Empirical Evaluation of Raking Ratio Adjustments for Nonresponse

Ismael Flores Cervantes[1] and J. Michael Brick[1]
[1]Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Abstract

Raking ratio adjustments are used to benchmark sampling weights to known control totals for a variety of reasons. Raking is used to reduce sampling error through the use of auxiliary variables correlated to survey response. In recent years, raking has been used to reduce non-response bias. It also produces weights that have face validity. The use of raking can expedite the creation of analysis weights by simplifying the number of weighting adjustments. This paper is an empirical evaluation of analysis weights created by raking sampling weights without nonresponse adjustments. The raked based weights are compared with analysis weights created by applying sequential weighting adjustments at each stage of nonresponse in a telephone survey (*i.e.*, screener interview nonresponse, extended interview nonresponse). As part of the evaluation of these weights, estimates and their standard errors are computed and compared to determine if there are significant differences for a wide range of variables.

Keywords: Raking, nonresponse adjustment, weighting.

1. Weighting Process

In survey sampling methodology, the two main methods to deal with nonresponse are weighting and imputation. In weighting, analysis weights are created and applied to the data of respondents to compensate for the data not observed due to nonresponse. The analysis weights are (generally) larger than the base weights. Respondents, *r*, are viewed as the result of a two-phase selection. In the first phase, the sample *s* is drawn from a population *U* with a known individual probability of inclusion in the sample, denoted as π_k. In the second phase, the set of respondents' *r* is the result of a selection from *s*. Each element in the sample *s* (and in the population) has its own response propensity θ_k, $(0 \le \theta_k \le 1)$ that has an unknown $q(r \mid s)$ distribution. This approach is known as "quasi-randomization theory" the selection of respondents in the second phase is assumed to be random (Oh and Sceuren, 1983).

In the weighting approach, the estimate of the total of characteristic y_k in the population, $Y = \sum_U y_k$ is computed as

$$\hat{Y} = \sum_r w_k y_k = \sum_r d_k \hat{\phi}_k y_k ,$$

where w_k is the analysis weight, d_k is the design base weight for the sampling unit *k*, $d_k = 1/\pi_k$. The estimated response influence is $\hat{\phi}_k = 1/\hat{\theta}_k$, and $\hat{\theta}_k$ is the estimate of the response propensity θ_k. The estimate $\hat{Y}$ is the extension of the Horvitz-Thompson estimator to a two-phase sampling. There are several ways to compute $\hat{\phi}_k$. A common method is the response homogeneity group (RHG) or weighting classes that assumes that the population can be classified into groups where all elements in the group respond independently with the same probability. The weighting classes are created using variables that are available for respondents and nonrespondents.

In practice, many surveys include additional subsampling procedures that are implemented to increase the efficiency of the data collection. In this case, the two-phase model described before is not directly applicable. However, the quasi-randomization idea can be generalized for each stage of data collection. In this case, the estimate of the total in the population, is expressed as

$$\hat{Y} = \sum_r \left(\prod_{i=1}^{I} d_k^{(i)} \prod_{j=1}^{j} \hat{\phi}_k^{(i)} \right) y_k = \sum_r D_k \; \Phi_k \; y_k = \sum_r w_k \; y_k \tag{1}$$

where $d_k^{(i)} = 1/\pi_k^{(i)}$ is the design weight computed as the inverse of the conditional probability of selection $\pi_k^{(i)}$ at the stage *(i)*, $D_k = \prod_i d_k^{(i)}$ is the product of all design based adjustments, $\phi_k^{(j)} = 1/\hat{\theta}_k^{(j)}$ is the inverse of the response propensity $\hat{\theta}_k^{(j)}$ of the observed respondent *k* at stage *j*, and $\hat{\Phi}_k = \prod_j \hat{\phi}_k^{(i)}$ is the product of all nonresponse adjustments applied to unit *k*.

2. Raking

In most surveys, the weight w_k is further adjusted (or benchmarked) to external information, so the analysis weight w_k^* is computed as $w_k^* = w_k\ g_k^*$ where g_k^* is the final adjustment factor. Procedures currently used in surveys to incorporate auxiliary information in estimation are postratification, raking, and in generalized regression estimator.

Raking was developed by Deming and Stephan (1940) as a tool for improving the face validity of survey estimates by matching the distribution of weights to known marginal totals. Raking is a special case of calibration methods (Deville, et al., 1993).

In raking, estimates are controlled to marginal population totals. It can be thought of as a multidimensional poststratification because the weights are poststratified to one set (a dimension) of control totals, and then these adjusted weights are poststratified to another dimension. The procedure continues until all dimensions are adjusted. The process is then iterated until the control totals for all dimensions are simultaneously satisfied (at least within a specified tolerance).

An important advantage of raking over other simpler adjustment methods such as poststratification is that it is suitable to bring in data from different sources, with multiple characteristics (*e.g.*, race, ethnicity, sex, and own/rent). Raking also allows bringing in data at different levels of geography so that adjustments to population totals at the state and the counties can be made simultaneously.

Generalizing, the expression (1) that includes a raking adjustment in the last step of weighting is

$$\hat{Y}_c = \sum\nolimits_r D_k\ \hat{\Phi}_k\ g_k y_k = \sum\nolimits_r w_k^* y_k \tag{2}$$

where g_k is the raking adjustment of $D_k\ \hat{\Phi}_k$ that meets the calibration equation

$$\sum\nolimits_r D_k\ \hat{\Phi}_k\ g_k\ \mathbf{x}_k = \sum\nolimits_r w_k^* \mathbf{x}_k = \mathbf{X}\,,$$

where $\mathbf{x}_k$ is the vector of auxiliary variables known for all respondents. The raked weights $w_k^* = D_k\ \hat{\Phi}_k g_k$ are computed by solving a nonlinear optimization problem (Deville and Särndal, 1994, Dupont, 1994).

3. Alternative Approach

The estimator (2) is a function of both the probability of selection of sampling units at different stages and estimates of response propensities of respondents at different stages. Operationally, these adjustments are sequentially applied to the base weights. Although most of the sample design weighting factors are determined at the design stage (*i.e.*, subsampling rates, probability of selection of sampling units), the estimation of the response propensities is computed using the observed data at each weighting step. As the number of stages increases, a survey can include a large number of factors.

There are advantages associated with simplifying the weighting process reducing the number of weighting adjustments. In particular, a single weighting adjustment is much faster to implement and may have a lower cost.

The alternative estimator is

$$\hat{Y}_c = \sum\nolimits_r D_k\ g_k^c\ y_k = \sum\nolimits_r w_k^c y_k \tag{3}$$

where g_k^c is the raking adjustment of D_k that meets the calibration equation

$$\sum\nolimits_r D_k\ g_k^c\ \mathbf{x}_k = \sum\nolimits_r w_k^c\ \mathbf{x}_k = \mathbf{X}\,.$$

Comparing the estimators (2) and (3), is equivalent to comparing the raking factor g_k to $g_k^c / \hat{\Phi}_k$. This comparison evaluates how well a raked weight without nonresponse adjustment does relative to a raked weight with additional nonresponse adjustments.

Sverchov et al (2005) examined a similar problem comparing regression estimators of totals with weights adjusted for nonresponse propensity to unadjusted weights. They proved that calibrated nonresponse adjusted estimates vary slightly from calibrated estimates that ignore the nonresponse adjustment. They suggested that the nonresponse adjustment step could be omitted from the weighting process. However, they assumed the same set of variables were used for nonresponse adjustments and raking. In our study, there are variables that are available only for nonresponse adjustments.

4. Survey Data

In this study, we used data from the 2007 California Health Interview Survey (CHIS). CHIS is a list assisted telephone sample that collects data on public health and access to health care in California. The CHIS sample is large (around 50,000 completed interviews), and was allocated to produce estimates by county. The final weight is produced by applying 12 adjustments to the base weight and then raking to 11 dimensions. For our study, we considered the interviews from counties as separate surveys (the sample was selected independently by county). We selected four counties to represent four surveys with different sample sizes and populations. Table 1 summarizes the counties or surveys for this study.

Table 1: Variables used for nonresponse adjustment and raking dimension

Survey	*Stratification*	*Completed interviews*	*Characteristics*
Los Angeles (LA)	Service Planning Area (6)	11,106	Large minorities
Sacramento (SC)	None	1,450	Medium size
San Francisco (SF)	None	907	Few child/teen interviews, large nonresponse
Marin (MR)	None	571	Small

Source: UCLA Center for Health Policy Research, 2007 California Health Interview Survey

The full weighting process is summarized in Table 2. For more detail, see California Health Interview Survey, 2008). Table 2 classifies the adjustments into two groups based on the purpose of the adjustment. The design-based group includes all adjustments for sampling. The second group has all adjustments for some form of nonresponse (*i.e.*, telephone, household, and person level).

Table 2: Weighting adjustments

			Adjustment	
	Weight/ Adjustment	*Mean*	*Design based*	*Non-response*
1	*Base weights*	35.52	✓	
2	*New work*	1.02		✓
3	*Refusal Conversion*	1.84	✓	
4	*Unknown residential status*	1.37		✓
5	*Unknown Eligibility status*	1.03		✓
6	*Unknown presence of children*	1.2		✓
7	*Screener interview nonresponse*	3.07		✓
8	*Multiple telephone*	0.97	✓	
9	*Section G adjustment*	1.77		✓
10	*Person base weight*	1.94	✓	
11	*Extended interview nonresponse rate*	1.63		✓
12	*Trimming*	0.99		
13	*Raking*	1.44		
14	*Raked weight*	444.16		

Source: UCLA Center for Health Policy Research, 2007 California Health Interview Survey.

We evaluated estimates that used raked weights without any of the nonresponse adjustments.

5. Auxiliary Variables

Table 3 shows the variables that were used in CHIS survey for nonresponse adjustments, raking, or for both. Some of the variables that are only one column are correlated at some degree. For example, the variable that identifies whether the screener respondent is the sampled adult is correlated to gender (woman are more likely to answer the telephone). The indicator for presence of children is correlated to the number children in the household. The only two variables without strong correlation with the variables used in the raking dimensions are the indicator of a mailing address and the interviewer assessment.

Table 3: Variables used for nonresponse adjustment and raking dimension

Variable	*Non-response adjustment*	*Raking*
Geographic area	✓	✓
Sex	✓	✓
Age categories	✓	✓
Having a mailing address	✓	
Screener respondent is sampled adult	✓	
Presence of children	✓	
Refusal conversion status	✓	
Interviewer assessment	✓	
Household tenure		✓
Education		✓
Race-ethnicity		✓
Number of children in household		✓
Number of adults in household		✓

Source: UCLA Center for Health Policy Research, 2007 California Health Interview Survey.

6. Alternative Weights

We created three sets of weights for evaluation purposes and these are shown in Table 4. All these weights were raked to the same set of control totals used in CHIS. The differences among the sets are the number and type of adjustment factors applied to the base weight before raking.

The first set of weights is called *FULL*. It is the weight used in CHIS. All adjustment factors in Table 3 are included. We do not assume that *FULL* is the "best", but it uses the largest set of variables for the adjustments.

The second weight is called *DSGN* and it only includes the design base adjustments listed in Table 2 in addition to raking. This alternative weight includes the five adjustments design-based adjustments. Since the design-based adjustments are known, the pre-raked weight does not need to be computed sequentially.

The third weight is called *BASE* and it only includes the household and person base weights. This weight is for reference, and it is not intended to substitute any of the previous weights. However, this weight can be used to separate the effect of the design and nonresponse adjustment.

7. Evaluation of Weights

We begin by examining the distribution of the weights. Figure 1 shows all scatter plot combinations produced with the three weights for two surveys (Los Angeles and San Francisco).The plots show the dispersion is greater between the *FULL* weight and the other two weights than between the *BASE* and *DSGN* weights. The plot for *BASE* and *DSGN* shows two lines in addition to the centerline; this pattern is the result of subsampling factors not included in *BASE*. The scatter plots for the other surveys follow the same pattern.

Table 5 shows the descriptive statistics for the three set of weights and surveys. The mean weights are equal for all three methods because the last step in each is raking to the same control totals. The distribution of the *DSGN* weights is similar to the *FULL* weights, but the FULL weights have greater dispersion (see the standard deviations).

Next, we examine the weights with indicators used in calibration proposed by Särndal and Ludström (2005). They proposed indicators to determine the best auxiliary data for the calibrated estimator. We studied these indicators to provide information on the properties of these weights.

The first indicator, $IND1$, measures the variability of the estimated inverse of the response propensities, υ_k, computed using raking. The expression of $IND1$ is

$$IND1 = \frac{\sum_r d_k \left(\upsilon_k - \bar{\upsilon}\right)^2}{\sum_r d_k},$$

where $\bar{\upsilon} = \sum_r d_k \upsilon_k / \sum_r d_k$ is the design-based weighted mean of υ_k. It is not straightforward to compute $IND1$ because we have three components: the design-based factor $\left(D_k\right)$, the nonresponse adjustments $\left(\hat{\Phi}_k\right)$ and the raking factor (g_k .or g_k^*). We modified the indicator to evaluate the weights with $\upsilon_k = g_k$ and $d_k = D_k \ \Phi_k$, for the weight *FULL*; and $\upsilon_k = g_k^*$, and $d_k = D_k$ for the weights *DSGN* and *BASE*.

The interpretation of the modified indicator is not straightforward. For weight *FULL*, this indicator measures the residual variation in the response propensities after adjusting for nonresponse. If the nonresponse adjustments had been done before raking, then the raking factor may only captures the residual variability of the nonresponse propensities. If the nonresponse adjustments are effective, then we expect this indicator to be small. For the weight DSGN, the indicator measures how much variability in the response propensities is captured by the raking factor when the weights have not been adjusted for nonresponse.

Table 4: Variables used for nonresponse adjustment and raking dimension

Weight	Description (all include the same raking adjustment)
FULL	Includes all weighting adjustments applied in CHIS
DSGN	Ignores all nonresponse adjustments and uses all design base adjustments
BASE	Uses only the household and person base weights

Table 5: Distribution of *FULL*, *DSGN*, and *BASE* weights by survey

Interview	*Survey*	*Sample size*	*Weight*	*Mean*	*Standard deviation*	*Minimum*	*Maximum*
Adult	*Los Angeles*	11,106	*FULL*	659.8	813	11	10,026
			DSGN	659.8	734	12	8,281
			BASE	659.8	713	23	8,364
	Sacramento	1,450	*FULL*	691.5	676	16	4,893
			DSGN	691.5	640	16	5,705
			BASE	691.5	620	17	4,914
	San Francisco	907	*FULL*	742.6	865	30	14,642
			DSGN	742.6	805	40	13,355
			BASE	742.6	691	76	7,434
	Marin	571	*FULL*	331.5	350	34	3,411
			DSGN	331.5	327	42	2,792
			BASE	331.5	310	40	3,005

The second indicator, $IND2$, is based on a predictive approach that assumes a model that relates the variable y_k and the auxiliary variables $\mathbf{x}_k$. The indicator $IND2$ is computed as

$$IND2 = 1 - \frac{\sum_r w_k \left(y_k - \hat{y}_k\right)^2}{\sum_r w_k \left(y_k - \bar{y}_k\right)^2}$$

where $\hat{y}_k$ is the predicted value and $\bar{y}_k$ is the sample-based estimate of the mean $\bar{Y}$ both computed using the calibrated weights.

Table 6: Variables used for nonresponse adjustment and raking dimension

		Weight	
Indicator/Survey	*FULL*	*DSGN*	*BASE*
IND1			
Los Angeles	0.4	29.5	33.3
Sacramento	0.6	14.0	16.3
San Francisco	0.5	28.3	35.2
Marin	0.4	8.2	9.9
IND2			
Los Angeles	71.7	71.0	70.8
Sacramento	71.2	70.9	70.6
San Francisco	74.2	73.4	73.3
Marin	57.6	57.0	56.6

Table 6 shows the values for the two indicators. The indicators were computed for a large number of variables. The table shows the average across the variables. We think that this approach is more relevant in multipurpose surveys where a large number of variables are collected. The very low values of *IND1* for *FULL* suggest that the pre-raked nonresponse adjusted weights capture most the variability, and the raking factor only accounts a very small variability of the adjustment. We expect that raking is largely correcting for undercoverage and increasing the precision of the estimates but not adjusting very much for nonresponse. If the implicit model used in the weighting classes is correct, then raking should not be adjusting for nonresponse bias in the *FULL* approach.

The values of the indicator *IND2* are very similar for the three weights across of the survey. Since *IND2* measures how well an assumed model fits the observed data, these results suggest the three weights have almost the same fit. These results suggest that when the same or highly correlated variables are used for nonresponse and raking, there is not much gain if the weights are not adjusted for nonresponse before raking.

In the last part of the analysis, we computed estimates of total and proportions using the three weights. Figure 2 shows all scatter plots combinations of 703 estimates of proportions and totals computed using the three weights for Los Angeles. The estimates of proportions and totals from the different weights are nearly identical including the BASE estimates. Figure 3 shows the scatter plots for the standard errors of the same proportions and totals in Figure 1. These plots show differences in the standard errors, being more noticeable for totals. The plots do not show a trend, that is, the standard errors of *DSGN* or *BASE* estimates are consistently larger or smaller than the FULL standard errors.

Table 7: Percentage of statistically different estimates of proportions and totals by survey

			Number of	*Percentage of estimates with significant differences*	
Type of estimate survey		*Sample size*	*estimates*	*DSGN %*	*BASE %*
Proportions	*Los Angeles*	11,106	733	8	14
	Sacramento	1,450	414	9	11
	San Francisco	907	332	7	11
	Marin	571	228	8	8
Totals	*Los Angeles*	11,106	733	7	17
	Sacramento	1,450	414	9	10
	San Francisco	907	332	11	12
	Marin	571	228	16	10

We tested if the difference of these estimates were statistically significant. These tests are not an indicator of bias but of differences between estimates. The results of these tests for proportions and totals for the four surveys are presented in Table 7. The table shows the number of estimates and the percentage that are statistically different with respect to the *FULL* weight. The percentage of differences between *FULL* and *DSNG* is lower for estimates of proportions than for totals. Except for estimates of totals for San Francisco and Marin, less than 10 percent of the *FULL* estimates are statistically different from the *DSGN* estimates.

We also examined the magnitude of the differences for the estimates that were statistically different. We calculated the relative difference between of the *DSGN* and *FULL* estimates with respect to the *FULL* estimate. These are not estimates of the relative bias since neither of the estimates is the true value. Table 8 shows the average of the relative differences across the estimates that were statistically different. The relative difference in proportions is less than 1 percent for all surveys except for Los Angeles. This means that for example, for an estimate of proportion of 0.50, the other estimate is 0.505. Although the estimates of proportions are statistically different, the differences are small. For totals, the average relative difference is less than 10 percentage points that may be important for some estimates.

6. Conclusions

Although it is difficult to generalize results from an empirical study, these results suggest that if the information used for both nonresponse adjustment sand raking is the same, it is likely that one single adjustment will produce estimates consistent with the one that has all full adjustments.

Table 8: Relative differences of statistically different estimates of totals and proportions

Type of estimate survey		*Number of estimates statistically different*	*Relative difference*	
			Estimates %	*Standard error %*
Proportions	Los Angeles	59	1.9	0.9
	Sacramento	37	0.9	-4.6
	San Francisco	23	0.3	-3.5
	Marin	18	0.2	-1.4
Totals	Los Angeles	51	0.6	1.9
	Sacramento	37	-7.3	1.5
	San Francisco	37	-4.4	0.4
	Marin	36	-2.1	0.4

References

California Health Interview Survey. (2008) C*HIS 2007 Methodology Series: Report 5 – Weighting and Variance Estimation*. Los Angeles, CA: UCLA Center for Health Policy Research, 2008. http://www.chis.ucla.edu/

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table *Annals of Mathematical Statistics* 1940; 14: 427-444.

Deville, J.C., and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87, 376382.

Deville, J.C., Särndal, C.E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Dupont, F. (1994) Calibration Used as a Nonresponse Adjustment, IN: Diday, E. (ed.) *New Approaches in Classification and Data Analysis,* Springer Verlag, pp.539-548.

Oh, H.L. and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In: W.G. Madow, I. Olkin and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2. New York: Academic Press, 143-184.

Särndal, C.E.,and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Sverchkov, M., Dorfman, A.H., Ernst, L.R., Moerhle, T.G., Paben, S.P., and Ponikowski C.H. (2005). On Non-Response Adjustment via Calibration, *Proceedings of the Survey Research Methods Section of the American Statistical Association* (CD-ROM).

Figures

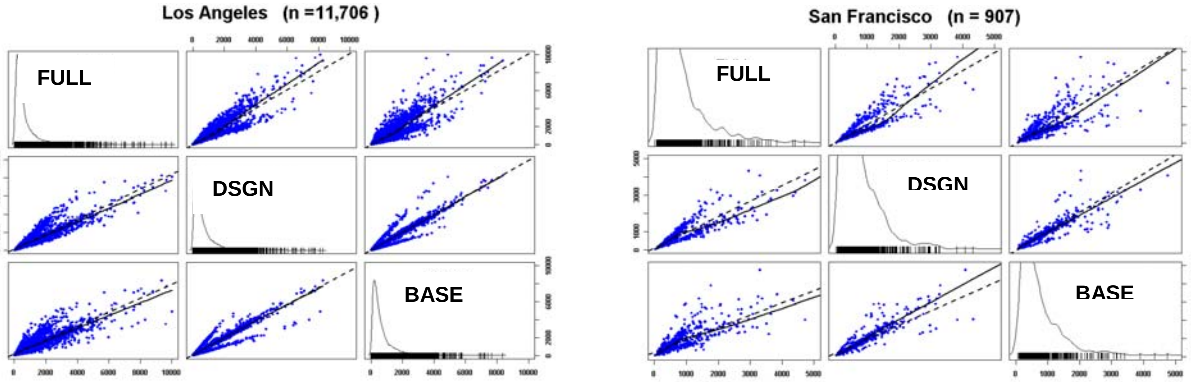

Figure 1: Scatter plots for *FULL*, *DSGN*, and *BASE* weights for Los Angeles and San Francisco

Proportions

Totals

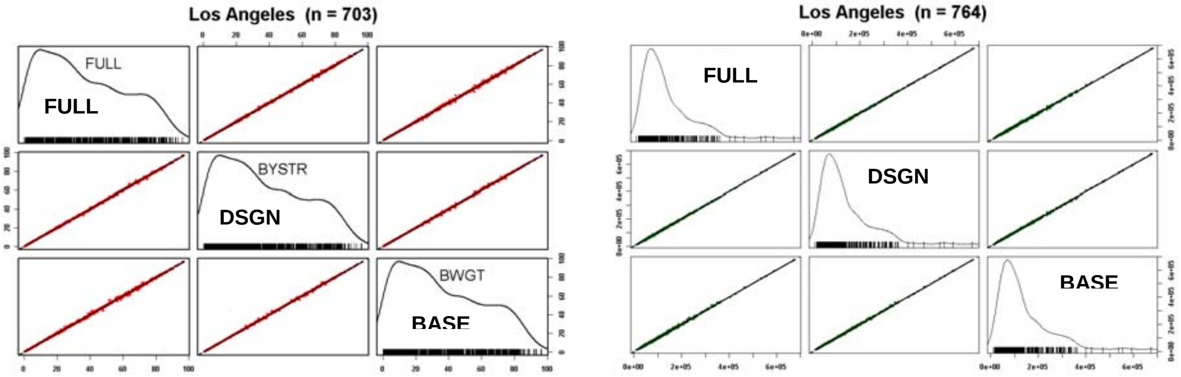

Figure 2: Scatter plots for estimates of proportions and totals for Los Angeles and San Francisco

Standard Errors of Proportions

Standard Errors of Totals

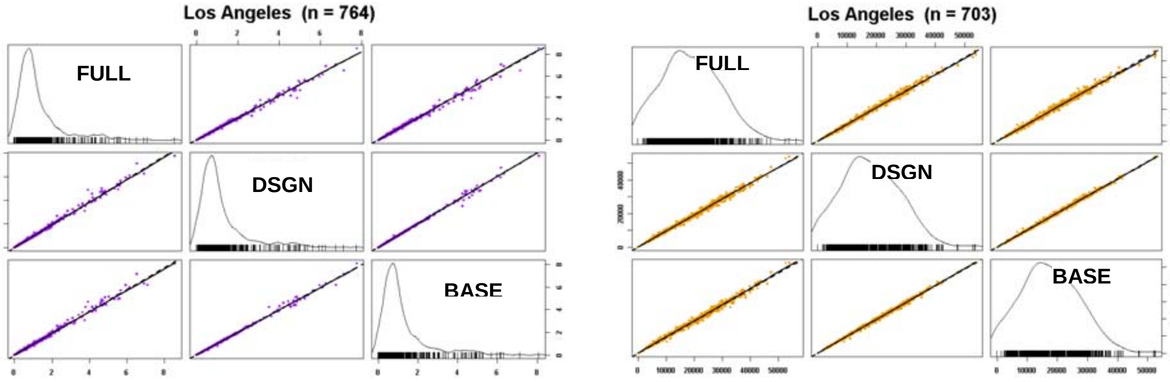

Figure 3: Scatter plots for estimates of standard errors of proportions and totals for Los Angeles and San Francisco