Minimizing Conditional Local Bias for Health Estimates from the Behavioral Risk Factor Surveillance System for U.S. Counties Contiguous to the United States-Mexico Border

Joe Fred Gonzalez, Jr.[a], Machell Town[b], Jay J. Kim[a]
Centers for Disease Control and Prevention
National Center for Health Statistics[a]
National Center for Chronic Disease Prevention and Health Promotion[b]

Summary[1]

The Behavioral Risk Factor Surveillance System (BRFSS) is a State telephone based survey of the civilian non-institutionalized adult (18 years and over) population residing in the United States. Consequently, the BRFSS final weights that are currently available in the data files are designed to produce unbiased estimates of socio-demographic and health characteristics for adults at the State level (Gonzalez, et al, 2005). In addition to State level BRFSS estimates, there is interest in the health status of adults residing in the 25 U.S. counties contiguous to the United States-Mexico Border region (Arizona, California, New Mexico, and Texas.) The purpose of this paper is to investigate alternative ways of arriving at post-stratification factors (ratio adjustments) by collapsing the weighting matrix by age-sex-ethnicity/race for producing final weights/estimates for this border region. A modified optimal approach which minimizes local (cell) squared bias was applied to BRFSS data (25 contiguous counties). Then, a conditional mean square error analysis was used to observe the effect of cell collapsing (in tandem with the optimal bias approach) on the absolute bias and variance estimators for several BRFSS socio-demographic and health characteristics.

Keywords: unbiased estimation, poststratification

1. Introduction

The BRFSS is a State telephone based survey of the civilian non-institutionalized adult (18 years and over) population residing in the United States. However, there is interest in another geographical subpopulation, the 25 U.S. counties contiguous to the United States-Mexico Border (Arizona, California, New Mexico, and Texas.). The map in Figure 1 displays the "sister cities" along both sides of the United States-Mexico Border. Figure 2 shows a map of the actual counties that are contiguous to the United States-Mexico Border.

It was determined that it would be worthwhile to produce BRFSS estimates for the adult population in the border region by certain age-sex-ethnicity/race cells. The desired six age groups were: 18-24, 25-34, 35-44, 45-54, 55-64, and 65 and over. The desired three ethnicity/race groups were: Hispanic, White Non-Hispanic, and Non-Hispanic Black/Multiracial and others. In previous work (Gonzalez, et al, 2005 and 2006), BRFSS sample counts were tabulated by age-sex-ethnicity/race within each border county. Although sample counts were insufficient for some cells within each border county for the current estimation research, BRFSS county level estimation techniques have been investigated (Jia, et al, 2004) and have been produced (Jia, et al, 2006). For detailed documentation for producing county level estimates, the reader is referred to: BRFSS's SMART (Selected Metropolitan/Micropolitan Area Risk Trends) data from metropolitan/micropolitan statistical areas. The URL for these data is http://apps.nccd.cdc.gov/brfss-smart/SelMMSAPrevData.asp. The SMART home page is http://apps.nccd.cdc.gov/brfss-smart/index.asp.

For the current estimation research, sample sizes were aggregated by the desired age-sex-ethnicity/race cells for the 25 counties contiguous to the United States-Mexico Border (Arizona, California, New Mexico, and Texas.). At the border region level, cell sizes were sufficiently large for the desired age-sex-ethnicity/race cells for both Hispanics and White Non-Hispanics, and in a few instances for Non-Hispanic Black/Multiracial and others. This level of geographical aggregation was defined as the United States-Mexico Border for the purpose of our paper. Hereafter, the United

[1] **Disclaimer:** This paper represents the views of the authors 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, or the National Center for Chronic Disease Prevention and Health Promotion.

States-Mexico Border Region will be simply referred to as the "border region." In addition, the same age-sex-ethnicity/race crosstabulation that was used for determining sample size sufficiency was also used as the weighting matrix for this investigation.

This paper will focus on a modified conditional local minimum bias strategy for calculating 2001 poststratification factors by investigating alternative ways of collapsing cells by age-sex-ethnicity/race for producing final weights/estimates for this border region. A modified optimal approach which minimizes local (cell) squared bias was applied to BRFSS data (25 contiguous counties). Then, a conditional mean square analysis was used to observe the effect of cell collapsing (in tandem with the modified optimal bias approach) on the bias and the root mean squared error (RMSE) of estimates of health characteristics of U.S. adults (18+ years).

2. Sample Weighting Procedures for the Border Region

Post-stratification is used for incorporating population distributions of key socio-demographic variables into survey estimates. One reference about post-stratification is Kim (2004) "Effect of Collapsing Rows/Columns of Weighting Matrix on Weights."

For this analysis, the variable _WT2, which is available in the 2001-2003 BRFSS data sets is the initial sample weight as follows:

$$\_WT2 = \_STRWT * NAD / NPH$$

where,

STRWT = within State stratum weight,
NAD = number of adults in household, and
NPH = number of phones in the household.

For purposes of this investigation, the initial sample weight (_WT2) was used to create the "initial poststratification factors (PSF)" which were calculated in the usual manner by age (6 groups)-sex(2)-ethnicity/race (Hispanic, White Non-Hispanic, and Non-Hispanic Black/Multiracial and Others) as follows:

PSF = Census pop. count within an i-th cell / sum of _WT2 within same i-th cell.

The "initial poststratified Final Weights" used in this investigation were calculated for the year 2001 as follows:

"Final_ Weight" = _WT2*PSF

where PSF is as previously defined.

The usual approach, c*onventional cell collapsing* was used. This approach is usually driven by sample size considerations (here, minimum cell count, raw cell count = 20), and maximum ratio criteria (original PSF) by domains, and row adjacency. The new method applies the previously mentioned criteria, and in addition, *the modified local minimum bias approach*. Table 1 shows the maximum ratio criteria used for conventional collapsing for 2001 data for the 25 contiguous counties.

The "Final_ Weights" were used to produce 2001 BRFSS percent estimates of adult characteristics using the following binary health variables for adults (18+ years of age):

- Ever had Asthma
- Ever had high blood pressure
- High cholesterol
- Diabetes
- Having health insurance
- Current smoker
- Any exercise.

3. Conditional Bias and Mean Square Error Analysis

First, we will introduce the notation involved in doing a mean square error (MSE) analysis as follows:

$$MSE\ (p) = [Bias\ (p)]^2 + [se\ (p)]^2$$

where p= percent estimator of a health characteristic, and se (p) = standard error estimator for the percent estimator of the same health characteristic.

The percent estimates of health characteristics using the "initial poststratified Final_Weights" are unbiased estimates and treated as "*parameters,*" that is, as true values of health characteristics for the adult population in the border region for this mean square error (MSE) analysis. So, in reality, the bias and RMSE analysis is *conditiona*l. The bias and RMSE analysis was performed by comparing these "*parameters*" of health

characteristics with corresponding percent estimates of health characteristics generated by: applying the *local (cell) minimum bias strategy* described later followed by investigating the effects on the RMSE of the same estimates.

"New" PSF, corresponding Final Weights, and corresponding percent estimates were produced by using the above approach.

Table 2 (Kim, 2004) defines the quantities that are involved for producing PSF using *conventional cell collapsing.*

Table 3 shows an example of initial PSFs for row 1 and row 2 where PSF = $f_i = N_i / W_i$ where all quantities are as previously defined in Table 2.

For the sake of illustration, suppose that we collapse row 1 and row 2 and assume that the row population counts are the same, that is, $N_1 = N_2$ (Kim, et al, 2006). What would be the revised PSF for each row? The revised PSF for row 1 would be

$$\frac{N_1 + N_2}{W_1 + W_2} = \frac{2N_1}{5W_1} = \frac{2}{5} f_1 \ .$$

That is, by collapsing rows 1 and 2, row 1 has lost 3/5 of its original population count.

Similarly, the revised PSF for row 2 would be

$$\frac{N_1 + N_2}{W_1 + W_2} = \frac{8N_2}{5W_2} = \frac{8}{5} f_2 \, .$$

That is, by collapsing rows 1 and 2, row 2 has gained 3/5 of its original population counts.

The ratios 2/5 and 8/5 are the *collapsing adjustment factors (CAFs)* for the above example.

A generalization of *CAFs* by Kim (2004) follows. Let $N_2 = c\, N_1$ where c > 0. The revised PSF in terms of the original PSF (f_1) for row 1 is:

$$\frac{N_1 + N_2}{W_1 + W_2} = \frac{f_2(1+c)}{cf_1 + f_2} f_1$$

where

$$\frac{f_2(1+c)}{cf_1 + f_2} = CAF_1 \ \ for\ row\ 1.$$

Similarly,

$$\frac{f_1(1+c)}{cf_1 + f_2} = CAF_2 \ \ for\ row\ 2.$$

What follows is a possible remedy to avoid shifting potentially large population counts from one row to another when collapsing rows. We refer to this as a modified local (cell) minimum bias strategy. In this strategy, an expression for the squared bias was developed by multiplying CAF_1 (for the under covered cell 1) by (1+k), and similarly, CAF_2 (for the over covered cell 2) by (1- k). Thus, we obtain the following expression for the conditional squared bias

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

which can be expressed as

$$\left[(1+k) CAF_1 \, N_1 \bar{x}_1 + (1-k)\, CAF_2 \, N_2 \bar{x}_2 - (N_1 \bar{x}_1 + N_2 \bar{x}_2) \right]^2 .$$

The above expression was minimized by differentiating with respect to k (Kim, 2007), setting the derivative equal to zero, and solving for k, we obtain:

$$k \ = \ \frac{(1 - CAF_1) \ + \ c \, \frac{\bar{x}_2}{\bar{x}_1} (1 - CAF_2)}{CAF_1 \ - \ c \, \frac{\bar{x}_2}{\bar{x}_1} CAF_2}$$

If we multiply the first term of the expression for the squared bias by (1 + k) and the second term by (1 - k), the weighted sum would not be $N_1 + N_2$. Thus, we need an adjustment by a factor as follows:

$$\frac{cf_1 + f_2}{(1+k) f_2 + (1-k) cf_1} .$$

Hence, the final PSF for cell 1 is

$$\frac{(1+k)(1+c) f_2}{(1+k) f_2 + (1-k) cf_1} f_1 ,$$

and the final PSF for cell 2 is

$$\frac{(1-k)(1+c) f_1}{(1+k) f_2 + (1-k) cf_1} f_2 .$$

4. Results

Tables 4–6 show the results of comparing the performance of conventional collapsing with the new method for 2001 white non-Hispanic, Hispanic, and Black/ Multiracial/Others for both sexes combined, all ages combined, respectively. Overall, Table 4 shows that the new method performed better for White Non-Hispanic, both sexes combined, and all ages combined, in terms of bias and RMSE. The seven (7) total in each of Tables 4-6 refers to the fact that we investigated seven (7) health variables listed in Section 2.

Tables 7–8 show the results of comparing the performance of conventional collapsing with the new method for 2001 white non-Hispanic, Hispanic, for individual age-sex groups, respectively. Overall, in terms of bias, the new method performed much better for white non-Hispanic and Hispanic, individual age-sex groups, as shown in Tables 7-8. In terms of RMSE, the new method performed much better for Hispanic, individual age-sex groups as shown in Table 8.

5. Concluding Remarks

In addition to providing state-level estimates of adult health characteristics, BRFSS data can be used to produce unbiased estimates for specific geographic areas, namely, for adults residing along the U.S.-Mexico Border.

A limitation of the new method is that the analysis in this investigation is conditional on estimates using original PSFs, instead of actual parameter values.

Another limitation of the new approach is that it is variable dependent, i.e., it depends on the ratio of cell means (included in expression for squared bias which was previously given) for a specific health variable. For this paper, analysis was done by optimizing with respect to the variable "percent any exercise."

In this paper, we optimized with respect to local bias, however, we examined overall bias and RMSE for both sexes and all age groups combined. Comparable health estimates for the U.S. counties contiguous to the United States-Mexico Border can be produced using the new methodology for race/ethnicities and age-sex groups.

The methodology for optimizing bias globally is currently available, but has not been implemented.

References

Gonzalez, Joe F.; Town, Machell; Kim, Jay J. (2005) Mean Square Error Analysis of Health Estimates from the Behavioral Risk Factor Surveillance System for Counties along the United States-Mexico Border Region, Proceedings of the American Statistical Association, Survey Research Methods Section.

Gonzalez, Joe F.; Town, Machell; Kim, Jay J. (2006) Estimation and Reliability Issues of Health Estimates from the Behavioral Risk Factor Surveillance System for U.S. Counties Contiguous to the United States-Mexico Border, Proceedings of the American Statistical Association, Survey Research Methods Section.

Jia, Haomiao; Muennig Peter; Borawski, Elaine. (2004) Comparison of Small-Area Analysis Techniques for Estimating County-Level Outcomes, American Journal of Preventive Medicine; 26 (5):453-60.

Jia, Haomiao; Link, Michael; Holt, James.; Mokdad, Ali H.; Li, Lee; Levy, Paul S. (2006) Monitoring County-Level Vaccination Coverage During the 2004-2005 Influenza Season, American Journal of Preventive Medicine; 31 (4):275-280.

Kim, Jay J. (2004) Effect of Collapsing Rows/Columns of Weighting Matrix on Weights, Proceedings of the ASA Survey Research Methods Section.

Kim, Jay J.; Valliant, Richard; Zha, Wenxing, (2006) *Cell Collapsing Strategies based on Collapsing Adjustment Factor*. Proceedings of the American Statistical Association, Survey Research Methods Section.

Kim, Jay J. (2007) Alternative Approach for Collapsing Using a Collapsing Adjustment Factor, NCHS Internal Memorandum

Figure 1.

Figure 2.

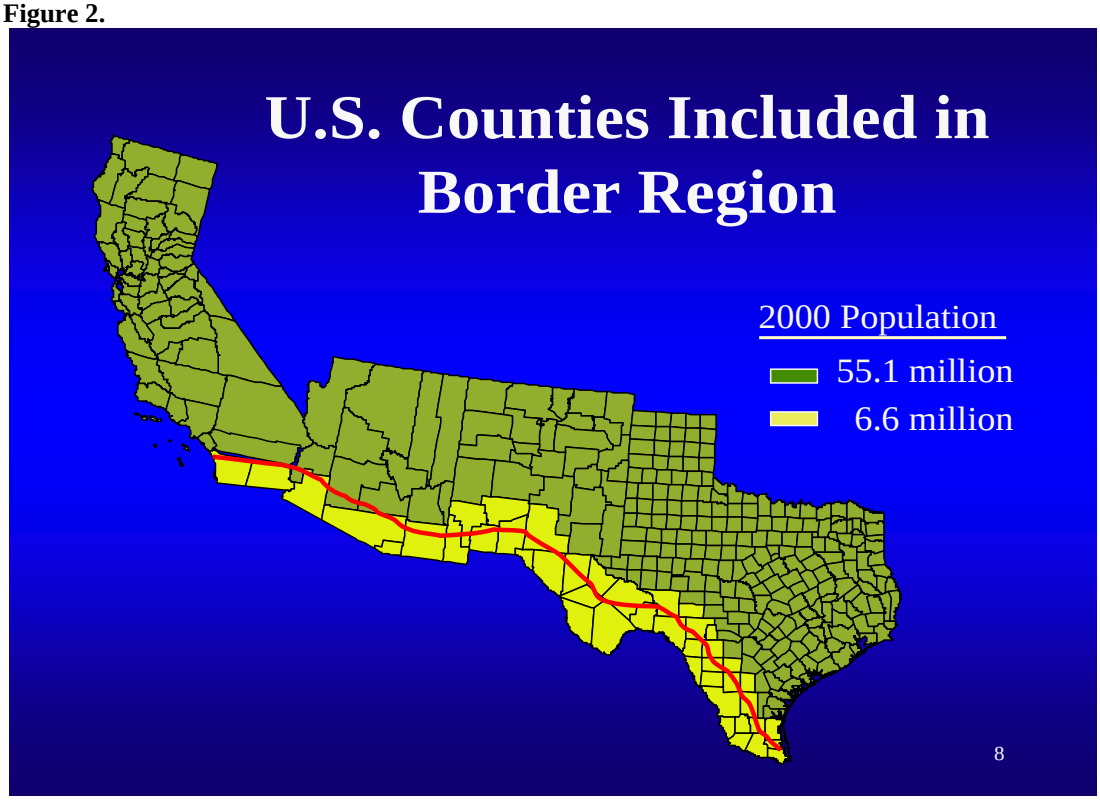

Table 1. Maximum Ratio Criteria Used for Conventional Collapsing for 2001 Data Year (25 contiguous counties) with a Minimum Cell Count of 20

Ethnicity/Race Group	Maximum Ratio
Hispanic	4
White (Non-Hispanic)	4
Non-Hispanic Black/Multiracial/Others	8

Table 2. Weighting Matrix for Calculating Usual PSF.

Rows	Raw Sample Count	Initially Weighted Sample Count	Control Count
Row 1	n_1	W_1	N_1
Row 2	n_2	W_2	N_2

Table 3. Conventional Collapsing Example (PSF = $f_i = N_i / W_i$)

Row 1	$f_1 = 4$
Row 2	$f_2 = 1$

Table 4. Performance of Conventional Collapsing vs. New Method for 2001 White-Non Hispanic , Both Sexes, All Ages.

Measure	Conventional Collapsing Better	New Method Better	Total
Bias	**0**	**7**	**7**
RMSE	**0**	**7**	**7**

Table 5. Performance of Conventional Collapsing vs. New Method for 2001 Hispanic , Both Sexes, All Ages.

Measure	Conventional Collapsing Better	New Method Better	Total
Bias	**5**	**2**	**7**
RMSE	**6**	**1**	**7**

Table 6. Performance of Conventional Collapsing vs. New Method for 2001 Black/Multiracial/Others, Both Sexes, All Ages.

Measure	Conventional Collapsing Better	New Method Better	Total
Bias	**6**	**1**	**7**
RMSE	**5**	**2**	**7**

Table 7. Performance of Conventional Collapsing vs. New Method for 2001 White-Non Hispanic, Individual Age-Sex Groups.

Measure	Conventional Collapsing Better	New Method Better	Total
Bias	**0**	**13 (1 tie)**	**14**
RMSE	**5**	**9**	**14**

Table 8. Performance of Conventional Collapsing vs. New Method for 2001 Hispanic, Individual Age-Sex Groups.

Measure	Conventional Collapsing Better	New Method Better	Total
Bias	**8**	**13**	**21**
RMSE	**14**	**7**	**21**