Minimizing Conditional Global MSE 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
[a]National Center for Health Statistics, 3311 Toledo Rd, Hyattsville, MD 20782
[b]National Center for Chronic Disease Prevention and Health Promotion, 4770 Buford Hwy, NE MS K-40
Atlanta, GA 30341-3717

Abstract[1]

The Behavioral Risk Factor Surveillance System (BRFSS) is a State-based telephone survey of the adult civilian non-institutionalized population residing in the United States. Consequently, the BRFSS final weights that are currently available in the public use data files are designed to produce unbiased estimates of health conditions by socio-demographic characteristics at the State level. 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. The purpose of this paper is to apply an alternative approach of poststratification by minimizing conditional global mean squared error of BRFSS health estimates for adults residing in the combined 25 counties contiguous to the United States - Mexico border.

Keywords: unbiased estimation, root mean squared error, poststratification

1. Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) is a State-based telephone survey of the civilian non-institutionalized adult (18 years and over) population residing in the United States. However, there is also interest in another geographical subpopulation, the combined 25 U.S. counties contiguous to the United States-Mexico Border States (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, Non-Hispanic White, and Non-Hispanic Black/Multiracial and others. In previous work (Gonzalez, et al, 2005, 2006, and 2007), BRFSS sample counts were tabulated for the year 2001 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 developed (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 (http://apps.nccd.cdc.gov/brfss-smart/SelMMSAPrevData.asp). For the current estimation research, sample sizes were aggregated by the target age-sex-ethnicity/race cells for the combined 25 counties contiguous to the United States-Mexico Border (Arizona, California, New Mexico, and Texas.). At the border region level, sample sizes were sufficiently large for the desired age-sex-ethnicity/race cells for both Hispanics and Non-Hispanic Whites, and in a few instances for Non-Hispanic Black/Multiracial and others. Hereafter, the United 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.

[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.

This paper will focus on a conditional global minimum MSE strategy for calculating 2001 poststratification factors by investigating alternative ways of collapsing cells by age-sex-ethnicity/race for producing final weights and estimates of health conditions of U.S. adults (18+ years) along the border region. Then, the effect of this alternative collapsing approach on the MSE and root mean squared error (RMSE) of health estimates will be determined.

2. Sample Weighting Procedures for the Border Region

Post-stratification is used for incorporating population distributions of key socio-demographic variables into survey estimates. For a reference on poststratification, see Kim (2004) "Effect of Collapsing Rows/Columns of Weighting Matrix on Weights." The variable _WT2, which is the initial sample weight from the 2001-2003 BRFSS data sets, is defined 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.

The initial sample weight (_WT2) was used to create the "initial poststratification factors (PSF)" which were calculated by age (6 groups)-sex (2)-ethnicity/race (Hispanic, Non-Hispanic White, and Non-Hispanic Black/Multiracial and Others) as follows:
PSF_i = 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:

$Final\_Weight_i = \_WT2*PSF_i$.

In this paper two approaches are compared. The first approach is c*onventional cell collapsing*. 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 second approach, *the conditional global minimum MSE method*, also employs the previously mentioned criteria.

The Final_Weights were used to produce the 2001 BRFSS prevalence estimates of the following conditions among adults (18+ years of age):

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

3. Background of Post Stratification Strategies and Previous Work

A.) Weighting class methods have traditionally been used based on the minimum sample counts for the cell/row/column and the ratio factor or post stratification factor (PSF) within a cell;
B.) The post stratification factor is the ratio of the control count to the initially weighted sample count or the inverse of the coverage ratio;
C.) Traditional approach combines homogeneous cells which are similar in content and characteristics.
D.) However, Kim (2004) discovered a potential problem with combining cells that have different coverage ratios.

E.) Gonzalez, et al (2005) investigated a heuristic approach and conditional local bias approach (2006, 2007) and found out that this approach performs better than the conventional approach in bias and MSE.

3.1 Conditional Global Minimum Mean Squared Error (MSE) Analysis

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

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

where p = percent estimator of a health condition, bias = P - p, where P= population parameter, and se (p) = standard error estimator for the percent estimator of the same health condition. Also, root mean squared error of p = RMSE (p) = $\sqrt{\text{MSE (p)}}$.

The prevalence of health conditions using the initial poststratified Final_Weights are unbiased estimates and treated as "*parameters,*" that is, as true values of health conditions for the adult population for this mean square error (MSE) analysis. Since we are using estimates as parameters, this bias and RMSE analysis is *conditiona*l on our sample. The bias and RMSE analysis was performed by comparing these "*parameters"* of health conditions with corresponding prevalence estimates of health conditions generated by *Minimizing Conditional Global MSE analysis* 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.

4. Methodology for Collapsing Two or More Cells

Let $N_2 = c\, N_1$, where c is a constant and N_i is the control count for cell *i*, *i* = 1, 2.

Let $f_i = N_i / \hat{N}_i$ be the post stratification factor (PSF) for cell *i* . $\hat{N}_i$ is the initially weighted sample count for cell *i.*

Collapsing Adjustment Factor (CAF) for cell 1 is defined as

$$\beta_1 = \frac{f_2(1+c)}{cf_1 + f_2},$$

and CAF for cell 2,

$$\beta_2 = \frac{f_1(1+c)}{cf_1 + f_2}.$$

4.1 Two Cells Are Collapsed

Suppose 2 cells are collapsed with another 2 cells.

Let $V_i = V\left[\beta_i N_i \bar{x}_i\right] = \beta_i^2 N_i^2 \frac{\sigma_i^2}{n_i}$, i = 1, 2. (1)

Note than σ_i^2 are the cell variances based on the global (over entire weighting matrix) mean and $\bar{x}_i$ is the sample mean for cell i. Let k_i, i = 1, 2, be further adjustment factors for the collapsing adjustment factors. The k_i's can be found by minimizing the following mean square error, assuming only two cells need to be collapsed with another two cells, as before.

$$k_1^2 V_1 + V_2 + k_2^2 V_3 + V_4 + \left(k_1\beta_1 N_1 \bar{x}_1 + \beta_2 N_2 \bar{x}_2 + k_2\beta_3 N_3 \bar{x}_3 + \beta_4 N_4 \bar{x}_4 - \sum_{i=1}^{4} N_i \bar{x}_i \right)^2 \quad (2)$$

Differentiating expression (2) with respect to k_1 and k_2 and solving a system of two equations in the two unknowns k_1 and k_2, we have

$$k_1 = \frac{\beta_1 N_1 \bar{x}_1 \left(\sum_{i=1}^{4} N_i \bar{x}_i - \beta_2 N_2 \bar{x}_2 - k_2 \beta_3 N_3 \bar{x}_3 - \beta_4 N_4 \bar{x}_4 \right)}{V_1 + \left(\beta_1 N_1 \bar{x}_1 \right)^2} \tag{3}$$

and

$$k_2 = \frac{\beta_3 N_3 \bar{x}_3 \left(\sum_{i=1}^{4} N_i \bar{x}_i - k_1 \beta_1 N_1 \bar{x}_1 - \beta_2 N_2 \bar{x}_2 - \beta_4 N_4 \bar{x}_4 \right)}{V_3 + \left(\beta_3 N_3 \bar{x}_3 \right)^2}. \tag{4}$$

To simplify further, let

$I_i = \beta_i N_i \bar{x}_i$, i = 1, 2, 3, 4,

Then from equation (3),

$$k_1 = \frac{I_1 \left(\sum_{i=1}^{4} N_i \bar{x}_i - I_2 - k_2 I_3 - I_4 \right)}{V_1 + \left(I_1 \right)^2} \tag{5}$$

Similarly, from equation (4), we have

$$k_2 = \frac{I_3 \left(\sum_{i=1}^{4} N_i \bar{x}_i - k_1 I_1 - I_2 - I_4 \right)}{V_3 + \left(I_3 \right)^2} \tag{6}$$

Denoting $\sum_{i=1}^{4} N_i \bar{x}_i - I_2 - I_4$ by *A*, and after some algebra, equations (5) and (6) can be simplified to as

$$k_1 = \frac{A I_1 V_3}{V_1 V_3 + I_3^2 V_1 + V_3 I_1^2} \tag{7}$$

$$k_2 = \frac{A I_3 V_1}{V_1 V_3 + I_3^2 V_1 + V_3 I_1^2} \tag{8}$$

4.2 Three or More Cells Are Collapsed

If there are six cells and three of which are collapsed with another three, we have

$$k_1^2 V_1 + V_2 + k_2^2 V_3 + V_4 + k_3^2 V_5 + V_6 + \left(k_1 \beta_1 N_1 \bar{x}_1 + \beta_2 N_2 \bar{x}_2 + k_2 \beta_3 N_3 \bar{x}_3 + \beta_4 N_4 \bar{x}_4 k_2 + k_3 \beta_5 N_3 \bar{x}_3 + \beta_6 N_4 \bar{x}_4 - \sum_{i=1}^{6} N_i \bar{x}_i \right)^2 \tag{9}$$

Differentiating expression (9) with respect to k_1, k_2, and k_3 and solving a system of three equations in three unknowns k_1, k_2, and k_3, we have

$$k_1 = \frac{A I_1 V_3 V_5}{V_1 V_3 V_5 + I_1^2 V_3 V_5 + I_3^2 V_1 V_5 + I_5^2 V_1 V_3} \tag{10}$$

$$k_2 = \frac{A I_3 V_1 V_5}{V_1 V_3 V_5 + I_1^2 V_3 V_5 + I_3^2 V_1 V_5 + I_5^2 V_1 V_3} \tag{11}$$

and

$$k_3 = \frac{AI_5V_1V_3}{V_1V_3V_5 + I_1^2V_3V_5 + I_3^2V_1V_5 + I_5^2V_1V_3}. \quad (12)$$

In this case, $A = \sum_{i=1}^{6} N_i\bar{x}_i - I_2 - I_4 - I_6$.

In general, when p cells are combined with another p cells, we have the following:

$$D = \prod_{i=1}^{p} V_{2i-1} + \sum_{i=1}^{p} I_i^2 \prod_{j\neq i}^{p} V_{2i-1}$$

and

$$k_i = \frac{AI_i \prod_{j\neq i}^{p} V_{2i-1}}{D}, \qquad \text{i = 1, 2, 3, . . . , p,}$$

where, $A = \sum_{i=1}^{2p} N_i\bar{x}_i - \sum_{i=1}^{p} I_{2i}$.

[NOTE: This MSE analysis was conducted in two ways, using original PSFs and truncating PSFs with a minimum value of 0.8 and a maximum value of 2.0. See Tables 2-3.]

5. Results

Table 1 shows the ratio of the RMSE of estimates applying the conditional global minimum MSE approach to the corresponding RMSE for estimates applying the conventional collapsing approach. [**NOTE**: Only domains where there was a difference in the RMSE are shown in Tables 1 and 2.] A ratio less than one indicates that the global minimum approach performed better than the conventional collapsing approach. A ratio greater than one indicates the opposite relationship. Table 2 is similar to Table 1, except that the PSFs were truncated with a minimum value of 0.8 and a maximum value of 2.0. Table 3 summarizes the results of Tables 1-2. In general, the conventional collapsing approach performed better in terms of RMSE than the proposed new method.

6. Conclusions

A.) A limitation of the new method: analysis is conditional on estimates using original PSFs, instead of parameter values.

B.) Another limitation of the new approach is that it is variable dependent, i.e., it depends on the ratio of cell means for a specific variable. For this paper, analysis was done by optimizing with respect to the variable "percent having health insurance."

C.) The approach we adopted here is to try to optimize globally in terms of mean square error. However, global optimization does not necessarily optimize at the local level and the statistics we compared were local ones, or at the domain level. Thus, this approach does not do as well as the conventional approach.

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.

Gonzalez, Joe F.; Town, Machell; Kim, Jay J. (2007) 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, 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. Map of United States-Mexico Border Showing Major "Sister Cities."

Figure 2. U.S. Counties along United States-Mexico Border.

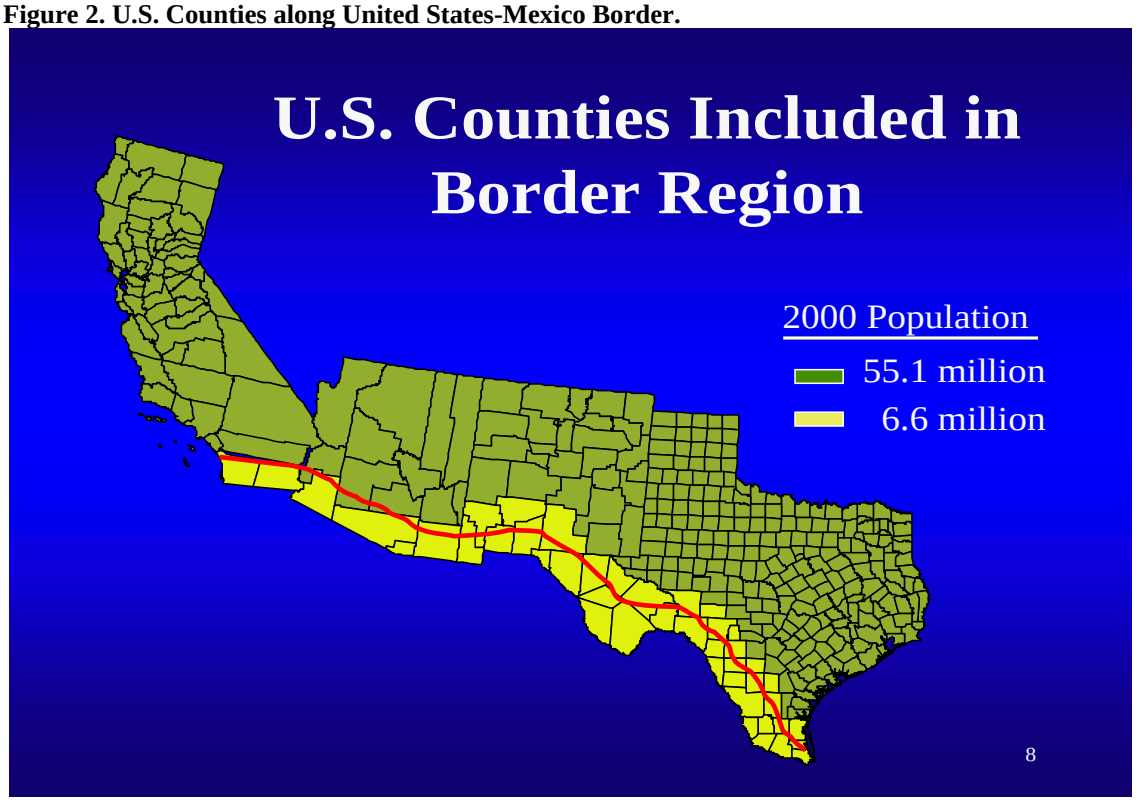

Table 1. Ratio of (RMSE Global Minimum, Using Original PSF) / (RMSE Conventional Collapsing) by Race-Sex-Age and Seven Health Conditions, 2001 BRFFS.

Race-Sex-Age	Asthma	High BP	High CHOL	Diabetes	Insurance	Current Smoker	Any Exercise
NH-WM							
25-44	1.121	0.953	1.221	1.270	0.937	0.928	1.094
45-64	1.003	1.017	0.985	1.004	1.053	1.007	0.930
NH-BM							
25-44	1.394	1.390	1.401	1.163	1.335	1.276	1.637
45-64	1.274	1.256	1.194	1.356	1.394	1.247	1.171
NH-BF							
25-44	1.392	0.793	1.267	1.284	1.171	1.196	1.216
45-64	1.011	1.016	0.938	1.195	0.923	0.940	1.062
HM							
25-44	0.863	1.067	1.147	0.940	1.054	1.075	1.092
45-64	0.932	1.002	0.976	1.038	1.004	1.022	0.975
HF							
45-64	1.028	0.981	0.987	1.005	0.986	1.007	0.981

[NOTE: Only domains where there was a difference in the RMSE are shown in Table 1.]

Table 2. Ratio of (RMSE Global Minimum, using Truncated PSF) / (RMSE Conventional Collapsing) by Race-Sex-Age and Seven Health Conditions, 2001 BRFSS.

Race-Sex-Age	Asthma	High BP	High CHOL	Diabetes	Insurance	Current Smoker	Any Exercise
NH-WM							
25-44	1.129	0.951	1.245	1.291	0.934	0.926	1.105
45-64	1.003	1.017	0.985	1.004	1.053	1.007	0.930
NH-BM							
25-44	1.063	1.063	1.054	0.967	1.049	1.035	1.057
45-64	0.987	1.029	1.020	1.068	1.005	1.039	1.026
NH-BF							
25-44	1.083	0.799	1.062	1.062	1.004	1.040	1.047
45-64	0.995	0.999	0.965	1.055	0.970	0.962	1.012
HM							
25-44	0.912	0.972	1.064	0.911	1.004	1.027	1.027
45-64	1.065	1.031	1.032	0.999	1.010	0.995	1.032
HF							
45-64	1.028	0.981	0.987	1.005	0.986	1.007	0.981

[NOTE: Only domains where there was a difference in the RMSE are shown in Table 2.]

Table 3. Performance of Global Minimum MSE Approach vs. Conventional Collapsing for Seven Health Conditions for Several Age-Sex-Ethnicity/Race Domains, 2001 BRFSS.

Health Condition	Using Original PSF		Using Truncated PSF	
	Global Better	Conventional Better	Global Better	Conventional Better
Asthma	2	7	3	6
H-BP	3	6	5	4
H-CHOL	4	5	3	6
Diabetes	1	8	3	6
Health Insurance	3	6	3	6
Current Smoker	2	7	3	6
Any Exercise	3	6	2	7