The Annals of Applied Statistics
2020, Vol. 14, No. 3, 1207–1241
https://doi.org/10.1214/20-AOAS1327

SIZE ESTIMATION OF KEY POPULATIONS IN THE HIV EPIDEMIC IN ESWATINI USING INCOMPLETE AND MISALIGNED CAPTURE-RECAPTURE DATA

BY ABHIRUP DATTA[1,*], ANDREW PITA[1,†], AMRITA RAO[2,‡], BHEKIE SITHOLE[3], ZANDILE MNISI[4] AND STEFAN BARAL[2,§]

[1]*Department of Biostatistics, Johns Hopkins University,* [*]*abhidatta@jhu.edu;* [†]*apita3@jhmi.edu*
[2]*Department of Epidemiology, Johns Hopkins University,* [‡]*arao24@jhu.edu;* [§]*sbaral@jhu.edu*
[3]*FHI360 LINKAGES Program,* *bhekie.sithole@gmail.com*
[4]*Research Department, Ministry of Health,* *zandimnisi@gmail.com*

In 2020, our understanding of the distributions of HIV risks in the most burdened settings, including eSwatini, remains limited. In part, this is driven by the limited availability of the size and burden of the populations at the greatest risk for HIV. Given pervasive social and healthcare stigmas, the size estimations of these populations often rely on the multiplier method—a variant of the capture-recapture approach where the first survey is replaced by an enumeration of population members who used some service or attended an event. To characterize the distributions of marginalized communities in eSwatini, multiple data sources are available at each region for the multiplier method. Current practices in such circumstances produce multiple population size estimates at each region ignoring the correlation among these estimates. We recast the multiple multiplier method as a special case of capture-recapture problem with incomplete data and propose a fully model based approach for size estimation using multiple capture-recapture data with arbitrary pattern of incompleteness. We use a data augmentation scheme that allows us to model the correlations in the data and produce a unified estimate of population size per region. A hierarchical model ties together the models for multiple regions, allowing us to borrow strength across the regions and enabling extrapolation to areas without data. In eSwatini we also encounter data misalignment where counts from some of the data sources are not available for each region but as an aggregate over few regions. We propose a solution to the general misalignment problem which considers data-source-specific patterns of misalignment. We use simulation studies to demonstrate the accurate inferential capabilities of our Bayesian multiplier method. This approach is then used to produce uncertainty-quantified population size estimates of key populations in eSwatini. Lastly, we propose a Bayesian nonparametric extension for incomplete capture-recapture that allows nonindependent data sources.

1. Introduction. In the broadly generalized HIV epidemics observed across Southern Africa, there has often been an assumption of the homogeneity of HIV-related risks giving rise to the term “general population.” In the last several years there has been a move to better characterize the distributions of HIV risks in generalized epidemics, including the burdens among *key populations* (KP) with well defined risks for the acquisition and transmission of HIV. Accurately characterizing attributable fractions of HIV among key populations at higher risk for HIV necessitates estimating both the burden of HIV and key population size. This manuscript focuses on regional size estimation with proper uncertainty quantification of MSM (men having sex with men) and FSW (female sex workers) populations in eSwatini.

Received March 2019; revised February 2020.
Key words and phrases. Bayesian, capture-recapture, epidemiology, HIV, misalignment, multiplier method.

In settings where key populations for HIV surveillance like MSM and FSW face stigma, discrimination and often criminalization, data collection endeavors pose challenges, including risks of unintended disclosure of identity, imprisonment and violence. Carefully designed surveys that take into account safety of the participants and communities and the legal framework and that also incentivize recruitment of participants are often conducted (respondent driven sampling (RDS), Heckathorn (1997)). Sampling methods also often take advantage of social networks of participants or the fact that certain populations may congregate at common meeting places (Priorities for Local AIDS Control Efforts (PLACE), Weir et al. (2005)). Accounting for sampling design is critical to improve size estimation for key-populations Edwards et al. (2018).

Population size estimation using survey data is one of the longest studied problems in statistics and demography. A popular tool for size estimation is the capture-recapture (also referred to as the mark-and-recapture) method where two surveys are conducted in the same population. In the context of human populations, the second survey, which is conducted after the first one is completed, usually asks the participants about their participation in the first one. Estimates of population size can be derived using the total count of participants in each of the two surveys and the overlap size between the two surveys (Lincoln et al. (1930), Petersen (1896)).

Conducting multiple surveys are often not feasible in contexts of key populations given the budgetary constraints and the aforementioned challenges of surveying such hidden populations. Hence, the standard capture-recapture is often replaced by a related method called *multiplier method* in budget-constrained and stigmatized settings. The multiplier method resourcefully uses only a single survey and other nonsurvey sources to estimate the population size (Fearon et al. (2017), WHO (2010)). The multiplier method proceeds by obtaining a count of the number of KP members who were involved in an activity or used a service. Examples of such activities include attending a special event for the specific community, enrolling with a community-based organization (CBO), visiting a clinic for medical services, etc. The event attendance logs, organization registers, patient records, etc., provide a count of the KP members who attended the event or used the service. Subsequently, one survey is conducted and the KP members participating in the survey are asked about their involvement in this past activity or service. So, the total number of survey participants, the total number of KP members who used that service or participated in that activity and the overlap between the survey and this past source are all known. The data now is exactly similar to the standard capture-recapture format, and we can obtain the estimator of population size.

Often, some regions have multiple listings (sources) each of which provides a total count of KP members using the respective services. Subsequently, when the survey is conducted, there is a separate question about the involvement of the participant in each of these activities or sources. If there are S total data sources for a region, then $S - 1$ separate multiplier method estimates are customarily generated by using pairs of data sources—one always being the survey and the other being any one of the $S - 1$ nonsurvey source. This practice is flawed. First, it produces multiple estimates of population size for the same region which is not desirable. Second, the correlation between the multiple estimates for a region, owing to the reliance on the common survey, is hard to estimate directly and is ignored in subsequent usage of these direct estimates. This leads to manifold issues. For example, aggregation of these multiplier method estimates into a unified estimate is often done by taking a simple mean or median of these estimates, and the correlation among the estimates are ignored while generating the confidence bounds of the unified estimate (Holland et al. (2016)). Also, subsequent extrapolation exercises, using these multiple direct estimates along with demographic covariates in a regression setup to predict population sizes in areas with no survey data, ignore this correlation and treat these estimates as independent data points (Datta et al. (2018)).

Akin to the equivalence between the multiplier method and capture-recapture with two sources, multiplier data with $S-1$ nonsurvey sources can be recast as a capture-recapture problem with S sources, for which one can obtain an unified estimate of population size (Darroch (1958)). However, the sufficient statistics for obtaining this estimator are the marginal counts for each of the S sources and the total count of participants (r) who were included in at least one of the sources. In a multiplier method setting, we cannot determine r as we have no participation history of the KP members who were not in the survey but were in any of the other $S-1$ listings, thereby proscribing direct application of the multiple capture-recapture approach.

A second problem that we encounter in the KP-related data for eSwatini is the spatial misalignment among the different data sources. To elaborate, some of the listings counts for the multiplier method were not available for the individual regions where size estimations were completed but were available for a larger geographical region encapsulating more than one of these smaller units. In general, this situation can arise if two or more areas are geographically proximal enough so that many of the KP counts are only available as aggregated over all of them, but, administratively, these units belong to different regions hence requiring separate size estimations, as resource allocation for HIV prevention programs is often divided along administrative lines.

In this manuscript we propose a general method for estimating population size based on incomplete and geographically misaligned multiplier data from numerous sources by recasting the problem in the capture-recapture framework with missing data. We propose a fully model based solution that:

(a) produces an unified estimate of population size for each region by jointly using all multiplier sources available for that region, thereby eliminating the quandary of dealing with multiple size estimates for the same region,

(b) accommodates for arbitrary source-specific geographical misalignment in data collection to produce size estimates at the desired geographical resolution of administrative units,

(c) can be seamlessly embedded in a hierarchical Bayesian setup that allows proper propagation of uncertainty into the final conclusions, simultaneous size estimation in multiple regions by borrowing of strength across regions to improve precision and extrapolation of population size for regions with no survey data.

Our solution involves a data-augmentation approach for the general problem of multiple capture-recapture with arbitrary patterns of incompleteness. We refer to the special case where the incompleteness is due to the use of multiple multiplier sources as the *Bayesian multiplier method*. We develop a novel Gibbs sampler using multivariate hypergeometric distributions effectuating fast and easy implementation. We demonstrate how this problem is related to the problem of sampling from contingency tables (Dobra, Tebaldi and West (2006)). In particular, our problem is equivalent to sampling $2 \times 2 \times \cdots \times 2$ contingency tables with known marginals, and, hence, our approach also offers a Gibbs sampler based solution to that problem.

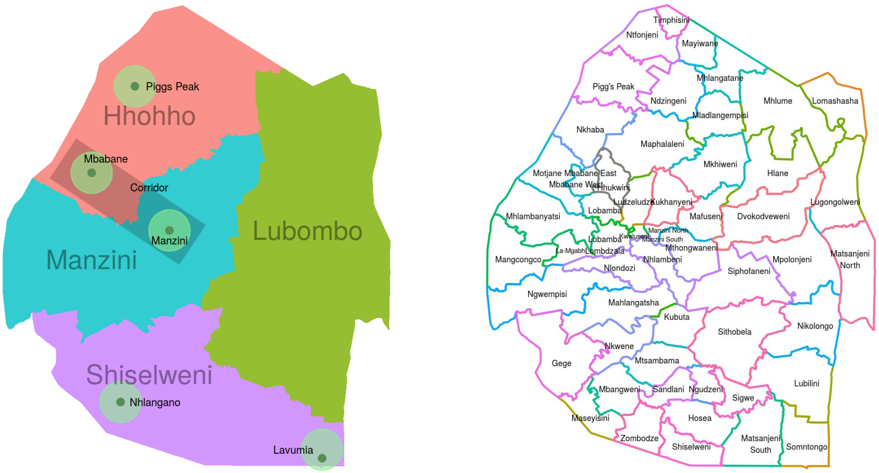
We extend our methodology to accommodate spatial misalignment in population size estimation problems. We propose a solution for the general problem where the misalignment pattern is arbitrary and can be specific to the source of the data. We once again leverage the data-augmentation strategy and formulate the Gibbs sampler using noncentral hypergeometric distribution. Finally, we extend our methodology to relax the assumption of independence data sources central to capture-recapture estimates. We accomplish this using a Bayesian nonparametric latent clustering of the population.

The rest of the manuscript is organized as follows. In Section 2 we present the KP-related data in eSwatini and discuss the analysis goals. Section 3 is dedicated to methods development, beginning with a review of the relevant capture-recapture literature and then describing the models. The Gibbs samplers for incompleteness and misalignment are presented in Section 4. Section 5 presents the data analysis in detail and presents the population size estimates of MSM and FSW for eSwatini. Sensitivity analyses using simulated data are conducted in Section 6. Section 7 presents the extension to nonindependent data sources. Section 8 concludes the manuscript with a discussion of the assumptions used in our analysis and some of the challenges in this area that will guide future research.

2. Key population size estimation in eSwatini.

2.1. *Importance.* eSwatini, a landlocked country in southern Africa with a population of approximately 1.4 million people, has one of the most broadly generalized HIV epidemics in the world. The most recent estimates suggest that more than one quarter of reproductive aged adults are living with HIV (The PHIA Project (2017)). Despite significant improvements over the last decade in testing access and uptake and in the provision of antiretroviral treatment (ART), an estimated 7000 adults were newly infected with HIV in 2017. While the HIV epidemic in eSwatini is broadly generalized, specific key populations have been shown to bear higher burdens of HIV than other similarly aged reproductive aged adults (The PHIA Project (2017), Baral et al. (2013, 2014)). Moreover, while data on both FSW and MSM are limited, studies have also demonstrated limited coverage of HIV prevention, diagnostic and treatment services (Baral et al. (2013, 2014), Berger et al. (2018), Brown et al. (2016), Fielding-Miller et al. (2014), Fonner et al. (2014), Grover et al. (2016), Kennedy et al. (2013), Logie et al. (2018a, 2018b), Risher et al. (2013), Yam et al. (2013)). Both sex work and same-sex practices are criminalized and, as a result, FSW and MSM commonly report social exclusion, stigma, discrimination and violence (Kennedy et al. (2013), Logie et al. (2018a, 2018b), Rao et al. (2017), Risher et al. (2013)). Consequently, there is limited disclosure of status as a sex worker or MSM in the context of HIV surveillance systems and more broadly health systems in eSwatini (Sabin et al. (2016)). Taken together, there are limited data highlighting the HIV prevention and treatment needs of key populations in eSwatini (Kennedy et al. (2013), Logie et al. (2018a, 2018b), Rao et al. (2017), Risher et al. (2013)). Moreover, size estimates among both FSW and MSM in eSwatini remain limited, with one size estimation activity being completed in 2014 among FSW in four locations (Mbabane/Manzini Corridor, Lavumisa, Piggs Peak, Nhlangano) and one in 2014 among MSM in three locations (Mbabane/Manzini Corridor, Piggs Peak, Nhlangano) (Rao et al. (2017)). These estimates are in specific urban areas but say very little about the rest of the country. Better data and methods are needed to estimate the population size of key populations in order to develop adequate programs that are responsive to the needs of those they intend to serve (Sabin et al. (2016)).

2.2. *Data.* eSwatini is divided into four administrative regions—Hhohho, Manzini, Shiselweni and Lubombo (Figure 1(a)), and each region is subdivided into several smaller administrative regions called *Tinkhundlas* (Figure 1(b)). The goal of the analysis was to estimate, with proper uncertainty quantification, the number of MSM and FSW in each of these four regions. Separate surveys were conducted for both of these populations in 2014 using the PLACE method. A modified version of the PLACE method was used to characterize venues where MSM and FSW meet new potential sexual partners (Weir et al. (2005)). Details of study design and recruitment have been described previously in Rao et al. (2017). Briefly, MSM and FSW were recruited using snowball sampling through outreach at mapped hotspot venues. Eligibility to participate in the survey for MSM was that the individual had to report receptive or insertive anal sex with another man in the last 12 months and for FSW was that the individual had to report more than half of their income in the last 12 months came from sex work.



(a) (b)

FIG. 1. *The left figure shows the four regions* (*four colors*) *of eSwatini and the sites of data collection* (*green circles*). *The grey rectangle indicates the Corridor region which encompasses both Mbabane and Manzini*. *The right figure shows the Tinkhundlas* (*smaller administrative divisions*) *within each region*.

The survey participants came primarily from in and around five sites—Piggs Peak, Mbabane, Manzini, Nhlangano and Lavumisa. These sites are highlighted in green in Figure 1(a). Among the sites, Piggs peak and Mbabane belong to the Hhohho region, Manzini is in the Manzini region, Nhlangano and Lavumisa are in the Shiselweni region. Table 1 provides a distribution of the surveyed MSM and FSW in these five sites, while Tables 10 and 11 in the Appendix detail the distribution of exact reported location of the participants which were then mapped to these five sites.

Overall 532 MSM were surveyed among which only one was from Lavumisa and only four were from Lubombo. Since these sample sizes were too low and, additionally, there was no other data in Lubombo, these five cases were excluded from the analysis. In total, 781 FSW were surveyed with only seven from Lubombo. These seven cases were again excluded because of the low sample size and lack of additional data in that region. For the other sites the count of survey participants was substantial, and we also had data from additional sources allowing us to use multiplier method. We describe these sources below:

Unique object identifier (UID): The unique object method is a way to enumerate members of the population of interest. Unique objects are distributed to members of the community for a certain period before the survey, serving as tags or marks. Therefore, the total number of unique objects distributed is known. Subsequently, in the survey participants were asked

TABLE 1
Location distribution of the survey participants

	Location	Region	*Count*		Location	Region	*Count*
MSM	Piggs Peak	Hhohho	57	FSW	Piggs Peak	Hhohho	127
	Mbabane	Hhohho	223		Mbabane	Hhohho	255
	Manzini	Manzini	177		Manzini	Manzini	257
	Nhlangano	Shiselweni	70		Nhlangano	Shiselweni	47
	Lavumisa	Shiselweni	1		Lavumisa	Shiselweni	88
		Lubombo	4			Lubombo	7
		Total	532			Total	781

whether they had received an unique object. This informs about the size of the overlap between the survey participants and the recipients of a unique object. In this case, the unique objects were a specially designed deck of cards.

Coupons (CPN): HIV service coupons were handed out to community members in some of the sites. Coupons serve the same purpose as unique objects in terms of obtaining a count of the coupon recipients, which along with the responses of the survey participants about them receiving coupons allows population size estimation using the multiplier method.

Rainbow night (RNB): Rainbow night was a social event for MSM. The total number of MSM attending the event was available from the event registration logs. The survey later asked the participants if they were attendees in that rainbow night to obtain the overlap size.

FLAS: The Family Life Association of eSwatini is an NGO that offers a variety of clinical and counseling services geared toward sexual health and HIV prevention. The FSW survey participants were asked if they had attended a FLAS mobile health clinic in the month of September.

2.3. *Incompleteness and misalignment*.

Not all of these data sources were available or usable for each site. Figure 2 uses Venn diagrams to demonstrate the data we have for MSM at each site. In Piggs Peak, we have total number of survey participants, the total number of unique object recipients and the overlap between the two. In Nhlangano, we have the total numbers for each of the survey participants, unique object recipients and rainbow night attendees. Since the survey participants were asked both about receiving the unique objects and attending the rainbow night, we also know the overlaps between survey and any of these two sources. However, the overlap between UID and RNB among those who did not participate in the survey is unknown. We will discuss later in Section 3.1 why this missing piece necessitates new methods development for population size estimation. For Mbabane and Manzini, we have three data sources—survey, UID and coupons. The marginal counts for survey and UID are available at both sites. However, Mbabane and Manzini are geographically close and are part of an area commonly referred to as the *Corridor* (Figure 1(a)). Due to this proximity, some data collection efforts take place jointly across the two cities and only provide the total count for the entire corridor instead of separate counts for the two individual cities. For MSM, this was the case for coupons, as the individual coupon numbers for Mbabane

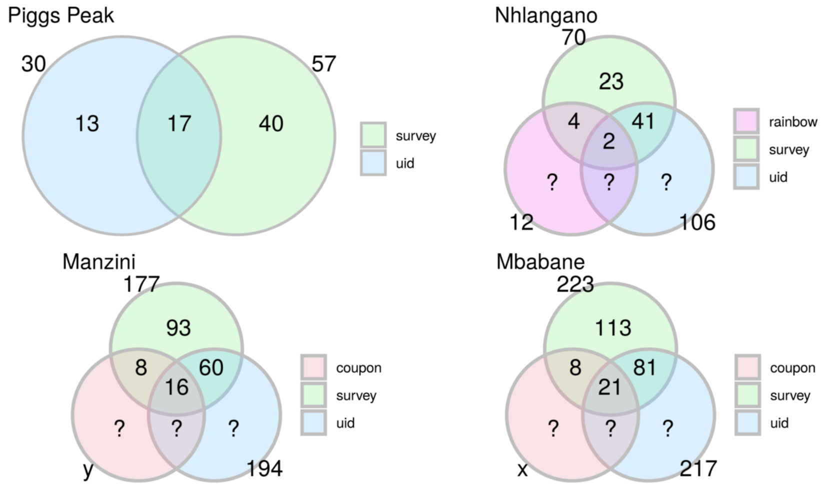


FIG. 2. *MSM counts and overlaps for the various data sources in each site. Note that the marginal total numbers of coupons distributed in Manzini and Mbabane* (y *and* x) *are not known, but the total number of coupons distributed is known such that* $y + x = 106$.

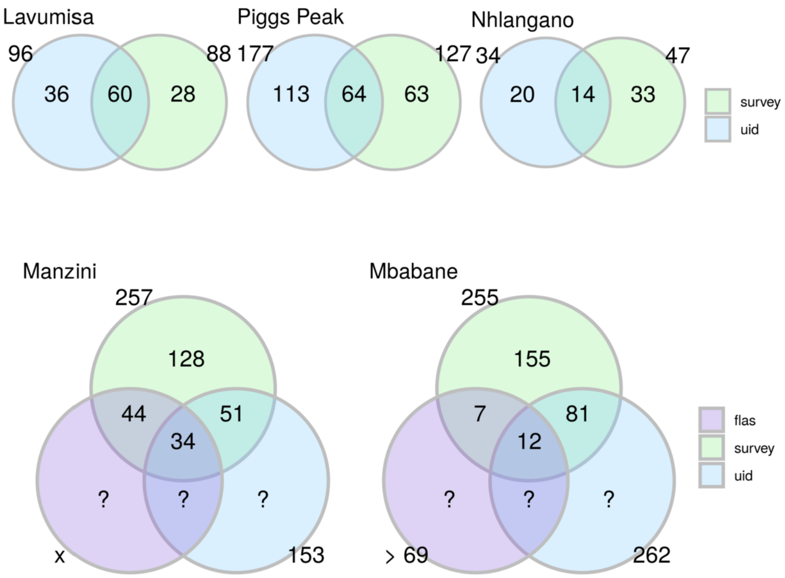


FIG. 3. *FSW counts and overlaps for the various data sources in each site. We know that the total number of FLAS attendees in Manzini and Mbabane is* 186 *and also that at least* 70 *of these were located in Mbabane.*

and Manzini were not known; we only knew that 106 total coupons were distributed in the Corridor, thereby leading to *misalignment*. Finally, like in Nhlangano, for both Mbabane and Manzini all overlap sizes among the survey participants are known, but the overlap between UID and CPN outside of the survey participants are not known.

Turning now to FSW (Figure 3), we see that in each of Piggs Peak, Nhlangano and Lavumisa, we have the number of survey participants, the total number of unique object recipients and the overlap between the two. For Mbabane and Manzini we again have three data sources—survey, UID and attendance at the FLAS mobile clinic. In both Mbabane and Manzini, we know the marginal counts of the survey participants and UID recipients. However, as with the MSM data and coupon counts, we do not know the marginal number of FLAS attendees separately in Mbabane and Manzini, only that 186 total FSW were served between both locations. Out of the 186, it was known that the program HC3 served 70 FSW in Mbabane, and the program PSI served 116 FSW in Mbabane and Manzini combined. This led to the additional constraint that, among the 186 FSW reached by FLAS, at least 70 of them were from Mbabane.

3. Model.

3.1. *Review of capture-recapture methodology.* We first provide a brief overview of the statistical principles underlying the related approaches of capture-recapture and multiplier method for estimating population size. Let N denote the unknown size of the target population (which is a subset of the total population). Let n_1 denote the total number of individuals of the KP included in the first survey or nonsurvey source (for multiplier method), n_2 denote the number of individuals marked in the second source (a survey) and n_{12} denote the size of the overlap between the two surveys. Then, the Lincoln–Petersen estimator (Lincoln et al. (1930), Petersen (1896)) for the total population size is given by

$$\widehat{N} = \widehat{N}_{\text{mult}} = \frac{n_1 n_2}{n_{12}}. \tag{1}$$

The MSM data for Piggs peak and the FSW data for Piggs peak, Nhlangano and Lavumisa all conform to this standard capture-recapture format. The fundamental assumption driving the derivation of this estimate is that the proportion of individuals in the target population, who were tagged in the first survey, are assumed to be same as the proportion of individuals in the second survey who were also tagged in the first survey. This holds if inclusion of an individual in the two surveys is assumed to be independent of each other. Under this assumption, a model based alternative to estimate the population size proceeds by denoting p_i, $i = 1, 2$, the homogeneous inclusion probability of any individual in the ith survey. Then, we can perceive of $n_1 - n_{12}, n_2 - n_{12}$ and n_{12} as a partial realization from Multinomial distribution, that is,

$$\begin{pmatrix} n_1 - n_{12} \\ n_2 - n_{12} \\ n_{12} \\ N - r \end{pmatrix} \Bigg| N, p_1, p_2 \sim \text{Multinomial}\left(N, \begin{pmatrix} p_1(1-p_2) \\ p_2(1-p_1) \\ p_1 p_2 \\ (1-p_1)(1-p_2) \end{pmatrix} \right), \tag{2}$$

where $r = n_1 + n_2 - n_{12}$ denote the total number of individuals marked in at least one of the surveys. Darroch (1958) demonstrated that the MLE estimate of N from this multinomial model coincides with the Lincoln–Petersen estimator presented in (1). However, the model based approach is substantially more versatile, as it allows easy extension to multiple surveys (Darroch (1958, 1959)) as well as enable Bayesian formulation via specifying priors (Castledine (1981), George and Robert (1992)). The Bayesian formulation is particularly useful in applications, such as ours, involving stratified size estimation where hierarchical modeling enables borrowing of strength across multiple regions or subgroups for parameter estimation.

Observe that the likelihood for (2) can be expressed as

$$\ell(N, p_1, p_2 \mid n_1, n_2, n_{12}) \propto \frac{N!}{(N-r)!} p_1^{n_1} (1-p_1)^{N-n_1} p_2^{n_2} (1-p_2)^{N-n_2}.$$

It is easy to conclude from the likelihood above that $(n_1, n_2, r)'$ is sufficient for estimating N, p_1 and p_2. This remains true even when we extend to the case where there are S surveys (Darroch (1958)). To elucidate, let $\mathcal{S} = \{1, 2, \dots, S\}$, p_i, $i \in \mathcal{S}$ denote the survey-specific inclusion probabilities, n_i denote the number of individuals included in the ith survey and r denote the total number of individuals marked in at least one survey. Also, for any two disjoint subsets A and B of $\mathcal{S}$, let $n_{A,B}$ denote the number of individuals marked in the surveys indexed by the members of A and not included in the surveys indexed by B. More formally, for the jth individual in the population $\mathcal{I}(j)$ denote the $S \times 1$ vector of inclusion history, that is, $\mathcal{I}(j)_i = 1$ if the jth individual is included in the ith survey and is 0 otherwise. Then, for $A, B \subset \mathcal{S}$, $A \cap B = \{\}$ we have

$$n_{A,B} = \sum_{j=1}^{N} \prod_{i \in A} \mathcal{I}(j)_i \prod_{i' \in B} (1 - \mathcal{I}(j)_{i'}). \tag{3}$$

Once again, assuming that all the S data sources are independent, the corresponding inclusion probability is given by

$$u_{A,B} = P(\mathcal{I}(j)_i = 1 \; \forall i \in A, \mathcal{I}(j)_{i'} = 0 \; \forall i' \in B) = \prod_{i \in A} p_i \prod_{i \in B} (1 - p_i). \tag{4}$$

The distribution for the counts of the finest partitions, specifying the complete inclusion history for the S listings, are given by

$$\{n_{A,A^c}\}_{A \subset \mathcal{S}} \sim \text{Multinomial}(N, \{u_{A,A^c}\}_{A \in \mathcal{S}}).$$

The likelihood for this can be simplified to

$$\ell(N, p_1, p_2, \dots, p_S \mid \{n_{A,A^c}\}) \propto \frac{N!}{(N-r)!} \prod_{i=1}^{S} p_i^{n_i} (1-p_i)^{N-n_i}. \tag{5}$$

This clearly demonstrates the sufficiency of $(r, n_1, n_2, \dots, n_S)'$ and that we do not need the counts n_{A,A^c} for each of the 2^S individual partitions A.

Castledine (1981) explored Bayesian inference for capture-recapture using priors for the sample size N and inclusion probabilities p_i's. George and Robert (1992) developed a Gibbs sampler for the model in (5) using conjugate Beta priors for p_i and Poisson prior or Jeffrey's prior ($\mathrm{pr}(N) \propto 1/N$) for N. When N itself is only the size of subpopulation embedded in a greater population of size P, we can also use a binomial(P, ϕ) prior for N which also leads to conjugacy. The binomial prior was used in Bao, Raftery and Reddy (2015), but the method only considered one capture-recapture or multiplier method estimate along with other data sources informing about population size. Extensions to more than one multiplier estimates with incompleteness and misalignment was not considered.

Using Beta(a_i, b_i) prior for p_i and a Jeffrey's prior for N, the Gibbs sampler steps are given by

$$\begin{aligned} p_i \mid \cdot &\sim \mathrm{Beta}(a_i + n_i, b_i + N - n_i), \\ N \mid \cdot &\sim r + \mathrm{NB}\left(r, 1 - \prod_{i=1}^{S}(1 - p_i)\right), \end{aligned} \tag{6}$$

where the $X \mid \cdot$ notation denotes the condition distribution of X, given all the other random variables, and NB(r, p) denote a negative binomial distribution with pdf

$$p(k) = \binom{r + k - 1}{r} p^r (1 - p)^k, \quad k = 0, 1, \dots.$$

In our application, for regions that have only two sources of data, one of which is the survey which provides the information about the overlap between the survey and the first source, one can calculate r and, hence, can directly implement this model.

3.2. *Multiplier method and incomplete data.* Complications arise when r is not observed or cannot be evaluated from the survey records. This can be due to several reasons, like when surveys were conducted independent of the knowledge of past surveys. In such cases we only have the marginal counts n_i's and cannot recover r. Another possibility is that all but one of the listings are not surveys but enrollment of participation in some events or services. This is a common occurrence in estimation of key population using the multiplier method where the survey participants are asked about their enrollment or participation in all of the $S - 1$ previous listings. Hence, we obtain the full participation history of every individual in the survey (assuming complete cases). However, no history is available for the individuals not surveyed but marked in any of the other $S - 1$ listings and, consequently, r is not available.

In our application this situation is exemplified for the MSM data in Nhlangano where we have three sources of information—unique objects that were handed out, participation in a rainbow night and, finally, the survey where participants were asked whether they had received an unique object in the past or had attended the rainbow night. As is evident from Figure 2, this information is insufficient to recover r—the total number of MSM who were involved in at least one of the three activities. Similar instances are encountered for both MSM and FSW estimation in Mbabane and Manzini; however, data for those regions, in addition to being incomplete, are also spatially misaligned, and we will address this in Section 3.3.

Without loss of generality we assume that the survey is the Sth listing. When generating direct estimates of population size using surveys and $S - 1$ other sources using the multiplier method, a common theme is to generating one estimate for a pair of sources: one of which is the survey, and the second is any of the $S - 1$ sources. Since the overlap between the survey and this chosen source is available from the data, one can use the Lincoln–Petersen estimator to obtain a direct estimate of the population size. Repeating this for each of the $S-1$ sources, one can obtain $S-1$ direct estimates for each region. Subsequently, an unified estimate is obtained by taking an average over the $S-1$ estimates. We refer to this estimator as the *average Lincoln–Petersen estimator* or *average multiplier-method* and, using the notation introduced in (3), it is given by

$$\widehat{N}_{\text{mult.avg}} = \frac{1}{S-1}\sum_{i=1}^{S-1}\frac{n_{i,\{\}}n_{S,\{\}}}{n_{\{i,S\},\{\}}}. \tag{7}$$

While using this estimator is popular, owing to its ease-of-use, it fails to incorporate the dependence among the estimates stemming from all of them using the survey data as one of the two sources. To elaborate, in (7) all the $S-1$ estimators use the same $n_{S,\{\}}$, and the quantities $n_{\{i,S\},\{\}}$ are correlated for different i's. Circumventing these complex dependencies among the estimates has repercussions, as we demonstrate in Section 6.2. We here propose an estimation based on the joint likelihood for such incomplete multiple capture-recapture or multiplier data. Our approach incorporates these dependencies and produces an unified direct estimate.

Let $\mathcal{S}^* = \{1, 2, \dots, S-1\}$. Using the notation defined in Section 3.1, the observed data $\mathcal{D}$ consists of the marginal counts $n_i = n_{\{i\},\{\}}$ as well as the counts $n_{A\cup\{S\},B}$ for any $A, B \subset \mathcal{S}^*$ and $A \cap B = \{\}$, as we know the complete inclusion history for each survey participant. Let $\mathcal{M}$ denote the set of missing counts $\{n_{A,A^c} : A \subset \mathcal{S}^*\}$. Given $\mathcal{D}$, let Θ denote the space of possible choices of $\mathcal{M}$ strictly compatible with the marginal counts n_i and the counts $n_{A\cup\{S\},A^c}$. Formally, the compatibility can be defined as the counts in $\mathcal{M}$ satisfying the following equations:

$$\sum_{\{A:i\in A\subset\mathcal{S}^*\}} n_{A,A^c} = n_i - \sum_{\{A:i\in A\subset\mathcal{S}^*\}} n_{A\cup\{S\},\mathcal{S}^*\setminus A} \quad \forall i \in \mathcal{S}^*. \tag{8}$$

Note that in (8), above the left-hand side concerns the missing counts in $\mathcal{M}$, whereas the constraints defined on the right-hand side are based on the observed counts. Let $\Theta = \{\mathcal{M} : \mathcal{M} \text{ satisfies } (8)\}$. The likelihood for the observed data $\mathcal{D}$ is then given by

$$\ell(N, p_1, \dots, p_S \mid \mathcal{D}) = \frac{N!\prod_{i=1}^{S} p_i^{n_i}(1-p_i)^{N-n_i}}{\prod_{A\subset\mathcal{S}^*} n_{A\cup\{S\},\mathcal{S}^*\setminus A}!}\sum_{\mathcal{M}\in\Theta}\frac{1}{\prod_{A\subset\mathcal{S}^*} n_{A,A^c}!}. \tag{9}$$

Clearly, this likelihood is intractable, because of the summation over a complicated parameter space, which rules out direct optimization to obtain the MLE. A natural solution to this would be using the joint likelihood for the complete data

$$\ell(N, p_1, \dots, p_S, \mathcal{M} \mid data) = \frac{N!\prod_{i=1}^{S} p_i^{n_i}(1-p_i)^{N-n_i}}{\prod_{A\subset\mathcal{S}^*} n_{A\cup\{S\},\mathcal{S}^*\setminus A}!\prod_{A\subset\mathcal{S}^*} n_{A,A^c}!} \tag{10}$$

to impute the missing cell counts $\mathcal{M}$ via a MCMC or EM algorithm. In Section 4 we detail an MCMC solution using Gibbs updates for this model.

3.3. *Misalignment.* Next, we consider the case of spatially misaligned data, where we intend to estimate population size for several regions, but some of the marginal counts are not known for each individual region. Instead, marginal counts for some of the listings are known for a larger geographical area which includes more than one area. The situation is epitomized in Mbabane and Manzini (Figures 2 and 3) where we have data from three listings: a survey, unique identifiers and coupons for MSM or NGO membership for FSW. However, the total number of coupons distributed or NGO members are not known individually for Mbabane and Manzini but were known for the Corridor region which encapsulates both of these areas.

We propose a solution for the misalignment problem in the most general form here, allowing misalignment specific to each data source. Let us consider the problem of size estimation in regions $1, 2, \ldots, K$. At each region, like in Section 3.2, there are S data sources (surveys or listings) for the population of interest. All the variables introduced in Section 3.2 are now given a region-specific index. For example, $N^{(k)}$ now denotes the total population in region k, for $k = 1, \ldots, K$.

In case of no spatial misalignment, the joint likelihood for the complete data from all the regions would be

$$\begin{aligned}&\ell(\{N^{(k)}, p_1^{(k)}, p_2^{(k)}, \ldots, p_S^{(k)}, \mathcal{M}^{(k)}\}_{\{k=1,\ldots,K\}} \mid data\}) \\ &\propto \prod_{k=1}^{K} \frac{N^{(k)}! \prod_{i=1}^{S} (p_i^{(k)})^{n_i^{(k)}} (1 - p_i^{(k)})^{N^{(k)} - n_i^{(k)}}}{\prod_{A \subset \{1,\ldots,S\}} n_{A,A^c}^{(k)}!}.\end{aligned} \tag{11}$$

This is simply an extension of (10) for K regions. As in Section 3.2, we assume that the survey is the Sth source due to which all counts of the form $n_{A\cup\{S\},A^c}^{(k)}$, $A \in \mathcal{S}^*$ are known for each region k. Now, to portray the misalignment, for the ith source ($i \in \mathcal{S}^*$), let $P_i^{(1)}, P_i^{(2)}, \ldots, P_i^{(g_i)}$ denote a partition of $\{1, 2, \ldots, K\}$ such that, instead of the individual region-specific survey counts $n_i^{(k)}$, we have the total survey counts for each of the larger regions indexed by the $P_i^{(j)}$'s, which we denote by $m_i^{(j)} = \sum_{k \in P_i^{(j)}} n_i^{(k)}$. Hence, in addition to the counts $\mathcal{M}^{(k)}$, we also have the set of marginal counts $\mathcal{H} = \{n_i^{(k)} : i \in \mathcal{S}^*, k \in 1, 2, \ldots, K\}$ as missing. Let Ψ denote the parameter space for $\mathcal{H}$ such that the missing marginal counts $n_i^{(k)}$ are compatible with the observed aggregate counts $m_i^{(j)}$. $\mathcal{H} \in \Psi$ implies that the counts $n_i^{(k)}$ satisfy

$$\begin{aligned}n_i^{(k)} &\geq \sum_{\{A : i \in A \subset \mathcal{S}^*\}} n_{A\cup\{S\},\mathcal{S}^*\setminus A}^{(k)}, \\ \sum_{k \in P_i^{(j)}} n_i^{(k)} &= m_i^{(j)}.\end{aligned} \tag{12}$$

Like in (8), the missing counts in the left-hand side of (12) need to satisfy the constraints specified by the observed counts on the right-hand side. If $\Theta^{(k)} = \{\mathcal{M}^{(k)}$ satisfies (8)$\}$, then the likelihood for the misaligned observed data is given by

$$\begin{aligned}&\ell(\{N^{(k)}, p_1^{(k)}, p_2^{(k)}, \ldots, p_S^{(k)}\}_{\{k=1,\ldots,K\}} \mid data\}) \\ &\propto \sum_{\mathcal{H}\in\Psi} \sum_{k=1}^{K} \sum_{\mathcal{M}^{(k)} \in \Theta^{(k)}} \prod_{k=1}^{K} \frac{N^{(k)}! \prod_{i=1}^{S} (p_i^{(k)})^{n_i^{(k)}} (1 - p_i^{(k)})^{N^{(k)} - n_i^{(k)}}}{\prod_{A \subset \mathcal{S}} n_{A,A^c}^{(k)}!}.\end{aligned} \tag{13}$$

This is again computationally intractable, and we will use the tractable complete multinomial likelihood in (11) and augment the data by imputing the missing $n_i^{(k)}$'s and $\mathcal{M}^{(k)}$'s to implement the Gibbs sampler.

4. Computation.

In this section we provide the Gibbs steps to sample all the parameters in the completed likelihoods (10) and (11), respectively.

4.1. *Incompleteness.*

Conditional on everything else, sampling for N and p_i's in (10) can be easily achieved using the Gibbs sampler in (6). Hence, our problem reduces to imputing the missing cell counts given N and $p_1, \ldots, p_S$.

The likelihood in (10) is the same as that for a $2 \times 2 \times \cdots \times 2$ contingency table (S-way contingency table with binary factors) with known marginals (n_i). The counts $\{n_{A,A^c} \mid A \subset \mathcal{S}\}$

correspond to the finest partitions which we also refer to as the atomic cells. There is substantial literature on sampling contingency tables with fully or partially known marginals and other constraints. A fundamental solution to this was proposed by Diaconis and Sturmfels (1998) where Metropolis-random walk on the space of contingency tables compatible with the data was constructed using *Markov bases*—set of moves which can explore the space of all compatible tables. The problem of finding the Markov bases for a given contingency table is generally difficult and computationally prohibitive; Dobra (2012) has considered local bases for Monte Carlo sampling of contingency tables. Dobra, Tebaldi and West (2006) provides a nice overview of hierarchical Bayesian sampling of contingency tables using Markov bases and other techniques like compositional sampling. Other solutions to this problem include the sequential Monte Carlo solution (Chen et al. (2005)) and a Gibbs sampling (Smith, Forster and McDonald (1996)), both of which were concerned only with two-way tables. The latter solution, that is, Gibbs sampling, is particularly attractive as, unlike the other approaches based on acceptance-rejection, it relieves the user of the task of carefully choosing Markov bases or proposal distributions.

Here, we construct a Gibbs sampler for drawing samples from the posterior distribution of $2 \times 2 \times \cdots \times 2$ contingency tables based on (10). We first illustrate the approach with $S = 3$ using Venn diagrams in Figure 4. Let $A\overline{B}$ denote cardinality of the set $A \cap B^c$, $\overline{A}B\overline{C}$ denote that of $A^c \cap B \cap C^c$ and so on. For $S = 3$, we know the marginal counts $A = n_1$, $AB + \overline{A}B = n_2$, $ABC + \overline{A}BC + \overline{AB}C + A\overline{B}C = n_3$. If the set of atomic (finest) partitions were known, the joint posterior from could be simplified as

$$p(N, p_1, p_2, p_3 \mid data)$$
$$= \frac{N! \prod_{i=1}^{3} p_i^{n_i}(1 - p_i)^{N-n_i}}{ABC!\overline{A}BC!\overline{AB}C!A\overline{B}C!AB\overline{C}!\overline{A}B\overline{C}!\overline{ABC}!A\overline{BC}!} p(N)p(p_1, p_2, p_3).$$

Instead of directly sampling the counts for the atomic (finest) partitions, we adopt a strategy illustrated in Figure 4 using Venn diagrams. We introduce, from left to right, the three sets (blue, red and green circles, resp.) one by one. Each set (circle) corresponds to inclusion in one of the surveys. At each step, we consider the counts of all the partitions within the circle that was newly added to the Venn diagram at that step. So step 0 gives N (when there is no partition), step 1 (left) gives A, step 2 (middle) gives AB and $\overline{A}B$ and step 3 (right) gives ABC, $\overline{A}BC$, $\overline{AB}C$ and $A\overline{B}C$. It is now easy to see that, based on these set of counts, one can recover all the atomic partition counts and vice versa. So sampling these set of sequential counts is equivalent to sampling the atomic counts and, hence, the contingency table.

Let $\mathrm{MVHG}_S(n, (N_1, N_2, \ldots, N_S)')$ denote the S-dimensional multivariate hypergeometric distribution with pmf

$$p((x_1, x_2, \ldots, x_S)') = \prod_{i=1}^{S} \binom{N_i}{x_i} \Big/ \binom{\sum_{i=1}^{S} N_i}{n}.$$

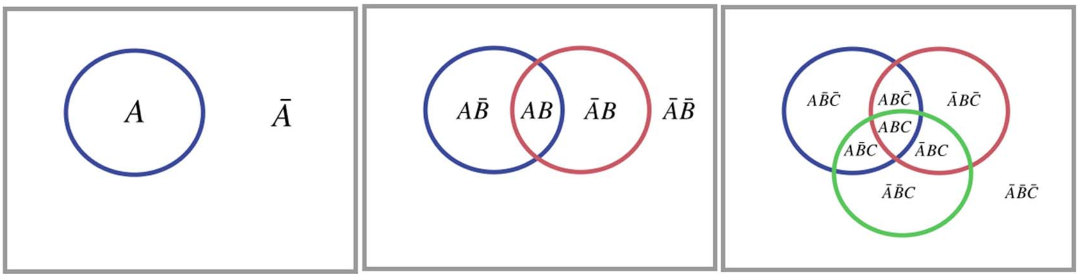


FIG. 4. *Venn diagram illustrating the Gibbs sampler for incomplete multiple capture-recapture data.*

From Figure 4 (right), we observe that the partitions within the green circles ABC, $\overline{A}BC$, $\overline{AB}C$ and $A\overline{B}C$, can be thought of as the color distributions of n_3 balls picked up from an urn which contains balls of four colors with counts AB, $\overline{A}B$, $\overline{AB}$ and $A\overline{B}$, respectively. Hence, we have

$$(ABC, \overline{A}BC, \overline{AB}C, A\overline{B}C)' \mid \cdot \sim \mathrm{MVHG}_4\big(n_3, (AB, \overline{A}B, \overline{AB}, A\overline{B})'\big).$$

On the other hand, conditional on everything else, the step 2 (middle figure) partition counts AB, $\overline{A}B$, $\overline{AB}$, $A\overline{B}$ have well-defined upper and lower bounds. For example, AB is bounded below by ABC and above by $A - A\overline{B}C$. Similarly, $\overline{A}B$ is bounded below by $\overline{A}BC$ and above by $\overline{A} - \overline{AB}C$. Hence, the only random components are the residuals $AB - ABC$ and $\overline{A}B - \overline{A}BC$ which needs to be picked from $A - ABC - A\overline{B}C$ and $\overline{A} - \overline{A}BC - \overline{AB}C$, respectively. The problem is equivalent to choosing $n_2 - (ABC + \overline{A}BC)$ balls from an urn containing $A - ABC - A\overline{B}C$ balls of one color and $\overline{A} - \overline{A}BC - \overline{AB}C$ balls of another color. Since the quantities n_2, A, $\overline{A} = N - A$, ABC, $A\overline{B}C$, $\overline{A}BC$ and $\overline{AB}C$ are known for this step of the Gibbs sampler, we have

$$\begin{pmatrix} AB \\ \overline{A}B \end{pmatrix} \Bigg| \cdot \sim \begin{pmatrix} ABC \\ \overline{A}BC \end{pmatrix} + \mathrm{MVHG}_2 \left(n_2 - (ABC + \overline{A}BC), \begin{pmatrix} A - ABC - A\overline{B}C \\ \overline{A} - \overline{A}BC - \overline{AB}C \end{pmatrix} \right).$$

This shows how we can leverage the multivariate hypergeometric distribution to specify the Gibbs sampling steps. We now formalize the algorithm for general S.

Using the notation introduced in Section 3.1, instead of directly sampling the finest partition counts, we intend to sample the counts $n_{\{12\},\{\}}$, $n_{\{2\},\{1\}}$, $n_{\{123\},\{\}}$, $n_{\{23\},\{1\}}$, $n_{\{13\},\{2\}}$, $n_{\{3\},\{12\}}$ and so on. We introduce some additional notation. Let $U_i = \{n_{A,B} \mid A \cup B = \{1, 2, \dots, i\}, A \cap B = \{\}\}$ and $V_i = \{n_{A \cup \{i\},B} \mid A \cup B = \{1, 2, \dots, i-1\}, A \cap B = \{\}\}$. Given N and n_i's, we will sample the cell counts in the sets $V_2, V_3, \dots, V_S$. We will also view the count sets U_i's and V_i's as vectors of lexicographically ordered counts. Note that there is a bijection between U_{i-1} and V_i, as each member $n_{A\cup\{i\},B}$ of V_i is a subset of its image $n_{A,B}$ in U_{i-1}. Consequently, $V_i \le U_{i-1}$. Also, $U_i = (V_i', U_{i-1}' - V_i')'$ for any i. Combining these results, we have $U_i! = V_i!(U_{i-1} - V_i)!$ and $U_{i-1}! = (V_i + (U_{i-1} - V_i))!$ where for any vector $x = (x_1, \dots, x_k)'$ let $x! = \prod_i x_i!$. The likelihood (10) can be expressed as

$$\frac{N! \prod_{i=1}^{S} p_i^{n_i} (1 - p_i)^{N - n_i}}{V_S!(U_{S-1} - V_S)!} = \prod_{i=1}^{S} p_i^{n_i} (1 - p_i)^{N - n_i} \binom{U_{i-1}}{V_i} \quad \text{where } U_0 = N.$$

Given N, the V_i's are a one-to-one function of the set of finest partition counts $\{n_{A,A^c}\}$. Also, $V_i \subset U_i$, and U_i can be recovered deterministically given N, $V_1, V_2, \dots, V_i$. Hence, U_{S-1} is determined by N, $V_1, \dots, V_{S-1}$ and we immediately then have the following update:

$$V_S \mid \cdot \sim \mathrm{MVHG}_{2^{S-1}}(n_S, U_{S-1}). \tag{14}$$

Next, we sample from the full conditional of V_i, given N, $V_1, \dots, V_{i-1}, V_{i+1}, \dots, V_S$. Note that, as V_S is known, all counts of the form n_{A,A^c} where $S \in A$ are known. Similarly, when V_{S-1} and V_S are known, counts of the form n_{A,A^c} where $S - 1 \in A$ are also known. To elucidate this with an example, let $A = \mathcal{S}^*$. Then, $n_{A,A^c} = n_{A,\{\}} - n_{A\cup\{S\},\{\}}$. Since $n_{A,\{\}} \in V_{S-1}$ and $n_{A\cup\{S\},\{\}} \in V_S$, n_{A,A^c} is known. Proceeding like this, it can be proved that, when $V_S, \dots, V_{i+1}$ is known, all the atomic counts of the form n_{A,A^c} are known when A contains at least one of $i+1, \dots, S$. Hence, the only unknown atomic counts are the sets $W_i = \{n_{A\cup\{i\},B\cup\{i+1,\dots,S\}} \mid A \cup B = \{1, \dots, i-1\}, A \cap B = \{\}\}$ and $\overline{W}_i = \{n_{A,B\cup\{i,i+1,\dots,S\}} \mid A \cup B = \{1, \dots, i-1\}, A \cap B = \{\}\}$. For $j > i$, let $V_{ij} = \{n_{A\cup\{i,j\},B\cup\{i+1,\dots,i+j-1\}} \mid A \cup B = \{1, 2, \dots, i-1\}, A \cap B = \{\}\}$, $\overline{V}_{ij} = \{n_{A\cup\{j\},B\cup\{i,i+1,\dots,i+j-1\}} \mid A \cup B = \{1, 2, \dots, i-1\}, A \cap B = \{\}\}$. Note that $V_{ij} \cap V_{ij'} = \{\}$ and $\overline{V}_{ij} \cap \overline{V}_{ij'} = \{\}$. Also, both V_{ij} and $\overline{V}_{ij}$ are subsets of

V_j and are known for this step, and we have $V_i = W_i + \sum_{j=i+1}^{S} V_{ij}$. Hence, the knowledge of $V_{i+1}, \dots, V_S$ helps determine the lower bound $\sum_{j=i+1}^{S} V_{ij}$ for V_i. On the other hand, knowledge of N, $V_1, \dots, V_{i-1}$ determines U_{i-1}, an upper bound for V_i. A sharper upper bound is given by $U_{i-1} - \sum_{j=i+1}^{S} \overline{V}_{ij}$ which is also known. Combining all this, we have

$$p(V_i \mid \cdot) \propto \frac{1}{W_i!\overline{W}_i!}, \quad \text{for} \sum_{j=i+1}^{S} V_{ij} \le V_i \le U_{i-1} - \sum_{j=i+1}^{S} \overline{V}_{ij}.$$

Since $W_i + \overline{W}_i = U_{i-1} - \sum_{j=i+1}^{S} (V_{ij} + \overline{V}_{ij})$, we immediately have the Gibbs update

$$V_i \mid \cdot \sim \sum_{j=i+1}^{S} V_{ij} + \mathrm{MVHG}_{2^{i-1}}\left(n_i - \sum_{j=i+1}^{S} 1' V_{ij}, U_{i-1} - \sum_{j=i+1}^{S} (V_{ij} + \overline{V}_{ij})\right). \tag{15}$$

Repeating (15) for $2 \le i \le S - 1$ completes the specification of the Gibbs sampler for a $2 \times 2 \times \cdots \times 2$ (2^S) contingency table with known marginals (n_i) and the grand total (N). In our application of size estimation using multiple multiplier method, V_S is known, as the complete inclusion histories of individuals participating in the survey (Sth listing) are known. Hence, the V_S-update in equation (14) is skipped, and our *Bayesian multiplier method* only performs the updates in (15) and, additionally, updates N and p_i's using (6).

4.2. *Misalignment.* We define few additional notations. Let $a_i^{(k)} = n^{(k)}_{\{i\},\{1,\dots,S\}\setminus\{i\}}$ and $b_i^{(k)} = n^{(k)}_{\{\},\{1,\dots,S\}}$. Note that $n_i^{(k)} = a_i^{(k)} + q_i^{(k)}$ and $a_i^{(k)} + b_i^{(k)} = N^{(k)} - t_i^{(k)}$, where

$$q_i^{(k)} = \sum_{A \subset \{1,\dots,S\}\setminus\{i\}, A \neq \{\}} n_{A\cup\{i\}, A^c\setminus\{i\}} \quad \text{and}$$

$$t_i^{(k)} = \sum_{A \subset \{1,\dots,S\}\setminus\{i\}, A \neq \{\}} n_{A, A^c\setminus\{i\}}.$$

We first sample the $n_1^{(k)}$'s in the Gibbs sampler. Given all the $N^{(k)}$'s and the $V_i^{(k)}$'s (region-specific analogs of the V_i's defined in Section 3.1) for $i \ge 2$, the $q_1^{(k)}$'s and $t_1^{(k)}$'s are known. The only unknown counts in the denominator of (11) are the $a_1^{(k)}$ and $b_1^{(k)}$'s, implying

$$p(\{n_1^{(k)}\} \mid \cdot) \propto \prod_{k=1}^{K} \frac{1}{a_1^{(k)}!(N^{(k)} - t_1^{(k)} - a_1^{(k)})!} \left(\frac{p_1^{(k)}}{1 - p_1^{(k)}}\right)^{a_1^{(k)}}. \tag{16}$$

Since $\{P_i^{(J)}\}_{1\le j\le g_i}$ is a partition of $\{1, 2, \dots, K\}$, for any set or vector R of K numbers, indexed by the K regions, let $P_i^{(j)}(R)$ denote the subvector of R corresponding to the indices in $P_i^{(j)}$. For example, if $R = \{r_1, \dots, r_K\}$, and $P_i^{(j)} = \{1, 3, 5\}$, then $P_i^{(j)}(R) = \{r_1, r_3, r_5\}$. We denote by $\mathrm{FNCMVHG}_k(n, (N_1, \dots, N_k)', (o_1, \dots, o_k))'$ the k-dimensional Fisher's noncentral multivariate Hyper-geometric distribution with pmf

$$p((x_1, \dots, x_k)') \propto \prod_{k=1}^{K} \binom{N_k}{x_k} o_k^{x_k} \Big/ \binom{\sum_{k=1}^{K} N_k}{n}.$$

Noncentral hypergeometric distributions arise when drawing $n = \sum x_k$ balls from a biased urn containing balls of K colors with odds o_k of picking a ball of the kth color. The R-package *BiasedUrn* (Fog (2015)) allows drawing random samples from FNCMVHG distributions.

Sampling from the full conditional for $n_i^{(k)}$'s are equivalent to sampling from that of the $a_i^{(k)}$'s, and we know that sum of the $a_i^{(k)}$'s for all i in a given partition $P_i^{(j)}$ is

$m_i^{(j)} - \sum_{k \in P_i^{(j)}} q_i^{(k)}$ which is known. It is now clear from (16) that the full conditional for $P_i^{(j)}(a_i^{(k)})$ follows a noncentral multivariate hypergeometric distribution, and we have the following update for $n_1^{(k)}$ within each partition:

$$\begin{aligned} &P_1^{(j)}(\{n_1^{(k)}\}) \mid \cdot \\ (17) \quad &\sim P_1^{(j)}(\{q_1^{(k)}\}) \\ &\quad + \text{FNCMVHG}_{|P_1^{(j)}|}\Biggl(m_1^{(j)} - \sum_{k \in P_1^{(j)}} q_1^{(k)}, P_1^{(j)}(\{N^{(k)} - t_1^{(k)}\}), P_1^{(j)}\biggl(\biggl\{\frac{p_1^{(k)}}{1 - p_1^{(k)}}\biggr\}\biggr)\Biggr). \end{aligned}$$

Once, all the $n_1^{(k)}$'s are updated, we update the $q_2^{(k)}$'s and $t_2^{(k)}$ and sample the $n_2^{(k)}$'s using the analog of (17) where all the subscripts are 2 instead of 1, and proceed like this to update all the $n_i^{(k)}$'s.

5. Size estimation in eSwatini.

5.1. *MSM.* Throughout the analysis we use the following conventions. The names of cities and areas are abbreviated to two letters. PP, Nh, Me, Mn, Lv and Co refers to Piggs Peak, Nhlangano, Mbabane, Manzini, Lavumisa and the Corridor, respectively. For any site or area $s \in \mathcal{L} = \{\text{PP}, \text{Me}, \text{Mn}, \text{Nh}\}$, let $D(s)$ denote the set of data sources available for that site, that is, for MSM, $D(\text{PP}) = \{\text{UID}, \text{SRV}\}$, $D(\text{Nh}) = \{\text{UID}, \text{SRV}, \text{RNB}\}$, etc. Also, let $N^{(s)}$ denote the total MSM population in that area in the age group of 18–32 years. This age group accounts for 90% of the MSM survey participants. We follow the notation introduced in Sections 3.2 and 3.3. For example, $n_{\{\text{UID}\}}^{(s)}$ denotes the total number of MSM in site s who received an unique object, $n_{\{\text{UID},\text{SRV}\},\{\text{CPN}\}}^{(s)}$ denotes the total number of MSM in site s who received an unique object, participated in the survey but did not receive a coupon, and so on. Similarly, $p_{\text{UID}}^{(s)}$ denotes the inclusion probability of receiving an unique object in site s, etc.

We use the data augmentation strategies developed in Sections 3.2 and 3.3 to augment and sample the missing finest partition counts for each site. The data-augmented likelihood is given by

$$(18) \qquad \prod_{s \in \mathcal{L}} \frac{N^{(s)}!}{\prod_{A \subset D(s)} n_{A,A^c}^{(s)}!} \prod_{x \in D(s)} (p_x^{(s)})^{N_x^{(s)}} (1 - p_x^{(s)})^{N^{(s)} - N_x^{(s)}}.$$

We use independent conjugate Beta(a, b) priors for all the inclusion probabilities $p_x^{(s)}$ with $a = b = 1$ (i.e., uniform distribution). Analysis assessing sensitivity to the choice of these hyperparameter values is conducted in Section S4.3 of the Supplement Material (Datta et al. (2020)) and confirms robust results. For the site-specific total MSM counts $N^{(s)}$, which are our main quantities of interest, we considered two prior choices—a Jeffrey's prior $p(N^{(s)}) \propto 1/N^{(s)}$ as well as a Binomial prior $N^{(s)} \mid \phi^{(s)} \sim \text{Binomial}(P^{(s)}, \phi^{(s)})$ where $P^{(s)}$ is the total male population in the age group of 18–32 years at site s. The Appendix provides the details on how the total male population numbers $P^{(s)}$ were derived. Both prior choices lead to conjugacy for $N^{(s)}$ in the Gibbs sampler; however, the Binomial prior offers the possibility of borrowing strength across sites in modeling the $\phi^{(s)}$'s which represent the proportion of total male population who are MSM at site s and are critical quantities of interest. For example, if we had data for many sites, we could model $\phi^{(s)}$ using demographic covariates. Since our application only involves four to five sites, we use a simpler exchangeable model for $\phi^{(s)}$, under the assumption that the proportion of total male population who are MSM are roughly similar across the sites. We model $\phi^{(s)} \mid a_\phi, b_\phi \overset{\text{ind}}{\sim} \text{Beta}(a_\phi, b_\phi)$ where (a_ϕ, b_ϕ) is given the vague prior $p(a_\phi, b_\phi) \propto 1/(a_\phi + b_\phi)^2 I(a_\phi > 1, b_\phi > 1)$ (Bao, Raftery and Reddy (2015)). The exchangeable model is also critical for predicting MSM population size in Lubombo (Lu) where there is no data. The expected proportion in Lubombo will be given by $E(\phi^{(\text{Lu})} \mid \mathit{data}) = E(a_\phi/(a_\phi + b_\phi) \mid \mathit{data})$ which is easily obtained using the postconvergence MCMC samples. Hence, we only present the results for the Binomial hierarchical prior for $N^{(s)}$'s in the main manuscript. The analogous numbers using the Jeffrey's prior are presented in Section S4.2 of the Supplementary Material (Datta et al. (2020)) which, in general, are in close agreement with the results in this Section. Letting

$$p_{A,A^c}^{(s)} = \prod_{x \in A} p_x^{(s)} \prod_{x \in D(s) \setminus A} (1 - p_x^{(s)}), \tag{19}$$

the full hierarchical model for the Binomial prior is given by

$$\begin{aligned} &\prod_{s \in \mathcal{L}} \Bigg(\text{Binomial}(N^{(s)} \mid P^{(s)}, \phi^{(s)}) \\ &\quad \times \text{Multinomial}(\{n_{A,A^c}^{(s)}\}_{A \subset D(s)} \mid N^{(s)}, \{p_{A,A^c}^{(s)}\}_{A \subset D(s)}) \\ &\quad \times \prod_{x \in D(s)} \text{Beta}(p_x^{(s)} \mid a, b) \times \text{Beta}(\phi^{(s)} \mid a_\phi, b_\phi) \Bigg) \times p(a_\phi, b_\phi). \end{aligned} \tag{20}$$

In the Gibbs sampler, full conditionals for all inclusion probabilities $p_x^{(s)}$ as well as the MSM proportions $\phi^{(s)}$ follow conjugate Beta distributions, while the MSM population sizes $N^{(s)}$ were sampled as

$$N^{(s)} \mid \cdot \sim r^{(s)} + \text{Bin}(P^{(s)} - r^{(s)}, x^{(s)}/(1 - \phi^{(s)} + x^{(s)})), \quad \text{where } x^{(s)} = \phi^{(s)} \prod_{x \in D(s)} (1 - p_x^{(s)})$$

and $r^{(s)}$, as defined earlier, is the total number of MSM included in at least one of the data sources in region s. Note that the full conditional distribution for $N^{(s)}$ is Binomial here, instead of being negative-Binomial in (6), because of the switch from a Jeffrey's prior to a Binomial prior for $N^{(s)}$ in order to exploit the fact that $N^{(s)}$ is the size of a subpopulation (MSM or FSW) within a larger general population (all adult males or females) of size $P^{(s)}$. The hyperparameters a_ϕ and b_ϕ were updated using a Metropolis random walk step within the Gibbs sampler. Turning to the data augmentation part, Piggs Peak is the simplest case where there are only two sources, UID and SRV, and the overlap between them is known. Hence, all the finest partitions are known (Figure 2), and no data augmentation is needed. However, for the other three sites (Nhlangano, Mbabane and Manzini) we sample the finest partition counts for those who did not participate in the survey. These are sampled in the Gibbs sampler using the multivariate hypergeometric distribution (equation (15)). Additionally, since the data is misaligned and we only know the total number of coupons distributed for the Corridor (which includes both Mbabane and Manzini) and not for these individual sites, we need to sample $n_{\text{CPN}}^{(\text{Mb})}$ and $n_{\text{CPN}}^{(\text{Mn})}$ from the noncentral hypergeometric distribution using equation (17). We used multiple parallel MCMC chains each of length 10,000 to ascertain convergence using the Gelman–Rubin diagnostic. We removed the first 5000 iterations as burn-in and used the latter 5000 iterations as draws from the posterior distribution for each parameter. Table 2 presents the posterior summaries of each parameter.

We first reemphasize that the size estimates $N^{(s)}$ in Table 2 are only the estimates of MSM population of age 18–32 years who are represented in the survey at that site. These denominators (total male population of the same age group) represented at each of these

TABLE 2
Posterior estimates of parameters in (20) *for MSM population size estimation*

	Region	Parameter	Mean	95% CI
MSM (18–32 years) size estimates for each site	Piggs Peak	$N^{(\mathrm{PP})}$	102	83–130
	Nhlangano	$N^{(\mathrm{Nh})}$	169	151–192
	Mbabane	$N^{(\mathrm{Me})}$	463	423–513
	Manzini	$N^{(\mathrm{Mn})}$	447	399–504
MSM percentages	Piggs Peak	$\phi^{(\mathrm{PP})}$	3.5%	2.7%–4.6%
	Nhlangano	$\phi^{(\mathrm{Nh})}$	4.1%	3.4%–4.9%
	Mbabane	$\phi^{(\mathrm{Me})}$	4.5%	3.9%–5.1%
	Manzini	$\phi^{(\mathrm{Mn})}$	2.8%	2.4%–3.2%
	Lubombo	$a_\phi/(a_\phi + b_\phi)$	3.8%	2.8%–5.3%
Inclusion probabilities	Piggs Peak	$p_{\mathrm{SRV}}^{(\mathrm{PP})}$	57%	42%–72%
	Piggs Peak	$p_{\mathrm{UID}}^{(\mathrm{PP})}$	30%	20%–42%
	Nhlangano	$p_{\mathrm{SRV}}^{(\mathrm{Nh})}$	42%	33%–51%
	Nhlangano	$p_{\mathrm{UID}}^{(\mathrm{Nh})}$	63%	53%–73%
	Nhlangano	$p_{\mathrm{RNB}}^{(\mathrm{Nh})}$	7.6%	4%–12%
	Mbabane	$p_{\mathrm{SRV}}^{(\mathrm{Me})}$	48%	42%–55%
	Mbabane	$p_{\mathrm{UID}}^{(\mathrm{Me})}$	47%	41%–54%
	Mbabane	$p_{\mathrm{CPN}}^{(\mathrm{Me})}$	12%	8.2%–16%
	Manzini	$p_{\mathrm{SRV}}^{(\mathrm{Mn})}$	40%	33%–46%
	Manzini	$p_{\mathrm{UID}}^{(\mathrm{Mn})}$	44%	37%–50%
	Manzini	$p_{\mathrm{CPN}}^{(\mathrm{Mn})}$	12%	8.1%–16%

sites are derived in the Appendix and are summarized in Table 12. The region-specific size estimates for MSM in the broader age group of 15–49 years are provided later in Table 3. We also see from Table 2 that the MSM percentages varies from around 2.8% in Manzini to 4.5% in Mbabane. The different site and data source specific inclusion probabilities varied from around 7.6% to 63%. By virtue of using the exchangeable model, we could extrapolate these results to obtain the expected MSM percentage in Lubombo which was 3.8%.

5.2. *Extrapolation to regional estimates.* We use the MCMC samples from the posterior distribution of the parameters in (20) to extrapolate to MSM population size estimates for the age group of 15–49 years for each of the four regions of eSwatini for the year of 2014 (when the survey was conducted). Size estimation for this broader age group assumes that the MSM percentage in the age group of 18–32 is same as the MSM percentage in the larger age group of 15–49 years. Since we do not have much data outside the age group of

TABLE 3
Region specific size estimates of MSM in the age group of 15–49 *years for* 2014

Region	Male 15–49 population	MSM size estimate	MSM Proportion
Hhohho	85,918	3508 (2913–4208)	4.1% (3.4–4.9%)
Manzini	95,721	2650 (2154–3194)	2.8% (2.3–3.3%)
Shiselweni	45,216	1868 (1440–2348)	4.1% (3.2–5.2%)
Lubombo	52,519	2016 (1408–2833)	3.8% (2.7–5.4%)

18–32 years, who constituted more than 90% of the survey participants, this assumption is indispensable.

We will obtain the posterior distribution of size estimate in each Tinkhundla which we then aggregate to get the regional size estimates. From Table 10, the following Tinkhundlas were well represented in the data: Manzini North, Manzini South and Kwaluseni for Manzini site, Piggs Peak for the Piggs Peak site, Mbangweni for the Nhlangano site and Motjane, Mbabane East, Mbabane West and Lobamba for the Mbabane site. For these Tinkhundlas the posterior distribution of the size estimate $N^{(t)}$ is given by $\{P^{(t)}\phi^{(t)}(x) | x = 1, \dots, 5000\}$ where $\phi^{(t)}(x)$ denotes the site specific MSM percentage for the xth post burn-in MCMC iteration.

To extrapolate to Tinkhundlas that were not well represented in the data, posterior predictive size estimate samples $N^{(t)}(x)$ were generated as $\mathrm{Binomial}(P^{(t)}, \phi^{(t)}(x))$ for x in $1, \dots, 5000$ and a suitably chosen $\phi^{(t)}$ (which we discuss below). The Binomial model for the extrapolation adds uncertainty to the posterior distribution of the size estimates to reflect the fact that there was no data for these Tinkhundlas.

In Manzini region for MSM, there is only one data collection site producing the estimate $\phi^{(\mathrm{MM})}$. Hence, for all Tinkhundlas in Manzini we use $\phi^{(t)} = \phi^{(\mathrm{Mn})}$ for generating Binomial samples. Similarly, for Tinkhundlas in Shiselweni we use $\phi^{(t)} = \phi^{(\mathrm{Nh})}$. For Tinkhundlas in Lubombo, we use the $\phi^{(t)}$ as the expected MSM percentage $a_\phi/(a_\phi + b_\phi)$.

The extrapolation for Tinkhundlas in Hhohho is a little more complicated, as there are two estimates of MSM percentages: $\phi^{(\mathrm{Me})}$ for the Mbabane region Tinkhundlas Motjane, Mbabane East, Mbabane West and Lobamba, and $\phi^{(\mathrm{PP})}$ for the Piggs Peak Tinkhundla. For Tinkhundlas other than these five, we first interpolate the MSM percentage based on their proximity to Mbabane or Piggs Peak. We do this using a spatial autoregression. Let $H = \{H(1), H(2), \dots\}$ denote an ordering of the Tinkhundlas in Hhohho such that $H(a)$ for $a \in \{1, \dots, 5\}$ corresponds to the five Tinkhundlas in Hhohho with data. Subsequently, we use a directed acyclic graph autoregressive model (Datta et al. (2019), DAGAR)—a spatial autoregression that allows directly modeling conditional means of areal units given the means for some of the units. Let for two Tinkhundlas t and t' in Hhohho, $t \sim t'$ mean that they share a geographical border. Then, DAGAR specifies the $\phi^{(t)}$'s for the Tinkhundlas in Hhohho without data recursively as

$$\phi^{(H(a))} = \frac{\sum_{x=1}^{a-1} \phi^{(H(x))} I(H(x) \sim H(a))}{\sum_{x=1}^{a-1} I(H(x) \sim H(a))}, \quad a > 5.$$

The above expression implies that $\phi^{(t)}$ for a Tinkhundla is simply the average of the neighboring Tinkhundlas listed before it in H. This, in turn, ensures that $\phi^{(t)}$ is a weighted average of $\phi^{(\mathrm{Me})}$ and $\phi^{(\mathrm{PP})}$ with the weights being based on the graphical distance between the Tinkhundla t and the five Tinkhundlas of Hhohho with data.

The regional estimates are given in Table 3. The mean estimate for the MSM percentage varied from 2.8% in Manzini to 4.1.% in Hhohho. The credible interval for the MSM percentage was widest for Lubombo which had no direct data.

5.3. *FSW.* The analysis for FSW broadly follows similar trajectory. The model is same as in (18) with $\mathcal{L} = \{\mathrm{PP}, \mathrm{Me}, \mathrm{Mn}, \mathrm{Nh}, \mathrm{Lv}\}$. There is no data augmentation involved in either of Piggs Peak, Nhlangano or Lavumisa as each of them only has two data sources with known overlap. For Mbabane and Manzini, as shown in Figure 3, we know the marginal total for the survey and uid as well as the overlap between those two listings. We do not know the marginal totals for those who visited a FLAS clinic in Mbabane and Manzini but know the total number of FLAS members in the Corridor to be 186. This is similar to the misalignment in the MSM data but has the additional constraint imposed by the knowledge that at least 70 of

TABLE 4
Posterior estimates of parameters for FSW population size estimation

	Region	Parameter	Mean	95% CI
FSW (18–35 years) size estimates for each site	Piggs Peak	$N^{(\mathrm{PP})}$	348	306–400
	Nhlangano	$N^{(\mathrm{Nh})}$	119	88–174
	Mbabane	$N^{(\mathrm{Me})}$	747	663–841
	Lavumisa	$N^{(\mathrm{Lv})}$	141	131–156
	Manzini	$N^{(\mathrm{Mn})}$	427	397–462
FSW percentages	Piggs Peak	$\phi^{(\mathrm{PP})}$	12%	10–14%
	Nhlangano	$\phi^{(\mathrm{Nh})}$	2.4%	1.7–3.6%
	Mbabane	$\phi^{(\mathrm{Me})}$	4.5%	3.9–5.1%
	Lavumisa	$\phi^{(\mathrm{Lv})}$	10%	8.4–12%
	Manzini	$\phi^{(\mathrm{Mn})}$	2.1%	1.9–2.4%
	Lubombo	$a_\phi/(a_\phi + b_\phi)$	7.7%	3.9–16%
Inclusion probabilities	Piggs Peak	$p_{\mathrm{SRV}}^{(\mathrm{PP})}$	37%	30–44%
	Piggs Peak	$p_{\mathrm{UID}}^{(\mathrm{PP})}$	51%	43–60%
	Nhlangano	$p_{\mathrm{SRV}}^{(\mathrm{Nh})}$	41%	26–58%
	Nhlangano	$p_{\mathrm{UID}}^{(\mathrm{Nh})}$	30%	18–43%
	Mbabane	$p_{\mathrm{SRV}}^{(\mathrm{Me})}$	34%	29–40%
	Mbabane	$p_{\mathrm{UID}}^{(\mathrm{Me})}$	35%	30–41%
	Mbabane	$p_{\mathrm{FLAS}}^{(\mathrm{Me})}$	9.7%	7.4–12%
	Lavumisa	$p_{\mathrm{SRV}}^{(\mathrm{Lv})}$	62%	52–71%
	Lavumisa	$p_{\mathrm{UID}}^{(\mathrm{Lv})}$	68%	58–77%
	Manzini	$p_{\mathrm{SRV}}^{(\mathrm{Mn})}$	60%	54–66%
	Manzini	$p_{\mathrm{UID}}^{(\mathrm{Mn})}$	36%	31–41%
	Manzini	$p_{\mathrm{FLAS}}^{(\mathrm{Mn})}$	27%	22–32%

this 186 FSW visited the FLAS clinic in Mbabane. Thus, the sampler in Section 3.3 is slightly modified to accommodate this constraint by using a truncated noncentral hypergeometric distribution in equation (17) with cut-off of 70.

Table 4 presents the posterior summaries of each parameter. Again, note that the size estimates $N^{(s)}$ in Table 4 are estimates of the size of the FSW population (18–35) at that location which represent 90% of the survey participants. We see that for FSW the absolute numbers are higher in Mbabane and Manzini, but the percentages are higher for Piggs Peak and Lavumisa. Region-specific size estimates for FSW in the age group of 15–49 years are given in Table 5.

TABLE 5
Region specific size estimates of FSW in the age group of 15–49 *years*

Region	Female 15–49 population	FSW size estimate	FSW Proportion
Hhohho	84,784	6085 (5170–7106)	7.2% (6.1–8.4%)
Manzini	10,0361	2113 (1741–2514)	2.1% (1.7–2.5%)
Shiselweni	50,488	3272 (2613–4041)	6.5% (5.2–8%)
Lubombo	52,762	4059 (1998–8454)	7.7% (3.8–16%)

TABLE 6

Posterior predictive p-values for assessing the independence assumption in the data analysis

	Piggs Peak	Nhlangano	Lavumisa	Mbabane	Manzini
MSM	0.54	0.53	NA	0.48	0.48
FSW	0.5	0.5	0.51	0.45	0.22

The extrapolation strategy mimicked that for MSM for Hhohho, Manzini and Lubombo. For Shiselweni, however, it was different, as (unlike for MSM) we had now two sites with data leading to two FSW percentages—$\phi^{(\mathrm{Nh})}$ for Nhalangano and $\phi^{(\mathrm{Lv})}$ for Lavumisa. So, akin to our strategy for Hhohho, we used a spatial autoregression using the DAGAR model to interpolate the $\phi^{(t)}$'s for the remaining Tinkhundlas of Shiselweni and generated the size estimates using $N^{(t)} \mid \phi^{(\mathrm{Nh})}, \phi^{(\mathrm{Lv})} \overset{\mathrm{iid}}{\sim} \mathrm{Binomial}(P^{(t)}, \phi^{(t)})$.

The regional FSW estimates are provided in Table 5. We see that the regional FSW percentage varies is once again lowest in Manzini (2.1%), whereas it is around 7.5% in Hhohho and Lubombo. Once again, the estimate for Lubombo had the highest uncertainty.

We assess the independence assumption for both the MSM and FSW data analysis. Since we use a hierarchical model and MCMC-based Bayesian computation, we use the model adequacy checks based on posterior predictive samples, as described in Chaper 6.3 of Gelman et al. (2013). Similar tests were performed for the capture-recapture data in Bao, Raftery and Reddy (2015). As in Section 3.2, we view the problem as an incomplete contingency table. In each MCMC iteration we first compute a χ^2 goodness of fit statistic using the table of *observed counts* (actual observed counts and imputed counts for that iteration) and the expected counts (total population size × cell probabilities for that iteration under the independence assumption). We then see if this statistic is less than the χ^2 statistic from a randomly generated (posterior predictive) contingency table using the same cell probabilities. The average of this binary variable over all the posterior samples gives us the tail-probability (p-value for testing the independence assumption). The p-values are provided in Table 6. We see that, for both MSM and FSW, p-values for all regions are sufficiently large and does not offer any evidence against the independence assumption.

6. Additional analyses.

We conducted simulation studies and additional analyses of the eSwatini data to assess the performance of the models and robustness of the estimates to various choices. We only present one representative set of results here, with the rest (sensitivity to prior and hyper-parameter choices) provided in Section S4 of the Supplementary Material (Datta et al. (2020)).

6.1. *Simulation studies.*

We first conducted a set of simulation studies to assess the performance of our data augmented Bayesian model for generating direct estimates of population size in multiple regions using incomplete and misaligned data. The simulation framework closely resembled the data available to us. For MSM, we used four sites representing Piggs Peak, Mbabane, Manzini and Nhlangano and simulated MSM population sizes at each of these sites as $N^{(s)} \overset{\mathrm{ind}}{\sim} \mathrm{Binomial}(P^{(s)}, \phi^{(s)})$. We used the true total populations $P^{(s)}$'s from Table 12 and the posterior means of $\phi^{(s)}$ from Table 2 to generate the data. This implies that the data generating mechanism is not similar to the model fitted which assumes an exchangeable model for the $\phi^{(s)}$ and, hence, gives us an opportunity to validate sensitivity to the exchangeability assumption. Subsequent to generating the $N^{(s)}$, we generate the complete

TABLE 7

Simulation results for MSM size estimation. True values and coverage probabilities were averaged over 100 *simulations. The estimates are the posterior mean. CP denotes coverage probability*

	Region	Parameter	Truth	Estimate	CP
MSM (18–32 years) size estimates for each site	Piggs Peak	$N^{(\mathrm{PP})}$	102	102	96%
	Nhlangano	$N^{(\mathrm{Nh})}$	166	162	92%
	Mbabane	$N^{(\mathrm{Me})}$	456	447	93%
	Manzini	$N^{(\mathrm{Mn})}$	449	463	92%
MSM proportions	Piggs Peak	$\phi^{(\mathrm{PP})}$	3.5%	3.5%	96%
	Nhlangano	$\phi^{(\mathrm{Nh})}$	4.1%	4.0%	92%
	Mbabane	$\phi^{(\mathrm{Me})}$	4.5%	4.3%	90%
	Manzini	$\phi^{(\mathrm{Mn})}$	2.8%	2.9%	98%
Inclusion probabilities	Piggs Peak	$p_{\mathrm{UID}}^{(\mathrm{PP})}$	30%	30%	95%
	Piggs Peak	$p_{\mathrm{SRV}}^{(\mathrm{PP})}$	57%	57%	98%
	Nhlangano	$p_{\mathrm{UID}}^{(\mathrm{Nh})}$	63%	65%	94%
	Nhlangano	$p_{\mathrm{RNB}}^{(\mathrm{Nh})}$	8%	8%	93%
	Nhlangano	$p_{\mathrm{SRV}}^{(\mathrm{Nh})}$	42%	42%	99%
	Mbabane	$p_{\mathrm{SRV}}^{(\mathrm{Me})}$	48%	49%	99%
	Mbabane	$p_{\mathrm{UID}}^{(\mathrm{Me})}$	47%	48%	98%
	Mbabane	$p_{\mathrm{CPN}}^{(\mathrm{Me})}$	12%	12%	92%
	Manzini	$p_{\mathrm{SRV}}^{(\mathrm{Mn})}$	40%	39%	91%
	Manzini	$p_{\mathrm{UID}}^{(\mathrm{Mn})}$	44%	43%	92%
	Manzini	$p_{\mathrm{CPN}}^{(\mathrm{Mn})}$	12%	11%	97%

inclusion history in the listings $D(s)$ for each individual, that is, we generate the finest partition counts from Multinomial$(N^{(s)}, \{p_{A,A^c}\}_{A\subset D(s)})$, where the inclusion probabilities p_{A,A^c} for the finest partitions are defined in (19). As before, we use the posterior mean estimates of $p_x^{(s)}$ from Table 2 to construct the p_{A,A^c}.

For estimating MSM size using the generated data, we do not use all the counts but only a subset, mirroring the incompleteness and misalignment in the eSwatini data, as highlighted in Figure 2. Table 7 presents the true parameter values and the posterior means averaged over 100 replicate datasets and the corresponding coverage probabilities of posterior 95% credible intervals. We see that estimates of all the parameters—size estimates, MSM percentages and inclusion probabilities—for all the four sites are very close to their simulated truths. The coverage probabilities for most of these parameters are also close to the nominal level of 95%.

We proceeded in the same manner described above for MSM in conducting simulation studies for the FSW analysis. The details are provided in Section S4.1 of the Supplementary Material (Datta et al. (2020)). The simulation results for FSW, like the MSM simulation, were extremely accurate in terms of parameter estimation and inference. Overall, the simulation studies demonstrate the ability to generate accurate direct estimates of key populations for multiple regions based on our Bayesian hierarchical model using multiple multiplier data disguised as incomplete and misaligned capture-recapture data. Given that the simulated data was generated independently for each region, using the MSM or FSW percentages estimated in Tables 2 or 4, and then analyzed using an exchangeable model, it also demonstrates propriety of the exchangeability assumption for the eSwatini data.

TABLE 8
Three Lincoln–Peterson (multiplier-method) estimates and our model based size estimate of the MSM population size in Nhlangano

Estimator type	Estimate (credible interval)
mult.uid	173 (148–197)
mult.rnb	140 (64–216)
mult.avg	156 (106–206)
mult.Bayes	169 (151–192)

6.2. *Comparison with average multiplier method.* In this section we analyze the MSM data from Nhlangano to show how the different Lincoln–Peterson estimates compare with our model based estimates. We use the MSM data from Nhlangano for this comparison, as it has three data sources (UID, RNB and SRV) and, therefore, yields two multiplier method estimates—one using UID and one using RNB (the SRV is always used).

We have the following four multiplier method estimates for MSM population size in Nhlangano:

$$\widehat{N}^{(\mathrm{Nh})}_{\mathrm{mult.uid}} = \frac{N^{(\mathrm{Nh})}_{\mathrm{UID}} N^{(\mathrm{Nh})}_{\mathrm{SRV}}}{N^{(\mathrm{Nh})}_{\{\mathrm{UID},\mathrm{SRV}\},\{\}}}, \qquad \widehat{N}^{(\mathrm{Nh})}_{\mathrm{mult.rnb}} = \frac{N^{(\mathrm{Nh})}_{\mathrm{RNB}} N^{(\mathrm{Nh})}_{\mathrm{SRV}}}{N^{(\mathrm{Nh})}_{\{\mathrm{RNB},\mathrm{SRV}\},\{\}}},$$

$\widehat{N}^{(\mathrm{Nh})}_{\mathrm{mult.avg}}$ which is the average of the two, and $\widehat{N}^{(\mathrm{Nh})}_{\mathrm{mult.Bayes}}$ which is the estimate (posterior mean) from our Bayesian multiplier method.

In Table 8 we compare these and our model based estimate (from Table 2) and the respective credible intervals. The variance estimate of the generic Lincoln–Peterson estimate in (1) is given by

$$\widehat{V(\widehat{N})} = \frac{n_1 n_2 (n_1 - n_{12})(n_2 - n_{12})}{n_{12}^3}. \tag{21}$$

The credible intervals for the Lincoln–Peterson estimators $\widehat{N}^{(\mathrm{Nh})}_{\mathrm{mult.uid}}$ and $\widehat{N}^{(\mathrm{Nh})}_{\mathrm{mult.rnb}}$ were calculated as estimate $\pm 1.96 * \sqrt{\text{respective variance estimates}}$. The credible interval for the average method was simply calculated as average of the credible intervals for $\widehat{N}^{(\mathrm{Nh})}_{\mathrm{mult.uid}}$ and $\widehat{N}^{(\mathrm{Nh})}_{\mathrm{mult.rnb}}$. Several interesting findings can be observed from Table 8. First, we note that the two multiplier-method estimates using UID and RNB, respectively, yield widely different results. The estimates differ by about 20%, and the estimate using RNB does not even lie within the 95% credible interval of the estimate using UID. The credible interval using RNB is also very wide (about three times the length of the credible interval using UID). This shows how different listings can yield very different multiplier-method estimates, leading to the quandary of which one to use. Interestingly, we see that our model based estimate aligns very closely with the estimate using UID instead of aligning with the average estimate. Even the credible intervals from our model were much closer to that using the UID data instead of the average method which has twice as wide credible interval.

To understand why our model, instead of giving equal weights to UID and RNB estimates, places more weight on the UID estimate, we look back at the inclusion probabilities of UID and RNB for Nhlangano in Table 2. We see that $p^{(\mathrm{Nh})}_{\mathrm{UID}}$ was estimated to be 63% while $p^{(\mathrm{Nh})}_{\mathrm{RNB}}$ was only 7.6%. This very large discrepancy between the two inclusion probabilities is perhaps the driving force in our model aligning with the UID estimate. The low-inclusion probability for RNB implies that the estimate will typically have high variance. Our model seems to incorporate this information thereby weighting the different sources in a data-driven manner while the average estimate places equal weights on all the sources ignoring the variances.

We could not repeat this exercise for the other regions, as the MSM data from Piggs Peak and the FSW data from Piggs Peak, Lavumisa and Nhlangano only have two data sources and yields a single estimate, whereas both the MSM and the FSW data for Mbabane and Manzini are misaligned, so the traditional multiplier-method cannot be computed.

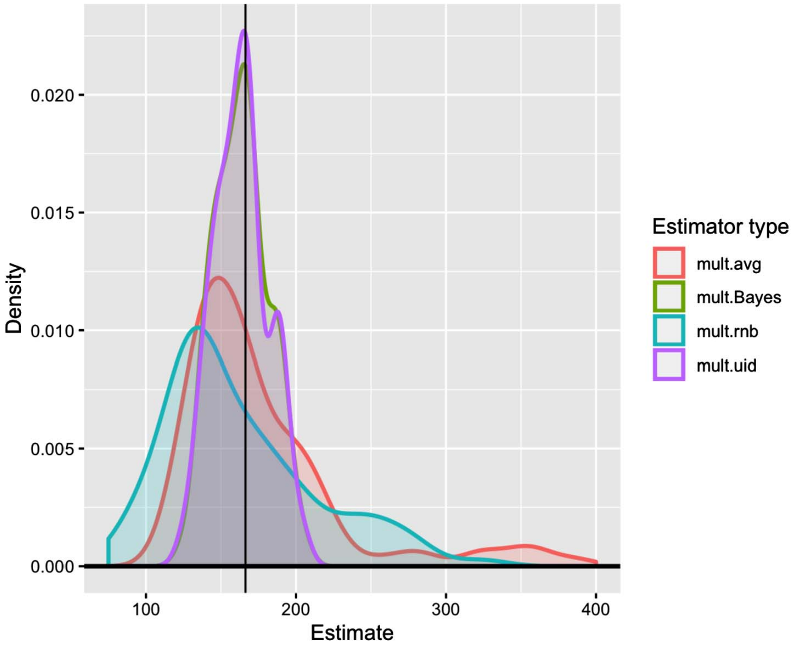
Next, we used the model based estimates for MSM in Nhlangano to simulate 100 datasets in the same fashion as in Section 6.1 but only restricted to Nhlangano. Figure 5 compares the performance of the methods. The densities of size estimates based on the 100 datasets for each of the methods are provided in Figure 5(a). We see that among the two multiplier methods based on single nonsurvey sources, mult.uid and mult.rnb, the former performs substantially better than the latter whose density is both more dispersed and biased away from the true value. The average multiplier method, which is usually used in such settings of multiple multiplier sources, performs better than mult.rnb but is still significantly biased. The density from the Bayesian multiplier method which also uses both sources (UID and RNB) aligns almost indistinguishably with the better performing method (UID). It is also centered around the true population size (black vertical line) and has much less variability than mult.avg. In Figure 5(b) we plot the average size estimate and 95% credible intervals over the 100 datasets. The point estimate once again clearly reveals the superiority of our Bayesian method over the average multiplier method which incurs bias from the mult.rnb part. Looking at the interval estimates, we see that the interval width for mult.rnb and mult.avg are an order of magnitude wider than the intervals for mult.uid and mult.Bayes.

We also calculated and compared the same four estimators used in Table 8 to the simulated true size estimates in terms of mean absolute error (MAE), mean square error (MSE), coverage probability (CP) and mean credible interval width (CIW) for these 100 replicate datasets. The results of the comparison are presented in Table 9. We observe that, like the real data analysis in Table 8, all the evaluation metrics from our model closely align with the multiplier method using UID. These two estimators produce substantially lower MAE and MSE than the multiplier method using RNB or the average multiplier method. The credible interval widths for our model and for $\widehat{N}_{\text{mult.uid}}$ are also an order of magnitude narrower than from the other two methods. In terms of coverage probability, our model produces slight under-coverage whereas the average multiplier method results in slight over-coverage. The effect of low inclusion probability for RNB is clearly evident with both the estimator using RNB or the average estimator, giving equal weights to RNB and UID producing worse error estimates and wider intervals.

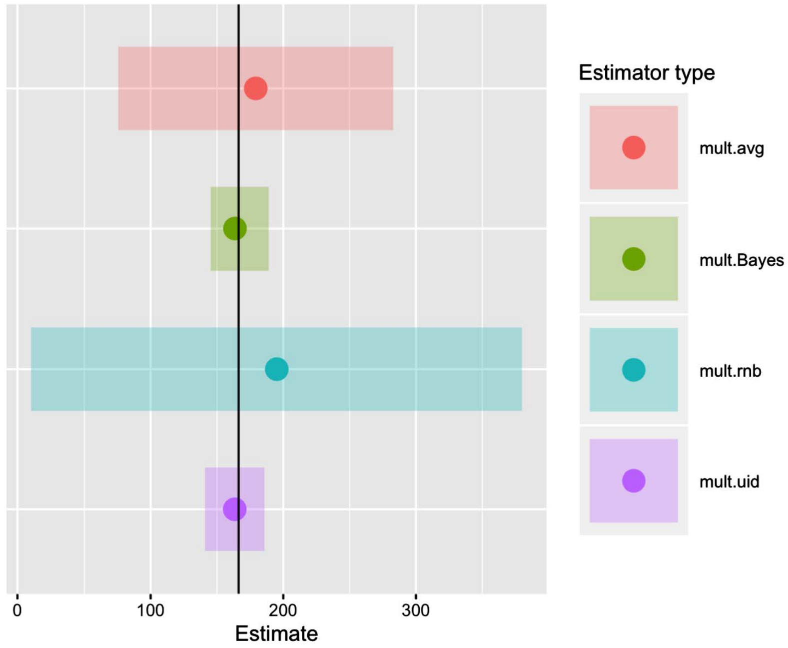
The results comprehensively demonstrate how our Bayesian multiplier method performs better than the average multiplier method. Despite not leveraging any prior distinction between the reliability of the two data sources UID and RNB, the Bayseian method weights them differently, aligning with the one (UID) which leads to the more accurate multiplier-method estimate.

7. Relaxing the independence assumption.

We propose a Bayesian nonparametric extension of our model to settings where the assumption of independent data sources is violated. We first discuss the related assumption of homogeneity of population also used in our model and all standard capture-recapture formulations. While several solutions to model the heterogeneity have been proposed (Manrique-Vallier and Fienberg (2008)), we focus on the finite mixture model as it offers an avenue to model nonindependent data sources. A finite mixture model represents the heterogeneous population as a union of homogeneous stratas such that we can use the independence assumption within each strata. Assuming population



(a) Density of estimates for 100 datasets.



(b) Mean estimate and 95% credible intervals

FIG. 5. *Comparison of the different multiplier methods for* 100 *simulated datasets. Black line indicates the mean true population size.*

stratification along a known and measured variable (covariate), various forms of stratified capture-recapture methods are available (Sekar and Deming (1949)).

If knowledge of such a covariate is not presumed, one proceeds via a latent discrete model for the covariate implying that every individual belongs to one of C possible homogeneous

TABLE 9

Model comparison metrics for the different multiplier method estimates for the 100 *simulated MSM datasets*

Estimator type	MAE	MSE	CP	CIW
mult.avg	35	61	97%	207
mult.rnb	67	120	93%	370
mult.uid	10	13	86%	45
mult.Bayes	10	12	89%	44

classes. We now have a two-stage hierarchical model

$$\begin{aligned} &\mathcal{I}(j)_i \mid \mathcal{C}(j) = c \overset{\text{ind}}{\sim} \text{Bernoulli}(p_i[c]), \quad i = 1, \dots, S, j = 1, \dots, N, \\ &\mathcal{C}(j) \overset{\text{iid}}{\sim} \text{Categorical}(\pi) \quad \text{such that } \pi[c] \geq 0 \text{ and } \pi'1 = 1, \end{aligned} \tag{22}$$

where $\mathcal{C}(j)$ is the class-membership for the jth individual and $p_i[c]$ is the inclusion probability for the ith source in class c, for $c = 1, \dots, C$, $i = 1, \dots, S$. Throughout this section we will use the notation $x[c]$ to denote the version of x restricted to class c. For example, $N_{A,B}[c]$ will denote the number individuals in class c, each of whom were included in all sources of A and no sources of B, and hence $\sum_{c=1}^{C} N_{A,B}[c] = N_{A,B}$.

Marginalizing over the latent class-memberships we have the inclusion probabilities

$$\begin{aligned} u^*_{A,B} &= P(\mathcal{I}(j)_i = 1 \,\forall i \in A, \mathcal{I}(j)_{i'} = 0 \,\forall i' \in B) \\ &= \sum_{c=1}^{C} \pi[c] \prod_{i \in A} p_i[c] \prod_{i \in B} (1 - p_i[c]) \\ &= \sum_{c=1}^{C} \pi[c] u_{A,B}[c]. \end{aligned} \tag{23}$$

It is evident that this is a generalization of the inclusion probabilities in (4) and no longer assumes inclusions in the different listings are independent. In fact, any discrete distribution can be represented as (23) for some choice of C (Dunson and Xing (2009), Vermunt et al. (2008)). Arnold, Hayakawa and Yip (2010), Basu and Ebrahimi (2001), Norris and Pollock (1996) applied such finite mixture models to relax the independence assumption for capture-recapture data. If C is considered unknown and assigned a prior, Bayesian computation will need to rely on reversible jump MCMC (Green (1995)) type algorithms due to change in parameter dimensionality for different values of C. A more sophisticated solution that circumvents RJMCMC was proposed in Manrique-Vallier (2016). Akin to ideas in Dunson and Xing (2009) and Manrique-Vallier and Reiter (2014), a Bayesian nonparametric approach was adopted using Dirichlet process mixtures.

Our extension of capture-recapture to incomplete and misaligned data can be harmonized with this state-of-the-art approach of Manrique-Vallier (2016) for relaxing the independence assumption. We now outline this extension that combines the incompleteness model of Section 3.2 with the Dirichlet process prior for π of Manrique-Vallier (2016).

If all the counts $\{n_{A,A^c}[c] : c \in 1, \dots, C, A \subset \mathcal{S}\}$ were known, the joint likelihood will be given by

$$N! \prod_{c=1}^{C} \frac{\pi[c]^{N[c]} \prod_{i=1}^{S} p_i[c]^{n_i[c]} (1 - p_i[c])^{N[c] - n_i[c]}}{\prod_{A \subset \mathcal{S}^*} n_{A \cup \{S\}, \mathcal{S}^* \setminus A}[c]! \prod_{A \subset \mathcal{S}^*} n_{A,A^c}[c]!}. \tag{24}$$

However, none of the finest partition counts $n_{A,A^c}[c]$ are known, as they correspond to partitions based on a latent clustering variable. Hence, we formulate a Gibbs sampler to impute them. Note that none of the known quantities, that is, the marginal counts $n_i = \sum_{c=1}^{C} n_i[c]$ and the counts within the survey participants $V_S = \sum_{c=1}^{C} \sum_{A \subset \mathcal{S}^*} n_{A\cup\{S\},\mathcal{S}^*\setminus A}$ feature directly in (24) above. However, we can rewrite (24) as

$$\prod_{i=1}^{S} \binom{U_{i-1}}{V_i} \prod_{A\subset\mathcal{S}} \left((u^*_{A,A^c})^{n_{A,A^c}} \binom{n_{A,A^c}}{n_{A,A^c}[1],\dots,n_{A,A^c}[C]} \prod_{c=1}^{C} p^*_{A,A^c}[c] \right), \tag{25}$$

where $p^*_{A,B}[c] = u_{A,B}[c]/u^*_{A,B}$, and U_i's and V_i's are same as in Section 4.1. Recall that $W_i = \{n_{A\cup\{i\},B\cup\{i+1,\dots,S\}} \mid A \cup B = \{1,\dots,i-1\}, A \cap B = \{\}\}$ with $W_S = V_S$ and W_0 being the set of KP members who aren't included in any of the S listings.

We group the latent finest partition counts $n_{A,A^c}[c]$ as follows. Let W_i^* denote the matrix formed by stacking up the row vectors $(n_{A,B}[1],\dots,n_{A,B}[C])'$ for all (A,B) such that $n_{A,B} \in W_i$. In the Gibbs sampler, we update the tuples (V_i, W_i^*) for $i = 0,\dots,S$ and the cluster-and-source-specific inclusion probabilities $p_1[1],\dots,p_S[C]$. First, for $i = 0$ we update:

1. Generate $W_0 \sim \mathrm{NB}(r, 1 - u^*_{\{\},\mathcal{S}})$,
2. Update $V_0 = N = W_0 + r$,
3. Generate $W_0^* \mid W_0, \cdot \sim \mathrm{Multinomial}(W_0, p^*_{\{\},\mathcal{S}})$,

where $p^*_{A,A^c} = (p^*_{A,A^c}[1],\dots,p^*_{A,A^c}[c]))'$ and, as before, r is the total number of KP members included in at least one of the listings which is known given $V_1,\dots,V_S$. This update is exactly identical to the analogous update in Step (v) of Manrique-Vallier (2016). Similarly, assuming a Beta(a,b) prior for the inclusion probabilities $p_i[c]$'s and a Stick-breaking prior (Sethuraman (1994)) for $(\pi[1],\dots,\pi[C])'$, the updates for these parameters remain same as the analogous updates in Steps (ii), (iii) and (iv) of Manrique-Vallier (2016).

The remaining part of the Gibbs sampler, that is, the updates for (V_i, W_i^*) for $i \ge 1$, is entirely different from the sampler in Manrique-Vallier (2016) and extends the sampler in Section 4.1. This is because Manrique-Vallier (2016) assumes knowledge of complete inclusion histories of any individual listed in one of the S listings, which is not the case in our application and many other applications involving multiplier method, where only the inclusion histories of individuals taking the survey is known. The sampler in this section is set under this more general setting without assuming knowledge of complete inclusion histories. Hence, the need for the following Gibbs imputations of (V_i, W_i^*)'s.

As in Section 4.1, $V_i = W_i + \sum_{j=i+1}^{S} V_{ij}$. Given $N, V_1,\dots,V_{i-1},V_{i+1},\dots,V_S$ in the Gibbs sampler, $\sum_{j=i+1}^{S} V_{ij}$ is known, and we only need to sample W_i.

Note that

$$\begin{aligned}
\mathrm{pr}(W_i, W_i^* \mid \cdot) \propto{}& \binom{U_{i-1} - \sum_{j=i+1}^{S} (V_{ij} + \overline{V}_{ij})}{W_i} \\
&\times \prod_{(A,B):n_{A\cup\{i\},B}\in W_i} \left(\frac{u^*_{A\cup\{i\},B}}{u^*_{A,B\cup\{i\}}} \right)^{n_{A\cup\{i\},B}} \\
&\times \prod_{A:n_{A,A^c}\in W_i} \binom{n_{A,A^c}}{n_{A,A^c}[1],\dots,n_{A,A^c}[C]} \prod_{c=1}^{C} p^*_{A,A^c}[c].
\end{aligned}$$

The first two rows correspond to a noncentral multivariate hypergeometric distribution for W_i with the vector of odds $O_i = \{\frac{u^*_{A\cup\{i\},B}}{u^*_{A,B\cup\{i\}}} : (A, B) \ni n_{A\cup\{i\},B} \in W_i\}$; the last row corresponds to products of multinomial distribution for the rows of W_i^* given W_i. Hence, for $i = 1, \dots, S$, the Gibbs update for (V_i, W_i^*) is given by:

1. Generate $W_i \sim \mathrm{FNCMVHG}_{2^{i-1}}(n_i - \sum_{j=i+1}^{S} 1' V_{ij}, U_{i-1} - \sum_{j=i+1}^{S}(V_{ij} + \overline{V}_{ij}), O_i)$,
2. Update $V_i = W_i + \sum_{j=i+1}^{S} V_{ij}$,
3. Generate $W_i^* \mid W_i, \cdot \sim \prod_{A:n_{A,A^c} \in W_i} \mathrm{Multinomial}(n_{A,A^c}, p^*_{A,A^c})$.

As $W_S = V_S$ and V_1 are known, so we skip their updates and only update W_S^* and W_1^*. Note that if there was only one cluster, that is, we were using the independence model, then the odds O_i becomes proportional to the vector of ones. So, the noncentral distribution in Step 1 reduces the the central multivariate hypergeometric distribution of (15). Thus, the nonindependence model and the Gibbs sampler is a coherent generalization of the independence model.

8. Conclusion. We have presented uncertainty quantified size estimates of MSM and FSW for all four regions of eSwatini. The size estimates derived here can be used to characterize distributions of HIV risks, current gaps in coverage of HIV prevention and treatment and support dynamic HIV transmission models to estimate the contribution of the unmet HIV prevention and treatment needs among key populations to the HIV epidemic in eSWatini (Sabin et al. (2016), WHO (2010)).

The complex nature of the dataset, owing to the small number of sites with direct data, incompleteness and misalignment, shaped most of these analysis and motivated the methods development. Importantly, given similar dynamics of stigma affecting key populations in many countries with generalized HIV epidemic settings, the small data problem is a common one. In response, We proposed a general Bayesian multiplier method for generating a unified size estimate using multiplier data from multiple sources via a data augmented model. In the process we offered a Gibbs sampler based solution to sample from the distribution of $2 \times 2 \times \dots \times 2$ contingency tables with known marginals. The Bayesian implementation is fast and ensures proper uncertainty propagation leading to meaningful credible intervals for size estimates. It also allows borrowing of strength across multiple regions, in modeling the region specific percentages of MSM and FSW, which is essential for extrapolation. In our setup we used a simple exchangeable model to achieve this, but if more sites have direct data, then one can easily expand to a regression model. We also proposed a solution for misaligned data which helped us to disentangle the data at the Corridor to produce separate estimates for Mbabane and Manzini.

Despite the modeling innovations proposed here, the analysis relied on a number of assumptions which we highlight to ensure that the estimates are understood in the proper context.

Independence. As we have highlighted multiple times in Section 3.1, the assumption that inclusion or exclusion in each of the listings of a region are jointly independent is central to the capture-recapture model and Lincoln–Peterson estimates, their model-based extension to multiple data sources and our extension to incomplete and misaligned data. In our analyses we conducted tests to see that the independence assumption was not violated in the data. However, this may not be the case in general. In general, if inclusion in two sources are positively correlated, then the Lincoln–Peterson estimate will underestimate the true population size, and, if the sources are negatively correlated then it will overestimate. Hence, in Section 7 we provided a solution for the setting when this assumption is violated. We showed how we can seamlessly harmonize our model with a state-of-the-art solution to modeling nonindependent capture-recapture data using Bayesian nonparametric (Dirichlet process) priors (Manrique-Vallier (2016)). This ensures that our model can accommodate nonindependence in the most general form and remain computationally tractable. A set of simulation studies comparing the independent and the Dirichlet-process based dependent model is presented in Section S3 of the Supplementary Material (Datta et al. (2020)).

Immigration. Throughout the analysis we assumed a closed population, that is, no migration between the regions and into or out of eSwatini from other countries. This is unrealistic, especially as the Corridor, which informs the size estimates for both Hhohho and Manzini region, witnesses a lot of movement. Also Piggs Peak is located near the border with South Africa. Therefore, frequent immigration and emigration is common. Migration has not been accounted for in the analysis due to lack of data.

We did some theoretical and simulation studies to assess the impact of immigration on our model. The results, detailed in Section S1 of the Supplementary Material (Datta et al. (2020)), show that if the counts for a source gets inflated by a factor of c due to immigration, the corresponding inflation in the Lincoln–Peterson size estimate is only cp where p is the inclusion probability for that source. Just as an example, even if immigration inflates counts by 20% but inclusion probability in source 1 is 15%, the estimates will only be inflated by $20\% \times 15\% = 3\%$. Hence, unless both the inclusion probability and the immigration inflation factor for a source are very high, the impact on the size estimates are moderate.

Age extrapolation. There are two age related assumptions in the analysis:

1. Engagement age-range: Our first assumption is about the age-specific participation of the MSM or FSW community in the different listings. We assume that the age distribution of survey participants matches age distribution of those included in the other sources. Based on the empirically predominant age-group in the survey, this amounted to assuming that for each data source, among the FSW (or MSM) members the same age-group (18–35 for FSW, 18–32 for MSM) participate in all the listings.

This assumption is a function of the structure and availability of the data. Most of the nonsurvey sources generally do not have age-resolved data but just a summary count. The survey is the only listing with information about age distribution. Hence, assuming the age distribution of participants in the other listings (which has no age-data) to be same as the age distribution of the survey is a convenient assumption.

However, this assumption is not unreasonable; as given similar methods of recruitment and engagement of MSM or FSW in research and program, it is very likely that community members included in the survey are very similar to those engaging in the other programs. Also, typically, the issue of engagement is not limited to the survey and tends to be the case in general for the program. This is to say that the age group dominating the survey likely represents the age group that are also the majority participants in the other listings. Hence, assuming that the different engagements have similar age coverage is not unreasonable in such data-scarce settings (see Weir et al. (2018), for more discussion on this).

We can impose this assumption directly on the empirical counts or use it via modeling of the population-level age-specific inclusion probabilities. We did the former, as the likelihood due to the latter adds more parameters and does not facilitate convenient Gibbs updates. However, the latter version of the assumption is weaker, as it postulates that, although the population-level age-specific odds ratios of being included in a listing are same across listings, the empirical age distributions may not be exactly same. We offer some heuristics in Section S2 of the Supplementary Material (Datta et al. (2020)) on why this weaker assumption would lead to very similar size estimates as the stronger one.

We reemphasize that we do not assume that the proportion of FSW (or MSM) who participate in a listing is constant across all ages. This would be an inappropriate assumption as participations vary by age.

Finally, to generalize our method to other real world applications in settings where we cannot make any assumption about similarity in age distributions among different listings, we could simply consider age to be the latent stratifying factor and use the extension we proposed to the mixture model framework of Section 7 to accommodate nonindependence and heterogeneity for obtaining the size estimates.

2. Extrapolation beyond the engagement age-range: The second assumption is related to extrapolation of estimates to age-groups other than those represented in the data. The assumption is that, among all women (or men), the proportion who are FSW (or MSM) are same within and outside the engagement age-range.

Since the survey participants predominantly belonged to the engagement age ranges of 18–32 for MSM and 18–35 for FSW, in order to extrapolate to the broader age group of 15–49 years, we rely on this assumption which in many settings is a reasonable one for MSM but may not be for FSW. This is because sexual orientation is a fixed trait whereas an occupation is less so. However, in this study there is no age-resolved data to model the fractions as functions of age for the extrapolation. Related studies suggest that the percent that remain selling sex is pretty consistent across age groups, and this assumption is used to extrapolate from the age group representing the data collection to the broader age group of 15–49 years (Holland et al. (2015, 2016)). If however, this is not true and the percentage of FSW in the age group 35–49 is less (or more) than the percentage in the age group 18–35, then these estimates are overestimating (or underestimating) the FSW population size.

Geographical extrapolation. We only have data from *four* or *five* geographically separated sites with which it is not possible to model a spatial correlation among the MSM or FSW fractions across Tinkhundlas that can then be used for geographical extrapolation. Hence, we resorted, to generate estimates for the entire regions based on direct estimates at the few sites. We hypothesized that the MSM or FSW proportions in the Tinkhundlas were similar to the respective proportions at their closest sites and used a spatial autoregression model where the mean estimate of MSM or FSW fraction for a Tinkhundla without data was modeled as the average of those of its *neighboring* Tinkhundlas (see Section 5.2 for more details).

In summary, all these assumptions were needed given limited primary data of key populations reinforcing the need to effectively study the needs of those at highest risk of HIV even in the context of generalized epidemics. The model-based approach also ensures impact of all these assumptions are well understood. These estimates are not authoritative, but they do offer significantly more statistical rigor over common practices, like wisdom-of-the-crowd estimates or average multiplier method, which we have shown to be inferior to our approach. Additionally, we offer a solution to the common problem of geographical misalignment in such datasets, in presence of which even these naive average multiplier method estimates cannot be evaluated. Future programmatic and survey data focused on key populations in eSwatini can be used to corroborate or refute these conjectures and ultimately better describe the distributions of HIV risks in countries with the most generalized HIV epidemics.

APPENDIX: TOTAL POPULATION NUMBERS

This section details the rationale and process of calculating the denominators $P^{(s)}$, the total male or female population in the 90th percentile age range for each location. To do this, we used preliminary results from the 2017 eSwatini census and interpolated the population number to get estimates for $P^{(s)}$ in 2014, when the survey was conducted.

This process is summarized in the following steps. We first identified where survey respondents came from for each of the data collection sites. These reported locations were then mapped to the corresponding Tinkhundlas which contain those locations. The total population representing each site was initially calculated as a weighted average of the 2017 population of these Tinkhundlas from which the participants came from, with the weights being proportional to the number of participants from that Tinkhundla. We then calculated the 90th percentile age of survey respondents for each survey site. This turned out to be 18–32 years for MSM and 18–35 years for FSW. We calculated the region-specific proportions of individuals in each region that fall in these age ranges and adjusted the total population using these proportions. Finally, using the 2017 population numbers, we extrapolated backward in time to obtain the population numbers in 2014. To do this, we used annual growth rate data for each region, using census data from 2007 and 2017, and used these average annual growth rates to interpolate our weighted population number from 2017 to 2014. All the information used in this process, including the region specific annual growth rates, proportion of population within the age range corresponding to 90% of the survey participants, the reported locations of the participants along with the corresponding Tinkhundlas and the Tinkhundla-specific weights used to obtain the site-specific total population numbers are in Tables 10 and 11. The final population numbers representative of each site are provided in Table 12.

Acknowledgments. We thank the Editors and two anonymous reviewers for their thoughtful comments and feedbacks which improved the manuscript. The funding for data collection in eSwatini was funded by PEPFAR through the USAID eSwatini (GHH-I-00-07-00032-00). Abhirup Datta and Andrew Pita's efforts for methods development and implementation were supported by Project SOAR (Supporting Operational AIDS Research), Cooperative Agreement AID-0AA-A-14–00060, is made possible by the generous support of the American people through the President's Emergency Plan for AIDS Relief (PEPFAR) and the United States Agency for International Development (USAID). Stefan Baral and Amrita Rao's efforts were supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01AI136664. Analyses were supported by Project SOAR (Supporting Operational AIDS Research), Cooperative Agreement AID-0AA-A-14–00060, is made possible by the generous support of the American people through the President's Emergency Plan for AIDS Relief (PEPFAR) and the United States Agency for International Development (USAID). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "Size estimation of key populations in the HIV epidemic in eSwatini using incomplete and misaligned capture-recapture data" (DOI: 10.1214/20-AOAS1327SUPP; .pdf). Contains discussions on impact of unaccounted immigration, age-related assumptions, simulation study for dependent data sources comparing the main method and the extension in Section 7 using Dirichlet processes, additional simulation studies parallel to those in Section 6.1 but for FSW, sensitivity of the data analysis with respect to choice of priors for the size estimates and choice of Beta hyperparameters.

TABLE 10

MSM: *Region-specific age-group proportion*, *annual growth rates*, *counts and weights used for reported locations assigned to each Tinkhundla in determining the total population representing the survey at each site*

Region	Percent of Males (18–32)	Annual Growth Rate	Site	Locations reported	Counts	Tinkhundla	Weight
Manzini	30%	1.1%	Manzini	Manzini/Matsapha	173	Manzini North and South, Kwaluseni	173/177
				Malkerns	3	Lobamba Lombdzala	4/177
				Mahlanya	1		
Hhohho	29%	1.3%	Mbabane	Mbabane/Ezulwini	155	Mbabane East and West, Lobamba	155/223
				Ngwenya	61	Motjane	68/223
				Oshoek	3		
				Motshane	4		
			Piggs Peak	Piggs Peak	54	Piggs Peak	54/57
				Matsamo	3	Timphisini, Ntfonjeni	3/57
Shiselweni	26%	−0.2%	Nhlangano	Nhlangano	70	Mbangweni	1

TABLE 11

FSW: Region-specific age-group proportion, annual growth rates, counts and weights used for reported locations assigned to each Tinkhundla in determining the total population representing the survey at each site

Region	Percent of Females (18–32)	Annual Growth Rate	Site	Locations reported	Counts	Tinkhundla	Weight
Manzini	35%	1.1%	Manzini	Manzini/Matsapha	249	Manzini North and South, Kwaluseni	249/257
				Malkerns	8	Lobamba Lombdzala	8/257
Hhohho	32%	1.3%	Mbabane	Mbabane/Ezulwini	198	Mbabane East and West, Lobamba	198/255
				Ngwenya	55	Motjane	57/255
				Oshoek	2		
			Piggs Peak	Piggs Peak	121	Piggs Peak	121/127
				Matsamo	2	Timphisini	2/127
				Buhleni	4	Mayiwane	4/127
Shiselweni	29%	−0.2%	Nhlangano	Nhlangano	47	Mbangweni	1
			Lavumisa	Lavumisa	81	Somntongo	81/88
				Hluthi	3	Hosea	3/88
				Matsanjeni	4	Matsanejni South	4/88

TABLE 12
Male and female population numbers representative of the survey participants at each of the data collection sites

Site	Male 18–32 population	Female 18–35 population
Lavumisa	–	1383
Piggs Peak	2940	2962
Mbabane	10,295	16,796
Manzini	16,324	20,323
Nhlangano	4038	4988

REFERENCES

ARNOLD, R., HAYAKAWA, Y. and YIP, P. (2010). Capture-recapture estimation using finite mixtures of arbitrary dimension. *Biometrics* **66** 644–655. MR2758846 https://doi.org/10.1111/j.1541-0420.2009.01289.x

BAO, L., RAFTERY, A. E. and REDDY, A. (2015). Estimating the sizes of populations at risk of HIV infection from multiple data sources using a Bayesian hierarchical model. *Stat. Interface* **8** 125–136. MR3322160 https://doi.org/10.4310/SII.2015.v8.n2.a1

BARAL, S. D., KETENDE, S., MNISI, Z., MABUZA, X., GROSSO, A., SITHOLE, B., MAZIYA, S., KERRIGAN, D. L., GREEN, J. L. et al. (2013). A cross-sectional assessment of the burden of HIV and associated individual- and structural-level characteristics among men who have sex with men in Swaziland. *J. Int. AIDS Soc.* **16** 18768.

BARAL, S., KETENDE, S., GREEN, J. L., CHEN, P. A., GROSSO, A., SITHOLE, B., NTSHANGASE, C., YAM, E., KERRIGAN, D. et al. (2014). Reconceptualizing the HIV epidemiology and prevention needs of female sex workers (FSW) in Swaziland. *PLoS ONE* **9** e115465.

BASU, S. and EBRAHIMI, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika* **88** 269–279. MR1841274 https://doi.org/10.1093/biomet/88.1.269

BERGER, B. O., GROSSO, A., ADAMS, D., KETENDE, S., SITHOLE, B., MABUZA, X. S., MAVIMBELA, M. J. and BARAL, S. (2018). The prevalence and correlates of physical and sexual violence affecting female sex workers in Swaziland. *J. Interpers. Violence* **33** 2745–2766. https://doi.org/10.1177/0886260516629385

BROWN, C. A., GROSSO, A. L., ADAMS, D., SITHOLE, B., KETENDE, S., GREENE, J., MAZIYA, S., MNISI, Z., MABUZA, X. et al. (2016). Characterizing the individual, social, and structural determinants of condom use among men who have sex with men in Swaziland. *AIDS Res. Hum. Retrovir.* **32** 539–546. https://doi.org/10.1089/AID.2015.0240

CASTLEDINE, B. J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika* **68** 197–210. MR0614956 https://doi.org/10.1093/biomet/68.1.197

CHEN, Y., DIACONIS, P., HOLMES, S. P. and LIU, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *J. Amer. Statist. Assoc.* **100** 109–120. MR2156822 https://doi.org/10.1198/016214504000001303

DARROCH, J. N. (1958). The multiple-recapture census. I. Estimation of a closed population. *Biometrika* **45** 343–359. MR0119360 https://doi.org/10.2307/2333183

DARROCH, J. N. (1959). The multiple-recapture census. II. Estimation when there is immigration or death. *Biometrika* **46** 336–351. MR0119361 https://doi.org/10.2307/2333531

DATTA, A., LIN, W., RAO, A., DIOUF, D., KOUAME, A., EDWARDS, J. K., BAO, L., LOUIS, T. A. and BARAL, S. (2019). Bayesian estimation of MSM population in Côte d'Ivoire. *Stat. Public Policy* **6** 1–13. https://doi.org/10.1080/2330443X.2018.1546634.

DATTA, A., BANERJEE, S., HODGES, J. S. and GAO, L. (2019). Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models. *Bayesian Anal.* **14** 1221–1244. MR4044851 https://doi.org/10.1214/19-BA1177

DATTA, A., PITA, A., RAO, A., SITHOLE, B., MNISI, Z. and BARAL, S. (2020). Supplement to "Size estimation of key populations in the HIV epidemic in eSwatini using incomplete and misaligned capture-recapture data." https://doi.org/10.1214/20-AOAS1327SUPP.

DIACONIS, P. and STURMFELS, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** 363–397. MR1608156 https://doi.org/10.1214/aos/1030563990

DOBRA, A. (2012). Dynamic Markov bases. *J. Comput. Graph. Statist.* **21** 496–517. MR2945478 https://doi.org/10.1080/10618600.2012.663285

DOBRA, A., TEBALDI, C. and WEST, M. (2006). Data augmentation in multi-way contingency tables with fixed marginal totals. *J. Statist. Plann. Inference* **136** 355–372. MR2211344 https://doi.org/10.1016/j.jspi.2004.07.002

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004 https://doi.org/10.1198/jasa.2009.tm08439

EDWARDS, J. K., HILEMAN, S., DONASTORG, Y., ZADROZNY, S., BARAL, S., HARGREAVES, J. R., FEARON, E., ZHAO, J., DATTA, A. and WEIR, S. S. (2009). Estimating sizes of key populations at the national level: Considerations for study design and analysis. *Epidemiology* **29** 795–803. https://doi.org/doi:10.1097/EDE.0000000000000906

FEARON, E., CHABATA, S. T., THOMPSON, J. A., COWAN, F. M. and HARGREAVES, J. R. (2017). Sample size calculations for population size estimation studies using multiplier methods with respondent-driven sampling surveys. *JMIR Public Health Surveill.* **3** e59. https://doi.org/10.2196/publichealth.7909

FIELDING-MILLER, R., MNISI, Z., ADAMS, D., BARAL, S. and KENNEDY, C. (2014). "There is hunger in my community": A qualitative study of food security as a cyclical force in sex work in Swaziland. *BMC Public Health* **14** 79.

FOG, A. (2015). BiasedUrn: Biased Urn Model Distributions. R package version 1.07.

FONNER, V. A., KERRIGAN, D., MNISI, Z., KETENDE, S., KENNEDY, C. E. and BARAL, S. (2014). Social cohesion, social participation, and HIV related risk among female sex workers in Swaziland. *PLoS ONE* **9** e87527. https://doi.org/10.1371/journal.pone.0087527

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*. CRC Press/CRC, Boca Raton.

GEORGE, E. I. and ROBERT, C. P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika* **79** 677–683. MR1209469 https://doi.org/10.2307/2337223

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810 https://doi.org/10.1093/biomet/82.4.711

GROVER, E., GROSSO, A., KETENDE, S., KENNEDY, C., FONNER, V., ADAMS, D., SITHOLE, B., MNISI, Z., MAZIYA, S. L. et al. (2016). Social cohesion, social participation and HIV testing among men who have sex with men in Swaziland. *AIDS Care* **28** 795–804. https://doi.org/10.1080/09540121.2015.1131971

HECKATHORN, D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Soc. Probl.* **44** 174–199.

HOLLAND, C. E., PAPWORTH, E., BILLONG, S. C., TAMOUFE, U., LEBRETON, M., KAMLA, A., ELOUNDOU, J., MANAGA, F., YOMB, Y. et al. (2015). Antiretroviral treatment coverage for men who have sex with men and female sex workers living with HIV in Cameroon. *J. Acquir. Immune Defic. Syndr.* **68** S232–S240.

HOLLAND, C. E., KOUANDA, S., LOUGUÉ, M., PITCHE, V. P., SCHWARTZ, S., ANATO, S., OUEDRAOGO, H. G., TCHALLA, J., YAH, C. S. et al. (2016). Using population-size estimation and cross-sectional survey methods to evaluate HIV service coverage among key populations in Burkina Faso and Togo. *Public Health Rep.* **131** 773–782. https://doi.org/10.1177/0033354916677237

KENNEDY, C. E., BARAL, S. D., FIELDING-MILLER, R., ADAMS, D., DLUDLU, P., SITHOLE, B., FONNER, V. A., MNISI, Z. and KERRIGAN, D. (2013). "They are human beings, they are Swazi": Intersecting stigmas and the positive health, dignity and prevention needs of HIV-positive men who have sex with men in Swaziland. *J. Int. AIDS Soc.* **16** 18749.

LINCOLN, F. C. et al. (1930). Calculating waterfowl abundance on the basis of banding returns.

LOGIE, C. H., DIAS, L. V., JENKINSON, J., NEWMAN, P. A., MACKENZIE, R. K., MOTHOPENG, T., MADAU, V., RANOTSI, A., NHLENGETHWA, W. et al. (2018a). Exploring the potential of participatory theatre to reduce stigma and promote health equity for lesbian, gay, bisexual, and transgender (LGBT) people in Swaziland and Lesotho. *Health Educ. Behav.* **46** 146–156.

LOGIE, C. H., PEREZ-BRUMER, A., JENKINSON, J., MADAU, V., NHLENGETHWA, W. and BARAL, S. (2018b). Marginalization and social change processes among lesbian, gay, bisexual and transgender persons in Swaziland: Implications for HIV prevention. *AIDS Care* **30** 33–40.

MANRIQUE-VALLIER, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics* **72** 1246–1254. MR3591609 https://doi.org/10.1111/biom.12502

MANRIQUE-VALLIER, D. and FIENBERG, S. E. (2008). Population size estimation using individual level mixture models. *Biom. J.* **50** 1051–1063. MR2649394 https://doi.org/10.1002/bimj.200810448

MANRIQUE-VALLIER, D. and REITER, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Statist.* **23** 1061–1079. MR3270711 https://doi.org/10.1080/10618600.2013.844700

NORRIS, J. L. and POLLOCK, K. H. (1996). Nonparametric mle under two closed capture-recapture models with heterogeneity. *Biometrics* 639–649.

PETERSEN, C. G. J. (1896). The yearly immigration of young plaice in the Limfjord from the German Sea. *Rep. Dan. Biol. Stn.* **6** 1–48.

RAO, A., STAHLMAN, S., HARGREAVES, J., WEIR, S., EDWARDS, J., RICE, B., KOCHELANI, D., MAVIMBELA, M. and BARAL, S. (2017). Sampling key populations for HIV surveillance: Results from eight cross-sectional studies using respondent-driven sampling and venue-based snowball sampling. *JMIR Public Health Surveill*. **3** e72. https://doi.org/10.2196/publichealth.8116

RISHER, K., ADAMS, D., SITHOLE, B., KETENDE, S., KENNEDY, C., MNISI, Z., MABUSA, X. and BARAL, S. D. (2013). Sexual stigma and discrimination as barriers to seeking appropriate healthcare among men who have sex with men in Swaziland. *J. Int. AIDS Soc*. **16** 18715. https://doi.org/10.7448/IAS.16.3.18715

SABIN, K., ZHAO, J., CALLEJA, J. M. G., SHENG, Y., GARCIA, S. A., REINISCH, A. and KOMATSU, R. (2016). Availability and quality of size estimations of female sex workers, men who have sex with men, people who inject drugs and transgender women in low- and middle-income countries. *PLoS ONE* **11** e0155150. https://doi.org/10.1371/journal.pone.0155150

SEKAR, C. C. and DEMING, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *J. Amer. Statist. Assoc*. **44** 101–115.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SMITH, P. W., FORSTER, J. J. and MCDONALD, J. W. (1996). Monte Carlo exact tests for square contingency tables. *J. Roy. Statist. Soc. Ser. A* 309–321.

THE PHIA PROJECT, I. (2017). Swaziland HIV incidence measurement survey 2: A population-based HIV impact assessment. Report.

VERMUNT, J. K., VAN GINKEL, J. R., VAN DER ARK, L. A. and SIJTSMA, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociol. Method*. **38** 369–397.

WEIR, S., TATE, J., HILEMAN, S., KHAN, M. and JACKSON, E. (2005). Place: priorities for local AIDS control efforts: A manual for implementing the place method.

WEIR, S. S., BARAL, S. D., EDWARDS, J. K., ZADROZNY, S., HARGREAVES, J., ZHAO, J. and SABIN, K. (2018). Opportunities for enhanced strategic use of surveys, medical records, and program data for HIV surveillance of key populations: Scoping review. *JMIR Public Health Surveill*. **4** e28. https://doi.org/10.2196/publichealth.8042

WHO (2010). Guidelines on estimating the size of populations most at risk to HIV. Report, UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance.

YAM, E. A., MNISI, Z., SITHOLE, B., KENNEDY, C., KERRIGAN, D. L., TSUI, A. O. and BARAL, S. (2013). Association between condom use and use of other contraceptive methods among female sex workers in Swaziland: A relationship-level analysis of condom and contraceptive use. *Sex. Transm. Dis*. **40** 406–412. https://doi.org/10.1097/OLQ.0b013e318283c16d