Published in final edited form as:
Soc Networks. 2022 May ; 69: 74–83. doi:10.1016/j.socnet.2020.05.001.

Using geographical data and rolling statistics for diagnostics of respondent-driven sampling

Brian Kim[a,*], **Moses Ogwal**[b], **Enos Sande**[c], **Herbert Kiyingi**[c], **David Serwadda**[b], **Wolfgang Hladik**[c]

[a]Joint Program in Survey Methodology, University of Maryland, 1218 LeFrak Hall, 7251 Preinkert Dr., College Park, MD 20742, USA

[b]Makerere University School of Public Health, Old Mulago Hill Road, New Mulago Hospital Complex, P.O. Box 7072, Kampala, Uganda

[c]Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333, USA

Abstract

Respondent-driven sampling (RDS) is commonly used to sample from key populations without a sampling frame since traditional methods are unable to efficiently survey them. Surveying these populations is often desirable to inform service delivery, assess effectiveness of programs, and determine prevalence of diseases. However, there are concerns about how RDS works in practice due to its many assumptions. To assess some of these assumptions, we develop diagnostics using geographical data and demonstrate their utility by identifying lack of convergence and characterizing RDS reach in surveys conducted among female sex workers and men who have sex with men in Kampala, Uganda.

Keywords

Network sampling; Key populations; Acquired immune deficiency syndrome; Convex Hull

1. Introduction

Key populations, such as female sex workers (FSW) and men who have sex with men (MSM), are difficult to survey because of the lack of a clearly defined sampling frame. In many cases, surveying these populations is desirable so that program planners, or health service providers can provide targeted services, assess the effectiveness of their programs, and determine the prevalence of diseases (Lansky et al., 2007; UNAIDS and Organization, 2010; Johnston et al., 2013). In 2014, the Joint United Programme on HIV/AIDS (UNAIDS) set three targets, called the 90-90-90 targets, to achieve by 2020: that 90% of all people living with HIV will know their status, 90% of all people with diagnosed HIV infection will receive sustained antiretroviral therapy, and 90% of all people receiving antiretroviral therapy will have viral suppression. Surveying key populations is essential to properly

[*]Corresponding author. kimbrian@umd.edu (B. Kim).

characterizing the progress toward reaching the 90-90-90 targets (UNAIDS, 2014; Hladik et al., 2017).

A link-tracing network sampling method called respondent-driven sampling (RDS) has been widely used to study hard-to-reach populations due to the hidden nature of these populations (Heckathorn, 1997; Johnston et al., 2006, 2015; Malekinejad et al., 2008). RDS starts with a small sample (usually a convenience sample) of initial respondents selected by investigators called seeds. These seeds are given a small number of coupons (typically 2–5) which they can give to other members of the population. Those who receive the coupons may present themselves at the survey office to be added to the study. After participating in the survey, individuals are given coupons to recruit others into the study. This process continues, growing the recruitment chain over successive waves (each step in the recruitment chain after the seed). Participants receive incentives both for participating in the study and for recruiting others into the study (Heckathorn, 1997).

RDS has become more widely adopted because it allows researchers to survey hard-to-reach populations without a sampling frame (Malekinejad et al., 2008). Unlike another link-tracing sampling method called snowball sampling in which respondents are asked to nominate possible recruits to the researcher, in RDS, respondents themselves are responsible for recruiting others into the study by handing out the coupons that they receive. This means that respondents are not required to divulge any information about others in the population, lessening confidentiality concerns, though more of the burden of recruitment falls on the respondent (Heckathorn, 1997).

RDS is not without its limitations. Because the respondents themselves are essentially in charge of the sampling, we must make many assumptions not only about the sampling process, but also about the network structure of the overall population (Heckathorn, 1997; Gile and Handcock, 2010; Gile et al., 2015). For example, respondents may choose to only give coupons to people they are closest to, which could oversample from certain subgroups due to homophily effects Paquette et al. (2011). Similarly, respondents may only know others who are similar to themselves, leaving few options in who they can recruit. Respondents may also recruit as is convenient, possibly resulting in samples that do not cover the study area. Since estimation methods using RDS make many sampling and network assumptions about the under-lying RDS surveys, violations of these assumptions might call into question the validity of estimates based on RDS data (Wang et al., 2005; Handcock and Gile, 2011; Salganik et al., 2011; White et al., 2012; Salganik, 2012; Gile et al., 2015). Gile et al. (2015) note that traditional statistical diagnostic tests are not appropriate because of the unknown dependence between recruiter and recruit. Therefore, we must develop new diagnostic measures that are both informative about the RDS process as well as easy to use during the actual data collection to help researchers as they run their studies. While previous work on RDS diagnostics has been done by Gile et al. (2015), Lansky et al. (2012), Liu et al. (2012), there are still gaps in the diagnostics, particularly in exploring the assumption that all units within the population are connected. To date, there have been no RDS diagnostics developed that use respondent latitude and longitude information.

The aim of this paper is to build on existing diagnostics of RDS performance by showing how diagnostics with geographical data can be used to provide more insight into the RDS process than information collected through typical survey questions. We start by taking some of the diagnostics first developed by Gile et al. (2015) — convergence plots which show the change in a weighted statistic as additional respondents are added, as well as bottleneck plots which show the same, except split up by the recruitment trees started by each seed – and expanding on them by looking at rolling values of statistics in addition to the cumulative values. Then, we explore the use of these diagnostics using geographic data, using the idea of a convex hull to track how much of the city an RDS survey traverses using both the cumulative and rolling average versions of the convergence and bottleneck plot. To demonstrate the utility of these diagnostic plots, we apply these to FSW and MSM in Kampala, Uganda, two survey populations for which geographic data were collected.

In Section 2, we introduce the data that we will be looking at. In Section 3, we discuss the assumptions in RDS. In Section 4, we develop the diagnostics and apply them to the sampled FSW and MSM populations. In Section 5, we include a discussion of some limitations of these diagnostics and in Section 6, we make concluding remarks.

2. Data

To develop the RDS diagnostics in this paper, we used data collected in the 2012 Crane Survey, a key population focused survey in Kampala, Uganda run in collaboration between the Makerere University School of Public Health, the Ministry of Health, and the Centers for Disease Control and Prevention and funded by the US Government President's Emergency Plan for AIDS Relief (Hladik et al., 2017; Doshi et al., 2018). The Crane Survey used RDS to survey key populations at high risk for HIV, including female sex workers (FSW), men who have sex with men (MSM), drug users (DUS), and other populations. For our applications, we focused on FSW and MSM since they had the largest sample sizes, and we were able to create more substantial examples using them. In this paper, we focus on two groups from the Crane Survey: FSW and MSM. Table 1 shows the characteristics of the RDS run as part of the Crane survey for both FSW and MSM. The eligibility criteria for FSW was women 15 years or older[1] residing in greater Kampala who had sold sex to men in the past 6 months, and the eligibility criteria for MSM were biological males 18 years or older residing in greater Kampala who had had anal sex with another man within the last six months. Exclusion criteria included being unable to understand the language (Luganda or English), coupon receipt from a stranger (rather than someone known to them), or being from outside the study area of greater Kampala.

For the majority of participants, the data were collected using audio computer assisted self-interview (ACASI) in Luganda or English. ACASI is an survey data collection system in which respondents listen to questions through headphones and mark answers on a computer. This allowed respondents to answer questions in a secure, private setting compared to face-to-face interviews, allowing them to answer sensitive questions with more candor. Data were collected on demographics, HIV-related behavior, STD symptoms, as well as

[1] Per study protocol, FSW under the age of 18 were referred to services for trafficked and exploited minors.

various HIV-related knowledge, attitudes and practices, and mental health questions. The key innovation that provides much of the focus of the paper was the geographical data collected in these surveys. In a separate face-to-face interview, respondents were also asked to place a pin on an offline Google Map of Kampala marking their primary location of occupation (FSW) or socialization (MSM). Staff ensured geographic orientation by pointing out the survey office location as well as well-known Kampala landmarks. The maps that the respondents were shown were clean and did not contain any pins placed by previous respondents. The latitude and longitude values of the pin locations were then recorded. We note that though the latitude and longitude values calculated were quite exact, the pin locations themselves might not be, since respondents were simply asked to place the pin by sight. We did not require a greater level of precision than this, because in our applications, we focused on how the recruitment throughout the city as a whole. If needed, staff assisted respondents who had trouble finding “their” locations on the map.

For our applications, respondents who had pin locations outside of the city boundaries of Kampala, Uganda were also omitted, even though members who resided in greater Kampala were included in the overall survey. Respondents were only asked to record publicly accessible locations such as commercial venues or open air street locations. No private residential information was recorded. Collected geographic data were stored on password restricted computers accessible to survey staff and investigators on a need-to-know basis. Further discussion of how researchers may choose to approach collecting geographic data in general is provided in Section 4.5.

3. Respondent-driven sampling assumptions

Even though RDS is useful for surveying hidden populations, there are many assumptions that must be made to use RDS data for prevalence and size estimates. In this paper, we will be focusing on two key types of assumptions: seed dependence and population structure.

One of the main ideas behind employing RDS was that having long chains of recruitment reduced the dependence of the samples on the initial seeds (Heckathorn, 1997). This seed dependence can be affected by the existence of recruitment homophily (where members of the social network tend to form ties with other members similar in characteristics to themselves) and of bottlenecks (where two parts of a social network are separated by just a few nodes). Though the samples might be influenced by the seeds to begin with, estimators typically assume that long chains remove seed dependence. The samples can then be considered to have been sampled independent of the initial seed, and assume the bias that comes with the initial convenience sample has been eliminated. That is, even though the seeds are selected from a convenience sample, the final sample is expected to be representative of the full population (Volz and Heckathorn, 2008; Gile and Handcock, 2010; Heckathorn, 2011; Gile et al., 2015).

Separate, but related, is the issue of having a clearly defined population that is being sampled. Notably, when using RDS, the geographical boundaries of the population can become uncertain. Since the respondents are put in charge of finding new candidate participants, they may recruit members from outside the study area, or only members within a certain subset of the study area (e.g. a particular community or neighborhood). While it is relatively easy to exclude those who report being outside of the geographical area of interest (e.g. a city or metropolitan area), finding out whether the sampling is able to reach everyone within the city is harder.

This can be thought of as a sort of selection error—the error that results as respondents self-select into the study. Even though RDS is designed specifically around self-selection as candidate participants choose to go to the study site and be included, we still have to be wary of a combination of coverage error and non-response error affecting our estimates. While we do not have a sampling frame, we do have a target population, and we are assuming that RDS could potentially reach everyone in this target population, and that we are able to exclude people who are not in this population.

In this way, we have some assumptions about the overall population structure that, in most applications, are simply assumed to be true (Gile et al., 2015). First, we assume that there is a connected network of individuals in the population. Sometimes, this is close to trivial—in the case of men who have sex with men (MSM), for example, we are only interested in the population as it relates to the heightened risk for HIV. Therefore, missing out on a pair of exclusive partners who do not have any contact with others and do not have HIV might be acceptable. However, it is possible that, due to some large social or geographical barriers, there are several large groups of people within the same city who do not have any connection to each other. In this case, an RDS chain that starts with a seed in one part of the city would (theoretically) not be able to traverse to everyone in the population, and the real population it would be drawing from would be the subset that was connected.

We note that this last point is important not only as it relates to estimating population parameters such as HIV and other STI prevalence measures, but also as it relates to population size estimation using RDS data (see Handcock et al., 2015; Crawford et al., 2018 for population size estimation methods using RDS data). While it is possible that, if the disjoint components in the population network have similar population characteristics such as HIV prevalence, the estimates could quite possibly be unaffected, population size estimation is definitely affected by the disjoint nature of the population, and identifying whether the RDS is capable of reaching all members of the population is crucially important in determining what exactly the size estimate represents. It is also quite unlikely that disjoint populations would also exhibit the same exact prevalence characteristics in reality.

4. Diagnostics for RDS performance

Gile et al. (2015) recommend using convergence plots and bottleneck plots to determine whether the RDS recruitment chains reach convergence (i.e. they reach a point at which there is little to no seed dependence) and whether the seeds converge to the same value (i.e. the bottleneck effect is weak enough that different seeds are not sampling from effectively different sub-populations). Convergence plots show the sample size on the horizontal axis and (typically) a weighted statistic on the vertical axis. In this way, the convergence plot shows the value of the sample statistic as each subsequent respondent is added to the study, and as such, the order in which the sample statistics are shown is the chronological order.

Because of this, we might have a case in which respondents in different waves (that is, number of steps removed from the seed) enter the study at a similar point in time. Bottleneck plots are similar to convergence plots, but plotted by seed so that differences across the various chains can be observed.

These are two of the plots recommended by Gile et al. (2015) and we refer to them as the cumulative convergence and bottleneck plots. In addition to these, we will look at sample statistics with a rolling window (as opposed to finding the cumulative statistic), helping us see whether most of the recruitment of people with certain characteristics is happening earlier or later in the sample. This achieves a similar goal as the convergence plots, but allows for more clarity in changes. The added information is particularly apparent when using the convex hull introduced in Section 4.4. In all cases, for the rolling average, we used a window of size 25. That is, each value of the average is calculated using the most recent 25 respondents at that point in time. We note that while this number was chosen arbitrarily, one could very easily use different values for the size of the rolling window, which would, in essence, change the smoothness of the rolling average plots. In our diagnostic plots, we did not observe any difference besides smoothness when varying the size of the rolling window. In applications with other RDS data, we suggest trying different values to find one that shows trends most clearly.

4.1. Diagnostics using respondent characteristics

In general, many different characteristics about respondents may be collected during the survey process. Demographic information such as age or sex is commonly measured, and these are also characteristics which we might worry about for homophily. The goal of these diagnostics is to check whether homophily effects may be biasing our results, and drawing samples from only a sub-group within the overall population. Previous studies have shown that RDS could be vulnerable to homophily effects by age, sex, or ethnicity (Uuskula et al., 2010; Paquette et al., 2011; Phillips II et al., 2014), which is why we would want to use these variables to create our diagnostics.

The choice of respondent characteristic for this section is quite population-dependent. That is, these should be variables on which researchers might expect a high level of recruitment homophily, and might be worried about. We use age in our application as an example, but there are many other demographic characteristics that could be used. For example, in a population that does not have a homogenous sex by definition (as is the case with both FSW and MSM), the proportion of females or males recruited might be used as well. Depending on the study area, local cultural characteristics might be considered, such as a measure of social class, because it is possible the recruitment might only happen in one direction. That is, people from a lower social class may be unable to recruit others in a higher social class, eventually resulting in samples from only the lowest social class.

In the next section, we first demonstrate how rolling statistics may be used by analyzing diagnostics with respondent age, as it was one of the key substantial demographic characteristics that was collected in these studies. In Section 4.4, we introduce diagnostics with geographical information and expand on how rolling statistics may be used in that context.

4.2. Applying RDS diagnostics with age

Fig. 1 shows the convergence and rolling plots for the average age of FSW, with the dashed line representing the final overall value of average age. From the convergence plot (bottom), we see that average age starts out relatively low, before going up and ending at the dashed line representing the overall average. The rolling plot (top) shows a similar trend, with the first half of the rolling averages generally below the dashed line and the second half generally above. However, we do see a portion in the first half of the rolling plot that rises above the dashed line, suggesting some older members are recruited for a short period early on in the RDS process which may not be immediately clear from only using a convergence plot.

Fig. 2 shows the convergence and rolling plots of the average age of MSM. Similar to FSW, we see a general trend of rising average age as additional respondents are recruited into the study. The convergence plot (bottom) shows this with all of the values lying below the dashed line until the very end. However, from just the convergence plot, we might miss one trend that we can see very clearly in the rolling plot (top). Between around the 300th and 400th person recruited, we see a very clear drop in average rolling age. There is a slight dip in the cumulative average age, but since it is in the second half of recruitment, the effect is less pronounced due to the high value of the denominator. The change in slope is easy to miss and could be ignored as just random variation. This might be indicative of recruiting happening within a group of relatively very young respondents. This trend is not something that might be of immediate concern for researchers, though it should be monitored to see whether the more recent respondents all seem to tend to be younger. In our example, we see that the rolling age does go back up, indicating that the recruitment chain did not get stuck in a group of young people. This is encouraging, because we may be worried about a sort of social structure, in which respondents end up only recruiting those who are lower on the hierarchy (and therefore younger).

Gile et al. (2015) also suggest using bottleneck plots to determine whether seeds show similar trends and convergence patterns. This is aimed at determining whether the seeds start out from different ends of a bottleneck and recruit only within the portion of the social network that they started out in. Fig. 3 shows the bottleneck plots with both the cumulative and rolling values. Intuitively, these plots represent the same thing as in Fig. 1 except split up by seed. We note that for FSW, there were only two seeds that provided the vast majority of the data, and these are the only two seeds that are shown in these plots as well as in further bottleneck plots.

In both the rolling and cumulative plots, we see very similar trends. Both seeds seem to have very similar average age values, and the rolling age seems to be stable throughout. This is supported by the variation in the rolling plot on either side of the dashed line, which represents the overall average age, and the proximity to the dashed line in the convergence plot. We do see a very slight upward trend in the convergence plot, but from the variance shown in the rolling plot, this does not give us reason to think that there are very different groups of ages being sampled from.

Fig. 4 shows the bottleneck plots with the 6 most fruitful seeds with both rolling (top) and cumulative (bottom) values for MSM. We can see a similar trend in the biggest chain as in the overall convergence plot — about halfway through the recruitment chain, we see a sharp decline in the rolling bottleneck plot as well as the cumulative bottleneck plot.

4.3. Discussion of diagnostics with age

Even though it may seem as though the rolling versions of the convergence and bottleneck plots show very similar information, the rolling version do have utility in showing trends that might not be quite as apparent in the cumulative plots. For example, in both of the FSW and MSM cumulative bottleneck plots (bottom plot in Figs. 3 and 4), the chain with the largest sample size generally lies on one side of the dashed line (below for FSW and above for MSM), moving slowly toward the overall average. However, in the case of FSW, the rolling plot shows that generally, there does not seem to be reason to worry about a very strong homophily effect by age. On the other hand, for MSM, the rolling plot shows a very clear trend, with older members of that chain being recruited earlier before a sharp dropoff in rolling average age. Even though both FSW and MSM had a chain with consistently negative or positive slope, the rolling plots show that the FSW chain does not seem to have a change in average age of recruited participants while the MSM chain does.

For both FSW and MSM, we did not have a known population age. If a population value is known by researchers (which would be quite rare in key populations such as these), then the known population value can be used as the baseline to compare these cumulative and rolling statistics to. In the absence of such a known value, however, we used the overall sample average. Even if this value is different from the true (unknown) population value, these diagnostic plots still have value as they show whether estimates have stabilized, and can show if there are idiosyncrasies in the recruitment trees of different seeds, such as converging to different values, which would suggest poor mixing.

4.4. Diagnostics with geographical information

A key innovation in the 2012 Crane Survey was the collection of geographical information. In this section, we develop diagnostics that use this type of data, showing how it can be used to investigate the effective reach of the RDS process. The key assumption that we wanted to try to explore was whether the RDS survey was able to reach members from all around the city of Kampala. Since our study area encompasses all of Kampala, we want to make sure that the RDS was not simply finding all its recruits from only a small subset of the city.

To this end, we utilized the concept of a convex hull, the smallest convex figure that contains all of the reported pin locations on the map. The convex hull has been used when trying to determine an animal’s “home-range” based on a collection of locations (Worton, 1995; Getz and Wilmers, 2004; List and Macdonald, 2003; Creel and Creel, 2002). In these applications, the locational information is collected through methods such as radio telemetry (Worton, 1995), and the convex hull is used to determine the range of various mammals such as the kit fox or African wild dog (List and Macdonald, 2003; Creel and Creel, 2002).

In our application, we are not trying to find the home-range of any individual person or animal. Instead, we aim to look at the “home-range” of the RDS process as it moves around the city. To do this, we take each participant's primary reported pin location and treat it as a location of the RDS, drawing the convex hull around these pin locations to represent the RDS range. By looking at size of the convex hull over time, we can determine whether recruiting happens in a very small area at a time, or whether the RDS recruits across different areas of the map.

Fig. 5 shows an illustrative example of how a convex hull might be drawn. The plot on the left shows examples of pin locations of respondents at a point in time, and the plot on the right shows the same pin locations along with the convex hull drawn around the points. For our applications, we then compute the area of the convex hull in square kilometers to find a measure of the overall reach of the RDS process.

We used the `chull` function in the `R` package `grDevices` to find the points that make up the vertices of the convex hull, and used the `areaPolygon` function in the `geosphere` package to find the area of the convex hull.

4.5. Applications of the Convex Hull

We first looked at both the cumulative and rolling convergence plots of the area of the convex hull. Figs. 6 and 7 show these for the FSW and MSM populations, respectively. We note that when looking at the cumulative convergence and bottleneck plots for the area of the convex hull, the plots will be monotonically increasing, as it is impossible to have a smaller convex hull as we add points to the map. Therefore, the stabilization that we are looking for in these plots is reaching close to a maximum convex hull area and showing little growth beyond that. For this reason, we omit the horizontal dashed line showing the final overall value as we had with age.

In our example, Kampala's total land area is 176 square kilometers, though in general the convex hull may be much smaller due to general city layout or even bigger since the city itself is not necessarily a convex figure. Kampala, in fact, is an example of this, as there is a large bay in the city. We provide area of Kampala, then, as a rough estimate, to get an idea of the context rather than using it as a maximum value or ideal goal. This, as before in our choice of age as a respondent characteristic, is highly dependent on the study area of interest. Researchers studying cities with highly concentrated populations should make note of this, as well as taking care to consider the overall makeup of the city. This is not restricted to simple population density; the location and density of venues such as clubs plays a large part in this and is more important for the purposes for determining how respondents might come into contact with each other. Researchers must also consider that in very large cities (for example, New York), the convex hull area is highly unlikely reach anywhere near the area of the city, and the size of the city provides essentially no information about what convex hull to expect. In these cases, though, we can still use the convex hull area in order to get a sense for how much of the city was able to be sampled from. In cities where there are geographic boundaries (either natural ones such as rivers running through the city, or social ones such as neighborhoods separated by large levels of income disparity), researchers can look for large jumps in the convex hull area as indications that the RDS has jumped across these boundaries. We encourage local experts to play a role in critically evaluating these diagnostics, as they are extremely context-dependent.

Furthermore, our pin locations were based on area of work (FSW) or socialization (MSM). This was because these are the locations in which the key populations would meet others within the population. In other applications, researchers may choose a different definition of location. For example, in the case of using RDS to count the homeless population, researchers may consider using a mix of different types of locations such as soup kitchens or other areas of socialization, perhaps even using multiple locations for each person.

With the context-specific aspects of these diagnostics in mind, we will discuss the diagnostics for our FSW and MSM data. Starting with the convergence plot of the FSW population in Fig. 6, we see that the convex hull size reaches an upper bound about halfway through the study. The convex hull area actually gets quite close to 200, which is possible due to the non-convex shape of Kampala. This suggests that the RDS is generally managing to explore all areas of the city, though, as the area is quite large. If we found that the convex hull area were maxing out at around 100 or lower, we might be more concerned, depending on the location of venues. For example, if we knew that the venues were all concentrated in a smaller area, then a smaller convex hull would not be a reason for concern. However, if we expected the venues to be generally spread out throughout the city including at the edges (for example, at the ports for a city bordering water), then we might want to investigate further as to why the convex hull area was so small.

The rolling convex hull plot for FSW does not show much of a trend, hovering close to around 30 square kilometers for most of the recruitment process. This suggests that there aren’t any small neighborhoods that the RDS gets stuck in, and that the recruitment happens consistently throughout various parts of the city.

The convergence plot of the MSM population as seen in Fig. 7 shows a similar effect, reaching an area of 120 square kilometers after a little more than 100 recruits while the final area was close to 140 square kilometers. This suggests that the RDS process is reaching the geographical edge of the MSM population in Kampala by around a sample size of 100, and that additional recruits do very little to push out the convex hull. Though neither FSW nor MSM showed this behavior, one additional pattern to watch for would be if there were no “leveling off” effect in the convergence plot, and the area of the convex hull were to keep showing an increase even near the end of the study. This would suggest that the full limits had not been reached, and the RDS was still in the process of exploring more areas of the city.

We are concerned with not just the range of the overall RDS, though. We want to be able to see whether there is seed dependence in the recruitment process. To investigate this, we can use bottleneck plots similar to how we used them in Section 4.2. Figs. 8 and 9 show the bottleneck plots with cumulative and rolling convex hulls for FSW and MSM.

In Fig. 8, the top plot showing the rolling convex hull by seed, we see that both seeds result in chains that maintain quite consistent convex hull areas. The smaller recruitment chain is quite short, and doesn’t contain much information, though the rolling convex hull areas are consistent with the longer chain. These rolling areas are quite large considering they all contain only 25 respondents, going up almost as high as 100, which would represent more than half of Kampala's area. This suggests that the recruiting is quite consistently happening from all around the city. In addition, we see from the cumulative plot on the bottom of Fig. 8 that the convex hull seems to reach close to its largest size early on for the larger recruitment chain, and they both reach above 100, with the longer chain going up above 175. This suggests that the two seeds are not recruiting from completely disjoint populations, as there would need to be some overlap in the maximum convex hulls for each chain. We note that, in this population, because there was essentially one dominant chain, the bottleneck plots actually show much of the same information as the convergence plots, and are less useful. This is generally the case in RDS studies with fewer productive seeds, as there must be long enough chains in order to calculate substantive rolling values.

Fig. 9 shows the bottleneck plots for MSM. Since there were only six seeds that resulted in chains with at least 25 respondents, we only included those chains in these bottleneck plots. We note that the areas of some of the rolling convex hulls are quite small, suggesting a high level of seed dependence and that recruitment is happening only within a small area. For example, in the two chains with the fewest respondents, we see that the area of the rolling convex hull generally stays below 10 square kilometers. In cases like this, researchers may choose to investigate why this is happening by consulting local experts (if possible), and encourage respondents to recruit from elsewhere. These chains ended quite early, but if they were longer, later respondents could be asked how many people they know who have been recruited into the study to determine whether this phenomenon is due to a local community repeatedly recruiting within itself, or whether separate groups of people from the same location are being recruited. If this type of issue were to exist with all seeds, adding additional seeds in other locations within the city may be considered. In our example, we do see that the rest of the chains do tend towards around 80 square kilometers. These are not significantly smaller than the maximum convex hull area from the convergence plot for MSM, and so we aren't worried about individual chains recruiting from completely different areas of the city. If all of the seeds had had cumulative areas much smaller than observed in the convergence plot, it would be an indication that there wasn't much overlap in convex hulls across chains.

4.6. Discussion of convex hull diagnostics

The geographic information allows us to create diagnostic plots using convex hulls to characterize RDS recruitment. Using cumulative and rolling convex hull areas with convergence plots, we were able to detect if it seemed like recruitment was leveling off, or starting to recruit from a relatively small area. The bottleneck versions of these same plots gave us additional information about individual seed behavior, providing insights into when specific chains might start recruiting from only a very small area. This can provide an indication for investigators to add more seeds in different locations as the RDS data collection is ongoing, or investigate why recruitment might be limited to certain geographic neighborhoods. Future respondents could also be asked to consider people from a location different from where they themselves were recruited.

We note that for many of these trends, it was necessary to use the rolling plot to observe how the convex hull changed over time. When using just the cumulative convex hull, once a certain large area is reached, it is hard to detect changes in how the RDS process is recruiting. The rolling plots, on the other hand, are able to capture whether the recruitments start only happening in a very small area, as was the case with the FSW in Fig. 8.

5. Limitations and future work

One limitation of the convex hull diagnostic plots is that they can be hard to interpret as the study is ongoing. Researchers may consider including questions in their surveys asking about how many people the respondents know in different neighborhoods to supplement these diagnostics and be more proactive in determining whether a dip in rolling convex hull area is cause for concern or not. For example, if researchers see such a dip and notice that the recent respondents generally said they do not know very many people outside of their own neighborhood, researchers might consider adding seeds from different geographical areas.

Investigators should balance knowledge of the geographical area of interest with the information provided by the diagnostics in this paper. Notably absent are the respondent pin locations displayed on a map. Due to privacy concerns, diagnostics with actual pin locations placed on a map were not deemed suitable to be shared with a larger audience. However, researchers during the RDS collection may use the actual locations placed on a map, and looking at these locations, along with the convex hull, on a map over time could provide additional insights into the recruitment process. A survey item asking about neighborhood of residence could achieve similar goals as the geographic data, but it would not be as fine-grained and could fail to detect cases in which disjoint networks exist in the same neighborhood.

In addition, for the diagnostics we have presented in this paper, we chose to omit those with pin locations outside of the city limits of Kampala, Uganda. However, one might be interested in exactly these cases, to identify when recruitment jumps outside of some geographical boundary.

The use of convex hulls is not perfect. Venues for socialization and work are not placed uniformly within a city, and the structure of streets and walkways can greatly affect the true distance between two points. In other words, there is no easy way to determine what a target convex hull size should be. Instead, we suggest looking at trends to determine whether there is reason to be concerned about recruitment getting stuck in certain areas, or whether the seed dependence is very high.

Furthermore, the usage of these convex hulls does not guarantee that researchers will be able to detect whether the population social network is connected. Disjoint networks might occupy the same geographic space, but be separated by social class or other demographic characteristics, such as education level. However, since part of how the population is defined is according to the geographic location (e.g. all FSW older than 15 years of age who had sold sex to a male within the last six months within the greater Kampala area), we believe examining the convex hull diagnostic plots is highly useful for evaluating the “selection error” that might occur with RDS.

For future extensions to these diagnostics, one area of interest might be in including study site. We did not consider study site in our diagnostics, but the location of whether respondents were required to go to take the survey could play a huge role in the recruitment. This is particularly pertinent in very large cities, where multiple study sites might be necessary, and researchers may be concerned about the level of geographic coverage of the RDS study.

6. Conclusion

Based on these results, we recommend using geographical data, when available, as another diagnostic measure for RDS, as they can help determine the effective reach of these surveys. We suggest using the area of the convex hull with convergence and bottleneck plots, as well as both cumulative and rolling versions of each of these plots, to better determine the range of the RDS and whether the seed dependence is overcome as the RDS progresses. In addition, we recommend using rolling plots alongside cumulative plots in general to more easily identify trends as they occur.

Acknowledgements

This survey was funded by the US Government's President's Emergency Plan for AIDS Relief (PEPFAR). We appreciate the courage of the survey participants to take part and the survey staff's commitment to their work. The Crane Survey Group included Wolfgang Hladik, David Serwadda, Moses Ogwal, Enos Sande, Joy Kusiima Bbaale, Mark Berry, Herbert Kiyingi, Alex Opio, and Michael Muyonga.

References

Crawford F, Wu J, Heimer R, 2018. Hidden population size estimation from respondent-driven sampling: a network approach. J. Am. Stat. Assoc 113, 755–766. [PubMed: 30828120]

Creel S, Creel N, 2002. The African Wild Dog: Behavior, Ecology, and Conservation. Princeton University Press.

Doshi R, Sande E, Ogwal M, Kiyingi H, McIntyre A, Kusiima J, Musinguzi G, Serwadda D, Hladik W, 2018. Progress toward unaids 90-90-90 targets: A respondent-driven survey among female sex workers in kampala, uganda. PLOS ONE 13.

Getz W, Wilmers C, 2004. A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. Ecography 27, 489–505.

Gile K, Handcock M, 2010. Respondent-driven sampling: An assessment of current methodology. Sociol. Methodol 40 (1), 285–327. [PubMed: 22969167]

Gile KJ, Johnston LG, Salganik MJ, 2015. Diagnostics for respondent-driven sampling. J. Royal Stat. Soc.: Series A 178, 241–269.

Handcock M, Gile K, 2011. Comment: On the concept of snowball sampling. Sociological Methodology.

Handcock M, Gile K, Mar C, 2015. Estimating the size of populations at high risk for HIV using respondent-driven sampling data. Biometrics 71, 258–266. [PubMed: 25585794]

Heckathorn D, 1997. Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems 44, 174–199.

Heckathorn D, 2011. Snowball versus respondent-driven sampling. Sociol. Methodol 41 (1), 355–366. [PubMed: 22228916]

Hladik W, Sande E, Berry M, Ganafa S, Kiyingi H, Kusiima J, Hakim A, 2017. Men who have sex with men in kampala, uganda: Results from a bio-behavioral respondent driven sampling survey. AIDS Behav. 21, 1478–1490. [PubMed: 27600752]

Johnston L, McLaughlin K, El Rhilani H, Toufik A, Bennani A, Alami K, Elomari B, Handcock M, 2015. Estimating the size of hidden populations using respondent-driven sampling data: Case examples from Morocco. Epidemiology 26 (6), 846–852. [PubMed: 26258908]

Johnston L, et al. , 2006. Assessment of respondent driven sampling for recruiting female sex workers in two vietnamese cities: Reaching the unseen sex worker. J. Urban Health 83 (7).

Johnston LG, Chen YH, Silva-Santisteban A, Raymond HF, 2013. An empirical examination of respondent driven sampling design effects among hiv risk groups from studies conducted around the world. AIDS Behav. 17, 2202–2210. [PubMed: 23297082]

Lansky A, Abdul-Quader AS, Cribbin M, Hall T, Finlayson TJ, Garfein RS, Lin LS, Sullivan PS, 2007. Developing an hiv behavioral surveillance system for injecting drug users: the national hiv behavioral surveillance system. Public Health Report 122, 48–55.

Lansky A, Drake A, Wejnert C, Pham H, Cribbin M, Heckathorn D, 2012. Assessing the assumptions of respondent-driven sampling in the national HIV behavioral surveillance system among injecting drug users. Open AIDS J 6, 77–82. [PubMed: 23049656]

List R, Macdonald D, 2003. Home range and habitat use of the kit fox (vulpes macrotis) in a prairie dog (cynomys ludovicianus) complex. J. Zool 259, 1–5.

Liu H, Li J, Ha T, Li J, 2012. Assessment of random recruitment assumption in respondent-driven sampling in egocentric network data. Social Netw. 1, 13–21.

Malekinejad M, Johnston L, Kendall C, Kerr L, Rifkin M, Rutherford G, 2008. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. AIDS Behav. 12, S105–S130. [PubMed: 18561018]

Paquette DM, Bryant J, Wit JD, 2011. Use of respondent-driven sampling to enhance understanding of injecting networks: A study of people who inject drugs in Sydney, Australia. Int. J. Drug Policy 22, 267–273. [PubMed: 21550790]

Phillips II G, Kuhns LM, Garofalo R, Mustanski B, 2014. Do recruitment patterns of young men who have sex with men (ymsm) recruited through respondent- driven sampling (rds) violate assumptions? J. Epidemiol. Commun. Health 68, 1207–1212.

Salganik M, 2012. Commentary: Respondent-driven sampling in the real world. Epidemiology 23, 148–150. [PubMed: 22157310]

Salganik M, Fazito D, Bertoni N, Abdo A, Mello M, Bastos F, 2011. Assessing network scale-up estimates for groups most at risk of HIV/aids: Evidence from a multiple-method study of heavy drug users in curitiba, brazil. Am. J. Epidemiol 174, 1190–1196. [PubMed: 22003188]

UNAIDS, 2014. 90-90-90: An ambitious treatment target to help end the AIDS epidemic. UNAIDS. Technical Report.

UNAIDS, Organization, W.H., 2010. Guidelines on Estimating the Size of Populations Most at Risk to HIV. UNAIDS and World Health Organization. Technical Report UNAIDS/00.03E.

Uuskula A, Johnston LG, Raag M, Trummal A, Talu A, Des Jarlais DC, 2010. Evaluating recruitment among female sex workers and injecting drug users at risk for hiv using respondent-driven sampling in estonia. J. Urban Health 87 (2), 304–316. [PubMed: 20131018]

Volz E, Heckathorn D, 2008. Probability-based estimation theory for respondent driven sampling. J. Official Stat 24, 79–97.

Wang J, Carlson RG, Falck RS, Siegal HA, Rahman A, Li L, 2005. Respondent-driven sampling to recruit mdma users: A methodological assessment. Drug and Alcohol Dependence 78, 147–157. [PubMed: 15845318]

White R, Lansky A, Goel S, Wilson D, Hladik W, Hakim A, Frost S, 2012. Respondent driven sampling-where we are and where should we be going? Sexually Transmitted Infect. 88, 397–399.

Worton B, 1995. A convex hull-based estimator of home-range size. Biometrics 51, 1206–1215.

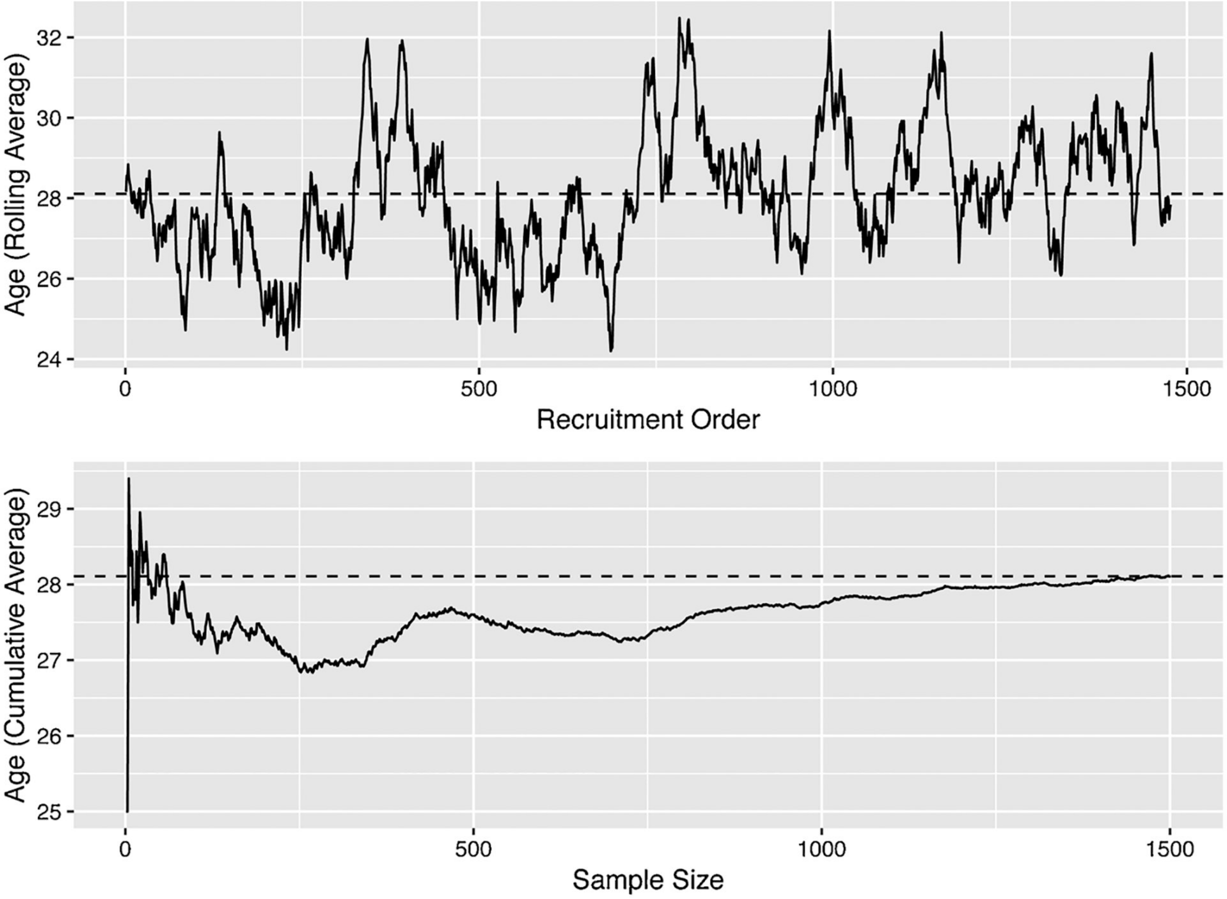

Fig. 1.

The rolling average age convergence plot (top) and the cumulative average age convergence plot (bottom) for FSW are shown. The cumulative plot shows an increasing trend, but the rolling plot suggests that this is not the result of recruiting moving to a predominantly older group.

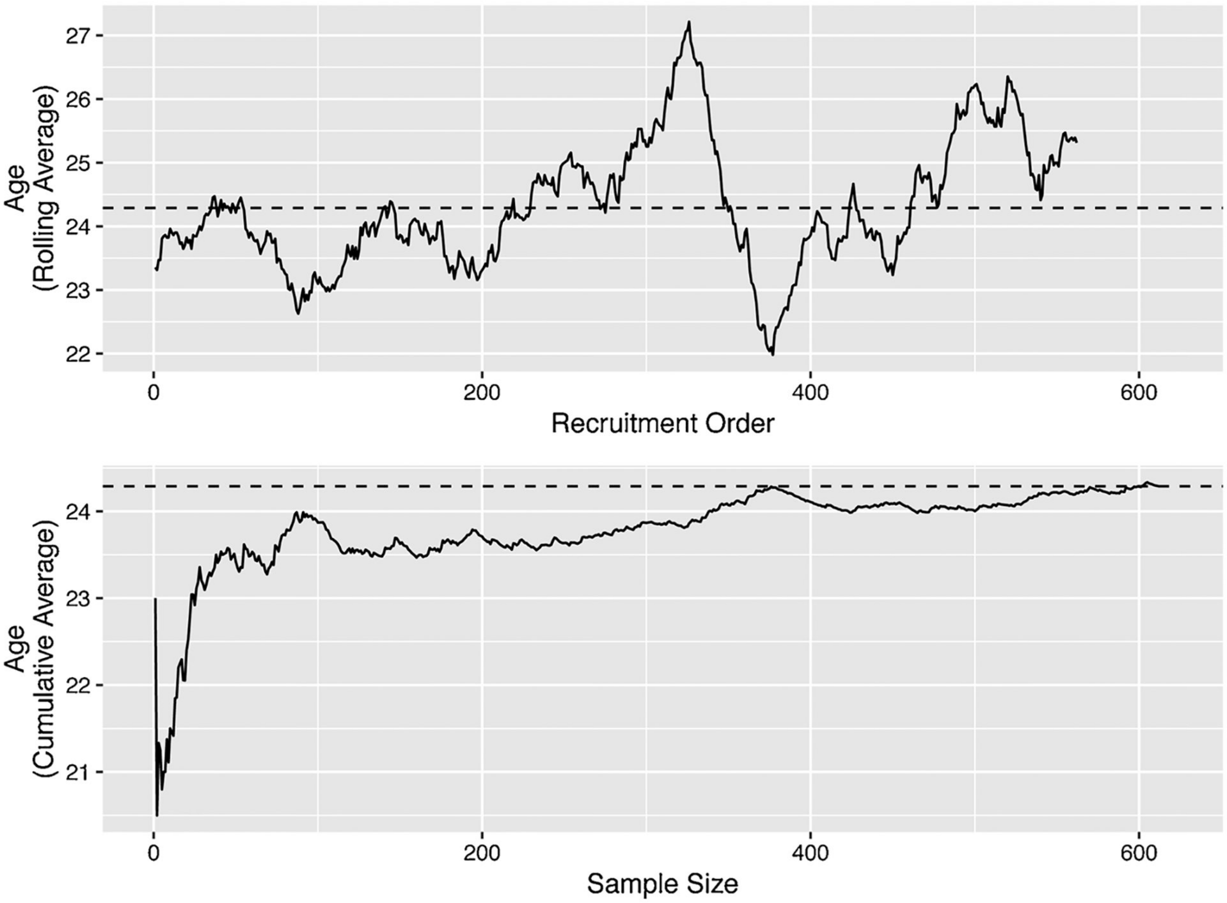

Fig. 2.

The rolling average age convergence plot (top) and the cumulative average age convergence plot (bottom) for MSM are shown. As with FSW, there is a generally increasing trend, and the rolling plot shows a sharp drop partway, suggesting recruitment moved to a younger group at that point in time.

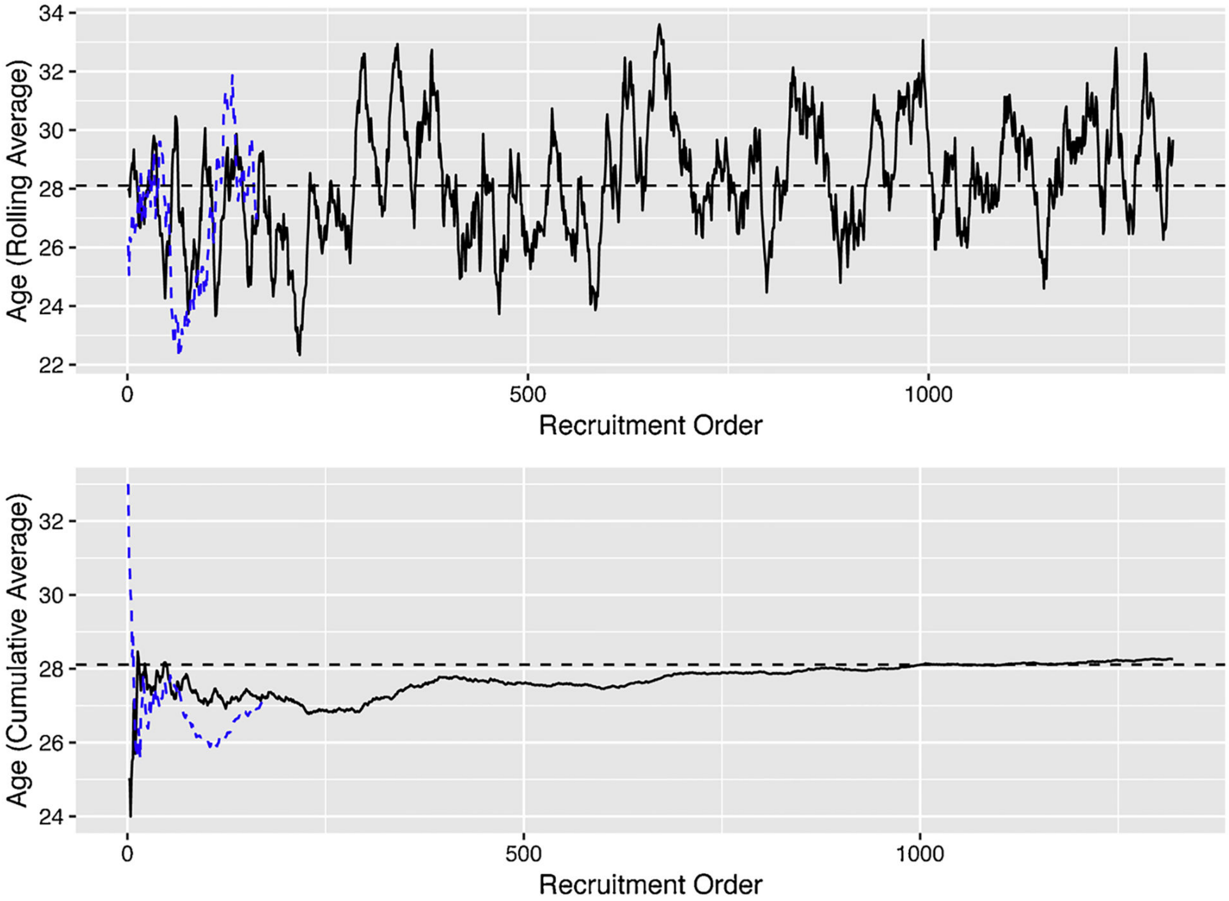

Fig. 3.

The rolling average age bottleneck plot (top) and the cumulative average age bottleneck plot (bottom) for the top two seeds of FSW are shown. Both show very little differences between the two seeds.

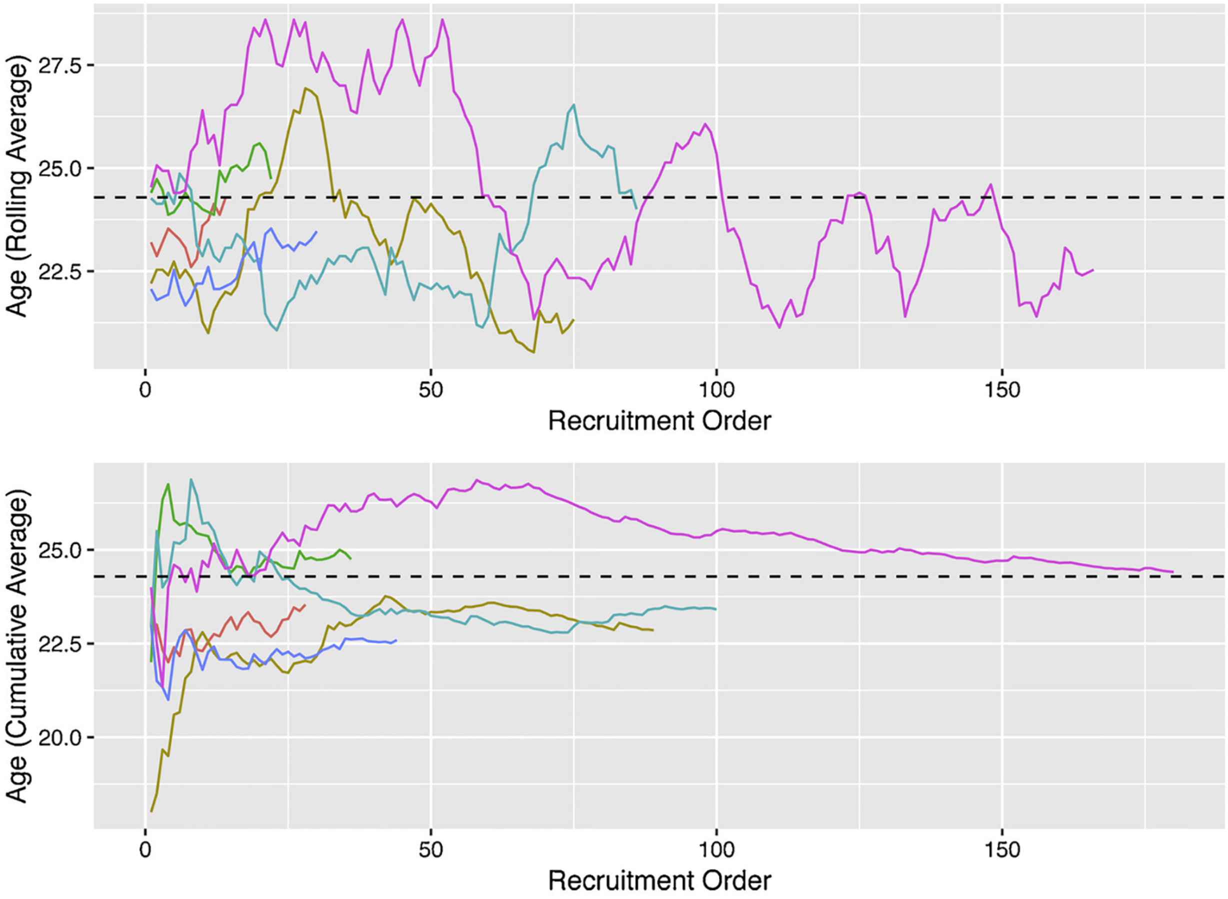

Fig. 4.

The rolling average age bottleneck plot (top) and the cumulative average age bottleneck plot (bottom) for the top 6 seeds for MSM are shown. The biggest recruitment chain shows a clear pattern of recruiting older respondents to start before moving to a younger group later on in the chain.

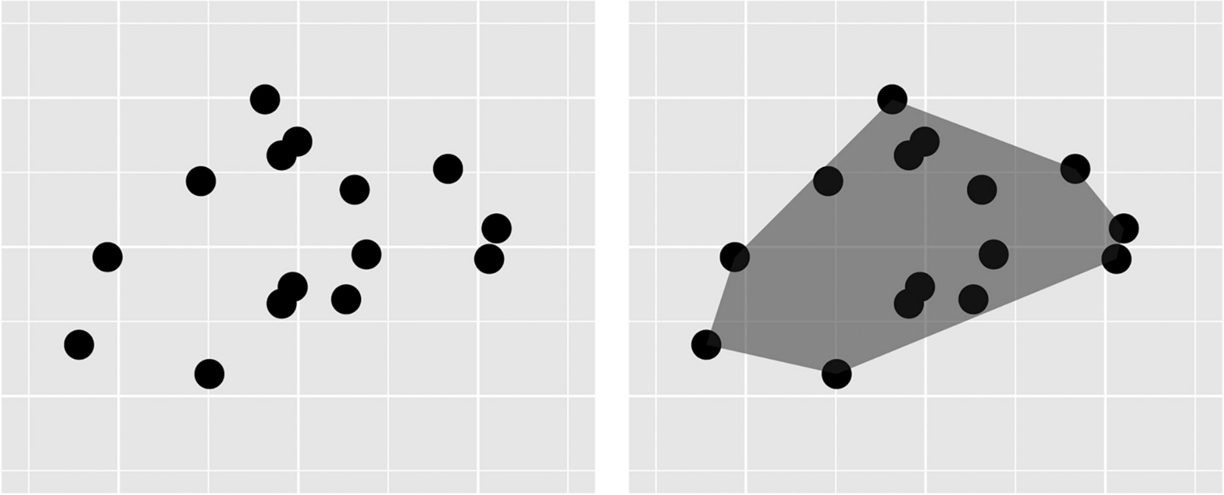

Fig. 5.

Illustrative example of how a convex hull might be drawn. The left figure shows example location points, and the right figure shows the convex hull drawn around those locations points.

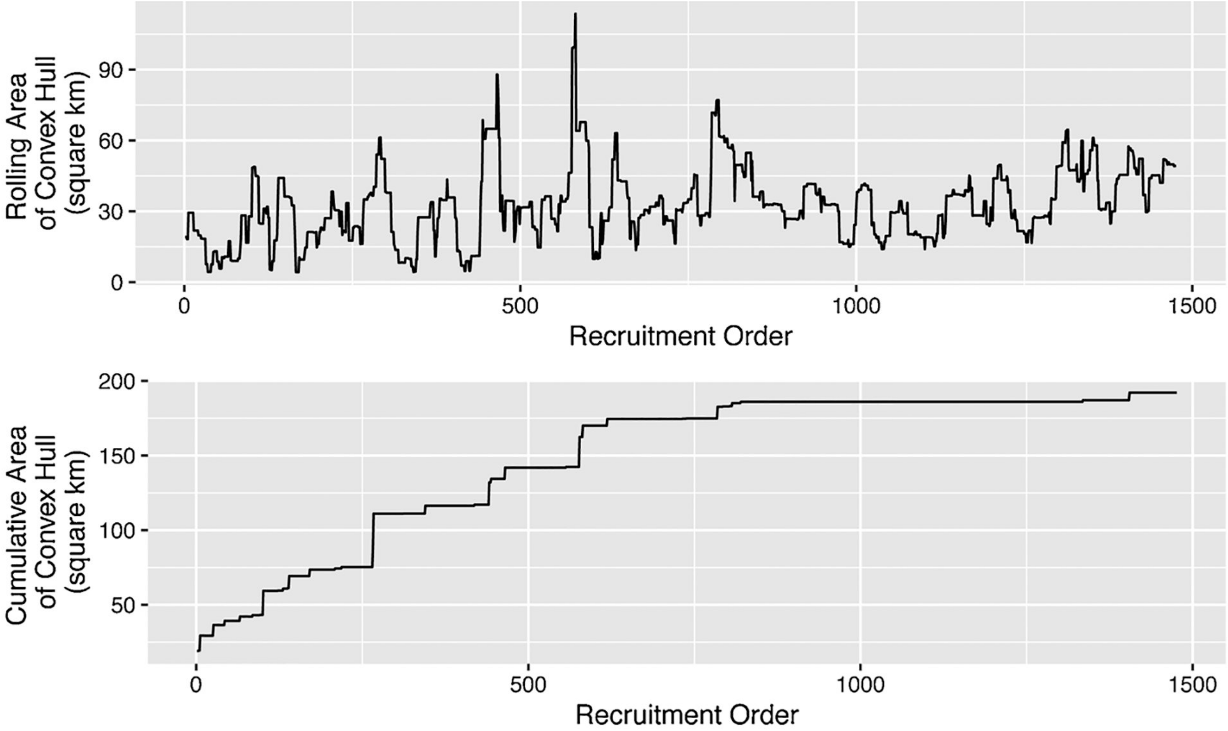

Fig. 6.

The rolling convex hull area convergence plot (top) and the cumulative convex hull area convergence plot (bottom) for FSW are shown. The rolling plot shows overall consistency, and the convex hull reaches near its maximum size at around 750 people recruited.

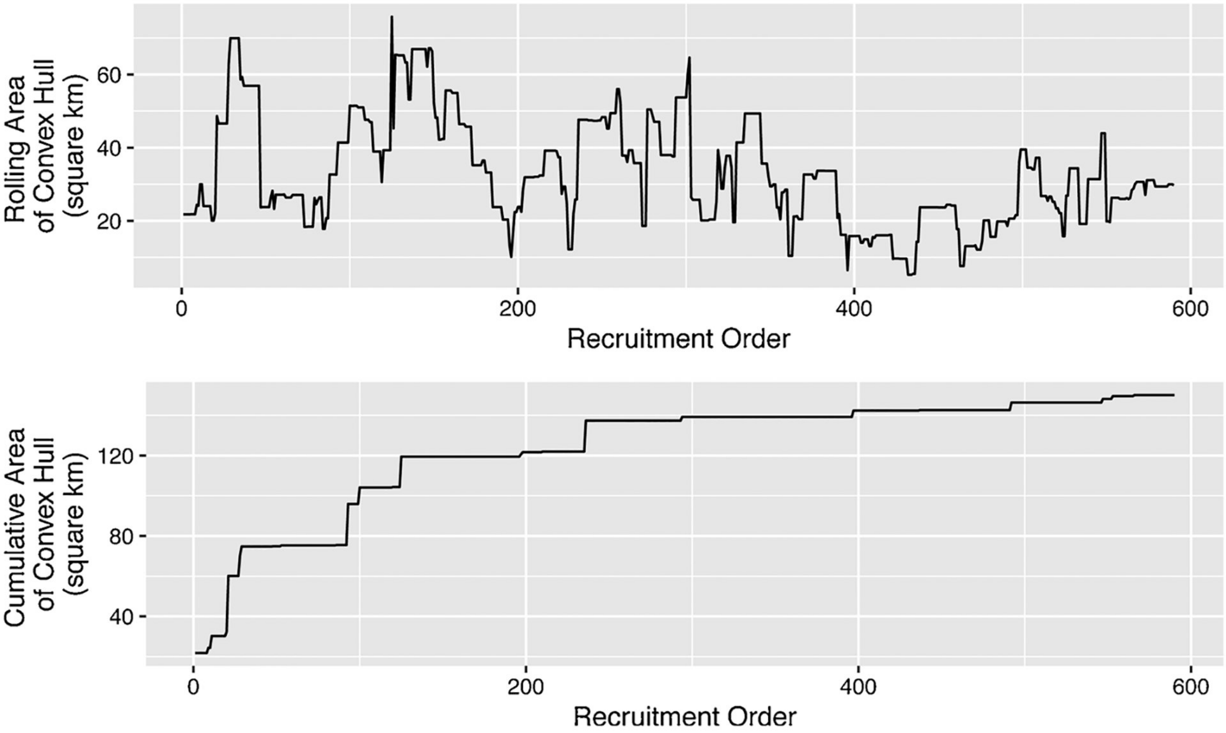

Fig. 7.

The rolling convex hull area convergence plot (top) and the cumulative convex hull area convergence plot (bottom) for MSM are shown. As opposed to FSW, there seems to be evidence of convergence in cumulative convex hull area as it reaches close to the maximum area early on in recruitment.

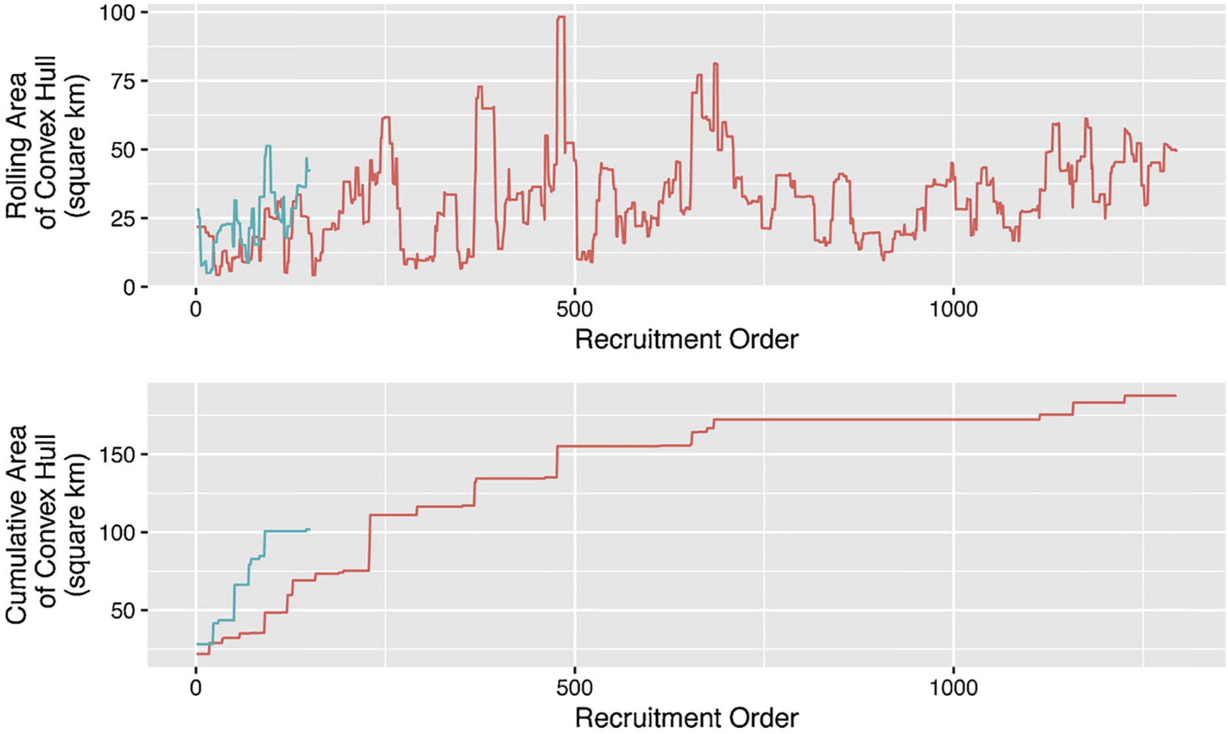

Fig. 8.

The rolling convex hull area bottleneck plot (top) and the cumulative convex hull area bottleneck plot (bottom) for FSW, with only the two seeds which resulted in chains with more than 25 respondents, are shown. All areas are in square kilometers.

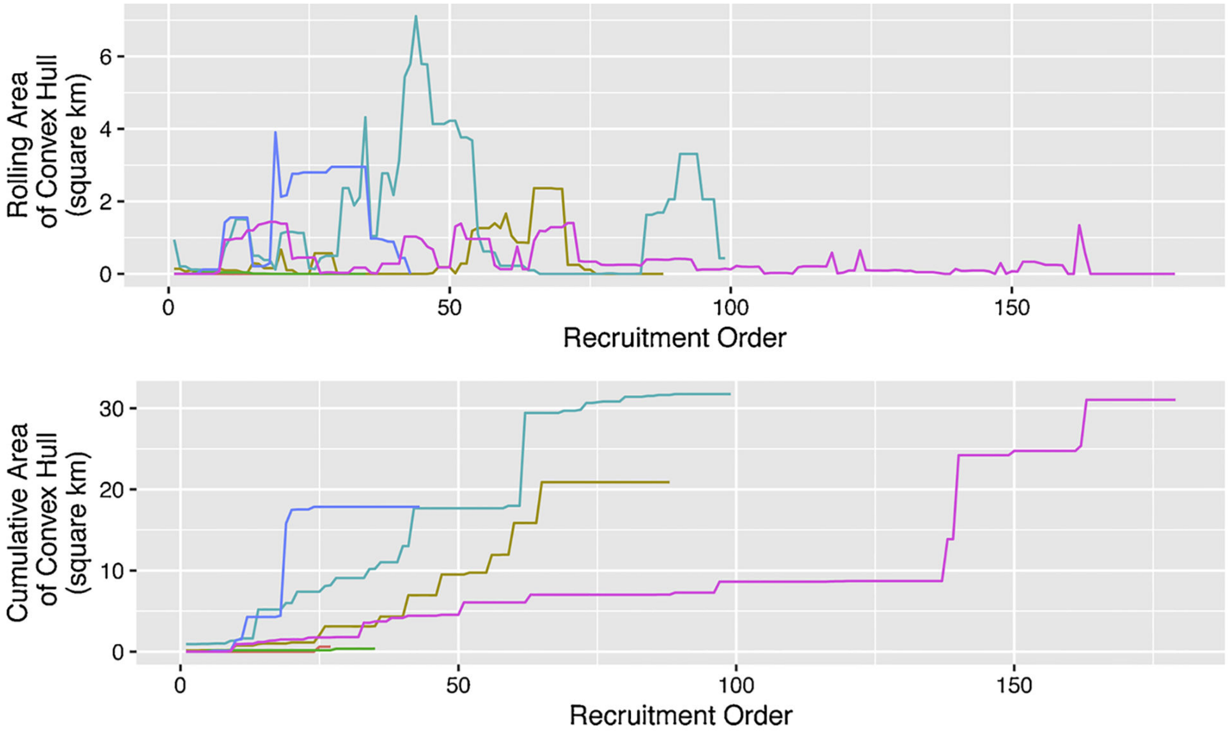

Fig. 9.

The rolling convex hull area bottleneck plot (top) and the cumulative convex hull area bottleneck plot (bottom) for MSM, with only the six seeds which resulted in chains with more than 25 respondents, are shown. The size of convex hulls indicates that different chains don't seem to be recruiting from disjoint subgroups. All areas are in square kilometers.

Table 1

Characteristics of the RDS run in Kampala, Uganda. See Doshi et al. (2018) and Hladik et al. (2017) for a more comprehensive description of the RDS studies.

	FSW	MSM
Sampling period (months)	10	16
Number of initial seeds	4	2
Total number of seeds	4	36
Coupons per person	Up to 4	Up to 8
Total coupons redeemed	1916	1035
Final sample size	1501	612