Small area estimation to correct for measurement errors in big population registers with application to Israel's census

Danny Pfeffermann and Dano Ben-Hur

Central Bureau of Statistics, Israel

UNECE Group of Experts

Population and Housing Censuses

Genève, September, 2019

Content

1 Propose a new method for running a census, which combines a **survey** with **administrative** (**big?**) data.

2 Propose a new way of integrating the survey information with the population register for constructing a single census estimator, **accounting for errors in the register**.

3 Consider alternative ways for testing the ignorability of the missing survey data and how to account for it.

4 Illustrate the procedures by use of data from Israel's 2008 census.

2008 Census in Israel

Israel has a fairly accurate population register; almost perfect at the **country level**.

Population register much **less accurate** for **small statistical areas**, with an **average** enumeration error of ~**15% (~8% in large cities)** and **95 percentile** of **~40%**.

Main reason for inaccuracy at statistical area level: people moving in or out of areas, often **report late** their change of address (or not at all).

2008 census in Israel (cont.)

In 2008, the **ICBS** conducted an **integrated census**, which consisted of the **population register**, corrected by estimates obtained from two **coverage samples** for each area. A **field (area) sample** of dwellings for estimating the ***register undercount*** (**U sample**), **&** a **telephone sample** of people **registered in the area** for estimating the **register over-count** (**O sample**).

- Register undercount is the number of people living in an area but not registered as living in that area.

- Register over-count is the number of people registered as living in an area but not living in that area.

Difficulties with the field U sample

- Requires listing all the apartments in each statistical area, or at least in a sample of cells or buildings in each area. **Very costly** and involves verifying that the listed apartments are dwelling units.
- Coverage problems in places where there are access restrictions such as closed floors, closed gates,...
- Problems in locating sampled apartments when collecting data, because not all the apartments are identified at the listing process.
- Response on internet is encouraged, but because of the problems above, not clear which households already responded on the Web.
- Many **logistic problems** in performing such a large scale operation.

New method planned for our 2020/21 census

- Select a **single sample** from the **population register** and obtain information from sampled units on residence of members of the household on census day **+ socio/economic/demographic infor.**
- The sample will be carried out by Internet → telephone → personal interview. (**No prior listing required!!**)
- If no nonresponse or if response is missing completely at random, the **sample estimates** computed from the sample will be **design-unbiased** (over all possible sample selections).
- Combine the sample estimate with the register count in each statistical area, to form an **empirical minimum design mean square error composite** estimator.

Computation of sample estimator

Denote by $P_i = \frac{N_i}{N}$; $N = \sum_i N_i$, the true proportion of people in the register living in area i , and by $\hat{P}_i$ the corresponding sample estimator.

Let $\mathbf{K} \cong N$ denote the size of the register on census day.

The sample estimator for the count of area i is: $\hat{N}_i = K \times \hat{P}_i$.

The design variance of this estimator is: $Var_D(\hat{N}_i \mid K) = K^2 Var_D(\hat{P}_i) = \sigma_{Di}^2$.

- We shall draw a stratified sample in each area with the strata defined by age, marriage status and other variables affecting residence, with different sample sizes in different areas (depending on the area size).
- Definition of estimator will depend on availability of "good" covariates.

Final Composite estimator

Combine the direct estimator with the register count by use of a **composite estimator** of the form,

$$\hat{N}_{i,Com} = \alpha_i \hat{N}_i + (1-\alpha_i) K_i .$$

- The register count is a **fixed number** in a given census day and therefore has no variance, but it can be wrong (**biased**).

- The sample estimator is (approximately) unbiased but has a variance.

$$MSE(\hat{N}_{i,Com}) = E_D(\hat{N}_{i,com} - N_i)^2 = \alpha_i^2 Var_D(\hat{N}_i) + (1-\alpha_i)^2 (K_i - N_i)^2 .$$

Determination of the weighting coefficient α_i

$$\hat{N}_{i,Com} = \alpha_i \hat{N}_i + (1-\alpha_i) K_i$$

$$MSE(\hat{N}_{i,Com}) = E_D(\hat{N}_{i,com} - N_i)^2 = \alpha_i^2 Var_D(\hat{N}_i) + (1-\alpha_i)^2 (K_i - N_i)^2 .$$

The coefficient α_i, minimizing the ***MSE*** is,

$$\alpha_{i,opt} = \frac{(K_i - N_i)^2}{(K_i - N_i)^2 + Var_D(\hat{N}_i)} .$$

In practice we don't know N_i, hence estimate $\alpha_{i,opt}$ as,

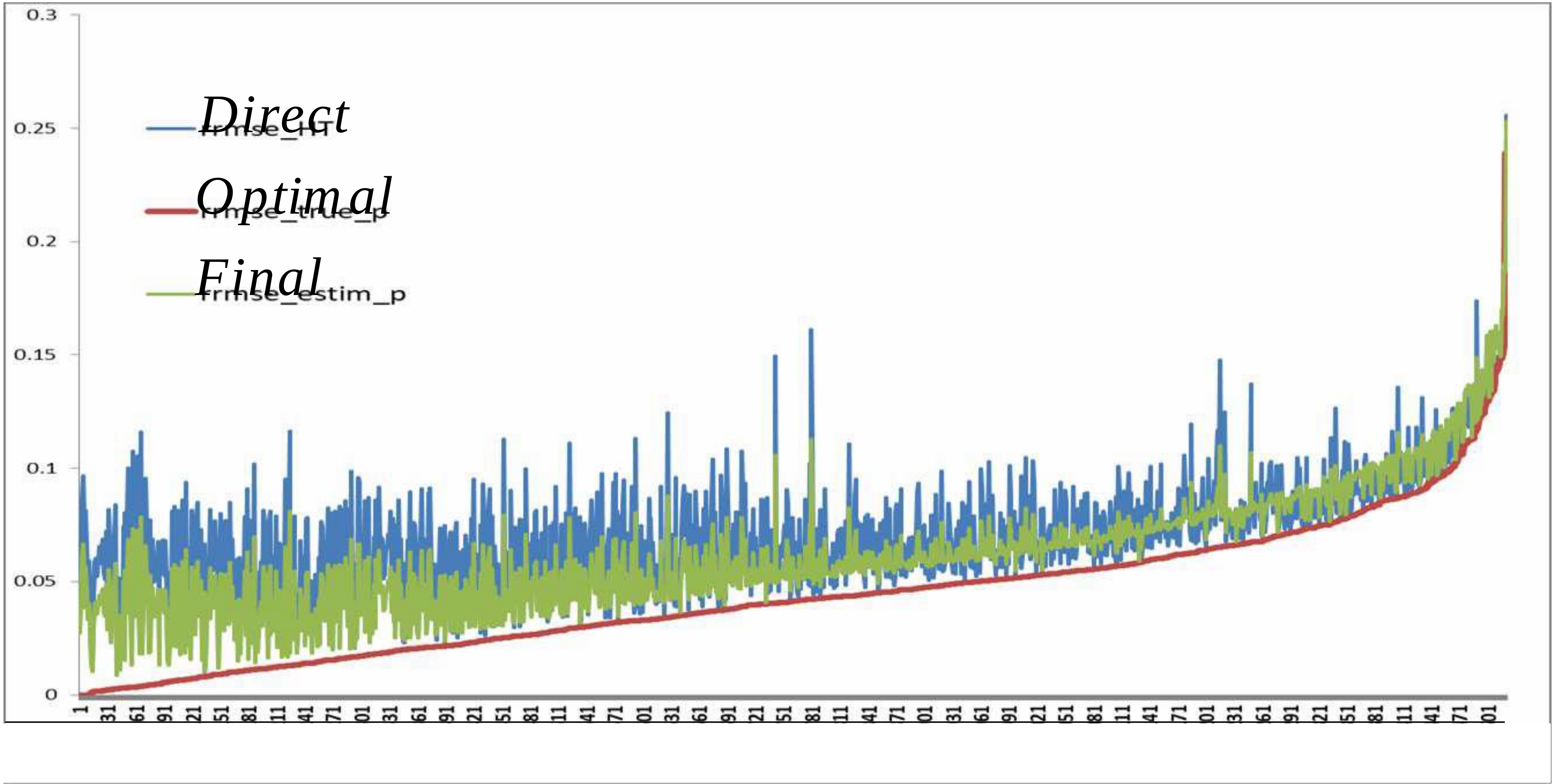
$$\hat{\alpha}_i = \frac{(K_i - \hat{N}_i)^2}{(K_i - \hat{N}_i)^2 + \hat{Var}_D(\hat{N}_i)} .$$

Final composite estimator: $\hat{N}_{i,fin} = \hat{\alpha}_i \hat{N}_i + (1 - \hat{\alpha}_i) K_i .$

Empirical results based on U- sample of 2008 Census

- **U- sample** consists of about **1,100,000** individuals.
- Consider U- sample in each area as **“true area population”**.
- Simple random sample of **10%** of individuals listed in the register from each statistical area.
- **More efficient sampling designs and estimators currently tested**!!
- Calculated for each statistical area the **direct** (standard) estimator, the **optimal** composite estimator (with optimal weights), and the final composite estimator with **estimated weights**.
- Sampling and estimation repeated **1,000** times.

Relative Root Mean Square Error in simulation experiment



Results shown for 1521 largest areas sorted by RRMSE of opt. est.

Results. Summary table

	Estimator	Mean	5th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	99th Pctl
Relative Root Mean Square Error	Direct	0.0700	0.0317	0.0548	0.0683	0.0833	0.0986	0.1448
	Optimal	0.0437	0.0043	0.0211	0.0414	0.0589	0.0808	0.1383
	Final	0.0623	0.0253	0.0442	0.0580	0.0751	0.0958	0.1443
	Register	0.1114	0.0044	0.0239	0.0578	0.1104	0.2236	0.8501
Absolute Relative Bias	Direct	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Optimal	0.0251	0.0040	0.0152	0.0250	0.0339	0.0414	0.0573
	Final	0.0160	0.0019	0.0075	0.0151	0.0227	0.0300	0.0469
	Register	0.1114	0.0044	0.0239	0.0578	0.1104	0.2236	0.8501
Relative Standard Error	Direct	0.0700	0.0317	0.0548	0.0683	0.0833	0.0986	0.1448
	Optimal	0.0329	0.0004	0.0086	0.0283	0.0484	0.0731	0.1348
	Final	0.0598	0.0268	0.0433	0.0561	0.0711	0.0908	0.1429
	Register	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Variance of composite estimator

The composite estimator is a **nonlinear function** of many estimators:

$$\hat{N}_{i,fin} = \hat{\alpha}_i \hat{N}_i + (1 - \hat{\alpha}_i) K_i$$

$$= \frac{(K_i - \hat{N}_i)^2 \hat{N}_i + \hat{\sigma}^2_{Di} K_i}{(K_i - \hat{N}_i)^2 + \hat{\sigma}^2_{Di}} \, .$$

The complexity of the composite estimator makes it very difficult to derive an explicit expression for its variance. Consequently, we use the **nonparametric bootstrap method**.

Nonparametric Bootstrap (BS) for variance estimation- Illustration

- Draw a **10%** sample of persons listed in the register from each statistical area - hereafter, the **original sample**.

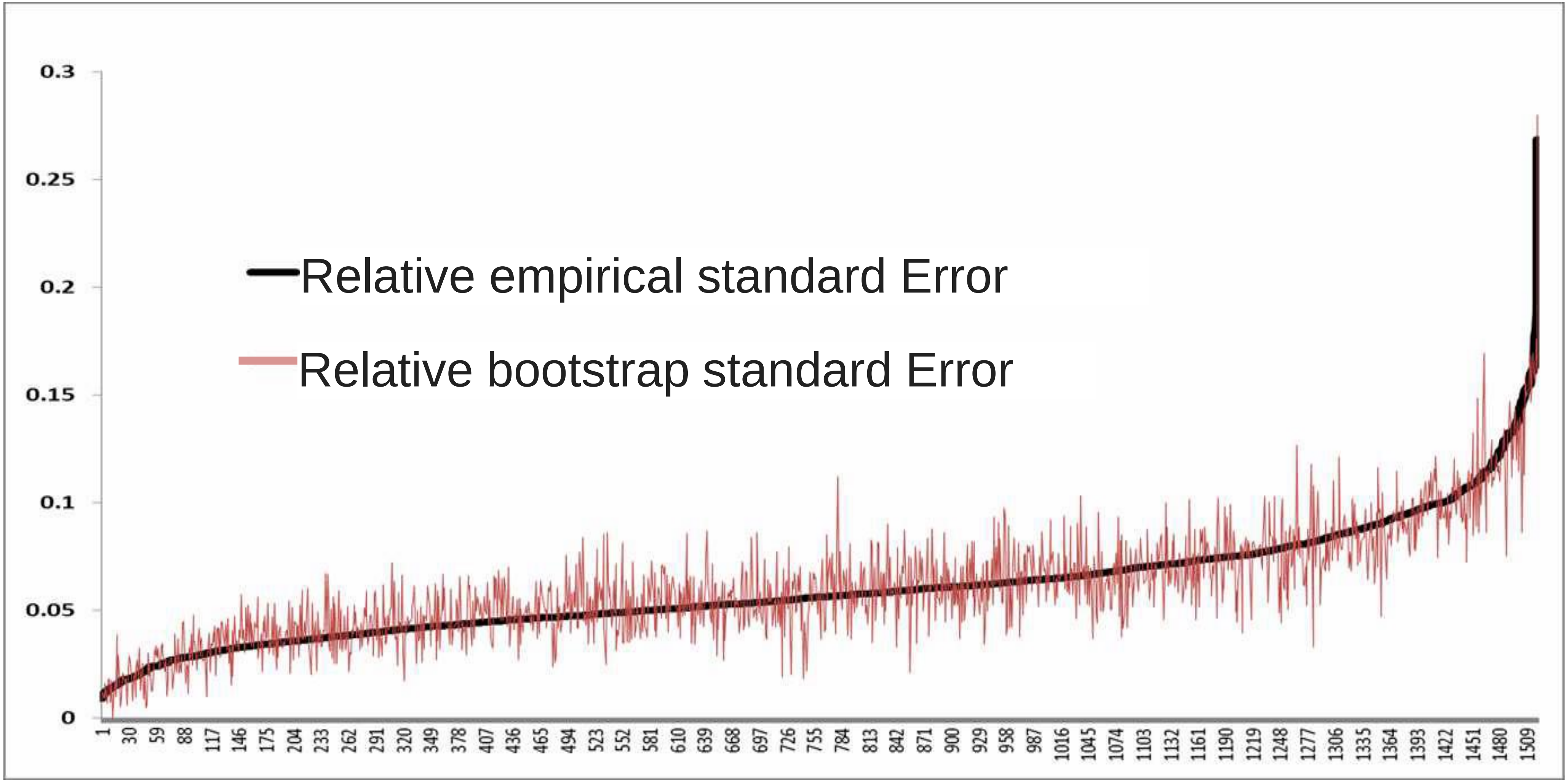
- Draw a sample **with replacement** from each original sample (**same sample size**), **1,000** times.

- For each statistical area and **BS** sample, compute the **final estimator**

- Calculate the **bootstrap variance** between the final estimators within each statistical area, over the **1000** **BS** samples.

- Compare the bootstrap variance with the empirical variance in each statistical area, as calculated from the **1,000** original samples.

Use of nonparametric Bootstrap for variance estimation (cont.)



Evaluation of proposed method, some initial thoughts

❖ Evaluation of census method **Integral part** of census planning.

1- Compare the design-based estimators with the register counts.

❖ Will help evaluating the potential of running an administrative census.

Example: Construct an interval around the true count in each area based on the sample estimate, **e.g.,** $K_i \overset{?}{\in} \hat{N}_i \pm C \times STD(\hat{N}_i)$, check if the interval covers the register count. **If not, extend the sample (allow a priory).**

❖ Evaluation used to **correct** the estimation, not just for "quality reports".

2Compare the final census estimates of **counts** and **socio-economic variables** with corresponding **administrative** or **survey estimates** at **aggregated levels**, for which the latter estimates are deemed **reliable**.

3 Special evaluation procedures for hard to count sub-populations.

Adjusting for not missing at random (NMAR) nonresponse

❖Like in any survey, a census is subject to nonresponse. In Israel's **O** (telephone) sample of the 2008 integrated census, the nonresponse rate was **24%**, even though response to the census is mandatory.

❖When the nonresponse is **"explained"** by the values of known covariates, adjusting for nonresponse is relatively easy. (Missing at random, **MAR**)

❖However, when the nonresponse depends, at least in part, on the value of the target outcome (not missing at random, **NMAR**), adjusting for nonresponse is complicated and generally requires, statistical modelling.

Can we **test** if the nonresponse is **NMAR?** Can we **adjust** for it**?**

Non-response in Telephone sample of 2008 Census in Israel

High rate of nonresponse in telephone survey, **average ~ 24%**.

Questions: Is the Nonresponse informative? Can we deal with it?

Total "O" (telephone) sample size 266,973

62,767 non-respondents

23.5%

204,206 Respondents

76.5%

29,257 Lived in a different area from their registered area, **14.3%**

174,949 Lived in their registered area, **85.7%**

Response in telephone sample by size of administrative family

Administrative family size	Total number approached by telephone	Percent persons for which information was obtained
1	39,003	53.79%
2	29,714	70.65%
3	31,023	74.79%
4	46,396	79.95%
5	47,430	84.35%
6	31,653	83.90%
7	19,175	84.91%
8	11,388	84.04%
9	6,855	84.77%
10	4,336	85.61%
Total	**266,973**	**76.49%**

The larger the family size, the larger the response.

Proportions of response by family size and number of telephones

Administrative family size	Number of telephone numbers per family				
	1	2	3	4	+5
1	0.44	0.6	0.7	0.74	0.76
2	0.56	0.63	0.76	0.78	0.82
3	0.47	0.62	0.71	0.76	0.82
4	0.53	0.66	0.75	0.78	0.85
+5	0.57	0.69	0.8	0.83	0.88

Proportions increase with family size and with number of telephones

Logistic regression to predict response in telephone survey

Variable	Odds ratio	Standard Error	Wald Chi-Square	Pr > ChiSq
# of telephones per family	1.7	0.003	24797.97	<0.0001
Administrative family size	1.12	0.004	748.29	<0.0001
Other ages	1	-	-	-
Age 20-29	0.97	0.014	3.64	0.0562
Age 30-39	0.82	0.015	178.32	<0.0001
Single	1	-	-	-
Married	1.22	0.012	261.67	<0.0001
Widower	1.21	0.029	44.29	<0.0001
Divorced	0.6	0.022	574.82	<0.0001
Jew	1.05	0.012	18.93	<0.0001
Other	1	-	-	-
Born in Israel	1.33	0.013	463.81	<0.0001
Other	1	-	-	-

Is the response ignorable when predicting residence?

Distribution of estimated response probabilities under the model

Residence in Register	n	Mean	5th Pctl	25th Pctl	75th Pctl
Lived in different area from registered area	29,257	0.81	0.49	0.82	0.88
Lived in registered area	174,949	0.81	0.49	0.82	0.88
Total	**204,206**	**0.81**	**0.49**	**0.82**	**0.88**

Intermediate conclusion: "Same" distribution of response probabilities in the two groups⇒**Response ignorable when predicting residence.**

What about prediction of other variables?

Suppose we want to estimate the **percentage of divorced people** in each area and we only know it for the respondents in the survey.

- The **O-sample** was drawn from the population register and the true number of divorced persons registered in each area is actually **known**.

We fit a logistic model for estimating the probability of response with all the covariates of the previous model, **except for the marriage status**.

Distribution of estimated response probabilities under the model

Marriage status	Sample size	Mean	5th Pctl	25th Pctl	75th Pctl
Other	195,773	0.815	0.489	0.822	0.885
Divorced	8,433	0.742	0.359	0.683	0.843
Total	**204,206**	**0.812**	**0.487**	**0.819**	**0.885**

Intermediate conclusion: Different distribution of response probabilities in the two groups$\Rightarrow$ **Response not ignorable when predicting marriage status.**

Can we account for NMAR nonresponse?

Sverchkov & Pfeffermann (JRSS-A, 2018) propose a method that uses the **Missing Information Principle** (**MIP**) for estimating the response probabilities in small areas.

Basic idea: Construct the likelihood that would be obtained if the missing **outcome** values **were also known** for the nonrespondents, and then integrate out the missing values with respect to the **distribution of the nonrespondents.** The latter distribution is defined by the **distribution of the respondents' outcomes** as fitted to the observed values.

- In what follows we show how the method performs when predicting the **true number of divorced persons registered in the area**.
- The **O-**sample is drawn from the population register and the true number of divorced persons registered in each area is actually **known**.

Notation and models considered

Outcome variable- $\boldsymbol{y_{ij}}$**;** $y_{ij} = 1$ if person j registered in area i is divorced, $y_{ij} = 0$ otherwise.

Covariates- $\mathbf{x}_{ij}$**;** same as before. **Response indicator-** $\boldsymbol{R_{ij}}$**;** $R_{ij} = 1$ if unit j in area i responds, $R_{ij} = 0$ otherwise.

Models fitted for **observed outcomes** of **responding units,** and for **response probability**:

$$Pr(y_{ij}|\mathrm{x}_{ij}, u_i, \boldsymbol{R_{ij}} = \mathbf{1}) = \frac{exp(\beta_0 + \mathrm{x}'_{ij}\beta + u_i)}{1 + exp(\beta_0 + \mathrm{x}'_{ij}\beta + u_i)};\quad u_i \sim N(0, \sigma_u^2)$$ **random effect,**

$$\Pr(\boldsymbol{R_{ij}} = \mathbf{1}|\mathrm{y}_{ij}, \mathrm{x}_{ij}; \gamma) = \frac{\exp(\gamma_0 + \mathrm{x}'_{ij}\gamma + \boldsymbol{\gamma}_{\mathbf{y}} y_{ij})}{1 + \exp(\gamma_0 + \mathrm{x}'_{ji}\gamma + \boldsymbol{\gamma}_{\mathbf{y}} y_{ij})}.$$

- If $\gamma_y \neq 0$, the nonresponse is **nonignorable** (**NMAR**).

Logistic model to predict response when estimating residence

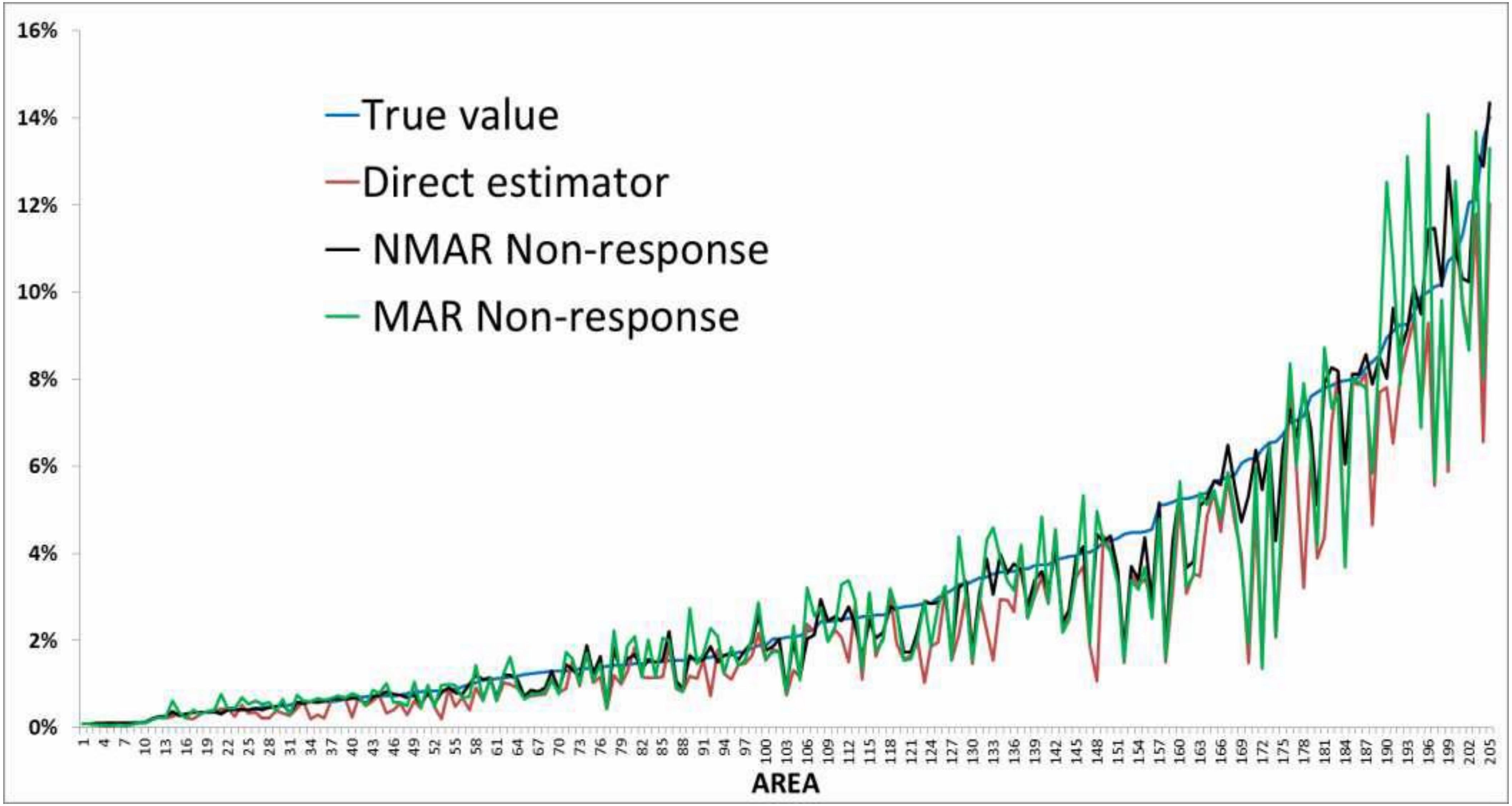
Covariates	Odds ratio MAR	Odds ratio NMAR
# of telephones per family	1.70	1.71
Administrative family size	1.12	1.11
Other ages	1.00	1.00
Age 20-29	0.97	0.97
Age 30-39	0.82	0.83
Single	1.00	1.00
Married	1.22	1.21
Widower	1.21	1.22
Divorced	0.60	0.59
Jew	1.05	1.05
Other	1.00	1.00
Born in Israel	1.33	1.32
Other	1.00	1.00
Lives in registered area	-	0.989 P.Value=0.472

Logistic model to predict response when estimating marriage status

Variable	Odds ratio	Odds ratio NMAR
# of telephones per family	1.70	1.83
Administrative family size	1.15	1.11
Other ages	1.00	1.00
Age 20-29	0.98	0.95
Age 30-39	0.87	0.86
Jew	1.04	1.05
Other	1.00	1.00
Born in Israel	1.27	1.25
Other	1.00	1.00
Divorced	-	**0.531 (P.Value<0.0001)**

- **Sampling design or response may be informative with respect to one outcome, and not with respect to another outcome.**

Percent of divorced persons in areas



Difference between true value and estimates (BIAS, marriage)

Estimate	Mean	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
Direct	0.0075	-0.0005	0.0006	0.0036	0.0099	0.0211
MAR	0.0033	-0.0077	-0.0018	0.0004	0.0057	0.0168
NMAR	0.0019	-0.0027	-0.0004	0.0001	0.0032	0.0094

Absolute relative distance of estimates from true value

Estimator	Mean	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
Direct	0.270	0.042	0.121	0.233	0.406	0.551
MAR	0.256	0.032	0.113	0.216	0.379	0.472
NMAR	0.118	0.004	0.022	0.055	0.156	0.362

Summary

1 Proposed a new method for running a census, combining **sample estimates** with **administrative** (**big?**) data √

2 Sample drawn from the (partly erroneous) administrative data √

3 Proposed a way of combining the survey information with the register into a single estimator, accounting for errors in the register √

4 Considered alternative ways to testing the informativeness of the missing sample data and how to account for it. √

5 Illustrated the three topics by use of real empirical data. √