*Statistics and Applications* {**ISSN 2452-7395 (online)**}
Volume 16 No. 1, 2018 (New Series), pp 289-303

# On Modifications to Linking Variance Estimators in the Fay-Herriot Model that Induce Robustness

**Snigdhansu Chatterjee**
*School of Statistics, University of Minnesota*

Received: Feb 19, 2018; Revised: March 01, 2018; Accepted: March 12, 2018

**Abstract**

We study the finite-sample properties of the traditional Prasad-Rao and the Fay-Herriot method for estimating the unknown linking variance component estimator in the Fay-Herriot small area model. Our study suggests that the traditional estimators are very sensitive to observations where the residuals after fitting the mean using a linear regression are too low, or too high. Accordingly, we propose some modified estimators. These estimators are shown to have much better finite sample properties. We study cases with two different patterns of the sampling variances in the Fay-Herriot model, and use three different sample sizes, exhibiting cases of high, moderate and low sample size situations, in this paper.

*Key words:* Fay-Herriot model, Prasad-Rao estimation, Fay-Herriot estimation, Robustness, Finite Sample Properties, Simulations.

## 1 Introduction

The Fay-Herriot (Fay III and Herriot, 1979) model is perhaps the most popular and widely-used model for small area statistics. In its current and popularly studied form, this is the following coupled framework:

1. **Sampling model:** Conditional on the unknown and unobserved *area-level effects* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$, the sampled and observed data $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^T$ follows a $n$-variate Normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\mathbf{D}$ with *known* diagonal entries $D_i > 0$ and off-diagonal entries 0. This component of the Fay-Herriot model captures the *sampling level* variability and distribution in the observations, conditioned on the inherent characteristics $\boldsymbol{\theta}$ of the various small areas under consideration.

Corresponding Author: Snigdhansu Chatterjee
Email id: chatt019 @ umn.edu

2. **Linking model:** The unobserved area-level effects $\boldsymbol{\theta}$ follows a $n$-variate Normal distribution with mean $X\beta$ for a known and non-random $n \times p$ matrix $X$ and unknown but fixed vector $\beta \in \mathbb{R}^p$. The covariance matrix is $\psi \mathbb{I}_n$, where the matrix $\mathbb{I}_n$ is the $n$ dimensional identity matrix and $\psi > 0$ is an unknown positive-valued constant. This component of the Fay-Herriot model links the various small areas together and provides effective shrinkage, by requiring that all the small areas share a common set of regression parameters $\beta \in \mathbb{R}^p$ and a common variance component $\psi > 0$.

In the above description of the Fay-Herriot model and in the rest of this paper, we consider all vectors to be column vectors, and for any vector or matrix $A$, the notation $A^T$ denotes its transpose. For any finite-length Euclidean vector $a$, the notation $|a|$ stands for its Euclidean norm, that is $|a| = (a^T a)^{1/2}$. In a slight abuse of notation, we will use $D$ to denote a vector in $\mathbb{R}^n$ whose entries are the diagonal elements of $\mathbf{D}$. The notation $a_i$ will denote the $i$-th element of vector $a$, similarly $A_{ij}$ will denote the $(i,j)$-th element of matrix $A$. The notation $\mathrm{tr}(A)$ denotes the trace of a matrix $A$, that is, $\mathrm{tr}(A) = \sum_i A_{ii}$. The notation $\mathcal{I}_A$ is the indicator function of measurable set $A$, that is, it takes the value one if $A$ holds and the value zero otherwise.

The notation $N_q(\mu, \Sigma)$ will be used to denote the $q$-dimensional Normal (Gaussian) distribution with mean $\mu \in \mathbb{R}^q$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^q \times \mathbb{R}^q$. When $q = 1$, it is dropped from the notation, thus $N(\mu, \sigma^2)$ corresponds to the one-dimensional Gaussian distribution. Other notations will be introduced and described as they arise.

Most surveys are constrained by feasibility, ethics and cost constraints and do not record adequate data on all variables of interest in as fine a scale as required by stakeholders. Models like the Fay-Herrot framework are consequently required to obtain accurate predictions at finer resolutions of spatial, demographic or other variables for planning, policy formulation and implementation purposes. A comprehensive recent monograph detailing small area methods, principles and procedures is Rao and Molina (2015), other resources on small area statistics include Jiang et al. (2002); Das et al. (2004); Pfeffermann and Glickman (2004); Datta et al. (2005); Rao (2005); Jiang and Lahiri (2006); Chatterjee et al. (2008); Li and Lahiri (2010); Salvati et al. (2012); Datta et al. (2011); Pfeffermann (2013); Yoshimori and Lahiri (2014a,b); Molina et al. (2015); Sugasawa and Kubokawa (2015); Rao (2015).

In the above Fay-Herriot model, the unknown parameters are the regression parameters $\beta \in \mathbb{R}^p$ and a common *linking variance component* $\psi > 0$. Define the diagonal matrix $\mathbf{B}$, whose $i$-th element is given by

$$B_i = D_i/(D_i + \psi),\ i = 1, \ldots, n. \tag{1.1}$$

The main quantity of interest in small area statistics centers around the properties of

$$[\boldsymbol{\theta}|\mathbf{Y}_n] = N_n\big(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{V}}\big),\ \text{where} \tag{1.2}$$

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= \Big[\mathbf{D}^{-1} + \psi^{-1}\mathbb{I}_n\Big]\big(\mathbf{D}^{-1}\mathbf{Y} + \psi^{-1}X\beta\big) \\ &= \big(\mathbb{I}_n - \mathbf{B}\big)\mathbf{Y} + \mathbf{B}X\beta \text{ and} \end{aligned} \tag{1.3}$$

$$\begin{aligned}\tilde{\mathbf{V}} &= \Big[\mathbf{D}^{-1} + \psi^{-1}\mathbb{I}_n\Big]^{-1} \\ &= \big(\mathbb{I}_n - \mathbf{B}\big)\mathbf{D}. \end{aligned} \tag{1.4}$$

For example, the conditional mean, $\mathbb{E}(\boldsymbol{\theta}|\mathbf{Y}_n) = \tilde{\boldsymbol{\theta}}$ in our notation above, often referred to the *BLUP* (Best Linear Unbiased Predictor), is a primary quantity of interest. Note that the BLUP depends on the unknown parameters $\beta$ and $\psi$, and as such cannot be predicted. Since $\tilde{\boldsymbol{\theta}}$ is a random variable, our preferred terminology to *predict* it, and not *estimate* it.

Estimators for the unknown parameters $\beta$ and $\psi$ may be obtained from the marginal distribution of $\mathbf{Y}$:

$$[\mathbf{Y}] = N_n\big(X\beta, \mathbf{D} + \psi\mathbb{I}_n\big). \tag{1.5}$$

If $\hat{\beta}$ and $\hat{\psi}$ are estimators of $\beta$ and $\psi$, we may consider using plugging-in these in place of the unknown parameters, and this yields the *EBLUP* (Empirical Best Linear Unbiased Predictor)

$$\hat{\boldsymbol{\theta}} = \big(\mathbb{I}_n - \hat{\mathbf{B}}\big)\mathbf{Y} + \hat{\mathbf{B}}X\hat{\beta}, \tag{1.6}$$

where $\hat{\mathbf{B}}$ is a diagonal matrix with the $i$-th diagonal element given by $\hat{B}_i = D_i/(D_i + \hat{\psi})$, $i = 1, \ldots, n$. A typical small area application revolves around using $\hat{\boldsymbol{\theta}}$ as a predictor, and the quality of prediction for the $i$-th small area is often evaluated using the *MSPE* (Mean Squared Prediction Error)

$$\Gamma_i(\beta, \psi) = \mathbb{E}\big(\hat{\boldsymbol{\theta}}_i - \theta_i\big)^2. \tag{1.7}$$

Alternatively, prediction intervals may also be used, as done in Chatterjee et al. (2008) and Yoshimori and Lahiri (2014b).

The quality of prediction, evaluated by MSPE $\Gamma_i$ or any other measure, naturally depends on the properties of the estimators $\hat{\beta}$ and $\hat{\psi}$. In this paper, we first present some evidence that the latter can be very biased, depending on the data at hand and the methodology used to estimate the linking variance $\psi$. We then present some strategies improve these estimators, so that they are robust against outlying observations and data quality issues. Apart from numeric studies, we also present some insights to understand the properties of the improved estimators. An improved estimator for $\psi$ typically naturally results in an improved estimator for $\beta$, and better quality EBLUP and other quantities of interest.

We consider two different classical estimation frameworks in this paper, both have been discussed and extensively studied earlier. First, we use the moment-based estimators that have been proposed in the seminal paper Prasad and Rao (1990), where much of the modern studies in small area statistics originated. Second, we consider the estimation methodology that is related to the original paper in this line of work, namely Fay III and Herriot (1979). A modern treatment and discussion on both these estimation methods may be found in Datta et al. (2005).

The Prasad-Rao (PR hereafter) estimation procedure begins by fitting an ordinary least squares regression on $\mathbf{Y}$ using the covariates $X$ in (1.5), thus obtaining

$$\hat{\beta}_{OLS} = \big(X^TX\big)^{-1}X^T\mathbf{Y}.$$

Let us use the notation $\mathbf{P_x}$ for the projection matrix on the column space of $X$, thus $\mathbf{P_x} = X\big(X^TX\big)^{-1}X^T$, this matrix is often referred to as the *hat matrix* in statistical literature. Based on the above estimator for $\beta$, we may define the vector of residuals

$$R = (\mathbb{I}_n - \mathbf{P_x})\mathbf{Y}.$$

We use the notation $R_i$ for the $i$-th element of $R$. The PR estimator for $\psi$is given by

$$\hat{\psi}_{PR} = (n-p)^{-1}\Big\{|R|^2 - \sum_{i=1}^{n} D_i + \sum_{i=1}^{n}(\mathbf{P_x D})_i\Big\}. \tag{1.8}$$

This estimator is positive almost surely when the intercept is the only covariate. Based on the discussion around equation (3.5) of Prasad and Rao (1990), we conclude the authors recommend using $\hat{\psi}_{PR}$ from (1.8) to obtain a new estimator of the regression coefficients:

$$\hat{\beta}_{PR} = \big(X^TWX\big)^{-1}X^TW\mathbf{Y}, \text{ where} \tag{1.9}$$

$$W = \big(\mathbf{D} + \psi\mathbb{I}_n\big)^{-1}. \tag{1.10}$$

We also would like to study an alternative estimator of $\psi$, that is a natural extension of equation (17) of Datta et al. (2005). This is given by

$$\hat{\psi}_{PR1} = (n-p)^{-1}|R|^2 - n^{-1}\sum_{i=1}^{n} D_i. \tag{1.11}$$

It can be easily seen that $\hat{\psi}_{PR1}$ has the same first order asymptotic properties as $\hat{\psi}_{PR}$, for any fixed $p$. One advantage of this estimator is its easier computation.

Based on the description of Datta et al. (2005), the Fay-Herriot (FH) method of parameter estimation involves simultaneously solving the following set of equations:

$$\beta = \big(X^T\mathbf{B}\mathbf{D}^{-1}X\big)^{-1}X^T\mathbf{B}\mathbf{D}^{-1}\mathbf{Y}, \tag{1.12}$$

$$1 = (n-p)^{-1}(\mathbf{Y} - X\beta)^T\big(\mathbf{D} + \psi\mathbb{I}_n\big)^{-1}(\mathbf{Y} - X\beta)^T. \tag{1.13}$$

The solution to the above equations are denoted by $\hat{\beta}_{FH}$ and $\hat{\psi}_{FH}$.

Note one issue of concern with $\hat{\psi}_{PR}$, $\hat{\psi}_{PR1}$ and $\hat{\psi}_{FH}$ as estimators of the linking variance $\psi$, which is a *positive number*. The issue is that none of these estimators are guaranteed to positive values only, and in fact in simulation experiments, these variance estimators are seen to take negative values with non-trivial frequencies. The reason for this concerning matter is easier to understand in $\hat{\psi}_{PR}$: this typically happens when a substantial number of $R_i^2$'s are below the corresponding $D_i$ values. In other words, the observations for which the regression line is a good fit and the corresponding residual is small, are poorer candidates for the problem of estimating the linking variance! In order to practically address the issue of potential negative numbers a estimates of variance, $\hat{\psi}_{PR}$, $\hat{\psi}_{PR1}$ and $\hat{\psi}_{FH}$ are typically truncated at a small positive number, which we set at $\epsilon = 10^{-4}$ throughout this paper.

## 2 A Large sample Numeric Study

To motivate the necessity for studying the properties of linking variance estimators, we present a large sample numeric study in this section. The simulation exercises reported here are based on the simulation framework presented in Datta et al. (2005). We consider $n = 5k$ small areas, where a set of known sampling variance components $(D_1, \ldots, D_5)$ are repeated $k$ times. We consider $p = 1$ and include the intercept as the only regression term, thus $X$ is a $n$-dimensional column of ones. We consider $\beta = 0$, thus the mean $\mathbb{E}\mathbf{Y} = \mathbb{E}\boldsymbol{\theta} = 0 \in \mathbb{R}^n$, although this is considered unknown during the estimation and prediction processes. We consider $\psi = 1$ in this study. All simulation experiments are replicated $K = 1000$ times, and the various summary measures and plots are based on these 1000 replications of each experiment.

In our first study, we use $k = 100$, that is, $n = 500$ small areas. We use $(D_1, \ldots, D_5) = (0.1, 0.4, 0.5, 0.6, 4.0)$, which is pattern (c) of Datta et al. (2005). The boxplot of $\hat{\psi}_{PR}$, $\hat{\psi}_{PR1}$ and $\hat{\psi}_{FH}$ are given in the left panel of Figure 1. In our second study, we used an identical framework except for setting all the $D_i$'s to be 1. The boxplot of $\hat{\psi}_{PR}$, $\hat{\psi}_{PR1}$ and $\hat{\psi}_{FH}$ are given in the right panel of Figure 1.

Note that the results of the classical linking variance estimators for the case of uneven $D$-values, as shown in the left panel, are disastrous. Essentially, all methods overestimate the linking variance, to the extent that the true value is not within the the whiskers of the boxplot. For the case of all $D_i = 1$, the results for the classical Prasad-Rao estimator (PR) are just as bad, however the Fay-Herrot (FH) and the first modified Prasad-Rao (PR.1) method as given in (1.11) perform well in this case.

Note that the above simulations do not show any kind of inconsistency of the well established PR, FH or PR.1 methods. The *asymptotic properties* of these methods are as discovered and reported in various publications earlier, some of which are cited above. The *finite sample* properties of these estimators may be affected by different conditions, as demonstrated in Figure 1, and samples of sizes even $500$ are not adequate to ensure that the asymptotic regime is in place for these linking variance estimators. A deeper study shows that the driving issue here is not the sample size, but the fact that residuals with very high absolue values are often found in data, and methods like PR are affected by outliers, leverage and influential points. This issue has been recognized in the classical literature on regression for decades, see for example Chatterjee and Hadi (1986, 2015). The linking variance estimators are particularly affected by residuals with high absolute values, and the use of the projection matrix (also called the "hat matrix") in PR may make the matter worse.

It is interesting to observe that for estimating the linking variance, two kinds of observations are problematic. First, those that fit the linear regression very well, and consequently have small residual values, which in turn leads to underestimation. Second, those that have very high residuals, which leads to overestimation. Since the distribution of $R_i^2$ in general is intractable, a universal prescription on how to handle very small and very high residuals is difficult. Also, the fact that the $D_i$ values may have a large spread contributes to the problem.

## 3 Modifications for Robustness

These paper is on proposals to address the lack of finite-sample robustness issues as illustrated by Figure 1, and thereby improve the performance of estimators of the linking variance.

First, note that while not explicitly recommended by Prasad and Rao (1990), perhaps owing to computational difficulties when this paper was published, the Prasad-Rao scheme lends itself to a system of equations, just like the Fay-Herrot method given in equations (1.12) and (1.13). Define

$$W = \big(\mathbf{D} + \psi \mathbb{I}_n\big)^{-1},$$
$$M_2 = \mathbb{I}_n - X\big(X^T W X\big)^{-1} X^T W.$$

This PR-system equations are as follows:

$$\hat{\beta}_{PR} = \big(X^T W X\big)^{-1} X^T W \mathbf{Y}, \text{ and} \tag{3.1}$$
$$\hat{\psi}_{PR2} = \mathrm{tr}(M_2^T M_2)^{-1}\Big\{|R|^2_{PR} - \mathrm{tr}(M_2^T M_2 \mathbf{D})\Big\}, \text{ where} \tag{3.2}$$
$$R_{PR} = \mathbf{Y} - X\hat{\beta}_{PR}. \tag{3.3}$$

The original PR estimators, given in (3.2) and (3.1) may be seen as a one-step version of the above system of equations. Our first proposed modification is to iterate between (3.2) and (3.1) to convergence, and consider the simultaneous solution to these equations as our parameter estimators.

Our simulation studies, some of which are reported later in this paper, show that $\hat{\psi}_{PR2}$ presents a substantial improvement in finite-sample performance over the original $\hat{\psi}_{PR}$, and also a noticeable but small improvement over $\hat{\psi}_{PR1}$. While these results are reasonably satisfactory and promising, we still noted the substantial positive bias in $\hat{\psi}_{PR2}$, as seen earlier for $\hat{\psi}_{PR}$ and $\hat{\psi}_{PR1}$. As noted earlier, residuals whose values are very close to zero, or very high in absolute value, are problematic for linking variance component estimation. To address this issue, our proposal is to only use those values for which $|\log(D_i^{-1}R_i^2|) < C$ for some cut-off value $C$. This eliminates observations with very small residual values, as well as very high residual values. Define the diagonal matrix $\tilde{W}$ with entries

$$\tilde{W}_{ij} = \mathcal{I}_{\{i=j\}}\mathcal{I}_{\{|\log(D_i^{-1}R_i^2)|<C\}}.$$

Based on this, our next version of the Prasad-Rao type of estimators is as follows:

$$\hat{\psi}_{PR3} = (n-p)^{-1}\Big\{R^T\tilde{W}R - \sum_{i=1}^{n} D_i + \sum_{i=1}^{n}(\mathbf{P_x D})_i\Big\}. \tag{3.4}$$

The estimators (3.2) and (3.4) are designed to make minimal changes to the original PR estimator $\hat{\psi}_{PR}$ for the linking variance given in (1.8). In the former case, we propose the natural optimization of simultaneously solving the equations that arise in Prasad and Rao (1990). In the

latter case, we essentially use a trimmed sum instead of the full sum, a well-established procedure for ensuring a reasonable breakdown value of the resulting estimator. We essentially use a multiplicative trimming in $\tilde{W}$ owing to the fact that both $D_i$ and $R_i^2$ are positive, and as a random variable $R_i^2$ has a skewed distribution.

Our final estimator in the PR-framework combines the features of $\hat{\psi}_{PR2}$ and $\hat{\psi}_{PR3}$. That is, we propose that in our next estimator, we solve simultaneously for the estimators of $\beta$ and $\psi$ as in (1.12) and (1.13), or as in (3.1) and (3.2). We also require that only those residuals be used that are neither too small or too large. Recall that the definition of $W$ is given in (1.10) and that of $R_{PR}$ is given in (3.3). Define

$$\check{W}_{ij} = \mathcal{I}_{\{i=j\}}\mathcal{I}_{\{|\log(D_i^{-1}R_{PRi}^2)|<C\}},$$
$$M_4 = \mathbb{I}_n - X\big(X^TWX\big)^{-1}X^TW.$$

Thus, our next set of estimators are given as the simultaneous solution to

$$\hat{\beta}_{PR} = \big(X^TWX\big)^{-1}X^TW\mathbf{Y}, \text{ and} \tag{3.5}$$
$$\hat{\psi}_{PR4} = \mathrm{tr}(M_2^T\check{W}M_2)^{-1}\Big\{|R|_{PR}^2 - \mathrm{tr}(M_2^T\check{W}M_2\mathbf{D})\Big\}. \tag{3.6}$$

In our simulations, it turned out that convergence was extremely fast, typically requiring only 2 iterations for the successive values of $\hat{\psi}_{PR4}$ to be within $10^{-4}$ of each other.

Based on the considerations discussed above, we propose a modified version of the FH estimator as well. Suppose we start with our initial estimator $\hat{\beta}_{FH}$, and define the initial $R_{FH} = \mathbf{Y} - X\hat{\beta}_{FH}$. Define

$$\vec{W}_{ij} = \mathcal{I}_{\{i=j\}}\mathcal{I}_{\{|\log(D_i^{-1}R_{FHi}^2)|<C\}}.$$

Based on this initial value, our modified FH estimators for $\beta$ and $\psi$ are the simultaneous solutions to

$$\beta = \big(X^T\vec{W}\mathbf{B}\mathbf{D}^{-1}X\big)^{-1}X^T\vec{W}\mathbf{B}\mathbf{D}^{-1}\mathbf{Y}, \tag{3.7}$$
$$1 = (n-p)^{-1}(\mathbf{Y}-X\beta)^T\vec{W}\big(\mathbf{D}+\psi\mathbb{I}_n\big)^{-1}(\mathbf{Y}-X\beta)^T, \tag{3.8}$$
$$R_{FH} = \mathbf{Y} - X\beta. \tag{3.9}$$

The solution to the above equations are denoted by $\hat{\beta}_{FH2}$ and $\hat{\psi}_{FH2}$. Note that at each iteration the residuals are updated, making the optimization problem interesting.

## 4 Results from Simulation Experiments

### 4.1 Study with Large Sample Size

Our first study is on comparing the proposed modified estimators, namely PR.2, PR.3, PR.4 and FH.2 in the cases described in Section 2. In Figure 2, we present the results for all the methods.

The framework of the studies are exactly as described in Section 2, namely, $n = 500$ and $\psi = 1$, the experiments are replicated $K = 1000$ times. As in Figure 1, the left panel corresponds to $(D_1, \ldots, D_5) = (0.1, 0.4, 0.5, 0.6, 4.0)$, the right panel corresponds to all $D_i$'s equal to one.

We notice from the left panel of Figure 2 that PR.2 is comparable to PR.1, and while PR.3 offers considerably reduced bias and reduced variability, its box and whiskers plot still does not cover the true value. On the other hand, PR.4 and FH.2 perform as desired. While there still is a bit of positive bias, it is not substantial and significant enough for the true value to be completely out of range. We report some additional details in Table 1, where we present the mean, bias, variance, the bias of the median, and the percentage of cases where the linking variance estimator was below $10^{-4}$, for this case. We also present the number of iterations required for the computation of PR.4 method to converge.

A natural question, given the performances of PR.3 and PR.4 in the left panel of Figure 2, is whether the iterated equation solving steps of PR.2 are needed at all, since the substantial reduction of bias seems to be driven primarily by the deletion of residuals with very high and very low values. The right panel of Figure 2 shows that just deleting cases with very high and very low residuals is not enough, and in this case FH, PR.1, PR.2, PR.4 and FH.2 perform well. In fact, PR.3 is a poor performer in this situation, though not as poor as the original PR. We report additional details in Table 2, where we present the mean, bias, variance, the bias of the median, and the percentage of cases where the linking variance estimator was below $10^{-4}$, for this case. We also present the number of iterations required for the computation of PR.4 method to converge. It is noticeable that in this case only a single iteration leads to convergence, and that PR.1 and PR.2 are identical up to high number of significant digits.

### 4.2 Study with Moderate Sample Size

We repeat the entire exercise, however, this time with sample size $n = 50$, thus making this a moderate sample size experiment. The boxplots for the linking variance component estimators are given in Figure 3. Additional details from the simulations are given in Table 3 and Table 4. From Figure 3, it is evident that all the methods except for the original PR method perform well, with FH, PR.4 and FH.2 performing exceedingly well under both layouts of the $D$-values, and PR.1, PR.2 and PR.3 performing excellently in at least one case. Note however from Table 3 that PR.4, and from Table 4 that FH, PR.1, PR.2, FH.2 can take negative (or very low) values when the sample size is not extremely high. Overall, the benefits for modifying the estimators to ensure robustness seem to be applicable in this case also.

### 4.3 Study with Low Sample Size

We again repeat the entire exercise, however, this time with sample size $n = 15$, thus making this a low sample size experiment. Note that this sample size now corresponds to the that of Datta et al. (2005). The boxplots for the linking variance component estimators are given in Figure 4. Additional details from the simulations are given in Table 5 and Table 6. From Figure 4, it is

evident that all the methods perform decently. Note that in this case the boxplot for the orignal PR estimator contains the true value of $\psi$. In terms of the plots alone, PR.4 and FH.2 stand out as being always reliable and accurate.

However, it is evident from Table 5 and Table 6 that several of the proposed new methods, along with FH and PR.1, can produce the occasional very low linking variance component estimator. This is most noticeable in the case of PR.1, PR.2 and PR.4 in the case of unequal $D$-values, and for FH, PR.1, PR.2, PR.4 and FH.2 in the equal $D$-values case.

### 4.4 On the Mean Squared Prediction Error

It is of interest to understand how the different estimators of $\psi$ affect the *mean squared prediction error* (MSPE), described in (1.7) earlier. As an illustrative example, we consider the case of moderate sample size $n = 50$, and $(D_1, \ldots, D_5)$ = $(0.1, 0.4, 0.5, 0.6, 4.0)$. In Table 7 we present the MSPE figures obtained out of $K = 1000$ replications of the simulation experiment. As can be seen from this table, the modified estimators of $\psi$ result is substantially reduced MSPE values. This is especially noticeable in the case of PR.3, PR.4 and FH.2. This aspect deserves further study and will be pursued later elsewhere. We do not report the MSPE values for the other cases owing to space, and because the findings do not differ in substance from that of Table 7.

Notice that the expression of MSPE in (1.7) is dependent on the unknown parameters, hence in practice this has to be estimated. The asymptotic expressions for the MSPE under different modified estimators require further study, and should relate to asymptotics of robust estimators. Another option is to use resampling methods, which we also propose to study in future.

## 5 Conclusions and Future Work

The various simulations reported in this paper, and other estensive simulations carried out and not reported here, suggest that $(i)$ simultaneously solving for estimators of $\beta$ and $\psi$, and $(ii)$ trimming out very high and very low residual values substantially improves the finite sample properties of the linking variance estimator. While the improvements in the PR estimators are dramatic, there is very significant improvement in the FH estimator for the linking variance as well.

Theoretical properties of the modified estimators need to be studied. Such properties would not just be about asymptotic efficiency, but robustness properties as well. Also, there seems to be room for improvement in the nature of the trimming, and it is possibly that asymmetric trimming, or winsorization or reweighting may produce even better answers. We have not eliminated the possibility of a negative valued estimator for the linking variance component.

**Acknowledgments**

This research is partially supported by the National Science Foundation (NSF) under grants # DMS-1622483 and # DMS-1737918. I also thank the reviewers and editors for their comments, which helped improve the paper.

**References**

Chatterjee, S. and A. S. Hadi (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*,**1**(3), 379 – 393.

Chatterjee, S. and A. S. Hadi (2015). *Regression Analysis by Example*. John Wiley & Sons.

Chatterjee, S., P. Lahiri, and H. Li (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics*, **36**(3), 1221 – 1245.

Das, K., J. Jiang, and J. N. K. Rao (2004). Mean squared error of empirical predictor. *The Annals of Statistic*,**32**(2), 818 – 840.

Datta, G. S., P. Hall, and A. Mandal (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, **106**(493), 362 – 374.

Datta, G. S., J. N. K. Rao, and D. D. Smith (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, **92**(1), 183 – 196.

Fay III, R. E. and R. A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**(366), 269–277.

Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation (with discussion). *Test*, **15**(1), 1 – 96.

Jiang, J., P. Lahiri, and S.-M. Wan (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics*, **30**(6), 1782 – 1810.

Li, H. and P. Lahiri (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, **101**(4), 882 – 892.

Molina, I., J. N. K. Rao, and G. S. Datta (2015). Small area estimation under a Fay–Herriot model with preliminary testing for the presence of random area effects. *Survey Methodology*, **41**(1), 1 – 19.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**(1), 40 – 68.

Pfeffermann, D. and H. Glickman (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. *ASA Section on Survey Research Methods Proceedings*, 4167–4178.

Prasad, N. G. N. and J. N. K. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**(409), 163 – 171.

Rao, J. (2005). Inferential issues in small area estimation: Some new developments. *Statistics in Transition*, **7**(3), 513 – 526.

Rao, J. N. K. (2015). Inferential issues in model-based small area estimation: Some new developments. *Statistics in Transition*, **16**(4), 491 – 510.

Rao, J. N. K. and I. Molina (2015). *Small Area Estimation*. John Wiley & Sons.

Salvati, N., N. Tzavidis, M. Pratesi, and R. Chambers (2012). Small area estimation via M-quantile geographically weighted regression. *Test*, **21**(1), 1 – 28.

Sugasawa, S. and T. Kubokawa (2015). Parametric transformed Fay–Herriot model for small area estimation. *Journal of Multivariate Analysis*, **139**, 295 – 311.

Yoshimori, M. and P. Lahiri (2014a). A new adjusted maximum likelihood method for the Fay–Herriot small area model. *Journal of Multivariate Analysis*, **124**, 281 – 294.

Yoshimori, M. and P. Lahiri (2014b). A second-order efficient empirical Bayes confidence interval. *The Annals of Statistics*, **42**(4), 1233 – 1261.

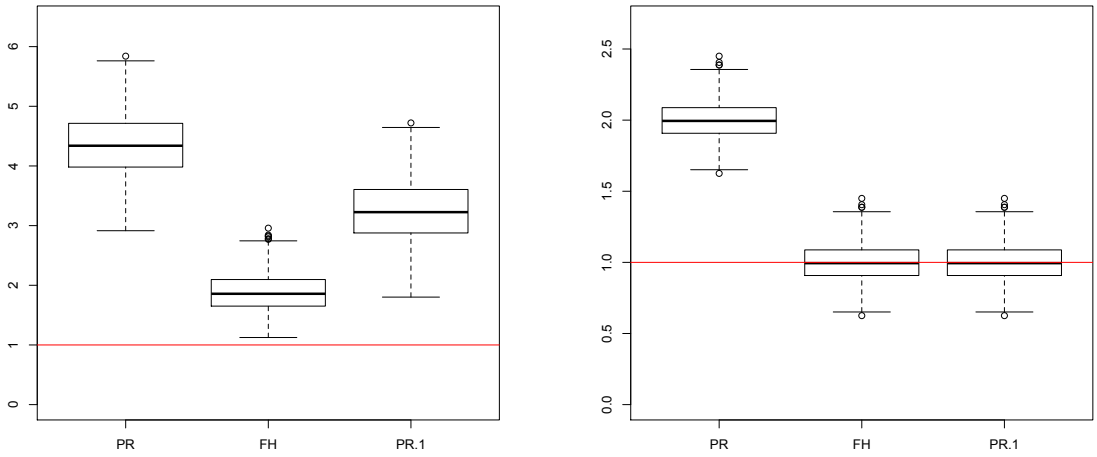


Figure 1: **Boxplots of the Prasad-Rao (PR), Fay-Herriot (FH), and the first modified Prasad-Rao (PR.1) linking variance estimators, based on $K = 1000$ replications of two experiments. The sample size is $n = 500$, and $\psi = 1$ for both panels. The left panel corresponds to $(D_1, \ldots, D_5)$ = $(0.1, 0.4, 0.5, 0.6, 4.0)$, the right panel corresponds to all $D_i$'s equal to one.**

Table 1: **Table of mean, percentage of very low (below $10^{-4}$) values, bias, variance mean squared error, and bias of the median for various linking variance estimators, based on $K = 1000$ replications. The sample size is $n = 500$, $\psi = 1$ and $(D_1, \ldots, D_5)$ = $(0.1, 0.4, 0.5, 0.6, 4.0)$.**

| | PR | FH | PR.1 | PR.2 | PR.3 | PR.4 | FH.2 | Iter |
|---|---|---|---|---|---|---|---|---|
| Mean | 4.357 | 1.8829 | 3.245 | 3.236 | 1.734 | 1.4272 | 1.2467 | 1.9090 |
| $< 10^{-4}$ | 0.000 | 0.0000 | 0.000 | 0.000 | 0.000 | 0.0000 | 0.0000 | |
| Bias | 3.357 | 0.8829 | 2.245 | 2.236 | 0.734 | 0.4272 | 0.2467 | |
| Variance | 0.245 | 0.0977 | 0.245 | 0.244 | 0.026 | 0.0434 | 0.0367 | 0.0828 |
| MSE | 11.517 | 0.8773 | 5.283 | 5.243 | 0.564 | 0.2259 | 0.0976 | |
| Median-Bias | 3.339 | 0.8576 | 2.226 | 2.217 | 0.727 | 0.4188 | 0.2268 | |

Table 2: **Table of mean, percentage of very low (below $10^{-4}$) values, bias, variance mean squared error, and bias of the median for various linking variance estimators, based on $K = 1000$ replications. The sample size is $n = 500$, $\psi = 1$ and all the $D_i$'s equal to one.**

| | PR | FH | PR.1 | PR.2 | PR.3 | PR.4 | FH.2 | Iter |
|---|---|---|---|---|---|---|---|---|
| Mean | 1.9995 | 0.9995 | 0.9995 | 0.9995 | 1.399 | 0.892 | 0.961 | 1 |
| $< 10^{-4}$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | |
| Bias | 0.9995 | -0.0005 | -0.0005 | -0.0005 | 0.399 | -0.1085 | -0.039 | |
| Variance | 0.0165 | 0.0165 | 0.0165 | 0.0165 | 0.006 | 0.008 | 0.015 | 0 |
| MSE | 1.015 | 0.0165 | 0.0165 | 0.0165 | 0.165 | 0.020 | 0.0169 | |
| Median-Bias | 0.9945 | -0.0054 | -0.0055 | -0.0055 | 0.398 | -0.1095 | -0.039 | |

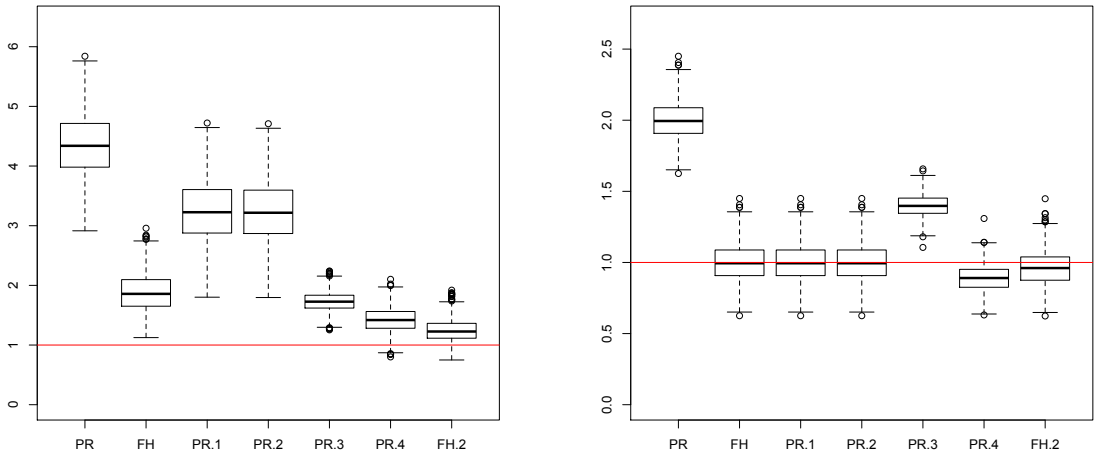


Figure 2: **Boxplots of the Prasad-Rao (PR), Fay-Herriot (FH), four modified Prasad-Rao (PR.1 - PR.4) and one modified Fay-Herriot (FH.2) linking variance estimators, based on $K = 1000$ replications of two experiments. The sample size is $n = 500$, and $\psi = 1$ for both panels. The left panel corresponds to $(D_1, \ldots, D_5) = (0.1, 0.4, 0.5, 0.6, 4.0)$, the right panel corresponds to all $D_i$'s equal to one.**

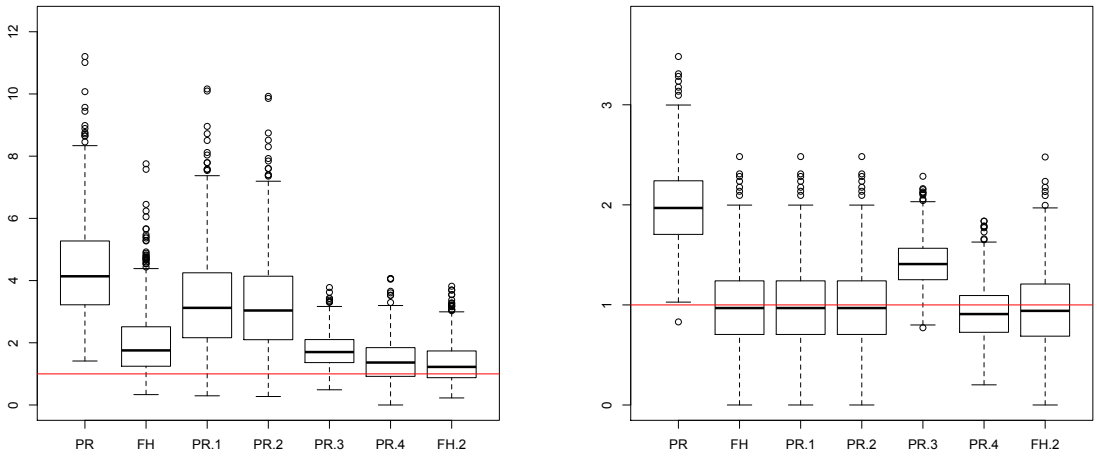


Figure 3: **Boxplots of the Prasad-Rao (PR), Fay-Herriot (FH), four modified Prasad-Rao (PR.1 - PR.4) and one modified Fay-Herriot (FH.2) linking variance estimators, based on $K = 1000$ replications of two experiments. The sample size is $n = 50$, and $\psi = 1$ for both panels. The left panel corresponds to $(D_1, \ldots, D_5) = (0.1, 0.4, 0.5, 0.6, 4.0)$, the right panel corresponds to all $D_i$'s equal to one.**

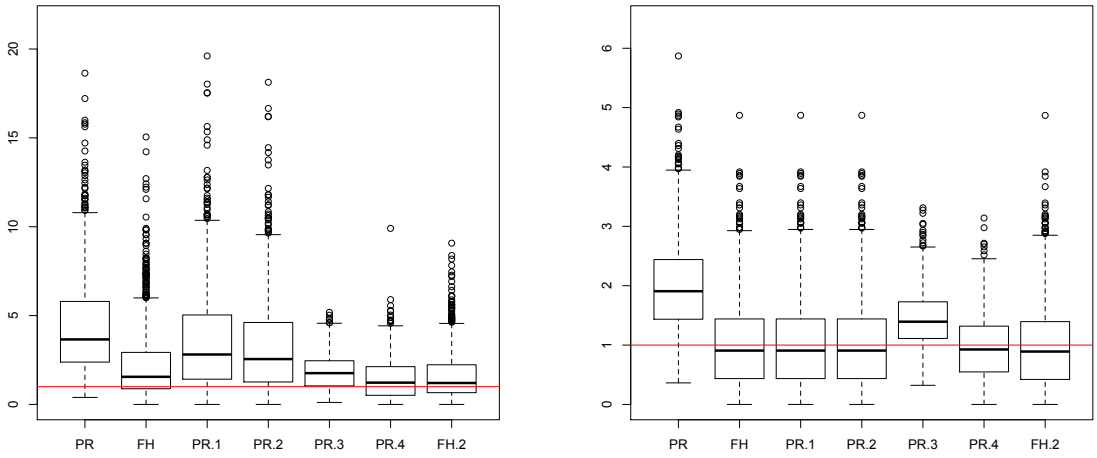


Figure 4: **Boxplots of the Prasad-Rao (PR), Fay-Herriot (FH), four modified Prasad-Rao (PR.1 - PR.4) and one modified Fay-Herriot (FH.2) linking variance estimators, based on $K = 1000$ replications of two experiments. The sample size is $n = 15$, and $\psi = 1$ for both panels. The left panel corresponds to $(D_1, \ldots, D_5) = (0.1, 0.4, 0.5, 0.6, 4.0)$, the right panel corresponds to all $D_i$'s equal to one.**

Table 3: **Table of mean, percentage of very low (below $10^{-4}$) values, bias, variance mean squared error, and bias of the median for various linking variance estimators, based on $K = 1000$ replications. The sample size is $n = 50$, $\psi = 1$ and $(D_1, \ldots, D_5) = (0.1, 0.4, 0.5, 0.6, 4.0)$.**

| | PR | FH | PR.1 | PR.2 | PR.3 | PR.4 | FH.2 | Iter |
|---|---|---|---|---|---|---|---|---|
| Mean | 4.336 | 1.983 | 3.292 | 3.204 | 1.757 | 1.415 | 1.357 | 1.967 |
| $< 10^{-4}$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.200 | 0.000 | |
| Bias | 3.336 | 0.983 | 2.292 | 2.204 | 0.757 | 0.415 | 0.357 | |
| Variance | 2.208 | 1.017 | 2.278 | 2.179 | 0.286 | 0.450 | 0.409 | 0.032 |
| MSE | 13.336 | 1.983 | 7.532 | 7.035 | 0.859 | 0.622 | 0.537 | |
| Median-Bias | 3.138 | 0.755 | 2.123 | 2.038 | 0.702 | 0.365 | 0.224 | |

Table 4: **Table of mean, percentage of very low (below $10^{-4}$) values, bias, variance mean squared error, and bias of the median for various linking variance estimators, based on $K = 1000$ replications. The sample size is $n = 50$, $\psi = 1$ and all the $D_i$'s equal to one.**

| | PR | FH | PR.1 | PR.2 | PR.3 | PR.4 | FH.2 | Iter |
|---|---|---|---|---|---|---|---|---|
| Mean | 1.991 | 0.991 | 0.991 | 0.991 | 1.413 | 0.915 | 0.962 | 1.000 |
| $< 10^{-4}$ | 0.000 | 0.100 | 0.100 | 0.100 | 0.000 | 0.000 | 0.100 | |
| Bias | 0.991 | -0.009 | -0.009 | -0.009 | 0.413 | -0.085 | -0.038 | |
| Variance | 0.162 | 0.162 | 0.162 | 0.162 | 0.059 | 0.080 | 0.152 | 0.000 |
| MSE | 1.143 | 0.162 | 0.162 | 0.162 | 0.230 | 0.087 | 0.153 | |
| Median-Bias | 0.969 | -0.031 | -0.031 | -0.031 | 0.408 | -0.091 | -0.059 | |

Table 5: **Table of mean, percentage of very low (below $10^{-4}$) values, bias, variance mean squared error, and bias of the median for various linking variance estimators, based on $K = 1000$ replications. The sample size is $n = 15$, $\psi = 1$ and $(D_1, \ldots, D_5)$ = $(0.1, 0.4, 0.5, 0.6, 4.0)$.**

| | PR | FH | PR.1 | PR.2 | PR.3 | PR.4 | FH.2 | Iter |
|---|---|---|---|---|---|---|---|---|
| Mean | 4.380 | 2.271 | 3.538 | 3.229 | 1.867 | 1.435 | 1.644 | 1.985 |
| $< 10^{-4}$ | 0.000 | 0.300 | 3.400 | 3.800 | 0.000 | 6.300 | 0.700 | |
| Bias | 3.380 | 1.271 | 2.538 | 2.229 | 0.867 | 0.435 | 0.644 | |
| Variance | 7.617 | 4.339 | 8.649 | 7.424 | 0.981 | 1.331 | 1.903 | 0.015 |
| MSE | 19.042 | 5.954 | 15.092 | 12.392 | 1.733 | 1.520 | 2.318 | |
| Median-Bias | 2.659 | 0.555 | 1.811 | 1.553 | 0.760 | 0.226 | 0.208 | |

Table 6: **Table of mean, percentage of very low (below $10^{-4}$) values, bias, variance mean squared error, and bias of the median for various linking variance estimators, based on $K = 1000$ replications. The sample size is $n = 15$, $\psi = 1$ and all the $D_i$'s equal to one.**

| | PR | FH | PR.1 | PR.2 | PR.3 | PR.4 | FH.2 | Iter |
|---|---|---|---|---|---|---|---|---|
| Mean | 2.009 | 1.022 | 1.022 | 1.022 | 1.447 | 0.962 | 0.996 | 1.000 |
| $< 10^{-4}$ | 0.000 | 6.900 | 6.900 | 6.900 | 0.000 | 3.400 | 7.100 | |
| Bias | 1.009 | 0.022 | 0.022 | 0.022 | 0.447 | -0.038 | -0.004 | |
| Variance | 0.614 | 0.583 | 0.583 | 0.583 | 0.224 | 0.316 | 0.549 | 0.000 |
| MSE | 1.631 | 0.584 | 0.584 | 0.584 | 0.424 | 0.317 | 0.549 | |
| Median-Bias | 0.907 | -0.093 | -0.093 | -0.093 | 0.394 | -0.074 | -0.109 | |

Table 7: **Table of the mean squared prediction error (MSPE), scaled by a factor of 100, for the EBLUP, based on $K = 1000$ replications. The sample size is $n = 50$, $\psi = 1$ and $(D_1, \ldots, D_5)$ = $(0.1, 0.4, 0.5, 0.6, 4.0)$. The MSPE values corresponding to small areas with identical $D_i$ values have been averaged.**

| D | PR | FH | PR.1 | PR.2 | PR.3 | PR.4 | FH.2 |
|---|---|---|---|---|---|---|---|
| 0.1 | 1.03 | 1.28 | 1.11 | 1.12 | 1.25 | 2.10 | 1.64 |
| 0.4 | 14.14 | 14.71 | 14.21 | 14.23 | 14.63 | 17.48 | 16.26 |
| 0.5 | 21.34 | 21.19 | 21.14 | 21.14 | 21.00 | 23.64 | 22.49 |
| 0.6 | 29.04 | 27.47 | 28.22 | 28.16 | 27.05 | 29.67 | 28.41 |
| 4.0 | 509.82 | 272.06 | 412.24 | 403.28 | 203.38 | 177.61 | 185.78 |