Workshop on Survey Methodology:

Big data in official statistics

Block 6: Big data as primary data source

20 May 2019,

Brazilian Network Information Center (NIC.br),

São Paulo, Brazil

Jan van den Brakel

Statistics Netherlands and Maastricht University

Introduction

Sampling theory:

- Data generating process known and controlled by the design of the probability sample
- Probability sampling offers a clear frame work to construct optimal sampling strategy, i.e. (design + estimator)
- Correct for under and over sampling via inclusion probabilities and calibration
- Measuring the uncertainty via variance estimation
- Growing interest in using alternative data sources that are generated as a by-product of processes not directly related to statistical production purposes. Referred to as non-probability data or big data

Non-probability data or big data:

- Unknown to which extend results can be generalized to an intended target population
- Data generating process is unknown
- Data driven research
- Many examples at CBDS:
  - Social media studies; Sentiment index
  - Propensity to move from registers
  - Web scraping / text mining from websites (innovative companies and sustainable companies)
  - Scanner data for price indices
  - Hay fever indicator based on scanner data
  - Mobile phone data for day time populations
  - Measuring increase of urbanization with satellite data
  - Measuring solar power panels with aerial images

  - Estimating solar power production indirect
- Problem: no clear frame work, apparently each application requires a different approach

Outline:

- Review of literature to correct for selection bias in nonprobability samples
- Some examples in more detail:
  - Estimating unmetered photovoltaic power
  - Two examples of satellite images and aerial images
  - Measuring road intensity with road sensors
  - Day time population with mobile phone data

Inference for nonprobability samples

Most methods often (but not always) use a reference sample that is based on a probability sample

- Machine learning algorithms to analyze the relation between images and survey or census data
  - Remotely sensed night-time light (via satellite images) as a proxy for poverty (Noor et al., 2008)
  - Day time satellite images to predict well-being (Engstrom et al., 2017)
  - Mobile phone data to predict poverty (Blumenstock et al., 2015)
- Weighting and calibration
  - Similar to weighting in sample surveys (Särndal et al., 1992)
  - Strong assumption MAR conditional on the covariates

- Quasi-randomization
  - Explicit model for estimating inclusion probabilities for the units in the nonprobability sample
  - Same covariates in the probability and nonprobability sample
  - Elliott and Valliant (2017); Valliant et al. (2013); Valliant and Dever (2011), based on propensity scores (Rosenbaum and Rubin, 1983, 1984)
  - Strong assumption MAR conditional on the covariates
- Superpopulation model approach
  - No reference sample
  - Explicit model for the observations in the sample $y_i = f(x_i)$
  - Predictions for the units not included in the sample $\hat{y}_i$

  - Prediction population total $\hat{t}_y = \sum_{i \in s} y_i + \sum_{i \in (U \setminus s)} \tilde{y}_i$
  - Valliant et al. (2000), based on strong assumption MAR conditional on the covariates

- Inverse sampling
  - Available:
    * Selective big data sample ($\mathcal{B}$) containing target variable y_i and auxiliary variable $\mathbf{x}_i$
    * Representative probability sample ($\mathcal{A}$) containing auxiliary variable $\mathbf{x}_i$
  - $\mathcal{A}$ is used to assess the selectivity of $\mathcal{B}$
  - Calculate importance weights for all units in $\mathcal{B}$
  - Draw a sample from $\mathcal{B}$ with unequal selection probabilities proportional to the important weight
  - $\rightarrow$ Simple random sample
  - Reference: Kim and Wang (2018)

- Data integration
  - Available:
    - Selective big data sample ($\mathcal{B}$) containing target variable y_i and auxiliary variable $\mathbf{x}_i$
    - Representative probability sample ($\mathcal{A}$) containing auxiliary variable $\mathbf{x}_i$
  - Imputation of y_i in $\mathcal{A}$ from $\mathcal{B}$ using $\mathbf{x}_i$ via nearest neighbour (Rivers, 2007)
  - Construct weights for all units in $\mathcal{B}$ based on a parametric model and apply unequal probability weighting (Kim and Wang, 2018)

Estimating unmetered photovoltaic power consumption

- Energy accounting requires coherent statistics on energy related issues
- Statistics on renewable energy for evaluating the agenda on energy transition and on climate policy
- Production of electricity by domestic photovoltaic installations
  - currently unknown
  - incomplete register of PV installations and assumptions about their average capacity
- Purpose of this project: approximate the amount of unmetered photovoltaic electricity indirectly

Estimating unmetered photovoltaic power consumption

Approach

- If PV installations produce a lot of electricity, less electricity will be taken from the high voltage grid
- Available data:
  - Time series data on electricity exchange on the high power grid
  - Meteorological time series data on solar irradiance
- Hidden signal on the amount of solar power produced by domestic PV installations

Data

Data

- Time series on electricity exchange from the high power voltage grid:
  - MWh at a daily frequency
  - January 1st 2004 through December 31th 2017
  - Downloaded from the website of the Dutch Transmission System Operator (Tennet)
- Meteorological time series data
  - Solar irradiance in J/cm^2 at a daily level
  - Temperature (in 0.1^oC) at a daily level
  - Day length
  - January 1st 2004 through December 31th 2017
  - Downloaded from the website of the Royal Netherlands Meteorological Institute for the same period.

Data

Time series

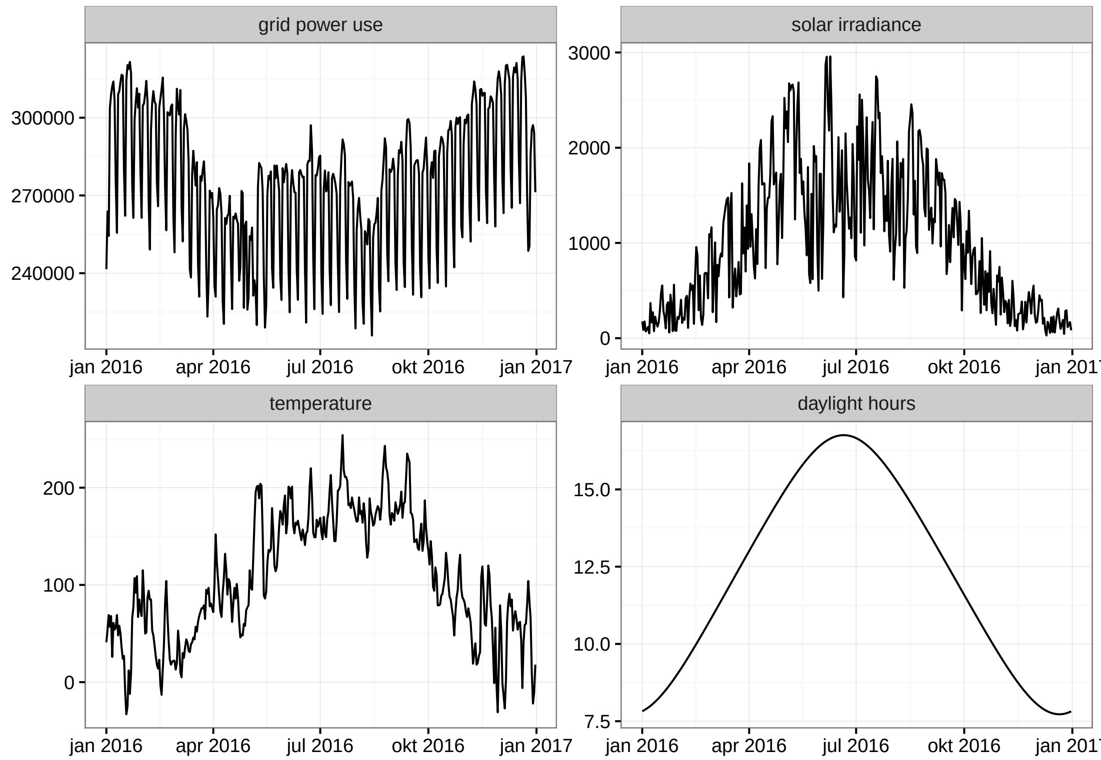

Figure 1: Available time series for 2016 on a daily frequency.

Model

Model

- Production of solar power (P_t):
  - Irradiance (I_t)
  - Temperature (T_t)
  - Day length (L_t)
  - Calendar effects (C_t)
- Problem: Electricity demand (Y_t) also depend on I_t, T_t, D_t, C_t

Model

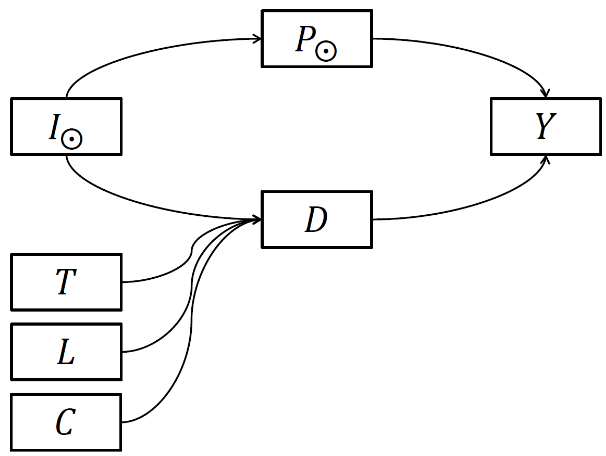

Figure 2: Directed acyclic graph (DAG) for the solar power causal model, with $I_{\odot t}$ solar irradiance, $P_{\odot t}$ solar power, Y grid power, D total demand, T temperature, L length of day and C calender effects.

Two causal paths between I_t and Y_t,

$$I_t \rightarrow P_t \rightarrow Y_t \tag{1}$$

$$I_t \rightarrow D_t \rightarrow Y_t \tag{2}$$

Model

Problem: how to isolate the effect of I_t on P_t:

- Causal modelling (Pearl, 1995)
- Assume independence between P_t and D_t
- Estimate the effect of I_t on demand D_t
  - ARIMAX model for period 2004 - 2010 (Box et al., 2015)
  - $Y_t = f(I_t, T_t, L_t, C_t)$
  - β_I: effect of I_t on demand

Model

Problem: how to isolate the effect of I_t on P_t (cont.):

- Estimate the effect of I_t on demand P_t
  - ARIMAX model for period 20013 - 2017
  - Correct Y_t for the effect of I_t on demand:

    $\tilde{Y}_t = Y_t - \beta_I I_t$
  - $\tilde{Y}_t = f(I_t, T_t, L_t, C_t)$
  - $\tilde{\beta}_{I,y}$: effect of I_t on $\tilde{Y}_t$ (year dependent)
- Estimate the daily production of solar power:

  $\hat{P}_t = \tilde{\beta}_{I,y} I_t$
- Annual estimates: aggregating the daily estimates $\hat{P}_t$

Results

ARIMAX(p,d,q) model:

- Modelselection based on AIC
- d=1
- AR lags p=6
- MA lages $q = 1$
- Selected covariates and their interactions: Buelens and van den Brakel (2018)

Results

Results of the ARIMAX model fit

Year	$\tilde{\beta}_{I,y}$	SE	$\hat{P}_t$ (MWh)	$\hat{D}$ (MWh)	Percentage solar
2013	-0.390	0.787	140,877	101,554,484	0.14%
2014	-1.296	0.797	485,381	99,549,220	0.49%
2015	-2.004	0.755	774,212	100,436,422	0.77%
2016	-3.409	0.828	1,275,643	102,065,655	1.25%
2017	-5.086	0.807	1,867,628	103,223,204	1.81%

- $\tilde{\beta}_{I,y}$ shows a clear increase in solar power production
- Demand ($\hat{D}$): grid power+solar power

Results

Estimated solar power

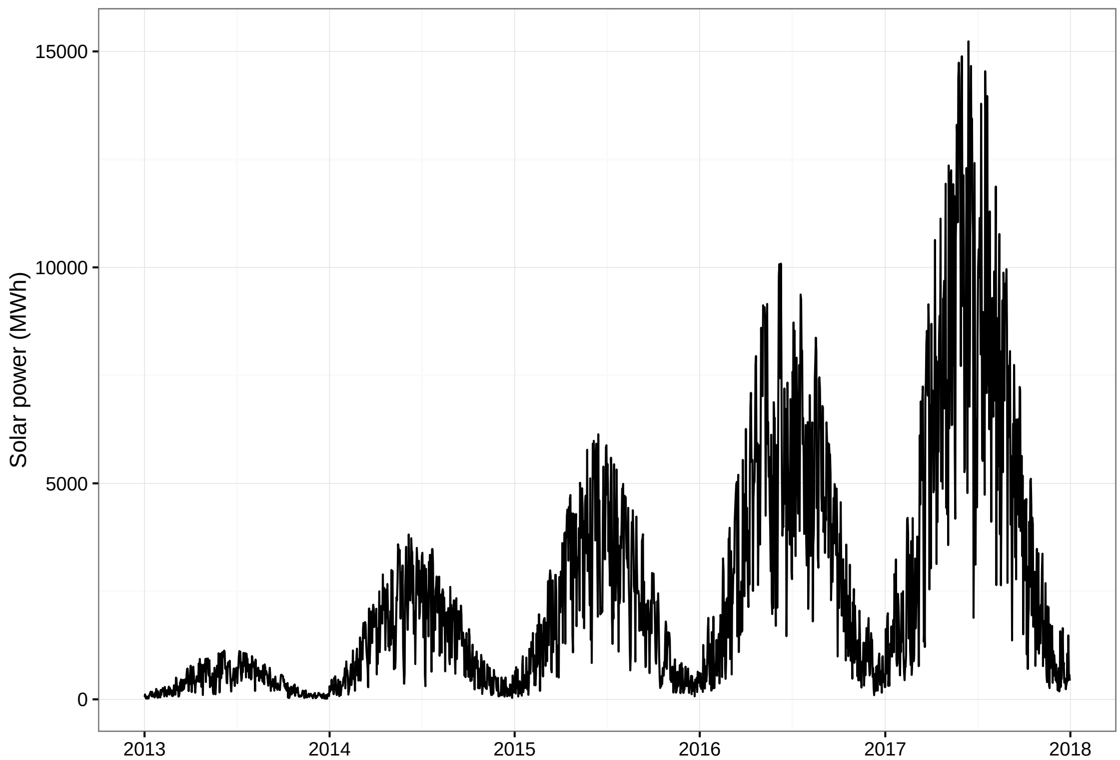

Figure 3: Estimated solar power for the years 2013—2017 in MWh.

Results

Model evaluation

- Standardized residuals
- Comparison with CBS publications on solar power production

Results

Table 1: Diagnostic checks on standardized residuals of the ARIMAX fit data set A and B.

Diagnostic	Data set A	Data set B
Skewness	-2.17	-1.88
Kurtosis	22.94	19.32
p-value Bowman-Shenton test on normality	0.00	0.00
p-value Box-Ljung test on autocorrelation	0.01	0.00
p-value F-test on heteroscedasticity	0.53	0.39

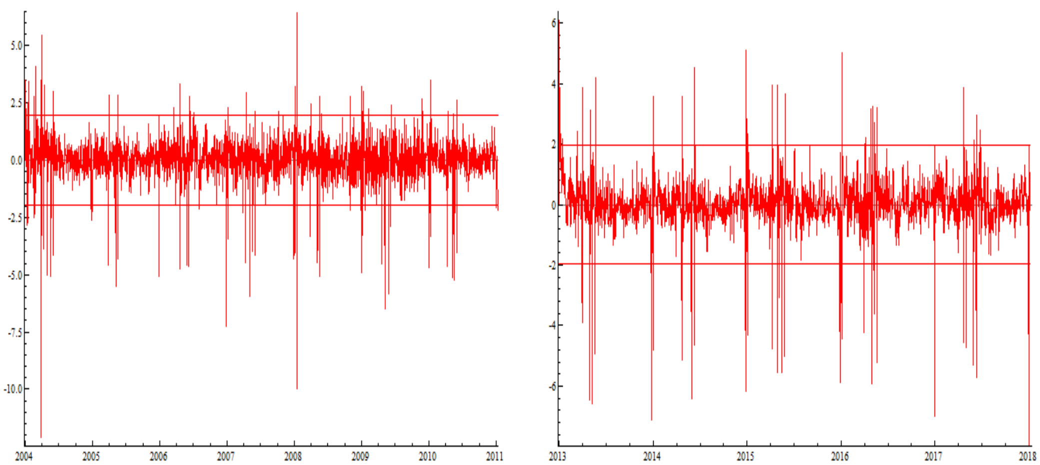

Figure 4: Standardized residuals of the ARIMAX model with a 95% confidence interval applied to data set A (left panel) and data set B (right panel).

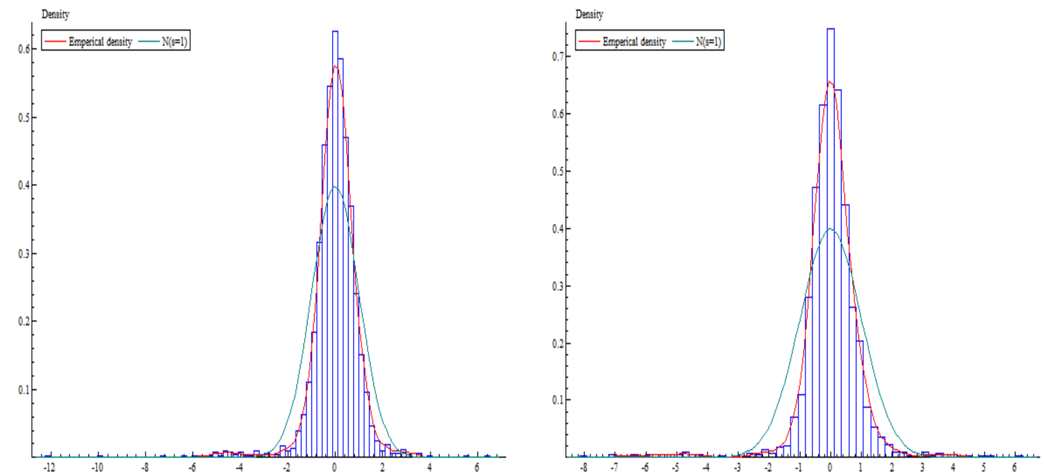

Figure 5: Histogram of the standardized residuals of the ARIMAX model with the empirical distribution and the standard normal distribution for data set A (left panel) and data set B (right panel).

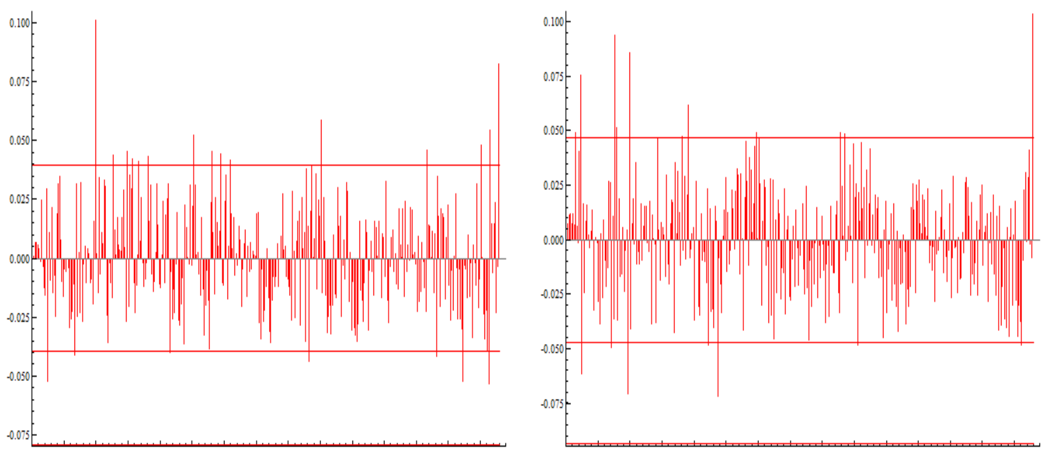

Figure 6: Correlogram of the residuals of the ARIMAX model with a 95% confidence interval for data set A (left panel) and data set B (right panel).

Results

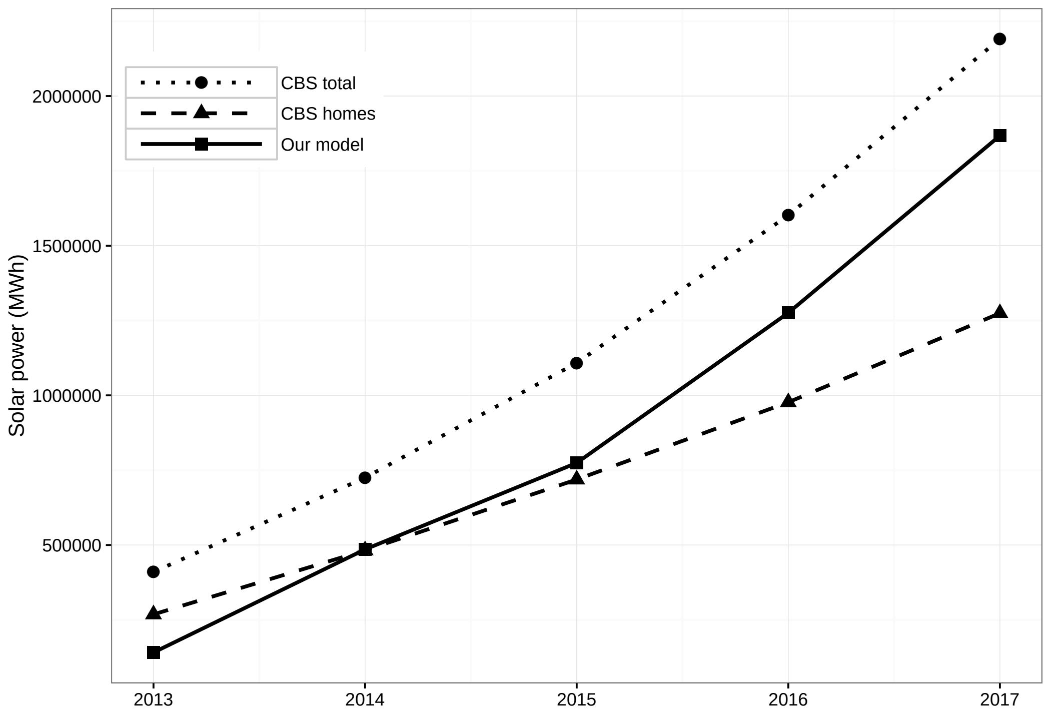

Figure 7: Comparison of our model results (solid line) with official statistics published by CBS on total solar energy consumption (dotted line) and the amount consumed by households (dashed line).

- Solid line ARIMAX estimate
- Dashed line: total solar power estimate (CBS) includes metered solar power by solar power farms
- Dotted line: solar power household PV installations

(CBS)

Divergence in 2016 and 2017 might be explained by unmetered PV installations of companies

Conclusions

- Statistical information on the use of renewable energy relevant for SDG indicators and energy transition
- Method to estimate unmetered solar power using data found on the internet
- Results do not disagree with CBS publications
- Improvements
  - Time series models (STM?)
  - More realistic modelling of interactions between temperature and production of solar power
  - Multivariate approach for regional estimates
  - Account for increase of unmetered wind energy

References

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

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Buelens, B. and van den Brakel, J. (2018). *Estimating unmetered photovoltaic power consumption sing causal models*. Technical report, Statistics Netherlands.

Elliott, M. R. and Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science* 32 (2), 249–264.

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

Kim, K. and Wang, Z. (2018). Sampling techniques for big data analysis in finite population inference. *International Statistical Review*.

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).

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* 82 (4), 669–688.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*.

Rosenbaum, P. and Rubin, D. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association* 79 (387), 516–524.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1), 41–55.

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

Valliant, R. and Dever, J. (2011). Estimating propensity adjustment for volunteer web surveys. *Sociological Methods & Research* 40, 105–137.

Valliant, R., Dever, J., and Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York: Springer Verlag.

Valliant, R., Dorfman, A., and Royall, R. (2000). *Finite population sampliong and inference: a prediction approach*. New York: Wiley and Sons.