Non-probability sampling and big data

EMOS Webinar 2022/23

Ralf Münnich

18. January 2023

What is the problem with survey data?

- Gallup opinion poll in 1948 on US elections
  - Dewey (Republicans) versus Truman (Democrats)
  - Use of quota sampling
  - Prediction: Dewey – but survey stopped early
  - Winner: Truman
- Johnson's red bus (Brexit), Trump election, etc.
- Huge debate in Germany:
  Market and opinion research versus internet surveys
  Non-response versus web selectivity
- What is a really good survey?

What is the problem with survey data?

- Gallup opinion poll in 1948 on US elections
  - Dewey (Republicans) versus Truman (Democrats)
  - Use of quota sampling
  - Prediction: Dewey – but survey stopped early
  - Winner: Truman
- Johnson's red bus (Brexit), Trump election, etc.
- Huge debate in Germany:
  Market and opinion research versus internet surveys
  Non-response versus web selectivity
- What is a really good survey?

What is the problem with survey data?

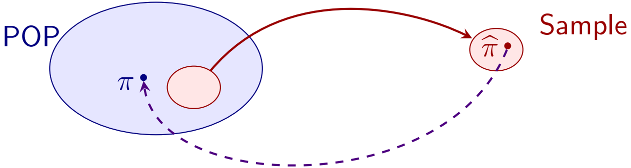
- Gallup opinion poll in 1948 on US elections
  - Dewey (Republicans) versus Truman (Democrats)
  - Use of quota sampling
  - Prediction: Dewey – but survey stopped early
  - Winner: Truman
- Johnson's red bus (Brexit), Trump election, etc.
- Huge debate in Germany:
  Market and opinion research versus internet surveys
  Non-response versus web selectivity
- What is a really good survey?

Content

General idea of *estimation*

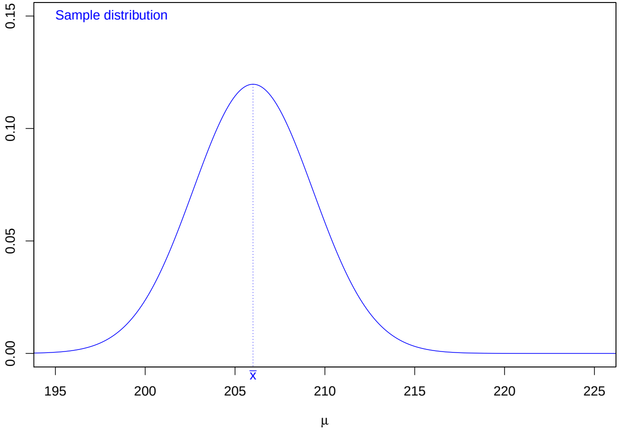
We are interested in population parameters which are generally unknown (here: π).

After analysing populations using methods of descriptive statistics, we now draw a sample of the population and evaluate the outcome ($\rightarrow$ point estimation).



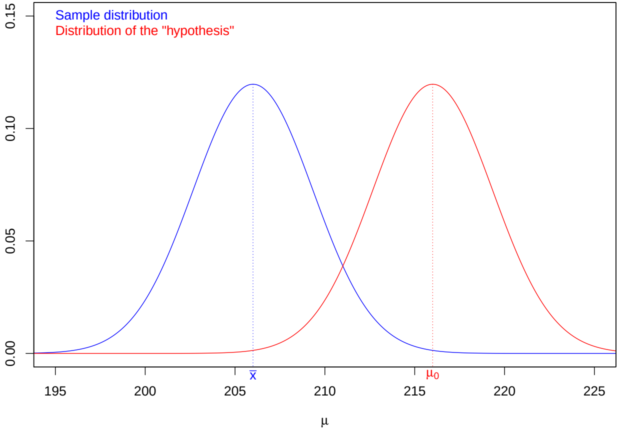
Additionally, we want to specify an interval of *plausible* values ($\rightarrow$ interval estimation in terms of providing quality information).

General idea of inference



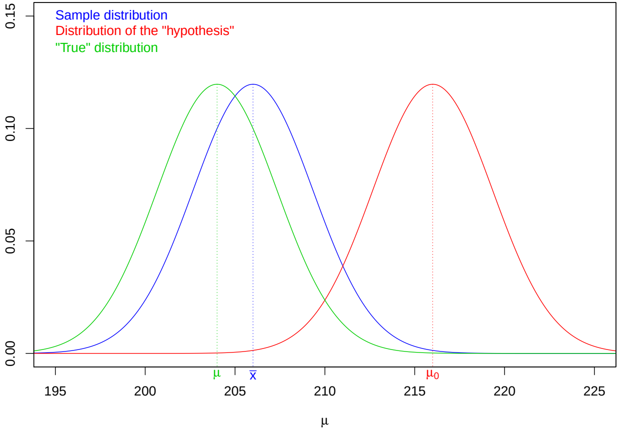
- Analysis of sample (estimation, e.g. using $\overline{x}$)
- Hypothesis for population (e.g. μ_0 for μ)
- True distribution *unknown* in reality

General idea of inference



- Analysis of sample (estimation, e.g. using $\overline{x}$)
- Hypothesis for population (e.g. μ_0 for μ)
- True distribution *unknown* in reality

General idea of inference



- Analysis of sample (estimation, e.g. using $\overline{x}$)
- Hypothesis for population (e.g. μ_0 for μ)
- True distribution *unknown* in reality

What is the impact of a quality concept?

Relevance of the statistical concept:
: End-user, *user needs*, hierarchical structure and contents

Accuracy and reliability:
- Sampling errors: standard error, CI coverage
- Non-sampling errors: nonresponse, coverage error, measurement errors

Timeliness and punctuality: Time and duration from data acquisition until publication

Coherence and comparability: Preliminary and final statistics, annual and intermediate statistics (regions, domains, time)

Accessibility and clarity: Publication of data, analysis and method reports

Completeness

See: European Statistics Code of Practice

How does this fit in a data science context?

Are all requirements of the previous slides met in a general framework of data collection in data science?

- Is the frame (core population) known and well-addressed?
- How is the data gathering process controlled?
  Is every unit separately drawable in a completely known way?
- Does every unit provide full information?
  Non-response is not solely an issue in public surveys!
- Are there any other sources of imprecision?
  Is the measuring process adequate/precise?

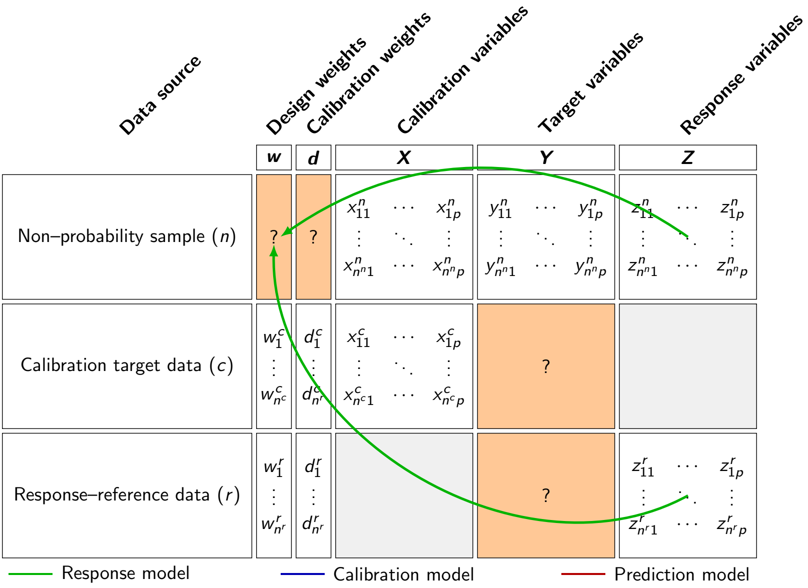
Note: possible sources of imprecision have to be controlled, and if possible, measured. The output has to be evaluated in light of the **data gathering process** including all these drawbacks. This includes also possible *corrections* of the results.

Non–probability samples

Main differences to classical probability samples:

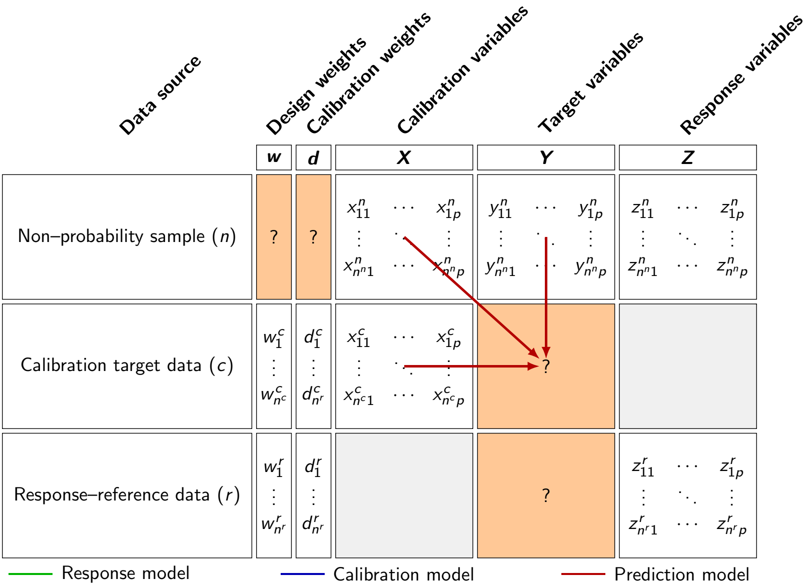
- Uncontrolled / non–random data–generating process
- Missing 'sampling'–information
  Inclusion / participation probability π_i unknown
  (and also π_{ij})
- Coverage of the target population not assured
  $\pi_i = 0$ possible (possibly also overcoverage)
- Possibly poor representativity & selection bias
- But: usually lower costs (probably also faster)
  Easier to obtain certain variables

Data source	Design weights	Calibration weights	Calibration variables	Target variables	Response variables
	$\boldsymbol{w}$	$\boldsymbol{d}$	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Z}$
Non–probability sample (n)	?	?	$\begin{matrix} x_{11}^n & \cdots & x_{1p}^n \\ \vdots & \ddots & \vdots \\ x_{n^n1}^n & \cdots & x_{n^np}^n \end{matrix}$	$\begin{matrix} y_{11}^n & \cdots & y_{1p}^n \\ \vdots & \ddots & \vdots \\ y_{n^n1}^n & \cdots & y_{n^np}^n \end{matrix}$	$\begin{matrix} z_{11}^n & \cdots & z_{1p}^n \\ \vdots & \ddots & \vdots \\ z_{n^n1}^n & \cdots & z_{n^np}^n \end{matrix}$
Calibration target data (c)	$\begin{matrix} w_1^c \\ \vdots \\ w_{n^c}^c \end{matrix}$	$\begin{matrix} d_1^c \\ \vdots \\ d_{n^r}^c \end{matrix}$	$\begin{matrix} x_{11}^c & \cdots & x_{1p}^c \\ \vdots & \ddots & \vdots \\ x_{n^c1}^c & \cdots & x_{n^cp}^c \end{matrix}$	?	
Response–reference data (r)	$\begin{matrix} w_1^r \\ \vdots \\ w_{n^r}^r \end{matrix}$	$\begin{matrix} d_1^r \\ \vdots \\ d_{n^r}^r \end{matrix}$		?	$\begin{matrix} z_{11}^r & \cdots & z_{1p}^r \\ \vdots & \ddots & \vdots \\ z_{n^r1}^r & \cdots & z_{n^rp}^r \end{matrix}$


Data source
Design weights
Calibration weights
Calibration variables
Target variables
Response variables
w
d
X
Y
Z
Non–probability sample (n)
Calibration target data (c)
Response–reference data (r)
?
Response model
Calibration model
Prediction model

Data source	Design weights $\boldsymbol{w}$	Calibration weights $\boldsymbol{d}$	Calibration variables $\boldsymbol{X}$	Target variables $\boldsymbol{Y}$	Response variables $\boldsymbol{Z}$
Non–probability sample (n)	?	?	$\begin{matrix} x^n_{11} & \cdots & x^n_{1p} \\ \vdots & \ddots & \vdots \\ x^n_{n^n1} & \cdots & x^n_{n^np} \end{matrix}$	$\begin{matrix} y^n_{11} & \cdots & y^n_{1p} \\ \vdots & \ddots & \vdots \\ y^n_{n^n1} & \cdots & y^n_{n^np} \end{matrix}$	$\begin{matrix} z^n_{11} & \cdots & z^n_{1p} \\ \vdots & \ddots & \vdots \\ z^n_{n^n1} & \cdots & z^n_{n^np} \end{matrix}$
Calibration target data (c)	$\begin{matrix} w^c_1 \\ \vdots \\ w^c_{n^c} \end{matrix}$	$\begin{matrix} d^c_1 \\ \vdots \\ d^c_{n^r} \end{matrix}$	$\begin{matrix} x^c_{11} & \cdots & x^c_{1p} \\ \vdots & \ddots & \vdots \\ x^c_{n^c1} & \cdots & x^c_{n^cp} \end{matrix}$	?	
Response–reference data (r)	$\begin{matrix} w^r_1 \\ \vdots \\ w^r_{n^r} \end{matrix}$	$\begin{matrix} d^r_1 \\ \vdots \\ d^r_{n^r} \end{matrix}$		?	$\begin{matrix} z^r_{11} & \cdots & z^r_{1p} \\ \vdots & \ddots & \vdots \\ z^r_{n^r1} & \cdots & z^r_{n^rp} \end{matrix}$

Response model — Calibration model — Prediction model


Data source
Design weights
Calibration weights
Calibration variables
Target variables
Response variables
w
d
X
Y
Z
Non–probability sample (n)
Calibration target data (c)
Response–reference data (r)
?
Response model
Calibration model
Prediction model

Data source	Design weights	Calibration weights	Calibration variables	Target variables	Response variables
	$\mathbf{w}$	$\mathbf{d}$	$\mathbf{X}$	$\mathbf{Y}$	$\mathbf{Z}$
Non–probability sample (n)	?	?	$x^n_{11} \cdots x^n_{1p}$ $\vdots \ddots \vdots$ $x^n_{n^n1} \cdots x^n_{n^np}$	$y^n_{11} \cdots y^n_{1p}$ $\vdots \ddots \vdots$ $y^n_{n^n1} \cdots y^n_{n^np}$	$z^n_{11} \cdots z^n_{1p}$ $\vdots \ddots \vdots$ $z^n_{n^n1} \cdots z^n_{n^np}$
Calibration target data (c)	w^c_1 $\vdots$ $w^c_{n^c}$	d^c_1 $\vdots$ $d^c_{n^r}$	$x^c_{11} \cdots x^c_{1p}$ $\vdots \ddots \vdots$ $x^c_{n^c1} \cdots x^c_{n^cp}$	?	
Response–reference data (r)	w^r_1 $\vdots$ $w^r_{n^r}$	d^r_1 $\vdots$ $d^r_{n^r}$		?	$z^r_{11} \cdots z^r_{1p}$ $\vdots \ddots \vdots$ $z^r_{n^r1} \cdots z^r_{n^rp}$

Response model — Calibration model — Prediction model

Response model

- Modelling the response process:

  Approximation of unknown inclusion / participation probability

  $$\pi_i \ = \ \mathrm{P}\left(i \in \mathcal{S}^n\right) \ = \ \mathrm{P}\left(R_i = 1\right)$$

  where

  $$R_i \ = \ \begin{cases} 1, & i \in \mathcal{S}^n \\ 0, & i \in \mathcal{S}^r \end{cases}$$

- Set of observations $\mathcal{S}^r$ outside non–probability sample required to estimate model

Response model

- by logistic regression model on response–variables $\boldsymbol{Z}$ using model parameters $\boldsymbol{\omega}$:

$$\pi_i \approx \widehat{\pi}_i := \left(1 + \exp\left(-\boldsymbol{z}_{i\cdot}^{\mathsf{T}}\boldsymbol{\omega}\right)\right)^{-1} \tag{1}$$

- Estimation via weighted binomial log-likelihood (pseudo-log-likelihood) of $\boldsymbol{R}$ given $\boldsymbol{\omega}$ and $\boldsymbol{Z}$

$$\log\left(\mathcal{L}_{\mathrm{o}}\left(R|\boldsymbol{\omega}, \boldsymbol{Z}^n, \boldsymbol{Z}^r, \boldsymbol{w}^n, \boldsymbol{w}^r\right)\right) = \sum_{i\in\mathcal{S}^n} w_i^n \cdot \log\left(\widehat{\pi}_i\right) + \sum_{i\in\mathcal{S}^r} w_i^r \cdot \log\left(1 - \widehat{\pi}_i\right)$$

- (Initial) weights $\boldsymbol{w}^n$ and $\boldsymbol{w}^r$ *might* be ones
  (cf. Fuller, 2009; Rosenbaum und Rubin, 1983; Valliant und Dever, 2011)

Response model: Individual weighting

- Non–probability sampling treated as (single or additional) stage of random sampling with unknown probabilites
- Assumption: Participation is a random phenomenon
- $\widehat{\pi}_i$ is treated like inclusion probability in random sampling

$$\widetilde{w}_i \; := \; w_i^n \cdot \widehat{\pi}_i^{-1} \quad \text{for all} \quad i \tag{2}$$

(cf. e.g. Little, 1988; Valliant und Dever, 2011)

Response model: Grouped weighting

- Model (2): Possibly high variance of estimators
- Replacing $\widehat{\pi}_i$ by mean of similar observations

$$\widetilde{w}_i \;=\; w_i^n \cdot \left(\sum_{j\in g} 1\right) \Big/ \left(\sum_{j\in g} \widehat{\pi}_j\right) \quad \text{for all} \quad i \in g \tag{3}$$

  g: class including observation i

- Aim:

  Reduce variability of weights

  Less vulnerability to model misspecification

Response model: Grouped weighting: Post–stratification

- Weighted class proportions are calibrated to those of $\mathcal{S}^r$
- Replacing $\widehat{\pi}_i$ by post–stratification weight of propensity classes

$$\widetilde{w}_i \;=\; w_i^n \cdot \left(\sum_{j \in (g \cup \mathcal{S}^r)} w_j^n\right) \Big/ \left(\sum_{j \in (g \cup \mathcal{S}^n)} w_j^r\right) \quad \text{for all} \quad i \in g \tag{4}$$

g: class including observation i

(cf. e.g. Little, 1986; Rosenbaum und Rubin, 1983; Valliant und Dever, 2011)

Calibration model

- Find weights such that estimates meet known totals

$$\boldsymbol{\tau}\left(\boldsymbol{X}^{n}, \widetilde{\boldsymbol{w}} \circ \boldsymbol{d}\right) \stackrel{!}{=} \boldsymbol{\tau}\left(\boldsymbol{X}\right) \qquad (5)$$

Different ways to achieve calibration constraints, e.g.

- Generalized regression estimator (GREG)
  (includes post–stratification)
- Raking

(cf. Deville und Särndal, 1992; Deming und Stephan, 1940)

Calibration model: Targets' quality

Known population totals as calibration targets

- Exact compliance (often) reasonable
- Possibly high variance in weights / estimates if $\boldsymbol{X}$ includes many variables

Estimated calibration targets

- Commonly, high–quality random samples are used
- Subject to (survey–)errors as well
- Exact compliance less reasonable
- Inexact (relaxed) calibration is considered

(cf. Chang und Kott, 2008; Deville und Särndal, 1992; Deville et al., 1993)

Calibration model: Relaxed constraints

- Exact compliance (equation (5)) is replaced by an adequate similarity:

$$\boldsymbol{\tau}\left(\boldsymbol{X}^{n}, \widetilde{\boldsymbol{w}}\right) \stackrel{!}{=} \boldsymbol{\tau}\left(\boldsymbol{X}^{c}, \boldsymbol{w}^{c}\right) \circ \boldsymbol{\epsilon} \qquad (6)$$

- $\boldsymbol{\epsilon}$ is a multiplicative error vector, determining the relation

$$\epsilon_k = \frac{\boldsymbol{\tau}\left(\boldsymbol{x}^{n}_{\cdot k}, \widetilde{\boldsymbol{w}}\right)}{\boldsymbol{\tau}\left(\boldsymbol{x}^{c}_{\cdot k}, \boldsymbol{w}^{c}\right)}$$

of the k–th total:

$$\begin{aligned} \epsilon_k &< 1: \quad \text{below target} \\ \epsilon_k &= 1: \quad \text{exact compliance} \\ \epsilon_k &> 1: \quad \text{above target} \end{aligned}$$

Calibration model: Relaxed constraints

Guggemos und Tillé (2010): 'penalized Calibration' resembles GREG:

$$\begin{aligned} \operatorname*{argmin}_{\boldsymbol{\omega},\boldsymbol{\epsilon}} & \left(\sum_{j=1}^{n^n} w_j^n \cdot \frac{(1-d_j)^2}{2} + \sum_{k=1}^{p} \mathrm{v}_k \cdot \frac{(1-\epsilon_k)^2}{2} \right) \\ \text{s.t.} \quad & \boldsymbol{\tau}\left(\boldsymbol{X}^n, \widetilde{\boldsymbol{w}}\right) \stackrel{!}{=} \boldsymbol{\tau}\left(\boldsymbol{X}^c, \boldsymbol{w}^c\right) \circ \boldsymbol{\epsilon} \\ & \mathrm{L}_{\epsilon_k} \leq \epsilon_k \leq \mathrm{U}_{\epsilon_k} \quad \text{for all} \quad k = 1, \ldots, p \end{aligned} \tag{7}$$

- Each ϵ_k is either fixed to 1 or unconstrained

Calibration model: Gelman bounds

- Limiting range / variation of weights: '*Gelman-bounds*':

$$\frac{\mathrm{Max}\,(\widetilde{\boldsymbol{w}})}{\mathrm{Min}\,(\widetilde{\boldsymbol{w}})} \qquad (8)$$

- Additional *boundary constraints* are introduced:

$$\mathrm{L}_{\boldsymbol{d}} \;\leq\; d_j \;\leq\; \mathrm{U}_{\boldsymbol{d}} \qquad \text{for all} \quad j \;=\; 1,\ldots,n^n \qquad (9)$$

- $\mathrm{L}_{\boldsymbol{d}}$ and $\mathrm{U}_{\boldsymbol{d}}$ are global lower and upper bounds for $\boldsymbol{d}$

(cf. Gelman, 2007; Meng et al., 2009; Münnich et al., 2012a)

Calibration model: Extensions

Münnich et al. (2012c):

- Arbitrary Box–constraints

  $L_{\widetilde{w}_j}$ and $U_{\widetilde{w}_j}$ for weights / Gelman factor and

  L_{ϵ_k} and U_{ϵ_k} for totals

Münnich et al. (2012b):

- Computational effective implementation
- Using duality–approach to
- Determine weights by Lagrange multipliers, thus reducing dimensions to number of constraints

Calibration model: Functional form approach

- Different distance functions often lead to very similar results
- Advantage of distance functions is questioned
- *Functional form approach* / *instrument vector approach*
- Correction depending on *instrument variables* $\boldsymbol{Z}$ and parameters $\boldsymbol{\omega}$:

$$\boldsymbol{d} := \mathbf{d}_{\circ}(\boldsymbol{\omega}, \boldsymbol{Z}^{n}) \tag{10}$$

- $\boldsymbol{Z}$ does not (necessarily) coincide with $\boldsymbol{X}$

(cf. Estevao und Särndal, 2000, 2006; Folsom und Singh, 2000)

Calibrated response model

- *Generalized raking model*

$$\mathbf{d}_{\mathrm{r}}(\omega, \boldsymbol{Z}^n) = \exp(\boldsymbol{Z}^n \omega) \tag{11}$$

- *Logit model* (inverse propensity score, cf. equation (1))

$$\mathbf{d}_{\mathrm{l}}(\omega, \boldsymbol{Z}^n) = 1 + \exp(-\boldsymbol{Z}^n \omega) \tag{12}$$

Calibrated propensity weights, but possibly differing parameter estimation:

- If $\dim(\boldsymbol{X}^n) = \dim(\boldsymbol{Z}^n)$: ω determined from constraints alone
- Otherwise, distance functions still needed

(cf. Folsom und Singh, 2000; Kott, 2003, 2006)

Calibrated response model: Distance functions & relaxed constraints

- Chang und Kott (2008) propose minimizing the distance to calibration targets for raking– or logit–model (11) and (12)

$$\underset{\boldsymbol{\omega}}{\operatorname{argmin}} \left(\sum_{k=1}^{p} \mathrm{v}_k \cdot \frac{(1-\epsilon_k)^2}{2} \right) \tag{13}$$

$$\text{s. t.} \qquad \boldsymbol{\tau}\left(\boldsymbol{X}^n, \widetilde{\boldsymbol{w}}\right) \overset{!}{=} \boldsymbol{\tau}\left(\boldsymbol{X}^c, \boldsymbol{w}^c\right) \circ \boldsymbol{\epsilon}$$

Prediction model: Linear Regression

- Regression equation

$$\widehat{y}_{il}^{n} = \beta_0 + \sum_{j=1}^{p} x_{ij}^{n} \cdot \beta_j \quad . \tag{14}$$

- $\boldsymbol{x}_{\cdot 0}^{n}$: intercept column,
- Parameters determined by least squares

$$\begin{aligned} \boldsymbol{\beta} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left((\boldsymbol{E}^{n})^{\mathsf{T}} \operatorname{diag}(\boldsymbol{w}^{n}) \, \boldsymbol{E}^{n} \right) \\ &= \left((\boldsymbol{X}^{n})^{\mathsf{T}} \operatorname{diag}(\boldsymbol{w}^{n}) \, \boldsymbol{X}^{n} \right)^{-1} (\boldsymbol{X}^{n})^{\mathsf{T}} \operatorname{diag}(\boldsymbol{w}^{n}) \, \boldsymbol{Y}^{n} \end{aligned} \tag{15}$$

Prediction model: Support vector machine

- Regression equation

$$\widehat{y}_{il}^{n} = \beta_0 + \sum_{j=1}^{p} x_{ij}^{n} \cdot \beta_j \tag{16}$$

- Parameters determined by

$$\begin{aligned} \boldsymbol{\beta} &= \underset{\boldsymbol{\beta}}{\text{argmin}} \left(\frac{1}{2} \cdot \sum_{j=1}^{p} \beta_j^2 + C \cdot \sum_{i=1}^{n^n} \xi_i + C \cdot \sum_{i=1}^{n^n} \xi_i \right) \\ \text{s. t.} & \qquad \xi_i,\ \xi_i \geq 0 \\ & \qquad |y_{il}^{n} - \widehat{y}_{il}^{n}| \leq e + \xi_i \end{aligned} \tag{17}$$

- Slack-variables ξ_i, ξ_i^* for violation of the
- *Maximum* distance e of points to the regression line

Further predictive methods

From the methodological point of view, any prediction model can be applied. However, they are relying on strong assumptions, and likely additional weighting, benchmarking or alignment methods might be considered.

A very detailed reading is:

Simon Lenau (2023): Statistical and Machine Learning Methods for Handling Selectivity in Non-Probability Samples. PhD dissertation, Trier University.

Further readings: InGRID deliverable at `https://www.inclusivegrowth.eu/project-output` and the references therein.

Graphical representation of missingness patterns

Two data sources. '✓' and '?' indicate observed and missing data

		X	Y
Probability Sample	1	✓	?
	⋮	⋮	⋮
	n	✓	?
Big Data Sample	1	✓	✓
	⋮	⋮	⋮
	N_B	✓	✓

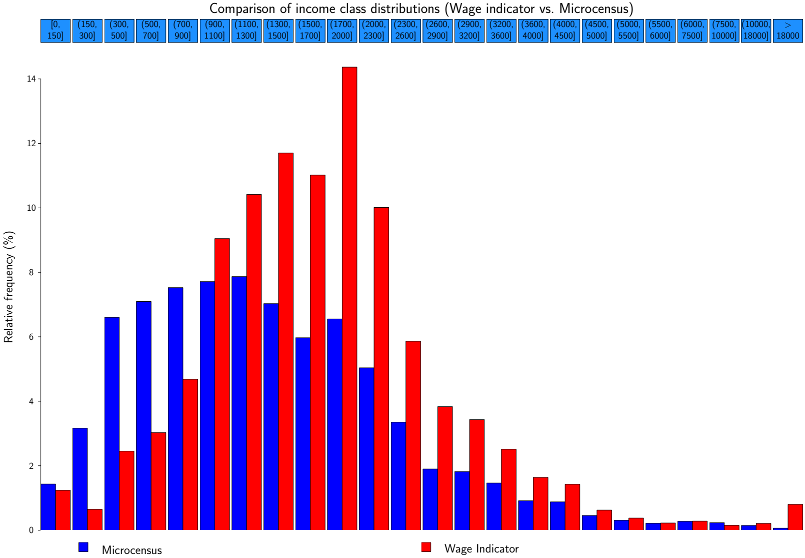
See Yang and Kim (2018)

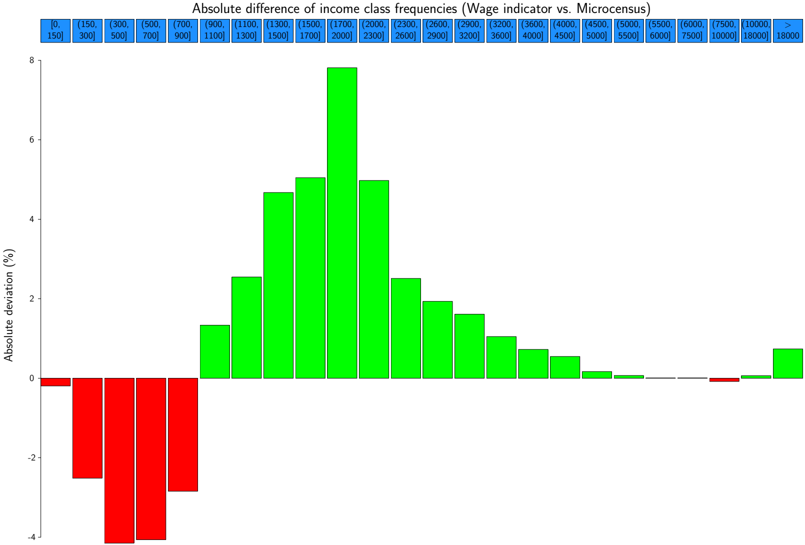
Combining information from individual records

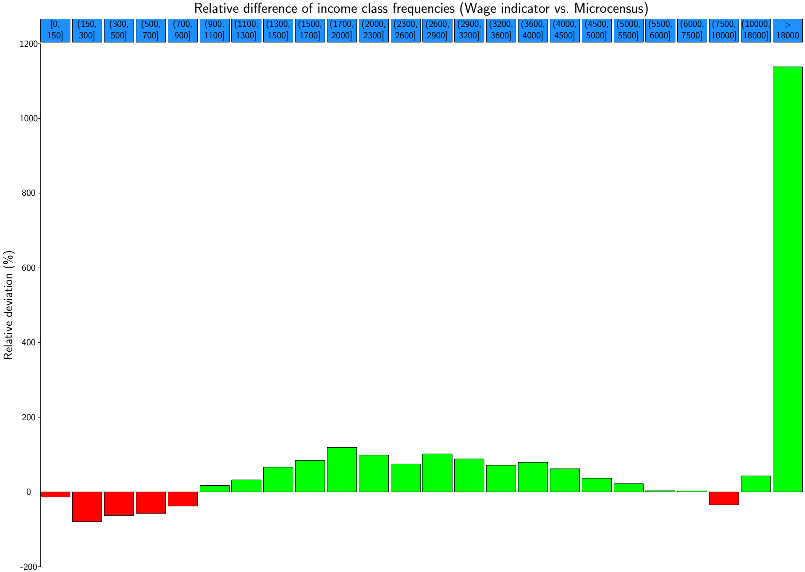
In some cases, data records for individuals can be merged from different sources. There are different methods applicable:

- Record linkage (possibly legal constraints)
- Statistical matching
- Multiple imputation
- Mass imputation

Again, model assumptions have to be met (e.g. CIA), some of which can be hardly verified.

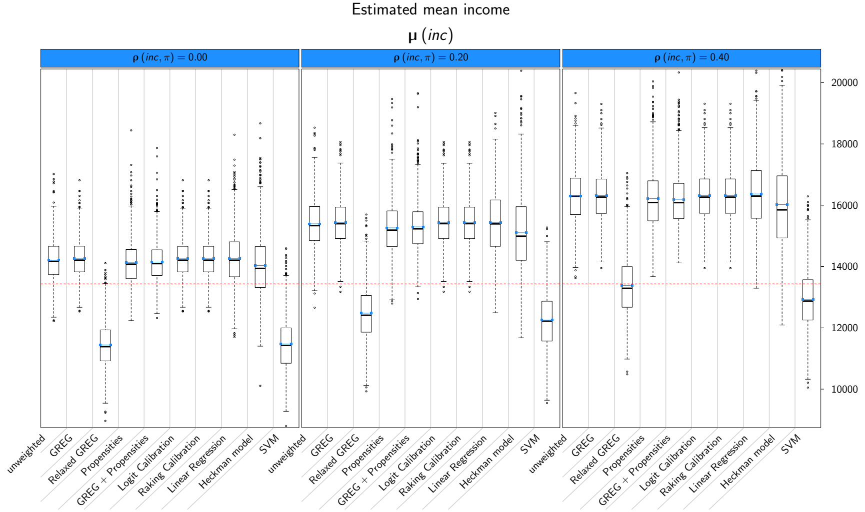

Comparison of income class distributions (Wage indicator vs. Microcensus)
[0, 150]
(150, 300]
(300, 500]
(500, 700]
(700, 900]
(900, 1100]
(1100, 1300]
(1300, 1500]
(1500, 1700]
(1700, 2000]
(2000, 2300]
(2300, 2600]
(2600, 2900]
(2900, 3200]
(3200, 3600]
(3600, 4000]
(4000, 4500]
(4500, 5000]
(5000, 5500]
(5500, 6000]
(6000, 7500]
(7500, 10000]
(10000, 18000]
> 18000
Relative frequency (%)
0
2
4
6
8
10
12
14
Microcensus
Wage Indicator


Absolute difference of income class frequencies (Wage indicator vs. Microcensus)
[0, 150]
(150, 300]
(300, 500]
(500, 700]
(700, 900]
(900, 1100]
(1100, 1300]
(1300, 1500]
(1500, 1700]
(1700, 2000]
(2000, 2300]
(2300, 2600]
(2600, 2900]
(2900, 3200]
(3200, 3600]
(3600, 4000]
(4000, 4500]
(4500, 5000]
(5000, 5500]
(5500, 6000]
(6000, 7500]
(7500, 10000]
(10000, 18000]
> 18000
Absolute deviation (%)
8
6
4
2
0
-2
-4


Relative difference of income class frequencies (Wage indicator vs. Microcensus)
[0, 150]
(150, 300]
(300, 500]
(500, 700]
(700, 900]
(900, 1100]
(1100, 1300]
(1300, 1500]
(1500, 1700]
(1700, 2000]
(2000, 2300]
(2300, 2600]
(2600, 2900]
(2900, 3200]
(3200, 3600]
(3600, 4000]
(4000, 4500]
(4500, 5000]
(5000, 5500]
(5500, 6000]
(6000, 7500]
(7500, 10000]
(10000, 18000]
> 18000
Relative deviation (%)
1200
1000
800
600
400
200
0
-200

Simulation setup

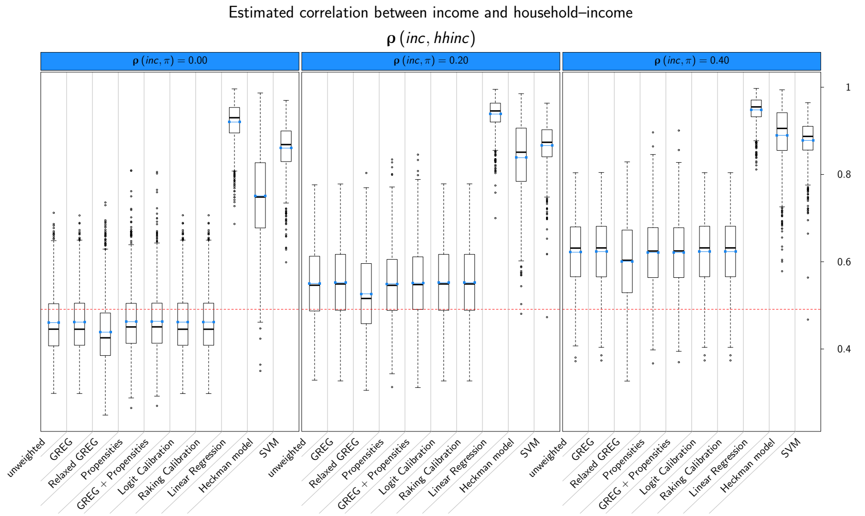
- Population: Subset of the AMELIA universe with $N = 20\,000$
- $R = 1\,000$ Poisson–samples of size $n = 1\,000$ (5%)
  Participation probabilities π
- Fixed correlations with variables
  - `gender`
  - `isced`
  - `bas`
  - `age`
- Varying correlation with income variable `inc`
- π assumed to be unknown
- Estimation with presented methods
  (cf. Burgard et al., 2017)



Calibration targets: Population

Calibration & prediction model variables: gender isced bas age

Response model variables: gender isced bas age hhinc

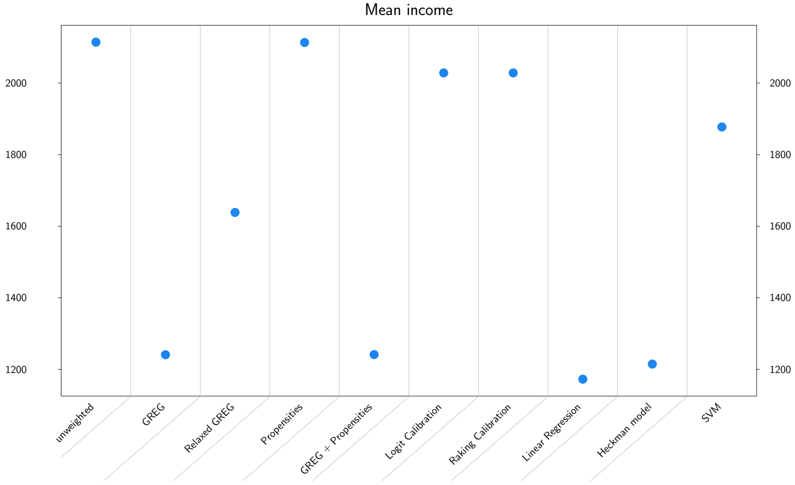


Calibration targets: Population

Calibration & prediction model variables: gender isced bas age

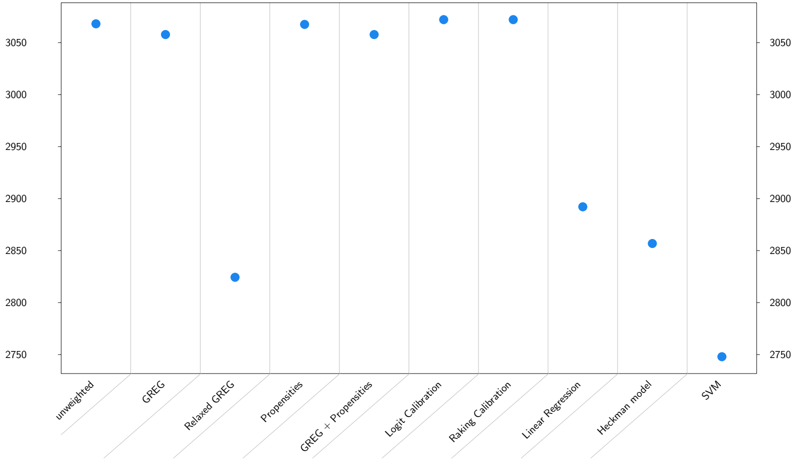
Response model variables: gender isced bas age hhinc

Estimates from the WageIndicator Survey:



Estimates from the WageIndicator Survey:

Mean household income



Two cases of use of Big Data

- Twitter sentiment analysis
- Quality measure of Satellite images

Twitter Sentiment Analysis

- Sentiment analysis is a natural language processing (NLP) technique used to determine whether data is positive or negative.
- Tweets collected using #*JoeBiden*, #*DonaldTrump*, #*Biden*, #*Trump* with the Twitter API
- 38,432,811 tweets were collected, employing streaming Tweepy API across the United States between 28 September 2020, and 20 November 2020.

See Chaudhry et al. (2021)

Figure: Twitter dataset

	created_at	tweet_id	tweet	likes	retweet_count	source	user_id	user_name
0	2020-10-15 00:00:01	1.316529e+18	#Elecciones2020 | En #Florida: #JoeBiden dice ...	0.0	0.0	TweetDeck	3.606665e+08	El Sol Latino News
1	2020-10-15 00:00:18	1.316529e+18	#HunterBiden #HunterBidenEmails #JoeBiden #Joe...	0.0	0.0	Twitter for iPad	8.099044e+08	Cheri A. us
2	2020-10-15 00:00:20	1.316529e+18	@IslandGirlPRV @BradBeauregardJ @MeidasTouch T...	0.0	0.0	Twitter Web App	3.494182e+09	Flag Waver
3	2020-10-15 00:00:21	1.316529e+18	@chrislongview Watching and setting dvr. Let's...	0.0	0.0	Twitter for iPhone	8.242596e+17	Michelle Ferg
4	2020-10-15 00:00:22	1.316529e+18	#censorship #HunterBiden #Biden #BidenEmails #...	1.0	0.0	Twitter Web App	1.032807e+18	the Gold State

Positivity proportion per state

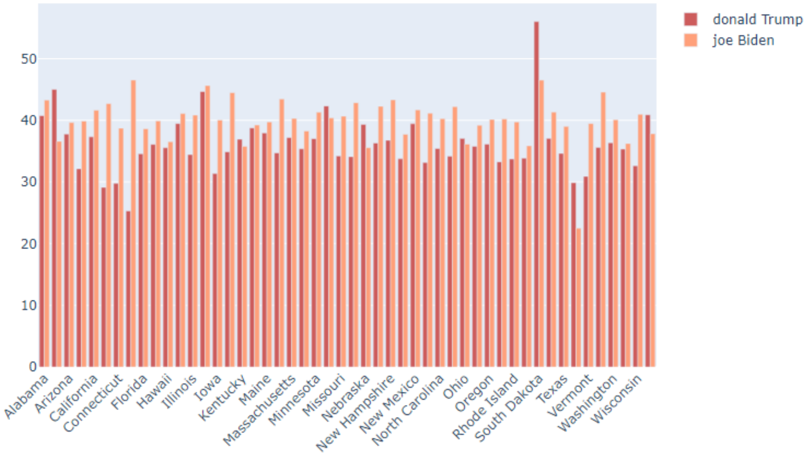


Figure: Sentiment analysis and real poll results for Joe Biden

	Maine	California	New York	Arkansas	Idaho
Sentiments	64.3%	64.2%	61%	62.1%	60.8%
Margin of Victory	9%	30%	23.2%	−27.6%	−30.7%

- Results in Maine, California, and New York aligned with the Twitter sentiment analysis
- Arkansas and Idaho although had positive Twitter sentiment, however, had opposing responses in the real poll.

Using satellite images

- Satellite-based recordings can have global coverage
- They are considered objective measurements
- Satellites take multiple recordings
- Some data are available for free
- Can act as a projection space to relate multiple measurements together with environmental information.

Improving Forest Inventory

Julian Wagner and Ralf Münnich et al. (2017) used satellite data to improve the Rhineland-Palatinate forest inventory.

- The standard German forest inventory (GNFI) selects some forest areas and collects high-quality information on about 150 variables.
- For local, small regions the sample size is insufficient.
- Airborn Laser scanning (ALS) data from 2002-2013 were available and combined with topographic maps.
- The forest canopy height minus the ground elevation results in normalized surface models.
- This measure of vegetation height is used as a proxy variable to improve the GNFI survey data in a small area estimation
- Problem: The canopy height is related non-linear to biomass.

Given:

- U a population of size N positions in D areas with populations sizes N_d
- A sample of n partitioned into subsamples S_d of size n_d.
- Estimating the small area mean θ for each area with:

$$\theta_d = \mu_{y,d} := \frac{1}{N_d} \sum_{i \in U_d} y_i$$

  n_d might be to small for reliable estimates of $\mu_{y,d}$ for each area.
- Canopy heights might be a good proxy to improve on the local estimates $\mu_{y,d}$
- Standard small area approaches are parametric model-based, implying a linear relationship of the dependent variable and predictors in the form of a random effect (see Battese et al. 1998).

$$y_{i_d} = \tilde{x}_{i,d}^T + u_d + e_{i,d}$$

Tree height and timber volume are not linear:

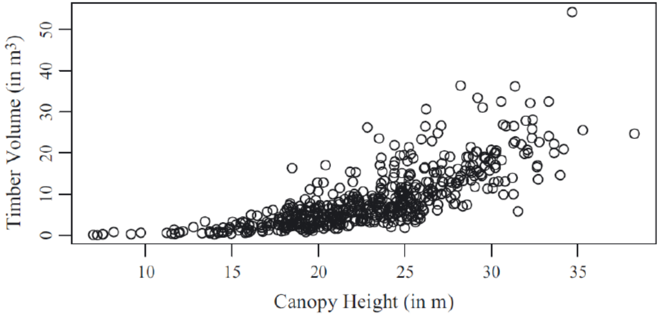


Figure 1 from: Wagner et al. (2017): Non-parametric small area models using shape-constrained penalized B-splines. in Royal Statistical Society, p.1089-1109

→ non-parametric, B-spline based estimation.

Data quality in NPS and big data

- Classical concepts of data quality or the total survey error do not really meet the needs for big data
- Often non-response measures are used (e.g. R indicators, cf. Shlomo, Schouten, and others)
- Fitness for use (user oriented, cf. Wang and Strong, 1996)
- Approaches are in development (e.g. Biemer, 2017)
- But inference (NPS and Big Data) is still an issue, and for big data additionally the complexity

Discussion and the references:

Münnich/Articus (2022): Big Data und Qualität - ist viel gleich gut? 85-99. In: B. Wawrzyniak/M. Herter (Eds.): Neue Dimensionen in Data Science, Berlin.

Summary

All approaches require auxiliary data:

Response models:

- Reference data with good response–predictors

Calibration models:

- Calibration targets, correlating with variables of interest

Prediction models:

- Totals / means (*linear* models), microdata (*non–linear* models)
- Good predictors for **every variable of interest**

Editing / Data cleaning of utmost importance

Issues with NPS and Big Data

- There is no general strategy but case-specific
- Weighting helps but in special cases, model-based methods might be better – but how to know?
- NPS have big advantages, when time matters, e.g. changes right before elections or special events (influential information)
- Big data can help a lot
  - mobile data: traffic control, or road use
  - satellite data: forest inventory, urban patterns, (independant) SDG indicators
- BUT: whenever we aim getting important information (e.g. for budget transfers), use proper high-quality survey data!

Issues with NPS and Big Data

- There is no general strategy but case-specific
- Weighting helps but in special cases, model-based methods might be better – but how to know?
- NPS have big advantages, when time matters, e.g. changes right before elections or special events (influential information)
- Big data can help a lot
  - mobile data: traffic control, or road use
  - satellite data: forest inventory, urban patterns, (independant) SDG indicators
- BUT: whenever we aim getting important information (e.g. for budget transfers), use proper high-quality survey data!

Thanks for your attention!

And thanks to the InGRID Research Infrastructure who financially supported the research on the wage indicator survey (https://www.inclusivegrowth.eu) as well as Simon Lenau and Abrar Ahmed for their support.

References I

Burgard, J. P., Kolb, J.-P., Merkle, H. und Münnich, R. (2017): *Synthetic data for open and reproducible methodological research in social sciences and official statistics*. In: AStA Wirtschafts- und Sozialstatistisches Archiv, 11 (3), S. 233–244.

Chang, T. und Kott, P. S. (2008): *Using calibration weighting to adjust for nonresponse under a plausible model*. In: Biometrika, 95 (3), S. 555–571.

Chaudhry, H. N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z. I., Shoaib, U. und Janjua, S. H. (2021): *Sentiment analysis of before and after elections: Twitter data of US election 2020*. In: Electronics, 10 (17), S. 2082.

References II

Deming, W. E. und Stephan, F. F. (1940): *On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known*. In: The Annals of Mathematical Statistics, 11 (4), S. 427–444.

Deville, J.-C. und Särndal, C.-E. (1992): *Calibration estimators in survey sampling*. In: Journal of the American statistical Association, 87 (418), S. 376–382.

Deville, J.-C., Särndal, C.-E. und Sautory, O. (1993): *Generalized raking procedures in survey sampling*. In: Journal of the American statistical Association, 88 (423), S. 1013–1020.

Estevao, V. M. und Särndal, C.-E. (2000): *A functional form approach to calibration*. In: Journal of Official Statistics, 16 (4), S. 379–399.

References III

Estevao, V. M. und Särndal, C.-E. (2006): *Survey estimates by calibration on complex auxiliary information*. In: International Statistical Review, 74 (2), S. 127–147.

Folsom, R. E. und Singh, A. C. (2000): *The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification*. In: Proceedings of the Survey Research Methods Section, American Statistical Association (2000), S. 598–603.

Fuller, W. A. (2009): Sampling Statistics. Wiley series in survey methodology.

Gelman, A. (2007): *Struggles with survey weighting and regression modeling*. In: Statistical Science, 22 (2), S. 153–164.

References IV

Guggemos, F. und Tillé, Y. (2010): *Penalized calibration in survey sampling: Design-based estimation assisted by mixed models*. In: Journal of statistical planning and inference, 140 (11), S. 3199–3212.

Kott, P. S. (2003): *A practical use for instrumental-variable calibration*. In: Journal of Official Statistics, 19 (3), S. 265.

Kott, P. S. (2006): *Using calibration weighting to adjust for nonresponse and coverage errors*. In: Survey Methodology, 32 (2), S. 133–142.

Little, R. J. (1986): *Survey nonresponse adjustments for estimates of means*. In: International Statistical Review, 54 (2), S. 139–157.

References V

Little, R. J. (1988): *Missing-data adjustments in large surveys*. In: Journal of Business & Economic Statistics, 6 (3), S. 287–296.

Meng, X., Duan, N., Chen, C. und Alegria, M. (2009): *Power-shrinkage: An alternative method for dealing with excessive weights*. In: Proceedings of the joint statistical meetings 2009.

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P. und Kolb, J.-P. (2012a): *Stichprobenoptimierung und Schätzung im Zensus 2011*. Statistik und Wissenschaft Bd. 21.

Münnich, R., Sachs, E. W. und Wagner, M. (2012b): *Calibration of estimator-weights via semismooth Newton method*. In: Journal of Global Optimization, 52 (3), S. 471–485.

References VI

Münnich, R. T., Sachs, E. W. und Wagner, M. (2012c): *Calibration of estimator-weights via semismooth Newton method*. In: Journal of Global Optimization, 52 (3), S. 471–485.

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

Valliant, R. und Dever, J. A. (2011): *Estimating propensity adjustments for volunteer web surveys*. In: Sociological Methods & Research, 40 (1), S. 105–137.