DISCRETE-TIME CLOSED CAPTURE-RECAPTURE MODELS FOR HARD-TO-REACH POPULATION SIZE ESTIMATION: APPLICATION TO KEY POPULATION FOR HIV PREVENTION IN RWANDA.

Author:
Elysée TUYISHIME

Supervisors:
Prof. Angela Unna Chukwu
Dr. Ignace Kabano

A research thesis submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy (PhD) in Data Science with specialization of Biostatistics.

African Center of Excellence in Data Science (ACE-DS),
College of Business and Economics

Kigali, Rwanda

July, 2024

Declaration

I, Elysée TUYSIME, hereby declare that this thesis is the result of my own work and has not been submitted for any degree at the University of Rwanda or any other institution.

July 20th, 2024

Elysée TUYISHIME Date

July 20th, 2024

Prof. Angela Unna Chukwu Date

July 20th, 2024

Dr. Ignace Kabano Date

Abstract

Background: The global share of new HIV infections due to key populations (KP) and their partners is steadily rising and was estimated at 70%, with 51% in sub-Saharan Africa, in 2021.Rwanda has a heterogeneous HIV epidemic, that is widespread in the adult population (age 15-49 years), with features of a concentrated epidemic among specific population subgroups, with 35% among female sex workers (FSWs) and 6.9% among men who have sex with men (MSM).) Members of these populations are often difficult to find, and the size of these populations is largely unknown, posing a substantial challenge to calculate epidemiologic measures of the disease and to evaluate the reach and coverage of public health programs in line with progress towards the UNAIDS 95-95-95 targets. Several methods have been used so far, each presenting both strengths and weaknesses. Capture-recapture is currently being recommended due to it mathematical ground and defensible results. However, there are still some methodological limitations, including dependencies between samples, inability to reach highly hidden key population subgroups, as well as loss of marks or tags that biases produced population size estimates. With this research project, we aim at addressing list-dependency between samples and tag loss bias that arises during capture-recapture implementation and develop an extended network-tracked capture-recapture approach able to account for harder to reach KP subgroups.

Methods: To achieve the research objectives, firstly, we derived and applied a Generalized capture-recapture (CRC) model for population size estimation (PSE) from Bayesian model averaging to address list-dependencies between samples; and secondly derived an extended network-tracked capture-recapture method for obtaining population size estimates from a single Respondent Driven Sampling (RDS) that addresses tag loss bias on population size estimates and able to account for usually missed KP subgroups in multiple CRCCRC studies. After derivation of the model and methods, we applied the concepts to three different national wide studies implemented in Rwanda involving FSWs aged 15 years and above and MSM aged 18 years and above, between 2021 and 2023. Data collection methods commonly used in the three studies included bio-behavioral survey (BBS), three-source capture-recapture (3S-CRC) and Respondent driven sampling was used in selecting participants. R-4.3 software was used for data analysis.

Results: The Generalized capture-recapture model from Bayesian model averaging demonstrates a 71% reduction in standard errors as compared to Bayesian Latent class model. Once applied to the MSM 2021 study, the estimated MSM PSE lies within credible sets ranging from 19,347 to 22,268 with a median of 20,787 vs 18,100 median PSE ranging from 11,265 to 29,708 if the Bayesian Latent class model is used. Whereas, for FSW 2022 study, the PSE of street- and venue-based FSWs in Rwanda was estimated to be within credible sets ranging from 31,873 to 43,354 with a median of 37,647 vs a

35,954 median PSE ranging from 14,736 to 55,215 once Bayesian Latent class model is used. A low tag retention was observed between consecutive capture rounds in CRC implementation corresponding to 59%. The FSW 2023 study, estimated FSW PSE was 98,587 ranging from 82,978 to 114,196 once network-traced capture-recapture method is used.

Conclusion: The results of the analyses featured in this dissertation demonstrate derivation of a Generalized capture-recapture model from Bayesian model averaging that overcomes sample dependencies that arises in capture recapture studies. The Generalized capture-recapture model demonstrates a 71% reduction in standard errors as compared to Bayesian Latent class model. Furthermore, the novel network-traced capture-recapture method developed, brings three estimators into practice namely, Cross-Sample, Cross-Alter, and Cross-Network that are free of tag-loss problems and demonstrated the ability to reach unreached KPs subgroups once CRC method is applied. This research brings a substantial contribution in the field of population size estimation concerning KPs and produces more reliable population size estimates, which have implications for the allocation of limited public health resources to marginalized populations.

Keywords: Capture-Recapture, Population Size Estimation, HIV, Hard-to-Reach, Key Population, Rwanda.

Contents

List of tables

List of figures

Publications

1. *Tuyishime E, Kayitesi C, Musengimana G, Malamba S, Moges H, Kankindi I, Escudero H, Habimana Kabano I, Oluoch T, Remera E, Chukwu A.* Population Size Estimation of Men Who Have Sex with Men in Rwanda: Three-Source Capture-Recapture Method
   JMIR Public Health Surveillance 2023;9:e43114
   URL: https://publichealth.jmir.org/2023/1/e43114
   DOI: 10.2196/43114

2. *Tuyishime E, Remera E, Kayitesi C, Malamba S, Sangwayire B, Habimana Kabano I, Ruisenor-Escudero H, Oluoch T, Unna Chukwu A.* Estimation of the Population Size of Street- and Venue-Based Female Sex Workers and Sexually Exploited Minors in Rwanda in 2022: 3-Source Capture-Recapture
   JMIR Public Health Surveillance 2024;10:e50743
   URL: https://publichealth.jmir.org/2024/1/e50743
   DOI: 10.2196/50743

1. *Tuyishime E, Kayitesi C, Remera E, Malamba S, Habimana Kabano I, Unna Chukwu A.* Estimating the Size of Hard to Sample Populations: A Comprehensive Study on Female Sex Workers and Sexually Exploited Minors in Rwanda using Private Network Sampling in 2023

   PLoS ONE 19(5): e0300637

   URL: https://doi.org/10.2471/journal.pone.2342748

Acronyms

2S-CRC:	Two-source capture-recapture
3S-CRC:	Three-source capture-recapture
CI:	Credible interval
CRC:	Capture-recapture
CS:	Credible Set
FSW:	Female Sex Workers
LCMCR:	Bayesian nonparametric latent-class capture-recapture package
MCMC:	Markov Chain Monte Carlo
MoH:	Ministry of Health
MSM:	Men Who Have sex with Men
NSP:	National Strategic Plan
PNS:	Privatized Network Sampling
PSE:	Population Size Estimation
PWID:	People Who Inject Drugs
RBC:	Rwanda Biomedical Center
RDS:	Respondent-Driven Sampling
RPHC5:	5th Rwanda Population and Housing Census
UR:	University of Rwanda
WHO:	World Health Organization

Acknowledgements

I want to thank God most of all for being with me in all difficulties and for allowing this study to be completed successfully.

I would also like to thank my family and friends for their endless words of encouragement and support. I will be forever grateful to my PhD support group for being very supportive and inspiring during my ups and downs.

I am gratefully to my supervisory team, Prof Angela Unna Chucku and Dr. Ignace Kabano for their priceless support and considerable guidance and advice through this PhD Journey. I would like also to thank the African Centre of Excellence in Data Science, University of Rwanda for providing everything needed during my study journey. My gratitude also goes to my work organization, the U.S Centers for Disease Control and Prevention, Rwanda for always allowing me to pursue my PhD and for giving me time to work on research as well as access to the needed materials and software.

I am also grateful to Rwanda Biomedical Center (RBC) for providing data access throughout this process. Thank you!

Chapter 1

Introduction

1.1. Background

1.1.1. Introduction

Globally, gay men and other men who have sex with men (MSM), Female sex workers (FSW), transgenders (TG), people who inject drugs (PWID), and people in prisons and other closed settings are considered the five main key populations (KPs) that are particularly vulnerable to HIV and frequently lack adequate access to health services [1]. Some studies conducted on the HIV/AIDS epidemic have revealed the high burden of HIV infection among KPs.

The World Health Organization (WHO) highlights the need for focused efforts on KPs who are particularly vulnerable and disproportionately affected by HIV due to some specific risk behaviors. The common risk factors include; marginalization, and structural factors such as stigma, discrimination, violence, human rights violations, and criminalization, which contribute to the lack of access to prevention and treatment services, and hence become the key drivers of new HIV transmission in the general population [1, 2]. During 2022, in Eastern and Southern Africa, 54% of the total number of new HIV infections were reported among the general population, and the remainder were reported among KPs, with 13% of those reported among FSW mainly because of multiple partners, a low rate of condom use, stigmatization, and marginalization [3].

Whether the receptive partner is male or female, unprotected receptive anal sex carries a far higher biological risk than unprotected receptive vaginal sex; the risk of HIV transmission during anal intercourse may be up to 18 times higher than that during vaginal intercourse [1].

Consequently, MSM have a heightened vulnerability to HIV infection. HIV prevalence among MSM is estimated to be 21% worldwide, 6% in sub-Saharan African nations, and ranging from 3.8% up to 31% [6,7]. Furthermore, due to systematic issues including discrimination, self-stigmatization, and limited access to resources, people with marginalized sexual or gender identities or behaviors may not always be able to protect themselves against HIV infection [8]. In low-prevalence nations and West and Central Africa, the HIV prevalence ratios are very high [9].

1.1.2. Trends of HIV among KPs in Rwanda

In Rwanda, FSWs and MSM are considered the top two key population groups for HIV prevention and treatment focus due to their high-risk behaviors for contracting and/or transmitting STIs/HIV. They are often stigmatized and marginalized and relatively disproportionately affected by HIV. Due to a continued observed high prevalence of HIV among FSW and MSM as compared to the general population, they are considered as persistent niche of HIV and are thus treated as the bridge of HIV infection to the general population [4-6]. The HIV prevalence among FSWs in Rwanda has exhibited a downward trend over the years [7]. In 2010, the HIV prevalence among FSW in Rwanda stood at 50.8% [5]. Subsequently, in 2015, it decreased to 45.8% [8], and the most recent survey conducted in 2019 indicates a further reduction to 35.5% [9]. HIV prevalence among MSM stood at 6.9% by 2021 from 4.0% back in 2015 [11]. According to Rwanda's HIV and AIDS national strategic plan [12], FSW and MSM are two of the main priority populations for HIV prevention and care.

Estimating the size of FSW and MSM populations is crucial because it generates denominators for program design, planning, and implementation. In addition, it informs interventions, public health initiatives, provides useful parameters inputs for modeling, and helps to determine the appropriate resource allocation, gaps and unmet needs to serve such communities.

1.2. Problem statement

Attempts to estimate the KPs population size are frequently thwarted in ways that limit traditional survey tactics including Census and enumeration due to several factors that lead to their rare visibility in communities and hence become hard-to-reach. Various methods have been used thus far to estimate the population sizes of hard-to-reach groups, each with its own set of strengths and drawbacks. Unfortunately, there is no gold standard approach, and various methodologies frequently result in contradictory conclusions [10]. The capture-recapture (CRC) method has been commonly used owing to its mathematical ground and defensible results [11-13]. However, two methodological concerns remain, including list dependencies between samples, inability to reach key population subgroups specifically those that do not attend specific venues at specific time, as well as loss of marks or tags during capture-recapture study implementation that biases produced population size estimates. With this research, remedial approaches and tools are developed and applied to the real-world capture-recapture data.

1.3. Objectives

1.3.1. Main Objective

The general objective of this research is to address list dependencies between samples, tag loss bias on PSE, and derive an extended network-tracked CRC method from a single Respondent-Driven Sampling (RDS) that is free from tag loss and capable to reach unreached KP subgroups once CRC method is used.

1.3.2. Specific objectives

- ❑ To derive a generalized capture-recapture (CRC) model for population size estimation (PSE) by applying the Bayesian model averaging to address list dependencies between samples.

- ❑ To address tag loss bias on population size estimates in capture-recapture studies.

- ❑ To derive an extended network-tracked capture-recapture method for obtaining population size estimates from a single Respondent-Driven Sampling survey.

1.4. Rationale of the study

For capture-recapture method, 4 major assumptions must be met to give reliable population size estimates. These assumptions include that individual captures should be independent, the population should be closed during the data collection period, each target population member's capture history should be correct (no tag or mark is lost), and the chance of getting captured should be homogeneous [17]. Sometimes, the method's underlined assumptions are not met mostly due to some uncontrollable circumstances, hence affects resulting population size estimates.

Researchers have been setting up preventive procedures to meet the CRC methodology underlined assumptions. To ensure that the population is closed, the investigators tend to limit the study implementation withing a short timeframe, and the KP tagging is done randomly to ensure that the probability of being tagged is homogeneous. However, to control that each capture history is correct for each individual, correctly identification of individual who were tagged is crucial. Sometimes it is challenging due to some tagged individuals lose their tags and becomes hard to correctly confirm that they have been tagged during a consecutive capture round, hence biasing the produced population size estimates.

During the implementation of a capture-recapture study, sampled individuals from one capture round to another might be related, yet this breaches the CRC underlined assumption, hence leading to a biased population size estimate. Researcher have been adopting some ad hoc preventive procedures to overcome the sample dependence issues, however it is not controllable and ascertain that the sample dependence is resolved and there is no need to be accounted in the estimation process.

In Rwanda, there have been three rounds of FSW PSE dated 2010 [22], 2018 [23] and 2022, all using Time Location Sampling (TLS). As the digital era emerges, with technologies reshaping and reorienting sex markets [24, 25], the use of venue-based sampling approaches might be missing a chunk of FSW who never congregate in the venues, including those practicing sex work at home (home-based) and those who get clients from internet platforms (internet-based).

With this research, we aim at providing remedial solutions to the above-mentioned limitations arising during the implementation of capture-recapture studies.

1.5. Scope of the study

With this research, we develop remedial solutions by deriving and applying a Generalized capture-recapture (CRC) model for population size estimation (PSE) using Bayesian model averaging to addresses list-dependencies between samples; and derived an extended network-tracked capture-recapture method for obtaining population size estimates from a single Respondent Driven Sampling (RDS) that addresses tag loss bias on population size estimates and able to reach usually missed KP subgroups in multiple CRC studies. The study only focuses on the population size estimation of female sex workers and men who have sex with men and applied the derived methods to the real-world data collected between 2021 and 2023 using both CRC and RDS methods in Rwanda.

1.6. Organization of the thesis

This thesis is organized into six chapters. It offers a general introduction to the study in chapter one. Furthermore, chapter one describes the objectives, research questions, motivation of the study, and scope of work. Chapter two highlights available literatures around the topic, discussing previous research, and key concepts related to the research topic. Chapter three focuses on the methods of the research to meet research set objectives. Chapter four presents the findings and results obtained. Chapter five discusses the results that have been obtained. Lastly, Chapter six concludes the thesis and provides recommendations for further research and published manuscripts.

1.7. Research motivation.

Global HIV prevention is increasingly prioritizing the estimation of the size of hidden populations. By estimating the population prevalence of HIV infection and computing an attributable percentage, epidemiologists can estimate the burden of disease in relevant populations. Furthermore, estimations of population size are helpful parameter inputs for population dynamics models that account for social and sexual networks in the spread of disease. Finally, estimations of the population size aid in assessing the coverage and reach of initiatives aimed at groups at risk.

To estimate the population sizes of hard-to-reach groups, a variety of techniques have been employed thus far, each with unique advantages and disadvantages. Capture-recapture[14], multiplier [15], network scale-up [16], and successive sampling [17] approaches are all common. Unfortunately, there is no gold standard approach, and various methodologies frequently result in contradictory conclusions[10].

Considering its' mathematically ground and plausible conclusions, the multiple sources capture-recapture method has been widely employed in epidemiology to estimate the number of key populations targeted by health interventions for certain health conditions [11, 12, 18]. The Three-source capture-recapture (3S-CRC) approach has been utilized in several studies to estimate the size of specific population categories, such as FSW, MSM, and PWID [19-22].

This research project aims to contribute to existing knowledge by deriving a Generalized CRC model for PSE from Bayesian model averaging that addresses list dependencies between samples, addressing tag loss bias on population size estimates in multiple capture-recapture studies, and to explore an extended network-tracked capture recapture method for obtaining population size estimates. All the proposed methods are applied to estimate the population size of hard to sample population groups in Rwanda, focusing on Female Sex workers and Men who have Sex with Men.

Chapter 2

Literature review

2.1. Introduction to the literature review section

The literature review chapter provides HIV epidemic context globally, regionally, and at the country level. Furthermore, it provides existing challenges in HIV care and treatment programs related to the missing critical data to inform policies and interventions. And lastly, this chapter describes the importance and existing limitations in line with the efforts to estimate the population size of key populations highly affected by the HIV epidemic.

2.2. Global and country specific HIV context

Two decades ago, the global AIDS pandemic seemed unstoppable. More than 2.5 million people were acquiring HIV each year, and AIDS was claiming two million lives a year. In parts of southern Africa, AIDS was reversing decades of gains in life expectancy. Effective treatments had been developed but were available only at prohibitively expensive prices, limiting their use to a privileged few people [1, 2]. As antiretroviral therapy (ART) availability and accessibility evolves over the years, numbers of new HIV infections and AIDS-related deaths have continued to decrease globally, bringing the AIDS response closer to achieving Sustainable Development Goal (SDG) 3.3 of ending AIDS as a public health threat by 2030.

Globally in 2022, out of the 39.0 million [33.1 million–45.7 million] people living with HIV, 86% [73%–98%] knew their HIV status, 76% [65–89%] were receiving antiretroviral therapy, and 71% [60–83%] were virally suppressed [1, 2].

HIV infection rates among some subpopulation groups are still disproportionately high [23]. HIV research in sub-Saharan Africa, is now interested in same-sex practices. Recent research findings show that MSM groups are common in sub-Saharan Africa and that their HIV infection rates are substantial [24]. Unprotected receptive anal sex has been found to carry a risk that is approximately 18 times higher biologically than unprotected receptive vaginal sex [23]. Because of this, MSM have increased chance of acquiring HIV infection. By this fact, MSM are at higher risk of contracting HIV infection. The global estimated HIV prevalence among MSM remains high, with the highest prevalence found in sub-Saharan African countries [25, 26]. In addition, 54% of all new HIV infections in Eastern and Southern Africa in 2022 were reported among KPs; 13% of these infections were reported among female sex workers (FSW), primarily due to the fact that they had multiple partners, a low rate of condom use, stigma, and marginalization [3].

Rwanda is located in East Africa, that has borders with: Tanzania, Uganda, the Democratic Republic of the Congo (DRC), and Burundi. The nation is organized into five administrative regions, which comprise the City of Kigali and four provinces, with thirty districts as additional subnational unit levels. Rwanda is facing a mixed HIV epidemic, with features of a concentrated pandemic among specific key population subgroups at increased risk of acquiring HIV infection, FSWs [6] and MSM [27], and a generalized epidemic across adults with an HIV prevalence stabilizing at about 2.7%. According to Rwanda’s HIV and AIDS national strategic plan (NSP2018-2024), MSM and FSW are ranked the top key populations for HIV prevention and care focus [28].

2.3. Challenges in HIV care and prevention programs

The focus of attention has switched to major impacted populations in sub-Saharan Africa as nations work toward controlling the HIV epidemic, such as FSW, PWID, TG, and MSM who are more likely to contract HIV [29]. The World Health Organization (WHO) highlights the need to focus on key population subgroups (KPs) who are particularly vulnerable and disproportionately affected by HIV due to several risk factors including behaviors, social norms, and political, which contribute to a lack of access to prevention and treatment services [1, 2] and hence become the remaining key drivers of new HIV transmission in the general population.

The members of key populations are usually hard to reach with HIV treatment and prevention programs due to country specific culture norms, political context, and social norms that consider sex work and homosexuality severely criminalized. At present, MSM in countries where homosexuality is criminalized face a nearly five-fold increased risk of contracting HIV in comparison to MSM in countries where homosexuality is not criminalized [30].

The willful neglect, denial and apparent ignorance about the epidemiological situation are the top line major obstacles to effective prevention and treatment for people from key populations. A surprising number of countries lack targeted programs, size estimates and HIV data for key populations. Occasionally, the lack of data results in the omission of crucial information that policymakers and planners rely on to monitor the management of the HIV epidemic. This information is used to assess needs, coverage, and the spread of new HIV infections.

Furthermore, increased risk for HIV transmission among KPs is highly associated with social marginalization, and those individuals who are socially marginalized may not identify themselves as such when accessing health services. This makes it difficult to track them in HIV program registers and impedes efforts to plan and have informed resource allocations for high impact.

2.4. Population size estimation for Key population groups highly affected by HIV.

For years, it has been difficult to estimate the size of hard-to-reach population group because of inability to construct sampling frame, being dynamic in both place and time, and the possibility that members may be reluctant to identify themselves a s such. Many techniques are proposed for estimating population sizes, each with pross and cons making it harder to get an optimal preference. These includes Venue-based sampling [31], Time-location sampling (TLS) [32], Respondent-Driven Sampling (RDS) [33], Multiplier Methods[34], Network Scale-Up Method (NSUM) [35], Successive Sampling-Population Size Estimation (SS-PSE) [36], and Capture-Recapture (CRC) [37]. Numerous methods have been employed, each with unique advantages and disadvantages, and there is currently no one gold standard.

The capture-recapture (CRC) methods have been shown to be useful in estimating the size of hard-to-reach populations [38-40]. CRC methods have a long history in wildlife population assessment studies, and later were adopted in human subject studies mostly population size estimation of hard-to-reach. CRC methods were for the very first time adopted in human studies, dating back to 1786, when Laplace attempted to estimate the population of France [41]. The traditional capture-recapture (CRC) method employs two captures and involves the use of the Peterson estimator [42]; First, n_1 subjects are sampled, marked and released back into the population of interest. After a certain amount of time, a second sample of size n_2 is drawn and the number of marked subjects m_2 is counted. The marked fraction in the second sample allows for the estimation of the initial population size.

$$\widehat{N} = \frac{n_1 n_2}{m_2} \tag{1}$$

With Peterson estimator from Equation (1), change in population during the study period is not accounted for. Capture-recapture models are broken down into two major categories- open and closed population models. Closed populations models assume that population size is constant over the study period, while Open population models do not, the scope of this thesis focuses on Closed population models. Under Closed population models, there are Discrete and continuous-time models, in a typical discrete-time model, the target population is sampled several times (or over a certain number of occasions), and for each occasion, any subject captured can be counted only once. Closed CRC models rely on four key assumptions in order to produce accurate population estimates: each capture is independent; the population is closed; the capture history of each member of the target population is accurate; and the probability of being caught is uniform [43]. Maximum count for each subject is the number of samples.

Intuitively, for independent samples, when recaptures in the subsequent samples are few, we know that the size is much larger than the number of distinct captures. On the other hand, if the recapture rate is high, then we are likely to have caught most of the subjects. CRC is an empirical population size estimation method (PSE) that yields population size estimates with greater precision because of its statistical basis. According to research, performing a CRC with additional (three or more) sampling/capture rounds enhances the design, yields more reliable estimates, and eases the sample independence requirement (assumption) in comparison to two-source CRC [44].

There are two fundamental concerns that arise when modeling CRC data, including List dependence and Capture heterogeneity. Every capture event should, in theory, constitute an independent draw from the population. List dependency, on the other hand, refers to the fact that various capture events may frequently be associated. On the other hand, each subject in the population should have the same probability of being captured; however, some subgroups may have increased propensity for capture. Furthermore, there are three effects influencing capture probability: temporal effect (subscript t), behavior effect (subscript b), and individual heterogeneity (subscript h) and all the possible interactions might occur, leading to eight fundamental Closed CRC models: M_0 (no variation effect), M_t, M_h, M_{th} , M_b , M_{tb} , M_{bh} , M_{tbh}. The analysis of closed CRC data amounts to finding the best fitting model and estimating the population size from the chosen model. The selection of the best fitting model is challenging since all three sources of variability might be present in varying degrees, and these reasons might lead to model selection bias.

The results of 3S-CRC are mathematically supported and tenable, which has made it a popular tool in epidemiology to estimate the size of the key population targeted by health interventions for certain health disorders [11, 12, 18]. Using a sampling frame-free approach, the 3S-CRC method has been employed in numerous studies to estimate the size of certain population subgroups [19-22]. To date, there have been four rounds of studies aimed at estimating the population size of FSW and only one study for MSM in Rwanda, with the commonality of all using Time Location Sampling (TLS) methodologies [45-47]. Two-sources capture-recapture, enumeration, and multiplier were listed among the methods used, with a commonly stated methodological limitation of inability to tackle within non-venue-based FSW and leading to potential underestimation of FSW.

With this research project, we aim at developing a Generalized Capture recapture model for PSE derived from Bayesian model averaging process and address list dependency bias on population size estimates in multisource capture-recapture. Furthermore, we explore the trust embedded in social ties as well as the strengths of capture recapture to come up with a novel approach known as Privatized network Sampling (PNS). PNS is a PSE method that is capable of reaching different FSW subgroups, including non-venue-based subgroups [48], free of tag loss, and produces financial resources effective and more credible inferences about population size, given that it is built into respondent driven sampling (RDS) with no additional cost added.

Chapter 3
Methods

3.1. Introduction

Chapter 3 describes methods used to achieve the objectives of the research project. Firstly, the chapter provides step-by-step development of the Generalized CRC models as well as the network-traced capture-recapture approach with the derivation of corresponding estimators. Furthermore, this chapter continues by demonstrating the application of the developed methods to real-world data by describing the study population, study design and settings, sample size calculation and sampling, data management, and statistical data analysis.

3.2. Study population.

Adult men aged at least 18 years who self-report as gay or bisexual or who have had anal intercourse with a man during the past year and who have lived predominantly in Rwanda for the past year are included in the study population. Living predominantly in Rwanda means that, despite the potential of leaving the nation, you have spent the majority of the last 12 months in Rwanda. Any MSM who was unwilling to participate voluntarily was omitted from the study. Furthermore, the study population includes biologically born females (girls or women), aged 15 years and above, who self-reported having any type of sex with men in exchange for goods, money, or services in the last 3 months and practicing sex work at street- and venue-based hotspots as well as at other non-venue-based places including but not limited to home-based, internet-based and using pimps. Those fulfilling the above criteria and who are under 18 years of age are here referred to as sexually exploited minors.

3.3. Study design and setting.

3.3.1. MSM Population Size Estimation using Three-source capture recapture method.

To estimate the size of MSM in Rwanda, the three-source capture-recapture (3S-CRC) approach was employed. In a typical CRC method, a random subset of the population of interest is marked during the initial encounter in the capture-recapture method. The number of people who were first marked is later observed by drawing in another segment of the population. The estimate of the population size decreases with increasing rate of observing tagged individuals in the second sample, tagging can be repeated as often as needed. The 3S-CRC is a robust method that has been demonstrated to be effective in estimating populations without a sample frame, such as FSWs, MSM, and people who inject drugs (PWID) [19]. It is further detailed in other sources [49]. The MSM 3S-CRC began with a capture stage. Members of the MSM population were "encountered" at this phase, and they were then "marked" by giving them a unique present that was difficult to buy on the local market. A second capture (recapture) was started a week later by providing MSM-friendly services to MSM across the nation. A Respondent-Driven Sampling (RDS) technique was employed during the third capture (recapture) [50, 51], with the inclusion of particular questions to identify those MSM met on the other capture occasions.

MSM community-based organizations (CBO) chose MSM key informants during capture one to help distribute unique objects across MSM's associations, groups, and non-members of any associations or MSM groups. A list of MSM associations and groups was created, together with the number of members for each in each province across the nation: Twelve from the City of Kigali, ten from the Southern province, eight from the Northern and Western provinces, and fourteen from the Eastern province. The distribution of distinct objects was allocated a specific color for each province. Probability proportional to the size of each association or group was used to determine the number of objects to be distributed within each association or group. Because MSM associations and groups already have a list of members, systematic sampling within an association or group was found to be an appropriate method for determining who gets the unique object. The method involves selecting a random start point and sampling interval from the list of members of MSM associations or groups.

Within an association or group, each picked MSM was given three unique objects: two for him and two to distribute to other MSM he knows and who are not members of any associations or groups. With the assistance of MSM key informants, a skilled team of distributors managed the distribution of unique objects inside an MSM association or group. As distinctive items, branded keychains costing little more than $3 US were utilized. The study goals were given to MSM, and they were counseled to retain the unique objects they had received in a secure location in case they were subsequently required to provide them for validation. The completion of this task required a week.

It was possible that some items would not be correctly distributed and returned, or that the same person would receive more than one object. A debriefing was conducted with the MSM key informant and the MSM association/group members to reduce any potential bias on the population size estimates. The debriefing covered study objectives, eligibility requirements for receiving the object, the object distribution process, and important points such as asking the object receiver to confirm if he had not been approached by another person in the same study context to prevent duplication. Furthermore, the $3 monetary worth of the unique object was chosen to reduce the likelihood that the distributor of the object would want to keep the items for themselves or that the recipient would be ready to accept more than one item. The object distribution procedure was tracked every day by the object distributor and the MSM key informant, who reported the quantity of objects that were successfully delivered and those that were unsuccessfully distributed and had to be physically returned.

The second capture was started the next week, and MSMs were tagged with certain MSM-friendly services. MSM key informants and their CBOs assisted in the selection process for assistance. Since MSM are provided with health services through standard medical facilities that are furnished with MSM-friendly environments and packages, these facilities were utilized to provide MSM-friendly services, such as condom and lubricant distribution. A key informant who is MSM and a healthcare professional who typically treats MSM at the same facility were assigned to give the chosen service to MSM and to document relevant data for the study at 23 health facilities chosen nationwide for this purpose. In order to provide a warm and inviting environment for MSM, the key informant's duty was to act as the receptionist and enable screening, while the health worker's function was to deliver services and document pertinent data. Community mobilizers worked on impending service provision to MSM prior to the second capture in order to raise awareness and dispel the stigma and fear associated with participation. After being counted as captured, those who accepted the offered services were asked about whether they had received the delivered unique object during the preceding week. Responses (yes/no) were noted based on the MSM's possession of the present and his ability to accurately identify the special object he had been given.

The Integrated Behavioral and Biological Surveillance Survey (IBBSS) employing RDS [52] , came after capture two, providing the means for the third capture. RDS is a type of sampling that uses the concepts of biased networks and Markov-chain theory to lessen biases that are typically present in chain-referral techniques [51]. It has been demonstrated that, unlike most chain-referral samples, sampling starts with a purposefully selected group of initial subjects, but the final sample's makeup is entirely independent of those initial subjects [53]. All MSM who were recruited by their peers through the RDS technique were counted as captured during this third capture. Eight study sites were dispersed across the administrative provinces: two in the Western province, three in the city of Kigali, and one in each of the Northern, Southern, and Eastern provinces. Each referral chain started with MSM seeds, those who satisfied the study eligibility requirements and were well-liked and regarded by their peers. Through the Implementing Partners, which include NGOs and CBOs that assist the MSM community, the investigator had contact with the seed during this study. For a total of 24 seeds, three were chosen at each study location.

Participants in the RDS survey were asked if, in the second week of the study's implementation, they had received either unique objects during capture one or supplied services during capture two. Several prompts were utilized to verify previous involvement, such as physically receiving the gift or properly identifying it on a laminated card that had pictures of several objects. MSM were prohibited from taking part in the study more than once in a single round. To ensure that each participant was only recorded once during capture three and to identify any duplication both inside and between study sites, an interoperable fingerprint system was employed.

Every study site had fingerprint machines installed and linked to the internet to facilitate simple, instantaneous data synchronization. Following the recording of a fingerprint, the data was automatically translated into alpha numerical codes and sent to the central server to synchronize all research site-level data. This established technology allowed us to recognize any MSM attempting to re-register his fingerprint at the same or a different study site. The fingerprint machines kept the alpha-numerical codes as participant IDs to protect participant confidentiality after converting recorded fingerprints into codes that could not be reversed.

3.3.2. FSW Population Size Estimation using Three-source capture recapture method.

Using the three-source capture-recapture approach (3S-CRC), a cross-sectional, countrywide FSW and sexually exploited minor population size estimation was conducted [54]. The method involved visiting hotspots where FSWs are known to congregate on three separate occasions, and sampling FSWs that are found at the hotspots on each occasion, calculating the degree to which FSWs samples overlapped across three consecutive occasions. In this framework, an encountered FSW at the visited hotspot is referred to as captured, and each encounter occasion is referred to as a capture round in the CRC method context. A resampled FSW at a subsequent capture round was referred to as recaptured, and the intuition is that the degree to which FSW samples overlap across the three consecutive capture rounds is inversely proportional to the population size.

The objects used to tag FSWs who were presented at hotspots were small, inexpensive, and branded with specific messages so that they would have a memorable design and only be available from the study staff who distributed them. During capture one, a small bag branded with the "imigongo” traditional art form was offered; for the second capture, a purse branded with a flower and the key message "Rinda ubuzima" (‘Protect your life’) was offered; and during the third capture, a hair comb branded with a tree picture as a key message was offered.

A stratified multistage sampling design was used, with administrative provinces considered as strata and FSW hotspots as primary sampling unit (PSU). Information from FSW’s hotspot mapping exercise was used as the sampling frame for this FSW PSE 2022.

Prior to this survey, the Rwanda Biomedical Center (RBC) conducted a FSW hotspot mapping exercise across the country from March to May 2022 to collect some key information that would inform future studies involving FSWs. Hotspot mapping consisted of teams going to the field to identify active venues and streets where FSWs congregate to find sexual clients. The FSWs hotspot mapping exercise was facilitated by key informants identified by implementing partners who provide health services to FSWs to guide mapping teams. The mapping exercise identified 668 hotspots (street- and venue-based) countrywide and collected some beneficial data, including hotspot name, hotspot size, pick days, pick hours, and corresponding geo-coordinates, to guide the sampling process.

The principal sampling processes were as follows: Using the national list of FSW hotspots resulting from the hotspot mapping exercise, FSW hotspots were stratified by administrative provinces and the City of Kigali, and then a specific number of hotspots was selected using probability proportional to the number of FSW hotspots within each of the 4 provinces and the City of Kigali. Hotspot sampling was performed using probability proportional to size (PPS) for generating PPS samples. In PPS sampling, the probability that a hotspot was sampled was proportional to the estimated size of FSWs observed at that hotspot during the hotspot mapping exercise. In practice, this means that hotspots with many FSWs are more likely to be sampled than hotspots with fewer FSWs.

To enhance the geographical representativeness of the sample, hotspots were listed by corresponding administrative provinces, and provinces were considered strata. To execute hotspot sampling, we listed all hotspots in order of the number of FSW observed during mapping exercise within a stratum (to reflect the relative sizes of the FSW populations), calculated the cumulative number of FSWs for each hotspot listed, determined the sampling interval, picked a random starting point, and finally selected a hotspot based on the random starting point, sampling interval, and cumulative FSW population size. This process was repeated at each capture round to minimize list dependency between capture occasions and resulted in selecting 62 hotspots countrywide at each capture round.

3.3.3. FSW Population Size Estimation using Privatize Network Sampling method.

To inform the development of this estimation, a formative assessment (FA) was conducted. A group meeting that included implementing partners (IP), stakeholders, and FSWs was convened, and focus group discussions (FGD) were conducted. FSWs from 5 provinces in Rwanda came to Kigali for a one-day meeting on March 10th, 2023. The objectives of the FA included the identification of sociocultural factors limiting or facilitating access to FSW, assessing the feasibility of the planned method and procedures, and identifying barriers and strategies to overcome them.

Representatives from implementing partners' institutions who usually serve FSW communities attended the FA, as well as 10 FSWs from 5 provinces presenting disparities in demographic characteristics and how they reach clients. Among the participating FSWs, 2 were under 20 years old, 2 were home-based, 1 uses internet-based platforms to reach clients, 2 belonged to a network of university FSWs, and the other 3 were street- and venue-based FSWs. The FA assessment focused on three main themes: study design and procedures, characteristics of the study population, and survey logistics.

For the study design and procedure's theme, we got confident in the proposed method (PNS), decided to use name initials and the last 5 digits of one's phone number combination to form the unique identification number of participants, informed study site preparation and setting, informed coupon (invitation) design, learned that there is a need for the engagement of local government, and received insights on the content development of training materials for data collectors. Regarding the second theme, we learned that some FSW subgroups are extremely hard to reach, including FSWs in the university students' networks, those using pimps, and middlemen, and this information guided us on the outreach strategies. Finally, the FA has informed the logistic component of the study, such as compensation for participation, including a FSW within each of the study teams to serve as a receptionist, and consideration for an electronic coupon (invitation).

This estimation utilized one single method of population size estimation, Privatized Network Sampling (PNS). This method utilized network data collected using the questions specifically developed for this purpose. Within a bio-behavioral survey (BBS) questionnaire that used RDS to sample FSW, more questions were added collecting information on the degrees at which FSW population are networked, hence used for population size estimation purposes.

Data were collected countywide in 10 study sites, which included Gihundwe Health Center (HC), Kibuye HC, Gisenyi HC in the West; Gitarama HC and Rango HC in the South; Muhoza in the North; Mukarange HC and Nyagatare HC in the East; Remera HC and WE-ACT FOR HOPE CLINIC in the City of Kigali. Data collection was performed between May 8th, 2023, and June 24th, 2023.

The PNS sampling followed the Respondent Driven Sampling (RDS), which is a probability-based chain-referral sampling methodology used to sample FSWs for a biobehavioral survey (BBS). For the RDS process, initial survey participants ("seeds") were purposively recruited by the survey team to start enrollment. Criteria to be a seed, one should have been engaged in commercial sex work at least 12 months prior to the estimation, well-connected within FSW social networks, well regarded by peers, able to communicate with data collectors, and supportive of estimation goals. Furthermore, three seeds by site were purposively recruited reflecting diversity in sociodemographic characteristics (e.g., age, sexual orientation and gender identity, education, area of residence, marital status, language, religion), HIV status, and affiliation with a KP organization or KP service provider.

Seeds and other subsequent recruiters were provided with a maximum of 3 coupons to distribute to their peers in their FSW social circle for recruitment. Instructions for peer recruitment using a recruitment process script were provided to seeds and participants by staff at the study site. When potential recruits came to the survey site, they were screened for eligibility and enrolled if they met the survey inclusion criteria and consented to participate; at this point, they were considered participants. After participating in the survey, these individuals were given their own recruitment coupons and asked to distribute them to their peers that they knew are FSW. This process continued until the target sample size and survey parameters were achieved.

However, RDS data contains limited information about participant's network. To collect major identifiable information about how individuals in the sample are related to one another, for each recruited respondent, using a cryptographic hash function, a hashed (anonymized) ID was created from the initials of the first and last name and the last 5 digits of the respondent's phone number using Tele funked coding [55].

Furthermore, each recruited respondent was asked how many of the total network size peers the recruit knows their name and phone numbers, and a hashed ID was also created for up to 5 peers in the respondent's personal network. If the respondent stated knowing 5 or fewer peers, a hashed ID was created for each of them. If the respondent knows more than 5 peers, then 5 peers were selected in a near-random fashion using an age-related selection process of peers with an age that was closest to the participant. The entered data (initials, last 5 digits of phone number) were not stored, and the hash ID was stored. The hashed ID could not be used to reconstruct the respondent's provided data. This information helped to evaluate the rate at which participants' networks contained other sampled participants.

3.4. Sample size and sampling.

3.4.1. MSM Population Size Estimation using Three-source capture recapture method.

The proportion of all males aged 18 and over who had at least one male sexual partner in the previous year by province (0.30% City of Kigali, 0.16% Eastern, 0.08% Northern, 0.24% Southern, 0.24% Western) was used to estimate the pooled and provincial level stratified sample size for the first two initial captures. This was based on Rwanda Population-Based HIV Impact Assessment (RPHIA,2019). We calculate the necessary minimum sample size by assuming a design effect of 1.5, a precision of 0.5%, and adjusting for the 15% loss of coupons from a prior study related to CRCs [47]. For each of the first two captures, we calculated that a total of 2,705 items would need to be redistributed across the provinces as follows: 803 in Kigali city, 586 in Western province, 658 in Southern province, 219 in Northern province, and 439 in Eastern province.

Based on the outcomes of the 2020 Rwandan Integrated Bio-Behavioral Survey (IBBS), the sample size for the RDS survey was determined. To calculate the sample size, the MSM HIV prevalence was 11.3% in Kigali, 6.4% in the western province, 1.4% in the southern province, 3.1% in the northern province, and 1.2% in the southern province, and that the non-response rate is 10%, the design effect is 1.5 and the precision ω is 0.025. The formula for determining the sample size for an RDS study was applied, according to *Salganik's* 2006 [56]. Estimated minimum total sample size across the nation is 2,210, broken down into the following provinces: 1027 in Kigali city, 613 in Western province, 141 in Southern province, 308 in Northern province, and 121 in Eastern province.

Using MS-CRC Power Analysis of the *shinyrecap* application, the estimated sample sizes for each capture and statistical power were verified [57].

3.4.2. FSW Population Size Estimation using Three-source capture recapture method.

Statistical power and the anticipated sample size for each capture round were determined using the shine recap application's MS-CRC Power Analysis [57]. Using the previously estimated size of FSWs in Rwanda of 23,495 [47], we set the application to simulate 500 capture recapture studies and report the amount of variability in the estimates based on the posited population and the sample size at each capture event to 2,000 at an alpha level of 0.05. We discovered that there is a 95% probability that the population size estimate from the CRC study will fall between 7.6% and the true value, or 1,780 absolute accuracies. Considering the 11% non-response rate from the previous study, 2,000 was found to be an appropriate sample size for every cycle of capture (*Appendix 1*). The number of objects distributed to each hotspot was proportional to the total number of FSWs estimated at the hotspot according to 2022 mapping data.

To select the number of FSWs to be offered unique objects (UOs) within a selected FSW hotspot, a systematic sampling approach was used for the distribution of UOs. The Unique Objects distribution process started with the FSW key informant conducting visual head counts of FSWs present at the hotspot, then estimating the distribution interval by dividing the head counts by the assigned hotspot U.Os. If the result of the division was one, every FSW present at the hotspot should receive the unique object; otherwise, a random start would be randomly selected within the distribution interval following the physical standing position of FSWs in the hotspot. *Table 1* below shows the provincial distribution of the sampled 62 hotspots and 2,000 UOs assigned at each capture round.

Table 1: Provincial level sample size replicated at each of the 3 capture rounds, Rwanda 2022.

	Information from hotspot mapping		**3S-CRC sampling**		
Province	Number of hotspots per province	Estimated total number of FSW at hotspots during mapping exercise	Number of FSW to be sampled	Number of hotspots to be selected and visited	Average number of FSWs to be sampled per hotspot
City of Kigali	100	3,883	346	9	39
Eastern	237	5,825	518	21	25
Northern	74	3,095	275	7	42
Southern	61	2,858	255	5	47
Western	225	6,810	606	20	30
Total	**697**	**22,471**	**2,000**	**62**	

3.4.3. FSW Population Size Estimation using Privatize Network Sampling method.

The computation of sample size relied on the FSW biobehavioral survey (BBS), which was sufficiently powered to estimate the provincial-level HIV. The sample size calculation was based on the province-specific prevalence of HIV among FSWs aged 15 years and above, estimated from the previous rounds of FSW BBS[9].

The design effect (*Deff*) for each province was estimated at 0.998 in East Province, 1.683 in West Province, 1.221 in North Province, 3.245 in South Province and 1.508 in Kigali City. The 95% Z score value was 1.96 with an alpha level of 0.05 (95% confidence), and the finite population correction was applied using the results of 2018 FSW population size estimation (PSE) [47]. As a result, the minimum sample size of 2,500 was estimated: East and South with 415, West with 623, North with 503 and City of Kigali with 544.

3.5. Data management

Data were collected electronically using an Open Data Kit (ODK)[55] installed on android tablets. All collected data were reviewed daily and checked for errors before submission. Daily, data from completed participant’s questionnaires were electronically pushed to a password-protected database to ensure data safety.

Data quality checks were conducted regularly to ensure that high-quality data were generated. Mainly for PNS purposes, the RDS plot was run to view the recruitment graph and check if the tree matches what actually was happening in recruitment, checked for duplicated hashed IDs (duplicates are expected, but here, we checked that these are indeed unique individuals), checked for cases where a single subject reports the same hashed ID value for two of their contacts, and checked for subjects who report the exact same network contacts.

At the end of data collection, all study site-level data in a CSV format (Comma Delimited) were merged with coupon recruitment information from RDS to track for chain referral aspects.

3.6. Statistical analysis

3.6.1. MSM Population Size Estimation using Three-source capture recapture method: Bayesian nonparametric latent class model.

To prepare for analysis, R-4.0.5 for Windows was used to export participant-level data from ODK, and cleaning procedures were carried out using pre-established exclusion criteria and data logical flow, data were analyzed by province. For every provincial-level subset, aggregated datasets with counts of every capture/recapture combination were created. Using aggregate data sets, a Bayesian non-parametric latent-class model was employed to generate the final PSE with 95% credible sets.

Here we introduce a Bayesian Nonparametric Latent Class model (NPLCM) for estimating the size of a closed population from multiple recapture data. This approach, which does not require a separate model selection step, is based on the Dirichlet process mixing of the product-Bernoulli distribution, which allows it to transparently vary its complexity and handle complicated patterns of heterogeneity of captures [58, 59]. Lastly, we describe an effective Markov chain Monte Carlo sampling approach (MCMC) for modeling posterior data from our model.

Expanding upon concepts from Manrique-Vallier and Fienberg, we formulate multiple-recapture estimation as a problem involving missing data [60, 61]. Considering a closed finite population of N individuals. Considering that every person can appear on one or more of the J lists that only include a portion of that population, or not appear at all, we write x_{ij} = 1 to indicate that individual $i \in \{1, \dots \dots \dots, N\}$ was captured by list $j \in \{1, \dots \dots \dots, J\}$, and x_{ij} = 0 to indicate otherwise. We group these capture indicators into individual capture vectors, x_i = $(x_{i1}, \dots \dots .. x_{ij}) \in \{0,1\}^J$. In this line, any individual with a capture vector composed uniquely of zeros, $\boldsymbol{O}$ = (0, ……., 0), is unobserved, and therefore cannot be present in any sample (list).

Let $n = \sum_{i=1}^{N} I(X_i \neq 0)$ be the number of observed individuals. Here, I (•) takes the value 1 if the condition in the argument is true and 0 otherwise. Our task is to determine the number of unobserved individuals, $n_0 = \sum_{i=1}^{N} I(X_i = 0)$ or equivalently the population size $N = n + n_0$.

By following missing data ideas, here we consider a complete data-generation process and a nonignorable missing data mechanism. Let the complete data-generation process be $f(X|\theta)$ for $X \in \{0,1\}^J$, such that $X_i \overset{iid}{\sim} f(X|\theta)$ for $i = 1, \dots\dots\dots, N$ with N Known. The corresponding missing data mechanism consists of not observing subjects with a capture vector **O.** Reordering the sequence of $X_i s$ so that all the unobservable capture vectors are grouped together at the end of the sequence, this is at positions $i = n + 1, \dots\dots\dots, N$, we get:

$$p(\chi|\,\theta, N) = \binom{N}{n} f(0|\theta)^{N-n} \prod_{i=1}^{n} f(X_i|\theta) I(N \geq n) \qquad (2)$$

Where $\chi = (X_i, \dots\dots\dots\dots, X_n)$. Here, the probability mass function of the argument—which may be inferred from the context—is represented by the symbol $p(\bullet)$. Both $N\ and\ \theta$ in the multiple-recapture scenario are unknown and require estimation. This is addressed by specifying a prior distribution $p\ (N, \theta)$ and computing $p(N, \theta|\chi) \propto p(N, \theta) p(\chi|\theta, N)$.

The next task is to account for data heterogeneity in our modeling process, we address this by using the Latent Class Model. Ideally, everyone in the population has the same probability of being captured; however, some subgroups may have increased propensity for capture. Using some sort of stratification is a suggested course of action in this situation. The idea is to split the population into classes that should be fairly homogeneous, as this is where basic models are expected to hold more reliably. The population size estimates are then obtained by applying those models independently to each stratum. The covariate data needed to create this type of stratification must be readily available and closely connected to the source of sample variation in order to use the stratification method. Independence model is one instance of a potential simple model to be taken into consideration:

$$f(X|\lambda_1, \dots\dots\dots, \lambda_J) = \prod_{j=1}^{J} \lambda_j^{x_j} (1 - \lambda_j)^{1-x_j} \qquad (3)$$

The data may be considered missing if an appropriate stratification mechanism is absent. Taking into account the possibility that the population divide would produce K homogeneous strata, and the assumption that each of them will remain independent. Let $\boldsymbol{\pi} = (\pi_1, \dots\dots\dots, \pi_K)$

with $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ be the vector of strata probabilities. Then, the probability mass function of the capture vector is:

$$P(X|\lambda, \pi) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}^{x_j} (1 - \lambda_{jk})^{1-x_j} \tag{4}$$

Where $\boldsymbol{\lambda} = (\lambda_{jk})$ with $\lambda_{jk} \epsilon (0,1)$. This mixture model has independent models for each of its components, such as (3), using settings unique to each stratum. It accepts an augmented data representation as the two-step procedure, much like any other mixture models:

$$x_j|\mathcal{Z} \overset{indep}{\sim} Bernoulli\left(\lambda_{j\mathcal{Z}}\right) \;\; for\ j = 1, \dots \dots \dots . J$$

$$\mathcal{Z} \sim Discrete\left(\{1,2, \dots \dots . . , K\}, (\pi_1, \dots \dots . . , \pi_K)\right) \tag{5}$$

In this case, stratum assignment is expressly represented by the latent variable $\mathcal{Z}$. The Latent Class Model is the mixture of product-Bernoulli distributions in equation (4) [62]. In the space $\{0,1\}^J$, the mixture in (4) can reflect any feasible discrete distribution. However, we still need to address the model selection problem of selecting a suitable number of latent classes, K, before it can be applied.

In order to avoid predetermining the amount of mixture components and to enforce data-learned sparsity in the mixture, Dunson and Xing suggested a Bayesian nonparametric modification to the LCM [63]. Using an infinite number of latent classes simultaneously in conjunction with a prior specification that introduces sparsity into the mixture by concentrating the majority of the probability mass into a small finite subset was suggested as an alternative to attempting to identify the “best” finite number of latent classes. The resultant model, which is an infinite-dimensional mixture of product-multinomial distributions, overcomes the model selection issue of needing to choose the proper number of latent classes, K, while retaining the expressiveness and simplicity of the original LCM. In addition, it serves as a tool for model averaging, propagating the dimensionality uncertainty of the model into estimates.

An example of a Dirichlet process mixing of product-Bernoulli distributions is the nonparametric LCM found in Dunson and Xing. The hierarchical generating method below helps to explain it:

$$x_j|\mathcal{Z} \overset{indep}{\sim} Bernoulli\left(\lambda_{j\mathcal{Z}}\right) \;\; for\, j = 1, \dots \dots \dots .J$$

$$\mathcal{Z} \sim Discrete\left(\{1,2, \dots \dots . .\}, (\pi_1, \pi_2, \dots \dots)\right)$$

$$\lambda_{jk} \overset{indep}{\sim} Beta\;(1,1) for\, j = 1, \dots . . J\; and\; k = 1,2, \dots \dots$$

$$(\pi_1, \pi_2, \dots \dots) \sim SB(\alpha)$$

$$\alpha \sim Gamma\;(a, b), \qquad (6)$$

where the stick-breaking procedure with parameter $\alpha > 0$ is denoted by $SB(\alpha)$ [64]. By focusing the majority of the probability mass onto the first few coordinates of the stick-breaking process for $\boldsymbol{\pi}$, sparsity is introduced into the mixture and overfitting is prevented. We may get the effective dimensionality of the mixture from the data by using the prior distribution on α [65]. The finite-dimensional stick-breaking prior is defined here, $(\pi_1, \dots \dots . ., \pi_{K^*}) \sim SB_{K^*}(\alpha)$, by making $\pi_k = V_k \prod_{h<k}(1 - V_h)$ for $V_{k^*} = 1$ and $V_1, \dots . ., V_{K^*-1} \overset{indep}{\sim} Beta\;(1, \alpha)$ where K^* is the upper bound on the number of classes for a finite-dimensional approximation that is large enough [66].

Going back to the initial problem of estimating the unknown size of a closed population, we plug the LCM probability mass function from EquationE (4) into the general multiple-recapture multinomial model shown in (2) to obtain a joint model for the observable sample (those units with capture patterns different from zero),

$$P(X|\lambda, \pi, N) \propto \binom{N}{n}\left[\sum_{k=1}^{K^*} \boldsymbol{\pi_k} \prod_{j=1}^{J} \lambda_{jk}^{x_j}(1 - \lambda_{jk})\right]^{N-n}$$

$$\times \prod_{i=1}^{n} \sum_{k=1}^{K^*} \boldsymbol{\pi_k} \prod_{j=1}^{J} \lambda_{jk}^{x_{ij}} (1-\lambda_{jk})^{1-x_{ij}} \qquad (7)$$

We can see that (7) is comparable to a marginalized form of the enriched data representation by using the latent variable from (5):

$$P(\mathrm{X}, \mathrm{Z}, \mathrm{Z}^0 | \lambda, \pi, N) \propto \binom{N}{n} \prod_{i=1}^{n_0} \pi_{\mathcal{Z}_i^0} \prod_{j=1}^{J} (1-\lambda_{j\mathcal{Z}_i^0})$$

$$\times \prod_{i=1}^{n} \pi_{\mathcal{Z}_i} \prod_{j=1}^{J} \lambda_{jk}^{x_{ij}} (1-\lambda_{jk})^{1-x_{ij}} \qquad (8)$$

Where $\mathrm{Z} = (\mathcal{Z}_1, \dots \dots .., \mathcal{Z}_n)$ and $\mathrm{Z}^0 = (\mathcal{Z}_1^0, \dots \dots .., \mathcal{Z}_{n_o}^0)$, and both $\mathcal{Z}_i$ and $\mathcal{Z}_i^0$ take values on the set $\{1, \dots \dots \dots, K^*\}$ for each $i = 1, \dots \dots .., n$. By selecting prior distributions for parameters π and N, we finalize a comprehensive Bayesian specification. Given the nature of latent variables, the enhanced data representation in (8) easily leads to Gibbs sampling techniques based on MCMC algorithms that take use of conditional independence.

Nevertheless, there are more challenges because the vector Z^0 length is precisely equal to $n_0 = N - n$. It is not feasible to design acceptable Gibbs sampling methods by only deriving entire conditional distributions for N and each $\mathcal{Z}_i^0$, since this would lead to a reducible Markov chain, since N is itself a parameter to estimate. Basu and Ebrahim highlighted this difficulty [67], who suggested using a conditional decomposition to sample N and the latent variables (Z^0 in this example) simultaneously in order to overcome it. Fienberg also took advantage of this concept [68], and Manrique-Vallier and Fienberg[69] using multiple-recapture and having been modified and expanded upon by Manrique-Valier and Reiter [70] as a standard procedure for sampling from the NPLCM while adhering to intricate structural zero constraints.

Theologically, the overall method that Manrique-Valier and Reiter suggested [70] is directly applicable to this problem. Still, more simplifications are possible due to the unique structure of the multiple-recapture issue, in which only one cell is unobservable. Let $\boldsymbol{\omega} = (\omega_1, \dots \dots .., \omega_{K^*})$ with $\omega_k = \sum_{i=1}^{n_0} I(\mathcal{Z}_i^0 = \kappa)$. Here ω_k denotes the number of unobserved individuals that belong to latent K. Then, we get the representation:

$$P(\mathcal{X}, \mathrm{Z}, \omega | \lambda, \pi, N) = \binom{N}{n, \omega_1, \ldots\ldots.., \omega_{K^*}} \prod_{k=1}^{K^*} \left(\pi_k \prod_{j=1}^{J} (1 - \lambda_{jk}) \right)^{\omega_k}$$

$$\times \prod_{i=1}^{n} \pi_{\mathcal{Z}_i} \prod_{j=1}^{J} \lambda_{j\mathcal{Z}_i}^{\mathcal{X}_{ij}} (1 - \lambda_{j\mathcal{Z}_i})^{1-\mathcal{X}_{ij}}$$

$$\times I \left(\sum_{k=1}^{K^*} \omega_k = N - n \right) \qquad (9)$$

Here (8) is comparable to (7) after marginalizing over **Z** and **ω.**

The structure of a gibbs sampler algorithm, which extracts samples from the posterior distribution of model (9), including N, the population size, is shown below. Manrique-Vallier and Reiter's first steps are comparable when using the previous distribution suggested in (8) [70]:

(i) Sample from $P(Z| \ldots..): For\ i = 1, \ldots..n.$

Sample $\mathcal{Z}_i \sim Discrete\ (\{1, \ldots.., K^*\}, (p_1, \ldots\ldots, p_{K^*})),$

with $p_k \propto \pi_k \prod_{j=1}^{J} \lambda_{jk}^{x_{ij}} \left(1 - \lambda_{jk}\right)^{1-x_{ij}}$

(ii) Sample from $P(\lambda| \ldots..): For\ j = 1, \ldots..J.$ and $k = 1, \ldots..K^*.$

Let $n_k = \sum_{i=1}^{n} I(\mathcal{Z}_i = k)$ and $n_{jk} = \sum_{i=1}^{n} I\left(x_{ij} = 1,\ \mathcal{Z}_i = k\ \right).$

Then, sample $\lambda_{jk} \sim Beta(n_{jk} + 1, n_k - n_{jk} + \omega_k + 1).$

(iii) Sample from $P(\pi| \ldots..): For\ k = 1, \ldots..K^* - 1.$

Sample

$$V_k \sim Beta\left(1 + v_k, \alpha + \sum_{h=k+1}^{K^*} v_h\right)$$

where $v_k = n_k + \omega_k$. Let $V_{K^*} = 1$ and make $\pi_k = V_k \prod_{h<k}(1 - V_h)$ for all $k = 1, \ldots \ldots, K^*$

(iv) Sample from $p(\alpha| \ldots.)$: $\alpha \sim Gamma(a - 1 + K^*, b - \log \pi_{K^*})$

(v) Sample from $p(N, \omega|\ldots.)$: The full joint conditional distribution of ω and N is:

$$P(\omega, N|\lambda, \mathrm{Z}, \alpha, \pi, \chi) \propto P(N) \frac{n_0!}{\omega_1!, \ldots \ldots, \omega_{K^*}!} \rho_1^{\omega_1} \ldots \ldots \ldots . \rho_{K^*}^{\omega_{K^*}}$$

$$\times I(N = n + n_0)$$

where $n_0 = \sum_{k=1}^{K^*} \omega_k$ and $\rho_k = \pi_k \prod_{j=1}^{J}(1 - \lambda_{jk})$. For $P(N) \propto \frac{1}{N}$, Since N is entirely defined by $\boldsymbol{\omega}$, this is a negative multinomial distribution that is not conditional on N. Therefore, by compounding a negative binomial with a multinomial distribution, we are able to collect samples from this distribution [71]:

(i) Sample $n_0 = NegBinomial\left(n, 1 - \sum_{k=1}^{K^*} \pi_k \prod_{j=1}^{J}(1 - \lambda_{jk})\right)$. Make $\mathrm{N} = \mathrm{n} + n_0$

(ii) Sample $(\omega_1, \ldots \ldots, \omega_{K^*}) \sim Multinomial\left(n_0(p_1, \ldots \ldots \ldots, p_{K^*})\right)$ for $p_k \propto \rho_k$

The above proposed Gibbs sampler algorithm consist only of sampling steps from standard distributions. Additionally, it only requires sampling $K^* \times (J + 2) + n + 1$ variates per iteration.

The latent-class model for capture-recapture (LCMCR) in R-4.0.5 for Windows was used for all analyses [72, 73]. The LCMCR approach is built on a Dirichlet process mixture, which allows it to transparently adjust its complexity without the need for a separate model selection phase and tolerate complicated patterns of heterogeneity of captures [58, 59]. For a Dirichlet process parameter, uninformative priors—those that dominate the likelihood function and have little bearing on the inference—were provided (α~Gamma (0.25, 0.25)). We used K=5 latent classes; 10,000 samples from the posterior distribution drawn with a burn-in of 10,000 iterations and a thinning interval of 1,000 iterations to specify the Markov Chain Monte Carlo (MCMC) sampling. Trace plots and the posterior probability distribution histogram for population size were used to evaluate the MCMC sampling's convergence.

For three-source capture-recapture, median population size estimates with 95% credible sets were generated both overall and per province. The highest density intervals (the HD Interval package in R) were provided to make it easier to interpret findings and apply estimates for programs.

3.6.2. **FSW Population Size Estimation using Three-source capture recapture method**: Bayesian Model Averaging

To overcome sample heterogeneity, the Bayesian nonparametric latent class (NPLC) approach divides the population into relatively homogeneous classes where simple models are expected to hold better. These models are then applied to each stratum individually to generate estimates of the population size. As a result, there are more models that may be fitted to the data due to the huge dimensionality, which makes the challenge of model selection even more urgent. There is a solution to this issue with Bayesian model averaging.

Based on a closed, finite population of N persons, we discuss the missing data problem here, where each individual might be listed or overlooked by any of the J lists that partly enumerate that population. In this case, the existence of a person's name on the particular list is deemed a capture. The information can be shown as a contingency table with one dimension for each list if it is feasible to identify the subjects or their capture histories individually. It is necessary to estimate the number of persons in the cell where none of the lists were able to locate them.

Before deriving the model, we introduce Bayesian graphical models. Consider there are three different variables representing presence or absence in three different lists, we denote cell probabilities and data counts by θ_{ijk} and D_{ijk}, where i indexes the first list, respectively. The indices take on values 0, indicating absence from a list; 1, indicating presence in a list. Each variable in a graphical model is shown as a node in the graph, and linkages between variables show the clear interdependence that exist between them [74]. If S, C_1 and C_2 represent sets of variables, and if S separates C_1 and C_2in the graph, i.e., all paths connecting both sets pass through S, then $C_1$is independent of C_2 given S.

A prior distribution for θ ought to be focused on values that meet the model's independence connections; a class of such priors known as "hyper-Dirichlet" distributions was presented. [75].

Hyper-Dirichlet can be constructed by moving through the graph according to a perfect ordering, $(C_1, C_2, \ldots\ldots, C_I)$, of the cliques, and placing a Dirichlet marginal distribution on θ_{C_I}, for each clique C_i in turn, subjects to the constraint that each marginal distribution must cohere with what has been specified for previous cliques [76].

We begin by enumerating a class of possible models, indexed by $\mathcal{M} = \{1,2, \ldots\ldots, k\}$, for the cell probabilities of the contingency table. We also specify a prior distribution, $pr(\mathcal{M})$, over $\mathcal{M}$. The prior $pr(\mathcal{M})$ should ideally be based on expert opinion. The cell probabilities for each model $m \in \mathcal{M}$ are parametrized by some vector $\theta(m)$; priors for these parameters are described below.

Let N be the total population size; we assume that N is independent of $\mathcal{M}$and $\theta(\mathcal{M})$ a priori. A typical choice for the prior of N when no information about it is available is the Jeffreys prior,

$$pr(N) \propto \frac{1}{N} \qquad (10)$$

An alternative is a noninformative prior for the integers proposed by Rissanen [77],

$$pr(N) \propto 2^{log^*(N)} \qquad (11)$$

Where $log^*(N)$ is the sum of the positive terms in $\{\log_2(N), \log_2\{\log_2(N),\}, \ldots\ldots\ldots\ \}$.

Recalling that N is independent of θ and $\mathcal{MM}$ a priori, using Bayes Theorem to arrive at the posterior distribution of $(N, \theta(\mathcal{MM}))$ conditional on the model M, and, integrating out $\theta(m)$, given the data D we obtain:

$$pr(N|D, \mathcal{M} = m) = \frac{pr(D|N, \mathcal{M} = m)\, pr(N)}{pr(D|\mathcal{M}=m)}\ , \qquad (12)$$

where,

$$pr(D|N, \mathcal{M} = m) = \int_{\theta\,(m)}^{\cdot} pr\{D|N, \theta(m), \mathcal{M} = m\}\, pr\{\theta(m)|\mathcal{M} = m\}\, d\theta\,(m)$$

$$= \binom{N}{D} \frac{\psi_m\{\alpha(m) + D^*\}}{\psi_m\{\alpha(m)\}}$$

$$(13)$$

Here D^*indicate the whole set of data including the unobserved cell's count. Recall that the unobserved count is determined by N and D. In case a model m has cliques $(C_1, \dots\dots, C_I)$ and separators $(S_2, \dots\dots, S_I)$, indicate by $\alpha_{C_I}(\alpha_{C_I}: C_I: \in \mathcal{C}_I)$ the Dirichlet distribution's parameters on θ_{C_I}, and in a similar manner specify α_{S_I}, where the sets of feasible configurations for $\mathcal{C}_I$ and $\mathcal{S}_I$, respectively, are denoted by C_I and S_I. Next, the function $\psi(\bullet)$ is described as follows:

$$\psi(\alpha) = \frac{\left\{\prod_{i=1}^{I} \prod_{C_I \in \mathcal{C}_I} \Gamma(\alpha_{C_I})\right\}}{\left\{\Gamma(\sum_{C_I \in \mathcal{C}_I} \alpha_{C_I}) \prod_{j=2}^{I} \prod_{S_I \in \mathcal{S}_I} \Gamma(\alpha_{S_I})\right\}} \qquad (14)$$

wherew, Γ stands for the gamma function.

The posterior for N is given by expression (3), conditional on a certain $\mathcal{MM} = m$. We construct an unconditional posterior distribution by averaging models in order to account for model uncertainty:

$$pr(N|D) = \frac{\sum_m pr(D|N, \mathcal{M} = m)\, pr(\mathcal{M}=m)}{pr(D)} \qquad (15)$$

where,

$$pr(D) = \sum_m \sum_N pr(D|\mathcal{M} = m, N)\, pr(N)\, pr(\mathcal{M} = m)$$

Participant-level data was imported into RStudio for analysis (R package: shinyrecap), and cleaning was performed based on pre-determined exclusion criteria (self-reporting to be a FSW and accepting the offered UO) and data logical flow following skip patterns. The dataset was subset by province to have provincial-level FSW population size estimates. For every subgroup, aggregated datasets with counts of every capture-recapture combination were created. *Table 2* below shows how data were aggregated by overall and provincial $2^k - 1$ contingency tables for analysis preparation, where k stands for the number of capture occasions and aggregated counts, n_i where i stands for specific capture occasion.

Table 2: Three-source capture-recapture aggregated dataset, Rwanda 2022.

Capture 1:	Capture 2:	Capture 3:	Total:
1	0	0	n_1
0	1	0	n_2
0	0	1	n_3
1	1	0	$n_{1\&2}$
1	0	1	$n_{1\&3}$
0	1	1	$n_{2\&3}$
1	1	1	$n_{1\&2\&3}$

The final PSE with credibility sets was created from aggregated data sets using Bayesian model averaging [78], which is adaptable and capable of handling different types of variability in capture probability. For 3S-CRC data, confidence intervals and the median population size with 95% credibility sets were generated for both the overall and province-specific.

3.6.3. FSW Population Size Estimation using Privatize Network Sampling method.

For inference on disadvantaged and difficult-to-reach groups, link-tracing approaches like respondent-driven sampling (RDS) are often utilized. An extension of RDS that allows for additional inferential processes is privatized network sampling (PNS), in which the identities of each subject's connections are gathered in a way that protects their privacy. For this research, we derive and implement two PNS population size estimators.

Let the degree of individual i in the population of interest be denoted by d_i for $i \in \{1, \ldots\ldots N\}$, where N is the population size. For the duration of the study, we assume that the population's size and degrees will remain constant. The entire degree of a set v is represented as $d_v =$

$\sum_{i \epsilon v} d_i$, the mean degree of a set represented by $\bar{d}_{v=} \frac{d_v}{|v|}$, and the population mean degree expressed as $\bar{d}$. First, the seeds are chosen, and the degree of the node is taken into consideration at each phase of the PNS sampling process. Following the sampling of a network node, the identities of all the alters or a randomly selected selection of them are gathered. Subsequently, a recruiter node is chosen from the sampled nodes. Next, a random recruit is chosen from among their alters that haven't been sampled yet.

Let t represent the total number of seeds and $S_i \in \{1, \ldots\ldots N\}$ the index of the ith recruited subject, with realization S_i and $R_i \in \{-1,1, \ldots\ldots N\}$ the recruiter with realization r_i, where $r_i = -1$ in the case of seeds. Up to and including the ith subject, we represent the set of sampled subjects and their recruiters using the notation $S_{\leq i}$ and $R_{\leq i}$ respectively. An individual seed is a tree's beginning. We refer to s_i^c as the ith sampled subject within the cth tree, and $s_i^{\backslash c}$as the ith sampled subject in all other trees omitting the cth one. $G_i \in \{1, \ldots\ldots t\}$ is the tree of the ith sampled subject, and g_i is the realization. In addition, let n represent the number of sampled subjects, n^c be the size of the cth tree and $n^{\backslash c} = n - n^c$. We omit the subscript index when referring to the full sample to make notation easier ($S = S_{\leq n}$ and $S^c = S^c_{\leq n^c}$).

With realization o_{s_i}, the neighbors reported by the ith sampled subject, apart from their recruiter and recruits, are represented as O_{S_i}. In a case indexed by a set, O is the multiset union over the elements of the set $O_S = \uplus_{x \epsilon S} O_x$. The alters reported by subjects in tree c are presented as $O^c = O_{S^c}$ for simplicity of notation, whereas those from other trees are represented as $O^{\backslash c} = O_{S^{\backslash c}}$. The size of any multiset x is $|x|$ and the ith element after ordering the members of a multiset is denoted as $x[i]$. Summations over all elements of a multiset are denoted as $\sum_{k \epsilon x}$.

The total number of cross tree matches between nominated alters and sampled subjects is defined to be

$$M = \sum_{i=1}^{n} q\left(S_i, O^{\backslash G_i}\right),$$

Where $q\left(S_i, O^{\backslash G_i}\right)$ is the number of times the ith sampled individual is nominated by sampled individuals in different trees.

$$q\left(S_i, O^{\backslash G_i}\right) = \sum_{k \epsilon O^{\backslash G_i}} I(S_i = k)$$

And I is the indicator function. Assuming that, conditional upon the recruitment trees observed thus far and the total number of nominated connections in each tree, free edge ends in the network are connected completely at random, we have that:

$$E(q\left(S_i, O^{\backslash G_i}\right)|S_{\leq i}, R_{\leq i}, |O^1|, \dots., |O^t|) =$$

$$\frac{d_{s_i} - I(R_{s_i} \neq -1)}{N\bar{d} - 2\left(i - \sum_{j=1}^{|R_{\leq i}|} I(R_{\leq i}[j] = -1)\right)} \left|O^{\backslash G_i}\right|$$

The numerator is the number of free edge ends incident upon node S_i, potentially excluding one edge end observed connecting the individual to their recruiter. The denominator is the total number of free edge ends in the graph, excluding those observed in the recruitment graph. $\left|O^{\backslash G_i}\right|$ is the number of nominated individuals in the other trees that could potentially lead to the sampled node.

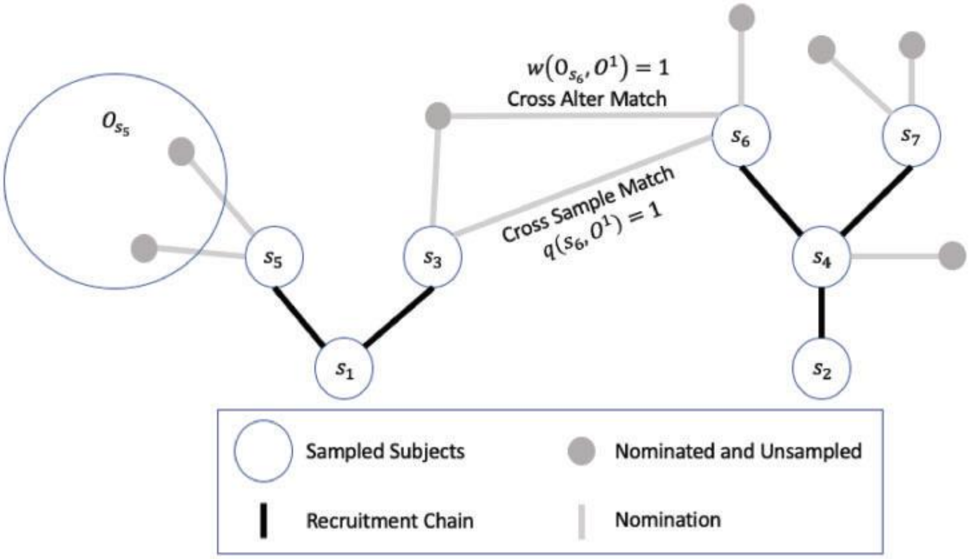


Figure 1: An example PNS recruitment Graph with Two Trees, one of Size 3 and the other of Size 4

Gray points indicate connections reported by sampled subjects who were not themselves sampled. Cross sample matches are reported connection between individuals in different trees. Cross Alter matches are shared connections between individuals in different trees who are not sampled.

We may now write the expected number of matches as:

$$E(M) = E\left(\sum_{i=1}^{n} \frac{d_{s_i} - I(R_{s_i} \neq -1)}{N\bar{d} - 2\left(i - \sum_{j=1}^{|R_{\leq i}|} I(R_{\leq i}[j] = -1)\right)} \left|O^{\setminus G_i}\right|\right)$$

$$\approx E\left(\sum_{i=1}^{n} \frac{d_{s_i} - I(R_{s_i} \neq -1)}{N\bar{d} - (n-t)} \left|O^{\setminus G_i}\right|\right)$$

$$= E\left(\sum_{c=1}^{t} \frac{\sum_{i=1}^{n^c} d_{s_i^c} - I(R_{s_i^c} \neq -1)}{N\bar{d} - (n-t)} \left|O^{\setminus c}\right|\right)$$

$$= E\left(\frac{\sum_{c=1}^{t}\left(\bar{d}_{S^c} - \frac{n^c - 1}{n^c}\right) n^c \left|O^{\setminus c}\right|}{N\bar{d} - n + t)}\right) \qquad (16)$$

Considering the classical form of population size estimation using capture-recapture, taking two independent random samples *A* and *B* from a population, the expected rate that individuals from sample *A* are seen in sample *B* is, on average, proportional to the size of sample *A* relative to the population size and so the expected rate of overlap is,

$$E(|A \cap B|) = \frac{|A||B|}{N}$$

Rearranging the equation, yields the classical Lincoln-Petersen estimator[42].

$$N = \frac{|A||B|}{(|A \cap B|)} \approx \frac{|A||B|}{|a \cap b|} \qquad (17)$$

Where a and b are the observed samples from A and B so that $|a \cap b|$ is the observed overlap. We now create an estimator by replacing the random variables M, S and O with their observed values m (observed number of matches), s (the sample), and o (nominated alters). Additionally, we use the Gile's successive sampling estimator [79].

$$\hat{\bar{d}} = \frac{\sum_{i \epsilon s} d_i w_i}{\sum_{i \epsilon s} w_i}$$

where w_i are the successive sampling weights, in place of the unobserved $\bar{d}$. We note that the successive sampling estimator is a function of the total population size because the successive sampling estimator uses population size to construct the weights.

The resulting estimating equation - **the Cross Sample population size estimator $\boldsymbol{\hat{N}_{cs}}$** is,

$$0 = \frac{\sum_{c=1}^{t}\left(\bar{d}_{S^c} - \frac{n^c - 1}{n^c}\right) n^c |O^{\backslash c}|}{N\bar{d} - n + t - m} \qquad (18)$$

One challenge is that when privatizing hashes are used, then m in (18) is not observed, only the hashed values from the nominated neighbors are observed. Suppose that each node is assigned a hashed identifier h_i such that the probability that two random nodes have the same identifier is ρ . Conditional upon the recruitment up to the ith node, and assuming that edge ends are connected randomly, the expected total number of edge ends incident upon other nodes with the same identifier is,

$$\rho\left(\bar{d}(N-1) - 2\left(i - \sum_{j=1}^{|R_{\leq i}|} I(R_{\leq i}[j] = -1)\right)\right)$$

$$\approx \rho(\bar{d}(N-1) - n + t)$$

Building on the One-step Estimator of Networked population size [80], the number of free edge ends for node S_i is $d_{s_i} - I(R_i \neq -1)$ and so the probability that a match is actually a true match is

$$P(O^c[k] = S_i|h(O^c[k]) = h(S_i), R_i, S_i) \approx \left(\frac{\rho(\bar{d}(N-1)-n+t)}{d_{s_i}-I(R_i \neq -1)}\right)^{-1}$$

And so, the expected total number of matches can be related to the number of hashed matches as

$$E(M) \approx E\left(\sum_{i=1}^{n} \sum_{k \epsilon O^{\setminus G_i}} P(O^c[k] = S_i|h(O^c[k]) = h(S_i), R_i, S_i)\, I(h(S_i) = h((k))\right)$$

We may then replace the m in (18) with a sample estimate adjusting for potential random clashes of the hash function,

$$\hat{m}(N) = \sum_{i=1}^{n} \sum_{k \epsilon o^{\setminus g_i}} \left(\frac{\rho(\hat{\bar{d}}(N-1)-n+t)}{d_{s_i^c} - I(r_i \neq -1)} + 1\right)^{-1} I\big(h(s_i) = h((k)\big)$$

With the Cross Sample estimator, the number of potential matches incident on each tree is equal to the product of the number of nominated individuals in the other trees and the sample size of the tree. In a well-connected population where individuals know many other members of the population, we expect the number of nominated individuals to greatly exceed the number of sampled individuals. Consequently, we anticipate the number of potential matches between the nominated alters from a tree and the nominated alters from other trees to be large and thus potentially a better target for inference. The total number of cross tree matches between nominated alter sets is defined as:

$$U = \sum_{i=1}^{n} W\left(O_{s_i}, O^{\backslash G_i}\right)$$

With realization u , where $W\left(O_{s_i}, O^{\backslash G_i}\right)$ is the number of times nominated alters of i$^{\text{th}}$ sampled individual are nominated by sampled individuals in the different trees.

$$W\left(O_{s_i}, O^{\backslash G_i}\right) = \sum_{j \epsilon O_{s_i}} \sum_{k \epsilon O^{\backslash G_i}} I(j = k)$$

The total number of free edge ends incident on nominated alters of individual s_i is $\left|O_{s_i}\right|\left(\bar{d}_{O_{s_i}} - 1\right)$, excluding the edges connecting the alters to individual s_i. We again assume that edges ends are conditionally connected completely at random to calculate the conditional expectation,

$$E\left(W\left(O_{s_i}, O^{\backslash G_i}\right) \middle| S_{\leq i}, R_{\leq i}, |O^1|, \ldots\ldots, |O^t|\right) = \frac{\left|O_{s_i}\right|\left(\bar{d}_{O_{s_i}} - 1\right)}{N\bar{d} - 2\left(i - \sum_{j=1}^{|R_{\leq i}|} I(R_{\leq i}[j] = -1)\right)} \left|O^{\backslash G_i}\right|$$

The expected number of matches is then,

$$E(A) = E\left(\sum_{i=1}^{n} \frac{\left|O_{s_i}\right|\left(\bar{d}_{O_{s_i}} - 1\right)}{N\bar{d} - 2\left(i - \sum_{j=1}^{|R_{\leq i}|} I(R_{\leq i}[j] = -1)\right)} \left|O^{\backslash G_i}\right|\right)$$

$$\approx E\left(\frac{\sum_{i=1}^{n}\left(\bar{d}_{O_{s_i}} - 1\right)\left|O_{s_i}\right|\left|O^{\backslash G_i}\right|}{N\bar{d} - n + t}\right)$$

$$= E\left(\frac{\sum_{c=1}^{t}\left|O^{\backslash c}\right|\left|O^{c}\right| \sum_{i \epsilon S^c} \frac{\left|O^{i}\right|}{\left|O^{c}\right|}\left(\bar{d}_{O_i} - 1\right)}{N\bar{d} - n + t}\right)$$

$$\approx E\left(\frac{(\tilde{d} - 1)\sum_{c=1}^{t}\left|O^{\backslash c}\right|\left|O^{c}\right|}{N\bar{d} - n + t}\right) \qquad (19)$$

where $\tilde{d} = \frac{\sum_{i=1}^{N} d_i^2}{\sum_{i=1}^{N} d_i}$ is the mean degree when degrees are sampled with probability proportional to degree. The approximation used in the second equation mirrors the approximation used in the Cross Sample estimator (16). The approximation in the fourth equation is obtained by noting that the alter degrees ($d_{O_{s_i^c}}$) are selected with probability proportional to degree, since edge ends are assumed to be connected completely at random. Hence, the expected value of $\bar{d}_{O_i}$ is $\tilde{d} - 1$. Replacing random variables by their realizations and means by their estimators in equation (19) we arrive at the **Cross Alter estimating equation,**

$$0 = \frac{\left(\hat{\tilde{d}}-1\right)\sum_{c=1}^{t}|O^{\setminus c}||O^{c}|}{N\hat{\tilde{d}}-n+t} - u, \qquad (20)$$

with $\hat{\tilde{d}} = \frac{\sum_{i\epsilon s} d_i^2 w_i}{\sum_{i\epsilon s} d_i w_i}$**.** The N that solves (2020) is the Cross Alter estimate $(\widehat{\boldsymbol{N}}_{\boldsymbol{ca}})$

Similar to the adjustment to the Cross Sample estimator, the probability of a tie existing given a connection to an individual with the same hash is,

$$P\left(O^{c}[k] = O^{c}[l] | h\left(O^{\setminus c}[k]\right) = h(O^{c}[l])\right) \approx \left(\frac{\rho(\bar{d}(N-1)-n+t)}{d_{O^{c}[k]}-1} + 1\right)^{-1}$$

However, unlike the Cross Sample estimator, we do not observe the degrees in the denominator and so must take a different approach to the adjustment. Under the random connection assumption, the nominated alters are selected with probability proportional to degree. If an observed match of hashes $\left(h\left(O^{\setminus c}[k]\right) = h(O^{c}[k])\right)$ is false, meaning that $O^{\setminus c}[k] \neq O^{c}[k]$, then the degree distribution of this false match will match the degree distribution of $O^{c}[k]$ since the hashing function is independent of the degree distribution. If an observed hash match is a true match on the other hand, the match is selected from among the nominated alters with probability proportional to degree, meaning that its degree distribution is proportional to degree squared since the alter were themselves selected with probability proportional to degree.

The degree distribution of the hash matches is therefore a mixture distribution. Let ϕ be the estimated proportion of hash matches that are true matches with,

$$p_1 = \frac{1}{\sum_{i\epsilon s} d_i^2 \omega_i} \sum_{i\epsilon s} \left(\frac{\rho\left(\hat{\hat{d}}(N-1) - n + t\right)}{d_i - 1} + 1 \right)^{-1} d_i^2 \omega_i$$

and

$$p_2 = \frac{1}{\sum_{i\epsilon s} d_i \omega_i} \sum_{i\epsilon s} \left(\frac{\rho\left(\hat{\hat{d}}(N-1) - n + t\right)}{d_i - 1} + 1 \right)^{-1} d_i \omega_i$$

Being the estimated probabilities of a hash match being a true match marginalizing over the degree distributions of true and false matches, respectively. Since the degree distribution is mixed with mixing probability equal to the probability that a hash match is a true match, ϕ is defined by the relationship,

$$\phi = p_1\phi + (1 - \phi)p_2$$

Or more simply,

$$\phi = \frac{p_2}{1 - p_1 + p_2}$$

We can then multiply the number of hash matches by ϕ to get an estimate of the number of true matches to substitute in for u in (5)

$$\hat{u}(N) = \ \phi \sum_{i=1}^{n} \sum_{j\epsilon o_{s_i}} \sum_{k\epsilon o^{\backslash g_i}} I(h(j) = h(k))$$

The **Cross Sample estimator** makes use of the number of matches into the sample (m) and the **Cross Alter estimator** makes use of the matches to the nominated alters (u). The **Cross**

Network estimator makes use of both of these quantities by combining equation (18) and (20).

$$\frac{\sum_{c=1}^{t}\left(\left(\hat{\bar{d}}-1\right)|O^{c}|+\left(\bar{d}_{s^{c}}-\frac{n^{c}-1}{n^{c}}\right)n^{c}|O^{\setminus c}|\right)}{N\hat{\bar{d}}-n+t-u-m} \tag{21}$$

The solution to which is the **Cross Network estimator** $(\boldsymbol{\hat{N}_{cn}})$**.** $-\hat{u}\ (N)-\ \hat{m}\ (N)$ may be substituted in for $-\mathrm{u}-\mathrm{m}$ in the case of hashed identifiers.

A complete dataset for analysis was composed of key PNS variables, including subject ID (unique identifiers for individuals in the dataset), recruiter ID (unique identifier for the recruiter of the subject), subject hash ID (the privatized identifier for the subject), degree (the network size of the individual, this excludes contacts for whom the individual does not know their identifiers), and contact hash IDs (privatized identifiers for each contact of the subject).

Each site-level dataset was converted into an RDS coupon data frame, which means that each recruiter is aligned alongside the corresponding direct recruits. With this data frame, each seed is regarded as the base of a tree that branches out as its recruits recruit more people, and each tree is its own sample. Three estimators were considered, including *Cross Sample Estimator, Cross Alter Estimator,* and Cross *Network Estimator,* which combines the Cross-Sample and Cross-Alter Estimators. The intuition for PNS is "the rate at which subjects' networks contain other sampled subjects recruited by other seeds is related to population size."

The PNS method relies on three assumptions, including: connections between recruits and recruiters are completely random; small sample fractions (i.e., small sample fractions lead to potentially large amounts of sampling error when estimating population size); and long recruitment chains. However, these assumptions do not always hold, imposing a methodological limitation in that case. The *Cross-Network Estimator* displays reduced volatility compared to both the *Cross-Alter* and *Cross-Sample estimators and* showed little biased results considering different levels sample fractions and network sizes as described elsewhere[48]. Consequently, to assess estimator robustness, variance was the main parameter used.

All data analysis processes were performed using RStudio-2023.06.2-561 with the “RDS” and “*pnspop*" packages. Confidence intervals were calculated using the bootstrap process with the number of samples set to 10,000.

Site-level estimates were aggregated at the national level, then the pooled national estimates were distributed by province using proportions of FSW from existing program data. The analysis outputs include point estimates with corresponding 95% credible intervals. Finally, each estimate was adjusted for the proportion of FSW who did not have cell phones.

3.7. Ethical considerations

All the three surveys reviewed ethical approvals from the Rwanda National Ethics Committee (RNEC). The studies ensured maximum anonymity and confidentiality to guarantee that study participants are not victimized for participation. In this line, no names or identifying information were collected from any survey participant, participants were only identified using unique IDs. Furthermore, the study sites were secure places within a health facility setting that usually offers HIV services to key population communities to minimize possibility of stigma. Finally, paper-based study documents were maintained by the team leaders and stored in a designated locked cabinet during field work. Access to data was restricted and closely monitored, and all electronic data collection tools were password protected.

Participation in the studies was voluntarily and free to withdraw at any time during the conduct of the study. Neither refusal to participate nor withdrawal will affect services they would normally receive. In Rwanda, children under 18 years require parental consent prior to participation in the survey. The only exception is for “emancipated minors” who are children head of household; these children are not required to provide the consent of a caregiver but can instead consent directly. For FSWs under the age of consent who are NOT the head of household, we request a waiver of informed consent. A written informed consent was obtained from the study participants to be part of the survey. Furthermore, a waiver of informed consent for participants aged 15 to 17- years was granted by RNEC. Children <18 years of age identified as being engaged in sex work, trafficked, or victim of violence, received a special post HIV test counseling and were referred for appropriate services to ensure their protection and well-being.

Data collection staff completed training on human subjects’ research and signed a confidentiality agreement before the start of enrollment. Participants were compensated with for transportation costs and time and for each successful referral enrolled in the survey for RDS specific recruitments. Compensation for transport was determined based on the areas' transport cost as stipulated by Rwanda Utilities Regulatory Authority (RURA) and delivered in cash by the study site accountant.

Prior to implementation, field staff received a one-week standardized training together in one site, followed by a half-day refresher training at their respective sites. These trainings focused on general knowledge of key population, ethical issues in human subject research, and standard operating procedures for the survey implementation.

Chapter 4

Results

4.1. Introduction of the results section

Chapter 4 presents the main results from three studies implemented in Rwanda between 2021 and 2023 by applying the developed methods to this research project. Firstly, it provides the population size of men who have sex with men using the capture-recapture method with a Bayesian non-parametric latent class model. Next, it provides the population size of female sex workers using the developed generalized capture-recapture from Bayesian model averaging. Lastly, it presents the population size estimate of female sex workers when the network-traced capture recapture (Privatized Network Sampling - PNS) method is used.

4.2. MSM Population Size Estimation using Three-source capture recapture method: Bayesian nonparametric latent class model.

Capture one Results: Key holders with the unique design were distributed to the MSM through their corresponding associations, groups, and key informants. The following are the results from capture one. A total of 2,465 out of the 2,723 objects (90.5%) were successfully distributed (*Table 3*).

Table 3: Provincial-level objects distribution breakdown

PROVINCE	ACTUAL NUMBER OF OBJECTS ASSIGNED	ACTUAL NUMBER OF OBJECTS SUCCESSFULLY DISTRIBUTED	ACTUAL NUMBER OF OBJECTS NOT DISTRIBUTED SUCCESSFULLY AND RETURNED
EASTERN	636	558	78
CITY OF KIGALI	894	885	9
NORTHERN	166	150	16
SOUTHERN	585	515	70
WESTERN	442	357	85
TOTAL	**2,723**	**2,465**	**258**

Capture two Results: MSM friendly services were provided at 23 health facilities usually having key population service package country wide. Lubricants and Condoms were distributed during the capture two period, a total of 1,340 out of the anticipated 2,705 MSM (49.5%) came for health services at health facilities. Out of 1,340 MSM who came for the services, 1,314 met inclusion criteria and were offered the services (98.1%), and of them 721 (54.9%) were identified as having received the distributed unique object during the previous week. *Table 4* provides provincial level breakdown.

Table 4: Provincial-level service provision among MSM

PROVINCE	NUMBER OF HEALTH FACILITY	ANTICIPATED MSM TO BE OFFERED SERVICES	MSM RECEIVED AT HF	MSM RECEIVED OFFERED SERVICES	MSM RECEIVED DISTRIBUTED UNIQUE OBJECT DURING THE PREVIOUS WEEK
EASTERN	5	439	343	337	211
CITY OF KIGALI	5	803	510	497	185
NORTHERN	1	219	28	25	4
SOUTHERN	7	658	291	291	195
WESTERN	5	586	168	164	126
TOTAL	**23**	**2,705**	**1,340**	**1,314**	**721**

Capture three results: During capture three, RDS was used for participant recruitment. Every MSM recruited during RDS was screened for eligibility and considered as captured during capture three once he consents to participation. A total of 2,211 MSM were captured during this capture occasion. Among those captured during capture three, 422 (19.1%) were identified as having received distributed unique objects during capture one whereas 415 (18.8%) were identified as having received MSM friendly services during capture 2. *Table 5* provides provincial level break down for capture three results.

Table 5: Provincial-level breakdown for capture three results

PROVINCE	ANTICIPATED MSM TO BE CAPTURED DURING CAPTURE THREE	MSM CAPTURED DURING CAPTURE THREE	MSM RECEIVED DISTRIBUTED UNIQUE OBJECT	MSM RECEIVED PROVIDED MSM FRIENDLY SERVICES	MSM RECEIVED BOTH UNIQUE OBJECT AND MSM FRIENDLY SERVICES
EASTERN	121	126	50	64	36
CITY OF KIGALI	1,027	1,021	128	124	42
NORTHERN	308	303	20	18	4
SOUTHERN	141	152	42	54	28
WESTERN	613	609	182	155	100
TOTAL	**2,210**	**2,211**	**422**	**415**	**210**

Overall, we sampled 2,465, 1,314, and 2,211 MSM in captures one, two, and three, respectively. There were 721 recaptures between captures one and two, 415 recaptures between captures two and three, and 422 recaptures between captures one and three. There were 210 MSM captured in all three captures. The Venn-diagram in *Figure 2* below, provides all the three capture occasions and overlaps detailed results:

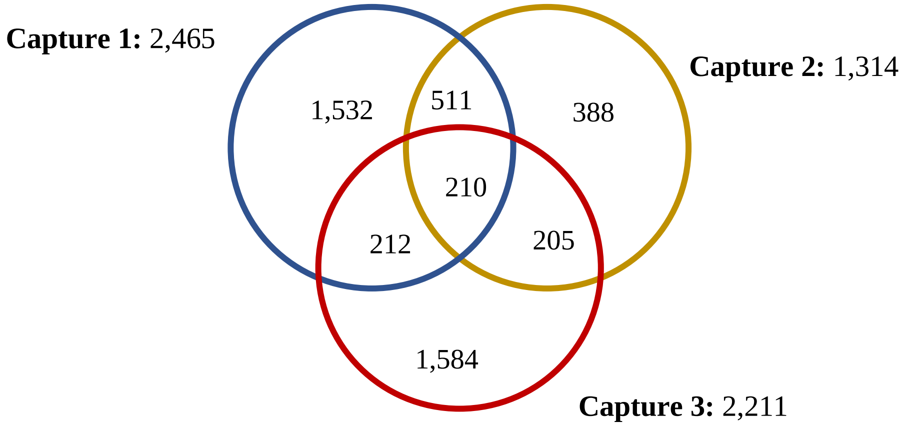


Figure 2: Venn-diagram representing individual capture results and overlaps between capture occasions.

Before conducting CRC analysis, we explored dependency between captures by testing for homophily in the RDS recruitment chain based on capture history variable, hence we found (Homophily = 1.016596) which shows a non-dependency.

The trace plot in *Figure 3* below, presents simulation results over 10,000 samples demonstrating the population size distribution. This indicates for a converging simulation result based on the random noise shape observed between the values 10,000 and 20,000 on the Y-axis.

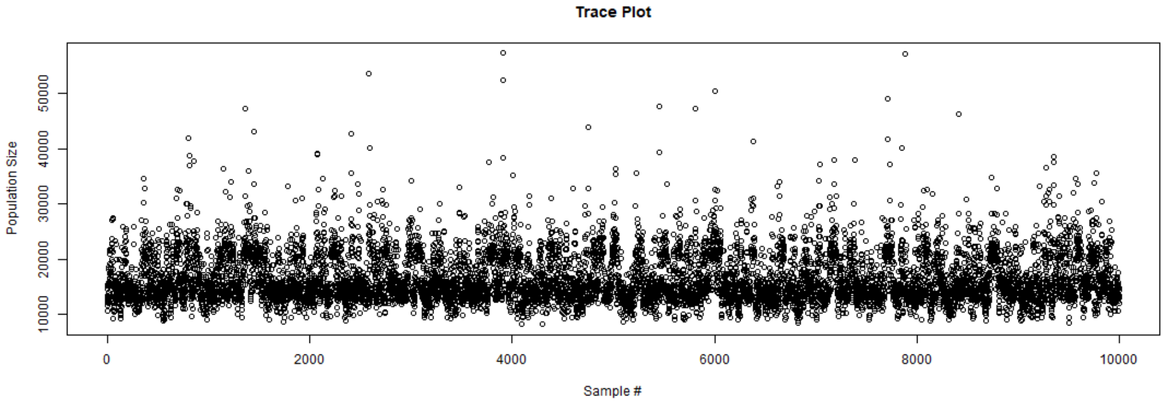


Figure 3: Trace plot showing the population size by sample number.

Figure 4 below, presents the posterior distribution (Population size distribution) provided by the model.

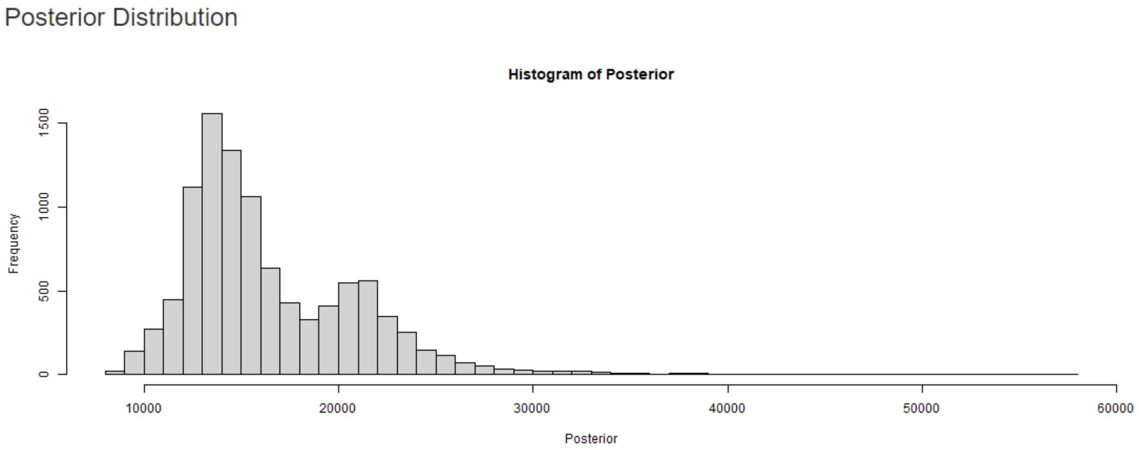


Figure 4: Histogram of posterior probability distribution for population size.

The final MSM population size estimates for the overall and the provincial levels are presented in the table below. The median from the posterior probability distribution is used as the point estimate and 95% probability intervals are used to describe uncertainty.

Table 6: Population size estimation of men who have sex with men, Rwanda 2022.

Province	Size estimate percent [a]	Median	95% credible sets [CS]
Eastern	0.3 [0.3 – 0.5]	2,287	[1,927 – 3,014]
City Of Kigali	2.7 [1.6 – 4.6]	7,842	[4,587 – 13,153]
Northern	0.5 [0.2 – 1.0]	2,375	[842 – 4,239]
Southern	0.4 [0.3 – 0.6]	2,109	[1,681 – 3,418]
Western	0.3 [0.3 – 0.5]	2,469	[1,994 – 3,518]
Overall	**0.7 [0.4 – 1.1]**	**18,100**	**[11,300 – 29,700]**

[a]*Population estimates of men were based on the 2012 Census data by National Institute of Statistics of Rwanda 2021 population size projections.*

We estimated the overall population of MSM in Rwanda to be 18,100 (95% CS: 11,300 - 29,700) where the majority lived in the City of Kigali 7,842 (95% CS: 4,587 – 13,153). The MSM population size estimates were almost similar for the remaining 4 provinces (Northern, Southern, Eastern, and Western provinces).

Considering the Bayesian model averaging process, that allows to flexibly account for list dependency by creating models for all possible dependencies, and averaging over them in a way that is proportional to the probability that the dependence is correct.

The first step in the analysis is to formulate a prior for population size. This is to represent the prior knowledge about population size along with uncertainty. By default, a "log-normal" prior is used. ***Figure 5*** below, presents the prior distribution as well as cumulative distribution.

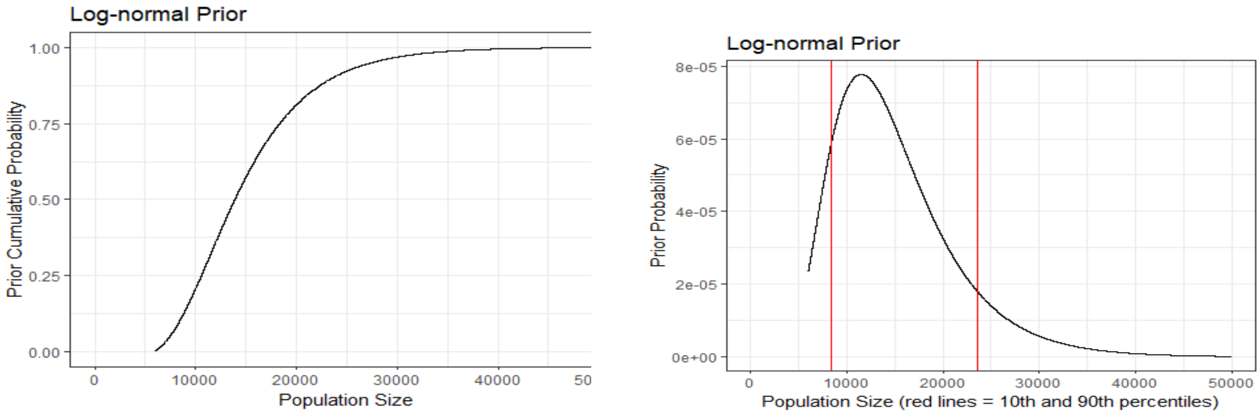


Figure 5: Prior distribution as well as cumulative distribution.

Once the prior is specified, the probability distribution of the population size is calculated. ***Figure 6*** below, presents the posterior population size estimates distribution, furthermore, it presents both individual model distribution as well as the average model distribution. In addition, ***Table 7*** describes the probability distribution for population size. Either the mean and/or median is used for point estimate. The (5% probability interval is used to summarize the estimate's uncertainty.

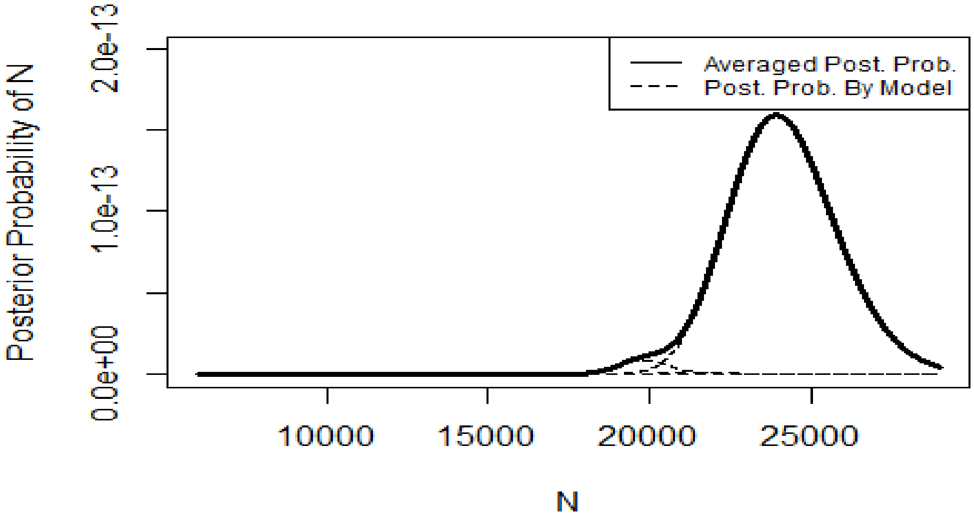


Figure 6: Posterior population size estimation distribution.

Table 7: BMA PSE distribution mean/median and uncertainty around estimates.

Mean	Median	95% Lower	95% Upper
21,188	20,787	19,347	22,268

4.3. FSW Population Size Estimation using Three-source capture recapture method: Bayesian Model Averaging

Of the 1,778 FSWs approached during Capture 1, 1,768 (99.4%) were newly captured, referring to not being captured elsewhere within the same week. Among those newly captured, unique object acceptance was high at 1,766 (99.9%). For 1,870 FSWs approached during Capture 2, 1,851 (98.9%) were newly captured within the second week of capture. Among those newly captured in Capture 2, UO acceptance was high at 1,848 (99.8%). During Capture 3, 1,910 FSWs were approached, and 1,867 (97.7%) were newly captured. The main reasons for unique object refusal documented were not being willing to receive the object and being willing to receive money instead of a unique object. *Table 7* below presents the results by capture round.

Table 8: Results of 3S-CRC by capture round, FSW PSE, Rwanda 2022

	Capture 1 **n (%)**	**Capture 2** **n (%)**	**Capture 3** **n (%)**
Approached FSWs	1,778	1,870	1,910
Already in current capture			
Yes	10 (0.6)	19 (1.1)	43 (2.3)
No	1,768 (99.4)	1,851 (98.9)	1,867 (97.7)
Unique object acceptance			
Accepted	1,766 (99.9)	1,848 (99.8)	1,865 (99.9)
Refused	2 (0.1)	3 (0.2)	2 (0.1)
Reason for refusal			
Doesn't want/Refused unique object	1	1	2
Wanted money, not objects	1	2	0

The majority of FSWs sampled were presumed to be 25 years old, while sexually exploited minors aged 15–17 had few captures across all three capture rounds. *Table 8* below describes sampled FSWs at each capture round by age and province.

Table 9: Sampled FSWs by capture round, Age Group and Province, FSW PSE, Rwanda 2022.

Capture	Age group	City of Kigali	Eastern Province	Northern Province	Southern Province	Western Province	Total
Capture 1							
	15 - 17	9	2	2	2	13	28
	18 - 24	122	109	162	106	129	628
	25+	127	142	222	247	372	1,110
							1,766
Capture 2							
	15 - 17	5	5	5	0	0	11
	18 - 24	116	130	206	126	329	911
	25+	239	164	162	124	237	926
							1,848
Capture 3							
	15 - 17	0	5	3	3	24	35
	18 - 24	184	155	85	92	335	851
	25+	131	152	276	142	278	979
							1,865

A total of 1,766 unique objects were distributed countrywide during Capture 1, 1,848 objects during Capture 2, and 1,865 objects during the third capture. In a three-week survey implementation exercise, 62 hotspots were visited countrywide in each capture round; however, bigger hotspots were resampled in the subsequent capture rounds. Two hotspots were resampled between Capture 1 and Capture 2; eight hotspots were resampled between Capture 2 and Capture 3; six hotspots were resampled between Capture 1 and Capture 3; and two hotspots were resampled in all three Capture rounds. Below are the maps for individual captures, highlighting venue and street hotspots visited (*Figure 5*).

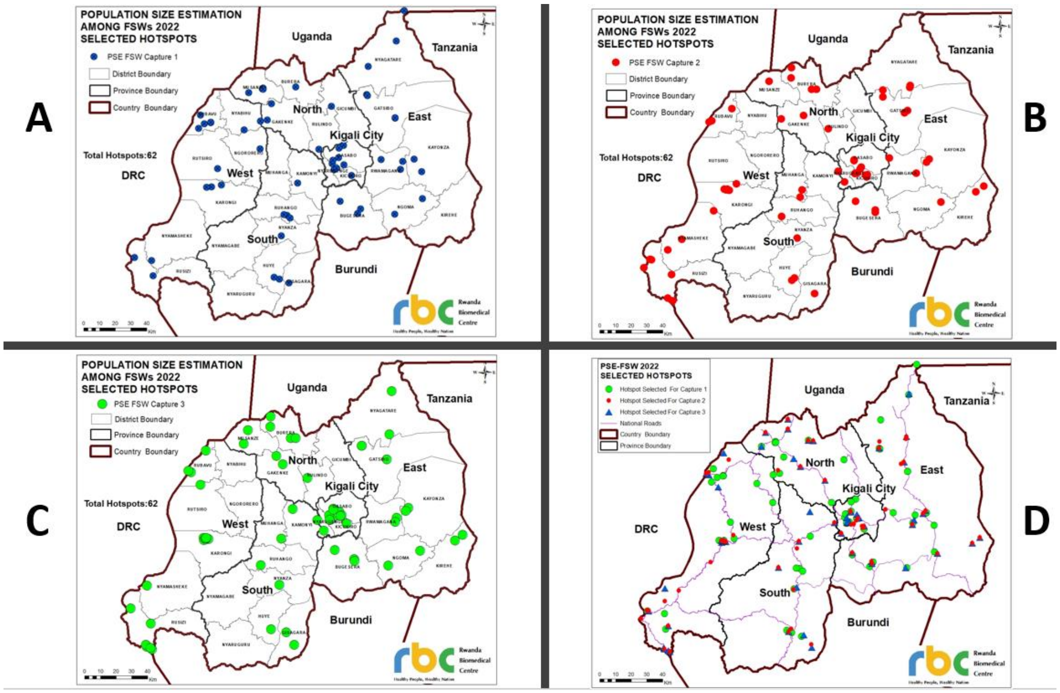


Figure 7: Maps of individual capture highlighting venue/street hotspots visited, Rwanda 2022.

The 3S-CRC aggregated, cleaned final dataset was imported into *shinyrecap* for analysis. For all three capture rounds, 1,766 FSWs, 1,848 FSWs, and 1,865 FSWs were sampled, where there were 1,408 FSWs, 1,471 FSWs, and 1,529 FSWs observed strictly during Capture 1, Capture 2, and Capture 3, respectively. There were 169 exclusive overlaps between Capture 1 and Capture 2, 210 exclusive overlaps between Capture 2 and Capture 3, and 65 recaptures between Capture 1 and Capture 3. Finally, 61 FSWs were recaptured in all three Capture rounds. *Figure 6* below presents the Venn diagram illustrating aggregated data of capture history results for single, double, and triple captures to construct a structured 3S-CRC dataset.

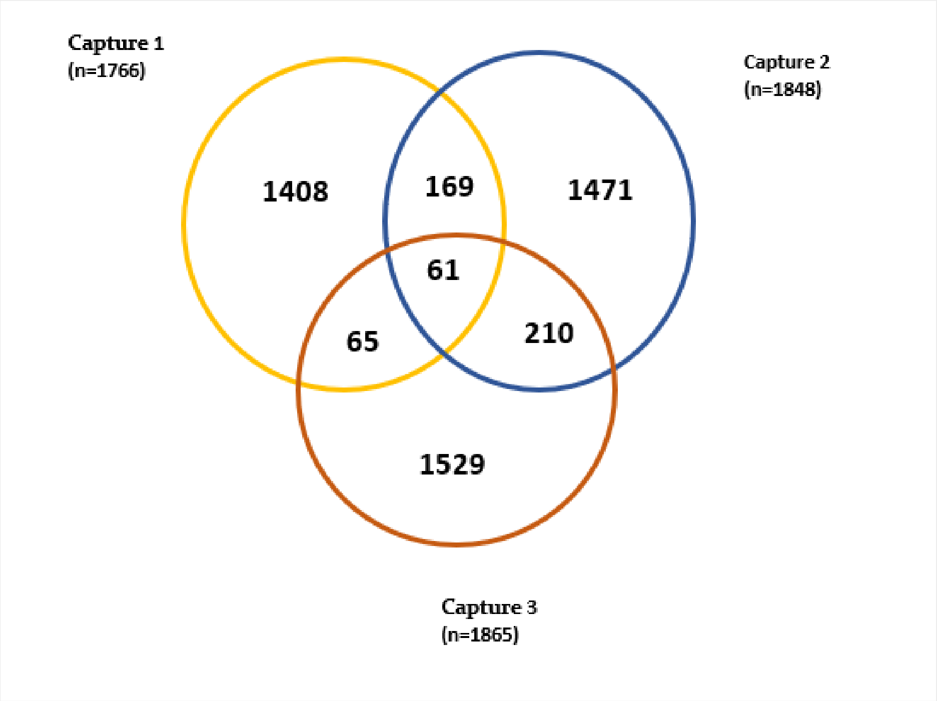


Figure 8: The Venn diagram presenting national aggregated data of capture history results for single, double, and triple captures, Rwanda 2022.

Out of 231 FSWs recaptured between Capture 1 and Capture 2, 96 physically presented UOs received during Capture 1, while out of 135 who did not have unique objects with them, 134 were able to correctly describe and identify the received unique object on a laminated card, bringing the total number of recaptures to 230. Of the 127 FSW recaptured between Capture 1 and Capture 3, 53 had provided unique objects with them, while of the 74 who did not have unique objects with them, 73 were able to correctly describe and identify the received unique object on a laminated card. Of the 272 FSW recaptured between Capture 2 and Capture 3, 111 had the received unique objects with them, while 160 who did not have unique objects with them were able to correctly describe and identify the received unique object on a laminated card. *Table 9* highlights the two methods used to record recapture histories.

Table 10: Recapture identification cascade, FSW PSE, Rwanda 2022.

	Re-capture Round	
	Capture 2	Capture 3
Capture 1 (C1)		
Total recaptured From C1	231	126
Showed C1 object	96	53
Did not have unique objects with them	135	74
Correctly identified C1 object	134	73
Capture 2 (C2)		
Total recaptured From C2	NA	271
Showed C2 object	NA	111
Did not have unique objects with them	NA	161
Correctly identified C2 object	NA	160

Only one FSW out of 135 FSWs who were claiming to have been offered capture one unique object during Capture 2 was unable to describe and correctly identify the object received on a laminated card. Out of 74 FSWs in Capture 3 who were claiming to have been offered a unique object but who did not have the objects with them, 73 were able to describe and correctly identify the object received on a luminated card. Only one FSW out of 272 FSWs who were claiming to have been offered Capture 2 unique object during Capture 3 was unable to describe and correctly identify the object received on a luminated card.

The FSW population size presented in *Table 10* below is based on three models: log-linear, Bayesian model averaging (using non-informative prior), and Bayesian nonparametric latent-class models.

Table 11: FSW National Population Size Estimates using 3S-CRC method, FSW PSE, Rwanda 2022

Model type	% of women 15 years and above who are FSWs*	Median PSE	95% Credible Set	
			Lower Bound	Upper Bound
Log-linear (Mth Poisson2)	1.0 (0.8–1.2)	34,370	28,164	42,246
Bayesian Model Averaging (non-informative prior)	1.1 (0.9–1.3)	37,647	31,873	43,354
Bayesian Latent Class	1.0 (0.4–1.6)	35,954	14,736	55,215

** Denominators are national total number of adult females aged 15 years and above from 5th Rwanda Population and Housing Census (PHC), 2022.*

Based on the outputs and model diagnostics (Appendices 2-4), the data was found to contain list dependence; therefore, Bayesian Model Averaging with non-informative prior was chosen, which best dealt with list dependence as it automatically detected potential dependencies in the data. After fitting the model, the population size of street- and venue-based FSWs in Rwanda was estimated to be within a credible set ranging from 31,873 to 43,354 with a median of 37,647, corresponding to 1.1% (0.9–1.3) of the general population of adult females aged 15-49 years in Rwanda (*Table 11*).

Relative to adult females in the general population, Western and Northern provinces rank first and second with a higher concentration of FSWs, respectively. The City of Kigali and Eastern Province rank third and fourth, respectively. The Southern province was identified as having a lower concentration of FSWs.

Table 12: FSW provincial Population Size Estimates produced using Bayesian Model Averaging with non-informative prior, FSW PSE, Rwanda 2022.

Province	% Of women 15-49 years who are FSWs *	Median	95% Credible Set	
		PSE	Lower Bound	Upper Bound
City of Kigali	0.8 (0.5–1.0)	3,974	2,815	5,197
Eastern Province	0.6 (0.3–1.0)	5,022	2,535	8,601
Northern Province	1.1 (0.7–1.6)	5,993	3,710	8,876
Southern Province	0.5 (0.2–0.9)	3,884	1,548	6,727
Western Province	1.2 (0.9–1.6)	8,983	6,536	11,791

* *Denominators are provincial total number of adult females aged 15years and above from 5th Rwanda Population and Housing Census (PHC), 2022.*

4.4. FSW Population Size Estimation using Privatize Network Sampling method.

In total, 30 FSWs were enrolled as seeds across 10 study sites countrywide to initiate referral chain recruitment. The maximum wavelength achieved during recruitment was 11 with a mode of 4. *Figure 7-11* below illustrates the RDS recruitment tree by province.

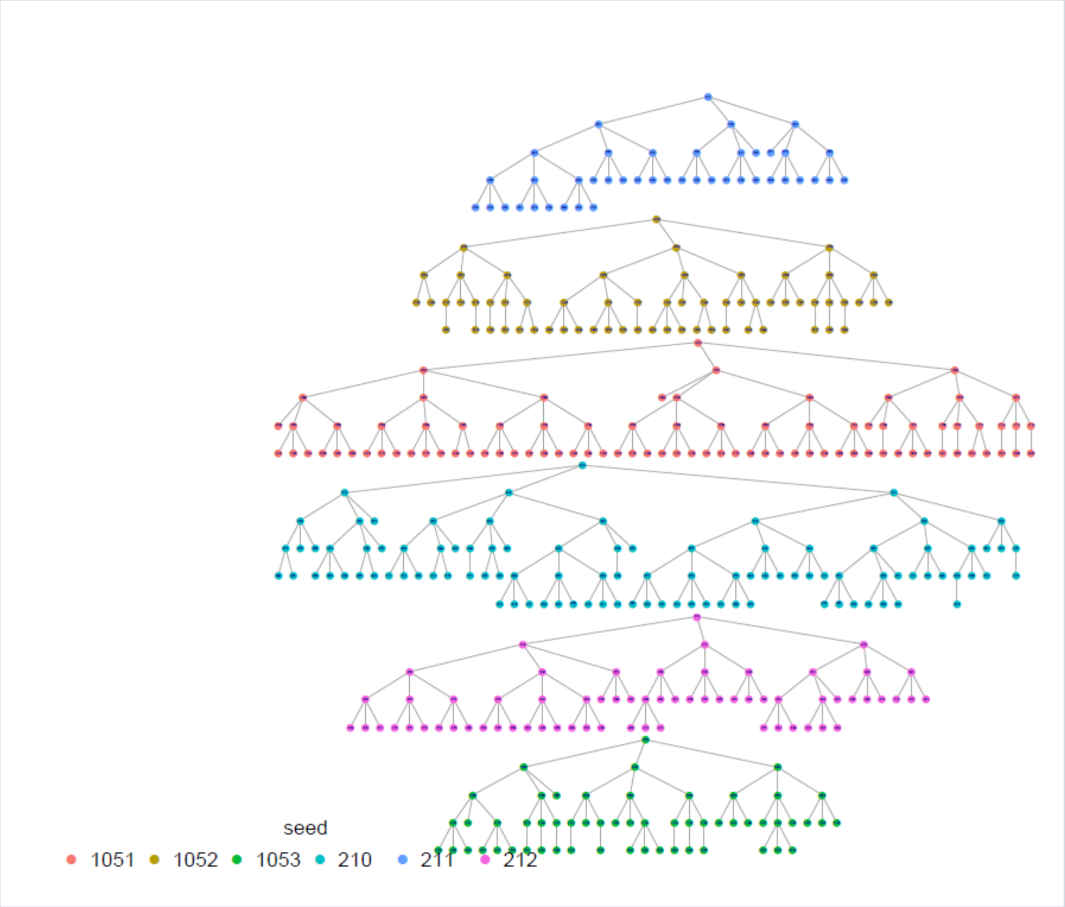


Figure 9: Eastern Province recruitment tree, Rwanda FSW BBS 2023.

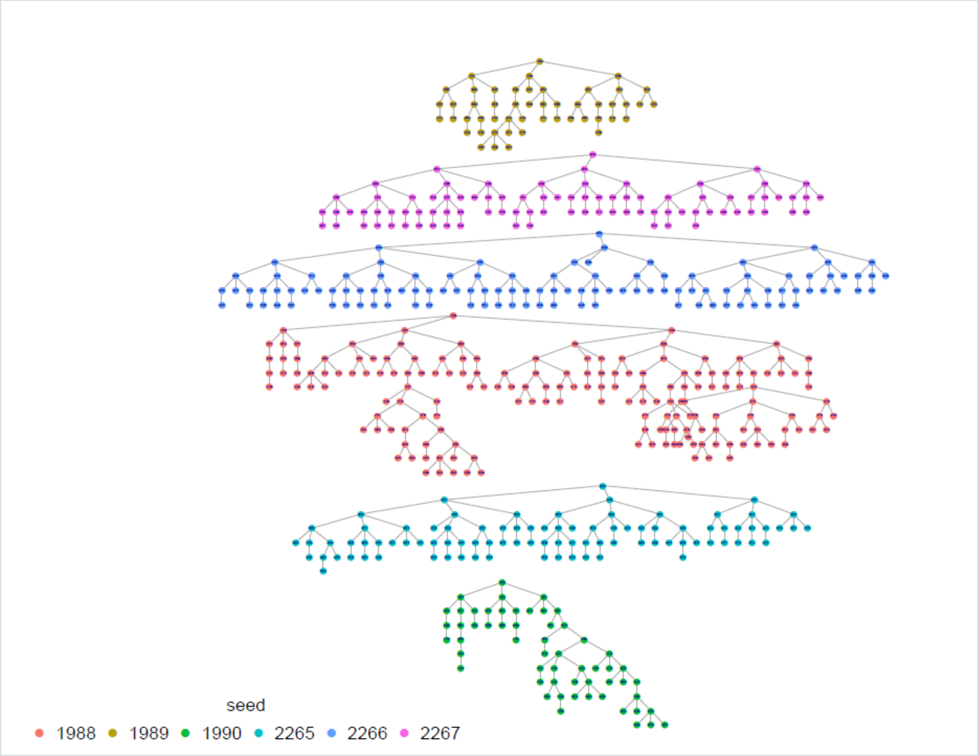


Figure 10: City of Kigali recruitment tree, Rwanda FSW BBS 2023.

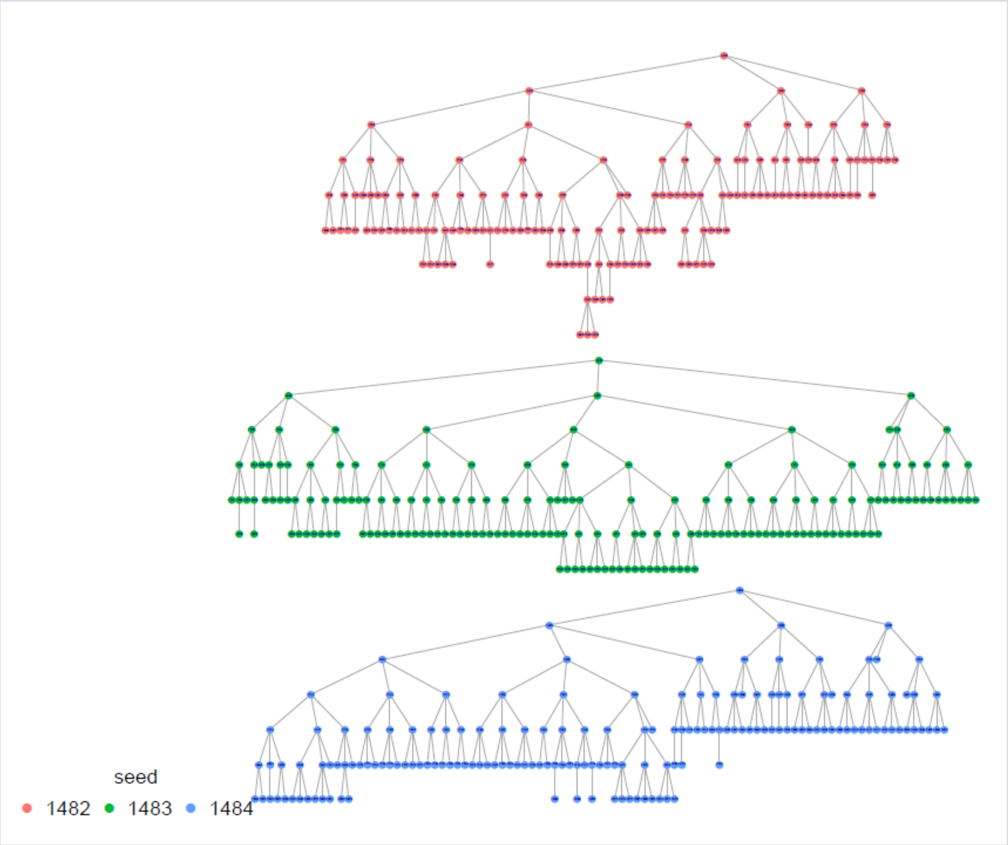


Figure 11: Northern province recruitment tree, Rwanda FSW BBS 2023.

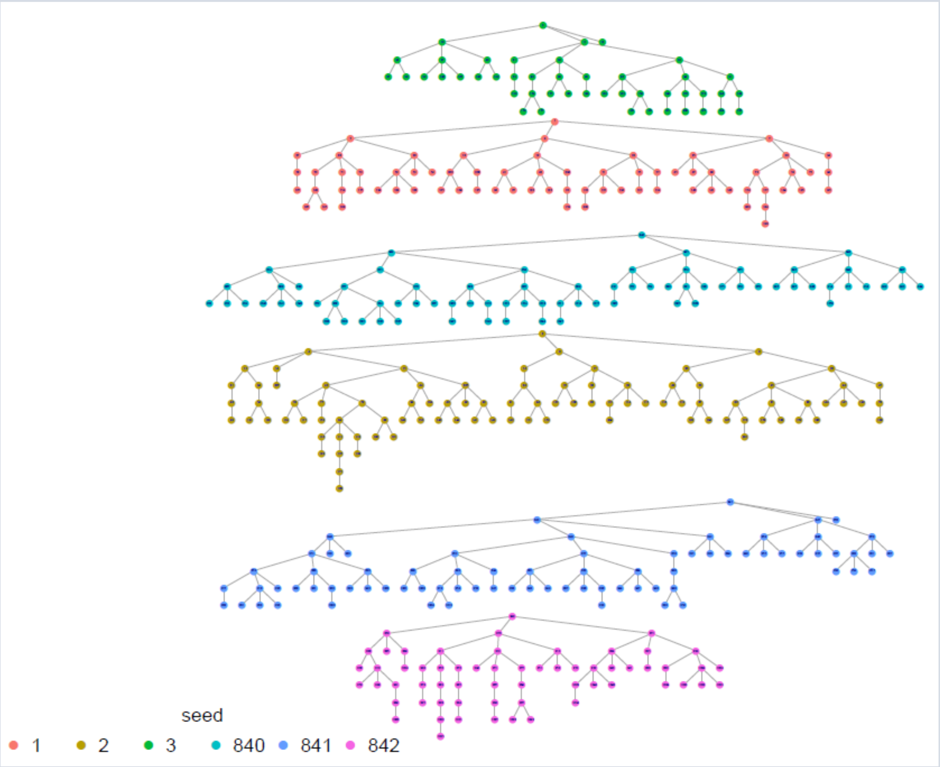


Figure 12: Southern province recruitment tree, Rwanda FSW BBS 2023.

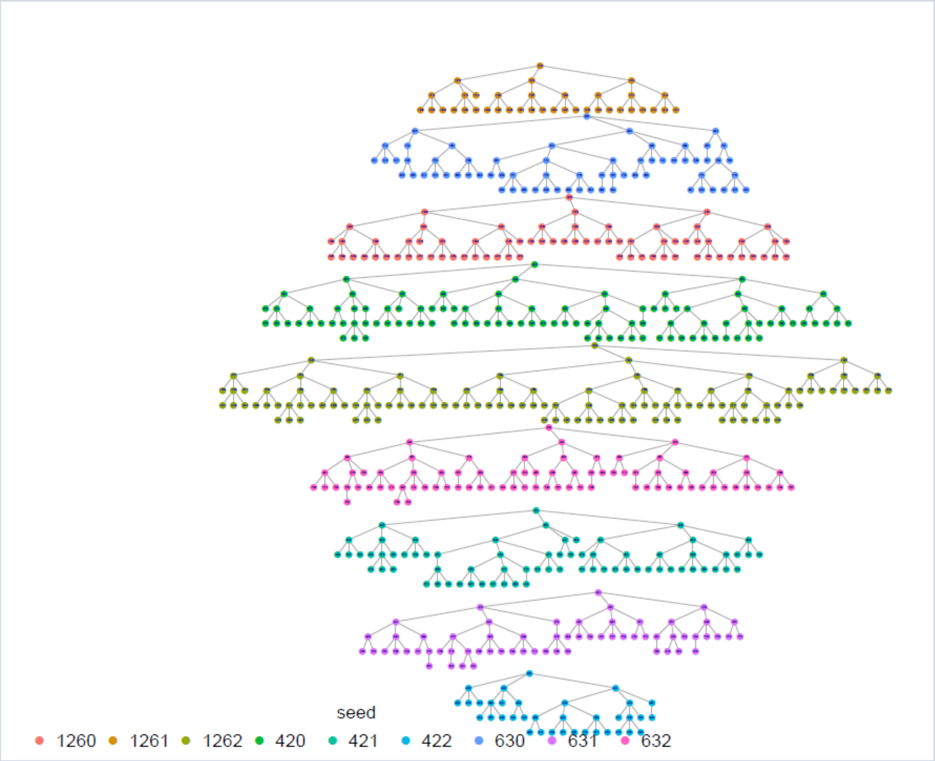


Figure 13: Western province recruitment tree, Rwanda FSW BBS 2023.

The highest proportion of FSW were estimated among those aged 30-39 years, with the smallest proportion among 15-17 age group (0.8%) which was anticipated. The results shows that the study recruitment tapped into all age categories of FSW, which is a great indication for a representative sample as far as age is concerned. Regarding marital status, the majority, 69.8% (95% CI: 66.5-72.9), 22.0% (95% CI:19.3-24.9) were single and divorced/separated respectively, and the same distribution remains across all provinces. In terms of where FSW meet or find clients, FSW participants presented diverse ways including previously untapped subgroups when using venue-based sampling approaches as Time Location Sampling (TLS). in this line, an estimated 14.3% (95% CI: 12.3-16.6) of FSW meet clients using the internet, phone brokers, or escort agency. Of all recruited FSW, 79.4% (95% CI: 76.4 -82.2) had a cell phone and a Sim card, and 81.4% (95% CI: 78.0 – 84.4) of them had at least nominated an FSW peer of whom she knows by name and has a telephone contact. This aligns with what formative assessment (FA) finds and provides confidence in the suggested way of creating the anonymized unique identification using the last 5 digits of one's phone number and name's initials combination to trace deidentified recruitment chains. (see *Table 12*)

Table 13: Social demographic characteristics of participants by province, Rwanda FSW PNS 2023.

		Province					
		Northern	Eastern	Western	City of Kigali	Southern	Overall
	N	Row % [95% CI]	Row % [95% CI]	Row % [95% CI]	Row % [95% CI]	Row % [95% CI]	Row % [95% CI]
Age group							
15 - 17	23	2.6 [1.4 - 4.8]	0.3 [0.1 - 1.5]	1.5 [0.5 - 4.1]	0.4 [0.1 - 1.7]	0.0 [0.0 - 0.2]	0.8 [0.5 - 1.3]
18 - 24	515	26.3 [22.1 - 30.9]	21.3 [16.5 - 27.1]	25.4 [20.8 - 30.7]	24.5 [18.3 - 31.9]	13.6 [9.6 - 19.1]	20.7 [17.9 - 23.9]
25 - 29	497	23.5 [19.4 - 28.0]	22.4 [17.3 - 28.5]	20.8 [16.8 - 25.5]	12.9 [9.1 - 18.0]	19 [14.3 - 24.7]	18.1 [15.7 - 20.8]
30 - 34	535	22.2 [18.1 - 26.9]	21.7 [16.6 - 27.9]	19.8 [15.9 - 24.3]	15.8 [12.3 - 20.2]	26.3 [20.2 - 33.5]	21.4 [18.6 - 24.6]
35 - 39	502	15.0 [11.9 - 18.9]	17.7 [13.4 - 22.9]	19.6 [15.6 - 24.3]	20.0 [15.6 - 25.1]	23.9 [18.3 - 30.6]	20.4 [17.7 - 23.4]
40+	439	10.5 [7.8 - 13.9]	16.6 [12.3 - 22.0]	12.9 [9.9 - 16.8]	26.4 [20.9 - 32.8]	17.1 [12.6 - 22.8]	18.6 [15.9 - 21.5]
Current marital status							
Single	1733	73.3 [68.5 - 77.5]	74.9 [69.0 - 80.0]	77.8 [73.0 - 81.9]	55.0 [48.3 - 61.5]	78.7 [72.4 - 83.9]	69.8 [66.5 - 72.9]
Married/Cohabitating	54	0.3 [0.0 - 2.4]	0.8 [0.3 - 2.5]	1.3 [0.3 - 5.0]	9.3 [6.6 - 13.0]	0.7 [0.2 - 2.6]	3.5 [2.5 - 4.8]
Divorced/Separated	633	24.5 [20.4 - 29.1]	22.0 [17.1 - 27.7]	16.4 [13.0 - 20.5]	27.8 [22.6 - 33.6]	16.8 [12.2 - 22.8]	22.0 [19.3 - 24.9]
Widow	88	1.9 [0.9 - 4.0]	2.3 [1.1 - 4.8]	4.2 [2.4 - 7.5]	7.9 [4.5 - 13.3]	3.5 [1.7 - 7.0]	4.6 [3.2 - 6.7]
Prefer not to answer	3	0	0	0.3 [0.1 - 1.4]	0	0.3 [0.0 - 2.1]	0.1 [0.0 - 0.7]
Education level							
None	490	16.7 [13.2 - 20.9]	25.9 [20.4 - 32.4]	22.9 [18.4 - 28.2]	16.5 [12.5 - 21.4]	18 [13.1 - 24.2]	18.0 [15.5 - 20.8]
Primary	1327	43.5 [38.5 - 48.5]	50.7 [44.1 - 57.2]	49.1 [43.7 - 54.5]	55.1 [48.4 - 61.7]	59.8 [52.6 - 66.6]	54.0 [50.5 - 57.6]
Secondary/vocational/higher education	662	39.8 [34.9 - 44.9]	23.4 [18.4 - 29.3]	24.5 [20.2 - 29.3]	28.4 [22.7 - 34.9]	20.5 [15.2 - 27.0]	27.1 [24.0 - 30.3]
Do not know/No answer	32	0	0	3.5 [1.9 - 6.4]	0	1.7 [0.6 - 4.8]	0.9 [0.4 - 1.9]
Where do you usually meet or find clients?							
Brothel/Guesthouse/Massage/Parlor	118	0.2 [0.0 - 1.1]	1.8 [0.6 - 5.1]	4.8 [3.2 - 7.0]	16.2 [12.2 - 21.3]	3.2 [1.5 - 6.5]	6.9 [5.4 - 8.9]
Hotel/Club/Bar/Restaurant	1110	47.8 [42.8 - 52.9]	54.6 [48.1 - 61.1]	58 [52.6 - 63.2]	36.1 [29.4 - 43.3]	75.5 [69.5 - 80.6]	55.3 [51.8 - 58.8]
Street/Park	720	22.2 [18.3 - 26.7]	7.5 [5.6 - 9.9]	21.2 [17.1 - 25.9]	34.5 [28.8 - 40.8]	8.3 [5.3 - 12.8]	20.3 [17.8 - 23.1]
Other public places	66	2.8 [1.3 - 6.2]	4.4 [2.1 - 9.0]	1.4 [0.6 - 3.0]	1.4 [0.6 - 3.2]	2.0 [1.0 - 4.0]	2.0 [1.4 - 3.0]
Internet, phone broker, escort agency	467	26 [21.7 - 30.8]	30.1 [24.4 - 36.6]	12.4 [9.3 - 16.3]	11.5 [8.0 - 16.3]	9.6 [6.6 - 13.7]	14.3 [12.3 - 16.6]
Other	29	1.0 [0.2 - 4.3]	1.6 [0.6 - 4.3]	2.2 [1.0 - 4.7]	0.2 [0.1 - 0.4]	1.5 [0.6 - 3.8]	1.0 [0.6 - 1.9]
Prefer not to answer	1	0	0	0.1 [0.0 - 0.9]	0	0	0.0 [0.0 - 0.1]
Have a Sim card and cell phone							
Yes	2002	75.2 [70.5 - 79.3]	95.9 [94.3 - 97.1]	76.1 [70.8 - 80.7]	86.6 [82.1 - 90.1]	74.0 [67.2 - 79.9]	79.4 [76.4 - 82.2]
No	509	24.8 [20.7 - 29.5]	4.1 [2.9 - 5.7]	23.9 [19.3 - 29.2]	13.4 [9.9 - 17.9]	26.0 [20.1 - 32.8]	20.6 [17.8 - 23.6]
Has at least one peer's contact							
Yes	1509	93.6 [88.3 - 96.6]	91.5 [89.2 - 93.4]	59.6 [53.5 - 65.5]	64.3 [56.6 - 71.4]	96.0 [94.5 - 97.1]	81.4 [78.0 - 84.4]
No	493	6.4 [3.4 - 11.7]	8.5 [6.6 - 10.8]	40.4 [34.5 - 46.5]	35.7 [28.6 - 43.4]	4.0 [2.9 - 5.5]	18.6 [15.6 - 22.0]
Number of peers from whom contacts are available							
One peer	188	3.7 [1.9 - 7.2]	2.4 [1.4 - 3.9]	15.7 [10.4 - 23.0]	37.3 [28.7 - 46.7]	1.8 [1.1 - 2.9]	12.9 [9.9 - 16.6]
Two peers	270	6.8 [4.5 - 10.1]	1.7 [0.9 - 3.3]	26.5 [20.9 - 33.0]	32.8 [25.6 - 41.0]	2.7 [1.8 - 4.1]	13.2 [10.9 - 16.0]
Three peers	548	36.6 [31.1 - 42.4]	95.6 [93.5 - 97.0]	37.4 [30.3 - 45.2]	16.6 [11.7 - 23.1]	4.1 [2.5 - 6.9]	21.3 [18.8 - 24.2]
Four Peers	128	16 [12.0 - 21.1]	0.2 [0.0 - 1.2]	9.9 [6.6 - 14.6]	7.1 [4.3 - 11.5]	5.8 [2.9 - 11.4]	8.1 [6.2 - 10.5]
Five peers	238	22.1 [17.6 - 27.4]	0.1 [0.0 - 0.3]	8.3 [5.3 - 12.9]	3.3 [1.6 - 6.6]	67.2 [59.3 - 74.3]	33.1 [28.7 - 37.8]
More than five peers	137	14.8 [11.3 - 19.1]	0.0 [0.0 - 0.1]	2.2 [1.0 - 4.6]	2.9 [1.3 - 6.3]	18.3 [12.7 - 25.7]	11.3 [8.7 - 14.6]

Table 14: presents both unadjusted and adjusted study site-level population size estimates by each estimator used.

		Unadjusted Estimates			Adjusted Estimates		
Study site	**Estimators**	PSE	Lower Bound	Upper Bound	PSE	Lower Bound	Upper Bound
GIHUNDWE HC	Cross-Sample	15,538	3,399	71,023	19,599	4,288	89,586
GISENYI HC	Cross-Sample	1,463	1,033	2,073	1,494	1,054	2,116
GITARAMA HC	Cross-Sample	6,256	2,884	13,571	7,839	3,613	17,005
KIBUYE HC	Cross-Sample	5,875	2,995	11,527	8,728	4,448	17,123
MUHOZA HC	Cross-Sample	104,226	41,554	261,421	140,369	55,964	352,075
MUKARANGE HC	Cross-Sample	1,021	773	1,349	1,371	1,038	1,811
NYAGATARE HC	Cross-Sample	3,905	1,945	7,842	3,905	1,945	7,842
RANGO HC	Cross-Sample	2,903	2,151	3,919	3,852	2,854	5,200
CITY OF KIGALI	Cross-Sample	11,314	8,029	15,942	14,414	10,229	20,310
GIHUNDWE HC	Cross-Alter	4,043	2,975	5,494	5,099	3,752	6,930
GISENYI HC	Cross-Alter	1,511	1,159	1,969	1,542	1,183	2,009
GITARAMA HC	Cross-Alter	6,006	3,830	9,418	7,526	4,799	11,801
KIBUYE HC	Cross-Alter	3,696	2,830	4,827	5,490	4,203	7,170
MUHOZA HC	Cross-Alter	27,472	21,272	35,479	36,999	28,649	47,783
MUKARANGE HC	Cross-Alter	354	226	556	476	303	746
NYAGATARE HC	Cross-Alter	10,166	5,196	19,892	10,166	5,196	19,892
RANGO HC	Cross-Alter	4,650	3,580	6,039	6,170	4,750	8,013
CITY OF KIGALI	Cross-Alter	18,457	13,978	24,373	23,514	17,807	31,051
GIHUNDWE HC	Cross-Network	5,479	3,996	7,512	6,911	5,040	9,476
GISENYI HC	Cross-Network	1,491	1,193	1,863	1,522	1,218	1,901
GITARAMA HC	Cross-Network	6,106	4,117	9,055	7,651	5,159	11,346
KIBUYE HC	Cross-Network	4,241	3,343	5,381	6,300	4,965	7,993
MUHOZA HC	Cross-Network	32,758	25,288	42,434	44,117	34,057	57,149
MUKARANGE HC	Cross-Network	597	454	785	801	610	1,053
NYAGATARE HC	Cross-Network	7,200	4,364	11,881	7,200	4,364	11,881
RANGO HC	Cross-Network	4,191	3,351	5,243	5,562	4,446	6,957
CITY OF KIGALI	Cross-Network	14,540	11,991	17,630	18,523	15,277	22,460

Site-level population size estimate variation was observed across estimators. The population size estimates produced by the cross-sample estimator were greater than those produced using the cross-Alter estimator at the Gihundwe HC, Kibuye HC, Muhoza HC, and Mukarange HC study sites. Population size estimates from Gisenyi HC and Gitarama HC remained almost the same across the two estimators. However, population size estimates produced by the cross-sample estimator were less than those produced using the cross-Alter estimator for the Nyagatare HC, City of Kigali, and Rango HC study sites. Site-level population size estimates produced using the cross-network estimator were observed to be between the estimates produced using cross-sample and cross-Alter estimators except for the Gisenyi HC study site. The same patterns were observed for both adjusted and unadjusted population size estimates; however, the differences are not statistically significant (see *Table 13*). Site-level estimates present high volatility; thus, their use and interpretation should be done cautiously.

Table 15: Overall Population Size Estimates by estimator, Rwanda FSW BBS 2023.

	Unadjusted Estimates			**Adjusted Estimates**		
Estimators	PSE	Lower Bound	Upper Bound	PSE	Lower Bound	Upper Bound
Cross-Sample	152,502	64,763	388,667	201,570	85,434	513,067
Cross-Alter	76,354	55,045	108,046	96,981	70,643	135,394
Cross-Network	76,603	58,097	101,782	98,587	82,978	114,196

Looking at the pooled population size estimate, the Cross-Sample estimator tends to produce larger estimates as compared to the Cross-Alter and Cross-Network estimators considering both unadjusted and adjusted estimates. On the other hand, population size estimates produced using cross-Alter and cross-network estimators go hand in hand considering both adjusted and unadjusted estimates (see *Table 14)*. The observed patterns in estimates across estimators aligns with several simulation study finding where Cros Alter and Cross Network estimators present similar level of performance, but Cross Alter shows an elevated variance level[48].

Table 15 presents provincial-level adjusted FSW population size estimates produced using the Cross Network estimator, which combines both cross-sample and cross-Alter estimator features and presents the lowest variance and bias as compared to other estimators.

Table 16: Cross-Network Estimator's Population Size Estimates by province, Rwanda FSW BBS 2023.

Province	Consensus Population Size Estimate [95%CI]	*PSE as % of female aged 15+ years of the general population
Northern	11,317 [9,526 -13,109]	1.6% [1.3-1.9]
Southern	15,826 [13,320 -18,331]	1.7% [1,4-1.9
Eastern	19,833 [16,693 -22,973]	1.7% [1.4-2.0]
Western	20,593 [17,332 -23,853]	2.2% [1.8-2.5]
City of Kigali	31,018 [26,107- 35,929]	5.3% [4.5-6.1]
Total	**98,587 [82,978- 114,196]**	**2.3% [1.9-2.6]**

**Calculated based on the 5th Rwanda Housing and Population census, 2022[81].*

The national level FSW PSE was estimated at 98,587 (95% CI: 82,978 – 114,196), corresponding to 2.3% of the total adult female population aged 15 years and above in Rwanda. The highest FSW PSE was observed in the City of Kigali with 31,018 (95% CI: 26,107 – 35,929), followed by West province with 20,593 (95% CI: 17,332 – 23,853), the East province with 19,833 (95% CI: 16,693 – 22,973), and the South province with 15,826 (95% CI: 13,320 – 18,331). The lowest FSW population size was estimated in the North province with 11,317 (95% CI: 9,526 – 13,109). These patterns go hand in hand with vibrating areas in Rwanda where most of FSW attractive activities are found, including but not limited to economic, tourism and leisure activities.

Chapter 5

Discussion

5.1. Introduction to the discussion section

This chapter discuss the results obtained when applying the methods described in chapter 3. It also provides the comparison of the efficiency of the statistical methods applied in this research with other similar literature. Furthermore, comparison between the research findings with the available literatures.

5.2. MSM Population Size Estimation using Three-source capture recapture method: Bayesian nonparametric latent class model.

The MSM population size estimation presents the first use of 3S-CRC to estimate MSM population size on a nationwide scale in Rwanda. The 3S-CRC method provided an estimate for MSM in the City of Kigali with 7,842 (95% CI: 4,587 – 13,153) that was similar to what was estimated by "Projet San Francisco (PSF)" (8,411 (6,760 – 11,151))[82]. Considering geographical coverage and difference in methodologies used, the 2018 and 2021 Kigali MSM population size estimations, we are able to understand the slight differences in estimates being observed. The distribution of MSM in each of the remaining 4 provinces was fairly uniform and lower than the estimates in Kigali. Differences in estimate of the MSM population size distribution across the country may also reflect long-term movement patterns among MSM, from rural to urban as well as from smaller to larger urbanized contexts[83].

To some degree, MSM size estimates are influenced by the proportion of MSM who may not participate in the study due to potential privacy concerns, a potentially significant element given the burden of stigma and heteronormative behavioral expectations (e.g. marriage and parenting) for MSM in Rwanda. Our overall MSM population size estimate represents 0.7% [0.4 - 1.1] of the total adult male population in Rwanda based on 2012 Census data by National Institute of Statistics of Rwanda 2021 population size projections. The 2020 WHO and UNAIDS technical brief recommends the revision of the MSM PSE for those countries using the MSM PSE less than 1% of total adult males based on the region [19]. UNAIDS monitoring system through 2019, estimated a global median proportion of adult men who had sex with another man in the previous year of 1.9% across 38 low or middle-income countries, this proportion is at 1.45% in Eastern and Southern Africa where Rwanda is located [84-86]. From that, the estimate from the current study aligns with the WHO recommendation regarding to MSM PSEs [26].

The final estimate for this study was based on 3S-CRC dataset. In summary, four major assumptions must be met for CRC to give reliable population estimates: individual captures are independent; the population is closed; each target population member's capture history is correct; and the chance of getting caught is homogeneous [43].

To minimize dependencies between captures, we used different distribution settings for each capture occasion. During the first capture, members of the MSM population were tagged by the key chain through their corresponding associations, groups, and key informants. In the second capture, MSM were tagged by being offered MSM friendly services through health facilities that usually serve MSM nationwide. In third capture, we used RDS method where all recruited MSM were considered as captured during capture three. For all the three captures, a one-week time interval was used between two consecutive capture occasions to minimize recall bias and fulfilling population closeness assumption. At each capture of the first two capture rounds; unique object distribution, MSM friendly services provision, procedures ensured a random aspect to ensure that the chance of getting caught is homogeneous. However, our estimates might be limited with missing a random sampling aspect during the third round where RDS was used.

There were other several plausible constraints to the design of our estimation activity. Possible limitation of the underlying 3S-CRC assumptions that might have influenced the validity of our findings, leading to reduced accuracy of population sizes and wide confidence ranges. To begin with, we employed unique objects as a tagging strategy to protect the anonymity of sampled populations. However, not all individuals were carrying the received object at the subsequent capture occasion, complicating the identification of recaptures. Furthermore, we had to assume that the person presenting the object is the person who received the object (an essential limitation present in anonymous sampling-based CRC). We tried to mitigate these limitations by offering the opportunity to identify the correct object from a laminated card with several pictures including the correct unique object for those presenting without unique objects with them. There was a possibility of guessing or having seen the object and therefore biasing the PSE downwards. To overcome possible participation duplicates enrollment during the third capture, a biometric system using fingerprint identification was installed and employed across all study sites. We also acknowledge any possible selection bias that might have been influenced by the study inclusion criteria set.

The key strength of our study is that it is powered to provide national and provincial level PSE for MSM in Rwanda for the very first time. Sampling considered administrative provinces as strata and targeted 28 (out of 30) districts in Rwanda which are nationally representative and reflective of the demography. The selected districts included key urban areas with a high likelihood of expanding the catchment to include participation by MSM based in rural areas. Furthermore, during the third capture occasion, RDS was used giving more confidence in reaching MSM with lower social visibility.

The final estimate of the MSM population size in Rwanda is based on a Bayesian model averaging approach to account for the complex patterns of heterogeneity between captures and the aggregation of homogeneous strata into latent classes. While other statistical techniques make reasonably strong assumptions about the structure of the joint distribution of capture patterns, the latent-class Bayesian method, on the other hand, is a model-averaging strategy that seeks to estimate the joint distribution as directly as feasible from the data [58].

5.3. FSW Population Size Estimation using Three-source capture recapture method: Bayesian Model Averaging

This study provides both national and provincial-level estimates of the population size of street- and venue-based FSWs and sexually exploited minors aged 15 and above in Rwanda. The population size of street- and venue-based FSWs and sexually exploited minors is estimated to be within a credible set ranging from 31,873 to 43,354, with a median of 37,647, corresponding to 1.1% (0.9–1.3) of adult females aged 15 years and older in the general population. These results indicate a significant difference in the FSW population size as compared to the 2018 population size of FSWs aged 15 and above, which was estimated to range from 8,328 to 22,806 credible sets with a median of 13,716 [47].The difference might be attributed to several factors, including but not limited to differences in estimation models used and geographical coverage.

Furthermore, this study provides a provincial-level population size estimate of street- and venue-based FSWs and sexually exploited minors aged 15 and above for the very first time. The largest population size estimate was found in the Western province, followed by the Northern and Eastern provinces. The City of Kigali and Southern Province were found to have relatively lower estimates of the FSW population as compared to other provinces. Differences in estimates distribution across the country may reflect long-term internal movement patterns among FSWs, from rural to more urbanizing areas as well as from smaller to larger urbanized contexts, as indicated by the Rwanda Population and Household Census 2022 [81].

The findings from the 2022 female sex workers population size estimation might not have considered high profile and those FSWs using web-based and social media platforms to reach their clients, leading to a slight possible underestimation of the true population size. Furthermore, we acknowledge possible methodological limitations that might influence the final FSW PSE from this study. Compared to program coverage data from the Rwanda Health Management Information System (RHMIS), the key strength of our study is that it is powered to provide national and provincial-level PSE for FSW in Rwanda for the very first time.

So far, three rounds of FSW population size estimation have been conducted in Rwanda since 2010 [45-47]. The 2010 FSW size estimation using capture recapture and multiplier methods estimated the national population size of FSWs to range from 2,998 to 3,412 with a median of 3,205. In 2012, the population size of FSWs was estimated to range from 23,000 to 39,000. Later, after 6 years, in 2018, the national population size of FSWs was estimated to range from 8,328 to 22,806, with a median of 13,716. These differences in the population size of FSWs might be attributed to different reasons, including but not limited to methodological or geographic coverage differences. Compared with the previous three rounds of FSW population size estimation exercises, we observed a difference in the FSW population size, which also might be attributed to the reasons stated above.

5.4. FSW Population Size Estimation using Privatize Network Sampling method.

The national population size of female sex workers and sexually exploited minors aged 15 and above using PNS approach in Rwanda was estimated at 98,587 (95% CI: 82,978 – 114,196), corresponding to 2.3% of the total females aged 15 years and above in Rwanda based on the 5th Rwanda Housing and Population census, 2022[81]. The highest FSW concentration was found in the City of Kigali, with 5.3% [4.5 – 6.1] and the lowest in the North province with 1.6% [1.3 – 1.9] as % of female aged 15 years and above of the general population.

The reported FSW population size estimate looks larger when compared with previously estimated size of FSW population in Rwanda. The variances between the current FSW PSE and the previous ones can be more potentially attributed to the methodological capability to reach non-venue based FSWs around the country. The National HIV annual report 2022-23 reported a total number of 60,460 FSW identified during the reporting year period in the HIV program [87]. Believing that the program reported number is not exhaustive and recognizing the limitations associated with program data, including the possible inability to deidentify individual-level data, give more confidence in the PSE resulted from the current study.

One important limitation associated to the use of PNS to estimate the population size in our context, is related to the coverage of SIM cards and cell phones among FSW within the survey sample, which was used to produce hashed ids to uniquely identify overlaps (alters) in the tracing network sampling. Access to mobile technology varies among individuals, and not all FSWs may possess a SIM card or a functioning cell phone. Furthermore, the estimates from this study might have been affected by those FSW who decide not to provide FSW phone numbers, the data collection tool used that did enable to delimit participants only for the province where the survey was being implemented, and the social desirability bias of the BBS. Lastly, we acknowledge methodological related limitations in line with underlining PNS assumptions.

To minimize the effects of the listed limitations on the estimates, the study has adjusted the estimates to account for the proportion of those FSW who did not have SIM card and cell phones. In addition, to have provincial-level estimates, we assumed that the provincial-level distribution of FSW population reported by HIV program in its' 2022-2023 HIV annual report remains the same for the current FSW population size estimation, therefore the pooled estimate was proportionally distributed across provinces using the percent proportion distribution from the annual HIV report 2022-2023. Regarding the PNS underlining assumptions, the sample size used was a significant portion of the previously estimated size of FSW, and a long recruitment chain was achieved by reaching 11 waves.

Chapter 6

Conclusion and future work

In conclusion, MSM PSE 2021 study provides for the first time, an estimated population size of MSM aged 18 years and above in Rwanda. The results will allow national programs and implementation partners to invest in HIV services at a level that is commensurate with need, coverage, and new infections. These data enable policymakers and planners to monitor HIV epidemic control nationwide, specifically, among the MSM population and to plan for other health services, such as prevention and treatment of STIs, among others.

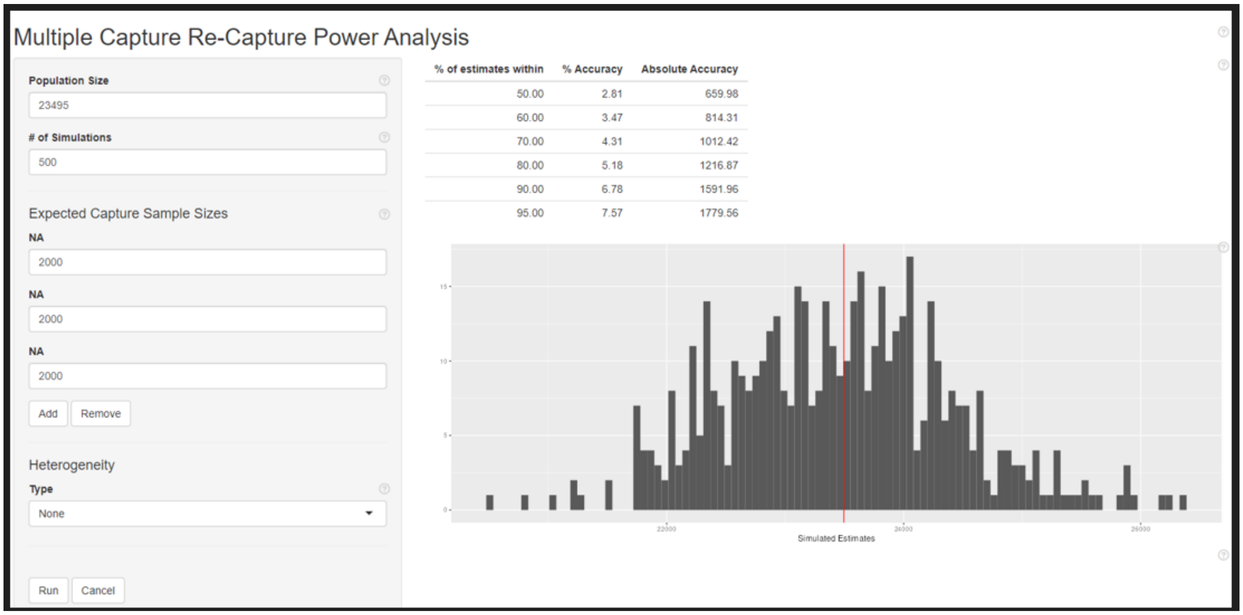
Furthermore, this study sheds light on critical aspects of the female sex workers (FSW) population in Rwanda, revealing a higher concentration compared to the regional average, with 2.3% identified as FSWs of the total adult females in the general population in contrast to the 1.1% reported in sub-Saharan Africa[88]. FSW population size estimate derived from these studies serves as the basis for targeted interventions, resource allocation, advocacy efforts, resources mobilization, and policy development aimed at improving access to preventive, care and treatment services for FSW in Rwanda.

While these estimates are usable at national and provincial levels, further work is needed on small area estimation to align PSE results with the intended HIV treatment and prevention interventions at sub-national levels among MSM. Furthermore, we acknowledge that there are still limitations of estimating some hard-to-reach MSM groups, this is a potential area for further research. Additionally, this is the first time that PNS is implemented for the estimation of the FSW population size estimate in Rwanda, adding to the emerging tools that we have in the hard-to-reach PSE field. Moving forward, future research endeavors could explore smaller area estimation techniques to ascertain the distribution of FSWs at the district level, enabling more localized and tailored interventions. Furthermore, we call for further research to investigate a way of using know previous data, literature information as priors instead of using non-informative priors while conducting Bayesian modeling.

Finally, this research project demonstrates how the Generalized CRC model for PSE derived from Bayesian model averaging process provides a remedial approach to overcome model selection bias and privatize network sampling overcomes tag loss bias as well as demonstrating its capability to reach harder-to-reach key population subgroups.

Appendices

Appendix 1: Multiple Capture Re-Capture Power Analysis



Multiple Capture Re-Capture Power Analysis

Population Size: 23495

# of Simulations: 500

Expected Capture Sample Sizes

NA: 2000

NA: 2000

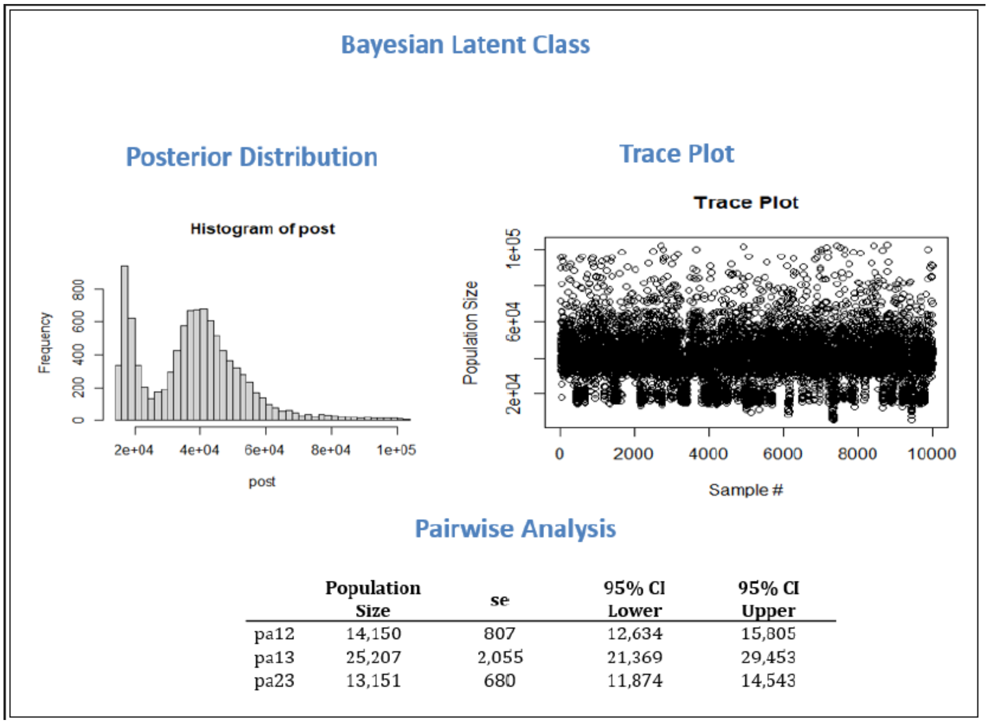
NA: 2000

Heterogeneity

Type: None

% of estimates within	% Accuracy	Absolute Accuracy
50.00	2.81	659.98
60.00	3.47	814.31
70.00	4.31	1012.42
80.00	5.18	1216.87
90.00	6.78	1591.96
95.00	7.57	1779.56

Appendix 2: Bayesian Latent Class



Bayesian Latent Class

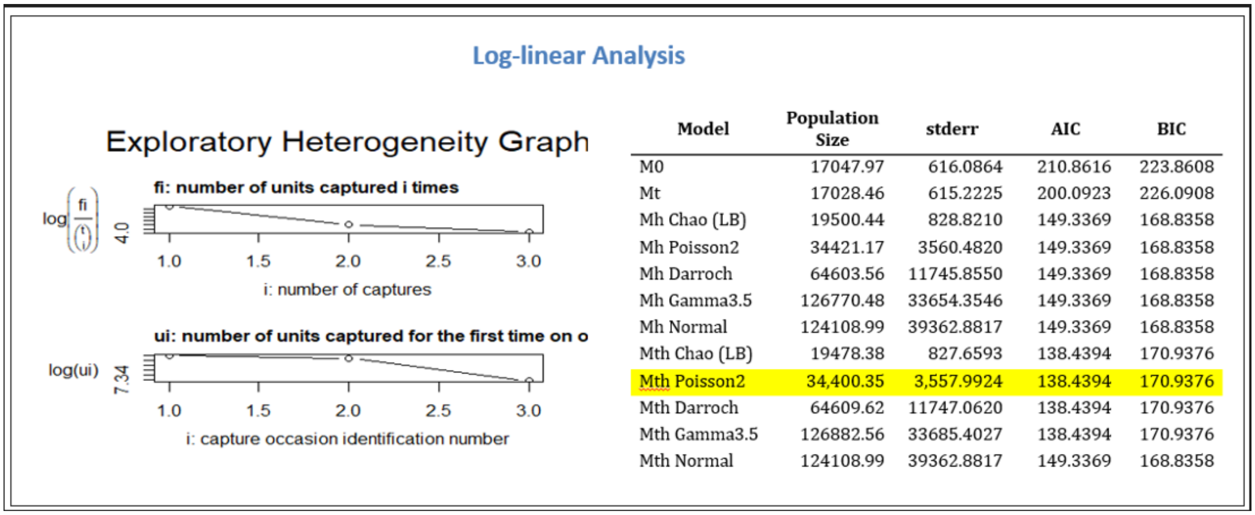
Posterior Distribution

Trace Plot

Pairwise Analysis

	Population Size	se	95% CI Lower	95% CI Upper
pa12	14,150	807	12,634	15,805
pa13	25,207	2,055	21,369	29,453
pa23	13,151	680	11,874	14,543

Appendix 3: Log-linear models

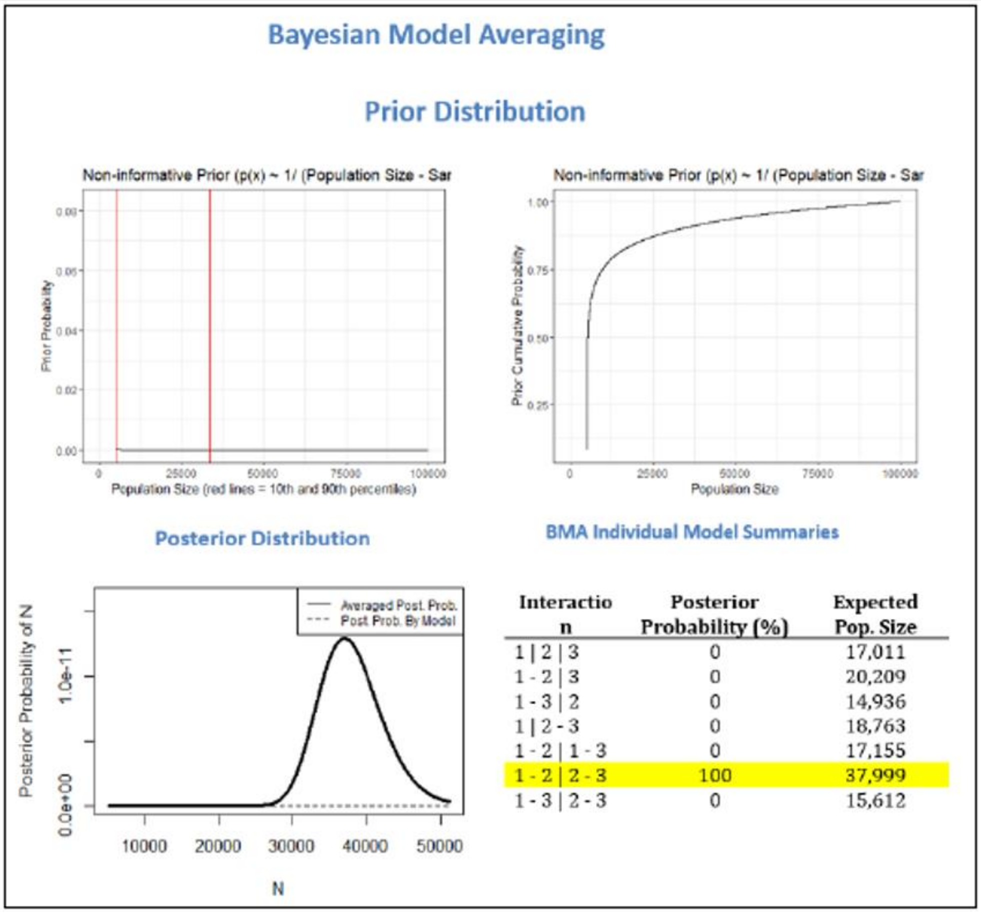


Log-linear Analysis

Exploratory Heterogeneity Graph
fi: number of units captured i times
log(fi / (3 i))
4.0
1.0
1.5
2.0
2.5
3.0
i: number of captures
ui: number of units captured for the first time on o
log(ui)
7.34
1.0
1.5
2.0
2.5
3.0
i: capture occasion identification number

Model	Population Size	stderr	AIC	BIC
M0	17047.97	616.0864	210.8616	223.8608
Mt	17028.46	615.2225	200.0923	226.0908
Mh Chao (LB)	19500.44	828.8210	149.3369	168.8358
Mh Poisson2	34421.17	3560.4820	149.3369	168.8358
Mh Darroch	64603.56	11745.8550	149.3369	168.8358
Mh Gamma3.5	126770.48	33654.3546	149.3369	168.8358
Mh Normal	124108.99	39362.8817	149.3369	168.8358
Mth Chao (LB)	19478.38	827.6593	138.4394	170.9376
Mth Poisson2	34,400.35	3,557.9924	138.4394	170.9376
Mth Darroch	64609.62	11747.0620	138.4394	170.9376
Mth Gamma3.5	126882.56	33685.4027	138.4394	170.9376
Mth Normal	124108.99	39362.8817	149.3369	168.8358

Appendix 4: Bayesian Model averaging



Bayesian Model Averaging

Prior Distribution

Non-informative Prior (p(x) ~ 1/ (Population Size - Sar
Prior Probability
0.08
0.06
0.04
0.02
0.00
0
25000
50000
75000
100000
Population Size (red lines = 10th and 90th percentiles)
Non-informative Prior (p(x) ~ 1/ (Population Size - Sar
Prior Cumulative Probability
1.00
0.75
0.50
0.25
0
25000
50000
75000
100000
Population Size

Posterior Distribution

Posterior Probability of N
1.0e-11
0.0e+00
Averaged Post. Prob.
Post Prob. By Model
10000
20000
30000
40000
50000
N

BMA Individual Model Summaries

Interaction	Posterior Probability (%)	Expected Pop. Size
1 | 2 | 3	0	17,011
1 - 2 | 3	0	20,209
1 - 3 | 2	0	14,936
1 | 2 - 3	0	18,763
1 - 2 | 1 - 3	0	17,155
1 - 2 | 2 - 3	100	37,999
1 - 3 | 2 - 3	0	15,612

Appendix 5: R Codes used for data analysis.

R Codes used for data analysis: Bayesian nonparametric latent-class model

```
##############################################################################
##      R PROGRAM: Elysee_PSE.R
##
##        PROJECT: FSW-PSE: Estimating the Size of MSM
##                 Population in Rwanda Using Three-Source
##                 Capture-Recapture Methods, 2021.
##
## INVESTIGATOR(S): Elysee TUYISHIME
##
## DISCLAIMER: Although this program has been developed and used for the purposes of data analysis
##                    in this thesis, no warranty, expressed or implied, is made by the investigator as to the
##                    accuracy and functioning of the program and related program material nor shall the fact
##                    of distribution constitute any such warranty, and no responsibility is assumed by the
##                    investigator in connection therewith.
##
##     CHANGE LOG: Date       Change
##                ----------  ----------------------------------------
##############################################################################

basepath <- "C:/Users  /IBBSS_PSE Among MSM 2021/Datasets/Final data files for Analysis/Final/Outputs"
setwd(basepath)
install.packages('Rcapture')
install.packages('LCMCR')
install.packages('lattice')

library(Rcapture)
library(LCMCR)
library(lattice)
##############################################################################
## Read the data exactly as provided by Ermias
##############################################################################
datafsw<-('
        ch1 ch2 ch3 Freq
         1   0   0   2465
         0   1   0   1314
         0   0   1   2229
         1   1   0    721
         1   0   1    422
         0   1   1    415
         1   1   1    210')
FSW <- read.table(textConnection(datafsw),header=TRUE)

desc <- descriptive(FSW, dfreq = TRUE)
plot(desc)
```

```
################################################################################
## Fit frequentist loglinear models
################################################################################
Res <- closedp(FSW, dfreq = TRUE)
print(Res)

################################################################################
## Heterogeneity graph
################################################################################
boxplot(Res)

################################################################################
## Construct results table
################################################################################
CI0 <- closedpCI.t(FSW, dfreq = TRUE, m=("M0"))
CIt <- closedpCI.t(FSW, dfreq = TRUE, m=("Mt"))
CIh.c <- closedpCI.t(FSW, dfreq = TRUE, m=("Mh"), h=c("Chao"))
CIh.p <- closedpCI.t(FSW, dfreq = TRUE, m=("Mh"), h=c("Poisson"))
CIh.d <- closedpCI.t(FSW, dfreq = TRUE, m=("Mh"), h=c("Darroch"))
CIh.g <- closedpCI.t(FSW, dfreq = TRUE, m=("Mh"), h=c("Gamma"))
CIth.c <- closedpCI.t(FSW, dfreq = TRUE, m=("Mth"), h=c("Chao"))
CIth.p <- closedpCI.t(FSW, dfreq = TRUE, m=("Mth"), h=c("Poisson"))
CIth.d <- closedpCI.t(FSW, dfreq = TRUE, m=("Mth"), h=c("Darroch"))
CIth.g <- closedpCI.t(FSW, dfreq = TRUE, m=("Mth"), h=c("Gamma") )
Est <- rbind(CI0$CI, CIt$CI, CIh.c$CI, CIh.p$CI, CIh.d$CI, CIh.g$CI,
            CIth.c$CI, CIth.p$CI, CIth.d$CI, CIth.g$CI)[, 1:3]
Est <- cbind(Est, Res$results[1:10, c(3, 5)])
N <- sum(FSW$Freq)
dfR <- N - Res$results[, 4]
CritChisq <- qchisq(0.025, df = dfR, lower.tail = FALSE)
pVal1 <- pchisq(Res$results[, 3], df = dfR, lower.tail = FALSE)[1:10]
Est <- cbind(Est, pVal1)
write.csv(Est, file = file.path(basepath, "Loglinear_results.csv"))
################################################################################
## ui fit Chi-squared values
################################################################################
uiChiSq <- (uifit(Res)$fit.stat)[1:10]
pVal2 <- pchisq(uiChiSq, Res$results[1:10, 4], lower.tail = FALSE)
cbind(uiChiSq, pVal2)

################################################################################
## Bayesian nonparametric latent-class model
################################################################################
x <- seq(0, 1, by = 0.01)
png(file.path(basepath, "priors.png"))
par(mfrow = c(2,3))
plot(x, dbeta(x, shape1 = 0.25, shape2 = 0.25), type = "l")
plot(x, dbeta(x, shape1 = 1, shape2 = 1), type = "l")
plot(x, dbeta(x, shape1 = 5, shape2 = 5), type = "l")
plot(x, dbeta(x, shape1 = 1, shape2 = 5), type = "l")
plot(x, dbeta(x, shape1 = 5, shape2 = 1), type = "l")
dev.off()
par(mfrow = c(1,1))

FSW[,c(1:3)] <- lapply(FSW[,c(1:3)] , factor)
## nlcm with Jeffrey's hyperprior
smplr.Jeff <- lcmCR(FSW, tabular = TRUE, K = 10, a_alpha = 0.025,
            b_alpha = 0.025, seed = 123, buffer_size = 10000,
            thinning = 100)
post.Jeff <- lcmCR_PostSampl(smplr.Jeff, burnin = 100000,
                 samples = 10000, thinning = 100,
                 output = FALSE)
CI.Jeff <- quantile(post.Jeff, c(0.025, 0.5, 0.975))
CI.Jeff <- as.data.frame(t(CI.Jeff))
CI.Jeff <- CI.Jeff[,c(2,1,3)]
colnames(CI.Jeff)[1:3]<-c("Median","LowerCI", "UpperCI")
CI.Jeff
## nlcm with uniform hyperprior
smplr.unif <- lcmCR(FSW, tabular = TRUE, K = 10, a_alpha = 1,
            b_alpha = 1, seed = 123, buffer_size = 10000,
            thinning = 100)
post.unif <- lcmCR_PostSampl(smplr.unif, burnin = 100000,
                 samples = 10000, thinning = 100,
                 output = FALSE)
CI.unif <- quantile(post.unif, c(0.025, 0.5, 0.975))
CI.unif <- as.data.frame(t(CI.unif))
CI.unif <- CI.unif[,c(2,1,3)]
colnames(CI.unif)[1:3]<-c("Median","LowerCI", "UpperCI")
CI.unif
## nlcm with symmetric 5,5 hyperprior
smplr.5_5<- lcmCR(FSW, tabular = TRUE, K = 10, a_alpha = 5,
            b_alpha = 5, seed = 123, buffer_size = 10000,
            thinning = 100)
post.5_5 <- lcmCR_PostSampl(smplr.5_5, burnin = 100000,
                 samples = 10000, thinning = 100,
                 output = FALSE)
CI.5_5 <- quantile(post.5_5, c(0.025, 0.5, 0.975))
CI.5_5 <- as.data.frame(t(CI.5_5))
CI.5_5 <- CI.5_5[,c(2,1,3)]
colnames(CI.5_5)[1:3]<-c("Median","LowerCI", "UpperCI")
CI.5_5
## nlcm with positively skewed 1,5 hyperprior
smplr.1_5<- lcmCR(FSW, tabular = TRUE, K = 10, a_alpha = 1,
            b_alpha = 5, seed = 123, buffer_size = 10000,
            thinning = 100)
```

```
post.1_5 <- lcmCR_PostSampl(smplr.1_5, burnin = 100000,
                  samples = 10000, thinning = 100,
                  output = FALSE)
CI.1_5 <- quantile(post.1_5, c(0.025, 0.5, 0.975))
CI.1_5 <- as.data.frame(t(CI.1_5))
CI.1_5 <- CI.1_5[,c(2,1,3)]
colnames(CI.1_5)[1:3]<-c("Median","LowerCI", "UpperCI")
CI.1_5
## nlcm with negatively skewed 5,1 hyperprior
smplr.5_1<- lcmCR(FSW, tabular = TRUE, K = 10, a_alpha = 5,
            b_alpha = 1, seed = 123, buffer_size = 10000,
            thinning = 100)
post.5_1 <- lcmCR_PostSampl(smplr.5_1, burnin = 100000,
                  samples = 10000, thinning = 100,
                  output = FALSE)
CI.5_1 <- quantile(post.5_1, c(0.025, 0.5, 0.975))
CI.5_1 <- as.data.frame(t(CI.5_1))
CI.5_1 <- CI.5_1[,c(2,1,3)]
colnames(CI.5_1)[1:3]<-c("Median","LowerCI", "UpperCI")
CI.5_1

post <- rbind(data.frame(Prior = rep("Jeffreys", length(post.Jeff)), Size = post.Jeff),
         data.frame(Prior = rep("Uniform", length(post.unif)), Size = post.unif),
         data.frame(Prior = rep("Beta(5,5)", length(post.5_5)), Size = post.5_5),
         data.frame(Prior = rep("Beta(1,5)", length(post.1_5)), Size = post.1_5),
         data.frame(Prior = rep("Beta(5,1)", length(post.5_1)), Size = post.5_1))
post$Prior <- as.factor(post$Prior)
png(file = file.path(basepath, "Prior_sensitivity.png"))
bwplot(Size ~Prior, data = post, layout = c(1,1),
    ylab = "Population size", xlab = "Stick-breaking hyperprior")
dev.off()
ests <- rbind(data.frame(Prior = "Jeffreys", Size = CI.Jeff[1],
                Lower = CI.Jeff[2], Upper = CI.Jeff[3]),
         data.frame(Prior = "Uniform", Size = CI.unif[1],
                Lower = CI.unif[2], Upper = CI.unif[3]),
         data.frame(Prior = "Beta(5,5)", Size = CI.5_5[1],
                Lower = CI.5_5[2], Upper = CI.5_5[3]),
         data.frame(Prior = "Beta(1,5)", Size = CI.1_5[1],
                Lower = CI.1_5[2], Upper = CI.1_5[3]),
         data.frame(Prior = "Beta(5,1)", Size = CI.5_1[1],
                Lower = CI.5_1[2], Upper = CI.5_1[3])
         )
write.csv(ests, file = file.path(basepath, "Posterior_summary_by_hyperprior.csv"))


############################ END OF THE PROGRAM ##################################
```

R Codes used for data analysis: Bayesian Model Averaging (MBA)

## Input Data

```
####################################################################################
        df <- structure(list(Capture_1 = c(1L, 0L, 0L, 1L, 1L, 0L, 1L), Capture_2 = c(0L,
        1L, 0L, 1L, 0L, 1L, 1L), Capture_3 = c(0L, 0L, 1L, 0L, 1L, 1L,
        1L), Total_counts = c(1408L, 1471L, 1529L, 169L, 210L, 65L,
        61L)), row.names = c(NA, -7L), class = “data.frame”)
        getData <- function(disag=FALSE){
          if(disag && “Aggregate”== “Aggregate”)
           df <- disaggregate(df[-length(df)],df[[length(df)]])
          df
        }
```

## Log-linear Analysis Report

```
####################################################################################
        library(Rcapture)
        library(shinyrecap)
```

# Model Comparison

```
          logli <- closedp(getData(TRUE), dfreq = FALSE)
          normTFit <- try(closedpCI.t(getData(TRUE), m="Mth",h="Normal"))
          normFit <- try(closedpCI.t(getData(TRUE), m="Mh",h="Normal"))
          results <- as.data.frame(logli$results[1:10,-c(3,4,7)])
          colnames(results)[1] <- "Population Size"
          if(!inherits(normTFit, "try-error")){
           results <- rbind(results, normFit$results[,c(1,2,7,8)])
           row.names(results)[11] <- "Mth Normal"
          }
          if(!inherits(normFit, "try-error")){
           results <- rbind(results[1:6,], normFit$results[,c(1,2,7,8)],results[7:nrow(results),])
           row.names(results)[7] <- "Mh Normal"
          }
```

# Model Results for Mth Poisson

```
          agg <-  "Aggregate"  == "Aggregate"
          ci <- closedpCI.t(getData(),
                      dfreq = agg,
                      m =  "Mth" ,
                      h =  "Poisson" )
          if( "Poisson"  == "Normal"){
           ci <- ci$results[c(1,3,4)]
          }else{
           ci <- ci$CI[1:3]
          }
          names(ci)<-
           c("Population Size", "Lower 95%", "Upper 95%")
          print(round(ci))
```

# Log-linear Descriptives

```
        freqstat <- descriptive(getData(TRUE), dfreq = FALSE)
        print(freqstat)

        plot(freqstat)
```

## Bayesian Model Averaging

```
######################################################################################
      if(!exists("input")) input <- list()
      input$DataType <- "Aggregate"
      input$dgaPriorType <- "lnorm"
      input$dgaPriorMedian <- 13714L
      input$dgaPriorDelta <- 0.125
      input$dgaNMax <- 50000L
      input$dgaPrior90 <- 23495L
      input$dgaSaturated <- FALSE
      library(dga)
      library(ggplot2)
      library(shinyrecap)
```

#Prior Distribution

```
        if( !is.null(getData())){
          dat <- getData()
          if (input$DataType == "Aggregate"){
            nobs <- sum(dat[[length(dat)]])
            ncap <- ncol(dat) - 1
          }else{
            nobs <- nrow(dat)
            ncap <- ncol(dat)
          }
        }else
          return(NULL)
        if(input$dgaPriorType == "lnorm"){

          mu <- log(input$dgaPriorMedian)
          ssd <- (log(input$dgaPrior90) - mu) / qnorm(.9)
          x <- 0:(input$dgaNMax - nobs) + nobs
          values <- dlnorm(x,mu,ssd)
        }else{
          x <- 1:(input$dgaNMax - nobs) + nobs
          values <- 1 / (1:(input$dgaNMax - nobs))
        }
        values <- values / sum(values)
        prior <- list(x=x, values=values)
        priorDist <- function() prior
        dgaPriorType <- input$dgaPriorType
        x <- prior$x
        values <- prior$values
```

```
if(dgaPriorType == "lnorm"){
  titl <- "Log-normal Prior"
}else{
  titl <- "Non-informative Prior (p(x) ~ 1/ (Population Size - Sample Size))"
}
lower90 <- x[min(which(cumsum(values) >= .1))]
upper90 <- x[min(which(cumsum(values) >= .9))]
p <- ggplot() +
  geom_line(aes(x=x,y=values)) +
  geom_vline(xintercept = lower90, color="red") +
  geom_vline(xintercept = upper90, color="red") +
  xlab("Population Size (red lines = 10th and 90th percentiles)") +
  ylab("Prior Probability") +
  ggtitle(titl) +
  theme_bw() +
  xlim(c(0,max(x)))
print(p)

p <- ggplot() +
    geom_line(aes(x=x,y=cumsum(values))) +
    xlab("Population Size") +
    ylab("Prior Cumulative Probability") +
    ggtitle(titl) +
    theme_bw() +
    xlim(c(0,max(x)))
print(p)
```

#Posterior Distribution

```
dat <- getData()
if (input$DataType == "Aggregate") {
  dat <- disaggregate(dat[,-ncol(dat)], dat[[ncol(dat)]])
}
if(ncol(dat) == 3){
  data(graphs3)
  graphs <- graphs3
}else if(ncol(dat) == 4){
  data(graphs4)
  graphs <- graphs4
}else{
  data(graphs5)
  graphs <- graphs5
}
nobs <- nrow(dat)
rec <- make.strata(dat, locations=rep("a",nrow(dat)))$overlap.counts
rec <- array(rec, dim=rep(2, ncol(dat)))

mu <- log(input$dgaPriorMedian)
ssd <- (log(input$dgaPrior90) - mu) / qnorm(.9)
```

```
nmax <- input$dgaNMax - nobs
delta <- input$dgaPriorDelta
prior <- priorDist()
  x <- prior$x
  post <- bma.cr(rec,
                 delta=delta,
                 Nmissing=x - nobs,
                 logprior = log(prior$values),
                 graphs = graphs)
dga <- list(prior=prior, post=post)
    post <- dga$post
if(!input$dgaSaturated){
  post <- post[-nrow(post), , drop=FALSE]
}
postN <- colSums(post)
postN <- postN / sum(postN)
x <- dga$prior$x
mn <- sum(x * postN)
med <- x[which(cumsum(postN) > .5)[1]]
```

# HDI

```
opt <- optimize(
  function(cut){
    abs(.05 - sum(postN*(postN <= cut)))
  },
  interval = c(0,max(postN))
)
inInterval <- which(postN > opt$minimum)
lower <- x[inInterval[1]]
upper <- x[inInterval[length(inInterval)]]

#lower <- x[which(cumsum(postN) > .025)[1]]
#upper <- x[which(cumsum(postN) > .975)[1]]
result <- data.frame(mn, med, lower, upper)
names(result) <- c("Mean","Median","95% Lower","95% Upper")
result %>% knitr::kable(digits=0)

postN <- colSums(post)
postN <- postN / sum(postN)
ind <- cumsum(postN)  < .995
plotPosteriorN(post[,ind], x[ind])
```

# BMA Individual Model Summaries

```
if(!input$dgaSaturated){
  graphs <- graphs[-length(graphs)]
}
mp <- rowSums(post)
means <- apply(post, 1, function(p){
  p <- p / sum(p)
  sum(p * x)
})
means <- as.integer(round(means))
mp <- mp / sum(mp)
mp <- round(mp * 100, 3)

data.frame(Interaction=formatGraphs(graphs),
        `Posterior Probability (%)` = mp,
        `Expected Pop. Size` = means,
        check.names=FALSE) %>% knitr::kable()
```

## Bayesian Latent Class

```
###############################################################################
          library(LCMCR)
          library(shinyrecap)
          dat <- getData()
          if ("Aggregate" == "Aggregate") {
            dat <- disaggregate(dat[,-ncol(dat)], dat[[ncol(dat)]])
          }
          input <- list(lcmcrShape = 0.25, lcmcrScale = 0.25, K = 10L, lcmcrThinning = 100L,
          lcmcrSamples = 1000L, lcmcrBurnin = 100000L)
          K <- input$lcmcrK
          shape <- input$lcmcrShape
          invScale <- input$lcmcrScale
          thinning <- input$lcmcrThinning
          samples <- input$lcmcrSamples
          burnin <- input$lcmcrBurnin
          d2 <- as.data.frame(lapply(dat, as.factor))
          sampler <- lcmCR(d2, tabular = FALSE, K = K, a_alpha = shape,
                       b_alpha = invScale, seed = "auto", buffer_size = samples*thinning + burnin + 1,
                       thinning = thinning)
          post <- lcmcrSample(sampler, burnin = burnin,
                           samples = samples, thinning = thinning,
                           output = FALSE, nMonitorBreaks=100, monitorFunc = func)
          result <- list(N=post)
          resultVal <- function() result
```

# Posterior Distribution

```
post <- resultVal()$N
quant <- quantile(post, c(0.50, .025, 0.975))
hdint <- HDInterval::hdi(post)
result1 <- data.frame(mean(post), quant[1], hdint[1], hdint[2])
names(result1) <- c("Mean","Median","95% Lower","95% Upper")
result1 %>% knitr::kable(digits=0)

hist(post, breaks=50)
```

# Trace Plot

```
  ess <- effectiveSize(post)

  plot(post,
    xlab="Sample #",
    ylab="Population Size",
    main="Trace Plot")
# Pairwise Analysis
        library(shinyrecap)
        library(CARE1)
        dat <- getData()
        if ("Aggregate" == "Aggregate") {
         dat <- disaggregate(dat[,-ncol(dat)], dat[[ncol(dat)]])
        }
        result3 <- estN.pair(as.record(dat))
        result3 <- result3[,-2]
        colnames(result3)<- c("Population Size", "se", "95% CI Lower","95% CI Upper")
        result3 %>% knitr::kable(digits=0)
############################ END OF THE PROGRAM ##################################
```

Privatize Network Sampling Analysis code

```
# Install packages

        install.packages('RDS')
        install.packages('pnspop')

# Import libraries
        library(RDS)
        library(pnspop)

# Load data
        data(fsw2023)
# Run plot on ”rds” to view the recruitment graph.

        Rds <-as.rds.data.frame(

                fsw2023,
                id = “subject”,
                recruiter.id = “recruiter”,
                network.size = “degree”
        )
        Plot (rds)

## Cross Network Analysis

        Cross_tree_pse(
        Subject = fsw2023$subject,
        Recruiter = fsw2023$recruiter,
        Subject_hash = fsw2023$subject_hash,
        Degree = fsw2023$degree,
        Nbrs = fsw2023[c(“friend_hash1” , “friend_hash2” , “friend_hash3” , “friend_hash4”
        ,“friend_hash5” , “friend_hash6” , “friend_hash7” , “friend_hash8” , “friend_hash9”,
        “friend_hash10”)],

        Method = “network”
        )
## Cross Sample Analysis

Cross_tree_pse(
        Subject = fsw2023$subject,
        Recruiter = fsw2023$recruiter,
        Subject_hash = fsw2023$subject_hash,
        Degree = fsw2023$degree,

         nbrs = fsw2023[c(“friend_hash1” , “friend_hash2” , “friend_hash3” , “friend_hash4”
,“friend_hash5” , “friend_hash6” , “friend_hash7” , “friend_hash8” , “friend_hash9”, “friend_hash10”)],

        Method = “sample”
        )
```

```
## Getting Confidence Intervals

boostrap_pse(

        Cross_tree_pse(
        Subject = fsw2023$subject,
        Recruiter = fsw2023$recruiter,
        Subject_hash = fsw2023$subject_hash,
        Degree = fsw2023$degree,
        Nbrs = fsw2023[c(“friend_hash1” , “friend_hash2” , “friend_hash3” , “friend_hash4”
        ,“friend_hash5” , “friend_hash6” , “friend_hash7” , “friend_hash8” , “friend_hash9”,
        “friend_hash10”)],

        Method = “network”
        rho = .0001
        n_boostrap = 5000
        )

boostrap_pse(

        Cross_tree_pse(
        Subject = fsw2023$subject,
        Recruiter = fsw2023$recruiter,
        Subject_hash = fsw2023$subject_hash,
        Degree = fsw2023$degree,
        Nbrs = fsw2023[c(“friend_hash1” , “friend_hash2” , “friend_hash3” , “friend_hash4”
        ,“friend_hash5” , “friend_hash6” , “friend_hash7” , “friend_hash8” , “friend_hash9”,
        “friend_hash10”)],

        Method = “sample”
        rho = .0001
        n_boostrap = 5000
        )

############################ END OF THE PROGRAM ###################################
```

References

1. WHO, *Focus on key populations in national HIV strategic plans in the WHO African Region*. 2018, World Health Organization. Regional Office for Africa.
2. UNAIDS, *Global AIDS Update*. 2023.
3. UNAIDS, *Global AIDS Update* 2022.
4. Mutagoma, M., et al., *High HIV prevalence and associated risk factors among female sex workers in Rwanda.* International journal of STD & AIDS, 2017. **28**(11): p. 1082-1089.
5. Mutagoma, M., et al., *Syphilis and HIV prevalence and associated factors to their co-infection, hepatitis B and hepatitis C viruses prevalence among female sex workers in Rwanda.* BMC infectious diseases, 2017. **17**(1): p. 1-9.
6. Nsanzimana, S., et al., *Prevalence and incidence of HIV among female sex workers and their clients: modelling the potential effects of intervention in Rwanda.* BMJ global health, 2020. **5**(8): p. e002300.
7. Mutagoma, M., et al., *Sexual risk behaviors and practices of female sex workers in Rwanda in over a decade, 2006–2015.* International journal of STD & AIDS, 2018. **29**(13): p. 1316-1323.
8. Mutagoma, M., et al., *Syphilis and HIV prevalence and associated factors to their co-infection, hepatitis B and hepatitis C viruses prevalence among female sex workers in Rwanda.* BMC Infect Dis, 2017. **17**(1): p. 525.
9. Rwanda Biomedical Center/Institute of HIV/AIDS, D.P., *Integrated Bio-Behavioral Surveillance Survey Among Female Sex Workers in Rwanda.* 2019.
10. Vadivoo, S., et al., *Appropriateness and execution challenges of three formal size estimation methods for high-risk populations in India.* Aids, 2008. **22**: p. S137-S148.
11. Des Jarlais, D., et al., *Using dual capture/recapture studies to estimate the population size of persons who inject drugs (PWID) in the city of Hai Phong, Vietnam.* Drug and Alcohol Dependence, 2018. **185**: p. 106-111.
12. Vuylsteke, B., et al., *Estimating the number of female sex workers in Côte d'Ivoire: results and lessons learned.* Tropical Medicine & International Health, 2017. **22**(9): p. 1112-1118.
13. Hook, E.B. and R.R. Regal, *Capture-recapture methods in epidemiology: methods and limitations.* Epidemiologic reviews, 1995. **17**(2): p. 243-264.
14. Chao, A., et al., *The applications of capture-recapture models to epidemiological data.* Statistics in medicine, 2001. **20**(20): p. 3123-3157.
15. Johnston, L.G., et al., *Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations.* Sexually transmitted diseases, 2013. **40**(4): p. 304-310.
16. Maltiel, R., et al., *Estimating population size using the network scale up method.* The annals of applied statistics, 2015. **9**(3): p. 1247.
17. Kim, B.J. and M.S. Handcock, *Population size estimation using multiple respondent-driven sampling surveys.* Journal of Survey Statistics and Methodology, 2021. **9**(1): p. 94-120.
18. Van Hest, N., *Capture-recapture methods in surveillance of tuberculosis and other infectious diseases*. 2007.
19. Doshi, R.H., et al., *Estimating the size of key populations in Kampala, Uganda: 3-source capture-recapture study.* JMIR public health and surveillance, 2019. **5**(3): p. e12118.

20. Baluku, M. and T. Wamala, *When and how do individuals transition from regular drug use to injection drug use in Uganda? Findings from a rapid assessment.* Harm Reduction Journal, 2019. **16**(1): p. 1-8.
21. Apodaca, K., et al., *Capture-recapture among men who have sex with men and among female sex workers in 11 towns in Uganda.* JMIR public health and surveillance, 2019. **5**(2): p. e12316.
22. Okiria, A.G., et al., *Novel approaches for estimating female sex worker population size in conflict-affected South Sudan.* JMIR public health and surveillance, 2019. **5**(1): p. e11576.
23. Beyrer, C., et al., *Global epidemiology of HIV infection in men who have sex with men.* the Lancet, 2012. **380**(9839): p. 367-377.
24. Sandfort, T.G., et al., *HIV testing and the HIV care continuum among sub-Saharan African men who have sex with men and transgender women screened for participation in HPTN 075.* PLoS One, 2019. **14**(5): p. e0217501.
25. HIV/AIDS, T.J.U.N.P.o., *UNAIDS Global AIDS Update* 2022.
26. Organization, W.H., *UNAIDS. Recommended Population Size Estimates of Men Who Have Sex with Men.* World Health Organization, 2020.
27. Remera, E., et al., *HIV and hepatitis B, C co-infection and correlates of HIV infection among men who have sex with men in Rwanda, 2021: a respondent-driven sampling, cross-sectional study.* BMC Infectious Diseases, 2024. **24**(1): p. 347.
28. Health, R.o.R.-M.o., *Rwanda HIV and AIDS National Strategic Plan (NSP 2018 - 2024)*. 2018.
29. Baral, S., et al., *Elevated risk for HIV infection among men who have sex with men in low-and middle-income countries 2000–2006: a systematic review.* PLoS medicine, 2007. **4**(12): p. e339.
30. Beyrer, C., et al., *The global response to HIV in men who have sex with men.* The Lancet, 2016. **388**(10040): p. 198-206.
31. Fanzana, B. and E. Srunv, *A venue-based method for sampling hard-to-reach populations.* Public Health Rep, 2001. **116**: p. 216-22.
32. Karon, J.M. and C. Wejnert, *Statistical methods for the analysis of time–location sampling data.* Journal of Urban Health, 2012. **89**(3): p. 565-586.
33. McCreesh, N., et al., *Evaluation of respondent-driven sampling.* Epidemiology, 2012: p. 138-147.
34. Medhi, G.K., et al., *Size estimation of injecting drug users (IDU) using multiplier method in five districts of India.* Substance abuse treatment, prevention, and policy, 2012. **7**(1): p. 1-5.
35. Bernard, H.R., et al., *Counting hard-to-count populations: the network scale-up method for public health.* Sexually transmitted infections, 2010. **86**(Suppl 2): p. ii11-ii15.
36. Johnston, L.G., et al., *Estimating the size of hidden populations using respondent-driven sampling data: case examples from Morocco.* Epidemiology (Cambridge, Mass.), 2015. **26**(6): p. 846.
37. Wesson, P.D., *If you are not counted, you don't count: Estimating the size of hidden populations*. 2016: University of California, Berkeley.
38. Pollock, K.H., *A Capture-Recapture Design Robust to Unequal Probability of Capture.* The Journal of Wildlife Management, 1982. **46**(3): p. 752-757.
39. Tilling, K., *Capture-recapture methods—useful or misleading?* International Journal of Epidemiology, 2001. **30**(1): p. 12-14.
40. Böhning, D., J. Bunge, and P.G. Heijden, *Capture-recapture methods for the social and medical sciences*. 2018: CRC Press Boca Raton.
41. Buckland, S.T., I.B.J. Goudie, and D.L. Borchers, *Wildlife population assessment: past developments and future directions.* Biometrics, 2000. **56**(1): p. 1-12.
42. Chao, A., H.Y. Pan, and S.C. Chiang, *The Petersen–Lincoln Estimator and its extension to estimate the size of a shared population.* Biometrical Journal: Journal of Mathematical Methods in Biosciences, 2008. **50**(6): p. 957-970.

43. Neugebauer, R. and J. Wittes, *Voluntary and involuntary capture-recapture samples--problems in the estimation of hidden and elusive populations.* American journal of public health, 1994. **84**(7): p. 1068-1069.
44. van Hest, R., *Capture recapture Methods in Surveillance of Tuberculosis and Other Infectious Diseases*. 2007, the Netherlands.
45. Mutagoma, M., et al., *Estimation of the size of the female sex worker population in Rwanda using three different methods.* International Journal of STD & AIDS, 2015. **26**(11): p. 810-814.
46. Rwanda Biomedical Center/Institute of HIV/AIDS, D.P., et al., *Estimating the size of populations through a household survey*. 2012, RBC/IHDPC, SPF, UNAIDS, and ICF International Calverton, Maryland, USA.
47. Musengimana, G., et al., *Female sex workers population size estimation in Rwanda using a three-source capture− recapture method.* Epidemiology & Infection, 2021. **149**.
48. Fellows, I.E., *Estimating Population Size from a Privatized Network Sample.* Journal of Survey Statistics and Methodology, 2022. **10**(5): p. 1346-1369.
49. Becker, N. and C. Heyde, *Estimating population size from multiple recapture experiments.* Stochastic processes and their applications, 1990. **36**(1): p. 77-83.
50. Gile, K.J. and M.S. Handcock, *7. Respondent-driven sampling: An assessment of current methodology.* Sociological methodology, 2010. **40**(1): p. 285-327.
51. Heckathorn, D.D., *Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations*.* Social Problems, 2014. **44**(2): p. 174-199.
52. Abdul-Quader, A.S., et al., *Implementation and analysis of respondent driven sampling: lessons learned from the field.* Journal of urban health, 2006. **83**(1): p. 1-5.
53. McCreesh, N., et al., *Evaluation of respondent-driven sampling.* Epidemiology (Cambridge, Mass.), 2012. **23**(1): p. 138.
54. White, G.C., *Capture-recapture and removal methods for sampling closed populations*. 1982: Los Alamos National Laboratory.
55. Hartung, C., et al. *Open data kit: tools to build information services for developing regions*. in *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*. 2010.
56. Salganik, M.J., *Variance estimation, design effects, and sample size calculations for respondent-driven sampling.* Journal of Urban Health, 2006. **83**(1): p. 98-112.
57. McIntyre, A.F., et al., *Population size estimation from capture-recapture studies using shinyrecap: design and implementation of a web-based graphical user interface.* JMIR Public Health and Surveillance, 2022. **8**(4): p. e32645.
58. Manrique-Vallier, D., *Bayesian population size estimation using Dirichlet process mixtures.* Biometrics, 2016. **72**(4): p. 1246-1254.
59. Teh, Y.W., *Dirichlet Process.* Encyclopedia of machine learning, 2010. **1063**: p. 280-287.
60. Fienberg, S.E. and D. Manrique-Vallier, *Integrated methodology for multiple systems estimation and record linkage using a missing data formulation.* AStA Advances in Statistical Analysis, 2009. **93**: p. 49-60.
61. Little, R.J. and D.B. Rubin, *Statistical analysis with missing data*. Vol. 793. 2019: John Wiley & Sons.
62. Goodman, L.A., *Exploratory latent structure analysis using both identifiable and unidentifiable models.* Biometrika, 1974. **61**(2): p. 215-231.
63. Dunson, D.B. and C. Xing, *Nonparametric Bayes modeling of multivariate categorical data.* Journal of the American Statistical Association, 2009. **104**(487): p. 1042-1051.
64. Sethuraman, J., *A constructive definition of Dirichlet priors.* Statistica sinica, 1994: p. 639-650.
65. Gelman, A., et al., *Bayesian data analysis*. 1995: Chapman and Hall/CRC.

66. Ishwaran, H. and L.F. James, *Gibbs sampling methods for stick-breaking priors.* Journal of the American statistical Association, 2001. **96**(453): p. 161-173.
67. Basu, S. and N. Ebrahimi, *Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence.* Biometrika, 2001. **88**(1): p. 269-279.
68. Fienberg, S.E., M.S. Johnson, and B.W. Junker, *Classical multilevel and Bayesian approaches to population size estimation using multiple lists.* Journal of the Royal Statistical Society Series A: Statistics in Society, 1999. **162**(3): p. 383-405.
69. Manrique-Vallier, D. and S.E. Fienberg, *Population size estimation using individual level mixture models.* Biometrical Journal: Journal of Mathematical Methods in Biosciences, 2008. **50**(6): p. 1051-1063.
70. Manrique-Vallier, D. and J.P. Reiter, *Bayesian estimation of discrete multivariate latent structure models with structural zeros.* Journal of Computational and Graphical Statistics, 2014. **23**(4): p. 1061-1079.
71. Sibuya, M., I. Yoshimura, and R. Shimizu, *Negative multinomial distribution.* Annals of the Institute of Statistical Mathematics, 1964. **16**(1): p. 409-426.
72. Manrique-Vallier, D., *LCMCR: Bayesian Non-Parametric Latent-Class Capture-Recapture.* R package version 0.4, 2017. **3**.
73. Biggeri, A., et al., *Latent class models for varying catchability and correlation among sources in Capture-Recapture estimation of the size of a human population.* Statistica Applicata, 1999. **11**(3): p. 1-14.
74. Darroch, J.N., S.L. Lauritzen, and T.P. Speed, *Markov fields and log-linear interaction models for contingency tables.* The Annals of Statistics, 1980: p. 522-539.
75. Dawid, A.P. and S.L. Lauritzen, *Hyper Markov laws in the statistical analysis of decomposable graphical models.* The Annals of Statistics, 1993: p. 1272-1317.
76. Golumbic, M.C., *Algorithmic graph theory and perfect graphs*. 2004: Elsevier.
77. Rissanen, J., *A universal prior for integers and estimation by minimum description length.* The Annals of statistics, 1983. **11**(2): p. 416-431.
78. Johndrow, J., K. Lum, and P. Ball, *dga: Capture-Recapture Estimation using Bayesian Model Averaging. R package version 1.2*. 2014.
79. Gile, K.J., *Improved inference for respondent-driven sampling data with application to HIV prevalence estimation.* Journal of the American Statistical Association, 2011. **106**(493): p. 135-146.
80. Khan, B., et al., *One-step estimation of networked population size: Respondent-driven capture-recapture with anonymity.* PloS one, 2018. **13**(4): p. e0195959.
81. Rwanda, N.I.o.S.o., *Fifth Population and Housing Census - 2022*. 2022: https://www.statistics.gov.rw/datasource/fifth-population-and-housing-census-2022.
82. Francisco, P.S., *Behavioral and Biological Assessment and Population size estimation for men who have sex with men (MSM) in Kigali, Rwanda.* 2018.
83. Lieb, S., et al., *Statewide estimation of racial/ethnic populations of men who have sex with men in the US.* Public health reports, 2011. **126**(1): p. 60-72.
84. (WHO), W.H.O., *KEY POPULATIONS STRATEGIC INFORMATION: RECOMMENDED POPULATION SIZE ESTIMATES OF MEN WHO HAVE SEX WITH MEN*. 2020.
85. Organization, W.H., *Key population strategic information: recommended population size estimates of men who have sex with men: technical brief.* 2020.
86. Baral, S., et al., *Population size estimation of gay and bisexual men and other men who have sex with men using social media-based platforms.* JMIR public health and surveillance, 2018. **4**(1): p. e9321.

87. RBC, *HIV Annual Report, Rwanda*. 2022-2023, Rwanda Biomedical Center.

88. Laga, I., et al., *Mapping the number of female sex workers in countries across sub-Saharan Africa.* Proceedings of the National Academy of Sciences, 2023. **120**(2): p. e2200633120.

Discrete-time closed capture-recapture models for hard-to-reach population size estimation: application to key population for HIV prevention in Rwanda, © July 2024

Author:

Elysée TUYISHIME

Supervisors:

Prof. Angela Unna Chukwu

Dr. Ignace Kabano

Institute:

African Center of Excellence in Data Science (ACE-DS),
College of Business and Economics

University of Rwanda, Kigali Rwanda.