CNSTAT Workshop on Future Directions for Social and Behavioral Science Methodologies in the Next Decade

September 25–26, 2024
Washington, DC

Future Directions for Post-Survey Adjustments and Statistical Disclosure Limitation in Official Statistics

Natalie Shlomo
Natalie.shlomo@manchester.ac.uk

MANCHESTER 1824
The University of Manchester

Topics Covered

1. Current state-of-the-art:
   - Survey data collection
   - Post survey adjustments
   - Statistical disclosure control (SDC)
2. Challenges
3. Research areas with recommendations:
   - Future of surveys
   - Use of generative AI in surveys
   - Multi-source statistics
   - Dissemination and confidentiality guarantees

State-of-the-Art

Survey Data Collection:

- Traditional approaches for designing and collecting survey data, eg., large government surveys (The Labor Force Survey, The American Community Survey, The Health Survey, Crime Victimization Survey)
- Data collection typically through mixture of modes: internet, phone and face-to-face, and using electronic interfaces

- Adaptive and responsive survey designs aim to target data collection to ensure a more representative sample
  - Balancing on auxiliary variables available at the frame level to reduce nonresponse bias at the source and ensure less variable survey weights
  - Loss function for optimization focus on contrast between responding and not responding units, eg.

$$R = 1 - 2S(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{n} d_i(\rho_i - \bar{\rho})^2} \quad \text{and } \bar{\rho} = \frac{1}{N}\sum_{i=1}^{n} d_i\,\rho_i$$

State-of-the-Art

Post-survey data collection:

Data cleaning: all units in-scope, other basic checks

Editing and imputation: Identify erroneous records using edit rules (use optimization algorithms to identify erroneous variables for imputation, eg. Felligi and Holt 1976)
In the digital era, most edits handled at the data collection stage

Coding: In-house programs for automatic coding: occupation, industry, ethnicity, with coders needed and quality assurance procedures

State-of-the-Art

Post-survey data collection:

Item missing data: Imputation based on parametric, nonparametric or mixture approaches (eg. PMM) including machine learning methods

- Imputation carried out by agency (using sensitive variables not released to the public), but leads to single imputation
- Assumption that item nonresponse is small and does not impact on variance estimation
- Replication weights can be released with the data and can account for uncertainty due to imputation

Imputed values flagged so analysts can use Multiple Imputation or other model-based imputation approaches

State-of-the-Art

Survey adjustments:

Weighting typically 3 stages: (1) design weights (2) unit nonresponse adjustment (inverse response rates/propensities using parametric or machine learning (classification trees)) (3) post-stratification/ calibration
Weight trimming to ensure CV below tolerable threshold
Unit non-response studies using auxiliary data to improve data collections

Variance estimation using linearization or replication methods

Small Area Estimation – model- based estimation combining direct and indirect estimates using auxiliary information to obtain estimates at unplanned lower geographical levels and their MSEs

- Area level or individual level models, temporal and/or spatial
- Mean square error calculations
- Benchmark small area estimates to design-based regional totals

State-of-the-Art

Statistical disclosure limitation:

Safe Data:

- Releasing public-use files from social survey data (small sampling fractions): recoding, deleting sensitive variables, sub-sampling, and possible 'light' perturbative methods such as top or bottom coding, adding noise, swapping
- Quantifying disclosure risk (typically using probabilistic modelling or record linkage) and data utility measures. Estimates of disclosure risk:

$$\tau_1 = \sum_k I(f_k = 1, F_k = 1)$$

$$\tau_2 = \sum_k I(f_k = 1)\frac{1}{F_k}$$

F_k population count in cell k

f_k sample count in cell k

- R-U confidentiality map to identify SDC method and parameters
- Complex survey data, business surveys, longitudinal data may have public-use data via synthetic data generation (some degree of success using statistical or machine learning approaches)

State-of-the-Art

Statistical disclosure limitation:

Safe Data: (cont.):

- Use of Differential Privacy in US census tables

Mechanism A satisfies ε-*differential privacy* if for *any* D, D' that differ by one row, for *any* output O: $P(A(D) = O) \leq e^{\varepsilon} P(A(D') = O)$

- Injected into model-based synthetic data generation 'On the Map' (see 2008 paper: https://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf)
- Differential privacy is an output perturbation method and works well for tabular data, but more challenging for input (microdata) perturbation

Safe Access:

- Data Enclaves and Secure Servers for trusted users to access sensitive data

Challenges

The world of surveys and survey dissemination is changing fast:

New approaches for engaging with respondents and data users, while survey response rates in decline and survey costs going up

Modern data ecosystem provides new data sources, including non-probability surveys, that provide granular data often very timely in comparison with surveys and censuses

Modernization calls for interoperability of data sources, including surveys, and integration of surveys among themselves and with censuses, geospatial data and administrative archives

Challenges

The world of surveys and survey dissemination is changing fast:

New methods and tools, such as machine learning and generative Artificial Intelligence, are active areas of research and have the potential to create new contexts in the development of survey statistics

Exploiting new data collection modes in the way we interact with smartphone devices, sensors and passive data collection, as well as opportunities for web scraping

Confidentiality and privacy in a landscape of increasing disclosure risks and privacy concerns

Research Areas

Future of surveys:

A random (household) survey is essential and here-to-stay:

- Use to assess coverage/representativeness in administrative and other found data sources (including nonprobability and online surveys)
- Allows for collecting (attitudinal) data on questions not found in alternative data sources
- Consider rotational panel designs to include a longitudinal component
- Employ census-type data collection strategies, using adaptive designs and targeted data collection, under mixed modes, including face-to-face follow-up
- Embed such surveys in legislation to make them mandatory, eg. The American Community Survey

Research Areas

Future of surveys:

National Statistical Institutes (NSIs) to consider push-to-web approaches (probability-based but including elements of quota sampling) to host a panel member survey (similar to other statistical organizations, such as YouGov, IPSOS, etc.)

- Requires a national population frame with addresses and characteristics
- All responding panel members to complete core questionnaire and then sub-sampling for other questionnaires
- Weighting would include inverse propensity weighting (based on reference sample) and using the core sample for calibration
- Make more use of mobile devices and passive data collections (apps) to collect information on items such as travel, expenditures, etc.

Research Areas

Future of surveys:

Non-probability online web-surveys useful for hard-to-capture populations (eg. ethnic minority groups) to ensure equitable statistics

- Respondent driven sampling allows for producing snowball samples to meet quotas with possibility of estimating inclusion probabilities
- Other methods of adjusting for selection bias in non-probability samples using a reference sample
- Recruitment and registration links, offering incentives, avoiding bots and fraud, see for example, EVENS (Survey on the Impact of COVID19 on Ethnic Minorities in the United Kingdom https://www.evensurvey.co.uk/)

Research Areas

Future of surveys:

Need more acceptance of model-based estimation including building statistical registers and mass imputation to (partially) replace censuses (versus traditional survey weighting)

- Dealing with large nonresponse and selection bias in probability-based surveys requires inverse propensity weighting
- More use of non-survey data and data integration to improve data collection strategies and compensate for nonresponse, selection bias and estimation
- Include small area estimation in national statistics

Recommendation 1: Consortium of NSIs, survey organizations and academics to research the landscape of Survey Futures and transforming data collections (see: https://surveyfutures.net/ for an example of funding from the United Kingdom Economic and Social Research Council (ESRC))

Research Areas

Use of Generative AI in Surveys

- Questionnaire development, testing and evaluation
- Dynamic interviewing eg., replacing interviewers
- Automatic Coding: Make more use of Large Language Models, neural networks and machine learning to implement automatic coding, eg. fastText by Meta provides a probability of a correct coding allowing for quality assurance processes
- Web scraping, eg. job adverts to replace Job Vacancy Surveys (but need to research compensating for selection bias in non-probability sources)
- Producing imputations and synthetic data (Generalized Additive Models (GAMS) showing promise)

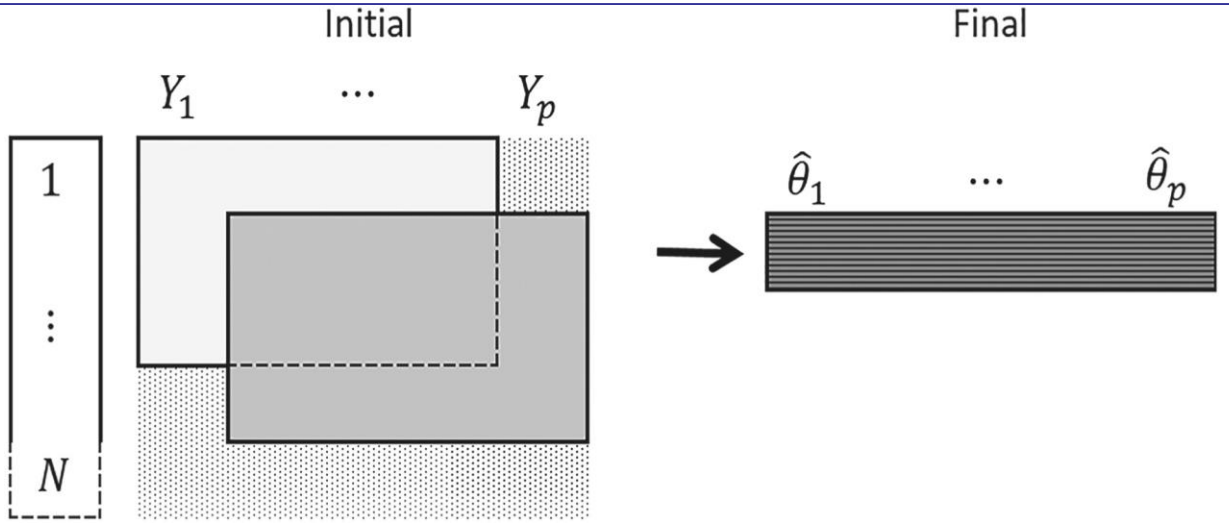
Recommendation 2: Introduce calls for grants in Advancing and Leveraging Generative AI in Survey Research

Research Areas

Multisource statistics

- Combining survey data with administrative data/big data aims to provide more detailed and timely statistics and reduce response burden and costs
- Different situations arise depending on coverage, overlapping variables and units, microdata vs aggregated data

In this setting below we use extended capture–recapture methods that account for over-coverage to estimate N :



De Waal,
van Delden,
Scholtus (2019)

Research Areas

Multisource statistics

- Research how to carry out quality assessments and frameworks accounting for all sources of error
- Develop quantitative measures for accuracy and coherence of an output, particularly for non-sampling errors
- Measure bias and confidence intervals depending on configuration of the data

Recommendation 3: Consortium of NSIs, academics to research and develop multi-source statistics, example European Statistical System (ESSNet) on Quality of Multi-source Statistics and other similar projects (see: https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2020_January_N81_03.pdf for a summary)

Research Areas

Confidentiality and Privacy

Public-use files are becoming more high-risk

- Focus on open web-based dissemination methods sing table builders and remote analysis servers
  - Approaches such as Differential Privacy, with formal privacy guarantees, should be included in SDC toolkit, particularly for output dissemination
  - Differential Privacy offers confidentiality guarantees for attribute disclosures for a given privacy budget
  - Privacy budgets can be adapted for survey data so combine SDC methods that focus on avoiding re-identification (sub-sampling, coarsening) with Differential Privacy, to allow for lowering privacy budgets
- Synthetic data production for complex data structures is a challenge: consider statistical approaches embedded with Differential Privacy

Research Areas

Confidentiality and Privacy

- Develop facilities for trusted users and researchers to access protected data, through Data Enclaves and Remote Access Servers
- Continue to develop methods for generating synthetic data to allow for investigating microdata prior to gaining access

    - Example: UK Data Archive and the ONS Secure Research Service (SRS), both give accredited researchers secure access to de-identified, unpublished data in order to work on research projects for the public good https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice

Recommendation 4: Consortium of NSIs and academics to research: access to data and remote servers, web-based dissemination methods and embedding Differential Privacy in the SDC tool-kit

Summary

Key gaps and recommendations: survey futures, use of generative AI, multi-source statistics and confidentiality for the dissemination of statistical data

Social and economic sciences rely on survey and non-survey data to inform research and evidence-based policies

Consider alternative ways to collect and find data (administrative data, big data, non-probability samples) to integrate with survey data

Random surveys are a staple of social science research, but need to make sure they are fit-for-purpose and of high-quality to assess coverage and representation in alternative data sources

Summary

Open access dissemination of statistical data requires more rigorous confidentiality protection, eg. Differential Privacy, combined with SDC methods, and using web-based platforms, synthetic data

Ensure trusted users and researchers have access to sensitive data whilst guaranteeing privacy concerns

National funding for collaborative research brings together academics, statisticians, data scientists, social scientists and methodologists to participate in joint research and ensuring the future of national statistical systems

Thank you for your attention