Cell Collapsing Strategies Based on Collapsing Adjustment Factors

Jay J. Kim, Office of Research and Methodology, National Center for Health Statistics
Richard Valliant, University of Michigan and Joint Program for Survey Methodology
Wenxing Zha, National Institute on Alcohol Abuse and Alcoholism, NIH

Abstract[1]

In poststratification, one of the cell collapsing criteria is a ratio criterion, where the ratio is the poststrafication factor or inverse coverage ratio. Quite often, if the ratio for a cell is greater than 2 or less than ½, then the cell is collapsed with another cell. However, this can introduce bias in a poorly covered group. Two censoring (or truncation) ratio approaches in collapsing were proposed and implemented in a simulation study (Kim, et al, 2005). The simulation study showed that the censoring approaches are better than the conventional approach mentioned above. In this paper, we propose four new collapsing strategies, two of which are based on the conditional bias and the other on the conditional mean square error.

Keywords: Cell collapsing criteria; coverage ratio; censoring; bias; mean square error; poststratification.

1. Introduction

Poststratification typically begins with a set of candidate cells that would all be used in estimation if the sample meets certain standards based on sizes of the sample and weight adjustments. If the standards are violated, cells are collapsed together. Although collapsing is common in practice, there is a limited literature on its effects. Little (1993) addressed cell collapsing in a Bayesian predictive modeling framework. We consider the problem from the design-based point-of-view, emphasizing the role of poststratification in correcting for undercoverage by the sample.

Cell collapsing has traditionally been performed based on the minimum raw sample counts for the cell/row/column and the ratio factor, i.e., the inverse coverage ratio for the cell/row/column. The ratio factor is also called the poststrafication factor. Kim, et al (2005) call the ratio factor the initial adjustment factor (IAF). Typically, if IAF is greater than 2 or less than .5 for a cell, it is collapsed with another cell. The traditional collapsing approach combines cells which are similar in content. However, Kim (2004) raised a potential problem of combining cells which are different in coverage ratios. Let f_i , i=1, 2, be the IAF for cells 1 and 2, i.e., $f_i = N_i / \hat{N}_i$, where N_i is the control count for cell i and $\hat{N}_i$ is the survey estimate prior to poststratification. $1/f_i$, i = 1, 2, is the coverage ratio for cells 1 and 2. Let $N_2 = c N_1$. Kim (2004) showed that when c = 10 and f_1 / f_2 = 4.0, cell 1 will lose 73 percent of its own weight to cell 2. For the same c, if f_1 / f_2 = 0.25, cell 1 will gain additional 214 percent of its own weight. This additional weight is from cell 2. Thus he showed that the current approach of cell collapsing can introduce bias, which can be large in certain occasions. He proposed methods for mitigating the adverse impact of the currently popular cell collapsing approaches. Kim, et al (2005) implemented in their simulation studies two approaches which involve censoring or truncation of large IAF's. One approach adopts hard censoring. That is, an IAF for a cell cannot exceed the censoring point. The other allows the IAF to go above the censoring point. The simulation showed that the latter, the soft censoring approach performs better than the former. Either one performs better than the most popular collapsing approach.

Kim's (2004) transfer factor, which was later renamed the collapsing adjustment factor (CAF) in Kim, et al (2005) deals with weight gain or loss. The collapsing adjustment factor for cell 1, CAF_1 is defined as

$$CAF_1 = \frac{f_2(1+c)}{cf_1 + f_2},$$

and the collapsing adjustment factor for cell 2, CAF_2 is

$$CAF_2 = \frac{f_1(1+c)}{cf_1 + f_2}.$$

[1] **Disclaimer:** This paper represents the views of the author and should not be interpreted as representing the views, policies or practices of the Centers for Disease Control and Prevention, National Center for Health Statistics.

Gonzalez, et al (2005) also applied a heuristic approach to BRFSS data for compensating for small f_i's, such as 0.52 to make the final adjustment factor of the cell to be somewhere near 0.8. The mean square error based on this approach was significantly lower than that based on the currently popular approach, or the do-nothing approach for revising the low IAF's. This paper treats Gonzalez, et al's approach in an optimal framework. That is, in this paper, we propose four new collapsing strategies, which are based on minimizing either bias or mean square error. The first is based on the conditional local bias, the second on the conditional local mean square error, the third on the conditional global bias and the last on the conditional global mean square error. In the above, "local" refers to the combined cells and "global" to all cells of the weighting matrix. These approaches are in line with Deville and Särndal (1992) in the sense that the final weights are developed such that they are closest to their initial weights.

2. Collapsing Approaches Based on Collapsing Adjustment Factors

Suppose we have the following table. We introduce some definitions.

Table 1 Notation for estimated counts, control counts, initial weighted and initial adjustment factors

	Estimated counts with initial weights	Control Count	IAF
Cell 1	$\hat{N}_1$	N_1	$f_1 = N_1 / \hat{N}_1$
Cell 2	$\hat{N}_2$	N_2	$f_2 = N_2 / \hat{N}_2$

We consider a general sample design that may include stratification and clustering, although some of the subsequent formulas are most easily understood in the context of simple random sampling.

There could be many cell collapsing approaches which utilize the collapsing adjustment factor (CAF). However, here we will concentrate on four approaches only, as mentioned above.

2.1 Collapsing Approach Based on Conditional Local Bias

Let x be a key characteristic of interest in a survey. Then, the estimated mean of the characteristic based on the combined cells is

$$\hat{\bar{x}}_c = \frac{1}{N_1+N_2}\left[\frac{f_2(1+c)}{cf_1+f_2}\hat{N}_1 f_1\bar{x}_1 + \frac{f_1(1+c)}{cf_1+f_2}\hat{N}_2 f_2\bar{x}_2\right],$$

where $\hat{\bar{x}}_c$ is the combined mean estimate of the characteristic and $\hat{\bar{x}}_i,\ i=1,2$ are the estimated cell means of the characteristic. The general form of an estimated mean is $\hat{\bar{x}}_i = \sum_{k\in s_i} w_k x_k \Big/ \sum_{k\in s_i} w_k$ where x_k is the value for unit *k*, s_i is the set of sample units for cell *i*, and w_k is the (unadjusted) survey weight.. Noting that $\hat{N}_i f_i = N_i$, i = 1, 2, $\hat{\bar{x}}_c$ can be re-expressed as follows.

$$\hat{\bar{x}}_c = \frac{1}{N_1+N_2}\left[\frac{f_2(1+c)}{cf_1+f_2}N_1\bar{x}_1 + \frac{f_1(1+c)}{cf_1+f_2}N_2\bar{x}_2\right] \qquad (1)$$

As mentioned before, $\frac{f_2(1+c)}{cf_1+f_2}$ in the above is the collapsing adjustment factor for cell 1 (CAF_1) and $\frac{f_1(1+c)}{cf_1+f_2}$ is the collapsing adjustment factor for cell 2 (CAF_2).

Note that if no collapsing is needed and two cells stand by themselves, the mean of the two cells is

$$\bar{x} = \frac{1}{N_1+N_2}\left[N_1\bar{x}_1 + N_2\bar{x}_2\right].$$

Assume that $\hat{\bar{x}}_i$ is approximately design-unbiased, i.e., $E\left(\hat{\bar{x}}_i\right) \doteq \bar{x}_i$, where $\bar{x}_i$ is the population cell mean. As mentioned in the papers by Kim (2004) and Kim, et al (2005), by combining two cells which have disparate coverage ratios, we can artificially shift weights, which could introduce bias, sometimes in great amounts.

Suppose cell 1 has a poorer coverage ratio than cell 2. In the spirit of Deville and Särndal's approach, we can readjust CAF_1 such that the mean of the collapsed cell due to a poor coverage ratio is as close as possible to that of the mean of the two independent cells. To do so, we multiply CAF_1 by "*k*." To get the *k*, we minimize the following squared conditional bias with respect to *k*.

$$\left[k\frac{f_2(1+c)}{cf_1+f_2}N_1\bar{x}_1 + \frac{f_1(1+c)}{cf_1+f_2}N_2\bar{x}_2 - (N_1\bar{x}_1 + N_2\bar{x}_2)\right]^2.$$

This is conditional in the sense that f_i is treated as fixed. The above can be expressed as

$$\left[k\,CAF_1\,N_1\bar{x}_1 + CAF_2\,N_2\bar{x}_2 - (N_1\bar{x}_1 + N_2\bar{x}_2)\right]^2$$

Differentiating the above with respect to *k* and setting the resulting expression equal to zero, we get

$$k = \frac{1}{CAF_1}\left[1 + c\,\frac{\bar{x}_2}{\bar{x}_1}(1 - CAF_2)\right], \qquad (2)$$

where $c = \dfrac{N_2}{N_1}$. *k* can also be expressed as follows.

$$k = \frac{cf_1 + f_2}{f_2(1+c)} + c\,\frac{\bar{x}_2}{\bar{x}_1}\,\frac{(f_2 - f_1)}{f_2(1+c)}, \qquad (3)$$

Theorem 1. If $\bar{x}_1 = \bar{x}_2$, $k = 1$.

Proof. In equation (2), if $\bar{x}_1 = \bar{x}_2$, $1 + c(1 - CAF_2)$ reduces to CAF_1.

Lemma 1. If $\bar{x}_2 > \bar{x}_1$, then $\dfrac{\bar{x}_2}{\bar{x}_1} = 1 + d$, where $d > 0$. Thus *k* can be expressed as

$$k = \frac{1}{CAF_1}\left[1 + c(1+d)(1 - CAF_2)\right]$$

$$= 1 + \frac{cd}{CAF_1}(1 - CAF_2)\ .$$

Since $CAF_2 > 1$, *k* is less than 1.

Lemma 2. If $\bar{x}_1 > \bar{x}_2$, then $\dfrac{\bar{x}_2}{\bar{x}_1} = 1 - e$, where $e > 0$. *k* can be expressed as

$$k = \frac{1}{CAF_1}\left[1 + c(1-e)(1 - CAF_2)\right]$$

$$= 1 + \frac{ce}{CAF_1}(CAF_2 - 1)\ .$$

Since $CAF_2 > 1$, *k* is greater than 1.

The above suggests that if cell 1 has a larger mean than cell 2, then using a number greater for cell 1 than CAF_1 will reduce local bias. Otherwise, using a smaller number for cell 1 than CAF_1 will reduce the bias.

After collapsing and weighting, the total weights for the combined cells should match the control total of those cells. That is, the following should hold.

$$\frac{f_2(1+c)}{cf_1 + f_2}\,f_1\hat{N}_1 + \frac{f_1(1+c)}{cf_1 + f_2}\,f_2\hat{N}_2 = \frac{f_2(1+c)}{cf_1 + f_2}\,N_1 + \frac{f_1(1+c)}{cf_1 + f_2}\,N_2$$

$$= N_1 + N_2$$

However, if we multiply CAF_1 by *k*, the sum would not be $N_1 + N_2$. That is,

$$k\,\frac{f_2(1+c)}{cf_1 + f_2}\,f_1\hat{N}_1 + \frac{f_1(1+c)}{cf_1 + f_2}\,f_2\hat{N}_2$$

$$= k\,\frac{f_2(1+c)}{cf_1 + f_2}\,N_1 + \frac{f_1(1+c)}{cf_1 + f_2}\,N_2$$

$$= (N_1 + N_2)\,\frac{cf_1 + kf_2}{cf_1 + f_2} \qquad (4)$$

$$\neq N_1 + N_2\ .$$

To restore equality in the above, we have to multiply equation (4) by $\dfrac{cf_1 + f_2}{cf_1 + kf_2}$.

That is,

$$k\,\frac{f_2(1+c)}{cf_1 + kf_2}\,f_1\hat{N}_1 + \frac{f_1(1+c)}{cf_1 + kf_2}\,f_2\hat{N}_2 = k\,\frac{f_2(1+c)}{cf_1 + kf_2}\,N_1$$

$$+ \frac{f_1(1+c)}{cf_1 + kf_2}\,N_2 = N_1 + N_2$$

In other words, the adjustment factor (post stratification factor) for cell 1 should be $\dfrac{kf_1 f_2(1+c)}{cf_1 + kf_2}$ and that for cell 2 $\dfrac{f_1 f_2(1+c)}{cf_1 + kf_2}$. The difference between this formula and that for the IAF for cell 1 is that this formula has a *k* in both the numerator and denominator, but for cell 2, the formula has a *k* in the denominator only.

Information is collected on many characteristics in a survey. It is thus recommended that *k* be computed for two or three key characteristics, then any one of them is picked or their average is used.

Example 1.

The poststratification factors adjusted for the survey non-response rate for Black males aged 18-24 and 25-34 for the 2001 BRFSS data (encompassing 44 counties which border with Mexico) are 10.22 and 3.86, respectively. In the following, we will observe what happens when the two cells are collapsed.

Table 2. Weighting Table

	N_i	$\hat{N}_i$	f_i
18-24	43,654	4,271.48	10.22
25-34	55,856	14,478.81	3.86

In this case,

c = 1.2795

CAF_1 is,

$$\frac{(3.85778)(2.2795)}{(1.2795)(10.2199)+3.85778} = .519\ .$$

What this factor indicates is, by combining the cells, the sample units in cell 1 (18-24) will lose approximately 48 percent of their own weights (the weights they would receive when the cell is not combined with the other cell).

CAF_2 is

$$\frac{(10.2199)(2.2795)}{(1.2795)(10.2199)+3.85778} = 1.376\,.$$

$CAF_2 =$ 1.376 implies that the units in cell 2 will get an additional 37.6 percent of above their own weight because of collapsing. The value of *k* is,

$$k = (.519^{-1})\left(1+1.2795\frac{\bar{x}_2}{\bar{x}_1}(1-1.376)\right)$$
$$= (1.92678)\left(1-.4807\frac{\bar{x}_2}{\bar{x}_1}\right).$$

If $\frac{\bar{x}_2}{\bar{x}_1}$ = 1.1, then *k* = .90743. The adjustment factor for cell 1 is

$$\frac{kf_1 f_2(1+c)}{cf_1+kf_2} = \frac{(.90743)(10.2199)(3.85778)(2.2795)}{(1.2795)(10.2199)+(.90743)(3.85778)}$$

$$= 4.92$$

The adjustment factor for cell 2 is $\frac{f_1 f_2(1+c)}{cf_1+kf_2}$ = 5.52.

If $\frac{\bar{x}_2}{\bar{x}_1}$ = .9, then *k* = 1.093. The adjustment factor for cell 1 is 5.679 and that for cell 2 is 5.198.

Without minimizing the local bias, the initial adjustment factor (IAF) for both cells is 5.307. As noted above, if the cells are kept separate, the adjustments would be 10.22 and 3.86.

In the practical case where there is more than 2 cells, the set of cells that are sparse are determined, i.e., the set of cells that must be collapsed. The analysis above then applies to a pair formed by a sparse cell and the nonsparse cell with which it is collapsed.

2.2 Collapsing Approach Based on Conditional Local Mean Square Error

In the case of simple random sampling (SRS), let σ_i^2, i = 1,2, be the variance of a survey variable in the cell i and n_i the sample size of cell i. The f_i's are variables. However, we again assume that the f_i's are constant to facilitate the derivation of the following variance formula. Under this assumption, $V\left(\hat{\bar{x}}_i\right)=\sigma_i^2/n_i$ either under SRS with replacement or SRS without replacement with a negligible sampling fraction. For a more general design, σ_i^2/n_i in the formulas below would be replaced by the more general $V\left(\hat{\bar{x}}_i\right)$. The variance of the mean of cell 1 after collapsing is

$$V_1 = V\left[\frac{f_2(1+c)}{cf_1+f_2}N_1\bar{x}_1\right]$$
$$= \frac{f_2^2(1+c)^2}{(cf_1+f_2)^2}N_1^2\frac{\sigma_1^2}{n_1}.$$

and that for cell 2 is

$$V_2 = V\left[\frac{f_1(1+c)}{cf_1+f_2}N_2\bar{x}_2\right]$$
$$= \frac{f_1^{\,2}(1+c)^2}{(cf_1+f_2)^2}N_2^2\frac{\sigma_2^2}{n_2}$$

We can derive a formula for a new "*k*" that minimizes the mean square error of the collapsed cell estimator. That is,

$$\underset{k}{Min}\left\{k^2V_1+V_2+\left[k\frac{f_2(1+c)}{cf_1+f_2}N_1\bar{x}_1+\frac{f_1(1+c)}{cf_1+f_2}N_2\bar{x}_2\right.\right.$$
$$\left.\left.-\left(N_1\bar{x}_1+N_2\bar{x}_2\right)\right]^2\right\}$$

As in section 2.1, the calculations below are conditional in the sense of treating f_i as fixed. Differentiating the above with respect to *k* and setting it to zero, we get

$$k=\frac{N_1\bar{x}_1+N_2\bar{x}_2\dfrac{f_2-f_1}{f_2(1+c)}}{\dfrac{f_2(1+c)}{cf_1+f_2}N_1\left(\dfrac{\sigma_1^2}{n_1\bar{x}_1}+\bar{x}_1\right)}\ . \qquad (5)$$

Using the fact that $N_2 = c\,N_1$, *k* can be expressed as

$$k=\frac{N_1\left[\bar{x}_1+c\bar{x}_2\dfrac{f_2-f_1}{f_2(1+c)}\right]}{\dfrac{f_2(1+c)}{cf_1+f_2}N_1\left(\dfrac{\sigma_1^2}{n_1\bar{x}_1}+\bar{x}_1\right)}=\frac{\bar{x}_1+c\bar{x}_2\dfrac{f_2-f_1}{f_2(1+c)}}{\dfrac{f_2(1+c)}{cf_1+f_2}\left(\dfrac{\sigma_1^2}{n_1\bar{x}_1}+\bar{x}_1\right)}\ .$$

Dividing both numerator and denominator above by $\bar{x}_1$, we have

$$k\ =\frac{(cf_1+f_2)+c\dfrac{\bar{x}_2}{\bar{x}_1}\dfrac{(f_2-f_1)(cf_1+f_2)}{f_2(1+c)}}{f_2(1+c)\left(\dfrac{\sigma_1^2}{n_1\bar{x}_1^2}+1\right)}. \qquad (6)$$

Equation (6) can also be expressed as

$$k\ =\frac{1+c\dfrac{\bar{x}_2}{\bar{x}_1}\left(1-CAF_2\right)}{CAF_1\left(\dfrac{\sigma_1^2}{n_1\bar{x}_1^2}+1\right)}\ .$$

In comparing the above with equation (2), we can note that the formula for *k* based on the mean square error has the additional term in the denominator, which is $CAF_1\dfrac{\sigma_1^2}{n_1\bar{x}_1^2}$.

Since the above expression is positive, the new *k* is smaller than the one in (2) that minimized the conditional local bias. The larger the relvariance of the cell 1 estimate, the smaller its adjustment.

As in the section 2.1, the adjustment factors (post stratification factors) for cells should be further modified to ensure that the total weights of the combined cell match its control total.

2.3 Collapsing Approach Based on Conditional Global Bias

This time we assume the overall weighted sample mean without collapsing is the population mean. Suppose only two cells need to be collapsed, which are not adjacent. For simplicity, we assume those two cells are cell 1 and cell 3. We also assume cell 1 is collapsed with cell 2, and cell 3 with cell 4.

Define

$$\beta_1=\frac{f_2(1+c_1)}{c_1f_1+f_2}\text{, where } c_1=N_2\,/\,N_1,$$

$$\beta_2=\frac{f_1(1+c_1)}{c_1f_1+f_2},$$

$$\beta_3=\frac{f_4(1+c_2)}{c_2f_3+f_4}\text{, where } c_2=N_4\,/\,N_3,$$

$$\beta_4=\frac{f_3(1+c_2)}{c_2f_3+f_4},$$

and

$$N=\sum_{i=1}^{L}N_i\ .$$

Note that $\beta_i = CAF_i$, i = 1, 2, 3, 4 in the above.

Suppose there are L cells. Then the conditional global bias is,

$$\frac{\beta_1 N_1 \bar{x}_1 + \beta_2 N_2 \bar{x}_2 + \beta_3 N_3 \bar{x}_3 + \beta_4 N_4 \bar{x}_4 + \sum_{i=5}^{L} N_i \bar{x}_i}{N} - \frac{\sum_{i=1}^{L} N_i \bar{x}_i}{N}$$

Note in the above, we assume that cells 1 and 3 have low coverage rates and thus lose weights when they are collapsed with cells 2 and 4, respectively. To alleviate the amount of lost weights in cell collapsing, CAF_1 is multiplied by k_1 and CAF_3 by k_2 as follows.

$$\frac{k_1 \beta_1 N_1 \bar{x}_1 + \beta_2 N_2 \bar{x}_2 + k_2 \beta_3 N_3 \bar{x}_3 + \beta_4 N_4 \bar{x}_4 + \sum_{i=5}^{L} N_i \bar{x}_i}{N}$$

$$- \frac{\sum_{i=1}^{L} N_i \bar{x}_i}{N} .$$

Since there are common terms in the above equation, it can be simplified as follows.

$$\frac{k_1 \beta_1 N_1 \bar{x}_1 + \beta_2 N_2 \bar{x}_2 + k_2 \beta_3 N_3 \bar{x}_3 + \beta_4 N_4 \bar{x}_4 - \sum_{i=1}^{4} N_i \bar{x}_i}{N} .$$

To obtain k_i, i = 1, 2, which minimize the squared conditional global bias, we have

$$\underset{k_1,\, k_2}{Min} \left[k_1 \beta_1 N_1 \bar{x}_1 + \beta_2 N_2 \bar{x}_2 + k_2 \beta_3 N_3 \bar{x}_3 + \beta_4 N_4 \bar{x}_4 - \sum_{i=1}^{4} N_i \bar{x}_i \right]^2 \qquad (7)$$

Partially differentiating the above equation with respect to k_1 and k_2, respectively, and setting the results to 0, we obtain,

$$k_1 = \frac{\sum_{i=1}^{4} N_i \bar{x}_i - \beta_2 N_2 \bar{x}_2 - k_2 \beta_3 N_3 \bar{x}_3 - \beta_4 N_4 \bar{x}_4}{\beta_1 N_1 \bar{x}_1} \qquad (8)$$

and

$$k_2 = \frac{\sum_{i=1}^{4} N_i \bar{x}_i - k_1 \beta_1 N_1 \bar{x}_1 - \beta_2 N_2 \bar{x}_2 - \beta_4 N_4 \bar{x}_4}{\beta_3 N_3 \bar{x}_3} \qquad (9)$$

Note the above two equations are derived from the same (one) equation. Thus k_1 and k_2 can not be uniquely solved by using equations (8) and (9) alone. Equation (2) can help. Note that, if there is only one cell which needs to be collapsed, that is, there is only one *k* to solve for, then *k* can be found, but if more than 3 cells need to be collapsed, this approach cannot provide the solutions. However, an iterative approach can be used in this situation. Suppose *p* cells need to be collapsed, that is, we have *p* unknown *k*'s (k_1, k_2, ... , k_p) to solve for. We will solve for one *k* each time and repeat the process until the solutions are reached. More specifically,

Iteration 1.

Collapse with another the first cell which violates the collapsing criteria. Leave the other cells not-collapsed. Solve for k_1. Collapse the next cell with another and solve for k_2 given k_1. Repeat until k_p given k_1, k_2 . . k_{p-1} is solved.

Iteration 2.

Using k_2, k_3, ... , k_p, solve for k_1. Solve for k_2 given k_1, k_3, ... , k_p. Repeat this process until the solution for k_p given k_1, k_2, ... , k_{p-1} is found.

If respective *k*'s in Iteration 1 and Iteration 2 do not differ more than a specified tolerance, stop. Otherwise, repeat Iteration 2 until successive solutions are near identical.

Note that in the above case we don't have unique solutions for *k*'s.

As before, for the two other approaches, the adjustment factors (post stratification factors) for cells 1 and 2 should be further modified to ensure that the total weights of the combined cell match its control total.

2.4 Collapsing Approach Based on Conditional Global Mean Square Error

In general, we have a number of sparse cells that are collapsed with different nonsparse cells. The CAFs for the sparse cells can be adjusted in such a way that the overall mean square of an estimated mean or total is minimized. The technique described in this section is a generalization of the one presented in section 2.2.

First, define the following:

S_{sp} = the set of L_{sp} sparse cells;

$S_{\overline{sp}}$ = the set of $L_{\overline{sp}}$ nonsparse cells;

$\mathbf{k}_{sp} = \left(k_1, \ldots, k_{L_{sp}}\right)^T$ = the vector of adjustments for the L_{sp} sparse cells;

$\mathbf{V}_{sp} = diag\left(V\left(\hat{\bar{x}}_i\right)\right)\ i \in S_{sp}$;

$\mathbf{V}_{\overline{sp}} = diag\left(V\left(\hat{\bar{x}}_i\right)\right)\ i \in S_{\overline{sp}}$;

$\mathbf{z}_{sp} = \left(z_i\right),\ i \in S_{sp}$ with $z_i = \beta_i N_i \bar{x}_i$; and

$\mathbf{z}_{\overline{sp}} = \left(z_i\right),\ i \in S_{\overline{sp}}$

If the weights of each unit in sparse cell *i* are adjusted by k_i while the weights in the nonsparse cells are unadjusted, then the bias of the full sample estimator of a total is

$$\sum\nolimits_{i \in S_{sp}} k_i z_i + \sum\nolimits_{i \in S_{\overline{sp}}} z_i - \sum\nolimits_{i=1}^{L} N_i \bar{x}_i =$$
$$\mathbf{k}_{sp}^T \mathbf{z}_{sp} + \mathbf{1}^T \mathbf{z}_{\overline{sp}} - t_+$$

where $t_+ = \sum\nolimits_{i=1}^{L} N_i \bar{x}_i$ and **1** is a vector of $L_{\overline{sp}}$ 1's. The values of k_i, *i* = 1, 2 can be found by minimizing the following mean square error:

$$\Phi = \mathbf{k}_{sp}^T \mathbf{V}_{\overline{sp}} \mathbf{k}_{sp} + \mathbf{1}^T \mathbf{V}_{\overline{sp}} \mathbf{1} + \left(\mathbf{k}_{sp}^T \mathbf{z}_{sp} + \mathbf{1}^T \mathbf{z}_{\overline{sp}} - t_+\right)^2 \quad (10)$$

The derivative of (10) with respect to $\mathbf{k}_{sp}$ is

$$\partial\Phi / \partial\mathbf{k}_{sp} \propto \mathbf{k}_{sp}^T \mathbf{V}_{\overline{sp}} + \left(\mathbf{k}_{sp}^T \mathbf{z}_{sp} + \mathbf{1}^T \mathbf{z}_{\overline{sp}} - t_+\right) \mathbf{z}_{sp}^T \quad (11)$$

Equating (11) to **0** and solving leads to

$$\mathbf{k}_{sp}^T = \left(t_+ - \mathbf{1}^T \mathbf{z}_{\overline{sp}}\right) \mathbf{z}_{sp}^T \left(\mathbf{V}_{sp} + \mathbf{z}_{sp} \mathbf{z}_{sp}^T\right)^{-1}.$$

As for the other three approaches, the adjustment factors (poststratification factors) should be further modified to ensure that the total of weights after collapsing match the grand total of all control counts. In particular, this can be accomplished by using

$$\mathbf{k}^* = \left(\mathbf{k} / \mathbf{k}^T \mathbf{u}\right) t_+,$$

where $\mathbf{k}^T = \left(\mathbf{k}_{sp}^T, \mathbf{1}^T\right)$ and $\mathbf{u} = \left(\mathbf{u}_{sp}, \mathbf{u}_{\overline{sp}}\right)$ with $\mathbf{u}_{sp} = \left(u_i\right),\ i \in S_{sp}$, $\mathbf{u}_{sp} = \left(u_i\right),\ i \in S_{\overline{sp}}$, and $u_i = \beta_i N_i$.

3. Concluding Remarks

Indiscriminately collapsing cells of a weighting matrix can introduce bias. Kim (2004), Kim, et al (2005) and Gonzalez, et al (2005) suggested how to correct this problem. Gonzalez, et al (2005) used a heuristic approach. This paper introduces four optimal approaches by which the adverse impact of the cell collapsing can be mitigated. The first is based on the conditional local bias, the second on the conditional local mean square error, the third on the conditional global bias and the last on the conditional global mean square error. The first approach has been implemented in Gonzalez, et al (2006) on BRFSS data. In terms of both bias and mean square error, the new approach proved to be significantly better than the traditional approach. The three other methods are yet to be tested empirically.

References

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

Gonzalez, J.F. Jr., Town, M. and Kim, J.J. (2005), Mean Square Analysis of Health Estimates from the Behavioural Risk Factor Surveillance System for Counties along the United States-Mexico Border Region, Proceedings of the American Statistical Association, Survey Methods Research Section, on CD.

Gonzalez, J.F. Jr., Town, M., Kim, J.J., Notzon, S. and Albertorio, J.R. (2006), Estimation and Reliability Issue of Health Estimates from the Behavioural Risk Factor Surveillance System for US Counties Contiguous to the US – Mexico Border. To appear on the Proceedings of the American Statistical Association, Survey Methods Research Section, on CD.

Kim, J.J. (2004), Effects of Collapsing Rows/Columns of Weighting Matrix on Weights, Proceedings of the American Statistical Association, Survey Methods Research Section, on CD.

Kim, J.J., Tompkins, L., Li, Jianzhu and Valliant, R. (2005), A Simulation Study of Cell Collapsing in

Poststratification, Proceedings of the American Statistical Association, Survey Methods Research Section, on CD.

Little, R.J.A. (1993), Post-stratification: A Modeler's Perspective, *Journal of the American Statistical Association*, **88**, 1001 – 1012.