

An overview of methods for treating selectivity in big data sources

Maciej Beręsewicz

Department of Statistics, Poznań University of Economics and Business
Centre for Small Area Estimation, Statistical Office in Poznań

Workshop of the Baltic-Nordic-Ukrainian Network
on Survey Statistics 2018



POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Table of contents

- 1 Introduction
 - Motivation
 - What is big data?
 - Missing data mechanisms
- 2 How to correct selection bias ?
 - Quasi-randomization approach
 - Prediction / superpopulation model approach
 - Sample selection models
- 3 Summary
- 4 References

Introduction

Table of contents

- 1 Introduction
- 2 A statistical approach to big data
 - Defining big data
 - Statistical characteristics of big data
 - Addressing selectivity in big data
- 3 Analysis of specific big data sources
 - Mobile network data
 - Online social networks — Twitter
 - Web activity data — Google Trends
 - Web activity data — Wikipedia usage
- 4 Unit-level methods to correct selectivity
- 5 Domain-level methods to correct selectivity
- 6 Conclusions

Motivation – new data sources for statistics

- Increasing unit non-response in sample surveys;
- Growing information needs at a low level of (spatial) aggregation;
- *A change of paradigm in official statistics*, which involves the adoption / reuse of existing data sources instead of creating new ones (cf. ESSnet on Big Data);
- Internet data sources (IDSs) and big data are still not recognized and their *suitability as statistical sources is often unknown*;
- *New data sources*, in particular big data and the Internet have become *an important issue* in official statistics and small area estimation (cf. Daas et al., 2015; Japiec et al., 2015; Pfeffermann et al., 2015, Marchetti et al., 2015; Schmid et al., 2017).
- Note that, *The Internet, moreover, not only generates a great deal of today's "big data", but also provides ordinary-size data in a more accessible way – for example, access to public opinion polls or to local property records* (Citro, 2014).

What is big data?

Happy families are all alike;
every unhappy family is
unhappy in its own way

Leo Tolstoy

...tidy datasets are all alike but
every messy dataset is messy in
its own way

Hadley Wickham (JSS, 2014)

What is big data?

Happy families are all alike;
every unhappy family is
unhappy in its own way

Leo Tolstoy

...tidy datasets are all alike but
every messy dataset is messy in
its own way

Hadley Wickham (JSS, 2014)

What is big data?

Happy families are all alike;
every unhappy family is
unhappy in its own way

Leo Tolstoy

...tidy datasets are all alike but
every messy dataset is messy in
its own way

Hadley Wickham (JSS, 2014)

What is big data?



Figure 1: Example base transceiver station (BTS) from with several antennas located close by this Faculty

What is big data?

What is big data?

```
> orange_bts
      mobile_number payer_id      mobile_number_2      start_date length cost service_id location bts_id
1: f631585c5cdfcdcc44f54764a87cfa43 8636554 c215a10de85fc75d12183f45a133d524 2013-01-15 12:00:02 0 2.0 6 19720
2: a95c35c1b6457e0ce1070a6b8bb36c08 9165023 8944cba2bafe6c7b943cf31866cda1b3 2013-01-15 12:00:13 0 2.0 6 19720
3: c3244b0c788736657d89bd03e355f854 7862577 c215a10de85fc75d12183f45a133d524 2013-01-15 12:00:25 0 2.0 6 19720
4: 70e617965a33700fba187bc275e4b992 1889161 4f4302b958d33ee33b137c0c13b046b6 2013-01-15 12:00:19 0 2.0 6 19720
5: 0490b8101b50f29f695ae7644e31ba04 4356384 4f4302b958d33ee33b137c0c13b046b6 2013-01-15 12:00:27 0 2.0 6 19720
----
700855: 5117d5b1f3ae453f01c4353e28279ea1 2752745 3df913615ac1fdca400e8fa310eaa4d 2013-01-15 12:04:32 67 0.0 3001 5001
700856: 4197cf577efcbce7d7903808e81f8a1a 9142277 1203e523b3cbd175413c66bcd06391d7 2013-01-15 12:13:56 88 0.0 3001 5001
700857: 2215fd79744d43974bd304b9dbffff77e 2752745 3df913615ac1fdca400e8fa310eaa4d 2013-01-15 12:14:30 31 0.0 3001 5001
700858: d68be8f267d035864f2bc9fad9899cc25 8527717 120adb4955ae23fc5fbcfd00d784fa0 2013-01-15 12:05:00 23 0.1 3001 5001
700859: 499c9c853c5e5e40956de87385937cc7 10002961 blackberry.net 2013-01-15 12:02:33 0 0.0 1282 50000
```

Figure 2: Example records from Call Details Records (CDR) data (contracts only)

Two type of mobile phone data (in most cases there is no link with CRM systems):

- active (CDR) – needs an action e.g. call, sms, internet connection (39 mln actions on one day) – pseudo-response.
- passive (signalling data) – no actions needed, e.g. switching between antennas, signal strength – unaware response?

What are the implications?

What is big data?

What is big data?

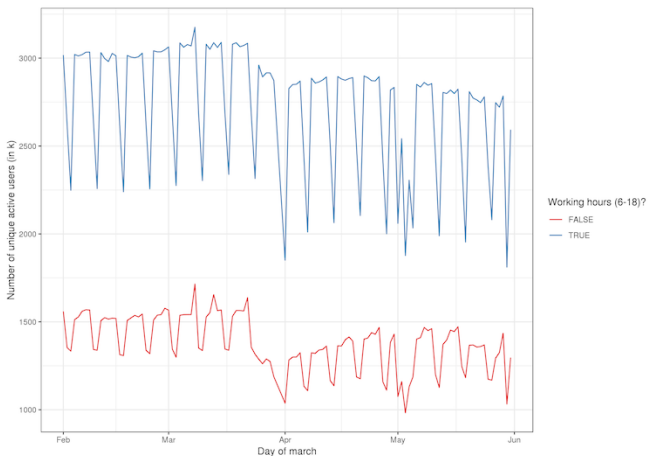


Figure 3: Number of active users in Poland based on CDR data (contracts only). Number of Polish citizens 15+ 32,7 mln

What is big data?


Table 1: Selected characteristics of user's activities between 1st February and 31th May

User ID	# contacts	# sms	# calls	calls duration	# bts	# days
891026	2549	4	12	0.28	2	27
1958650	683	–	454	4.24	7	74
4009251	461	–	302	7.25	4	81
3086158	460	–	305	7.22	5	81
4632841	460	–	300	7.13	5	81

Note: # contacts – daily average number of unique contacts, sms – daily median number of send sms, calls – daily median number of calls made, calls duration – daily average calls duration (in hours), # bts – number of unique bts, # day – number of days with activity (max 81).

Signalling data do not allow to calculate such statistics, thus may be harder to remove units that do not belong to the population of interest.

What is big data?



Pielęgniarka NOWA


DOM POMOCY SPOLECZNEJ

Białystok, PL

📅 Opublikowana mniej niż godzinę temu

🕒 Bądź jednym z pierwszych 10 aplikujących.

Aplikuj w witrynie firmy



Opis oferty pracy
Firma

Opis stanowiska

sprawowanie bezpośredniej opieki nad mieszkańcami, udzielanie pomocy przedlekarskiej w stanach zagrożenia życia oraz przy urazach, zranieniach, oparzeniach, składanie sprawozdania o stanie zdrowia podopiecznych, ich zachowaniu, podawanie leków , towarzyszenie mieszkańcom podczas konsultacji w poradniach specjalistycznych

Branża

Szpitalne i opieka zdrowotna

Forma zatrudnienia

Pełny etat

Doświadczenie

Partner

Obowiązki służbowe

Opieki zdrowotna

Figure 4: Example job offer from LinkedIn seen from the web browser

What is big data?

```

"decoratedJobPostingsModule": {
  "elements": [
    {
      "isInApply": false,
      "decoratedJobPosting": {
        "jobPosting": {
          "listDate": 1534860416000,
          "countryCode": "pl",
          "companyName": "Dom Pomocy Społecznej w Białymstoku",
          "listingType": "BASIC",
          "id": 816002536,
          "applicationRouting": "COMPANY_WEBSITE",
          "sourceDomain": "oferty.praca.gov.pl",
          "title": "Pielęgniarka"
        },
        "formattedDescription": "Sprawowanie bezpośredniej opieki nad mieszkańcami, udzielanie pomocy przedlekarskiej w stanach zagrożenia życia oraz przy urazach, zranieniach, oparzeniach, składanie ...",
        "formattedTitle": "",
        "cityState": "Białystok, PL",
        "companyName": "Dom Pomocy Społecznej w Białymstoku",
        "formattedLocation": "Białystok, PL",
        "decoratedCompany": {
          "company": {
            "companyId": 8069002,
            "names": [
              {
                "name": "DOM POMOCY SPOLECZNEJ",
                "active": true,
                "locale": "en_US",
                "type": "CANONICAL"
              }
            ],
            "universalName": "dom-pomocy-spolecznej"
          },
          "hasPaidLcp": false,
          "canonicalName": "DOM POMOCY SPOLECZNEJ"
        }
      }
    }
  ]
}

```

Figure 5: Example job offer from LinkedIn seen from a different perspective

How new are big data for statistics?

Non-statistical definitions of big data

Big data are highly detailed exhaust data automatically captured by sensors or generated during the use of IT systems.

Big data in statistics (survey methodology)

- an imperfect frame(s),
- non-probability character (cf. non-probability samples),
- **creation mechanism involves a self-selection process** (cf. non-response in surveys; opt-in panels),
- paradata-designed to capture the process and all possible data (cf. paradata in survey data collection).

How new are big data for statistics? – Big data survey

Big data survey – why?

- another type of secondary data source (collected for other purposes; cf. register-based surveys),
- a new kind of Internet survey (automated data collection made via the Internet and IoT),
- an opt-in panel (non-probability selection, longitudinal observation and data collection).

Possible ways of using big data

We can distinguish the following ways of using big data for statistics

- **as auxiliary variables** – cf. Marchetti et al. (2015), van den Brakel et al. (2017) – question: is this effort worth it?
- **as extension to existing sources** – e.g. demand for labour survey + online job vacancies (e.g. Beręsewicz et. al (2018b)).
- **as main source for statistics** – e.g. road statistics in the Netherlands.

How new are big data for statistics? – selectivity

(...) selectivity as a general term for selection errors resulting from:

- (self-selection) decisions of individuals (e.g. whether to tweet or use a particular mobile provider),
- decisions of the owners of the electronic platforms where data are captured (e.g. in terms of business concept, technical infrastructure), or
- the limitations of the technologies.

As a result, selectivity causes coverage, measurement and non-response (or missingness) errors, which introduce potential bias in estimates based on big data sources.

Not an easy task to distinguish from coverage and (non-response) selection bias.

Quality issues

Statistica Neerlandica (2012) Vol. 66, nr. 1, pp. 41–63
doi:10.1111/j.1467-9574.2011.00508.x

Topics of statistical theory for register-based statistics and data integration

Li-Chun Zhang*

Statistics Norway, Kongensgt. 6 Pb 8131 Dep, N-0033 Oslo, Norway

Journal of Official Statistics, Vol. 33, No. 2, 2017, pp. 477–511, <http://dx.doi.org/10.1515/JOS-2017-0023>

Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ

Giles Reid¹, Felipa Zabala², and Anders Holmberg³

Missing data mechanisms

Biometrika (1976), **63**, 3, pp. 581–92

Printed in Great Britain

581

Inference and missing data

BY DONALD B. RUBIN

Educational Testing Service, Princeton, New Jersey

SUMMARY

When making sampling distribution inferences about the parameter of the data, θ , it is appropriate to ignore the process that causes missing data if the missing data are ‘missing at random’ and the observed data are ‘observed at random’, but these inferences are generally conditional on the observed pattern of missing data. When making direct-likelihood or Bayesian inferences about θ , it is appropriate to ignore the process that causes missing data if the missing data are missing at random and the parameter of the missing data process is ‘distinct’ from θ . These conditions are the weakest general conditions under which ignoring the process that causes missing data always leads to correct inferences.

Missing data mechanisms

Some notation

- U – the target population, U_I is the Internet population, U_{NI} is the non-Internet population (in practice more complicated; cf. Zhang 2012, Beręsewicz et al. 2018a, ch. 3).
- N – size of the target population, $k = 1, \dots, N$,
- Y – target variable (e.g. number of job vacancies, support for given party); Y_{obs} observed data, Y_{mis} not observed (missed) data,
- $I_k \in \{0, 1\}$ – an indicator variable; if given unit from U has access to the Internet; $\sum_k (I_k = 1) = N_I$, $\sum_k (I_k = 0) = N_{NI}$
- $R_k \in \{0, 1\}$ – an indicator variable; if given unit from U or U_I participate in survey, study, or use given social media, or provide information on social media (e.g. tweet). This variable might be used to generally define any interaction with the Internet / IoT.
- $\rho_k = E(R_k)$ – the **response probability** of element k .

Missing data mechanisms

Rubin (1976) proposed the following classification of missing data mechanisms (we drop I_k for simplicity):

- Ignorable:

- Missing Completely at Random (MCAR)

$$Pr(R = 1 | Y_{obs}, Y_{mis}, \psi) = Pr(R = 1 | \psi) = const. \quad (1)$$

- Missing at Random (MAR) – depends solely on the observed data

$$Pr(R = 1 | Y_{obs}, Y_{mis}, \psi) = Pr(R = 1 | Y_{obs}, \psi). \quad (2)$$

- Non-ignorable:

- Not missing at random (NMAR) – depends both on the observed and non-observed data

$$Pr(R = 1 | Y_{obs}, Y_{mis}, \psi). \quad (3)$$

Real life issue – *Swiss cheese* data (full of item non-responses).

How to correct selection bias ?

Econometrica, Vol. 47, No. 1 (January, 1979)

SAMPLE SELECTION BIAS AS A SPECIFICATION ERROR

BY JAMES J. HECKMAN¹

This paper discusses the bias that results from using nonrandomly selected samples to estimate behavioral relationships as an ordinary specification error or “omitted variables” bias. A simple consistent two stage estimator is considered that enables analysts to utilize simple regression methods to estimate behavioral functions by least squares methods. The asymptotic distribution of the estimator is derived.

How to correct selection bias ?

Statistical Science

2017, Vol. 32, No. 2, 249–264

DOI: 10.1214/16-STS988

© Institute of Mathematical Statistics, 2017

Inference for Nonprobability Samples

Michael R. Elliott and Richard Valliant

Abstract. Although selecting a probability sample has been the standard for decades when making inferences from a sample to a finite population, incentives are increasing to use nonprobability samples. In a world of “big data”, large amounts of data are available that are faster and easier to collect than are probability samples. Design-based inference, in which the distribution for inference is generated by the random mechanism used by the sampler, cannot be used for nonprobability samples. One alternative is quasi-randomization in which pseudo-inclusion probabilities are estimated based on covariates available for samples and nonsample units. Another is superpopulation modeling for the analytic variables collected on the sample units in which the model is used to predict values for the nonsample units. We discuss the pros and cons of each approach.

Key words and phrases: Coverage error, hierarchical regression, quasi-randomization, reference sample, selection bias, superpopulation model.

How to correct selection bias ?

So, how we can correct the bias due to self-selection character of big data source?

In the statistical literature, we can find two approaches:

- **Quasi-randomization approach** – where we somehow weight our sample to known population totals, including modelling propensity to respond.
- **Prediction / superpopulation model approach** – where we build a model on a sample and then predict for out-of-sample units.
- ... and a mix of these two.

However, there is one general statistically sound (always working) method, nor established methodology that removes 100% of bias due to non-response.

How to correct selection bias ?

	Low Association (X, Y)	High Association (X, Y)
Low association (X, R)	Little effect on bias; Little effect on variance	Little effect on bias; Variance reduction
High association (X, R)	Little effect on bias; Variance inflation	Bias reduction; Variance reduction

Figure 6: Effect of bias correction method; Zhang, Thomsen and Kleven (2013)

All approaches requires Y or/and X, Z (auxiliary, proxy) to be present in data source and some known / estimated population totals/means/quantiles...

...but do we have (good) population totals or even defined population (e.g. day-night population, job vacancies)?

How to correct selection bias ?

	Low Association (X, Y)	High Association (X, Y)
Low association (X, R)	Little effect on bias; Little effect on variance	Little effect on bias; Variance reduction
High association (X, R)	Little effect on bias; Variance inflation	Bias reduction; Variance reduction

Figure 6: Effect of bias correction method; Zhang, Thomsen and Kleven (2013)

All approaches requires Y or/and X, Z (auxiliary, proxy) to be present in data source and some known / estimated population totals/means/quantiles...

...but do we have (good) population totals or even defined population (e.g. day-night population, job vacancies)?

How to estimate propensities or to reweight?

In the literature we can find the following approaches:

- **Reference survey.** One approach is to use a reference survey in parallel to the non-probability survey / big data. The underlying idea is to:
 - combine the reference sample and the sample of volunteers and fit a model to predict the probability of being in the non-probability sample / big data.
- **Sample matching** – be done on an individual or aggregate level
 - *individual-level matching* – matching cases from non-probability sample / big data to probability sample – e.g. using propensity scores (Rosenbaum and Rubin 1983)
 - *aggregate level consists* – making the frequency distribution of the non-probability sample / big data the same as that of the population – e.g. using post-stratification, calibration (cf. Devill and Särndal, 1992)).

Quasi-randomization approach

Ideally, we would like to construct weights that capture:

- Undercoverage error (denote by c_k)
- Different propensities to respond (denote by ρ_k)
- Differences in distributions of socio-demographic variables (denote by d_k)

In the end we are interested in some quantity, say, total:

$$\hat{Y} = \sum_k d_k \rho_k^{-1} c_k^{-1} y_i. \quad (4)$$

Calibration

- 1 This technique was proposed by Devill and Särndal (1992) and is a method of searching for so called **calibrated weights** by minimizing distance measure between **the sampling weights** and **the new weights**, which satisfy certain calibration constraints.
- 2 As a consequence when the new weights are applied to the auxiliary variables in the sample, they reproduce the known population totals of the auxiliary variables exactly.
- 3 It is also important that the new weights should be as close as possible to sampling weights in sense of chosen distance measure (Särndal C-E., Lundström S. 2005, Särndal C-E. 2007).
- 4 In case of big data **the sampling weights** can be replaced by either pseudo-weights equal to one or N/n .

Calibration – how to find weights

(C1) Find the minimum of distance function:

$$D(\mathbf{w}, \mathbf{d}) = \frac{1}{2} \sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i} \longrightarrow \min, \quad (5)$$

(C2) Calibration equations:

$$\sum_{i=1}^n w_i x_{ij} = \mathbf{X}_j, \quad j = 1, \dots, k, \quad (6)$$

(C3) Calibration constraints:

$$L \leq \frac{w_i}{d_i} \leq U, \quad \text{where: } L < 1 \text{ i } U > 1, \quad i = 1, \dots, n. \quad (7)$$

Can be easily done in R with `survey::calibrate`, `sampling::calib` or `laeken::calibWeights`.

Alternative approach – no auxiliary variables

Recently, Matei and Renalli (2015) suggested using latent trait models (item response theory) to deal with non-ignorable item non-response, i.e. missing data depends on the study variable itself. This is also known as missing not at random (MNAR) mechanism.

The proposed approach assumes that there is an underlying latent variable that drives respondents to answer questions.

The approach has two steps:

- **modelling non-response** based on item non-response in order to estimate probability to respond to survey denoted by p_i ,
- **modelling item non-response** to estimate probability to respond to given question denoted by q_{ik}

In the presentation we focus on the first part i.e. modelling probability to respond denoted by p_i .

Alternative approach – no auxiliary variables

Let assume that there is a latent variable θ_i denoting 'willing to respond' by unit i . Assuming that θ_i is known for all units in U , in absence of auxiliary variables we can estimate θ_i and thus, p_i by

$$p_i = Pr(R_i = 1|\theta_i) = \frac{1}{1 + \exp(-\alpha_0 - \alpha_1\theta_i)}. \quad (8)$$

Matei and Renalli (2015) proposed to estimate p_i in two steps:

- estimate θ_i using latent trait model based on non-response patterns created by item non-response,
- estimate p_i based on logistic regression using estimated $\hat{\theta}_i$ as a covariate.

Matei and Renalli (2015) estimator

Matei and Renalli (2015) proposed the following estimator of total:

$$\hat{Y} = \sum_i \frac{y_{ik}}{\pi_i p_i q_{ik}}, \quad (9)$$

where i denotes respondent, k denotes given question, π_i is the inclusion probability, p_i is the estimated propensity to respond, q_{ik} is the probability to respond to question k .

Note: Tested on small (<500) data, with bigger data calculation of p_i or q_{ik} might be problematic (near zero).

Prediction / superpopulation model approach

The general framework is as follows:

- We assume that model build on a sample holds for the population:

$$f_s(y_k|x_k) = \frac{Pr(R_i = 1|x_i, y_i) f_P(y_i|x_i)}{Pr(R_i = 1|x_i)}. \quad (10)$$

We do not assume any form of the model (e.g. machine learning could be used).

- Then we predict for the non-sampled units

$$\hat{y} = \sum_{k \in s} y_k + \sum_{k \notin s} \hat{y}_k. \quad (11)$$

Recently: LASSO regression for non-probability samples (Chen, 2016; Chen, Valliant and Elliott 2018), other types of model-calibration / model-assisted estimators (see Statistical Science) or multilevel regression and post-stratification (see Andrew Gelman works).

Sample selection models – Heckman

The basic idea is as follows, we model jointly selection (S) and outcome (O) equation, given by

$$\begin{aligned}y_i^{S*} &= \boldsymbol{\beta}^{S'} \mathbf{x}_i^S + \varepsilon_i^S \\y_i^{O*} &= \boldsymbol{\beta}^{O'} \mathbf{x}_i^O + \varepsilon_i^O\end{aligned}$$

where errors are correlated assuming normal distribution

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon^O \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix} \right)$$

then

$$y_i^O = \boldsymbol{\beta}^{O'} \mathbf{x}_i^O + \text{E} \left[\varepsilon^O | \varepsilon^S \geq -\boldsymbol{\beta}^{S'} \mathbf{x}_i^S \right] + \eta_i \equiv \boldsymbol{\beta}^{O'} \mathbf{x}_i^O + \rho \sigma \lambda \left(\boldsymbol{\beta}^{S'} \mathbf{x}_i^S \right) + \eta_i.$$

Sample selection models – Heckman and others

- Joint modelling selection mechanism and Y variable,
- Requires \mathbf{X} to formulate a model,
- Basic model works only for normally distributed data and probit link,
- Extensions and packages:
 - `SampleSelection` – binary and continuous,
 - `SemiParSampleSel` – binary, count and Continuous with copulas,
 - `miceMNAR` – multiple imputation based on Heckman's selection model for binary and continuous variables.
 - `GJRM` – Generalised Joint Regression Modelling – the most comprehensive software.

Other approaches: see Pfefferman & Sverchkov works, Sikov (2018), **Riddles et al. (2016)**.

Online job vacancies – auxiliary variables

Table 2: Quality of coding measured by number of job offers with digits based on pooled data from 2011, 2013 and 2014 for occupation (max 6 digits) and 2011-2014 for NACE (max 3 digits)

Number of digits	Occupancy	NACE
6 digits	1 693	–
5 digits	53 533	–
4 digits	5 311	–
3 digits	799	30 934
2 digits	313	15 609
1 digits	1	7 610

Online job vacancies – target variables

Table 3: Share of competences included in job offers by data source based on pooled data 2011, 2013 and 2014

Competences	Online data	District Labour Offices
Artistic	15.99	2.25
Availability	20.86	2.74
Cognitive	20.28	1.57
Computer	32.03	8.52
Interpersonal	54.34	6.84
Managerial	26.04	1.74
Mathematical	0.37	0.06
Office	3.82	1.72
Physical	5.29	1.85
Self-organization	59.08	7.80
Technical	4.27	5.02

Source: Beręsewicz et al. (2018).

Online job vacancies – auxiliary variables

Table 4: Cramer's V between competences and occupation, NACE and voivodeship based on the pooled data for 2011, 2013 and 2014

Competences	Occupation	NACE	Voivodeship
Artistic	0.24	0.11	0.08
Availability	0.19	0.10	0.05
Cognitive	0.27	0.16	0.10
Computer	0.47	0.21	0.13
Interpersonal	0.50	0.33	0.11
Managerial	0.37	0.19	0.06
Mathematical	0.05	0.04	0.03
Office	0.14	0.07	0.04
Physical	0.11	0.06	0.04
Self-organization	0.44	0.29	0.12
Technical	0.23	0.10	0.06

Online job vacancies – share of selected competences

Table 5: Share of selected competences in 2014 based on three estimators and online data

Competence	Naive	Calibrated	Heckman
Interpersonal	61%	45%	55%
Technical	4%	7%	14%
Managerial	31%	23 %	29%

Naive – simple average; *Calibrated* – adjusted for occupancy totals; *Heckman* – selection model based on occupancy and voivodeship, outcome model based solely on occupancy, based on multiple imputation with miceM-NAR. Population totals – estimates from the Demand of Labour survey.

Summary

- Strong auxiliary variables are needed – need to be extracted, if any exist.
- Population level information is required – definition of population? do we actually have these information or some proxy?
- One should consider that in big data NMAR is more common than MAR – however methods that account for this type of error are quite complex – consider bias exploration, if possible (cf. Zhang 1999).

Conclusion

Inference from probability samples are all alike but inference from messy non-probability sample (big data) is messy in its own way.

Are we actually prepared for big data?

Conclusion

Inference from probability samples are all alike but inference from messy non-probability sample (big data) is messy in its own way.

Are we actually prepared for big data?

References (selected) I

- Beręsewicz, M., Lehtonen, R., Reis, F., Di Consiglio, L., & Karlberg, M. (2018a). An overview of methods for treating selectivity in big data sources.
- Beręsewicz, M., Białkowska, G., Marcinkowski, K., Maślak, M., Opiela, P., & Zadroga, K. (2018b), Extending the Demand for Labour Survey by qualifications from online job offers, working paper.
- Beręsewicz, M. & Hinc, T., (2018), Approximate nearest neighbours imputation, working paper.
- Belsby, L., Bjornstad, J., & Zhang, L.-C. (2005). Modeling and Estimation Methods for Household Size in the Presence of Nonignorable Nonresponse Applied to the Norwegian Consumer Expenditure Survey. *Survey Methodology*, (12).
- Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, 78(2), 161-188.
- Buelens, B., Daas, P. J. H., Burger, J., Puts, M., & van den Brakel, J. (2014). Selectivity of Big Data, (11). Retrieved from <http://www.cbs.nl/NR/rdonlyres/457A097A-DA43-4006-AFE0-A8E8316CFEF0/0/201411x10pub.pdf>
- Carroll, P., Murphy, T., Hanley, M., Dempsey, D., & Dunne, J. (2018). Household Classification Using Smart Meter Data. *Journal of Official Statistics*, 34(1), 1–25.
- Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J. A., . . . Thompson, M. (2017). Approaches to Improving Survey-Weighted Estimates. *Statistical Science*, 32(2), 227–248. <http://doi.org/10.1214/17-STS609>
- Diaz, F., Gamon, M., Hofman, J. M., Kıcıman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PloS one*, 11(1), e0145406.

References (selected) II

- Elliott, M. R., & Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, 32(2), 249–264. <http://doi.org/10.1214/16-STS598>
- Galimard, J. E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in medicine*, 35(17), 2907-2920.
- Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1-23.
- Haziza, D., & Beaumont, J.-F. (2017). Construction of Weights in Surveys: A Review. *Statistical Science*, 32(2), 206–226. <http://doi.org/10.1214/16-STS608>
- Heckman, J. Sample Bias As A Specification Error, *Econometrica*, 1979, 47(1), 153-162.
- Sikov, A. (2018). A brief reiew of approaches to non-ignorable non-response. *International Statistical Review*, 0(0), 1–27. <http://doi.org/10.1111/insr.12264>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*. <http://doi.org/10.1093/biomet/63.3.581>
- Toomet, O., & Henningsen, A. (2008). Sample selection models in R: Package sampleSelection. *Journal of statistical software*, 27(7), 1-23.
- Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980-991.
- Wojtys, M., Marra, G., & Radice, R. (2016). Copula regression spline sample selection models: the R Package SemiParSampleSel. *Journal of Statistical Software*, 71(6).
- Zhang, L. C. (1999). A note on post-stratification when analyzing binary survey data subject to nonresponse. *Journal of Official Statistics*, 15(2), 329-334.
- Zhang, L. C., Thomsen, I., & Kleven, Ø. (2013). On the Use of Auxiliary and Paradata for Dealing With Non-sampling Errors in Household Surveys. *International Statistical Review*, 81(2), 270-288.

Thank you!