

MASS IMPUTATION FOR TWO PHASE SAMPLING: USE OF SMALL AREA ESTIMATION AND CALIBRATION TECHNIQUES

H.J. Mantel, A.C. Singh and M. Yu¹

ABSTRACT

We consider the problem of imputing study variables (y -variables) to a first phase sample by mass imputation from a regression model based on a second phase sample. This problem arises in the redesigned Survey of Employment, Payroll and Hours (SEPH) where in the first phase x -variables are obtained from an administrative source and in the second phase y -variables are collected directly from a sub-sample of establishments. It is convenient to produce estimates for various domains from the completed first phase sample which would be quite large; however, it is desirable that these estimates be close to those obtained by the usual small area techniques, at least for a selected set of domains. This can be achieved by calibrating the initial regression imputed y -values so that the expansion estimates for the selected domains match the desired small area estimates. Here calibration is used in an unusual manner in that the y -values are being calibrated rather than the sampling weights. The approach will be empirically evaluated and compared to alternatives using SEPH data.

RÉSUMÉ

Les auteurs étudient le problème de l'imputation des variables étudiées (variables y) à un échantillon de première phase par imputation massive à partir d'un modèle de régression basé sur un échantillon de seconde phase. Ce problème se pose dans le cas de l'enquête remaniée sur l'emploi, la rémunération et les heures de travail (EERH). Dans le cadre de celle-ci, on obtient, à la première phase, les variables x d'une source administrative, et on collecte, à la seconde phase, les variables y auprès d'un sous-échantillon d'établissements. Il est commode de produire des estimations pour divers domaines à partir de l'échantillon de première phase complet de grande taille; cependant, il est souhaitable que ces estimations, et celles obtenues par les techniques habituelles d'évaluation des petites régions, ne s'écartent pas trop, au moins pour un ensemble choisi de domaines. On peut y arriver en étalonnant les valeurs initiales de y , imputées par régression, de sorte que les estimations obtenues par expansion pour les domaines choisis coïncident avec les estimations désirées relatives aux petites régions. L'étalonnage est ici utilisé de façon non conventionnelle, en ce sens que ce sont les valeurs de y que sont étalonnées, et non les poids de l'échantillon. La méthode sera évaluée empiriquement et comparée à d'autres; on utilisera pour cela les données de l'EERH.

1. INTRODUCTION

The research described in this paper is motivated by a problem of estimation for Statistics Canada's Survey of Employment, Payroll and Hours (SEPH). SEPH is a survey of business establishments which aims to produce estimates of total employment, total payroll and other variables for industry groups by province and for the whole of Canada.

Recently two of the key SEPH variables have become available on an administrative source. Every business in Canada is required to remit to Revenue Canada monthly deductions of income tax from employees' pay, together with accompanying PD7 forms which now contain information on the total number of employees and the total payroll associated with the remittance.

In order for these data on employment and payroll to be used to meet the objectives of SEPH, they must first be associated to particular industrial classifications and locations. This is a relatively simple matter for data from businesses of simple structure, with a single location and

a single industrial classification; however, the problem is much more difficult for more complexly configured businesses. Consequently, the frame for SEPH has been divided into two portions: an establishment portion covering units which are large or complexly structured, and an administrative portion covering smaller units of simple structure. Development work to take advantage of the payroll deduction (PD) data for the establishment portion is ongoing; more details are given in Hidiroglou, Latouche, Armstrong and Gossen (1995). The focus of the present paper is estimation for the administrative portion of the frame.

The objectives in this paper are to propose some reasonable alternatives to the current estimation method for the administrative portion of SEPH and to empirically evaluate the performance of the current method relative to those alternatives. The organization of the paper is as follows. We first describe, in Section 2, the two phase sampling design used for the administrative portion of SEPH, and the current estimation strategy which involves mass imputation of secondary SEPH variables, such as

¹ H.J. Mantel, A.C. Singh and M. Yu, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

hours, to first phase sample units. In Section 3 we describe how some ideas from small area estimation, namely sample size dependant (SSD) estimation and generalized SSD estimation, might be used to improve the mass imputation. The use of calibration to further improve the imputations is described in Section 4. Section 5 describes an empirical comparison of these alternatives to the current SEPH estimation strategy. Concluding remarks are given in Section 6.

2. CURRENT SEPH DESIGN AND ESTIMATION

The sample design for the administrative portion of SEPH is a two phase design. In the first phase a very large simple random sample of PD accounts is drawn from the administrative source. The reason for sampling the administrative data is that the files are too large to be handled in their entirety, and the decrease in precision due to sampling rather than taking a census is not of practical concern.

For the second phase of sampling the administrative portion of the SEPH frame has been partitioned into model groups which are predefined collections of 1980 Standard Industrial Classification (SIC) groups with similar regressions of the secondary SEPH variables onto employment and payroll. The model groups are usually collections of SIC2 (two digit) groups within provinces, though some are collections of SIC3 groups or cross province boundaries. The PD accounts selected in the first phase are matched to their corresponding units on the Business Register (a comprehensive list frame of Canadian businesses used in many of Statistics Canada's business surveys). The second phase sample is then a stratified simple random sample of the first phase units from the Business Register, with strata corresponding to model groups. The minimum second phase sample size in a model group is 40, though some model groups have a much larger sample size. All of the SEPH variables are then collected directly from the establishments in the second phase sample.

Within model groups secondary SEPH variables are regressed onto employment and payroll using a weighted regression; for example, the weights used for the hours regressions are the inverses of employment. These regressions are then used to impute the secondary variables to the first phase sample. Estimates for various domains are then obtained as expansion estimates using the imputed data and the first phase sampling weights.

Two estimation issues of concern are considered in proposing alternatives to the current SEPH estimation method. First, there was some concern that the sizes of the samples for estimating the regression coefficients in some of the model groups are quite small, and that the use of small area estimation techniques might be helpful.

Second, the model groups would generally not be of analytical interest; rather, estimates for industry groupings which are smaller than model groups or which cross model groups would be desired. However, estimators of secondary SEPH variables for these industry groupings would be essentially synthetic, and therefore biased, in that the regression which applies within a model group need not hold within a subdomain of the model group. This issue is partially addressed by calibrating the imputed data within predefined calibration groups (which are generally smaller than the model groups) to efficient estimators for the calibration group totals.

3. GENERALIZED SAMPLE SIZE DEPENDENT ESTIMATION

In this section we describe generalized sample size dependent (SSD) estimation which was first proposed in Singh and Mian (1995). We first describe the original SSD estimator.

The basic idea of SSD estimation for a domain is that if the sample within the domain is too small then we will borrow strength from a larger domain containing the domain of interest by taking a convex combination of the direct estimator for the domain with a synthetic estimator based on the data from the larger domain; however, if the sample in the domain is large enough then we would use just the direct estimator. This idea was introduced by Drew, Singh and Choudhry (1982). The original approach to decide whether or not the sample within the domain was too small was based on the ratio of the observed sample size to the expected. This approach was questioned by Kalton (1994) since the expected sample size within domains of interest may vary considerably. Comparison of the realized sample size to a fixed cutoff may be more reasonable. In our empirical study later we consider a version of SSD suggested by Singh, Gambino and Mantel (1994) in response to Kalton's comment. For imputation in model group g we then have the estimated regression coefficient

$$\hat{\beta}_{SSD,g} = \left\{ \lambda n_g^{-1} X_g^T W_g X_g + (1-\lambda) n^{-1} X^T W X \right\} \times \left\{ \lambda n_g^{-1} X_g^T W_g y_g + (1-\lambda) n^{-1} X^T W y \right\} \quad (1)$$

$$\lambda = \begin{cases} 1 & n_g \geq n_{\min} \\ n_g/n_{\min} & n_g < n_{\min} \end{cases}$$

where X_g , X are the second phase sample matrices of regressors (employment and payroll) in the domain of interest and in the larger domain, respectively, W_g , W are diagonal regression weight matrices, y_g , y are vectors of secondary SEPH variates (e.g., hours), n_g is the observed

second phase sample size in model group g , and n is the total second phase sample size.

For the generalized SSD estimator we construct, for each model group g , distance weights $d_{gg'}$ for borrowing from other model groups g' , where $0 \leq d_{gg'} < 1$ if $g \neq g'$ and we set $d_{gg} = 1$. These distance weights imply a preference ordering for borrowing data with model group g' being preferred to g'' as a donor for g if $d_{gg'} > d_{gg''}$. Then if $n_g < n_{\min}$ we would borrow from the model groups g^j for which d_{gg^j} is largest, borrowing just enough to achieve the target n_{\min} ; furthermore, the distance weights from the last donor group needed to achieve n_{\min} , say $g'' \neq g$, would be discounted by a factor $n_{\text{need}}/n_{g''}$ where n_{need} was the number of units still needed to attain n_{\min} . We then have

$$\hat{\beta}_{\text{GSSD},g} = \{X_{Bg}^T W_{Bg} D_{Bg} X_{Bg}\}^{-1} X_{Bg}^T W_{Bg} D_{Bg} y_{Bg} \quad (2)$$

where the subscript Bg denotes the data from model group g together with that from all of the groups required to achieve the minimum sample size, and D_{Bg} is the diagonal matrix of distance weights.

There still remains the problem of defining the distance weights $d_{gg'}$. For our empirical study we used historical data to look at the differences $||\beta_{\xi} - \beta_{\xi'}||$ (after suitable normalization of the covariates). The weight $d_{gg'}$ would be a decreasing function of these differences, truncated at zero. If these weights are stable over time then the approach can be expected to work well.

4. CALIBRATION

As noted above, SEPH produces estimates for domains, such as SIC2 by province or SIC3 by Canada, which do not generally correspond to model groups; so most of the estimators for secondary SEPH variables are synthetic and therefore biased to some extent. This issue may be partially addressed by forming calibration groups and using small area techniques to develop good estimators at the calibration group level, and then calibrating the imputed data within calibration groups to those estimators. The calibration groups would be chosen to be generally smaller than the model groups but to contain similar types of industries. This approach has two advantages. First, at the calibration group level we can use the best estimation techniques we have. Secondly, since the calibration groups are generally smaller than the model groups, so that sub-domains of a calibration group would tend to be more like the calibration group than the corresponding model group, estimates for sub-domains of the calibration groups would also often be improved.

The calibration here is different from the traditional calibration (Deville and Särndal, 1992) in that it is the

imputed data that are being adjusted, rather than the estimation weights. The data on the primary SEPH variables, employment and payroll, from the first phase sample are considered to be of very high quality and we did not want to distort their distribution in any way. Adjustment of the imputed data, which is highly model dependent, is much more acceptable.

For simplicity, in the empirical study described later we use a sample size dependent combination of generalized regression (GREG) and regression synthetic estimators as the benchmark estimator within calibration groups; however, the calibration groups were chosen to be large enough, in terms of expected sample size, that the GREG estimator would usually be acceptable and little or no weight would be given to the synthetic part. The general form of the estimator for calibration group a is then

$$\hat{y}_{\text{SSD},a} = \sum_{\xi} x_{a\xi}^T \hat{\beta}_{\xi} + (\lambda N_a / \hat{N}_{\text{EXP},a}) \sum_{\xi} (\hat{y}_{\text{EXP},a\xi} - \hat{x}_{\text{EXP},a\xi}^T \hat{\beta}_{\xi}) \quad (3)$$

$$\lambda = \begin{cases} 1 & n_a \geq n_{\min} \\ n_a / n_{\min} & n_a < n_{\min} \end{cases}$$

where $x_{a\xi}$ is the population total of x in calibration group a and model group g , N_a is the population total number of business establishments in calibration group a , $\hat{N}_{\text{EXP},a}$ is the expansion estimator of N_a based on the second phase sample in calibration group a , $\hat{y}_{\text{EXP},a\xi}$ and $\hat{x}_{\text{EXP},a\xi}$ are the expansion estimators of the population totals of y and x , respectively, in calibration group a and model group g , and $\hat{\beta}_{\xi}$ may be any of the alternative estimators of β_{ξ} . If $x_{a\xi}$ or N_a are not known then they can be estimated by their expansion estimators based on the first phase sample. Note that n_{\min} in (3) need not be the same as n_{\min} in (1) or the n_{\min} used for the generalized SSD estimator in (2).

The SSD estimators for the calibration groups are themselves first calibrated to the unbiased regression estimator of the provincial total, i.e., $\hat{y}_{\text{REG}} = \sum_{\xi} x_{\xi} \hat{\beta}_{\xi}$. Calibration of this sort was previously considered by Mantel, Singh and Bureau (1993). In this calibration we want to take account of the fact that the SSD estimators for different calibration areas have different accuracies. To this end we make a working assumption that the MSEs of $\hat{y}_{\text{SSD},a}$ are proportional to $\nu_a = (1/n_a - 1/N_a)$ and the calibrated estimator is then given by

$$\hat{y}_{\text{CAL},a} = \hat{y}_{\text{SSD},a} + (\nu_a / \sum_i \nu_i) (\hat{y}_{\text{REG}} - \sum_i \hat{y}_{\text{SSD},i}). \quad (4)$$

Finally, the imputed y -values within calibration group a are simply ratio calibrated to the estimated group total $\hat{y}_{\text{CAL},a}$. This may be loosely rationalized by making a working assumption that the MSEs of the imputed y -values, as estimators of the true y -values, are proportional to the imputed y -values.

5. EMPIRICAL STUDY

The estimation methods described in Sections 2, 3 and 4 were compared empirically by means of a Monte Carlo simulation of SEPH sampling from a pseudo-population constructed from SEPH survey data for the province of Ontario. This pseudo-population contains records for 1379 distinct units from 26 model groups and 68 SIC2 groups. Each record also has a replicate weight, which corresponds to the estimation weight from the SEPH data used to construct the pseudo-population, so that the pseudo-population is representative of the Ontario population of business establishments covered by the administrative portion of the SEPH frame. The sum of the replicate weights is 96100. The variables included for each record are employment, payroll and hours for six consecutive months, as well as an SIC3 code from which the model group and calibration group can be derived. Simple random sampling without replacement was applied within each model group, taking account of the replicate weight for each distinct unit. The sample sizes within model groups were chosen to match the second stage sample sizes from the current SEPH survey; thus the total sample size was 2537 and the sample size within model groups varied from 40 to 335. For simplicity we did not simulate the first phase of sampling, since the first phase sample is very large and is not thought to contribute significantly to overall survey error. Sampling and estimation from this pseudo-population was simulated 10,000 times.

The variable of interest for estimation is taken to be hours, using employment and payroll as covariates. The regression weights for the regressions of hours onto employment and payroll are the inverses of employment. These are the weights currently used in SEPH estimation for hours.

Four estimators of the regression coefficient β_g are considered: the current SEPH estimator, the sample size dependent estimator (1), the generalized sample size dependent estimator (2), and the true β_g , *i.e.*, the population value of the weighted coefficient of regression of y onto x in group g using the same regression weights w . This last estimator would not be available in practice, but is given here for comparison as the best possible given the basic paradigm of regression imputation within the given model groups. The value of n_{\min} used in calculation of the sample size dependent estimator in (1) and the generalized sample size dependent estimator described in Section 3 was taken to be 100. The value of n_{\min} used in calculation of the sample size dependent estimators for calibration groups in (3) was taken to be 40. The first five months of data from our pseudo-population were used to determine the distance weights $d_{gg'}$ for use in the generalized sample size dependent estimator. All of the sampling, estimation and evaluation was then done using the data from the sixth month.

The imputations obtained from each of the estimators are also calibrated within calibration groups as described in Section 4, yielding a total of eight sets of imputations. The population was partitioned into 30 calibration groups, with expected sample size ranging from 40 to 201. The calibration groups were chosen to have a minimum expected sample size of 40, and they usually correspond exactly to model groups; however, four very large model groups were split into two or three parts along SIC2 boundaries, and one SIC2 which consisted of two smaller model groups was made into a single calibration group.

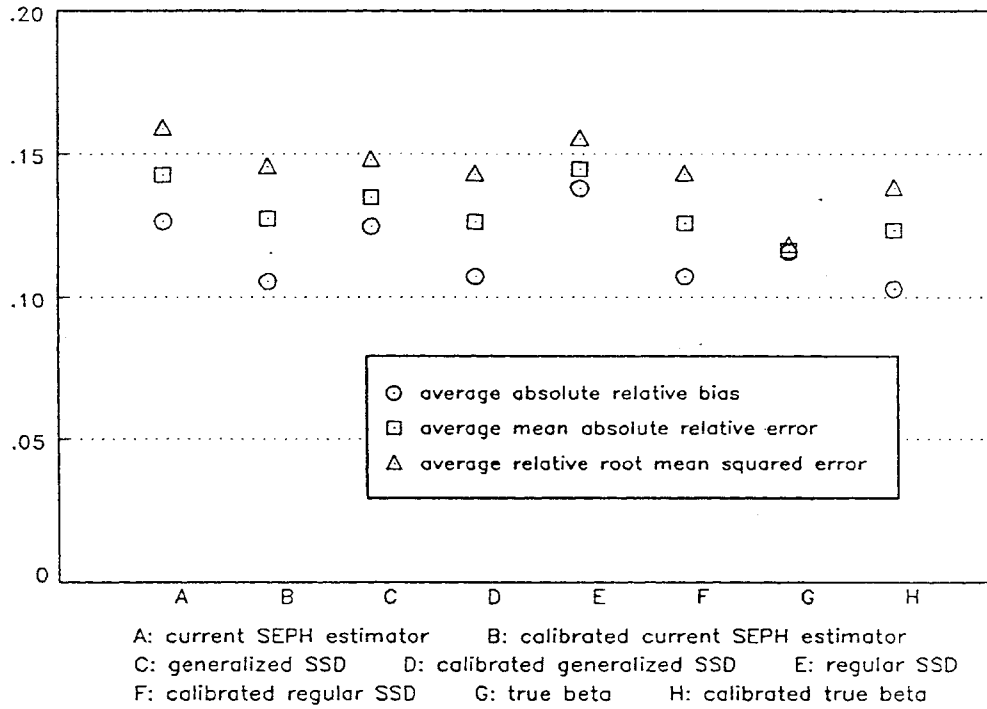
Three different measures of performance are calculated to evaluate the various imputation methods. All of these may be applied within various domains such as model groups or SIC2 by province groups. The basis for the evaluation measures is comparison to the known true total hours from the pseudo-population. The first measure is bias which is the Monte Carlo mean of the estimate of the total minus the true total. Second is the mean absolute error which is the Monte Carlo mean of the absolute difference between the estimated total and the true total. Third is the root mean squared error which is the square root of the Monte Carlo mean of the squared difference between the estimated total and the true total. All three measures are made into relative measures by dividing by the known true total hours. Explicit formulas for these evaluation measures are given in Mantel, Singh and Bureau (1993).

Average SIC2 level evaluation measures are given in Figure 1; here the word "average" refers to the unweighted average over 67 of the SIC2 groups (SIC2 46 was excluded because it had only four units in the pseudo-population and a total of 0 hours worked).

We first consider the uncalibrated estimators. Note that, at least on average, use of the generalized sample size dependent estimator leads to a slight improvement over the current SEPH estimator at the SIC2 level with respect to the variability measures mean absolute relative error and relative root mean squared error, whereas the performance of the regular sample size dependent estimator is somewhat worse than that of the current SEPH estimator. Comparison to the results for the "true" β shows that the current SEPH estimator is already doing close to as well as is possible, given the basic paradigm of regression imputation within the given model groups.

Calibration leads to improvement on all three evaluation measures, with the exception of the estimator based on the "true" β for which only the bias is improved. Note that after calibration there is little difference among the average evaluation measures for the different imputation methods.

Figure 1: evaluation measures, SIC2



As noted above, the calibration groups most often correspond exactly to model groups; in these cases calibration does not lead to any significant gains since, at the model group level the estimators are already doing very well. However, for those model groups that were split into two or more calibration groups very significant gains are observed; these gains are shown in Table 1. In

most of these cases the absolute relative bias is significantly reduced, apparently at some cost in terms of variability since the reductions in mean absolute relative error and relative root mean squared error are not as large. In a few isolated cases the calibrated estimator does not perform as well; however, improvements in the evaluation measures are much more common and larger.

Table 1. SIC2 level evaluations for the current SEPH estimator and its calibrated version

calibration group	model group	SIC2	expected sample size	relative bias		mean absolute relative error		relative root MSE	
				A	B	A	B	A	B
7	4435	42	201.2	.031	.001	.037	.030	.044	.037
8	4435	40	91.0	.213	.008	.213	.081	.219	.101
9	4435	41	24.4	-.213	.004	.213	.050	.214	.063
9	4435	44	18.4	-.317	-.124	.317	.126	.319	.141
13	5235	55	14.6	.014	.056	.025	.058	.032	.069
13	5235	56	35.9	-.060	-.018	.060	.030	.066	.038
14	5235	57	67.5	.057	.011	.067	.057	.080	.071
17	6535	60	62.5	.032	.007	.052	.061	.064	.076
18	6535	64	9.9	-.131	-.097	.132	.102	.143	.122
18	6535	69	65.7	-.018	.019	.042	.056	.054	.069
20	7235	75	71.8	-.099	.001	.099	.037	.105	.046
21	7235	76	75.2	.108	.004	.108	.043	.115	.054

A: current SEPH estimator B: calibrated current SEPH estimator

6. CONCLUSIONS

As seen from our empirical evaluation, the current SEPH imputation procedures work quite well and there is little room for improvement, given the basic paradigm of regression imputation based on a regression of hours onto employment and payroll within the given model groups; nevertheless, limited borrowing of strength from carefully selected related model groups can lead to slight improvement with respect to variability of the resultant estimators.

The fact that use of the "true" β leads to an average absolute relative bias of more than 11% at the SIC2 level indicates that the dependence of hours on employment and payroll is not homogeneous within model groups, and procedures which are approximately unbiased at a finer level than the model groups can perform significantly better. In particular, dividing model groups with large sample size into two or more subgroups of interest, and calibrating the imputed hours for these subgroups to approximately unbiased estimators can lead to large reductions in bias and smaller reductions in mean absolute error and mean squared error.

For smaller subgroups of interest and for domains that cross calibration groups some synthetic bias would still remain. The requirement of having good quality estimators for the calibration groups means that they cannot be too small. One approach to reduce the minimum acceptable size of calibration groups that might be developed in the future is the application of time series methods using the data from a number of months of the survey.

ACKNOWLEDGEMENT

We are grateful to Marcel Bureau for assistance in construction of the pseudo-population used in the empirical study and to Michel Latouche for details of the current SEPH sampling design.

REFERENCES

- Deville, J.-C., and Särndal, C.-E. (1992). "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87, 376-382.
- Drew, J.D., Singh, M.P., and Choudhry, G.H. (1982). "Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 545-550.
- Hidiroglou, M.A., Latouche, M., Armstrong, B., and Gossen, M. (1995). "Improving Survey Information Using Administrative Records: The Case of the Canadian Employment Survey", *Proceedings of the 1995 Annual Research Conference*, U.S. Bureau of the Census, to appear.
- Mantel, H.J., Singh, A.C., and Bureau, M. (1993). "Benchmarking of Small Area Estimators". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, June 1993, 920-925.
- Kalton, G. (1994). "Comment", discussion of Singh, Gambino and Mantel (1994), *Survey Methodology*, 20, 18-20.
- Singh, A.C., and Mian, I.U.H. (1995). "Generalized Sample Size Dependent Estimators for Small Areas", *Proceedings of the 1995 Annual Research Conference*, U.S. Bureau of the Census, to appear.
- Singh, M.P., Gambino, J., and Mantel, H. (1994). "Issues and Strategies for Small Area Data", *Survey Methodology*, 20, 1-22, (with discussion).