Remote mode of data collection during the pandemic

Santanu Pramanik

NCAER National Data Innovation Centre

Faculty Development Program on Research Methodology & Publication Ethics

Organized by Islamic University of Science and Technology, Awantipora

10th Feb, 2022

Importance of data and survey

- Data-based research is getting increasingly popular
- Data are being used for measuring or informing policy responses
- Good quality data collected in a scientific manner are useful for many reasons:
  - It helps to draw inference about the population in an unbiased and objective way with respect to certain indicators of interest
  - It helps to monitor different programs and schemes
  - It helps to evaluate the effectiveness of different programs and schemes
  - It helps to understand the (causal) association between different phenomenon/events
  - It can explain the reasons behind the occurrence of an event
- The above mentioned functionalities of data influence policy decisions

Sampling vs Complete enumeration

- One of the most common ways to satisfy the demand for data is to conduct a survey
  - Surveying the fraction of the population is known as sampling
  - When an entire population is enumerated, it's known as complete enumeration or census (e.g., population census of India- 15th population census 2011, 7th economic census 2019-20)
- While using surveys to collect data, sampling is often preferred over the complete enumeration for various reasons

Advantages of sampling over census

- Less time consuming- leading to more timely reporting
- Less costly
- Sampling method of data collection leads to sampling error in survey estimates
- Sampling error emerges from the fact that a fraction of the population is surveyed to draw inference about the population and not the entire population
  - estimates from sample survey still can be more accurate as non-sampling error component can be smaller in sampling compared to complete enumeration
- Non-sampling error includes coverage error, nonresponse error, measurement errors, among others
- Countries like US and Canada are gradually switching from census to surveys because of the reasons mentioned above
  - Few basic questions (short-form) are collected in census and for all other information the census (long-form) is replaced by American Community Survey in US and National Household Survey in Canada

Key terminologies relevant for surveys

- Target population
  - is the entire set of units for which the survey data are to be used to make inferences
  - The geographic location and demographic and temporal characteristics of the target population need to be specified
  - Inclusion and exclusion criteria for the types of units being part of the target population should be clearly defined
- Sampling frame
  - Sampling frame is the list of units from where the sample is selected
  - Usually the latest census data is used as sampling frame (e.g., list of districts, list of villages and urban wards)
  - Sometimes it's not readily available, particularly for the ultimate sampling units (e.g., households, individuals)
    - We need to construct the sampling frame through houselisting

Different sources of errors in surveys

- Sampling error
  - a fraction of the population is surveyed to draw inference about the population and not the entire population
  - standard methods available to measure
    - Standard Error (SE)
    - Coefficient of variation (CV)
    - 95% confidence interval (CI)
- Non-sampling errors
  - Coverage error- imperfect sampling frame
  - Nonresponse error- respondent burden and fatigue, attrition
  - Measurement error- iwer, respondent, data entry and processing
  - Often difficult to quantify and usually ignored

Nonobservation error

Observation error

Design of a survey

- The design of a survey involves many interrelated components
- Mode of data collection
  - how to collect data
  - minimize bias and variance in data
  - Mixed mode surveys are very common in many countries
- Questionnaire design
  - tool to collect data
  - minimize bias and variance in data
- Sampling design
  - how to select sample
  - representativeness of sample
- Sample size
  - reliability of survey estimates

Mode of data collection

- Face-to-face interviewing
  - Requires an interviewer collecting data and involves direct interaction between interviewer and respondent
  - In low- and middle-income countries (LMICs) like India, historically, surveys have been carried out by conducting face-to-face interviews
- Two alternative versions for such surveys:
  - PAPI (Pen and paper interviewing)
  - CAPI (Computer assisted personal interviewing)

Advantages of CAPI over PAPI

- In CAPI, data are uploaded to the server on a regular basis (daily or twice a week)
- Uploaded data can be used for monitoring data quality and sharing feedback with the field team to avoid errors going forward
- Data collection and data processing phases tend to be merged
  - processing errors may be reduced
- Interviewers do not need to understand the complicated skip pattern and logical routing
  - it’s pre-programmed using CAPI software (Blaise, CSPro, SurveyCTO, Google form, Survey Monkey, LimeSurvey etc.)
  - iwers can focus only on the content and interviewing skills

Remote mode surveys

- Face-to-face surveys as a method of data collection were commonplace prior to the COVID-19 crisis, particularly in developing countries
- Fear of contracting infection and non-pharmaceutical interventions such as physical distancing and mobility restrictions to contain the spread of infection made it infeasible
- Overwhelming demand for data to respond to economic and health emergencies
  - this forced the remote modes of data collection such as mobile and web surveys to come to the forefront

Telephone survey

- Household surveys conducted using face-to-face interviews tend to be costly, resource-intensive, and time-consuming
  - Sometimes it becomes imperative to conduct more frequent surveys, which in turn, limits the use of F2F surveys
  - Sometimes it's not possible to conduct F2F surveys (e.g., during natural disaster, pandemic, lockdown)
- The widespread use of mobile phones in India offers a new opportunity to remotely conduct surveys
  - Telephone surveys minimize direct interaction between interviewer and respondent
- Telephone surveys can also be computer-assisted
  - CATI: computer-assisted telephone interviewing

Limitations of telephone surveys

- Major challenge: to obtain a reliable sampling frame of telephone numbers
  - Often phone numbers are collected through earlier F2F surveys and then used for telephone follow-ups
- Response rates tend to be lower compared to F2F surveys

Delhi NCR Coronavirus Telephone Survey (DCVTS)

- The fourth and the last quarterly follow up and the endline survey of Delhi Metropolitan Area Study (DMAS) started in March 2020
  - data collection was suspended mid-March following cases of Coronavirus
- We used the available resources for conducting telephone surveys to understand
  - people's knowledge, perception and behavior with respect to COVID-19
  - the impact of the pandemic on people's life and livelihood and their coping mechanisms
- 4 rounds of DCVTS at different stages of epidemic and the lockdown
  - Data analysis and report preparation were done urgently and disseminated widely through sharing with key stakeholders, webinar, press release

Coverage and nonresponse bias in telephone surveys

- In the pre-COVID era, telephone survey was not a common method of conducting surveys in India
  - The effect of phone mode of data collection on household survey estimates is not fully known
- Multiple sources of nonobservation error
  - Lack of access to mobile phones
  - Non-contacts because of non-working phone numbers
  - Refusal or drop-offs in phone surveys
- It is crucial to compare the distributions of included (respondents) and excluded units (nonrespondents) with respect to key socio, economic and demographic characteristics which are often associated with most outcome variables

Phone surveys in the absence of phone numbers

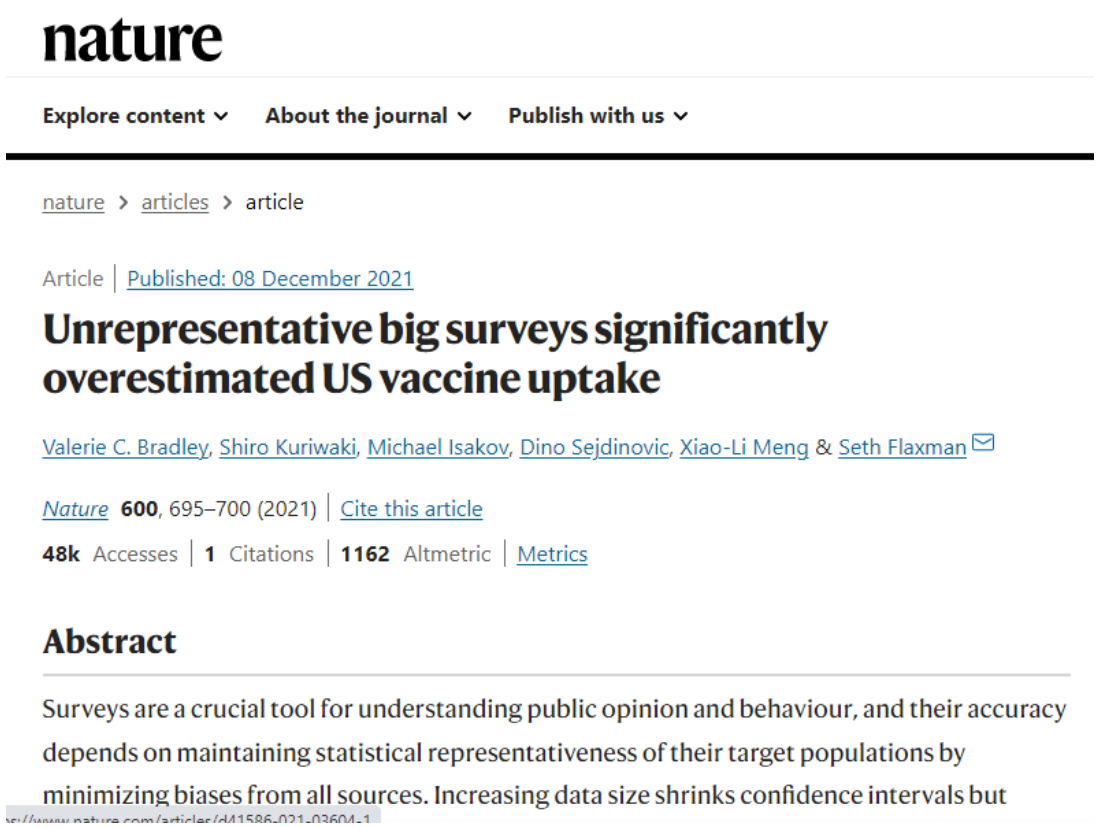
- Rice institute: Mobile phone survey for measuring social discrimination
- Social Attitudes Research, India (SARI) built representative samples of adults ages 18 to 65 by using probability weighted random digit dialing (RDD) and within-household respondent selection
- DOT assigns mobile phone companies 5-digit "series" to use at the beginning of the 10-digit mobile phone numbers that they sell in a particular mobile circle
- The SARI team generated a sampling frame of potentially active numbers in each mobile circle
  - Equally divide the number of series across the number of subscribers a company reports
  - Number of subscribers per company, per state are obtained from Telecom Regulatory Authority of India (TRAI) reports
  - Then add a randomly generated five-digit number to each series to form a 10 digit mobile number
- SARI surveyors called these numbers in a random order

Accuracy of survey estimates

- Poorly designed survey questionnaires
- Improperly designed sampling plan
- Inaccurate (or absence of) sampling frame
- Insufficient sample size
- Nonresponse rate and composition
- Lack of experience and commitment of field agencies collecting data
- Limited involvement of research team during the training of interviewers and monitoring of data collection activities
- Iwer's lack of understanding and/or faulty interviewing techniques
- Comprehension error on respondent's part or hiding of the truth
- Inaccurate recording of data

Does large sample protect against bias?

- The primary concerns with remote mode surveys are undercoverage of target population and self-selection of respondents resulting in biased estimates
- Large sample sizes may not protect against bias, rather can even make the estimates more biased



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 08 December 2021

Unrepresentative big surveys significantly overestimated US vaccine uptake

Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng & Seth Flaxman

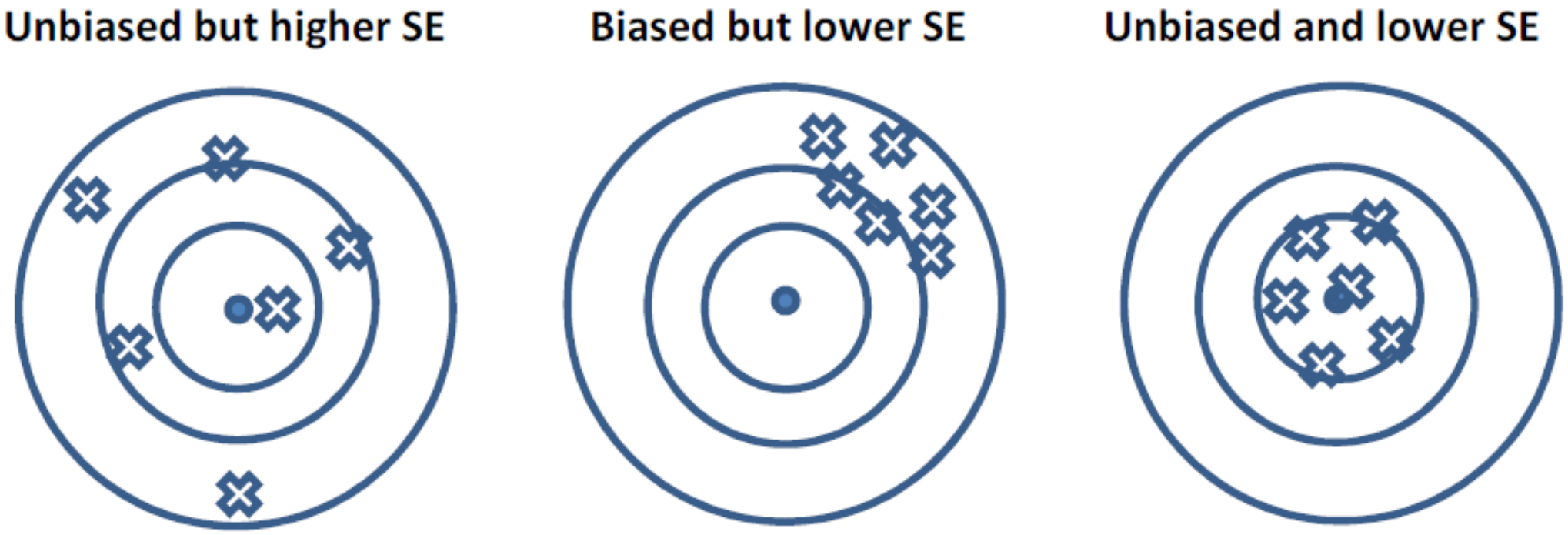
Nature **600**, 695–700 (2021) | Cite this article

48k Accesses | **1** Citations | **1162** Altmetric | Metrics

Abstract

Surveys are a crucial tool for understanding public opinion and behaviour, and their accuracy depends on maintaining statistical representativeness of their target populations by minimizing biases from all sources. Increasing data size shrinks confidence intervals but

Accuracy of survey estimates: bull's eye



Decomposing the total error of a survey estimate

- Total error = Data quality defect × Data scarcity × Inherent problem difficulty
- Data quality: correlation between the event that an individual's response is recorded and its value
- Data scarcity: function of the sample size n and the population size N
- Inherent problem difficulty: heterogeneity of the outcome variable in the population
- Bradley et al. 2021

Formerly, COVID-19 Symptom Survey (CSS)

COVID-19 Trends and Impact Survey (CTIS)

Methodology

COVID-19 Trends and Impact Survey (CTIS)

- CTIS is a world-wide web survey in the times of COVID-19
- University of Maryland and Carnegie Mellon University in partnership with Facebook have been conducting this survey daily since April 2020
  - > 200 countries and > 50 languages
- FB users are invited to take part in the surveys
  - self-report COVID-19-related symptoms
  - experience with COVID tests
  - contacts with others
  - mental health and economic security
  - disruptions in routine health services
  - Vaccination and vaccine hesitancy
- Designed to help monitor and forecast how COVID-19 may be spreading, without trading off the privacy of the people who took the surveys
- FB does not share background information of the respondents (FB users) with their academic partners, and they in turn do not share survey responses with FB

CTIS Methodology

- Sampling frame: Facebook Active User Base (FAUB) aged 18+ in India
- Study design: Repeated cross section
  - Daily new random sample of users
  - There could be repetition of survey request to the same users (particularly in small states), but should be treated as a new sample
- Sampling design: To select daily sample, stratified random sampling is considered (strata being the states) in order to provide representative sample at the national level
- Survey weights: Survey weights are constructed for each respondent so that respondents of CTIS better represent the target population

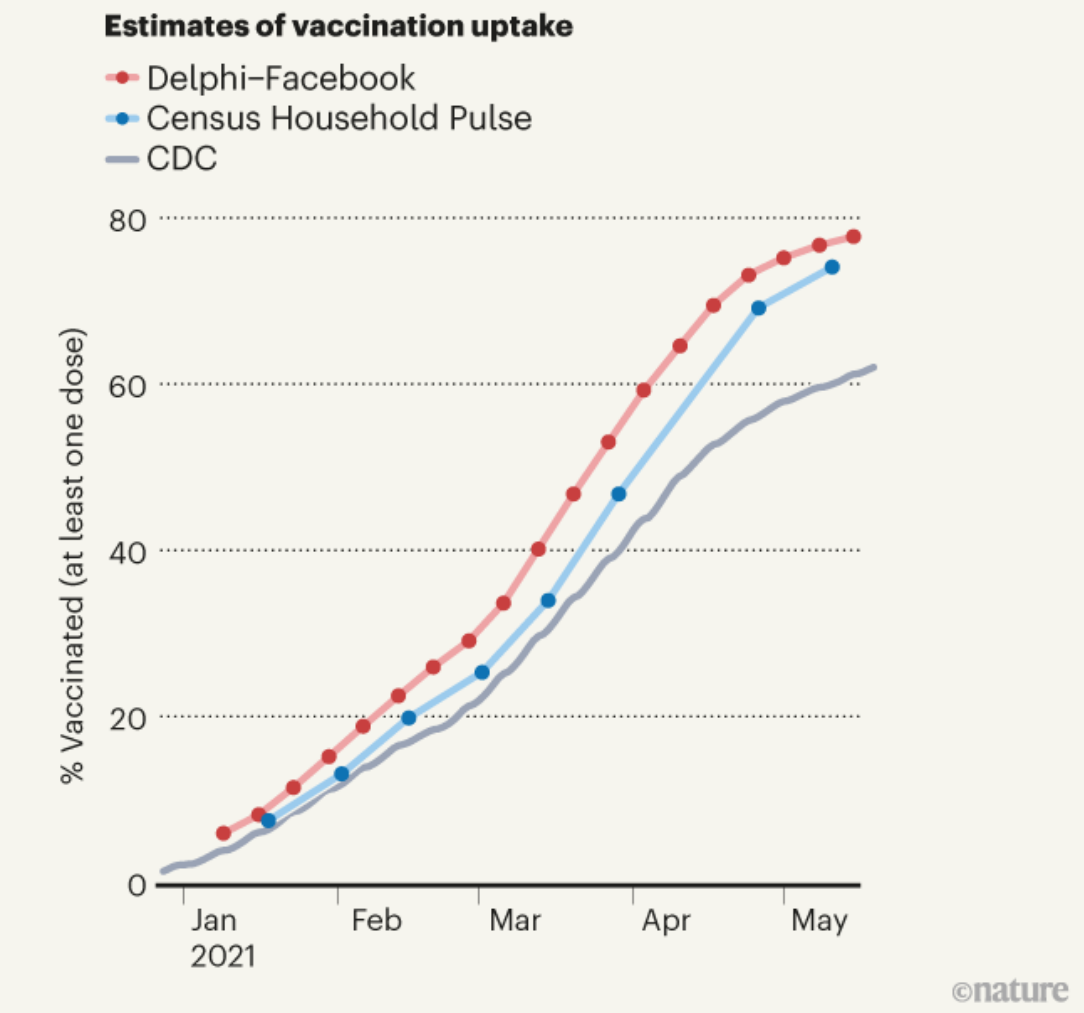
Weighting: Step 1

- Two types of adjustment are done to construct final weights so that the sample on each day represents the adult population at the state level and India level
  - minimize nonresponse error and coverage error in survey estimates
- Step 1: From CTIS respondents to the FAUB
- Not everyone selected for CTIS responded- nonresponse
- Nonresponse error is minimized by using Inverse Propensity Score Weighting (IPSW) method to make the sample more representative of FB users
- Covariates in the IPSW model: FB team included user characteristics (??) collected from FB users

Weighting: Step 2

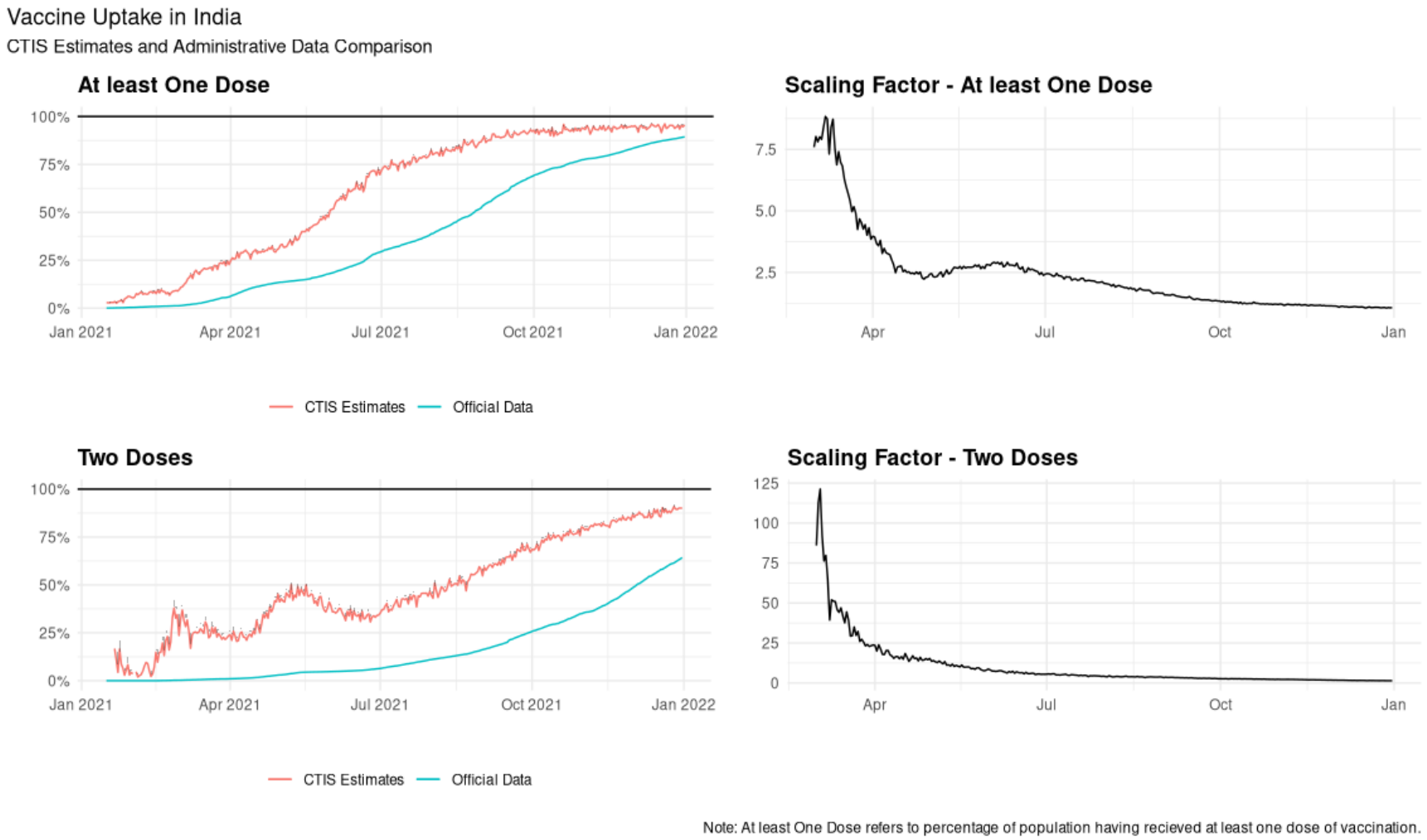
- Step 2: From FB users to general adult population
- Not all adults are FB users- undercoverage
- Coverage error is minimized by adopting poststratification adjustment of weights with Step 1 weights as inputs
- Poststrata are defined over age (4 categories: 18-24, 25-44, 45-64, 65+) and gender (2 categories) within each state
  - Urban-rural area of residence was not considered as a poststratum
- External benchmarks are obtained from the UN 2019 world population projections
- Conclusion: The final weights can be interpreted as the (estimated) number of adults in the general adult population represented by a CTIS respondent for that day

Overestimation of vaccine uptake: US CTIS



- Bradley et al. (2021) compared estimates of the uptake of COVID-19 vaccines among US adults
- Vaccination estimates from US CTIS (Delphi-Facebook) are certainly inflated compared to the official numbers (US CDC estimates)
- They claim that the errors in CTIS estimates of vaccine uptake increase over time

Overestimation of vaccine uptake: India CTIS



Bias in vaccine uptake estimates

- Bias in CTIS estimates of vaccine uptake is not constant over time
  - Increases up to a certain point of time and then decreases
  - CTIS estimates of vaccination coverage and the true population quantities converge with time
  - This contradicts the results in Bradley et al. paper which found that the errors have increased over time
- Our findings are explained by the fact that the level of variation in the outcome of interest in the population (problem difficulty in Bradley et al.'s terminology) first increases with time and then goes downward after more than 50% of the population are vaccinated

Discussion

- CTIS sample is likely to be biased towards more educated, internet savvy, urban respondents even after weighting adjustment to account for survey non-response and individuals who are not on the Facebook platform
- The objective of constructing weights is to provide a weight per respondent so that respondents of CTIS better represent the target population
- The two-step weighting procedure partially corrects the bias in gender and age groups (to a lesser extent) represented in the CTIS samples
- However, urbanicity a key demographic variable, is not used in the second stage (poststratification adjustment) of weighting procedure
- Most of the respondents in FAUB are from cities and as a result of not accounting for this, the estimates are likely to be skewed towards urban population

Reference books on survey sampling

1. Introduction to Survey Sampling
   - Graham Kalton
   - SAGE publication
2. Model Assisted Survey Sampling
   - Carl-Erik Särndal, Bengt Swensson, Jan Wretman
   - Springer
3. Survey Methodology
   - Robert M. Groves, Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, Roger Tourangeau
   - Wiley