243–251, 1948.

[218] Gray PB. HIV and Islam: is HIV prevalence lower among Muslims? . *Social Science and Medicine* , 2003.

[219] Muechrcke PC and Muechrcke JO. *Map use; reading, analysis and intepretation*. Madison, wis: JP publications, 3rd edition, 1992.

[220] Ghys PD, WalKer N, McFarland W, Miller R, and Garnett GP. Improved data, methods and tools for the 2007 HIV and AIDS estimates and projections. *Sexually Transimitted Infections*, 84(1):i1–i4, 2008.

[221] Ghys PD, Brown T, Grassly NC, Garnett G, Stanecki KA, Stover J, and Walker N. The UNAIDS Estimation and Projection Package: a software package to estimate and project national HIV epidemics. *Sexually Transimitted Infections*, 80(1):i5, 2004.

[222] Diggle PJ, Tawn JA, and Moyeed RA. Model-based geostatistics (with discussion). *Applied Statistics*, 47:299–350, 1998.

[223] Bastani R, Erickson PA, Marcus AC, and et al. AIDS related attitudes and rik behaviours: a survey of a random sample of California heterosexuals. *Preventive Medicine*, 25:105–117, 1996.

[224] Bellocco R and Marschner I. C. Joint analysis of HIV and AIDS surveillance data in back-calculation . *Statistics in Medicine*, 19:297–311, 2000.

[225] Brookmeyer R. Reconstruction and Future trends of the AIDS epidemic in the United States . *Science*, 253:37–42, 1991.

[226] Brookmeyer R and Damiano A. Statistical Methods for short -term projections of AIDS incidence . *Statistics in Medicine*, 8:23–34, 1989.

[227] Brookmeyer R and Gail M H. Minimum Size of the acquired Immunodeficiency Syndrome (AIDS) epidemic in the United States . *The Lancet*, ii:1320–1322, 1986.

[228] Brookmeyer R and Gail M H. Biases in prevalene cohorts. *Biometrics*, 43:739–749, 1987.

[229] Brookmeyer R and Gail M H. A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of American Statistical Association*, 83(402):301–308, 1988.

[230] Brookmeyer R and Gail M H. *AIDS EPidemiology. A Quantitative Approach*. Oxford University Press, 1994.

[231] Brookmeyer R and Goedert J J. Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics*, 45:325–335, 1989.

[232] Brookmeyer R and Liao J. The analysis of delays in disease reporting: Methods and results for the acquired immunodeficiency syndrome. *American Journal of Epidemiology*, 132:355–365, 1990.

[233] Brookmeyer R and Liao J. Statistical modelling of AIDS epidemic for forecasting health care needs. *Biometry*, 46:1151–1163, 1990.

[234] Carroll R. and Rupert D. A comparison between maximum likelihood and generalized lerast squares in a heteroscedastic linear model . *Journal of the American Statistical Association*, 77:878–882, 1982.

[235] Geary R. The contiguity ratio and statistical mapping. *Incorporated Statistician*, 5:115–145, 1954.

[236] Healy M J R and Tillett H E. Short-term Extrapolation of AIDS Epidemic . *Journal of Royal Statistical Association A*, 151(1):50–61, 1988.

[237] Kaluzny S. R., Vega S.C., Cardoso T. P, and Shelly A. A. *S+ Spatial Stats. User's Manual for Windows and UNIX*. Springer, New York , 1998.

[238] *Künsch* H R. Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika*, 74:517–524, 1987.

[239] Lande R, Engen S, and Saether BE. Extinction times in finite metapopulation models with sochastic local dynamics . *Oikos*, 83:383–389, 1998.

[240] Lyerla R, Gouws E, and Garcia-Calleja. The quality of sero-surveillance in low-and middle-income countries: status and trends through 2007. *Sexually Transimitted Infections*, 84(1):i85–i91, 2008.

[241] Stall R and Leigh B. Understanding the relationship between drug or alcohol use and high risk sexual activity for HIV transmission: where do we go from here? *Addiction*, 89:131–134, 1994.

[242] Schinazi RB. On role of Social Clusters in the Transmission of infectious diseases. *Theoretical Population Biology*, 61:163–169, 2002.

[243] Downer RG. An introduction to Smoothing Incidence Rates by Penalized Likelihood . *Statistics in Medicine*, 15:907–917, 1996.

[244] Marshall RJ. A Review of Methods for the Statistical Analysis of Spatial Patterns of Disease . *Journal of the Royal Statistical Society, Series A, Statistics in Society*, 154:421–441, 1991.

[245] Marshall RJ. Mapping Disease and Mortality rates using Empirical Bayes Estimators . *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 40(2), 1991.

[246] Tsutakawa RK. Estimation of Cancer Mortality Rates in a Bayesian Analysis of Small Frequencies . *Biometrics*, 41(1):69–79, 1985.

[247] Tsutakawa RK. Mixed Model for analysing Geographic variability in Mortality Rates . *Journal of American Statistical Association*, 83(407):637–650, 1988.

[248] Tsutakawa RK, Shoop GL, and Mariennfield CJ. Empirical Bayes Estimation of Cancer Mortality Rates . *Statistics in Medicine*, 4:201–212, 1985.

[249] Banerjee S, Carlin B. P., and Gelfand A. E. *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall/CRC, New York , 1999.

[250] Cui J. S and Becker N. G. Estimating HIV incidence using dates of both HIV and AIDS diagnosis . *Statistics in Medicine*, 19:1165–1177, 2000.

[251] Mbiti J. S. African Religions and Philosophy. *Heinemann Educational Book Incorporated*, 2nd ed:138–143, 1989.

[252] Richardson S. *Statistical methods for geographical correlation studies. In Geographical and environmental epidemiology: Methods for small area studies.* P. Elliott, J. Cuzick, D. English, and R. Stern eds. Oxford University Press, 1992.

[253] Richardson S. and Monfort C. *Ecological correlation studies. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, editors. Spatial Epidemiology: Methods and Applications* . Oxford: Oxford University Press, 2000.

[254] Rosenberg P S. A simple correction of AIDS surveillance data for reporting delays. *Journal of AIDS*, 3:49–54, 1990.

[255] Rosenberg P. S. Backcalculation models of age-specific HIV incidence rates . *Statistics in Medicine*, 13:1975–1990, 1994.

[256] Rosenberg P S and Gail M H. Uncertainty in estimates of HIV prevalence derived backcalculation. . *Annals of Epidemiology*, 1:105–115, 1990.

[257] Rosenberg P S and Gail M H. Backcalculation of Flexible Linear Models of the HIV Infection Curve . *Applied Statistics*, 40:269–282, 1991.

[258] Rosenberg P S, Gail M H, and Pee D. Mean square error of estimates of HIV prevalence and short-term AIDS projections derived from backcalculation. *Statistics in Medicine*, 10:1167–1180, 1991.

[259] Rosenberg P S, Gail M H, and Carrol R J. Estimating HIV prevalence and AIDS incidence in the United States: a model that accounts for therapy and changes in the Surveillance definition of AIDS . *Statistics in Medicine*, 11:1633–1655, 1991.

[260] Rosenberg P S, Gail M H, Schrager L. K., Vermund S. H, and et al. National AIDS incidence trend and the extent of zidovudine therapy in selected demographic and transmission groups . *Journal of AIDS*, 4:392–401, 1991.

[261] Rosenberg P S, Biggar R J, Goedert J J, and Gail M H. Backcalculation of the number with human immunodeficiency virus infection in the United States. *American Journal of Epidemiology*, 133:276–285, 1991.

[262] Zhong S, Xue Y, Coa C, Li X, Guo J, and Fang L. Explore disease mapping of Hepatitis B using Geostatistical Analysis Techniques. *ICCS*, LNCS 3516:464–471, 2005.

[263] Ogbuagu SC and Charles JO. Survey of Sexual networking in Calabar. *Health Transition Review*, 3:105–119, 1993.

[264] Lerman S.E. and Liao J.C. Neonatal circumcision. *Pediatric Clinics of North America*, (48):1539–1557, 2001.

[265] Edwards Sharon. Having multiple sexual partners is linked to age of first sex and birthplace. *Family Planning Perspective*, 1994.

[266] Mbulaiteye SM, Ruberantwari A, Nakiyingi JS, Carpenter LM, Kamali A, and Whitworth JAG. Alcohol and HIV: a study among sexually active adults in rural southwest Uganda . *International Journal of Epidemiology* , 29:911–915, 2000.

[267] Obi SN. Extramarital sexual activity among infertile women in southeast Nigeria. *Journal of Obstetrics and Gynecology, India*, 56:72–75, 2006.

[268] Utulu SN and Lawoyin TO. Epidemiological features of HIV infection among pregnant women in Makurdi, Benue State, Nigeria. *Journal of Biosoc. Science*, pages 1–12, 2006.

[269] Brooks SP. Markov chain monte carlo and its application. *The Statistician*, 47:69–100, 1998.

[270] Malamba SS, Wagner HU, Maude G, and et al. Risk factors for HIV infection in adults in a rural Ugandan community: a case -control study. *AIDS*, 8:253–257, 1994.

[271] Brown T, Salomon JA, Alkema L, Raftery AE, and Gouws E. Progress and challenges in modelling country-level HIV/AIDS epidemics: the UNAIDS Estimation and Projection Packages. *Sexually Transimitted Infections*, 84(1):i5–i10, 2008.

[272] Chandola T, Clark P, Wiggins RD, and Bartley M. Who you live with and where you live: setting the context for health using multiple membership multilevel models. *Journal of Epidemiology Community Health*, 59:170–175, 2005.

[273] Tango T. Estimation of Haemophilia-associated AIDS incidence in Japan

using individual dates of diagnosis. *Statistics in Medicine*, 8:1509–1514, 1989.

[274] Bailey TC and Gatrell AC. *Iterative Spatial Data Analysis*. Harlow: London, 1995.

[275] Hagenaars TJ, Donnelly CA, and Ferguson NM. Spatial heterogeneity and the persistence of infectious diseases. *Journal of Theoretical Biology*, 229:349–359, 2004.

[276] Akinbobola TO and Saibu MOO. Income inequality, Unemployment and Poverty in Nigeria: a Vector Autoregressive Approach. *Policy Reform*, 7(3):175–183, 2004.

[277] Isiugo-Abanihe UC. Extramarital relations and perceptions of HIV/AIDS in Nigeria. *Health Transition Review*, 4:111–125, 1994.

[278] UNAIDS. Epidemiological fact sheets on HIV/AIDS and Sexually Transmitted Infections: Nigeria . *www.unaids.org/hmtl/pub/publications/fact-sheets01/Nigeria*, 2004.

[279] UNAIDS. Epidemiological fact sheets on HIV/AIDS and Sexually Transmitted Infections: Nigeria. *http://www.who.int/globalatlas/predefinedReports/EFS2008/short/EFSCountryProfiles2008-NG.pdf*, 2008.

[280] UNAIDS Epidmiology Reference Group. Recommended methodology for the estimation and projection of HIV prevalence and AIDS mortality in the short-term. *Meeting held at La Mainaz Hotel, Gex France*, 2001.

[281] UNAIDS Epidmiology Reference Group. Technical report and recommendations: Improving parameter estimation, projection methods, uncertainty

estimation and epidemic classification. *Meeting held in Prague, Czech Republic*, 2006.

[282] UNAIDS Epidmiology Reference Group. Technical report and recommendations: Improving EPP and spectrum estimation tools for the 2008-9 round of national estimates with specific attention to prevalenc fits, uncertainty, etc. *Meeting held in London*, 2008.

[283] UNDP. Human Development Report. *United Nations Development Program*, 141, 2004.

[284] UNDP. 2007/2008 Human Development Reports. 2008.

[285] UNFPA. Donor support for contraceptives and condoms for STI/HIV prevention 2005. 2005.

[286] UNFPA. Worldwide indicators. *http://www.unfpa.org/worldwide/indicator.do?filter=getIndicatorValues*, 2007.

[287] UNICEF. Infomation by country. At a glance: Nigeria Statistics . *www.unicef.org/inforbycountry/nigeria-statistics.html*, 2007.

[288] DeGruttola V and Lagakos SW. The value of AIDS incidence data in assessing the spread of HIV infection . *Statistics in Medicine*, 8:35–44, 1989.

[289] DeGruttula V., Tu X. M., , and Pagano M. Pediatric AIDS in New York City: Estimating the distribution of Infection, latency and reporting delays and projecting future incidence . *Journal of American Statistical Association*, 87:633–640, 1992.

[290] DeGruttula V. and Lagakos S. W. Analysis of doubly censured survival data, with application to AIDS . *Biometrics*, 45:1–11, 1989.

[291] Diez Roux A V and Aiello A E. Multilevel Analysis of Infectious Disease . *Journal of Infectious Disease*, 191(1):S25–S33, 2005.

[292] Saphonn V, Hor LB, Ly SP, and Chhuon S. How well do antenatal clinic (ANC) attendees represent the general population? *International Journal of Epidemiology*, 31(2):449–455, 2002.

[293] Bowman A. W and Azzalini A. *Applied smoothing Techniques for data analysis*. Oxford University Press Inc., New York , 1997.

[294] Browne W, Goldstein H, and Rabash J. Multiple membership multiple classification (MMMC) models . *Statistical Modelling*, 1(2):103–124, 2001.

[295] Hill P W and Goldstein H. Multilevel modelling of educational data with cross-classification and missing identification of units. *Journal of Educational and Behavioural Statistics*, 23:117–128, 1998.

[296] WHO. Methods for estimation/projection of HIV infectionand AIDS cases/deaths. *HIV/AIDS in Asia and Pacific Region. Annex 4*, pages 102–109, 2003.

[297] WHO. World Health statistics. *http://www.who.int/countries/nga/nga/en/*, 2008.

[298] WHO, UNAIDS and UNICEF . Towards universal access: scaling up priority HIV/AIDS interventions in the health sector. 2007.

[299] Browne WJ and Drapper D. A comparison of bayesian and likelihood based methods for fitting multilevel models. *Nottingham Statistics Research Reports*, 2004.

[300] Getz WM, Lloyd-Smith, Cross PC, and et al. Modeling the invasion and spread of contagious diseases in heterogeneous populations. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Math. Soc* , 2005.

[301] Gilks WR, Richard S, and Spiegelhalter DJ. *MCMC in Practice*. Chapman and Hall, New York, 1996.

[302] Sakamoto Y., Ishiguro M., and Kitagawa G. *Akaike Information Criterion Statistics*. KTK Scientific Publishers, Tokyo., 1986.

[303] Yasui Y, Lui H, Benach J, and Winget M. An empirical evaluation of various priors in the empirical bayes estimation of small area disease risks. *Statistics in Medicine*, 19(17–18):2409–2420, 2000.

[304] Wai yaun T. *Stochastic Modeling of AIDS epidemiology and HIV pathogenesis*. World Scientific Publishing Co. Pte. Ltd, 2000.

[305] MacNab YC. Hierarchical bayesian modelling of spatially correlated health service outcome and utilization rates. *Biometrics*, 59(2):305–316, 2003.

[306] Macnab YC, Kmetic A, Gustafson P, and Shep S. An innovative application of Bayesian disease mapping methods to patient safety research: A Canadian adverse medical event study . *Statistics in Medicine*, 25:3960–3980, 2006.

[307] MacNab YC and Dean CB. Parametric bootstrapping and penalizes quasi-likelihood inference in conditional autoregressive models. *Statistics in Medicine*, 19(17/18):2421–2435, 2000.

[308] Macnab YC and Dean CB. Autoregressive Spatial Smothing and Temporal Spline Smoothing for Mapping Rates . *Biometrics*, 5:949–956, 2001.

[309] Macnab YC and Dean CB. Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine*, 21:347–358, 2002.

[310] Macnab YC, Farrel PJ, Gustafson P, and Sijin W. Estimation in Bayesian Disease Mapping. *Biometrics*, 60:865–873, 2004.