Proceedings of Statistics Canada Symposium 2018
Growth in Statistical Information: Challenges and Benefits

Small Area Estimation to Correct for Measurement Errors In Big Population Registers

Danny Pfeffermann[1] and Dano Ben-Hur[2]

Abstract

Like in many countries, Israel has a fairly accurate population register at the national level, consisting of about 9 million persons. However, the register is much less accurate for small geographical (statistical) areas, with an average area enumeration error of about 13%. The main reason for the inaccuracy at the area level is that people moving in or out an area are often late in reporting their change of address. In order to correct the errors at the area level in our next census, we investigate the following three-step procedure:

A- Draw a sample from the register to obtain initial direct sample estimates for the number of persons residing in each area on "census day",
B- Apply the Fay-Herriot model to the direct estimates in an attempt to improve their accuracy,
C- Compute a final estimate for each statistical area as a linear combination of the estimate obtained in Step B and the register figure.

We also consider a procedure to deal with not missing at random (NMAR) nonresponse in Step A. We illustrate the proposed procedures using data from the 2008 Census in Israel.

Key Words: Direct estimator; Fay Herriot model; Missing Information Principle; NMAR nonresponse

1. Introduction

In this article, we propose a new method of running a census, which combines a survey with big administrative data. We consider alternative ways of integrating the survey information with the administrative data for forming a single census estimate in small geographical areas, accounting for errors in both data sources and for not missing at random (NMAR) nonresponse. We illustrate our proposed method using data from the 2008 Census in Israel.

1.1 Description of last census in Israel (2008)

Israel has a fairly accurate population register; almost perfect at the country level. However, the population register is much less accurate for small statistical areas, with an average enumeration error of 13% and a 95 percentile of 40%. Israel is divided into about 3,000 statistical areas, and census information such as counts and socio-economic information is required for every area. The main reason for the inaccuracy in the register counts at the area level is that people moving in or out of areas, often report late their change of address. In 2008, the Israel Central Bureau of Census (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 (the "U sample"), and a telephone sample of people registered to the area for estimating the register over-count (the "O sample"). The U sample was also used for collecting the socio-economic information.

The final, census estimate has been computed as follows: Denote by N_i the true number of persons residing in area i on census day and by K_i the number of persons registered as living in the area. Let $p_{i,L|R}$ represent the proportion of persons living in area i among those registered as living in the area and $p_{i,\mathrm{R}|\mathrm{L}}$ represent the proportion of persons registered in area i among those living in the area. Then,

$$N_i \times p_{i,R|L} = K_i \times p_{i,L|R} \Rightarrow \hat{N}_i = K_i \times \frac{\hat{p}_{i,L|R}}{\hat{p}_{i,R|L}}. \quad (1)$$

[1]Danny Pfeffermann, National Statistician and Head of Central Bureau of Statistic, Jerusalem , Israel, Professor, Hebrew University of Jerusalem, Israel and University of Southampton, UK.

[2]Dano Ben-Hur, Central Bureau of Statistic, 66 Kanfey Nesharim street, Jerusalem , Israel, 9546456

By use of Taylor expansion, the conditional (design-based) variance of $\hat{N}_i$ can be approximated as,

$$\mathrm{Var}(\hat{N}_i \mid K_i) \cong K_i^2 \left[\frac{Var(\hat{p}_{i,L|K})}{[E(\hat{p}_{i,K|L})]^2} + \frac{[E(\hat{p}_{i,L|K})]^2}{[E(\hat{p}_{i,K|L})]^4} \times Var(\hat{p}_{i,K|L}) \right]. \quad (2)$$

The use of the U- and O samples with the resulting estimates looks very appealing but the actual implementation of the area U (field) samples was far from being straightforward. As major difficulties we mention:

1. The method requires listing all the apartments in each statistical area, or at least in a sample of cells or buildings in each area. This is a very costly operation and it requires additionally verifying that the apartments listed are dwelling units.
2. Coverage problems in any place where there are access restrictions such as closed floors, closed gates, etc.
3. Problems in locating sampled apartments when collecting the data, because not all the apartments are identified at the listing stage.
4. Response by internet is encouraged, but because of the above problems, it is not always clear, which household actually responded.
5. Many logistic problems in performing such a large scale field operation.

1.2 New method planned for the next census in Israel

In view of the difficulties with the 2008 census listed above, we plan a different method for our 2021 census (31 December, 2020 as census reference day). The census will combine information from a single sample taken from the population register, with information available from the register and other administrative files. The sample will collect information on residence of all members of the administrative household on census day, as well as socio-economic information. It is planned to obtain the information by the Internet, then by phone from people not responding via the internet, and in cases of nonresponse by either of the two modes, by personal interviews.

The direct estimates obtained from the sample will be improved by use of the Fay-Herriot (F-H) estimator, employing relevant covariate information known at the area level, such as the number of buildings and the total volume of all the buildings in the area, with the volume defined as the building roof area times its height. Other covariates will be used for estimating the area socio-economic means of interest.

For estimating the area counts, we shall combine the F-H estimator with the corresponding register count, to obtain our final, composite, census estimator (see below).

2. Proposed three-stage census estimator

2.1 Direct count estimate (Stage 1)

Denote by N the Number of residents in the country on census day and by N_i the Number of residents in area i, such that $N = \sum_i N_i$. Let $p_i = N_i / N$ – the true proportion of residents in the register living in area i, and denote by $\hat{p}_i$ the corresponding direct sample estimator, e.g., the sample proportion in the case of simple random sampling. (More efficient sampling designs and direct estimators are presently studied.) Let $K \cong N$ denote the size of the register on census day. The direct estimator for the count of area i is then $\hat{N}_i = K \times \hat{P}_i$. The variance is: $Var_D(\hat{N}_i \mid K) = K^2 Var_D(\hat{P}_i) = \sigma_{Di}^2$.

2.2 "Improved" Fay-Herriot estimates (Stage 2)

The (standard) Fay Herriot (F-H) (1979) model is:

$$\hat{N}_i = \alpha + \mathrm{x}_i'\beta + u_i + e_i, \quad (3)$$

where $\hat{N}_i$ is the direct sample estimator, x_i represents the area covariates (number of residential buildings in the area and total volume of all the residential buildings in our empirical illustrations; we are presently searching for more powerful covariates), u_i is a random effect and e_i is the sampling error of the direct estimator.

Under the model (3), the improved, empirical best linear predictor (EBLUP) of the true count is,

$$\hat{N}_{i,IMP} = \hat{\gamma}_i \hat{N}_i + (1-\hat{\gamma}_i)\mathrm{x}_i'\hat{\beta}; \quad \hat{\gamma}_i = \hat{\sigma}_u^2(\hat{\sigma}_u^2 + \hat{\sigma}_{D_i}^2)^{-1}, \tag{4}$$

where $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_{D_i}^2$ are appropriate sample estimates.

2.3 Final census count estimates

The final count estimate in area i, will be obtained as a weighted average of the improved F-H estimate in (4), and the population register count. For this, we assume $K_i \sim Possion(N_i) \Rightarrow Var(K_i) = N_i$. The final composite census estimator is thus,

$$\hat{N}_{i,COM} = \hat{\alpha}_i K_i + (1-\hat{\alpha}_i)\hat{N}_{i,IMP}; \ \hat{\alpha}_i = \frac{\hat{\sigma}_{i,FH}^2}{\hat{\sigma}_{i,FH}^2 + \hat{Var}(K_i)}. \tag{5}$$

3. Alternative estimation of census counts

3.1 Model extension

Rather than computing the composite estimator (5), include the register count as an additional covariate in the F-H model (3). Fitting this model "as is", implies conditioning on the known register count, ignoring its possible error.

3.2 Model extension, accounting for the errors of the register errors

Following Ybarra and Lohr (2008), we account for the measurement errors of the register counts by assuming,

$K_i \sim N(N_i, Var(K_i))$. Denote $\tilde{\mathrm{x}}_i = (\mathrm{x}_i', K_i)$. Assuming that all the other covariates are measured without error,

$$C_i = Var(\tilde{x}_i) = \begin{bmatrix} 0 \ldots 0, & \ldots, & 0 \\ 0 \ldots 0, & \ldots, & 0 \\ \ldots & , & \cdot \\ \ldots & , & \cdot \\ \ldots & , & \cdot \\ 0 \ldots 0, & \ldots, & V(K_i) \end{bmatrix}, \text{ and}$$

$$\hat{N}_{i,YL} = \hat{\delta}_i \hat{N}_i + (1-\hat{\delta}_i)\tilde{\mathrm{x}}_i'\hat{\beta}; \quad \hat{\delta}_i = \frac{\hat{\sigma}_u^2 + \hat{\beta}' C_i \hat{\beta}}{\hat{\sigma}_u^2 + \hat{\beta}' C_i \hat{\beta} + \hat{\sigma}_{Di}^2}. \tag{6}$$

4. Empirical illustrations

To illustrate the method, we use the Over-count (O) sample taken for the 2008 census. The total sample size is approximately 600,000 persons. We consider the 205 areas of sizes 1,000-10,000 as estimated in the 2008 census, because these area sizes correspond to the size of the statistical areas of interest. The sample has been drawn by stratified simple random sampling. The covariates used for the models are the number of residential buildings in the area and the total volume of all the residential buildings. The F-H model parameters have been estimated by MLE, using the PROC mixed procedure in SAS, assuming normality of the random effects and the sampling errors. The 2008 census estimates (based on the O and U samples) are taken as the true counts (referred to in the figures as the "Census values").

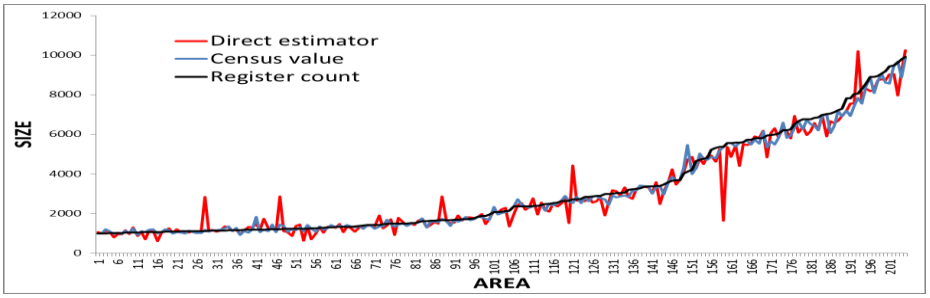

Figure 4-1. *Direct estimator, Census value and Register count for the 205 areas, ordered by their size in the register.*

As can be seen, the direct estimator is unbiased, but with large variance.

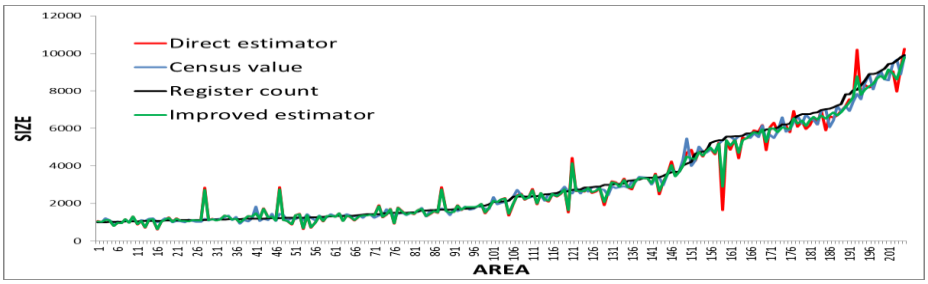

Figure 4-2. Direct *estimator, Census value, Register count and Improved* (F-H) *estimator*.

The improved F-H estimator reduces only slightly the variance of the direct estimator. We are presently searching for more powerful covariates.

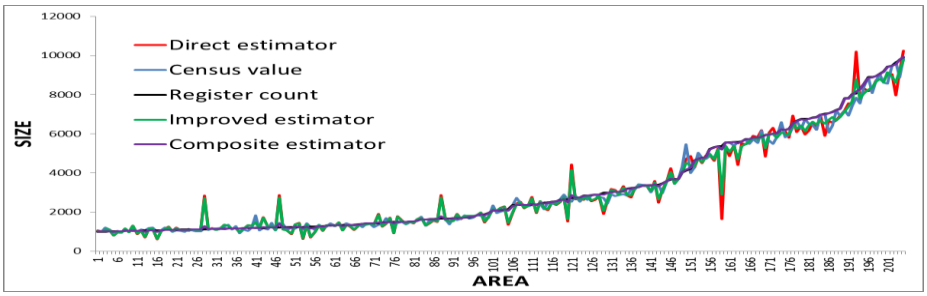

Figure 4-3 Direct *estimator, Census value, Register count, Improved estimator and Composite estimator*.

The Composite estimator is seen to estimate the true counts much more precisely than the other estimators. Table 4.1 exhibits some summary statistics of the performance of the various estimators considered so far.

Table 4-1 Absolute *relative distance of estimates from census values*

Estimate	Mean	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
Direct	0.1047	0.0101	0.0243	0.0556	0.1084	0.2202
Register count	0.0616	0.0010	0.0151	0.0507	0.0912	0.1344
Improved	0.0946	0.0112	0.0275	0.0573	0.0956	0.1959
Composite	0.0598	0.0056	0.0189	0.0469	0.0834	0.1257

Finally, Figure 4-4 and Table 4-2 exhibit he results obtained when adding the register count as an additional covariate in the F-H model, with (FH_WME) and without (FH_NME) accounting for its measurement error. In the latter case, we estimated σ_u^2 and β by the method of modified least squares (Ybarra and Lohr, 2008).

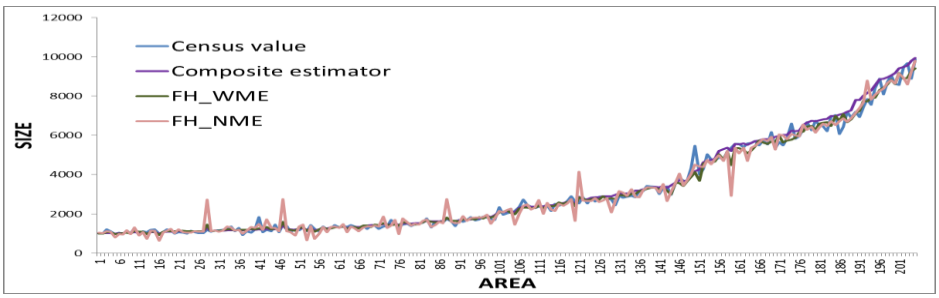

Figure 4-4 Estimates *when adding the register count to the covariates of the Fay-Herriot model, with and without accounting for its measurement error*.

Table 4-2 Absolute *relative distance of estimates from census values*

Estimate	Mean	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
Direct	0.1047	0.0101	0.0243	0.0556	0.1084	0.2202
Register count	0.0616	0.0010	0.0151	0.0507	0.0912	0.1344
Improved	0.0946	0.0112	0.0275	0.0573	0.0956	0.1959
FH_NME	0.0893	0.0100	0.0261	0.0540	0.0931	0.1877
Composite	0.0598	0.0056	0.0189	0.0469	0.0834	0.1257
FH_WME	0.0603	0.0094	0.0227	0.0498	0.0793	0.1230

As clearly seen, not accounting for the measurement error of the register count yields a census estimator with only minor improvement over the variable direct estimator. Accounting for the error of the register count improves the performance of the F-H estimator very significantly, but quite surprisingly, the composite estimator performs somewhat better, despite of the EBLUP property of the Ybarra and Lohr (2008) estimator. Although only based on a single empirical study, a possible explanation for this result is that in the latter estimator, the same weight is assigned to the register count and the other (fixed) covariates, whereas the composite estimator is more flexible, allowing for different weights for the register count and the other covariates. Further theoretical research and empirical illustrations are required to validate this result.

5. Accounting for Not Missing At Random (NMAR) nonresponse

Sverchkov and Pfeffermann (2018) propose a method that uses the Missing Information Principle of Orchard and Woodbury (1972) for estimating the response probabilities in small areas. The basic idea is as follows: first construct the likelihood that would be obtained if the missing outcome values were known also for the nonrespondents. However, since the missing outcomes are practically unknown, replace the likelihood by its expectation with respect to the distribution of the missing outcomes, given all the observed data. The latter distribution is obtained from the distribution of the observed outcomes, as fitted to the observed values. See Sverchkov and Pfeffermann (2018) for the relationship between the distributions of the observed- and the missing outcomes, for given covariates and response probabilities.

Ideally, we would want to show how the method performs in estimating the true number of persons residing in each area on census day, but this information is practically unknown for our test data (the O-sample used so far). Consequently, in what follows we illustrate instead the performance of the method when predicting the true number of divorced persons registered in each area. The O-sample is drawn from the population register and the true number of divorced persons registered in each area is known.

Define the outcome variable, y_{ij}, to be 1 if person j registered in area i is divorced, and 0 otherwise, and the response indicator, R_{ij}, to be 1, if unit j in area i responds and 0 otherwise. We restrict the analysis to persons aged 20+. The models fitted for the observed outcomes of responding units and for the response probabilities are defined in Equations (7) and (8). The covariates used for this illustration are listed in Table 5.1.

$$\Pr(y_{ij}=1\,|\,\mathrm{x}_{ij},u_i,R_{ij}=1)=\frac{\exp(\beta_0+\mathrm{x}'_{ij}\beta+u_i)}{1+\exp(\beta_0+\mathrm{x}'_{ij}\beta+u_i)};\quad u_i\sim N(0,\sigma_u^2), \tag{7}$$

$$\Pr(R_{ij}=1\,|\,y_{ij},\mathrm{x}_{ij},\mathrm{u}_i;\gamma)=\frac{\exp(\gamma_0+\mathrm{x}'_{ij}\gamma+\gamma_y y_{ij})}{1+\exp(\gamma_0+\mathrm{x}'_{ij}\gamma+\gamma_y y_{ij})}. \tag{8}$$

Clearly, for $\gamma_y\neq 0$, Equation (8) defines an informative response mechanism.

We first impose $\gamma_y = 0$, thus assuming that being divorced does not affect the probability of response, which corresponds to assuming missing at random (MAR) nonresponse. This is implemented by omitting the marriage status, y_{ij}, from the response model (8).

Table 5-1 *Odds ratios of estimated Logistic model of response probabilities assuming MAR nonresponse*

Variable	Odds ratio in case of MAR non-response
# of telephones per family	1.70
Administrative family size	1.15
Age 20-29	0.98
Age 30-39	0.87
40+	1.00
Jew	1.04
Other	1.00
Born in Israel	1.27
Other	1.00

As expected, the odds ratio for responding increases as the number of telephones belonging to the administrative family increases, and similarly for the administrative family size. The age group with the smallest response probability is 30-39 (odds ratio=0.87), and people born in Israel have a much higher odds ratio to respond than people born abroad. From this logistic regression we can estimate for each person the probability to respond.

Table 5-2 *Distribution of estimated response probabilities under the model in Table 5-1.*

Marriage status	Mean	5th Pctl	25th Pctl	75th Pctl
Other	0.815	0.489	0.822	0.885
Divorced	0.742	0.359	0.683	0.843
Total	0.812	0.487	0.819	0.885

Clearly, the assumption that $\gamma_y = 0$ is incorrect. The probability of responding among divorced persons is significantly lower than for other persons.

Next we estimate the response probabilities by including the binary variable "divorced" as an explanatory variable.

Table 5-3 *Odds ratios of estimated Logistic model of response probabilities allowing for NMAR nonresponse*

Variable	Odds ratio in case of MAR non-response	Odds ratio in case of NMAR non-response
# of telephones per family	1.70	1.83
Administrative family size	1.15	1.11
Age 20-29	0.98	0.95
Age 30-39	0.87	0.86
Other age	1.00	1.00
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**

As already implied by Table 5-3, the odds ratio for responding among divorced persons is about twice smaller than for other persons. Interestingly, the odds ratios of the other covariates are very similar to the odds ratios obtained when assuming MAR nonresponse.

Once the response probabilities have been estimated, they can be used for predicting the true area means of the target variable (proportions of divorced persons in the present illustration), using the approximately design-unbiased estimator,

$$\hat{\bar{Y}}_i^{HB} = \sum_{j,(i,j)\in R} (y_{ij} / \tilde{\pi}_{j|i}) / \sum_{j,(i,j)\in R} (1/\tilde{\pi}_{j|i});\ \tilde{\pi}_{j|i} = \pi_{j|i}\,\hat{p}_r(y_{ij}, x_{ij}; \hat{\gamma}), \quad (9)$$

where $\pi_{j|i}$ denotes the sampling probability. Sverchkov and Pfeffermann (2018) derive also the empirical best predictor under the models (7) and (8), but we don't consider this predictor in the present paper.

Figure 5-1 and Tables 5-2 and 5-3 compare the performance of the following three predictors of the true proportions of divorced persons in the various areas: The proportion of divorced persons in the observed sample, ignoring the non-response (hereafter the direct estimator), the estimator obtained when assuming MAR nonresponse, and the estimator obtained when allowing for NMAR nonresponse (Equation 8).

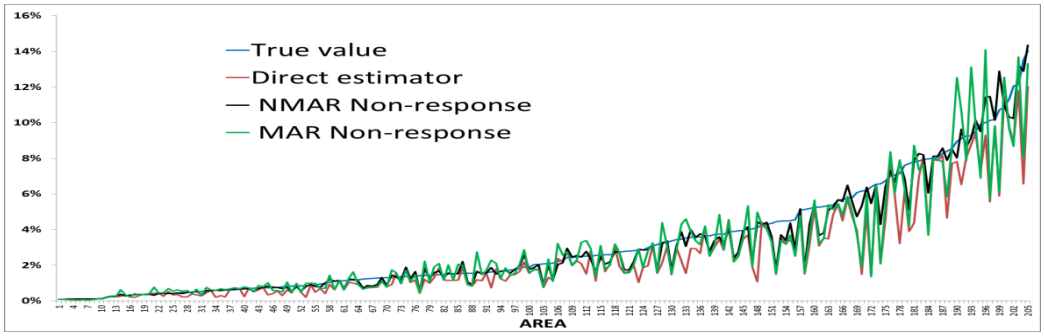

Figure 5-1 *Percent of divorced persons in areas*: *true value, direct estimator, and estimators obtained when assuming MAR and NMAR non-response.*

Table 5-4 *Difference between true values and estimates (BIAS) over all the areas*

Estimator	*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

Table 5-5 *Absolute relative distance of estimates from true values*

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

As clearly indicated by Figure 5.1 and Tables 5-4 and 5-5, the estimates obtained when accounting for NMAR nonresponse have by far, the smallest bias and the smallest absolute relative distance from the true values. The direct estimates, which ignore the non-response have large bias and large relative distance from the true values.

6. Concluding Remarks

In this article we consider a new method for running a census, combining sample estimates with big administrative data. A major advantage of this method is that it does not require the use of personal interviews, except in the case of nonrespondents. Israel still does not have a sufficiently reliable housing register, and the use of a field sample requires prior listing of all the dwelling apartments in a sample of cells in each statistical area, which is rather complicated logistically and very expensive. It also requires verifying that each of the apartments is a dwelling unit.

Under the new method, a single sample of persons is drawn from the register, which is known to be generally accurate at the national level, except for some small "outlying" sub populations, such as Bedouins or illegal immigrants. We consider alternative ways of combining the survey information with the population register to form a single final census estimator, accounting for the sampling errors in the survey, and address errors in the register. We also propose a simple descriptive procedure of testing the informativeness of the missing sample data, and a way of accounting for NMAR nonresponse. We illustrate all the above topics by use of real empirical data.

We are currently planning a "pilot census" for next year in two statistical regions of Israel, which will hopefully provide us another opportunity to test the ideas discussed in the present article, with more up-to-date data.

References

Fay, R. E. and Herriot, R. A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 697-715.

Sverchkov, M., and Pfeffermann, D. (2018), "Small area estimation under informative sampling and not missing at random non-response". *Journal of the royal statistic society*, Series A, 181, 981-1008.

Ybarra, L.M.R., and Lohr, S.L. (2008), "Small area estimation when auxiliary is measured with error", *Biometrika*, 95, 919-931.