Workshop on Survey Methodology:

Big data in official statistics

Block 1: Introduction

20 May 2019,

Brazilian Network Information Center (NIC.br),

São Paulo, Brazil

Jan van den Brakel

Statistics Netherlands and Maastricht University

Introduction

Official statistics:

1. Purpose: provide reliable statistical information about finite target populations

   - Target population U containing N elements $i = 1, \ldots, N$.
   - Variable of interest: y_i
   - Interest in:
     - population totals $Y = \sum_{i=1}^{N} y_i$,
     - population means $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$
   - National level but also for breakdowns w.r.t. regional or socio-demographic classifications
   - $\Rightarrow$ Information for domains: Y_d and $\bar{Y}_d$

2. Common approach for NSI's to collect this information: survey sampling

   - Draw a sample s of size n from target population U with $n << N$
   - Examples: simple random sampling,
     stratified simple random sampling,
     cluster sampling, two-stage sampling, sampling with unequal inclusion probabilities, etc.
   - Collect data among the sampling units: observe values $y_i, i = 1, \ldots, n$
   - Estimates for the unknown population parameter
   - Mode of inference traditionally design-based:
     - Horvitz-Thompson estimator:
       - $\hat{Y} = \sum_{i=1}^{n} d_i y_i$
       - design weights: $d_i = \frac{1}{\pi_i}$
       - π_i: inclusion probability sampling unit i

- General regression estimator:
  - Improves HT estimator with auxiliary information, say $\mathbf{x}_i$, for which the population totals, say $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$ are known
  - Calibrate the design weights (d_i) such that $\hat{\mathbf{X}} = \sum_{i=1}^{n} w_i \mathbf{x}_i = \mathbf{X}$
  - GREG estimator: $\hat{Y}_r = \sum_{i=1}^{n} w_i y_i$
  - Motivation: $y_i = \beta^t x_i + e_i$
- Details: Särndal et al. (1992)

3. Design-based or model-assisted inference (expectation and variance with respect to the sample design)

   - Advantages:

     – Approximately design-unbiased estimator based on relative small samples.

       Data generating process is known and controlled through the sample design and its estimator (sampling strategy).

     – Uncertainty quantified via variance calculation

     – Robust for model miss specification

     – Auxiliary information reduces design variance and corrects for selective non-response

   - Disadvantages:

     – Large design variances in case of small sample sizes

     – Data collection expensive

     – Surveys are not very timely

- Non response
- Response burden
- ...

4. National statistical institutes: increasing interest to use alternative data sources like registers and “big data”

Big data:

1. Large data sets that are generated as a by-product of processes not directly related to statistical production purposes.

2. Examples of these include:

   (a) time and location of network activity available from mobile phone companies,

   (b) social media messages from Twitter and Facebook

   (c) internet search behavior from Google Trends

   (d) information found on the internet

   (e) scanner data

   (f) sensor data, e.g. satellite images, aerial images and road sensor data

   (g) administrative data like tax registers

Use of Big data in official statistics:

1. Primary data source
2. Covariates in small area estimation models or models for nowcasting
   (a) Area level model (Fay and Herriot, 1979):
   - Uses cross-sectional correlations
   - Avoids matching unstructured big data sources with survey data on the unit level
   - Marchetti et al. (2015) uses mobility of cars tracked with GPS as a covariate for predicting poverty in a Fay-Herriot model

   (b) Official statistics:
   - Repeated surveys
   - Therefore time series models are more appropriate
   - For this course we focus on structural time series models

Outline course:

- Block 2: Small area estimation
- Block 3: Introduction structural time series models
- Block 4: Bivariate state space model for nowcasting
- Block 5: Dynamic factor models
- Block 6: Big data as primary data source
- Block 7: Remote sensing data

References

Fay, R. and Herriot, R. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association* 74 (366), 269–277.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Perdreschi, D., Rinzivillo, Pappalardo, L., and Gabrielli, L. (2015). Small area model-based estimators using Big data sources. *Journal of Official Statistics* 31, 263–281.

Särndal, C., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.