Comparing Alternative Weight Adjustment Methods

Kimberly Henry[1] and Richard Valliant[2]
[1]Statistics of Income, Internal Revenue Service
[2]Joint Program in Survey Methodology, University of Michigan.
[1]77 K Street NW, Washington DC, 20001, Kimberly.A.Henry@irs.gov.

Abstract: Several design-based, model-based, and model-assisted methods have been developed to adjust survey weights for nonresponse or coverage errors, to reduce variances through the use of auxiliary data or by restricting the range of the weights themselves. Some methods directly change the weights, like calibration weighting and design-based ad hoc weight trimming methods. Other methods implicitly adjust the weights, like robust superpopulation modeling approaches. The generalized design-based method models the weights as a function of the survey response variables and using the smoothed weights predicted from the model to estimate finite population totals. This paper provides empirical examples of how several adjustment methods change a given sample's weights and the resulting impact on estimates.

Key words: model-assisted estimation, model-based estimation, weight trimming, p-splines, weight smoothing

1. Introduction: Differential Sample Weights

Survey weights are important since they reflect the various sample design decisions made to select the sample units and a number of practical issues that arise when the data are collected and cleaned. These issues can be both planned and unplanned (Henry and Valliant 2012) and occur in the data collection and post-data collection stages in sampling. Practitioners also tend to think more in terms of weights, even though thinking in terms of estimators makes more statistical sense. There are many weighting methods that differ based on the desired type of inference. We compare the weights produced from some of these methods in a particular problem.

For all alternatives considered here, we will write estimators of totals in the form, $\hat{T} = \sum_{i \in s} w_i y_i$ where *s* is the set of sample units, w_i is a weight for unit *i*, and y_i is a data value. Practitioners often think of weighting as a distinct step in survey processing—not entirely divorced from estimation but somewhat removed from it. In contrast, some of the "robust" alternatives we cover are geared toward improving estimates for specific *y*'s. As we illustrate, the weights for those alternatives depend on *y*. Consequently, writing them as weighted sums of *y*'s may seem awkward but allows comparisons to be made with other weighting approaches.

2. Alternative Weighting Methods

2.1. Design-based Methods

HT Estimator. We start with base weights, or Horvitz-Thompson (HT, 1952) weights. These weights are the inverse of the probability of selection for each sample unit i, i.e., $w_i = \pi_i^{-1}, \pi_i = P(i \in s)$ where s denotes a probability sample of size n drawn from a population of N units. The HT estimator for the finite population total for a finite population total of the variable of interest y is then

$$\hat{T}_{HT} = \sum_{i \in s} y_i / \pi_i = \sum_{i \in s} w_i y_i \ . \qquad (1)$$

This estimator is unbiased for the finite population total in repeated πps sampling, but can be quite inefficient due to variation in the selection probabilities if π_i and y_i are not closely related. Alternative sample designs, such as probability proportional to a measure of size, introduce variable probabilities of selection in (1). The variability in selection probabilities can increase under complex multistage sampling and multiple weighting adjustments. Influential observations in estimating a population total using (1) can arise simply due to the combination of probabilities of selection and survey variable values. Thus, the HT-based estimates from one particular sample may be far from the true total value, particularly if the probabilities of selection are negatively correlated with the characteristic of interest (see discussion in Little 2004).

Poststratification. Here the HT weights are adjusted such that they add up to external population counts by available domains. This adjustment is used to correct an imbalance than can occur between the sample design and sample completion, i.e., if the sample respondent distribution within the external categories differs from the population (e.g., subgroups respond or are covered by the frame at different rates), as well as reduce potential bias in the sample-based estimates. Denoting the poststrata by $d=1,\dots,D$, the poststratification estimator for a total involves adjusting the HT-weighted domain totals ($\hat{T}_d$) by the ratio of known (N_d) to estimated ($\hat{N}_d = \sum_{i\in s}\pi_i^{-1}$) domain sizes. Here the case weights are $N_d / \hat{N}_d \pi_i$ and the estimator is $\hat{T}_{PS} = \sum_{d=1}^{D} N_d \hat{T}_d / \hat{N}_d$.

Ad hoc Trimming/Weight Redistribution. There is limited literature and theory on design-based weight trimming methods, most of which are not peer-reviewed publications and focus on issues specific to a single survey or estimator. Potter (1988; 1990) presents an overview of alternative procedures and applies them in simulations. All design-based methods involve establishing an upper cutoff point for large weights, reducing weights larger than the cutoff to its value, then "redistributing" the weight above the cutoff to the non-trimmed cases. This ensures that the weights before and after trimming sum to the same totals (Kalton and Flores-Cervantes 2003). The methods vary by how the cutoff is chosen.

Chowdhury *et al.* (2007) describe the weight trimming method used to estimate proportions in the U.S. National Immunization Survey (NIS). The "current" (at the time of the article) cutoff value was $\text{median}(w_i)+6IQR(w_i)$, where $IQR(w_i)$ denotes the inter-quartile range of the weights. Versions of this cutoff (e.g., a constant times the median weight or other percentiles of the weights) have been used by other survey organizations (Battaglia *et al.* 2004; Pedlow *et al.* 2003; Appendix A in Reynolds and Curtin 2009).

2.2. Model-based Methods

We apply two model-based approaches, the super population (e.g., Valliant *et al.* 2000) and the generalized design-based (e.g., Beaumont 2008). Both methods incorporate models into estimation in very different ways. Each has an associated set of case weights though often then are implicitly defined. For the superpopulation approach, we consider estimation using the Best Linear Unbiased Predictor (BLUP) and some robust-BLUP methods. The generalized design-based approach uses a model between the weights and survey values, then weights are replaced with their predictions from the model.

Best Linear Unbiased Prediction. This approach assumes that the population survey response variables $\mathbf{Y}$ are a random sample from a larger ("super") population and assigned a probability distribution $P(\mathbf{Y}|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$. Typically the BLUP (e.g., Royall 1976) method is used to estimate the model parameters. Here, for observation i , we assume that the population values of $\mathbf{Y}$ follow the model

$$E_M\left(y_i|\mathbf{x}_i\right)=\mathbf{x}_i^T\boldsymbol{\beta},\ Var_M\left(y_i|\mathbf{x}_i\right)=\sigma^2 D_i, Cov_M\left(y_i,y_j\right)=D_{ij}\sigma^2,\ i\neq j \tag{2}$$

where $\mathbf{x}_i$ denotes a p-vector of benchmark auxiliary variables for unit i , which is known for all population units, and $D_i>0$ is a constant associated with population unit i . We consider only the case where the *y*'s are uncorrelated. For $\mathbf{Y}=\left(y_1,\dots,y_N\right)^T$ denoting the vector of population y-values, the total is $T=\mathbf{1}^T\mathbf{Y}$, where**1** is a vector of *N* 1's.The population total can also be written as $T=\mathbf{1}_s^T\mathbf{y}_s+\mathbf{1}_r^T\mathbf{y}_r$ where $\mathbf{1}_s$ and $\mathbf{1}_r$ are vectors of *n* and *N-n* 1's. Denote a linear estimator of the total as $\hat{T}=\mathbf{w}_s^T\mathbf{y}_s$, where $\mathbf{w}_s=\left(w_1,\dots,w_n\right)^T$ is a n-vector of coefficients. The population matrix of covariates is $\mathbf{X}=\begin{bmatrix}\mathbf{X}_s\\ \mathbf{X}_r\end{bmatrix}$, where $\mathbf{X}_s$ is the $n\times p$ matrix for sample units and $\mathbf{X}_r$ is the $(N-n)\times p$ matrix for nonsample units. Under the general prediction theorem (Thm. 2.2.1 in Valliant *et al.* 2000), the optimal estimator of a total is then

$$\hat{T}_{opt}=\mathbf{1}_s^T\mathbf{y}_s+\mathbf{1}_r^T\mathbf{X}_r^T\hat{\boldsymbol{\beta}}\ , \tag{3}$$

where $\hat{\boldsymbol{\beta}} = \mathbf{A}_s^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{y}_s$, $\mathbf{A}_s = \mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s$, and $\mathbf{V}_{ss}$ is the part of **V** for the sample units. The optimal value of the weight vector is $\mathbf{w}_s = \mathbf{V}_{ss}^{-1}\mathbf{X}_s\mathbf{A}_s^{-1}\mathbf{X}_r^T\mathbf{1}_r + \mathbf{1}_s$.

Note that the BLUP is also variable-specific since a separate model may be formulated for each y-variable. The case weights depend on the covariate matrix **X**, the variance $Var_M\left(\mathbf{Y}\right)$ (and thus indirectly on the variable y), and how the sample and non-sample units are designated. They can be less than one, even negative. However, this expression is a standard form for writing the case weights, which is beneficial when comparing between the alternatives. The estimator for the total using these weights is model-unbiased under the BLUP model, and is only design-unbiased under special circumstances (e.g., Hansen et. al 1983).

Since the efficiencies of the BLUP method depend on how well the associated model holds, these methods can be susceptible to model misspecification. When comparing a set of candidate weights to a preferable set of weights, the difference in the estimated totals under the "preferable" model attributed to model misspecification is a measure of design-based inefficiency or model bias. To overcome the bias, the superpopulation literature has developed a few robust alternatives. Generally, each approach involves using a preferable alternative model to adjust the BLUP estimator for model misspecification and/or influential observations. We describe two examples.

Chambers et al.'s Robust BLUP. Chambers *et al.* (1993) proposed an alternative to the BLUP approach that applies a model-bias correction factor to linear regression case weights. This bias correction factor is produced using a nonparametric smoothing of the linear model residuals against frame variables known for all population units is applied to the BLUP estimator (3). Suppose that the true model is $y_i\left|\mathbf{x}_i = m\left(\mathbf{x}_i\right) + v_i e_i\right.$, with working model variance $Var\left(y_i\left|\mathbf{x}_i\right.\right) = \sigma^2 D_i$. The model bias in the BLUP total under this model is $E_M\left(\hat{T}_{BLUP} - T\right) = \sum_{i\in r}\delta\left(\mathbf{x}_i\right)$, where $\delta\left(\mathbf{x}_i\right) = \mathbf{x}_i^T E_M\left(\hat{\boldsymbol{\beta}}\right) - m\left(\mathbf{x}_i\right)$. Since the residual $\hat{e}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ (under the preferred model) is an unbiased estimator of $-\delta\left(\mathbf{x}_i\right)$, the sample-based residuals can estimate the nonsample $\delta\left(\mathbf{x}_i\right)$ values. This leads to the *nonparametric calibration estimator* for the finite population, given by

$$\hat{T}_C = \sum_{i\in s} y_i + \sum_{i\in r}\mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \sum_{i\in s}\hat{e}_i = \hat{T}_{BLUP} + \sum_{i\in s}\hat{\delta}\left(\mathbf{x}_i\right), \tag{4}$$

In general, nonparametric case weights are $\mathbf{w}_s = \mathbf{V}_{ss}^{-1}\left[\mathbf{V}_{sr} + \mathbf{X}_s\mathbf{A}_s^{-1}\left(\mathbf{X}_r^T - \mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\right)\right]\mathbf{1}_r + \mathbf{1}_s + \mathbf{m}_s$, where $\mathbf{m}_s$ contains the sample residual-based estimates of $-\delta\left(\mathbf{x}_i\right)$. Here, the case weights are $w_i = w_{i(BLUP)} - \hat{e}_i / y_i$ if $y_i \neq 0$, where $w_{i(BLUP)}$ are the original BLUP weights. The estimator for the total using these weights is model-unbiased under the preferred model, when the BLUP is not, and is also approximately design-unbiased. The weights depend directly on y and, like the BLUP weights, they have no size restriction. Chambers *et al.* (1993) used ridge regression to produce $\hat{e}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{ridge}$. Chambers (1996) proposed further extensions. The robust BLUP is model-unbiased under the preferred model when the BLUP is not and the $\hat{\boldsymbol{\beta}}$ parameter estimates are less influenced by extreme observations.

Difference Estimator. Firth and Bennett (F&B 1998) produce a similar bias-correction factor to Chambers *et al.* (1993,) for a difference estimator (Särndal *et al.* 1992) as follows:

$$\hat{T}_D = \hat{T}_{BLUP} + \sum_{i\in s}\left(\pi_i^{-1} - 1\right)\left(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right). \tag{5}$$

The case weights are $w_i = w_{i(BLUP)} + \left(\pi_i^{-1} - 1\right)\hat{e}_i / y_i$ if $y_i \neq 0$, where $\hat{e}_i$ comes from the BLUP model. Again, these case weights depend directly on y and have no size restrictions. Estimator (5) is model-unbiased under the BLUP model, but it smoothes the effects of influential observations, and is also approximately design-unbiased.

Generalized Design-Based Method. A recently developed weight smoothing approach uses a model to trim large weights on highly influential or outlier observations. The general framework and theory for estimating finite population totals was developed later by Beaumont (2008). Generally, within a given observed sample, we fit a model between the weights and the survey response variables. Denote M as the model proposed for the weights, conditional on the sample y-values $\mathbf{Y}$, sample inclusion indicators $\mathbf{I}$, and the design used to select the sample π. The model M trims weights by removing variability in them. This is different from the superpopulation model-based approach, where the model describes the relationships between a survey response variable and a set of auxiliary variables. Here only one model is fit and one set of smoothed weights is produced for all y-variables. The weights predicted from the model then replace the weights and are used to estimate the total. The hope is that using regression predictions of the weights will eliminate extreme weights.

One example of a weights model that is appropriate in *pps* samples (used in the empirical example) is the inverse model: $E_M\left(w_i^{-1}\middle|\mathbf{I},\mathbf{Y}\right)=\mathbf{H}_i^T\boldsymbol{\beta}+v_i^{1/2}\varepsilon_i$, where $\mathbf{H}_i$ and $v_i>0$ are known functions of the y-variables, the errors are $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \left(0,\sigma^2\right)$, and $\boldsymbol{\beta},\sigma^2$ are unknown model parameters. This model produces the smoothed weight $\hat{w}_i=\left[\mathbf{H}_i^T\hat{\boldsymbol{\beta}}\right]^{-1}$, where $\hat{\boldsymbol{\beta}}$ is the generalized Least Squares estimate of $\boldsymbol{\beta}$. Since $\tilde{w}_i=E_M\left(w_i\middle|\mathbf{I},\mathbf{Y}\right)$ is unknown, it is estimated with $\hat{w}_i$, found by fitting a model to the sample data. The estimator for the finite population total is then $\hat{T}_B=\sum_{i\in s}\hat{w}_i y_i$, with case weights $\hat{w}_i$.

2.3. Model-Assisted Weighting Methods

Here two model-assisted approaches are discussed: the generalized regression (GREG) estimator and a robust estimator produced using penalized spline models. Both are special forms of calibration estimators, where we incorporate an underlying model for the survey and auxiliary variables, but evaluate estimators with respect to their design-based properties.

Generalized Regression (GREG). Case weights resulting from calibration on benchmark auxiliary variables can be defined with a global regression model for the survey variables (Kott 2009). Deville and Särndal (1992) proposed the calibration approach that involves minimizing a distance function between the base weights and final weights to obtain an optimal set of survey weights. Here "optimal" means that the final weights produce totals that match external population totals for the auxiliary variables $\mathbf{X}$ within a margin of error.

Specifying alternative calibration distance functions produces alternative estimators. A least squares distance function produces the *general regression estimator (GREG)* $\hat{T}_{GREG}=\hat{T}_{HT}+\hat{\mathbf{B}}^T\left(\mathbf{T}_X-\hat{\mathbf{T}}_{XHT}\right)=\sum_{i\in s}g_i y_i/\pi_i$, where $\hat{\mathbf{T}}_{XHT}=\sum_{i\in s}w_i\mathbf{x}_i=\sum_{i\in s}\mathbf{x}_i/\pi_i$ is the vector of Horvitz-Thompson totals for the auxiliary variables, $\mathbf{T}_X=\sum_{i=1}^N\mathbf{x}_i$ is the corresponding vector of known totals, $\hat{\mathbf{B}}^T=\mathbf{A}_s^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{\Pi}_s^{-1}\mathbf{y}_s$ is the regression coefficient, with $\mathbf{A}_s=\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{\Pi}_s^{-1}\mathbf{X}_s$, $\mathbf{X}_s^T$ is the matrix of $\mathbf{x}_i$ values in the sample, $\mathbf{V}_{ss}=diag\left(v_i\right)$ is the diagonal of the variance matrix specified under the model, and $\mathbf{\Pi}_s=diag\left(\pi_i\right)$ is the diagonal matrix of the probabilities of selection for the sample units. In the second expression for the GREG estimator, $g_i=1+\left(\mathbf{T}_X-\hat{\mathbf{T}}_{XHT}\right)^T\mathbf{A}_s^{-1}\mathbf{x}_i v_i^{-1}$ is called the "g-weight." Thus, the case weights here are $w_i=g_i/\pi_i$.

The GREG estimator for a total is model-unbiased under the associated working model and is approximately design-unbiased when the sample size is large (Deville and Särndal 1992). When the model is correct, the GREG estimator achieves efficiency gains. If the model is incorrect, then the efficiency gains will be dampened (or nonexistent) but the GREG estimator is still approximately design-unbiased. However, the case weights can be negative or less than one. Calibration can also introduce considerable variation in the survey weights. To overcome the first problem, extensions to limit the range of calibration weights have been developed that involve either using a bounded distance function (Rao and Singh 1997; Singh and Mohl 1996) or bounding the range of the weights using an optimization method (such as quadratic programming, Isaki *et al.* 1992). Chambers (1996) proposed penalized calibration optimization function to produce non-negative weights and methods that impose additional constraints on the calibration equations.

p-spline (Robust GREG). Recent survey methodology research has focused on a class of estimators based on penalized (*p*-) spline regression to estimate finite population parameters (Zheng and Little 2003, 2005; Chen *et. al* 2010; Breidt and Opsomer 2000; Breidt *et al.* 2005). Breidt *et al.* (2005) develop a model-assisted *p*-spline estimator similar to the GREG estimator. In application, they showed their *p*-spline estimator is more efficient than parametric GREG estimators when the parametric model is misspecified, but the *p*-spline estimator is approximately as efficient when the parametric specification is correct.

Assuming that quantitative auxiliary variables x_i are available and known for all population units, Breidt *et al.* (2005) propose the following superpopulation regression model: $y_i = m(x_i) + \varepsilon_i,\ \varepsilon_i \overset{\text{ind}}{\sim} N(0, D_i)$. Treating $\{(x_i, y_i) : i \in U\}$ as one realization from this model, the *spline function* using a linear combination of truncated polynomials is $m(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \cdots \beta_p x^p + \sum_{q=1}^{Q} \beta_{q+p} \left(x - \kappa_q\right)_+^p, i = 1, \ldots, N$, where the constants $\kappa_1 < \ldots < \kappa_L$ are fixed "knots," and the term $(u)_+^p = u^p$ if $u > 0$ and zero, otherwise, p is the degree of the spline, and $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_{p+Q}\right)^T$ is the coefficient vector. For $m_i = m(\mathbf{x}_i, \boldsymbol{\beta}_U), i \in U$ denoting the *p*-spline fit obtained from the hypothetical population fit at $\mathbf{x}_i$, Breidt *et al*. (2005) incorporate m_i into survey estimation by using a difference estimator $\sum_{i \in U} m_i + \sum_{i \in s} (y_i - m_i)/\pi_i$. Given a sample, m_i here can be estimated using a sample-based estimator $\hat{m}_i$. For $\mathbf{W}_s = diag(1/\pi_i), i \in s$ and for fixed α, the π-weighted estimator for the *p*-spline model coefficients is $\hat{\boldsymbol{\beta}}_\pi = \left(\mathbf{X}_s^T \mathbf{W}_s \mathbf{X}_s + \mathbf{D}_\alpha\right)^{-1} \mathbf{X}_s^T \mathbf{W}_s \mathbf{y}_s = \mathbf{G}_\alpha \mathbf{y}_s$, such that $\hat{m}_i = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_\pi\right)$. Their estimator is

$$\hat{T}_{mpsp} = \sum_{i \in U} \hat{m}_i + \sum_{i \in s} \frac{y_i - \hat{m}_i}{\pi_i} \square \sum_{i \in s} \left[\frac{1}{\pi_i} + \sum_{j \in U} \left(1 - \frac{I_j}{\pi_j} \right) \mathbf{x}_j^T \mathbf{G}_\alpha e_i \right] y_i = \sum_{i \in s} w_i^* y_i, \quad (6)$$

where $I_j = 1$ if $j \in s$ and zero otherwise and $e_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\pi$ and $w_i^* = \pi_i^{-1} - \left(\frac{\hat{m}_i}{\pi_i} - \frac{N\hat{\bar{m}}_U}{n} \right) \Big/ y_i$, with $N\hat{\bar{m}}_U = \sum_{i \in U} \hat{m}_i$. In *penalized* (*p*)*-spline regression*, the influence of the knots is bounded using a constraint on the Q-spline coefficients. One constraint with the truncated polynomial model is to bound $\sum_{q=1}^{Q} \beta_{q+p}^2$ by some constant, while leaving the polynomial coefficients $\beta_0, \ldots, \beta_p$ unconstrained. This smoothes the $\beta_{p+1}, \ldots, \beta_{p+Q}$ estimates toward zero, reducing the possibility of over-fitting the model. Adding the constraint as a Lagrange multiplier, for a fixed constant $\alpha \geq 0$ in we have $\hat{\boldsymbol{\beta}} = \arg_{\boldsymbol{\beta}} \min \sum_{i \in U} \left(y_i - m(\mathbf{x}_i, \boldsymbol{\beta})\right)^2 + \alpha \sum_{q=1}^{Q} \beta_{q+p}^2$. The smoothing of the resulting fit depends on α; larger values produce smoother fits; $\alpha = 0$ corresponds to the Chambers' ridge regression model.

Breidt and Opsomer (2000) and Breidt *et al.* (2005) proposed and developed a model-assisted *p*-spline estimator that was more robust to misspecification of the linear model, resulting in minimum loss in efficiency compared to other calibration estimators. The *p*-spline estimators are a specific case of a robust model-prediction estimator. It is approximately model-unbiased under the *p*-spline model, when the linear GREG model does not hold and it can also smooth the effects of influential observations. However, this method assumes that the covariates are known for all population units and applies when these are quantitative (vs. categorical) variables.

2.4. Comparing Theoretical Properties

Before comparing the alternatives in an empirical example, we consider the theoretical bias properties of the alternative estimators, which correspond to comparing the assumed underlying structural models. The design- and model-bias properties are summarized in Table 1. It is apparent that no one particular estimator will theoretically outperform the other alternatives in every scenario. Theoretically, the model-assisted approaches aim for a compromise between the design- and model-based approaches. However, the performance of both the GREG and robust-GREG estimators depends on the fit of associated underlying model to the sample data. We include an empirical example to gauge how the alternatives perform when applied to data collected in single-stage samples.

Table 1: Theoretical Properties of Alternative Estimators

Method	Design-Unbiased?	Model-Unbiased?
Design-based		
HT	Yes	Under the model $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$, and $\pi_i = nx_i / N\bar{x}_U$
PS	Approximately	Under the model $y_i = \mu_g + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$ where g is a poststratum
Model-based		
BLUP	Not here	Under the BLUP model $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2 x_i^\gamma)$.
Chamber's Robust	Not here	Under the BLUP model; model-unbiased if BLUP model is wrong and model $y_i = m(\mathbf{x}_i) + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2 x_i^\gamma)$ fits better.
F&B Difference	Yes	Under BLUP model; effect of influential observations is reduced.
Beaumont	Not here	Design- and model-unbiased under the model $w_i^{-1} = \beta_1 y_{1i} + \beta_2 y_{2i} + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2 y_{1i})$
Model-assisted		
GREG	Approximately	Under the model $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$.
Robust GREG	Approximately	Under the *p*-spline model $y_i = m(\mathbf{x}_i) + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2 x_i^\gamma)$; effect of influential observations is reduced.

3. Empirical Comparisons

The data used in this illustration come from the 1998 Survey of Mental Health Organizations (SMHO; Manderscheid and Henderson 2002). The survey dataset with non-zero hospital beds was randomly replicated up to a pseudopopulation of 10,000 units. The original SMHO98 sample is stratified by the type of the organization, with sample sizes in collapsed strata given with the pseudopopulation counts in Table 2.

Table 2: Number of SMHO98 Sample Units, by Organization Type

Organization Type (stratum)	SMHO Subsample	Pseudopopulation
Psychiatric Hospital	215	3,242
Residential	64	959
General Hospital	216	3,329
Military Veterans	38	522
Multi-service or Substance Abuse	131	1,948
Total	664	10,000

The variables of interest are the total (y_1) and a count (y_2) of medical expenditures an individual organization incurred during a calendar year, and an artificial dependent variable (y_3), where $y_3 = 10 + 1.5 \ln(x) + \varepsilon, \varepsilon \sim N(0,1)$. The auxiliary variable (x) is the number of beds in a given hospital. The SMHO98 file was modified by removing hospitals with zero beds, five Partial Care of Outpatient hospitals, and one extremely large hospital, leaving 664 hospitals. We expanded this dataset to 10,000 units by selecting additional units with replacement from the 664. The values of x, y_1, y_2, and y_3 were slightly randomly perturbed to eliminate duplicate values. Figure 1 shows the plots of these variables in the pseudopopulation.

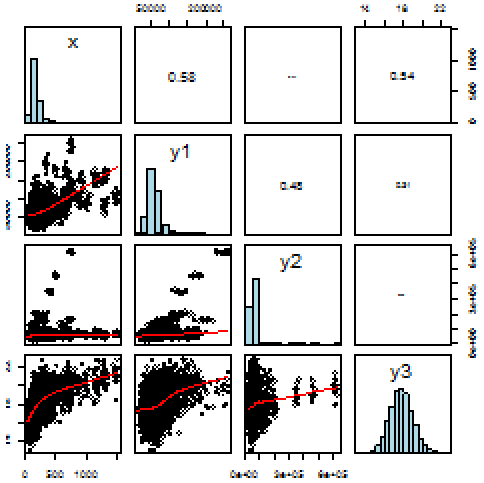

Figure 1: Plots of y-Variables vs. Number of Beds, SMHO98 Pseudopopulation

We see the relationship between the number of beds and the expenditures from Figure 1. The number of beds is more correlated with the total expenditures ("y1", 0.58) and y3 (0.54). Figure 2 shows the *x*-*y* plots in more detail, with points in different colors by hospital type. In each plot, linear (in blue) and loess smoother (in green) prediction lines are also shown. The difference in the two lines indicates that there are some influential points and/or curvature in the data. The relationships between number of beds and the *y*-variables vary by hospital type, so we apply weighting adjustments within the type of hospital.

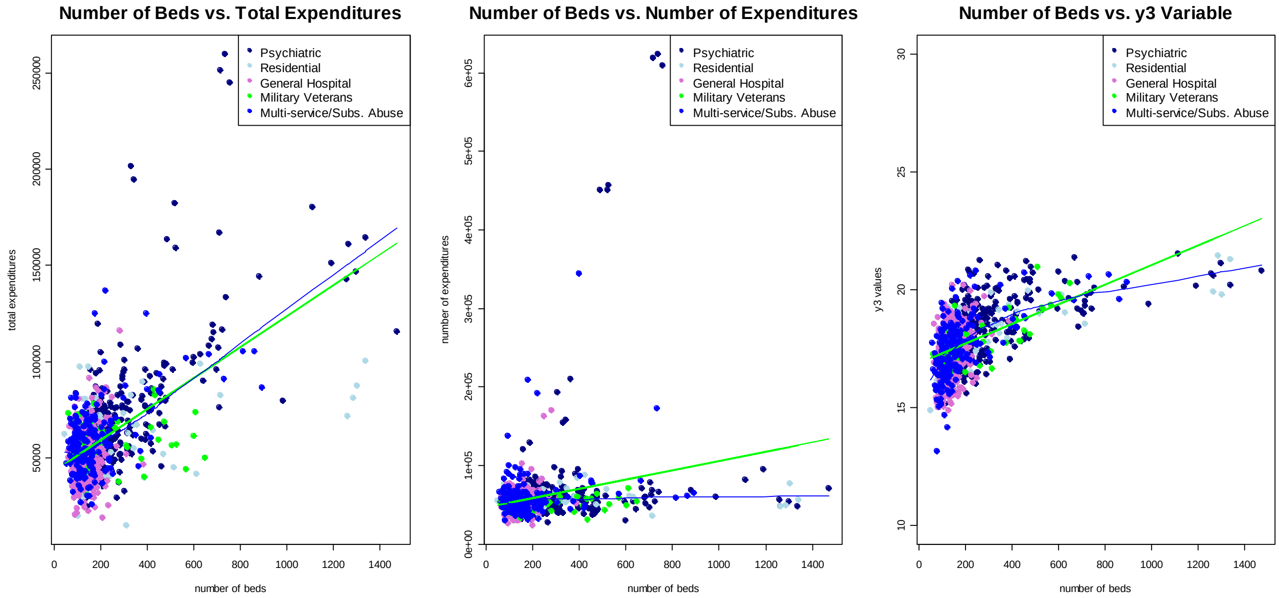

Figure 2: Plots of Total Expenditures and Number of Expenditures vs. Number of Beds for SMHO Populations, With Linear Regression and Nonparametric Smoother Lines

3.1. One-Sample Results

Using the number of beds as the auxiliary variable, a sample of size 500 was drawn from the non-zero bed SMHO psuedopopulation with probability proportional to the number of beds (the measure of size, or MOS). Two design-based methods were used:

(i) the HT estimator, with base weights varying under the pps sampling, and

(ii) post-stratification adjustments to the population size by the hospital type in Table 2.

Four model-based methods were used:

(iii) the BLUP using the working model described below within each of the 5 hospital types,

(iv) Chambers' robust estimator, using ridge regression within each of the 5 hospital types,

(v) the F&B difference estimator within each of the 5 hospital types, and

(vi) Beaumont's estimator assuming an inverse model within each hospital type, with the weight as the dependent variable and y_1, y_2, and y_3 as independent variables in the model.

Last, two model assisted methods are also included:

(vii) the GREG estimator using the working model within the 5 hospital types, and

(viii) the robust *p*-spline GREG within each of the 5 hospital types.

An ad hoc design-based weight trimming method is also used to trim the design-based PS and GREG model-assisted weights, using the empirical 95th percentile as the cutoff (choices (ii) and (viii)). A single cutoff was used for the entire sample, without regard to hospital type. The excess weight exceeding the cutoff for a particular unit was then equally distributed to all units within the same hospital type (vs. redistributing the total excess weight to all non-trimmed cases in the sample, disregarding the hospital type).

The working model for most strata is: $E_M\left(y_i|\mathbf{x}_i\right)=\mathbf{x}_i^T\boldsymbol{\beta},\ Var_M\left(y_i|\mathbf{x}_i\right)=\sigma^2 x_i^{\gamma_h}$, where $\mathbf{x}=\left[\sqrt{x_i} \quad x_i\right]$ for x denoting the number of hospital beds, y the total expenditure, and the variance measure of size γ_h is estimated iteratively from the population and rounded (Henry and Valliant 2006), giving $\gamma_h \approx \left(1.75, 0.50, 0.75, 0.25, 0.50\right)$ for y_1, $\gamma_h \approx \left(2.00, 0.75, 1.00, 0.75, 1.00\right)$ for y_2, and $\gamma_h \approx \left(0.00, 2.00, 0.00, 0.25, 0.00\right)$ for y_3. Upon further inspection of the data, the same model was used, but without the $\sqrt{x_i}$ -component in $\mathbf{x}$ for stratum 2 with y_1 and strata 2 and 4 for y_2 and y_3. The calibration and robust calibration methods use the same models, but with a constant variance. Since Figure 2 shows different relationships by hospital type, the model was fit within each stratum. For consistency, the Beaumont smoothing model was fit within each stratum. In all estimators, including Beaumont's, the working models did not include an intercept.

Figure 3 on the following page shows plots of the case weights produced using each method and plots of the alternative weights vs. the original HT base weights, for each variable. For this sample, the number of sample units by hospital type were (213, 63, 111, 26, 87). The plots of the PS and GREG weights before and after the trimming are also shown. Each plot also contains a 45 degree reference line shown in red. We see that the different methods can produce very different weights.

The PS and GREG weight adjustments do not drastically change the HT weights. We see that trimming and redistribution adjustment the PS and GREG weights results in slight changes since we only adjust 5 percent of the sample weights. The model-based methods produce very different weights; this is not necessarily detrimental since they use a different estimation strategy. However, in terms of case weights, some of the BLUP, robust BLUP, F&B difference, and *p*-spline weights are more varied, some even negative. The Beaumont method here applies severe trimming to the HT weights. And we see how the ad hoc trimming and redistributing modifies the PS and GREG weights in the last two plots. The impact of all these alternative weights on estimation of totals is examined next.

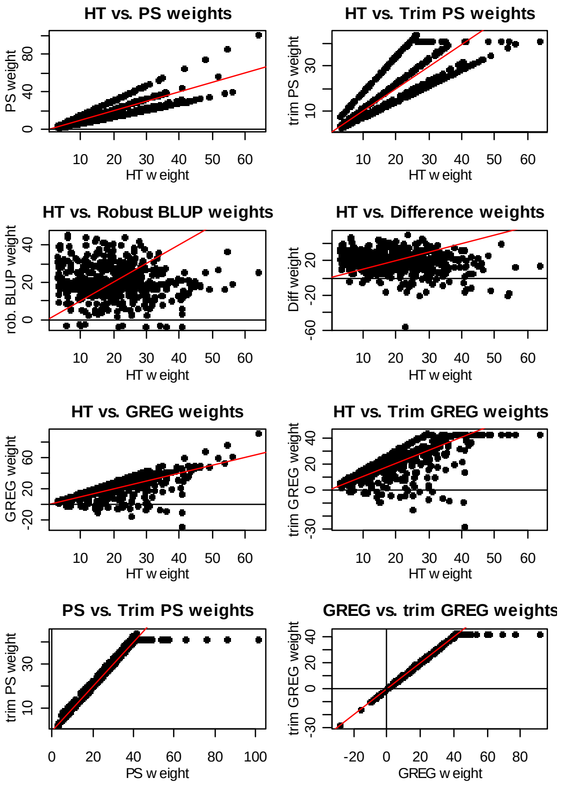

Figure 3: Plots of Alternative Weights vs. HT Weights, Example Sample of n=500 Drawn from SMHO Pseudopulation y_1-Variable

3.2. Simulation Results

We replicated the one-sample results, selecting 10,000 samples of size 500 using the same design as described above from the SMHO pseudopopulation. Table 3 on the following page shows the bias of each alternative estimator relative to the true population totals (Relbias) the estimator's root mean square error (RMSE), relative to the RMSE of the HT estimator (RMSE Ratio).

For the totals estimated from the *pps* 10,000 samples, we see that the HT and PS estimators have very low bias. The PS estimator also has lower RMSE than the HT estimator for all three variables, lowest for the variable y_3. However, the trimming and redistribution of the largest 5 percent PS weights introduces a positive bias in the totals for all three variables. As larger weights are associated with larger *x*-values, and thus generally associated with smaller *y*-values, redistributing the weight equally to the non-trimmed cases (which have higher *y*-values than the trimmed cases) causes this bias. The same occurred for the calibration weights, with the exception for y_3. Thus, in these *pps* samples, weight trimming is generally not appropriate.

The model-based estimators have a slight bias, but interestingly the bias for the robust and difference estimators, which include so-called "bias correction factors," do not outperform the BLUP. The BLUP and robust BLUP do outperform the HT estimator for the variable y_3 in terms of the RMSE. This is expected since the BLUP model incorporates terms that account for the curvature we introduced in this variable. The Beaumont estimator is the poorest performer, in terms of both the bias and RMSE of estimated totals. In modeling the inverse weights, the estimated totals appear to be very sensitive to changes in the small π_i 's, which in turn create large differences between the HT and predicted smooth weights. The model-assisted estimators have both low bias and RMSE's. Both the calibration and p-spline calibration estimators are among the best performers for all three variables; the *p*-spline being relatively the "best' in terms of the relative bias and RMSE relative to the HT estimator's RMSE.

Table 3: Relative Bias and RMSE Ratios of Estimated Totals, from SMHO98 Simulation

	Total Expenditures		*Number of Expenditures*		y_3 *-Variable*	
Method	**RelBias (%)**	**RMSE Ratio**	**RelBias (%)**	**RMSE Ratio**	**RelBias (%)**	**RMSE Ratio**
Design-based						
HT	0.05	1.00	0.03	1.00	0.03	1.00
PS	0.08	0.41	0.04	0.46	0.02	0.02
Trimmed PS	10.85	25.95	6.29	7.52	2.10	0.93
Model-based						
BLUP	4.12	3.99	2.55	1.54	1.48	0.50
Chamber's Robust	4.38	4.47	2.52	1.52	1.33	0.42
F&B Difference	4.62	5.11	2.90	2.47	0.93	0.29
Beaumont (Inverse)	-25.03	134.57	-28.91	124.43	-33.35	228.78
Model-assisted						
GREG	0.93	0.71	2.75	2.06	0.67	0.11
Trimmed GREG	1.23	0.84	2.93	2.22	0.73	0.13
Robust GREG	-0.01	0.53	-0.17	0.71	0.01	0.02

Figure 4 at the end of this paper shows the empirical box plots of the totals, for each variable. These boxplots in show the Table 3 results visually. Generally, we see how the Beaumont estimator has the largest bias and variance, the trimmed estimators and model-based estimators have a positive bias, and the model-assisted estimators have relatively low bias and variance.

4. Discussion: Implications for Inference and Practice

4.1. Practical Considerations

In practice, there is usually only one realized sample. The simulations we conducted are generally not feasible. Even when they are, long-run simulation properties may not be a good reflection of the quality of an particular sample that is selected. However, it is possible, within a one-sample comparison, to express each of the alternative estimators in a form with a "base component" of some weight (either the HT or BLUP, depending on the estimator) multiplied by y_i and an "adjustment factor component" multiplied by y_i .

Table 4 shows the breakdown of these weighted components within each estimator. For example, the robust BLUP can be written as $\hat{T}_{BLUP} - \sum_{i \in s} \hat{e}_i = \sum_{i \in s} \left(w_{i(BLUP)} - \hat{e}_i / y_i \right) y_i$. This makes it clear that what the effect is on different weights. We then compare different components to each other to gauge their contribution to the weighted total.

As an example, Figure 4 shows plots of the different components for the y_1 –variable in Table 2. The x-axis of each plot is the base component multiplied by y_1 and the y-axis is the adjustment factor component multiplied by y_1 .

Table 4: Base and Adjustment Factor Contributions, by Estimator

Estimator Name	Estimator Form	Base Component	Adjustment Factor Component
Poststratification	$\sum_{d=1}^{D} N_d \hat{T}_d \big/ \hat{N}_d$	π_i^{-1}	$\pi_i^{-1}\left(N_d \big/ \hat{N}_d - 1\right)$
Robust BLUP	$\hat{T}_{BLUP} - \sum_{i\in s}\hat{e}_i$	$w_{i(BLUP)}$	$-\hat{e}_i \big/ y_i$
Difference	$\hat{T}_{BLUP} + \sum_{i\in s}\left(\pi_i^{-1}-1\right)\hat{e}_i$	$w_{i(BLUP)}$	$\left(\pi_i^{-1}-1\right)\hat{e}_i \big/ y_i$
Beaumont	$\sum_{i\in s}\hat{w}_i y_i$	π_i^{-1}	$\hat{w}_i - \pi_i^{-1}$
GREG	$\sum_{i\in s}\left[\pi_i^{-1} + \pi_i\left(w_{i(cal)}-1\right)\big/ y_i\right] y_i$	π_i^{-1}	$\pi_i\left(w_{i(cal)}-1\right)\big/ y_i$
Robust GREG (pspline)	$\sum_{i\in s}\left[\pi_i^{-1} - \left(\frac{\hat{m}_i}{\pi_i} - \frac{N\hat{\bar{m}}_U}{n}\right)\Big/ y_i\right] y_i$	π_i^{-1}	$-\left(\frac{\hat{m}_i}{\pi_i} - \frac{N\hat{\bar{m}}_U}{n}\right)\Big/ y_i$
Adhoc Trimming	$\sum_{i\in s} w_{i(trim)} y_i$	π_i^{-1}	$w_{i(trim)} - \pi_i^{-1}$

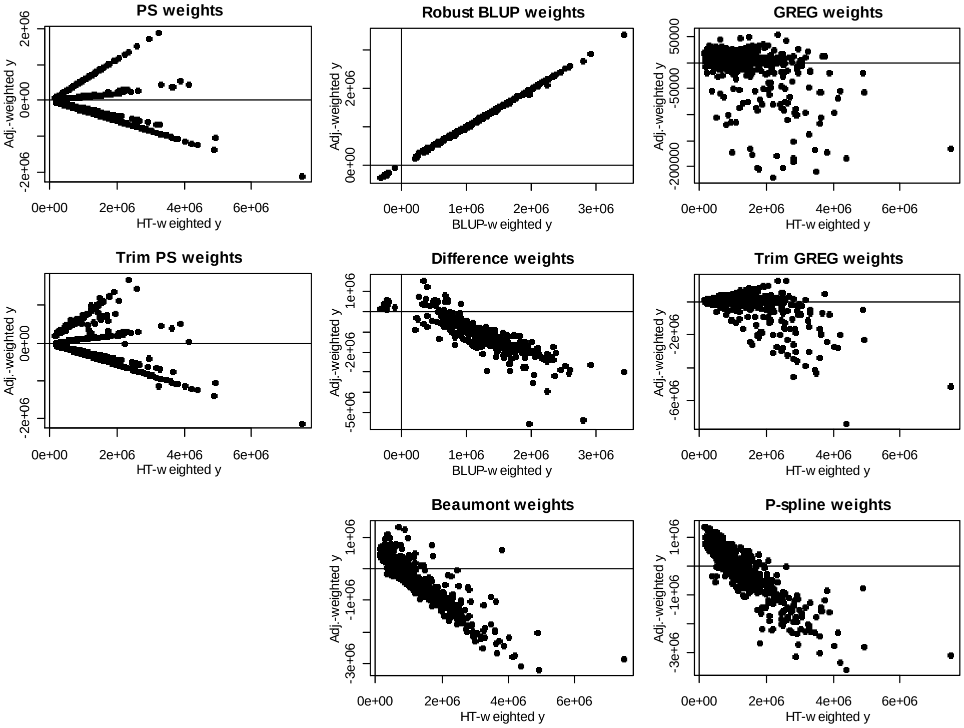

Figure 4: Plots of Alternative Weighted y- Components vs. HT/BLUP- Weighted y- Components, Example Sample of n=500 Drawn from SMHO Pseudopulation, y_1 -Variable

As is apparent from Figure 4, the alternative estimators perturb the base component weights substantially, both positively and negatively. When the base component is the π -weight, the HT-estimator is unbiased and the positive and negative perturbations need to balance out to maintain the unbiasedness in repeated sampling. The PS, GREG, and robust GREG do have this property and were unbiased in the simulation study. The Beaumont estimator is unstable for having many, large negative adjustments, leading to its large negative bias in the simulations.

4.2. Summary

The goal behind all of the methods that we have summarized is to somehow make inferences robust to anomalous values of weights or *y*'s or both. There are many variations on how this can be attempted. Among them are:

- Smooth or trim the weights;
- Smooth or trim the *y*'s;
- Use nonparametric estimators that are minimally affected by outlying weights, *y*'s, or combinations of the two.

In some cases, explicit formulas for weights are obtained; in others the smoothing must be done using iterative methods that give only implicit sets of weights. Practitioners fighting deadlines gravitate toward methods where weights are trimmed or smoothed without consideration of the analysis variables with which the weights will be used. This is pragmatic because the process of weight computation often proceeds on a parallel track from the editing of the analytic variables. However, the weight-trimming approach can be inefficient for some variables. If an outlying y_i or $w_i y_i$ product causes an estimator to have an unnecessarily large variance, weight trimming alone may not correct the problem and may, in fact, make it worse. Plus, values of weights or *y*'s that are innocuous for full population estimates may be quite influential for some domain estimates. The pros and cons of the different approaches are summarized below.

Methods that incorporate realistic models will improve the estimates of totals. By incorporating the relationship between the survey variable and some known auxiliary information, estimates of totals can have lower mean square errors. When the model is correctly specified, the associated estimators are optimal (e.g., the BLUP in Valliant *et. al* 2000). However, when the model does not hold or the sample contains outliers, several robust alternative estimators have been developed. While the superpopulation model-based and model-robust approaches introduce implicitly defined weights, their impact on estimation varies based on the method used. For example, the "robust" alternative methods incorporate a residual-based adjustment to improve estimates of the finite population total by reducing the bias. These methods can handle both categorical and quantitative auxiliaries. In our simulation study, the BLUP was somewhat biased but often had root mean square errors smaller than the HT-estimator. The robust alternatives did not improve substantially on the BLUP in our study.

The generalized design-based method (Beaumont 2008) smoothes weights by modeling them as functions of the *y*'s. The weight for each unit is then replaced by its regression prediction. Although the method may be an improved weight trimming method in some applications, much of the associated theory and its effectiveness in practice need to be further studied. While the variability in estimated totals may be reduced through a reduction of variance in the weights, this method seems easy to misapply. In addition, this method modifies all survey weights (perhaps substantially), while the typical design-based approaches aim to make sizeable changes to only a small number of cases. The wholesale changing of all weights by the generalized design-based approach may damage some estimates for domains even if overall population estimates are improved. In the simulations reported here, weight smoothing was extremely inefficient—introducing bias and dramatically inflating RMSEs.

In our empirical study, the GREG and robust GREG (*p*-spline) estimators were the most efficient choices, being nearly unbiased and having RMSEs substantially less than the basic HT-estimator. The poststratified estimator was also competitive even though it did not explicitly account for the relationship between the *y* and *x* variables we used. In contrast, trimming and redistributing the weights in the poststratified and GREG estimators was completely ineffective. Trimming added a slight to relatively large bias in the estimated totals, and thus increased RMSEs.

Generally, all weight trimming or modification methods have the potential to "undo" the effects of previous steps in weight calculation, like base weighting, nonresponse adjustment, and calibration to external controls. Nonresponse and calibration adjustments are designed to reduce biases and/or variances. In some cases, variable weights can be more efficient and their beneficial bias/variance reductions could be needlessly removed through arbitrary trimming of large weights. Thus, there is a need for diagnostic measures of the impact of weight trimming or modification on survey inference that extend past the existing "design effect" type of summary measures, most of which do not incorporate the survey variable of interest. The current methods do not quantify such "loss of information;" i.e., there is no indication of how various methods' distortion of the original weight distribution potentially impacts inference about full population or domain estimates.

References

Battaglia M.P., Izrael, D., Hoaglin D.C., and Frankel, M.R. (2004), "Tips and Tricks for Raking Survey Data (a.k.a. Sample Balancing)," *American Association for Public Opinion Research*. Available at: http://www.amstat.org/sections/srms/proceedings/y2004/files/Jsm2004-000074.pdf.

Beaumont, J.P. (2008), "A new approach to weighting and inference in sample surveys," *Biometrika*, **95** (3), 539-553.

Breidt, F.J., Claeskens, G., and Opsomer, J.D. (2005), "Model-assisted estimation for complex surveys using penalized splines," *Biometrika,* **92**, 831-846.

Breidt, F.J., and Opsomer, J.D. (2000), "Local polynomial regression estimators in survey sampling," *The Annals of Statistics*, **28**, 1026–1053.

Chambers, R. L., Dorfman, A.H., and Wehrly, T.E. (1993), "Bias Robust Estimation in Finite Populations Using Nonparametric Calibration," *Journal of the American Statistical Association*, **88**, 260-269.

Chambers, R. L. (1996), "Robust Case-Weighting for Multipurpose Establishment Surveys," *Journal of Official Statistics*, **12** (1), 3-32.

Chen, Q., Elliott, M. R., and Little, R. J. A., (2010), "Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling," *Survey Methodology*, **36**, 23-34.

Chowdhury, S., Khare, M., and Wolter, K. (2007), "Weight Trimming in the National Immunization Survey," *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association*.

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

Firth, D. and Bennett, K.E. (1998), "Robust models in probability sampling," *Journal of the Royal Statistical Society,* Series B, **60**, 3-21.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983), An evaluation of model-dependent and probability-sampling inferences in sample surveys, *Journal of the American Statistical Association*, **78**, 776-793.

Horvitz, D., and Thompson, D. (1952), **"**A Generalisation of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, **47**, 663-685.

Henry, K., and Valliant, R. V. (2012). "Methods for Adjusting Survey Weights when Estimating a Total," *Proceedings of the 2012 Federal Committee on Statistical Methodology's Research Conference*, http://www.fcsm.gov/12papers/Henry_2012FCSM_V-A.pdf.

Kalton, G., and Flores-Cervantes, A. (2003), "Weighting Methods," *Journal of Official Statistics*, **19** (2), 81-97.

Kott, P. (2009), "Calibration weighting: combining probability samples and linear prediction models," in D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics*, *Sample Surveys: Design, Methods and Application,* **29B**, Amsterdam: Elsevier BV.

Little, R.J.A. (2004), "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling," *Journal of the American Statistical Association*, **39**, 546-556.

Manderscheid, R.W. and Henderson, M.J. (2002). *Mental Health, United States, 2002*. DHHS Publication No. SMA04-3938. Rockville MD USA: Substance Abuse and Mental Health Services Administration. Available at http://mentalhealth.samhsa.gov/publications/allpubs/SMA04-3938/AppendixA.asp

Pedlow, S., Porras, J., O.Muircheartaigh, C., and Shin, H. (2003), "Outlier Weight Adjustment in Reach 2010," *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* American Statistical Association, 3228-3233.

Potter, F.A. (1988), "Survey of Procedures to Control Extreme Sampling Weights," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 453-458.

Potter, F. A. (1990), "Study of Procedures to Identify and Trim Extreme Sample Weights," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 225-230.

Rao, J.N.K. and Singh, A.C. (1997), "A ridge-shrinkage method for range-restricted weight calibration in survey sampling," *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Washington, D.C., 57-65.

Reynolds, P.D., and Curtin, R.T. (2009), *Business Creation in the United States: Initial Explorations with the PSED II Data Set.* New York: Springer.

Royall, R. M. (1976), "The Linear Least-squares Prediction Approach to Two-stage Sampling," *Journal of the American Statistical Association*, **71**, 657–664.

Särndal, C.E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer: Berlin, New York.

Singh, A.C., and Mohl, C.A. (1996), "Understanding Calibration Estimators in Survey Sampling." *Survey Methodology*, **22**, 107-115.

Valliant, R., Dorfman, A., and Royall, R. M. (2000), *Finite Population Sampling and Inference*, New York: Wiley & Sons.

Zheng, H. and Little, R.J.A. (2003), "Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples," *Journal of Official Statistics*, **19**, 99-117.

Zheng, H., and Little, R.J.A. (2005), "Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model," *Journal of Official Statistics*, **21**, 1-20.

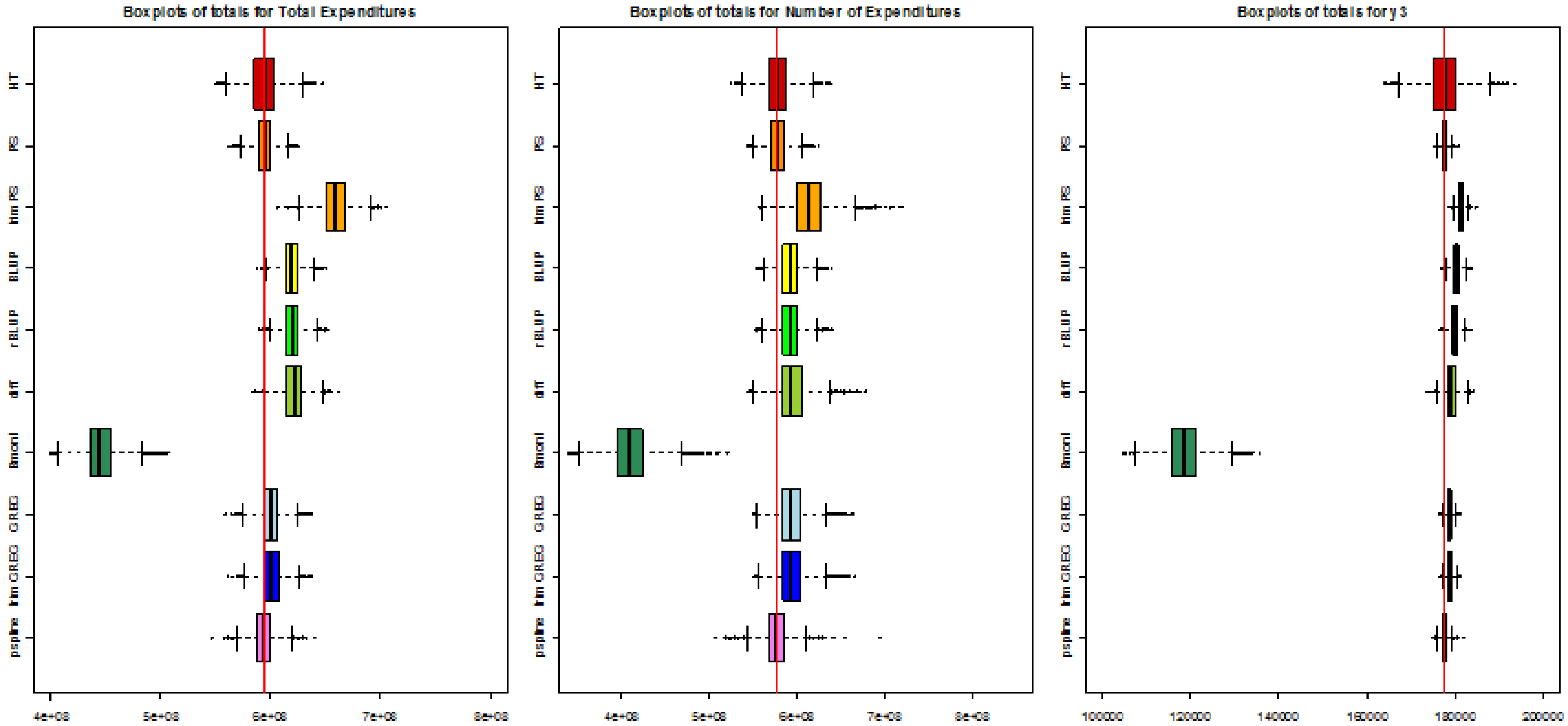

Figure 4: Boxplots of Alternative Totals, 10,000 Simulated Samples of n=500