THE USE OF GENERALIZED RAKING PROCEDURES TO IMPROVE THE QUALITY OF SMALL DOMAIN ESTIMATION

Guylaine Dubreuil and Johanne Tremblay, Statistics Canada
Guylaine Dubreuil, Statistics Canada, 11-F R.-H. Coats, Ottawa, Canada, K1A 0T6

KEY WORDS: Auxiliary information, Variance estimation, Poststratification.

1. INTRODUCTION

Following a wide variety of ad hoc requests concerning the Canadian Farm Financial Survey, a need to improve the quality of estimates for small domains surfaced. The use of auxiliary information at the estimation level, such as the number of farms in the various domains, has been proposed. Two approaches to improve the quality of estimates for small domains were considered for the 1993 Farm Financial Survey. Both approachs used information coming from the 1991 Census of Agriculture. The first approach was a one-way poststratification that has shown promising results but has also posed the problem of additivity at the provincial level, an important constraint. Another approach was a multi-way poststratification using generalized raking procedures. This last approach was suggested with the aim to improve existing estimates but respecting the constraint of additivity. This paper investigates both approaches and compares them in terms of the improvement in the variance and in the coefficients of variation (CV's).

2. METHODOLOGY OF THE CANADIAN FARM FINANCIAL SURVEY

The objective of the Canadian Farm Financial Survey (FFS) is to collect financial information such as revenues, expenses, assets and liabilities. Since the data to be collected are very sensitive, the selected farms are primarily contacted by personal interview. Very few farms are contacted by telephone.

The target population of the 1993 FFS consists of all Canadian farms which were active during reference year 1992. Due to operational constraints, farms with less than $2,000 in sales from agricultural activities, institutional farms, community pastures, farms on Indian Reserves or farms that are part of multiholding companies were excluded.

Two frames were available for the 1993 FFS: an area frame and a list frame. The area frame is used to compensate for the undercoverage of the 1991 Census of Agriculture and identify new operators since the Census. As only area sample farms which are not on the list frame contributes to the area frame portion, the estimates produced from both frames are completely independent. In this paper, only the estimates coming from the list frame are studied. The design of the survey associated with the list frame is a stratified sample with simple random sampling within each stratum. The strata are defined by province, farm type and farm size in terms of total assets. In 1993, the survey was conducted in the provinces of Manitoba, Saskatchewan, Alberta and British Colombia. For the four provinces, a sample of 5,947 farms was selected from a target population of 150,823 farms.

3. HORVITZ-THOMPSON ESTIMATION METHOD

Let y_k be the value of the variable of interest, Y, associated with k^{th} population element. For the FFS, the method of Horvitz-Thompson (H-T) estimation is used to produce unbiased domain estimates of the population total $t_d = \sum_{k \in U_d} y_k$ where U_d represents the population in the domain d. The H-T estimator of the domain total for a stratified random sample is given by

$$\hat{t}_{d(HT)} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_{hd}} y_k ,$$

where N_h and n_h represent respectively the population size and the sample size in stratum h and s_{hd} represents the sample in the $(hd)^{th}$ cell, that is in the intersection of stratum h and domain d. An estimator of the variance is then given by

$$\hat{V}(\hat{t}_{d(HT)}) = \sum_{h=1}^{H} \frac{N_h}{n_h}(N_h - n_h) S^2_{y_{s_{hd}}} ,$$

where $S^2_{y_{s_{hd}}}$ represents the sample variance in the $(hd)^{th}$ cell. The CV is defined by

$$CV_{d(HT)} = \sqrt{\hat{V}(\hat{t}_{d(HT)})} / \hat{t}_{d(HT)} .$$

4. ONE-WAY POSTSTRATIFICATION METHOD

The first approach considered was a one-way

poststratification method. This approach uses as auxiliary information, the number, N_d, of farms in the domain U_d. This information comes from the Census of Agriculture for which exclusions have been made to get the same target population as for the survey. The one-way poststratified estimator of the domain total for a stratified random sample is given by (Särndal, Swensson and Wretman, 1992)

$$\hat{t}_{d(post)} = \frac{N_d}{\hat{N}_d} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_{hd}} y_k ,$$

where

$$\hat{N}_d = \sum_{h=1}^{H} \frac{N_h}{n_h} n_{s_{hd}}$$

and $n_{s_{hd}}$ represents the sample size in the $(hd)^{th}$ cell.

An estimator of the variance is

$$\hat{V}(\hat{t}_{d(post)}) = \frac{N_d^2}{\hat{N}_d^2} \sum_{h=1}^{H} \left[\frac{N_h(N_h - n_h)}{n_h(n_h - 1)} \sum_{k \in s_{hd}} (y_k - \bar{y}_{s_{hd}})^2 + \sum_{h=1}^{H} n_{s_{hd}} \left(1 - \frac{n_{s_{hd}}}{n_h}\right) \left(\bar{y}_{s_{hd}} - \frac{\hat{t}_{d(post)}}{N_d}\right)^2 \right],$$

where $\bar{y}_{s_{hd}}$ is the usual mean of y_k in the $(hd)^{th}$ cell.

The CV is defined by

$$CV_{d(post)} = \sqrt{\hat{V}(\hat{t}_{d(post)})} / \hat{t}_{d(post)} .$$

5.COMPARISON BETWEEN THE H-T AND THE POSTSTRATIFIED ESTIMATORS

A comparison was performed between the H-T estimator and the poststratified estimator using Manitoba data from the 1993 Canadian Farm Financial Survey and totals from the 1991 Census of Agriculture. Three sets of domains of interest were considered: the farm types (defined at a different level than the strata), the classes of sales and the crop districts (see Appendix I for the population sizes and the sample sizes). These three sets of domains are treated separately because the crossed domains produced too many empty cells. The results of the study based on six different variables of interest are presented in Table 1 in terms of the improvement in the variance and in the CV. The *Ratio of variances* and the *Improvement in the CV's* are measured respectively by the average of

$$\hat{V}(\hat{t}_{d(post)}) / \hat{V}(\hat{t}_{d(HT)})$$

and the average of

$$\frac{CV_{d(post)} - CV_{d(HT)}}{CV_{d(HT)}}$$

over each set of domains, within the province.

Comparing the poststratified estimator with the H-T estimator, we observe that for each variable of interest, there is an improvement in the variance in the sets of domains *classes of sales* and *crop districts*. However, this is not the case when the set of domains is farm types, despite the fact that these domains are highly related to the stratum definition. In fact, this is due to a particular farm type, with a sample size of 3, for which the ratio of the poststratification variance estimate and the H-T variance estimate is very high (varying from 4.19 to 8.40 depending on the variable of interest). Otherwise, the ratio is less than 1 or close to 1. If we look at the *CV*'s there is an improvement in each case and sometimes this improvement is quite substantial. Nevertheless, a problem arises with the poststratification method; from the last column of Table 1, we can see that an estimate of the provincial total for a given variable changes and can vary a lot depending on the set of domains used to estimate it. This issue, called problem of additivity in this paper, is problematic in our case due to the difficulty for users to reconcile different estimates of provincial totals for the same variable. For this reason, the one-way poststratification method does not satisfy all of our needs.

6. GENERALIZED RAKING PROCEDURES (MULTI-WAY POSTSTRATIFICATION)

To satisfy the constraint of additivity at the provincial level, one solution is to consider generalized raking procedures (Deville, Särndal and Sautory, 1993) also called *multi-way poststratification*. These procedures can be used for estimation with auxiliary information in two or more dimensions. That is, these procedures can simultaneously treat the auxiliary information of diverse types of domains to provide a series of adjustment weights, called *g-weights*. The advantage of using the generalized raking procedure is that the estimate of the provincial total of a particular variable of interest is the same whatever the set of domains. This is due to the fact that the new weights are

TABLE 1: COMPARISON BETWEEN THE H-T AND THE ONE-WAY POSTSTRATIFIED ESTIMATORS IN MANITOBA

Variable of interest	Estimation Method	Set of domains	*Ratio of variances*	*Improvement in the CV's* (%)	Provincial Total Estimate
Assets	Poststratification	Farm types	1.15	-20.84	10,748,706,319
		Classes of sales	0.32	-42.18	10,176,728,514
		Crop districts	0.62	-22.20	10,289,714,482
	H-T	-	-	-	10,376,868,008
Sales	Poststratification	Farm types	1.66	-13.37	2,315,359,322
		Classes of sales	0.14	-62.16	2,141,989,326
		Crop districts	0.78	-12.06	2,214,241,220
	H-T	-	-	-	2,237,664,715
Expenses	Poststratification	Farm types	1.61	-12.71	1,883,766,100
		Classes of sales	0.28	-46.78	1,745,161,209
		Crop districts	0.79	-11.80	1,803,473,044
	H-T	-	-	-	1,823,416,244
Short Term Liabilities	Poststratification	Farm types	1.23	-13.23	458,587,322
		Classes of sales	0.77	-15.35	429,540,232
		Crop districts	0.92	-5.37	439,297,701
	H-T	-	-	-	446,366,763
Long Term Liabilities	Poststratification	Farm types	1.68	-8.29	1,266,462,557
		Classes of sales	0.73	-16.63	1,153,200,949
		Crop districts	0.85	-8.62	1,191,170,984
	H-T	-	-	-	1,203,681,425
Total Liabilities	Poststratification	Farm types	1.65	-9.65	1,725,049,879
		Classes of sales	0.66	-21.24	1,582,741,181
		Crop districts	0.84	-9.08	1,630,468,685
	H-T	-	-	-	1,650,048,189

calculated only once for the three sets of domains. In the one-way poststratification estimation, the *g-weights* are given by $N_d/\hat{N}_d$ and consequently, they vary with the set of domains.

By using the generalized raking procedures, the *g-weights* are calculated by an iterative process which minimizes the distance between the original weights and the final weights while satisfying the constraints. To compute the *g-weights*, we used the SAS macro CALMAR developed at I.N.S.E.E. in France (Sautory, 1991). The macro CALMAR can compute different series of *g-weights* depending on the distance measure we specify. Four different distance measures can be used: the linear method, the raking ratio method, the logit method with a *g-weight* lower threshold L and a *g-weight* upper threshold U and the truncated linear method with a *g-weight* lower threshold L and a *g-weight* upper threshold U. The lower and the upper thresholds should be such that $L < 1 < U$ and should not be too restrictive, otherwise the *g-weights* will have

a tendency to accumulate at the thresholds.

To choose the appropriate distance measure for our case and to avoid extreme weights which may produce unrealistic estimates for some domains, we considered two additional constraints: positive final weights (i.e. positive *g-weights*; $L > 0$) and avoid extreme final weights. In that regard, our choice fell on the two truncated methods. After some tests on both methods using a lower threshold $L=0$ and an upper threshold $U=4$, the logit method was preferred because the *g-weights* provided by this method generally showed a better behaviour. By good behaviour, we mean that the mean and the median of the *g-weights* are close to one and the *g-weights* are concentrated around one with only few *g-weights* around the thresholds. Note that with the logit method, the lower threshold $L=0$ was unnecessary because this method always provides positive *g-weights*. The distance function associated with the logit method with thresholds L and U is as follows (Sautory, 1991)

$$G(x)=\frac{1}{A}\left((x-L)\log\frac{x-L}{1-L}+(U-x)\log\frac{U-x}{U-1}\right) \quad if\ L<x<U$$
$$=\infty \quad otherwise$$

with x corresponding to the *g-weight* and

$$A=\frac{U-L}{(1-L)(U-1)}.$$

Now, consider $X_k=(x_{k1}, \ldots, x_{kd}, \ldots, x_{kD})^T$ where D corresponds to the total number of domains of the study and

$$x_{kd}=1 \quad if\ k \in d$$
$$=0 \quad otherwise.$$

Note that $\Sigma_U X_k=(N_1, \ldots, N_d, \ldots, N_D)^T=X$, which is a vector of known totals (the number of farms in each domain coming from the census). Let d_k and w_k respectively be the original weight and the final weight, for the k^{th} population element. The problem then consists of finding w_k, a solution to (Sautory,1991)

$$\min_{w_k} \sum_{k\in s} d_k G\left(\frac{w_k}{d_k}\right)$$
$$under \sum_{k\in s} w_k X_k$$

where s represents the sample. For our study, $d_k=N_h/n_h$ for k in stratum h. Also, the D domains correspond to the three sets of domains discussed in the previous sections. That is, a subset of X_k is associated to each set of domains.

Once CALMAR has calculated the final weights, the estimator of the domain population total for a stratified random sample is then given by

$$\hat{t}_{d(cal)}=\sum_{h=1}^{H}\sum_{k\in s_{hd}} w_k y_k=\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k\in s_{hd}} g_k y_k,$$

where

$$g_k=w_k/d_k$$

are the corresponding *g-weights*. From Deville and Särndal (1992), for any distance measure under mild constraints, $\hat{t}_{d(cal)}$ is asymptotically equivalent to the regression estimator. Consequently, the two estimators share the same asymptotic variance which is given by (Hidiroglou, 1991)

$$\hat{V}(\hat{t}_{d(cal)})=\sum_{h=1}^{H}\frac{N_h-n_h}{n_h-1}\frac{n_h}{N_h}\sum_{k\in s_h}(u_{hk}-\bar{u}_h)^2,$$

where

$$u_{hk}=\frac{N_h}{n_h}g_k e_{dk}, \quad \bar{u}_h=\frac{\sum_{k\in s_h} u_{hk}}{n_h},$$

with s_h representing the sample in the stratum h. Here, the residuals e_{dk} are given by

$$e_{dk}=y_{dk}-X_k^T\left(\sum_{k\in s}\frac{N_h}{n_h}X_k X_k^T\right)^{-1}\left(\sum_{k\in s}\frac{N_h}{n_h}X_k y_{dk}\right)$$

with

$$y_{dk}=y_k \quad if\ k \in d$$
$$=0 \quad otherwise.$$

As usual, the CV is defined by

$$CV_{d(cal)}=\sqrt{\hat{V}(\hat{t}_{d(cal)})}/\hat{t}_{d(cal)}.$$

TABLE 2: COMPARISON BETWEEN THE H-T AND THE MULTI-WAY POSTSTRATIFIED ESTIMATORS

Variable of interest	Set of domains	Manitoba		Saskatchewan		Alberta	
		Ratio of variances	*Improvement in the CV's (%)*	*Ratio of variances*	*Improvement in the CV's (%)*	*Ratio of variances*	*Improvement in the CV's (%)*
Assets	Farm types	1.08	-22.56	1.16	-8.76	0.60	-22.39
	Classes of sales	0.35	-41.45	0.40	-33.96	0.37	-36.57
	Crop districts	0.56	-25.88	0.41	-34.15	0.55	-23.24
Sales	Farm types	1.51	-17.59	0.90	-17.08	0.64	-16.69
	Classes of sales	0.14	-62.70	0.16	-61.10	0.19	-57.91
	Crop districts	0.71	-15.97	0.50	-27.05	0.64	-12.75
Expenses	Farm types	1.46	-16.86	0.87	-17.93	0.73	-11.91
	Classes of sales	0.30	-46.06	0.24	-49.18	0.41	-33.49
	Crop districts	0.72	-15.76	0.48	-27.83	0.66	-12.40
Short Term Liabilities	Farm types	1.14	-15.59	1.17	-3.26	0.91	-4.71
	Classes of sales	0.77	-16.02	0.74	-12.55	0.87	-5.63
	Crop districts	0.86	-7.98	0.68	-11.67	0.78	-3.35
Long Term Liabilities	Farm types	1.56	-9.32	0.88	-12.46	0.77	-12.78
	Classes of sales	0.73	-17.34	0.64	-20.12	0.71	-14.22
	Crop districts	0.86	-9.55	0.66	-14.30	0.71	-12.64
Total Liabilities	Farm types	1.53	-11.04	0.89	-12.53	0.74	-13.61
	Classes of sales	0.64	-22.40	0.61	-22.23	0.69	-15.50
	Crop districts	0.83	-10.92	0.64	-15.63	0.71	-12.03

7. COMPARISON BETWEEN THE H-T AND THE MULTI-WAY POSTSTRATIFIED ESTIMATORS

The comparison performed with the one-way poststratification method was also done with the generalized raking procedure. This time, the study included three provinces: Manitoba, Saskatchewan and Alberta. The results are presented in Table 2.

If we compare the results of Manitoba in Table 2 to those presented in Table 1, we can see that the *Ratio of variances* and the *Improvement in CV's* are close. For the other provinces, Saskatchewan and Alberta, the variance improves almost everywhere except for two variables of interest in Saskatchewan. However, the *Ratio of variances* is not too far from 1 and there is again an improvement in the CV's. This improvement in the variance when the set of domains is the farm type could be explained by the fact that there are less types of farms in Saskatchewan and Alberta, so the domains are larger.

8. CONCLUSION

In this study, the generalized raking procedures provide satisfactory results. However, we should be careful if the domains are very small because we use an asymptotic variance. In this study, the sample size was generally large enough to get relatively precise results, except for some farm types.

In fact, given that in practice we are often interested in smaller domains, we have to determine what domain sample size is necessary to obtain relatively precise results by using the asymptotic variance. And if the domain sample size is not large enough, we have to

find another way to estimate the variance.

APPENDIX I: POPULATION SIZE AND SAMPLE SIZE IN THE DOMAINS

Set of domains	Description	Manitoba		Saskatchewan		Alberta	
		N_d	n_d	N_d	n_d	N_d	n_d
Farm types	Dairy	1,194	48	754	16	1,385	39
	Cattle	5,071	187	8,916	191	22,307	585
	Hogs	1,228	100	768	16	1,646	60
	Poultry & Eggs	303	28	-[1]	-[1]	402	10
	Potatoes	86	13	-[1]	-[1]	-[1]	-[1]
	Grains & Oilseeds	12,662	621	43,603	1,071	18,696	417
	Fruits & Vegetables	158	5	-[1]	-[1]	-[1]	-[1]
	Greenhouse & Nursery	177	3	-[1]	-[1]	-[1]	-[1]
	Others	2,940	156	4,141	94	9,188	204
Classes of sales	sales ≤ 9,999	3,656	163	6,218	144	8,824	187
	10,000 ≤ sales ≤ 24,999	4,203	120	10,311	169	11,090	189
	25,000 ≤ sales ≤ 49,999	4,471	141	13,509	243	9,800	186
	50,000 ≤ sales ≤ 99,999	5,287	223	16,286	357	10,513	223
	100,000 ≤ sales ≤ 249,999	4,680	338	10,096	341	9,620	289
	250,000 ≤ sales ≤ 499,999	1,134	86	1,412	93	2,515	114
	sales ≥ 500,000.	388	90	350	41	1,262	127
Crop districts[2]	1	1,840	94	2,392	49	3,192	86
	2	2,274	125	2,311	48	6,632	191
	3	2,327	107	1,948	41	6,350	213
	4	1,113	42	3,232	77	9,869	225
	5	1,102	53	1,547	40	11,742	257
	6	2,150	64	2,756	69	8,464	216
	7	2,821	165	2,861	87	7,375	127
	8	3,465	203	1,664	43	-	-
	9	2,579	138	1,134	36	-	-
	10	727	33	1,394	36	-	-
	11	1,476	73	4,998	114	-	-
	12	1,945	64	5,284	104	-	-
	13	-	-	4,092	90	-	-
	14	-	-	3,703	94	-	-
	15	-	-	2,359	63	-	-
	16	-	-	2,331	51	-	-
	17	-	-	3,072	84	-	-
	18	-	-	3,310	83	-	-
	19	-	-	4,717	101	-	-
	20	-	-	3,077	78	-	-

REFERENCES

Deville, J.-C. and Särndal, C.E. (1992), "Calibration Estimators in Survey Sampling". Journal of the American Statistical Association 87, 376-382.

Deville, J.-C., Särndal, C.E. and Sautory, O. (1993), "Generalized Raking Ratio Procedures in Survey Sampling". Journal of the American Statistical Association 88, 1013-1020.

Hidiroglou, M.A. (1991), "Structure of the Generalized Estimation System". Statistics Canada Technical Report, September 1991.

Särndal, C.E., Swensson B. and Wrettman, J. (1992), Model Assisted Survey Sampling, New York: Springer-Verlag.

Sautory, O. (1991), "Redressement d'Échantillons d'Enquêtes auprès des Ménages par Calages sur Marges". I.N.S.E.E. Technical Report,

[1]Included in the farm type 'Others'

[2]The crop districts are differently defined in each province.