



Faculty of Sciences
Department of Mathematics & Statistical Sciences

Hierarchical Resource Allocation Procedures for HIV Prevention in Botswana

by

Atamelang Mareka

Reg. No: 13001057

BSc (Pure and Applied Mathematics), BIUST, Botswana

A Dissertation/Thesis Submitted to the Faculty of Sciences in Partial Fulfilment of the
Requirements for the Award of the Degree of Master of Science in Pure and Applied
Mathematics of BIUST

Supervisor(s):

Professor Semu Kassa
Department of Mathematics & Statistical Sciences,
Faculty of Sciences, BIUST
E-mail: kassas@biust.ac.bw

January 31, 2023

DECLARATION REGARDING THE WORK AND COPYRIGHT

Candidate Atamelang Mareka

Student ID: 13001057

Thesis Titled: Hierarchical Resource Allocation Procedures for HIV Prevention in Botswana

I **Atamelang Mareka**, certify that the Thesis is all my own original work and that I have not obtained a degree in this University or elsewhere on the basis of any of this work.

This dissertation/thesis is copyright material protected under the Berne Convention, the Copyright and Neighbouring Rights Act, Act. No. 8 of 2000 and other international and national enactments, in that behalf, on intellectual property. It must not be reproduced by any means, in full or in part, except for short extracts in fair dealing; for researcher private study, critical scholarly review or discourse with an acknowledgement, without the written permission of the office of the Postgraduate School, on behalf of both the author and the BIUST.

Signed: 

Date: Januray 29, 2023

Supervisor Professor Semu M. Kassa

I **Professor Semu M. Kassa**, hereby confirm that I have inspected the above titled thesis and, to the best of my knowledge, it is based on the original work of the candidate.

Signed: 

Date: Januray 31, 2023

Abstract

In the 1980s the world was hit by the HIV/AIDS pandemic which continues to claim a lot of lives especially in Africa. While a lot has been done since its outbreak in an effort to have it under control the need for funding continues to grow especially in developing countries. Due to economic constraints, funds intended for HIV prevention in countries such as Botswana have declined. In turn, challenges of finding an efficient way to allocate funds optimally emerged. This study intends to first understand the dynamic of the infection of the disease in the population of Botswana. Furthermore to analyse two resource allocation techniques commonly used, these being the equity based model and the cost effectiveness analysis model, in comparison to a mathematical model called hierarchical optimization allocation method. Cost effectiveness is a type of economic analysis that assesses the cost and the outcome or benefit of a program. Another resource allocation model which is commonly used is the equity-based criteria where funds are proportionally allocated across states or districts in a country according to reported cases. Optimization is a technique where the best possible solution is chosen from the available or feasible set of solutions. [The focus of this study](#) is on the bi-level optimization where the decision to get the best possible solution passes through two levels. The number of new infections and HIV prevalence is used as a measure to compare the outcome of the investment. The results in the study reveal that bi-level optimization is the best criteria to use to allocate resources as it has proven to be more efficient in terms of reducing the HIV prevalence at a faster pace, it is also more cost effective as compared to equity and CEA allocation methods. It was therefore concluded that bi-level optimization is the recommended technique for allocating HIV prevention resources and funds.

Acknowledgements

I would like to express my deepest gratitude to my supervisor professor Semu Kassa for his massive input. I would also like to thank my family and friends for their support.



Contents

Contents	v
List of Figures	v
List of Tables	vi
1 Introduction	1
1.1 Problem statement	3
1.2 Significance of the study	3
1.3 Justification	3
1.4 Objectives	4
1.5 Limitations of the Study	4
2 Literature review	5
2.1 Equity model	5
2.2 Cost Effectiveness Analysis	6
2.3 Optimization	8
3 Preliminaries in Optimization Problems	10
3.1 Optimality Conditions	11
3.1.1 Necessary Conditions	12
3.1.2 Sufficient Conditions	12
3.1.3 Classification of Optimization Problems	12
3.1.4 Unconstrained Optimization Problems	12
3.1.5 Constrained Optimization	12
3.2 Bilevel Optimization	14
3.2.1 Reduction to Single level	16
3.2.2 Using the KKT approach	16
3.2.3 Double Penalty Function approach	17
3.2.4 Using the Optimal-Value reformulation approach	18
3.2.5 Solution Algorithms	19
3.2.6 Descent Method	19
3.2.7 Trust Region Method	19
3.2.8 Heuristic Methods	20
3.2.9 Systematic Sampling Method	22
4 Resource Allocation Through Cost Effectiveness Analysis and Equity model	23
4.1 Equity allocation model	25
4.2 Solution Procedure for Equity	25
4.3 Cost Effectiveness Analysis	26

4.4	Solution Procedure for Cost effectiveness analysis	27
5	Allocation through Bi-level Optimization Procedure	30
5.1	Lower level resource allocation	30
5.2	Upper level resource allocation	31
5.3	Formulation of the bi-level programming problem	33
5.4	Solution procedures for bi-level optimization	34
5.5	Solution Approach	36
6	Results and Comparison	38
7	Conclusion and Recommendations	41
	Appendix	43
A	Matlab Codes for Results	43
A.1	The main partitioning MatLab code	43
A.2	The Particle Swarm Optimisation (PSO) MatLab code	48
A.3	Matlab Code for the Results in Tables 6.1, 6.2, 6.3	51
A.4	Additional Matlab Codes for the Investment figures In Bi-level Allocation	65
	Bibliography	71

List of Figures

1.1	Allocation of HIV funds in Botswana [56]	2
2.1	Allocation of HIV funds through equity in Botswana	6
4.1	diagram representing resource allocation	23
4.2	The mathematical model	24
6.1	Allocation of HIV funds in the Central District through equity, CEA and Bilevel optimization.	38
6.2	Allocation of HIV funds in the North East District through equity, CEA and Bilevel optimization.	39
6.3	Allocation of HIV funds in the South East District through equity, CEA and Bilevel optimization.	39



List of Tables

4.1	Upper level equity allocation	26
4.2	Lower level equity allocation-Low risk community	26
4.3	Lower level equity allocation-High Risk community	26
4.4	Cost effectiveness analysis per person(BWP)	29
4.5	Lower level CEA allocation-Low Risk community	29
4.6	Lower level CEA allocation-High Risk community	29
5.1	Disease parameters assumed to be constant	35
5.2	Parameters assumed to vary from district to district.	35
5.3	Initial risk population	35
5.4	Partitioning the decision space	36
5.5	Lower level Bi-level optimization allocation-Low Risk community	37
5.6	Lower level Bi-level optimization allocation-High Risk community	37
5.7	Parametric values used for the PSO simulation	37
6.1	DISTRICT INVESTMENT FOR HIV PREVENTIVE PROGRAMS	40

1. Introduction

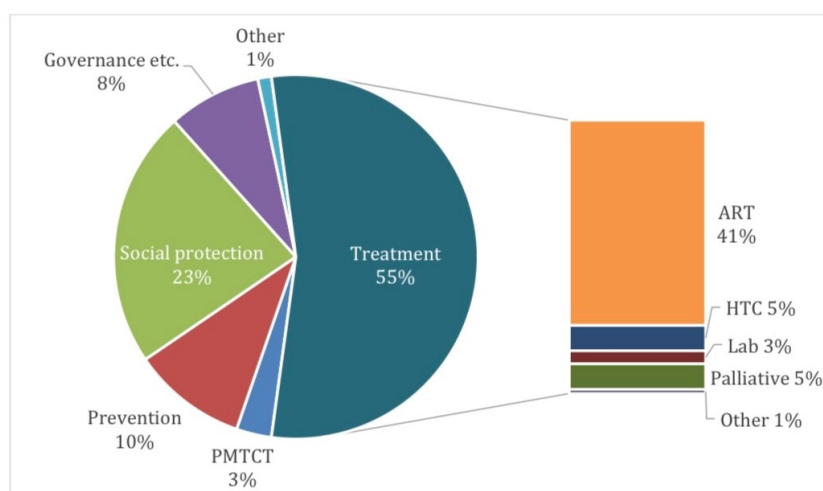
HIV/AIDS [pandemic](#) is a crisis that befell the world in the 1980s and has since spread rapidly across the world. It continues to claim a lot of lives. In the last 20 years, over 60 million people have been infected with HIV, and of those cases 95% are in developing countries [\[34\]](#) while today around 37 million people worldwide live with HIV, some of these infected people are not on treatment.

Over the years a lot has been done to have this epidemic under control. It was only in the late 80's that the very first treatment which was called Retrovir was discovered [\[3\]](#). Before this other mechanisms such as raising awareness, providing technical and financial support and many other programs were used as preventive measures. All this has led to millions of dollars being spent. Although this is the case in 1999 the World Health Organization (WHO) announced AIDS caused by HIV as the fourth biggest cause of death worldwide and the number one biggest killer in Africa [\[14\]](#). In 2001 the United Nations General Assembly called for the creation of a global fund to support efforts by countries and organizations to try and combat the spread [\[6\]](#). This goes to show that from the outbreak of the HIV/AIDS epidemic the level of needing funding has continued to grow. This is because [a](#) lot of resources and funds have been sourced and distributed as a way of investing in the prevention of HIV and the negative impact it has made. Various organizations have raised funds to help fight this epidemic especially in developing countries. In the year 2014 alone USD 19.2 billion was made available globally to finance the response to AIDS in such countries [\[6\]](#). There has been provision of free condoms in some countries and with free antiretroviral therapy which amongst many other things reduces the viral load of HIV. Further there has been a lot of efforts put in raising awareness about HIV, all of these done to control the epidemic.

[Most, if not all, governments](#) especially in developing countries are aware of the need to allocate funds and resources optimally for the betterment of its people's health. That is because currently a challenge encountered in these countries are declining funds from global organizations and in turn they are unable to reach the desired impact. According to [\[4\]](#) recent years have seen a regression, donor funding for the HIV response in low- and middle-income countries declined by 7% between 2015 and 2016. At the end of 2018, *US\$19.0 billion* was available for the AIDS response in low- and middle-income countries, almost 1 billion less than in 2017 [\[55\]](#).

Below is a graph depicting Botswana's allocation of the HIV prevention funds approach in the years 2013/14 to 2017/18 as shown by [\[56\]](#)

From the graph in [Figure 1.1](#) it is evident that in the years indicated Botswana allocated more



Source: NAHPA, Analysis of HIV Investment in Botswana, 2019

Figure 1.1: Allocation of HIV funds in Botswana [56]

than half of its HIV funds towards treatment some percentage towards prevention of mother to child transmission. This sheds some light towards how Botswana may still be allocating its HIV prevention funds.

Botswana being an upper middle income country faces a huge challenge in sustaining its impressive HIV response as donors increasingly focus on low-income countries. Although Botswana is known for being the first country to provide free antiretroviral treatment in the Sub Saharan region it is the fourth highest HIV prevalence in the world. However, with only 370 000 people living with HIV recorded in 2018, 200 000 are women while the remaining is their counterparts [57]. Young people in Botswana are also highly affected by this pandemic with a reflection of low levels of HIV knowledge among young people, with less than half (47%) of those between 15 and 24 able to answer basic questions on HIV [6]. Overall, there has been a significant decrease in the number of new infections per year from 18 000 in 2005 to 8500 in 2018 [6].

According to [20] commendable progress has been made towards accelerating access to HIV prevention, treatment, care and support services for those in need. However, the high incidence rate suggests that even though it has been 40+ years since the epidemic and even though the efforts that have been put into fighting it have been undeniably significant, HIV continues to outpace these efforts. Since it remains that the country has a high HIV prevalence it goes to show that there is still a need for the country to work on reducing their new infections, that is, to efficiently prevent the virus. With the funds the country allocates it needs to see to it that they are distributed efficiently.

Resource allocation is the procedure in which funds are allocated to the intervention programs addressing and solving a similar or common challenge. It should be done in a manner that supports an organization or in this case a country's goals. There are several types of models

that can be used to support the decision-making process for HIV/AIDS resource allocation, these include equity, cost effectiveness and optimization. Allocation based on equity is in a criteria where resources are distributed proportional to the number of HIV/AIDS in different target groups, while cost effectiveness examines and compares the costs and the health benefits of each HIV prevention intervention. With optimization, the problem is usually stated as one of choosing the amount to be invested in several interventions to optimize total health benefits subject to a budget constraint [35].

1.1 Problem statement

Ideally it is every country's wish to have zero HIV new infections. Even with declining funds, Botswana still aims to have their rate of new infections significantly reduced. Botswana had targetted to have zero new infections by 2016. However in the year 2016, Botswana had 10 000 new infections [6]. Currently in Botswana the target is to have at most 3000 new infections per year [44]. However the reality is that although the number of new infections has gone down from 15000 in 2005 to 9500 in 2019 [5] there is still a lot that needs to be done to reduce this number to the targeted 3000. One of Botswana's greatest challenges in responding to HIV is funding the intervention programs. International donors are withdrawing funding and this leaves the country with the challenge of having to effectively allocate the little that they have to reach their goals. It is therefore of great importance to evaluate the resource allocation techniques that are a contribution to the country's obstacle in reaching their goals. There are different methods which can be used to allocate the scarce funds. However the most crucial thing is that the applied method be the most effective to achieve desirable results. That is, for this study, identifying the most effective method among the existing ones is the target. These techniques include Cost effectiveness analysis, equity and allocation through optimization.

1.2 Significance of the study

This study is relevant since Botswana is still one of the countries with a very high HIV prevalence even though it has invested a lot in its prevention. Evaluating and understanding the available options in the effective or rather optimal allocation of HIV prevention resources will lead to the country being at a better position to decide on the best possible allocation technique which will as a result lead to the country experiencing a significant and vital decrease in the number of new HIV infections. Also this study will give researchers in Botswana an insight and a basis for further investigation into matters surrounding effective allocation of HIV prevention resources.

1.3 Justification

With Botswana being faced with the challenges of being one of the countries with the highest HIV prevalence and having their HIV prevention funds being reduced in spite of this all. It becomes vital to see to it that the resource allocation process being used is one that will lead to the most satisfactory results being achieved. The most commonly used methods of resource

allocation in different countries include equity model and cost effectiveness analysis, however [31] has shown that these are not entirely efficient and have suggested the use of optimization to get more significant results. Therefore in this study there will be an analysis of these methods using Botswana data, for the optimization technique the bilevel optimization model will be used. Not enough research has been done in regards to this and therefore by adding to the little that exists some light will be shone on what will ultimately get the country more positive results.

1.4 Objectives

Main Objective

The main objective is to compare the efficiency of different resource allocation techniques used for HIV prevention purposes in Botswana with allocation through optimization, particularly the bilevel optimization.

Specific objectives

1. To investigate the impact of the previously employed methods of allocation of HIV prevention funds and resources and determine the outcome of optimal allocation techniques, in particular two level resource allocation may have in Botswana.
2. To find out how efficient Cost effectiveness analysis and the equity model are in allocating HIV prevention resources in Botswana
3. To compare the results of Cost Effective Analysis and equity model with allocation by optimization for the case of Botswana.

1.5 Limitations of the Study

The limitations experienced during this study include

- The process to access data collected from the ministry was lengthy mostly because of the interference by the Corona Virus pandemic and eventually we did not manage to get any data at all we however resorted to using other data available online which was nonetheless resourceful.
- Another one was the age of the secondary data that was found, the data was mostly from three or four years back nothing too recent. However the data was still relevant and useful for the research.

2. Literature review

A lot of countries especially those that are developing are struggling with distributing the limited resources or funds that they have optimally in order to achieve the greatest possible impact. There are different kinds of funds and resources, those acquired from global organizations and those raised at national level. The challenge currently encountered in these countries are declining funds from global organizations and them not being able to reach an impact that is desired. That is, according to [6] recent years have seen a regression, donor funding for the HIV response in low- and middle-income countries declined by 7% between 2015 and 2016 and hence the urgency in seeing that the funds and resources that are there are distributed efficiently.

National governments in some low- and middle-income countries are now coming forward to fund their own responses, with the aim to fill the funding gaps and to also work towards a more sustainable response to the HIV pandemic. This is because though through the years recognizable progress has been made there is still important milestones and breakthroughs to be achieved when it comes to HIV infection aversion. The world has agreed to meet a set of global targets by 2020 as part of UNAIDS Fast-Track strategy to end the AIDS epidemic as a public health threat [55]. This is a target that all countries especially the developing ones should aim to reach at an optimal level.

Funds spent on HIV prevention go through different levels of distribution. In each level there are decision makers who commonly use heuristic methods and ratio based methods to distribute resources and funds [31]. In [34] it is highlighted that for allocation to take place in Pakistan it is absolutely dependent on intuition and incremental budgeting. Furthermore because of the technical skills required by the official decision making tools, they are not used. The decision making for the HIV prevention resource allocation methods under particular organizations, including large NGOs and also for governments should begin with determining the programs that will be implemented in a community. Upon selecting these programs then the decision makers can focus on the allocation techniques. Below is an overview according to some literature, of the HIV prevention allocation techniques that this study will be focusing on.

2.1 Equity model

This method is also known as the equity model where a budget B is allocated proportionally, based on the number of the total infected population. Funds targeted towards HIV interventions are commonly allocated using the equity-based criteria and often go through several levels of distribution. According to [34] equity can be applied under each level. We define equity as proportional to the number of HIV cases in each region. Equity is viewed as essential since it represents fairness and justice in distribution. Simple resource allocation models can be

based on equity criteria such as an allocation proportional to the number of HIV/AIDS cases in different target groups [30]. Equity based allocation method is the most commonly used resource allocation method for distributing funds towards HIV prevention. It is said to conceivably reward areas which report higher caseloads and use their resources ineffectively [30].

Equity is a critical and controversial factor when deciding how to allocate public resources [46]. Therefore it is crucial to clearly define equity in order to have a model that will cover all needs. For all the definitions of equity that have been put forward, one element remains common, that is, across all sets of individuals, resources should be allocated equally. In any case, this commonality leaves much to be chosen especially how to characterize equity in a particular decision-making setting. [46].

According to [56] in the year 2019 Botswana allocated its HIV prevention funds using the equity model. Districts are ranked into three priority groups depending on the estimated number of new infections and the level of HIV prevalence. That is, these districts are separated into high, medium and low priority groups. Below is a graph taken from [56] illustrating how equity allocation is applied in allocating HIV funds in Botswana.

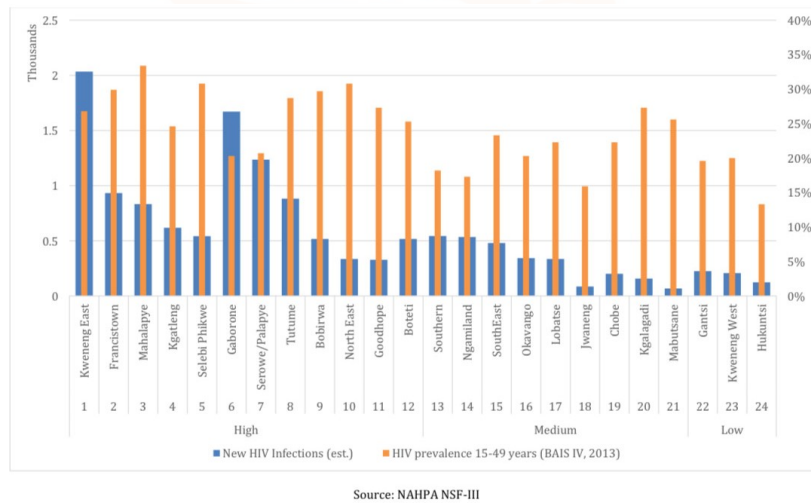


Figure 2.1: Allocation of HIV funds through equity in Botswana

2.2 Cost Effectiveness Analysis

Cost Effectiveness Analysis (CEA) in [24] is defined to be a specific type of economic analysis in which all costs are related to a single, common effect. Cost Effectiveness is calculated using a ratio where the sum of all benefits is divided by the sum of all costs. Decision makers can use CEA to compare various resource allocation methods. They can also use CEA to consider alternatives in a comparable and objective way that provides support for the ultimate choice. However it is stated in [24] that CEA will not always reveal a distinct choice.

On the other hand, [27] defines cost effectiveness as the minimum cost for a given health benefit, or equivalently, maximal health benefit for given expenditures on health care. This paper goes on to show how one can measure the outcomes using 'quality adjusted life years' (QALY). QALY combines the number of lives lived and the health quality of these years. Outcomes can also be measured using the DALY-(Disability-Adjusted Life Years). DALYs for a disease or health condition are calculated as the sum of the Years of Life Lost (YLL) due to premature mortality in the population and the Years Lost due to Disability(YLD) [14].

Cost Effectiveness Ratio (CER)- will inform the decision makers about the cost of an intervention in correspondence to the health benefits that is brought about by the intervention. For instance, in a case where an intervention costing \$ 30,000 per patient and providing 5 QALYs (5 years of perfect health) has a CER of $\$ 30,000/5 = \$ 6000$ per QALY. By measuring an intervention in this manner, one is able to directly compare the cost effectiveness of different programs [32].

Incremental Cost-Effectiveness Ratio (ICER)- The incremental cost effectiveness ratio (ICER) is a measure put in place to allow for comparison among two separate programs based on the cost as per the benefit acquired. ICER is calculated by finding the difference between the cost of program one and two and dividing it by the difference of their effectiveness [32].

Cost-effectiveness analysis is instrumental as it empowers decision makers focusing on HIV prevention programs to make knowledgeable choices about resource allocation. The efficiency of various programs can be assessed by measuring and analysing their costs and outcomes. This way even an approximation on the future resource and fund requirements can be made [7].

In [7] it is highlighted that, the problems that managers and planners face while using CEA. These among others include, insufficient information on the costs and impact of an intervention, availability of expertise, recognising and analysing appropriate pointers to weigh the impact. They also classify cost effectiveness analysis as an economic evaluation tool and proceed to define the techniques of this evaluation as follows,

- Cost analysis- costs are measured and the average or incremental cost is calculated.
- Cost-effectiveness analysis- impact is measured using indicators related to the change in health status.
- Cost-benefit analysis- impacts of the intervention are translated into monetary terms, in order to obtain a ratio.
- Cost-utility analysis- impact is measured in terms of gains in the quality-adjusted life-years (QALYs) of an individual.

While [42] states that health organizations over the world are required to set priorities and allocate resources within the constraint of limited funding. That is, regardless of the scarcity

of resources it is still of paramount importance to be conscious of where funds are distributed to. Here they describe an approach which is used for setting context for priority called program budgeting and marginal analysis (PBMA).

2.3 Optimization

In [31] Semu Kassa also analysed the allocation of funds and proposes a hierarchical mathematical optimization model to solve the problem of failing to optimally distribute funds or resources to control or prevent HIV. This paper concentrates on developing a model where the decision for resource allocation passes through various levels. In the problem formulation they assume that major decisions concerning resource allocation is covered in three hierarchical groupings where the upper is the international funding organization the middle one is the regional representatives like the Sub Sahara Africa and finally the lower ones being the country representatives. The lower level minimizes the total number of individuals who become newly infected while the middle level maximizes the total number of HIV infections averted in the region and the upper level maximizes the global gain in health.

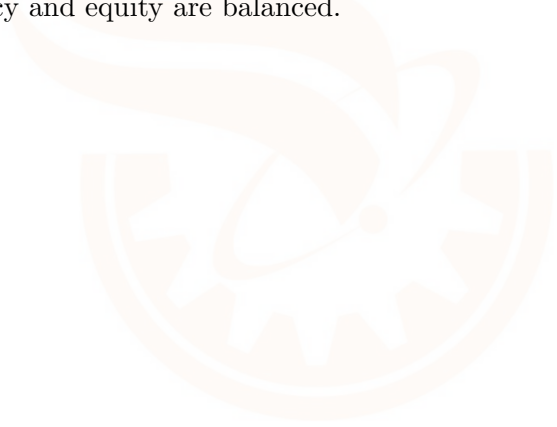
The difference in the numbers of HIV infections averted according to how resources were allocated, either through proportion to HIV incidence (equity) or through efficient distribution is revealed in [59]. The optimization model is based on the assumption that with a particular constraint placed on the total budget, the objective is to minimize the number of new infections. Hence, observing the time between zero and time T using discrete time, the collective number of new infections will be estimated.

A. Lasry et al. in [34] also evaluates the impact of simple resource allocation heuristics versus optimal allocation, given two levels of decision-making, in the context of Sub-Saharan Africa. They model a two-level resource allocation process with two regions of high risk and low risk populations and they conclude that applying optimality on both two levels yields an improvement in HIV prevention. Also [59] models a two level resource allocation model in a similar way Lasry and team did, that is, analyzing the allocation of HIV prevention resources at two decision making levels. Furthermore [59] also had four approaches which were optimal - optimal, optimal - proportional, proportional - optimal, and proportional - proportional. They concluded that it is not enough to only allocate resources optimally at the upper level while neglecting the lower level in that regard. Another one of their conclusions is that if optimization is to be applied on only one level of decision making it is the lower level as it will yield great results.

In [31] a hypothetical example was used to show the procedure of this three level programming problem since there is no actual data from the field. However this author unlike [34] who recommended an equity-optimal model recommends that the use of optimal – optimal approach over the other possible approaches.

S. R. Earnshaw et al. in their article [19] develop a linear programming model for allocating HIV prevention funds which they highlighted to be the most easily understood mathematical programming technique to use. Their linear programming model determines an allocation that maximizes the weighted number of potential infections averted subject to the availability of funds. They used it to determine efficient allocations of the HIV prevention funds at Florida under a series of eight scenarios. These scenarios gave different results though some were similar to others. However it was concluded that overall the linear programming method over these scenarios and their varying constraints performs better than the resource allocation method that was being applied at Florida at the time.

However, Kaplan and Merson in [30] argue that the success of all the work that goes into HIV prevention is dependent on the decision of the magnitude and level at which the interventions are funded. Furthermore, [30] has indicated that what is more pivotal than increasing the total budget meant for prevention is being cautious of how HIV prevention funds are channelled into various interventions. In this study [30] different allocation procedures are compared, that is, proportional allocation and equity and cost effectiveness and efficiency and they also offer an alternative where efficiency and equity are balanced.



3. Preliminaries in Optimization Problems

Optimization techniques are guaranteed to assist more in the effective allocation of limited resources through the apparent need for more than what is available. Optimization comes from the word "optimal" which means best. When you optimize something you are making it best. Optimising can be done by either maximizing or by minimizing a solution. A couple of researchers have shown where proportionality models which are commonly used to allocate funds fall short while highlighting the impact mathematical optimization might have in controlling the HIV pandemic.

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a real valued function, and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ be vector valued functions of appropriate dimensions. Using these functions a basic optimization problem can be formulated as follows,

$$\begin{aligned} &\text{optimise} && f(x_1, \dots, x_n) \\ &\text{s.t} && g_j(x_1, \dots, x_n) \leq 0, \quad j = 1 : m \\ &&& h_i(x_1, \dots, x_n) = 0, \quad i = 1 : p \end{aligned} \tag{3.1}$$

Since optimizing can either be minimizing or maximizing, in the case that the objective function $f(x_1, \dots, x_n)$ is to be minimized then the formulation of the general optimization problem becomes,

$$\begin{aligned} &\min && f(x_1, \dots, x_n) \\ &\text{s.t} && g_j(x_1, \dots, x_n) \leq 0 \quad j = 1 : m \\ &&& h_i(x_1, \dots, x_n) = 0 \quad i = 1 : p \end{aligned} \tag{3.2}$$

where

- f is the objective function - which is a real valued function whose value is to be minimized over a set of possible alternatives.
- $x_1, \dots, x_n$ are decision variables which are the inputs, values that the decision maker can control. These represent values or quantities to be determined.
- The functions g and h are constraints - which are usually represented through equations or inequalities. Constraint $g_j(x_1, \dots, x_n) \leq 0$ is what is called the inequality constraint while $h_i(x_1, \dots, x_n) = 0$ is the equality constraint.

To define the concept of the solution for the optimization problem (3.2) the following terms are defined,

Definition 1 (Local minimum). [18] Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a real valued function and let $x^* \in \mathbb{R}^n$, then x^* is a local minimum of f if there is an $\varepsilon > 0$ in a neighborhood $N(x^*; \varepsilon)$ of x^* such that

$$f(x^*) \leq f(x) \quad \forall x \in N(x^*; \varepsilon). \quad (3.3)$$

A local maximum is defined similarly but with the inequality reversed.

Definition 2. [15] A point x^* satisfying the constraints $h_i(x_1, \dots, x_n) = 0$ and $g_j(x_1, \dots, x_n) \leq 0$ is said to be a regular point of the constraints if the gradient vectors $\nabla h_i(x_1, \dots, x_n)$, $\nabla g_k(x_1, \dots, x_n)$, where $g_k(x_1, \dots, x_n) = 0$, are linearly independent.

Definition 3 (Convex Function). [18] If a set $C \subset \mathbb{R}^n$ is a convex set then the function $f : C \rightarrow \mathbb{R}$ is said to be convex if for any $x_1, x_2 \in C$

$$f(\varphi x_1 + (1 - \varphi)x_2) \leq \varphi f(x_1) + (1 - \varphi)f(x_2) \quad (3.4)$$

for all $\varphi \in \mathbb{R}$ such that $0 \leq \varphi \leq 1$.

A global optimum is needed to get a more precise answer. However for the existence of such a solution one may need to make some assumptions. That is,

1. The set $C \subset \mathbb{R}^n$ is compact;
2. $f(x)$ is smooth - That is f should be continuously differential for certain order especially for the first and second order;
3. $f(x)$ is convex.

Definition 4 (Global minimum). [9] Let $x^* \in \mathbb{R}^n$ then x^* is a global minimum of a function f over the set $C \subset \mathbb{R}^n$ if it satisfies

$$f(x^*) \leq f(x) \quad \forall x^* \in C. \quad (3.5)$$

For a global maximum all else is similar except the inequality is reversed. Now suppose that all the involved functions are differentiable in the interior of their domains.

Definition 5 (Lagrange multipliers). [33] Lagrange method is used for maximizing or minimizing a general function $f(x)$ subject to a constraint of the form $g(x) = 0$. If the extreme values exist then $\nabla f(x^*) = -\lambda \nabla g(x^*)$, where λ is called the Lagrange multiplier.

3.1 Optimality Conditions

An optimality condition is a condition that a solution x^* to (3.2) must fulfil. Optimal solutions can be characterised according to these conditions. These conditions are called the necessary and sufficient conditions.

3.1.1 Necessary Conditions

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable at x^* . Also given that $x^* \in \mathbb{R}^n$ is a local minimum then

$$\nabla f(x^*) = 0.$$

The condition above uses the gradient vector whose components are the first partial derivatives, it is therefore called the first-order condition.

For the second order condition, in [18] it is supposed that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable at x^* and that $x^* \in \mathbb{R}^n$ is a local minimum where $\nabla f(x^*) = 0$ then the hessian matrix of f at x^* , that is $\mathbf{H}(x^*) = \nabla^2 f(x^*)$ is a positive semi definite matrix.

These are the necessary conditions which must be true for every local minimum.

3.1.2 Sufficient Conditions

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice differentiable at x^* . If $\nabla f(x^*) = 0$ and $\mathbf{H}(x^*)$ is positive semi definite then x^* is a local minimum of f [18].

Furthermore, suppose in addition that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex function then $\mathbf{H}(x)$ is positive semi definite for all x in the domain of f . In such a case, if $\nabla f(x^*) = 0$ we can simply conclude that x^* is a global minimum [18].

3.1.3 Classification of Optimization Problems

As highlighted above, optimization problems can either be constrained or unconstrained. Below is a brief explanation of both constrained and unconstrained optimization problems.

3.1.4 Unconstrained Optimization Problems

The optimization (in this case minimization) of a function f with the absence of constraints is written as

$$\min_{x \in \mathbb{R}^n} f(x). \quad (3.6)$$

The solutions to an unconstrained optimization problem can be obtained through the above conditions.

3.1.5 Constrained Optimization

1. **Problems with Equality Constraints** : This problem is defined as

$$\begin{aligned} \min_{x \in X} \quad & f(x) \\ \text{s.t} \quad & h_j(x) = 0 \text{ for all } j = 1, \dots, p \end{aligned} \quad (3.7)$$

where $h_j(x) = 0$ is an equality constraint.

Define the Lagrangian of (3.7) as:

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^p \lambda_i h_i, \quad \lambda \in \mathbb{R}^p. \quad (3.8)$$

In accordance with [9], let f have a minimum at x^* then there exist a unique λ^* such that

$$\begin{aligned} \nabla_x \mathcal{L} &= \nabla f(x^*) + \sum_{i=1}^p \lambda_i^* \nabla h_i(x^*) = 0 \\ \nabla_\lambda \mathcal{L} &= h_j(x^*) = 0, \forall j = 1, \dots, p. \end{aligned}$$

These make the first order necessary conditions as stated by [9]. The function $\mathcal{L}$ is the Lagrangian and the components λ_i are called the Lagrange multipliers or dual variables.

In vector forms, the above expression can be written as follows. Suppose that x^* is a local minimum of f subject to $h(x) = 0$ (with $h(x) = (h_1(x), \dots, h_p(x))$), then there exist $\lambda \in \mathbb{R}^p$ such that

$$\nabla f(x^*) + \lambda^T \nabla h(x^*) = 0. \quad (3.9)$$

Denote $L(x^*) = \nabla_x^2 \mathcal{L}$, $F(x^*) = \nabla^2 f(x^*)$ and $H(x^*) = \nabla^2 h(x^*)$. If x^* is a minimum point, then the matrix $L(x^*) = F(x^*) + \lambda^T H(x^*)$ is positive semi-definite on the tangent plane $T = \{y \in \mathbb{R}^n : \nabla h(x^*)^T y = 0\}$. This being the second order necessary condition according to [9].

For the second order sufficiency conditions according to [15] let x^* be a local minimum of the optimization problem (3.7). Suppose that x^* is regular, then $\nabla f(x^*) + \lambda^T \nabla H(x^*) = 0$ and the matrix $L(x^*) = F(x^*) + \lambda^T H(x^*)$ is positive semi definite on $T = \{y \in \mathbb{R}^n : \nabla h(x^*)^T y = 0\}$ then x^* is a strict local minimum of f .

2. Problems with Inequality Constraints : This problem is defined as

$$\begin{aligned} \min_{x \in X} \quad & f(x) \\ \text{s.t} \quad & g_i(x) \leq 0 \quad i = 1, \dots, m \end{aligned} \quad (3.10)$$

where $g_i(x) \leq 0 \quad i = 1, \dots, m$ is an inequality constraint.

To derive first and second order optimality conditions for the problem above, it is assumed that some regulatory conditions are satisfied.

Let x^* be a point satisfying the constraint $g(x^*) = (g_1(x^*), \dots, g_m(x^*)) \leq 0$ and let I be a set of indices i such that $g_i(x^*) = 0$, then x^* is said to be a regular point if the gradient vectors $\nabla g_i(x^*)$ are linearly independent for $i \in I$ [9]. If $g_i(x) = 0$ for some $i \in \{1, \dots, m\}$, then the constraint g_i is called an active constraint at x .

As written in [52], let x^* be a local minima and a regular point for problem (3.10) then there exist a unique λ^* such that,

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0 \quad (3.11)$$

$$g(x^*) \leq 0$$

$$(\lambda^*)^T g(x^*) = 0 \quad (3.12)$$

$$\lambda^* \geq 0. \quad (3.13)$$

These conditions are called the Karush Kuhn Tucker(KKT) Conditions. In the above formulation, condition (3.12) is called the complementarity condition and according to [9] it enforces either λ^* to be zero or $g(x^*)$ to be zero. It is equivalent to saying that a component of λ may be non-zero only if the corresponding constraint is active. An active constraint being one where $\lambda^* > 0$ making condition (3.12) $g(x^*) = 0$.

To derive the second order necessary conditions, suppose that the functions f and g are twice continuously differentiable at x^* and that x^* is a regular point of the constraints in (3.10). If x^* is a relative minimum of f then there exist a vector $\lambda \in \mathbb{R}^m$ such that (3.11) - (3.13) holds and $L(x^*) = F(x^*) + \lambda^T H(x^*)$ is positive semi-definite.

For a point x^* satisfying the constraints in (3.10) to be a strict relative minimum of problem (3.10) the following conditions must be satisfied:

- (a) x^* satisfies all the constraints and there exist $\lambda \in \mathbb{R}^m$ such that conditions (3.11) - (3.13) hold,
- (b) the Hessian matrix $L(x^*) = F(x^*) + \lambda^T G(x^*)$ where $G(x^*) = \nabla^2 g(x^*)$ is positive semi definite.

This is the second order sufficiency conditions according to [9].

3.2 Bilevel Optimization

Bilevel optimization is a type of optimization where one problem is embedded in another problem, that is one optimization problem is contained in another. The outer problem is called the upper level problem while the one embedded in it is called the lower level problem.

The general bilevel programming problem is as follows,

$$\begin{aligned}
 & \min_{x \in X} F(x, y) \\
 & \text{s.t. } G(x, y) \leq 0 \\
 & \text{where } y \text{ solves} \\
 & \min_{y \in Y} f(x, y) \\
 & \text{s.t. } g(x, y) \leq 0.
 \end{aligned} \tag{3.14}$$

The variables of problem (3.14) are divided into two classes namely the upper-level variables where $x \in X$ and the lower level variables where $y \in Y$. Similarly the functions $F(x, y)$ and $f(x, y)$ are the upper and lower level objective functions respectively, while $G(x, y)$ and $g(x, y)$ are the upper level and lower level constraint functions.

With regard to the bi-level problem above (3.14) the leader who is the upper level decision maker can find themselves in an uncertain position if the follower or lower level decision maker has multiple optimal solutions. This uncertainty is brought by the fact that it hence becomes unclear which decision the follower might take. Given this case the leaders often find themselves in one of two common positions, optimistic or pessimistic position.

1. **Optimistic Position:** According to [51], given that the leader is in this position and also provided that the lower level has multiple optimal solutions, the follower will be expected to opt for a solution among the optimal set, which will yield the **best** objective function value at the upper level. That is there is an assumption that there is some cooperation between the leader and follower. The choice function of the follower therefore becomes

$$\Psi(x_u) = \operatorname{argmin}_{x_l \in X_l} [F(x_u, x_l)],$$

where the pair (X_l, Y_l) is the solution set of the lower level problem. This function is thereafter included in the constraints under the upper level.

2. **Pessimistic Position:** Furthermore [51] goes on to define pessimistic position as a case where the leader optimizes for the worst possible scenario, that is the leader assumes that the follower will opt for a solution among the optimal set which in this case will yield the worst objective function value at the upper level.

$$\chi(x_u) = \operatorname{argmax}_{x_l \in X_l} [F(x_u, x_l)],$$

where the pair (X_l, Y_l) is the solution set of the lower level problem. Again, this function is thereafter included in the constraints under the upper level.

Definition 6. *The Constraint region of the bilevel problem is such that $S = \{(x, y) : x \in X, y \in Y, G(x, y) \leq 0, g(x, y) \leq 0\}$.*

For each fixed $x \in X$, we assume that there is a feasible set in the lower level such that

$S(x) = \{y \in Y : g(x, y) \leq 0\}$. That means, we assume that the set $S(x)$ is nonempty for each fixed $x \in X$.

Definition 7. *The follower's reaction set is such that $P(x) = \{y \in Y : y \in \operatorname{argmin}[f(x, y) : y \in S(x)]\}$.*

Definition 8. *The inducible region is the set $\mathcal{R} = \{(x, y) \in S, y \in P(x)\}$.*

The general solution for a bi-level problem is such that, let $(x^*, y^*) \in S$ be the optimal solution for the upper level objective function F then $y^* \in Y$ is the optimal solution for f in the lower level. This can be describe more precisely using the inducible set as, $(x^*, y^*) \in S$ is a solution to the original bilevel problem (3.14) if it solves

$$\min_{(x, y) \in \mathcal{R}} F(x, y).$$

According to [49], besides the concepts of optimistic and pessimistic considerations local and global solutions can also be defined as in the case of classical optimization.

Definition 9. *A feasible point (x^*, y^*) of the bi-level problem (3.14) is a local optimum if there exists an $\epsilon > 0$ such that $F(x, y) \geq F(x^*, y^*)$ holds for all $(x, y) \in \mathcal{R}$ with $\|(x, y) - (x^*, y^*)\| < \epsilon$. A local optimum solution is called a global optimal solution if $\epsilon > 0$ can be chosen arbitrary large.*

3.2.1 Reduction to Single level

3.2.2 Using the KKT approach

There is a possibility and chance that the lower level optimization problem can be replaced with its Karush-Kuhn-Tucker (KKT) conditions, provided that the lower level problem is convex and sufficiently regular.

KKT conditions are necessary conditions which play an important role in constrained optimization. They are defined as first order conditions that provide a unified treatment of constrained optimization [9]. The KKT conditions should allow for inequality constraints, also for any number of constraints, the constraints may be binding or not at the solution [9].

Given an optimization problem (3.14) where it is assumed that there are $m > 1$ inequality constraints, the Lagrangian of the lower level problem is given by,

$$\mathcal{L}(x, y, \lambda) = f(x, y) + \sum_{j=1}^m \lambda_j g_j(x, y). \quad (3.15)$$

Then with the satisfaction of the regularity conditions, for a local minimum of the lower level problem, there exist a unique λ^* such that

$$\begin{aligned} \nabla_y \mathcal{L}(x^*, y^*, \lambda^*) &= 0 \\ g_j(x^*, y^*) &\leq 0, j = 1 : m \\ \lambda_j^* g_j(x^*, y^*) &= 0, j = 1 : m \end{aligned} \quad (3.16)$$

$$\lambda_j^* \geq 0, j = 1 : m$$

Then according to [52], the KKT conditions as shown above appear as Lagrangian and complementarity constraints, and reduce the overall bilevel optimization problem (3.14) to a single-level constrained optimization problem of the following form,

$$\begin{aligned} & \min_{x \in X, y \in Y, \lambda \in \mathbb{R}_+^m} F(x, y) \\ & \text{subject to} \\ & G_k(x, y) \leq 0, k = 1 : K \\ & \nabla_y L(x, y, \lambda) = 0 \\ & g_j(x, y) \leq 0, j = 1 : m \\ & \lambda g_j(x, y) = 0, j = 1 : m \\ & \lambda_j \geq 0, j = 1 : m \end{aligned} \tag{3.17}$$

$$\text{where} \\ L(x, y, \lambda) = f(x, y) + \sum_{j=1}^m \lambda_j g_j(x, y)$$

If the lower problem is strictly convex, then problem (3.17) is equivalent to the original bilevel problem (3.14). However, this method has its challenges. The Lagrangian constraints can lead to non-convexities even when suitable convexity assumptions are made on all the objectives and constraints in the bi-level formulation. Moreover, the complementarity condition makes the single optimization problem a mixed integer problem [51].

3.2.3 Double Penalty Function approach

With the penalty function method the bilevel optimization problem is handled by solving a series of unconstrained optimization problems. Basically, the idea is to transform the two level problem into a sequence of single level unconstrained problem. According to [8] to solve a general bilevel programming problem (BLPP) it is assumed that for any fixed $x \in X$ there exist the optimal solution $y = f(x)$ which is uniquely determined. Also in [8] the penalty function approach was proposed in which the follower's problem was replaced by an unconstrained problem with an augmented objective function,

$$p(x, y, r) = f(x, y) + r\phi(g(x, y)), \quad r > 0, \tag{3.18}$$

where ϕ is a continuous penalty function and r is a parameter.

The general optimization problem (3.2) becomes,

$$\begin{aligned}
 & \min_{x,y} F(x, y) \\
 & \textbf{subject to} \\
 & G(x, y) \leq 0 \\
 & \nabla_y p(x, y, r) = 0 \\
 & g(x, y) < 0,
 \end{aligned} \tag{3.19}$$

where the last constraint prescribes the domain of the function p .

The next step is to solve (3.19) using a given parameter r and introduce a new overall penalty function augmented objective function

$$Q(x, y; t, s, r) = F(x, y) + t\Phi(G(x, y)) + r\phi(g(x, y)) + s\Psi(\|\nabla_y p(x, y, r) = 0\|), \tag{3.20}$$

where Φ is a continuous penalty function and Ψ is a continuous monotone increasing penalty function.

Hence we can approximate the problem (3.14) by the following single level unconstrained problem,

$$\min_{x,y} Q(x, y; t, s, r). \tag{3.21}$$

Further [51] highlights that to achieve the single level problem both the upper and lower level objective functions are penalized. That is, the penalized lower level problem is replaced with its KKT conditions and thereafter the remaining single level unconstrained problem is solved by the penalty approach.

3.2.4 Using the Optimal-Value reformulation approach

By defining the value function for the lower level problem as in [45] by

$$h(x) = \inf_{y \in S(x)} f(x, y), \tag{3.22}$$

one can also equivalently formulate the bilevel problem (3.14) into a single level optimization problem [22],

$$\begin{aligned}
 & \min_{x \in X, y \in Y} F(x, y) \\
 & \textbf{subject to} \\
 & G_k(x, y) \leq 0, k = 1 : K \\
 & f(x, y) \leq h(x) \\
 & g_j(x, y) \leq 0, j = 1 : m.
 \end{aligned} \tag{3.23}$$

Problem (3.23) is equivalent to the original bilevel problem (3.14) in the optimistic case, whether the lower level problem is convex or not. However, due to the infimum (or minimum) definition of $h(x)$ in (3.22), the single level problem (3.23) becomes a non-convex and non-differentiable

problem. A Gauss–Newton-type method is proposed for these type of reformulation in [22].

3.2.5 Solution Algorithms

Once the bilevel problem (3.14) is formulated into a single level form, one may apply some of existing type of solution methods to solve the equivalent single level problem. Some of the well known methods for nonlinear bilevel problems are described below.

3.2.6 Descent Method

A descent direction in bilevel optimization leads to decrease in upper level function value while keeping the new point feasible [9]. Considering that a point is viewed as feasible provided that it is lower level optimal, finding the descent direction can be very difficult. With the steepest descent method, the most important elements include the descent direction and the step size (or how far to descend). The gradient descent method utilizes the steepest gradient in order to search for an optimum of any given function [16].

Let us consider an unconstrained single level optimization problem of a function $f(z) : \mathbb{R}^n \rightarrow \mathbb{R}$ where $f(z)$ is a convex function and also differentiable. For z^* to be a minimizer it is necessary that the gradient of f be equal to zero at this point. This point is searched for iteratively starting at an initial guess. So generally we move from z^k to z^{k+1} by choosing a direction Δ^k to move from and then setting

$$z^{k+1} = z^k + \alpha^k \Delta^k, \quad (3.24)$$

where α^k is the appropriate step-size. Then choosing

$$\Delta^k = -\Delta f(z^k) \quad (3.25)$$

means that we are moving in the direction of “steepest descent” at each iteration.

The steepest descent algorithm for a single level unconstrained problem as written by [9] is as follows

1. Given x^0 , set $k = 0$
2. $d^k = -\Delta f(x^k)$. If $d^k = 0$ then stop
3. Solve $\min_{\alpha} f(x^k + \alpha d^k)$ for the step size α^k perhaps chosen by an exact or inexact line search.
4. Set $x^{k+1} \leftarrow x^k + \alpha^k d^k$, $k \leftarrow k + 1$. Go to step 1

Note that, for a bi-level problem, the candidate solution x^{k+1} should satisfy the optimality of the lower level problem (or its KKT condition) at each step.

3.2.7 Trust Region Method

In trust-region methods, the algorithms approximate a certain region of the objective function with a model function [58]. A region around the current best solution is defined, while a

particular model is used to estimate the original objective function. If the approximation is good the region is expanded and if it is not good then it is contracted.

The essential elements of a trust region algorithm include a guide on computing the trust region trial step and another on choosing whether a trial step should be accepted. An iteration of a trust region algorithm has the following form as seen in [58], a trust region is available at the beginning. Then an approximate model is constructed, and it is solved within the trust region, giving a solution which is called the trial step. A merit function is chosen, which is used for updating the next trust region and for choosing the new iterate point [58]. Trust region methods can be used to solve unconstrained optimization problems.

However, the method suits to problems whose lower level functions are smooth and strictly convex that satisfy strict complementarity conditions.

3.2.8 Heuristic Methods

These are methods based on concepts and notions found in nature. Heuristics are rules we develop based on experiences from the past. They are tools that help one to make decisions. The goal is to explore the search space in order to find optimal solutions. A heuristic method cannot be guaranteed to produce an optimum, however it will likely yield a good solution [41].

Genetic Algorithms (GAs)

These are subcategories of evolutionary algorithms, it is inspired by the theory of evolution with features such as mutation, crossover and selection [41]. Genetic algorithms are commonly used to generate high quality solutions to optimization by relying on biologically inspired operators [1].

To solve an optimization problem using genetic algorithms the aim is to make the candidate solution referred to as phenotypes sometimes to become or evolve to be better solutions. From the population, random individuals are generated and thereafter put through iterations which modify them to test fitness. The algorithm will terminate if a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population [9].

According to [9] general framework for a genetic algorithm can be described as follows

1. $k = 0$
2. Initialize $\Pi(k)$ (population at generation k)
3. Evaluate $\Pi(k)$
4. While termination condition not satisfied, do
 - (a) Begin
 - (b) $k \leftarrow k + 1$
 - (c) Select $\Pi(k)$
 - (d) Recombine $\Pi(k)$

(e) End

One of the limitations of genetic algorithms is that in many problems, GAs have a tendency to converge towards local optima or even arbitrary points rather than the global optimum of the problem [29]. On the other hand, bi-level problems are generally non-convex problems with possibly many local optimal solutions. This therefore means that it does not "know how" to sacrifice short-term fitness to gain longer-term fitness [29].

Simulated annealing

This is a probabilistic technique for approximating the global optimum of a given function, that is, it can give a reasonable approximation for a function with a large search space [41]. Similarly according to [10] simulated annealing is a method proposed to find the global minimum of a cost function that may possess several local minima. It is often used when a search space is discrete. Let Ω be the solution space (set of all possible solutions). Let $f : \Omega \rightarrow \mathbb{R}$ be an objective function defined on the solution space. The goal is to find a global minimum $w^* \in \Omega$ such that $f(w) \geq f(w^*)$ for all $w \in \Omega$. The basic elements of simulated annealing as stated by [10] are as follows

1. A finite set S .
2. A real valued cost function J defined on S . Let $S^* \subset S$ be the set of global minima of the function J assumed to be the proper subset of S .
3. For each $i \in S$, a set $S(i) \subset S - \{i\}$ called the set of neighbours of i .
4. For every i , a collection of positive coefficients q_{ij} , $j \in S(i)$ such that $\sum_{j \in S(i)} q_{ij} = 1$. It is assumed that $j \in S(i)$ if and only if $i \in S(j)$.
5. A non increasing function $T : N \rightarrow [0, \infty]$, is called the cooling schedule. Here N is the set of positive integers and $T(t)$ is called the temperature at time t .
6. An initial state $x(0) \in S$.

Thereafter, its performance is based on whether $x(t)$ converges to the set S^* and also how fast the convergence takes place. These are explained further in [10]

However with simulated annealing [10] mentions that its limitation is that it lacks rigorous theoretical justification.

Particle Swarm Optimization

This is a high level heuristic algorithm that is based on the concept of swarm intelligence capable of solving complex mathematical problems [11]. It is appropriate in solving non linear continuous function. According to [11] PSO is based on the study of the behaviour of a flock of birds and that of fish schooling. Each solution is considered as a bird and is called a particle. The particles have a fitness value which is calculated through the objective function. Since this algorithm is based on the birds being a flock, each particle is not only expected to be preservative of their individual best performance, but to also be aware of their best performance as a group.

Hence adjust their velocity considering their best performance and also considering the best performance of the best particle.

The general particle swarm algorithm as shown in [11] is as follows,

- i Initialize particle
- ii Evaluate fitness of each particle
- iii Modify velocity based on previous on previous best and global best positions
- iv Repeat until stopping criterion is satisfied
- v Terminate criteria.

The particle swarm algorithm formula as written by [11] is as follows,

$$V_{i+1} = wV_i + c_1r_1(p_i - x_i) + c_2r_2(p_i - x_i) \quad (3.26)$$

$$x_{i+1} = x_i + V_{i+1} \quad (3.27)$$

The formula (3.26) is called the velocity change of PSO where c_1 and c_2 in (3.26) are the learning factor or learning rate and w is the inertia weight. This formula has three components, the momentum the learning and the social components [26]. The second formula (3.27) represents the position of the particles.

3.2.9 Systematic Sampling Method

This is a probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed periodic interval. Suppose a sample size n is desired from a population of size $N = nk$, also suppose that the N units in the population are numbered 1 to N in some order [50].

To draw a sample size n , [50]

- i Select a random number between 1 and k .
- ii Suppose that number is i .
- iii Select the first unit whose serial number is i .
- iv Select every k^{th} unit after i .
- v The sample will contain $i, i + k, i + 2k, \dots, i + (n - 1)k$ serial number units.

So, the first unit is chosen at random while other units are selected systematically. This systematic sample is called the k^{th} systematic sample and k is termed as a sampling interval.

The advantages of systematic sampling is that it is easier to draw a sample and often easier to execute without mistakes. Further it is less costly and the selection of units is simple [50].

4. Resource Allocation Through Cost Effectiveness Analysis and Equity model

In Botswana there are ten districts and within these districts are different programs that are intended to decrease the number of new HIV infections. For the three allocation methods to be studied some general assumptions will be made. Firstly, it is assumed that there is no migration between the districts. Furthermore it is assumed that there are only three districts in Botswana. Since the allocation of funds for HIV prevention is done through different levels of decision making, an assumption that there are two levels of decision making being the upper and lower level is made. On the lower level, decisions on allocation of resources need to be made at district level. These districts decide on how to best distribute the funds among their available HIV preventive programs that suit the respective community. The upper level however decides how to allocate funds to the different districts. The upper level is taken as the national funding. Below is a picture that depicts the information above.

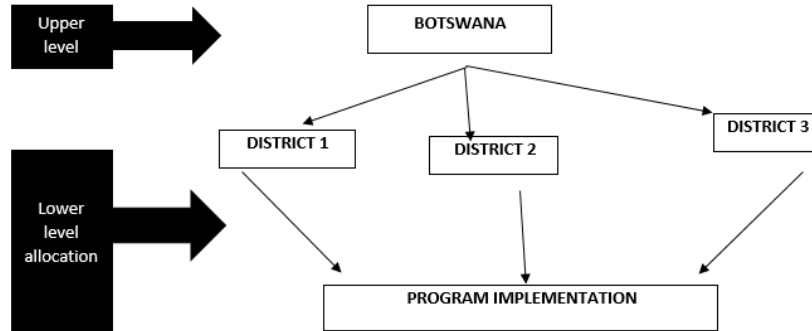


Figure 4.1: diagram representing resource allocation

A mathematical model is developed to determine the outcome of the proposed allocation of funds. The model diagram is illustrated in Figure 4.2.

The populations are assigned to different compartments as seen above, that is, S, I, T, V and A . The compartment S is for the susceptible population while I, T, V represents the infected, those undergoing treatment and infected children respectfully. Finally the A compartment is for the population that have AIDS.

The parameters represent infectious rate, λ , natural death rate, μ , recruitment to getting ARVs (treatment), ρ , mother to child transmission, m . There is also γ which is for the rate of recruitment for children to receive ART, β being the transmission probability ϵ_1 for rate of disease transmission due to ART and ϵ_2 for rate of MTCT due to treatment. Lastly the parameters δ_V and δ_A are for the mortality of infected children and for additional death rate due to AIDS respectfully.

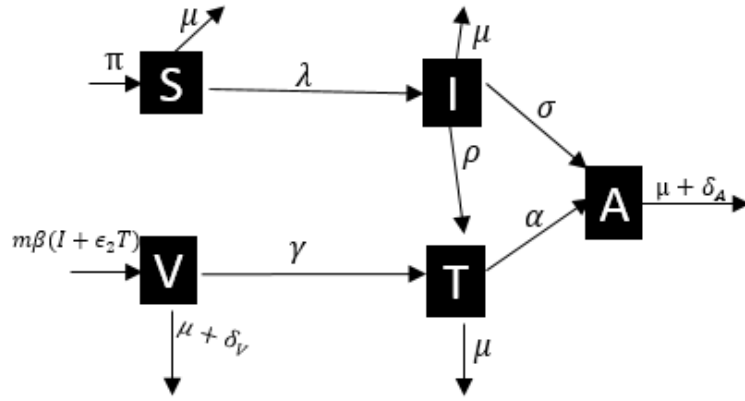


Figure 4.2: The mathematical model

The corresponding differential equations are

$$\begin{aligned}
 \frac{dV}{dt} &= m\beta(I + \varepsilon_2 T) - (\rho\gamma + \mu + \delta_V)V \\
 \frac{dS}{dt} &= \pi - \lambda \frac{S(I + \varepsilon_1 T)}{Q} - \mu S \\
 \frac{dI}{dt} &= \lambda \frac{S(I + \varepsilon_1 T)}{Q} - (\rho + \mu + \sigma)I \\
 \frac{dT}{dt} &= \rho(I + \gamma V) - (\mu + \alpha)T \\
 \frac{dA}{dt} &= \alpha T + \sigma I - (\mu + \delta_A)A
 \end{aligned} \tag{4.1}$$

where $Q_j(t) = S_j(t) + I_j(t) + T_j(t)$ represents the total proportion of the sexually active population in district j . The allocation of resources is to influence some of the parameters in the model system (4.1). As it is described in [31] the model system can be rewritten using the production functions that corresponds to these rates (parameters) as follows,

$$\begin{aligned}
 \frac{dV}{dt} &= \theta_m(v^2)\beta(I + \varepsilon_2 T) - (\theta_\rho(v^3)\gamma + \mu + \delta_V)V \\
 \frac{dS}{dt} &= \pi - \frac{\theta_\lambda(v^1)S(I + \varepsilon_1 T)}{Q} - \mu S \\
 \frac{dI}{dt} &= \frac{\theta_\lambda(v^1)S(I + \varepsilon_1 T)}{Q} - (\theta_\rho(v^3) + \sigma + \mu)I \\
 \frac{dT}{dt} &= \theta_\rho(v^3)(I + \gamma V) - (\theta_\alpha(v^4) + \mu)T \\
 \frac{dA}{dt} &= \theta_\alpha(v^4)T + \sigma I - (\mu + \delta_A)A
 \end{aligned} \tag{4.2}$$

The information above will be used for all the three resource allocation methods for HIV prevention in Botswana.

4.1 Equity allocation model

Equity model is a way of allocating HIV resources or funds. According to [46] Equity is a critical and controversial factor when deciding how to allocate public resources. Equity may be taken as synonymous with fairness, however, [46] elaborates on the definition of equity so that different models of equity may be carefully compared, contrasted, and analysed. It uses a ratio of the number of all HIV cases in the program implementing area over the total number of HIV cases in the decision area. Basically, with this model, for funds to be allocated this ratio is applied. Equity in resource distribution requires that individuals with similar needs ought to have access to the same resources (horizontal equity) and that the greater the need for an individual then subsequently that individual will get more resources to address the need (vertical equity) [2].

The need for equalizing resources when dealing with the equity model has been established. Hence according to [2] equity is categorized by three elements, the first being, identifying the resources to be allocated, secondly, specifying who the resources will be allocated to and lastly, over which time period. They also go on to highlight that for equity to be applied, the period for which equity is applied should be observed and integrated into the scope.

Love-Koh in [38] gives an outline of the significant tools that have been created to measure, assess and advance health equity. Their suggestion is that benefit incidence analysis can be utilized to approximate the allocation of public health resources. Benefits in this context is defined as the financial value of the services used.

On the other hand, [34] defines an equity model where they assume equity to be relative or proportional to the quantity of HIV infections. Under this model, funds are allocated to regions proportionally based on their share of the total infected population in each respective region.

Zaric and Brandeau in [59] considered proportional or equity method where at the upper level, the proportional allocation method distributes to regions proportional to HIV incidence in each region while at the lower level this method allocates funds to each program in proportion to HIV incidence in the corresponding risk groups.

4.2 Solution Procedure for Equity

To integrate this method in Botswana's HIV prevention resource allocation we assume that the Central district has the highest HIV prevalence followed by North East and lastly South East. Central has an HIV prevalence of 22.6 and the north east 21 while the south east has 16.8 according to [53]. It is also assumed that for both levels, allocation through equity method will be applied. For each risk group the group size, HIV prevalence and HIV incidence is estimated. Further, we assume that the budget is 1.6 million pula and that all of it is to be allocated proportionally. This therefore means that the upper level decision makers will allocate 37.4% of the budget to the central district and 34.8% , 27.8% for the north east and south east respectively. The values in Table 4.1 illustrate the information provided above.

For the lower level where there are two community groups under each district, being the high

	CD	NE	SE
HIV prevalence ratio	22.6	21	16.8
Percentage (%)	37.4	34.8	27.8
BWP	1683000000	1566000000	1251000000

Table 4.1: Upper level equity allocation

risk (HR) and low risk (LR) group, it is assumed that for the budget intended for each district 60% will go towards the high risk while the low risk group will be allocated 40% of the budget. In this level the decision makers in each district decide on how to best allocate funds among the programs. Since this method of allocation is based on proportionality, it is assumed that the rates associated with each program will be the guide on how to allocate the HIV prevention resources. These rates as previously stated include the contact rate λ , the mother to child transmission rate m and the recruitment to getting ART treatment rate ρ and lastly the rate of defaulting in the use of ART α . It is also assumed that these rates are the same in each district, $\lambda = 0.17$, $m = 0.8$, $\rho = 0.019$ and $\alpha = 0.011$. Tables 4.2 and 4.3 represent the information above.

Districts	λ	m	ρ	α
CD(239 360 000)	40 691 200	191 488 000	4 547 840	2 632 960
NE(222 720 000)	37 862 400	178 176 000	4 231 680	2 449 920
SE(177 920 000)	30 246 000	142 336 000	3 380 480	1 957 120

Table 4.2: Lower level equity allocation-Low risk community

Districts	λ	m	ρ	α
CD(359 040 000)	60 526 800	287 232 000	6 821 760	3 949 440
NE(334 080 000)	56 793 600	267 264 000	6 347 520	3 674 880
SE(266 880 000)	45 369 600	213 504 000	5 070 720	2 935 680

Table 4.3: Lower level equity allocation-High Risk community

4.3 Cost Effectiveness Analysis

Cost effectiveness is a type of economic analysis that is designed to assess the cost and the benefit or outcome of a health intervention [24]. The benefits are thereafter measured using a ratio. This proportion or ratio according to [24] is as thus;

$$\frac{\sum \text{All-benefits}}{\sum \text{All-costs}}, \quad (4.3)$$

where the benefits are the health outcomes of the program and the cost is the total cost of everything that has to do the program.

Also, [13] defines a cost-effectiveness ratio as the net cost divided by changes in health outcomes. An examples of this definition can be the cost per case of disease prevented or the cost per death averted. To calculate the net cost [13], the cost per implementation of intervention and cost of

aversion (treatment or whatever it takes to avert infection or disease) were added while changes in health outcome is the number of cases averted.

According to [28] in cost effectiveness studies, there are different model types that are used, these comprise of decision trees, deterministic and Markov models, dynamic and static models and individual-based and population based models. Decision trees and Markov models are known for their capability to firstly, monitor specific clinical incidents or events and secondly tracking the resources linked with those events. For these two reasons, cost-effectiveness modelling studies is commonly used.

Cost-effectiveness modelling determines if an intervention is of good value [28]. It provides a comparison between alternative interventions. That is, each HIV prevention intervention will be characterised by the average cost of each program and the expected reduction of HIV incidence among the program participants. Cost-effectiveness analyses evaluate a health intervention by asking, “How much health benefit do we get for our money?” [28]. Most countries in the Sub Saharan Africa use the cost effectiveness model to allocate HIV prevention resources [28]. To work on the productivity of programming for HIV prevention, optimizing limited financial funds and resources is a critical element needed to increase top calibre, cost-effective interventions so that HIV prevention is maximized. Determining the Cost-effectiveness determines whether the HIV prevention interventions are ideal for efficient resource allocation and for identifying incompetent preventive programs [48].

Outcome measures are then presented as number of HIV infections averted (HIA) for a specific HIV prevention intervention by reporting quality-adjusted life years (QALYs) gained or disability-adjusted life years (DALYs) averted. That is, the cost-effectiveness ratio (CER) is most useful when expressed as dollars per life-year gained (with appropriate adjustments for quality of life, usually expressed as QALYs), this is according to [47]. In QALY it is supposed that 1 QALY (1 Year of Life * 1 Utility = 1 QALY) is equivalent to a year of life lived in great wellbeing and that a year of life lived in a condition below this outstanding wellbeing is worth less than 1 QALY [47]. The CER outcomes can be weighed looking at the quality of life gained or through the Disability Adjusted life years (DALY). DALYs are a gauge of something lost as opposed something acquired or can be defined also as the amount of years of life lost prematurely due to disease or life lived with disability because of sickness. According to [23] the impact of interventions on DALYs (i.e. the DALYs averted by an intervention) is measured by calculating DALYs in two scenarios: with and without the intervention.

Semu Kassa in [31] states that Allocation by Cost-Effectiveness (ACE) does not consider a few significant factors for example, expanding or diminishing marginal returns to scale, mutual exclusivity of programmes, and interaction of program outcomes, which should be urgently considered. This leaves a gap for a more efficient model that will take these factors into consideration.

4.4 Solution Procedure for Cost effectiveness analysis

To implement the CEA one would need to firstly start by asking necessary or important questions such as, the number of HIV infected individuals, what programs are put in place to prevent HIV,

how much is spent on these programs and how much does the community benefit from these programs. Secondly, one would then develop a model based on key points covered on questions asked from there, data should be entered to the developed model. This data is answered questions or collected information. Lastly results. Results are calculated from the provided data, this can be done by plotting cost vs QALY and finally analysing these results.

Another way to implement the CEA is to define the set of possible choices of interventions. Secondly, identify the outcomes of the interventions. These outcomes should be as broad as possible from the consequences of different times at which the intervention is applied to how it affects different sectors of the society. From there one should proceed to choose a subset that makes sense, that is, one should narrow their scope of outcomes. Next one should identify the relevant constrained resources (expressed in terms of dollars). Furthermore, one will need to identify the benefits and resource needs(costs)for each intervention of choice. This can be done by creating a decision tree, assigning probabilities to events and values to outcomes and then calculating the expected value of each alternative.

Since the interventions for HIV prevention have already been mentioned now we proceed to finding out their estimated or calculated costs in Botswana. First we start with the cost reducing the number of new infections which in it includes testing, provision of condoms, circumcision and behaviour change programs among other things. Their total estimated cost will fall within the range of \$56 – \$112.60. Furthermore the cost of treatment which includes ARV drugs, coupled with laboratory tests and human resource, then according to [17], the cost per person per year becomes \$283. For the mother to child transmission the cost is calculated using the treatment cost plus the frequent clinic or hospital visits that the expecting mother needs to make and according to [60] the estimated cost becomes \$319 per person per year. While for those that default on taking the treatment as they should it is assumed that they eventually get AIDS and are subject to hospital stays and more human resource services as they may need to get home visits. Therefore the estimated cost of caring for an AIDS patient are \$300.

Now to identify the benefits that these interventions have in the prevention of HIV/AIDS in Botswana, statistics from the year 2018 and 2017 and 2016 were used. Ultimately these benefits should see to the number of new infections being reduced. In 2018, 370000 people were living with HIV, and in that year 8500 new infections were recorded, with 4800 AIDS related deaths and only 84% of the infected people taking treatment, this being according to [21]. Furthermore the estimated number of children living with HIV was 13 500. While according to [54] in 2016 the number of people living with HIV in Botswana was 360 000 and 83% of these people were taking treatment. However there were 3900 AIDS related deaths with the number of new infections being 10000 and the estimated number of children living with HIV was 12000. The investment made on the years 2015/2016 according to [56] was BWP 1 476 000 000. Also according to [56] the budget for both the years 2017/2018 was BWP 1 571 000 000.

The formula used to calculate the cost effectiveness ratio in Table 4.4 is

$$\frac{\text{net cost}}{\text{changes in health outcomes}}. \quad (4.4)$$

Program	Total cost (BWP)	Total Benefit	CER
Contact rate	84.30	0.48	175.63
Treatment	283	0.32	884.38
MTCT	319	0.01	31900
Defaulting in the use of ART	300	0.22	1363

Table 4.4: Cost effectiveness analysis per person(BWP)

Using the assumed budget total of *BWP*1 600 000 000, the allocation towards the districts which is done by the decision makers in the upper level will be done using equity while the allocation in districts will be done using cost effectiveness. Therefore for upper level allocation a reference to (4.1) should be made. For the lower level since there are two community groups for each district, being the low risk and high risk community, it is assumed that the high risk community will get 60% and the low risk group 40%. With this given, the value for the v_j^l (where j represents the district and l the respective programs) will be calculated as follows.

$$v_j^l = D_{ji} * \frac{r_l}{\sum ratio}, \quad (4.5)$$

where r represents the ratio for each program and i represents the community groups. Therefore the v_j^l will be calculated as in Tables 4.5 and 4.6.

Districts	λ	m	ρ	α
CD(239 360 000)	1 224 799.25	6 167 442.68	222 462 540	9 505 217.64
NE(222 720 000)	1 139 652.78	5 738 689.98	200 997 230	8 844 427.11
SE(177 920 000)	910 412.27	4 584 355.79	165 359 856	7 065 375.68

Table 4.5: Lower level CEA allocation-Low Risk community

Districts	λ	m	ρ	α
CD(359 040 000)	1 837 198.87	9 251 164.02	333 693 811	14 257 826.50
NE(334 080 000)	1 709 479.16	8 608 034.97	310 495 845	13 266 640.70
SE(266 880 000)	1 365 618.41	6 876 533.68	248 039 784	10 598 063.50

Table 4.6: Lower level CEA allocation-High Risk community

5. Allocation through Bi-level Optimization Procedure

The bi-level optimization technique also consists of two levels of decision making namely the upper level (leader) and the lower level (follower). The two levels are thoroughly explained below.

5.1 Lower level resource allocation

This is where HIV preventive programs are to be implemented. It is assumed that the programs address the main dynamics, reducing or minimizing the rate of infection λ , and of MTCT m , increasing the rate of recruitment ρ and decreasing the rate of failure in committing to taking treatment (ARV) properly α . These are the main incentives for HIV control. In the content of HIV epidemic control, production functions are intended to convert dollar amounts invested into changes in the sufficient contact rate λ and m [34]. The functions help to derive the cost needed for new infections in a particular population at time t .

It is assumed that the objective here is to minimize the number of new infections given a total budget constraint at time t [37]. The new infections are from the newly born children and also from sexually active individuals. *It is also assumed that the effect of other transmission possibilities such as sharing a sharp object with an infected person are negligible.* Hence the cumulative number become,

$$\begin{aligned}
 f_j = & \sum_{j=1}^n \int_0^T e^{-rt} \left\{ \theta_{m_j}(v_j^2) \beta [I_j(t) + \varepsilon_2 T_j(t)] \right. \\
 & \left. + \theta_{\lambda_j}(v_j^1) S_j(t) \left(\frac{I_j(t) + \varepsilon_1 T_j(t)}{Q_j(t)} \right) \right\} dt,
 \end{aligned} \tag{5.1}$$

where the variable v^2 is a decision variable representing the amount reducing m and v^1 represents the investment spent on reducing λ while ρ and α are investments represented by v^3 and v^4 respectively and j is for the districts. Since treatment correlates positively with the reduction of contact rate it should hence be incorporated in the equation. Also according to [31] previous investments on ρ and α have an impact on the number of the individuals in the treated cohort and due to reduced infectiousness of these individuals their contribution to the emergence of new infections is significant at the current time. The value of T is therefore approximated using the equation (4.3) for a reasonably small change in time $\delta(t)$ at any given time t as follows,

$$T(t+\delta t) = T(t) + [\theta_\rho(v^3)(I(t) + \gamma V(t)) - (\theta_\alpha(v^4) + \mu)T(t)]\delta t \tag{5.2}$$

$$= [\theta_\rho(v^3)(I(t) + \gamma V(t))]\delta t - [(\theta_\alpha(v^4) + \mu)\delta t - 1]T(t) \tag{5.3}$$

By making T the subject we get,

$$T(t) = [\theta_\rho(z)(I(t - \delta t) + \gamma V(t - \delta t))]\delta t - [(\theta_\alpha(w) + \mu)\delta t - 1]T(t - \delta t), \quad (5.4)$$

where δt represents the unit change in time t hence $t - \delta t$ is the time before the current time t .

Now the resulting objective function from 5.1 becomes

$$\begin{aligned} f_j(\mathbf{v}_j) = & \sum_j \int_0^T e^{-rt} \left\{ [(\theta_{m_j}(v_j^2)\beta(I_j(t) + \varepsilon_2[\theta_\rho(v_j^3)(I_j(t - \delta t) + \gamma V_j(t - \delta t))])]\delta t \right. \\ & - [(\theta_{\alpha_j}(v_j^4) + \mu_j)\delta t - 1]T_j(t - \delta t)) \\ & + \theta_{\lambda_j}(v_j^1) \frac{S_j(t)}{Q_j(t)} I_j(t) \{ \varepsilon_1[\theta_{\rho_j}(v_j^3)(I_j(t - \delta t) + \gamma V_j(t - \delta t))]\delta t \} \\ & \left. - [(\theta_{\alpha_j}(v_j^4) + \mu_j)\delta t - 1]T(t - \delta t)] dt \right\}. \end{aligned} \quad (5.5)$$

Let B_j be the budget constraint allocated to a district j and $N(t)$ be the total population of the country at time t making $N_j(t)$ is the total population of a district at time t . The variables $(v^1, v^2, v^3, v^4) = \mathbf{v}_j^l$ are the decision variables representing the amount invested in the respective interventions in district j . These decision variables are constrained by the maximum budget B_j and the assumption that they are non negative.

Hence the lower level allocation model for a district j is as follows

$$\begin{aligned} \min_{v_j} \quad & f_j(y, v) \\ \text{s.t} \quad & \sum_{j=1}^3 (v_j^1 + v_j^2 + v_j^3 + v_j^4) \leq B_j \\ & B_j \leq y_j \text{ for all } j = 1 : 3 \\ & v_j^l \geq 0, \quad j = 1 : 3; \quad l = 1 : 4 \end{aligned}$$

5.2 Upper level resource allocation

The upper level problem will be to allocate funds $B_1, B_2, \dots, B_n$ optimally to districts in order to minimize the cumulative number of new infections in the country. Therefore the upper level

allocation problem will be as follows

$$\begin{aligned}
 \min_{B_j} \quad & F(y, v) = h_1 y_1 + h_2 y_2 + h_3 y_3 \\
 \text{s.t} \quad & y_1 + y_2 + y_3 = \mathbf{B} \\
 & \sum_{j=1}^3 (B_j) \leq \mathbf{B} \\
 & B_j \geq d_j y_j \\
 & v_{j1}^l + v_{j2}^l \geq z_{ji}^l y_j \text{ for all } j = 1, 2, 3. \ l = 1, 2, 3, 4. \\
 & y \geq 0,
 \end{aligned}$$

$\mathbf{B}$ is the total amount available to fund districts for their HIV prevention programs and B_j represents the amounts allocated to the regions while h_j is the estimated number of new infections for district j over a particular time period for each money allocated to that district. The values of h_j can be estimated using the formulation in [31] which is done by choosing appropriate weights w_1 and w_2 , to get

$$h_j = w_1 \frac{r_j^1 \times e_j^1}{N_j \times c_j^1} + w_2 \frac{r_j^2 \times e_j^2}{N_j \times c_j^2}, \quad (5.6)$$

where,

- c_j^1 = an average estimated cost per person to implement one intervention aiming to avert new infections in district j .
- c_j^2 = an average estimated cost per person to implement one intervention aiming to avert new progression to AIDS in district j .
- r_j^1 = the baseline number of new infections that will occur in district j over time τ
- r_j^2 = the baseline number of new AIDS cases that will progress in district j over time τ .
- e_j^1 = total number of expected potential AIDS cases in district j
- N_j = The size of population to be addressed during the intervention in district j .

5.3 Formulation of the bi-level programming problem

When we combine all the two problems described above in sequential order the general resource allocation problem becomes,

$$\begin{aligned}
& \min_{B_j} F(y, v) = h_1 y_1 + h_2 y_2 + h_n y_3 \\
& \text{s.t. } y_1 + y_2 + y_3 = \mathbf{B} \\
& \quad \sum_{j=1}^3 (B_j) \leq \mathbf{B} \\
& \quad B_j \geq d_j y_j \\
& \quad v_{j1}^l + v_{j2}^l \geq z_j^l y_j \text{ for all } j = 1, 2, 3. \ l = 1, 2, 3, 4. \\
& \quad y \geq 0, \\
& \quad \text{where } v \text{ solves,} \\
& \min_{v_j} f_j(y, v) \\
& \text{s.t. } \sum_{j=1}^3 (v_j^1 + v_j^2 + v_j^3 + v_j^4) \leq B_j \\
& \quad B_j \leq y_j \text{ for all } j = 1 : 3 \\
& \quad v_{ji}^l \geq 0, \ i = 1 : 2, j = 1 : 3, l = 1 : 4
\end{aligned} \tag{5.7}$$

The total budget $\mathbf{B}$ is allocated nationally to put in place the country's interventions. The upper level decision makers decides on the optimal distribution of the budget B_j for each district j . The lower level decision makers solves its problem to obtain an optimal vector $\mathbf{v}_j(B_j)$, which is dependent on the choice of the upper level decision maker. Since the lower level decision makers must wait for the upper level decision makers to make a choice before they can allocate their funds optimally this decision process is called a hierarchical structure.

To arrive at the optimal allocation we apply the solution techniques for multilevel programming problems. These techniques however depend on the production functions for investment on each HIV preventive program. According to [34] production functions can be defined as the level of outputs that can be achieved from inputs amounts, they are intended to convert dollar amounts invested into changes in the infectious rate, mother to child transmission rate, etc. That is, They help to derive the number of new infections that occur in a particular population for a given level of investment. In [31] it is highlighted that it is it is advisable to use a combination of models depending on the nature of interventions to formulate corresponding production functions. In some of the intervention programs each incremental unit of money spent may produce the same incremental reduction or upsurge in epidemic component. In such type of programs the production function could be taken to be linear. In our model, for example, to formulate the proportion of investment in increasing the rate of recruitment for treatment, ρ we may use $\theta_{\rho_i} = a_i + b_i y$ where a_i and b_i are constants and in the same way the production

function for modelling the proportion of investment in decreasing the rate m of MTCT can also be formulated as $\theta_{m_i} = \sigma_i - \beta_i y$ where σ_i and β_i are constants.

However in some programs which deal with behavioural change in communities an increase in investment may not necessarily result in an increase in better behaviour but may rather worsen the behaviour of the community, that is the increase in investment in such programs is not guranteed to always be directly proportional to aspired results or impact. Therefore according to [31] In such cases we may model the corresponding production function by a function which is convex with respect to the investment variables. These interventions include the investment in reducing the rate of infection due to unsafe sexual contacts λ and the rate of failure in adherence to proper treatment of ARV α .

In [31] these were modelled as follows,

$$\begin{aligned}\theta_{\lambda_i}(y) &= \lambda(1 + e^{-(cy)}) \\ \theta_{\alpha_i}(y) &= \alpha(1 + e^{-(cy)}).\end{aligned}\tag{5.8}$$

5.4 Solution procedures for bi-level optimization

To solve the bi-level resource allocation problem some assumptions are made for simplicity. It is assumed that there are three districts in Botswana and that the total budget available for HIV prevention measures is $\mathbf{B} = BWP1.6$ billion. This budget is therefore to be allocated as y_1, y_2, y_3 among the three districts respectively and hence $y_1 + y_2 + y_3 \leq \mathbf{B}$.

Again with each district it is assumed that there are two community(or risk) groups to which a budget v_{j1} and v_{j2} is to be allocated to fight the epidemic ($j = 1, 2, 3$). In total it is therefore assumed that there are six community groups in Botswana. Each allocated budget for a single community group has four components $(v_{ji}^1, v_{ji}^2, v_{ji}^3, v_{ji}^4) = v_{ji}^l$ corresponding to each component categories.

The lower level must determine the total investment corresponding to the four categories in each community group $(1, 2)$ for each district j .

Assuming that each of the districts act independently, then for each district the lower level problem will be solved and combined in the upper level. Another assumption is that

$$h_j = \begin{pmatrix} 0.39 \\ 0.32 \\ 0.29 \end{pmatrix}, \quad d_j = \begin{pmatrix} 20\% \\ 40\% \\ 25\% \end{pmatrix}\tag{5.9}$$

where the rows represent the district and the column the country Botswana, d_j is the minimum percentage of the country's budget for the district j . Furthermore it is assumed that the minimum percentage of the total allocation for any one of the interventions categories is $z_j^l = 7.5\%$ of the district budget.

Then the leader's problem will be

$$\begin{aligned}
 \min_B \quad & F = [0.2y_1] + [0.31y_2] + [0.25y_3] \\
 \text{s.t} \quad & B_1 + B_2 + B_3 \leq \mathbf{B} \\
 & y_1 + y_2 + y_3 = \mathbf{B} \\
 & v_j \geq d_j y_j \text{ for all } j = 1, 2, 3 \\
 & \sum_{l=1}^4 [v_{j1}^l + v_{j2}^l] \geq z_j^l y_j \text{ for all } j = 1, 2, 3 \text{ and } l = 1, \dots, 4 \text{ } i = 1, 2.
 \end{aligned}$$

In the lower level problems there are various parameters to be used. Some of these are disease parameters that remain constant globally while others are country and district specific, these include the birth and death rate as taken from [39] while the rest of the parameter values are taken from [31] These are shown in Tables 5.1 and 5.2.

Parameter	Description
$\delta_V = 0.15$	mortality rate for infected children
$\delta_A = 0.2$	additional death rate due to AIDS
$\sigma = 0.1$	rate of progression to AIDS if not treated
$\gamma = 1.3$	preferential rate of recruitment for children to receive ART
$\beta = 0.12$	transmission probability
$\epsilon_1 = 0.08$	factor of reduction on rate of disease transmission due to ART
$\epsilon_2 = 0.167$	factor of reduction on rate of MTCT due to treatment

Table 5.1: Disease parameters assumed to be constant

Parameter	Description(each is for the three districts)
$\pi = [0.0098, 0.0078, 0.0065]$	Birth rate for the districts
$\mu = [0.0032, 0.0015, 0.0011]$	Death rate for the respective districts
$\lambda_H = [0.16, 0.17, 0.13]$	Initial unsafe contact rates for high risk groups
$\lambda_L = [0.11, 0.14, 0.09]$	Initial unsafe contact rates for low risk groups
$m_H = [0.31, 0.37, 0.22]$	Initial rate of MTCT for high risk groups
$m_L = [0.19, 0.22, 0.15]$	Initial rate of MTCT for high low groups
$\alpha_A = [0.15, 0.32, 0.12]$	Initial rate of defaulting in the use of ART
$\rho_H = [0.15, 0.15, 0.20]$	Initial rate of recruitment for ART in High risk groups
$\rho_L = [0.075, 0.075, 0.10]$	Initial rate of recruitment for ART in High risk groups

Table 5.2: Parameters assumed to vary from district to district.

The initial size of the risk population is taken from [40] three districts were picked at random, these being the central, North East and South East districts in Botswana. The other initial population groups have been reasonably estimated.

Population size	D1	D2	D3
$S(0)$	576064	316606	159225
$V(0)$	24360	20010	13075
$I(0)$	67524	59787	85817
$T(0)$	55 369	49 025	70 369
$A(0)$	20121	21600	18071

Table 5.3: Initial risk population

The production functions for the model are chosen as taken from [31] to be,

$$\begin{aligned}
 \theta_{m_j}(v_j^2) &= m(1 - a_2 v_j^2) \\
 \theta_{\rho_j}(v_j^3) &= \rho(1 + a_1 v_j^3) \\
 \theta_{\lambda_j}(v_j^1) &= \frac{1}{2}\lambda(1 + e^{-(c_1 v_j^1)}) \\
 \theta_{\alpha_j}(v_j^4) &= \frac{1}{2}\alpha(1 + e^{-(c_2 v_j^4)}),
 \end{aligned} \tag{5.10}$$

with $a_1 = 1.65 \times 10^{-8}$, $a_2 = 5.43 \times 10^{-8}$, $c_1 = 4.56 \times 10^{-1.55}$, $c_2 = 3.35 \times 10^{-2}$.

Then given the leader's variable B_j for district j and the investment variable v_{ji} for district j and community group i , the decision maker at each level optimizes its resource allocation problem for the desired epidemic control independently. Their costs add up to the total cost with the same weight and therefore the lower level optimization problem is assumed to be of the following form,

$$\begin{aligned}
 \min_v \quad & f_v = \sum_{j=1}^3 f_j(v_j) \\
 \text{s.t} \quad & (v_{j1}^1 + v_{j1}^2 + v_{j1}^3 + v_{j1}^4) + (v_{j2}^1 + v_{j2}^2 + v_{j2}^3 + v_{j2}^4) \leq B_j.
 \end{aligned}$$

5.5 Solution Approach

To solve the bi-level problem 5.7 a method called systematic sampling will be used. Systematic sampling is a procedure that is a derivative free meta heuristic algorithm comprised of realizing randomness and partitioning the leader's decision space. We also use the particle swarm optimization procedure to solve the lower level problem. As explained previously PSO is a computational method that gives the best solution for a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. This method is ideal since the bi-level problem is a dynamic system and is also non convex.

Firstly, the leader's decision space will be partitioned as in Table 5.4.

a_1^j	a_2^j	a_3^j	a_4^j	...	
---------	---------	---------	---------	-----	--

Table 5.4: Partitioning the decision space

For this partition to take place an assumption is made about the minimum amount that is to be allocated to each district as indicated above and likewise to the maximum amount as well. The minimum amount is assumed to be 7.5% of the budget $\mathbf{B}$ while the maximum amount given that the three districts is 85% for each district. From there in the first partition a_1^j an amount y_n is picked at random for each of the districts as the leader's action, the next points will then be selected using systematic sampling and emphasising a fixed interval. Secondly all the leader's representative actions from each partition are substituted into the follower's problem to obtain their corresponding optimal reactions. The follower after receiving the leader's representative action finds an optimal solution v_n^j of his own problem using or applying particle

swarm optimization(PSO). Once the reaction from the lower level is obtained, the values for each representative from each of the partitions (together with their corresponding reaction values) will be used to optimize the leader's objective. Furthermore, another set of points from these

Districts	λ	m	ρ	α
CD	620414.8999574	7016787.558969	232807297.5664	7225835.372894
NE	616039.7883554	4299641.331937	200568849.9683	5430078.590092
SE	1191937.654057	3760048.986512	170497857.7882	7939779.422819

Table 5.5: Lower level Bi-level optimization allocation-Low Risk community

partitions will be picked and assessed such that the point giving the best solution is stored. This process will repeat until stopping conditions are satisfied. The leader after obtaining the reaction for each of his or her representative chooses the best solution among the generated decision variables in terms of the leader's objective [25]. According to [25] since the current solution for a given iteration is obtained from a random choice of a representative from each district, it may not be the best solution for the problem. Therefore another set of leader's representative actions from each district will be selected again from the leader's decision space. However we have to avoid repetition of representative actions from any of the districts by further subdividing each partition into η subdivisions, where η is the maximum number of iterations. Finally for each leader's distinct action we obtain a global optimal reaction and we choose the

Districts	λ	m	ρ	α
CD	750005.2774157	8960465.892998	331857885.1059	8139061.653344
NE	1590358.822622	8202545.54276	305669317.1293	10905995.5218
SE	1567670.953512	6587867.98425	258542739.2618	9389489.845524

Table 5.6: Lower level Bi-level optimization allocation-High Risk community

best solution from among the previously selected best and any of the best solution from the current iteration based on the leader's objective.

Lastly, the optimal values (v_{ji}^l) calculated will be as given in Tables 5.5 and 5.6. The parametric

parameter variable names	parametric values
number of populations	100
Alpha	0.75
Epsilon	10^{-6}
c1 and c2 (learning factors)	2
w(weight of the inertia)	5
Maximum iterations	100

Table 5.7: Parametric values used for the PSO simulation

values that were used for the PSO are indicated in Table 5.7.

6. Results and Comparison

The decision process has been assumed to have two levels, being, the upper (leader) and lower (follower) level. It is through these two levels that allocation of HIV prevention resources is made. This allocation was made through equity, CEA and bi-level optimization. The criteria that is commonly used to gauge the performance of an allocation technique is keeping count of the number of new infections averted and also the quality of life gained. That is, the more the number of new infections averted through a particular method the more efficient that method is considered. Similarly, when the quality of life is improved for infected individuals then that method is seen to be very efficient.

In this study, the resultant number of new HIV infections and the HIV prevalence is used to compare the three resource allocation techniques. HIV prevalence is indirectly proportional to a better economy and better functioning community therefore using it for comparison purposes will lead to identification of a much efficient allocation method. That is, the lower the HIV prevalence then the better chance there is for a country's economy to perform better and vice versa. In the lower level the goal is to minimize the number of new infections in each district which will in turn see to the HIV prevalence rate being reduced. While in the upper level the goal is to minimize the costs associated with preventing HIV/AIDS through the programs indicated with the given budget of 1.6 billion and over a time period of 10 years. In the lower level, the objective function to be minimized is the function (5.6).

This function (5.6) is the cumulative number of new infections which incorporates different parameters, including those representing the main programs as listed and the differential equations as previously derived from (4.2).

The results of minimizing (5.6) are found in Figures 6.1, 6.2 and 6.3.

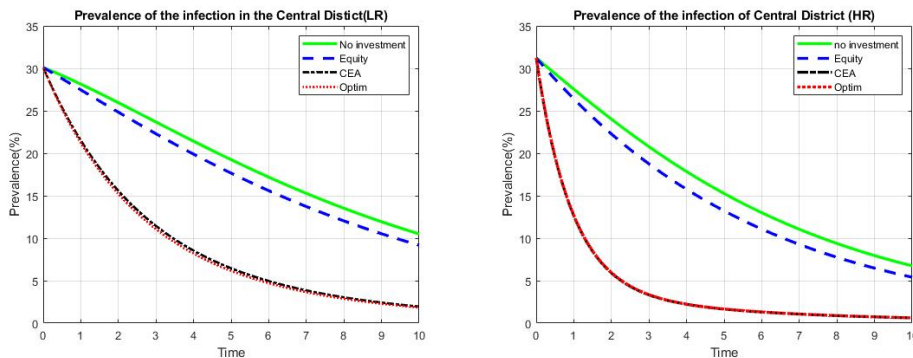


Figure 6.1: Allocation of HIV funds in the Central District through equity, CEA and Bilevel optimization.

The graphs in Figures 6.1, 6.2, and 6.3 are a representation of the impact that each technique has on the HIV prevalence over a period of 10 years. This impact is as a result of the optimal investment values as calculated and illustrated in tables [4.2, 4.3, 4.5, 4.6, 5.5, 5.6,] respectfully.

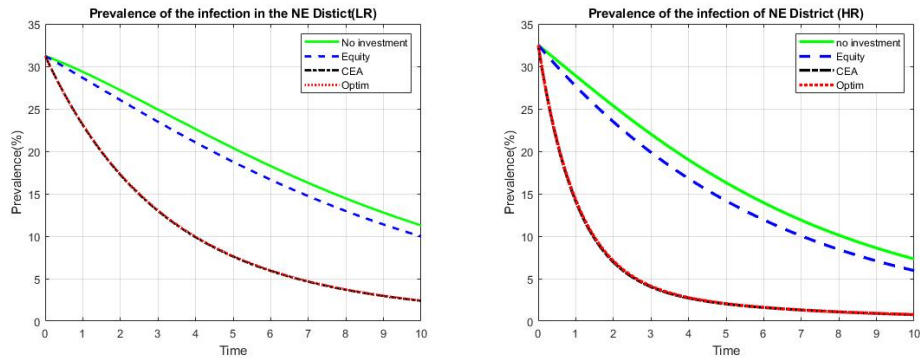


Figure 6.2: Allocation of HIV funds in the North East District through equity, CEA and Bilevel optimization.

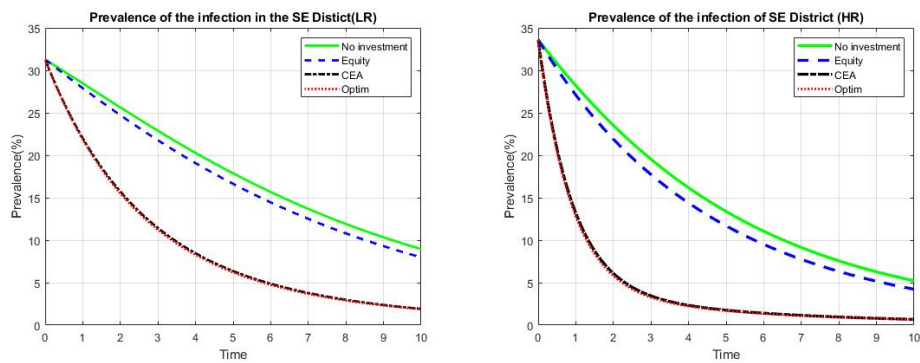


Figure 6.3: Allocation of HIV funds in the South East District through equity, CEA and Bilevel optimization.

HIV prevalence is the number of persons living with HIV in a defined population. These graphs have also indicated the rate of HIV prevalence when there is no investment. Not investing in the HIV prevention programs is shown to reduce the HIV prevalence at a much lower rate.

However allocating HIV prevention resources through the equity model is shown to give better results compared to when no investment is made. In spite of that, the equity model in every instance of Figures 6.1, 6.2 and 6.3 proves that it is desirably out performed by the CEA.

The CEA efficiently reduces the HIV prevalence by a very significant margin as clearly indicated in every district. The graphs are reflecting that the CEA and bi-level are somewhat performing at a similar rate. However looking at the investment proportions for all these methods it is apparent that while CEA and equity allocation utilize all the funds or the estimated budget, bi-level optimization achieves it's objective at significantly lower sum. That is, the total investment allocated to each district which has been labelled as y_j as seen in the leader's problem (5.10) is lower in bi-level optimization (see Table 6.1).

Allocation Techniques	y_1	y_2	y_3	Total
Equity	598 400 000	556 800 000	444 800 000	1 600 000 000
CEA	598 400 000	556 800 000	444 800 000	1 600 000 000
Bi-level Optimization	597 380 000	537 280 000	459 480 000	1 594 100 000

Table 6.1: DISTRICT INVESTMENT FOR HIV PREVENTIVE PROGRAMS

By using bi-level optimization the government would have saved a good 5.9 million as indicated in Table 6.1 which can be used for other purposes. This is a great save and also evidence that bi-level optimization is in this form a much more efficient method of allocating HIV prevention funds and resources.

7. Conclusion and Recommendations

In this paper different resource allocation techniques used for HIV prevention in Botswana were investigated, that is equity, CEA and bilevel optimization. Resource allocation is the process of assigning, planning and preparing project implementation or achieving goals. Allocation techniques are useful and resourceful because it is through them that the best possible outcome can be attained. Therefore for HIV prevention in Botswana we analyse these methods with their impacts on the number of new HIV infections. That is because new HIV infections has been found to impact the country's economic state. More new HIV infections mean reduced labour supply and productivity since there may be more HIV related deaths and hence loss of skilled personnel in key sectors. Further, more infections may mean constant sickness and as a result less productivity. Therefore it can be concluded that using the number of new infections as a measure of outcomes or effectiveness of the resource allocation methods is crucial.

In our scenario there are only two levels of decision making. It is also assumed that there are three districts in Botswana and each of them have two communities groups being the high risk and low risk group. It was further assumed that for allocation in the lower level, 60% of the budget particular to a district will be used on the low risk group while 40% will be allocated to the high risk group. With these allocation methods the goal is to find the v_j^l (where j is for the districts and l the programs) needed to minimize the function (5.6), that is, the investment needed for the HIV prevention programs that will minimize the number of new infections in the lower level. Upon minimizing (5.6) a comparison is done looking at the results.

For equity allocation method, HIV prevalence of the three districts as taken from [53] were used. These HIV prevalence ratios determined the percentage each district will have in getting the funds. This is clearly depicted by the table (4.1). Since there are two community groups with their respective ratio in terms of allocation. Other ratios which are used are those linked with the four HIV prevention programs.

For the CEA allocation technique, under each HIV preventive program relevant data for the years 2018, 2017 and 2016 were used. The differences which reflected health gain or lack thereof were put against the HIV prevention investments made for those years. With this, the health benefit particular to each program was calculated. Different literature was used to advice on the total cost particular to each program. Now with this the cost effectiveness ratio (4.4) was calculated. It was assumed that for this method the upper level uses equity for allocation and then cost effectiveness analysis for the lower level. It is on the lower level where the v_j^l s are calculated using the formula (4.5).

Finally for the bi-level optimization, the bi-level programming problem was formulated such that we got (5.7). In the upper level the goal was to minimize the cumulative number of new infections by optimizing the allocation of the HIV prevention funds. This was achieved by partitioning the leader's decision space, randomly picking a point or decision then selecting the next point using systematic sampling. While, for the lower level the goal is also to minimize the number of new infections in each district by optimizing the allocation of HIV prevention funds. Here the leader's action or decision is taken and through it the follower determines their best possible solution by applying the PSO method.

All these methods have proven to help in the reduction of the number of new infections. That is because where no money is allocated then there will be damage or a worsening of the HIV pandemic, however if allocation methods are applied there is immediate improvement.

From the table [6.1, 6.2, 6.3] it can be concluded that if bi-level optimization is used to allocate HIV prevention funds then a more significant impact or gain will be attained. It was also found that the CEA performs better than the equity allocation technique. These results and conclusion are only true in a model where there are three districts and two communities under each district.

With this study it is evident that the resource allocation techniques used for HIV prevention have a hefty impact on the direction and outcome of this pandemic. It is especially apparent that the bi-level optimization method is the most effective of the three allocation methods studied. It should therefore be given serious consideration.

A. Matlab Codes for Results

A.1 The main partitioning MatLab code

```
1 %% Systematic Evolutionary Algorithm %%
2 clc; clear ; clear var;
3
4 %
5 %% Problem Definition %%
6 %
7
8 % ObjFunU = @(x) ObjectiveFunctionU; % First level Decision maker
   objective function
9 % ConsFunU = @(x) myconstraintU;      % First level Decision maker
   constraint function
10 % ObjFunL = @(x) ObjectiveFunctionL; % Second level Decision maker
   objective function
11 % ConsFunL = @(x) myconstraintL;      % Second level Decision maker
   constraint function
12
13 BT = 1600000000;
14 MinVar_U = ((3/1000)*BT)*ones(1,2);
15 MaxVar_U = ((5/7)*BT)*ones(1,2);
16 noVar_U = 2; % Number of upper level
   variables , note y(3) = BT - (y(1)+y(2)).
17 VarSize_U = [1 noVar_U]; % Decision space
18 Range = MaxVar_U - MinVar_U; % Range of decision space(
   Length of upper problem decision space(?))
19
20
21 nvar_L = 24; % Number of variables in the
   lower level problem
22 MinVar_L = (0.031*(7/nvar_L)*BT)*ones(1,nvar_L); % Lower bound for
   the variables of the lower level problem
23 MaxVar_L = ((9/nvar_L)*BT)*ones(1,nvar_L); % Upper bound for the
   variables of the lower level problem
```

```

24 Npop = 100; % Number of population for
    the PSO % 50
25 Max_iter_PSO = 100; % Maximum iteration of the
    PSO % 50
26
27 %
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
28 %% Parameter Definition %%
29 %
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
30
31 %% Sampling Parameter for partitioning decision space at each
    Iteration
32 epsilon = 0.00; % Tolerance of a reaction
33 DELTA = 8000000000; % delta
34 a(1) = 2*ceil(Range(1)/(2*DELTA)); % Number of spies on first level
35 a(2) = 2*ceil(Range(2)/(2*DELTA)); % Number of spies on first level
36
37 b(1) = floor(a(1)/2); % Number of subintervals on first
    level
38 b(2) = floor(a(2)/2); % Number of subintervals on first
    level
39
40
41
42 MaxIter = 50; % Maximum number of iteration
43
44
45 Delta = DELTA/MaxIter; % Step size of the partition of
    the first subinterval
46 perm = randperm(MaxIter); % Random representative of the
    first subintervals
47 empty_Spies= [];
48 lambda = 10000000000000; % Penalty Factor
49
50
51
52
53
54 %

```

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

55 %% Main Iteration  %%
56 %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

57 for IT=1:MaxIter
58
59 %% Initialization
60 % Distribution of partition by first decision variables on first
    level
61
62 flsf(1) = MinVar_U(1)+(perm(IT)-1)*Delta+rand(1)*Delta;
                                % fls: first level partition
63 if flsf(1) < 0
64     flsf(1) = 0;
65 end
66 flsf(2) = MinVar_U(1)+DELTA+(MaxIter-perm(IT)+1)*Delta+rand(1)*Delta;
    %
67 if flsf(2) < 0
68     flsf(2) = 0;
69 end
70 for q=3:2:2*b(1)-1
71     flsf(q) = flsf(q-2)+2*DELTA;
72     if flsf(q) > MaxVar_U(1)
73         flsf(q) = MaxVar_U(1);
74     end
75     flsf(q+1) = flsf(q-1)+2*DELTA;
76     if flsf(q+1) > MaxVar_U(1)
77         flsf(q+1) = MaxVar_U(1);
78     end
79 end
80
81 flss(1) = MinVar_U(2)+(perm(IT)-1)*Delta+rand(1)*Delta;
                                % flss: first level partition of the second
    variable
82 if flss(1) < 0
83     flss(1) = 0;
84 end
85 flss(2) = MinVar_U(2)+DELTA+(MaxIter-perm(IT)+1)*Delta+rand(1)*Delta;
    %
86 if flss(2) < 0

```

```

87     flss(2) = 0;
88 end
89 for q=3:2:2*b(2)-1
90     flss(q) = flss(q-2)+2*DELTA;
91     if flss(q) > MaxVar_U(2)
92         flss(q) = MaxVar_U(2);
93     end
94     flss(q+1) = flss(q-1)+2*DELTA;
95     if flss(q+1) > MaxVar_U(2)
96         flss(q+1) = MaxVar_U(2);
97     end
98 end
99
100 % Distribution of first level candidate solution as a Matrix
101 for i=1:a(1)
102     for j=1:a(2)
103         Candidate01(i,j).Sol = [ flsf(i) flss(j) BT-(flsf(i)+flss(j))
104                                 ];
105     end
106 end
107 % Distribution of the above candidate solution as a vector
108 A = sortrows(cat(1,Candidate01.Sol));
109
110 for i=1:a(1)*a(2) % number of partition on the
111     first level
112
113     if (A(i,3) >= 0)
114         BeST(i).SolU = A(i,:);
115
116         % Call the PSO to obtain the corresponding best Lowervariable
117         % for each choice of Best(i).Sol
118
119         [xL,fL] = PSO_Mareka(BeST(i).SolU,MinVar_L,MaxVar_L,nvar_L,Npop
120                               ,Max_iter_PSO);
121
122         Best(i).SolL = xL;
123         Best(i).SolU = BeST(i).SolU;
124         Best(i).CostL = fL;
125         Best(i).CostU = ObjectiveFunctionU(BeST(i).SolU,Best(i).SolL)
126             +...

```

```

124         lambda*(sum(max(myconstraint_U_Ineq(Best(i).SolU,Best(i)
125             ).SolL),0)));
126     if i == 1
127         Best_T = Best(1); % Current
128         iteretion best upper
129
130     elseif i~= 1 && Best(i).CostU <= Best(i-1).CostU
131         Best_T = Best(i);
132
133     else
134         Best_T = Best(i-1);
135
136     end % if i==1
137 end % if condn on y(1)+y(2) = y(3)
138 end % for i=1:a(1)*a(2)
139
140 BEST(IT).COSTU= ObjectiveFunctionU(Best_T.SolU,Best_T.SolL);
141 BEST(IT).COSTL= Best_T.CostL; %only lower level constraints
142 BEST(IT).SOLU = Best_T.SolU;
143 BEST(IT).SOLL = Best_T.SolL;
144
145 % Selecting best solution for the current iteration
146
147 if IT == 1
148     BEst = BEST(IT);
149 elseif IT~= 1 && BEST(IT).COSTU <= GBEST(IT-1).COSTU
150     BEst = BEST(IT);
151
152 else
153     BEst = GBEST(IT-1);
154
155 end
156
157 GBEST(IT) = BEst;
158
159 % Displaying results on Command Window
160
161 disp(['Iteration ' num2str(IT) ': Upper Best Cost = ' num2str(GBEST
162     (IT).COSTU) ': Lower Best Cost = ' num2str(GBEST(IT).COSTL)]); %
163     ': Optimal Solution = ' num2str(GBEST(IT).SOLU)
164
165 end % for IT=1:MaxIter

```

```
162
163
164     disp(['The Optimal Solution for the upper level is = ' num2str(
        GBEST(MaxIter).SOLU) ': For the lower level problem is = '
        num2str(GBEST(MaxIter).SOLL)]);
```

A.2 The Particle Swarm Optimisation (PSO) MatLab code

```
1
2 %% Adapted and modified from the code produced by
3 % Jogendra Karthika Sri Sai Balaji Dulipala
4 % Last updated:- 19/4/2020
5 % contact details:- balajijogendra@gmail.com
6 %% Non-linear Inequality and equality constrained PSO
7 function [Xmin,Fmin]=PSO_Mareka(y, LB,UB, nvars ,Npop, max_it)
8 %function [Xmin,Fmin,SUM_Constraints]=PSO_non_linear_constraint(
    objective_function ,constraints ,constraints_eq ,LB,UB,nvars ,Npop,
    max_it)
9 %%
10 % Objective_function      function to be optimized
11 % nvars                   number of variables to be optimized
12 % npop                    number of populations
13 % LB                      Lower bounds of the problem
14 % UB                      Upper bounds of the problem
15 % max_it                  Maximum iterations
16 % constraints              Non-linear or linear inequality
    constraints
17 % constraints_eq           Non-linear or linear equality constraints
18 %%
19 format long
20     w=5;
21     c1=2;
22     c2=2;
23     alpha=0.75;
24     UppVar = y;
25
26     epss=1e-6;
27     for i=1:Npop
28         pop(i).position=LB+(UB-LB).*rand(1,nvars);
29         pop(i).cost=ObjectiveFunctionL(UppVar,pop(i).position);
30         c = myconstraint_L_Ineq(UppVar,pop(i).position);
31
32         pop(i).const=sum(c(c>epss));
```

```

33     pop(i).velocity=zeros(1,nvars);
34 end
35 %—————Sort to selection the global best and local best for
    1st iteration —————
36 X_Minus=[];
37 aa=[pop.const];
38 COST_MINUS=[pop(aa<=epss).cost];
39 if ~isempty(COST_MINUS)
40     X_Minus=pop(aa<=epss);
41     [~,INDEX_M]=sort(COST_MINUS);
42     X_Minus=X_Minus(INDEX_M);
43 end
44 X_PLUS=[];
45 SUM_C_PLUS=aa(aa>epss);
46 COST_PLUS=[pop(aa>epss).cost];
47 if ~isempty(SUM_C_PLUS)
48     AD=unique(SUM_C_PLUS);
49     if size(AD,2)==Npop
50         X_PLUS=pop(aa>epss);
51         [~,INDEX_P]=sort(SUM_C_PLUS);
52     else
53         X_PLUS=pop(aa>epss);
54         [OR,INDEX]=sort(SUM_C_PLUS);
55         COST_PLUS=COST_PLUS(INDEX);
56
57         kk=1;
58         N=0;
59
60         for m=1:size(AD,2)
61             B=length(find(AD(m)==OR));
62             [~,IND]=sort(COST_PLUS(kk:N+B));
63             INDEX_P(kk:N+B)=IND;
64             kk=B+1;
65             N=N+B;
66         end
67     end
68     X_PLUS=X_PLUS(INDEX_P);
69 end
70 pop=[X_Minus;X_PLUS];
71 %————— Forming the global best
    _____
72 global_best=pop(1);

```

```

73
74
75 for i=1:max_it
76     if i==1
77         local_best=global_best;
78     end
79     %————— Moving particals to the local and global best
80     positions—————
81     for j=1:Npop
82         pop(j).velocity = w.*pop(j).velocity+c1.*rand(1).*(
83             global_best.position-pop(j).position)+c2.*rand(1).*(
84             local_best.position-pop(j).position);
85         pop(j).position = pop(j).velocity+pop(j).position;
86         pop(j).position=min(pop(j).position,UB);
87         pop(j).position=max(pop(j).position,LB);
88
89         pop(j).cost= ObjectiveFunctionL(UppVar,pop(i).position); %
90         objective_function(pop(j).position);
91         c = myconstraint_L_Ineq(UppVar,pop(i).position);
92
93         pop(j).const=sum(c(c>epss));
94     end
95
96     local_best=global_best;
97     w=w*alpha;
98
99     for j=1:Npop
100         if (global_best.const<=epss && pop(j).const<=epss && pop(j).
101             cost<global_best.cost) || (global_best.const>epss && pop(j).
102             const<=epss)
103
104             global_best=pop(j);
105
106         elseif global_best.const>epss && pop(j).const>epss &&
107             global_best.const>pop(j).const
108
109             global_best=pop(j);
110
111     end
112 end

```

```

108
109 end
110
111
112
113 Xmin=global_best.position;
114 Fmin=ObjectiveFunctionL(UppVar,Xmin);
115
116 end

```

A.3 Matlab Code for the Results in Tables 6.1, 6.2, 6.3

```

1 function func_Graph_Hrisk(v1,v2,v3,v4,Count)
2
3 %%% get the parameter values %%%
4
5 %%% Intitial values of some varying parameters in countries or
   corresponding to risk (community) groups
6 N = [576064; 316606; 159225]; %Population size in the three districts
   , 1, 2, and 3 respectively
7
8 B = [0.0098; 0.0078; 0.0065]; %% Birth rates for the three districts ,
   respectively
9 D = [0.0032; 0.0015; 0.0011]; %% Death rates for the three districts ,
   respectively
10
11 LambdaH = [2.16; 3.17; 2.13]; % unsafe contact rate for high risk
   groups in the three districts
12 LambdaL = [1.11; 2.14; 1.09]; % unsafe contact rate for low risk
   groups in the three districts
13
14 MH = [0.31; 0.37; 0.28]; % rate of MICT for high risk groups in the
   three districts
15 ML = [0.19; 0.22; 0.15]; % rate of MICT for low risk groups in the
   three districts
16
17 AlphaA = [0.15; 0.32; 0.11]; % Rate of defaulting of the ART —
   assumed the same for both risk groups in each of the districts
18
19 RhoH = [0.15; 0.15; 0.20]; % rate of recruitment for ART in High risk
   groups in the three districts
20 RhoL = [0.075; 0.075; 0.10]; % rate of recruitment for ART in risk
   groups in the three districts

```

```

21
22 j = Count;
23
24 NN = N(j); %% Population size for district j
25
26 BB = B(j); %% Birth rates for district j % *NN
27 DD = D(j); %% Death rates for district j
28 LambdaHH = LambdaH(j); % unsafe contact rate for the High risk group
    in district j
29 LambdaLL = LambdaL(j); % unsafe contact rate for the Low risk group
    in district j
30 MHH = MH(j); % rate of MICT for the High risk group in district j
31 MLL = ML(j); % rate of MICT for the Low risk group in district j
32 Alpha = AlphaA(j); % Rate of defaulting of the ART in Country j —
    assumed the same for both risk groups
33 RhoHH = RhoH(j); % rate of recruitment for ART from the High risk
    group in district j
34 RhoLL = RhoL(j); % rate of recruitment for ART the Low risk group in
    district j
35
36 %%% The other parameter values are assumed to be the same across the
    district and risk groups
37
38 Prop = [0.44; 0.48; 0.32]; %% Proportions or percentage of
    population in High risk groups in the three districts
39
40 %%% Defining initial size of population for each risk groups in
    each of
41 %%% the districts
42
43 N11 = (1-Prop(j))*NN; % percentage of low risk population in district
    j
44 N12 = Prop(j)*NN; % percentage of high risk population in district j
45
46 %%% Table of proportion for initial populations in each of the three
    districts %%%
47 InitPop = [0.489 0.008 0.29 0.21 0.003 0.497 0.007 0.30 0.19 0.005;
    ...
48             0.495 0.004 0.30 0.19 0.007 0.505 0.006 0.31 0.17 0.009;
    ...
49             0.4903 0.0054 0.299 0.197 0.0083 0.504 0.008 0.321 0.16
    0.007];

```

```

50
51 %%% initial population size for the Low risk group in district j
    %%%%%%%%%%%%%%%
52 S011 = InitPop(j,1) * N11;
53 V011 = InitPop(j,2) * N11;
54 I011 = InitPop(j,3) * N11;
55 T011 = InitPop(j,4) * N11;
56 A011 = InitPop(j,5) * N11;
57
58 %%%%%%%%%%%%%%%
59
60 %%% initial population size for the High risk group in district j
    %%%%%%%%%%%%%%%
61 S012 = InitPop(j,6) * N12;
62 V012 = InitPop(j,7) * N12;
63 I012 = InitPop(j,8) * N12;
64 T012 = InitPop(j,9) * N12;
65 A012 = InitPop(j,10) * N12;
66
67 %%%%%%%%%%%%%%%
68
69
70 %%%% disease parameters — Assumed to be Constant across regions
    %%%%%% %%%%%%%%%%%
71 deltaV = 0.15; %
72 deltaA = 0.2; %
73 sigma = 0.1; %
74 gama = 1.3; %
75 beta = 0.12; %
76 epsilon1 = 0.08;
77 epsilon2 = 1/6;
78
79 %%%% Parameters that can vary from region to region or from risk
    groups to risk groups
80 %%% demographic parameters %%%%%%%%%%%
81
82 p = BB; %%%%0.03219; % birth rate % p = 0.099000000000000000000004357;
    %0.95;
83 mu = DD; %%0.02; % death rate
84
85 %%%% Disease parameters — to vary between risk groups or districts
    %%%%

```

```

86 %m = MLL; %%%0.31;
87 m = MHH; %%%0.31;
88 %lambda = LambdaLL; %0.16;
89 lambda = LambdaHH; %0.16;
90 alpha = Alpha; %%%0.4;
91 %rho = RhoLL; %%%0.075; %rho2 = ell1*rho1;
92 rho = RhoHH; %%%0.075; %rho2 = ell1*rho1;
93
94 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
95
96 % Some cost functions parameters – some of these values may vary from
    district to district %%%
97 a1 = 1.65 * 10−8; a2 = 5.43 * 10−8;
98 c1 = 4.56*10−10.55; c2 = 3.35*10−20;
99 r = 0.0375;
100
101 %%%%%%%%% %%%%%%%%% %%%%%%%%%
102
103 %%% initial values for HH communities %%%%%%%%%
104 %%% initial values for the ODE system %%%%%%%%%
105 V0 = V012; %%1017;
106 S0 = S012; %%48485;
107 I0 = I012; %%7407;
108 T0 = T012; %%2026;
109 A0 = A012; %%317;
110
111
112 % done with parameter values %%
113
114 %%% initialization and stoppage values ...
115
116 M = 120;
117 tfinal = 10;
118
119 h=tfinal/M;
120 h2 = h/2;
121
122 t=linspace(0,tfinal,M+1);
123
124 V = zeros(1,M+1);
125 S = zeros(1,M+1);
126 I = zeros(1,M+1);

```

```

127 T = zeros(1,M+1);
128 A = zeros(1,M+1);
129 Q = zeros(1,M+1);
130
131 %fun = zeros(1,M+1);
132
133
134 TN0 = V0 + S0 + I0 + T0 + A0;
135 V(1) = V0/TN0;
136 S(1) = S0/TN0;
137 I(1) = I0/TN0;
138 T(1) = T0/TN0;
139 A(1) = A0/TN0;
140 Q(1) = S(1)+I(1)+T(1);
141
142 %%%% Production functions to be implemented in the runge kuta
143 theta_v1 = max(0,lambda*(1 - exp(-(c1*v1))));
144 theta_v2 = max(0,m*(1 - a2*v2));
145 theta_v3 = rho*(1 + a1*v3);
146 theta_v4 = max(0,alpha*(1 - exp(-(c2*v4))));
147
148
149 %%%% Runge-Kuta 4th order method for solving the ODE ststem
    %%%%%%%%%%
150
151     for i=1:M
152
153         k11 = (0.0*p) + theta_v2*beta*(I(i)+epsilon2*T(i)) - (gama*
            theta_v3 + mu + deltaV)*V(i);
154         k12 = (1*p) - theta_v1*(S(i)/(S(i)+I(i)+T(i)))*(I(i)+epsilon1
            *T(i)) - mu * S(i);
155         k13 = theta_v1*(S(i)/Q(i))*(I(i)+epsilon1*T(i)) - (theta_v3 +
            sigma + mu)*I(i);
156         k14 = theta_v3*(gama*V(i)+I(i))-(theta_v4 + mu)*T(i);
157         k15 = sigma*I(i) + theta_v4*T(i)-(mu + deltaA)*A(i);
158
159
160
161         k21 = (0.0*p) + theta_v2*beta*((I(i)+h2*k13) + epsilon2*(T(i)
            +h2*k14)) - (gama*theta_v3 + mu + deltaV)*(V(i)+h2*k11);
162         k22 = (1*p) - theta_v1*((S(i)+h2*k12)/(S(i)+h2*k12+I(i)+h2*
            k13+T(i)+h2*k14))*(I(i)+h2*k13+epsilon1*(T(i)+h2*k14)) -

```

```

mu * (S(i)+h2*k12);
163 k23 = theta_v1*((S(i)+h2*k12)/(S(i)+h2*k12+I(i)+h2*k13+T(i)+
      h2*k14))*(I(i)+h2*k13+epsilon1*(T(i)+h2*k14)) - (theta_v3
      + sigma + mu)*(I(i)+h2*k13);
164 k24 = theta_v3*(gama*(V(i)+h2*k11)+I(i)+h2*k13)-(theta_v4 +
      mu)*(T(i)+h2*k14);
165 k25 = sigma*(I(i)+h2*k13) + theta_v4*(T(i)+h2*k14)-(mu +
      deltaA)*(A(i)+h2*k15);
166
167
168 k31 = (0.0*p) + theta_v2*beta*((I(i)+h2*k23) + epsilon2*(T(i)
      +h2*k24)) - (gama*theta_v3 + mu + deltaV)*(V(i)+h2*k21);
169 k32 = (1*p) - theta_v1*((S(i)+h2*k22)/(S(i)+h2*k22+I(i)+h2*
      k23+T(i)+h2*k24))*(I(i)+h2*k23+epsilon1*(T(i)+h2*k24)) -
      mu * (S(i)+h2*k22);
170 k33 = theta_v1*((S(i)+h2*k22)/(S(i)+h2*k22+I(i)+h2*k23+T(i)+
      h2*k24))*(I(i)+h2*k23+epsilon1*(T(i)+h2*k24)) - (theta_v3
      + sigma + mu)*(I(i)+h2*k23);
171 k34 = theta_v3*(gama*(V(i)+h2*k21)+I(i)+h2*k23)-(theta_v4 +
      mu)*(T(i)+h2*k24);
172 k35 = sigma*(I(i)+h2*k23) + theta_v4*(T(i)+h2*k24)-(mu +
      deltaA)*(A(i)+h2*k25);
173
174
175 k41 = (0.0*p) + theta_v2*beta*((I(i)+h*k33) + epsilon2*(T(i)+
      h*k34)) - (gama*theta_v3 + mu + deltaV)*(V(i)+h*k31);
176 k42 = (1*p) - theta_v1*((S(i)+h*k32)/(S(i)+h*k32+I(i)+h*k33+T
      (i)+h*k34))*(I(i)+h*k33+epsilon1*(T(i)+h*k34)) - mu * (S(i)
      +h*k32);
177 k43 = theta_v1*((S(i)+h*k32)/(S(i)+h*k32+I(i)+h*k33+T(i)+h*
      k34))*(I(i)+h*k33+epsilon1*(T(i)+h*k34)) - (theta_v3 +
      sigma + mu)*(I(i)+h*k33);
178 k44 = theta_v3*(gama*(V(i)+h*k31)+I(i)+h*k33)-(theta_v4 + mu)
      *(T(i)+h*k34);
179 k45 = sigma*(I(i)+h*k33) + theta_v4*(T(i)+h*k34)-(mu + deltaA
      )*(A(i)+h*k35);
180
181
182
183
184 V(i+1) = V(i) + (h/6)*(k11 + 2*k21 + 2*k31 + k41);
185 S(i+1) = S(i) + (h/6)*(k12 + 2*k22 + 2*k32 + k42);

```

```

186         I(i+1) = I(i) + (h/6)*(k13 + 2*k23 + 2*k33 + k43);
187         T(i+1) = T(i) + (h/6)*(k14 + 2*k24 + 2*k34 + k44);
188         A(i+1) = A(i) + (h/6)*(k15 + 2*k25 + 2*k35 + k45);
189
190         Q(i+1) = S(i+1)+I(i+1)+T(i+1);
191         TN(i+1) = V(i+1)+S(i+1)+I(i+1)+T(i+1)+A(i+1);
192
193         end %for
194
195     y(1,:) = V;
196     y(2,:) = S;
197     y(3,:) = I;
198     y(4,:) = T;
199     y(5,:) = A;
200
201     Nj = TN0*TN;
202
203     prev = 100.*(I+V+A);
204
205     figure;
206     plot(t,prev,'g-.','LineWidth',2.5);
207     title('Prevalence of the infection');
208     xlabel('Time');
209     ylabel('Prevalence(%)');
210     grid on
211     box on
212     hold on
213
214
215     %%% plots for the dynamics of the model
216
217     % subplot(1,2,2);
218     figure
219     plot(t,I,'r-',t,T,'g—',t,A,'b-','LineWidth',2.5);
220     title('The disease dynamics in the population');
221     xlabel('(d) Time (in years)');
222     ylabel('Population');
223     grid on
224     box on
225     legend('I','T','A'); %'Sc','Sp',
226
227 1 function func_Graph_Lrisk(v1,v2,v3,v4,Count)
228 2

```

```

3  %%% get the parameter values %%%
4
5  %%%%%%%%% Intitial values of some varying parameters in countries or
    corresponding to risk (community) groups
6  N = [576064; 316606; 159225]; %Population size in the three districts
    , 1, 2, and 3 respectively
7
8  B = [0.0098; 0.0078; 0.0065]; %% Birth rates for the three districts ,
    respectively
9  D = [0.0032; 0.0015; 0.0011]; %% Death rates for the three districts ,
    respectively
10
11 LambdaH = [2.16; 3.17; 2.13]; % unsafe contact rate for high risk
    groups in the three districts
12 LambdaL = [1.11; 2.14; 1.09]; % unsafe contact rate for low risk
    groups in the three districts
13
14 MH = [0.31; 0.37; 0.28]; % rate of MICT for high risk groups in the
    three districts
15 ML = [0.19; 0.22; 0.15]; % rate of MICT for low risk groups in the
    three districts
16
17 AlphaA = [0.15; 0.32; 0.11]; % Rate of defaulting of the ART —
    assumed the same for both risk groups in each of the districts
18
19 RhoH = [0.15; 0.15; 0.20]; % rate of recruitment for ART in High risk
    groups in the three districts
20 RhoL = [0.075; 0.075; 0.10]; % rate of recruitment for ART in risk
    groups in the three districts
21
22 j = Count;
23
24 NN = N(j); %% Population size for district j
25
26 BB = B(j); %% Birth rates for district j % *NN
27 DD = D(j); %% Death rates for district j
28 LambdaHH = LambdaH(j); % unsafe contact rate for the High risk group
    in district j
29 LambdaLL = LambdaL(j); % unsafe contact rate for the Low risk group
    in district j
30 MHH = MH(j); % rate of MICT for the High risk group in district j
31 MLL = ML(j); % rate of MICT for the Low risk group in district j

```

```

32 Alpha = AlphaA(j); % Rate of defaulting of the ART in Country j —
    assumed the same for both risk groups
33 RhoHH = RhoH(j); % rate of recruitment for ART from the High risk
    group in district j
34 RhoLL = RhoL(j); % rate of recruitment for ART the Low risk group in
    district j
35
36 %%% The other parameter values are assumed to be the same across the
    district and risk groups
37
38 Prop = [0.44; 0.48; 0.32]; %% Proportions or percentage of
    population in High risk groups in the three districts
39
40 %%%% Defining initial size of population for each risk groups in
    each of
41 %%%% the districts
42
43 N11 = (1-Prop(j))*NN; % percentage of low risk population in district
    j
44 N12 = Prop(j)*NN; % percentage of high risk population in district j
45
46 %%%% Table of proportion for initial populations in each of the three
    districts %%%%
47 InitPop = [0.489 0.008 0.29 0.21 0.003 0.497 0.007 0.30 0.19 0.005;
    ...
48             0.495 0.004 0.30 0.19 0.007 0.505 0.006 0.31 0.17 0.009;
    ...
49             0.4903 0.0054 0.299 0.197 0.0083 0.504 0.008 0.321 0.16
    0.007];
50
51 %%%% Table of proportion for initial populations in each of the
    three districts %%%%
52 % InitPop = [0.606 0.05 0.205 0.12 0.019 0.156 0.09 0.45 0.28 0.024;
    ...
53 %             0.506 0.05 0.245 0.16 0.029 0.156 0.09 0.45 0.28 0.024;
    ...
54 %             0.60 0.056 0.205 0.12 0.019 0.156 0.09 0.45 0.28 0.024];
55
56 %%% initial population size for the Low risk group in district j
    %%%%%%%%%%
57 S011 = InitPop(j,1) * N11;
58 V011 = InitPop(j,2) * N11;

```

```

59 I011 = InitPop(j,3) * N11;
60 T011 = InitPop(j,4) * N11;
61 A011 = InitPop(j,5) * N11;
62
63 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
64
65 %%% initial population size for the High risk group in district j
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
66 S012 = InitPop(j,6) * N12;
67 V012 = InitPop(j,7) * N12;
68 I012 = InitPop(j,8) * N12;
69 T012 = InitPop(j,9) * N12;
70 A012 = InitPop(j,10) * N12;
71
72 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
73
74
75 %%%% disease parameters — Assumed to be Constant across regions
    %%%%%%%%%% %%%%%%%%%%
76 deltaV = 0.15; %
77 deltaA = 0.2; %
78 sigma = 0.1; %
79 gama = 1.3; %
80 beta = 0.12; %
81 epsilon1 = 0.08;
82 epsilon2 = 1/6;
83
84 %%%% Parameters that can vary from region to region or from risk
    groups to risk groups
85 %%% demographic parameters %%%%%%%%%%
86
87 p = BB; %%%0.03219; % birth rate % p = 0.0990000000000000000000004357;
    %0.95;
88 mu = DD; %%0.02; % death rate
89
90 %%%% Disease parameters — to vary between risk groups or districts
    %%%
91 m = MLL; %%%0.31;
92 %m = MHH; %%%0.31;
93 lambda = LambdaLL; %0.16;
94 %lambda = LambdaHH; %0.16;
95 alpha = Alpha; %%%0.4;

```

```

96 rho = RhoLL; %%%0.075; %rho2 = ell1*rho1;
97 %rho = RhoHH; %%%0.075; %rho2 = ell1*rho1;
98
99 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
100
101 % Some cost functions parameters – some of these values may vary from
    district to district %%%
102 a1 = 1.65 * 10−8; a2 = 5.43 * 10−8; % a2 = 1.5*a1; %
103 c1 = 4.56*10−10.55; c2 = 3.35*10−20;
104 r = 0.0375;
105
106 %%%%%%%%% %%%%%%%%% %%%%%%%%%
107
108 %%% initial values for LL communities %%%%%%%%%
109 %%% initial values for the ODE system %%%%%%%%%
110 V0 = V011; %%%1017;
111 S0 = S011; %%%48485;
112 I0 = I011; %%%7407;
113 T0 = T011; %%%2026;
114 A0 = A011; %%%317;
115
116 %%% initialization and stoppage values ...
117
118 M = 120;
119 tfinal = 10;
120
121 h=tfinal/M;
122 h2 = h/2;
123
124 t=linspace(0,tfinal,M+1);
125
126 V = zeros(1,M+1);
127 S = zeros(1,M+1);
128 I = zeros(1,M+1);
129 T = zeros(1,M+1);
130 A = zeros(1,M+1);
131 Q = zeros(1,M+1);
132
133 %fun = zeros(1,M+1);
134
135
136 TN0 = V0 + S0 + I0 + T0 + A0;

```

```

137 V(1) = V0/TN0;
138 S(1) = S0/TN0;
139 I(1) = I0/TN0;
140 T(1) = T0/TN0;
141 A(1) = A0/TN0;
142 Q(1) = S(1)+I(1)+T(1);
143
144 %%%% Production functions to be implemented in the runge kuta
145 theta_v1 = max(0,lambda*(1 - exp(-(c1*v1))));
146 theta_v2 = max(0,m*(1 - a2*v2));
147 theta_v3 = rho*(1 + a1*v3);
148 theta_v4 = max(0,alpha*(1 - exp(-(c2*v4))));
149
150
151 %%%% Runge-Kuta 4th order method for solving the ODE ststem
152 %%%%%%%%%%
153
154 for i=1:M
155
156     k11 = (0.0*p) + theta_v2*beta*(I(i)+epsilon2*T(i)) - (gama*
157         theta_v3 + mu + deltaV)*V(i);
158     k12 = (1*p) - theta_v1*(S(i)/(S(i)+I(i)+T(i)))*(I(i)+epsilon1
159         *T(i)) - mu * S(i);
160     k13 = theta_v1*(S(i)/Q(i))*(I(i)+epsilon1*T(i)) - (theta_v3 +
161         sigma + mu)*I(i);
162     k14 = theta_v3*(gama*V(i)+I(i))-(theta_v4 + mu)*T(i);
163     k15 = sigma*I(i) + theta_v4*T(i)-(mu + deltaA)*A(i);
164
165     k21 = (0.0*p) + theta_v2*beta*((I(i)+h2*k13) + epsilon2*(T(i)
166         +h2*k14)) - (gama*theta_v3 + mu + deltaV)*(V(i)+h2*k11);
167     k22 = (1*p) - theta_v1*((S(i)+h2*k12)/(S(i)+h2*k12+I(i)+h2*
168         k13+T(i)+h2*k14))*(I(i)+h2*k13+epsilon1*(T(i)+h2*k14)) -
169         mu * (S(i)+h2*k12);
170     k23 = theta_v1*((S(i)+h2*k12)/(S(i)+h2*k12+I(i)+h2*k13+T(i)+
171         h2*k14))*(I(i)+h2*k13+epsilon1*(T(i)+h2*k14)) - (theta_v3
172         + sigma + mu)*(I(i)+h2*k13);
173     k24 = theta_v3*(gama*(V(i)+h2*k11)+I(i)+h2*k13)-(theta_v4 +
174         mu)*(T(i)+h2*k14);
175     k25 = sigma*(I(i)+h2*k13) + theta_v4*(T(i)+h2*k14)-(mu +
176         deltaA)*(A(i)+h2*k15);

```

```

168
169
170 k31 = (0.0*p) + theta_v2*beta*((I(i)+h2*k23) + epsilon2*(T(i)
      +h2*k24)) - (gama*theta_v3 + mu + deltaV)*(V(i)+h2*k21);
171 k32 = (1*p) - theta_v1*((S(i)+h2*k22)/(S(i)+h2*k22+I(i)+h2*
      k23+T(i)+h2*k24))*(I(i)+h2*k23+epsilon1*(T(i)+h2*k24)) -
      mu * (S(i)+h2*k22);
172 k33 = theta_v1*((S(i)+h2*k22)/(S(i)+h2*k22+I(i)+h2*k23+T(i)+
      h2*k24))*(I(i)+h2*k23+epsilon1*(T(i)+h2*k24)) - (theta_v3
      + sigma + mu)*(I(i)+h2*k23);
173 k34 = theta_v3*(gama*(V(i)+h2*k21)+I(i)+h2*k23)-(theta_v4 +
      mu)*(T(i)+h2*k24);
174 k35 = sigma*(I(i)+h2*k23) + theta_v4*(T(i)+h2*k24)-(mu +
      deltaA)*(A(i)+h2*k25);
175
176
177 k41 = (0.0*p) + theta_v2*beta*((I(i)+h*k33) + epsilon2*(T(i)+
      h*k34)) - (gama*theta_v3 + mu + deltaV)*(V(i)+h*k31);
178 k42 = (1*p) - theta_v1*((S(i)+h*k32)/(S(i)+h*k32+I(i)+h*k33+T
      (i)+h*k34))*(I(i)+h*k33+epsilon1*(T(i)+h*k34)) - mu * (S(i)
      +h*k32);
179 k43 = theta_v1*((S(i)+h*k32)/(S(i)+h*k32+I(i)+h*k33+T(i)+h*
      k34))*(I(i)+h*k33+epsilon1*(T(i)+h*k34)) - (theta_v3 +
      sigma + mu)*(I(i)+h*k33);
180 k44 = theta_v3*(gama*(V(i)+h*k31)+I(i)+h*k33)-(theta_v4 + mu)
      *(T(i)+h*k34);
181 k45 = sigma*(I(i)+h*k33) + theta_v4*(T(i)+h*k34)-(mu + deltaA
      )*(A(i)+h*k35);
182
183
184
185
186 V(i+1) = V(i) + (h/6)*(k11 + 2*k21 + 2*k31 + k41);
187 S(i+1) = S(i) + (h/6)*(k12 + 2*k22 + 2*k32 + k42);
188 I(i+1) = I(i) + (h/6)*(k13 + 2*k23 + 2*k33 + k43);
189 T(i+1) = T(i) + (h/6)*(k14 + 2*k24 + 2*k34 + k44);
190 A(i+1) = A(i) + (h/6)*(k15 + 2*k25 + 2*k35 + k45);
191
192 Q(i+1) = S(i+1)+I(i+1)+T(i+1);
193 TN(i+1) = V(i+1)+S(i+1)+I(i+1)+T(i+1)+A(i+1);
194
195 end %for

```

```
196
197 y(1,:) = V;
198 y(2,:) = S;
199 y(3,:) = I;
200 y(4,:) = T;
201 y(5,:) = A;
202
203 Nj = TN0*TN;
204
205 prev = 100.*(I+V+A);
206
207 %figure;
208 plot(t,prev,'y—','LineWidth',2.5);
209 title('Prevalence of the infection');
210 xlabel('Time');
211 ylabel('Prevalence(%)');
212 grid on
213 box on
214 hold on
```

A.4 Additional Matlab Codes for the Investment figures In Bi-level Allocation

```
1 function ff = LowerObjfuncMod(v1,v2,v3,v4,Count) %,N1,N2,P,D,Lambda,M
   ,Alpha,Rho
2
3 %% v1 represents a vector of two investment variables targeted to
   reduce lambda (unsafe contact rate) corresponding to the three
   districts
4 %% v2 represents a vector of two investment variables targeted to
   reduce m (MICT) corresponding to the three districts
5 %% v3 represents a vector of two investment variables targeted to
   increase rho (rate of recruitment for treatment) corresponding to
   the three districts
6 %% v4 represents a vector of two investment variables targeted to
   reduce alpha (rate of defaulting the treatment) corresponding to
   the three districts
7
8 %% Count represents a counter for the district , the dynamics is to be
   calculated
9
10 %%%%%%%%% Intitial values of some varying parameters in countries or
   corresponding to risk (community) groups
11 N = [576064; 316606; 159225]; %Population size in the three districts
   , 1, 2, and 3 respectively
12
13 B = [0.0098; 0.0078; 0.0065];%% Birth rates for the three districts ,
   respectively
14 D = [0.0032; 0.0015; 0.0011]; %% Death rates for the three districts ,
   respectively
15
16 LambdaH = [2.16; 3.17; 2.13]; % unsafe contact rate for high risk
   groups in the three districts
17 LambdaL = [1.11; 2.14; 1.09]; % unsafe contact rate for low risk
   groups in the three districts
18
19 MH = [0.31; 0.37; 0.28]; % rate of MICT for high risk groups in the
   three districts
20 ML = [0.19; 0.22; 0.15]; % rate of MICT for low risk groups in the
   three districts
21
22 AlphaA = [0.15; 0.32; 0.11]; % Rate of defaulting of the ART —
```

```

    assumed the same for both risk groups in each of the districts
23
24 RhoH = [0.15; 0.15; 0.20]; % rate of recruitment for ART in High risk
    groups in the three districts
25 RhoL = [0.075; 0.075; 0.10]; % rate of recruitment for ART in risk
    groups in the three districts
26
27 j = Count;
28
29 NN = N(j); %% Population size for district j
30
31 BB = B(j); %% Birth rates for district j    % *NN
32 DD = D(j); %% Death rates for district j
33 LambdaHH = LambdaH(j); % unsafe contact rate for the High risk group
    in district j
34 LambdaLL = LambdaL(j); % unsafe contact rate for the Low risk group
    in district j
35 MHH = MH(j); % rate of MICT for the High risk group in district j
36 MLL = ML(j); % rate of MICT for the Low risk group in district j
37 Alpha = AlphaA(j); % Rate of defaulting of the ART in Country j —
    assumed the same for both risk groups
38 RhoHH = RhoH(j); % rate of recruitment for ART from the High risk
    group in district j
39 RhoLL = RhoL(j); % rate of recruitment for ART the Low risk group in
    district j
40
41 %%% The other parameter values are assumed to be the same across the
    district and risk groups
42
43 Prop = [0.44; 0.48; 0.32]; %% Proportions or percentage of
    population in High risk groups in the three districts
44
45 %%% Defining initial size of population for each risk groups in
    each of
46 %%% the districts
47
48 N11 = (1-Prop(j))*NN; % percentage of low risk population in district
    j
49 N12 = Prop(j)*NN; % percentage of high risk population in district j
50
51 %%% Table of proportion for initial populations in each of the three
    districts %%%

```

```

52 InitPop = [0.489 0.008 0.29 0.21 0.003 0.497 0.007 0.30 0.19 0.005;
    ...
53         0.495 0.004 0.30 0.19 0.007 0.505 0.006 0.31 0.17 0.009;
    ...
54         0.4903 0.0054 0.299 0.197 0.0083 0.504 0.008 0.321 0.16
        0.007];
55
56
57 %%% initial population size for the Low risk group in district j
    %%%%%%%%%%
58 S011 = InitPop(j,1) * N11;
59 V011 = InitPop(j,2) * N11;
60 I011 = InitPop(j,3) * N11;
61 T011 = InitPop(j,4) * N11;
62 A011 = InitPop(j,5) * N11;
63
64 %%%%%%%%%%
65
66 %%% initial population size for the High risk group in district j
    %%%%%%%%%%
67 S012 = InitPop(j,6) * N12;
68 V012 = InitPop(j,7) * N12;
69 I012 = InitPop(j,8) * N12;
70 T012 = InitPop(j,9) * N12;
71 A012 = InitPop(j,10) * N12;
72
73 %%%%%%%%%%
74
75
76 y11 = v1(1); y12 = v2(1); y13 = v3(1); y14 = v4(1);
77 y21 = v1(2); y22 = v2(2); y23 = v3(2); y24 = v4(2);
78
79
80 %%%%%%%%%% %%%%%%%%%% %%%%%%%%%%
81
82
83 fkj = NN *( func(y11,y12,y13,y14,BB,DD,LambdaLL,MLL,Alpha,RhoLL,S011,
    V011,I011,T011,A011) + ...
84         func(y21,y22,y23,y24,BB,DD,LambdaHH,MHH,Alpha,RhoHH,
    S012,V012,I012,T012,A012) );
85
86

```

```

87  ff = fkj;

1  function const = myconstraint_L_Ineq(y,x)
2
3      % Given a two dimensional vector y, and a 24 dimensional variable
4      % vector x, the third component of y is calculated as  $y(3) = BT -$ 
5      %  $(y(1)+y(2))$ .
6      % The lower level constraints are given as inequality constraints
          and c
7      % is a 6 x 1 column vector and the RHS is zero. That means  $c \leq$ 
          0.

8
9      %      BT = 1600000000;
10     %       $y(3) = BT - y(1) - y(2)$ ;
11
12     B(1) = x(1)+x(2)+x(3)+x(4)+x(5)+x(6)+x(7)+x(8);
13     B(2) = x(9)+x(10)+x(11)+x(12)+x(13)+x(14)+x(15)+x(16);
14     B(3) = x(17)+x(18)+x(19)+x(20)+x(21)+x(22)+x(23)+x(24);
15
16     % Here the inequality constraint c of the lower level problem is
17     % a column vector with 27 components, with a form  $g(x) \leq 0$  for
18     % each of the rows.
19
20     const = [B(1)-y(1);          %% Budget used should not exceed
              what is allocated to the region.
              B(2)-y(2);
              B(3)-y(3);
              0.3*y(1)- B(1);      %% equity constraints
              0.25*y(2)-B(2);
              0.2*y(3)-B(3)
              0.05*y(1)-(x(1)+x(2));
              0.03*y(1)-(x(3)+x(4));
              0.05*y(1)-(x(5)+x(6));
              0.02*y(1)-(x(7)+x(8));
              0.05*y(2)-(x(9)+x(10));
              0.03*y(2)-(x(11)+x(12));
              0.05*y(2)-(x(13)+x(14));
              0.02*y(2)-(x(15)+x(16));
              0.05*y(3)-(x(17)+x(18));
              0.03*y(3)-(x(19)+x(20));
              0.05*y(3)-(x(21)+x(22));
              0.02*y(3)-(x(23)+x(24))
21     ];
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

```

```

39 end

1 function const = myconstraint_U_Ineq(y,x)
2
3     % Given a two dimensional vector y, and a 24 dimensional variable
4     % vector x, the third component of y is calculated as  $y(3) = BT -$ 
5     %  $(y(1)+y(2))$ .
6     %The values B(1),B(2),B(3) are the  $v(j)$ 
7     %The upper level constraints are given as inequality constraints.
8
9     BT = 16000000000;
10    %       $y(3) = BT - y(1) - y(2)$ ;
11
12    B(1) = x(1)+x(2)+x(3)+x(4)+x(5)+x(6)+x(7)+x(8);
13    B(2) = x(9)+x(10)+x(11)+x(12)+x(13)+x(14)+x(15)+x(16);
14    B(3) = x(17)+x(18)+x(19)+x(20)+x(21)+x(22)+x(23)+x(24);
15
16    % Here the upper level constraint, represented by c, is a column
17    % vector of size 19, each inequality is of the form  $g(x) \leq 0$ 
18    % form.
19    const = [B(1) + B(2) + B(3) - BT;
20            0.3*y(1)- B(1);
21            0.25*y(2)-B(2);
22            0.2*y(3)-B(3)
23            0.05*y(1)-(x(1)+x(2));
24            0.03*y(1)-(x(3)+x(4));
25            0.05*y(1)-(x(5)+x(6));
26            0.02*y(1)-(x(7)+x(8));
27            0.05*y(2)-(x(9)+x(10));
28            0.03*y(2)-(x(11)+x(12));
29            0.05*y(2)-(x(13)+x(14));
30            0.02*y(2)-(x(15)+x(16));
31            0.05*y(3)-(x(17)+x(18));
32            0.03*y(3)-(x(19)+x(20));
33            0.05*y(3)-(x(21)+x(22));
34            0.02*y(3)-(x(23)+x(24))
35            ];
36 end

1 function z = ObjectiveFunctionL(y,x)
2
3     % A 2 dimensional vector y, and a 24 dimensional vector x are to

```

```

        be
4      % input. y(3) is to be calculated as  $BT - y(1) - y(2)$ .
5      % The upper level variables are to be used only in the
        constraints.
6      % Given x, the matlab function LowerObjfuncMod calculates the
        total
7      % number of of infections in each region.
8      % The objective is to minimize the total number of infection in
        all the
9      % three regions.
10
11      BT = 1600000000;
12      y(3) = BT - y(1) - y(2);
13
14      z = LowerObjfuncMod([x(1),x(2)], [x(3),x(4)], [x(5),x(6)], [x(7),x
        (8)],1)+...
15      LowerObjfuncMod([x(9),x(10)], [x(11),x(12)], [x(13),x(14)], [x
        (15),x(16)],2)+...
16      LowerObjfuncMod([x(17),x(18)], [x(19),x(20)], [x(21),x(22)], [x
        (23),x(24)],3);
17 end

1 function z = ObjectiveFunctionU(y,x)
2
3      % Given a two dimensional variable vector y, and a 24 dimensional
        vector x,
4      % (1) the 3rd component of y can be calculated from the the total
        % budget BT and the sum of y(1) and y(2).
5      % (2) The variable x only appears in the constraint of the upper
        level
6      % problem.
7      Lvar = x;
8
9
10     z = 0.39*y(1) + 0.32*y(2) + 0.29*y(3);
11 end

```

Bibliography

- [1] Alam, T., Qamar, S., Dixit, A. and Benaida, M. (2020). Genetic Algorithm: Reviews, Implementations, and Applications. 10.36227/techrxiv.12657173.v1.
- [2] Anselmi, L., Lagarde, M. and Hanson, K. (2015). Equity in the allocation of public sector financial resources in low- and middle-income countries: a systematic literature review, *Health Policy and Planning* 30(4): 528–545. <https://doi.org/10.1093/heapol/czu034>
- [3] Avert News. (2017). History of HIV and AIDS overview. Available from: <https://avert.info/who-we-are/our-annual-reports>. Accessed on October 2020.
- [4] Avert news report. (2018). Funding for HIV and AIDS. Available from: <https://avert.info/who-we-are/our-annual-reports>. Accessed on October 2020.
- [5] Avert news. Understanding the HIV Pandemic. Available from: <https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/botswana>. Accessed on October 2020.
- [6] Avert news. Understanding the HIV Pandemic. Available from: <https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/botswana>. Accessed on February 2021.
- [7] Azimi, N. A., and Welch, H. G. (1998). The effectiveness of cost-effectiveness analysis in containing costs. *Journal of general internal medicine*, 13(10), 664–669. <https://doi.org/10.1046/j.1525-1497.1998.00201.x>
- [8] Bard, J. F. (1998). Practical bilevel optimization: Algorithms and applications. Dordrecht: Kluwer Academic Publishers. DOI 10.1007/978-1-4757-2836-1.
- [9] Bard, J. F. (1998). Practical bilevel optimization: Algorithms and applications. Dordrecht: Kluwer Academic Publishers. DOI 10.1007/978-1-4757-2836-1.
- [10] Bertsimas, D. and Tsitsiklis, J. (1993). Simulated Annealing. *Statistical sciences* 8(1):10-15. <https://doi.org/10.1214/ss/1177011077>
- [11] Bhattacharjya, R. K. Introduction to particle swarm optimization (lecture slides). Department of Civil Engineering Indian Institute of Technology Guwahati. Available from: <https://www.iitg.ac.in/rkbc/CE602/CE602/Particle%20Swarm%20Algorithms.pdf>. Accessed on September 2021
- [12] Broder S. (2010). The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. *Antiviral research*, 85(1), 1–18. <https://doi.org/10.1016/j.antiviral.2009.10.002>
- [13] Center for Disease Control and Prevention, Office of the Associate Director for Policy and Strategy <https://www.cdc.gov/policy/polaris/economics/cost-effectiveness.html>

- [14] Chaisson, R. E. (1999). HIV becomes world's leading infectious cause of death. The Hopkins HIV report : a bimonthly newsletter for healthcare providers, 11(4), 1.
- [15] Chong, E. and Zak, S. (2013). An introduction to Optimization. John Wiley & Sons, INC Hobocan New Jesrsey ISBN 978-1-118-27901-4
- [16] Christensen, J. and Bastien, C. (2016). Gradient Descent Method. Nonlinear Optimization of Vehicle Safety Structures. Elsevier B.V. <https://doi.org/10.1016/C2013-0-00069-2>
- [17] Cogswell, H., Nakhimovsky, S., Gutierrez, J., Faye, S., Cintron, C., Haider, R. and Avila, C. (2016) Antiretroviral Therapy in Botswana comparing costs, service utilization and quality at three levels of care. Health Finance and Governance. <https://www.hfgproject.org/antiretroviral-therapy-botswana-comparing-costs-service-utilization-quality-three-levels-care/>.
- [18] Durea, M. and Strugariu, R. (2014). An introduction to nonlinear optimization theory. De Gruyter Open (Lty) Berlin, Germany.
- [19] Earnshaw, S. R., Hicks, K., Richter, A., and Honeycutt, A. (2007). A linear programming model for allocating HIV prevention funds with state agencies: a pilot study. Health care management science, 10(3), 239–252. <https://doi.org/10.1007/s10729-007-9017-8>
- [20] Edwin, Patricia (18 October 2012). Botswana falls short in AIDS fight. MmegiOnline. Available from: [http : //www.mmegi.bw/index.php?sid = 1&aid = 175&dir = 2012/October/Thursday18](http://www.mmegi.bw/index.php?sid=1&aid=175&dir=2012/October/Thursday18). Accessed on September 2020.
- [21] Enane, L. A., Mokete, K., Joel, D., Daimari, R., Tshume, O., Anabwani, G., Mazhani, L., Steenhoff, A. P., and Lowenthal, E. D. (2018). "We did not know what was wrong"-Barriers along the care cascade among hospitalized adolescents with HIV in Gaborone, Botswana. PloS one, 13(4), e0195372. <https://doi.org/10.1371/journal.pone.0195372>
- [22] Fliege, J., Tin, A. and Zemkoho, A. (2021). Gauss–Newton-type methods for bilevel optimization. Computational Optimization and Applications 78:793 – 824. doi: 10.1007/s10589-020-00254-3
- [23] Fox-Rushby, J. A. and Hanson, K. (2001). Calculating and presenting disability adjusted life years (DALYs) in cost-effectiveness analysis. Health policy and planning, 16(3), 326–331. <https://doi.org/10.1093/heapol/16.3.326>
- [24] Garcia, G. (2004) Cost-Effectiveness Analysis: An Employer Decision Support Tool. Issue Brief: August 2004. Available on: <https://www.slideshare.net/gerardgap/cost-effectiveness-analysis-in-health-economics-presentation>. Accessed on February 2020.
- [25] Goshu, N.N. and Kassa, S.M. (2020). A Systematic Sampling evolutionary(SSE) method for stachastic bilevel programming problems. Computers and operations research 120 (104942). <https://doi.org/10.1016/j.cor.2020.104942>.

- [26] He, Y., Wei, M. and Ji, Z. (2016). The Parameters Selection of PSO Algorithm influencing On performance of Fault Diagnosis MATEC Web of Conferences 63, 02019. Doi: 10.1051/mateconf/20166302019.
- [27] Hoel M. (2003). Allocating health care resources when people are risk averse with respect to life time. Health economics, 12(7), 601–608. <https://doi.org/10.1002/hec.733>
- [28] Jacobsen, M. M., and Walensky, R. P. (2016). Modeling and Cost-Effectiveness in HIV Prevention. Current HIV/AIDS reports, 13(1), 64–75. <https://doi.org/10.1007/s11904-016-0303-2>
- [29] Jain, S. (2021). Week 4-1 Genetic Algorithm. Skill Link. Available from: <https://skill-lync.com/student-projects/week-4-1-genetic-algorithm-5>. Accessed on September 2021.
- [30] Kaplan, E. H. and Merson, M. H. (2002). Allocating HIV-prevention resources: balancing efficiency and equity. American journal of public health, 92(12), 1905–1907. <https://doi.org/10.2105/ajph.92.12.1905>
- [31] Kassa, S. M. (2018) Three-Level global resource allocation model for HIV control: A Hierarchical decision system approach Mathematical biosciences and engineering 15: 255-273. <http://dx.doi.org/10.3934/mbe.2018011>
- [32] Komorowski, M., and Raffa, J. (2016). Markov Models and Cost Effectiveness Analysis: Applications in Medical Research. In MIT Critical Data (Ed.), Secondary Analysis of Electronic Health Records. (pp. 351–367). Springer.
- [33] Lagrange Multipliers. Academic Resource Center. Illinois Institute of Technology. Available from: https://web.iit.edu/~Lagrange_Multipliers.pdf. Accessed on February 2022.
- [34] Lasry, A., Zaric, G. S. and Carter, M. W. (2007) Multi-level resource allocation for HIV prevention: A model for developing countries. European Journal of Operational Research 180: 786-799.
- [35] Lasry, A., Carter, M.W. and Zaric, G.S. (2008) S4HARA: System for HIV/AIDS resource allocation. Cost Eff Resour Alloc 6, 7. <https://doi.org/10.1186/1478-7547-6-7>
- [36] Lasry, A., Carter, M. W., and Zaric, G. S. (2011). Allocating funds for HIV/AIDS: a descriptive study of KwaDukuza, South Africa. Health Policy and Planning, 26(1), 33–42. <http://www.jstor.org/stable/45090702>.
- [37] Lasry, A., Stephanie, L., Hicks, K. and Uzunangelov, V. (2011). A model for allocating CDC’s HIV prevention resources in the United States. Health Care Management Science 14(1): 115–124 <https://link.springer.com/article/10.1007/s10729-010-9147-2>
- [38] Love-Koh, J., Griffin, S., Kataika, E., Revill P., Sibandze, S. and Walker S. (2020). Methods to promote equity in health resource allocation in low- and middle-income countries: an overview. Global Health 16(6). <https://doi.org/10.1186/s12992-019-0537-z>

- [39] Majelantle, A.N. (2014). Vital Statistic Report 2011. Statistics Botswana Available from: [https://statsbots.org.bw/sites/default/files/publications/Vital %20%20Statistics%20Report%202011.pdf](https://statsbots.org.bw/sites/default/files/publications/Vital%20%20Statistics%20Report%202011.pdf). Accessed on November 2021.
- [40] Mgumi, go q B. S. (2020) Cities, Towns and Villages Projections 2020, Statistics Botswana. Available from: [https://www.statsbots.org.bw/sites/default/files/Cities %20Towns%20and%20Villages%20Projections%202020.pdf](https://www.statsbots.org.bw/sites/default/files/Cities%20Towns%20and%20Villages%20Projections%202020.pdf). Accessed on November 2021.
- [41] Mijwel, M. M. (2015). Heuristic Algorithms. Computer science, college of science, University of Baghdad, Iraq. *https : //www.researchgate.net/publication/322162995HeuristicAlgorithms*.
- [42] Mitton, C., Donaldson, C. (2004). Health care priority setting: principles, practice and challenges. Cost Eff Resour Alloc 2, 3. *https : //doi.org/10.1186/1478 – 7547 – 2 – 3*
- [43] NEOS Guide. (2020). Types of Optimization Problems. Available from: <https://neos-guide.org/guide/algorithms/>. Accessed on January 2022
- [44] Oaitse Mathala, (2021), New HIV infections still high, The Patriot on Sunday, Jan 16 2021, <https://thepatriot.co.bw/new-hiv-infections-still-high/>
- [45] Outrata, J.V. (1990). On the numerical solution of a class of Stackelberg problems. ZOR - Methods and Models of Operations Research 34, 255–277. <https://doi.org/10.1007/BF01416737>.
- [46] Philip, D. L., Laura, A. M. and Maria E. M. (2012). Modeling Equity for Allocating Public Resources. International Series in Operations Research and Management Science, in: Michael P. Johnson (ed.), Community-Based Operations Research, chapter 0: 97-118. DOI: 10.1007/978 – 1 – 4614 – 0806 – 2%₄
- [47] Prieto, L. and Sacristán, J. A. (2003). Problems and solutions in calculating quality-adjusted life years (QALYs). Health and quality of life outcomes, 1, 80. <https://doi.org/10.1186/1477-7525-1-80>
- [48] Sarkar, S., Corso, P., Ebrahim-Zadeh, S., Kim, P., Charania, S., and Wall, K. (2019). Cost-effectiveness of HIV Prevention Interventions in Sub-Saharan Africa: A Systematic Review. EClinicalMedicine, 10, 10–31. <https://doi.org/10.1016/j.eclinm.2019.04.006>
- [49] Schmidt, M. and Beck, Y. (2021). A Gentle and Incomplete Introduction to Bilevel Optimization. Bilevel Optimization lecture notes. Trier University. Available from: <https://optimization-online.org/2021/06/8450/>.
- [50] Shalab, H. Sampling theory: Systematic sampling Chapter 11. Department of Mathematics and Statistics. Indian Institute of Technology Kanpur. Available from: <http://home.iitk.ac.in/~shalab/course1.htm>. Accessed October 2021.
- [51] Sinha, A., Malo, P. and Deb, K. (2017). A review on bilevel optimization: From Classical to evolutionary approaches and applications. arXiv. <https://doi.org/10.48550/arXiv.1705.06270>

- [52] Sinha, A., Soun, T. and Deb, K. (2017). Evolutionary Bilevel Optimization Using KKT - Proximity Measure. IEEE Press 2412–2419. <https://doi.org/10.1109/CEC.2017.7969597>
- [53] Statistics Botswana. (2013). Botswana AIDS Impact survey IV, National AIDS Coordinating Agency. Available from: <https://www.statsbots.org.bw/sites/default/files/publications/BOTSWANA%20AIDS%20IMPACT%20S> Accessed on November 2021.
- [54] UNAIDS. (2017). Botswana- HIV country profile:2016. Global AIDS Monitoring. Available from: https://indicatorregistry.unaids.org/sites/default/files/2017-global-aids-monitoring_en0.pdf. Accessed on 18 September 2019.
- [55] UNAIDS. (2021). Global HIV and AIDS statistics — Fact sheet. Available from: <https://www.unaids.org/en/resources/fact-sheet>. Accessed on August 2021.
- [56] UNICEF, (2019). Botswana Budget brief HIV/AIDS 2019. Available from: <https://www.unicef.org/esa/documents/botswana-budget-briefs-2019>. Accessed on January 2022.
- [57] UNICEF- Botswana, <https://www.unicef.org/botswana/hiv#:~:text=Botswana%20is%20ranked%20among>
- [58] Yuan, Y. (1999). A Review of Trust Region Algorithms for Optimization. Proceedings of the Fourth International Congress on Industrial and Applied Mathematics 99(1): 271-278. <https://www.researchgate.net/publication/2304086>
- [59] Zaric, G. S. and Brandeau, M. L.(2007) A little planning goes a long way: Multilevel allocation of HIV prevention resources, Medical Decision Making, 27: 71–81.
- [60] Zegeye, E. A., Mbonigaba, J., Kaye, S., and Johns, B. (2019). Assessing the cost of providing a prevention of mother-to-child transmission of HIV/AIDS service in Ethiopia: urban-rural health facilities setting. BMC health services research, 19(1), 148. <https://doi.org/10.1186/s12913-019-3978-4>.