TWO METHODS OF MEASURING CORRELATED RESPONSE VARIANCE

Karol P. Krótki and A. MacLeod, Statistics Canada

1. Introduction

The correlated response variance (CRV) is one component of the overall or total variance of a survey estimator. Its definition is based upon considering a survey process as theoretically repeatable so that a series of independent trials of the survey exists. Response deviation is defined as the difference between an individual's response in the actual survey trial and the expected value of that individual's response, with the expectation taken over all conceptual survey trials. The CRV is then defined as a measure of the variability in response caused by the correlation of these response deviations for any pair of individuals within the given trial of the survey process. It has been shown empirically that, in a setting where enumerators help the respondents fill out the questionnaire, the CRV is an important component of total variance [3].

To estimate the CRV for various variables, two methods were employed. The first of these, hereafter referred to as the old method, is described in [2]. It is the method upon which Statistics Canada based its published estimates of total variance in the 1971 and 1976 Censuses. The other method, hereafter identified as the new method, is defined and derived in [4].

This paper investigates and compares these two methods, both in terms of the estimators themselves (Section 2) and the corresponding estimates that are yielded (Section 3). These estimates will be provided and discussed for data collected on the basis of a 1/3 sample. In addition, the variance of both estimators is presented in Section 4 and Section 5 discusses the effects of outliers. Section 6 consists of the development of an improved estimator with lower variance than either the old or new estimators. The paper concludes with Section 7 which contains summary comments, an overall evaluation of the improved method and some research ideas.

2. A Theoretical Comparison of the Estimators

Since a survey process cannot in practice be repeated a number of independent times, a technique of interpenetration of enumerators and enumeration areas (EAs[1]) is used for both CRV estimators. In normal Census practice, one enumerator is assigned to handle one EA. With interpenetration, however, neighbouring EAs are paired and within these pairs the households in each EA are randomly split into two equal groups. One enumerator is then randomly assigned to one of these two groups in each EA of the pair while the second enumerator is given the other group from each EA of the pair.

For the old estimator, only a sample of interpenetrated EA pairs is needed. The new estimator, on the other hand, requires not only the sample of interpenetrated EA pairs but also a sample of non-interpenetrated EA pairs. In theory, all of the pairs to which the interpenetration technique is not applied could be used, instead of just a sample of such pairs. However, practical budgetary constraints dictate the use of a sample.

The correlated response variance for a Census estimator X based on sample data[2] is given by the following expression:

$$CRV(X) = \sum_{k=1}^{2P} N_k^2 \frac{(n_k-1)}{n_k} \rho_k \sigma_k^2 \tag{1}$$

where k is the EA index denoting the EA,

P is the number of EA pairs in Canada,
N_k is the total number of households in the kth EA and
n_k is the number of sample households in the kth EA (n_k is approximately equal to $N_k/3$).

The entity $\rho_k \sigma_k^2$ is the key term in (1). It is defined as the average value of $E(x_{kh}-X_{kh})(x_{kh'}-X_{kh'})$, for pairs of households (h,h';h≠h') in the kth EA that are enumerated by the same enumerator. x_{kh} is the actual Census response and X_{kh} is the mean of all conceptual responses so that $(x_{kh}-X_{kh})$ is the response deviation in the hth household of the kth EA. The expectation is taken over all conceptual responses, including enumerator assignments, for a given household.

For 100% data, $n_k = N_k$ and so (1) reduces to:

$$CRV(X) = \sum_{k=1}^{2P} N_k(N_k-1) \rho_k \sigma_k^2 \tag{2}$$

Dealing first with sample characteristics, the old estimator is given by:

$$\widehat{CRV}(X) = \frac{P}{p_I} \sum_{k=1}^{2p_I} N_k^2 \frac{(n_k-1)}{n_k} (C_k-D_k) \tag{3}$$

where p_I is the number of interpenetrated EA pairs in the sample, C_k is the between enumerator sum of squares and D_k is the within enumerator sum of squares. Expressions for C_k and D_k are given in [2]. Under certain conditions, (C_k-D_k) is an unbiased estimator of $\rho_k\sigma_k^2$.

For 100% data, the expressions for C_k and D_k are modified slightly. The CRV estimator in this situation is given by:

$$\widehat{CRV}(X) = \frac{P}{p_I} \sum_{k=1}^{2p_I} N_k(N_k-1)(C_k-D_k) \tag{4}$$

It is shown in [2] that under certain conditions both estimators are unbiased for correlated response variance; i.e., (3) is an unbiased estimator of (1) and (4) is an unbiased estimator of (2).

A weighted average of the $(C_k - D_k)$ expression over the interpenetrated EAs has been used as an estimator of CRV and total variance. For example, Statistics Canada employed such a procedure for the 1971 and 1976 Censuses. The methodology and results are discussed in [5] and [6] . The U.S. Bureau of the Census also used a weighted combination of $(C_k - D_k)$'s based on 1960 data, as discussed in [7]. Bailey, Moore and Bailar [1] calculated averages of expressions equivalent to $(C_k - D_k)$ over all pairs of assignment areas in a national crime survey.

In the new method, the non-interpenetrated and interpenetrated samples are considered separately at the EA pair level. For each pair m, the quantity $A_m = (t_{m1} - t_{m2})^2$ is determined where t_{mj} is the number of units in some category in the jth EA of the mth pair (j=1,2). The A_m are next averaged separately for the non-interpenetrated pairs and for the interpenetrated pairs. Finally, a weighted difference of these two resulting averages is obtained. This difference is shown by Fellegi [4] to be a biased estimator of $\rho_m \sigma_m^2$, which is the average value of $E(x_{kh} - X_{kh})(x_{k'h'} - X_{k'h'})$, (k=1,2; if k=k' then h≠ h'). The expectation is over all conceptual responses for a given household. The average is taken over all pairs of households in the mth EA pair which were interviewed by the same enumerator and over both enumerators in the EA pair. Fellegi [4] anticipates that the bias of this estimator is small.

Since estimates based on formulae (3) and (4) were already available, it was decided to obtain a form of the new estimator with the same expected value as the estimators in (3) and (4). This was somewhat complicated since the old estimator is based on calculations within individual EAs while the new estimator is based upon calculations within pairs of EAs. The problem was solved by looking at the expected values of the old and new estimators and by assuming that $\rho_k \sigma_k^2 = \rho_{k'} \sigma_{k'}^2 = \rho_m \sigma_m^2$, where k and k' are the two EAs of any pair m. For 100% data, (2) can be rewritten as:

$$CRV(X) = \sum_{m=1}^{P} [N_{m1}(N_{m1}-1) + N_{m2}(N_{m2}-1)] \rho_m \sigma_m^2$$

where m is the index denoting the pair under consideration and N_{mi} is the total number of private households in the ith EA (1=1 or 2) of the mth pair. For sample data, (1) can be similarly rewritten.

For sample data, the revised new estimator was found to be:

$$\hat{CRV}(X) = 2P \left\{ \sum_{m=p_I+1}^{P} W_m \frac{\left[\frac{N_{m1}}{n_{m1}} t_{m1} - \frac{N_{m2}}{n_{m2}} t_{m2} \right]^2}{P-p_I} - \sum_{m=1}^{pI} W_m \frac{\left\{ \frac{N_{m1}}{n_{m1}} t_{m1} - \frac{N_{m2}}{n_{m2}} \right\} t_{m2}}{p_I}^{2} \right\} \qquad (5)$$

$$\text{where } W_m = \frac{N_{m1}^2 (n_{m1}-1)}{N_m^2 \; n_{m1}} + \frac{N_{m2}^2 (n_{m2}-1)}{N_m^2 \; n_{m2}}$$

For 100% data, the revised new estimator is:

$$\hat{CRV}(X) = 2P \frac{1}{P-p_I} \sum_{m=p_I+1}^{P} \frac{N_m^2 (N_m-1) - 2N_{m1}N_{m2}}{N_m^2} (t_{m1}-t_{m2})^2$$

$$- \frac{1}{p_I} \sum_{m=1}^{p_I} \frac{N_m^2 (N_m-1) - 2N_{m1}N_{m2}}{N_m^2} (t_{m1}-t_{m2})^2 \qquad (6)$$

For algebraic convenience, it is assumed in both (5) and (6), without loss of generality, that the first p_I EA pairs are interpenetrated. All future uses of the term "new method" or "new estimator" shall refer to (5) for sample data and to (6) for 100% data.

3. An Empirical Comparison of the Estimators

The old and new estimates are based on sample sizes of 375 interpenetrated EA pairs and 564 non-interpenetrated EA pairs. The samples were obtained through a two-stage process. In the first stage, 188 Census Commissioner Districts (CCDs[3]) were selected by PPS (probability proportional to size) sampling, where the measure of size was the number of EAs per CCD. These CCDs were chosen independently from within sixteen strata formed by cross-classifying eight geographical regions of Canada with two types of enumeration methods (pick-up and mail-back). In the second stage, within each selected CCD, all EAs were first paired. Then two pairs were randomly selected for the interpenetrated sample, followed by three additional pairs for the non-interpenetrated sample. Selection of EAs within CCDs was done without replacement.

In order to be paired, two EAs had to be contiguous and of the same enumeration type and had to possess similar linguistic, agricultural, and density characteristics. In addition, each EA of a non-interpenetrated pair had to be a "full-load" EA. That is, it was assigned to an enumerator who was assigned no other EAs. These pairing criteria tend to make the two EAs in a pair alike so the assumption in Section 2 that $\rho_k \sigma_k^2 = \rho_k \sigma_{k'}^2 = \rho_m \sigma_m^2$ for each pair m is supported.

Using the old estimator, the correlated response variance was calculated for a large number of categories based on 100% data (age, sex, mother tongue, marital status) and sample data (education, labour force, mibility). A similar selection was made for use with the new estimator. The selection of variables and categories was made to cover a wide variety of results and subject-matter interests. Resulting CRV estimates for both estimators are given for sample data in Table 1. Results based on 100% data are not included in this paper due to lack of space. The corresponding coefficients of variation are also given where the coefficient of variation is defined as the square root of the absolute value of the CRV estimate for that category divided by the number of people in Canada belonging to the category.

The old coefficients of variation tend to be smaller in magnitude than the new ones. One possible reason for this general phenomenon is that there is an unmeasured effect confounded with new CRV estimates. The quantity $(t_{m1}-t_{m2})^2$ in the new estimator is affected by the fact that the number of households in the first and second EA of pair m is not equal.

It is reported in [7] that relatively high response variances existed in the 1960 U.S. Census for data pertaining to low education levels and unemployment (among other subject-matter areas). An examination of Table 1 tends to support this view, insofar as the coefficients of variation for these categories are among the highest displayed.

4. The Variance of the CRV Estimators

The next step in this study was to attempt to estimate the variance of the CRV estimators so that some measure of reliability could be attached to the CRV estimates. The jacknife technique was used. Estimated variances of the old and new CRV estimators are given for selected categories based on sample data in Table 2.

TABLE 1
ESTIMATES OF CORRELATED RESPONSE VARIANCE USING THE OLD AND NEW ESTIMATORS WITH CORRESPONDING COEFFICIENTS OF VARIATION FOR SELECTED CATEGORIES BASED ON SAMPLE DATA

CATEGORY	OLD CRV ESTIMATE	COEFFICIENT OF VARIATION	NEW CRV ESTIMATE	COEFFICIENT OF VARIATION
Highest Degree Received:				
High School Certificate	2,344,591	4.56×10^{-4}	11,293,155	1.00×10^{-3}
Bachelors	- 899,953	1.20×10^{-3}	1,058,321	1.30×10^{-3}
Masters	-1,928,039	8.24×10^{-4}	-147,189	2.28×10^{-3}
Non-university Cert.	-1,230,405	6.05×10^{-4}	485,724	3.80×10^{-4}
Highest Grades Completed:				
Less Than Grade 5	932,373	1.12×10^{-3}	1,398,493	1.37×10^{-3}
Grades 5 - 8	1,648,121	3.60×10^{-4}	-4,066,516	5.66×10^{-4}
Grades 9 - 10	1,502,907	3.68×10^{-4}	-3,689,625	5.76×10^{-4}
Not Attending School	4,197,176	1.43×10^{-4}	10,793,519	2.29×10^{-4}
Attending School Full-time	495,245	4.20×10^{-4}	618,369	4.69×10^{-4}
Unemployed:	-2,171,495	2.11×10^{-3}	-756,508	1.24×10^{-3}
On Temporary Layoff	-5,263,069	2.24×10^{-2}	-88,831	2.91×10^{-3}
Waiting to Start New Job	-2,761,460	1.39×10^{-2}	-33,105	1.53×10^{-3}
Looked for Work	-2,275,514	3.15×10^{-3}	-609,951	1.63×10^{-3}
Employed	3,357,282	1.92×10^{-4}	41,506,918	6.74×10^{-4}
Not In Labour Force	4,241,919	3.01×10^{-4}	-966,135	1.46×10^{-4}
Mobility Status (in last 5 years):				
Non-Mover	6,204,320	2.28×10^{-4}	-122,571,179	1.01×10^{-3}
Mover	3,093,314	1.71×10^{-4}	218,080,472	1.43×10^{-3}
Non-Migrant[a]	1,567,852	2.51×10^{-4}	26,889,868	1.04×10^{-3}

a. A non-migrant is a person living in the same municipality as five years ago.

The table shows that the variance estimates are almost all higher for the new CRV estimator than for the old one. However, a comparison of old and new coefficients of variation does not seem to reveal any clear patterns. Even for codes of a single variable, there is no obvious pattern for old coefficients. The most important observations then, to come from the estimates of the coefficient of variation is their large magnitude. With many of them greater than one, there is a strong evidence of high variance in both the old and new CRV estimates.

5. The Effects of Outliers on the CRV Estimates

In order to pinpoint some of the causes of the large coefficients of variation, a study was made of the effects of outliers on both the old and new CRV estimates. There was initially a strong suspicion that a few outliers were heavily influencing some of the new CRV estimates. The first problem in such a study was to develop a systematic criterion for defining an outlier.

For the new method, any pair for which $(t_{m1}-t_{m2})$ was more than five standard deviations from the mean was declared an outlying pair. The criterion developed for the old method was that, if either C_k or D_k was more than eight standard deviations from the mean, then the EA k would be called an outlying EA (for the particular characteristic).

The next stage was to recalculate the CRV estimates from some categories with the outliers deleted (and the sample size accordingly reduced). Looking at the old estimator first, there was little observed change in the CRV estimates. The removal of outlying EA pairs did have a more noticeable affect on the new CRV estimates. For some categories, the tendency of the CRVs to approach zero was quite pronounced.[3]

Another method for dealing with outliers with the new estimator was to replace in each outlying pair the value of $(t_{m1}-t_{m2})$ by the appropriate mean, depending on whether the pair was non-interpenetrated or interpenetrated. This approach gave

TABLE 2
VARIANCE ESTIMATES (WITH COEFFICIENTS OF VARIATION) OF THE OLD AND NEW CRV ESTIMATORS FOR SAMPLE DATA

CATEGORY	VARIANCE OF OLD CRV	COEFFICIENT OF VARIATION (OLD)	VARIANCE OF NEW CRV	COEFFICIENT OF VARIATION (NEW)
Highest Degree Received:				
High School Certificate	1.088×10^{12}	0.445	2.473×10^{13}	0.440
Bachelors	4.388×10^{11}	0.736	5.801×10^{11}	0.720
Masters	8.694×10^{11}	0.484	9.277×10^{9}	0.654
Non University Cert.	2.424×10^{11}	0.400	7.545×10^{12}	5.655
Highest Grade Completed:				
Less Than Grade 5	3.535×10^{12}	2.017	4.193×10^{12}	1.464
Grades 5 - 8	3.623×10^{12}	1.155	3.608×10^{13}	1.477
Grades 9 - 10	1.834×10^{12}	0.901	4.435×10^{13}	1.805
Not Attending School	1.013×10^{13}	0.758	2.968×10^{15}	5.047
Attending School Full Time	5.706×10^{11}	1.525	4.688×10^{12}	3.501
Unemployed:	4.147×10^{11}	0.297	4.515×10^{11}	0.888
On Temporary Layoff	1.747×10^{13}	0.794	1.629×10^{10}	1.437
Waiting to Start New Job	8.143×10^{11}	0.327	1.764×10^{9}	1.269
Looked for Work	2.620×10^{11}	0.225	2.139×10^{11}	0.758
Employed	7.000×10^{12}	0.788	7.739×10^{14}	0.670
Not In Labour Force	1.357×10^{13}	0.868	2.544×10^{14}	16.012
Mobility Status (in last 5 years):				
Non-Mover	4.309×10^{13}	1.058	8.912×10^{14}	0.400
Mover	9.542×10^{13}	3.158	2.110×10^{13}	0.402
Non-Migrant	7.176×10^{12}	1.709	2.512×10^{14}	0.589

very similar results to the approach of removing the outliers completely.

In conclusion, although removing outliers had more effect on the new CRV estimator than on the old one in reducing both the CRV estimates and their variances, the effects were still not significant. With the exception of the small mother tongue groups, the new CRVs remained farther from zero than their old counterparts and the variance of new CRVs remained larger than those of the old CRVs.

6. An improved estimator

It was suggested in Fellegi [4,p.500] that a weighted combination of the old and new estimators would provide an improved estimator. Such an estimator was calculated where the weights were inversely proportional to the variances of the estimators. The results are presented in Table 3.

For the first seven categories both the old and improved CRVs are negative and the interpenetration of the last column is not obvious. Nevertheless, it can be seen that in three of the seven cases major reduction has occurred and, in three of the remaining four cases, substantial reduction has occurred. For the categories with positive old CRVs, the mean reduction is 15.4% and for individual categories the reduction rate ranges from a low of 0.3% to a high of 81.9%. In conclusion, based on the categories chosen and the data studied, it appears that substantial improvements in the reliability of the estimators can be achieved by resorting to the improved method.

7. Conclusion

It has been shown how an improved estimator can be derived as the weighted combination of the old and new estimators. The estimator is improved in the sense that its variance is lower than that of either of the two original estimators. For certain variables the reduction in variance and consequent increase in stability is quite substantial. There are no extra field costs in determining the improved estimator. However, it

TABLE 3
ESTIMATES USING THE IMPROVED METHOD

CATEGORY	IMPROVED CRV	VARIANCE OF IMPROVED CRV	VARIANCE DECREASE OVER OLD CRV(%)
Highest degree received			
Bachelors	-56,602	2.498×10^{11}	43.1
Masters	-165,990	0.092×10^{11}	98.9
Non-university certificate	− 1,176,987	2.348×10^{11}	3.1
Labour force status			
Unemployed	− 1,494,059	2.162×10^{11}	47.9
On temporary layoff	-93,651	0.002×10^{13}	99.9
Waiting to start new job	-39,002	0.017×10^{11}	99.8
Looked for work	− 1,358,562	1.178×10^{11}	55.0
Highest degree received			
High school certificate	2,721,694	1.042×10^{12}	4.2
Highest grade completed			
Less than grade 5	1,145,589	1.918×10^{12}	45.7
Grades 5- 8	1,126,646	3.292×10^{12}	9.1
Grades 9-10	1,296,708	1.761×10^{12}	4.0
School attendance			
Not attending school	4,219,613	1.010×10^{13}	0.3
Attending school full time	508,604	5.087×10^{11}	10.6
Labour force status			
Employed	3,699,256	6.937×10^{12}	0.9
Not in the labour force	3,978,183	1.288×10^{13}	5.1
Mobility status (in last 5 years)			
Non-mover	2,476,082	4.110×10^{13}	4.6
Mover	9,917,528	1.728×10^{12}	81.9
Non-migrant[a]	2,271,132	7.176×10^{12}	2.7

[a] A non-migrant is a person living in the same municipality as five years age.

should be pointed out that the improved estimator requires the non-interpenetrated EAs to be paired.

The importance of increase in stability should not be underestimated. One of the problems involved in calculating and disseminating estimates of correlated response variance is that the estimates are small and have large variance. This poses the problem of how to present estimates to users in a meaningful way. Developments which produce estimators with improved stability should be pursued.

The investigation of outliers in this paper was one such an attempt to derive a more stable estimator. Future research might investigate this problem more thoroughly. It has been suggested that one source of instability is the variable sizes of the EAs. To compensate for this phenomenon, a new model could be developed in which the mean for an EA, rather than the total count, would be used as the basic unit of analysis. The current model has sample design complexities which are ignored in the estimators. Future work might consider incorporating these complexities into the calculation of CRV. Finally, this empirical investigation should be replicated on other datasets for a variety of variables in order to provide a more solid foundation on which to base evaluation of the improved method.

Footnotes

1. The EA is a spatial unit composed of a cluster of geographically contiguous households and assigned to one enumerator. In 1976, Canada was divided into 35,154 EAs.

2. In the 1976 Census, basic demographic data and mother tongue were collected on a 100% basis. Data on migration, education and labour force activity were based on a 1/3 sample.

3. A CCD on the average has 21 EAs.

References

[1] Bailey, L., T.F. Moore, and B.A. Bailar. "An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample". JASA, Vol. 73, 1978, pp. 16-23.

[2] Brackstone, G.J. and C.J. Hill. "The Estimation of Total Variance in the 1976 Census", Survey Methodology, Vol. 2, Number 2, 1976, pp. 195-208

[3] Fellegi, I.P. "Response Variance and its Estimation". JASA, Vol. 59, 1964, pp. 1016-1041.

[4] Fellegi, I.P. "An Improved Method of Estimating Correlated Response Variance", JASA, Vol. 69, 1974, pp. 496-501.

[5] Krótki, K.P. and C.J. Hill "Estimation of Correlated Response Variance". Paper prepared for annual meeting of ASA, 1978.

[6] Krótki, K.P., "1976 Census parametric evaluation: measurement of total variance for the 1976 Census - methodology and results report". Final report no. 18, Census Survey Methods Division, Statistics Canada, November 1978.

[7] U.S. Bureau of the Census. Evaluation and Research Program of the U.S. Censuses of Population and Housing (1960): Effects of Interviewers and Crew Leaders. Series ER 60, Number 7, Washington, D.C., 1968.