CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Epidemic tracking using network-based scaling

An Innovative Approach for Real-time Epidemic Surveillance and Control in East Africa

Master's thesis in Computer science and engineering

Daniele Murgolo & Tomas Vu

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's thesis 2023

Epidemic tracking using network-based scaling

An Innovative Approach for Real-time Epidemic Surveillance and Control in East Africa

Daniele Murgolo & Tomas Vu

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

Epidemic tracking using network-based scaling
An Innovative Approach for Real-time Epidemic Surveillance and Control in East Africa
Daniele Murgolo, Tomas Vu

Supervisor: Philippas Tsigas, Department of Computer Science and Engineering
Examiner: Marina Papatriantafilou, Department of Computer Science and Engineering

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Description of the picture on the cover page (if applicable)

Typeset in LaTeX
Gothenburg, Sweden 2023

Epidemic tracking using network-based scaling
An Innovative Approach for Real-time Epidemic Surveillance and Control in East Africa
Daniele Murgolo, Tomas Vu
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Estimating the sizes of subpopulations enables the potential to optimally allocate resources and funding for people in need, e.g., subpopulations that are affected by epidemics such as HIV/AIDS. We propose a method based on related literature on networking to accomplish this, namely the survey-based Network scale-up method (NSUM). This is done by collecting aggregated relational data (ARD) by asking participants "How many X do you know?", where X can be any subpopulation, such as people named Michael or doctors. The internal workings behind NSUM is that with the use of participants' social networks, the sizes of hard-to-reach subpopulations can be estimated by extrapolating and scaling up the total population.

The aim of this thesis is to build a system based on a Web questionnaire and the use of NSUM in hopes of estimating the sizes of hard-to-reach subpopulations. Different models within NSUM such as the random degree model (RD-model), barrier effect model (BE-model), and transmission bias model (TB-model) which account for various errors and biases, were also explored where comparisons and analysis of their performance were conducted.

Data sets from Uganda, pertaining to occupational distribution, and Rwanda, focusing on the age distribution were used. The results from the two data sets are presented in a forest plot. The metric of choice is the difference in magnitude between the true value and the estimated values. The RD-model produces estimates that are close to the true value with small variances whereas the TB-model overestimates with a large dispersion for both datasets. Lastly, the BE-model produces conflicting estimates between the two datasets.

Even though it is difficult to affirm the estimated value, we conclude that, while these estimates are consistent to a certain degree, the various biases and errors may produce less than satisfactory results.

Keywords: Computer science, data engineering, estimation, epidemics, network scale-up method.

Acknowledgements

I would like to express my heartfelt gratitude to all those who have supported me throughout this journey. Your unwavering belief in me has been a constant source of motivation. Thank you for your invaluable support.

Daniele Murgolo, Göteborg, 2023-12-08

I want to express my gratitude to those who supported me throughout this thesis. All of this work and years of studies could not have been done without them, and for that, I am very grateful. Thank you!

Tomas Vu, Göteborg, 2023-12-08

Contents

List of Figures

List of Tables

List of Abbreviations

1NF First Normal Form. 25

2NF Second Normal Form. 25

3NF Third Normal Form. 25

AIDS Acquired Immunodeficiency Syndrome. v, 1–3, 13, 32, 38

API Application Programming Interface. 15

ARD Aggregated Relational Data. v, ix, 6, 10–12, 28

BE Bias Effect. v, 9, 39, 41, 43, 44, 46, 51, 52, 55

CSRF Cross-Site Request Forgery. 20

DBSM Database Management System. 24

DRY Dont Repeat Yourself. 19

GDPR General Data Protection Regulation. 17

HIV Human Immunodeficiency Viruses. v, 1–3, 13, 29, 31, 32, 38, 39, 41, 44, 49, 50, 53–55

HTML Hypertext Markup Language. 20

IQR Interquartile Range. 54

MCMC Markov Chain Monte Carlo. 31, 38

MVC Model-View-Controlle. 19

NSUM Network Scale-Up Method. v, 2–7, 15, 18, 19, 21, 31, 38, 43

ORM Object-Relational Mapping. 19, 20, 28, 29

RD Random Degree. v, 8–10, 39, 41, 43, 44, 46, 50, 51, 53, 55, 56

RDBMS Relational Database Management System. 17

RDS Respondent Driven Sampling. 10

SEIR Susceptible, Exposed, Infectious, Recovered. 27

SIR Susceptible, Infectious, Recovered. 27

SQL Structured Query Language. 17, 19, 20

STI Sexually Transmitted Infections. 13

TB Transmission Bias. v, 9, 10, 39, 41, 43, 44, 46, 52, 53, 55

UNAIDS Joint United Nations Programme on HIV/AIDS. 38

WCAG Web Content Accessibility Guidelines. 29

WHO World Health Organization. 13

XSS Cross-Site Scripting. 20

1

Introduction

As part of their commitment to the third Sustainable Development Goal of the United Nations, East African countries are working to end epidemics, such as HIV/AIDS, that affect them by 2030. To optimally reach this goal, proper precautions and allocation of resources are a necessity; thus, the topic of estimating the sizes of subpopulations affected by an epidemic is an important issue. Digital technologies will play a crucial role in reaching this goal as existing health infrastructure and rapidly growing populations may not be sufficient.

Measurement of the sizes of subpopulations affected by epidemics can be difficult depending on the sensitivity of the topic. For example, in the case of HIV/AIDS, the reason why estimating the number of people infected with HIV/AIDS is difficult to assess is that people fear the stigma associated with the disease. Even if people choose to participate and answer surveys, it can still be a problem with respect to reporting accurate responses [1]. These unknown subpopulations are defined as hard-to-reach subpopulations. In addition to the challenges of estimating the sizes of subpopulations, there are other obstacles to keep in mind, such as time constraints when collecting data.

1.1 Purpose

The purpose of this thesis is to implement a comprehensive system that uses a data collection tool and estimates the true number of a hidden population. Specifically, we suggest using a Web-based questionnaire, which would contain a series of questions related to different groups. Participants would be able to access the questionnaire and provide their responses, thus facilitating the collection of relevant data. Once the data have been collected, it would then be utilized in a mathematical model to estimate the number of affected individuals.

However, in addition to the challenges of estimating the sizes of populations, there are other obstacles to take into account. A common problem within evaluating a data collection system is that not only do we have to build a system, but it takes time to convince people to participate and fill out a survey so that enough data can be collected. Furthermore, when the system is deployed in low-income African countries, the economic restrictions in these regions can significantly impact the successful implementation of the system. Factors such as limited resources and financial constraints may pose obstacles to both the initial setup and the ongoing function-

ing of the system. Addressing these challenges requires a thoughtful approach that takes into consideration the socio-economic conditions of the target communities, emphasizing the need for adaptability and resilience in the implementation strategy.

Firstly, to resolve the issues of estimating hard-to-reach subpopulations, we propose the use of the Network Scale-up Method [1] [2] [3]. The Network Scale-up Method, NSUM, makes use of the social networks of the respondents to estimate the size of any subpopulations. In other words, the method makes use of the sizes of known subpopulations to make assumptions about the sizes of unknown subpopulations. Therefore, instead of asking respondents direct questions about their personal attributes, they are asked how many people they know in a specific subpopulation. This method of collecting data gives it an advantage over many other data collection methodologies, such as the enumeration method or the synthetic estimation and multivariate indicator. Methods such as the synthetic estimation and multivariate indicator are computationally heavy and may require a lot of resources, for example, data that can be hard to obtain. Whereas NSUM is easy to deploy and very cost-efficient in terms of economic resources, making it superior. NSUM was developed in 1986 by Bernard, Hallett, Iovita, *et al.* [1] in an attempt to estimate the number of casualties in Mexico caused by an earthquake and it will be further explained in Section 2.2. Since then, numerous studies have used this method and extended the basic model by accounting for errors and biases to enable estimation of other hard-to-reach subpopulations, e.g., people affected by HIV/AIDS [2].

Secondly, in regards to obstacles with data collection mentioned above, such as convincing people to participate so that enough data can be collected for estimation. To bypass this, we manually collect data instead given the time frame of this thesis. Data are collected from official registries of countries such as Rwanda and Uganda, which are then used to estimate an already known subpopulation and treated as if it were unknown. The resulting estimates are then compared to the true values using the difference in magnitude.

Regarding the obstacle with data collection, in this thesis, we decided to manually gather data from official registries of East African countries, i.e., Rwanda and Uganda. We treat this data as if it were from an unknown subpopulation and we use it to estimate known subpopulations. Finally, we compare the resulting estimates to the true value, assessing the difference in magnitude.

This approach would involve scaling up from known cases to estimate the number of cases that have not been reported or detected. By utilizing NSUM and its extensions, we hope to gain a more accurate understanding of the extent of the issue at hand and, in turn, make more informed decisions about how to address it. Therefore, this system will provide valuable information that can help to efficiently and effectively manage the problem.

In summary, the purpose and main research of this thesis are the following:

1. **Implementation of a system for data collection**
2. **Analysis and performance comparisons of different models within NSUM**

2

Theory

Estimation of the size of subpopulations is a cumbersome task in itself, and depending on the topic, it can even be difficult. As mentioned in the introduction, the estimation of subpopulation sizes affected by certain epidemics can be difficult depending on the sensitivity of the topic, e.g., HIV / AIDS. These subpopulations are called hard-to-reach because information concerning them is hard to collect because individuals in these groups may be difficult to encounter, or they report inaccurate responses for fear of the stigma surrounding these epidemics. Thus, with the use of NSUM, one hopes to alleviate these issues in order to gather more accurate information.

The general idea behind this method is that everyone has an equal probability of knowing someone else in a specific subpopulation and with the use of individuals' social networks, one can extrapolate this to the overall population. A simple example would be if an individual considers 100 people in their social network and 2 of them belong to any subpopulation (e.g HIV), i.e. 2%, then 2% of the entire population have HIV, [1]. This idea is mathematically formulated in Equation [2.1] [4].

$$\frac{y_{ik}}{d_i} = \frac{N_k}{N} \tag{2.1}$$

Where y_{ik} is the number of people that person i knows in subpopulation k and d_i is the size of this person's social network. Also, N_k is the size of the subpopulation of interest and N is the size of the entire general population.

Although this method works with some notable results, the idea of being able to extrapolate in the aforementioned manner is quite prone to biases and errors. Newer methods have been developed in an attempt to counteract these obstacles, and in the following sections, we will go into more detail about the Network scale-up method, its internal workings, its weaknesses, and how to account for these. However, before that, we need to first introduce some mathematical statistics that are necessary and are the basis of the Network Scale-up method.

2.1 Basic Statistics

Network Scale-up Methods, like many estimation approaches, involve mathematical statistics. Since the general idea behind NSUM is that everyone has an equal

probability of knowing someone else in any subpopulation then this corresponds to a binomial distribution. However, the idea that everyone has an equal probability of knowing someone else in any subpopulation is quite a generous assumption. It would make more sense if people had a higher probability of knowing someone else who was more similar to them. To resolve this issue, one can make use of "prior" knowledge to get better estimates by accounting for these assumptions. The following sections will go through the Bayes theorem and how NSUM makes use of this theorem. Thereafter, the maximum likelihood estimation will be covered since it is the method of choice to estimate the size of the hard-to-reach subpopulations.

2.1.1 Conditional Probability and Bayes Theorem

Before going into Bayes' theorem and how it relates to NSUM, conditional probability needs to be covered, as it will help to understand the theorem. Conditional probability deals with probability that changes with additional information. When considering the case of smoking and lung cancer, it would make sense for people who are smokers to have a higher probability of contracting lung cancer. Another example that involves dice could be the probability of an odd outcome given that the outcome is at least 4 [5]. This can be formulated using the following notations:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3} \quad \text{where} \tag{2.2}$$

$$\begin{aligned} A &= \{\text{Event of odd number of dots}\} \\ B &= \{\text{Event of at least 4 dots}\} \\ A \cap B &= \{\text{Event of odd number AND at least 4 dots}\} \\ P(A) &= \text{Probability of odd number of dots} \\ P(B) &= \text{Probability of at least 4 dots} \\ P(A \cap B) &= \text{Probability of odd number AND at least 4 dots} \end{aligned}$$

In the numerator, the only possible outcome of "odd number AND at least 4 dots" is 5, then the probability of this is event 1/6. The denominator has outcomes {4,5,6}, which is half of the possible outcomes. This results in a probability of 1/3 when computing the probability of an odd event given the outcome of at least 4.

Based on previous conditional probability, the Bayes theorem is computed "backwards". Instead of A given B, the desired result is B given A, which is formulated in the following manner:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad \text{where} \tag{2.3}$$

$$P(B|A) = \text{Posterior}$$
$$P(A|B) = \text{Likelihood}$$
$$P(B) = \text{Prior}$$

The idea behind this theorem is that in the posterior, the distribution of B changes with the added data A. Similarly to the previous example concerning smokers and lung cancer, assume that the prior $P(B)$ is the probability of contracting lung cancer. Then the posterior $P(B|A)$ is the probability of contracting lung cancer given that a person is a smoker, which should be greater than $P(B)$ and thus change our conclusions about B [5].

As mentioned above, NSUM builds on the generous assumption idea that everyone has the probability of knowing someone else in a subpopulation. This assumption leads to errors and biases when trying to model the real world and estimate the sizes of hard-to-reach subpopulations. To reduce these errors and biases, one can add prior knowledge about certain subpopulations to get better estimates, which will be covered in later sections.

2.1.2 Maximum Likelihood Estimation

Mathematical statistics is not only about computing probabilities but also about estimating the parameters of a distribution, and the maximum likelihood estimation is one such method. The underlying methodology behind the maximum likelihood is to find the parameter that maximizes a likelihood function given some data. Consider an example where an event has possible outcomes 0 and 1 with probabilities $1 - p$ and p, respectively. The probability P of n number of events $\{x_i; i = 1, ..., n\}$ can be computed as follows:

$$P(x_1, x_2, ..., x_n) = p^k * (1 - p)^{n-k} \tag{2.4}$$

where each x_i has outcome 0 or 1 and was recorded n-k and k times, respectively. Furthermore, it is known that the value of p is either $\frac{1}{3}$ or $\frac{2}{3}$ and that the recorded observations are $\{1, 1, 0, 1\}$. Then the probability of these events, given that $p = \frac{1}{3}$ or $p = \frac{2}{3}$ is 0.025 and 0.099, respectively. Since 0.099 is greater than 0.025 it would make sense to conclude that $p = \frac{2}{3}$ [5].

Extending this idea to the general case where the probability p of an event x_i is not known, it can be derived using the likelihood function which is formulated in the following manner:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i) \tag{2.5}$$

NSUM Models	Variability in Degree	Non-Random Mixing	Lack of Awareness/ Stigma
Scale-Up Estimates			
Random Degree Model	✓		
Barrier Effects Model	✓	✓	
Transmission Effects Model	✓		✓

Table 2.1: Scale-Up models and their extensions

where θ is an unknown parameter of interest, in the previous case $\theta = p$, is estimated by maximizing the likelihood function $L(\theta)$. The maximum is computed by differentiation of the likelihood function $L(\theta)$, which is the product of the probability mass function (or probability density function) $f_\theta(X_i)$ of the events X_i. In the case of the previous example, $f_\theta(0) = 1 - p$ and $f_\theta(1) = p$.

This process is easier to do by taking the logarithm of the likelihood function, which produces the so-called log-likelihood function, as it is easier to differentiate and maximize a sum than a product. Thus, the maximum likelihood estimation method can be summarized with the following 2 steps [5].

1. Find L(θ) and transform it to l(θ) = log(L(θ))

2. Maximize l(θ) by differentiation with respect to θ, setting it equal to 0 and solving it for θ

The maximum likelihood estimation method is used to estimate the parameters of a model produced from NSUM. Since NSUM is built on a binomial distribution Bin(n,p), the parameters of interest are the number of trials n and the probability p. In the case of NSUM, these parameters are presented in a different manner and will be covered in the following sections.

2.2 Network Scale-up

The Network Scale-up Method uses the degree (social network) of individuals to extrapolate to the general population to estimate hard-to-reach populations. Aside from being able to estimate the sizes of hard-to-reach subpopulations, this method has the advantage of being relatively inexpensive and can be easily integrated into surveys. It is estimated that collecting Aggregated Relational Data (ARD) leads to a reduction in economic costs of $70 - 80\%$ compared to traditional data collection methods. Not only that, when it comes to ethical concerns, NSUM has another advantage in that it does not violate the respondents' privacy. This is done with the use of aggregated relational data, ARD, where respondents are asked "How many X do you know" where X can be any subpopulation of interest, such as people infected with HIV or people named Michael [6].

Once enough ARD has been collected, it can be utilized in a scale-up model which is formulated in the following manner:

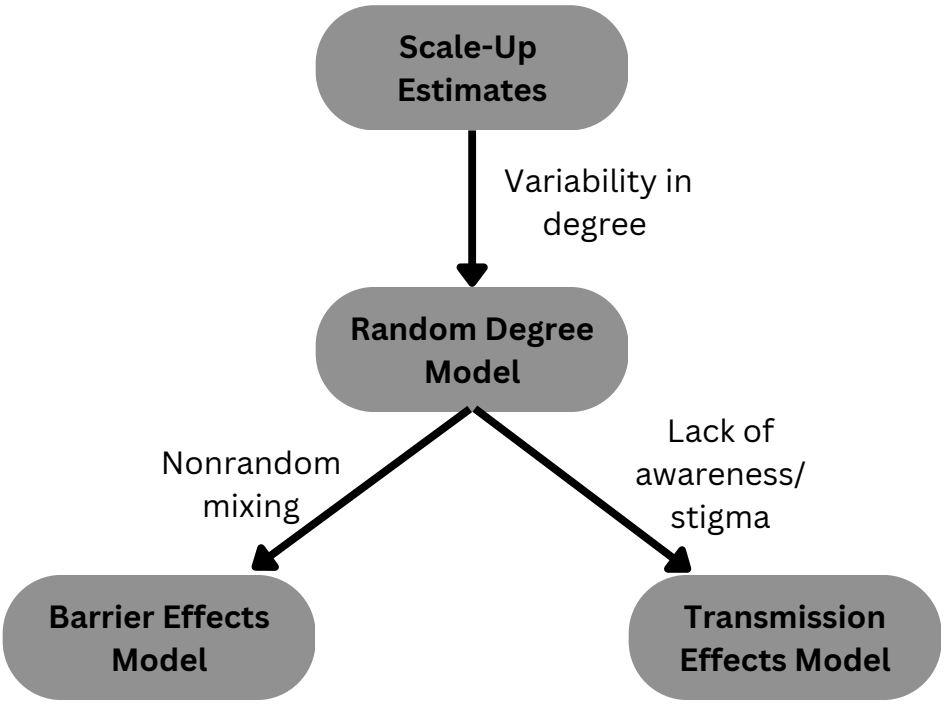


Figure 2.1: Scale-Up models

$$y_{ik} \sim Binom(d_i, \frac{N_k}{N}) \tag{2.6}$$

where y_{ik} is the number of people that individuals $i = 1, ..., n$ know in subpopulations $k = 1, ..., K$, which follows a binomial distribution. The sizes of the subpopulations $k = 1, ..., K-1$ are known, whereas group K is the unknown subpopulation of interest. Using this, the internal workings of how NSUM works are divided into two steps.

1. Estimate the size of the degree, d_i, of individual $i = 1, ..., n$
2. With the estimated degree of individuals, estimate the subpopulation of interest, N_K

In the first step, the degree d_i of individual i can be estimated using the maximum likelihood estimator in the following manner:

$$\hat{d}_i = N \frac{\sum_{k=1}^{K-1} y_{ik}}{\sum_{k=1}^{K-1} N_k} \tag{2.7}$$

In the second step, conditioning on $\hat{d}_i$, the maximum likelihood estimator of the size of the unknown subpopulation, N_K, is then:

$$\hat{N}_K = N \frac{\sum_{i=1}^{n} y_{iK}}{\sum_{i=1}^{n} \hat{d}_i} \tag{2.8}$$

This estimate can produce fairly good results but it heavily relies on the assumption that the proportion of people that are in an individual's social network and in a subpopulation is equal to the proportion of said set to the general population. As previously mentioned, everyone has the same probability of knowing someone else in a subpopulation. This assumption is easily violated because people have different tendencies to know others in different subpopulations. For example, Asians have a higher probability of knowing other Asians compared to non-Asians. This is called a barrier effect. As can be seen in Figure [2.1] and in Table [2.1], Maltiel, Raftery, McCormick, *et al.* [2] attempted to build extensions of the basic scale-up model to account for these biases and other errors, which will be covered in the following subsections.

2.2.1 Random Degree

The first extension of the basic scale-up model is the random degree model (RD-model), developed by Maltiel, Raftery, McCormick, *et al.* [2], to reduce overfitting. In the basic scale-up model, the degree of an individual would be overestimated and skewed if said individual reported knowing a large number of people in any subpopulation. Thus, by adding a random effect for the degree, the estimates for the degree are regularized and overfitting is reduced. The RD-model is presented as follows:

$$\begin{aligned} y_{ik} &\sim Binom(d_i, \frac{N_K}{N}) \\ d_i &\sim LogNormal(\mu, \sigma^2) \end{aligned} \tag{2.9}$$

$$\begin{aligned} \pi(N_K) &\propto \frac{1}{N_K} \leq N \\ \mu &\sim U(3, 8) \\ \sigma &\sim U(1/4, 2) \end{aligned} \tag{2.10}$$

where the degree d_i is modeled as a log-normal distribution, presented in Equation [2.9]. With Bayesian statistics, the parameters of the RD-model are estimated using the prior distributions presented in Equation [2.10]. Maltiel, Raftery, McCormick, *et al.* [2] determined these priors from multiple studies spanning multiple regions.

2.2.2 Barrier Effect

The barrier effect is a social phenomenon in which individuals are more prone to know someone in a subpopulation due to their own characteristics. Barrier effects can be influenced by both geographical and social factors and can affect hard-to-reach and easy-to-reach populations. This violates the constant proportion assumption in which everyone has the same probability of knowing someone else in any subpopulation, and as a result, the distribution is overdispersed in the number of people known in a specific population, which can lead to inaccurate estimates [2].

The barrier effect is not covered by the basic scale-up and the RD-model. Maltiel, Raftery, McCormick, *et al.* [2] developed an extension to the RD-model, the barrier effect model (BE-model). Previously, the probability that the respondent i knows someone in group k was fixed and equal to N_k/N. To account for overdispersion, this probability noted as $q_{i\,k}$, is modified and instead follows a Beta distribution. In total, the BE-model has the following structure:

$$\begin{aligned} y_{ik} &\sim Binom(d_i, q_{ik}) \\ d_i &\sim LogNormal(\mu, \sigma^2) \\ q_{ik} &\sim Beta(m_k, \rho_k) \end{aligned} \tag{2.11}$$

$$\begin{aligned} \mu &\sim U(3, 8) \\ \sigma &\sim U(1/4, 2) \\ \tau(m_K) &\propto \frac{1}{m_K} \\ \rho_k &\sim U(0, 1) \end{aligned} \tag{2.12}$$

where m_k is the prior mean of $q_{i\,k}$ which is computed as $E(q_{i\,k}) = m_k = \frac{N_k}{N}$ and where ρ_k represents the dispersion [2]. Since this model is an extension of the RD-model, it uses the same priors which are presented in Equation [2.10]

2.2.3 Transmission Bias

Transmission bias occurs when an individual is unaware or reluctant to acknowledge which subpopulation the people in their social networks belong to. For example, in the case of HIV, a person may be reluctant to disclose their HIV status for fear of the stigma surrounding this disease. Thus, individuals would not count the said person because they are unaware of the person's status. This violates the assumption that respondents have perfect knowledge of the subpopulations their contacts belong to and can lead to inaccurate results [2].

The transmission bias model (TB-model) is also an extension of the RD-model in which transmission bias is accounted for by including a transmission bias parameter,

denoted by $\tau_k \sim \text{Beta}(\eta_k, \nu_k)$. This parameter represents the portion of respondents' contacts in group k that the respondents reported knowing about. To account for transmission bias, it is assumed that $\tau_k = 1$ for groups of known size. For groups of unknown size, the transmission parameter is assumed to be less than or equal to 1. In practice, it is useful to incorporate external information on the transmission bias parameter, which can help improve the accuracy of the model [2]. In total, the TB-model is presented as follows:

$$\begin{aligned} y_{ik} &\sim Binom(d_i, \tau_k \frac{N_K}{N}) \\ d_i &\sim LogNormal(\mu, \sigma^2) \end{aligned} \tag{2.13}$$

$$\begin{aligned} \pi(N_K) &\propto \frac{1}{N_K} \leq N \\ \mu &\sim U(3, 8) \\ \sigma &\sim U(1/4, 2) \\ \tau_k &\sim Beta(\eta_K, \nu_K) \end{aligned} \tag{2.14}$$

The priors presented in Equation [2.14] are the same as the priors in the RD-model since the TB-model is an extension of it.

2.3 Aggregated Relational Data (ARD)

In this section, we delve into the theoretical concepts of Aggregated Relational Data (ARD). The section aims to provide a comprehensive understanding of the theoretical underpinnings that form the basis of ARD and its integration with the methods described above. As described by Breza, Chandrasekhar, Lubold, *et al.* [7], ARD is a powerful technique that allows us to gather information about a group of individuals by leveraging the social network of a subset of respondents. It provides a unique approach to estimating population-level characteristics, particularly in cases where direct observation or traditional sampling methods are unfeasible.

As a general rule, ARD does not involve observation of any links in the network and is collected using probability sampling techniques [7]. In this way, ARD differs from other methods that exploit subjects' networks to reach certain populations. For example, Respondent Driven Sampling (RDS) relies on connections between individuals to access members of a certain population. In RDS and other snowball sampling schemes, additional participants enter the surveys as they are recruited from previous respondents' networks. This is clearly a violation of privacy and needs to be addressed. However, this feature, while appealing from an efficiency point of view, fundamentally depends on social networks. On the other hand, ARD, as stated above, can be collected without observing any links and can be incorporated into standard survey platforms. As discussed in Section 2.2, the network still plays

a key role in the responses that individuals give. However, the sampling process is independent of the population's social networks. This also ensures a level of anonymity for those who complete the questionnaire.

Originally, ARD was proposed to estimate hard-to-reach populations, such as the number of HIV-positive men in the U.S (Killworth, Johnsen, McCarty, *et al.* [8]). Since then, ARD has expanded its application drastically, particularly in social sciences. As per how to analyze ARD, McCormick, Salganik, and Zheng [3] connects a model for ARD responses to network models of the fully observed graph. Particularly, they established a connection between ARD and the latent distance model, a common statistical approach to model fully observed networks. The result is that ARD is sufficient to identify parameters in a generative model for graphs also allowing inference in the distribution of them.

We can now formally define ARD. Let us take the undirected and unweighted graph: $g = (V, E)$ with vertex set V and edge set E. We have that our graph has $n = |V|$ nodes. We have that $g_{ij} = 1\{i, j \in E\}$ indicates that node i and node j are connected. Let's suppose that each node has one of K traits, where K is fixed and $K > 3$. Let G_k denote the nodes with trait k and $n_k = |G_k|$ the number of nodes with trait k. We suppose that the traits are binary and mutually exclusive so that every node has one of K traits. To collect Aggregated Relational Data (ARD) researchers ask m randomly chosen nodes the question "How many people with trait k are you linked to?" for each of these K traits. The definition of linking varies based on the research goal and on the application ARD is used, a typical definition is "having interacted with the person in the past 2 years". To simplify, we set $m = n$, meaning we have a response from all of our nodes but the results also apply to $m << n$ as in real-world scenarios. In such cases, we would need to impute parameters for nodes without ARD or make assumptions on node equivalence. Let y_{ik} denote node i response to the question about trait k, it follows that $y_{ik} = \sum_{j \in G_k} g_{ij}$. Fundamentally, when researchers are collecting ARD they do not observe any edges, just how many edges are present between the given node and people with trait k. [7]

Since K traits are mutually exclusive ARD counts distinct alters across trait groups. To model the network we consider the general graph model $\mathbb{P}(g_n|\theta^*)$ where edges form independently in the network, conditional on the unknown parameter vector θ^* these models are known as conditional-edge independent graph models. The number of elements in θ^* depends on the graph size but we omit this dependency. In certain cases, the distribution of θ_i^*, which are i.i.d., depends on the traits possessed by node i and is denoted by $\theta_i^*|t_i^* = k \sim F_k$. This conditional independence representation is based on exchangeability among nodes and implies that the resulting asymptotic sequence of networks generated by these models are dense, meaning that the average degree for a given n is a constant time n [9].

We define $p_{ij} = p_{ij}(\theta^*)$ as the probability that nodes i and j connect, given the model parameters. The response variable $y_{ik} = \sum_{j \in G_k} g_{ij}$ is then a sum of independent, but not identically distributed, Bernoulli(p_{ij}) random variables, which is known in various disciplines as either the Poisson's Binomial random variable or the Poisson Binomial random variable [10]. The probability mass function of y_{ik} given θ in

conditional edge-independent models is:

$$f_{i\,k}(y|\theta^*) = \sum_{A \subseteq A_y} \prod_{j \in A} p_{i\,j}(\theta) \prod_{j \in A^c} \{1 - p_{i\,j}(\theta)\} \tag{2.15}$$

Let A_y denote the set of subsets of (1, ..., n_k) that contain exactly y elements. When $p_{ij} = p$, the expression simplifies to the probability mass function of the Binomial(n_k, p) random variable. Modeling the ARD begins with analyzing the likelihood of the data, given by $L_n(y|\theta) = \prod_{i=1}^{n} \prod_{k=1}^{k} f_{ik}(y_{ik}|\theta)$. Here, we assume mutually exclusive traits, which allows us to write the likelihood of observing (y_{i1}, . . . , y_{iK}) as $\prod_{k=1}^{k} f_{ik}(yik|\theta)$. The conditional independence of edges, given θ, enables us to express the joint distribution of ARD responses over all individuals as a product, irrespective of whether traits are mutually exclusive.

Proving the consistency of $\hat{\theta_n} := argmax_\theta L_n(y|\theta)$ is challenging because the log-likelihood is complex, and each θ_i appears in n terms of the likelihood. Therefore, an alternative approach is to estimate θ differently. Instead of considering $f_{ik}(y_{ik}|\theta)$, where θ includes the parameters of node i as well as those of all other nodes with trait k (which are not observed with ARD), we examine the probability that node i connects to an arbitrary node with trait k, P_{ik}. This is expressed as follows:

$$P_{ik} := \mathbb{P}(g_{ij} = 1|\theta_i^*, j \in G_k) = \int_{\Theta_k} \mathbb{P}(g_{ij} = 1|\theta_i^*, \theta_j) dF_k(\theta_j) \tag{2.16}$$

In the context of network analysis, the model assumes that nodes with trait k are distributed according to a distribution $F_{\theta,k}$, where $\theta_j \sim F_{\theta,k}$. The support of F_k is denoted by Θ_k.

The probability that a node i connects with another node of trait k is denoted by P_{ik} and the number of nodes with trait k is denoted by n_k. To understand the utility of analyzing P_{ik}, instead of the full log-likelihood $L_n(y|\theta)$, it is noted that for any node i, $\frac{y_{ik}}{n_k}$ converges to P_{ik} as n_k approaches infinity. This is based on the assumption that the weak law of large numbers applies to the average $\frac{y_{ik}}{n_k}$, as is the case for conditionally edge-independent graphs.

Using this information, an estimating equation approach can be employed to estimate the model parameters by equating the vector of normalized aggregated relational data (ARD) responses with their respective edge probabilities $P_{i\,k}(\theta^*)$. This approach allows for the derivation of estimators of model parameters and uniform convergence of these estimators in a variety of network models.

2.3.1 Medical Questions and Information Content

Medical questions serve as the primary means to access indirect information on the prevalence of infection within a target subpopulation. To design effective medical questions, it is important to include inquiries about the occurrence of symptoms related to the infection. These symptoms can provide valuable insights into potential cases, ranging from overt manifestations, such as fever, coughing, or rash, to

more subtle indicators that might go unnoticed by infected individuals. Including a comprehensive set of symptom-related inquiries increases the accuracy of estimates by capturing a broad spectrum of potential cases.

For instance, in the context of estimating the number of individuals infected with HIV/AIDS, medical questions may include inquiries about common symptoms such as persistent fatigue, weight loss, night sweats, and recurrent infections. These symptoms have been identified by the World Health Organization (WHO) as indicative of HIV/AIDS infection (WHO, 2016) [11], and their inclusion in medical questions can help identify undiagnosed cases within hard-to-reach subpopulations.

In addition to symptoms, medical questions should aim to capture both diagnosed and undiagnosed cases of the infection. Diagnosed cases can be obtained through medical records or official surveillance systems, but undiagnosed cases often remain hidden. Including these queries allows for a more comprehensive understanding of the true prevalence of the disease, which is crucial for public health planning and resource allocation.

Another example, in the context of estimating the number of individuals infected with tuberculosis, medical questions may include a history of chronic cough lasting more than two weeks, unexplained weight loss, night sweats, and chest pains. These symptoms are commonly associated with tuberculosis infections (WHO, 2020) [12].

To enhance the precision and granularity of estimates, medical questions should gather additional information about the infected individuals and their social network connections. This may include demographic characteristics, such as age, gender, and socioeconomic status, as well as the nature and frequency of interactions among network members. Such information provides valuable insights into the differential risk of infection across various demographic strata and the potential transmission dynamics within the social network.

For instance, in the context of estimating the number of individuals infected with sexually transmitted infections (STIs), medical questions may include inquiries about demographic factors such as age, gender, sexual orientation, and educational level. Additionally, questions about the number of sexual partners, condom use, and engagement in high-risk behaviors can help identify potential patterns of transmission within the social network [13].

Moreover, it is essential to consider the context in which the medical questions are administered. The wording, order, and format of the questions can significantly impact respondent comprehension and response accuracy. Utilizing clear and unambiguous language, avoiding leading or suggestive phrasing, and employing standardized response options can help minimize potential biases and errors in the data collected.

In conclusion, when designing medical questions, it is advisable to follow established guidelines, such as those provided by reputable health organizations like the World Health Organization (WHO). Additionally, consulting with medical experts can provide valuable insights into the selection of relevant symptoms and the overall design of the questions. Careful attention to the design and administration of these ques-

tions is vital to ensure valid and reliable estimates that inform effective public health interventions.

3

Methods

In this day and age, with the large variety of programming languages, fancy libraries, frameworks, and APIs, there are several ways to build a system that can store responses to a questionnaire and apply these data with NSUM to extract something useful. Out of all the alternatives available, we chose to use Python and a framework called Django to build our system and the necessary tools in order to estimate the sizes of hard-to-reach subpopulations.

In the following sections, we will go into more detail as to why we chose Python and Django. We will also go into detail about the system architecture, such as how the questions are formed and their responses.

3.1 Software Requirements Specification

This section outlines the software requirements for the development and implementation of the web application and database, which will facilitate the collection and use of data to estimate the number of infected people in populations difficult to reach. Given that the project aims to address the context of African countries, it is imperative to consider software solutions that are accessible, open source, easy to deploy, and not overly cumbersome.

3.1.1 Web Application Development Framework

The selection of an appropriate web application development framework is crucial for the successful implementation of the project. The chosen framework should meet the specific requirements of the questionnaire-based application, while also considering the context of African countries. The following requirements guide the selection of a suitable development framework.

3.1.1.1 Scalability and Performance

To ensure that the web application can handle a potentially large number of users and data, the chosen framework, must demonstrate scalability and offer robust performance. It should be capable of efficiently processing user requests, handling concurrent connections, and seamlessly scaling up as the application usage grows.

3.1.1.2 Flexibility and Extensibility

The framework should provide a flexible and extensible architecture that allows for the addition of new features, modules, and functionalities in the future. As the project evolves and potential improvements or modifications arise, the framework should enable easy integration and customization without requiring significant rework or compromising the stability of the existing system. This flexibility is vital to accommodate future data analysis techniques.

3.1.1.3 Compatibility with Open Source Solutions

Considering the project's emphasis on utilizing open-source software, the chosen web development framework should seamlessly integrate with other open-source tools and technologies. This compatibility facilitates the use of open-source databases, libraries, and server infrastructure, ensuring a cost-effective and easily maintainable system. Moreover, the availability of a strong open-source community and extensive documentation can provide valuable support and resources for the development process.

3.1.1.4 Ease of Deployment and Management

Given the intended deployment in African countries, it is essential that the framework be easy to install, configure, and manage. The framework should have clear and comprehensive documentation, providing step-by-step instructions and guidelines on various server environments. Simplifying the deployment process reduces the barriers to entry, making the application accessible to users with limited technical expertise and minimizing the need for extensive technical support.

3.1.1.5 Security and Stability

The framework must prioritize security and stability to protect the integrity of the collected data and ensure the reliability of the web application. It should offer robust security features, such as built-in authentication and authorization mechanisms, to safeguard users' information. Additionally, regular updates and a strong community support system contribute to the stability of the framework, ensuring that vulnerabilities are promptly addressed and that the application remains reliable and resilient.

3.1.2 Questionnaire Design

Questionnaire design is a fundamental aspect that significantly impacts the user, data collection, and the overall success of the project. Considering the diverse user base in African countries, we need to ensure that users can provide accurate responses to the questionnaire. The presentation of information should be clear and unambiguous. Using a consistent and logical layout, employing appropriate typography, and using easily understandable language are critical factors in enhancing comprehension. Well-structured forms with concise instructions and contextual

help can assist users in comprehending the questions and provide a seamless data collection experience.

3.1.3 Data Storage and Retrieval

Efficient data storage and retrieval are important aspects of the project. Ensuring that the data collected can be effectively managed and accessed is vital. The design of a well-structured database schema plays a pivotal role in achieving optimal data organization and retrieval. Some of the key aspects that we should consider are listed below.

3.1.4 Database Schema Design

The database schema design serves as the foundation for organizing and structuring the collected data. A well-designed schema is essential for efficient data storage and retrieval. It should accurately represent the entities, relationships, and attributes relevant to the questionnaire data. By defining appropriate tables, columns, and relationships, the schema ensures data integrity and minimizes redundancy to the overall effectiveness and performance of the system.

3.1.5 Data Types and Optimization

The questionnaire data may encompass various data types, including text, numerical values, dates, and categorical values. Selecting the appropriate data types for each attribute optimizes storage efficiency and facilitates accurate representation. Additionally, employing indexing techniques, such as primary keys, foreign keys, and composite indexes, enhances data retrieval speed by minimizing the need for full-table scans and enabling efficient filtering and sorting operations.

3.1.5.1 Querying Capabilities

Efficient data retrieval is critical for the subsequent analysis and estimation processes. The chosen RDBMS should provide robust querying capabilities to handle complex queries and aggregations effectively. Employing optimized SQL queries, utilizing appropriate indexes, and leveraging the RDBMS's query optimization mechanisms can significantly improve query performance. Ensuring that the database schema is well-aligned with the expected query patterns enables fast and accurate retrieval of the required data.

3.1.5.2 Security and Data Privacy

Given the sensitive nature of health-related data, ensuring data security and privacy is of utmost importance. Implementing appropriate security measures, such as encrypted connections, user authentication, and access control mechanisms, protects the confidentiality and integrity of the collected data. Compliance with relevant data protection regulations, such as the General Data Protection Regulation (GDPR), is

essential to safeguard the privacy rights of individuals and maintain ethical data handling practices.

3.2 System Specification and Questionnaire Implementation

As previously mentioned, we chose to use the programming language Python [14] to build our system and its tools to estimate hard-to-reach subpopulations. We could have just as easily chosen R programming language [15] or Matlab [16] for their extensive applications in mathematical, especially since there is a ready-made NSUM package in R-programming language, but we prefer Python because of its ease of use. Also, Python is one of the most popular programming languages in the world with a large community that can aid us in our thesis, which is another reason why we chose Python.

To code the necessary tools for the network scale-up method and the different models, we took the NSUM package that was readily coded in R by Maltiel [2] language and translated it into Python. The justifications for this roundabout way were mentioned above. We also chose this manner of execution because of "Django" [17], a very popular framework in Python, which we will go into more detail in the next section.

The implementation of our system is not different from most other web-based database system applications. Similarly, we have a user who enters the web application and fills in a questionnaire. This information is then stored in a database which can then be processed and analyzed where NSUM can be applied. An overview of the system and the flow of information is presented in Figure [3.1] below.

Figure 3.1: General flow of information of system

This can be summarized in the following steps.

1. User enters web application and answers questionnaire
2. Data from the web application is then sent to the database
3. Data is then processed by the admin of the system using NSUM to get estimates

The first step is dependent on the country and its economic status, i.e., some countries may not have as easy access to computers and are more prone to use mobile

phones. This step is one of the reasons why NSUM is also chosen over other estimation methods, which will be covered more in a later section. The second step is developed using Django, where we can build the entire architecture within one programming language, i.e., both the web application and the database. Other reasons why we chose Django will be covered in the next section.

3.2.1 Django Framework

Django, a widely recognized Python framework, plays a crucial role in facilitating the efficient and reliable development of web applications. By adhering to the well-established Model-View-Controller (MVC) architectural pattern, Django not only ensures seamless code organization but also prioritizes essential software engineering principles such as code reusability, modularity, and the principle of "Don't Repeat Yourself" (DRY) [17].

At its core, Django's design philosophy centers around promoting rapid web development without compromising the security and maintainability of the resulting applications [17]. With its rich set of built-in features and libraries, Django empowers developers to swiftly create dynamic and interactive websites, significantly reducing the time required to translate conceptual ideas into function implementations.

The adherence to the MVC architectural pattern lies at the heart of Django's robustness. By enforcing a clear separation of concerns, Django distinguishes between the presentation layer (View), the business logic, and data manipulation (Model), and the user interaction (Controller). This separation enables developers to handle each component independently, facilitating code maintenance, readability, and reusability. Consequently, Django-based applications can scale effortlessly and accommodate evolving requirements without sacrificing stability or introducing unnecessary complexities.

The concept of code reusability is a fundamental principle in Django and it greatly enhances the development process. By encouraging developers to encapsulate common functionalities as reusable components, Django eliminates redundant code and promotes modular design. The modular approach not only enhances the overall project structure but also allows for rapid prototyping and iterative development. Developers can leverage Django's extensive collection of reusable models, known as "apps", to incorporate pre-built components into their projects, thereby accelerating development cycles and fostering a collaborative ecosystem within the Django community [17].

In line with the DRY principle, Django advocates for avoiding code duplication through abstraction and automation. Django's built-in features, such as its powerful object-relational mapping (ORM) system, alleviate the need for developers to write boilerplate code when interacting with databases. By automatically generating the underlying SQL queries, Django eliminates the repetitive and error-prone process of manually crafting database interactions, enabling developers to focus on higher-level functionalities. Furthermore, Django's template engine facilitates the separation of content and presentation, ensuring that developers can define reusable templates

and avoid duplicating markup or layout across different pages or views.

3.2.1.1 Alternatives

Django and Flask are the two most popular Python frameworks, each with its own strengths and characteristics. Flask is known for its simplicity and minimalistic approach, on the other hand, Django offers a more comprehensive and feature-rich framework. In section 3.2.1.2 we are going to discuss the advantages that Django offers with respect to Flask.

3.2.1.2 Why Django?

- **Batteries**; one of the key advantages of Django is its batteries included philosophy. Django offers a wide range of pre-built functionalities and components, including an ORM (Object Relational Mapping), authentication system, admin interface, and form handling. These features come bundled with Django, allowing developers to start building complex applications immediately. On the other hand, Flask takes a more minimalist approach, providing developers only with essential components, requiring them to choose and integrate additional libraries for various functionalities.

- **Robust security features**; it incorporates several built-in security features that help developers mitigate common security risks. The authentication system provides secure user authentication and password hashing, guarding against unauthorized user access and password breaches. It also includes protection against cross-site scripting (XSS) and cross-site request forgery (CSRF) attacks by implementing measures like automatic HTML escapes and CSRF tokens. Additionally, Django encourages the use of prepared statements and query parameterization to prevent SQL injections.

- **Strong community support**; Django has thriving community support with a vast ecosystem of packages, plugins, and extensions. The community actively participates in the development of the framework, ensuring regular updates, bug fixes, and security patches. This extensive ecosystem provides developers with a wide range of choices to enhance their projects. Conversely, even though Flask has a strong community as well, its ecosystem is smaller compared to Django's, requiring developers to search for and integrate additional components from various sources.

- **Scalability**; when it comes to handling large and complex web applications Django excels. Django's strong emphasis on modularity, separation of concerns, and reusable components makes it well-suited for handling complex codebases. It provides a strong and scalable structure that allows developers to build and maintain large applications with ease. Flask, while being lightweight and more flexible, may require more manual configuration and organization as the application grows in size and complexity.

3.3 Database Implementation

In this section, we are going to explain the implementation of the database in our project. The database has the task of collecting and storing the data needed for the project. Without a well-designed and well-implemented database, issues with data quality, data management, and data analysis can arise. This may negatively impact the validity and reliability of the study. A database is pivotal for storing and organizing large amounts of data collected from the sources. It allows us to manage data efficiently, retrieve data easily, and ensure data consistency. Additionally, it can help prevent data loss, errors, and inconsistencies, which occur if data is stored in multiple files or spreadsheets.

In our specific case, the database is collecting data from surveys, and passing the data to the network scale-up methods. The accuracy and completeness of the data stored in the database are critical for ensuring that the results from the NSUM methods are valid and reliable. If the database is not well-designed and implemented, it may lead to issues such as missing data, data errors, or data inconsistencies. which can reduce the accuracy and reliability of the results.

Overall, the database is a fundamental part of our study because it enables us to manage the data effectively, ensuring that the data collected is reliable, consistent, and of high quality, which is essential for drawing accurate conclusions and making meaningful recommendations.

3.3.1 Database Design

In this subsection, we are going to talk about the design choices we have made for the database.

3.3.1.1 Data Model

The data model for the questionnaire is a relational data model. In this model, the data is organized into tables and the relationships between them are defined by the use of foreign keys. In our case, the main tables are:

- Survey Table: this table would contain information about each survey created, including a unique survey ID, the name of the survey, a description of the survey, and the date it was created.
- RadioQuestion Table: a radio question typically allows respondents to choose a single option from a predefined set of choices. This table would contain information about each radio question in the questionnaire, including a unique question ID, the text of the question, and a foreign key to the Survey Table to associate the question with the appropriate survey.
- RadioAnswer Table: this table would contain information about each radio answer for each radio question, including a unique answer ID, the text of the answer, and a foreign key to the RadioQuestion table to associate the answer with the corresponding question.

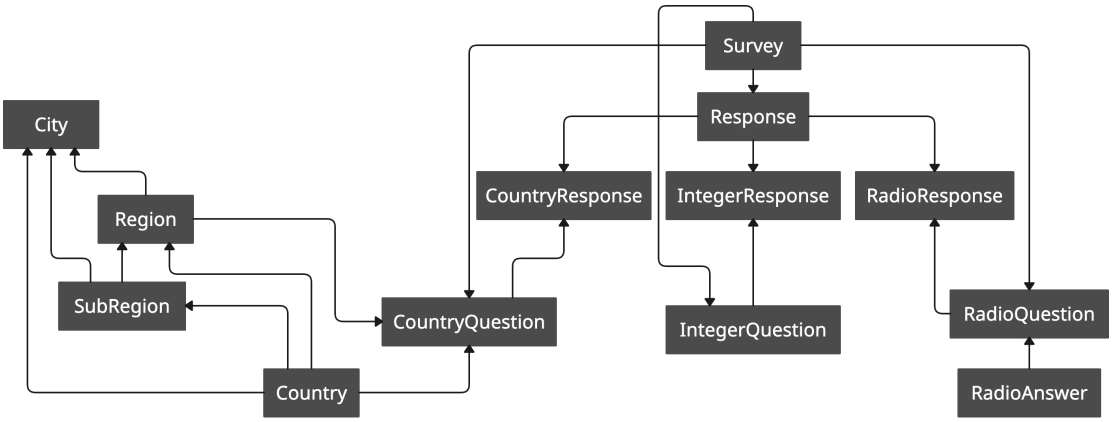


Figure 3.2: Database design

- IntegerQuestion Table: this table would contain information about each integer question in the survey, including a unique question ID, the text of the question, and a foreign key to the Survey table to associate the question with the appropriate survey.

- CountryQuestion Table: this table would contain information about each country question in the survey, including a unique question ID, the text of the question, and a foreign key to the Survey table to associate the question with the appropriate survey.

- Response Table: this table would contain information about each survey response, including a unique response ID, and a foreign key to the Survey table to associate the response with the appropriate survey.

- RadioResponse Table: this table would contain information about each radio response, including a foreign key to the Response table to associate the response with the correct response, a foreign key to the RadioQuestion to associate the response with the appropriate question, and a foreign key to the RadioAnswer table to associate the response with the appropriate answer.

- IntegerResponse Table: this table would contain information about each integer response, including a foreign key to the Response table to associate the response with the right response, a foreign key to the IntegerQuestion table to associate the response with the appropriate question, and an integer value to store the integer response.

- CountryResponse Table: this table would contain information about the country response, including a foreign key to the Response table to associate the response to the correct response, a foreign key to the CountryQuestion table to associate the response to the appropriate question, a foreign key to the Country table to associate the response with the appropriate country, and a foreign key to the Region table to associate the response to the appropriate region.

3.3.2 Schema Design

As previously stated, the database used in this study was designed to collect and store data from a survey and subsequently pass this data to a network scale-up method. The schema design outlines the structure of the database, including tables, columns, and relationships.

The database, as seen in Figure 3.1, consists of five main tables: Survey, RadioQuestion, RadioAnswer, IntegerQuestion, and Response. The Survey table contains the basic information about each survey, including the survey name and description. The RadioQuestion table contains the text of each radio question in the survey, while the RadioAnswer table contains the text of each possible answer for each radio button question. Similarly, the IntegerQuestion table contains the text of each integer in the survey, while the Response table contains the response data for each survey.

In addition to the five main tables, there are three more tables: IntegerResponse, CountryQuestion, and CountryResponse. The IntegerResponse table stores the value of each response for each integer question in the survey. The CountryQuestion table contains the text of each country question in the survey, while the CountryResponse table stores the response data for each country question.

The relationships between the tables are as follows: each survey can have many radio buttons and integer questions, while each radio button and integer question can only belong to one survey. Each radio button question can have many possible answers, while each possible answer can only belong to one radio button question. Each response belongs to one survey and can have many radio buttons and integer responses. Each radio button response belongs to one response and can only belong to one radio button question and one possible answer. Each integer response belongs to one response and can only belong to one integer question. Finally, each country's response belongs to one response and can only belong to one country question, one country, and one region.

The schema design was chosen based on the requirements of the research question and the structure of the survey. It allows efficient storage and retrieval of survey data, as well as the ability to easily pass these data to the network scale-up method for analysis. The use of separate tables for different question types allows for more efficient storage and retrieval of data, while the use of relationships between tables ensures data integrity and accuracy. Overall, the schema design helps to organize and structure the data in a way that is conducive to the research question.

3.3.3 Data Types

When designing a database, choosing the appropriate data types for each field is critical to ensure that the data is stored accurately ad efficiently. In this study, the following data types were chosen:

- CharField; this data type was used for fields that require a small to medium amount of text, such as survey question text and answer choices. CharField

allows for the storage of up to 255 characters.

- TextField; this data type was used for fields that require a large amount of text, such as survey descriptions. TextField allows for the storage of long text strings
- IntegerField: This data type was used for fields that require whole numbers, such as integer survey questions. IntegerField allows for the storage of integer values from -2147483648 to 2147483647. In the scope of this study, only positive numbers are allowed.
- ForeignKey: This data type was used to establish relationships between tables, such as the relationship between survey questions and their associated survey. ForeignKey allows for the storage of a reference to a primary key field in another table.
- DateField: This data type was used for fields that require dates, such as survey start and end dates. DateField allows for the storage of date values from January 1, 1900, to December 31, 9999.
- DateTimeField: This data type was used for fields that require both a date and time value, such as survey response submission times. DateTimeField allows for the storage of date and time values from January 1, 1900, to December 31, 9999.

The choice of data types was based on the nature of the data being stored and the requirements of the study. For example, CharField and TextField were chosen for fields that require text because they allow for the storage of a variable amount of text, while IntegerField was chosen for fields that require integers because it ensures that the data is stored as whole numbers. DateField and DateTimeField were chosen for fields that require dates and times because they allow for the storage of date and time values in a format that can be easily compared and manipulated.

3.3.4 Database Management System (DBMS)

For our study, we chose to use a relational DBSM, specifically SQLite, to manage our database. SQLite is a widely used open-source relational DBSM that provides efficient and lightweight data management [18]. It is particularly suitable for small to medium-sized databases, making it a good choice for our study. SQLite is also known for its reliability and stability, as well as its compatibility with a variety of programming languages.

One of the main benefits of using SQLite for our study is that it requires minimal setup and configuration. It does not require a separate server process, as the database engine is integrated into our application. This makes it easy to deploy and manage our database, particularly for smaller projects or single-user applications.

Another benefit of SQLite is that it supports a wide range of data types, including integers, floating-point numbers, strings, dates, and more. This allows us to store

different types of data in our database and to choose the appropriate data types for each field, based on the characteristics of the data.

Overall, SQLite is a reliable and efficient choice for our study's data management needs. Its lightweight design, compatibility with various programming languages, and support for a wide range of data types make it well-suited to our study's requirements.

3.3.5 Normalization

Normalization is an important process in database design, which aims to organize the data in a way that minimizes redundancy and dependency issues. The ultimate goal of normalization is to enhance the accuracy and consistency of the data in the database.

In the given Django implementation, we can see that the Survey model is the central model, with all other models having a foreign key relationship with it. This design is in compliance with the first normal form (1NF) of normalization, which requires that all columns in a table contain atomic values (i.e., no repeating groups or arrays). In this case, the Survey model has only one instance for each survey, with all other models referencing it.

The RadioQuestion, RadioAnswer, IntegerQuestion, and CountryQuestion models are all related to the Survey model via a foreign key, and each of them represents a single question in the survey. This design is in compliance with the second normal form (2NF) of normalization, which requires that the table is in 1NF and that all non-key attributes in the table are fully dependent on the primary key. In this case, the Survey model is the primary key, and all other models reference it as a foreign key.

The RadioResponse, IntegerResponse, and CountryResponse models are related to the Response model via a foreign key, and each of them represents a single response to a question in the survey. This design is in compliance with the third normal form (3NF) of normalization, which requires that the table is in 2NF and that all non-key attributes are independent of each other. In this case, the Response model is the primary key, and each of the other models has a foreign key reference to it.

Furthermore, the use of the pre-existing Country and Region models from the cities_light library instead of defining new models for these entities is another normalization technique. By using these existing models, we are avoiding the creation of redundant tables in our database, which enhances data consistency and reduces the likelihood of data anomalies.

Overall, the normalization of the database in this Django implementation ensures that the data is organized in a way that minimizes redundancy and dependency issues, leading to improved accuracy and consistency of the data.

3.4 Network Scale-Up Method

There are many methods available to estimate subpopulations with specific traits. However, some methods may not be suitable for certain countries depending on their characteristics, e.g. economic status. For example, the enumeration methods can be extensive and require a lot of resources by involving months of fieldwork and can still fail to produce results [1].

In this study, the network scale-up method, described in [2.2] was selected as the primary approach for estimating the number of infected people within a given population. The network scale-up method is a statistical technique that leverages social network data to estimate the prevalence of hard-to-reach or stigmatized populations, such as individuals affected by infectious diseases. This method has gained significant attention in epidemiological research due to its ability to provide estimates when traditional survey methods fall short.

3.4.1 Rationale for Choosing NSUM

Several factors influenced the decision to utilize the network scale-up method for estimating the number of infected individuals

- **Limited Accessibility to Data**: In situations where data collection is challenging due to various constraints, such as privacy concerns or stigmatization of the target population, the network scale-up method offers an alternative approach. By relying on social network connections rather than direct observations, it provides a way to indirectly estimate the prevalence of a specific characteristic, such as infection status.
- **Expanding Beyond Traditional Surveillance**: Traditional surveillance methods, such as clinical case reporting or seroprevalence surveys, may not capture the full extent of infectious diseases, particularly among populations that are marginalized or hidden. The network scale-up method allows for the inclusion of individuals who may not be captured by these conventional methods, leading to a more comprehensive estimate of the infected population.
- **Utilizing Social Connections**: The network scale-up method acknowledges the social nature of human interactions. By considering the connections between individuals, it leverages the assumption that people tend to have accurate knowledge about the characteristics of their close contacts. This provides an avenue for estimating the prevalence of infections by extrapolating information from network members to the wider population.

3.4.2 Alternatives to NSUM

While the network scale-up method (NSUM) provides valuable advantages, it is crucial to consider alternative approaches for estimating the number of infected individuals. These alternatives serve different purposes and come with their own

set of strengths and limitations, and understanding their context and applicability is essential:

- **Direct Sampling**: This method involves randomly selecting individuals from the target population and directly testing for the presence of the infection. Direct sampling can offer accurate prevalence estimates and serves as a real-time data collection method. However, it may be time-consuming, expensive, and logistically challenging, especially for large populations or diseases with low prevalence rates. Direct sampling can be a viable alternative to NSUM when immediate and precise prevalence data are needed.
- **Model-Based Approaches**: Various mathematical models, such as compartmental models (e.g., SIR or SEIR models) or agent-based models, simulate the spread of infectious diseases within a population. These models rely on assumptions about disease transmission dynamics and population characteristics to estimate the number of infected individuals. While powerful, they often require extensive data and complex parameterization. Model-based approaches are suitable alternatives when a detailed understanding of disease dynamics and potential interventions is the primary goal.
- **Serological Surveys**: Seroprevalence surveys involve testing blood samples from a representative subset of the population to detect specific antibodies to the infection. This method provides insights into past exposure and prevalence of the disease but may not capture current infections or offer real-time estimates. Serological surveys are valuable alternatives when understanding historical exposure and immunity levels is critical.

These alternatives are not necessarily interchangeable with NSUM but offer distinct approaches to studying disease prevalence. Direct sampling provides immediate, real-time data, model-based approaches offer a nuanced understanding of disease dynamics, and serological surveys contribute insights into historical exposure. The choice among these methods depends on the specific goals of the study, available resources, and the desired level of precision in estimating the number of infected individuals. By adopting NSUM, this study aims to leverage social network data to gain unique insights into disease prevalence within the population [2.2].

3.5 Ethical considerations

As technology continues to advance and shape our modern society, ethical considerations play a pivotal role in the development and deployment of software. While exploring the potential and the advantages of Django for web application developments, it is crucial to address the ethical implications associated with its usage. It is important to note that the ethical landscape surrounding technology is dynamic and constantly evolving. The ethical considerations outlined in this section are meant to be a starting point for discussion and critical analysis. However, they should not be exhaustive or definitive. This section aims to explore specific ethical considerations related to data privacy and protection, securities vulnerabilities, accessibility and inclusion, algorithmic bias, open-source collaboration, and social responsibil-

ity. Through a thoughtful examination of these topics, we aim to shed light on the multi-faced ethical landscape surrounding software development and encourage the adoption of ethical practices that prioritize the well-being interests of users and the wider community.

3.5.1 Data Privacy and Protection

Data privacy and protection are of vital importance in the context of our project. Django provides several features and mechanisms to handle user data. The utilization of Django's ORM (Object-Relational Mapping) [3.2.1.2] layer enables developers to work with relational databases effectively and efficiently. However, the ethical implications surrounding the use of aggregated relation data deserve careful consideration. ARD refers to the process of collecting, combining and analyzing multiple data points to derive insights and patterns at a broader scale, as described in the section [2.3]. While aggregated data can provide valuable information for various purposes, it also raises concerns related to individual privacy and the responsible use of personal data.

When working with data collection in Django, it is pivotal to ensure that privacy is adequately protected. We must consider how data is anonymized, aggregated, and stored to prevent the re-identification of individuals. Careful consideration should be given to data retention policies, ensuring that aggregated data are retained only as long as necessary and securely disposed of when no longer needed. That is why we resort to the use of ARD. The aggregation allows for hiding any information about the respondent. In the scope of this project, we decided to aggregate up to the region of each country to ensure anonymity of the respondents while still being being able to draw relevant conclusions.

Transparency and informed consent are also important ethical considerations when dealing with ARD. Users should be made aware of how their data is collected, anonymized, and used in aggregate. Providing clear and accessible policies and obtaining informed consent from users helps establish trust and respect for individual privacy rights.

Additionally, ethical considerations arise when determining the boundaries of data aggregation. Striking a balance between extracting meaningful insights from aggregated data and preserving the privacy and anonymity of individuals is significant. Developers must be cautious not to disclose sensitive or personally identifiable information inadvertently, even in aggregated form.

By recognizing the aforementioned ethical considerations surrounding ARD, developers and organizations can ensure that privacy rights are respected, data is used responsibly, and potential risks to individuals are mitigated.

3.5.2 Security Vulnerability and Responsible Disclosure

When encountering security vulnerabilities within the Django framework, it is vital to approach them responsibly. Responsible disclosure involves following a set of

ethical guidelines to report vulnerabilities to the Django development team or the appropriate security channels without exploiting or publicly disclosing the vulnerability before a fix is found. Responsible disclosure helps protect the user and the community by ensuring that vulnerabilities are patched before malicious actors can exploit them. It is essential to stay informed about security best practices, e.g., to participate in security discussions within the Django community and to remain vigilant for potential bugs. By embracing these ethical obligations, one can make valuable contributions to the security and integrity of the framework.

3.5.3 Accessibility and Inclusivity

Adhering to web accessibility standards and guidelines, Web Content Accessibility Guidelines (WCAG), is crucial when developing web applications like in this project. WCAG provides a comprehensive set of guidelines for making web content accessible to people with disabilities. These guidelines cover aspects such as permeability, operability, understandability, and robustness, offering a framework to ensure that web applications are usable by individuals with various impairments. Providing accessibility and inclusivity in web applications is an ethical responsibility. Accessibility not only benefits individuals with disabilities but also improves the user experience for all users. Since we are limited by time, these important aspects are going to be included in future research and work.

3.5.4 Bias and Fairness in Algorithms

Using Django's features such as the ORM layer, described in Section [3.2.1.2] can set off unwanted ethical concerns. Bias can be introduced when collecting, processing, or analyzing data within this layer. For example, if historical data used for training models within Django's ORM reflects biased or discriminatory patterns, it can perpetuate unfair treatment or reinforce existing inequalities. Testing and mitigating biases is a vital ethical responsibility when using Django. Developers should incorporate robust testing methodologies to identify and address potential biases in their applications. This includes evaluating the quality and representativeness of training data, analyzing the impact of algorithmic decisions on different user groups, and employing techniques such as fairness-aware machine learning algorithms to mitigate biases. Additionally, conducting regular audits and involving diverse perspectives in the testing process can help uncover biases that may have been overlooked.

3.5.5 Social Impact and Responsible Use

The scope of this study is to build a reliable data collection tool and estimate the number of infected people in hard-to-reach subpopulations. We have to consider the possibility of uncovering a strong presence of HIV in populations where there is a significant stigma associated with the disease. This stigma can contribute to discrimination, social exclusion, and negative attitudes toward individuals with underlying diseases and the study as well. There may be instances where the population might reject the possibility of carrying out the questionnaire for fear of rejection from family and friends, limited access to healthcare, and overall poor quality of life. We

must establish trust and build a relationship with the hard-to-reach populations. Engaging with local organizations, such as universities and hospitals, is crucial for gaining acceptance from the local community. Additionally, the research can identify the prevalence of the disease, by understanding the extent of the problem local authorities could develop targeted interventions and customized policies. This would contribute to reducing health disparity and promote equitable access to healthcare for all individuals, irrespective of their social and geographical circumstances.

4

Results

This chapter describes the tests performed to prove that our implementation works and provides a baseline estimation for infected HIV people. We are going to explore each model implemented in this field. We estimate all models using Markov Chain Monte Carlo (MCMC). For all models, μ and σ were sampled using closed-form Gibbs steps while we used random walk Metropolis steps with normal proposals for other parameters, as shown in Maltiel et al. 2015 [2].

Accurate estimation of the number of individuals infected with a disease is extremely important. This section presents the results of a study that aims to estimate the number of people infected with HIV using a web-based questionnaire and scale-up methods. The study focuses specifically on African countries, with a particular emphasis on Rwanda and Uganda, as future collaborations with official organizations are planned. The prevalence of HIV in sub-Saharan Africa is a significant public concern, and obtaining an accurate estimation of the infected population is crucial for effective policy-making, resource allocation, and intervention strategies.

In this section, we present the result of our study aimed at creating a data collection tool and estimating the number of people infected by a disease, HIV in this particular case. For the evaluation of the system, we will consider metrics beyond the performance. Other relevant metrics may include accessibility. Additionally, the methods for the estimation are going to be tested separately in section [4.3].

The significance of this study lies in its contribution to the field of HIV epidemiology and the development of improved methods for estimating HIV prevalence. By comparing and evaluating different network scale-up methods, we aim to improve our understanding of the strengths and limitations of these approaches in estimating the number of HIV-infected individuals in African countries. As a result, the research question guiding this study is: "How do different network scale-up methods compare in estimating the number of HIV-infected people in Rwanda and Uganda?". By exploring this question, we seek to identify the most reliable and accurate method for estimating HIV, and other diseases, the prevalence in resource-limited settings, where traditional survey-based data collection may be time-consuming and challenging.

Furthermore, our hypothesis posited that NSUM would yield estimates that differ from the official HIV prevalence data obtained from official sources. We anticipated that each method would introduce specific biases and assumptions that could in-

fluence the accuracy of the estimates By analyzing and comparing the results, we aim to gain insights into the performance of these methods and their applicability in estimating epidemics predominance in the interested countries.

By addressing these research objectives and evaluating the hypothesis, this study contributes to the existing knowledge of estimating HIV prevalence and provides valuable information to policymakers, public health professionals, and researchers working in the field of HIV/AIDS in African countries.

In the next subsection, we will present the official HIV prevalence data obtained from official sources in Rwanda and Uganda, establishing a baseline for comparison with the estimates derived from the network scale-up methods.

4.1 System Performance Evaluation

The performance evaluation of a web-based questionnaire must ensure a seamless user experience and maximize the efficiency of data collection. In this section, we present the results obtained from the Google Lighthouse scoring calculator, which assesses various metrics related to the performance of the website [19]. The evaluation was conducted separately for desktop and mobile platforms, providing a comprehensive understanding of the questionnaire performance across different devices.

The Lighthouse Score Calculator is a tool provided by Google to evaluate the performance and quality of a website or web application. It analyzes various metrics and provides a score based on the performance, accessibility, best practices, and search engine optimization (SEO) of the site. Here are the main metrics used by the Lighthouse Score Calculator and their units of measure:

- Performance (0-100); this metric measures the overall performance of the website based on various factors such as page load time, resource usage, and rendering speed [20]. Upon the compilation of performance metrics by Lighthouse, mostly denoted in milliseconds, the subsequent step involves the conversion of each raw metric value into a metric score within the range of 0 to 100. This transformation is executed by determining the metric value's position within the Lighthouse scoring distribution (a percentile), which is derived from a log-normal distribution based on actual performance metrics obtained from real website performance data present in the HTTP Archive [21]. The total final score is a weighted average of the below metric scores. For each metric, we indicate in the parentheses the unit and the weight used.
  - First Contentful Paint (FCP) (seconds, 0.10); FCP measures the time from when the page starts loading to when any part of the page's content is rendered on the screen. This includes any element (text, image, background image).
  - Largest Contentful Paint (LCP) (seconds, 0.25): LCP measures the time taken for the largest content element (such as an image or a block of text) to be rendered on the screen.

  - First Input Delay (FID) (milliseconds, 0.10): FID measures the time delay between a user's first interaction with the website (such as clicking a button) and the website's response.
  - Cumulative Layout Shift (CLS) (0.15): CLS measures the visual stability of the website by quantifying how much the content layout shifts while the page is loading. It is represented as a score between 0 and 1, where lower values indicate less layout shifting.
  - Total Blocking Time (TBT) (milliseconds, 0.30): TBT measures the total amount of time during page load when the main thread was blocked and unable to respond to user input.
  - Time to Interactive (TTI) (seconds, 0.10): TTI measures the time taken for the website to become fully interactive and responsive to user input.

- Accessibility (0-100): The accessibility score evaluates how well a website conforms to web accessibility standards, ensuring that people with disabilities can navigate and interact with the site effectively [22]. It measures factors such as proper use of headings, alternative text for images, keyboard navigation support, and color contrast.

  The Lighthouse Accessibility score is a composite measure derived from various accessibility audits [22], with each audit assigned a specific weight based on axe user impact assessments [23]. Unlike Performance audits, accessibility audits are binary, resulting in a pass-or-fail outcome. Partial compliance with an accessibility audit does not accrue points. For instance, if some buttons on a page possess accessible names while others do not, the page receives a score of 0 for the "Buttons do not have an accessible name" audit. The following list illustrates the weighting assigned to some accessibility audits, the sum of each weight adds to 100. Audits with higher weights exert more influence on the overall score.

  - Alternative Text (Alt Text), Weight: 10; Images on a website should have descriptive alternative text (alt text) associated with them. Alt text provides a textual description of the image content, enabling users who cannot see the images to understand their context.
  - Semantic HTML, Weight: 7; Proper use of semantic HTML elements is crucial for accessibility. This includes using appropriate tags for headings, lists, tables, and form elements, ensuring the structure and purpose of the content are conveyed accurately.
  - Keyboard Navigation, Weight: 7; Websites should be navigable using only a keyboard, without relying on mouse or touch interactions. Keyboard focus should be appropriately managed, allowing users to navigate through interactive elements and menus easily.
  - Color Contrast, Weight: 7; Sufficient color contrast between foreground text and background colors is essential for readability, especially for people with visual impairments. The accessibility score checks if the color

contrast meets the recommended standards for legibility.

- ARIA Roles and Attributes, Weight: 1; Accessible Rich Internet Applications (ARIA) roles and attributes can enhance the accessibility of dynamic web content and interactive elements. The score considers whether these ARIA roles and attributes are implemented correctly to improve accessibility for assistive technologies.
- Form Accessibility, Weight: 7; Web forms should be designed to be accessible, including proper labeling of form fields, providing clear instructions, and indicating any errors or required fields.
- Heading Structure, Weight: 3; Properly structured headings help users navigate and understand the content hierarchy. The accessibility score checks if headings are used logically and hierarchically throughout the web page.

- Best Practices (0-100): The best practices score assesses whether a website follows industry-standard best practices for web development and performance [24]. Lighthouse evaluates various criteria to ascertain compliance, such as optimized code structure, secure protocols (HTTPS), proper use of meta tags, and the absence of deprecated HTML features. The computation of the final score out of 100 incorporates a formula that integrates these factors. The specific details of the formula involve assigning weights to each criterion based on its significance in adhering to best practices. The cumulative impact of these weighted factors contributes to the overall score, reflecting the extent to which the webpage aligns with contemporary web development standards. Below is a list of some of the audits.
  - Optimized Code and Assets: The score evaluates whether the website's HTML, CSS, and JavaScript code are optimized for performance and maintainability. This includes considerations such as minification (removing unnecessary characters) and compression of code and assets to reduce file sizes and improve loading speed.
  - HTTPS Usage: Websites that implement secure connections using HTTPS (Hypertext Transfer Protocol Secure) are favored by search engines and considered more trustworthy. The best practices score checks if the website is configured to use HTTPS properly.
  - Proper Use of Meta Tags: Meta tags provide metadata about a web page, such as the page title, description, and author. The score verifies if these meta tags are implemented correctly and helps search engines understand the page's content and purpose.
  - Absence of Deprecated Features: The score identifies the usage of deprecated HTML features or obsolete practices that are no longer recommended. It encourages using modern and supported web technologies to ensure compatibility, security, and future-proofing.
  - Mobile-Friendliness: With the growing number of mobile users, websites

should be designed and optimized for a seamless mobile experience. The best practices score considers factors such as responsive design, viewport configuration, and mobile-friendly interactions.

  - Accessibility Considerations: While there is a separate accessibility score, the best practices score also takes into account some accessibility-related practices, such as the use of proper headings, semantic HTML, and alternative text for images.

  - Performance Considerations: Although performance is primarily evaluated in the performance score, the best practices score may include some performance-related aspects as well. For example, it may check if resources are efficiently cached, critical resources are inlined, or render-blocking resources are minimized.

- Search Engine Optimization (SEO) Score (0-100): The SEO score assesses the effectiveness of a website's optimization for search engines, with the ultimate goal of enhancing its visibility in search results [25]. This comprehensive evaluation takes into account various factors, including meta tags, structured data markup, page titles, headings, URL structure, mobile-friendliness, and site performance. The final score is calculated by analyzing the collective impact of these elements, with higher scores indicating better optimization practices. The assessment considers the alignment of meta information, the clarity and relevance of page titles and headings, the implementation of structured data for enhanced search result snippets, the user-friendly nature of the URL structure, and the overall performance and mobile responsiveness of the website. This holistic approach provides a nuanced understanding of the website's SEO strengths and areas for improvement, resulting in a numerical score that reflects its overall search engine optimization effectiveness.

  - Meta Tags: The score analyzes the presence and quality of meta tags such as the title tag and meta description. These tags provide concise and relevant information about the web page's content and help search engines understand and index the page accurately.

  - Heading Structure: Proper usage of headings (H1, H2, etc.) helps search engines understand the content hierarchy and relevance. The score evaluates if headings are used appropriately and reflect the structure and topic of the page.

  - URL Structure: Search engines prefer clean and descriptive URLs that reflect the page's content. The SEO score checks if the URLs are optimized with relevant keywords and are readable and user-friendly.

  - Keyword Usage: The score assesses the strategic use of relevant keywords throughout the content, including in headings, paragraphs, and image alt text. It ensures that the keywords are used naturally and not excessively (keyword stuffing).

  - Mobile-Friendliness: Since mobile devices account for a significant por-

Metric	Score ↑
Performance	100
Accessibility	81
Best Practices	83
SEO	80

Table 4.1: Performances of the desktop website.

tion of web traffic, the SEO score considers the mobile-friendliness of the website. It evaluates if the site is responsive, has a mobile-friendly design, and provides a seamless user experience on different screen sizes.

- Page Speed: The loading speed of a web page is an important ranking factor. The score considers the page's performance metrics, such as First Contentful Paint (FCP), Largest Contentful Paint (LCP), and Total Blocking Time (TBT), to assess the overall speed and responsiveness of the website.
- Structured Data Markup: The presence of structured data markup, such as Schema.org, helps search engines understand and display the content in more meaningful ways. The score checks if structured data is implemented correctly to enhance search engine visibility.
- Mobile Usability: The score assesses the mobile usability of the website, including factors such as touch-friendly elements, font sizes, and clickable elements' spacing. A website that provides a smooth and user-friendly experience on mobile devices is favored by search engines.

4.1.1 Desktop Performance Evaluation

The desktop performance evaluation yielded the scores in Table 4.1

The "Performance" metric measures the overall speed and responsiveness of the website. A score of 100 indicates that the questionnaire website performs exceptionally well in terms of loading speed, interactivity, and visual stability. This high score suggests that users accessing the questionnaire on desktop devices will experience minimal delays, enabling them to navigate through the website efficiently.

The "Accessibility" metric gauges the website's compliance with accessibility guidelines, ensuring that individuals with disabilities can use the platform effectively. Although the questionnaire website obtained a score of 81, there may be areas for improvement to ensure inclusivity and accommodate users with different accessibility needs.

The "Best Practices" metric evaluates adherence to recommended development practices, security measures, and overall website optimization. With a score of 83, the questionnaire website demonstrates a commendable implementation of best practices, indicating a robust and well-structured development approach.

Metric	Score ↑
Performance	99
Accessibility	81
Best Practices	83
SEO	67

Table 4.2: Performances of the mobile website.

Lastly, the “SEO” metric focuses on optimizing the website’s visibility in search engine results. A score of 80 indicates that the questionnaire website is well-optimized for search engines, potentially leading to higher organic traffic and improved discoverability.

4.1.2 Mobile Performance Evaluation

We also evaluated the mobile version of the system to get a more comprehensive assessment. The mobile performance evaluation resulted in Table [4.2]

The “Performance” metric, with a score of 99, showcases exceptional mobile performance for the questionnaire website. This indicates that the website is highly optimized for mobile devices, providing users with a fast and responsive experience on their smartphones or tablets.

Similar to the desktop evaluation, the "Accessibility" metric scored 81, implying that further enhancements can be implemented to ensure inclusivity and accessibility for users with disabilities.

The “Best Practices” metric, with a score of 83, demonstrates consistent adherence to industry standards and recommended development practices, leading to a well-optimized and secure mobile website.

Lastly, the “SEO” metric obtained a score of 67, indicating potential areas for improvement in optimizing the questionnaire website for search engines on mobile platforms. Enhancing the website’s mobile SEO practices can potentially enhance its visibility and reach a wider audience.

Overall, the performance evaluation of the web-based questionnaire using the Google Lighthouse scoring calculator reveals excellent results in terms of performance, best practices, and accessibility. However, there is room for improvement in mobile SEO and certain aspects of accessibility to ensure a fully inclusive and optimized user experience. Addressing these areas will enhance the usability and effectiveness of the questionnaire website, ultimately contributing to the accuracy and reliability of the collected data.

Note: It is important to note that the Google Lighthouse scoring calculator evaluates the website based on predefined metrics and guidelines. While these scores provide valuable insights, they should be considered alongside other usability testing methods and user feedback to obtain a comprehensive understanding of the questionnaire website’s performance and user experience.

4.2 Official HIV Prevalence Data

In this section, we are going to present the official HIV data obtained from official sources in Rwanda and Uganda. These figures serve as the reference point for comparing and evaluating the estimates derived from the network scale-up methods.

According to data sourced from the United Nations Programme on HIV/AIDS (UNAIDS), the official HIV prevalence rate in Rwanda is reported as 3% [26] or in other words about 230 000 people. Although the prevalence rate in Rwanda is comparatively lower than in many other sub-Saharan countries, HIV remains a significant cause of mortality for people aged 5 years or older, ranking second only to malaria [27].

Similarly, in Uganda, the official prevalence rate is reported as 5.2% [28] which sums up to 1 400 000 infected people. This sets Uganda as the 11th country by prevalence rate in the world [29].

The official numbers of HIV prevalence data obtained from Rwanda and Uganda serve as the gold standard against which we will compare the estimates derived from the network scale-up methods. by examining the differences between these estimates and the official data, we can evaluate the reliability and effectiveness of the NSUMs in estimating the number of HIV-infected individuals in these African countries.

Next, we will present the results obtained through the network scale-up methods, including the estimates derived from the random degree, transmission bias, and barrier effect methods, and compare them to the official HIV data.

4.3 Estimates

In this section, we compare the estimates derived from the network scale-up methods with the official HIV prevalence data mentioned in the latter section. By analyzing the difference between these estimates and the official data, we can evaluate the performance of the NSUMs.

The evaluation of the estimates produced by the models was guided by the findings and methodologies outlined in Maltiel 2015 [2], providing valuable insights into the assessment process. We employed Markov chain Monte Carlo (MCMC) to estimate all the models. In the case of μ and σ, we used closed Gibbs steps for sampling, while random walk Metropolis steps with normal proposals were employed for the other parameters.

The MCMC algorithms were implemented following the methodology outlined in Raftery and Lewis (1996) [30]

Verifying NSUM estimation results poses challenges due to the unknown actual size of hard-to-reach subpopulations. Consequently, we conducted several simulations to demonstrate the necessity of our models and their improvements in accounting for

biases. We examine our data using the RD-model, as well as the BE and TB-models. Additionally, we computed back estimates on the data from known population sizes.

To begin the comparison, we calculate the difference between the estimates obtained through the models and the official numbers for both Rwanda and Uganda. These differences provide information on the degree of variation between estimated values and known prevalence rates.

4.3.1 Uganda

In this subsection, we present the application of the models for estimating the number of HIV-infected people in Uganda. To provide context for our analysis, we first need to outline the demographic composition of Uganda and the distribution of its population across various sectors. The data used in our study are derived from the Uganda Bureau of Statistics [31]. Specifically, we used information on the number of people employed in different sectors as a bias to estimate the size of subpopulations relevant to our analysis.

Table [4.3] presents a comprehensive list of the sizes of these subpopulations. The values presented in the table offer information on the composition of the workforce in Uganda and help us estimate the degrees of the network. By incorporating these data into our analysis, we can establish a foundation for the estimation. To capture the data of the Ugandan population, we designed a fictitious questionnaire [A.1]. The questionnaire aims to collect data on employment patterns.

Subpopulation (Sectors)	Size
Agriculture	11 159 537
Trade	1 552 205
Education	767 933
Transport	588 204
Manufacturing	359 458
Construction	392 136
Other	1 519 527

Table 4.3: Known sizes for Seven reference subpopulations of Uganda (N=46 000 000).

4.3.1.1 Distribution of Estimations

Figure 4.1 shows the distribution of estimations obtained from each model. Each model was run multiple times to generate a range of estimations, which are represented as histogram plots. The x-axis represents the estimated number of HIV-infected individuals, while the y-axis represents the estimation counts.

As observed in Figure 4.1, the TB-model produces a skewed distribution with a peak around 1 800 000 estimations. This indicates that the TB-model tends to concentrate its estimates toward higher values. The skewed distribution suggests that the

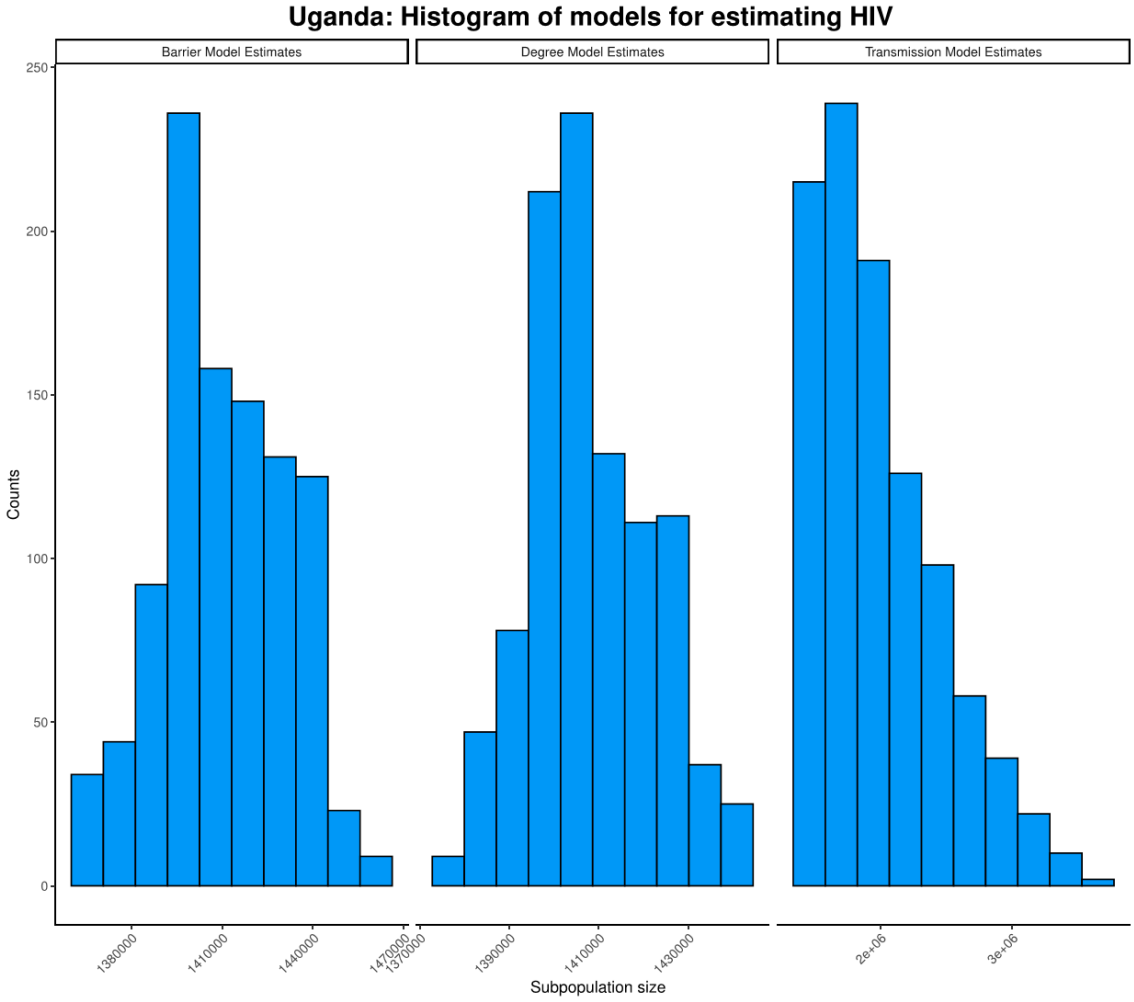
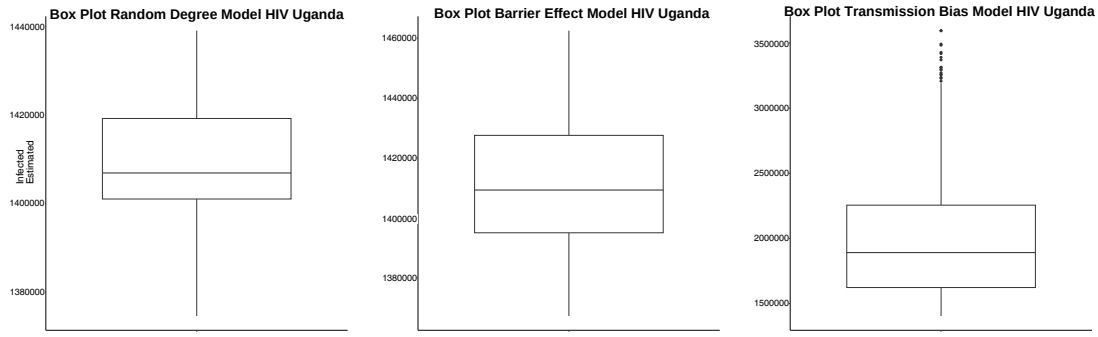


Figure 4.1: Distributions of the estimation for the number of HIV-infected people in Uganda for each model presented in a histogram. The x-axis reflects the estimated subpopulation size where each bin covers a range, while the y-axis displays the frequency of each range.



(a) Random Degree (b) Barrier Effect (c) Transmission Effect

Figure 4.2: Box Plots of estimates of HIV infected in Uganda show the distribution of the estimation of infected people. The box represents the interval of the 25-75 percentile and the line inside the boxes represents the median of the distributions. The y-axis represents the amount of infected estimates.

TB-model may be overestimating the number of HIV-infected individuals. It is important to note that the skewness of the distribution implies a potential bias in the model's estimation methodology, which may be further investigated and adjusted.

On the other hand, the BE-model exhibits a more symmetric distribution of the estimations, with a peak around 1 400 000. The symmetric nature of the distribution suggests that the BE-model provides a more balanced estimate throughout the range of values. This indicates that the model may offer a more accurate estimate of the number of HIV infections in Uganda compared to the TB-model.

In contrast, the RD-model displays a relatively similar distribution to the BE-model with a peak around 1 400 000 as well. The only noticeable difference is that the distribution is more spread out compared to the BE-model. The wider range of estimation suggests a higher level of variability and potential inconsistencies in the model results.

4.3.1.2 Model Performance

As mentioned in the introduction, to thoroughly evaluate the performance of each model in estimating the number of HIV-infected individuals we calculated the magnitude (estimated size divided by the true size) between the estimated values generated by the model and the known sizes of each subpopulation. Figure 4.3 provides a forest plot that visualizes the deltas for all subpopulations.

In Figure 4.3 each plot represents a different subpopulation. The x-axis represents the magnitude of the difference for each estimate compared to the true size, where the true size is represented by a dashed line in the plots. Each tick on the y-axis represents the different models considered in the evaluation. The black square represents the overall mean of the magnitude for each model and the length of the line represents the 95% confidence interval, offering a summary of their performance. For example, the transmission bias model in the agriculture subplot has an estimate

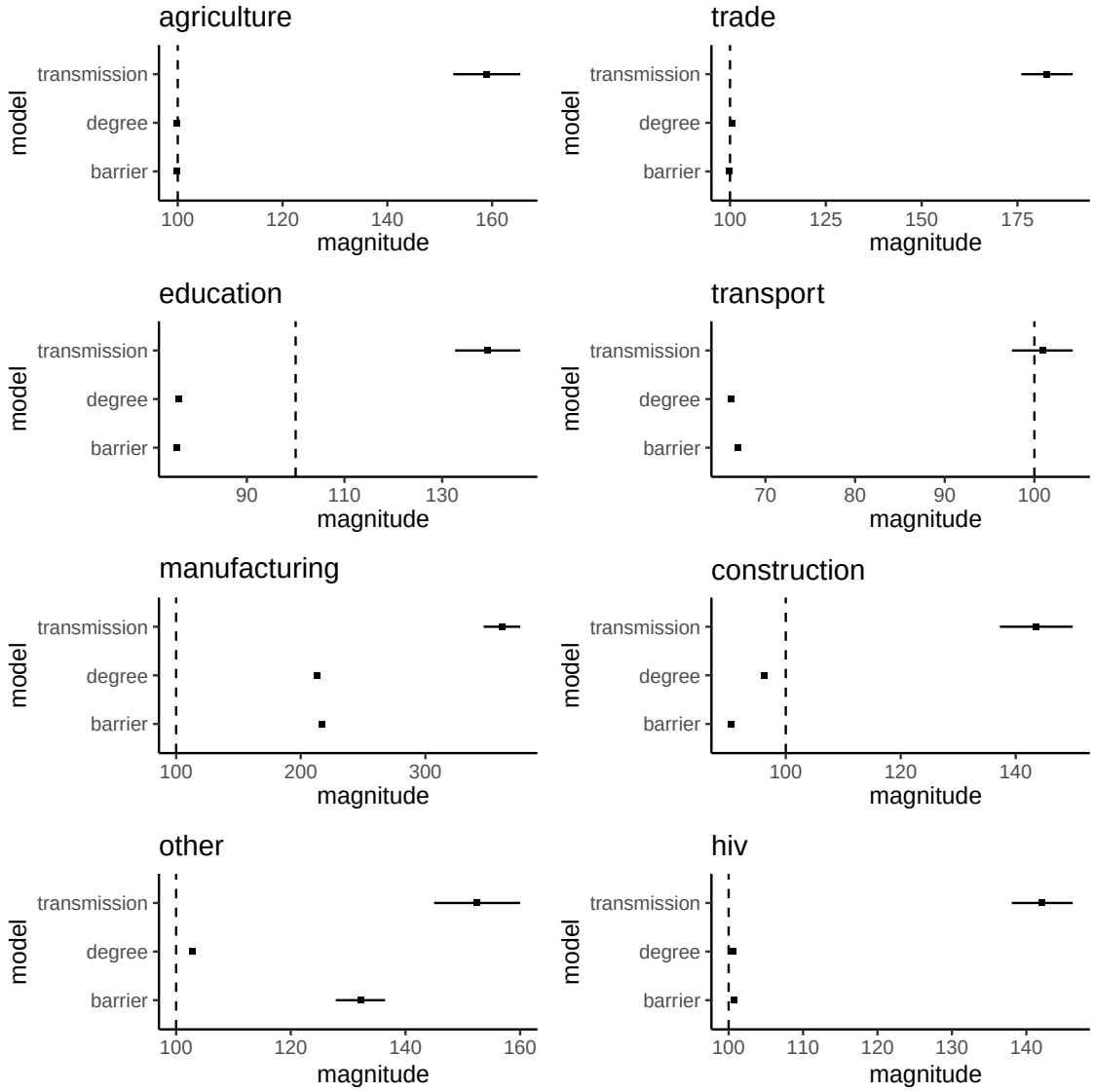


Figure 4.3: Forest plot depicts the disparities in estimated and actual sizes for different models in Uganda. The x-axis reflects the magnitude of differences between each estimate and the true size, while the y-axis lists the models. This visualization allows for an examination of how models estimate hidden populations, revealing tendencies towards overestimation or underestimation.

of around 160. The interpretation of this value is that it estimates a population size of 60% bigger compared to the true size.

The forest plot represents an overview of the variations in model performance across different subpopulations. Analyzing the results, we observe that the TB-model tends to consistently overestimate the true size of the subpopulation, particularly in the manufacturing sector it tends to estimate subpopulation sizes of around 300% bigger than the true number.
In contrast, the RD-model demonstrates a more balanced performance compared to the transmission effect model. The model exhibits closer magnitudes overall, indicating a closer alignment between estimated and known sizes.
In contrast, the random BE-model shows higher variability in its performance. For example, the "other" sector, shows a wider range of magnitudes within the 95% confidence interval. The BE-model appears to be less consistent compared to the RD-model, indicating challenges in accurately estimating the true sizes.

For supplementary visualizations, kindly refer to Section A.2 within the appendix.

4.3.2 Rwanda

In this subsection, we conduct the same experiments and simulations as in the previous subsection on Uganda, where we apply NSUM on data from Rwanda. Data used in the analyses are collected from the National Institute of Statistics of Rwanda [32], where we used population sizes of different ages in groups of 5 years. By doing so, we hope to exclude bias and error potentially embedded in the data when estimating the prevalence of HIV since the data is from an official government registry. Not only that, the topic of age is a socially trivial matter, and thus the concept of barrier effect and transmission bias should hypothetically not hold much weight in our estimation. A possible exception to this is the scenario where a person is more prone to know other people similar to their own age, which is the very definition of the barrier effect. The distribution of the data is presented in table [4.4]. As we did before, we propose a candidate questionnaire [A.3] that would gather the data shown in table [4.4]

Subpopulation (Ages)	Size
0-4	1 708 460
5-9	1 697 005
10-14	1 551 347
15-19	1 509 341
20-24	1 174 549
25-29	1 007 307
30-34	950 747
35-39	869 983
40-44	724 954
45-49	479 225
50-54	393 788
55-59	316 729
60-64	311 001
65-69	214 001

Table 4.4: Known sizes for 14 subpopulations of Rwanda divided in ages of 5 years with a total population size of N=13 460 000.

4.3.2.1 Distribution of Estimates

Again, to estimate the size of the subpopulation affected by HIV in Rwanda, each model was run multiple times with many iterations, and the resulting output is visualized in a histogram and a box-plot in Figure [4.4] and Figure [4.5], respectively.

Similarly to the previous section, the x-axis represents the estimated number of people affected by HIV, whereas the y-axis represents the counts of estimates within an interval. As can be seen from the histogram in Figure [4.4], the estimates seem to display a bell-shaped curve similar to that of a normal distribution, with a very slight skew, however. This is to be expected since the values were sampled from a binomial distribution since samples from a binomial distribution can be approximated by a normal distribution under specific conditions. This is confirmed by inspecting the box plot in Figure [4.5], where the median only slightly deviates from the center with symmetric tails. Although the models have somewhat similar shapes, they produce different estimates. By further examining Figure [4.4] and Figure [4.5], we can see that for the RD-model, the estimates accumulate around 240 000. Not only that, by looking at the intervals of the bins on the x-axis, the values in which the model outputs produce a small variance. For the BE-model, the values gather around 300 000, which is a lot higher compared to the RD-model. By closely inspecting the histogram for the BE-model, the intervals in which the bins are displayed are very spaced out, which indicates a large variance. These values range from 100 000 to 500 000 which, in comparison to the RD-model range of 210 000 to 260 000, is very large. Lastly, the TB-model also seems to peak around 300 000. However, the bins have a slightly smaller range compared to the BE-model but are still larger than the RD-model. These observations are confirmed by examining the box plot in Figure [4.5], where the median and quantiles differ for each model according to our observations. As a first intuition, these results should be expected as both the

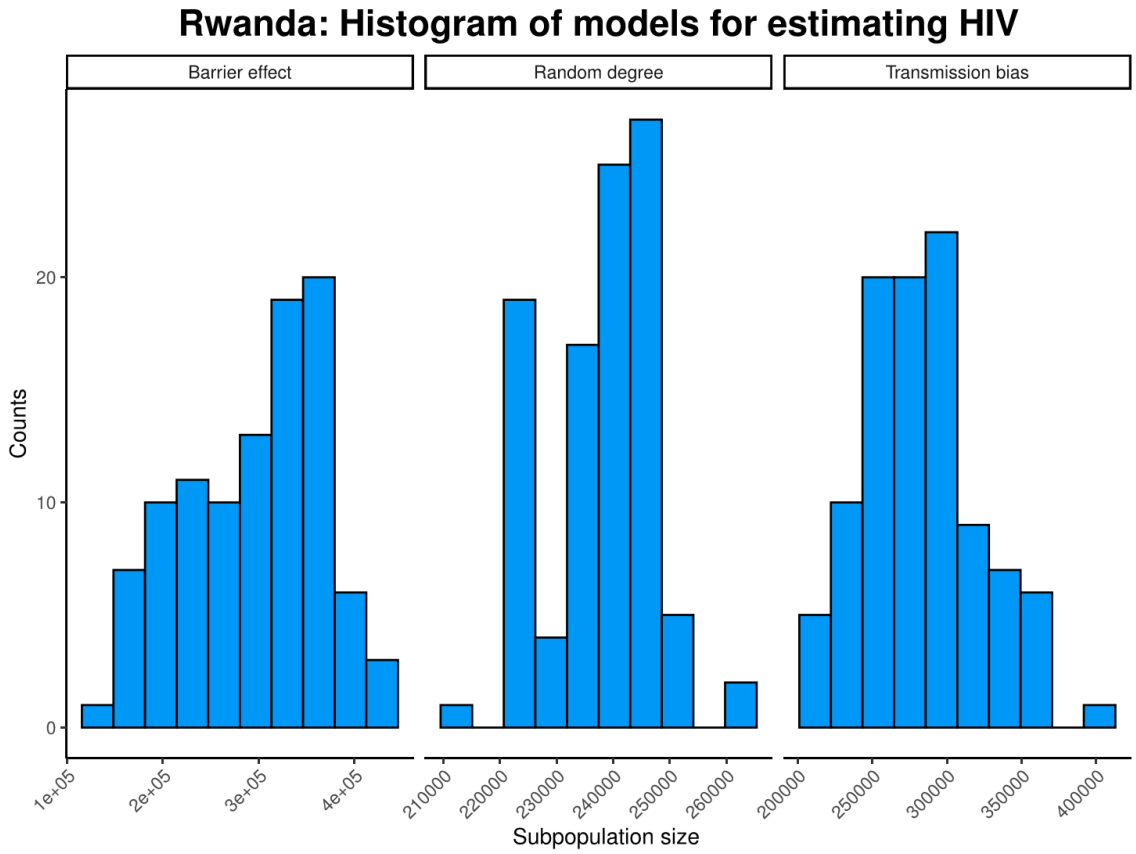
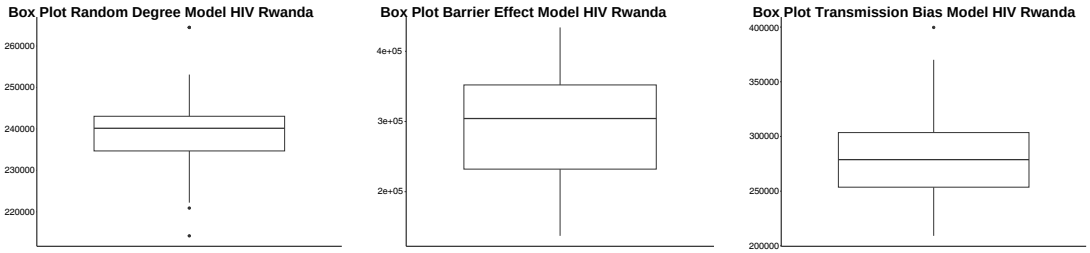


Figure 4.4: Distributions of the estimation for the number of HIV-infected people in Rwanda for each model presented in a histogram. The x-axis reflects the estimated subpopulation size where each bin covers a range, while the y-axis displays the frequency of each range.



(a) Random Degree (b) Barrier Effect (c) Transmission Effect

Figure 4.5: Box Plots of estimates for the number of people affected by HIV in Rwanda show the distribution of the estimation of infected people. The box represents the interval of the 25-75 percentile and the line inside the boxes represents the median of the distributions. The y-axis represents the amount of infected estimates.

BE-and TB-model overestimates in an attempt to account for possible bias and error embedded in the data.

4.3.2.2 Model Performance

To further assess the models, a forest plot for the first 8 known subpopulations was also produced for the different ages so that proper comparisons can be made, which is presented in Figure [4.6].

The forest plots present the performance for the different models by displaying the mean and the lower and upper bounds of a 95% interval. The first plot, A0_4, in the upper left corner represents the estimates of models for subpopulations with ages ranging from 0-4 years old. Whereas plot A5_9 represents the subpopulation with ages ranging from 5-9, and so on.

In all the plots representing different age groups, the RD-model seems to be the best-performing model and also the most consistent with the true value with very few deviations compared to the other models. For the BE-model, it underestimates all estimations, with the exception of subpopulation in ages 20-24. Not only that, for that specific group the variance is also larger compared to estimates in other subpopulations. Lastly, the TB-model overestimates all subpopulations. However, compared to the BE-model, the TB-model exhibits a more consistent variance across all subpopulations. These observations do not align with our previous statement that the BE-model, like the TB-model, overestimates to account for potential bias and error hidden in the data. Reasons as to why this is the case are unknown.

For supplementary visualizations, kindly refer to Section A.4 within the appendix.

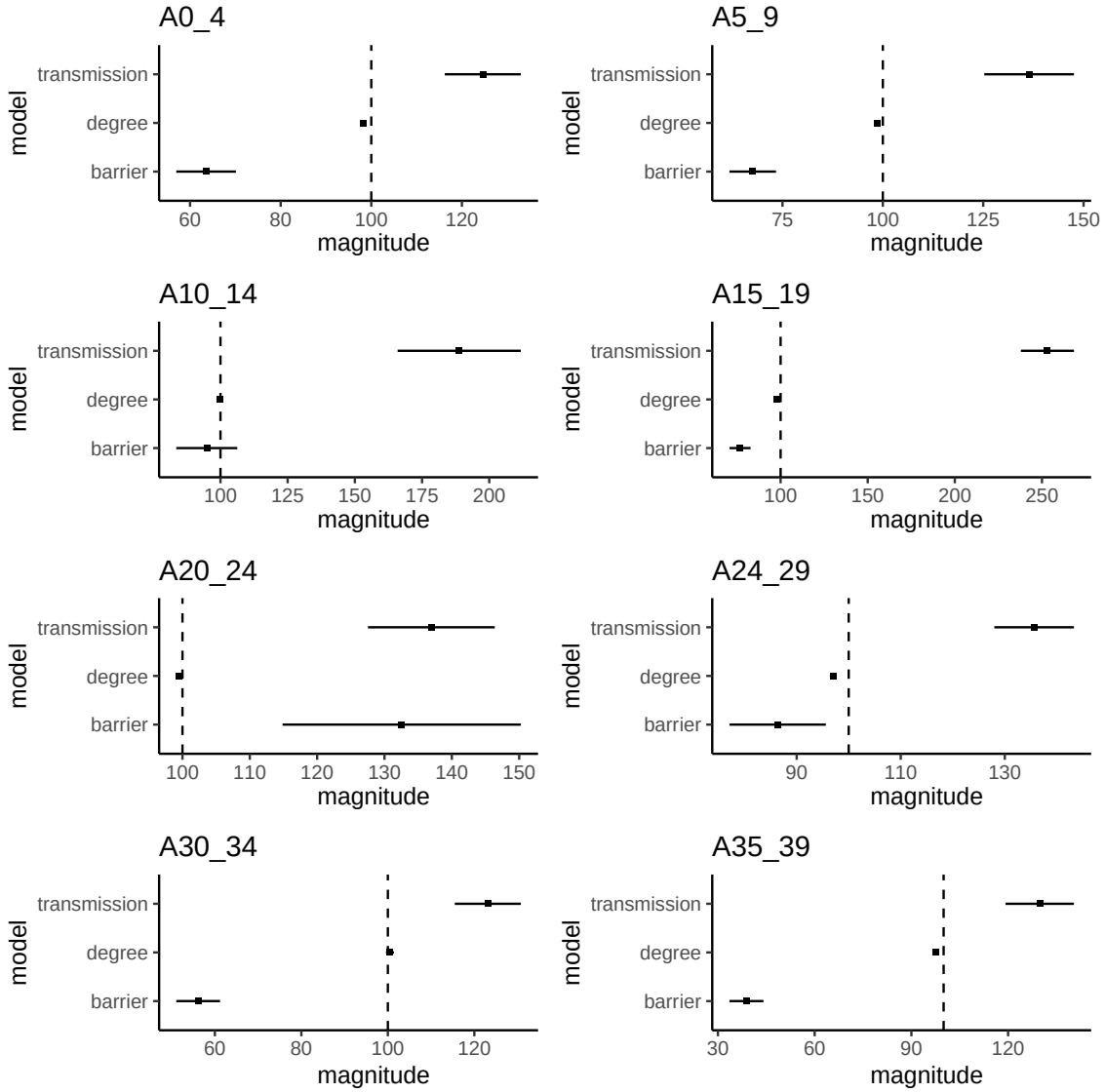


Figure 4.6: Forest plot depicts the disparities in estimated and actual sizes for different models in Rwanda pertaining to ages. The x-axis reflects the magnitude of differences between each estimate and the true size, while the y-axis lists the models. This visualization allows for an examination of how models estimate hidden populations, revealing tendencies towards overestimation or underestimation.

5

Discussion

In this chapter, we are going to provide an analysis and an interpretation of the results obtained from our study in the previous chapter. We estimate the number of HIV-infected individuals in two African countries, namely Rwanda and Uganda. We delve into the implications of our findings in the context of our research objective stated at the beginning of this report. We are going to examine the strength and the limitations of each network scale-up model employed in Section [4.3]. In addition, we are going to consider the broader implications of our findings for estimating epidemic prevalence, taking into account the challenges and complexities of the data sources and methodologies employed. Through this discussion, we aim to provide a deeper understanding of the precision, reliability, and applicability of network scale-up methods in estimating disease prevalence and offer insights for further research and improvements in this field.

5.1 Discussion of System Performance

5.1.1 System Desktop Version Performance

The desktop performance evaluation of the web-based questionnaire revealed exceptional results in terms of performance, best practices, and SEO. With a score of 100 in the "Performance" metric, the questionnaire website ensures a seamless user experience by delivering fast loading speeds, smooth interactivity, and visual stability. This signifies that desktop users will encounter minimal delays and can navigate through the website efficiently.

Furthermore, the questionnaire website obtained 83 in both "Best Practices" and "SEO" metrics, indicating a robust development approach and effective optimization for search engine visibility. However, the accessibility score of 81 suggests that there are areas where the website can be improved to ensure inclusivity for users with disabilities.

5.1.2 System Mobile Version Performance

The mobile performance evaluation of the web-based questionnaire demonstrated excellent results in terms of performance and best practices. With a score of 99 in the "Performance" metric, the questionnaire website delivers a highly optimized

experience for mobile users, ensuring fast loading speeds and smooth interactivity. Mobile device users can complete the questionnaire efficiently, enhancing the overall data collection process.

Similar to the desktop evaluation, the "Best Practices" metric received a score of 83, indicating consistent adherence to recommended development practices and website optimization techniques. However, an SEO score of 67 suggests that further improvements are necessary to boost the website's visibility in search engine results on mobile platforms.

While the questionnaire excels in performance and best practices for mobile devices, there is room for improvement in terms of mobile SEO. Addressing these areas, the website's discoverability can be enhanced, and users with disabilities can access and interact with the questionnaire more efficiently.

It is worth noting, that while the questionnaire excels in performance and best practices for mobile devices, there is a need for consideration regarding accessibility in low-income African countries where smartphones might not be as prevalent. Addressing these concerns would not only improve the website's discoverability but also ensure a broader range of users, including those regions with limited access to smartphones, can efficiently access and interact with the questionnaire. This approach aligns with a more inclusive and globally accessible design, fostering equitable participation in the data collection process.

5.2 Discussion of Estimation Results and Models

In this section, we will delve into a comprehensive discussion of the results obtained in Section [4.3], where we compare three different models: random degree, barrier effect, and transmission bias on their ability to estimate the number of HIV infections in African countries such as Uganda and Rwanda. Firstly, we should remember the strengths and weaknesses of each model:

5.2.1 Random Degree Model

The RD-model assumes that individuals in the hidden population have similar average numbers of contacts as individuals in the general population [6]. It is a simpler method that does not assume differential mixing patterns. Its strengths include ease of implementation and reduction of biases related to mixing assumptions. However, it may fail to capture specific network characteristics, have a limited representation of hidden population dynamics, require reliable data, and lack context-specific information.

5.2.1.1 Strengths

- **Simplicity**: it assumes the individuals in the hidden population have a similar degree of contacts as the general population, this makes it easier to apply when information is limited.

- **Reduced Bias**: by not taking any assumptions, the model reduces potential biases that may not hold in other models. This can enhance the accuracy of the estimation.

5.2.1.2 Weaknesses

- **Failure to capture network characteristics**: the model assumes homogeneity in contact patterns. However, in reality, there can be a significant variation in social and contact networks.
- **Limited representation of hidden population dynamics**: the model assumes that the hidden population and the general population have similar average numbers of contacts. This assumption may not hold if the hidden population engages in behaviors or activities that lead to different contact patterns, resulting in underestimation or overestimation of the hidden population size.
- **Lack of context-specific information**: The RD-model does not explicitly incorporate context-specific factors or barriers that may affect the size of the hidden population. It assumes that the hidden population behaves similarly to the general population without considering specific characteristics, circumstances, or constraints that might impact the hidden population's interactions and size.

5.2.2 Barrier Effect Model

The BE-model assumes that barriers restrict or limit interactions between the hidden population and the general population [6]. It offers a framework to estimate the size of hidden populations considering contextual factors and specific barriers. Its strengths include the incorporation of contextual factors, adaptability to different populations, and understanding of marginalized communities. However, it also has weaknesses, such as assumptions about barrier effectiveness, data limitations, biases, and ethical considerations.

5.2.2.1 Strengths

- **Hidden population estimation**: The BE-model offers a method to estimate the size of hidden populations considering barriers that restrict or limit the interactions between the hidden population and the general population. It provides a framework to understand populations such as undocumented immigrants, individuals engaged in illegal activities, or marginalized communities.
- **Incorporation of contextual factors**: The BE-model takes into account contextual factors that influence the interactions between the hidden population and the general population. It considers various barriers such as legal, social, or cultural factors that may affect the movement and engagement of individuals, providing a more nuanced understanding of the hidden population size.

- **Flexibility and adaptability**: The BE-model can be flexible and adaptable to different populations and contexts. It allows researchers to incorporate specific barriers relevant to the hidden population under study, thereby capturing the unique dynamics and complexities associated with the population.

5.2.2.2 Weaknesses

- **Assumption validity**: The accuracy of the BE-model relies on the assumption that barriers between the hidden population and the general population are effective in limiting interactions. However, the actual impact and effectiveness of barriers can vary and the model's estimates may be sensitive to the assumptions made regarding the strength and permeability of the barriers.
- **Data limitations and availability**: The BE-model requires reliable data on both the hidden population and the general population, including information on barriers and their impact. Obtaining comprehensive and accurate data on hidden populations can be challenging due to various factors such as limited access, fear of disclosure, or social desirability bias. The availability of high-quality barrier data can also pose challenges.
- **Potential biases**: Similar to the TB-model, the BE-model can introduce biases if assumptions regarding barriers and their impact are not valid or if the data used for estimation are biased or incomplete. Biases can result in underestimation or overestimation of the hidden population size.

5.2.3 Transmission Bias Model

The TB-model assumes that individuals in the hidden population are more likely to have contact with individuals from the general population [6]. It is a method used to estimate the size of hidden populations. While it has strengths such as providing insights into hard-to-reach populations and cost-effectiveness, it also has weaknesses, including reliance on assumptions, limited data availability, potential biases, and ethical concerns.

5.2.3.1 Strengths

- **Hidden population estimation**: The TB-model provides a framework to estimate the size of hidden populations, which are often challenging to study directly due to various social, cultural, or legal barriers. It allows researchers to gain insights into populations such as illicit drug users, sex workers, or individuals with stigmatized conditions.
- **Cost-effectiveness**: Compared to other methods like respondent-driven sampling or venue-based sampling, the TB-model can be more cost-effective. It uses existing data sources, such as population surveys or social network data, without requiring direct contact with the hidden population.
- **Generalizability**: The TB-model can potentially provide estimates at a larger scale. It leverages the assumption of differential mixing patterns be-

tween the hidden population and the general population, making it possible to estimate the hidden population size for a broader geographic area.

5.2.3.2 Weaknesses

- **Assumption validity**: The accuracy of the TB-model is heavily based on the assumption of differential mixing patterns between the hidden population and the general population. If this assumption does not hold true, the estimated size of the hidden population may be biased or incorrect.
- **Limited data availability**: The TB-model requires reliable data on both the hidden population and the general population. However, obtaining comprehensive and representative data for the hidden population can be challenging due to its hidden nature, leading to potential data limitations and inaccuracies.
- **Potential biases**: The TB-model assumes that individuals in the hidden population have a higher probability of contact with the general population. However, this assumption might introduce biases, as it may not hold uniformly across all hidden populations or individuals within the hidden population. Biases can result in over- or underestimation of the hidden population size.

5.3 Analysis Of Histograms and Boxplots

In this section, we will discuss the key characteristics of the histograms and box plots that were presented in Section [4.3] to visualize the distributions of the estimates of the number of HIV-infected individuals in Uganda and Rwanda using three different network scale-up methods: random degree, barrier effect, transmission bias. We will also examine any notable patterns or differences among these methods.

The histogram provides a visual representation of the frequency distribution of the estimates obtained from each network scale-up method. It consists of a series of bars where the height of each bar represents the frequency or count of estimates that fall within a particular range or bin. The shape of the histogram can provide insights into the underlying distribution of the estimates.

Upon analyzing the histograms for Uganda in Figure 4.1, we observed that the distributions of the estimates for the TB-model present a strong right-skewed pattern. This indicates that the majority of the estimates were concentrated toward the higher end of the range with a tail extending to higher values suggesting a possible bias. On the other hand, the random degree and the barrier effect showed a more symmetric shape indicating a more precautious estimation. However, the same can not be said about the estimates from the Rwanda data which can be seen in the histogram presented in Figure [4.4]. All the models exhibited a bell-shaped curve with only a slight skew. When it comes to the estimates themselves, the RD-model seems to be the most consistent model with a small variance. Whereas the barrier effect- and TB-model, tend to overestimate in hopes of accounting for potential biases and errors. Both of these models also displayed a large variance compared to the RD-model, which can be an indication of inaccurate estimates.

The boxplots, on the other hand, provide a summary of the distributions' key statistical measures, including median, quartiles, and any outliers. They help to compare the central tendency and dispersion measures among different models.

Examining the box plots in Figure 4.2, we found that the median estimate for the random degree and the barrier effect have similar values. On the contrary, the transmission bias showed a higher median, and the interquartile range (IQR) was larger compared to the latter models, suggesting a wider spread of estimates. The box plots for both Rwanda and Uganda confirm our previous analyses for the histogram.

Another notable difference between the methods was the presence of outliers. The transmission bias exhibited a higher number of outliers when compared to the other two methods, further indicating a higher degree of variability or potential inaccuracies.

In general, the histogram and box-plot analyses revealed distinct patterns and differences between the three models and the two data sets. These observations highlight the importance of considering the strengths and weaknesses of each method when estimating the number of HIV-infected people. More analysis is necessary to fully understand the factors that contribute to these factors.

5.4 Interpretation of Results

The result obtained from applying the three models provided valuable insights into the spread of the pandemic in the countries of Uganda and Rwanda. In this section, we will interpret these results and discuss their implications.

The estimates obtained through the network scale-up methods offer an approximation of the true number of HIV-infected individuals in Uganda and Rwanda, considering that direct measurement of such a hidden population is challenging. Using indirect data and leveraging relationships within social networks, these methods provide a means to estimate the prevalence of HIV infection.

For both Uganda and Rwanda, the estimates obtained from the transmission bias method, characterized by a higher central tendency and wider dispersion, suggest that this method may have captured a larger proportion of HIV-infected individuals in Uganda and Rwanda compared to the other methods. However, the presence of outliers in the estimates indicates potential uncertainties or biases in certain subpopulations or data sources. It is crucial to further investigate the factors contributing to these outliers and assess their impact on the overall estimation

On the other hand, the estimates derived from the barrier effect and random degree methods for Uganda exhibit relatively lower central tendencies and narrower dispersions. This indicates that these methods may have provided more conservative estimates of the number of HIV-infected individuals. The barrier effect method is for people who may not be fully integrated into their social networks. These methods, although potentially underestimating the true prevalence, provide insights into the limitations and biases that exist within the social network data. For the data

from Rwanda, he estimates that the RD-model displayed similar behavior as the RD-model for Uganda, i.e., smaller mean and variance compared to the other models. Whereas the estimates from the BE-model exhibited a behavior more similar to the transmission models. These findings made our findings and intuition regarding our models inconclusive.

The implications of these estimates extend beyond the immediate estimation of HIV-infected individuals. They contribute to our understanding of the dynamics of the pandemic and its spread in Uganda. By estimating the hidden population, we can gain insights into the overall burden of the disease and identify areas where interventions and resources can be targeted most effectively.

Estimates obtained through network scaling methods can help in public health planning and resource allocation. They provide a basis for assessing the effectiveness of existing interventions, identifying gaps in healthcare provision, and formulating strategies to prevent new infections and provide care and support to those affected.

However, it is important to acknowledge the limitations of the network scale-up methods and the associated uncertainties in the estimates. These methods rely on assumptions such as network homogeneity and accurate reporting, which may not hold in all cases. Estimates are also influenced by the quality and representativeness of the data sources used.

5.5 Comparison with Known Subpopulations

In order to assess the performance and applicability of the three network scale-up methods (random degree, barrier effect, and transmission bias) in estimating the number of HIV-infected individuals in Uganda and Rwanda, we compared the estimates obtained for various known subpopulations within the country. This section will discuss the selection of these subpopulations, the insights provided by the estimates, and any variations or consistencies observed among the three methods.

The selection of known subpopulations was based on existing data and research that provided a reliable reference point for the number of HIV-infected individuals in specific groups. We focused on subpopulations that are presented in Table [4.3] and Table [4.4].

By comparing the estimates obtained for these subpopulations using the three network scale-up methods, we gained insights into the performance of each method in capturing the known prevalence rates within these specific groups. Additionally, this comparison allowed us to assess the consistency or variability of the estimates across the different methods.

The estimates, visible in forest plots presented in Figure [4.3] and Figure [4.6] obtained for every subpopulation using the transmission bias method showed a higher prevalence rate compared to the barrier effect and random degree methods. This observation may indicate that the TB-model is not suited for a known population size as it assumes that the numbers reported are intrinsic underestimating the true size. For this reason, we also decided to create a forest plot without the TB-model.

Overall, the comparison of estimates for known subpopulations provided valuable insights into the performance of the three network scale-up methods. While each method exhibited variations in estimating the prevalence rates for different subpopulations, there were also consistencies observed, particularly in the estimates from the RD-model.

It is worth noting that the comparison with known subpopulations was limited to specific groups with well-documented prevalence rates. Therefore, caution should be exercised when generalizing the results to the broader population. Further research and validation using larger and more representative samples are necessary to strengthen the reliability of these estimates and ensure their suitability for informing public health policies and interventions.

6

Conclusions

In conclusion, this thesis presents a novel computational system developed within the framework of computer science to estimate the number of individuals infected with a pre-determined epidemic in Uganda. Leveraging the power of web technologies, the questionnaire provides a user-friendly interface that allows for efficient data collection from diverse individuals and populations. The Django framework ensures the stability and security of the system, enabling seamless data transmission and storage. By using this system, researchers and public health organizations can gather valuable information on the spread and the impact of epidemics, facilitating comprehensive estimation and informed decision-making for effective interventions.

By employing network scale-up methods, including the random degree, barrier effect, and transmission bias methods, we have demonstrated the effectiveness of computational techniques in estimating the number of infected individuals during a pandemic. Through the analysis of social network information and indirect data, our methods offer valuable insights into the hidden population and contribute to our understanding of epidemic spread in the region.

One of the primary contributions of this research is the development of a robust system that harnesses the capabilities of the Django framework and web-based questionnaires. This system serves as a powerful tool for collecting data from diverse individuals and populations, enabling a comprehensive estimation of the epidemic's impact.

The findings of this research not only validate the viability of network scale-up methods as an estimation approach but also shed light on the computational aspects of the analysis. Our analysis of histograms, box plots, central tendency measures, and dispersion measures provides deeper insights into the characteristics of estimates obtained through these computational methods. Specifically, the random degree method exhibits higher central tendency and wider dispersion, capturing a larger proportion of the infected individuals, while the barrier effect and transmission bias methods offer more conservative estimates with lower central tendencies and narrower dispersions.

Furthermore, the comparison of estimates for known subpopulations reveals intriguing variations and consistencies among the three network scale-up methods. While each method exhibited variations in estimating prevalence rates for different subpopulations, we observed notable consistencies, particularly in the estimates generated

by the random degree method. This comparison underscores the significance of considering specific characteristics and network dynamics when applying network scale-up methods within a computer science context.

The significance of this research lies in its pioneering contributions to the estimation of epidemics in East African countries from a computer science perspective. By developing a computational system that leverages web-based questionnaires and the Django framework, and by evaluating the performance of network scale-up methods in estimating the number of infected individuals, this thesis advances the field of epidemic monitoring. Our findings offer valuable insights for public health planning, resource allocation, and targeted interventions while emphasizing the importance of understanding the limitations, assumptions, and biases associated with computational methods. Moreover, our research points toward future directions for advancing these estimation techniques and further improving their accuracy and applicability.

In conclusion, this research exemplifies the transformational role that computer science can play in monitoring and estimating epidemics. By combining computational approaches, sophisticated algorithms, and innovative data-gathering techniques, we contribute to the development of cutting-edge tools and methodologies that enhance our understanding of epidemic dynamics and empower decision-making for public health interventions.

6.1 Implications for Future Research

The findings of this master thesis shed light on the estimation of the number of infected people during a pandemic in East African countries using network scale-up methods. The implications of these findings provide valuable directions for future research in this field. In this section, we will discuss the potential areas of improvement and modifications to network scale-up methods that could enhance their accuracy and applicability in estimating the number of infected people in the region.

1. Refinement of Sampling Strategies: Future research should prioritize the optimization of sampling techniques to reduce biases introduced during the sampling processes. This could involve exploring innovative statistical methodologies and advanced sampling algorithms to improve participant selection and enhance the representativeness of the collected data.

2. Real-Time Data Collection: Future studies could investigate the feasibility of implementing real-time data collection in the questionnaire by integrating automated data capture systems, such as APIs or data scraping tools. This would enable the collection of up-to-date information from online sources, enhancing the timeliness and accuracy of the data.

3. Incorporation of Network Dynamics: To gain a deeper understanding of the research subject, future work should focus on developing network models that capture temporal changes and dynamic interactions within communities. By leveraging emerging network analysis techniques and integrating data from multiple sources, researchers can better reflect changing patterns of infection

and transmission, considering factors such as individual behavior and social connections.

4. Mobile Compatibility: In the future, it would be beneficial to adapt the questionnaire to be more mobile-responsive or develop a dedicated mobile application. This would cater to the increasing prevalence of mobile devices and provide a convenient platform for a wider range of participants to access and complete the questionnaire.

5. Contextual Considerations: Future research should pay careful attention to social, cultural, and structural characteristics specific to East African countries. It is crucial to adapt research methods accordingly for different settings within the region, taking into account local contexts and engaging with stakeholders to ensure the questionnaire captures the nuances and complexities of the target population effectively.

Bibliography

[1] H. R. Bernard, T. Hallett, A. Iovita, *et al.*, "Counting hard-to-count populations: The network scale-up method for public health," *Sexually Transmitted Infections*, vol. 86, no. Suppl 2, pp. ii11–ii15, 2010, ISSN: 1368-4973. DOI: 10.1136/sti.2010.044446. eprint: https://sti.bmj.com/content/86/Suppl_2/ii11.full.pdf. [Online]. Available: https://sti.bmj.com/content/86/Suppl_2/ii11.

[2] R. Maltiel, A. E. Raftery, T. H. McCormick, and A. J. Baraff, "Estimating population size using the network scale up method," *The annals of applied statistics*, vol. 9, no. 3, p. 1247, 2015.

[3] T. H. McCormick, M. J. Salganik, and T. Zheng, "How many people do you know?: Efficiently estimating personal network size," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 59–70, 2010.

[4] I. Laga, L. Bao, and X. Niu, "Thirty years of the network scale-up method," *Journal of the American Statistical Association*, vol. 116, no. 535, pp. 1548–1559, 2021. DOI: 10.1080/01621459.2021.1935267. eprint: https://doi.org/10.1080/01621459.2021.1935267. [Online]. Available: https://doi.org/10.1080/01621459.2021.1935267.

[5] P. Olofsson and M. Andersson, *Probability, statistics, and stochastic processes*. John Wiley & Sons, 2012.

[6] I. Laga, L. Bao, and X. Niu, "Thirty years of the network scale-up method," *Journal of the American Statistical Association*, vol. 116, no. 535, pp. 1548–1559, Jul. 2021. DOI: 10.1080/01621459.2021.1935267. [Online]. Available: https://doi.org/10.1080%5C%2F01621459.2021.1935267.

[7] E. Breza, A. G. Chandrasekhar, S. Lubold, T. H. McCormick, and M. Pan, *Consistently estimating network statistics using aggregated relational data*, 2022. arXiv: 1908.09881 [stat.ME].

[8] P. D. Killworth, E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard, "A social network approach to estimating seroprevalence in the united states," *Social networks*, vol. 20, no. 1, pp. 23–50, 1998.

[9] P. Orbanz and D. M. Roy, "Bayesian models of graphs, arrays and other exchangeable random structures," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 2, pp. 437–461, 2014.

[10] E. Rosenman and N. Viswanathan, "Using poisson binomial glms to reveal voter preferences," *arXiv preprint arXiv:1802.01053*, 2018.

[11] W. H. Organization *et al.*, "Global health sector strategy on hiv 2016-2021. towards ending aids," World Health Organization, Tech. Rep., 2016.

[12] J. Chakaya, M. Khan, F. Ntoumi, *et al.*, "Global tuberculosis report 2020–reflections on the global tb burden, treatment and prevention efforts," *International journal of infectious diseases*, vol. 113, S7–S12, 2021.

[13] W. H. Organization *et al.*, "Global health sector strategies on, respectively, hiv, viral hepatitis and sexually transmitted infections for the period 2022-2030," 2022.

[14] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[15] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: `https://www.R-project.org/`.

[16] T. M. Inc., *Matlab version: 9.13.0 (r2022b)*, Natick, Massachusetts, United States, 2022. [Online]. Available: `https://www.mathworks.com`.

[17] Django Software Foundation, *Django*, version 2.2, May 5, 2019. [Online]. Available: `https://djangoproject.com`.

[18] SQLite Development Team, *When to Use SQLite*, `https://www.sqlite.org/whentouse.html`, Accessed: May 28, 2023, n.d.

[19] Google, *Google lighthouse calculator*, `https://developers.google.com/web/tools/lighthouse`, 2021.

[20] Google Developers, *Chrome developer documentation: Lighthouse performance scoring*, Accessed: 2023-12-08, 2023. [Online]. Available: `https://developer.chrome.com/docs/lighthouse/performance/performance-scoring/`.

[21] *HTTP Archive*, `https://httparchive.org/`, Accessed on June 28, 2023, Accessed 2023.

[22] Google Developers. "Chrome developer documentation: Lighthouse accessibility." Accessed: 2023-12-08. (2023), [Online]. Available: `https://developer.chrome.com/docs/lighthouse/accessibility/`.

[23] Google Developers. "Chrome developer documentation: Axe assessment." Accessed: 2023-12-08. (2023), [Online]. Available: `https://github.com/dequelabs/axe-core/blob/develop/doc/rule-descriptions.md`.

[24] Google Developers. "Chrome developer documentation: Lighthouse best practices." Accessed: 2023-12-08. (2023), [Online]. Available: `https://developer.chrome.com/docs/lighthouse/best-practices/`.

[25] Google Developers. "Chrome developer documentation: Lighthouse seo." Accessed: 2023-12-08. (2023), [Online]. Available: `https://developer.chrome.com/docs/lighthouse/seo/`.

[26] *Rwanda | UNAIDS*, `https://www.unaids.org/en/regionscountries/countries/rwanda`, Accessed: May 28, 2023, n.d.

[27] UN in Rwanda, *HIV/AIDS, Malaria and Other Diseases*, `https://web.archive.org/web/20160515093834/http://rw.one.un.org/theme-groups/hiv-aids-malaria-and-other-diseases/`, Archived: May 15, 2016. Accessed: May 28, 2023, 2014.

[28] *Uganda | UNAIDS*, `https://www.unaids.org/en/regionscountries/countries/uganda`, Accessed: May 28, 2023, n.d.

[29] Central Intelligence Agency, *The World Factbook | HIV/AIDS - Adult Prevalence Rate*, `https://www.cia.gov/the-world-factbook/field/hiv-aids-`

adult-prevalence-rate/country-comparison, Accessed: May 28, 2023, n.d.

[30] A. Raftery and S. Lewis, *Implementing mcmc, wr gilks, st richardson and dj spiegelhalter, eds., markov chain monte-carlo in practice*, 1992.

[31] Uganda Bureau of Statistics (UBOS), *2022 Statistical Abstract*, https://www.ubos.org/2022-statistical-abstract/, Accessed: May 28, 2023, 2022.

[32] National Institute of Statistics of Rwanda (NISR), *5th Popoulation and Housing Report*, https://statistics.gov.rw/publication/main_indicators_2022, Accessed: May 28, 2023, 2022.

A

Appendix 1

A.1 Fictitious Questionnaire for Uganda

Introduction

Hello! We are conducting a survey to understand the prevalence of HIV infections within Uganda. Your responses will help us gather important information. Your participation is voluntary and confidential.

Demographic Information

1. What is your sector of activity? (Please choose one)
    - Agriculture
    - Trade
    - Education
    - Transport
    - Manufacturing
    - Construction
    - Other

Network Statistics

For the sector you mentioned above, please indicate how many individuals you personally know. Please provide an estimate, even if it's approximate.

1. How many individuals do you know in the Agriculture sector?

   Number: _____
2. How many individuals do you know in the Trade sector?

   Number: _____
3. How many individuals do you know in the Education sector?

   Number: _____

4. How many individuals do you know in the Transport sector?

   Number: _____

5. How many individuals do you know in the Manufacturing sector?

   Number: _____

6. How many individuals do you know in the Construction sector?

   Number: _____

7. How many individuals do you know in the Other sectors?

   Number: _____

8. How many individuals do you know have been infected with HIV?

   Number: _____

A.2 Additional Plots for Uganda

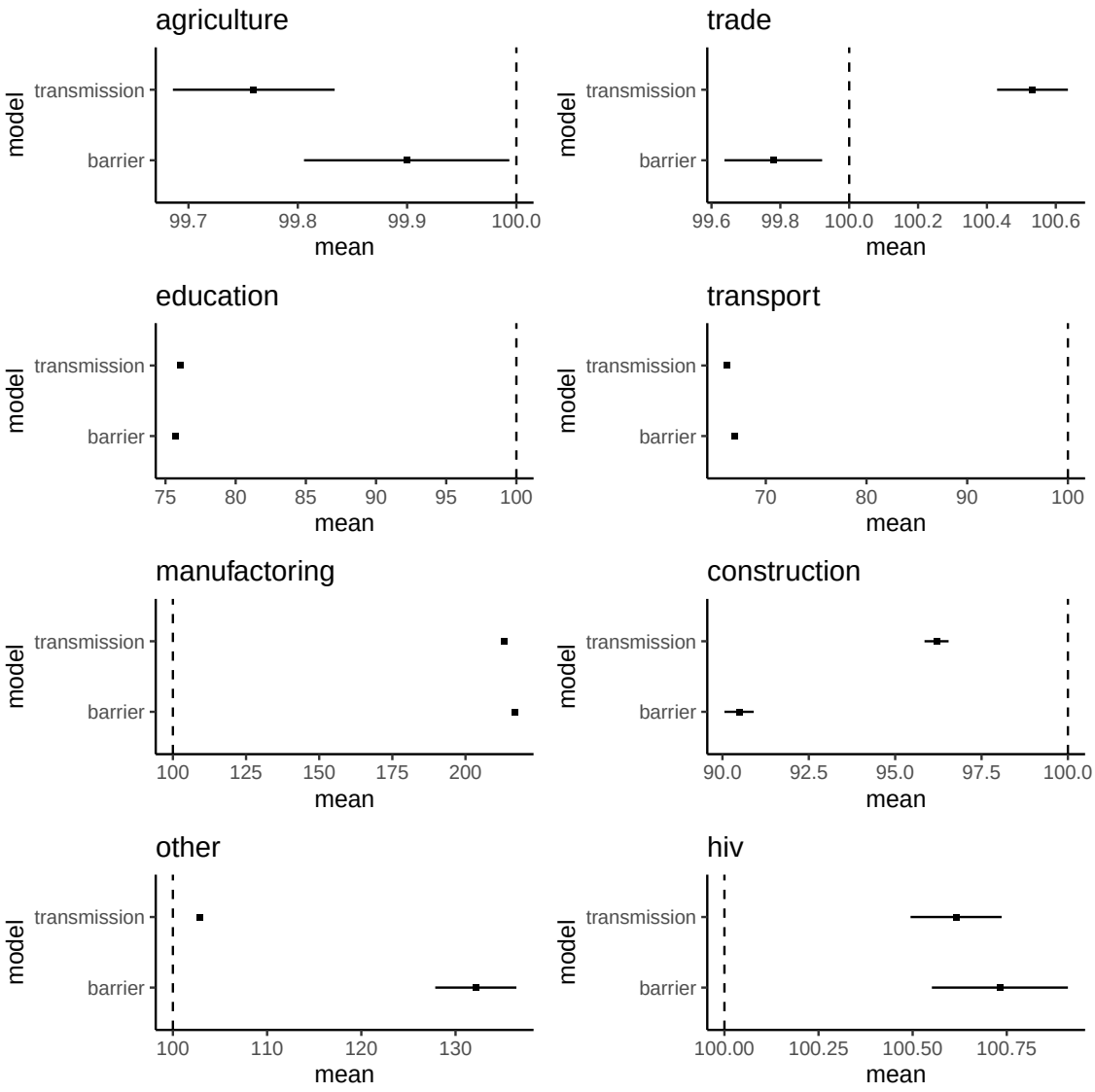


Figure A.1: Forest Plot Uganda without Transmission Bias Model

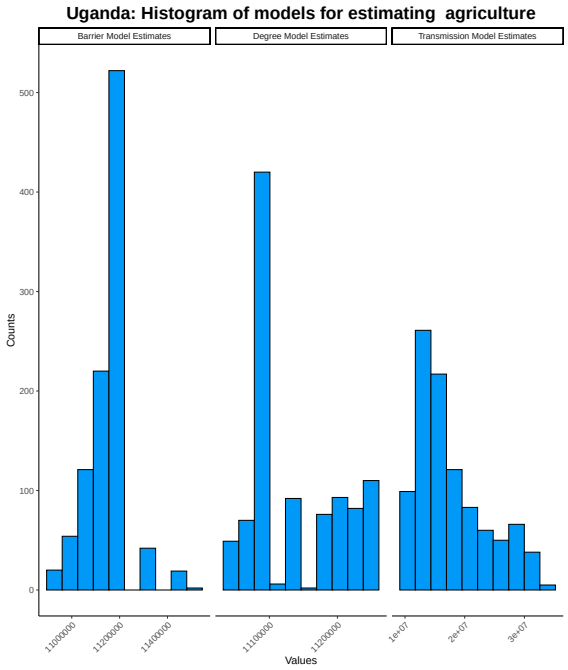


Figure A.2: Distributions of estimates for agriculture sector

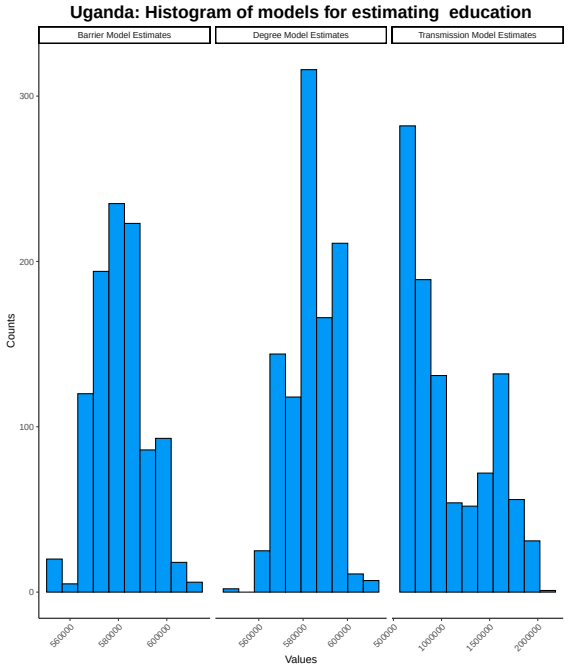


Figure A.3: Distributions of estimates for education sector

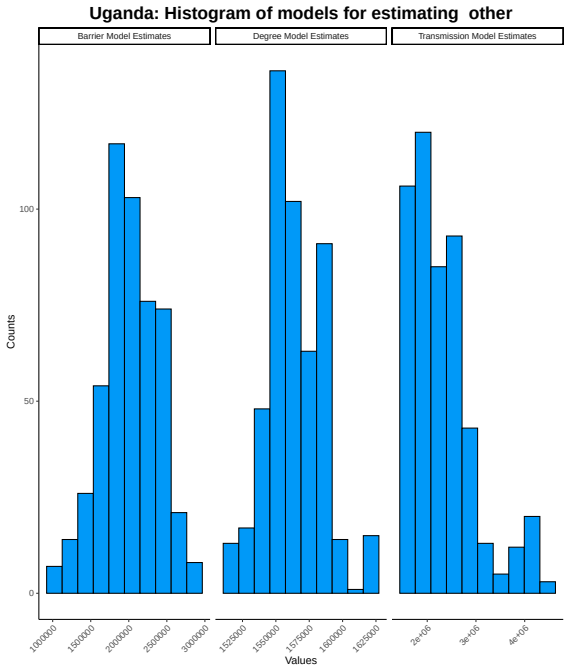


Figure A.4: Distributions of estimates for other sectors

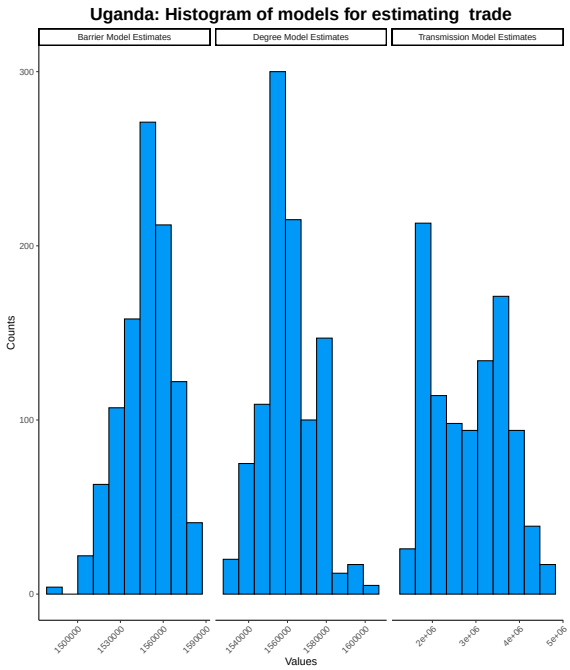


Figure A.5: Distributions of estimates for trade sector

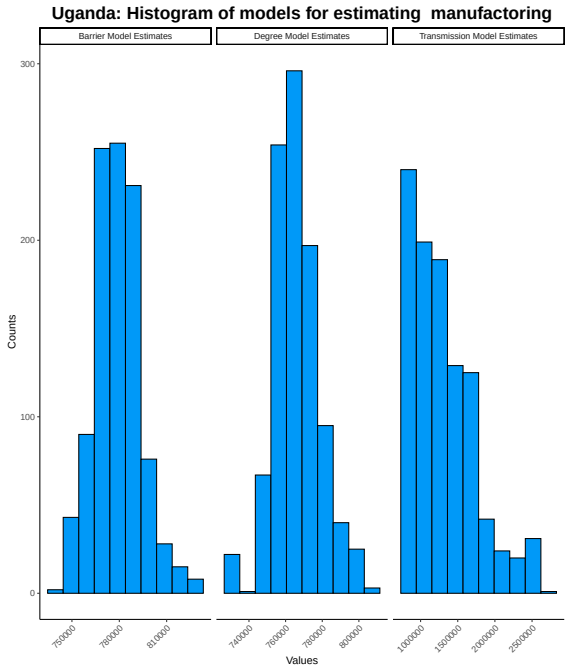


Figure A.6: Distributions of estimates for manufacturing sector

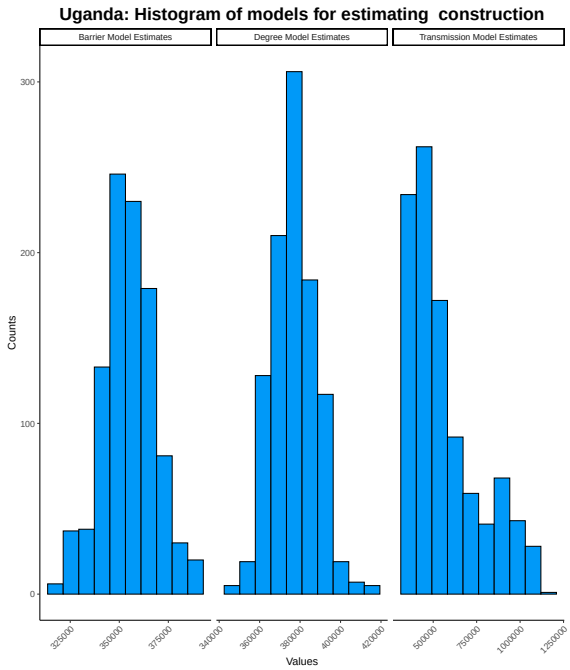


Figure A.7: Distributions of estimates for construction sector

A.3 Fictitious Questionnaire for Rwanda

Introduction

Hello! We are conducting a survey to understand the prevalence of health conditions within specific age groups in our community. Your responses will help us gather important information. Your participation is voluntary and confidential.

Demographic Information

1. What is your age group? (Please choose one)
    - 0-4
    - 5-9
    - 10-14
    - 15-19
    - 20-24
    - 25-29
    - 30-34
    - 35-39
    - 40-44
    - 45-49
    - 50-54
    - 55-59
    - 60-64
    - 65-69

Health Conditions Network

For the age group you mentioned above, please indicate how many individuals you personally know who may have the following health conditions. Please provide an estimate, even if it's approximate.

2. How many individuals do you know in the 0-4 age group?

    Number: _____

2. How many individuals do you know in the 5-9 age group?

    Number: _____

2. How many individuals do you know in the 10-14 age group?

Number: _____

2. How many individuals do you know in the 15-19 age group?

   Number: _____

2. How many individuals do you know in the 20-24 age group?

   Number: _____

2. How many individuals do you know in the 25-29 age group?

   Number: _____

2. How many individuals do you know in the 30-34 age group?

   Number: _____

2. How many individuals do you know in the 35-39 age group?

   Number: _____

2. How many individuals do you know in the 40-44 age group?

   Number: _____

2. How many individuals do you know in the 45-49 age group?

   Number: _____

2. How many individuals do you know in the 50-54 age group?

   Number: _____

2. How many individuals do you know in the 55-59 age group?

   Number: _____

2. How many individuals do you know in the 60-64 age group?

   Number: _____

2. How many individuals do you know in the 65-69 age group?

   Number: _____

Additional Information

3. How many individuals do you know have been diagnosed with HIV in the past year?

   Number: _____

A.4 Additional Plots for Rwanda

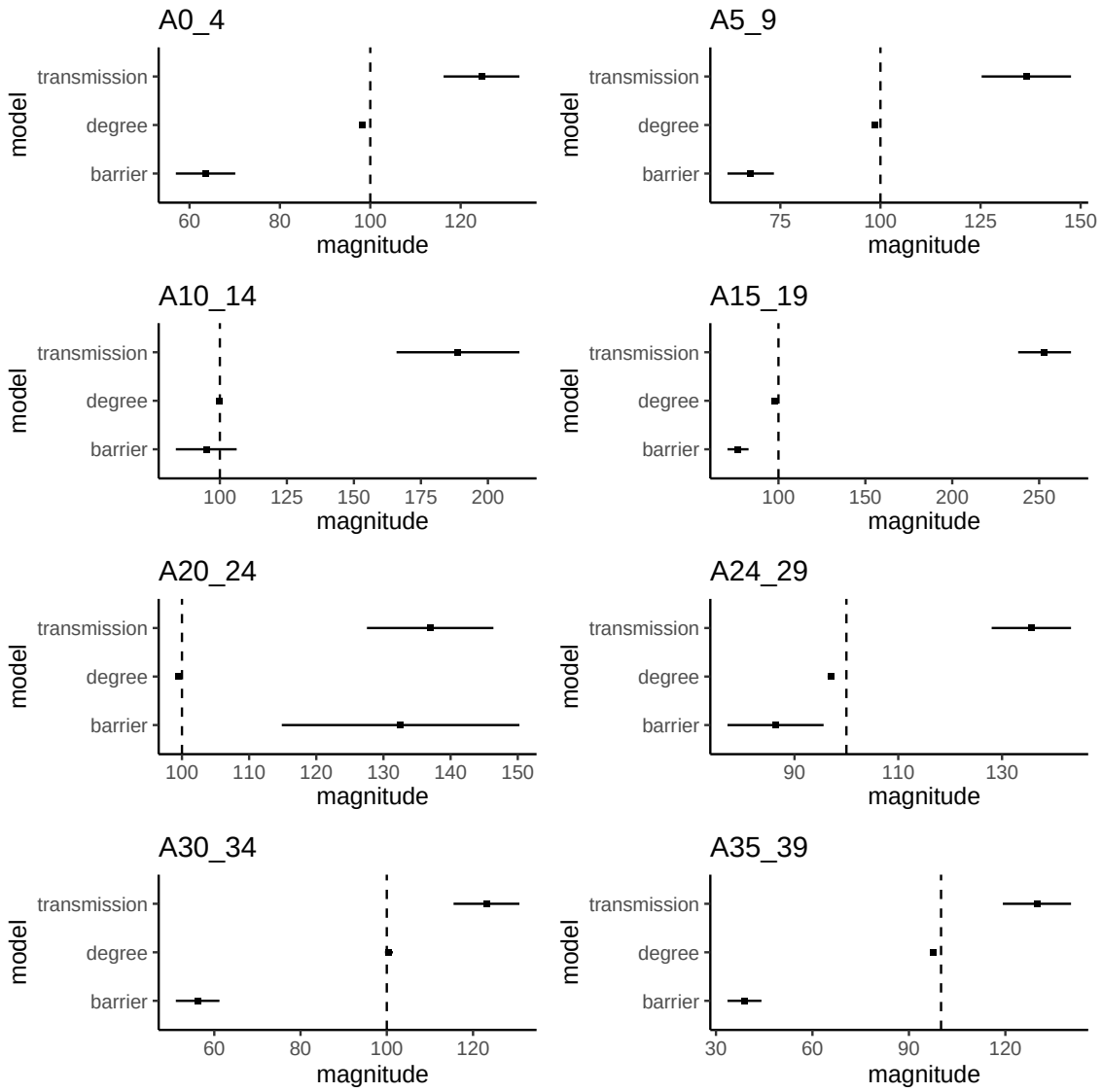


Figure A.8: Forest plot for Rwanda grouped by ages of 5

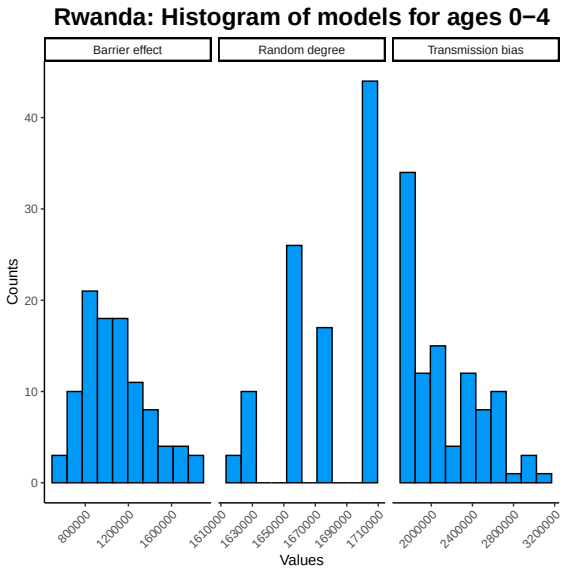


Figure A.9: Distributions of estimates for ages 0-4

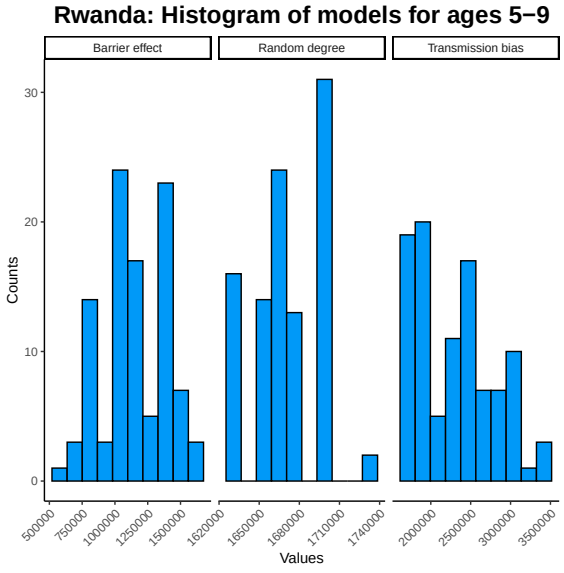


Figure A.10: Distributions of estimates for ages 5-9

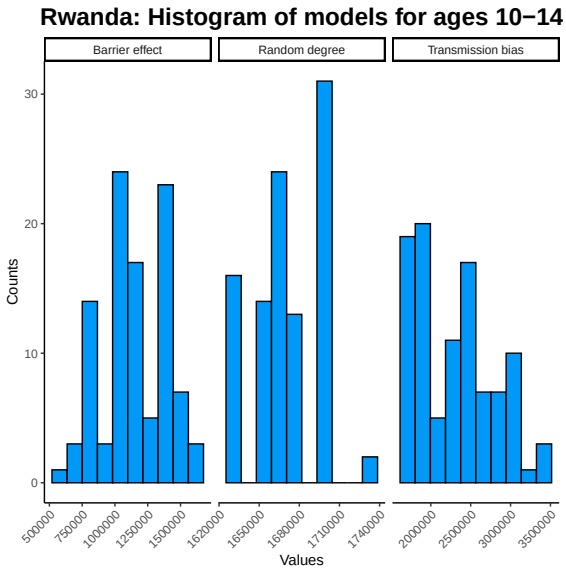


Figure A.11: Distributions of estimates for ages 10-14

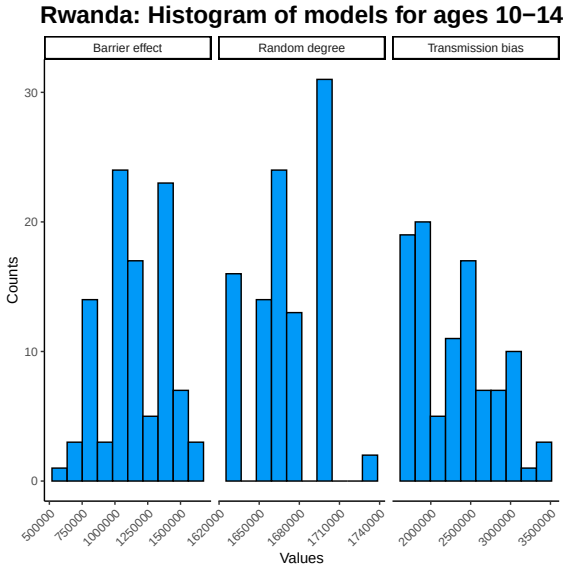


Figure A.12: Distributions of estimates for ages 15-19

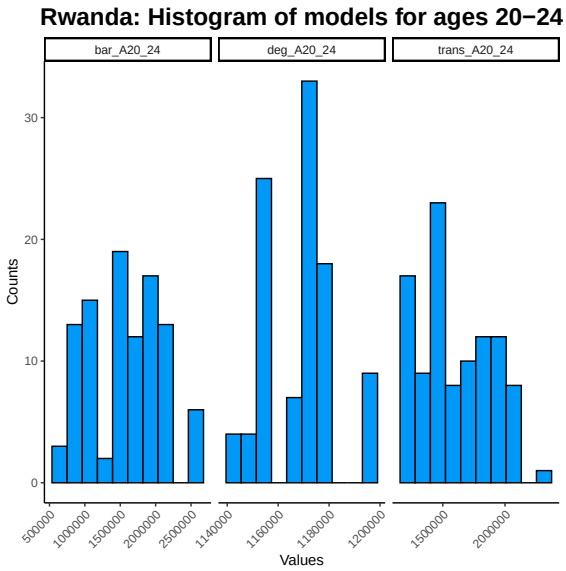


Figure A.13: Distributions of estimates for ages 20-24

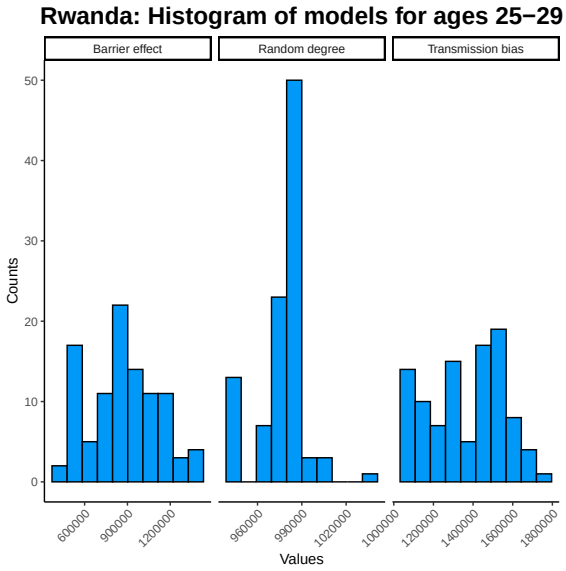


Figure A.14: Distributions of estimates for ages 25-29