United Nations ECE/CES/GE.41/2019/15

Economic and Social Council

Distr.: General
10 July 2019

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twenty-first Meeting
Geneva, 18–20 September 2019
Item 2 of the provisional agenda
Results of tests with regard to methodology, technology, participation, and other aspects

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

Note by Israel Central Bureau of Statistics*

Summary

As 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 domains, with an average domain enumeration error of about 13 per cent. The main reason for the inaccuracy at the domain level is that people moving in or out an area are often late in reporting their change of address. This document discusses means for correcting the estimates obtained from the register for small domains by mean of a survey taken from the register. In particular, it addresses the following questions:

(a) How to best draw the sample?

(b) How to combine the sample information with the register data?

(c) How to correct for non-response that is not missing at random? and

(d) How to assess the root mean square error of the resulting census estimators?

Empirical illustrations based on data from the 2008 census are presented.

* Prepared by Danny Pfeffermann, Dan Ben-Hur and Olivia Blum
(See also Pfeffermann et al, Statistics in Transition (SiT), Vol. 20, No. 1, Polish Statistical Association (PTS) and Statistics Poland (GUS)).

I. Introduction

1. This document proposes a new method of running a census, which combines a survey with administrative data. It considers alternative ways of integrating the survey information with administrative data in order to form a single census estimate in small geographical areas, accounting for errors in both data sources and for not missing at random (NMAR) nonresponse. The proposed method is illustrated using data from the 2008 census in Israel.

A. Description of last census in Israel (2008)

2. Israel has a fairly accurate Central Population Register (CPR); almost perfect at the country level. However, the CPR is much less accurate for small statistical areas, with an average enumeration error of 13 per cent and a 95th percentile of 40 per cent. 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 their change of address late, while others who have an interest in maintaining a particular address (e.g. tax benefits, school area, parking, etc.) do not report their change of address as long as the interest persists. 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 people living in the area on census day for estimating the register undercount (the "U sample"), and a sample of people registered in the same area for estimating the register over-count (the "O sample"). The U sample was also used for collecting socio-economic information.

3. 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 $p_{i,\mathrm{R|L}}$ represent the proportion of persons registered in area i among those living in the area. Then,

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

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

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

5. Over the last decade, Israel, as many other countries, has experienced an accelerated process of using administrative data for the production of official statistics in general, and in particular, it has improved its abilities to use administrative data for census purposes. As a result, the 2020 census methodology in Israel will use new data sources, and for the first time a geo-demographic administrative file (GDAF) will serve as the sampling frame for a sample that will be used to correct the administrative data for small statistical areas.

6. Two key facts enable the shift in the planned methodology: a) entries and departures to and from the country are well recorded; b) all people in the country have administrative records; citizens are registered in the CPR and the foreigners are reported in functional records such as work permits and visas. A conceptual and practical leap towards fully administrative censuses in the future can be envisaged, although unfortunately not yet for the next census of Israel in 2021, for which reference census day is defined as 31/12/2020.

B. New method planned for the next census in Israel

7. For the 2021 census a different method is planned, which will hopefully bring ICBS closer to the use of fully administrative censuses in the future. The census will combine information from a single sample taken from the GDAF, with information available from the register and other administrative files, mainly to correct the counts obtained from the GDAF. The sample will collect geo-demographic information on all members of households on census day, as well as socio-economic information. It is planned to obtain the information via the Internet, then by telephone from people not responding via the Internet, and, in cases of nonresponse by either of these two modes, by personal interviews.

8. The direct estimates obtained from the sample will be improved by the 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 multiplied by its height. Other covariates will be used for estimating the area socio-economic means of interest.

9. For estimating the area counts, the F-H estimator will be combined with the corresponding GDAF count, to obtain the final, composite, census estimator (see below).

II. Proposed three-stage census estimator

A. Direct count estimate (Stage 1)

10. 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 $p_i = N_i / N$ denote the true proportion of residents in the GDAF living in area i, and $\hat{p}_i$ denote 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 currently being studied). Finally, denote by $K \cong N$ the size of the GDAF on census day. The direct estimator for the count of area i is then $\hat{N}_i = K \times \hat{P}_i$. The conditional design-based variance of $\hat{N}_i$ is $Var_D(\hat{N}_i \mid K) = K^2 Var_D(\hat{P}_i) = \sigma^2_{Di}$

B. "Improved" Fay-Herriot estimate (Stage 2)

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

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

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 the empirical illustrations; more powerful covariates are currently being sought), u_i is a random effect and e_i is the sampling error of the direct estimator.

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

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

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

C. Final census count estimates

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

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

III. Alternative estimation of census counts

A. Including the register count among the covariates as fixed numbers

14. Rather than computing the composite estimator (5), include the GDAF 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. The final census count estimate is in this case the F-H estimate.

B. Accounting for the errors of the register errors

15. Following Ybarra and Lohr (2008), the GDAF count is added to the set of covariates but its possible measurement error is accounted for 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

$$C_i = Var(\tilde{\mathrm{x}}_i) = \begin{bmatrix} \mathbf{O}\ldots\mathbf{O}, & \ldots, & \mathbf{O} \\ \mathbf{O}\ldots\mathbf{O}, & \ldots, & \mathbf{O} \\ \cdot\ \cdot\ \cdot\ \cdot & , & \cdot \\ \cdot\ \cdot\ \cdot\ \cdot & , & \cdot \\ \cdot\ \cdot\ \cdot\ \cdot & , & \cdot \\ \mathbf{O}\ldots\mathbf{O}, & \ldots, & \boldsymbol{V}(\boldsymbol{K}_i) \end{bmatrix}, \text{ and}$$

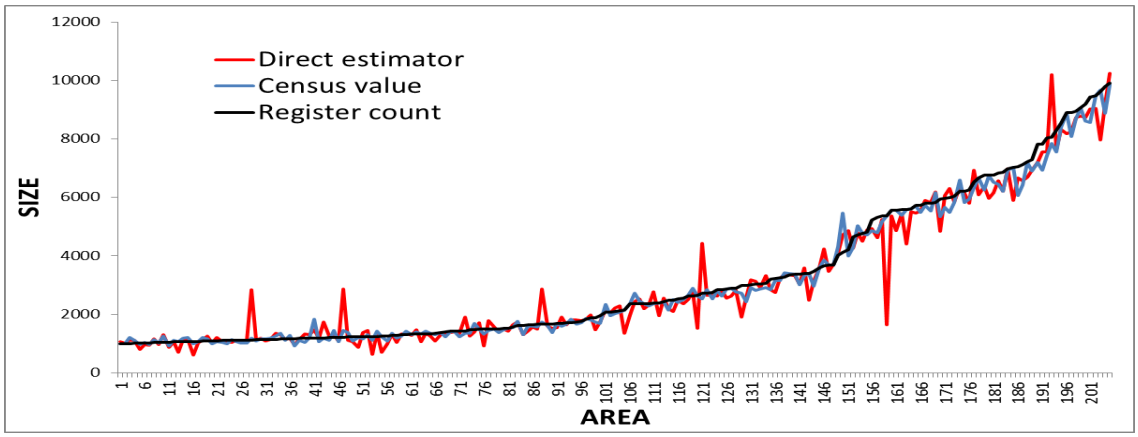
$$\hat{N}_{i,YL} = \hat{\delta}_i \hat{N}_i + (1-\hat{\delta}_i)\tilde{\mathrm{x}}'_i\hat{\beta};\quad \hat{\delta}_{\mathrm{i}} = \frac{\hat{\sigma}^2_u + \hat{\beta}' C_i \hat{\beta}}{\hat{\sigma}^2_u + \hat{\beta}' C_i \hat{\beta} + \hat{\sigma}^2_{Di}}. \quad (6)$$

IV. Empirical illustrations

16. To illustrate the method and its various options, the Over-count (O) sample is drawn from the central population register for the 2008 census. The total sample size is approximately 600,000 persons. The 205 areas of size 1,000-10,000 as estimated in the 2008 census are considered, because these area sizes correspond to the sizes of the statistical areas of interest in the next census. The sample was 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 maximum likelihood estimation, using the "PROC mixed" procedure in the software program SAS, which assumes normality of the random effects and the sampling errors. The 2008 census estimates (based on the O and U the samples) are taken as the true counts (referred to in the figures below as the "census values").

Figure I

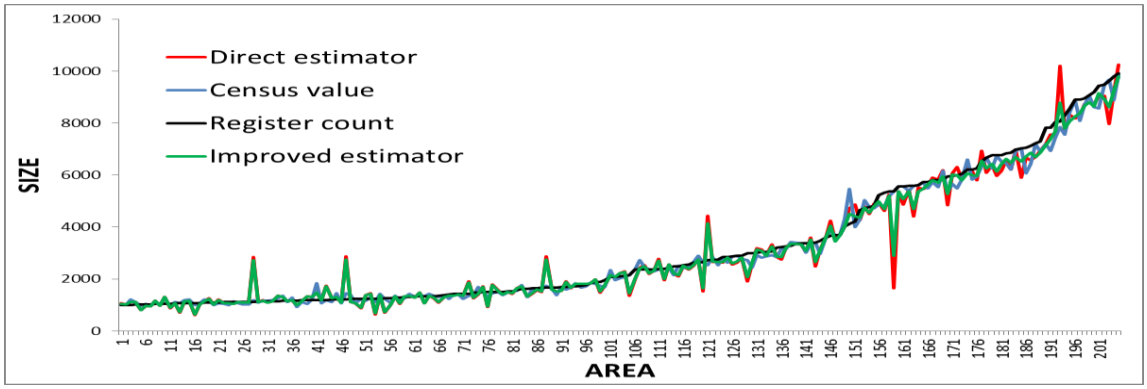
Direct estimator, Census value and Register count for the 205 small areas, ordered by their size in the register



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

Figure II

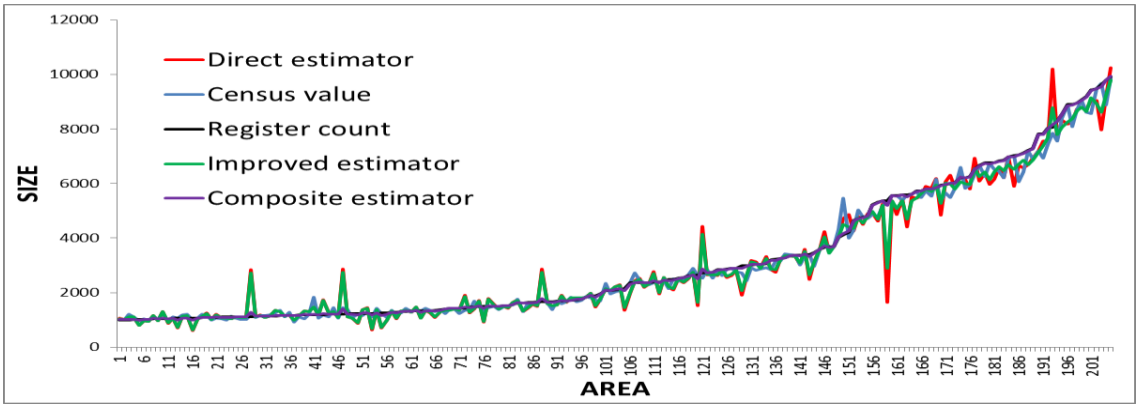
Direct estimator, Census value, Register count and Improved (F-H) estimator



18. The improved F-H estimator reduces only mildly the variance of the direct estimator. ICBS is in the process of searching for more powerful covariates. In particular, it is expected that a list will be obtained from the electricity company listing all dwelling apartments and houses in Israel, which should improve the F-H estimator very significantly compared to the use of only the number of buildings.

Figure III

Direct estimator, Census value, Register count, Improved estimator and Composite estimator



19. The Composite estimator is seen to estimate the true counts much more precisely than the other estimators. Table 1 exhibits some summary statistics of the performance of the various estimators considered so far.

Table 1
Absolute relative distance of estimates from census values I

Estimate	*Mean*	*10th per centile*	*25th per centile*	*50th per centile*	*75th per centile*	*90th per centile*
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

20. Finally, Figure IV and Table 2 exhibit the 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, σ_u^2 and β are estimated by the method of modified least squares (Ybarra and Lohr, 2008).

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

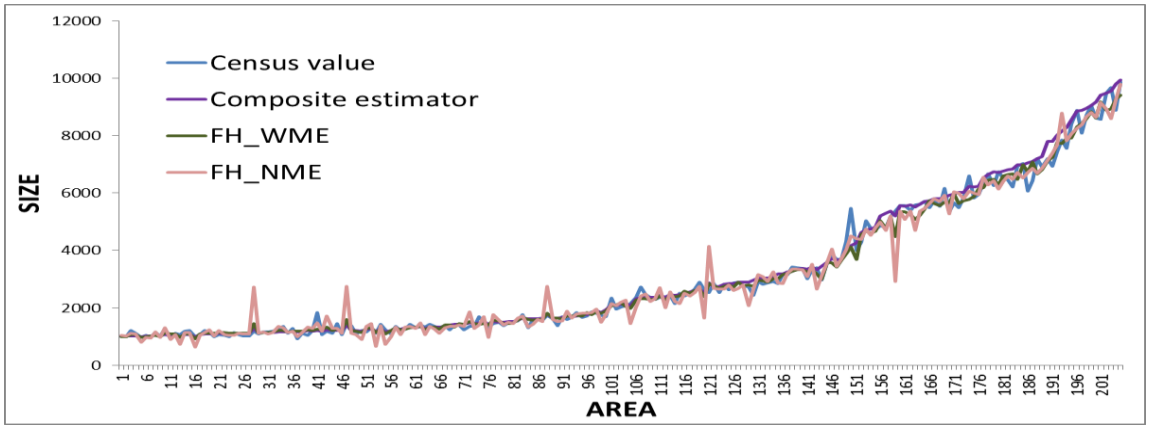


Table 2
Absolute relative distance of estimates from census values II

Estimate	*Mean*	*10th per centile*	*25th per centile*	*50th per centile*	*75th per centile*	*90th per centile*
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

21. As can be seen clearly, not accounting for the measurement error of the register count yields a census estimator with only minor improvement over the direct sample 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 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.

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

22. 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 also known 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.

23. It would be ideal 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 the test data (the O-sample used so far). Consequently, what follows illustrates 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 central population register and the true number of divorced persons registered in each area is therefore known.

24. Define the outcome variable, y_{ij}, to be 1 if person j registered in area i is divorced, and 0 otherwise. Let the response indicator, R_{ij}, take the value 1 if unit j in area i responds, and 0 otherwise. The analysis is restricted to persons aged 20 years and above. The model fitted for the observed outcomes of the responding units and the model assumed for the response probabilities are defined in Equations (7) and (8). The covariates used for this illustration are listed in Table 3.

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

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

25. Clearly, for $\gamma_y \neq 0$, Equation (8) defines an informative response mechanism. First $\gamma_y = 0$ is imposed, thus presuming 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). The results are shown in Tables 3 and 4.

26. Table 3 displays the odds ratios of the estimated logistic model of the response probabilities for this case. 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.

Table 3
Odds ratios of estimated logistic model of response probabilities assuming MAR nonresponse

Variable	*Odds ratio in case of MAR non-response*
Number 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

27. Table 4 shows the distribution of the estimated response probabilities under the model of Table 3.

Table 4
Distribution of estimated response probabilities under the model exhibited in Table 3

Marital status	*Mean*	*5th per centile*	*25th per centile*	*75th per centile*
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

28. It is quite clear from Table 4 that the supposition $\gamma_y = 0$ is incorrect. The probability of responding among divorced persons is significantly lower than for other persons. Hence, it the response probabilities were estimated by including the binary variable "divorced" as an additional explanatory variable.

Table 5
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*
Number 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

29. It can be seen in Table 5 that the odds ratio for responding among divorced persons is about twice as small as for other persons, in correspondence with the results in Table 6. Interestingly, the odds ratios of the other covariates are very similar to the odds ratios obtained when assuming MAR nonresponse.

30. The estimated models in Tables 3 and 5 permit estimation of the response probability for each responding person in the sample. By viewing the response as an additional stage of

sampling, the estimated response probabilities will be used for predicting the true area means of the target variable (proportion of divorced persons in the present illustration) using standard sampling theory, for example, by employing the approximately design-unbiased estimator,

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

where $\pi_{j|i}$ denotes the sampling probability. Sverchkov and Pfeffermann (2018) also derive the empirical best predictor under models (7) and (8), but this predictor is not considered in the present document.

31. Figure V and Tables 6 and 7 compare the performance of the following three predictors of the true proportion 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 V

Per cent of divorced persons in areas: true value, direct estimator and estimators obtained when assuming MAR and NMAR non-response

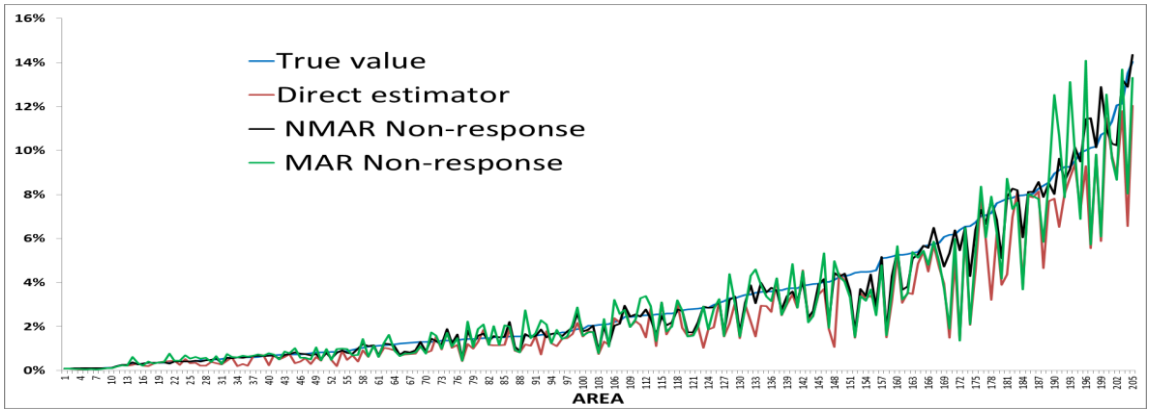


Table 6

Difference between true values and estimates over all the areas

Estimator	*Mean*	*10th per centile*	*25th per centile*	*50th per centile*	*75th per centile*	*90th per centile*
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 7

Absolute relative distance of estimates from true values

Estimator	*Mean*	*10th per centile*	*25th per centile*	*50th per centile*	*75th per centile*	*90th per centile*
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

32. As indicated by Figure V and Tables 6 and 7, 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 nonresponse, have large bias and large relative distance from the true values.

VI. Root MSE estimation under NMAR nonresponse

33. As in any publication of official statistics, there is a requirement to publish counts and other socio-demographic estimates, but also to evaluate their precision. One set of evaluation measures (but definitely not the only one), is to estimate for each area the root mean square error (RMSE) of the final estimate. This is relatively simple if the direct sample estimates were to be used as the final estimates and if all sample units would respond, but this obviously will not happen. Estimation of the RMSE is more complicated when using the Fay-Herriot estimates and accounting for NMAR nonresponse, and even more so, when using the composite estimator described in section II.

34. Sverchkov and Pfeffermann (2018) propose a bootstrap method for estimation of the RMSE of small area estimates under NMAR nonresponse, which accounts for the random processes assumed to generate the population values, and the sampling and response processes. This implies that the target area parameters (the true proportion of persons residing in the area on census day, out of all the persons registered in the CPR in the application), are considered as random, which is different from classical survey sampling applications under which the finite population values and hence the target parameters are viewed as fixed values. Users of sample survey (official statistics) estimates are familiar with measures of error, which only account for the variability originating from the randomness of the sample selection (known as the randomization distribution), and the nonresponse. In other words, users are accustomed to estimates of the design-based (randomization) MSE (denoted hereafter as DMSE), over all possible sample selections, with the population values of the survey variables (and hence the values of the target parameters) held fixed. Estimation and publication of the DMSE (or its square root) is a common routine in national statistical offices all over the world.

35. In a recent article, Pfeffermann and Ben-Hur (2019) propose a new procedure for estimating the DMSE of model-based small area predictors, which is shown to perform well in an extensive simulation study and to outperform other procedures for DMSE estimation proposed in the literature. This procedure is currently being extended for estimating the DMSE of the proposed composite estimators, accounting, in particular, for the inevitable NMAR nonresponse and the use of the composite estimator, which combines the survey estimate with the administrative population register count.

VII. Concluding Remarks

36. In this article a new method for running a census is considered, combining sample estimates with administrative data. A major advantage of this method is that it does not require the use of personal interviews, except in the case of non-respondents. Israel still does not have a sufficiently reliable dwelling register, and the use of a field sample requires prior listing of all the dwelling units 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 inhabited.

37. Under the new method, a single sample of persons is drawn from the GDAF, which is known to be generally accurate at the national level, except for some small “outlying” sub-populations, such as illegal immigrants. Alternative ways are considered of combining the survey information with the GDAF to form a single final census estimator, accounting for the sampling errors in the survey, and errors in the addresses in the GDAF. A simple descriptive procedure of testing the informativeness of the missing sample data is also proposed, and a way of accounting for NMAR nonresponse. All the above topics are illustrated by the use of real empirical data.

38. A census rehearsal is currently being planned for next year in two statistical regions of Israel, which will hopefully provide ICBS with another opportunity to test the ideas discussed in the present article, with more up-to-date data.

39. Leaning more about administrative sources of information opens the way for new opportunities for the census process and outcomes. Referring to an identified population in a known population frame implies a substantial change in the concept of a census. Area

boundaries become, in a way, a virtual entity rather than the main physical entity in a census. The theoretical and socio-economic implications of this change, and the influence on policymaking, should be further investigated.

References

Blum, O. (2018). From physical area to virtual lists: Toward an administrative census in Israel. UNECE group of experts on population and housing censuses. Geneva.

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.

Pfeffermann, D. and Ben-Hur, D. (2018). Estimation of Randomization Mean Square Error in Small Area Estimation. International Statistical Review, 0, 0, 1–19 doi:10.1111/insr.12289.

Sverchkov, M., and Pfeffermann, D. (2018), "Small area estimation under informative sampling and not missing at random non-response". Journal of the Royal Statistical 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.