Better understanding of Demographic and Migration statistics

Instructions: Click on the link to access each author's presentation.

Chair: Inga Masiulaityte-Sukevic

Participants:

Elio Villaseñor: Mapping Migration Trends: Leveraging Data Science and Web-Scraped News for Comprehensive Analysis of Transitory Populations

Maria Luiza Guerra de Toledo: Estimation of Life-Space Assessment Score for Brazilian Elderly Population Combining Probability and Nonprobability Samples

Mutale Sampa: Prevalence and Associated Factors of Preterm Birth Among Reproductive-Aged Women in Zambia: Examining Socioeconomic Inequalities

Andrea Pellandra:* Mapping indicators on Forcibly Displaced Persons (FDPs): the UNHCR journey.

Giorgio Alleva: Respondent Driven Sampling strategies for hard-to-reach populations

* Work presentation not available or non-existent

IAOS-ISI 2024 MEXICO CONFERENCE

Mapping Migration Trends: Leveraging Data Science and Web-Scraped News for Comprehensive Analysis of Transitory Populations

16 May 2024

Elio A. Villaseñor G., Olinca D. Paez D., Alejandra Figueroa M., Victor Silva C.,

Instituto Nacional de Estadística y Geografía (INEGI), México

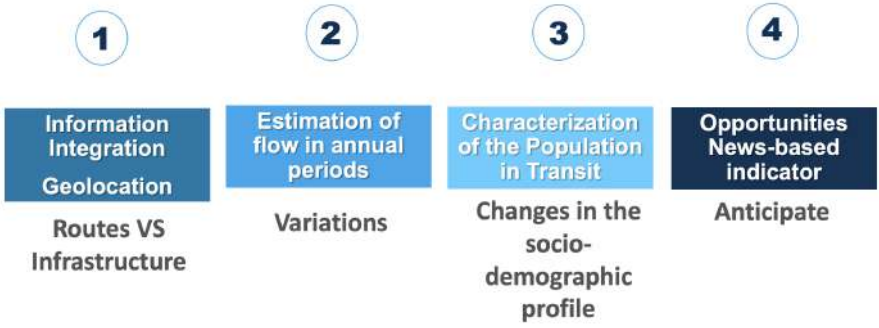
INEGI International Statistical Institute IAOS IMPROVING OFFICIAL STATISTICS

Contents

Hypothesis and goal

- By integrating traditional sources with alternative sources, it is possible to design a methodology and obtain indicators to size the population of migrants in transit.
- The goal is to have a methodology to measure the flow of the migrant population in transit through Mexican territory.

Analytical Strategy



Data Sources

Traditional

DENUE
Census 2020
Deaths

Survey on Migration on Mexico's Northern Border

Records of the INM

Other Records

+

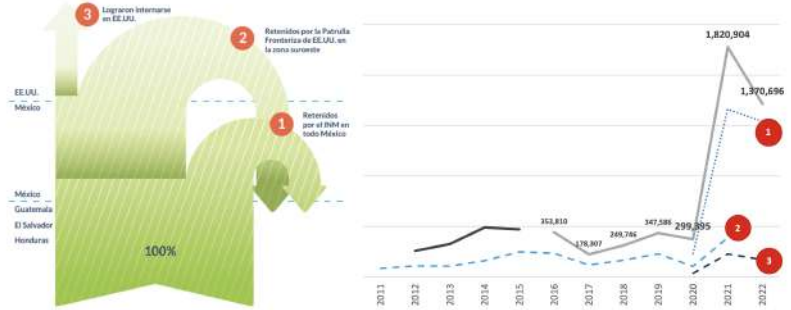
U.S. Custom´s & border protection

American Community Survey

Alternatives

Estimation of Flow

Integrating Data from Traditional Sources



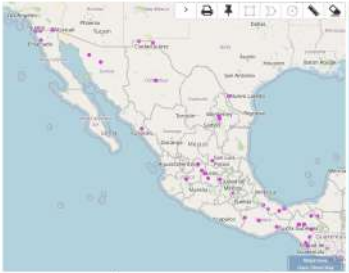
Source: Taken from Rodriguez Chavez (2016). Central American Migration in Irregular Transit through Mexico. New figures and trends. Policy Brief #14 (ANAMIG).

Opportunities

- Indicator based on news
- Routes vs. Infrastructure
- Variations
- Changes in the Sociodemographic Profile

Geolocation: Integrating Information Layers

How many migrant shelters and where they are located
2019 – 2020 - 2022

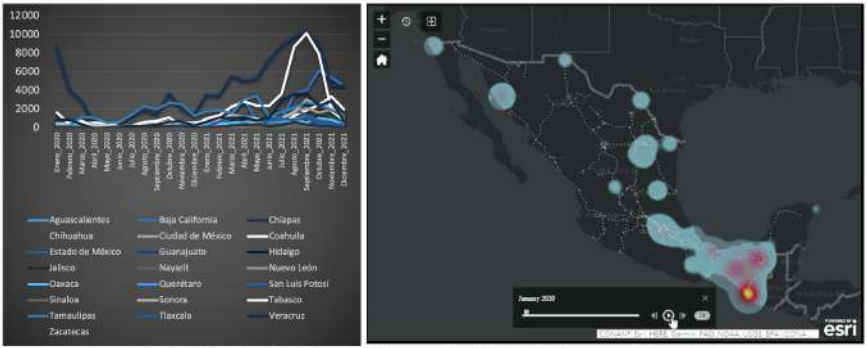


Source: DENUE (Economic Census 2019).

Sources: 2020 Population and Housing Census; BBVA Research 2020; Jesuit Network 2022.

Geolocation: Integrating Information Layers

Seasonal Pattern of Irregular Migratory Flow



Source: Migration Policy and Registration and Identity of Persons Unit of the SEGOB, from January 2020 to October 2022.

Geolocation: Integrating Information Layers

Hostels

Migration control

Population and Housing Census 2020.

Migration Policy and Registration and Identity of Persons Unit (SEGOB), 2020 and 2021.

Deaths 2020.

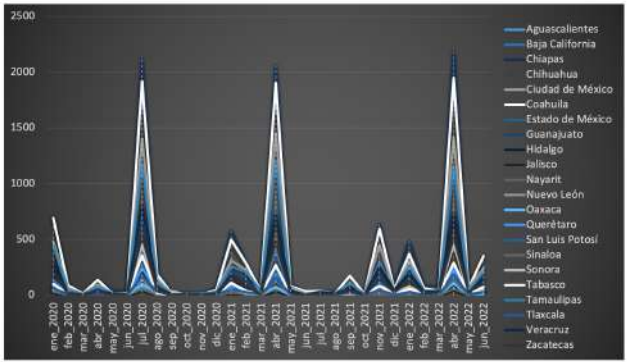
News on the Internet 2020-2022.

Deaths

Public Perception

Geolocation: Integrating Information Layers

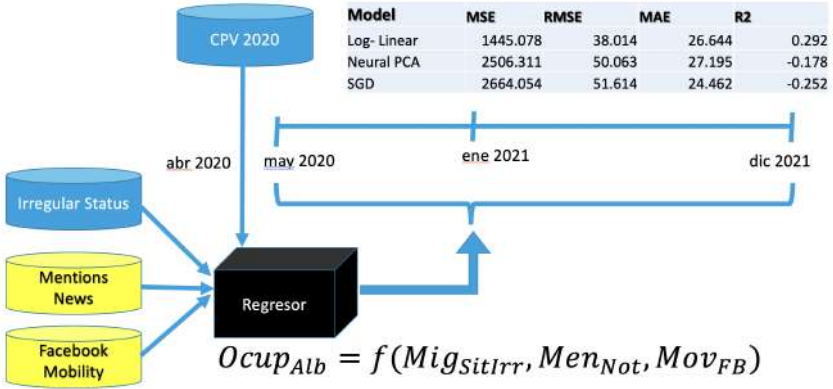
Alternative Sources



Web scraping of migration news

Flow model

Estimation of Migrants in Shelters



Model	MSE	RMSE	MAE	R2
Log- Linear	1445.078	38.014	26.644	0.292
Neural PCA	2506.311	50.063	27.195	-0.178
SGD	2664.054	51.614	24.462	-0.252

$$Ocup_{Alb} = f(Mig_{SitIrr}, Men_{Not}, Mov_{FB})$$

Flow model

Estimation of Migrants in Shelters

Conclusions

- High complexity of the problem
- Need to integrate various sources
- Usefulness and volatility of alternative sources
- Desirability of having mobile phone data
- Results obtained must be validated

THANK YOU

Thank You for your attention!
Questions?

IAOS-ISI 2024
MEXICO CONFERENCE

Thank you

INEGI

International Statistical Institute

IAOS
IMPROVING OFFICIAL STATISTICS

IOS-ISI 2024
MEXICO CONFERENCE

Estimation of Life-Space Assessment Score for Brazilian Elderly Population Combining Probability and Nonprobability Samples

Maria Luíza Guerra de Toledo
National School of Statistical Sciences
Brazilian Institute of Geography and Statistics (Brazil)

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

Presentation agenda

- Introduction: Life-space mobility survey
  - Study design, setting, and participants
  - Measures, objectives and hypothesis
- Methodology: Quasi-randomization to combine the probability and nonprobability samples
- Reference sample: PNS 2019
- Results
- Conclusions and future work

IOS-ISI 2024 MEXICO CONFERENCE

Introduction

Background:

- The COVID-19 pandemic hit Brazil in a scenario of substantial socioeconomic and health inequalities.
- Experts agree that older people are the group most affected by the COVID-19 pandemic (MORLEY; VELLAS, 2020).
- Social restriction recommendations have been set up as population-level measures to suppress community transmission of COVID-19 (LEWNARD, 2020).
- It is unknown the immediate impact of social restriction recommendations (i.e., lockdown, stay-at-home) on the *life-space mobility* of older people, particularly for those living in low-resource settings.

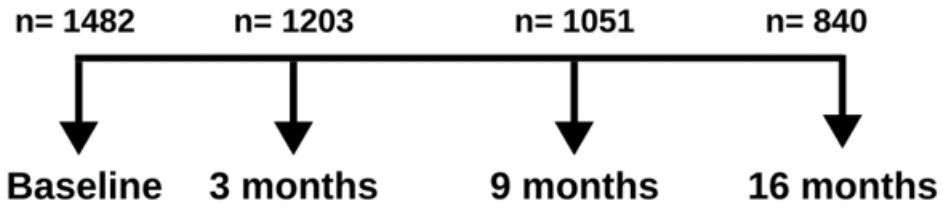
Introduction

Life-space mobility:

- Corresponds to how people engage in, maintain social relationships and roles, and participate in meaningful activities within their communities (RANTAKOKKO et al., 2013).
- It is recognized as a practical measure to capture older people's functional ability for moving around in their environments in a specific period of time (PEEL et al., 2005).
- Restriction of life-space mobility occurs due to a combination of losses in individuals' intrinsic capacity, limited personal resources, and difficulty dealing with environmental challenges, resulting in potentially health adverse outcomes (XUE et al., 2008).

Study Design, Setting, and Participants

- Prospective cohort survey to investigate life-space mobility during COVID-19 pandemic.
- Subject: A convenience snowball sample of participants aged 60 and older (n = 1,482) living in 22 (82%) states in Brazil, using an online platform.
- Baseline data collection between May and July, 2020.



- At the baseline, participants were asked about the places they reached both before the COVID-19 pandemic and a week before evaluation.

Study Design, Setting, and Participants

Objectives:

- To investigate the immediate impact of COVID-19 pandemic on life-space mobility of community-dwelling Brazilian older adults.
- To examine the social determinants of health associated with change in life-space mobility.

Hypothesis:

- Levels of life-space mobility throughout the pandemic will exhibit different trajectories according to social determinants.

IOS-ISI 2024 MEXICO CONFERENCE

Measures

Life-Space Mobility:

- Life-space mobility was assessed using a Brazilian Portuguese version of the Life-Space Assessment (SIMÕES *et al*., 2018).
- The LSA comprises:
  - five life-space levels
  - how often within the week they attained that level (frequency)
  - whether they needed any help to move to that level (independency)
- Composite score: each life-space level reached x frequency x independency.
- Score range from 0 to 120 points; higher scores represent greater mobility in space.

Measures

Social Factors, Comorbidities and Reported Social Restriction

- Gender
- Age group (60–69, 70–79, and ≥80 years)
- Self-report of skin color/race/ethnicity categorized according to official Brazilian classification (white, black, “pardo”, “amarelo”, and indigenous)
- Marital status (single, married, divorced, widowed)
- Education level (illiterate, 1–4 years, 5–8 years, and ≥9 years of schooling)
- Income level presented as the minimum wage per month guaranteed by law in Brazil (<1, 2–3, 4–7, 8–10, and >10 minimum wage salaries)
- Employment (active, inactive, or unemployed)

Preliminary analysis showed that the sample was biased when representing Brazilian elderly people in terms of sociodemographic characteristics.

Methodology

IOS-ISI 2024
MEXICO CONFERENCE

Inference from nonprobability samples

Recent resurgence in interest in making inferences from nonprobability samples for several reasons (VALLIANT, 2020):

- Response rates in probability surveys have been decreasing. A sample initially selected randomly can hardly be called a probability sample from the desired population.
- Nonprobability sources may either replace probability samples or be combined with them for inference (Social media and other data that can be scraped from the web).

Valliant, Dever and Kreuter (2018) review some of the problems that probability samples have encountered in the last decade.

Methodology: Estimating from a nonprobability sample (VALLIANT, 2020)

Quasi-Randomization

- The sample is treated as if it were obtained via a probability mechanism (unknown).
- Pseudo selection probabilities of being in the sample are estimated by using the sample in combination with some external data set that covers the desired population.

Let $\pi(i \in s|\boldsymbol{x_i}; \boldsymbol{\Phi})$ be the inclusion probability of unit i in the sample s, which depends on:

- A vector of covariates, $\boldsymbol{x_i}$,
- and an unknown parameter, $\boldsymbol{\Phi}$, that must be estimated.

Methodology: Estimating from a nonprobability sample (VALLIANT, 2020)

Given estimates of the pseudo-inclusion probabilities, $\pi\left(i \in s|\boldsymbol{x_i}; \widehat{\boldsymbol{\Phi}}\right)$, an estimator of a total of an analysis variable, y_i, is

$$\hat{t}_y = \sum_{i \in s} w_i y_i,$$

where the weight is defined as $w_i = 1/\pi\left(i \in s|\boldsymbol{x_i}; \widehat{\boldsymbol{\Phi}}\right)$.

A mean is estimated as $\hat{\bar{y}} = \sum_s w_i y_i / \sum_s w_i$.

Such estimators are approximately unbiased for target population values in the sense of repeated inclusion in the sample under the pseudo-probability distribution.

The difference from pure design-based inference is that we do not have control over the $\pi(i \in s|\boldsymbol{x_i}; \boldsymbol{\Phi})$'s.

Methodology: Estimating from a nonprobability sample (VALLIANT, 2020)

Reference sample:

- Must “represent” the full target population: the weights in the reference sample must inflate it to the target population.
- Must include the same covariates as the nonprobability survey.
- Is combined with the nonprobability sample, so the pseudo-inclusion probabilities for the nonprobability cases are estimated using binary regression model.

Methodology: Estimating from a nonprobability sample (VALLIANT, 2020)

Estimation procedure:

(1) Code the cases in the reference sample as 0 and the cases in the nonprobability sample as 1.

(2) Reference sample cases receive their probability sample weight. Assign a weight of 1 to each nonprobability case.

(3) Fit a weighted binary regression to predict the probability of being in the nonprobability sample.

This weighted regression will approximately estimate the census model that would be fit if the reference sample were the entire population, excluding the nonprobability sample.

Reference sample: PNS 2019

- National Survey of Health (PNS) from IBGE - the official statistics agency in Brazil.
- Information on the performance of the national health system. It also investigates the population's health conditions, records of chronic non-communicable diseases and respective risk factors.
- Probabilistic sample.
- It produces estimates disaggregated by sex and age groups, level of schooling, color or race, employment status, for Brazil, Major Regions and Federation Units.
- It includes the same covariates (social factors) as the nonprobability survey.

Results

IOS-ISI 2024
MEXICO CONFERENCE

Results

TABLE 1 | Frequency distribution for social determinants

Characteristic	Non-probability sample	Non-probability sample with pseudo-inclusion probabilities	Reference sample (populational estimates)
	N = 1.482	N = 32.886.581	N = 36.871.002
Female gender	73,9%	52,0%	56,6%
Age groups (years)			
60-69	56,1%	61,5%	54,4%
70-79	28,4%	23,9%	30,1%
80 and over	15,5%	14,6%	15,5%
Ethnicity			
White	61,7%	49,3%	51,1%
Black	6,8%	12,2%	10,1%
"Pardo"	29,6%	36,0%	37,0%
"Amarelo"	1,6%	2,0%	1,3%
Indigineous	0,3%	0,5%	0,5%
Marital status			
Single	10,3%	15,5%	14,0%
Married	53,7%	42,6%	45,6%
Divorced	12,4%	9,3%	6,1%
Widowed	23,6%	32,5%	34,4%

Results

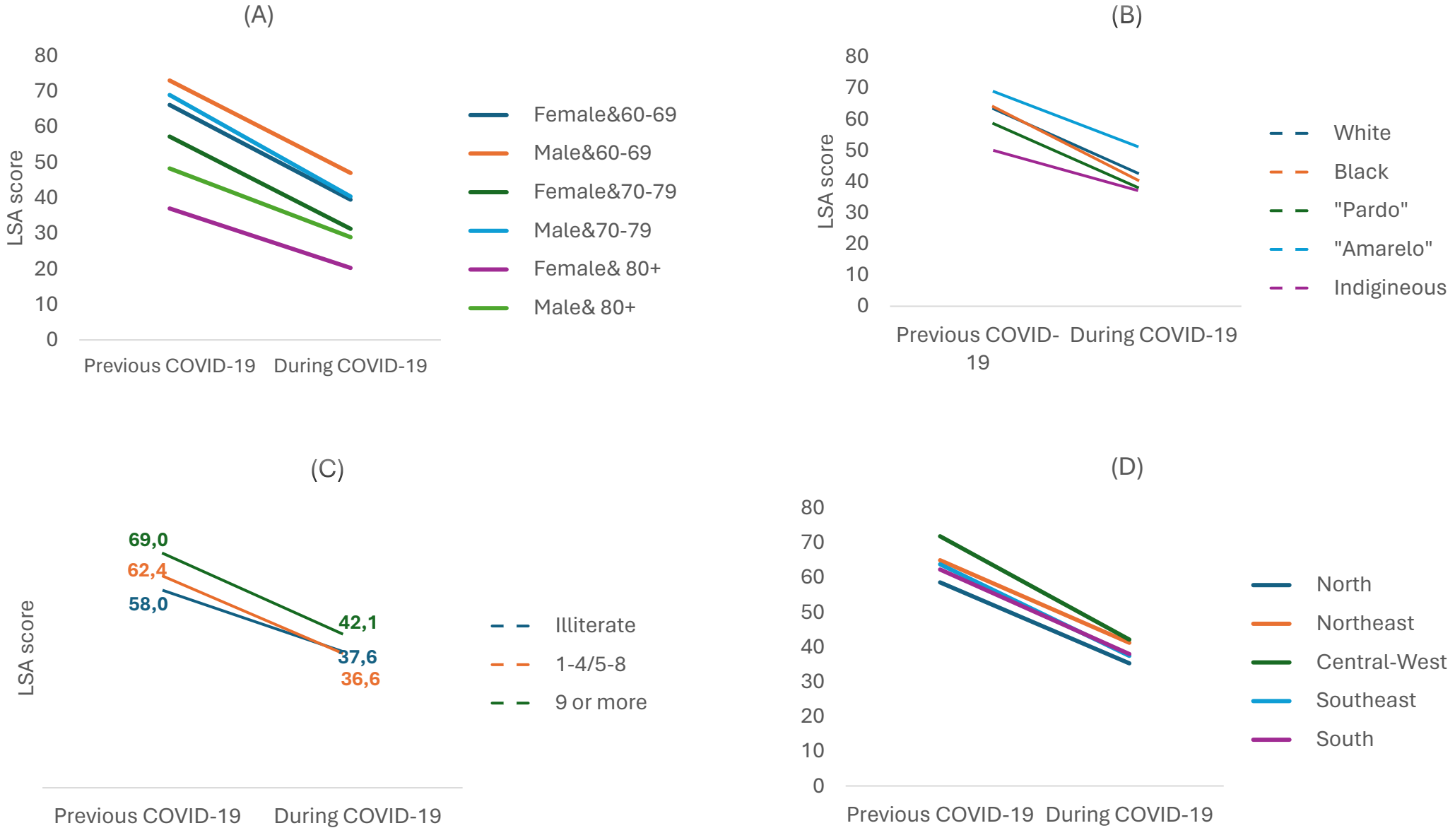
TABLE 1 | Continued

Characteristic	Non-probability sample	Non-probability sample with pseudo-inclusion probabilities	Reference sample (populational estimates)
	N = 1.482	N = 32.886.581	N = 36.871.002
Educational level (years of schooling)			
Illiterate	7,9%	15,0%	20,1%
1-4/5-8	31,2%	50,3%	53,9%
9 or more	60,9%	34,7%	26,0%
Income (minimum wage salary)[a]			
<1	34,5%	47,6%	
2-3	27,9%	33,3%	
4-7	18,0%	11,9%	
8-10	7,7%	2,9%	
10 or more	11,9%	4,3%	
Employment			
In labor force (active/unemployed)	43,6%	23,9%	8,5%
Not in labor force (inactive)	56,4%	76,1%	91,5%
Region of Brazil			
North	6,9%	8,9%	6,1%
Northeast	42,7%	28,2%	25,4%
Central-West	3,6%	5,9%	6,4%
Southeast	43,0%	48,1%	46,4%
South	3,8%	8,9%	15,7%

[a]Brazilian minimum wage salary 1,045.00 BRL (corresponding to 189.3 USD; 1st May 2020)

Results

FIGURE 1| Estimated life-space mobility scores before and since COVID-19 pandemic by (a) gender and age group, (b) ethnicity, (c) educational level and (d) region of Brazil.



Conclusions and future work

When combining the (biased) nonprobability sample and IBGE PNS:

- The sociodemographic characteristics in the nonprobability survey were corrected and closer to the populational distribution.
- Life-space mobility scores were estimated before and since COVID-19 pandemic, with pseudo-inclusion probabilities.

Future work focus on using the pseudo-inclusion probabilities for fitting regression models for LSA and other mobility and health measures collected in the survey.

Bibliographic references

Lewnard JA, Lo NC. **Scientific and ethical basis for social-distancing interventions against COVID-19**. *Lancet Infect Dis*. (2020) 20:631–3.

Morley JE, Vellas B. **Editorial: COVID-19 and older adults**. *J Nutr Health Aging*. (2020) 24:364–5.

Peel C, Sawyer Baker P, Roth DL, Brown CJ, Brodner EV, Allman RM. **Assessing mobility in older adults: the UAB Study of Aging Life-Space Assessment**. *Phys Ther.* (2005) 85:1008–119.

Rantakokko M, Portegijs E, Viljanen A, Iwarsson S, Rantanen T. **Life-space mobility and quality of life in community-dwelling older people.** *J Am GeriatrSoc*. (2013) 61:1830–2.

Simoes M, Garcia IF, Costa LDC, Lunardi AC. **Life-Space Assessment questionnaire: Novel measurement properties for Brazilian Community dwelling older adults**. *Geriatr Gerontol Int*. (2018).

Valliant R. **Comparing alternatives for estimation from nonprobability samples**. *Journal of Survey Statistics and Methodology* (2020) 8, 231–263.

Valliant R, Dever JA, Kreuter F. **Practical Tools for Designing and Weighting Survey Samples** (2018), *New York: Springer.*

Xue QL, Fried LP, Glass TA, Laffan A, Chaves PH. **Life-space constriction, development of frailty, and the competing risk of mortality: the Women's Health And Aging Study** *I. Am J Epidemiol.* (2008) 167:240–8.

IOS-ISI 2024
MEXICO CONFERENCE

Thank you

maria.toledo@ibge.gov.br

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

IAOS-ISI 2024
MEXICO CONFERENCE

Prevalence and Associated Factors of Preterm Births Among Reproductive-age women in Zambia

By Mutale Sampa-Kawana
University of Zambia

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

Background

- Preterm birth is defined as babies born alive before 37 weeks of pregnancy (WHO, 2020)

- Preterm birth is a huge global public health concern with an estimated 15 million babies born prematurely annually, with over 90% of these births occurring in LMICs.

- The preterm birth rates vary across regions, from as low as 5% in some developed countries to as high as 18% in LMICs, with the highest rates in SSA.

- In Zambia annually, there are approximately 77,600 preterm births.

Objectives

1. To estimate the prevalence of preterm births in Zambia using the 2018 ZDHS data.

2. To estimate the distribution of preterm birth by wealth quintile and rural and urban regions.

3. To determine the factors associated with preterm birth in Zambia.

Methods

IOS-ISI 2024
MEXICO CONFERENCE

Study Design,Data Source, Population and Sample Size

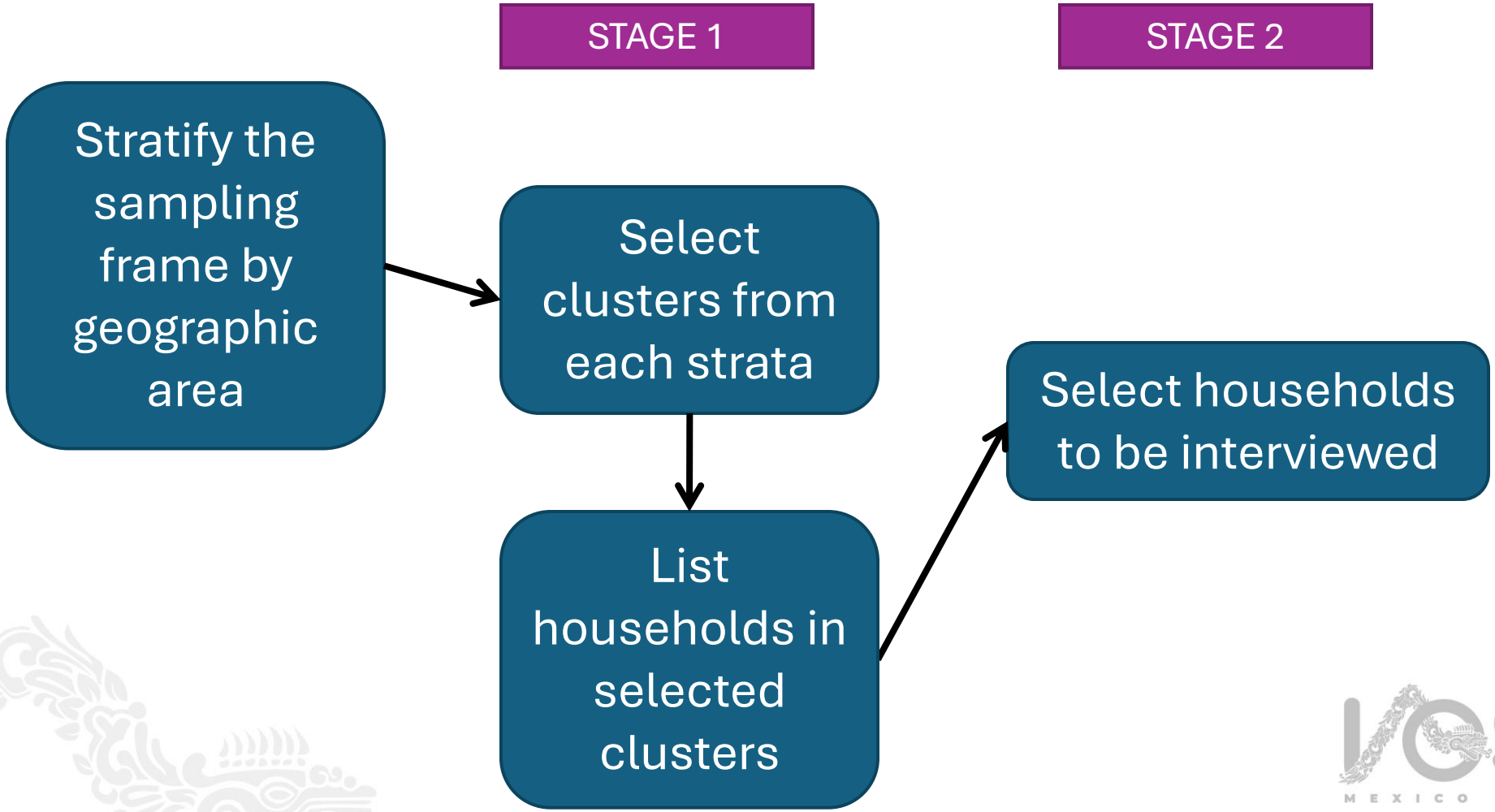
Study Design: Cross-sectional

Data source: Zambia Demographic Health Survey (ZDHS)

Study Population: women of reproductive age (15-49 years) who had given birth in the five years preceding the survey

Sample Size: 10,962 women

ZDHS Sampling Technique



Study Variables

❖ **Duration of pregnancy** was used to generate the outcome variable **preterm birth**

- Preterm birth if duration ≤8 months
- Full-term birth if Duration ≥9 months

Explanatory Variable

1. Age
2. Level of Education
3. Marital status
4. Wealth quintile
5. Employment status
6. Number of ANC visits
7. Contraceptive use
8. History of a terminated pregnancy
9. previous delivery by cesarean section

Data Management and Statistical Analysis

- Survey analysis methods were used to account for the complex survey design.

- Setting up data for survey command in Stata

```
svyset cluster [pweight=weight], strata(strata) vce(linearized)
```

- **Chi-square** test to check for an association between preterm birth and explanatory variables

- **Equiplots** to show the distribution of preterm births by wealth quintile and region

Chi-square test to check association

Survey Logistic regression to determine factors associated with preterm births

Model Diagnostics

❖ Pearson or Homsmer-Lemeshow goodness-of-fit test

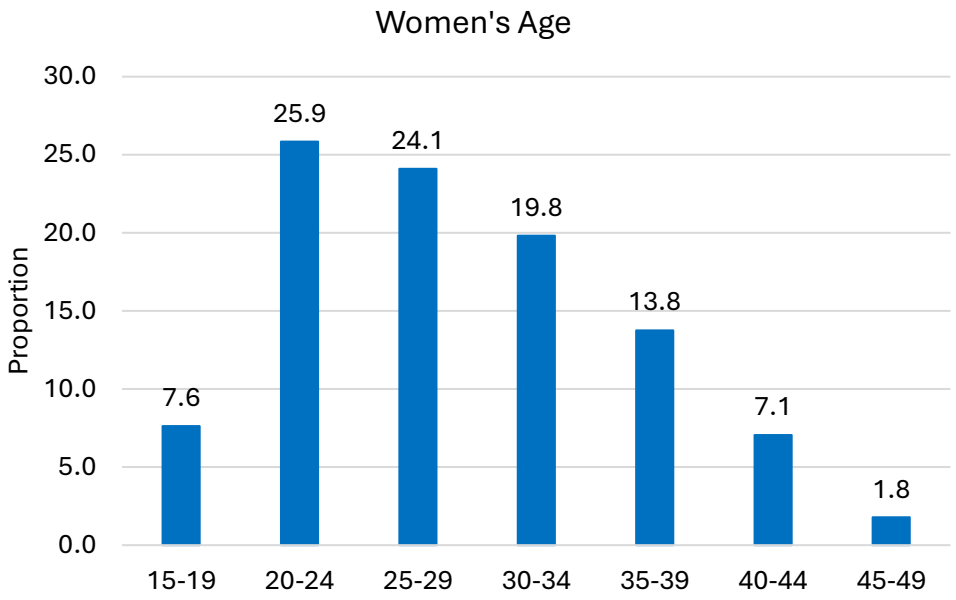
- Small Chi-squared values (with a larger p-value closer to 1) indicate a good logistic regression model fit.

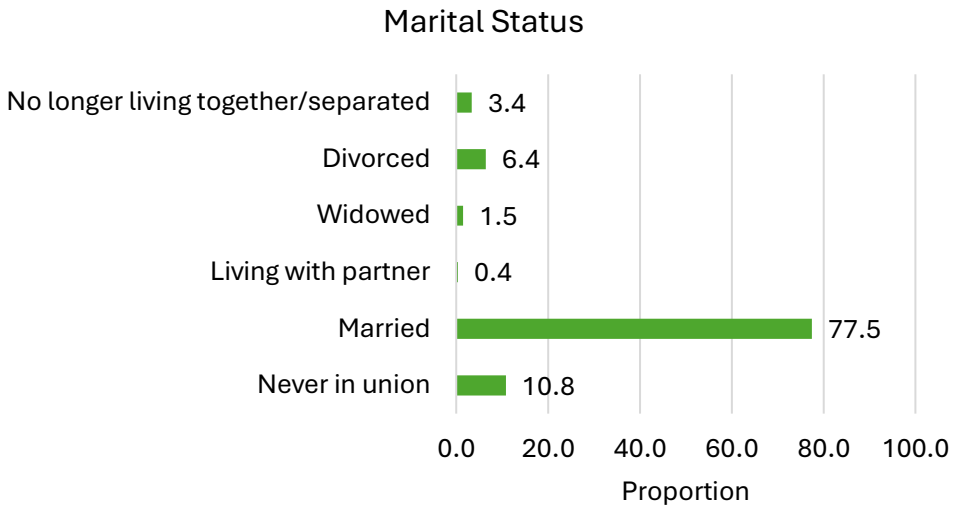
1. 2.66, P-value <0.0051 Full model
2. 1.25, P-value=0.26 Model without Marital Status
3. 1.30, P-value=0.2318 Model without Marital Status and Employment
4. 1.45, P-value=0.1655 Model without Marital Status, Employment and Education
5. 1.46, P-value= 0.1598 Model without Marital Status, Employment, Education, and Residence

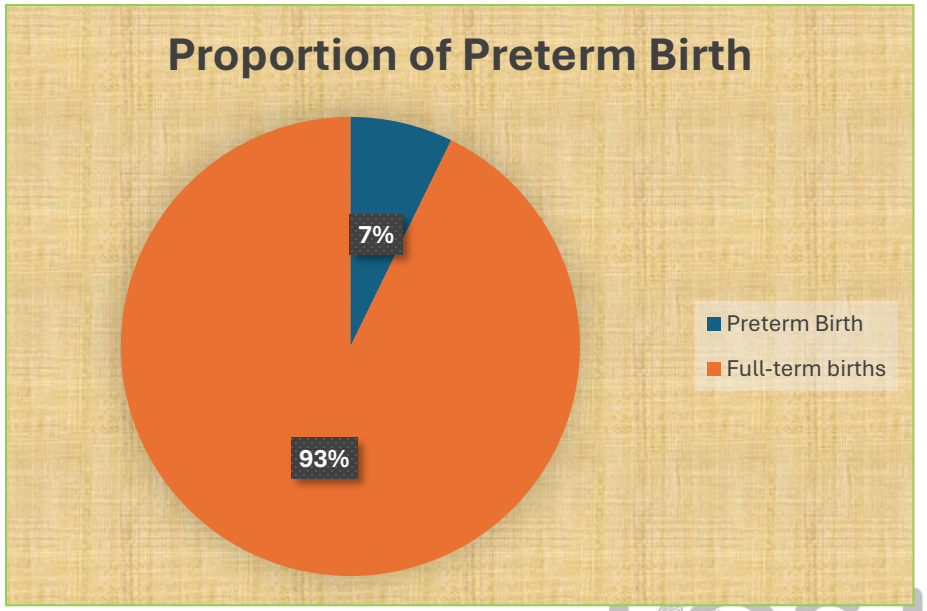
Results

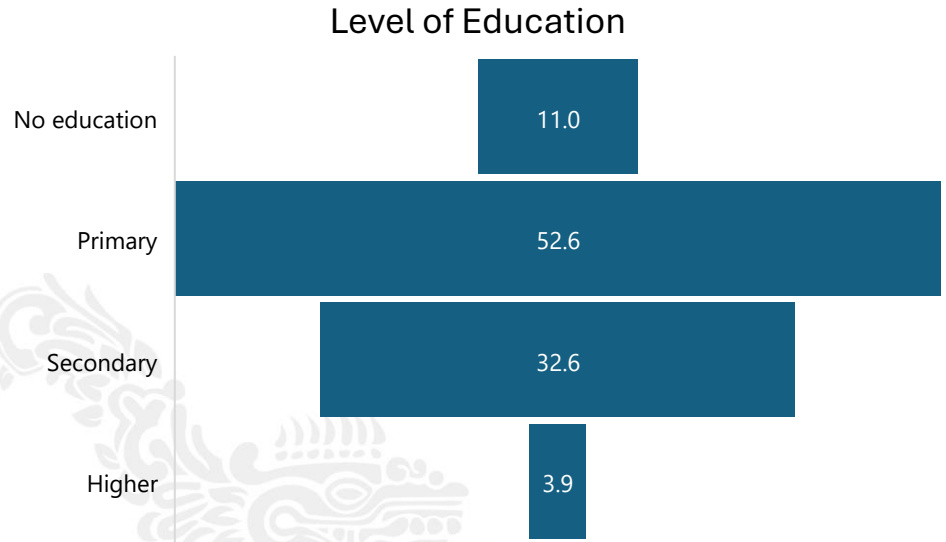
IOS-ISI 2024
MEXICO CONFERENCE

Participant Characteristics

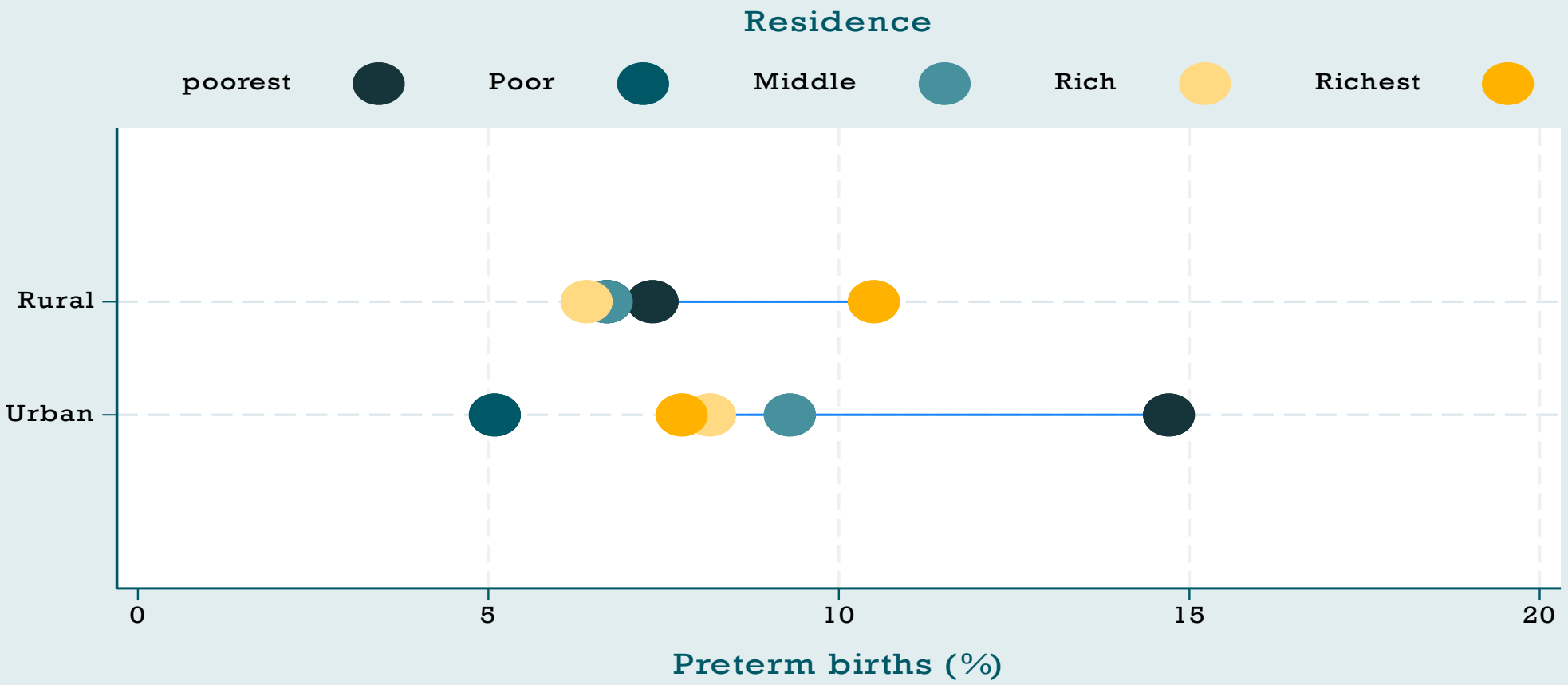








Distribution of Preterm Births by Residence and Wealth Quintile



Association of Preterm Birth and Demographic Characteristics

Variable	Full term Birth	Preterm Birth	P-Value
Marital Status			
Never in union	1095 (10.78%)	91 (11.36%)	
Married	7891 (77.66%)	601 (75.03%)	
Living with partner	39 (0.38%)	2 (0.25%)	0.2482
Widowed	148 (1.46%)	17 (2.12%)	
Divorced	652 (6.42%)	54 (6.74%)	
No longer living together/separated	336 (3.31%)	36 (4.49%)	
Highest educational level			
no education	1126 (11.1%)	79 (9.9%)	
primary	5373 (52.9%)	387 (48.3%)	0.013
secondary	3270 (32.2%)	299 (37.3%)	
Higher	392 (3.9%)	36 (4.5%)	
Age in 5-year groups			
15-19	752 (7.40%)	84 (10.49%)	
20-24	2597 (25.56%)	237 (29.59%)	
25-29	2454 (24.15%)	189 (23.60%)	0.001
30-34	2024 (19.92%)	148 (18.48%)	
35-39	1416 (13.94%)	92 (11.49%)	
40-44	734 (7.22%)	40 (4.99%)	
45-49	184 (1.81%)	11 (1.37%)	
Distance to the health facility			
Big problem	3756 (36.96%)	313 (39.08%)	0.234
Not a big problem	6405 (63.04%)	488 (60.92%)	
Region			
urban	2981 (29.3%)	257 (32.1%)	0.101
rural	7180 (70.7%)	544 (67.9%)	

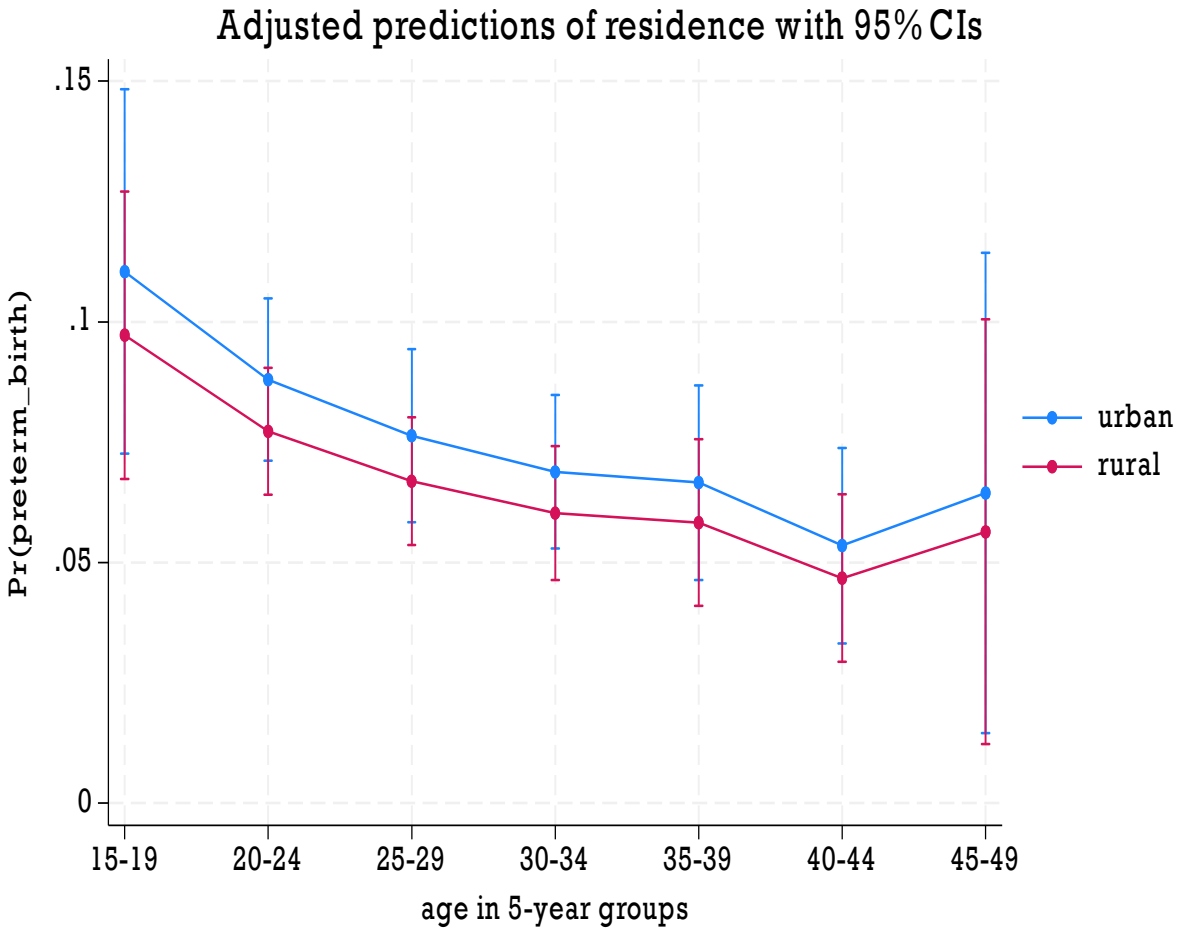
Association of Preterm Birth and Pregnancy-Related Characteristics

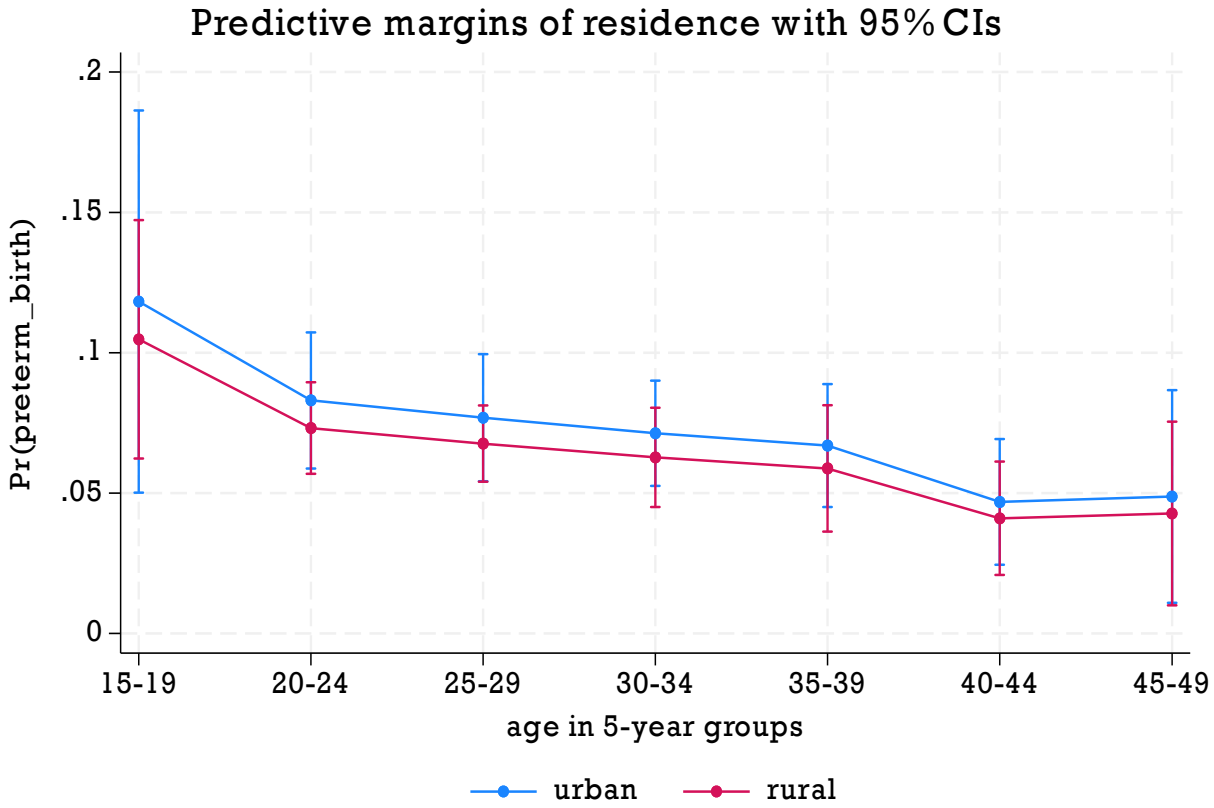
Variable	Full term Birth	Preterm Birth	P-Value
Number of ANC Visits			
No ANC visits	79 (1.16%)	10 (1.87%)	
Less than 4	2222 (32.49%)	254 (47.57%)	<0.0001
4 or more	4537 (66.35%)	270 (50.56%)	
Parity			
Less than Five	6722 (66.15%)	574 (71.66%)	
Five to Nine	3186 (31.36%)	202 (25.22%)	0.001
Ten or more	253 (2.49%)	25 (3.12%)	
Previous delivery by cesarean section			
No	9698 (95.44%)	754 (94.13%)	0.090
Yes	463 (4.56%)	47 (5.87%)	
History of a terminated pregnancy			
No	9256 (91.09%)	705 (88.01%)	0.004
Yes	905 (8.91%)	96 (11.99%)	
ANC in the first trimester			
No	4185 (61.92%)	336 (64.12%)	0.316
Yes	2574 (38.08%)	188 (35.88%)	

Factors Associated with Preterm Birth

Variable	Odds Ratio (95% Confidence)	p-value
Residence		
Urban	Ref	
Rural	0.829 (0.597-1.151)	0.262
Age in 5-year groups		
15-19	Ref	Ref
20-24	0.655 (0.384-1.119)	0.121
25-29	0.608 (0.382-0.968)	0.036
30-34	0.549 (0.305-0.988)	0.045
35-39	0.517 (0.260-1.026	0.059
40-44	0.352 (0.170-.0728)	0.005
45-49	0.365 (0.141-0.945)	0.038
Distance to the health facility		
Big problem	Ref	Ref
Not a big problem	0.856 (0.683-1.071)	0.174
Parity		
Less than 5	Ref	Ref
5 to 9	0.899 (0.653-1.236)	0.511
10 or more	2.609 (1.25-5.445)	0.011
ANC in the first trimester		
No	Ref	Ref
Yes	1.146 (0.912-1.439)	0.242
Number of ANC Visits		
Less than 4	Ref	Ref
4 or more	0.480 (0.385-0.599)	<0.0001
Ever had a terminated Pregnancy.		
No	Ref	Ref
yes	1.454 (1.063-1.988)	0.019

Age-Specific Probabilities of Preterm Birth by Residence





Conclusion

- The prevalence of preterm birth was found to be 7%

- The study found some inequalities in the distribution of preterm births.
  - In the urban areas, preterm birth is highest among the poorest.
  - In rural areas, preterm birth rates are highest among the richest.

- Age, parity, number of ANC visits , and history of a terminated pregnancy were factors associated with PB.

Limitations

- The definition of preterm birth was not very accurate, as the data did not include information on gestation age. Therefore, the proportion of preterm births may have been underestimated.

Recommendations

- Qualitative research should be conducted to understand the observed inequalities of preterm births in Rural and Urban Areas.

- The gestation age of the pregnancy should be included as a variable in the DHS data for a more accurate estimates

IOS-ISI 2024
MEXICO CONFERENCE

Thank you

INEGI
International Statistical Institute
IAOS
IMPROVING OFFICIAL STATISTICS

IAOS-ISI 2024, Mexico City
Improving Decision-Making for All
May 15, 2024 – May 17, 2024

Session 5: Better Understanding Of Demographic And Migration Statistics

Respondent Driven Sampling strategies for hard-to-reach populations

Giorgio Alleva, Piero Falorsi, Sapienza University of Rome, Italy
Stefano Falorsi, Paolo Righi, Istat, Italy
Andrea Fasulo, Eurostat

Outline

1. Lack of information on hidden and hard-to-count population groups
2. Background: the traditional RDS strategy and the VH estimator
3. RDS Data Collection
4. Proposed Estimators
5. A quasi unbiased strategy for a large scale survey.

Lack of information on hidden and hard-to-count population groups

In this paper, we focus on respondent-driven sampling (RDS) to estimate the size and characteristics of hidden or stigmatized populations or hard-to-measure population groups:

- homeless people, undocumented immigrants;
- women who have suffered violence, forced workers, HIV's;
- ethnic minorities, indigenous population, or transgender population.

They are finite populations, the size and composition of which is unknown, and it is not possible to investigate them through list (or area) sampling.

The principle of "leaving no one behind" is at the heart of the 2030 Agenda, and a key requirement for many Sustainable Development Goals (SDG) indicators is to be available for the most vulnerable and marginalised population groups.

Nevertheless most SDG indicators are still not available at the needed level of disaggregation to monitor the socioeconomic conditions of hidden and hard-to-count population groups.

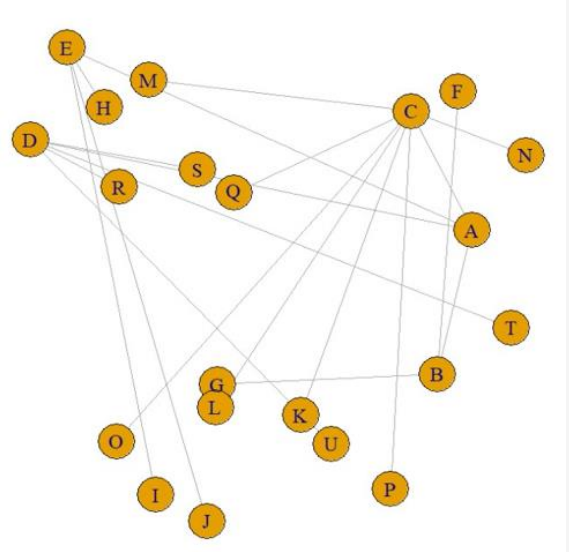
The disaggregation of data for SDG indicators on various hard-to-reach populations presents several critical issues that are difficult to overcome in the current context of official statistics in different countries.

It is very complex (and often impossible) to estimate the totals of variables related to those populations through models as in other situations.

Therefore, defining and implementing sampling strategies that can quickly improve this situation becomes necessary.

The respondent-driven sampling (RDS) method, exploiting existing connections among individuals of the target population, can be a helpful sampling tool to survey these populations.

Moreover, the effectiveness of the RDS can be further increased by employing an integrated approach in which the RDS is used in conjunction with other information sources, such as administrative or geographical data.



Background: the RDS method

The Respondent Driven Sampling (RDS) method (Heckathorn, 1997) is a network-based sampling technique.

Since its establishment, RDS has been employed in countless investigations of such populations across many nations (White et al., 2015).

It starts with a small sample of participants with which the researchers are familiar. Each participant is then given a small number of coupons with unique identifiers to distribute to their contacts in the target population, enrolling them in the study and increasing the sample size until the sample includes the desired number of respondents.

The sample evolves (adapts) with the progress of the interviews.

While the first selection is generally non-random, the selection of subsequent contacts is by random choice.

Background: the Volz and Heckathorn (VH) estimator of the total Y

$$\hat{Y}_{VH} = \sum_{k \in S} y_k w_{VH,k} \quad \text{where} \quad w_{VH,k} = \frac{N a_k L_k^{-1}}{\sum_{\ell \in S} a_\ell L_\ell^{-1}}$$

$a_k \rightarrow$ Number of times that unit k is selected in the RDS search process

$L_k \rightarrow$ Number of contacts of unit k.

$N \rightarrow$ Total number of people in the population

Feasibility

The major obstacle to estimating the total population is the need to know N. However, the VH estimator can be used to estimate the mean value of a characteristic y.

The need for a *Central unit* to avoid duplication of interviews and to record the a_k number implies good field organization of the survey process.

Objective of this presentation

The RDS method suffers lack of an estimation methodology that is sufficiently robust concerning varying conditions under which it is applied.

Even if it is advantageous for estimating mean and proportion values, the accuracy of the total estimates (total of variables or total unit of the population) depends on several features, including the nature of the network connecting the individuals in the population.

Below, we address the estimation problem by approaching the **RDS method as a particular indirect sampling technique** (Lavallé, 2007).

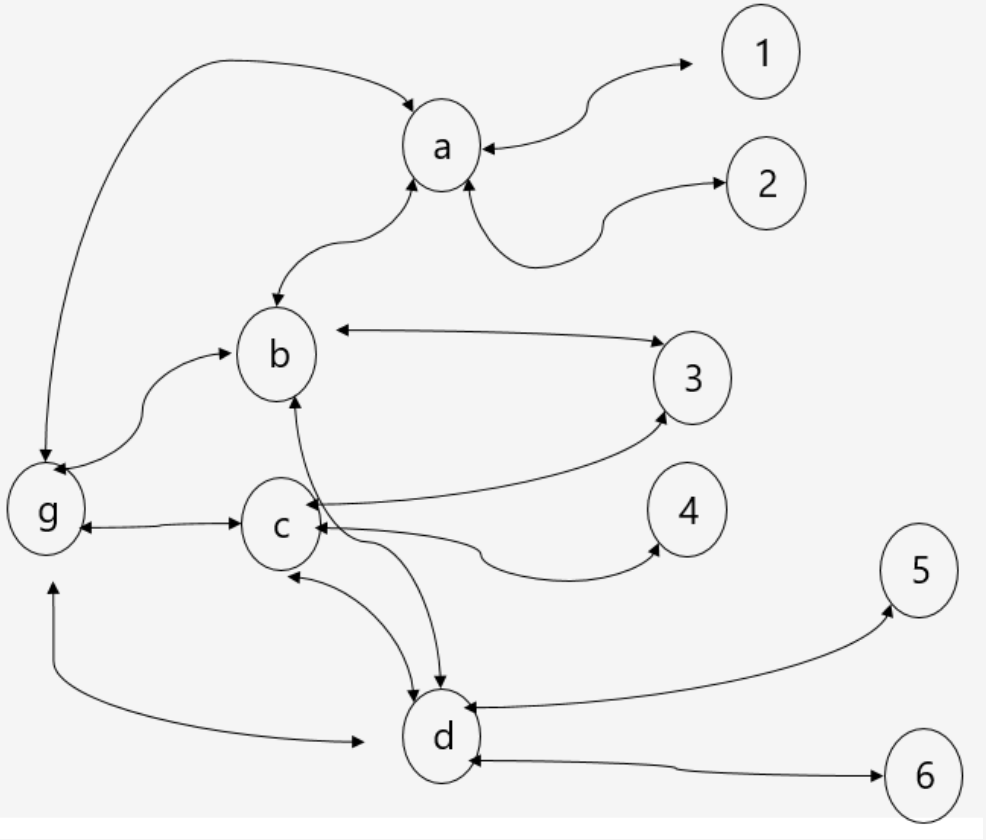
We address the estimation problem, and by approaching the RDS methodology as a particular indirect sampling technique, we propose *three unbiased estimation meth*ods as possible solutions.

In particular:

- the first method assumes a *random sampling* of the initial individuals;
- the second method considers *purposive sample selection into all the clusters of networks* that characterise the population of interest;
- the third method, leveraging the generalised capture-recapture estimation approach, consider an *estimator that accounts for the non-coverage of two independent indirect samplings*.

RDS Data Collection: an example of the research chain:

Consider the following graph representing the relationships connecting the units g, a, b, c, d, 1, 2, 3, 4, 5, 6.

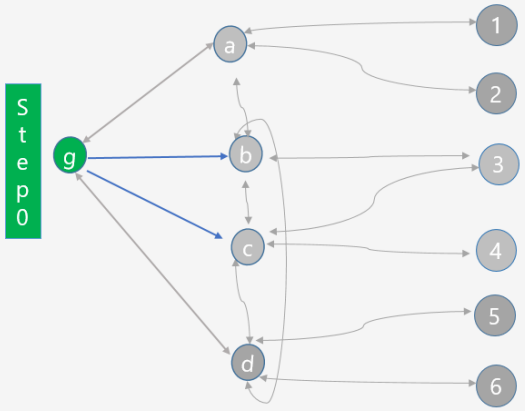


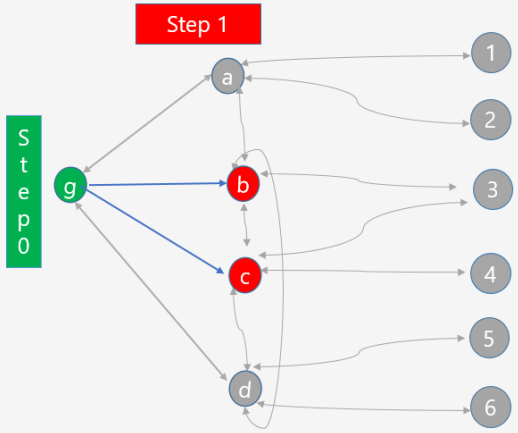
The relationship between two participants can be **direct** or **indirect**.

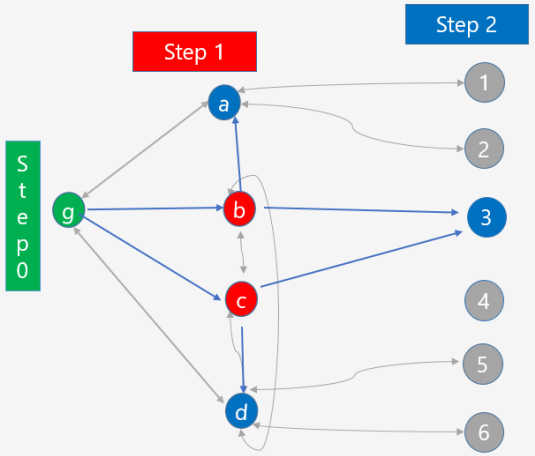
Direct Relations in the graph are **bi-univocal**.

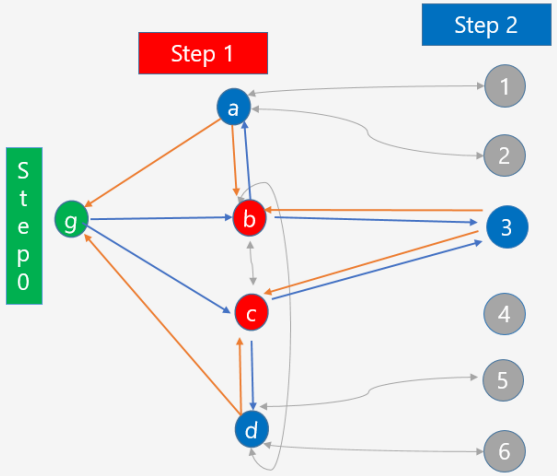
RDS data collection: procedures with and without “memory”

The traditional research chain (scheme 1)







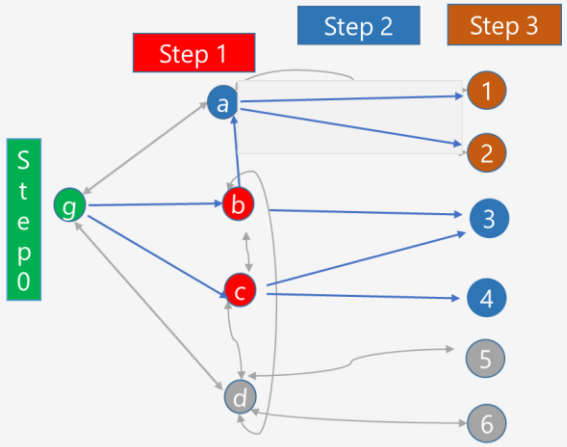


From units **a**, **3** and **d** we select units already interviewed and the process closes.

A more efficient research chain (scheme 2)

In every selection step each unit is randomly selected by excluding the units chosen in the previous steps.

So we can avoid possible “loops” in the search.



In Step 3 the RDS process stops since all the links of units 1 and 2 have already been involved in the sample.

RDS data collection: A more efficient research chain. Feasibility

To make the selection feasible, it is essential to know the *number of contacts that have not been selected in sample* S_r. Operationally, this quantity can be obtained in different ways.

Suppose not-identifiable but unique information about contacts of units included in the S_r sample is available in the data-collection APP used by the interviewer.

In that case, a specific software application can be launched that identifies units not included in S_r and proceeds to select units to be included in sample S_{r+1}randomly.

Alternatively, the same software application can be run by the *study centre* that supports the survey operations, and the results can be reported and provided in real-time to the interviewer who makes the S_{r+1} sample selection.

The indirect sampling mechanism

In indirect sampling, we have a U^A population of N^A units from which the research starts, and a U^B population of N^B units that constitute the study's target population.

The target parameter

$$Y = \sum_{k \in U^B} y_k.$$

may be viewed as the total

$$Y = \sum_{j \in U^A} \bar{y}_j^A$$

of the population U^A of the variables $\bar{y}_j^A$ where

$\bar{y}_j^A = \sum_{k \in U^B} \frac{\lambda_{j,k}}{L_k^B} y_k$ being $L_k^B = \sum_{j \in U^A} \lambda_{j,k}$

the total of direct links ($\lambda_{j,k}$) of Unit $k \in U^B$ with Unit $j \in U^A$.

The indirect sampling mechanism

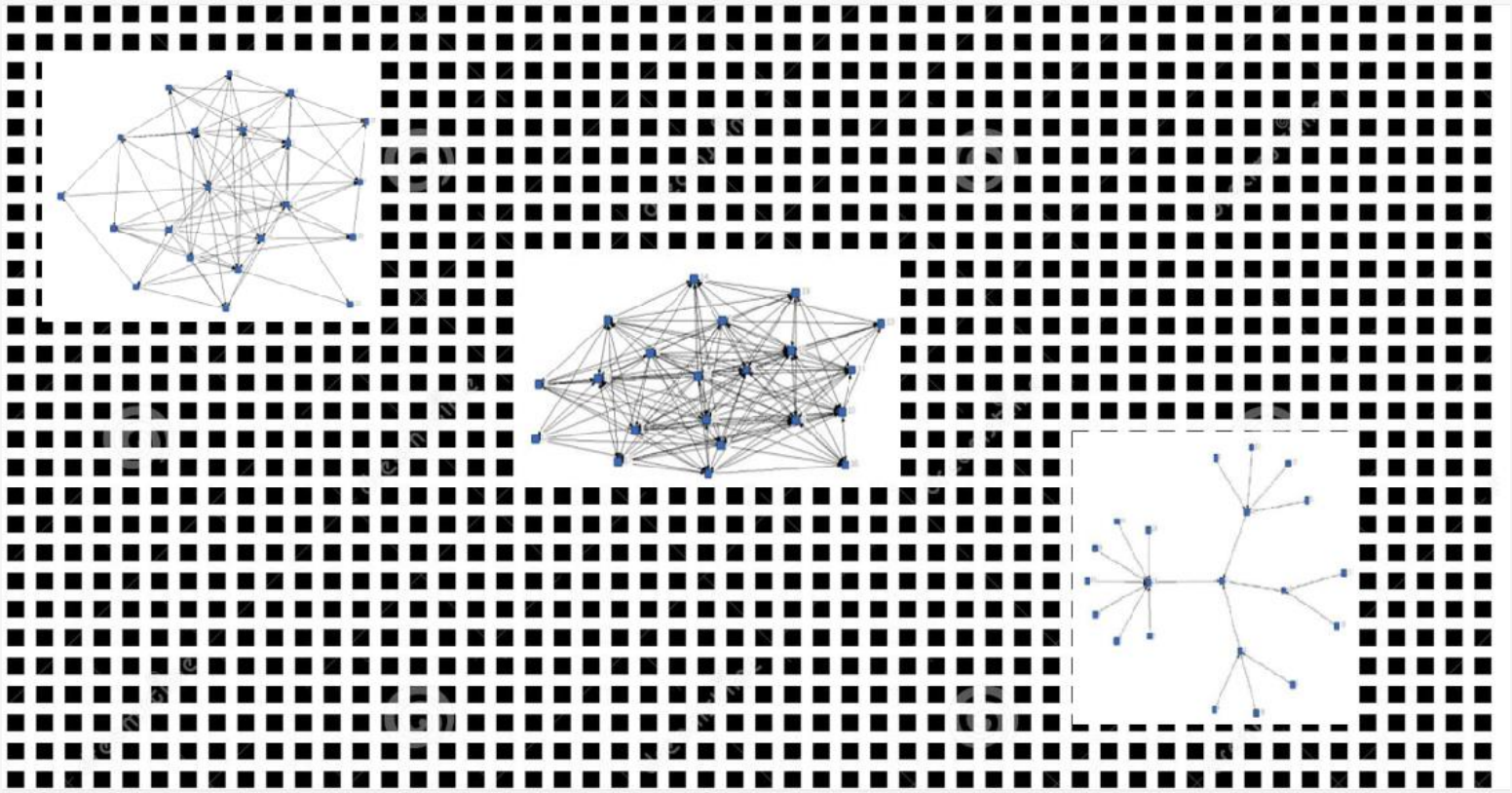
If the sample S_0 is selected non-randomly, it is possible to estimate the total $Y_{\overleftrightarrow{S_0}}$ of units directly or indirectly linked to the initial sample S_0

$$Y_{\overleftrightarrow{S_0}} = \sum_{k \in U^B_{S_0}} y_k = \sum_{j \in U^A_{S_0}} \bar{y}^A_j$$

$$= \sum_{j \in U^A_{S_0}} \sum_{k \in U^B_{S_0}} \frac{\lambda_{j,k}}{L^B_k} y_k \,.$$

Example. Three groups of separate units

$Y_{\overleftrightarrow{S_0}} < Y$ if S_0 does not cover all three groups



In the example, we assume that people of the target population belong to three disjoint clusters.

Since people of the indigenous population are grouped geographically, it is important to consider in S_0 sample all locations where people of the indigenous population are known to belong.

Therefore, observing each cluster's units in S_0 would be appropriate.

Estimator of the total $Y_{\overleftrightarrow{S_0}}$

Let r be the step where the RDS process stops.

The unbiased estimator $\hat{Y}_{\overleftrightarrow{S_0}}$ of $Y_{\overleftrightarrow{S_0}}$ can be obtained as:

$$\hat{Y}_{\overleftrightarrow{S_0}} \cong \sum_{j \in S_0} \dots \sum_{j_{r-1} \in S_{r-1}} \sum_{k \in S_r} y_k \frac{\lambda_{j,j_1}}{L^B_{j_1}} \times \dots \times \frac{\lambda_{j_{r-1},k}}{L^B_k} \left(\frac{1}{\tau_{j_1|j \in S_0}} \times \dots \times \frac{1}{\tau_{j_{r-1}|j_{r-2} \in S_{r-2}}} \frac{1}{\tau_{k|j_{r-1} \in S_{r-1}}} \right)$$

$$= \sum_{k \in S_r} y_k w_{(S_0)k_r} \quad \text{where}$$

$$w_{(S_0)k_r} \cong \sum_{j \in S_0} \dots \sum_{j_{r-1} \in S_{r-1}} \frac{\lambda_{j,j_1}}{L^B_{j_1}} \times \dots \times \frac{\lambda_{j_{r-1},k}}{L^B_k} \left(\frac{1}{\tau_{j_1|j \in S_0}} \times \dots \times \frac{1}{\tau_{j_{r-1}|j_{r-2} \in S_{r-2}}} \frac{1}{\tau_{k|j_{r-1} \in S_{r-1}}} \right)$$

The estimator $\hat{Y}_{\overleftrightarrow{S_0}}$ is unbiased for $Y_{\overleftrightarrow{S_0}}$ if r is greater than the maximum of the minimum paths between any pair of nodes in each cluster of the units of S_0.

Proposal of a strategy for large scale survey

Select Primary Sampling Units (PSU) geographically spread and with probability proportional to the expected size of the target population (PPS), e.g.:

- **Method (1):** Probability Proportional to Size (PPS) sampling with probabilities proportional to degree of concentration of area units (if relevant information is available in the area frame).
- **Method (2):** Disproportionately sampling in the strata with high concentrations.
- **Method (3):** Ranking the area units by broad categories of concentration and using the ranks in giving each area unit a score equal to its rank.
- **Method (4)**: Optimal first stage sampling (Falorsi and Righi, 2015), based on proxy information with measurement error on the first stage units.

Find seeds in each PSU, aiming to represent all the key socioeconomic subpopulations that researchers anticipate may exist in the target population, seeds are selected to be as varied as possible.

The RDS sampling search is carried out within each sample PSU. This improves the feasibility of the RDS search.

Proposal of a strategy for large scale survey

The total Y may be estimated as

$$\hat{Y} = \sum_{i \in S_{PSU}} \frac{1}{\pi_i} \hat{Y}_{i,\overleftrightarrow{S_0}}$$

Conclusions

- The disaggregation of data for SDG indicators on hard-to-reach populations presents several critical issues that are difficult to overcome in the current context of official statistics in different countries. In this context, it is impossible to estimate the characteristics of indigenous people through models as in other situations.
- Therefore, defining and implementing a sampling strategy that can quickly improve this situation becomes necessary. It is helpful to consider sampling designs which maximise the number of observed individuals of the target population.
- The respondent-driven sampling (RDS) method, based on existing connections among individuals of the target population, can be a helpful sampling tool to survey these populations.

What we have presented here represents ongoing research, the initial results of which are encouraging.

Open Issues

The research team is currently running experiments on simulated data and the empirical results will be presented in a new paper.

Some Open Issues

- How many steps we need to produce good estimates of the total?
- How to take into account in the stopping rule of the intensity and the structure of the connections?

For the feasibility of the proposed sampling strategy it is fundamental:

- to collect and exploit useful information for defining the initial seeds (people and geographical areas) from administrative sources, previous surveys or case studies;
- a clear protocol of data collection on the field and the recruitment of professional enumerators;
- the support of a *Central unit* for monitoring contacts and identificative variables (names and addresses, others) and to interact with enumerators during the field operation.

References

Johnston LG. 2013. Introduction to HIV/AIDS and sexually transmitted infection surveillance. Module 4. Introduction to respondent-driven sampling. World Health Organ., Geneva. http://www.lisagjohnston. com/respondent-driven-sampling/respondent-driven-sampling.

Johnston LG. 2007. Conducting respondent driven sampling studies in diverse settings: a manual for planning RDS studies. Cent. Dis. Control Prev., Atlanta, GA.

Gile K, Handcock MS. 2015. Network model-assisted inference from respondent-driven sampling data. J. R. Stat. Soc. A 178(3): 619–39.

Gile K, Johnston LG, Salganik MJ. 2015. Diagnostics for respondent-driven sampling. J. R. Stat. Soc. A 178(1):241–69.

Gile K, Beaudry. I S., Handcock M.S. and Miles Q. 2018. Methods for Inference from Respondent-Driven Sampling Data. Annu. Rev. Stat. Appl. 2018. 5:4.1–429

Goodman LA. 1961. Snowball sampling. Ann. Math. Stat. 32:148–70.

Handcock MS, Gile KJ. 2011. Comment: on the concept of snowball sampling. Sociol. Methodol. 41(1):367–71.

Hansen MH, Hurwitz WN. 1943. On the theory of sampling from finite populations. Ann. Math. Stat. 14(4):333–62.

Heckathorn DD. And Cameron, J. 2017. Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling. Annual Review of Sociology · August 2017.

Heckathorn DD. and Jeffri. 2001. Finding the Beat. Using respondent-driven sampling to study jazz musicians. Poetics. 28 2001. P. 307-329. Elsevier.

Heckathorn DD. 2008. Assumptions of RDS: analytic versus functional assumptions. Presented at CDC Consult. Anal. Data Collect. Respond.-Driven Sampl., Atlanta, GA.

Heckathorn DD. 2002. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Soc. Probl. 49:11–34.

Heckathorn DD. 1997. Respondent driven sampling: a new approach to the study of hidden samples. Soc. Probl. 44(2):174–99.

Lavallée P. 2007. Indirect Sampling. Springer. New York.

Lavallée P., Rivest L. P., 2012. Capture–Recapture Sampling and Indirect Sampling. Journal of Official Statistics, Vol. 28, No. 1, 2012, pp. 1–27.

Matthew J. Salganik; Douglas D. Heckathorn. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. Sociological Methodology, Vol. 34. (2004), pp. 193-239.

Salehi M.M. and Seber G. A. F. 2002. Unbiased Estimators for Restricted Adaptive Cluster Sampling. Australian and New Zealand Journal of Statistics. 44. 63-74.

Salehi M.M. and Seber G. A. F 2001. A New Proof of Murthy Estimator which applies to Sequential Sampling. Australian and New Zealand Journal of Statistics. 43. 281-286.

Verdery AM, Merli MG, Moody J, Smith J, Fisher JC. 2015a. Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China. Epidemiology 26:661.

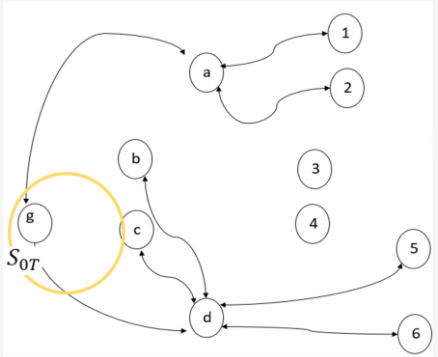
Volz E, Heckathorn DD. 2008. Probability based estimation theory for respondent driven sampling. J. Official Statistics. 24:79.

The three proposed estimators

First estimator: random selection of the initial sample

In each step of the RDS mechanism, an unbiased estimator of the total *Y* can be obtained. In the initial sample the total *Y* for the target people may be estimated with the standard Horvitz-Thompson estimator:

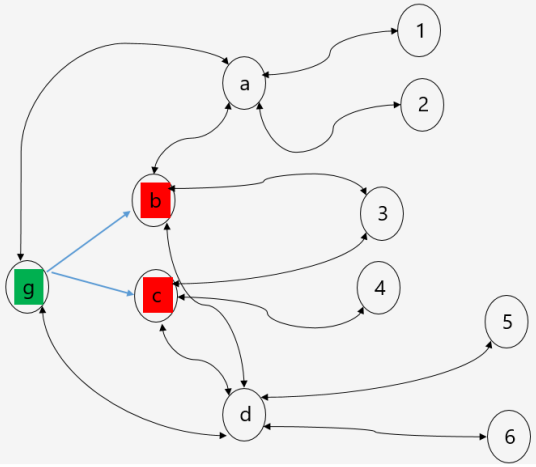
$\hat{Y}_0 = \sum_{j \in S_{0T}} y_j \frac{1}{\pi_j} = \sum_{k \in S_{0T}} y_k w_k$ where π_j is the inclusion probability.



Sample S_1 is formed by taking all the participants of sample S_{0T} plus the set S_1^+ including the participants randomly selected from the links of S_{0T}.

$$\boldsymbol{S_1 = S_{0T} + S_1^+}$$

S_1^+ is formed selecting, independently, $\bar{m}$ units (e.g. 2 or 3) for each unit in S_{0T} from the L_j^A units that are their direct contacts.



The unbiased estimator of *Y* based on $\boldsymbol{S_1}$ through the RDS process can be expressed in the standard *weighted form:*

$$\hat{Y}_1 = \sum_{k \in S_1} y_k w_k \text{ , where } w_k = \sum_{j \in S_{0T}} \frac{\lambda_{j,k}}{L_k^B} \left(\frac{1}{\pi_j} \frac{1}{\tau_{k|j \in S_{0T}}} \right) \quad \text{and} \quad \tau_{k|j \in S_{0T}} = \begin{cases} 1 & \text{if } j = k \\ \frac{\bar{m}}{L_j^A} & \text{otherwise} \end{cases}$$

Continuing the above illustrated process recursively, in the rth step, we form the sample $\boldsymbol{S_r}$ by taking all the participants of sample S_{r-1}, to which we add the participants randomly selected from the links of S_{r-1}^{+}.

The **conditional** probability that unit k is selected in sample S_r, given $j_{r-1} \in S_{r-1}$ is:

$$\tau_{k|j_{r-1}\in S_{r-1}} = \begin{cases} 1 & \text{if } k = j_{r-1} \\ \dfrac{\bar{m}}{L_{j_1}^{A}} & \text{otherwise} \end{cases}$$

The unbiased estimator of Y in $\boldsymbol{S_r}$ is: $\hat{Y}_r = \sum_{k\in S_r} y_k w_k$

where $w_k = \sum_{j\in S_{0T}} \cdots \sum_{j_{r-1}\in S_{r-1}} \frac{\lambda_{j,j_1}}{L_{j_1}^{B}} \times \cdots \times \frac{\lambda_{j_{r-1},k}}{L_k^{B}} \left(\frac{1}{\pi_j} \frac{1}{\tau_{j_1|S_0}} \times \cdots \times \frac{1}{\tau_{k|S_{r-1}}} \right)$

Conclusively, in this first sampling scheme, the design should maximize in the initial sample S_0 the number of observed individuals of the target population by adopting proper choices. In particular:

- to oversample areas where we have *a priori* information of a high concentration of the target population;
- to take into account auxiliary variables predictive of membership in the target population.

Second estimator: non-random selection of the initial sample

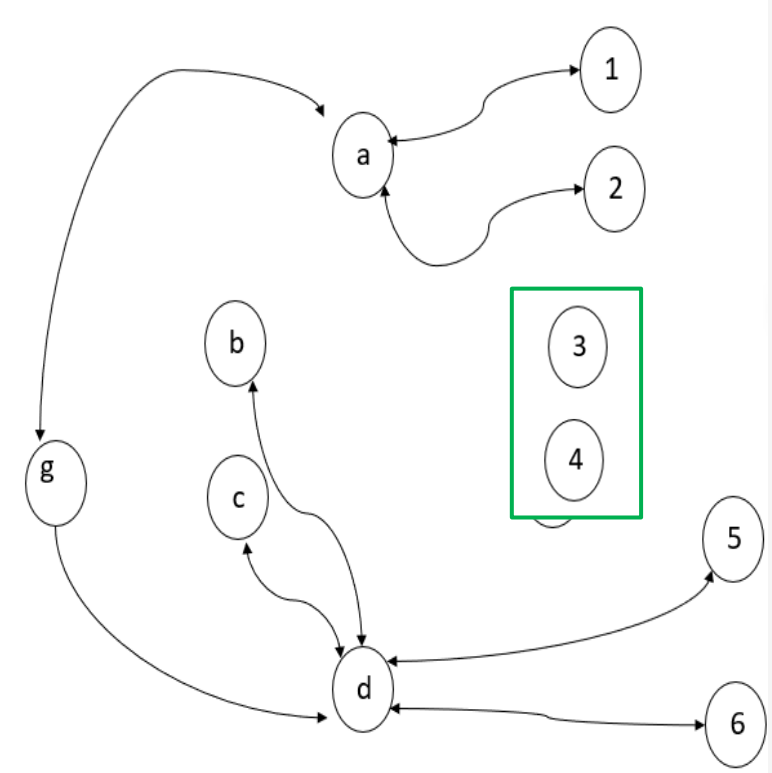
The S_0 sample is selected in a non-random mode:

In this case, we can only obtain a correct estimate of the set of units directly or indirectly connected with the participants of S_0.

We denote this total as $Y_{S_0\rightarrow}$

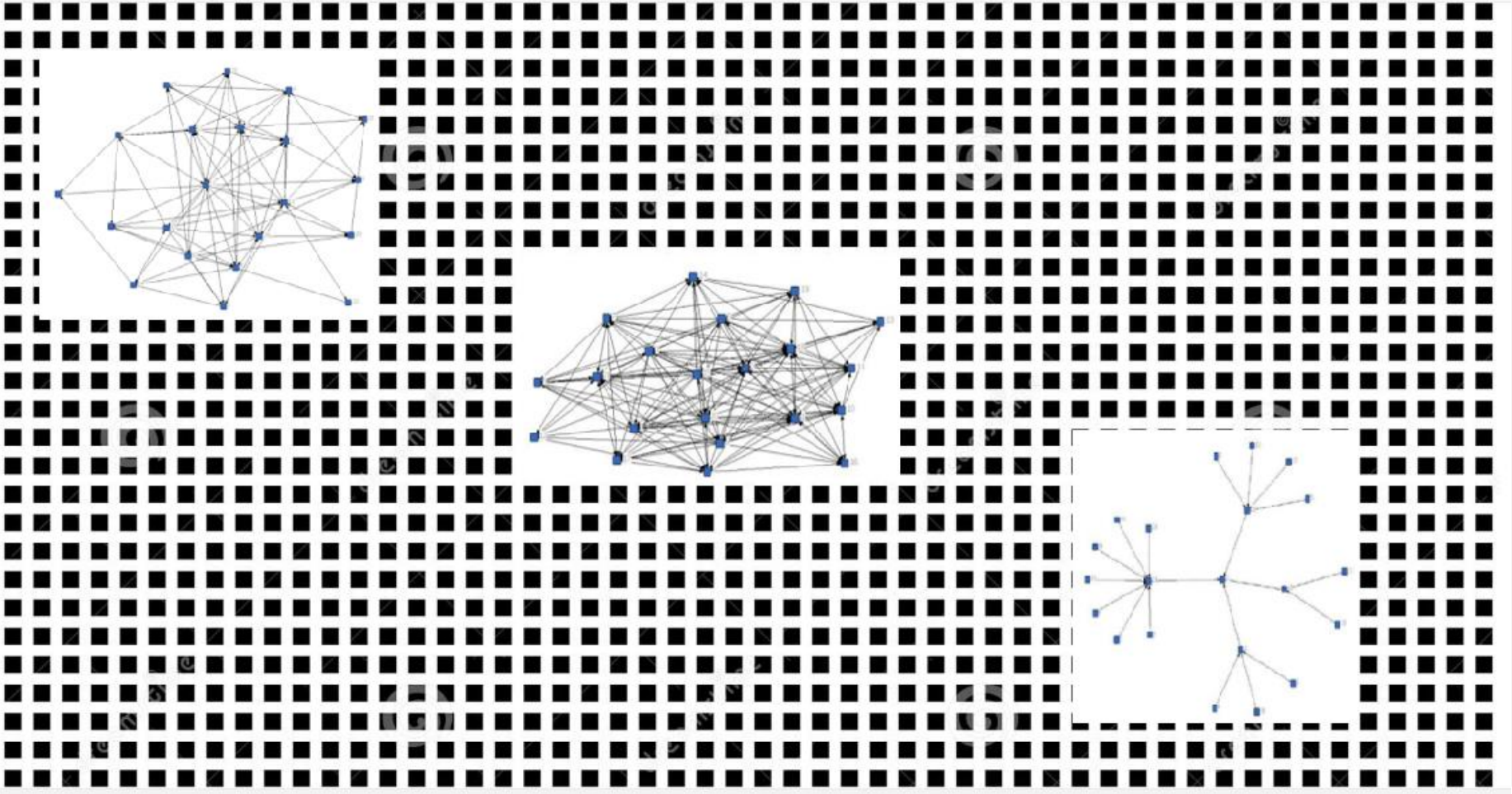
In the example we are considering $Y_{S_0\rightarrow}$ as the sum of the variable y of all the units excluding 3 and 4.

- If there are clusters that include people of the target population *unconnected* with those in S_0, we have $Y_{S_0\rightarrow} < Y$.
- If the participants of S_0 fall into all *disjointed* clusters in which the population of interest is organised, $Y_{S_0\rightarrow}$ coincides with the total Y.



Example of three groups of separate units

$Y_{S_0\rightarrow} < Y$ if S_0 does not cover all the following three groups



Let r be the step where the RDS process stops.

In this second scheme of sample design the unbiased estimator $\hat{Y}_{(S_0)r}$ of $Y_{(S_0)}$ can be obtained as:

$$\hat{Y}_{(S_0)r} = \sum_{k \in S_r} y_k w_{(S_0)k}$$

where $w_{(S_0)k} = \sum_{j \in S_0} \cdots \sum_{j_{r-1} \in S_{r-1}} \frac{\lambda_{j,j_1}}{L^B_{j_1}} \times \cdots \times \frac{\lambda_{j_{r-1},k}}{L^B_k} \left(\frac{1}{\tau_{j_1|j \in S_0}} \times \cdots \times \frac{1}{\tau_{k|j_{r-1} \in S_{r-1}}} \right)$

Note: The estimator $\hat{Y}_{(S_0)r}$ is unbiased for $Y_{S_0 \rightarrow}$ if r is greater than the maximum of the shortest paths between any pair of nodes in each cluster of the units of S_0.

Third estimator for dealing under-coverage

Even if the S_0 sample is randomly selected, the *first estimator* $\hat{Y}_r$ may be biased:

under-coverage may occur if respondents do not trust the interviewers and tend to hide their status.

Likewise, if the S_0 sample is non-randomly chosen, the *second estimator* can be affected by under-coverage if total $Y_{S_0\rightarrow}$ does not coincide with Y.

The Generalised Capture-Recature estimator (CReG) (Lavallé and Rivest, 2012), allows us to overcome both of the above mentioned forms of under coverage leveraging on a capture-recapture perspective

$$\hat{Y}_{CReG} = \frac{\hat{Y}_r \times \hat{Y}_{(S_0)r}}{\hat{Y}_{intersect}}$$

where

$$\hat{Y}_{intersect} = \sum_{k \in S_{intersect}} w_k \, w_{(S_0)k} \, y_k$$

where $S_{intersect}$ is the sample that includes the common units in the random and non-random samples.

Pierre Lavalle‘ suggests that the two basic samples are non-random but with a different mechanism of under-coverage of the two respondent groups.