Workshop on Survey Methodology:

Big data in official statistics

Block 2: Cross-sectional small area estimation models

20 May 2019,

Brazilian Network Information Center (NIC.br),

São Paulo, Brazil

Jan van den Brakel

Statistics Netherlands and Maastricht University

Introduction

Official statistics:

- Data source: traditionally probability samples in combination with registers
- Inference: traditionally design based or model assisted
- Main reason: free from model assumptions
- Drawback: large design variances in case of small sample sizes
- Relevance of data increases with the level of detail, its timeliness and frequency
- Interest in reliable domain estimates
- E.g. domain totals $Y_j = \sum_{i=1}^{N_j} y_i$
- Direct estimator $\hat{Y}_j = \sum_{i=1}^{n} w_i y_i \delta_{ij}$

  with δ_{ij} an indicator equal to one if i is an element of domain j and zero other wise

- Domains and areas are graphical or socio-demographic breakdowns of the population

Small Area Estimation

- Refers to model-based inference procedures that use a statistical model to improve the effective sample size in a particular domain with sample information of neighboring domains
- Overview: Rao and Molina (2015)
- NSI's:
  - Reserved to apply model based methods in the production of official statistics
  - It is however a solution for
    - small domain problems
    - use of non-probability data sources instead of survey data only
  - Increasing interest among NSI's, e.g. Statistics Netherlands

- Mainstream approaches in SAE:
  - Area level model or Fay-Herriot model (Fay and Herriot, 1979)
    multilevel model for the direct estimates at the domain level
  - Unit level model or Battese-Harter-Fuller model (Battese et al., 1988)
    multilevel model for the sampling units
- Success of both models depends on the available covariates
- Traditionally:
  - Registers
  - Census

- New non-probability data sources
  - Potential covariates in SAE models
  - Particularly for countries without registers and censuses
  - Area level model most appropriate since it avoids problems with matching fuzzy big data sources at the micro level
- Area level model:
  - Measurement error model: $\hat{Y}_j = \theta_j + e_j$
  - Linear model for population parameter: $\theta_j = \beta^t \mathbf{x}_j + u_j$
    - u_j: random domain effect
    - e_j: sampling error
  - Multi level model for direct estimator: $\hat{Y}_j = \beta^t \mathbf{x}_j + u_j + e_j$
  - Used to construct a prediction for θ_j (Rao and Molina, 2015)

Relevant literature

Literature on the use of big data sources for estimating poverty and wealth

- Marchetti et al. (2015) uses mobility of cars tracked with GPS as a covariate for predicting poverty in a Fay-Herriot model
- Noor et al. (2008) uses remotely sensed night-time light (via satellite images) as a proxy for poverty.
  - Analyse correlation between house hold survey data on income with night-time light intensity
  - Propose night-light intensity as a measure for poverty.
- Engstrom et al. (2017) uses day time satellite images to predict well-being.
  - Applied deep learning to extract features related to well-being (number of cars, building type, roof type, etc).
  - Applied machine learning methods to combine

survey data with satellite image features

- Used this to predict well being in other areas

- Blumenstock et al. (2015) used mobile phone data to predict poverty
  - Applied machine learning methods to combine survey data with mobile phone data
  - Used this to predict well being and poverty in other areas

- Steele et al. (2017) used mobile phone data and satellite images to predict poverty
  - Combine survey data with mobile phone data and satellite data in a generalized linear model to predict poverty for small spatial areas
  - Comes close to SAE methodology

- Schmid et al. (2017) uses mobile phone data for estimating literacy
  - Combine survey data with mobile phone data as covariates in an Area level model or Fay Herriot model

References

Battese, G., Harter, R., and Fuller, W. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association* 83, 28–36.

Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science* (350).

Engstrom, R., Hersh, J., and Newhouse, D. (2017). *Poverty from space: Usign high resolution satellite imagery for estimating economic well-being*. Technical report.

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.

Noor, A., Alegana, V., Gething, P., Tatem, A., and Snow, R. (2008). Using remotely sensed night-time light as a proxy for poverty in Africa. *Population and Health Metrics* (6:5).

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation, 2nd Edition*. John Wiley, New York.

Schmid, T., Bruckschen, F., Salvati, N., and Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society, Series A* 178, 239–257.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J., and Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface* 14 (127).