UNIVERSITY OF MINES AND TECHNOLOGY
TARKWA

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

A THESIS REPORT ENTITLED

DATA-DRIVEN ASSESSMENT OF IMPACTS AND INTERACTIVITIES OF SOCIO-ECONOMIC HIV DRIVERS – A CASE STUDY OF GHANA

BY

WILLIAM AKOTAM AGANGIBA

Submitted in Fulfilment of the Requirement for the Award of the Degree of Doctor of Philosophy in Computer Science and Engineering

TARKWA, GHANA

JUNE 2019

UNIVERSITY OF MINES AND TECHNOLOGY
TARKWA

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

KNOWLEDGE, TRUTH AND EXCELLENCE

A THESIS REPORT ENTITLED

DATA-DRIVEN ASSESSMENT OF IMPACTS AND INTERACTIVITIES OF SOCIO-ECONOMIC HIV DRIVERS – A CASE STUDY OF GHANA

BY

WILLIAM AKOTAM AGANGIBA

Submitted in Fulfilment of the Requirement for the Award of the Degree of Doctor of Philosophy in Computer Science and Engineering

TARKWA, GHANA

JUNE 2019

DECLARATION

I declare that this thesis is my own work. It is being submitted for the degree of Doctor of Philosophy in Computer Science and Engineering in the University of Mines and Technology (UMaT), Tarkwa. It has not been submitted for any degree or examination in any other university.

……………………………………….

(Signature of Candidate)

………….…… Day of …………………2019

KNOWLEDGE, TRUTH AND EXCELLENCE

ABSTRACT

Human Immunodeficiency Virus (HIV) is the virus responsible for the condition called Acquired Immunodeficiency Syndrome (AIDS). Since the first report of HIV in the early 1980s, its adverse effects on humanity cannot be overemphasised. HIV infection has led to the loss of many lives and adversely affected businesses across the globe. As a result, it has had very significant negative impacts on the world's economy in many adverse ways. Sub-Saharan Africa (SSA) is known to be most affected by this virus; accounting for approximately 70 per cent of the world's HIV infected population. Annual death records and new infections caused by the virus are highest in SSA. Attempts by Researchers, Governments and intergovernmental bodies have been largely successful in combating its spread but not enough to eliminate it since SSA remains a region of a generalised epidemic. Many researchers have identified fundamental Socio-economic drivers of the virus in the Sub-Saharan African context in various population subgroups. These drivers are usually identified as a list of prevailing factors characterising the spread of the virus in specific communities without further details regarding their measures of impact. Minimal efforts have been made in an attempt to assess, in quantitative measures, the contribution of individual drivers to the spread of the virus. Furthermore, not much has been done in an attempt to investigate possible interdependencies or interactivities among such Socio-economic drivers. Such details, if known, would give policymakers a deep insight into the behaviour of such drivers. With detailed insights, the objective to halt further new infections of the virus could be achieved much faster. Ghana is the contextual setting for this research. Ghana is situated in Sub-Saharan Africa where several institutions have been set up to formulate policies and to roll out campaigns to combat the spread of the virus. This research develops a data-driven computational model for assessing the degrees of impact and interdependences of Socio-economic HIV drivers using Feature Maximization, which is an emerging data mining technique. Feature Maximization first splits the dataset into several clusters before assessing the effect (degree of impact) of each driver contained in the dataset. This approach gives equal opportunity for the impact of each driver to be assessed adequately. This approach makes it possible to report each Socio-economic HIV driver together with its degree of impact for the given context of the study, which was not the case in earlier works. It is clear from the outcome that, the degrees of impact and interactivities of a given HIV driver depends on the setting (rural or urban), education level, marital status and occupation. The results obtained shows that, low (formal)

KNOWLEDGE, TRUTH AND EXCELLENCE

education and disrupted marital statuses such as divorce, separation and widowhood are strong drivers of the epidemic. Rural males of age groups 40-44 and females of age groups 30-34 and 40-49 are most prone to the epidemic. In the urban settings, the male age group most infected are 30-34 and 50-54 while the most infected age groups for females are 25-29, 35-39 and 40-44. Elementary occupations such as Crafts and Related Trades Workers were found to be strongly associated with the epidemic in urban areas while technicians and associate professionals are most at risk of getting infected by the rural setting. The developed model runs in linear time and is therefore suitable for large datasets. It would be very useful for stakeholders and policymakers in their quest to curb the HIV epidemic. It is recommended that future research works devise ways to establish the direction of causality among HIV drivers.

KNOWLEDGE, TRUTH AND EXCELLENCE

DEDICATION

To my Dad, Mr A. G. Akom:

Your words of profound wisdom continue to inspire me to be resilient, persistent and focused for excellence in life.

To my Mum, Mrs Mary Akom (of Blessed Memory):

Your wise counsel is what has brought me thus far. I wish you were here to see this day!

****

To my family:

To Millicent – my wife, soul mate and best friend. Thanks for the great inspiration and support. Others say "the sky is the limit", but you say "the sky the starting point"!

To my children Gabriel, Benedict and Milcah – you make my purpose on earth complete!

I love you all.

KNOWLEDGE, TRUTH AND EXCELLENCE

ACKNOWLEDGEMENTS

This PhD dream was first conceived in 2014. Through it all, there have been many challenging as well as painful and disappointing moments. This dream could have been cut short if the Almighty God had not brought me into contact with certain great personalities whose roles I wish to acknowledge here. My heartfelt gratitude goes firstly to my PhD advisors; Prof Jean-Charles Lamirel (University of Strasbourg, France) and Prof B. K. Alese (Federal University of Technology, Akure, Nigeria). Prof Jean-Charles Lamirel was instrumental in shaping the key ideas and discourse of this thesis. He travelled to Ghana twice solely for the sake of this work and later welcomed me for an internship at the Lorraine Research Laboratory in Computer Science and its Applications (LORIA) from July to December 2017. The key ideas expressed in this work were hatched during this period. On the other hand, the logical organisation, structure and layout of this work would not have been possible without the expert directorship of Prof B. K. Alese. His motivation, experienced guidance and critical analysis were integral to the completion of this work.

To my beloved wife, Millicent Agangiba, I am grateful for the sacrifices, your extraordinary support and inspiration and for painstakingly editing various drafts of this thesis. To my dearest children Gabriel, Benedict and Milcah, who sometimes had to endure my long absence to enable me to complete this thesis, I love you all.

My heartfelt gratitude goes to my parents, Mr A. G. Akom and Mrs Mary Akom of blessed memory for nurturing me and inspiring me to achieve academic excellence. To my siblings, Barakatu Akom, Raliatu Akom, Nurudeen Akom and Sumayatu Akom thank you all for the amazing support. I am grateful to the members of Kontoh and Duah families for their support and prayers. Special mention must be made of Mrs Cecilia Kontoh and Mrs Joyce Duah, who always gave me a reason to persevere harder in life.

My deepest gratitude is expressed to Dr Mrs Cynthia Boye (Head, Computer Science and Engineering Department, UMaT) and Assoc Prof Kumi-Boateng (Former Head, Computer Science and Engineering Department, UMaT). These individuals played key roles in creating the necessary administrative channels and support needed for the progress and success of this work.

Prof J. S. Y. Kuma (Vice Chancellor, UMaT) is uniquely remembered in this work for his utmost passion in seeing this dream come to pass successfully. Out of his passion, he

KNOWLEDGE TRUTH AND EXCELLENCE

created the necessary external avenues to make sure my colleagues and I got the best possible professors in Computer Science to supervise our PhD works. Through one of those avenues, I met Prof Jean-Charles Lamirel at the University of Cape Coast in 2016.

I also wish to acknowledge Prof Amoakoh Gyasi-Agyei (Melbourne Institute of Technology, Australia) for being the first person to suggest this line of research to me. Even though the idea was very blurred and not easily understood at that time; it has now yielded into a beautiful piece as expressed in this thesis.

The following personalities played various essential roles in my life and family during the execution of this work. Their wise counsel, prayers and motivation cannot be overemphasised. These include Prof R. K. Amankwah (Mentor and Dean, Faculty of Integrated Management Sciences, UMaT), Assoc. Prof Anthony Simons (Dean, Faculty of Engineering, UMaT), Assoc. Prof Grace Ofori-Sarpong (Dean, Faculty of Mineral Resources Technology, UMaT), Dr Richard Amorin (Chairman of Chaplaincy Board, UMaT), Dr Issaka Yakubu (Head, Mathematics Department) and all members of the Computer Science and Engineering Department, UMaT.

Special acknowledgement also goes to the National AIDS Control Programme (Accra) and the Ghana Statistical Service (Accra) for giving me access to the data which was necessary for this work. I am grateful to Madam Jacqueline Anum (Ghana Statistical Service, Accra) and Mr Ekow Wiah (National AIDS Control Program, Accra) for their assistance during my data collection.

Finally, I wish to acknowledge Humphrey Appiah Mensah (Syracuse University, USA), Emmanuel Saake (USA), Jonas Agyapong, Oscar Deho Blessed, Ebenezer Afrifa and Jewel Apeko Donkor who were teaching assistants at the Computer Science and Engineering Department of UMaT at the time this research was at its peak. These gentlemen contributed in various to make it comfortable for me to concentrate on the write-up.

TABLE OF CONTENTS

KNOWLEDGE, TRUTH AND EXCELLENCE

KNOWLEDGE, TRUTH AND EXCELLENCE

LIST OF FIGURES

KNOWLEDGE, TRUTH AND EXCELLENCE

LIST OF TABLES

KNOWLEDGE, TRUTH AND EXCELLENCE

INTRODUCTION

Background

Human Immunodeficiency Virus (HIV) is the causal agent of Acquired Immune Deficiency Syndrome (AIDS) (Arrehag *et al.*, 2006; Blattner *et al.*, 1988). AIDS was first reported in 1981 in the United States and some parts of Africa (Blattner *et al.*, 1988). HIV, however, was discovered in 1983 and was confirmed as the cause of AIDS in 1984 (Blattner *et al.*, 1988). Since their emergences, HIV and AIDS have constituted a major public health problem across the globe (Gaigbe-Togbe and Weinberger, 2004; Bain *et al.*, 2017). Two types of the virus are known; namely HIV-type 1 and HIV-Type 2. HIV-type 1 is the most dominant type of the virus known worldwide, while HIV-type 2 is common only in West Africa (Arrehag *et al.*, 2006). HIV-type 1 develops much faster into AIDS than HIV-type 2.

Even though HIV is a worldwide epidemic with many adverse consequences, its impact is extremely more devastating to Africa than any other part of the world. Africa houses about 70% of the world's people living with HIV (Were and Nafula, 2003; Arrehag *et al.*, 2006). HIV prevalence, death rate and yearly recorded new infections are extremely much higher in Africa compared to each of the other continents across the globe. Most infected are usually the economically active age groups; thereby imposing very adverse consequences on the African economy (Arrehag *et al.*, 2006). Socio-economic factors such as poverty, hunger, conflict and inadequate infrastructure are critical drivers of the spread of the disease in Africa (Arrehag *et al.*, 2006; Were and Nafula, 2003).

As a result of the severe threats of HIV to humanity, the Joint United Nations Programme on HIV (UNAIDS) and partners set out the 90-90-90 initiative in 2014. This initiative aimed to diagnose 90% of all HIV positive persons, provide antiretroviral therapy (ART) for 90% of those diagnosed, and achieve viral suppression for 90% of those treated by 2020 (Bain *et al.*, 2017). However, about 2.1 and 1.8 million people were newly infected in 2015 and 2017, respectively (Bain *et al.*, 2017; Anon., 2018). This posed a major setback to the 90-90-90 initiative. Out of the 2.1 million new infections recorded in 2015, Sub-Saharan Africa (SSA) alone accounted for about 1.5 million (approximately 70% of the global total) (Anon., 2016). SSA is known to be the most heavily infected region with HIV (Anon., 2006). Table 1.1 shows a summary of how the HIV epidemic has trended globally since 2005. For the three randomly picked years illustrated in Table 1.1 (2005, 2010 and 2016), SSA can be seen to be the most heavily burdened in terms of HIV infection; recording twice as high as the rest of the regions of the world put together in the following categories: (i) Number of people living with HIV (ii) New infections (iii) HIV related deaths. SSA is defined as the region of the world consisting of all African countries that are wholly or partially located south of the Sahara.

This study concentrates on Ghana; an instance of a Sub-Saharan African Country. The socio-economic conditions of Ghana are similar to that of most parts of SSA. The first case of HIV in Ghana was detected in 1986 and HIV has had a significant adverse impact on the country's economy since then. Historically, there have been varied and dwindling trends of prevalence of the epidemic in the country. Currently, due to governmental and intergovernmental organisational support, key indicators such as prevalence and incidence rates of the epidemic are diminishing, although the country remains in the generalised epidemic status.

There are well established institutions in the country with defined responsibilities for policy making, public campaigning and surveys which are necessary for acquiring knowledge and data for fighting the spread of the disease. These institutions are the National AIDS/STI Control Programmes and the Ghana AIDS Commission (GAC) which operates under the office of the President. These two institutions work closely together to tackle the epidemic in a variety of ways. They also provide the necessary data support for research works in the field of HIV /AIDS.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 1.1 Summary of the HIV situation of the World from 2005 to 2015

	Living with HIV (Millions)			Newly Infected (Millions)			HIV Related Deaths (Millions)		
Year	2005	2010	2016	2005	2010	2016	2005	2010	2016
Eastern and southern Africa	24.5	17.20	19.40	2.70	1.10	0.79	2	0.76	0.42
Western and central Africa		6.30	6.10		0.45	0.37		0.37	0.31
Middle East and North Africa	0.44	0.19	0.23	0.06	0.02	0.02	0.04	0.01	0.01
Asia and the Pacific	8.30	4.70	5.10	0.93	0.31	0.27	0.60	0.23	0.17
Latin America	1.60	1.80	1.80	0.14	0.1	0.09	0.06	0.06	0.04
Caribbean	0.03		0.31	0.04		0.02	0.27		0.01
Eastern Europe and central Asia	1.50	1.00	1.60	0.22	0.12	0.19	0.05	0.04	0.04
Western and central Europe and North America	2.00	2.10	2.10	0.07	0.09	0.07	0.03	0.03	0.02
World Total	38.37	33.29	36.70	4.16	2.19	1.83	3.05	1.50	1.00
Total for SSA	24.50	23.50	25.50	2.70	1.55	1.16	2.00	1.13	0.73
World Total Excluding SSA	13.87	9.79	11.20	1.47	0.64	0.66	1.05	0.37	0.27

(Sources: Anon. (2006), Anon. (2016), Anon.(2017a))

KNOWLEDGE TRUTH AND EXCELLENCE

Problem Statement

From the socio-economic and socio-demographic perspectives, several researchers have made concerted efforts to understand why the epidemic is trending so high in SSA and to provide recommendations for policy makers and stakeholders to act towards mitigating the situation. Several approaches have been used over the years for modelling and understanding various aspects of HIV infection in SSA. The commonest approaches used are multivariate methods such as regressions (linear and logistic), chi-square analysis as can be seen in (Nagoli *et al.*, 2010; Nattrass, 2009; Bogale *et al.*, 2009; Rogan et *al.*, 2010; Hargreaves *et al.*, 2007).

Nagoli *et al.*, (2010), developed a logistic regression model to study the factors characterising HIV vulnerability of people in two fishing villages in Mangochi District, Malawi. Their findings show that low education level and type of occupation (specifically fish trade and non-diversified livelihood) are key drivers of the epidemic in the chosen communities. Similarly, Bogale *et al.*, (2009), in their quest to study the level of HIV awareness and transmission modes among rural women in Ethiopia, highlight low education as a strong driver of HIV. These findings were made through analysis of questionnaire data using frequency distribution, cross-tabulation and chi-square analyses.

High education (Literacy), on the other hand, is equally a strong determinant of HIV infection. Using multivariate least squares regression models, Gummerson, (2013) determined that, literacy favours HIV risk behaviours such as multiple sexual partners and early marriage. Alternative methods such as Linear Regression used by De Walque (2009), Multivariable survival Analysis used by Bärnighausen *et al*. (2007) and Non-parametric Regression used by Fortson (2008) in their respective studies also linked literacy to HIV transmission-favoured behaviours such as premarital sex and multiple sexual partners. Parkhurst (2010) used a retrospective ecological comparison, trend analysis and chi-square test to investigate the relationships between the HIV infection and poverty and wealth across several African countries. The results identify both wealth and poverty as essential drivers of the epidemic. The findings show that, whereas wealth favours risky sexual behaviours such as premarital sex and multiple sexual partners, poverty drives young females into transactional sex and early marriage. Using Non-Parametric Regression, Fortson (2008) set out to estimate the relationship between HIV

KNOWLEDGE, TRUTH AND EXCELLENCE

and socio-economic status using online data in Burkina Faso, Cameroon, Ghana, Kenya and Tanzania. Wealth was found to be strongly associated with HIV positive status.

Kalichman *et al.*, (2006) applied chi-square tests, ANOVA and MANOVA to determine the relationships between social stressors and HIV/AIDS infection. Unemployment stood out as a strong driver of the epidemic. Other socio-economic factors also identified in the literature using methods already mentioned such as Linear and Logistic Regressions are being married and being a victim of disrupted marriages (De Walque, 2009) as well as being a female (Rogan et al., 2010). Urban lifestyle, including having access to media such as radio, television, newspapers, among others, are also strongly linked with HIV infection (Isiugo-Abanihe and Oyediran, 2004).

All these researches, however, have a very similar focus and findings; which is the identification of HIV drivers. At least, two issues significantly stand out which the existing researches have not addressed in the context of HIV drivers in SSA in the socio-economic perspective. These are: (1) the possibility to assess or measure the degree of impact of each of such drivers in driving the epidemic in a given population sub-group, (2) the possibility to establish whether any two or more of such factors interact or influence or drive each other in a given population sub-group. This research hypothesises that all known dominant HIV drivers in a given population do not have an equal impact on the epidemic. Furthermore, the impacts of such drivers and their relationships or interactivities are likely to vary according to socio-economic factors such as place of residence of the population under consideration.

The degree of impact is a term used by the researcher to denote a numerical assessment of how important one factor is, compared to another in a given population sub-group while the term interactivity in this research denotes a relationship between any two HIV drivers. The population subgroup as used in this research means a given class of people (for instance married people) living in a given area (for example urban area). This research is motivated by hints given by some researches of the possible complex relationships among HIV driving factors and the need for further work to establish such relationships (Whiteside, 2002; Shefer *et al.*, 2012; le Booysen, 2004; Shandera, 2007).

The aim of this research, therefore, is to develop a specific computational model for assessing the degrees of impact and the interactivity among HIV drivers using knowledge representation and data mining techniques.

Research Questions

To achieve the aim of this study, the following questions were posed:

What are the main socio-economic drivers of HIV in SSA?

How much impact does each driver have on various population subgroups?

What are the relationships among the various features of socio-economic HIV drivers in specific population subgroups?

Research Objectives

The specific objectives of this research are to:

use the Frame-based Knowledge Representation technique model and identify socio-economic HIV drivers.

develop a Computational Model to determine the impact and interactivity of the identified factor on HIV infection using a new combination of emerging techniques like Feature Maximization; and

evaluate the model using standard metric.

Research Methods and Materials

To address the stated research questions and achieve the objectives of this research, the first step involved an exhaustive systematic literature review of published research articles regarding the relationships among HIV drivers in Sub-Saharan Africa (SSA). Findings from the literature were synthesised using Frame-Based Knowledge Representation technique to depict the hierarchical relationships between socio-economic drivers and HIV infection in the context of SSA. Formally, a frame is a data structure consisting of a name for identifying the frame (frame name), slots (attributes) and facets (values). In the context of this study, the frame name denotes a given population subgroup; for example, *youth in rural areas*. The slot component takes an attribute of the given driver; for instance: *employment status.* The facet component, on the other hand, denotes the prevailing state of the stated driver. For instance, *employment status* as an HIV-driver could be in one of two states; namely *employed* or *unemployed*. Table 1.2 summarises the general structure of a single frame. Frames are powerful knowledge representation tools used commonly in the context of Artificial Intelligence.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 1.2 Structure of a Frame

	Slot	Facet
myFrame	Attr_1	Val_1
	Attr_2	Val_2
	. . . Arr_n	. . . Val_n

Attr_1, Attr_2, . . ., Attr_n are the various possible attributes of the Frame named "myFrame", whereas Val_1, Val_2, . . ., Val_n are the corresponding values of the attributes. In the Prolog programming language, this can be presented as:

```
frame(name(myName),
[Attr_1(Val_1)],
[Attr_2(Val_2)]
.
.
.
[Attr_n(Val_n)])
```

This makes it possible to query the model for knowledge it stores. Frames allow the application of important concepts such as inheritance; thereby allowing the possibility of generating networks of frames. A network of frames consists of a collection of frames connected in such a way that, there are possible hierarchies such as super Frames, Sub Frames, Individuals Instances, and so on.

The second step involved the design of computational architecture and logic of the system. The reasoning of the model is based on the Growing Neural Gas (GnG) clustering technique and Feature Maximization Metric (Lamirel *et al*., 2014). The dataset used is:

HIV specific dataset collected from twenty-eight districts in Ghana.

A neutral dataset (non-HIV specific) of approximately the same quantity and from the same districts as in (i).

To achieve the second and third objectives, the model executes in two phases. Firstly, it performs GnG on the dataset to create clusters. Secondly, thanks to Feature Maximization, the model computes the weight of each driver based on which it generates bipartite undirected graphs (contrast graphs) and ensures that only salient features appear in the formed clusters. The computed value of contrast of a driver represents its degree of impact, whereas the contrast graphs depict the interactivity of the drivers.

The combination of GnG and Feature Maximisation compute and output degrees of impact and interactivity of the socio-economic HIV drivers as follows:

Cluster 0

XX	Var0
XX	Var1
XX	Var2

Cluster 1

XX	Var1
XX	Var2
XX	Var3

Where XX are the computed degrees of impact or weights, Var0, Var1, Var2 are instances of the HIV-drivers (referred to as features) in cluster 0, and Var1, Var2, Var3 are features of HIV-drivers in cluster 1.

Formally, a bipartite graph G = (L ∪ R, E) is a graph composed of two disjoint sets of vertices L and R such that every edge from E ⊆ L × R connects one vertex of L and one vertex of R. In the context of this research, the two sets of vertices are respectively the clusters and their respective elements; thereby showing the relationship between clusters and their elements while depicting cluster to cluster relationship as depicted in Figure 1.1.

KNOWLEDGE, TRUTH AND EXCELLENCE

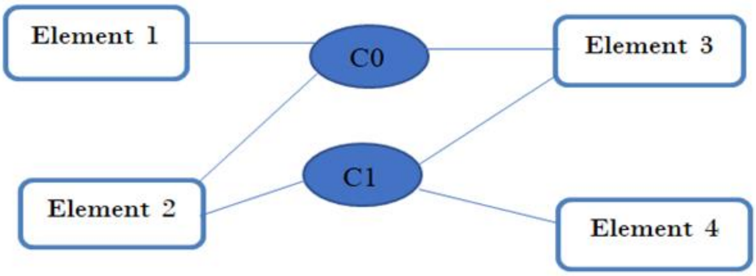


Figure 1.1 Undirected Bipartite Graph

Contribution to Science and Knowledge

The research proposes a computational model for determining the degrees of impact and interactivities of socio-economic HIV drivers. By the logic of Feature Maximization, the degree of impact (contrast) of a factor in a dataset is an indicator of the salience of that driver amongst other drivers in the dataset (Lamirel and Al Shehabi, 2015). With this reasoning, the concept of degree of impact can assess in quantitative terms the level of contribution of each driver to the spread of the epidemic in a given sub-population using data.

On the other hand, the concept of interactivities makes it possible to highlight underlying associations among drivers of the epidemic within the given dataset. Theoretically, this study goes beyond merely identifying socio-economic HIV drivers in previous works (as demonstrated in Section 1.2) to measure their impacts and interactivities. Practically, the developed model would guide stakeholders and policymakers who are concerned with the fight against HIV to make informed decisions in fashioning out strategies to combat the spread of the HIV epidemic.

Traditionally, the common concept used for assessing disease similar to the concept of degree of impact is the relative risk (risk ratio) (Irvine, 2004). The exploitation of this method, however, requires the use of background knowledge such as the size of exposed and the unexposed populations to the given driver. Such background knowledge is not easily accessible in many population sub-groups. Both concepts proposed in this work are data-driven and are therefore independent of such background facts.

KNOWLEDGE, TRUTH AND EXCELLENCE

Organisation of the Thesis

This thesis is organised into five chapters. Chapter one lays the background and outlines the problem necessitating this study. The research questions and objectives, as well as a quick summary of the methods and approaches used, are stated in this chapter. In chapter two, an extensive literature review is conducted covering the state of the epidemic in Sub-Saharan Africa, existing research findings and commonly used methods. The first part of chapter three is dedicated mainly to representing existing knowledge of relationships between the HIV epidemic and its socio-economic drivers using frame-based knowledge representation techniques. By so doing, the socio-economic drivers are identified, classified and data used for the research discussed. In the second part of chapter three, the proposed computational model is designed for determining the interactivities and degrees of impact of HIV socio-economic drivers. The model designed is implemented, analysed, and the results obtained are presented and discussed in chapter four. Finally, in chapter five, the methods, approaches, the problem being solved and results are reviewed. The chapter concludes the research by discussing the contributions and implications of the thesis.

KNOWLEDGE, TRUTH AND EXCELLENCE

LITERATURE REVIEW

Background

HIV remains a major health crisis across Africa; particularly, Sub-Saharan Africa (SSA) where it accounts for about 70 per cent of deaths (Anon., 2016; 2017a). This chapter focuses firstly on discussing the nature and impact of the epidemic across the Sub-Saharan African region. It highlights the extent to which the socio-economic aspect of the Sub-Saharan African population has wobbled in the past and is projected to still wobble in the future as a result of the HIV epidemic. The second focus of this chapter is to highlight the extent to which socio-economic and socio-demographic factors in SSA have favoured the flourish of the epidemic in the region. It further undertakes a detailed review of major scientific works that have been carried out with the aim of understanding the role of socio-economic and socio-demographic Factors in impeding the eradication of the epidemic from SSA. Lastly, the chapter identifies specific key gaps which the existing researches have missed to address.

HIV /AIDS

HIV is the abbreviation for Human Immunodeficiency Virus. It is a special kind of virus which suppresses the immune system, thereby slowing down and leading to the body's inability to fight infections and certain cancers (Anon., 2004b). HIV infection evolves into a syndrome called AIDS (Acquired Immune Deficiency Syndrome) if not diagnosed early and given effective treatment.

Mode of Transmission

HIV is known to be transmissible in the following ways (Anon., 2004b; Bertozzi *et al.*, 2006):

Through unprotected sexual intercourse with an infected partner. Transmission rate through anal sex is high than through penile-vaginal sex. Moreover, male to female transmission is more common than female to male transmission.

Through injections with syringes or needles contaminated with HIV infected blood. It can also be transmitted through blood transfusion where the blood is contaminated.

From mother to baby during pregnancy or through breastfeeding.

Transmission through sexual intercourse (unprotected sex with an infected partner) is the most dominant; accounting for about 80 per cent of cases worldwide and 90 per cent of cases in Sub-Saharan Africa (Askew and Berer, 2003; Bertozzi *et al.*, 2006).

State of HIV in Sub-Saharan Africa

Though the impact of HIV infection is greatly felt across the world, certain regions have been disproportionately negatively affected, and SSA is the most burdened. About 70 per cent of people living with HIV worldwide live in SSA with women representing about 57 per cent, and 75 per cent of the HIV infected being young people as at 2001(Bertozzi et al., 2006, Were and Nafula, 2003). In 2001, an estimated number of about 20.3 million people were HIV positive in Sub-Saharan Region, and for more a decade no much change has been seen as about 24.7 million people were still known to be suffering from the disease in 2013 (Anon., 2006; Anon., 2014). Compared with the eight major regions of the world, the impact of HIV on the Sub-Saharan Region, which has just about 10 per cent of the world's population, is exceptionally high.

Impact of HIV on Sub-Saharan Africa

Over the years, the impact of HIV has significantly been felt in various spheres of human life across the world. Its impact on SSA is particularly massive and greatly devastating right from the national level down to the individual level (Arrehag *et al.*,2006). The social and economic consequences of HIV are felt in the areas of education, industry, agriculture, transport, human resources development and many more. A good number of research works have been done in the Sub-Saharan African region with the aim of evaluating the various kinds of impacts which HIV have had on the lives of the people. These impacts are generally socio-economic (Booysen *et al.*, 2003; Arrehag *et al.*, 2006; Were and Nafula, 2003).

The following points summarise the various ways in which HIV has and continue to influence lives in SSA.

Even though the HIV epidemic does not lead to instant deaths, its patients, especially where not given early treatment, gradually become weak, less efficient and finally completely unable to work. HIV-related weaknesses and deaths of the economically active population, therefore, lead to a drastic decline in the labour force; thereby negatively affecting the demand and supply aspects of the economy (Booysen *et al.*, 2003). When people get infected, they gradually begin to lose their energy and productivity as a result of reduced body mass, energy, motivation and morale until they completely lose work capacity when the disease is fully developed (Cornia and Zagonari, 2007). The impact of this to the corporate organisation is increased expenditure and loss of revenue. This is because, when employees become morbid or die, firms lose revenue through absenteeism, healthcare cost, burial fees, payments of employee benefits, and so on (Bollinger and Stover, 1999). Firms may also lose experienced workers and consequently lose much capital, trying to replace those (Arrehag *et al.*, 2006). Figure 2.1 summarises the impact of HIV on Firms in SSA.

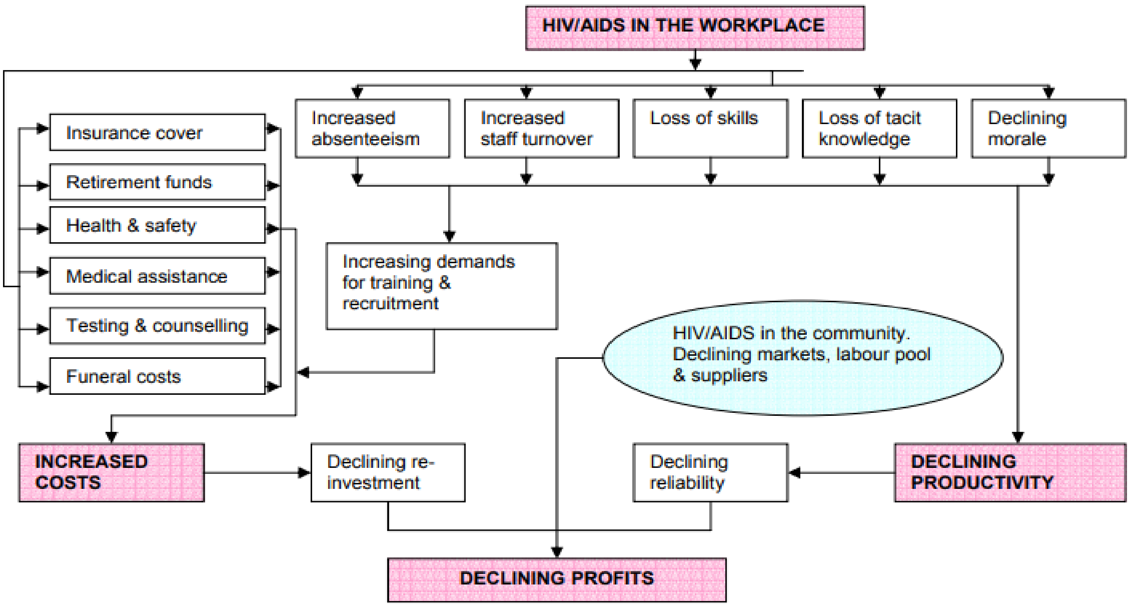


Figure 2.1 Impact of HIV on Firms (Daly, 2000)

At the household or family level, the stress due to HIV is equally enormous. HIV-related deaths and morbidity lead to high household expenditure and loss of household income as a result of medical and possibly funeral expenses for infected family members. This translates to changes in expenditure patterns and affects savings and investments (Bollinger and Stover, 1999). In some cases, the epidemic can result in the demise of both parents; thereby producing orphans and increasing economic strains of the extended family (Arrehag *et al.*, 2006). In many of such cases, households have to sell properties and in some cases borrow monies in order to survive (Arrehag *et al.*, 2006) which in turn leads to poverty, vulnerability and social burden (Were and Nafula, 2003).

High HIV infection also means high demand for health services as a result of increased demand for health treatment. This impacts directly on government expenditure as it leads to the high cost of employment and low productivity (Bollinger and Stover, 1999). In the education sector, HIV-related deaths and morbidity make it practically impossible to train and produce future work-force (Arrehag *et al.*, 2006).

Agriculture is another sector, which suffers greatly in SSA as a result of the HIV impact. The rate of loss of labour force in the agricultural sector was so significant that the Food and Agricultural Organisation (FAO) far back in the year 2000 projected a labour loss of between 10.7% and 26% by the year 2020 (Anon., 2004a). Taking into consideration that, agriculture plays a significant role on the economy of SSA (contributing to about 12% of GDP and providing employment for about 60% of its population) (Anon., 2004b), such a loss is very significant to the Sub-Saharan African region. Loss of productivity due to HIV infections of economically active people in farming communities commonly result in food shortages and malnutrition (Arrehag *et al.*, 2006).

Table 2.1 illustrates the state of agricultural labour force loss due to HIV in 2000 and projected loss by 2020.

Table 2.1 Project loss of Agricultural Labour Force due to HIV by 2020

Country in descending order of Labour Force Loss	**Projected Agricultural Labour Force Loss (%) by year**	
	2000	2020
Namibia	3	26
Botswana	6.6	23.2
Zimbabwe	9.6	22.7
Mozambique	2.3	20.0
South Africa	3.9	19.9
Kenya	3.9	16.8
Malawi	5.8	13.8
Uganda	12.8	13.7
Tanzania	5.8	12.7
Central African Republic	6.3	12.6
Cote d'Ivoire	5.6	11.4
Cameroon	2.9	10.7

(Source: Anon., (2004b)

The State of HIV in Ghana

In Ghana, the first case of HIV was reported in 1986. As a result, the government began major initiatives to track and curb the spread of the disease in the country (Akwara *et al.*,

KNOWLEDGE, TRUTH AND EXCELLENCE

2005). Amongst the first initiatives to combat the disease was the establishment of the National AIDS/STI Control Program (NACP) in 1987 (Akwara *et al.*, 2005).

Later in the year 2000, the Ghana AIDS Commission (GAC) came into being (Akwara *et al.*, 2005). GAC is responsible for the formulation of policies and strategies to provide High-level advocacy and preventive control and to lead in National policy and planning programmes in response to the HIV epidemic in Ghana. NACP, on the other hand, uses the Sentinel system to conduct surveys in antenatal clinics. This provides data and information for research and planning purposes. Currently, the sentinel survey service covers every district in Ghana.

Key modes of transmission of the epidemic in Ghana are heterosexual contact, mother to child transmission and transmission through blood and blood-related products (Akwara *et al.*, 2005; Abrefa-Gyan *et al.*, 2016). Heterosexual intercourse alone accounts for between 75 to 80 per cent of all HIV cases in Ghana. Transmission from mother to child is the second most dominant transmission mode; accounting for approximately 15 per cent of all cases while transmission through blood and blood products accounts for 5 per cent. HIV prevalence is relatively higher in densely populated areas, mining and border towns, and towns along main transportation routes.

Since 2013, the prevalence rate of HIV in the country has been on the downward trend; decreasing from 1.85% in 2013 through 1.74% in 2015 to 1.67% in 2017 (Anon., 2017b). It further projects that; prevalence would reach 1.51 by the year 2022. Despite the observed downward trend, the present prevalence state is still a matter of concern, because, by the World Health Organization (WHO) standards, countries with prevalence rates above 1% are considered as having a generalised epidemic (Anon., 2004c); signifying a high emergency. Figure 2.2 shows the estimated and projected prevalence trend of the epidemic in Ghana from 2013 to 2022. The overall HIV population in the country is projected to increase gradually from about 309,918 in 2013 through 316, 610 in 2018 to 328,364 in 2022. This increase as a result of increased survival rate from use of Antiretroviral Therapy (ART) over the last two decades (Anon., 2017b).

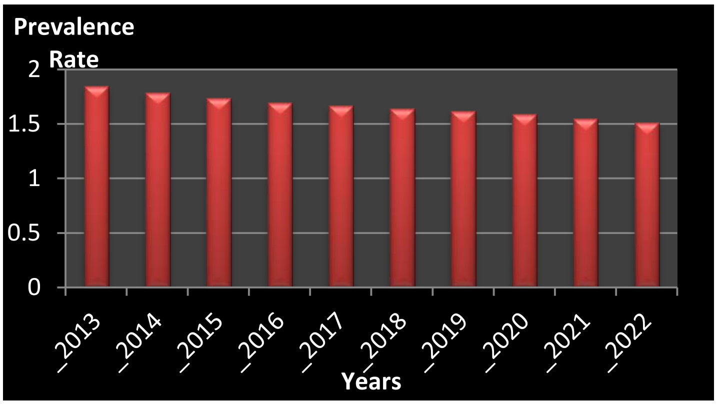


Figure 2.2 Projected Adult HIV Prevalence in Ghana (Source: Anon. (2017b)

The adverse social and economic consequences of the HIV pandemic in Ghana, like other parts of the Sub-Saharan African sub-region, are enormous. Firstly, the growing numbers of HIV infections are gradually overwhelming the health care system in the country. As mentioned earlier in the case of another part of SSA, the morbidity effect of the epidemic in the country results in very poor productivity in all sectors of the economy. The social implications of the epidemic, such as children losing parents to the pandemic in Ghana are not different from the rest of the Sub-Saharan African sub-region.

Stigma and discrimination are major contributory factors in the continued persistence of the pandemic in the country. Many people shy away from testing for the virus; thereby leading to several people not knowing their HIV statuses. As a result, due precautions are not taken, and the disease is spread blindly to others. Some of those who know about their HIV-positive statuses; however, do not open up for treatment. Such people end up becoming morbid and eventually dying.

Review of Related Works

This subsection focuses on identifying key drivers of the HIV epidemic using the systematic literature review. To do this, literature was obtained from high ranking HIV/AIDS/Health-specific databases and Google Scholar. The databases searched include AIDS and Behaviour, African Journal of AIDS Research and PubMed. Each selected article was reviewed in two dimensions. The first dimension is the Place of Residence. This refers to the level of socio-economic organisation of the place in which the research was conducted. With respect to places of residence, the papers were categorised into Rural, Urban and Mixed (Cross Residence). The cross residence is used in a context where the findings in the article are neither restricted to urban nor rural. The contexts were directly highlighted in the articles reviewed. The second dimension encompassed the Gender and Age group category. Some of the findings were youth specific while others were general. In this regard, the findings relating to gender and age group were categorised into Youth (both sexes), Youth (male), Youth (female), Male (general), Female (general) or the general population. Definitions of each of these categorisations are given as follows:

Youth: Persons who are in the age range of 15 to 24;
Youth (both sexes): This is where the findings in the given article refer to youth, irrespective of their gender;
Youth (female): This is where the findings in the given article refer to only female youth;
Youth (male): This is where the findings in the given article refer to only male youth;
Male (general): This is where the findings in the given article refer to males in general (without mention whether youth or adults);
Female (general): This is where the findings in the given article refer to females in general (without mention whether youth or adults);
Both (sexes): This is where the findings in the given article are generalised; not male or female or youth-specific.

Table 2.2 to Table 2.15 show the outcomes of the systematic literature review. In each Table, the identified HIV drivers are in bold and underlined for the sake of emphasis. The outcomes are given in Tables with the following column headings:

Article: In this column, the article being reviewed is cited;
Outcomes: This refers to the main summarised finding(s) from the reviewed article with respect to the socio-economic/socio-demographic factor being studied. The drivers are underlined and bolded in Table 2.2 to 2.14;
Method/Approach Used: This refers to the primary method or approach used in the reviewed article;
Limitation: This can either refer to what the research failed to address or weakness in the method used.
The findings in this subsection are used in building the Frame-based knowledge representation model in Section 3.2.2.

Table 2.2 Youth (Both Sexes), Cross Residence

Article	Method	Outcomes/Findings	Limitation
(Kleinschmidt *et al*., 2007)	Univariate logistic regression, Markov chain, Monte Carlo simulation, Bayesian kriging	**Unemployment** is positively associated with HIV infection; **Urban residence** is more associated with HIV infection than a rural residence.	Associations/interactivity among HIV drivers not determined; Imprecise about how much each driver contributes to the HIV /AID prevalence.
(Babalola, 2011)	Multilevel logistic regression	The **female gender** is more significantly infected with HIV than the male gender.	Imprecise about how much being a male or female contribute to HIV infection in the given population sub-group.
(Kembo, 2012)	Multivariate binary logistic regression	Persons of **age group 15-24** who are **divorced**, **widowed** or **separated** are more at risk of HIV than the unmarried.	Imprecise about how much the HIV drivers such as divorce, widowhood *etc*. identified in the study contributes to the epidemic.
(Smith, 2010)	Review of Published articles	**Poverty** and **violence** put females at higher risk of HIV infection than males.	The finding shows an association between drivers (poverty, violence, female gender and HIV /AIDS) but imprecise on how much each of them contributes to the epidemic relative to other drivers in the given sub-population.

Table 2.2 (Continued) Youth (Both Sexes), Cross Residence

Article	Method	Outcomes/Findings	Limitation
(Luke, 2005)	Logistic regression	Level of HIV infection is higher **female gender** than the male gender.	Imprecise about the risk ratio between the male and female genders.
(Pettifor *et al.*, 2007)	Maximum likelihood, probability model, sensitivity analyses	**Female gender** is at greater risk of HIV infection than the male gender.	Imprecise about the risk ratio between the male and female genders.
(Wilson *et al.*, 2011)	Review of published works	**Female gender** infected at younger ages than males.	Imprecise about the risk ratio between the male and female genders.
(Joesoef *et al.*, 2003)	Statistical analysis	**Female gender** is more likely than the male gender be HIV infected.	Imprecise about the risk ratio between the male and female genders.
(Glynn *et al.*, 2001)	Statistical analysis	HIV infection is much higher in female gender than in the male; **Being married** (Marriage) is an HIV risk factor.	Imprecise about the risk ratio between the male and female genders; Imprecise on how much marriage contributes to the epidemic relative to other drivers.
(Edelstein *et al.*, 2015)	Multivariate linear analyses	**Marriage** is associated with high HIV infection for those aged 15-19; **Age group 20-24** is strongly associated with HIV infection.	The measure of how much marriage contributes to the epidemic in the given sub-population is unknown.

Table 2.3 Youth (Female) in Rural Places of Residence

Reference	Method	Outcome	Limitation
(Yahya-Malima *et al.*, 2006)	Chi-square, logistic regression	High HIV prevalence was among **females** aged between 15–19 years; **Disrupted marriage** was associated with high HIV infection compared to other marital statuses; Formal **education** was associated with high HIV infection.	Imprecise on how much impact each of the drivers such as female gender, disrupted marriage and formal education contribute to the epidemic in the given sub-population; Imprecise on whether the identified factors interact or associate with or influence each other.
(Michelo, Sandøy and Fylkesnes, 2006)	Logistic regression	Higher **education** is less associated with HIV infection than lower education.	Imprecise on how much higher education contributes to the epidemic relative to other drivers in the given sub-population.
(Clark, 2004)	Logistic regression	**Age group 15-19** who are **married** have greater odds of HIV infections than their unmarried counterparts.	Imprecise on how much the married and the unmarried contribute to the epidemic relative to each other and relative to other drivers in the given sub-population.

Table 2.4 Youth (Female) in Urban Places of Residence

Reference	Method	Outcome	Limitation

(Kayeyi, Sandøy and Fylkesnes, 2009)	Stratified random cluster sampling, Multi-level mixed-effects regression models	Lower **education** is more associated with high HIV prevalence than higher education; There is a higher HIV risk in **urban** than in rural areas.	The precise measure of impact on HIV of Lower education relative to higher education is not known.
(Edelstein *et al.*, 2015)	Multivariate linear regression analysis	HIV infection is positively associated with being **divorced**, **widowed**, **separated.**	Imprecise on how much impact each of the identified factors (divorced, widowhood and separation make on the epidemic.
(Michelo, Sandøy and Fylkesnes, 2006)	Logistic regression	Higher **education** is less associated with HIV infection than lower **education**.	Imprecise on how much higher education contributes to the epidemic relative to other drivers in the given sub-population.
(Gabrysch, Edwards and Glynn, 2008)	Multivariate linear regression analyses	Low and **middle socio-economic** Status (SES) is associated with HIV; Higher education is positively associated with HIV prevalence.	The relative measure of how much impact both low and middle SES as well as high education makes on HIV infection.

Table 2.5 Youth (Female) Cross Residence

Article	**Method**	**Outcome**	**Limitation**

(Johnson *et al.*, 2009)	Logistic regression models	**Higher education** is less associated with HIV infection than lower education.	No precise measure of how much higher education contributes to the epidemic relative to low education.
(Karim *et al.*, 2012)	Multivariable proportional hazards model	HIV infection among **adolescents** is high.	The word **high** does not give a precise measure of the extent to which adolescents are prone to the epidemic relative to others.
(MacPhail, Williams and Campbell, 2002)	Review of published articles	**Gender inequality** exposes females to HIV.	The extent to which gender inequality exposes females to HIV infection relative to other drivers is not expressed.
(Mbirimtengerenji, 2007)	Review of published articles	**Poverty** influence **early marriage,** which predicts the high possibility of HIV in females.	The precise measure of contribution by identified drivers unknown.
(Smith, 2010)	Review of published articles	Level of HIV infection is higher among young **females** than young males; **Poverty** influences the HIV status of young females.	The extent to which poverty influences HIV in females is not measured.

Table 2.6 General Population (Rural)

Articles	**Method**	**Outcome**	**Limitation**

(Gómez-Olivé *et al.*, 2013)	Stratified random sampling, probit regression	The most HIV infected age group for both sexes is **<u>35-40.</u>**	The measure of association between age group 35-39 and HIV infection in the given population subgroup not precise.
(Welz *et al.*, 2007)	Unconditional logistic regression.	HIV infection is higher in **<u>females</u>** than males.	The precise measure of HIV infection on female gender relative to male gender unknown.
(Rosen *et al.*, 2008)	Unconditional logistic regression	**<u>Females</u>** are more infected with HIV than males.	The precise measure of HIV infection on female gender relative to male gender unknown.
(Odimayo, *et al.*, 2010)	Clinical screening	**<u>Divorce</u>**, **<u>farming occupation</u>** and **<u>low education</u>** are strongly associated with HIV infection.	Imprecise on how much each identified factor contributes to HIV infection in the given population subgroup; Unclear whether the identified factors interact or not.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 2.6 (Continued) General Population (Rural)

Reference	**Method**	**Assertion**	**Limitation**
(Wallrauch, Baernighausen and Newell, 2010)	Multivariable logistic regression	Both **stable marriage** and **disrupted marriage** were strongly associated with HIV infection.	The precise association between the identified factors and HIV infection is not expressed.
(Barnighausen *et al.*, 2007)	Statistical analysis	**Low education** predicts High HIV infection.	The measure of how much low education contributes to the epidemic is not expressed.
(Zulu, Kalipeni and Johannes, 2014)	Inverse distance weighting (ArcGIS)	**Rural residence** is associated with high HIV prevalence for the age group 30-34.	The association of rural residents to the epidemic is not precise.
(Magadi, 2013)	Logistic regression	**Wealth** is associated with a higher likelihood of HIV infection than poverty.	The extent to which the identified factors drive the epidemic is vaguely quantified.
(Hajizadeh *et al.*, 2014)	Relative and generalised concentration indices	HIV is concentrated among the **wealthy**.	The extent to which wealth drives the epidemic is vaguely quantified.
(De Walque *et al.*, 2005)	Statistical analysis	**High education** is associated with less HIV infection.	The linguistic variable **less** is not able to express the precise relationship between HIV infection and high education.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 2.7 General Population (Urban)

Articles	**Method**	**Outcome**	**Limitation**
(Madise *et al*., 2012)	Logistic regression models	**Divorce** and **widowhood** are more positively associated with HIV infection than the unmarried and the Married.	The extent to which each of the identified factors contributes to the epidemic not quantified.
(Auvert *et al*., 2001)	Logistic regression	The **married** and those with **disrupted marriages** are strongly associated with HIV infection than the never married. **Ages 25-49** are associated with higher HIV prevalence than those of 15-24.	The extent to which each of the identified factors contributes to the epidemic not quantified. The possible interactivities among identified drivers not determined.
(Magadi, 2013)	Logistic regression	**High educated** poor people are more associated with higher HIV positivity than the higher educated wealthy; The **educated poor** have a higher likelihood of HIV positivity than the **uneducated poor;** The **urban poor** have a higher likelihood of HIV than the non-poor.	It shows interactivity between educations, place of residence and wealth in driving the epidemic but the extent to which wealth and education each contributes to the epidemic is not quantified.
(González *et al*., 2012	Clinical screening, Statistical analysis	**Age range 28–47** compared to age group 18–27 is more infected.	A measure of HIV infection among the age group 28–relative to age group 18–27 is unknown.

Table 2.8 General Population Cross Residence

Articles	Method	Outcome	Limitation
(Barankanira *et al.*, 2016)	Geostatistics (kernel density estimation, Spatial cluster estimation), logistic regression.	The **female gender**, **high education**, **widowhood**, **wealth**, both **disrupted** and **stable marriage** are strong predictors of HIV infection.	The relative contribution of each identified factor not quantified; Possible interactivity of identified factors not determined.
(Brodish, 2015)	Multilevel logistic regression models	**High wealth** is a strong predictor of HIV status.	The extent to which Wealth contributes to the epidemic vaguely expressed.
(Opio, Muyonga and Mulumba, 2013)	Chi-square test, bivariate and multivariate analysis	HIV infection was highest among **widows**, **age group 25 years and older** and the **married.**	The relative contribution of each identified factor not quantified; Possible interactivity of identified factors not determined.
(Mill and Anarfi, 2002)	Thematic analysis	**Poverty** is a strong determinant of HIV.	The extent to which Poverty contributes to the epidemic vaguely expressed.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 2.8 (Continued) General Population Cross Residence.

Reference	**Method**	**Outcome**	**Limitation**
(Asiedu, Asiedu and Owusu, 2012)	Logistic regression	HIV infection is higher in **female gender** as compared to the male gender; **Urban residence** has a higher likelihood of being HIV positive than their rural counterparts.	The extent to which female gender contributes to the epidemic relative to the male gender not precisely expressed
(Shisana *et al.*, 2004)	Logistic regression	**Poor unmarried** people are at higher risk of HIV than poor married people; **Wealthy married** people are more at risk of HIV infection than wealthy unmarried.	The extent to which marriage and wealth contribute to the epidemic not precise.
(Rehle *et al.*, 2007)	Clinical screening	HIV prevalence among **the widowed** is higher than those who were either married or single.	The extent to which widowhood contributes to the epidemic is not precise.
(Buvé *et al.*, 2001)	Clinical screening	The prevalence of **female gender** is higher than in men.	The extent to which female gender contributes to the epidemic relative to the male gender not precisely expressed.
(Tanser *et al.*, 2009)	Kulldorff spatial scan statistic (Bernoulli model), Geolocation, 2D Gaussian kernel	Prevalence is higher in more **urban areas** than in the inaccessible rural area.	Factors responsible for the difference, not quantified.

Table 2.8 (Continued) General Population Cross Residence

Articles	Method	Outcome	Limitation
(Todd *et al.*, 2006)	Logistic regression	**Divorce** and **widowhood** predict a higher likelihood of HIV	Individual impact of each factor on HIV infection is not determined.
(Mermin *et al.*, 2008)	Logistic regression	The **female gender**, broken marriage, **low education**, **urban residence** and **high wealth** are strongly associated with the HIV epidemic.	The relative contribution of each identified factor not quantified; Possible interactivity of identified factors not determined.
(Kimanga, Ogola and Umuro, 2014)	Clinical screening, Bivariate and multivariate linear regression analysis	HIV prevalence is higher in **female gender** than males; HIV prevalence is higher among the **disrupted married** than the stable married or never married or the cohabiting; There is a higher HIV infection in **urban centres** than in rural places.	The relative contribution of each identified factor not quantified; Possible interactivity of identified factors not determined.

Table 2.8 (Continued) General Population Cross Residence.

Articles	**Method**	**Outcome**	**Limitation**
(Coburn, Okano and Blower, 2013)	Multivariate logistic regression	**Urban life**, **female gender**, **age group 30-39**, **low education** and **low education** are strong predictors of the HIV infection.	The relative contribution of each identified factor to HIV infection not quantified; Possible interactivity among the identified factors not determined.
(Hoshi *et al.*, 2016)	Clinical screening, Kulldorff statistic	**Age group 30–34** compared to age group 15–19 is more infected; More **Females** are infected than males.	The precise impact of age group 30-34 on HIV infection relative to 15-19; as well as the impact of female gender relative to male gender is not determined.
(Boerma *et al.*, 2002)	Statistical analysis	**Marital Status** Predicts high HIV infection.	Imprecise on the impact of marital status on HIV infection.
(Temam and Ali, 2012)	Chi-square statistics, Logistic regression	**Disrupted marriage** predicts higher HIV prevalence than a stable marriage.	Imprecise on the impact of disrupted marriage on HIV infection relative to others.
(Buor, 2005)	Linear regression	**High education** is an important determinant of HIV infection.	Imprecise on the impact of high education on HIV infection.

Table 2.8 (Continued) General Population Cross Residence

Reference	**Method**	**Assertion**	**Limitation**
(Kalichman *et al.*, 2006)	Linear Regression, Chi-square, ANOVA, MANOVA	**Low education** and **Unemployment** are associated with HIV infection.	The relative contributions of each identified driver to HIV infection is not quantified; Possible interactivity among the identified factors not determined.
(Lau and Muula, 2015)	Review of published work	**High wealth** directly correlates with high HIV infection.	Imprecise on the impact of high wealth on HIV infection.
(Agwu *et al.*, 2011)	Logistic linear regression, randomized block design and Pearson's Chi-square	**Widowhood**, Peasant farming, **illiteracy** and **low literacy** are strong predictors of HIV infection.	The impact of each identified factor on the epidemic not precisely measured; Possible interactivity among identified factors not studied.
(Piot, Greener and Russell, 2007)	Review of published articles	**Gender inequality** exposes females to HIV than males.	The extent to which Gender inequality exposes females to HIV imprecise
(Shelton, Cassell and Adetunji, 2005)	Statistical analysis	**Wealth** is a strong driver of HIV infection compared to poverty.	Impact of wealth on the epidemic not precisely measured.

Table 2.8 (Continued). General Population Cross Residence.

KNOWLEDGE, TRUTH AND EXCELLENCE

Article	**Method**	**Outcome**	**Limitation**
(Hargreaves and Glynn, 2002)	Review of published articles	**Low education** is associated with high HIV infection.	The precise impact on low education on HIV not captured.
(Temah, 2009)	Review of published articles	**Poverty**, **urban life** and **gender inequality** exposes females to HIV than males	The impact of each identified factor on the epidemic not precisely measured; Possible interactivity among identified factors not studied.
(Higgins, Hoffman and Dworkin, 2010)	Review of published articles	**Female gender** is more vulnerable to HIV than the male gender.	The impact of Female gender on the epidemic not precisely measured.
(Fox, 2010)	Review of published articles	**Poverty** is a strong driver of HIV in Africa	The impact of poverty on the epidemic not precisely measured.
(Wabiri and Taffa, 2013)	Univariate and multivariate logistic regression	**Poverty** and **female gender** are strong drivers of HIV	Impact of each driver on the epidemic not precise.

Table 2.8 (Continued) General Population Cross Residence

References	**Method**	**Assertions**	**Limitation**
(Gillespie *et al.*, 2007)	Review of published articles	The influence of **poverty** in HIV acquisition is not greater than the influence of **wealth.**	The precise measure of impact poverty and wealth on HIV infection is not captured.
(Glynn *et al.*, 2004)	Statistical analysis	The **more educated** are at increased risk of HIV infection.	The level of impact of more education on HIV infection is not expressed.
(Lopman *et al.*, 2007)	Chi-squared tests, Logistic regression	Low **socio-economic Status (SES)** is associated with HIV infection.	The level of association of low SES on HIV infection is not expressed.
(Tladi, 2006)	Linear regression, Chi-squared tests	**Poverty** compared to wealth is a stronger predictor of HIV infection.	The precise measure of the impact of poverty and wealth on HIV infection is not expressed.
(Joesoef *et al.*, 2003)	Statistical analysis	**Males and females older than 44** years almost have an equal likelihood of getting infected.	The precise risk levels of both sexes to HIV infection is not expressed.

Table 2.8 (Continued) General Population Cross Residence

References	**Method**	**Assertion**	**Limitation**
(Gumbe *et al*., 2016)	Clinical screening, Multivariate logistic regression	The **female gender** and **low literacy** were observed to have a strong association with HIV /AIDS.	The precise measure of the impact of female gender and low literacy on HIV infection is not expressed.
(Oluoch *et al*., 2011)	Statistical analysis	**Widowhood** and **urban** life are strongly associated with higher HIV infection.	The precise measure of the impact of widowhood and urban life on HIV infection is not expressed.
(Tanser *et al*., 2009)	Kulldorff spatial scan statistic (Bernoulli model), Geolocation, 2D Gaussian kernel	The more **urban places** are more associated with higher HIV infection than the more rural areas.	The association of urban life with HIV infection is vaguely stated.
(Aulagnier *et al*., 2011)	Logistic Regression	The **urban community** is strongly associated with HIV infection.	The association of urban life with HIV infection is vaguely stated.

Table 2.8 (Continued) General Population Cross Residence.

Articles	**Method**	**Outcome**	**Limitation**
(Hargreaves *et al.*, 2008)	Review of published articles	**Urban residence** is appositively associated HIV infection.	The association of urban life with HIV infection is vaguely stated.
(Messina *et al.*, 2010)	Poisson-based spatial scan statistic, inverse distance weighting	**Urban residence** is positively associated with HIV infection.	The association of urban life with HIV infection is vaguely stated.
(Hajizadeh *et al.*, 2014)	Relative and generalised concentration indices (RC and GC)	**The urban residence** itself is a factor, contributing to the concentration of HIV among the **wealthy**.	The association of urban life with HIV infection is vaguely stated.
(Baidoo *et al.*, 2012)	Statistical Analysis	**Females** are more infected with HIV than men.	The contribution of the female gender to the epidemic not precise.
(Chijioke and Akani, 2014)	Statistical Analysis	The **Females** are more infected with HIV than men; **Poverty** influences HIV infection among women than men.	The contribution of female gender and poverty to the epidemic not precise.
(Msamanga *et al.*, 2006)	Clinical screening, Multivariate Binomial Regression	**Urban residence** is associated with a higher prevalence of infection.	The association of urban life with HIV infection is vaguely stated.

Table 2.9 Female (General) in Rural areas

Articles	Method	Outcome	Limitation
(Obi *et al.*, 2011)	Statistical analysis	Women of **Age Group 25-34** Have the highest HIV prevalence.	The degree of HIV-risk of age group 25-34 is imprecise.
(Hargreaves *et al.*, 2007)	Logistic regression	**Women of low education** are particularly at risk of HIV infection.	The extent to which low education contributes to HIV infection is not precise.
(Welz *et al.*, 2007)	Unconditional logistic regression	**Age group 25-29** is most at risk of HIV infection.	The degree of HIV-risk of age group 25-29 is imprecise.
(Rosen *et al.*, 2008)	Unconditional logistic regression	**Females age group 25-29** are most HIV-infected.	The degree of HIV-risk of age group 25-29 is imprecise.
(Nel *et al.*, 2012)	Descriptive statistics, Poisson distribution	**Low education** level is strongly associated with high HIV infection	The extent to which low education contributes to HIV infection is not precise.
(Boerma *et al.*, 2003)	Statistical analysis	**Age group 25-29** has the highest odds of HIV infection for women; **Disrupted marital status** has the highest odds of HIV infection	The degree of HIV-risk of age group 25-29 is imprecise; The extent to which disrupted marriage contributes to HIV infection is imprecise.

Table 2.10 Female (General) in Urban Areas

References	Method	Outcome	Limitation

(Amornkul *et al.*, 2009)	Clinical screening	**Widowhood**, **employment** and age **group 25-29** strongly predict high HIV status.	The impact of each of the identified factors on HIV infection imprecise.
(Yahya-Malima *et al.*, 2006)	Chi-square, Logistic regression	**Disrupted marriages** and **formal education** are associated with very high HIV infection.	The impact of each of the identified factors on HIV infection imprecise.
(Ramjee *et al.*, 2016)	Clinical screening	**Unmarried Females younger than 25** years are at high risk of HIV-infected.	The extent to which not being married contributes to HIV infection is not precise.
(Kimani *et al.*, 2011)	Descriptive statistics and multivariate logistic regression	Attainment of **high education** predicts high HIV infection.	The extent to which high education contributes to HIV infection is not precise.
(Sing and Patra, 2015)	Bivariate and multivariate logistic regression	Age **group 35-39, 40-44** and **45-49** are more HIV -infected compared to others; **No or low-education** have a higher likelihood of HIV infection.	The extent to which the identified drivers contribute to HIV infection is not precise.
(Clark, Bruce and Dude, 2016)	Statistical analysis	**Early Marriage** predicts low **education** level in young Females, which leads to the risk of HIV infection.	Imprecise how much early marriage contributes to the epidemic.

Table 2.11 Female (General) Cross Residences

Articles	**Method**	**Outcome**	**Limitation**

(Morison, 2001)	Review of published articles	**Poverty** predicts a high prevalence of HIV in females; **Low education** predicts a high prevalence of HIV in females.	The impact of poverty and Low education on HIV infection is imprecise; Possible interactivity between Poverty and low education is not studied.
(Schur *et al.*, 2015)	Multivariate logistic random-effects models	**Poverty** predicts high HIV prevalence for women.	The impact of poverty on HIV infection is imprecise.
(Kiptoo *et al.*, 2009)	Clinical/Lab procedures	**Age group 31-35** years and widows had the highest prevalence compared to other marital statuses.	The impact of age group, 31-35 and widowhood on HIV infection, is imprecise.
(Rehle *et al.*, 2007)	Clinical/Lab procedures	HIV prevalence is highest in females of **ages in the range 20-29**.	The impact of the age group 20-29 on HIV infection is imprecise.
(Luke, 2005)	Logistic regression	Gender inequality exposes females to HIV infection.	The extent to which gender inequality drives the epidemic is not quantified.
(Msisha *et al.*, 2008)	Weighted logistic regression	Professional women are at high risk of HIV acquisition.	The degree of being a professional on HIV infection not determined.

Table 2.11 (Continued) Female (General) Cross Residences

References	**Method**	**Outcome**	**Limitation**

(Bertrand, 2016)	Logistic regression	**Disrupted marriage**, urban residence, Low education, no education, **unemployment** and Poverty predict High risk of HIV infection.	The degree to which each factor drives the epidemic is imprecise; Possible interactivities among the identified factors not determined.
(Agüero and Bharadwaj, 2014)	Review of published papers	**Low education** is strongly associated with higher HIV infection.	The degree to which each low education drives the epidemic is imprecise.
(Rodrigo and Rajapakse, 2010)	Review of published papers	**Poverty** puts females at a disadvantage for HIV infection.	The extent to which poverty drives the epidemic is not quantified.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 2.12 Male (General) Rural

Articles	**Method**	**Outcome**	**Limitation**
(Okiria *et al.*, 2014)	Poisson regression analysis	Men with **disrupted** marriages are at higher risk of HIV infection.	The impact of disrupted marriages on HIV not quantified/measured.
(Welz *et al.*, 2007)	Unconditional logistic regression	Highest HIV prevalence is found among **men aged group 30-34**.	The extent to which Age group 30-34 is associated with the epidemic not determined.
(Abebe *et al.*, 2003)	Logistic regression, clinical screening	**Higher education**, **age group 25-29**, **merchants**, **administrators**, **factory workers** and **supportives** (labourers and artisans) are associated with high HIV prevalence.	The degree to which each listed factor drives the epidemic is imprecise; Possible interactivities among the listed factors not determined.
(Boerma *et al.*, 2003)	Logistic regression	Age **group 30-34** and **men with disrupted marital** status have the highest odds of HIV infection.	The extent to which age group 30-34 and disrupted marriage contribute to the epidemic is imprecise.
(Amornkul *et al.*, 2009)	Clinical screening, logistic regression	**Age group 30-34** and those employed are associated with high HIV infection.	The extent to which age group 30-34 and employment contribute to the epidemic is imprecise.

Table 2.13 Male (General) in Urban

Articles	**Method**	**Assertions**	**Limitation**
(Abebe *et al.*, 2003)	Logistic regression, clinical screening	**Age group 25-29** is most associated with HIV infection	The degree to which Age group 25-29 associates to the epidemic relative to other factors not quantified.
(Michelo, Sandøy and Fylkesnes, 2006)	Logistic regression	There is a reduced risk of HIV infection for the more educated than, the **less educated young rural Males**.	By what degree are both the more and the less educated still associated with the epidemic.

Table 2.14 Male (General) Cross Residence

Articles	**Method**	**Assertion**	**Limitation**
(Abebe *et al.*, 2003)	Logistic regression, clinical screening	HIV infection is higher in **urban centres** than in rural areas.	The question "how high is urban centres associated with HIV than rural centres?" cannot be answered.
(J. Madise *et al.*, 2012)	Logistic regression	Men **aged 35–39** years old had the highest odds of being HIV positive.	The degree to which Age group 35-39 contributes to the epidemic is quantified.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 2.14 (Continued) Male (General) Cross Residence

Articles	Method	Outcome	Limitation
(Rehle *et al.*, 2007)	Clinical screening	HIV prevalence is highest in **males of ages** in the **range 30-39**.	The impact of age group 30-39 on HIV infection is imprecise.
(Buvé *et al.*, 2001)	Laboratory procedures	HIV prevalence is highest in **males of ages** in the **range 30-39**.	The impact of age group 30-39 on HIV infection is imprecise.
(Tanser *et al.*, 2009)		HIV prevalence is highest among **men aged 30-34**.	The impact of age group 30-34 on HIV infection is imprecise.
(Msisha *et al.*, 2008)	Kulldorff spatial scan statistic (Bernoulli model), Geolocation, 2D Gaussian kernel	**Unemployed** men are at greater risk of HIV than men with other occupations.	The impact of age group unemployment on HIV infection is imprecise.
(Kimanga, Ogola and Umuro, 2014)	Clinical screening, bivariate and multivariate analysis	HIV prevalence peaks at **45-49 year group** for men.	The impact of age group 45-39 on HIV infection is imprecise.

KNOWLEDGE, TRUTH AND EXCELLENCE

Summary of Systematic Literature Review

The articles reviewed spanned two decades; ranging from 2001 to 2016. From the literature, several methods have been used to study the HIV epidemic in SSA with various focuses. In general, the focus has been to identify the relationships between HIV and specific social and economic factors such as gender, Age, Education Level, Place of Residence, amongst others. Methods used could broadly be classified as mathematical, non-mathematical and combined. The mathematical methods include Logistic Regression, Linear Regression, Chi-Square and others. The non-mathematical methods include questionnaires, interviews, focus group discussions, thematic analysis and systematic literature reviews. Some of the researches also used a combination of both mathematical and non-mathematical methods where non-mathematical methods are used to gather facts and data before applying the mathematical methods for modelling and analysis.

The limitations across all reviewed articles are as follows:

Lack of precise way of measuring/assessing/quantifying the effect or impact of individual socio-economic HIV drivers. Majority of the authors relied on traditional measures such as "prevalence" and "incidence" to make conclusions with regards to the population subgroups, which are most infected with the epidemic. HIV prevalence is defined by the UNAIDS as the number of infections at a particular point in time expressed as a percentage whereas HIV incidence is the number of new HIV infections arising in a given period in a specified population (Anon., 2011). These measures are expressed relative to the population size under study and are therefore inadequate to highlight the underlying influence or impact of factors driving the epidemic.

The relative risk is a concept which makes it possible to assess the level of risk associated with a particular risk factor or driver of the disease. To use this measure, however, one needs to know the number of individuals exposed to the risk factor and the number who are not. This information is not feasible when the study concerns a large community of people, and only samples and assumptions could be used, which may be less accurate. Also, risk ratio only shows the relationship between a risk factor and the disease but not the relationship between the risk factor and other risk factors. As a result, the findings in the majority of the reviewed articles turn to the use of linguistic variables such as "high", "most", "less" to express the degree of impact of the observed HIV drivers which are imprecise.

Inability to determine the interactivity of observed drivers. A good number of the reviewed articles identified and listed several drivers acting on the given population at the same time. It is, however, not clear whether or not some of them interact or influence each other.

Review of Common Methods used in Related Works

Statistical Analysis

Statistics deals with techniques for collecting, organising, computing, analysing, interpreting and presenting numerical data (Larson, 2006; Jaggi., 2003). The two major aspects of statistics are descriptive and inferential. Descriptive statistics use graphical and numerical methods for describing data while inferential statistics involve procedures for inferring meaning from the data (Larson, 2006; Jaggi, 2003).

Characteristics of interest for which data is collected for the study of an entity are termed as variables. For instance, in a study involving human individuals, variables of interests may include measurable qualities such as gender, height and weight. Variables are broadly classified as *discrete (categorical)* or *continuous*: Discrete variables can assume an only a specific set of values, whereas continuous variables can assume an infinite set of values (Larson, 2006). Further, a discrete variable can be described as nominal if its set of possible values do not have any form of the natural order. The variable “gender” may take the values “male” or “female”. These are not ordered in any way. On the other hand, the variable “class” (of a graduating student) may take the values “first”, “second” or “third”. Such type of variables is termed “ordinal”.

Three common measures on a variable are distribution, central tendency and dispersion. (Jaggi., 2003). Distribution for a given variable is the possible values of that variable and their corresponding frequencies. The distribution can be measured by skewness or kurtosis. The Mean, Mode and Median are the key indicators of the central tendencies of a variable.

Dispersion, on the other hand, is an indicator of how closely packed together or scattered apart the data points are describing a given variable occurs. The conventional measure for this is the standard deviation.

In the literature, several HIV research works employed statistical analyses. These include Clark *et al*., (2006), Chijioke and Akani (2014), González *et al*., (2012) and Baidoo *et al*., (2012).

Analysis of Variance (ANOVA)

ANOVA is a statistical technique for testing the similarities or differences between samples based on their means. Several modifications of this technique exist to take care of various cases in terms of the number of dependent/independent variables affected (Watkins, 2016; Lane, 2017). One-way ANOVA is applied when the dependent variable is affected by one independent variable. However, when the dependent variable is affected by two independent variables, two-way ANOVA; a modified variant of the one-way ANOVA is applied.

Let there be m samples S_1, S_2 and S_3, as shown in Table 2.15 where each sample is characterised with values x_1, x_2, . . . , x_n. Let the notation S_{xi}^{j} denote the i^{th} value of the j^{th} sample.

Table 2.15 Generalised Data in Support for ANOVA

Row	**S_1**	**S_2**	**. . .**	**S_m**
1	x_1	x_1	. . .	x_1
2	x_2	x_2	. . .	x_2
	. . .	. . .	. . .	. . .
n	x_n	x_n		x_n

ANOVA can be used to test the statistical difference between the samples (if any) by the following steps:

State null hypothesis (H_0) and the alternate hypothesis (H_0) to be tested as follows:

(H_0) : There is no significant statistical difference among the *m* samples given.

(H_1) : There is a significant statistical difference among the *m* samples given.

Compute $S_{\bar{x}}^{j}$: the mean value of the j^{th} sample as follows:

KNOWLEDGE, TRUTH AND EXCELLENCE

$$S_{\bar{x}}^{j} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2.1)$$

Compute Grand (Overall) Mean μ of all observations as follows:

$$\mu = \frac{1}{m}(S_{\bar{x}}^{1} + S_{\bar{x}}^{2} + \ldots S_{\bar{x}}^{m}) = \frac{1}{m}\sum_{j=1}^{m} S_{\bar{x}}^{j} \qquad (2.2)$$

Compute the Sums of Square Between (SSB) as follows:

$$\text{SSB} = n(S_{\bar{x}}^{1} - \mu)^2 + n(S_{\bar{x}}^{2} - \mu)^2 + \ldots + n(S_{\bar{x}}^{m} - \mu)^2 = (\sum n\left(S_{\bar{x}}^{j} - \mu\right)^{2}$$

(2.3)

Compute Sum Squared Error (SSE) for each sample; Find the sum of SSE across samples:

$$\text{SSE} = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(S_{xi}^{j} - S_{\bar{x}}^{j}\right)^{2} \qquad (2.4)$$

Compute the degrees of freedom (df). The df between groups (samples) is computed as the total number of samples minus 1. That is:

$$df_{SSB} = m - 1. \qquad (2.5)$$

The df of the within groups (samples) is computed as a number of observations minus the number of samples. That is:

$$df_{SSE} = m * n - m \qquad (2.6)$$

(given that n is constant for each group)

Compute Mean Squares (MS). MS between groups is SSB divided by df_{SSB}. That is:

$$MS_{SSB} = \frac{SSB}{df_{SSB}} \qquad (2.7)$$

MS of SSE=Sum Square Error divided by df_SSE. That is:

$$MS_{SSE} = \frac{SSE}{df_{SSE}} \qquad (2.8)$$

Finally, compute the F-statistic as follows:

$$F = \frac{MS_{SSB}}{MS_{SsE}} \qquad (2.9)$$

To take a decision, the F statistic is compared with the critical value in the table of probabilities for the F distribution. A null hypothesis is rejected if F is greater or equal to critical value; else it is sustained. Kalichman *et al.* (2006) used ANOVA to investigate relationships between socio-economic stressors on HIV transmission in urban South Africa.

Logistic Regression

Logistic regression is a powerful technique for expressing *the probability of occurrence of an event given* a particular value of a predictor variable (Sommet and Morselli, 2017; Park, 2013). It is useful in modelling the relationship between the independent variable(s) and the dependent variable(s) where the desired outputs of the dependent variable are strictly categorical (Sommet and Morselli, 2017; Park, 2013; Sperandei, 2014). Park (2013), distinguishes between two classes of logistic regression; namely binary and multinomial. Binary logistic regression is applicable when the relationship being modelled is between a dichotomous dependent variable and an independent continuous or categorical variable. Multinomial logistic regression, on the other hand, is used when the dependent variable is non-dichotomous and comprises of more than two categories. Given a dependent variable Y and an independent variable X with k categories such that X=X$_1$, X$_2$, ... X$_k$, the probability of Y given X is given in the case of simple (univariate) logistic regression as shown in Equation (2.10) and complex (multivariate) logistic regression as shown in Equation (2.11) (Park, 2013):

$$P(Y|X) = \frac{1}{1+e^{-(\alpha+\beta\chi)}} \tag{2.10}$$

$$P(Y|X) = \frac{1}{1+e^{(\alpha+\beta1\chi1\,+\,\ldots+\beta k\chi k)}} \tag{2.11}$$

Where *α* and *β* are the parameters of logistic regression. This model fits an S-shaped logistic curve through the data; indicating the probability of the dependent variable occurring given the predictor variable. A logistic curve with *α=0* and *β=1* is shown in Figure 2.4. Instances of previous HIV researches which used univariate Logistic regression include Kleinschmidt *et al.* (2007), Wallrauch *et al.* (2010), Wabiri and Taffa (2013) and Kimanga *et al.*, (2014). Several of the researches reviewed in section 2.4 also made use of multivariate binary logistic regression (Kembo, 2012), (Coburn *et al.*, 2013) (Wabiri and Taffa, 2013), (Gumbe *et al.*, 2015) and (Kimanga *et al*., 2014).

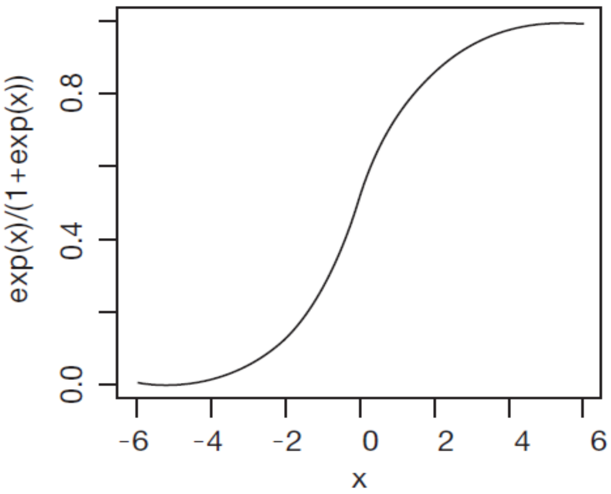


Figure 2.3 Logistic curve (source Park, 2013)

Several other variants of logistic regression used in related works as demonstrated in Section 2.6 include unconditional logistic regression (Welz *et al*., 2007) and (Rosen *et al*., 2008), weighted logistic regression (Msisha *et al*., 2008) and multilevel logistic regression (Babalola, 2011; Brodish, 2015).

Linear Regression

Linear Regression Analysis investigates the relationship between a single dependent variable and one or more independent variables (Shi and Conrad, 2009). It depicts the relationship between a variable whose value is being predicted or estimated (dependent variable) and the variable whose values are used to predict or estimate it (independent variables). Regression analysis is commonly used to perform such tasks as modelling, prediction and estimation (Shi and Conrad, 2009).

Given a data set with two labels X and Y where X={x_1,x_2, . . .,x_n} and Y={y_1, y_2, . . .,y_n} and Y depends on X, a simple linear regression modelling the relationship between X and Y using a line of best fit is given as follows (Shi and Conrad, 2009):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (2.12)$$

Where β_0 is the intercept of the linear model, β_1 is the slope and ϵ_i is the error term (the residual). The optimal values of β_0 and β_1 can be estimated by Ordinary Mean Squares (OMS) (Lantz, 2013). The errors are the vertical distances between the predicted value of y and the actual y values, as shown in Figure 2.3.

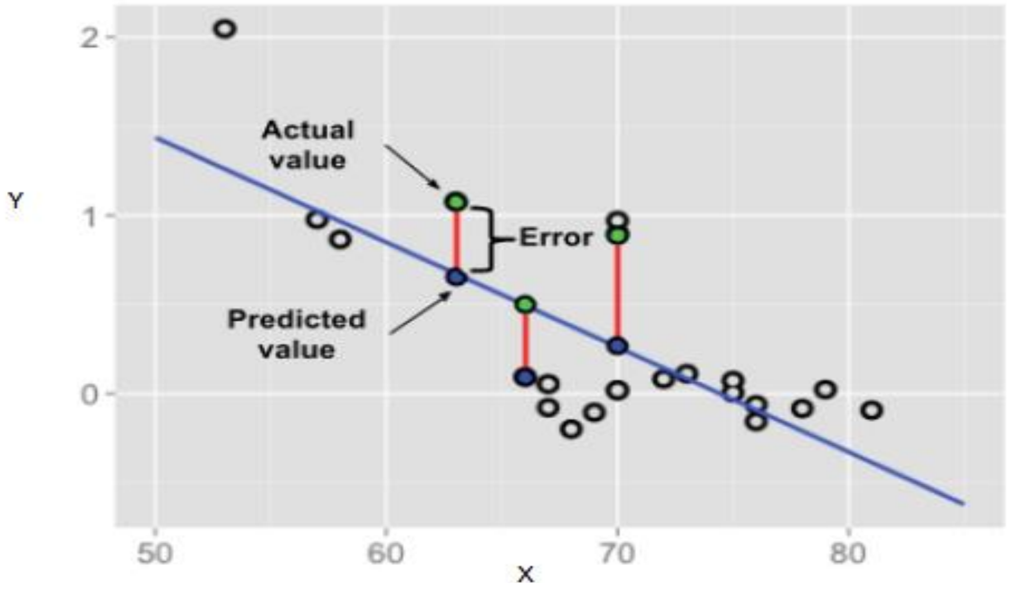


Figure 2.4 Linear Regression Model (Source Lantz, 2013)

Mathematically, the value of β_0 and β_1 that minimises the squared error are given by Equations (2.13) and (2.14) respectively:

$$\beta_0 = \bar{y} - b\bar{x} \tag{2.13}$$

$$\beta_1 = \frac{\sum(x_{i-}\bar{x})(y_i - \bar{y})}{\sum(x_i - \overline{x)}^2} \tag{2.14}$$

$\bar{x}$ and $\bar{y}$ are, respectively, the mean values of X and Y.

When Y depends on more than one value of X such as X1, X2,. . .,X3 , we have what is commonly termed as multivariate linear regression. The general form of multivariate linear regression is shown in Equation (2.15):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \epsilon_i \tag{2.15}$$

β_0 is the intercept; this is the value of y when the independent variables are zero. β_i is the change in y for a corresponding unit change in each x_i. The validity of inferences of a linear regression model depends on certain very important assumptions; some of which are stated as follows:

The regression function must be linear as far as possible to ensure the validity of results;

The error must normally be, independently and identically distributed with a mean of zero and a constant variance.

Linear regression is one of the frequently used methods in related works such as Kimanga *et al.*, (2014), Gabrysch *et al.*, (2008) and Kalichman *et al.*, (2006) and Tladi, (2006).

One adverse effect that may occur in linear regression analysis is a confounding effect; caused by confounding variable(S). A confounding variable is one which is independently associated with both the independent and dependent variables but not accounted for (Jager *et al.*, 2008; Braga *et al.*, 2012). The effect of such variable(s) is the distortion of association between the independent and dependent variable. Confounding has the adverse effects of increasing various and introducing bias. It could be minimised through random sampling and introducing control variables to control for the confounding variable(s).

Chi-Square Analysis

Chi-square (x^2) is a test of significance used mainly to test the goodness of fit, independence and homogeneity (Onchiri, 2013; Franke and Christie, 2012). In each of the cases, there must be a null and alternative hypothesis, one of which must be upheld and the other discarded at the end. The null hypothesis, H_0 always assumes that the observed is consistent with the expected; while the alternative hypotheses (H_1) proposes the opposite. Chi-square test for the goodness of fit is when it (chi-square) is used to decide whether a given observation is consistent with theoretically expected outcomes. In this case, chi-square is computed as follows, as shown in Equation (2.16).

$$x^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (2.16)$$

where O_i is the observed frequency for category i and E_i is the expected frequency for category i.

Test of independence is the next common use of chi-square analysis. This tests a possible relationship between two categorical observations. To proceed, a two-way contingency is constructed between the two observations and the expected value for each row-column pair is computed according to Equation (2.17).

$$E_{ij} = \frac{R_i C_j}{N} \qquad (2.17)$$

Where *R=Row*, *C=Column* and *N=Total* for i^{th} row and j^{th} column. Subsequently, x^2 is computed according to Equation (2.16).

Closely related to the chi-square test for independence is the test for homogeneity; where samples of two or more variables are tested for possible similarity. The contingency table approach is applied in such cases as discussed.

For the final determination of the test outcome, the degree of freedom is computed. For the case of the test for goodness of fit, the degree of freedom is computed, as shown in equation 18.

$$df = k - 1 \qquad (2.18)$$

Where k is the number of categories.

For the tests for independence and homogeneity, the degree of freedom is computed, as shown in equation 19.

$$df = (m - 1)(n - 1) \qquad (2.19)$$

Where m and n are respectively number of rows and number of columns of the contingency Table. Chi-square test is summarised as follows:

Determination of the *critical chi-Square value* by looking up the critical chi-square statistic value at 95% confidence level with *df* from a *table of critical Chi-Square values*.

Determination of significance. If x^2 is less than the critical value, then $H_0$is upheld otherwise H_1 is upheld.

Essentially, chi-Square is useful for testing the relationship between pairs of variables in data. Chi-square is observed in Section 2.4 as one of the most commonly used methods in existing HIV researches as seen in (Yahya-Malima *et al.*, 2006; Opio *et al.*, 2013; Temam and Ali, 2012)

Clustering

Clustering Algorithms

Clustering is an important data mining technique for the discovery of patterns in large datasets. We define clustering as a process of dividing a dataset into groups of similar objects (Berkhin, 2006; Sehgal and Garg, 2014). Objects in each cluster are similar among themselves and dissimilar to objects of other clusters. Clustering is important for data exploration, information retrieval and text mining, spatial database applications, Web analysis amongst others (Berkhin, 2006). Depending on how clusters are formed, clustering algorithms could be classified into the partition, hierarchical or density-based (Saraswathi and Sheela, 2014; Sisodia *et al.*, 2012).

Partition based clustering techniques

Partition-based clustering algorithms generally split a dataset of size n into *k* subgroups called clusters; ensuring that the data elements in a given cluster are as similar to each other as possible and are as different as possible to data elements in other clusters with respect to a given criterion (Saraswathi and Sheela, 2014; Sisodia *et al.*, 2012). Commonly, the criterion of similarity between elements in a given cluster is the centroid of that cluster; implying that elements within a given cluster have a common centroid (which is referred to as the centre of gravity of the cluster). With this technique, the value of *k* must usually be set in advance. In the initial step, this class of clustering algorithms split the data into the preset k clusters. It then proceeds recursively to check and reassign elements to more appropriate clusters in order to optimise the criterion function. Common examples of partition-based clustering algorithms are k-means and k-medoids.

Even though this class of clustering algorithms are scalable and simple, it has several disadvantages stated as follows (Sisodia *et al.*, 2012):

Suitable for datasets with compact spherical clusters that are well-separated;
Relies on the user to specify the number of clusters in advance;
Highly sensitive to noise and outliers;
Easily entrapped into local optima;
Unable to deal with non-convex clusters of varying size and density.

Hierarchical based clustering techniques

Hierarchical algorithms split a large dataset of size *n* recursively to form a nested sequence (or hierarchy) of clusters (Saraswathi and Sheela, 2014; Sisodia *et al.*, 2012). The hierarchy of clusters forms a tree-structure commonly termed as *dendrogram* with its root node representing the whole dataset and each leaf node being a single data of the dataset (Saraswathi and Sheela, 2014). The two types of hierarchical clustering approaches are agglomerative and divisive (Narmadha *et al.*, 2016). Agglomerative hierarchical clustering, also known as bottom-up starts with each element in the dataset as individual clusters and, at each step, merge the closest pair of data to form a single cluster. A visual illustration of hierarchical clustering is shown in Figures 2.5 and 2.6, respectively. In Figure 2.5, for instance, five clusters are initially formed; each cluster consisting of one of the elements each in the dataset (1, 2, 3, 4 and 5). In the next step, clusters closest to each other are merged to form one cluster. For instance, clusters 1 and 2 merge into one, while 3 and 4 also merge. This continues iteratively until one large cluster consisting of the smaller hierarchically formed cluster is obtained. On the contrary, divisive hierarchical clustering (**top-down** approach) starts with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual data remain.

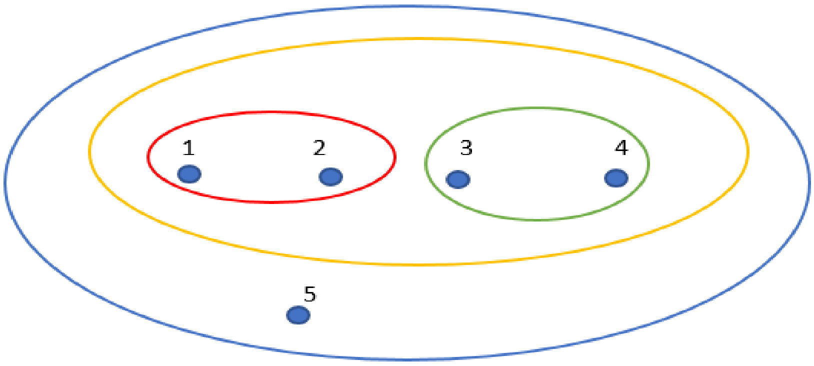


Figure 2.5 Visual Illustration of Hierarchical Agglomerative clustering

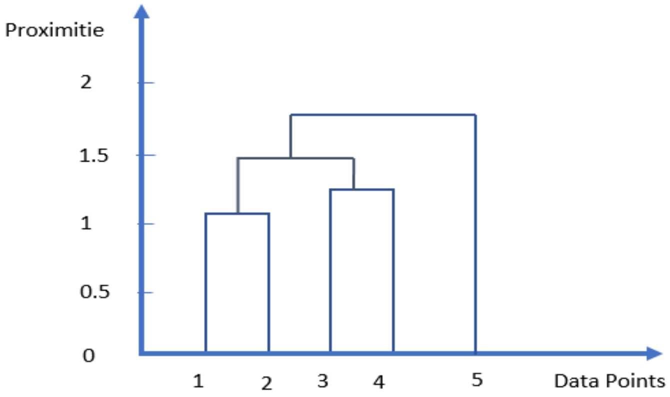


Figure 2.6 Dendogram Corresponding to Figure 2.5

The main advantage of hierarchical clustering is its suitability for problems involving point linkages such as taxonomy trees (Sisodia *et al.*, 2012).

It, however, has the following known disadvantages (Sisodia *et al.*, 2012):

It is unable to make corrections once the splitting/merging decision is made;
Its termination criterion is not clear;

Very expensive when applied to high dimensional and massive datasets;

Severe effectiveness degradation in high dimensional spaces due to a phenomenon called cursed of dimensionality.

Examples of hierarchical clustering algorithms include Clustering Using Representatives (CURE), Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) and CHAMELEON.

Density-based clustering techniques

This class of algorithms identify clusters as dense regions within data; usually separated by noise (Saraswathi and Sheela, 2014; Sisodia *et al.*, 2012). The noise represents regions of low data density. They are scalable and effectively handle outliers. The main disadvantages, however, include the following:

High sensitivity to the set of input parameters (for ex. density threshold);

Poor cluster descriptors;

Unsuitable for high-dimensional datasets because of the curse of dimensionality phenomenon.

Examples of algorithms in this class include DBSCAN and OPTICS.

Neural network based / Topology preserving clustering techniques

This is a class of clustering techniques which can be best categorised as unsupervised competitive and winner-take-most learning neural networks.

The key feature of techniques in this class is their ability to preserve data topology (Pena *et al.*, 2008; Ngo *et al.*, 2014). These techniques are, therefore, commonly referred to as Topology Preserving Approach (TPA). They are therefore capable of handling multidimensional data, effective identification of outliers and are less sensitive to initial conditions than usual partition-based methods. Self-Organising Maps (SOM) is the most known approach. It uses explicit topology, usually in the form of 2-dimensional maps, which permits to combine clustering and visualisation in the same process. Its related techniques such as Neural Gas (NG) and Growing Neural Gas (GnG) belong to the same class of clustering algorithms but conversely to SOM these two latter methods identify the implicit data topology during the clustering process. The next sub-section discusses GnG.

Growing Neural Gas (GnG)

Starting with just two nodes in n-dimensional input space R^n, GnG incrementally creates a network of nodes, where each node in the network has a position in R^n (Holmström, 2002). Using Competitive Hebbian Learning, GnG keeps topological relations between neighbouring nodes in the network (Holmström, 2002). Because of its incremental behaviour, GnG does not require a cluster size to be necessarily preset (Fritzke, 1995). The network is continually built until its maximum size is met, or an error stopping criterion is met (Fritzke, 1995). It is commonly used for clustering and vector quantization. GnG (Holmström, 2002), is summarised in pseudo-code as shown in Algorithm 2.1.

Algorithm 2.1

```
Create two randomly positioned nodes p and q;
Set Age of an edge between p and q to 0;
Set error of p to 0;
Set error of q to 0;
While there are still unread nodes do
        Generate random input vector $\bar{x}$
        Locate nodes s and t such that $\|\bar{w}_s - \bar{x}\|^2$ is smallest and $\|\bar{w}_t - \bar{x}\|^2$ second
        smallest
        Update local error of s as follows: $error_s = error_s + \|\bar{w}_s - \bar{x}\|^2$
        Move s, and it is topological neighbours towards x such that:
        $\bar{w}_s \leftarrow \bar{w}_s + e_w(\bar{x} - \bar{w}_s)$ and $\bar{w}_n \leftarrow \bar{w}_n + e_n(\bar{x} - \bar{w}_n)\ \forall n \in Neighbour(s)$;
 $e_w, e_w \in [0; 1]$
Update age of edges between s and its topological $Neighbour(s)$
If s and t are connected by an edge:
    Update age of edge to 0.
Else create an edge between them.
Remove old edges
Remove dead nodes
If the current iteration is an integer multiple of $\beta$
     Find the node u with largest error.
     Find node v with the largest error; $v \in Neighbour(u)$
     Insert the new node r such that $\bar{w}_r \leftarrow \frac{\bar{w}_u + \bar{w}_v}{2}$
```

```
        Create edges between $u$ and $r$, and $v$ and $r$
        Remove edge between $u$ and $v$.
        Decrease the error-variables of $u$ and $v$ such that
        $error_u \leftarrow error_u$ and $error_v \leftarrow error_v$
        set the error of node $r$ such as $error_r \leftarrow error_u$
      Decrease all error-variables of all nodes k by a factor $\delta$such that
          $error_k \leftarrow error_k - \delta * error_k \leftarrow error_k$
End while
```

Cluster Quality Evaluation

The need for evaluation of cluster quality is evident in very large datasets since there is usually no ground truth to facilitate work on such datasets in a supervised manner. In such cases, clusters must, therefore, be evaluated using the quality index. A quality index is a criterion which makes can acceptably be used in deciding:

(i) the clustering method to use;

(ii) the optimal number of clusters which can be generated from a given dataset.

To ensure efficient clustering results, several cluster quality index evaluation methods such as Dunn index (Dunn, 1974), the Davies-Bouldin index (Davies and Bouldin, 1979), the Silhouette index (Rousseeuw, 1987) amongst others exist. Such methods, however, are less suitable for high dimensional datasets as most of them are parametric, sensitive to noise, rely on Euclidean distance; thus, giving the same importance to each vector dimension, best suited for detecting Gaussian clusters and not independent of the clustering method used (Lamirel and Al Shehabi, 2015).

Even though methods proposed in (Bock, 1996), (Gordon, 1998) and (Halkidi, 2001) do not have these same stated drawbacks, they still do not work well with real-world data (Lamirel and Al Shehabi, 2015). Feature Maximization (Lamirel *et al.*, 2014) is a better method for addressing the above-stated problems while efficiently evaluating cluster quality. It is based on feature selection approach and is underlined by the fact that the more salient features a cluster contain, the better the quality of the cluster. Feature maximization is non-parametric, independent of the method, not sensitive to the noise issue; deals well with high-dimensional data and well detects degenerated clustering results.

Knowledge Representation Techniques

Knowledge Representation techniques are formal languages capable of structuring and reasoning about knowledge in a particular domain. A good Knowledge Representation technique must be expressive; having clearly defined syntax and semantics to allow knowledge in a given domain to be unambiguously represented and allow inference of new knowledge from existing knowledge (Tanwar et *al.*, 2010; Clark, 1996). Broadly, the various Knowledge Representation techniques are classified as logical, procedural, network-based and structured. With logical representation techniques (Tanwar *et al*., 2010), knowledge is represented as declarative statements and inference rules and proof procedures are used to reason on them. Example of this class of languages is predicate logic. Procedural representation techniques, on the other hand, allow knowledge to be represented as a list of instructions such as production rules (also known as if-then rules) (Tanwar *et al*.,2010). The third category of knowledge representation techniques is based on graphs (Rashid, 2015). Commonly known as network-based representation techniques, these represent objects or concepts as nodes and relationships between them the objects or concepts are edges. Typical examples include semantic networks. Another knowledge representation technique which has found much use is the structured technique (Rashid, 2015; Tanwar *et al.*). Such a technique commonly store knowledge in the attribute-value format. Examples include scripts and frames.

In this section, three major knowledge representations are discussed. These are:

First Order Logic

Semantic Networks

Frames

KNOWLEDGE, TRUTH AND EXCELLENCE

First Order Logic

First Order Logic (FOL) is otherwise known as Predicate Logic. The term “predicate” refers to a group of words which either binds a *term* to its attribute or denotes the relationship of such terms in a sentence. A term can be either a constant of a variable. Consider the following two sentences:

Socrates is a Man;

x is an even number

In these two sentences, Socrates is a constant term while x is a variable term. “is a man” and “is an even number” are respectively predicates.

A constant refers to a concrete object in a given set while a variable is a name that can denote any element in a given defined set.

The arity of a predicate

The number of arguments or individual terms a predicate combines is the *arity* of that predicate. The following common types of predicates exist:

Unary predicate – A predicate with one argument;

Binary predicate – A predicate with two arguments;

Ternary predicate – A predicate with three arguments.

In principle, however, there can be up to *n-ary* predicate (where n is a positive integer denoting the number of arguments the predicate takes). This is summarised in Table Table 0.21 with an example.

Table 2.16 Types of Predicates with Examples

Sentence	**Sentence in Predicate Logic Notation**	**Type of Predicate**
Socrates is a Man	Man (Socrates)	Unary
Mammals Drink Milk	Drink (Mammals, Milk)	Binary
Socrates feed goats with grass	Feeds (Socrates, goat, grass)	Ternary

Syntax and semantics of Predicate Logic

The syntax of FOL consist of the following symbols:

A set $\rho = (P, Q, R, ...)$ of predicate symbols;

An infinite set VAR = {x, y, z, . . .} of variables;

A set of constants CONS = {a, b, c, . . .} to represent concrete individuals such as John, Socrates, *etc* ;

Connectives – Conjunction (∧), Disjunction (∨), Implication (→), Equivalence (↔) and Negation (¬);

Brackets and parentheses - [,], (and);

Comma;

Quantifiers – Universal (∀) and Existential (∃).

A sequence of symbols of Predicate Logic which is grammatically correct is termed well-formed formula (commonly abbreviated as *wff*) or just formula. A set of *wffs* are inductively defined as:

If p ∈ ρ is an n-ary predicate symbol, and $x_1, x_2, . . . , x_n$ ∈ VAR are individual variables, then $p(x_1, x_2, . . . , _{xn})$ is a wff;

If A is a wff, the $\neg A$ is a wff;

If A and B are wffs, then $[A \wedge B]$, $[A \vee B]$, $[A \rightarrow B]$, and $[A \leftrightarrow B]$ are wffs;

If A is a wff and $x \in VAR$ is an individual variable, then the formulas $\forall xA$ and $\exists xA$ are wffs.

Knowledge Representation with Predicate Logic

Predicate Logic, as a language is useful for knowledge representation. Consider the following scenario expressed in the natural language:

Smarty is a parrot;

Parrot is a bird

Kofi is a person

Kofi owns Smarty

A bird is an animal

A person is a mammal

A mammal is an animal

Animals feed on plants

Animals have sound

This can be represented as a knowledge base in predicated logic as follows:

Parrot(Smarty)

$\forall x Parrot(x) \rightarrow Bird(x)$

Person(Kofi)

Owns(Kofi, Smarty)

$\forall x Bird(x) \rightarrow Animal(x)$

$\forall x Person(x) \rightarrow Mammal(x)$

$\forall x Mammal(x) \rightarrow Animal(x)$

$\forall x$ Animal (x) $\rightarrow$ feeds_on(x,Plants)

$\forall x$ Animal (x) $\rightarrow$ have(x,Sound)

Inference rules

Inference rules are principles which can be applied to reason on a Knowledge Base to derive new knowledge (commonly used for decision making). The commonly used inference rules are given in Table 2.18. For instance, in the knowledge base given, there is no explicit knowledge that, “Smarty is an animal” or “all parrots are animals”. These can however be derived from the Knowledge Base (KB). Using just the following three lines from the knowledge base and the inference rules, the knowledge that “all parrots are

KNOWLEDGE, TRUTH AND EXCELLENCE

animals” is derived as follows:

$\forall x Bird(x) \rightarrow Animal(x)$	Given in KB
$\forall x Parrot(x) \rightarrow Bird(x)$	Given in KB
$Parrot(Smarty)$	Given in KB

$Parrot(a) \rightarrow Bird(a)$	UI from line 2
$Bird(a) \rightarrow Animal(a)$	UI from line 1
$Parrot(a) \rightarrow Animal(a)$	HS from lines 4 and 5
$\forall x\ Parrot(x) \rightarrow Animal(x)$	UG from line 6

Table 2.17 Inference Rules

Rule	**Name of Rule**	**Rule**	**Name of Rule**
$p\rightarrow q$ p ——— $\therefore q$	Modus Ponens (MP)	p q ——— $\therefore p \wedge r$	Conjunction
$p\rightarrow q$ $\neg q$ ——— $\therefore \neg p$	Modus Tollens (MT)	$p \vee q$ $\neg p \vee r$ ——— $\therefore q \vee r$	Resolution
$p\rightarrow q$ $q\rightarrow r$ ——— $\therefore p\rightarrow r$	Hypothetical Syllogism (HS)	$\forall x P(x)$ ——— $\therefore P(c)$	Universal Instantiation (UI)
$p \vee q$ $\neg q$ ——— $\therefore p$	Disjunctive Syllogism (DS)	$P(c)$ ——— $\therefore \forall x P(x)$	Universal Generalisation (UG)
p ——— $\therefore p \vee q$	Addition	$\exists x Px$ ——— $\therefore P(c)$	Existential Instantiation (EI)
$p \wedge q$ ——— $\therefore p$	Simplification	$P(c)$ ——— $\therefore \exists x P(x)$	Existential Generalisation (EG)

Semantic Networks

The concept of Semantic Networks allows knowledge representation in the form of directed graphs, in which nodes represent objects in the domain being modelled, and arcs (edges) show relationships between the nodes (objects) (Lehmann, 1992; Huntback, 1996). The original concept was proposed by Quillian Ross in 1967 (Quillian, 1967) and

KNOWLEDGE, TRUTH AND EXCELLENCE

has since found extensive application in the field of Artificial Intelligence. The following are the main types of relationships used in Semantic Networks:

A kind of (ako): A relationship between a subclass and a superclass from which the subclass is derived;

Is-a: A relationship between an instance of a specific object to the class to which the object belongs;

Part of: A relationship between an object to a part of that object;

Has-a: A relationship between an object and property or attribute which that object has.

Apart from these defined terms for showing relationships, it is common to find other meaningful words being used to show the relationship. One strong capability of semantic networks is inheritance; the ability of a derived class to possess features of a superclass. Inheritance is denoted by the ako and is-a relationships. A semantic network example is shown in Figure 2.7.

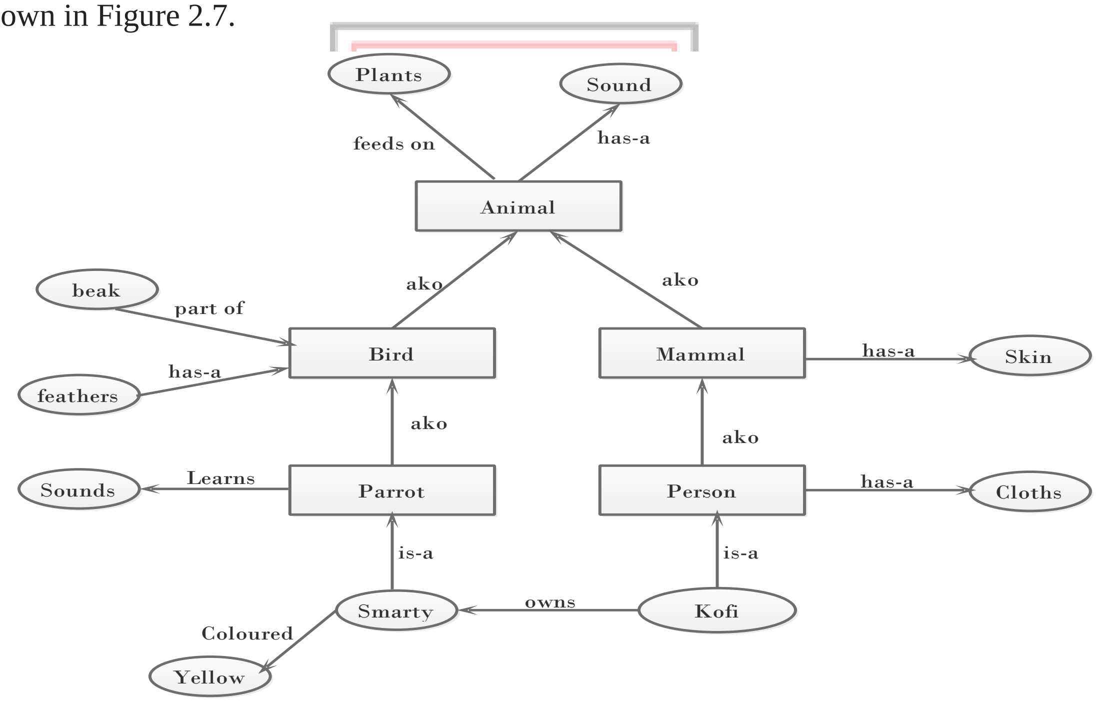


Figure 2.7 Semantic Network Example

In this example, two kinds of nodes are used: The rectangles represent concepts (classes) whereas the ovals represent attributes and specific (concrete) instances of classes. For example, Smarty is a concrete instance of a parrot, which is a class of birds with several species. By inheritance, it possesses all the attributes of all the superclasses from which it inherits in addition to its attribute; “yellow colour”. By this, it can be inferred that Smarty can learn sounds, has feathers, has a beak, has sound and feeds on plants.

Frames

The concept of Frames as a knowledge representation technique was proposed by Minsky in 1974 (Minsky, 1974). A frame is composed of a name, a set of slots and fillers or facets. The name identifies the frame (analogical to a class name in Object Oriented Programming). The slots are identifiers of attributes or properties of the object being represented. Each slot has a corresponding filler or value. For example, a slot “Coloured” can have the value “Yellow”. Table 2.18 is an example of a frame named “Animal” with two slots. Filler to a given slot can be a value or a pointer to other objects. In this example, the first slot is filled with a string while the second has binary filler; whose possible values are either True or False. Frames also allow inheritance; permitting one frame to inherit features from another; thereby forming a network of frames. Table 2.19 illustrates a frame “Bird” inheriting the attributes “Feeds on” and “Moves” from the frame animal:

Table 2.18 Example of a Frame

Animal	**slot**	**Fillers**
	Feeds on	*Plants*
	Moves	*T*

Table 2.19 Illustration of Inheritance in Frames

Animal	**slot**	**Fillers**
	Feeds on	*Plants*
	Moves	*T*

↑

Bird	**slot**	**Fillers**
	Feathered	*T*
	Wings	*2*

The illustration in Table 2.19 shows that a bird is an instance or a subclass of the concept “Animal”. Depending on the situation, slots may be filled at the class level, or instance level. Attributes which are common to all members of a class are filled at the class level. Table 2.19, for example, illustrates that all instances of animals feed on plants and move and this is inherited by the “Bird” frame. Where it is filled at the instance level, it indicates that the value of that attribute varies among members of that class.

KNOWLEDGE, TRUTH AND EXCELLENCE

SYSTEM DESIGN

Preamble

The objectives of this research are achieved in three broad, interconnected steps. The first step involves the identification of HIV drivers in Sub-Saharan Africa (SSA). Frame-based knowledge representation technique is used to represent knowledge obtained from the literature (Section 2.6) regarding the influence of such factors on HIV in SSA. This step leads to the identification and classification of HIV driving factors and guides the organisation of a dataset for the second step of the research. The second step involves the identification of districts from where the dataset would be collected and put in the right format for the modelling process. The last step applies clustering and Feature Maximization for the modelling and analysing of the collected data. This last step leads to the establishment of the degrees of impact and interactivity of HIV driven factors.

Identification of HIV Drivers in Sub Saharan Africa

The purpose of this subsection is to identify socio-economic and socio-demographic factors of HIV for this research. Firstly, the theory of social production of disease is discussed in order to introduce the role of societal factors in disease causation in the broader view. Secondly, knowledge obtained from the systematic literature review in Tables 2.5 to 2.14 is represented using frames in order to show a clear relationship between such driving factors and HIV. Finally, the HIV drivers for this research are extracted from the frame model and classified.

Theory of Social Production of Diseases

Social Production of Disease is a theory of Social Epidemiology concerned with explaining the economic and political determinants of Health. The Diderichsen model (Diderichsen et *al.*, 2001), which strongly supports and explains this theory is discussed here. The Diderichsen model provides four mutually exclusive social components (three are discussed here) necessary for understanding and explaining the link between one's social position and equity of health. These components are (i) Factors affecting Social Stratification (ii) Differential Exposure to health-damaging Factors (iii) Differential vulnerabilities leading to unequal Health Outcomes (iv) Differential Social Consequences of illness.

Factors affecting Social Stratification

This, amongst other things, defines one's position in society (social position or class). Individual's social positions are defined by factors such as gender, education, occupation, and so on. Individuals with more privileged social positions (such as male gender, high education, well-paying job, power, and so on) benefit more from valued social resources than the less privileged (such as female gender, people with low education, low paying jobs, no power). Hence, the less privileged may have little opportunity to education and employment than their more privileged counterparts. The way individuals are sorted in any society according to the social position is at the centre of the social differential of health.

Differential Exposure to health-damaging Factors

Each social group suffer from a specific pattern of health risk; with individuals of lower social positions having excess risk associated with ill health regarding a wide range of diseases. Unskilled people, for example, earn a low income, have little opportunity of choosing healthy lifestyles and may live in places which can easily expose them to various health hazards.

Differential Vulnerability

An equal distribution of a given risk factor across social groups can have different health impacts in the different social groups due to underlying differences with respect to the risk factor(s) in terms of vulnerability. For instance, vulnerability to damaging exposures is commonly observed in different amounts according to different age groups and gender**.** An important study which supports the idea of differential vulnerability observes that African women are more vulnerable to ill health due to lack of education, greater work burden and minimal income generation possibilities which drives some of them into commercial sex work (Kalipeni, 2000). In general, educated individuals are more likely to gain good paying and less hazardous jobs and are also more likely to avoid health risks and practice preventive behaviours than those with less privileged social positions.

Frame-Based Representation of HIV Drivers

The knowledge represented with frames in this section is extracted from the "outcome" column of Tables 2.5 to 2.19 in section 2.6. The "outcome" column represents knowledge extracted from the findings of each reviewed article in the Systematic Literature review process. The frame models are shown in Figure 3.1 to 3.4.

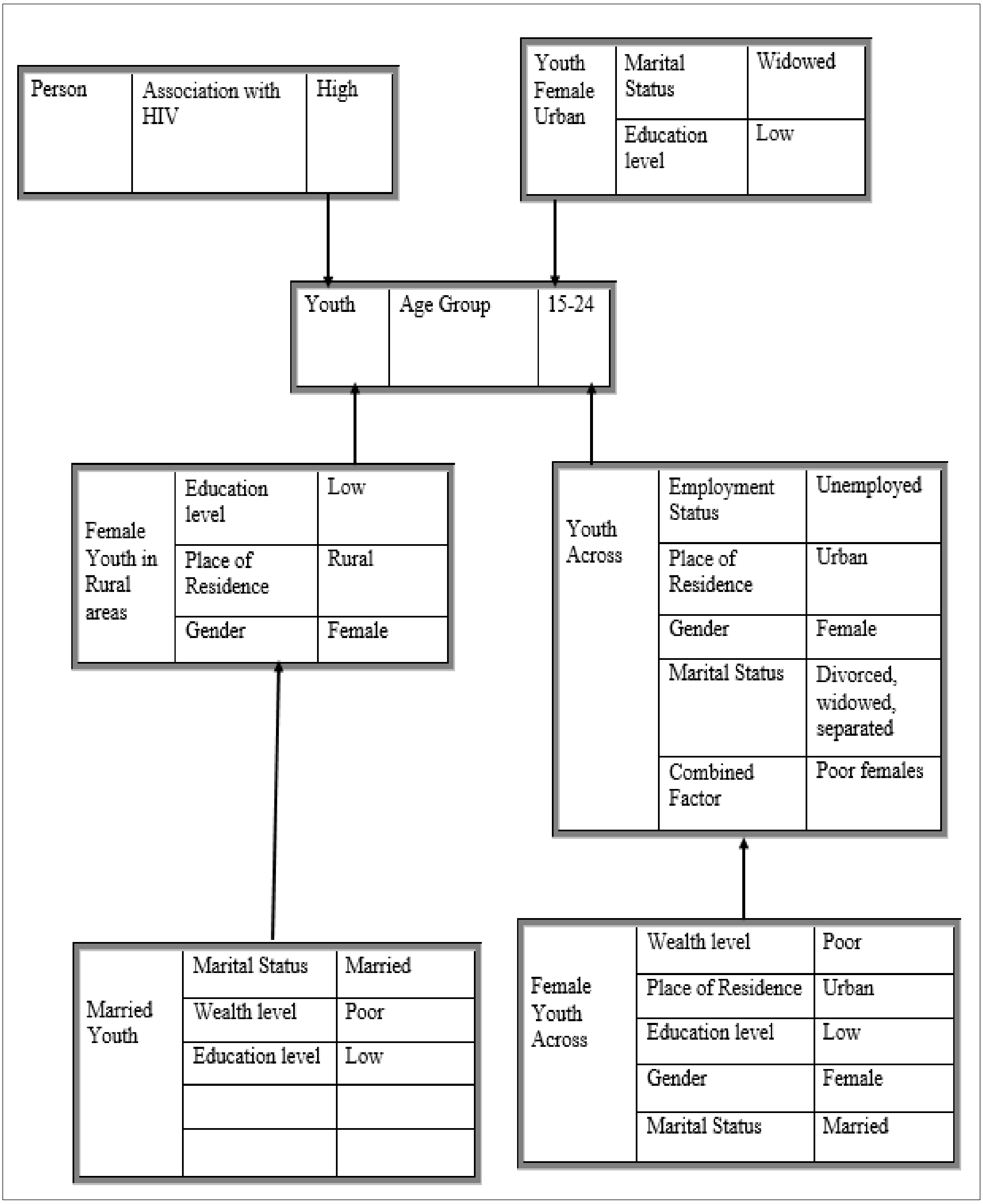
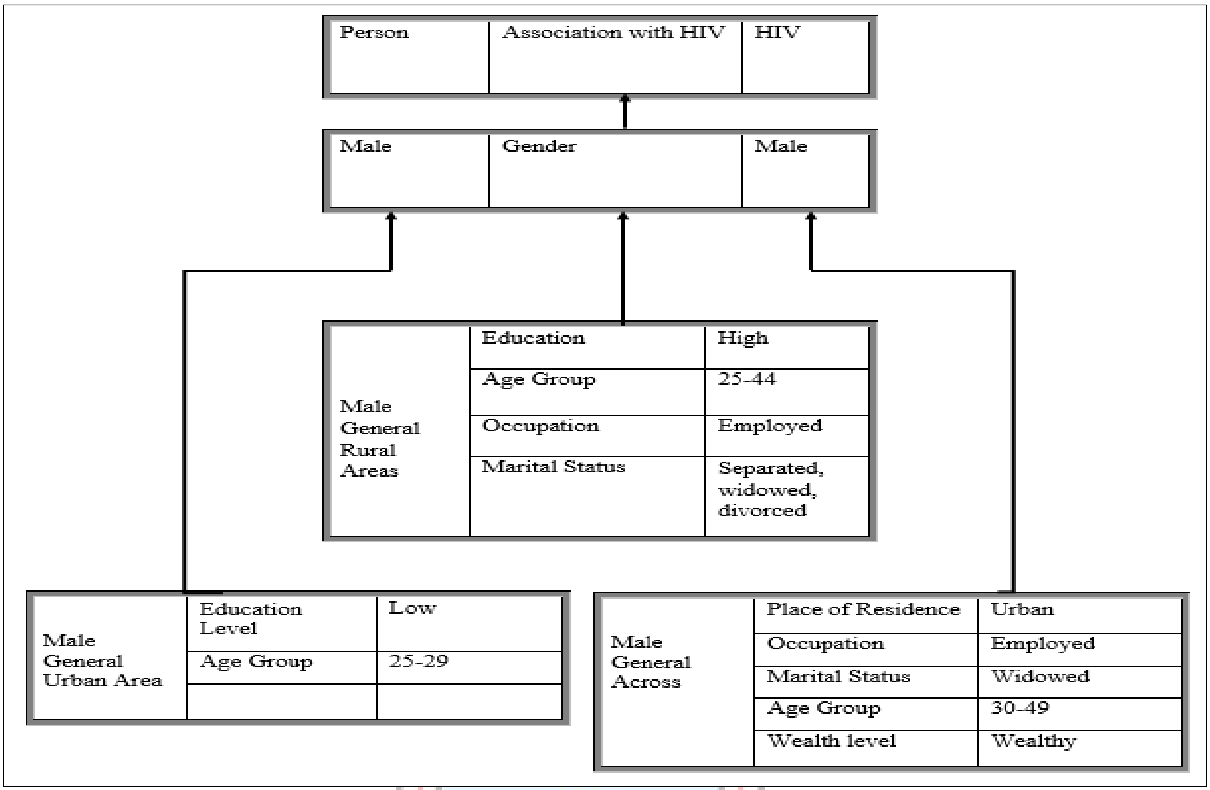


Figure 3.1 Frame Model identifying HIV Drivers Status of the Female Youth.



Person	Association with HIV	HIV

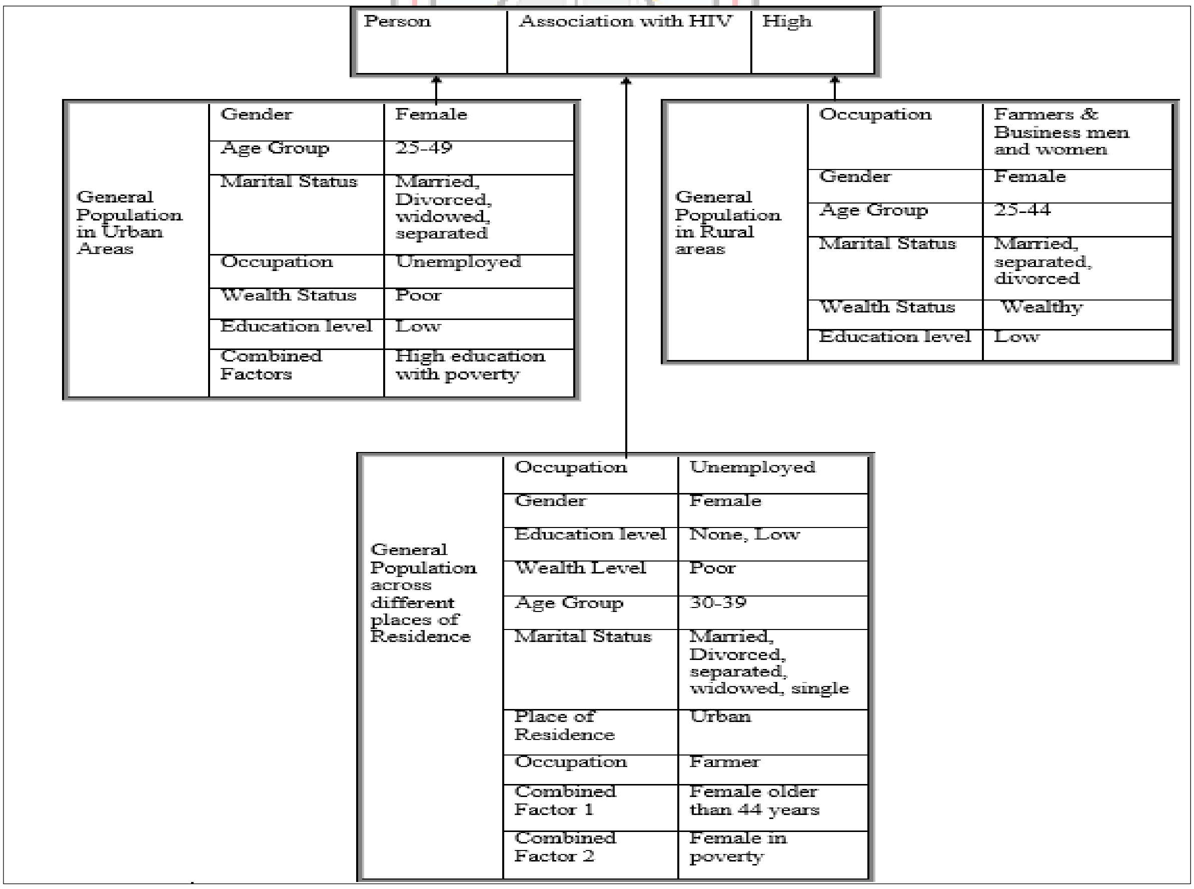
Male	Gender	Male

Male General Rural Areas	Education	High
	Age Group	25-44
	Occupation	Employed
	Marital Status	Separated, widowed, divorced

Male General Urban Area	Education Level	Low
	Age Group	25-29

Male General Across	Place of Residence	Urban
	Occupation	Employed
	Marital Status	Widowed
	Age Group	30-49
	Wealth level	Wealthy

Figure 3.2 Frame Model identifying HIV Drivers Status of Males.



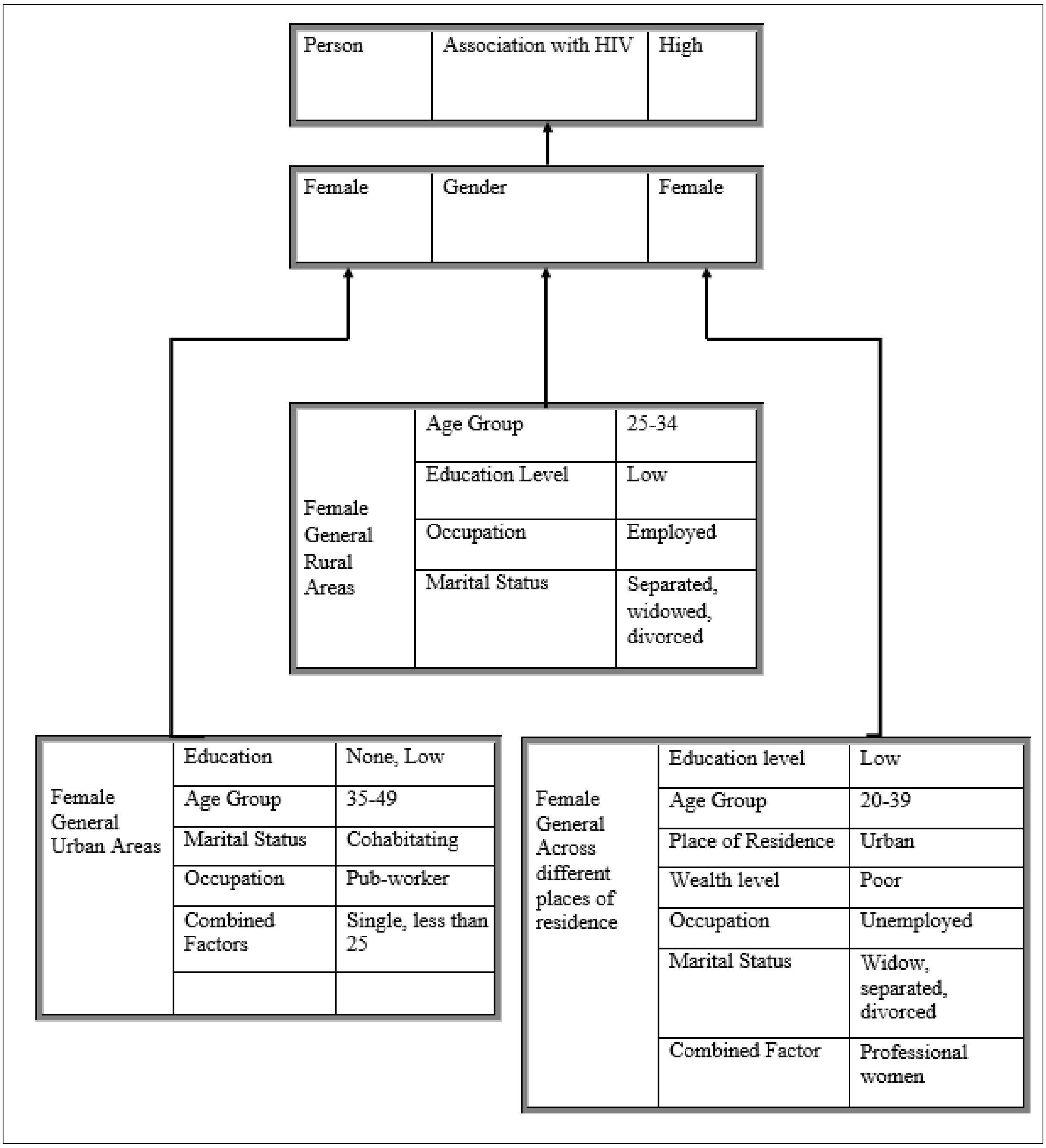
Person	Association with HIV	High

General Population in Urban Areas	Gender	Female
	Age Group	25-49
	Marital Status	Married, Divorced, widowed, separated
	Occupation	Unemployed
	Wealth Status	Poor
	Education level	Low
	Combined Factors	High education with poverty

General Population in Rural areas	Occupation	Farmers & Business men and women
	Gender	Female
	Age Group	25-44
	Marital Status	Married, separated, divorced
	Wealth Status	Wealthy
	Education level	Low

General Population across different places of Residence	Occupation	Unemployed
	Gender	Female
	Education level	None, Low
	Wealth Level	Poor
	Age Group	30-39
	Marital Status	Married, Divorced, separated, widowed, single
	Place of Residence	Urban
	Occupation	Farmer
	Combined Factor 1	Female older than 44 years
	Combined Factor 2	Female in poverty

Figure 3.3 Frame Model identifying HIV Drivers Status of the General Population.



Person	Association with HIV	High

Female	Gender	Female

Female General Rural Areas	Age Group	25-34
	Education Level	Low
	Occupation	Employed
	Marital Status	Separated, widowed, divorced

Female General Urban Areas	Education	None, Low
	Age Group	35-49
	Marital Status	Cohabitating
	Occupation	Pub-worker
	Combined Factors	Single, less than 25

Female General Across different places of residence	Education level	Low
	Age Group	20-39
	Place of Residence	Urban
	Wealth level	Poor
	Occupation	Unemployed
	Marital Status	Widow, separated, divorced
	Combined Factor	Professional women

Figure 3.4 Frame Model identifying HIV Status Drivers of Females.

Classification of drivers and their alignment to the theory

With insight from the frame models, the HIV drivers are identified and classified according to meaning into 7 categorical variables namely Education Level, Wealth Level, Age Group, Marital Status, Occupation, Gender and place of Residence. The details of this classification are in Table 3.1. Each of these variables can perfectly be located in the framework of the Diderichsen model (Diderichsen et *al.*, 2001); supporting the theory of Social Production of Disease / Political Economy of Health.

Table 3.1 Classification of HIV Drivers

Education Level	**Wealth Level**	**Age Group**
Lower Education	Poverty	Age Range 25-34
No Education	Wealth	Age Range 15-24
	Lower socio-economic Status (SES)	Age Range 30-39
Marital Status		Age Group 20-24
Widowhood		Age less than 18
Divorce	**Occupation**	Age Range 28-47
Separation	Unemployment	Age Range 25-49
Stable Marriage	Farmers	Age Range 35-49
Cohabitation	Employment	Age Group 15-19
Never Married	Businessmen	Age Range 25-44
	Working in a public House	
Gender	Professional Women	
Being a Female		
Place of Residence		
Urban Residence		

KNOWLEDGE, TRUTH AND EXCELLENCE

Data Collection and Organisation

Two different datasets were required for this study, namely the *HIV dataset* and a *reference* or *control dataset*. The HIV dataset contains the socio-economic and socio-demographic information of HIV patients (anonymous) of twenty-eight districts across Ghana. Guided by the fact that, *place of residence* is an important variable in this study, the districts were carefully selected to ensure a fair urban and rural representation. The Ghana National AIDS Control Programme collected it from 2008 to 2016. With reference to the classification scheme of districts in terms of percentage rural or percentage urban in the 2010 Population and Housing Census of Ghana, districts, which are more than 50% urban, were classified as *urban places of residence* and those 50% or less urban as *rural places of residence*. The data was then organised into tables with headings being the selected variables: *Age Group, Education Level, Marital Status and Occupation.* Details and categorisation of each variable are discussed next.

Age Groups

For this research, the ages of the HIV patients in the data were reorganised into the categories shown in Table 3.2.

Table 3.2 Age Groups categories

Category Label	**Age Group**
A0_14	Ages from 0 to 14
A15_19	Ages from 15 to 19
A20_24	Ages from 20 to 24
A25_29	Ages from 25 to 29
A30_34	Ages from 30 to 34
A35_39	Ages from 35 to 39
A40-44	Ages from 40 to 44
A45_49	Ages from 45 to 49
A50_54	Ages from 50 to 54
A55_59	Ages from 55 to 59
A60_64	Ages from 60 to 64
old_age	Ages greater than 64

KNOWLEDGE, TRUTH AND EXCELLENCE

Education level

This refers to the highest education level obtained by the patients at the time the data was collected. The various education levels identified in the datasets are:

Nil: This refers to those who have no formal education of any form.

Formal but low education levels: Primary, Middle School Leaving Certificate, Junior Secondary School (JSS) or Junior High School as their highest level of education.

Formal Secondary Education levels: Secondary School (Advanced Level or Ordinary Level), Senior Secondary School (SSS), Secondary Technical or Senior High School.

Post Secondary/Tertiary: Persons with a minimum of a certificate, Diploma or a Degree from a college, polytechnic or a university.

Marital Status

Six possible marital statuses were identified and labelled for this research, as shown in Table 3.3.

Table 3.3 Marital Statuses

Category Label	**Description**
Married	Refers to a person who is legally married to a wife or husband.
Single	Refers to a person who has never been married to a wife or husband.
Widow(er)	Widow refers to a woman whose husband has died. A widower is a man whose wife has died.
Divorced	Refers to a person who has divorced his or her spouse.
Cohabiting	Refers to a person living with a partner to whom he or she is not legally married.
Separated	Refers to a person who has separated from his or her spouse.

Occupations

The original data contained more than 200 different occupations. For this research work, the occupations were classified into 9 categories using the International Standard Classification of Occupations 2008 (ISCO-08). The categories of occupations and their respective descriptions are shown as follows:

Occ_1: Managers

Chief Executives, Senior Officials and Legislators

Administrative and Commercial Managers

Production and Specialized Services Managers

Hospitality, Retail and Other Services Managers

Occ_2: Professionals

Science and Engineering Professionals

Health Professionals

Teaching Professionals

Business and Administration Professionals

Information and Communications Technology Professionals

Legal, Social and Cultural Professionals

Occ_3: Technicians and Associate Professionals

Science and Engineering Associate Professionals

Health Associate Professionals

Business and Administration Associate Professionals

Legal, Social, Cultural and Related Associate Professionals

Information and Communications Technicians

Occ_4: Clerical Support Workers

General and Keyboard Clerks

Customer Services Clerks

Numerical and Material Recording Clerks

Other Clerical Support Workers

Occ_5: Services and Sales Workers

Personal Services Workers

Sales Workers

Personal Care Workers

Protective Services Workers

Occ_6: Skilled Agricultural, Forestry and Fishery Workers

Market-oriented Skilled Agricultural Workers

Market-oriented Skilled Forestry, Fishery and Hunting Workers

Subsistence Farmers, Fishers, Hunters and Gatherers

Occ_7: Craft and Related Trades Workers

Building and Related Trades Workers (excluding Electricians)

Metal, Machinery and Related Trades Workers

Handicraft and Printing Workers

Electrical and Electronic Trades Workers

Food Processing, Woodworking, Garment and Other Craft and Related Trades Workers

Occ_8: Plant and Machine Operators and Assemblers

Stationary Plant and Machine Operators Assemblers

Drivers and Mobile Plant Operators

Occ_9: Elementary Occupations

Cleaners and Helpers

Agricultural, Forestry and Fishery Labourers

Labourers in Mining, Construction, Manufacturing and Transport

Food Preparation Assistants

Street and Related Sales and Services Workers

Refuse and Other Elementary Workers

Galamsey Miners

Occ_0: Other Occupations

The Unemployed, Students, Children, Housewives, Refugees and Prisoners belong to this class. Occupations ***Occ_1*** through to ***Occ_9*** are classifications predefined by ***ISCO-08*** and were adopted as such for this work. ***ISCO-08***, however, does not make room for occupational statuses such as Unemployed, Students, Child, Housewife, Refugees and Prisoners. **Occ_0** was therefore carved to take care of those.

The reference datasets were obtained from the Ghana Statistical Service. The reference dataset was taken from precisely the same districts and similar sizes as the main dataset. Fig. 3.5 displays the map of Ghana; showing the various districts from which the data was collected.

Description of the Degree of Impact and Interactivity Model (DIIM)

This section presents the design of the model for computing and establishing degrees of impact and interactivities of socio-economic drivers of HIV. The model is based on the data-driven computing paradigm. As a data-driven system, the performance of the model depends absolutely on the available data. By design, it functions according to the centralised architecture. It is designed to be used by only specialised groups or institutions tasked to make informed policies and take initiatives to fight the HIV epidemic. Examples of such institutions include the Ghana AIDS Commission and the National AIDS/HIV Control Programme. The choice of centralised architecture is, therefore, necessary to ensure maximum privacy and security.

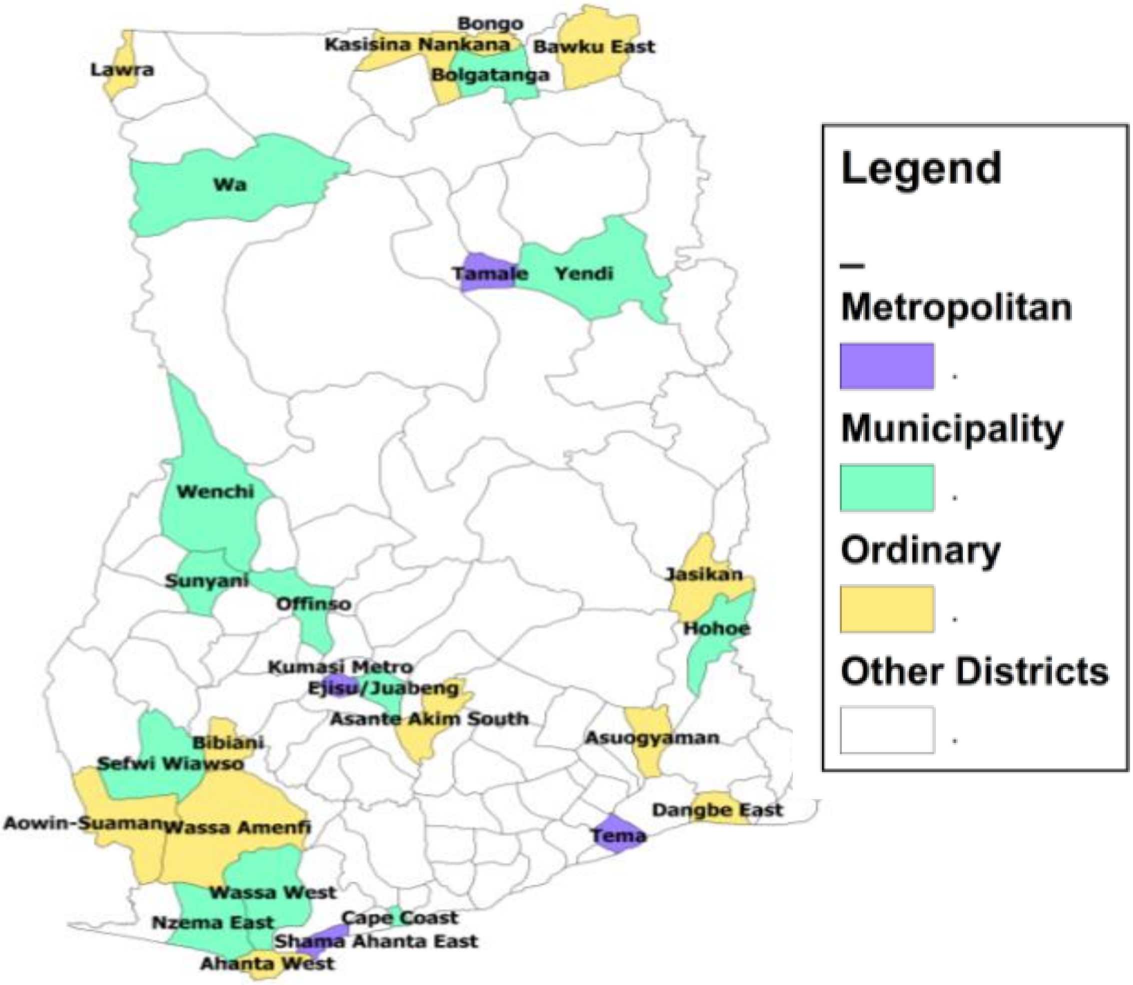


Figure 3.5 Map of Ghana showing Districts from which Data was collected

The main components of the system are the Data repository, the Pre-processing/Processing unit, the Centralised server, as well as the output and storage units. Summarised architecture is shown in Figure 3.6.

KNOWLEDGE, TRUTH AND EXCELLENCE

The data repository contains the data collected from the field in a binarized form. To use the system, the end user logs on to the centralised server and queries it. The centralised server then loads the appropriate data files from the data repository and gives the user the possibility to make further choices regarding preferred pre-processing and processing methods. If the requested dataset is found in the right format and the pre-processing/processing method is invoked successfully, then execution proceeds and the results are displayed to the user. Otherwise, the user receives an error message. Upon the user's choice, the results could also be saved in the storage unit for future reference. The summarised use case is shown in Figure.3.7.

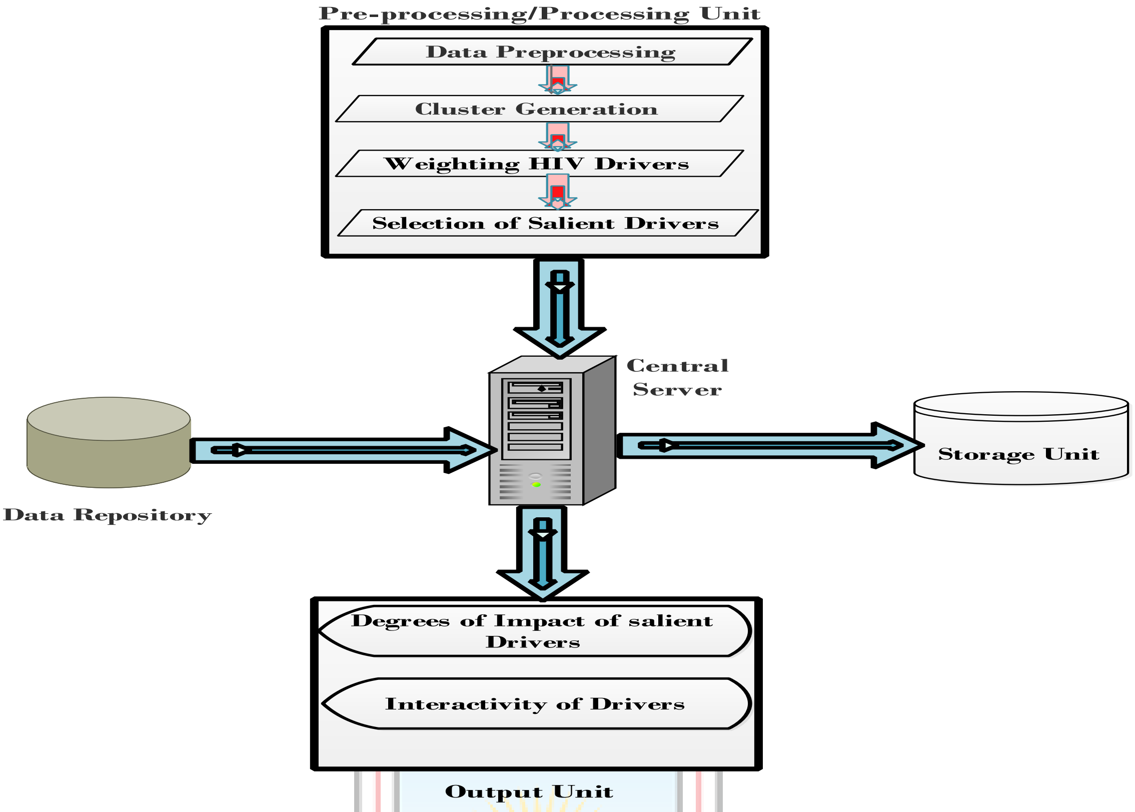


Figure 3.6 General System Architecture

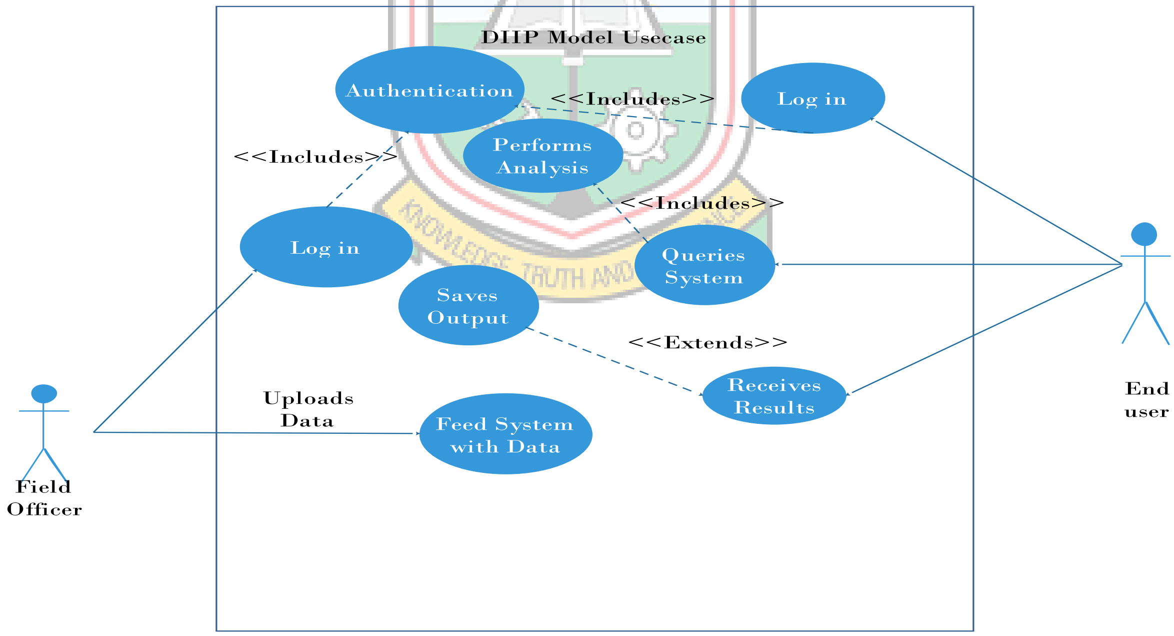


Figure 3.7 System Use Case

The system could be implemented either locally (offline) or on a network where some of the components could be placed at different locations. In the case of local implementation, the login and authentication (as shown in Figure 3.7) would provide security for the system and privacy of data. In the case of network implementation, however, security would be taken care of by placing a strong firewall between the two components concerned (for instance, between the server and the storage unit) in order to authenticate users.

Binarisation

The raw data collected from the field are nominal or categorical. For mathematical processing, the nominal is converted into numerical (binary) through a process termed binarisation. Given a dataset D of m-rows and n-columns, let each attribute in a column n_i take a finite number of possible nominal values. If the set of all distinct nominals of all attributes in D is $k = \{k_1, k_2, . . ., k_n\}$ then each row can be converted to a binary string of length n by the following scheme:

$$f(k_i) = \begin{cases} 1\,; if\ k_i\ is\ present \\ 0; \qquad otherwise \end{cases} \tag{3.1}$$

In the binary representation, each attribute's nominal value corresponds to a specific feature. This yields an m by k dimensional dataset. For the sake of clarity, consider a sample dataset with 2 attributes (MaritalStats and EduLevel) each having 3 nominal values (Table 3.4). The total number of distinct nominal values (modalities) in D is 6; that is k= {MaritalStats=m, MaritalStats=s, MaritalStats=w, EduLevel=JSS, EduLevel=T, EduLevel=P}.

KNOWLEDGE TRUTH AND EXCELLENCE

Table 3.4 Sample Dataset for Binarisation Example

Attribute **Row**	**MaritalStats**	**EduLevel**
1	m	JSS
2	m	JSS
3	s	T
4	m	JSS
5	w	P
6	w	P
7	m	T
8	w	JSS
9	m	JSS
10	s	P

By definition (Equation (3.1)), a binary string of length 6 is generated from each row, keeping the order in which, the nominal values occur (set k). In the first row, MaritalStats=m and EduLevel=JSS both appear, so they are replaced with 1 in k. The other nominal values do not appear in this row, so their places are taken by 0 in k. In effect, the following binary string is generated from the first row: 1 0 0 1 0 0. A full binarised form of Table 3.1 is shown in Table 3.5.

Table 3.5 Binarised Form of Sample Data.

Label	**MaritalStats =married**	**MaritalStats= single**	**MaritalStats =widow**	**EduLevel =JSS**	**EduLevel =Tertiary**	**EduLevel =Primary**
1	1	0	0	1	0	0
2	1	0	0	1	0	0
3	0	1	0	0	1	0
4	1	0	0	1	0	0
5	0	0	1	0	0	1
6	0	0	1	0	0	1
7	1	0	0	0	1	0
8	0	0	1	1	0	0
9	1	0	0	1	0	0
10	0	1	0	0	0	1

In the binary form, each row is a binary vector that represents a concatenation of binary strings. For each string, a ‘1’ corresponding to a given attribute nominal value is an indication of the presence of that nominal value in the string whiles a ‘0’ is an indication of its absence. For instance, the binary vector in row 1 is 100100. The only nominal values present are **‘MaritalStats=married’** and **‘EduLevel=JSS’**; the rest of the attributes nominal values are absent.

KNOWLEDGE, TRUTH AND EXCELLENCE

Cluster Generation

Cluster generation is a process to explore the data in order to establish the underlying population subgroups with a similar level of risk to HIV infection. Considering the multidimensional nature of the data used in this research and the limitations characterising hierarchical, partition and density-based clustering techniques (as discussed in Section 2.4.1), Growing Neural Gas (GnG) was used for the cluster generation. Relatively it is faster than Self Organising Maps (SOM) and has less chance to local minima compared to the traditional k-means. Considering each binary string as a node (point) in the real coordinate space of n-dimensions (R^n) each with a reference vector $\overline{w}_k$, cluster-generation process with GnG is presented under subsection 2.7.1.

The weighting of HIV Drivers

Let the clustering process be used to split the dataset resulting in a partition ***C*** with n disjoint (crisp) clusters (each cluster consists of similar binary strings). Feature Maximization (Lamirel and Al Shehabi, 2015) is used to assess the performance (degree of impact) of each socio-economic driver computed as follows:

$$FF_c(f) = 2\left(\frac{FR_c(f) * FP_c(f)}{FR_c(f) + FP_c(f)}\right) \tag{3.2}$$

Where,

$$FP_c(f) = \frac{\sum_{d\in c} W_d^f}{\sum_{f'\in F_c,\ d\in c} W_d^{f'}} \tag{3.3}$$

and

$$FR_c(f) = \frac{\sum_{d\in c} W_d^f}{\sum_{c\in C}\sum_{d\in c} W_d^f} \tag{3.4}$$

FF_c is referred to as ***feature F-measure*** while FR_c and FP_c are called ***Feature Recall*** and ***Feature Predominance,*** respectively. W_d^f is the weight of the feature *f* for the data (binary string) *d* and F_c represents all the features present in the dataset associated with the cluster *c*. Feature Predominance measures the ability of *f* to describe cluster ***c***, while Feature Recall characterises *f* according to its ability to discriminate ***c*** from other clusters.

The approach allows the most typical and most representative features to appear in a given cluster. The features judged relevant for a given cluster are those whose F-measure are better than their average F-measure in clusters in which they occur and better than the average F-measure of all the features in the partition (Lamirel and Al Shehabi, 2015).

If a feature *f* occurs in some cluster(s) c'in partition ***C*,** then the average F-measure of *f* in those clusters is given as the sum of all F-measures of *f* divided by the number of clusters in which the feature occurs. This is computed as follows:

$$\overline{FF}(f) = \sum_{c'\in C} \frac{FF_{c'}(f)}{|C_{/f}|} \tag{3.5}$$

Where $C_{/f}$ represents the subset of C in which the feature *f* occurs. On the other hand, the average F-measure of all the features in the partition is the sum of F-measures of all feature divided by the total number of features in the partition. This is computed as follows:

$$\overline{FF}_D = \sum_{f \in F} \frac{\overline{FF}(f)}{|F|} \qquad (3.6)$$

F is the total number of features in the partition. Therefore, membership of a set *Sc* of relevant features to appear in a given cluster *c* is therefore defined as follows:

$$S_c = \{f \in F_c | FF_c(f) > \overline{FF}(f) \text{ } and \text{ } FF_c(f) > \overline{FF}_D\}$$

(3.7)

Features not respecting the second condition (Equation (3.7)) in any cluster are discarded.

For each retained feature *f* in each cluster *c*, the *contrast* which is an indicator of the performance or strength of *f* in *c* is computed. Contrast is defined as the ratio between the *F*-measure of *f* (i.e. *FF*c(*f*)) and the average *F*-measure *FF* of *f* for the whole partition. The contrast of a feature *f* for a cluster *c* is expressed as shown in equation 3.8.

$$G_c(f) = \frac{FF_c(f)}{\overline{FF}(f)} \qquad (3.8)$$

Active features are those for which the contrast is greater than 1. Moreover, the higher the contrast of a feature for a cluster, the better its performance in describing the cluster content. For clearer elucidation, let the sample dataset in Table 3.5 be split into two clusters (C_1 and C_2) as shown in Table 3.6 by the clustering process.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 3.6 Sample Data in Two Clusters

	MaritalStats =m	**MaritalStats =s**	**MaritalStats =w**	**EduLevel =JSS**	**EduLevel =T**	**EduLevel =P**
C_1	1	0	0	1	0	0
	1	0	0	1	0	0
	0	0	1	1	0	0
	0	1	0	0	0	1
C_2	1	0	0	1	0	0
	0	1	0	0	1	0
	0	0	1	0	0	1
	0	0	1	0	0	1
	1	0	0	0	1	0
	0	1	0	0	0	1

The feature recall (Equation (3.4)) is computed stepwise as follows: Firstly, the weight of each feature in each cluster is computed and stored in a single dimensional array as follows: Let m be a number of features in the cluster (length of the binary vector). In Table 3.6, for instance, the number of features or binary vector length is 6. Let n be the number of data or binary vectors in the cluster. In Table 3.6, for instance, the number of binary vectors in cluster C1 is 4. Let W_f be the weight of feature f. The weight of MaritalStats in C_1, for instance, is 2. The weights are computed and stored sequentially in a single dimensional array, W_f_C_1[] as shown in Algorithm 3.1.

Algorithm 3.1

```
$W_f$=0
$W_f$_C1[]
For i ← 0 to n-1
    For j ← 0 to m-1
       $W_f$ ← $W_f$+ $C_1$[j][i]
    End for
    $W_f$_$C_1$[i] ←$W_f$
    $W_f$=0;
End for
```

So for clusters C_1, C_2, . . ., C_k, the procedure would compute and store the weights of their features in arrays W_f_C_1, W_f_C_2, . . .,W_f_C_k.

The second step involves a procedure to compute the total weight of each feature across all clusters. For each array W_f_C_i containing the weights of the features W_f in a cluster C_i, the weights are ordered and zero-indexed such that W_f_C_i[0], W_f_C_i[1], W_f_C_i[3], . . ., W_f_C_i[m-1] refers to the weight of the first, second, third, up to the mth feature respectively in the cluster C_i. The procedure to compute the weight of each feature in not only one cluster but across all clusters in the partition is given in Algorithm 3.2.

Algorithm 3.2

```
For i ← 0 to clustSize-1
    $W_f$_sum[i]←$W_f$_$C_1$[i]+$W_f$_$C_2$[i] + . . .+ $W_f$_Cm[i]
```

Where clustSize is the total number of clusters. This computes the total weight of each feature across all clusters in the partition in linear time. The total weights are also ordered and zero-indexed such that W_f_sum[0], W_f_sum[1], W_f_sum[3], . . ., W_f_sum[m-1] refers to the total weight of the first, second, third, up to the mth feature respectively across all clusters. For instance, the weight of **MaritalStats=m** in across all clusters is 4.

Feature recall for the first feature (*i.e.* MaritalStats=m) for instance is computed as FR(MaritalStats=m)=2/4=0.5 (sum of values in red divided by the sum of values with a blue border (Table 4.2).

In order to compute the feature predominance (Equation (3.3)), there is a need for a procedure, to sum up, the weights of all the features for each cluster. This procedure is executed in Algorithm 3.3.

Algorithm 3.3

```
                    $C_l$_weight=0
   For i ← 0 to n-1
    For j ← 0 to m-1
      $C_l$_weight ← $C_1$_weight+ C1[i][j]
    End for
            End for
```

Finally, feature predominance is computed, as shown below:

```
          For j ← 0 to m-1
FP_$C_i$[]← $\frac{\text{Wf\_Ci[i]}}{\text{Cl\_weight}}$
```

From Table 3.6, feature predominance of **MaritalStats=m** is done manually as **FP(MaritalStats=m)=2/8** =0.25 (sum of values in red divided by the sum of values with green border). Finally, the feature F-measure (Equation (3.2)) for each feature in a given cluster is computed as:

```
For j ← 0 to m-1
   F-Measure← $\frac{\text{FP\_Ci} * \text{FR\_Ci}}{\text{FP}_{\text{Ci}} + \text{FR\_Ci}}$
```

The Feature F-measure of **MaritalStats=m** is obtained as

FF(MaritalStats=m)= $2\left(\frac{0.5*0.25}{0.5+0.25}\right)$ = 0.33 (to 2 decimal places).

The full computations of Feature F-measures (FF) of individual features, average Feature F-measure across both clusters and overall average Feature-F measure are shown in Table 3.7. Features whose F-measures in both clusters are less than the overall average F-measure are discarded (stricken out in red in Table 3.7). This ensures that only features with significant impact are retained.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 3.7 Computations of Feature F-measures of all Features in the Sample Dataset

	Feature	FF in C1	FF in C2	Average FF	
	MaritalStats=m	0.33	0.25	0.29	
	~~MaritalStats=s~~	~~0.00~~	~~0.27~~	~~0.13~~	
	~~MaritalStats=w~~	~~0.18~~	~~0.27~~	~~0.22~~	
	EduLevel=JSS	0.50	0.13	0.31	
	EduLevel=T	0.00	0.29	0.14	
	EduLevel=P	0.17	0.38	0.27	
		Overall Average FF:		0.28	

Using equation 3.8, the contrasts of the retained features are then computed by dividing the Feature F-measure of each retained feature by its average. For **MaritalStats=m**, the contrast is $\frac{0.33}{0.29}$ for C_1 and $\frac{0.25}{0.29}$ and for C_2. The full computations of contrasts of retained features are shown in Table 3.8. The higher the contrast of a feature, the better its performance.

Table 3.8 Contrast of Retained Features

	C_1	C_2
MaritalStats=m	1.14	0.86
EduLevel=JSS	1.6	0.4
EduLevel=T	0.00	2.07
EduLevel=P	0.62	1.38

KNOWLEDGE, TRUTH AND EXCELLENCE

An element rightfully belongs to a given cluster if its contrast in that cluster is highest (commonly greater than 1). For this reason, ***Marital status =M*** and ***Edulevel=JSS*** are in C_1 and ***EduLevel=P,*** and ***EduLevel=T*** belong to C_2.

Computation of Optimal Model

Computation of the optimal model makes it possible to determine the optimal number of clusters (with the most salient features) which is appropriate for the given dataset. Optimal Model is the cluster size for which the data is most explained. The number of clusters for which both Positive Contrast (PC) and External Contrast (EC) are at peak is considered optimum. Where it is not possible for both to be at peak simultaneously for the same number of clusters, the optimal is taken at where PC is at a peak. The significance of PC in this modelling approach is in analogy with the commonly used measure of intra-cluster inertia. It is measured as a maximization of the averagely weighted contrast of active features for optimal partition. EC, on the other hand, is analogical to the combination of intra-cluster inertia and inter-cluster inertia in the usual models. It is measured as the maximization of the averagely weighted compromise between the contrast of active features and the inverted contrast of passive features for optimal partition. For a partition comprising *k* clusters, PC and EC are expressed respectively, as shown in Equation (3.9) and (3.10) (Lamirel and Al Shehabi, 2015).

$$PC = argmax\ k\left(\frac{1}{k}\sum_{i=1}^{k}\frac{i}{|s_i|}\sum_{f\in s_i} G_i\,(f)\right) \tag{3.9}$$

$$EC = argmax\left[\frac{1}{k}\sum_{i=1}^{k}\left(\frac{|s_i|\sum_{f\in s_i} G_i(f)+\overline{|s_\iota|}\sum_{h\in\bar{s}_i}\frac{1}{G_i h}}{|s|+|\bar{s}|}\right)\right] \tag{3.10}$$

Where, s_i is the number of data elements associated with a cluster *i*, $|s_i|$is the number of active features in *i*, and $\overline{s_\iota}$ the number of passive features in same the cluster.

The optimal model (optimal number of clusters) is the number of clusters for which both PC and EC are at peak or at least only the PC at peak. Detail steps for computation of optimal Model are shown below:

Compute the sum of PC and EC as S.

Arrange the various possible clusters by increasing the value of S.

Using the increasing value of S as a guide, scan through a file (arranged in ascending order cluster sizes) together with their respective ECs and PCs.

The optimal model is found where the PC and EC are simultaneously at peak.

Contrast Graphs

Contrast graphs are undirected bipartite graphs in which the clusters form one set of nodes, and their elements form another. Each cluster node connects to elements constituting it. Contrast graphs illustrate a clear relationship among its elements. Given the following two clusters, (C1) and (C2), the contrast graph shown in Figure 3.8 is generated.

Cluster 1 (C1)

****

1.918213	Age Group=A45_49
1.344814	Marital Status=Married
1.255502	Education Level=Middle

Cluster 2 (C2)

****

1.898805	Age Group=A50_54
1.158457	Marital Status=Married
1.100326	Education Level=Middle

In the graph in fig. 3.8 for instance, we identify relationship (interactivity) among Age Group 45-49, Middle education level and Married Marital Status. We can also identify interactivity among Age Group 50-54, Middle education level and Married Marital Status.

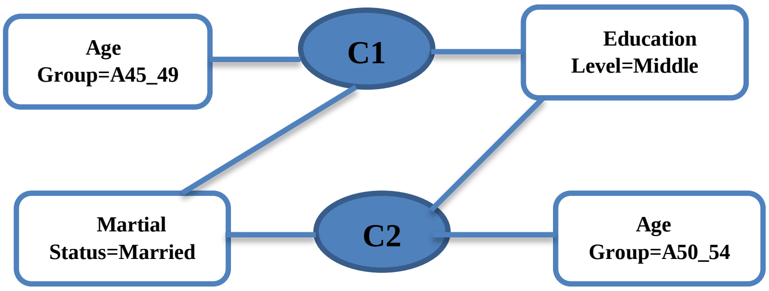


Figure 3.8 Contrast Graph

Implementation Design and Programming platform

The component diagram of the computational model is shown in Figure 3.9. A component diagram shows the structural relationships between the components of a system for easy implementation (Bell, 2004).

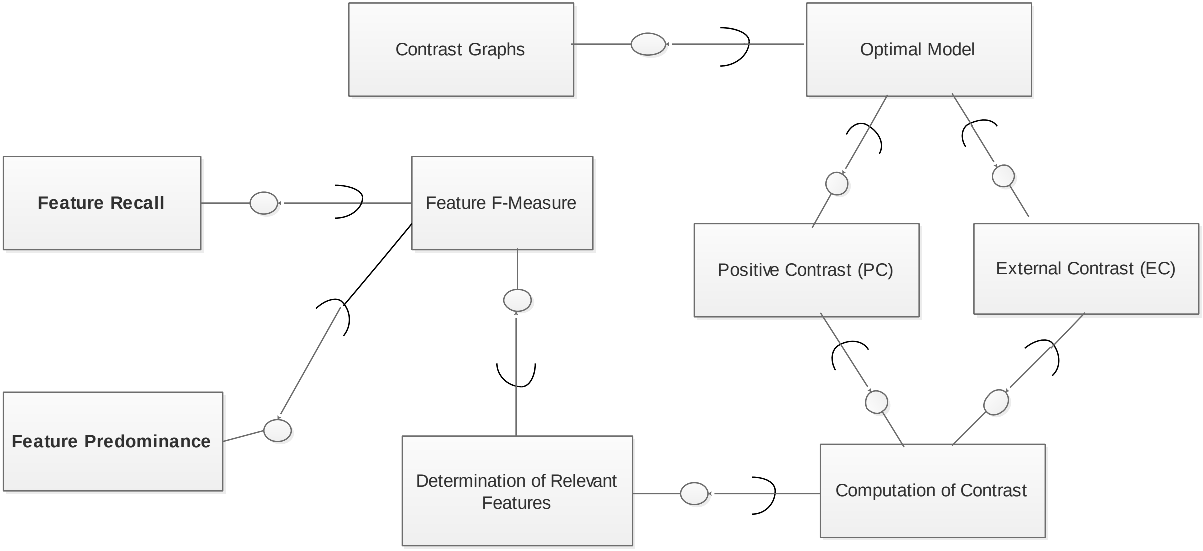


Figure 3.9 Component Diagram of Model

The model was implemented in the C++ Programming Language and ran on the Linux Shell.

SYSTEM IMPLEMENTATION

Preamble

In this chapter, four key aspects of the research are discussed. Firstly, how the data is organised for the research is discussed. To avoid generalised findings, the data is organised in such a way to aid findings which are peculiar to specific population subgroups. Secondly, the mode of operation of the implemented system is presented. This includes an evaluation of the computational efficiency of the system. Thirdly, the output of the model using the data is presented. Two approaches were applied separately to generate the outputs. Outputs of the first and second approaches are given and discussed under Table 4.5 and 4.6, respectively.

As discussed in Section 3.3, two datasets; namely HIV dataset and control (non-HIV) dataset were required for this research. By using these two distinct datasets, the researcher can compare and distinguish behaviours of HIV-driving factors in the HIV situation (using results from HIV dataset) from the non-HIV situation (using results from the non-HIV).

For the implementation, each of the two datasets was divided into *Rural* and *Urban*. The *Rural dataset* was further divided into *Rural Male* and *Rural Female.* Analogically, the *urban dataset* was further divided into *Urban Male* and *Urban Female*. As a result, the distinct datasets obtained from the *HIV dataset* and used for the implementation of the model are *Rural Male HIV*, *Rural Female HIV*, *Urban Male HIV*, and *Urban Female HIV*. Similarly, the distinct datasets obtained from the *control dataset* and used for the implementation of the model are *Rural Male control*, *Rural Female control*, *Urban Male control* and *Urban Female control.* Executing the model with each of these datasets separately allowed us to achieve deep insight into the possible different degrees of impact of the different HIV driving factors in different population subgroups. Table 4.1 shows the number of records (sizes) of the various datasets used for the study.

Table 4.1 Structure of used Data

HIV Dataset	Size	Control Dataset	Size
RURAL		**RURAL**	
RURAL MALE HIV	6257	REF RURAL MALE	6676
RURAL FEMALE HIV	5310	REF RURAL FEMALE	5202
ALL RURAL HIV	11 567	REF ALL RURAL	11 877
URBAN		**URBAN**	
URBAN MALE HIV	10092	REF URBAN MALE	9350
URBAN FEMALE HIV	9814	REF URBAN FEMALE	8215
ALL URBAN HIV	19906	REF ALL URBAN	17565

Design Implementation

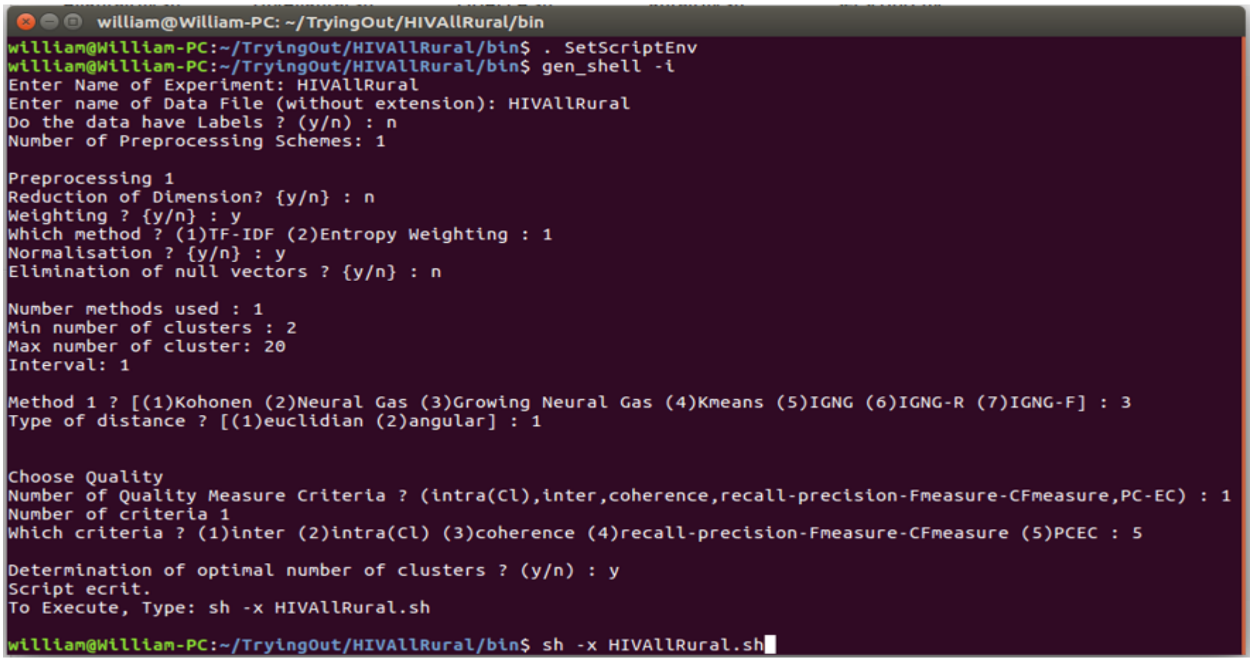
The system was implemented on a machine with the following specification:

Processor: Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz

Installed RAM: 8.00 GB

Operating System: Ubuntu Linux 16.0

The codes were implemented in C++ and launched through the Linux Shell. The screenshot in the Figure. 4.1 illustrates the system used. Once the user specifies the necessary information, the system begins to execute automatically.



```
william@William-PC: ~/TryingOut/HIVAllRural/bin
william@William-PC:~/TryingOut/HIVAllRural/bin$ . SetScriptEnv
william@William-PC:~/TryingOut/HIVAllRural/bin$ gen_shell -i
Enter Name of Experiment: HIVAllRural
Enter name of Data File (without extension): HIVAllRural
Do the data have Labels ? (y/n) : n
Number of Preprocessing Schemes: 1

Preprocessing 1
Reduction of Dimension? {y/n} : n
Weighting ? {y/n} : y
Which method ? (1)TF-IDF (2)Entropy Weighting : 1
Normalisation ? {y/n} : y
Elimination of null vectors ? {y/n} : n

Number methods used : 1
Min number of clusters : 2
Max number of cluster: 20
Interval: 1

Method 1 ? [(1)Kohonen (2)Neural Gas (3)Growing Neural Gas (4)Kmeans (5)IGNG (6)IGNG-R (7)IGNG-F] : 3
Type of distance ? [(1)euclidian (2)angular] : 1


Choose Quality
Number of Quality Measure Criteria ? (intra(Cl),inter,coherence,recall-precision-Fmeasure-CFmeasure,PC-EC) : 1
Number of criteria 1
Which criteria ? (1)inter (2)intra(Cl) (3)coherence (4)recall-precision-Fmeasure-CFmeasure (5)PCEC : 5

Determination of optimal number of clusters ? (y/n) : y
Script ecrit.
To Execute, Type: sh -x HIVAllRural.sh

william@William-PC:~/TryingOut/HIVAllRural/bin$ sh -x HIVAllRural.sh
```

Figure 4.1 Screenshot of System Execution

Assessment of System Performance with Varying Data Sizes

Three experiments were performed to test the speed of execution of the systems with varying input data sizes. In the first experiment, the number of records was held constant at 10,000 while the number of features varied from 50 to 200 with an interval of 50.

Thus, four datasets were created with the following dimensions:

Dataset 1: 10 000 records with 50 features; equivalent to 500 000 elements.

Dataset 2: 10 000 records with 100 features; equivalent to 1000 000 elements.

Dataset 3: 10 000 with 150 features; equivalent to 1 500000 elements.

Dataset 4: 10 000 with 200 features; equivalent to 2000 000 elements.

These datasets were run in turns, and the time used to execute was recorded. The result showed a linear curve with a gentle slope, as shown in Figure 4.2. This represents an average increase of 0.00003 seconds per unit increase in a number of elements.

The second experiment kept the number of features fixed at 50 while varying the number of records from 10 000 to 25 000 with an interval of 5000. The datasets thus used had the following dimensions:

Dataset 1: 10 000 records with 50 features; equivalent to 500 000 elements.

Dataset 2: 15 000 records with 50 features; equivalent to 750 000 elements.

Dataset 3: 20 000 records with 50 features; equivalent to 1000 000 elements.

Dataset 4: 25 000 records with 50 features; equivalent to 1 250 000 elements.

Running each dataset in turn and recording the time of execution, the results shown in Figure 4.3 was obtained. It shows a gentle slope representing an average of 0.00004 seconds per unit increase in a number of elements.

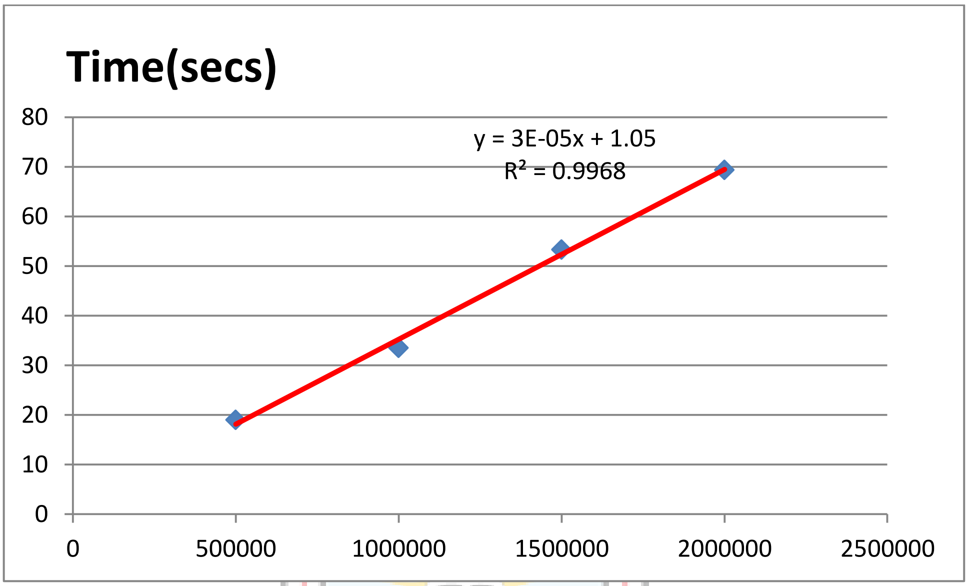


Figure 4.2 Time for processing Constant Records Size; varying Feature Size

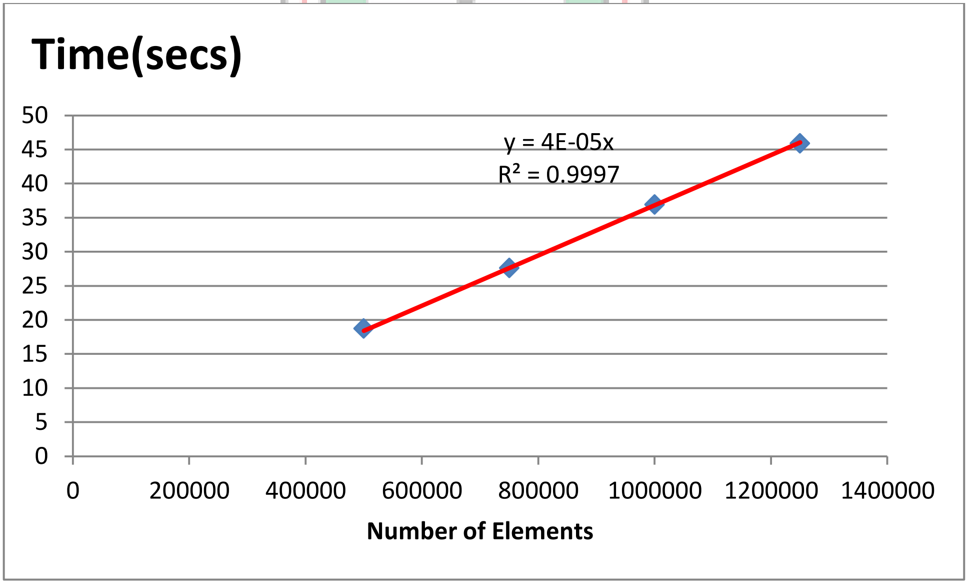


Figure 4.3 Time for processing Feature Size with varying Record Size

In the third experiment, both the number of records and the number of features were varied at the same time. The first dataset had 10 000 records with 50 features (500 000 elements); the second had 15 000 records with 100 features (1 500 000 elements), the third had 20 000 records with 150 features (3 000 000 elements), and the fourth had 25 000 records with 200 features (5 000 000). These datasets were executed in turns, and the time for executing each was observed. The result is shown in Figure 4.4.

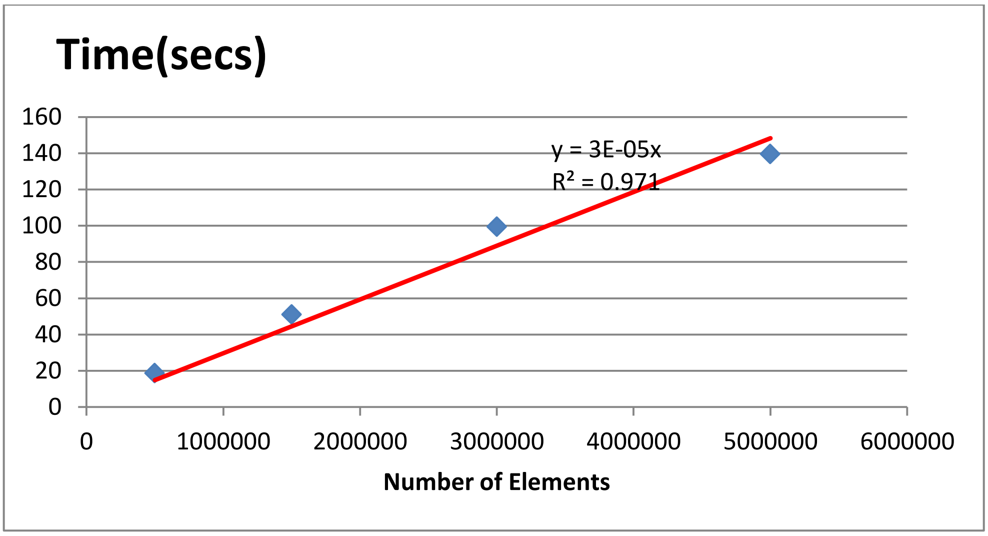


Figure 4.4 Time for varying intervals of both Number of Records and Features

It shows an average change of about 0.00003 seconds per unit change in a number of elements.

Observations from the experiments show that an increase in a number of elements has very little influence on the system execution time. In the average case, the system runs in linear time.

Degrees of Impact and Interactivities of HIV Drivers

The results generated are in the form of clusters. Only the optimal number of clusters as determined by the model is shown for each dataset. Each cluster is composed of only elements (HIV driving factors) which have been determined by the model to be interacting with each other. Each element in the cluster is preceded by a numerical value (the contrast) which shows the level of importance (degree of impact) of that element. The contrast of a given element can be interpreted as the weight or importance of that element in the cluster in which it appears relative to its importance in the rest of the cluster where it may appear.

Assumption for the Analysis and Criteria of Judgement

Key to this study is the assumption that the ideal (normal and acceptable) relationships and degrees of impact of HIV drivers are obtained using the model on the control (HIV-uninfected) dataset. This means that model results obtained with the HIV-infected dataset, which significantly deviate from the one obtained using the control data are abnormal due to HIV infection. Hence, to identify factors which are primarily associated with HIV infection, model results of both the control dataset and the HIV dataset are run side-by-side and compared and analysed. For the rest of this thesis, results obtained from the HIV dataset is referred to as *infected results*, whereas the ones obtained using the control dataset is referred to as the *uninfected results*.

Factors which appear in only the infected results with contrast value greater than 2 are considered unique drivers of the epidemic in the given population sub-group under consideration.

Degrees of Impact and Interactivities of HIV-Drivers in Rural Areas

Males in Rural Areas

The Rural Male HIV dataset contains data about only males infected with HIV in the selected Rural Districts, whereas the male reference dataset contains data about males who are uninfected by the epidemic in the same districts. The optimal model is found as 10 (clusters) for the Rural Male HIV infected dataset and 6 for the control dataset, as shown in Figure 4.5 and Figure 4.6, respectively. The elements in each cluster, as determined by the model, are shown in Table 4.2.

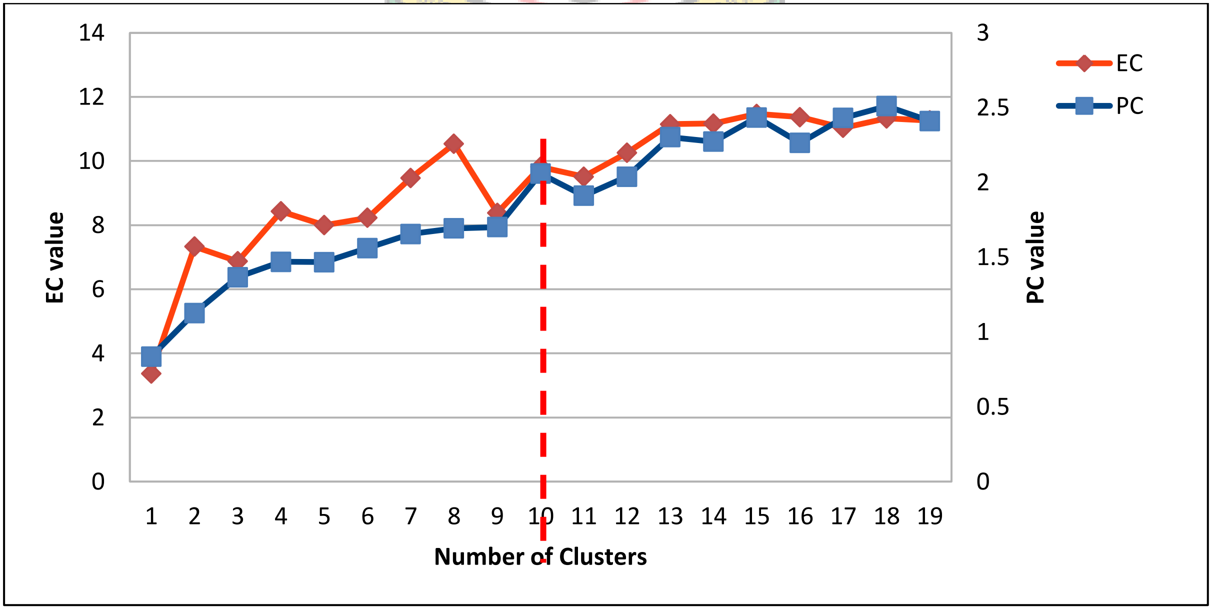


Figure 4.5 Optimal Model obtained for the Rural Male HIV Dataset

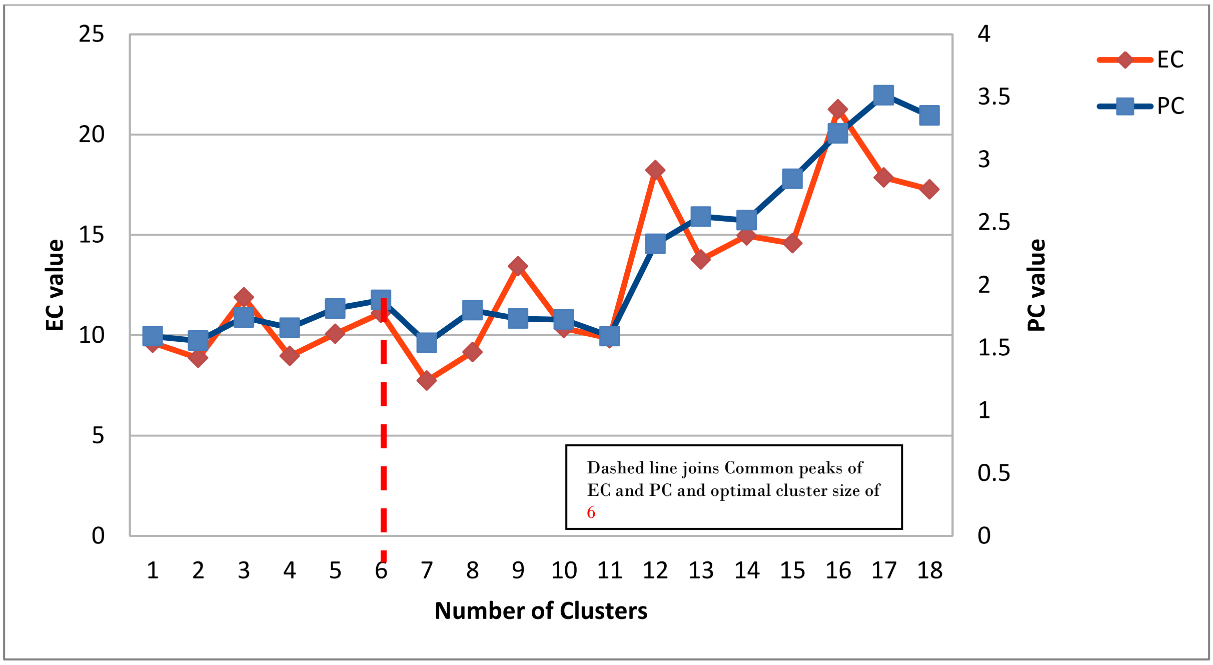


Figure 4.6 Optimal Model obtained for the Rural Male Control Dataset

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.2 Results of Running the Model with the Rural Male Dataset

Infected Results		**Uninfected Results**	
Cluster 0		**Cluster 0**	
4.008629	Education Level=Nil	1.913248	Marital Status=Single
2.300894	Age Group=A50_54	1.277313	Occupation=Agriculture,
2.245059	Marital Status=Widow(er)		Forestry and Fishery Workers
2.155972	Occupation =Agricultural,	1.228703	Education Level=Middle
	Forestry and Fishery Workers	1.126042	Education Level=Primary
2.083435	Age Group=A55_59		
2.054051	Age Group=A40_44	**Cluster 1**	
1.848949	Age Group=A45_49	2.405345	Marital Status=Married
1.844052	Marital Status=Married	1.969190	Age Group=A35_39
1.311808	Education Level=Middle	1.935236	Age Group=A55_59
		1.878215	Age Group=A50_54
Cluster 1		1.697582	Age Group=A30_34
3.608845	Occupation =Technicians and	1.624600	Education Level=Middle
	Associate Professionals	1.230837	Age Group=A45_49
2.074402	Age Group=A30_34	1.134562	Occupation=Agriculture,
2.043963	Age Group=A35_39		Forestry and Fishery Workers
1.914294	Marital Status=Single		
1.677940	Education Level=Post	**Cluster 2**	
	Secondary/ Tertiary	2.197001	Occupation=Crafts and
1.461117	Education Level=Middle		Related Trade Workers
1.107252	Marital Status=Married	1.970279	Education Level=Secondary
1.034693	Occupation =Agricultural,	1.926631	Age Group=A20_24
	Forestry and Fishery Workers	1.522213	Marital Status=Single
		1.406580	Age Group=A25_29
Cluster 2		1.329095	Education Level=Middle
5.097117	Education Level=Primary	1.056298	Occupation=Agriculture,
1.232031	Marital Status=Married		Forestry and Fishery Workers
1.213097	Age Group=A40_44		
1.189078	Age Group=A30_34	**Cluster 3**	
1.145683	Occupation =Agricultural,	2.856715	Education Level=Primary
	Forestry and Fishery Workers	2.024051	Marital Status=Single
1.048741	Age Group=A35_39	1.397865	Occupation=Agriculture,
			Forestry and Fishery Workers
Cluster 3			
1.762636	Occupation =Service and	**Cluster 4**	
	Sales Workers	1.795235	Marital Status=Married
1.347790	Age Group=A40_44	1.250272	Education Level=Middle
1.202222	Age Group=A30_34		
1.150557	Education Level=Middle		
1.092768	Marital Status=Single		
1.069541	Marital Status=Married		

Table 4.2 (continued) Results of Running the Model with the Rural Male Dataset

Infected Results		**Uninfected Results**	
Cluster 4		**Cluster 5**	
2.846262	Occupation =Other Occupations	2.015232	Education Level=Post Secondary/Tertiary
2.308932	Age Group=A25_29	1.918973	Education Level=Secondary
2.263804	Marital Status=Single	1.712579	Occupation=Plant and Machine Operators, and Assemblers
1.900282	Age Group=A15_19	1.178478	Age Group=A60_64
1.727758	Age Group=A0_14	1.112197	Age Group=old_Age
1.234015	Education Level=Nil		
Cluster 5			
1.736297	Occupation =Crafts and Related Trade Workers		
1.640472	Age Group=A25_29		
1.356735	Marital Status=Single		
1.319224	Age Group=A40_44		
1.265303	Education Level=Middle		
1.142095	Age Group=A35_39		
1.126917	Age Group=A30_34		
Cluster 6			
4.343273	Marital Status=Divorced		
1.258477	Age Group=A45_49		
Cluster 7			
2.918809	Occupation =Plant and Machine Operators, and Assemblers		
1.287491	Age Group=A35_39		
1.284873	Education Level=Middle		
1.143473	Age Group=A40_44		
1.126238	Age Group=A30_34		
Cluster 8			
2.735283	Education Level=Post Secondary/Tertiary		
1.420250	Age Group=A50_54		

Key Observations from the Model output of the Rural Male Dataset

Several unique drivers of HIV with contrast greater than or equal to 2 can be highlighted in Table 4.2. These unique drivers, together with their respective contrast, are shown in Table 4.3.

Table 4.3 List of Uniquely Significant Drivers of HIV among Men in Rural Areas

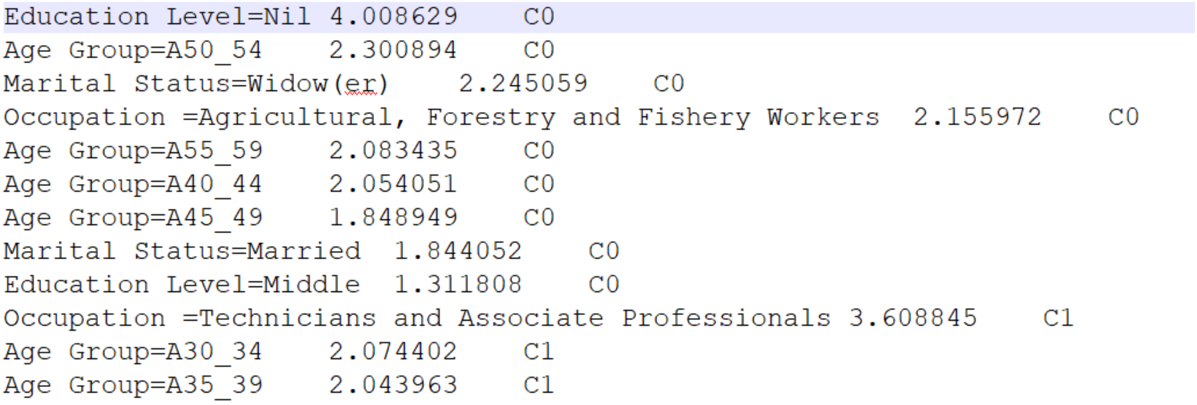
HIV-Driver	Degree of Impact
Marital Status=Divorced	4.343273
Education Level=Nil	4.008629
Occupation =Technicians and Associate Professionals	3.608845
Occupation =Other Occupations	2.846262
Marital Status=Widow(er)	2.245059
Age Group=A40_44	2.054051

The contrast graphs/ Interactivities of Factors among Rural Males

An element e_i with contrast value c_v belonging to a cluster c_k can be defined descriptively as: e_i c_v c_k. Hence a set of elements e_1, e_2, e_3 with respective contrast values c_1, c_2, c_3 belonging to the same cluster c_k can be described textually in a graph file as follows:

e_1 c_1 c_k
e_2 c_2 c_k
e_3 c_3 c_k

If there are several clusters c_1, c_2, . . ., c_n, with each comprising of several elements (some of which may belong to same cluster), then a complex bipartite graph is formed such that, c_1, c_2, . . ., c_n represent one set of nodes and their respective sets of elements along with their contrast values form another set of nodes. We term such a bipartite graph as contrast graphs. It illustrates interactivities of factors in the given population subgroup. As part of the design of the model created in this work, a graph file is created during its execution to describe the structure of the contrast graph that can be derived according to the clusters formed with the given dataset. Figure 4.7 shows a screenshot describing content of sample content graph file.



```
Education Level=Nil 4.008629    C0
Age Group=A50_54     2.300894    C0
Marital Status=Widow(er)     2.245059    C0
Occupation =Agricultural, Forestry and Fishery Workers  2.155972    C0
Age Group=A55_59     2.083435    C0
Age Group=A40_44     2.054051    C0
Age Group=A45_49     1.848949    C0
Marital Status=Married   1.844052    C0
Education Level=Middle   1.311808    C0
Occupation =Technicians and Associate Professionals 3.608845    C1
Age Group=A30_34     2.074402    C1
Age Group=A35_39     2.043963    C1
```

Figure 4.7 Part of graph file for rural male infected result

Figure 4.8 and Figure 4.9 summarises the scheme of interactivities among factors in the results obtained in both male rural HIV infected and uninfected datasets, respectively. In the figures, cluster 0 is labelled "C0"; cluster 1 is labelled "C1", cluster 2 is labelled "C2" and so on.

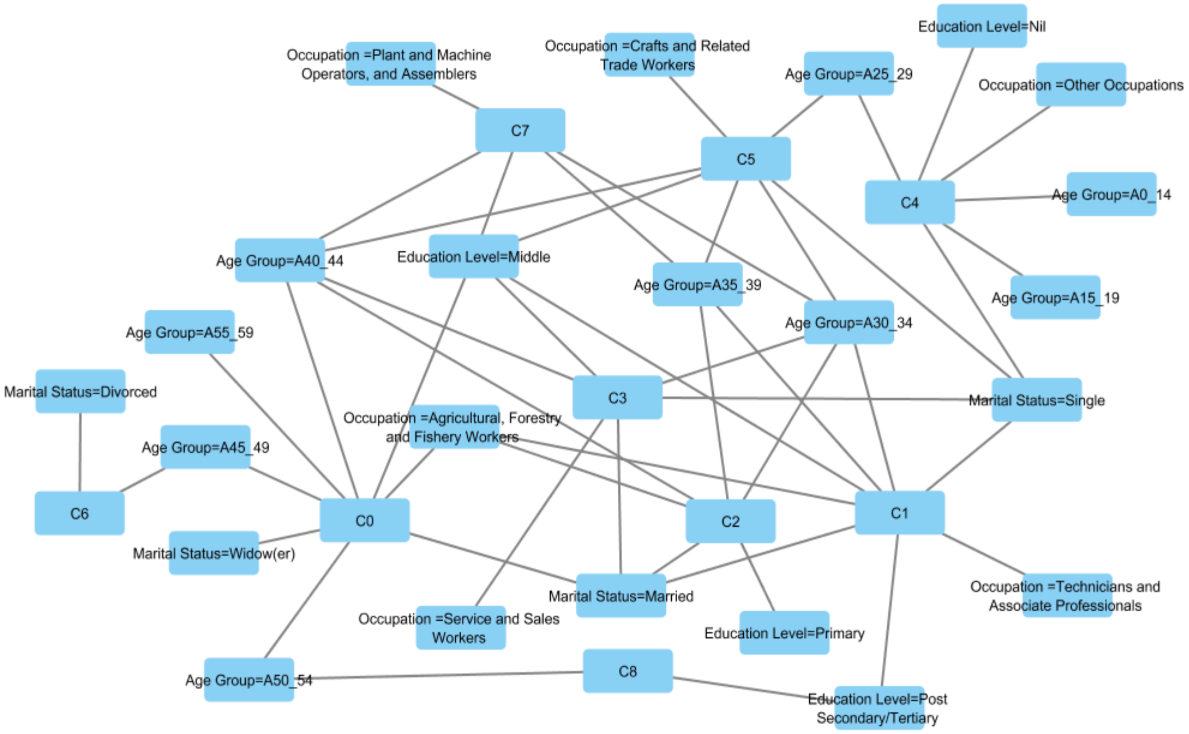


Figure 4.8 Interactivities of Factors in Rural HIV Male Dataset

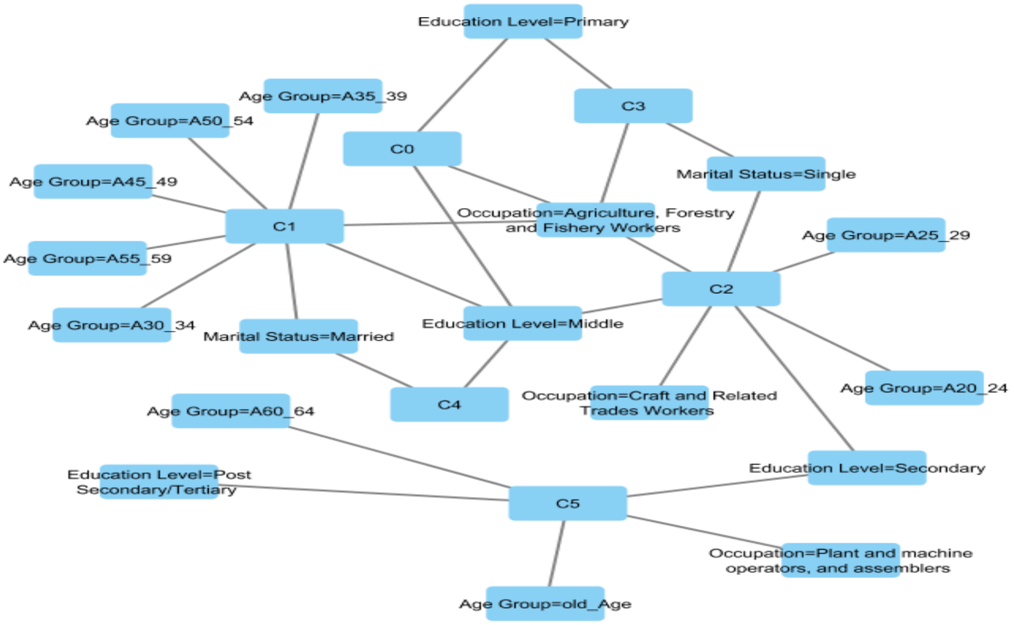


Figure 4.9 Interactivities of Factors in the Rural Male Control Dataset

Contrast Graph of Unique HIV Drivers

A contrast graph showing the interactivity of factors which are prevalent in only the infected result (and not appearing in the uninfected result) is referred to here as the contrast graph of unique HIV drivers. To generate this graph, we follow the strategy described hereafter:

Scan through the graph files of infected results and uninfected results at the same time.

Identify and copy elements which appear only in the graph file of infected results together with their clusters and contrast values and use them to generate a third graph file that will represent the interactions of the drivers that are unique to HIV.

A contrast graph of unique HIV drivers does not exclude factors with contrast less than 2. Figure 4.10 shows the interactivities of Contrast graph of unique HIV drivers associated with HIV among males in the rural areas.

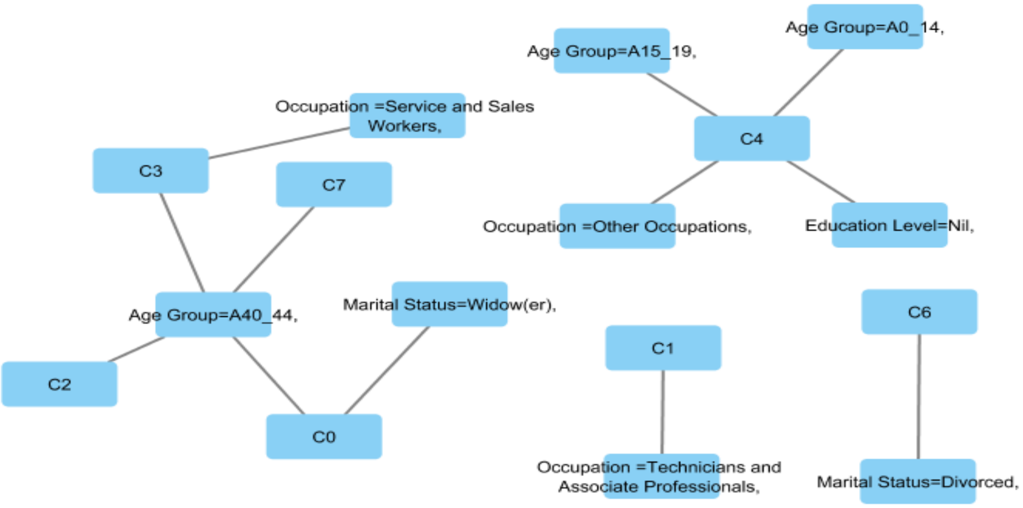

Age Group=A0_14,
Age Group=A15_19,
Occupation =Service and Sales Workers,
C4
C3
C7
Occupation =Other Occupations,
Education Level=Nil,
Age Group=A40_44,
Marital Status=Widow(er),
C6
C1
C2
C0
Occupation =Technicians and Associate Professionals,
Marital Status=Divorced,

Figure 4.10 Unique Interactivities of Factors among Males in Rural Areas

From Figure 4.10, it can be observed that age groups uniquely associated with HIV infection among males in the rural areas are in the range 0 to 19 and 40 to 44. It is observed that Age group 40-44 is associated with several different clusters; meaning that it is the age group most strongly associated infection of the epidemic than all the others. The figure also illustrates evident interactivity between age group 40-44 and marital status “Widow(er)” as well as occupation class: service and sales workers. Further observation also shows another interesting pattern where Age group 0 to 19 has similar likelihood of HIV infection like that of people with nil education and *other occupations* (Unemployed, Students, Refugees, and Prisoners).

Females in Rural Areas

The Rural Female HIV dataset contains data about only females infected with HIV in the selected Rural Districts whereas the Ref Rural female dataset (uninfected dataset) contains data about females who are uninfected by the epidemic in the same districts. Table 4.4 shows the results obtained for executing these two datasets. The rural female infected result is shown in column 1, and the uninfected female result is shown in column 2. From Table 4.4, a list of factors which appear uniquely in the infected results of the Rural HIV Female Dataset with contrast values greater than 2 are identified and organised into Table 4.5. These are factors which uniquely associate with HIV infection among females in rural areas. Female professionals are most at risk, followed by those with Post-Secondary and (or) tertiary education and others. "Nil" education level appears in as many as five out of the ten clusters of the infected female results but is completely missing in the uninfected results. "Nil" education level is, therefore, a driver of the epidemic in several groups of females in rural areas.

Table 4.4 Results of Running the Model with the Rural Female Dataset

Infected Results		**Uninfected Results**	
Cluster 0		**Cluster 0**	
2.644986	Age Group=A35_39	3.292952	Marital Status=Widow(er)
1.560635	Occupation =Agricultural, Forestry and Fishery Workers	2.002031	Marital Status=Divorced
1.536757	Occupation =Service and Sales Workers	**Cluster 1**	
1.448853	Marital Status=Married	2.019548	Marital Status=Divorced
1.443738	Education Level=Nil	1.783556	Age Group=A50_54
1.260808	Education Level=Primary	1.281842	Marital Status=Married
1.247137	Education Level=Middle	1.088769	Education Level=Middle
1.208602	Marital Status=Widow(er)	**Cluster 2**	
Cluster 1		1.945107	Age Group=A35_39
2.357719	Age Group=A25_29	1.651637	Marital Status=Married
2.217790	Occupation =Crafts and Related Trade Workers	1.517987	Occupation=Service and Sales Workers
1.703877	Marital Status=Married	1.323701	Age Group=A40_44
1.593675	Occupation =Service and Sales Workers	1.214918	Education Level=Middle
1.567044	Education Level=Middle	**Cluster 3**	
1.435670	Education Level=Primary	1.487057	Marital Status=Married
1.426798	Occupation =Other Occupations	1.244502	Occupation=Crafts and Related Trade Workers
1.400950	Marital Status=Single	1.212065	Occupation=Service and Sales Workers
1.268848	Education Level=Nil	1.071556	Education Level=Middle
1.145078	Occupation =Agricultural, Forestry and Fishery Workers		

Table 4.4. (Continued) Results of Running the Model with the Rural Female Dataset

Infected Results		Uninfected Results	
Cluster 2		**Cluster 4**	
2.800669	Age Group=A50_54	1.988860	Age Group=A25_29
2.713302	Age Group=A55_59	1.655739	Occupation=Crafts and Related Trade Workers
2.641037	Marital Status=Widow(er)		
1.776654	Marital Status=Divorced	1.632822	Education Level=Secondary
1.717280	Age Group=old_Age	1.550731	Occupation=Service and Sales Workers
1.496566	Age Group=A60_64		
1.397997	Occupation =Agricultural, Forestry and Fishery Workers	1.480767	Marital Status=Married
		1.251490	Education Level=Middle
1.011374	Education Level=Nil		
		Cluster 5	
Cluster 3		2.164617	Education Level=Secondary
3.027541	Age Group=A30_34	1.992732	Age Group=A20_24
1.966803	Occupation =Service and Sales Workers	1.858071	Marital Status=Cohabiting
1.962421	Marital Status=Married	1.533250	Occupation=Crafts and Related Trade Workers
1.677350	Education Level=Middle		
1.536707	Occupation =Agricultural, Forestry and Fishery Workers	1.385329	Occupation=Service and Sales Workers
1.478522	Education Level=Nil	1.207144	Education Level=Middle
1.437030	Education Level=Primary	1.201082	Marital Status=Single
1.177244	Occupation =Other Occupations	1.102539	Marital Status=Married
Cluster 4		**Cluster** 6	
1.986353	Marital Status=Cohabiting	3.302146	Marital Status=Single
		2.657815	Education Level=Primary
Cluster 5		1.815044	Occupation=Agriculture, Forestry and Fishery Workers
6.569744	Occupation =Professionals		
4.565594	Education Level=Post Secondary/ Tertiary		
		Cluster 7	
1.331375	Marital Status=Single	2.356793	Marital Status=Single
		1.335006	Occupation=Agriculture, Forestry and Fishery Workers
Cluster 6			
2.779346	Age Group=A40_44	1.167254	Education Level=Middle
1.573141	Marital Status=Widow(er)		
1.343558	Occupation =Agricultural, Forestry and Fishery Workers		
1.187571	Education Level=Nil		
1.066560	Occupation =Service and Sales Workers;		
1.034652	Education Level=Primary		
Cluster 7			
2.470072	Age Group=A20_24		
1.972271	Occupation =Crafts and Related Trade Workers		
1.501043	Marital Status=Single		
1.419691	Occupation =Other Occupations		
1.208839	Education Level=Primary		
1.141579	Marital Status=Married		
1.088079	Education Level=Middle		

Table 4.4. (Continued) Results of Running the Model with the Rural Female Dataset

Infected Results	Uninfected Results
Cluster 8	
2.876099 Age Group=A45_49	
1.819904 Marital Status=Divorced	
1.559062 Marital Status=Widow(er)	
1.247307 Occupation =Agricultural, Forestry and Fishery Workers	
Cluster 9	
3.209884 Age Group=A15_19	
2.582715 Marital Status=Separated	
2.330475 Age Group=A60_64	
1.951291 Age Group=A0_14	
1.544978 Marital Status=Single	
1.256283 Occupation =Other Occupations	

Table 4.5 HIV Drivers uniquely affecting Female in Rural Areas

HIV-Driver	Degree of Impact
Professionals	6.569744
Post-Secondary/ Tertiary	4.565594
Age Group 30 - 34	3.027541
Age Group 45 - 49	2.876099
Separated	2.582715

Contrast graph of unique HIV drivers is shown in Figure 4.11. This figure is an illustration of interactivity of factors which are uniquely associated with HIV among females in the rural areas.

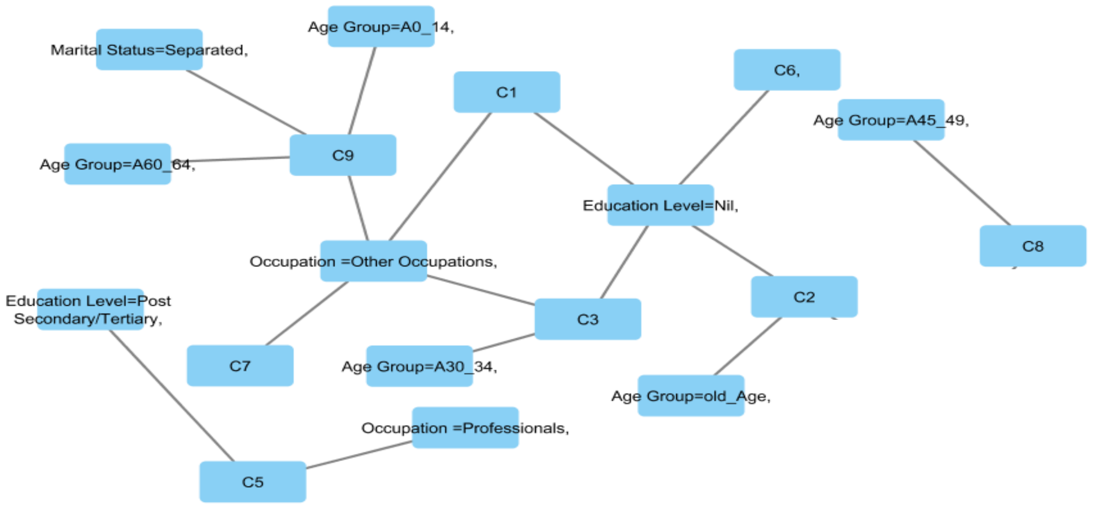


Figure 4.11 Unique Interactivities of Factors among Females in Rural Areas

From Figure 4.11, the “other occupations” and “Nil” education level are associated with as many as four different clusters. The more there are associations of a factor with other clusters, the more there are interactivities of that factor with other factors which are associated with those clusters. As can be observed from Figure 4.11, “other occupations” interacts with factors such as age group 60-64, separated marital status, age group 0-14, age group 30-34 and education level “Nil”. Similarly, education level “Nil” interacts with age group 30-34, old age and marital status divorced. Finally, marital status divorced interacts with education level “Nil”, old age and age group 45-49.

Males in Urban Areas

The Urban Male HIV data contains data about only males infected with HIV in the selected Urban Districts, whereas the urban male uninfected dataset contains data about males who are uninfected by the epidemic in the same districts. The optimal model found is 7 (clusters) for the urban Male HIV infected dataset and 12 for the uninfected dataset, as shown in Table 4.6.

Only one feature is observed to be uniquely characterising HIV infection among the urban males, as shown in Table 4.7, together with its degree of impact.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.6 Results of Running the Model with the Urban Male Dataset

Infected Results		**Uninfected Results**	
Cluster 0		**Cluster 0**	
1.951542	Marital Status=Single	2.783034	Age Group=A30_34
1.444146	Education Level=Primary	1.645974	Education Level=Secondary
1.418801	Age Group=A30_34	1.626574	Education Level=Post Secondary/Tertiary
1.183340	Education Level=Post Secondary/Tertiary		
		1.575225	Occupation=Crafts and Related Trade Workers
Cluster 1		1.569387	Marital Status=Married
1.977912	Occupation=Elementary Occupations	1.527306	Occupation=Professionals
		1.333472	Education Level=Middle
1.790506	Occupation=Professionals	1.278315	Marital Status=Single
1.733778	Education Level=Post Secondary/Tertiary	1.276014	Occupation=Service and Sales Workers
Cluster 2		**Cluster 1**	
2.614475	Education Level=Nil	2.681458	Occupation=Elementary Occupations
1.957127	Occupation=Agricultural, Forestry and Fishery Workers		
		2.498171	Occupation=Service and Sales Workers
1.561263	Age Group=A50_54		
1.377661	Age Group=A45_49	2.489594	Occupation=Plant and Machine Operators, and Assemblers
1.342483	Education Level=Primary		
1.325181	Marital Status=Married	1.581500	Education Level=Secondary
1.269767	Age Group=A40_44	1.547735	Marital Status=Single
1.193935	Education Level=Middle	1.249468	Age Group=A25_29
1.054822	Age Group=A35_39		
		Cluster 2	
Cluster 3		3.137993	Education Level=Primary
1.468653	Age Group=A30_34	2.554903	Marital Status=Single
1.414936	Education Level=Middle	1.967819	Occupation=Agriculture, Forestry and Fishery Workers
1.315123	Age Group=A35_39		
1.256056	Marital Status=Single		
1.234072	Education Level=Primary	**Cluster 3**	
1.093622	Age Group=A45_49	1.818594	Age Group=A50_54
1.050292	Marital Status=Married	1.661933	Marital Status=Married
1.035341	Age Group=A40_44	1.616069	Age Group=A55_59
		1.365181	Education Level=Post Secondary/Tertiary
		1.356440	Occupation=Professionals
		1.249542	Education Level=Middle
		1.087985	Occupation=Agriculture, Forestry and Fishery Workers

Table 4.6 (Continued) Results of Running the Model with the Urban Male Dataset

Infected Results		**Uninfected Results**	
Cluster 4		**Cluster 4**	
1.208944	Education Level=Middle	1.910659	Age Group=A60_64
1.181253	Age Group=A30_34	1.878042	Age Group=old_Age
1.178488	Education Level=Post Secondary/Tertiary	1.123444	Occupation=Agriculture, Forestry and Fishery Workers
1.163290	Marital Status=Single		
1.119377	Age Group=A35_39	**Cluster 5**	
1.116107	Age Group=A40_44		
1.077785	Age Group=A45_49	2.979633	Marital Status=Single
1.074752	Marital Status=Married	1.957534	Age Group=A20_24
Cluster 5		1.579618	Occupation=Crafts and Related Trade Workers
		1.577696	Education Level=Secondary
1.435585	Education Level=Middle	1.420892	Education Level=Primary
1.371735	Education Level=Primary	1.315464	Education Level=Middle
1.245463	Age Group=A35_39	1.148962	Occupation=Service and Sales Workers
1.168048	Age Group=A40_44		
1.154811	Age Group=A45_49		
1.102582	Marital Status=Married	**Cluster 6**	
1.061563	Age Group=A30_34		
		1.839500	Age Group=A35_39
		1.715176	Marital Status=Married
Cluster 6		1.455403	Education Level=Secondary
		1.300027	Education Level=Middle
1.286674	Education Level=Post Secondary/Tertiary	1.290296	Education Level=Post Secondary/Tertiary
		1.280874	Occupation=Plant and Machine Operators, and Assemblers
		1.173846	Occupation=Service and Sales Workers
		1.144157	Occupation=Crafts and Related Trade Workers
		Cluster 7	
		1.838972	Age Group=A45_49
		1.526139	Marital Status=Married
		1.262921	Education Level=Middle
		1.082092	Occupation=Agriculture, Forestry and Fishery Workers
		Cluster 8	
		1.967487	Education Level=Post Secondary/Tertiary

Table 4.6 (Continued) Results of Running the Model with the Urban Male Dataset

Infected Results	Uninfected Results
	Cluster 9 1.849523 Age Group=A40_44 1.763477 Marital Status=Married 1.486773 Occupation=Plant and Machine Operators, and Assemblers 1.369530 Education Level=Middle 1.003274 Occupation=Agriculture, Forestry and Fishery Workers **Cluster 10** 2.208235 Occupation=Crafts and Related Trade Workers 2.063315 Marital Status=Single 1.682385 Age Group=A25_29 1.588207 Occupation=Professionals 1.379750 Education Level=Post Secondary/Tertiary 1.288257 Education Level=Middle 1.279457 Education Level=Secondary 1.139419 Occupation=Agriculture, Forestry and Fishery Workers **Cluster 11** 5.163927 Marital Status=Widow(er) 3.829701 Marital Status=Divorced 1.047873 Age Group=A55_59

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.7 Features Unique to Urban Males

HIV-Driver	Degree of Impact
Education Level=Nil	2.614475

Females in Urban Areas

The Urban female HIV data contains data about only females infected with HIV in the selected Urban Districts, whereas the urban female control dataset contains data about females who are uninfected by the epidemic in the same districts. The optimal model found is 16 (clusters) for both datasets, which is illustrated in Table 4.8.

Table 4.8 Model Results of Urban Female Dataset

Infected Results		**Uninfected Results**	
Cluster 0		**Cluster 0**	
4.479073	Age Group=A35_39	5.208280	Age Group=A25_29
2.125906	Marital Status=Married	2.329849	Marital Status=Single
2.090858	Education Level=Nil	2.252027	Educational=Secondary
2.046750	Occupation=Service and Sales Workers	2.159574	Occupation=Crafts and Related Trade Workers
1.885887	Education Level=Middle	2.056489	Educational=Middle
1.712614	Education Level=Primary	1.994610	Marital Status=Married
1.627540	Occupation=Agricultural, Forestry and Fishery Workers	1.840436	Occupation=Service and Sales Workers
1.296361	Marital Status=Widow(er)	1.589225	Educational=Primary
1.233212	Occupation=Other Occupations	**Cluster 1**	
1.026104	Marital Status=Single	2.808761	Marital Status=Separated
Cluster 1		2.759210	Occupation=Other Occupations
3.107016	Occupation=Plant and Machine Operators, and Assemblers	**Cluster 2**	
3.607877	Occupation=Elementary Occupations	3.622670	Marital Status=Single
		2.852130	Age Group=A15_19
		2.576733	Educational=Primary
Cluster 3		1.972238	Age Group=A0_14
5.084776	Marital Status=Divorced	1.963223	Occupation=Agriculture, Forestry and Fishery Workers
1.143858	Age Group=A40_44	**Cluster 3**	
		4.770411	Occupation=Plant and Machine Operators, and Assemblers
Cluster 4		1.976307	Occupation=Technicians and Associate Professionals
5.567606	Marital Status=Separated	**Cluster 4**	
4.271327	Age Group=A60_64	4.012081	Marital Status=Widow(er)
1.501883	Marital Status=Widow(er)	2.967006	Age Group=old_Age
1.379150	Occupation=Agricultural, Forestry and Fishery Workers		
Cluster 5		**Cluster 5**	
4.068506	Occupation=Elementary Occupations	8.856977	Marital Status=Cohabiting
3.531359	Marital Status=Single	**Cluster 6**	
1.907711	Age Group=A0_14	7.288379	Educational=Post Secondary/Tertiary
1.361243	Age Group=A30_34	6.908644	Occupation=Professionals
1.272279	Education Level=Primary	**Cluster 7**	
		4.434245	Age Group=A20_24
		3.672093	Marital Status=Single
		2.225325	Educational=Secondary
		1.737209	Educational=Middle
		1.618529	Occupation=Crafts and Related Trade Workers
		1.593670	Occupation=Service and Sales Workers

Table 4.8 (Continued) Model Results of Urban Female Dataset

Infected Results		Uninfected Results	
Cluster 6		**Cluster 8**	
4.540652	Age Group=A25_29	5.067677	Age Group=A60_64
2.244929	Marital Status=Married	4.198468	Age Group=A55_59
2.223801	Education Level=Middle	3.379664	Marital Status=Widow(er)
2.172247	Occupation=Service and Sales Workers	1.947699	Marital Status=Divorced
2.106956	Education Level=Nil	**Cluster 9**	
1.888451	Occupation=Other Occupations	2.972932	Occupation=Clerical Support Workers
1.715412	Marital Status=Single	2.865662	Educational=Post Secondary/Tertiary
1.706651	Education Level=Primary		
		Cluster 10	
Cluster 7		4.831453	Age Group=A35_39
1.345212	Marital Status=Widow(er)	2.040092	Marital Status=Married
		1.810877	Educational=Middle
Cluster 8		1.724405	Educational=Primary
7.382702	Occupation=Professionals	1.691423	Occupation=Elementary Occupations
7.189729	Education Level=Post Secondary/Tertiary	1.506523	Occupation=Service and Sales Workers
3.972005	Occupation=Technicians and Associate Professionals	1.457205	Educational=Secondary
1.464605	Marital Status=Single	1.249839	Occupation=Crafts and Related Trade Workers
1.366550	Occupation=Other Occupations		
		Cluster 11	
1.352321	Age Group=A25_29	4.970760	Age Group=A30_34
1.286083	Marital Status=Married	2.187810	Marital Status=Married
		2.071633	Educational=Secondary
Cluster 9		1.913061	Occupation=Crafts and Related Trade Workers
3.348366	Age Group=A30_34	1.820305	Educational=Middle
2.915837	Marital Status=Married	1.630584	Educational=Primary
2.425462	Education Level=Nil	1.629663	Occupation=Service and Sales Workers
2.221923	Occupation=Service and Sales Workers		
2.168720	Education Level=Middle	**Cluster 12**	
1.802480	Occupation=Agricultural, Forestry and Fishery Workers	1.986022	Occupation=Managers
		1.566846	Educational=Post Secondary/Tertiary
1.618627	Occupation=Other Occupations		
		Cluster 13	
1.600540	Education Level=Primary	4.762494	Age Group=A40_44
		1.745923	Marital Status=Married
Cluster 10		1.722872	Educational=Middle
		1.260057	Occupation=Service and Sales Workers
4.603950	Marital Status=Cohabiting		
		1.228394	Occupation=Agriculture, Forestry and Fishery Workers

Table 4.8 (Continued) Model Results of Urban Female Dataset

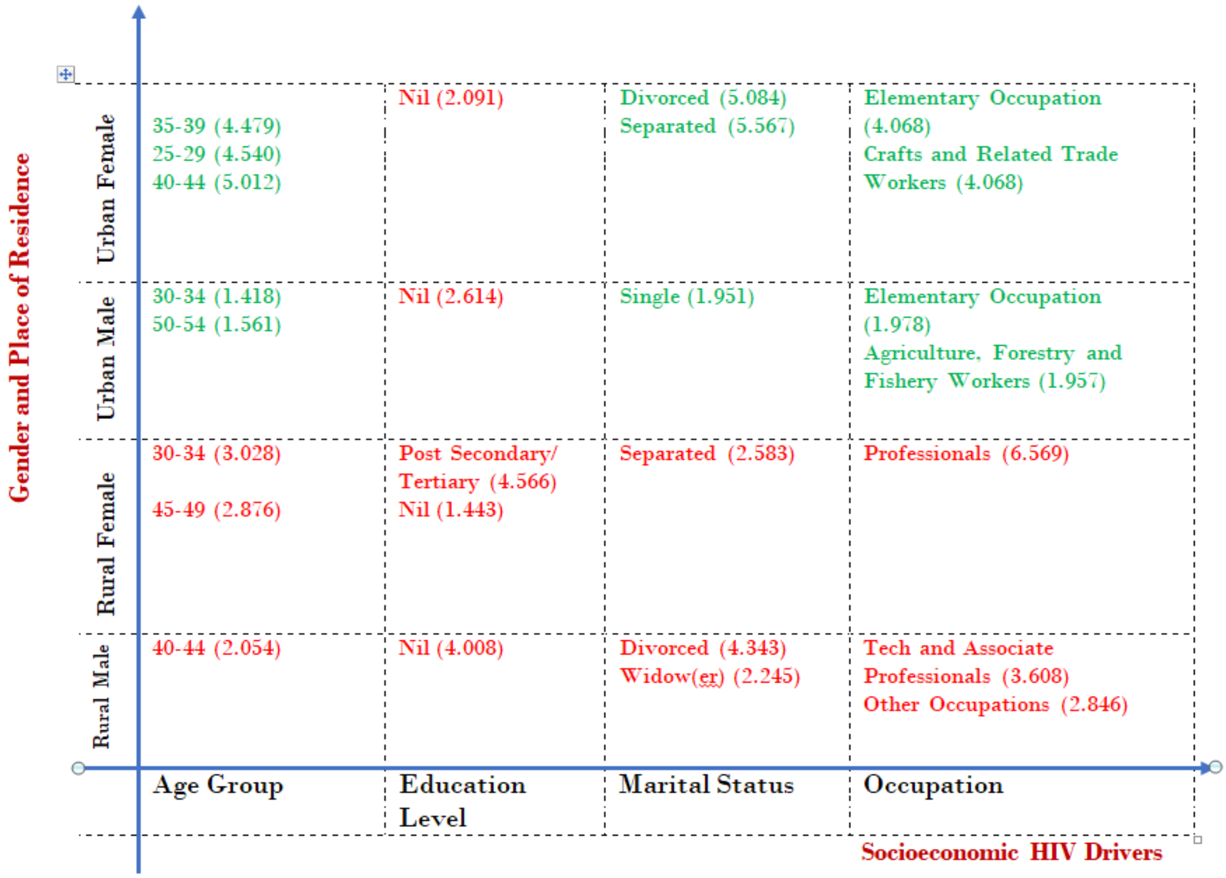
Infected Results		Uninfected Results	
Cluster 11		**Cluster 14**	
3.436676	Age Group=A20_24		
2.033201	Occupation=Other Occupations	4.804387	Age Group=A50_54
1.756621	Marital Status=Single	2.373115	Marital Status=Widow(er)
1.193674	Education Level=Nil	2.019994	Marital Status=Divorced
1.153984	Education Level=Post Secondary/Tertiary	1.291998	Educational=Middle
1.141551	Marital Status=Married	**Cluster 15**	
1.008355	Occupation=Crafts and Related Trade Workers	4.882922	Age Group=A45_49
Cluster 12		1.791028	Marital Status=Divorced
1.993904	Marital Status=Widow(er)	1.451504	Educational=Middle
Cluster 13		1.324768	Marital Status=Married
6.957918	Occupation=Crafts and Related Trade Workers		
1.274194	Education Level=Middle		
1.177730	Marital Status=Married		
Cluster 14			
3.865011	Age Group=A45_49		
1.840956	Marital Status=Widow(er)		
1.607357	Occupation=Agricultural, Forestry and Fishery Workers		
1.030765	Marital Status=Divorced		
Cluster 15			
5.012184	Age Group=A40_44		
2.119546	Marital Status=Widow(er)		
1.693490	Occupation=Agricultural, Forestry and Fishery Workers		
1.355035	Education Level=Nil		
1.352531	Occupation=Service and Sales Workers		
1.320563	Education Level=Primary		
1.313169	Marital Status=Married		
1.296695	Education Level=Middle		

KNOWLEDGE, TRUTH AND EXCELLENCE

Only one feature appears to be unique in the clusters formed from the female Urban HIV dataset (and not in the control dataset). That is the “Nil” education level. Also, certain factors appear not unique to the clusters of the female Urban HIV dataset, but they appear more significantly in it. These include people with Separated Marital Status and those with Elementary Occupations.

Summary of Findings

The approach taken in this first study makes it possible to identify essential drivers of the HIV epidemic while associating their influence on specific population sub-groups. The results obtained illustrate the degrees of impact of the various drivers in the different population sub-groups; namely Rural Male, Rural Female, Urban Male and Urban Female. For instance, even though “Nil” education level is observed to be an important driver of the epidemic across the four population sub-groups, its impact is greater among Rural males (degree of impact of 4.566) than all the others. Figure 4.12 is an illustration of how the various drivers are associated with the different population sub-groups. Each cell in the figure is an intersection between an HIV driver and a population sub-group. The degree of impact is indicated in brackets. Drivers appearing red represent those which uniquely characterise the given population sub-group with a contrast value greater than 2. Those coloured green, on the other hand, have high contrast values compared to other drivers in the HIV situation but not unique (to HIV situation).



	Age Group	Education Level	Marital Status	Occupation
Urban Female	35-39 (4.479) 25-29 (4.540) 40-44 (5.012)	Nil (2.091)	Divorced (5.084) Separated (5.567)	Elementary Occupation (4.068) Crafts and Related Trade Workers (4.068)
Urban Male	30-34 (1.418) 50-54 (1.561)	Nil (2.614)	Single (1.951)	Elementary Occupation (1.978) Agriculture, Forestry and Fishery Workers (1.957)
Rural Female	30-34 (3.028) 45-49 (2.876)	Post Secondary/ Tertiary (4.566) Nil (1.443)	Separated (2.583)	Professionals (6.569)
Rural Male	40-44 (2.054)	Nil (4.008)	Divorced (4.343) Widow(er) (2.245)	Tech and Associate Professionals (3.608) Other Occupations (2.846)

Figure 4.12 Associations of Important HIV Drivers and Population Sub-groups

Discussion

The findings of the study show that Age, Education Level, Marital Status and Occupation are critical socio-economic drivers (with varying degrees of impact) which potentially put one at risk of HIV infection. Further, these drivers are closely linked to one's setting (Urban or Rural) (Abebe *et al.*, 2003; Michelo *et al.*, 2006; Kleinschmidt, *et al.*, 2007) and gender (Male or Female) (Morison, 2001; Luke, 2005; Welz *et al.*, 2007; Barankanira *et al.*, 2016). The results of this study are corroborated in the extant literature on socio-economic drivers of HIV in Africa and particularly Sub-Saharan Africa and discussed in the following subsections in line with the aforementioned socio-economic drivers.

Age

The results obtained in this study indicate that males in rural residence within the age range 40-44 and the case of females, 30-34 and 45-49 are more prone to HIV infection. These findings on the age groups are similar to previous studies (Rosen *et al.*, 2008; Amornkul *et al.*, 2009; Obi *et al.*, 2011; Boerma, 2003; Abebe *et al.*, 2003; Welz *et al.*, 2007; Gómez-Olivé *et al.*, 2013) reported from some countries in Sub-Saharan Africa. For instance, Obi *et al.* (2011), found 25-34 to be the most-at-risk age group for HIV infection for females in rural parts of Nigeria. Similarly, studies by Gómez-Olivé *et al.*, (2013) and Amornkul *et al.*, (2009), found the age group 30-40 to more prone to HIV infections for both males and females in rural parts of South Africa and Kenya respectively. The finding on the age group 30-34 is consistent with a study by Obi *et al.*,(2011), which found that HIV prevalence was high among rural women of the same age group. Many rural males over 40 years are often involved in transgenerational (example marrying adolescents) or polygamous marriages increasing their risk of HIV infection (Luke, 2002; Kebaneilwe, 2011). Also, poverty has been found as a contributing factor to rural females' involvement in transactional sex especially with migrants putting them at risk of contracting HIV (le Booysen, 2004; Nattrass *et al.*, 2012). Additionally, many rural women are less empowered, and with their economic dependence on men, normative factors tend to render a lot of them powerless in negotiating safe sex (le Booysen, 2004; Welz *et al.*, 2007)

The results of the study suggest that for urban settings male within the age groups 30-34 and 50-54 are highly at risk of HIV infection while same is true for females within the age groups 25-29, 35-39 and 40-44. These findings are substantiated in previous studies (Amornkul *et al.*, 2009; Welz *et al.*, 2007; Tanser *et al.*, 2009; Sing and Patra, 2015) in the Sub-Saharan African region. For instance, Welz *et al.*, (2007), and Tanser *et al.*, (2009), found in their studies of HIV prevalence in urban settings of South Africa that the epidemic was high among men in the age group 30-34. Similarly, the findings on the age group 40-44 for urban females are echoed by Sing and Patra, (2015), who found HIV to be more prevalent among urban Tanzanian women in the age group 40-49. Urban self-style makes it easy for people to have access to media elements such as newspapers, television and radio. These media outlets promote HIV-prone behaviours such as early sexual debut in women, encourages multiple sexual partners in both sexes and increases the likelihood of extra-marital sex, especially among men (Hargreaves *et al.*, 2008).

Marital Status

The study found a close relationship between one's marital status and the likelihood of contracting HIV infection. Thus, different martial statuses promote certain behaviours that put one at risk of HIV infection. Results of the study indicate that divorce and widowhood are significant drivers of HIV infection among rural males. These findings resonate with the results of previous studies (Abebe *et al.*, 2003; Boerma, 2003; Okiria, 2014) in Sub-Sahara Africa. For instance, Boerma (2003), in a comparative study of HIV prevalence in Tanzania and Zimbabwe, found that disruptions in marriages such as divorce increase the risk of HIV infection among rural males in both countries. Also, rural females who are separated (i.e., not divorced but no longer living with a spouse) are prone to the infection. This confirms earlier findings by Yahya-Malima *et al.*, (2006) in Tanzania, Hargreaves *et al.*, (2007) in South Africa and Amornkul *et al.*, (2009), in Kenya. For males in urban areas, the study found that singleness is a potential HIV driver. This finding is corroborated by Amornkul et al. (2009) in a Kenyan study. On the other hand, marital statuses " Separated" and "Divorce" are potential drivers for females in urban settings contracting the epidemic. Similar findings are echoed by Bertrand (2016), who found that HIV prevalence was highest among Cameroonian women who were either divorced or separated in urban settings.

The findings on HIV and marital status can be explained from several perspectives, which include culture and adopted the lifestyle. For instance, ritual sex as part of widowhood rights and inheriting widows are cultural practices much sustained in rural settings in Africa which may explain why rural men are most at risk to HIV infection (Luke, 2002; Amornkul *et al.,* 2009). This is because if a partner dies of AIDS, surviving partners are potential carriers of the virus. Also, singleness due to late marriage promotes the lifestyle of frequent relationships and high rate of changing sex partners, which put single males at risk of infection (Bongaarts, 2007).

Occupation

The results obtained in this study indicate that the risk of HIV infection has a relationship with one's occupation. The movement of people from one place to another away from family for work has the potential to facilitate the spread of HIV through multiple sexual partnerships and transactional sex (Ojini and Coker, 2007; Nagoli et al., 2010).

The study found that both males and females in Elementary occupations living in urban areas are prone to the HIV epidemic. Also, females in Craft and Related Trade have a potential likelihood of contracting the infection. Skilled Agriculture, Forestry and Fishery workers who are males are most at risk of HIV infection. These findings are corroborated in previous studies (Arrehag *et al.*,2006; Ojini and Coker, 2007; Odimayo *et al.*, 2010; Nagoli *et al.*, 2010; Hüsken and Heck, 2012; Serbessa *et al.*, 2016) in some African countries. For example, a study conducted by Nagoli *et al.*, (2010), revealed that HIV prevalence was high among male fishing folks in Malawi. Similarly, Ojini and Coker, (2007), reported high HIV prevalence in fishing communities in Nigeria. Also, Odimayo *et al.* (2010), found that HIV infection in farming communities in Nigeria was high, leading to high morbidity and low productivity.

Further findings of the study show that Professional female workers in rural settings and males employed as "Technicians and Associate Professionals" and "Other occupations" are most at risk of the infection. These findings are consistent with earlier studies (Abebe *et al.*, 2003; Shisana *et al.*, 2004; Ncayiyana *et al.*, 2004; Amornkul *et al.*,2009; Bowa *et al.*,2016) conducted on HIV infection among persons with similar occupations. For instance, a study by Shisana et *al.*, (2004), argued that because HIV prevalence is higher in rural than urban, professionals who relocate to rural areas are at risk of infection. Notable among the professionals associated with occupational HIV risk are health workers (Ncayiyana *et al.*, 2004; Bowa *et al.*,2016). Many health facilities in Africa are under-resourced exposing health workers to several risks, including HIV infection, especially during surgery (Bowa *et al.*,2016).

Education Level

The results of the study show that one's level of education is an important driver of HIV infection. Further, the findings show that both males and females with "Nil" education either in a rural or urban setting are at risk of HIV infection. Several studies (De Walque 2009; Michelo *et al.*, 2006; Hargreaves *et al.*, 2007; Bogale *et al.*, 2009; Nel *et al.*, 2012; Sing and

Patra, 2015; Gumbe *et al.,* 2016) in the past have shown that "Nil" and low education are significant HIV driving factors. For instance, Bogale *et al.,* (2009), found that illiterate and those with low education in rural Ethiopia lacked knowledge on preventive measures of HIV, which put them at risk of contracting it. Also, the trend was similar for both males and females in urban settings. Likewise, a study by Michelo *al.,*(2006), among Zambian males living in urban areas reported high HIV prevalence. Low education, both in rural and urban settings makes one less likely to be aware of HIV and its routes of transmission (Johnson *et al.*, 2009). On the contrary, highly educated people tend to demonstrate better HIV preventive behaviour such as condom use and increased knowledge about HIV transmission and are more responsive to HIV/AIDS campaigns.

Further findings (as shown in Figure 4.12) suggest that post-secondary and tertiary education are driving factors of HIV in rural females. This finding mirror those of studies in other Sub-Saharan Africa countries; South Africa, Cameroon, Burkina Faso and Kenya (Hargreaves *et al.*, 2007; De Walque, 2009). High level of education tends to promote HIV-prone behaviours such as pre-marital sex, extra-marital sex, multiple sex partners (Hargreaves *et al.,* 2008; Bogale, 2009, Fortson, 2008; Fox, 2010). To this end, education appears to have a dual effect of potentially driving both HI-risky and HIV-protective behaviours (Parkhurst *et al.*, 2010).

Results obtained in this research, though are corroborated in literature are unique in several ways. Firstly, previous research works only identified HIV driving factors and reported them with absolutely no precision. For instance, with the earlier studies, it was not possible to state in comparative terms which of two HIV driving factors is a stronger driver of the epidemic in a given sub-population group than the other. For instance, from Figure 4.12, using the contrast or degree of impact measure; it is easy to conclude that, "Nil" education level is a stronger driver of the epidemic among rural males than it is among rural females or urban population. Furthermore, even though divorce and widowhood are both drivers of the epidemic among rural males, it is possible to compare the strengths or weights of the two using the contrast or degree of impact measure. More precisely, the method used in this study introduces novelty in the study of HIV epidemic by extending the ability to merely identify driving factors to measuring and reporting the impact of such factors with much more precision to ensure more explicit judgement decision making.

Secondly, previous studies could not determine the interactivity or associations among driving factors in various population sub-groups. The approach and method in this study have demonstrated the interactivities and associations of factors using contrast graph.

Combined Interactivity of Infected and Uninfected HIV drivers

The method in section 4.4 highlights the most salient and unique drivers of the HIV epidemic and how they interact in various population subgroups. This section highlights how drivers of the infected population interact with the drivers in the uninfected population in a given population subgroup using the following steps:

Approximately the same amount of records were taken from the infected and uninfected datasets and put together. As a result, the following datasets were produced:

Male Rural Mixed dataset – A dataset comprising of a portion of male HIV-infected dataset from the rural setting mixed with a portion of male HIV-uninfected dataset from the rural setting.

Male Urban Mixed dataset – A dataset comprising of a portion of male HIV-infected dataset from the urban setting mixed with a portion of male HIV-uninfected dataset from the urban setting.

Female Rural Mixed dataset – A dataset comprising of a portion of female HIV-infected dataset from the rural setting mixed with a portion of female HIV-uninfected dataset from the rural setting.

Female Urban Mixed dataset – A dataset comprising of a portion of female HIV-infected dataset from the urban setting mixed with a portion of female HIV-uninfected dataset from the urban setting.

These datasets were preprocessed and clustered using the Growing Neural Gas and further manipulated using Feature Maximization; paying attention to both contrast value and Feature F-measure. The frequency of features representing both the infected and uninfected datasets was recorded. For each cluster, features with F-measure greater than overall average F-measure and those with high contrast are shown in the Table 4.9 to Table 4.15. Table 4.9 shows clusters formed for Rural Males (Male Rural Mixed dataset) and Urban Males (Male Urban Mixed dataset). Table 4.10 is a comparison of HIV drivers (elements) characterising the two settings in Table 4.9. Table 4.11 shows clusters formed for Rural Females (Female Rural Mixed dataset) and Urban Females (Female Urban Mixed dataset). Table 4.12 is a comparison of HIV drivers (elements) characterising the two settings in table 4.11. Table 4.13 shows clusters formed for the Rural Males and Rural Females. Table 4.14 compares drivers (elements) characterising the two settings in Table 4.13. Table 4.15 shows clusters formed Urban Male and Urban Females. Finally, table 4.16 compares drivers (elements) characterising the two settings in Table 4.15.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.9 Clusters characterising Rural and Urban Males

Male Rural	**Male Urban**
Cluster 0 Freq. = 412 - HIV Freq. = 306 - No_HIV Age Group=A30_34 Occupation =Other Occupations Occupation =Technicians and Associate Professionals 1.553221 Education Level=Nil 1.501411 Occupation =Plant and Machine Operators, and Assemblers 1.448907 Occupation =Crafts and Related Trade Workers 1.428351 Marital Status=Married **Cluster 1** Freq. = 475 - No_HIV Freq. = 64 - HIV Age Group=A20_24 Age Group=A15_19 Marital Status=Single Education Level=Secondary **Cluster 2** Freq. = 131 - No_HIV Freq. = 20 - HIV Age Group=A0_14 Education Level=Primary Occupation =Elementary Occupations 2.011316 Marital Status=Single **Cluster 3** Freq. = 33 - HIV Freq. = 21 - No_HIV Occupation =Clerical Support Workers Marital Status=Widow(er) Occupation =Managers Marital Status=Separated	**Cluster 0** Freq. = 906 - No_HIV Freq. = 201 - HIV Age Group=A20_24 Marital Status=Single Age Group=A15_19 Age Group=A0_14 Education Level=Primary 1.727737 Education Level=Secondary **Cluster 1** Freq. = 19 - HIV Freq. = 6 - No_HIV Marital Status=Separated **Cluster 2** Freq. = 296 - No_HIV Freq. = 274 - HIV Age Group=A60_64 Age Group=old_Age Occupation=Other Occupations Marital Status=Widow(er) **Cluster 3** Freq. = 1039 - HIV Freq. = 651 - No_HIV Age Group=A35_39 Education Level=Post Secondary/Tertiary Occupation=Plant and Machine Operators, and Assemblers Occupation=Service and Sales Workers Education Level=Nil 1.561280 Occupation=Crafts and Related Trade Workers 1.464786 Marital Status=Married

Table 4.9 (Continued) Clusters Characterising Rural and Urban Males

Male Rural	**Male Urban**
Cluster 4 Freq. = 297 - HIV Freq. = 248 - No_HIV Age Group=A55_59 Age Group=old_Age Age Group=A60_64	**Cluster 4** Freq. = 816 - No_HIV Freq. = 312 - HIV Age Group=A25_29 Education Level=Secondary Marital Status=Cohabiting Occupation=Elementary Occupations Occupation=Clerical Support Workers 2.240200 Marital Status=Single 1.448680 Occupation=Crafts and Related Trade Workers

Cluster 5
Freq. = 346 - No_HIV
Freq. = 195 - HIV
 	
Age Group=A25_29
Occupation =Plant and Machine Operators, and Assemblers
Occupation =Crafts and Related Trade Workers
Marital Status=Cohabiting
1.740070 Marital Status=Single
	Cluster 5
	Freq. = 764 - HIV
	Freq. = 405 - No_HIV
	
	Age Group=A45_49

Cluster 6 Freq. = 570 - HIV Freq. = 238 - No_HIV Age Group=A35_39 Education Level=Nil Occupation =Agricultural, Forestry and Fishery Workers Education Level=Post Secondary/Tertiary 1.746948 Occupation =Plant and Machine O perators, and Assemblers 1.717551 Marital Status=Married 1.470118 Occupation =Crafts and Related Trade Workers	**Cluster 6** Freq. = 988 - HIV Freq. = 528 - No_HIV Age Group=A40_44 Marital Status=Married Education Level=Middle Occupation=Agricultural, Forestry and Fishery Workers Occupation=Technicians and Associate Professionals 1.720815 Occupation=Plant and Machine Operators, and Assemblers
Cluster 7 Freq. = 400 - HIV Freq. = 197 - No_HIV Age Group=A45_49 1.490862 Marital Status=Married	**Cluster 7** Freq. = 796 - No_HIV Freq. = 725 - HIV Age Group=A30_34 Occupation=Crafts and Related Trade Workers Occupation=Professionals 1.557639 Education Level=Secondary 1.503458 Education Level=Post Secondary/Tertiary 1.469876 Marital Status=Single

Table 4.9 (Continued) Clusters Characterising Rural and Urban Males

Male Rural	**Male Urban**
Cluster 8 Freq. = 521 - HIV Freq. = 246 - No_HIV Age Group=A40_44 Marital Status=Married Education Level=Middle Occupation =Service and Sales Workers 1.733479 Education Level=Nil	**Cluster 8** Freq. = 511 - HIV Freq. = 346 - No_HIV Age Group=A50_54 Marital Status=Divorced

Cluster 9
Freq. = 305 - HIV
Freq. = 147 - No_HIV
Age Group=A50_54
Marital Status=Divorced
Occupation =Professionals
	Cluster 9
	Freq. = 259 - HIV
	Freq. = 250 - No_HIV
	Age Group=A55_59
	Occupation=Managers

The first column of Table 4.9 shows the results obtained after running the **Male Rural Mixed dataset,** while the second column shows the result obtained from the **Male urban Mixed dataset.** The frequencies of features in each cluster coming from the HIV-infected data (shown simply as "HIV") and that of HIV-uninfected data (shown simply as "No_HIV") respectively are indicated. The ratio of the frequency of HIV-infected features in a cluster to that of HIV-uninfected features in the given cluster represents the *purity* of that cluster. The logic adopted for the sake of analysis in this section is stated as follows: *High purity (>=2) is an indication that the classes of people whose features (socio-economic drivers) are represented in that cluster are at high risk of HIV infection. Purity of 1 or less is an indication of low risk. Purity greater than 1 and less than 2 is an indication of some level of risk between low and high*.

Clusters with purities greater than or equal to 2 for males in the rural setting (Column 1 of Table 4.9) are cluster 6, cluster 7, cluster 8 and cluster 9. Cluster 6, for instance, is an indication that people of age group 35-39 with the following characteristics are prone to HIV infection:

Education Level=Nil

Occupation =Agricultural, Forestry and Fishery Workers

Education Level=Post Secondary/Tertiary

Occupation =Plant and Machine Operators, and Assemblers

Marital Status=Married

Occupation =Crafts and Related Trade Workers

In the case of males in the urban setting (Column 2 of table 4.9), only cluster 1 has purity greater than 2. Cluster 3, cluster 5, cluster 6 and cluster 8, however, are observed to have high purity; though less than 2. Because the total number of features in cluster 1 is extremely few compared to the size of the dataset, it is not emphasised.

Comparative analysis from Table 4.9

There are similar observable trends between the rural and the urban settings for the males.

Cluster 1 and cluster 2 of the rural males have very strong similarity with cluster 0 of the urban males. It exhibits that, age groups 24 and below (youth and children) with marital status single and low or no education (secondary, primary, nil) in both rural and urban settings have lower chances of contracting the disease.

The observation in cluster 5 of the rural setting is almost a subset of cluster 4 of the urban setting. It shows that the chance of contracting HIV among males of age group 25-29 are similar in both the rural and urban settings.

Cluster 9 of male rural and cluster 8 of male urban demonstrate a similar trend across both the rural and urban settings. They show that divorce is a strong driver of the epidemic for the age group 50-54.

Some good degree of similarity is also observed between cluster 8 of rural males and cluster 6 of urban males with respect to HIV drivers for the age group 40-44. The “married” marital status and “middle” education level are strong drivers in both settings. The concerned occupations differ between rural and urban people but they all concern low-level occupations.

Cluster 6 of the rural males have very similar observations also found in cluster 3 of the urban setting. Almost two-thirds of the total number in both clusters represent HIV infection for the age group 35-39. Again, the concerned occupations differ between rural and urban people, but they all concern low-level occupations.

Interestingly, some basic occupation like “plants and machine operators” occupation which is highly connected with HIV infection, seems to appear to influence HIV infection at an earlier age in a rural area (25-29 in cluster 5) than in urban area (35-39 in cluster 3).

Table 4.10 is a summary of the important HIV drivers grouped by age group across both rural and urban settings. Only clusters with purities greater than or equal to 2 are shown.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.10 Important HIV drivers prevailing in Both Rural and Urban Males

Age Groups	Rural	Urban	Drivers in both settings
25-29	**Occupation**=Plant and Machine Operators, and Assemblers, **Occupation**= Craft and Related Trade Workers	**Education Level**=Secondary **Occupation**=Clerical Support Workers **Occupation**=Elementary Occupations	**Marital Status**=Single; **Marital Status**=Cohabiting

35-39	**Occupation** =Agricultural, Forestry and Fishery Workers;	**Occupation**=Service and Sales Workers	**Education Level**=Nil;
			Education Level=Post Secondary/ Tertiary;
			Occupation=Plant and Machine Operators, and Assemblers;
			Occupation=Crafts and Related Trade Workers;
			Marital Status=Married

40-44	**Occupation** =Service and Sales Workers
	Education Level=Nil
		Occupation=Agricultural, Forestry and Fishery Workers; Occupation=Technicians and Associate Professionals; Occupation=Plant and Machine Operators, and Assemblers	**Marital Status**=Married;
			Education Level=Middle;

45-49	**Marital Status**=Married		
50-54	**Occupation**= Professional		**Marital Status**=Divorced;

Table 4.11 Clusters characterising Rural and Urban Females

Female Rural	**Female Urban**
Cluster 0 Freq. = 1508 - HIV Freq. = 439 - No_HIV Age Group=A30_34 Marital Status=Married Occupation =Service and Sales Workers Education Level=Nil 1.867247 Education Level=Middle 1.748347 Occupation =Crafts and Related Trade Workers 1.532082 Education Level=Primary 1.464499 Occupation =Agricultural, Forestry and Fishery Workers **Cluster 1** Freq. = 201 - HIV Freq. = 131 - No_HIV Marital Status=Cohabiting Occupation =Technicians and Associate Professionals Education Level=Secondary Occupation =Elementary Occupations Occupation =Plant and Machine Operators, and Assemblers Occupation =Clerical Support Workers Occupation =Managers **Cluster 2** Freq. = 351 - HIV Freq. = 97 - No_HIV Marital Status=Separated **Cluster 3** Freq. = 471 - HIV Freq. = 345 - No_HIV Age Group=A55_59 Age Group=A60_64 Age Group=old_Age Marital Status=Widow(er) 1.782844 Marital Status=Divorced	**Cluster 0** Freq. = 897 - HIV Freq. = 734 - No_HIV Age Group=A40_44 1.696954 Marital Status=Divorced 1.605439 Marital Status=Widow(er) 1.446666 Education Level=Nil 1.401224 Marital Status=Married **Cluster 1** Freq. = 35 - No_HIV Freq. = 33 - HIV Occupation=Managers Occupation=Plant and Machine Operators, and Assemblers Marital Status=Separated Occupation=Technicians and Associate Professionals **Cluster 2** Freq. = 863 - No_HIV Freq. = 466 - HIV Age Group=A55_59 Age Group=old_Age Age Group=A60_64 Age Group=A0_14 Marital Status=Widow(er) Occupation=Agricultural, Forestry and Fishery Workers 1.469891 Marital Status=Divorced 1.440069 Education Level=Primary **Cluster 3** Freq. = 1095 - HIV Freq. = 886 - No_HIV Age Group=A25_29 Marital Status=Married Marital Status=Cohabiting Occupation=Crafts and Related Trade Workers 1.705788 Education Level=Middle 1.502427 Occupation=Service and Sales Worker 1.440463 Education Level=Primary 1.404536 Education Level=Nil

Table 4.11(Continued) Clusters characterising Rural and Urban Females

Female Rural	**Female Urban**
Cluster 4 Freq. = 1443 - HIV Freq. = 678 - No_HIV Age Group=A25_29 Education Level=Middle Occupation =Crafts and Related Trade Workers Occupation =Other Occupations 2.030653 Marital Status=Married 1.521765 Education Level=Primary 1.494377 Education Level=Nil	**Cluster 4** Freq. = 1803 - HIV Freq. = 1064 - No_HIV Age Group=A30_34 Occupation=Service and Sales Workers Education Level=Middle Education Level=Nil 2.169248 Marital Status=Married 1.555977 Occupation=Crafts and Related Trade Workers 1.467998 Education Level=Primary
Cluster 5 Freq. = 1283 - HIV Freq. = 457 - No_HIV Age Group=A35_39 1.812025 Occupation =Service and Sales Workers 1.793653 Marital Status=Married 1.789657 Education Level=Nil 1.523655 Education Level=Middle	**Cluster 5** Freq. = 1455 - HIV Freq. = 898 - No_HIV Age Group=A35_39 Education Level=Primary Occupation=Elementary Occupations 1.904816 Education Level=Nil 1.899796 Marital Status=Married 1.621375 Education Level=Middle 1.532374 Occupation=Service and Sales Workers 1.409826 Marital Status=Divorced

Cluster 6
Freq. = 562 - HIV
Freq. = 199 - No_HIV
Education Level=Post Secondary/Tertiary
Occupation =Professionals
	Cluster 6
	Freq. = 512 - No_HIV
	Freq. = 433 - HIV
	Age Group=A50_54
	Marital Status=Divorced
	2.416490 Marital Status=Widow(er)

Cluster 7 Freq. = 837 - HIV Freq. = 333 - No_HIV Age Group=A40_44 1.622725 Marital Status=Divorced 1.484052 Education Level=Nil 1.431678 Marital Status=Widow(er)	**Cluster 7** Freq. = 625 - HIV Freq. = 557 - No_HIV Age Group=A45_49 2.120777 Marital Status=Divorced 1.975501 Marital Status=Widow(er)
Cluster 8 Freq. = 807 - HIV Freq. = 736 - No_HIV Age Group=A20_24 2.333102 Occupation =Crafts and Related Trade Workers 1.592644 Education Level=Middle 1.519592 Marital Status=Single 1.479482 Occupation =Other Occupations	

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.11(Continued) Clusters characterising Rural and Urban Females

Female Rural	Female Urban
Cluster 9 Freq. = 695 - No_HIV Freq. = 89 - HIV Age Group=A0_14 Marital Status=Single Education Level=Primary Occupation =Agricultural, Forestry and Fishery Workers	**Cluster 8** Freq. = 1072 - No_HIV Freq. = 686 - HIV Age Group=A20_24 Education Level=Secondary Occupation=Other Occupations 2.315341 Marital Status=Cohabiting 2.009030 Marital Status=Single 1.578683 Occupation=Crafts and Related Trade Workers 1.518718 Occupation=Other Occupations
Cluster 10 Freq. = 586 - HIV Freq. = 294 - No_HIV Age Group=A45_49 1.904848 Marital Status=Divorced 1.472433 Marital Status=Widow(er)	**Cluster 9** Freq. = 483 - HIV Freq. = 480 - No_HIV Marital Status=Single Occupation=Clerical Support Workers 1.890318 Education Level=Post Secondary/Tertiary 1.456958 Age Group=A25_29 1.417992 Occupation=Other Occupations
Cluster 11 Freq. = 535 - No_HIV Freq. = 188 - HIV Age Group=A15_19 2.487113 Marital Status=Single	**Cluster 10** Freq. = 393 - No_HIV Freq. = 137 - HIV Age Group=A15_19 1.478640 Marital Status=Single

Cluster 12
Freq. = 424 - HIV
Freq. = 263 - No_HIV
Age Group=A50_54
Marital Status=Divorced
1.632489 Marital Status=Widow(er)
	Cluster 11
	Freq. = 721 - No_HIV
	Freq. = 187 - HIV
	Occupation=Professionals
	Education Level=Post Secondary/Tertiary

Several clusters in the female rural setting (Table 4.11, column 1) has purity greater than 2. These include cluster 0, cluster 2, cluster 4, cluster 5, Cluster 6, Cluster 7 and cluster 10. On the contrary, none of the clusters in the females in the urban setting (Table 4.11 column 2) has purity greater or equal to 2. This is an indication that the HIV infection situation is worse in the rural setting than in the urban setting. Cluster by cluster comparison of the two settings with regards to the different age groups is given next.

For age groups, 15-19 of the rural female setting, the frequency of HIV elements in the cluster (cluster 11) are almost twice as much as the frequency of the no-HIV elements in the same cluster. In the case of Female HIV in the urban setting, the opposite is observed; showing the frequency of no-HIV elements appearing twice as much as the HIV elements.

Similar to (i) above, there are more HIV elements appearing in the cluster characterising females of age group 20-24 (cluster 8) in the rural setting than the no-HIV elements. The opposite is recorded for the same age group in cluster 8 for the females in the urban setting.

Age groups 25-29 (in cluster 4), 30-34 (cluster 0), 35-39 (cluster 5), 40-44 (cluster 7) and 45-49 (cluster 10) in the females rural setting have higher HIV elements than no-HIV elements. Similar trends are also observed for females in urban setting (cluster 3, cluster 4, cluster 5, cluster 0 and cluster 7).

For the age group 55-59, to be a widow seems to be a serious driver for infection in a rural area (cluster 3), while having much less influence in urban areas (cluster 2).

Being separated in the rural area (cluster 2) multiply the risk of infection critically. This situation seems to be not met in the urban area.

Table 4.12 shows that, while several drivers characterise HIV in rural setting with purity greater than 2, no single element with purity greater than or equal to 2 exists in the urban setting.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.12 Important HIV Drivers prevailing on Urban and Rural Females

Age Groups	Rural	urban	Drivers in both settings
25-29	**Education Level**=Middle **Occupation** =Crafts and Related Trade Workers **Occupation** =Other Occupations **Marital Status**=Married **Education Level**=Primary **Education Level**=Nil	None	None
30-34	**Marital Status**=Married **Occupation** =Service and Sales Workers **Education Level**=Nil **Education Level**=Middle **Occupation** =Crafts and Related Trade Workers **Education Level**=Primary **Occupation** =Agricultural, Forestry and Fishery Workers	None	None
35-39	**Occupation** =Service and Sales Workers **Marital Status**=Married **Education Level**=Nil **Education Level**=Middle	None	None
40-44	**Marital Status**=Divorced **Education Level**=Nil **Marital Status**=Widow(er)	None	None
45-49	**Age Group**=A45_49 1.904848 Marital Status=Divorced 1.472433 Marital Status=Widow(er)	None	None

Table 4.13 Clusters Characterising both Rural Males and Rural Females

Male	Female

Cluster 0
Freq. = 412 - HIV
Freq. = 306 - No_HIV
Age Group=A30_34
Occupation =Other Occupations
Occupation =Technicians and Associate Professionals
	Cluster 0
	Freq. = 1508 - HIV
	Freq. = 439 - No_HIV
	Age Group=A30_34
	Marital Status=Married
	Occupation =Service and Sales Workers
	Education Level=Nil

Cluster 1
Freq. = 475 - No_HIV
Freq. = 64 - HIV
Age Group=A20_24
Age Group=A15_19
Marital Status=Single
Education Level=Secondary
	Cluster 1
	Freq. = 201 - HIV
	Freq. = 131 - No_HIV
	Marital Status=Cohabiting
	Occupation =Technicians and Associate Professionals
	Education Level=Secondary
	Occupation =Elementary Occupations
	Occupation =Plant and Machine Operators, and Assemblers
	Occupation =Clerical Support Workers
	Occupation =Managers

Cluster 2
Freq. = 131 - No_HIV
Freq. = 20 - HIV
Age Group=A0_14
Education Level=Primary
Occupation =Elementary Occupations
2.011316 Marital Status=Single
	Cluster 2
	Freq. = 351 - HIV
	Freq. = 97 - No_HIV
	Marital Status=Separated

Cluster 3
Freq. = 33 - HIV
Freq. = 21 - No_HIV
Occupation =Clerical Support Workers
Marital Status=Widow(er)'
Occupation =Managers
Marital Status=Separated
	Cluster 3
	Freq. = 471 - HIV
	Freq. = 345 - No_HIV
	Age Group=A55_59
	Age Group=A60_64
	Age Group=old_Age
	Marital Status=Widow(er)

Cluster 4
Freq. = 297 - HIV
Freq. = 248 - No_HIV
Age Group=A55_59
Age Group=old_Age
Age Group=A60_64
	Cluster 4
	Freq. = 1443 - HIV
	Freq. = 678 - No_HIV
	Age Group=A25_29
	Education Level=Middle
	Occupation =Crafts and Related Trade Workers
	Occupation =Other Occupations

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.13 (Continued) Clusters Characterising both Rural Males and Females

Male	Female
Cluster 5 Freq. = 346 - No_HIV Freq. = 195 - HIV Age Group=A25_29 Occupation =Plant and Machine Operators, and Assemblers Occupation =Crafts and Related Trade Workers Marital Status=Cohabiting 1.740070 Marital Status=Single **Cluster 6** Freq. = 570 - HIV Freq. = 238 - No_HIV Age Group=A35_39 Education Level=Nil Occupation =Agricultural, Forestry and Fishery Workers Education Level=Post Secondary/Tertiary 1.746948 Occupation =Plant and Machine Operators, and Assemblers 1.717551 Marital Status=Married **Cluster 7** Freq. = 400 - HIV Freq. = 197 - No_HIV Age Group=A45_49 **Cluster 8** Freq. = 521 - HIV Freq. = 246 - No_HIV Age Group=A40_44 Marital Status=Married Education Level=Middle Occupation =Service and Sales Workers 1.733479 Education Level=Nil	**Cluster 5** Freq. = 1283 - HIV Freq. = 457 - No_HIV Age Group=A35_39 **Cluster 6** Freq. = 562 - HIV Freq. = 199 - No_HIV Education Level=Post Secondary/Tertiary Occupation =Professionals **Cluster 7** Freq. = 837 - HIV Freq. = 333 - No_HIV Age Group=A40_44 **Cluster 8** Freq. = 807 - HIV Freq. = 736 - No_HIV Age Group=A20_24 2.333102 Occupation =Crafts and Related Trade Workers **Cluster 9** Freq. = 695 - No_HIV Freq. = 89 - HIV Age Group=A0_14 Marital Status=Single Education Level=Primary Occupation =Agricultural, Forestry and Fishery Workers **Cluster 10** Freq. = 586 - HIV Freq. = 294 - No_HIV Age Group=A45_49 **Cluster 11** Freq. = 535 - No_HIV Freq. = 188 - HIV Age Group=A15_19 2.487113 Marital Status=Single

Table 4.13 (Continued) Clusters Characterising both Rural Males and Females

Male Rural	**Female Rural**

Cluster 9
Freq. = 305 - HIV
Freq. = 147 - No_HIV
Age Group=A50_54
Marital Status=Divorced
Occupation =Professionals
	Cluster 12
	Freq. = 424 - HIV
	Freq. = 263 - No_HIV
	Age Group=A50_54
	Marital Status=Divorced

Observations from Table 4.13

Age groups 0 to 24 (clusters 1 and clusters 2; purities <=1) of the rural male are at least risk of infection with HIV. For rural female, those at least risk of infection is those in age group 0 to 19 (cluster 9 and cluster 11; purities <=1).

Male age groups most prone to HIV infection in the rural setting are age groups 35 to 39 (cluster 6), 40 to 44 (cluster 8), 45 to 49 and 50 to 54 (cluster 9). On the other hand, female age groups most prone to HIV infection in the rural setting are 25 to 29 (cluster 4), 30 to 34 (cluster 0), 35 to 39 (cluster 5), 40 to 44 (cluster 7) and 45 to 49 (cluster 10).

Rural females with “separated” marital status (cluster 2) are at very high risk of infection.

There is a strong similarity between cluster 9 (of rural males) and cluster 12 (of rural females. This similarity indicates that divorce is a strong HIV driver for males of age group 50 to 54 and females of the same group in the rural setting.

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.14 Important HIV Drivers prevailing in both Rural Males and Females

Age Groups	Male	Female	Both males and Females

25-29	Occupation=Plant and Machine Operators, and Assemblers,
	Status=Cohabiting, Single
		Education Level=Middle
		Occupation =Other Occupations
		Education Level=Primary
		Education Level=Nil
			Marital Status=Marital
			Occupation =Crafts and Related Trade Workers

30-34	None	Marital Status=Married Occupation =Service and Sales Workers Education Level=Nil Education Level=Middle Occupation =Crafts and Related Trade Workers Education Level=Primary Occupation =Agricultural, Forestry and Fishery Workers	None

35-39	Education Level=Post Secondary/Tertiary;
	Occupation=Plant and Machine Operators, and Assemblers;
	Occupation =Agricultural, Forestry and Fishery Workers;
	Occupation=Crafts and Related Trade Workers;
		Occupation =Service and Sales Workers
		Education Level=Middle
			Education Level=Nil;
			Marital Status=Married

40-44	Occupation =Service and Sales Workers	Marital Status=Divorced Marital Status=Widow(er)	Education Level=Nil
45-49	Marital Status=Married	Marital Status=Divorced Marital Status=Widow(er)	
50-54	Occupation= Professional		Marital Status=Divorced;

KNOWLEDGE, TRUTH AND EXCELLENCE

Table 4.15 Clusters Characterising Urban Males and Urban Females

Male	Female
Cluster 0 Freq. = 906 - No_HIV Freq. = 201 - HIV Age Group=A20_24 Marital Status=Single Age Group=A15_19 Age Group=A0_14 Education Level=Primary 1.727737 Education Level=Secondary **Cluster 1** Freq. = 19 - HIV Freq. = 6 - No_HIV Marital Status=Separated **Cluster 2** Freq. = 296 - No_HIV Freq. = 274 - HIV Age Group=A60_64 Age Group=old_Age Occupation=Other Occupations Marital Status=Widow(er) **Cluster 3** Freq. = 1039 - HIV Freq. = 651 - No_HIV Age Group=A35_39 Education Level=Post Secondary/Tertiary Occupation=Plant and Machine Operators, and Assemblers Occupation=Service and Sales Workers Education Level=Nil	**Cluster 0** Freq. = 897 - HIV Freq. = 734 - No_HIV Age Group=A40_44 **Cluster 1** Freq. = 35 - No_HIV Freq. = 33 - HIV Occupation=Managers Occupation=Plant and Machine Operators, and Assemblers Marital Status=Separated Occupation=Technicians and Associate Professionals **Cluster 2** Freq. = 863 - No_HIV Freq. = 466 - HIV Age Group=A55_59 Age Group=old_Age Age Group=A60_64 Age Group=A0_14 Marital Status=Widow(er) Occupation=Agricultural, Forestry and Fishery Workers **Clusters 3** Freq. = 1095 - HIV Freq. = 886 - No_HIV Age Group=A25_29 Marital Status=Married Marital Status=Cohabiting Occupation=Crafts and Related Trade Workers **Cluster 4** Freq. = 1803 - HIV Freq. = 1064 - No_HIV Age Group=A30_34 Occupation=Service and Sales Workers Education Level=Middle Education Level=Nil 2.169248 Marital Status=Married

Table 4.15 (Continued) Clusters Characterising both Urban Males and Females

Male	Female
Cluster 4 Freq. = 816 - No_HIV Freq. = 312 - HIV Age Group=A25_29 Education Level=Secondary Marital Status=Cohabiting Occupation=Elementary Occupations Occupation=Clerical Support Workers 2.240200 Marital Status=Single **Cluster 5** Freq. = 764 - HIV Freq. = 405 - No_HIV Age Group=A45_49 **Cluster 6** Freq. = 988 - HIV Freq. = 528 - No_HIV Age Group=A40_44 Marital Status=Married Education Level=Middle Occupation=Agricultural, Forestry and Fishery Workers Occupation=Technicians and Associate Professionals 1.720815 Occupation=Plant and Machine Operators, and Assemblers **Cluster 7** Freq. = 796 - No_HIV Freq. = 725 - HIV Age Group=A30_34 Occupation=Crafts and Related Trade Workers Occupation=Professionals **Cluster 8** Freq. = 511 - HIV Freq. = 346 - No_HIV Age Group=A50_54 Marital Status=Divorced	**Cluster 5** Freq. = 1455 - HIV Freq. = 898 - No_HIV Age Group=A35_39 Education Level=Primary Occupation=Elementary Occupations **Cluster 6** Freq. = 512 - No_HIV Freq. = 433 - HIV Age Group=A50_54 Marital Status=Divorced 2.416490 Marital Status=Widow(er) **Cluster 7** Freq. = 625 - HIV Freq. = 557 - No_HIV Age Group=A45_49 2.120777 Marital Status=Divorced **Cluster 8** Freq. = 1072 - No_HIV Freq. = 686 - HIV Age Group=A20_24 Education Level=Secondary Occupation=Other Occupations 2.315341 Marital Status=Cohabiting **Cluster 9** Freq. = 483 - HIV Freq. = 480 - No_HIV Marital Status=Single Occupation=Clerical Support Workers **Cluster 10** Freq. = 393 - No_HIV Freq. = 137 - HIV Age Group=A15_19

Table 4.15 (Continued) Clusters Characterising both Urban Male and Female

Male	Female

Cluster 8
 	
Freq. = 511 - HIV
Freq. = 346 - No_HIV
 	
Age Group=A50_54
Marital Status=Divorced
 	
Cluster 9
 	
Freq. = 259 - HIV
Freq. = 250 - No_HIV
Age Group=A55_59
Occupation=Managers
	Cluster 11
	
	Freq. = 721 - No_HIV
	Freq. = 187 - HIV
	Occupation=Professionals
	Education Level=Post Secondary/Tertiary

Observations from Table 4.15

Male Age groups with the least risk of HIV infection in the urban settings are 20 to 24 (cluster 0) and 25 to 29 (cluster 4). On the other hand, female age groups with the least risk of HIV infection in the urban setting is 15 to 19 (cluster 10).

For both urban males and females, no age group is associated with a cluster with purity up to or greater than 2. In fact, a few of the age groups are associated with clusters with purities less than 0.5. Therefore, it is possible to suggest that, HIV situation of males in urban settings is evenly spread among age groups 30 and above with associated clusters of purities ranging from 0.53 to 1.03. Those younger than 30 however, are associated with clusters with purities far less than 0.5 (cluster 0) and (cluster 4). For urban females, all age groups are associated with clusters with purities in the range 0.53 to 1.68 except age group 15-19, which is associated with a cluster 10 of purity 0.35. It is also an indication that the HIV situation in urban settings is not as high as in rural settings.

Cluster 8 (of the urban males) and clusters 6 and 7 of the urban females are very similar in content. This similarity indicates that divorce is a stronger HIV driver for both males and females of age groups 45 to 54 in the urban setting.

The overall trend of HIV infection is summarised in Table 4.17, according to age.

Table 4.16 Important Factors prevailing on Urban Males and Females

Age Groups	Females	Males	Both Males and Females

25-29	None	Education Level=Secondary
		Occupation=Clerical Support Workers
		Occupation=Elementary Occupations
			None

35-39	None	Occupation=Service and Sales Workers	None
40-44	None	Occupation=Agricultural, Forestry and Fishery Workers; Occupation=Technicians and Associate Professionals; Occupation=Plant and Machine Operators, and Assemblers	None

Table 4.17 Summary of Overall HIV Infection Trends according to Age

		0-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	Old age
Rural	Male	x	x	x	x	*					*	*	*
	Female	x	x	*							*	*	*
Urban	Male	x	x	x	x	x	*	*	*	*	*	x	x
	Female	x	x	x	*	*	*	*	*	x	x	x	x

Key:

– Purity greater than or equal to 2; x – Purity less than 1; * – Purity greater than or equal to 1 but less than 2;

KNOWLEDGE, TRUTH AND EXCELLENCE

Observations from Table 4.17

Clusters with purity greater than 2 pertain only to the rural population. None of such clusters pertains to the urban population. This is an indication that the HIV infection trends in the rural areas are generally higher than the urban.

Males of age groups 30 to 54 are most at risk of HIV infection in the rural areas.

Females of age groups 25 to 54 are most at risk of HIV infection in the rural areas are.

Age group 0 to 29 has the least risk of HIV infection for males in the rural areas.

Age group 0 to 19 have the least risk of HIV infection for females in the rural areas.

Age group 0 to 34 and those older than sixty years have the least risk of HIV infection for males in the urban areas.

Age group 0 to 24 and those older than fifty years have the least risk of HIV infection for females in the urban areas.

Across both rural and urban areas, females are at most risk.

These observations conform well with those observed in sub-sections 4.4.3 and discussed in sub-section 4.4.4.

KNOWLEDGE, TRUTH AND EXCELLENCE

CONCLUSIONS AND RECOMMENDATION

Preamble

This study noted that existing research works in sub-Saharan Africa (SSA) have identified several HIV socio-economic drivers but the extent to which these drivers impact on a population within a specific context is not known; hence the need to measure their degrees of impact. To address this proposition, three research questions were posed:

What are the main socio-economic drivers of HIV in Ghana?

How much impact does each socio-economic HIV-driving factor have on various population subgroups?

What are the relationships among socio-economic drivers of HIV in specific population subgroups?

This chapter concludes the study by reviewing its key aspects by summarising the study method and approaches used to address the research questions. It also discusses the implication of this research. Finally, it highlights the limitations of the study and makes recommendations for future studies.

Conclusions

This research set out to address three research questions regarding the degrees of impact and interactivities of socio-economic drivers of HIV in Ghana. The first objective was to use the Frame-based Knowledge Representation technique to identify key socio-economic HIV drivers in SSA. To achieve this, a rigorous systematic literature review was conducted on a wide range of published articles to obtain and document findings regarding known drivers of the pandemic in SSA. The systematic literature review analysis approach was chosen because it follows systematic and explicit procedures in conducting a review, which makes it easier for other researchers to reproduce following the same approach on the same topic. The articles reviewed span two decades; ranging from 2001 to 2017. This was relevant to ensure that longstanding socio-economic determinants of HIV are captured and to obtain rich knowledge about the disease across various parts of sub-Saharan Africa.

KNOWLEDGE, TRUTH AND EXCELLENCE

In order to achieve a more compact and more precise representation of knowledge obtained from the literature, the frame-based knowledge representation technique was chosen over possible alternatives such as First Order Logic, Semantic Graphs and Conceptual Graphs. The representation illustrates an easy-to-visualise relationship between socio-economic factors and HIV.

Addressing the first objective made it possible to answer the research question: *What are the main socio-economic drivers of HIV in SSA*? The drivers identified include low education level, divorce, widowhood, elementary occupation, the female gender, and being a rural resident.

Further objectives of this research included fashioning out and implementing a Computational Model to determine the degrees of impact and interactivities of the identified drivers on HIV infection. This was achieved through Feature Maximization along with Growing Neural Gas (GnG) as a clustering technique. With this combination, each driver was assessed, and only the most relevant was retained and weighted. This made it possible to identify important clusters of drivers, thereby showing interactivities of such factors in a weighted manner. Widely used methods for HIV studies in the literature such as Linear Regression, Logistic Regression, Chi-Square and many others are not able to derive this kind of detailed insight into individual features. Besides, the complexity of the system is at most quadratic. This makes the implemented system suitable for analysing large datasets. This made it possible to answer the research questions: "*How much impact does each socio-economic HIV driver have on various population subgroups*"? and "*What are the relationships among socio-economic drivers of HIV in specific population subgroups*"? Feature Maximization leads to the derivation of what is termed in this work as the contrast value for each feature. The contrast value represents the weighted strength, which alternatively is the degree of impact of each feature. The population subgroups of concern for this research were Rural Males, Rural Females, Urban Males and Urban Females. The different trends of degrees of impact of each feature were derived and discussed in Sections 4.3.2 to 4.4.4.

This research theoretically, extends efforts of earlier studies and goes beyond the identification of socio-economic factors driving the continual spread of the HIV pandemic in SSA to assessing their impacts on the populace. It, therefore, paves the way for a new way of looking at the HIV driving factors with a certain level of detail and precision. The implication of this research is the increased probability of explaining the influence of instances of each HIV-driving factor in relative terms for various population sub-groups.

Some existing researches have shown the interactivity or relation between specific HIV -drivers with HIV infection but not relationships among the factors themselves. One of the key focuses of this research was to make it possible for relationships among the driving factors in the context of HIV to be highlighted; thereby opening a new research direction. In practice, the findings of this study have implications for policymakers and stakeholders to enable them to make target-specific prevention strategies towards the elimination of the pandemic.

Limitations and Recommendation of Research

Even though the modelling approach used in this work resolves the stated problem efficiently, there are still certain observed limitations. Firstly, the approach would produce very good results only for large datasets. It is therefore not suitable where data is in limited quantity. Secondly, the interactivity between factors is undirected; meaning it is not clear which factor drives the other. It is therefore recommended that; future research works devise ways to establish the direction of causality among factors.

KNOWLEDGE, TRUTH AND EXCELLENCE

REFERENCES

Abebe, Y., Schaap, A., Mamo, G., Negussie, A., Darimo, B. and Wolday, D. (2003), "HIV prevalence in 72 000 urban and rural male army recruits, Ethiopia", *Aids,* Vol. 17, No. 12, pp. 1835-1840.

Abrefa-Gyan, T., Cornelius, L. J. and Okundaye, J. (2016), "Socio-Demographic Factors, Social Support, Quality of Life, and HIV in Ghana", *Journal of Evidence-Informed Social Work,* Vol. 13, pp. 206-216.

Agüero, J. M. and Bharadwaj, P. (2014), "Do the more educated know more about health? Evidence from schooling and HIV knowledge in Zimbabwe", *Economic Development and Cultural Change,* Vol. 62, No. 3, pp. 489-517.

Agwu, E., Pazos, V., Ihongbe, J.C. and Ssengendo, J. (2011), "Appraisal of the inherent socio-demographic dynamics of HIV epidemic in four districts of South-Western Uganda", *Journal of Social Aspects of HIV /AIDS,* Vol. 8, No. 3, pp. 150-155.

Akwara, P. A., Fosu, G. B., Govindasamy, P., Alayón, S. and Hyslop, A. (2005), "An In-Depth Analysis of HIV Prevalence in Ghana: Further Analysis of Demographic and Health Surveys Data". Calverton, MD: ORC Macro.

Amornkul, P.N., Vandenhoudt, H., Nasokho, P., Odhiambo, F., Mwaengo, D., Hightower, A., Buvé, A., Misore, A., Vulule, J., Vitek, C. and Glynn, J. (2009), "HIV prevalence and associated risk factors among individuals aged 13-34 years in Rural Western Kenya", *PloS one,* Vol. 4, No. 7, pp. e6470.

Anon. (2004a), "Report on the Global AIDS Epidemic", *http://files.unaids.org/ en/media/unaids/contentassets/documents/unaidspublication/2004/GAR2004_en.pdf.* Accessed: January 22, 2019.

Anon. (2004b), "HIV and AIDS: The Science Inside*", http://ehrweb01.aaas.org/ science-inside /files/2012/03/AIDSbook.pdf.* Accessed: November 24, 2018.

Anon. (2004b), "Report on the Global AIDS Epidemic", *http://files.unaids.org/ en/media/unaids/contentassets/documents/unaidspublication/2004/GAR2004_en.pdf.* Accessed: January 22, 2019.

Anon. (2004c), "HIV and AIDS: The Science Inside*", http://ehrweb01.aaas.org/ science-inside /files/2012/03/AIDSbook.pdf.* Accessed: November 24, 2018.

Anon. (2004c). “HIV and work: global estimates, impact and response”, *https://www.ilo.org/aids/Publications/WCMS_116379/lang--en/index.htm*. Accessed: November 19, 2018.

Anon. (2004d). “HIV and work: global estimates, impact and response”, *https://www.ilo.org/aids/Publications/WCMS_116379/lang--en/index.htm*. Accessed: November 19, 2018

Anon. (2006), “2006 Report on the global AIDS epidemic”, Joint UN Programme on HIV /AIDS. *http://data.unaids.org/pub/report/2006/*, Accessed: October 30, 2018.

Anon. (2011), “UNAIDS Terminology Guidelines”, *www.unaids.org*. Accessed: January 2, 2018.

Anon. (2014a), “The Gap Report”, *http://www.unaids.org/sites/default/ files/ media_asset/*. Accessed: August 24, 2018.

Anon. (2016), “Global AIDS Update 2016”, *www.unaids.org*. Accessed: August 24, 2018

Anon. (2017a), “Data 2017 Programme on HIV /AIDS”, *www.unaids.org*. Accessed: August 24, 2018.

Anon. (2017b), “National and Sub-National HIV and AIDS Estimates and Projections 2017 Report”, *www.ghanaids.gov.gh*. Accessed: January 22, 2018.

Anon. (2018), “UNAIDS Data 2018”, *http://www.unaids.org/ sites /default/files/media_ asset/unaids-data-2018_en.pdf*. Accessed: November 1, 2018.

Arrehag, L., Durevall, D. and Sjöblom, M. (2006), “The impact of HIV on the economy, livelihoods and poverty of Malawi”, *https://www.sida.se/20061-the-impact-of-HIV aids-on-the-economy-livelihoods-and-poverty-of-malawi_1371.pdf*. Accessed: October 30, 2018.

Asiedu, C., Asiedu, E. and Owusu, F. (2012), “The Socio-Economic Determinants of HIV Infection Rates in Lesotho, Malawi, Swaziland and Zimbabwe”, *Development Policy Review*, Vol. 30, No. 3, pp. 305-326.

Askew, I. and Berer, M. (2003), “The contribution of sexual and reproductive health services to the fight against HIV /AIDS: A review”, *Reproductive Health Matters*, Vol. 11, No. 22, pp. 51-73.

Aulagnier, M., Janssens, W., De Beer, I., van Rooy, G., Gaeb, E., Hesp, C., van der Gaag, J. and de Wit, T.F.R. (2011), “Incidence of HIV in Windhoek, Namibia: Demographic and socio-economic associations” *PloS one*, Vol. 6, No. 10, pp. e25860.

Auvert, B., Buvé, A., Ferry, B., Caraël, M., Morison, L., Lagarde, E., Robinson, N.J., Kahindo, M., Chege, J., Rutenberg, N. and Musonda, R. (2001), "Ecological and individual level analysis of risk factors for HIV infection in four urban populations in sub-Saharan Africa with different levels of HIV infection", *Aids*, Vol. 15, pp. 15-30.

Babalola, S. (2011), "Factors associated with HIV infection among sexually experienced adolescents in Africa: a pooled data analysis", *African Journal of AIDS Research*, Vol. 10, No. 4, pp. 403-414.

Baidoo, I., Boatin, R.R., Adom, T., Datohe, D., Voure, T., Bansa, D., Brown, C. and Diaba, A. (2012), "Socio-Demographic Characteristics of Patients Diagnosed with HIV in Accra and Kumasi Metropolis", *African Journal of Clinical and Experimental Microbiology*, Vol. 13, No. 3, pp.161-169.

Bain, L.E., Nkoke, C. and Noubiap, J.J.N. (2017), "Can the UNAIDS 90–90–90 target be achieved? A systematic analysis of national HIV treatment cascades". *BMJ Global Health*, Vol. 2, No. 2, p. e000227.

Barankanira, E., Molinari, N., Niyongabo, T. and Laurent, C. (2016), "Spatial analysis of HIV infection and associated individual characteristics in Burundi: indications for effective prevention", *BMC public health*, Vol. 16, No. 1, pp.118-127.

Bärnighausen T, Hosegood V, Timaeus I, Newell M. (2007), "The Socio-economic determinants of HIV incidence: evidence from a longitudinal, population-based study in rural South Africa", *AIDS*, Vol. 21, No. 7, pp. 29–38.

Bell, D. (2004), "Uml basics: The component diagram", *www.ibm.com*. Accessed: February 8, 2019.

Berkhin, P., (2006), "A survey of clustering data mining techniques", *In Grouping multidimensional data*. Springer, Berlin, Heidelberg, pp. 25-71.

Bertozzi, S., Padian, N.S., Wegbreit, J., DeMaria, L.M., Feldman, B., Gayle, H., Gold, J., Grant, R. and Isbell, M.T. (2006), "HIV prevention and treatment, Disease control priorities in developing countries", Oxford University Press, Oxford, UK, pp. 331-370.

Bertrand, N.A.S. (2016), "The Socio-economic Determinants of the Prevalence HIV among Women in Cameroon", *African Journal of Economic Review*, Vol. 4, No. 2, pp. 188-202.

Blattner, W., Gallo, R.C. and Temin, H.M. (1988), "HIV causes aids", *Science*, Vol. 241, No. 4865, pp. 515-516.

Bock, H.H., (1996), “Probabilistic models in cluster analysis”, *Computational Statistics & Data Analysis*, Vol. 23, No. 1, pp. 5-28.

Boerma, J.T., Gregson, S., Nyamukapa, C. and Urassa, M. (2003), “Understanding the uneven spread of HIV within Africa: comparative study of biologic, behavioral, and contextual factors in rural populations in Tanzania and Zimbabwe”, *Sexually Transmitted Diseases*, Vol. 30, No. 10, pp. 779-787.

Bogale, G. W., Boer, H. and Seydel, E. R. (2009), “HIV -prevention knowledge among illiterate and low-literate women in rural Amhara, Ethiopia”, *African Journal of AIDS Research*, Vol. 8, No. 3, pp. 349–357.

Bollinger, L. and Stover, J. (1999), “The economic impact of AIDS in South Africa”, www.planet.uwc.ac.za. Accessed: February 18, 2018.

Bongaarts, J., 2007. Late marriage and the HIV epidemic in sub-Saharan Africa. Population studies, Vol. 61, No. 1, pp.73-83.

Booysen, F., Geldenhuys, J.P. and Marinkov, M., (2003), “The impact of HIV/AIDS on the South African economy: a review of current evidence”. *Development Policy Research Unit*, School of Economics, University of Cape Town.

Bowa, K., Kawimbe, B., Mugala, D., Musowoya, D., Makupe, A., Njobvu, M., & Simutowe, C. (2016), “A review of HIV and surgery in Africa”, *The open AIDS journal*, Vol.10, pp. 16-23.

Braga, L.H.P., Farrokhyar, F. and Bhandari, M. (2012), “Practical Tips for Surgical Research: Confounding: What is it and how do we deal with it?”, *Canadian Journal of Surgery*, Vol. 55, No. 2, 132 pp.

Brodish, P.H. (2015), “An association between neighbourhood wealth inequality and HIV prevalence in sub-Saharan Africa”, *Journal of biosocial science*, Vol. 47, No.3, pp. 311-328.

Buor, D. (2005), “Social class and HIV prevalence in Sub-Saharan Africa”, *Journal of Science and Technology*, Vol. 25, No. 2, pp. 66-79.

Buvé, A., Caraël, M., Hayes, R.J., Auvert, B., Ferry, B., Robinson, N.J., Anagonou, S., Kanhonou, L., Laourou, M., Abega, S. and Akam, E. (2001), “Multicentre study on factors determining differences in rate of spread of HIV in sub-Saharan Africa: methods and prevalence of HIV infection”, *Aids*, Vol. 15, pp. 5-14.

Chijioke, I. and Akani, Y. (2014), “Socio-Demographic profile of People Living with HIV (PLWAs) in Port Harcourt, Nigeria”, *The Nigerian Health Journal*, Vol. 14, No. 3, pp. 119-128.

Clark, C., Bruce, J. and Dude A. (2006), "Protecting Young Women from HIV /AIDS: The Case against Child and Adolescent Marriage", *International Family Planning Perspectives*, Vol. 32, No. 2, pp. 79-88.

Clark, P. (1996), "Requirements for a Knowledge Representation System", *www.cs. utexas. edu*. Accessed: November 23, 2018.

Clark, S. (2004), "Early marriage and HIV risks in sub-Saharan Africa", Studies in family planning", Vol. 35, No. 3, pp. 149-160.

Coburn, B.J., Okano, J.T. and Blower, S. (2013), "Current drivers and geographic patterns of HIV in Lesotho: implications for treatment and prevention in Sub-Saharan Africa", *BMC medicine*, Vol. 11, No. 1, pp. 224.

Cornia, G. and Zagonari, F. (2007), "The HIV and AIDS impact on the rural and urban economy", *In: (Cornia, G., Ed) AIDS, Public Policy and Child Well-being (2nd Ed)*, UNICEF Innocenti Research Centre, Florence pp. 185–205.

Daly, K., (2000) "The business response to HIV/AIDS: impact and lessons learned", *www.popline.org/node/263557*. Accessed: March 3, 2018

Davies, D.L. and Bouldin, D.W. (1979), "A cluster separation measure", *IEEE transactions on pattern analysis and machine intelligence*, No. 2, pp.224-227.

De Walque, D. (2009), "Does education affect HIV status? Evidence from five African countries", *The World Bank Economic Review*, Vol. 23, No. 2, pp. 209-233.

Diderichsen, F., Evans, T., and Whitehead M. (2001), "The social basis of disparities in health", *Challenging inequities in health*, pp. 1-27.

Dunn, J.C. (1974), "Well-separated clusters and optimal fuzzy partitions", *Journal of cybernetics*, Vol. 4, No. 1, pp.95-104.

Edelstein, Z. R., Santelli, J. S., Helleringer, S., Schuyler, A. C., Wei, Y., Mathur, S. and Wawer, M. J. (2015), "Factors associated with incident HIV infection versus prevalent infection among youth in Rakai, Uganda", *Journal of Epidemiology and Global Health*, Vol.5, No. 1, pp. 85-91.

Fortson, J. G. (2008), "The gradient in sub-Saharan Africa: Socio-economic status and HIV /AIDS", *Demography*, Vol. 45, No. 2, pp. 303-322.

Fox, A.M. (2010), "The social determinants of HIV serostatus in sub-Saharan Africa: an inverse relationship between poverty and HIV ?", *Public Health Reports*, Vol. 125, (supplement 4), pp. 16-24.

Franke, T.M., Ho, T. and Christie, C.A. (2012), "The chi-square test: Often used and more often misinterpreted", *American Journal of Evaluation,* Vol. 33, No. 3, pp. 448-458.

Fritzke, B. (1995), "A growing neural gas network learns topologies", *Advances in neural information processing systems,* pp. 625-632.

Gabrysch, S., Edwards, T. and Glynn, J.R. (2008), "The role of context: neighbourhood characteristics strongly influence HIV risk in young women in Ndola, Zambia", *Tropical Medicine & International Health,* Vol. 13, No. 2, pp. 162-170.

Gaigbe-Togbe, V. and Weinberger, M.B. (2004), "The social and economic implications of HIV /AIDS", *African population studies,* Vol. 19, (supplement B), pp. 31-59.

Gillespie, S., Greener, R., Whiteside, A. and Whitworth, J. (2007), "Investigating the empirical evidence for understanding vulnerability and the associations between poverty, HIV infection and AIDS impact", AIDS 21(supplement 7), pp. S1-S4.

Glynn, J.R., Caraël, M., Auvert, B., Kahindo, M., Chege, J., Musonda, R., Kaona, F., Buvé, A. and Study Group on the Heterogeneity of HIV Epidemics in African Cities. (2001), "Why do young women have a much higher prevalence of HIV than young men? A study in Kisumu, Kenya and Ndola, Zambia.", *Aids,* Vol. 15, pp. S51-S60.

Glynn, J.R., Carael, M., Buve, A., Anagonou, S., Zekeng, L., Kahindo, M. and Musonda, R. (2004), "Does increased general schooling protect against HIV infection? A study in four African cities", *Tropical medicine & international health*, Vol. 9, No. 1, pp.4-14.

Gómez-Olivé, F.X., Angotti, N., Houle, B., Klipstein-Grobusch, K., Kabudula, C., Menken, J., Williams, J., Tollman, S. and Clark, S.J. (2013), "Prevalence of HIV among those 15 and older in rural South Africa", *AIDS care,* Vol. 25, No. 9, pp. 1122-1128.

Gonzalez, R., Munguambe, K., Aponte, J., Bavo, C., Nhalungo, D. and Macete, E. (2012), "High HIV prevalence in a southern semi-rural area of Mozambique: a community-based survey", *HIV Med,* Vol. 13, No. 10, pp. 581-588.

Gordon, A. D. (1998), "Response to comments*", Bulletin of the International Statistical Institute,* Vol. 51, No. 3, pp. 414-415.

Gumbe, A., McLellan-Lemal, E., Gust, D.A., Pals, S.L., Gray, K.M., Ndivo, R., Chen, R.T., Mills, L.A., Thomas, T.K. and KICoS Study Team. (2015), "Correlates of prevalent HIV infection among adults and adolescents in the Kisumu incidence cohort study", Kisumu, Kenya, *International journal of STD & AIDS,* Vol. 26, No. 13, pp. 929-940.

Gummerson, E. (2013), “Have the educated changed HIV risk behaviours more in Africa?”, *African Journal of AIDS Research*, Vol. 12, No. 3, pp. 161–172.

Hajizadeh, M., Sia, D., Heymann, S.J. and Nandi, A. (2014), “Socio-economic inequalities in HIV prevalence in sub-Saharan African countries: evidence from the Demographic Health Surveys”, *International journal for equity in health*, Vol. 13, No. 1, pp.18.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001), “On clustering validation techniques”, Journal of Intelligent Information Systems, Vol. 17, No. 2, pp. 147-155.

Hargreaves, J. R., Bonell, C. P., Morison, L. A., Kim, J. C., Phetla, G., Porter, J. D., Watts, C. and Pronyk, P.M. (2007), “Explaining continued high HIV prevalence in South Africa: Socio-economic factors, HIV incidence and sexual behaviour change among a rural cohort, 2001–2004”, AIDS 21(supplement 7), pp. 39-48.

Hargreaves, J.R. and Glynn, J.R. (2002), “Educational attainment and HIV -1 infection in developing countries: a systematic review”, Tropical Medicine & International Health, Vol. 7, No. 6, pp. 489-498.

Hargreaves, J.R., Bonell, C.P., Boler, T., Boccia, D., Birdthistle, I., Fletcher, A., Pronyk, P.M. and Glynn, J.R. (2008) “Systematic review exploring time trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa”, *Aids*, Vol. 22, No. 3, pp. 403-414.

Higgins, J.A., Hoffman, S. and Dworkin, S.L. (2010), “Rethinking gender, heterosexual men, and women's vulnerability to HIV /AIDS”, American journal of public health, Vol. 100, No. 3, pp. 435-445.

Holmström, J. (2002), “Growing Neural Gas--Experiments with GNG--GNG with Utility and Supervised GNG”, Unpublished Master's Thesis, Uppsala University, Uppsala, Sweden. 197 pp.

Hoshi, T., Fuji, Y., Nzou, S.M., Tanigawa, C., Kiche, I., Mwau, M., Mwangi, A.W., Karama, M., Hirayama, K., Goto, K. and Kaneko, S. (2016), “Spatial distributions of HIV infection in an endemic area of western Kenya: guiding information for localized HIV control and prevention.”, PloS one, Vol. 11, No. 2, pp. e0148636.

Huntbach, M. (1996), “Artificial Intelligence I”, Notes on semantic nets and frames”, http://www.eecs.qmul.ac. uk/~mmh/ AINotes /AINotes4.pdf., Accessed: November 3, 2018.

Hüsken, S. M., and Heck, S. (2012), “The ‘Fish Trader+’model: Reducing female fish traders’ vulnerability to HIV”, *African Journal of AIDS Research*, Vol. 11, No.1, pp. 17-26.

ILO. (2004), "International Labor Organization (ILO) Report of ILO, 2004)", *https://www.ilo.org/public/english/bureau/stat/download/cpi/prefcpi.pdf* , Accessed: January 21, 2018.

Irvine E. J. (2004), "Measurement and expression of risk: optimizing decision strategies". Am J Med, (117 supplement 5A), pp. 2S-7S

Isiugo-Abanihe, U. C. and Oyediran, K. A. (2004), "Household Socio-economic status and sexual behaviour among Nigerian female youth", African Population Studies, Vol. 19, No. 1, pp. 81-98.

Jager, K.J., Zoccali, C., Macleod, A. and Dekker, F.W. (2008), "Confounding: what it is and how to deal with it, *Kidney international*, Vol. 73, No. 3, pp. 256-260.

Jaggi, S., (2003), "Descriptive statistics and exploratory data analysis", *Indian Agricultural Statistics Research Institute*, Vol. 1, pp.1-18.

Joesoef, M.R., Cheluget, B., Marum, L.H., Wandera, C., Ryan, C.A., DeCock, K.M. and Chebet, K. (2003), "Differential of HIV prevalence in women and men who attended sexually transmitted disease clinics at HIV sentinel surveillance sites in Kenya, 1990–2001", International journal of STD & AIDS, Vol. 14, No. 3, pp. 193-196.

Johnson, L.F., Dorrington, R.E., Bradshaw, D., du Plessis, H. and Makubalo, L. (2009), "The effect of educational attainment and other factors on HIV risk in South African women: results from antenatal surveillance, 2000–2005", *Aids*, Vol. 23 No. 12, pp.1583-1588.

Kalichman, S. C., Simbayi, L. C., Kagee, A., Toefy, Y., Jooste, S., Cain, D. and Cherry, C. (2006), "Associations of poverty, substance use, and HIV transmission risk behaviors in three South African communities", *Social Science & Medicine,* Vol. 62, No. 7, pp.1641-1649.

Kalipeni, E. (2000), "Health and disease in southern Africa: a comparative and vulnerability perspective", *Social Science & Medicine*, Vol. 50, No. 7, pp.965-983.

Karim, Q. A., Kharsany, A. B., Frohlich, J.A., Werner, L., Mlotshwa, M., Madlala, B.T. and Karim, S.S.A. (2012), "HIV incidence in young girls in KwaZulu-Natal, South Africa-public health imperative for their inclusion in HIV biomedical intervention trials", AIDS and Behavior, Vol. 16, No. 7, pp. 1870-1876.

Kayeyi, N., Sandøy, I.F. and Fylkesnes, K. (2009), "Effects of neighbourhood-level educational attainment on HIV prevalence among young women in Zambia", *BMC Public Health*, Vol. 9, No. 1, pp.310-317.

Kebaneilwe, M.D., (2011), "The Vashti paradigm: resistance as a strategy for combating HIV" *The Ecumenical Review,* Vol. 63, No. 4, pp.378-384.

Kembo, J. (2012), "Risk factors associated with HIV infection among young persons aged 15–24 years: Evidence from an in-depth analysis of the 2005–06 Zimbabwe Demographic and Health Survey", *Journal of Social Aspects of HIV /AIDS,* Vol. 9, No. 2, pp. 54-63

Kimanga, D.O., Ogola, S. and Umuro, M. (2014), "Prevalence and incidence of HIV infection, trends, and risk factors among persons aged 15–64 years in Kenya: results from a nationally representative study", *Journal of acquired immune deficiency syndromes*, Vol. 66 No. 1, pp. 13-26.

Kimani, J.K., Ettarh, R., Ziraba, A.K. and Yatich, N. (2013), "Marital status and risk of HIV infection in slum settlements of Nairobi, Kenya: results from a cross-sectional survey", *African Journal of Reproductive Health*, Vol. 17, No. 1, pp. 103-113.

Kiptoo, M., Mpoke, S., Mueke, J., Okoth, F. and Songok, E. (2009), "Survey on prevalence and risk factors on HIV -1 among pregnant women in North-Rift, Kenya: a hospital based cross-sectional study conducted between 2005 and 2006", *BMC International Health and Human Rights*, Vol. 9, No. 1, pp.10.

Kleinschmidt, I., Pettifor, A., Morris, N., MacPhail, C. and Rees, H. (2007), "Geographic distribution of human immunodeficiency virus in South Africa", *The American journal of tropical medicine and hygiene*, Vol. 77, No. 6, pp.1163-1169.

Lamirel, J.C. and Al Shehabi, S., (2015), "Feature maximization based clustering quality evaluation: a promising approach". *In Trends and Applications in Knowledge Discovery and Data Mining,* Springer, pp. 210-222.

Lamirel, J.C., Cuxac, P., Chivukula, A.S. and Hajlaoui, K., (2014), "Optimizing text classification through efficient feature selection based on quality metric", *Journal of Intelligent Information Systems,* Vol. 45, No. 3, pp. 379 -396.

Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D. and Zimmer, H., (2017), "An Introduction to Statistics", *www.citeseerx.ist.psu.edu*. Accessed: March 9, 2018

Lantz, B. (2013), *Machine learning with R,* Packt Publishers, Birmingham, UK, 396 pp.

Larson, M.G., (2006), "Descriptive statistics and graphical displays", *Circulation,* Vol. 114, No. 1, pp. 76-81.

Lau, C. and Muula, A.S. (2004), "HIV in sub-Saharan Africa", *Croatian Medical Journal*, Vol. 45, No. 4, pp. 402-414.

Le Booysen, F., (2004), "HIV /AIDS, poverty and risky sexual behaviour in South Africa", *African Journal of AIDS Research*, Vol. 3, No. 1, pp. 57-67.

Lehmann, F. (1992), "Semantic networks", *Computers & Mathematics with Applications*, Vol. 23, No. 2-5, pp. 1-50.

Lopman, B., Lewis, J., Nyamukapa, C., Mushati, P., Chandiwana, S., and Gregson, S. (2007), "HIV incidence and poverty in Manicaland, Zimbabwe: is HIV becoming a disease of the poor?", *AIDS*, Vol. 21, No. 7, pp. 57-66.

Luke, N. (2005), "Confronting the'sugar daddy'stereotype: age and economic asymmetries and risky sexual behavior in urban Kenya", *International family planning perspectives*, Vol. 31, pp. 6-14.

Luke, N., (2002), "Widows and 'Professional inheritors': Understanding AIDS risk perceptions in Kenya", *In Population Association of America Annual Meetings*, pp. 8-11.

MacPhail, C., Williams, B.G. and Campbell, C. (2002), "Relative risk of HIV infection among young men and women in a South African township", *International Journal of STD & AIDS*, Vol. 13, No. 5, pp. 331-342.

Madise, N.J., Ziraba, A.K., Inungu, J., Khamadi, S.A., Ezeh, A., Zulu, E.M., Kebaso, J., Okoth, V. and Mwau, M. (2012), "Are slum dwellers at heightened risk of HIV infection than other urban residents? Evidence from population-based HIV prevalence surveys in Kenya", *Health & place*, Vol. 18, No. 5, pp. 1144-1152.

Magadi, M.A. (2013), "The disproportionate high risk of HIV infection among the urban poor in sub-Saharan Africa", *AIDS and Behavior*, Vol. 17, No. 5, pp. 1645-1654.

Mbirimtengerenji, N.D. (2007), "Is HIV epidemic outcome of poverty in sub-Saharan Africa?", *Croatian Medical Journal*, Vol. 48, No. 5, pp. 605.

Mermin, J., Musinguzi, J., Opio, A., Kirungi, W., Ekwaru, J.P., Hladik, W., Kaharuza, F., Downing, R. and Bunnell, R. (2008), "Risk factors for recent HIV infection in Uganda", *Jama*, Vol. 300, No. 5, pp. 540-549.

Messina, J.P., Emch, M., Muwonga, J., Mwandagalirwa, K., Edidi, S.B., Mama, N., Okenge, A. and Meshnick, S.R. (2010), "Spatial and socio-behavioral patterns of HIV prevalence in the Democratic Republic of Congo", *Social Science & Medicine*, Vol. 71, No. 8, pp. 1428-1435.

Michelo, C., Sandøy, I.F. and Fylkesnes, K. (2006), "Marked HIV prevalence declines in higher educated young people: evidence from population-based surveys (1995–2003) in Zambia", *Aids*, Vol. 20, No. 7, pp. 1031-1038.

Mill, J. E. and Anarfi, K. (2002), "HIV risk environment for Ghanaian women: challenges to prevention", *Social Science & Medicine*, Vol. 54, No. 3, pp. 325-337.

Minsky, M. (1974), "A Framework for Representing Knowledge", *The Psychology of Computer Vision*, New York: McGraw-Hill, P. Winston (ed.), pp. 211-277

Morison, L. (2001), "The global epidemiology of HIV /AIDS", *British Medical Bulletin*, Vol. 58, No. 1, pp. 7-18.

Msamanga, G., Fawzi, W., Hertzmark, E., McGrath, N., Kapiga, S., Kagoma, C., Spiegelman, D. and Hunter, D. (2006), "Socio-economic and demographic factors associated with prevalence of HIV infection among pregnant women in Dar es Salaam, Tanzania", *East African Medical Journal*, Vol. 83, No. 6, pp. 311-321.

Msisha, W.M., Kapiga, S.H., Earls, F. and Subramanian, S.V. (2008), "Socio-economic status and HIV seroprevalence in Tanzania: a counterintuitive relationship", *International Journal of Epidemiology*, Vol. 37, No. 6, pp. 1297-1303.

Nagoli, J., Holvoet, K. and Remme, M. (2010), "HIV and AIDS vulnerability in fishing communities in Mangochi district Malawi" *African Journal of AIDS Research*, Vol. 9, No. 1, pp. 71–80.

Narmadha, D., alias Balamurugan, A., Sundar, G.N. and Priya, S.J. (2016), "Survey of clustering algorithms for categorization of patient records in healthcare", *Indian Journal of Science and Technology*, Vol. 9, No. 8, pp. 15-22

Nattrass, N. (2009), "Poverty, sex and HIV", *AIDS and Behaviour*, Vol. 13, No. 5, pp. 833- 840.

Nattrass, N., Maughan-Brown, B., Seekings, J. and Whiteside, A., (2012), "Poverty, sexual behaviour, gender and HIV infection among young black men and women in Cape Town, South Africa", *African Journal of AIDS Research*, Vol. 11, No. 4, pp. 307-317.

Ncayiyana, Dan J. (2004), "Doctors and nurses with HIV and AIDS in sub Saharan Africa", *British Medical Journal*, Vol. 329, No.(7,466), pp. 584-585.

Nel, A., Mabude, Z., Smit, J., Kotze, P., Arbuckle, D., Wu, J., van Niekerk, N. and van de Wijgert, J. (2012), "HIV incidence remains high in KwaZulu-Natal, South Africa: evidence from three districts", *PLoS One*, Vol. 7, No. 4, pp. e35278.

Ngo, P., Kenmochi, Y., Passat, N. and Talbot, H. (2014), "Topology-preserving conditions for 2D digital images under rigid transformations", *Journal of Mathematical Imaging and Vision*, Vol. 49, No. 2, pp. 418-433.

Obi, A., Osaro, E. and Nnenna, F.P. (2011), "Socio-demographic characteristics of adults screened for human immunodeficiency virus infection in Ahoada–East local government area in the Niger delta of Nigeria", *Journal of Global Infectious Diseases*, Vol. 3, No. 4, pp. 334.

Odimayo, M.S., Adediran, S.O. and Araoye, M.O. (2010), "Socio-demographic characteristics of adults screened for HIV in a rural community in Benue state, Nigeria", *African Journal of Clinical and Experimental Microbiology*, Vol. 11, No. 1, pp. 129-136.

Ojini, F.I. and Coker, A., (2007), "Socio-demographic and clinical features of HIV-positive outpatients at a clinic in south-west Nigeria", *African Journal of AIDS Research*, Vol. 6, No.2, pp. 139-145.

Okiria, A.G., Okui, O., Dutki, M., Baryamutuma, R., Nuwagaba, C.K., Kansiime, E., Ojamuge, G., Mugweri, J., Fleuret, J., King, R. and Bazeyo, W. (2014), "HIV incidence and factors associated with seroconversion in a rural community home based counseling and testing program in Eastern Uganda", *AIDS and Behavior*, Vol.18, No. 1, pp. 60-68.

Oluoch, T., Mohammed, I., Bunnell, R., Kaiser, R., Kim, A.A., Gichangi, A., Mwangi, M., Dadabhai, S., Marum, L., Orago, A. and Mermin, J. (2011), "Correlates of HIV infection among sexually active adults in Kenya: a national population-based survey", *The open AIDS journal*, Vol. 5, pp.125-134.

Onchiri, S., (2013), "Conceptual model on application of chi-square test in education and social sciences", *Educational Research and Reviews*, Vol. 8, No.15, pp. 1231-1241.

Opio, A., Muyonga, M. and Mulumba, N. (2013), "HIV infection in fishing communities of Lake Victoria Basin of Uganda–a cross-sectional sero-behavioral survey", *PloS One*, Vol. 8, No. 8, pp. e70770.

Park, H. (2013), "An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain", *Journal of Korean Academy of Nursing*, Vol. 43, No. 2, pp. 154-164.

Parkhurst, J.O. (2010), "Understanding the correlations between wealth, poverty and human immunodeficiency virus infection in African countries", *Bulletin of the World Health Organization*, Vol. *88*, 519-526.

Pena, M., Barbakh, W. and Fyfe, C. (2008), “Topology-preserving mappings for data visualisation”, *In Principal Manifolds for Data Visualization and Dimension Reduction*, Springer, Berlin, Heidelberg, pp. 131-150.

Pettifor, A.E., Hudgens, M.G., Levandowski, B.A., Rees, H.V. and Cohen, M.S. (2007), “Highly efficient HIV transmission to young women in South Africa”, *AIDS*, Vol. 21, No. 7, pp. 861-865.

Piot, P., Greener, R. and Russell, S. (2007), “Squaring the circle: AIDS, poverty, and human development”, *Plos Medicine*, Vol. 4, No. 10, pp. e314.

Quillian, M.R. (1967), “Word concepts: A theory and simulation of some basic semantic capabilities”, *Behavioural Science*, Vol. 12, No. 5, pp. 410-430.

Ramjee, G., Wand, H., Whitaker, C., McCormack, S., Padian, N., Kelly, C. And Nunn, A., (2012), “HIV incidence among non-pregnant women living in selected rural, semi-rural and urban areas in Kwazulu-Natal, South Africa*”, AIDS and Behavior*, Vol. 16, No. 7, pp. 2062 - 2071.

Rashid, P.Q. (2015), “Semantic network and frame knowledge representation formalisms in artificial intelligence”, *Unpublished Doctoral dissertation*, Eastern Mediterranean University, Cyprus, 60pp.

Rehle, T., Shisana, O., Pillay, V., Zuma, K., Puren, A. and Parker, W. (2007), “National HIV incidence measures-new insights into the South African epidemic”, *South African Medical Journal*, Vol. 97, No. 3, pp. 194-199.

Rodrigo, C. and Rajapakse, S. (2010), “HIV, poverty and women”, *International Health*, Vol. 2, No. 1, pp. 9-16.

Rogan, M., Hynie, M., Casale, M., Nixon, S., Flicker, S., Jobson, G. and Dawad, S. (2010), “The effects of gender and Socio-economic status on youth sexual-risknorms: evidence from a poor urban community in South Africa” *African Journal of AIDS Research*, Vol. 9, No. 4, pp. 355-366.

Rosen, S., Ketlhapile, M., Sanne, I. and DeSilva, M.B. (2008), “Characteristics of patients accessing care and treatment for HIV at public and nongovernmental sites in South Africa”, *Journal of the International Association of Physicians in AIDS Care*, Vol. 7, No. 4, pp. 200-207.

Rousseeuw, P.J. and Leroy, A.M. (1987), “Robust regression and outlier detection”, Vol. 1, *Wiley*, New York.

Saraswathi, S. and Sheela, M.I., (2014), "A comparative study of various clustering algorithms in data mining", *International Journal of Computer Science and Mobile Computing*, Vol. 11, No. 11, pp. 422-428.

Schur, N., Mylne, A., Mushati, P., Takaruza, A., Ward, H., Nyamukapa, C. and Gregson, S. (2015), "The effects of household wealth on HIV prevalence in Manicaland, Zimbabwe–a prospective household census and population-based open cohort study", *Journal of the International AIDS Society*, Vol. 18, No. 1, p. 20063.

Sehgal, G. and Garg, D.K. (2014), "Comparison of various clustering algorithms", *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 3, pp. 3074-3076.

Serbessa, M.K., Mariam, D.H., Kassa, A., Alwan, F. and Kloos, H., (2016), "HIV/AIDS among pastoralists and refugees in north-east Africa: a neglected problem", *African Journal of AIDS Research*, Vol. 15, No.1, pp.45-54.

Shandera, W. X. (2007), "Key determinants of AIDS impact in Southern sub-Saharan Africa", *African Journal of AIDS Research*, Vol. 6, No. 3, pp. 271–286.

Shefer, T., Strebel, A. and Jacobs, J. (2012), "AIDS fatigue and university students' talk about HIV risk", *African Journal of AIDS Research*, Vol.11, No. 2, pp. 113-121.

Shelton, J.D., Cassell, M.M. and Adetunji, J. (2005), "Is poverty or wealth at the root of HIV ?", *The Lancet*, Vol. 366, No. 9491, pp. 1057-1058.

Shi, R. and Conrad, S.A. (2009), "Correlation and regression analysis", Annals of Allergy, *Asthma & immunology*, Vol. 103, No. 4, pp. 35 – 41.

Shisana, O., Hall, E. J., Maluleke, R., Chauveau, J., and Schwabe, C. (2004), "HIV/AIDS prevalence among South African health workers", *South African Medical Journal*, Vol. 94, No. 10, pp. 846-850.

Shisana, O., Zungu-Dirwayi, N., Toefy, Y., Simbayi, L.C., Malik, S. and Zuma, K. (2004), "Marital status and risk of HIV infection in South Africa", *South African medical Journal*, Vol. 94, No. 7, pp. 537-543.

Sing, R.K. and Patra, S. (2015), "What Factors are Responsible for Higher Prevalence of HIV Infection among Urban Women than Rural Women in Tanzania?", *Ethiopian Journal of Health Sciences*, Vol. 25, No. 4, pp. 321-328.

Sisodia, D., Singh, L., Sisodia, S. and Saxena, K. (2012), “Clustering techniques: a brief survey of different clustering algorithms”, *International Journal of Latest Trends in Engineering and Technology*, Vol. 1, No. 3, pp. 82-87.

Smith, M.K. (2010), “Gender, poverty, and intergenerational vulnerability to HIV /AIDS”, *Gender & Development*, Vol. 10, No. 3, pp. 63-70.

Sommet, N., and Morselli, D. (2017), “Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS”, *International Review of Social Psychology*, Vol. 30, No. 1, 16 pp.

Sperandei, S. (2014), “Understanding logistic regression analysis”, *Biochemia Medica*, Vol. 21, No. 1, pp. 12-18.

Tanser, F., Bärnighausen, T., Cooke, G.S. and Newell, M.L. (2009), “Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic” *International Journal of Epidemiology*, Vol. 38, No. 4, pp. 1008-1016.

Tanwar, P., Prasad, T.V., and Aswal, M.S. (2010), “Comparative Study of Three Declarative Knowledge Representation Techniques”, *International Journal on Computer Science and Engineering*, Vol. 2, No. 7, pp. 2274-2281.

Temah, C.T. (2009), “What drives HIV epidemic in sub-Saharan Africa?”, *Revue d'économie du développement*, Vol. 17, No. 5, pp. 41-70.

Temam, G. and Ali, A. (2012), “Prevalence of HIV and discordant rate and their associated factors among premarital Voluntary Counseling and Testing (VCT) clients in Addis Ababa public VCT centers”, Addis Ababa, Ethiopia, *Ethiopian Journal of Health Development*, Vol. 26, No. 3, pp. 160-168.

Tladi, L.S. (2006), “Poverty and HIV in South Africa: an empirical contribution”, *Journal of Social Aspects of HIV /AIDS*, Vol. 3, No. 1, pp. 369-381.

Todd, J., Grosskurth, H., Changalucha, J., Obasi, A., Mosha, F., Balira, R., Orroth, K., Hugonnet, S., Pujades, M., Ross, D. and Gavyole, A. (2006), “Risk factors influencing HIV infection incidence in a rural African population: a nested case-control study”, *The Journal of infectious diseases*, Vol. 193, No. 3, pp. 458-466.

UNAIDS (2006), “2006 Report on the global AIDS epidemic”, *A UNAIDS 10th anniversary special edition.*

UNAIDS. (2018), "UNAIDS DATA 2018", *https://www.unaids.org/sites/default/files/media_asset/unaids-data-2018_en.pdf*, Accessed: January 21, 2018.

Wabiri, N. and Taffa, N. (2013), "Socio-economic inequality and HIV in South Africa", *BMC Public Health*, Vol. 13, No. 1, pp. 1037.

Wallrauch, C., Bärnighausen, T. and Newell, M.L. (2010), "HIV prevalence and incidence in people 50 years and older in rural South Africa", *South African Medical Journal*, Vol. 100, No. 12, pp. 812-813.

Walque, D., Nakiyingi-Miiro, J.S., Busingye, J. and Whitworth, J.A. (2005), "Changing association between schooling levels and HIV -1 infection over 11 years in a rural population cohort in south-west Uganda", *Tropical Medicine & International Health*, Vol. 10, No. 10, pp. 993-1001.

Watkins, J.C., (2016), "An introduction to the science of statistics: From theory to implementation", *www.math.arizona.edu/~jwatkins/statbook.pdf*. Accessed: December 10, 2018.

Welz, T., Hosegood, V., Jaffar, S., Bätzing-Feigenbaum, J., Herbst, K. and Newell, M.L., (2007), "Continued very high prevalence of HIV infection in rural KwaZulu-Natal, South Africa: a population-based longitudinal study", *AIDS*, Vol. 21, No. 11, pp. 1467-1472.

Were, M. and Nafula, N. (2003), "An assessment of the impact of HIV on Economic Growth: The case of Kenya", *www.econstor.eu*. Accessed: October 30, 2018.

Whiteside, A. (2002), "Poverty and HIV in Africa", *Third World Quarterly*, Vol. 23, No. 2, pp. 313-332.

Wilson, C.M., Wright, P.F., Safrit, J.T. and Rudy, B. (2010), "Epidemiology of HIV infection and risk in adolescents and youth", *Journal Of Acquired Immune Deficiency Syndromes*, Vol. 51, No. 1, pp. 5-12.

Yahya-Malima, K.I., Olsen, B.E., Matee, M.I. and Fylkesnes, K. (2006), "The silent HIV epidemic among pregnant women within rural Northern Tanzania", *BMC Public Health*, Vol. 6, No. 1, pp.109.

Zulu, L.C., Kalipeni, E. and Johannes, E. (2014), "Analyzing spatial clustering and the spatiotemporal nature and trends of HIV prevalence using GIS: the case of Malawi, 1994-2010", *BMC Infectious Diseases*, Vol. 14, No. 1, pp. 285.

Index

KNOWLEDGE, TRUTH AND EXCELLENCE

KNOWLEDGE, TRUTH AND EXCELLENCE

KNOWLEDGE, TRUTH AND EXCELLENCE

KNOWLEDGE, TRUTH AND EXCELLENCE

KNOWLEDGE, TRUTH AND EXCELLENCE

KNOWLEDGE, TRUTH AND EXCELLENCE

KNOWLEDGE, TRUTH AND EXCELLENCE