International Scholarly Research Network
ISRN Probability and Statistics
Volume 2012, Article ID 103860, 15 pages
doi:10.5402/2012/103860

Research Article

A Two-Parameter Ratio-Product-Ratio Estimator Using Auxiliary Information

Peter S. Chami,[1,2] Bernd Sing,[2] and Doneal Thomas[2]

[1] *The Warren Alpert Medical School of Brown University, Box G-A1, Providence, RI 02912, USA*
[2] *Department of Computer Science, Mathematics and Physics, Faculty of Science and Technology, The University of the West Indiesat Cave Hill, P.O. Box 64, Cave Hill, Bridgetown, St. Michael BB11000, Barbados*

Correspondence should be addressed to Doneal Thomas, donealt@gmail.com

Received 15 August 2012; Accepted 2 September 2012

Academic Editors: P. E. Jorgensen and V. Makis

We propose a two-parameter ratio-product-ratio estimator for a finite population mean in a simple random sample without replacement following the methodology in the studies of Ray and Sahai (1980), Sahai and Ray (1980), A. Sahai and A. Sahai (1985), and Singh and Espejo (2003).The bias and mean squared error of our proposed estimator are obtained to the first degree of approximation. We derive conditions for the parameters under which the proposed estimator has smaller mean squared error than the sample mean, ratio, and product estimators. We carry out an application showing that the proposed estimator outperforms the traditional estimators using groundwater data taken from a geological site in the state of Florida.

1. Introduction

We consider the following setting. For a finite population of size N, we are interested in estimating the *population mean* $\overline{Y}$ of the *main variable* y (taking values y_i for $i = 1, \ldots, N$) from a *simple random sample* of size n (where $n < N$) drawn without replacement. We also know the population mean $\overline{X}$ for the *auxiliary variable* x (taking values x_i for $i = 1, \ldots, N$). We use the notation $\overline{y}$ and $\overline{x}$ for the *sample means*, which are unbiased estimators of the population means $\overline{Y}$ and $\overline{X}$, respectively.

We denote the *population variances* of Y and X by

$$S_Y^2 = \mathbb{V}(Y) = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2, \qquad S_X^2 = \mathbb{V}(X) = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2, \tag{1.1}$$

respectively. Furthermore, we define the *coefficient of variation* of Y and X as

$$C_Y = \frac{S_Y}{\overline{Y}}, \qquad C_X = \frac{S_X}{\overline{X}}, \tag{1.2}$$

respectively, and the *coefficient of correlation* C between the two variables as

$$C = \rho \cdot \frac{C_Y}{C_X}, \tag{1.3}$$

where $\rho = S_{XY}/S_X S_Y$ denotes the *population Pearson correlation coefficient*.

As estimators of the population mean $\overline{Y}$, the usual *sample mean* $\overline{y}$, the *ratio estimator* $\overline{y}_R = (\overline{y}/\overline{x})\cdot\overline{X}$, and the *product estimator* $\overline{y}_P = (\overline{x}\cdot\overline{y})/\overline{X}$ are used. Murthy [1] and Sahai and Ray [2] compared the relative precision of these estimators and showed that the ratio estimator, sample mean, and product estimator are most efficient when $C > 1/2$, $-1/2 \leq C \leq 1/2$ and $C < -1/2$, respectively. In other words, when the study variate y and the auxiliary variate x show high positive correlation, then the ratio estimator shows the highest efficiency; when they are highly negative correlated, then the product estimator has the highest efficiency; when the variables show a weak correlation only, then the sample mean is preferred. (In the paper, we say an estimator is "most efficient" or has the "highest efficiency," if it has the lowest mean squared error (MSE) of all the estimators considered.)

For estimating the population mean $\overline{Y}$ of the main variable, we proposed the following two-parameter ratio-product-ratio estimator:

$$\overline{y}_{\alpha,\beta} = \alpha\left[\frac{(1-\beta)\overline{x}+\beta\overline{X}}{\beta\overline{x}+(1-\beta)\overline{X}}\right]\overline{y} + (1-\alpha)\left[\frac{\beta\overline{x}+(1-\beta)\overline{X}}{(1-\beta)\overline{x}+\beta\overline{X}}\right]\overline{y}, \tag{1.4}$$

where α, β are real constants. Our goal in this paper is to derive values for these constants α, β such that the bias and/or the mean squared error (MSE) of $\overline{y}_{\alpha,\beta}$ is minimal. In fact, in Section 5 we are able to use the two parameters α and β to obtain an estimator $\overline{y}^*(C)$ that is (up to first degree of approximation) both unbiased and has minimal MSE $((N-n)/(N\cdot n))S_Y^2(1-\rho^2)$; it was Srivastava [3, 4] who showed that this is the minimal possible MSE up to first degree of approximation for a large class of estimators (to which the one in (1.4) also belongs). The estimator $\overline{y}^*(C)$ thus corrects the limitations of the traditional estimators $\overline{y}$, $\overline{y}_R$, and $\overline{y}_P$ which are to be used for a specific range of the parameter C and, in addition, outperforms the traditional estimators by having the least MSE.

Note that $\overline{y}_{\alpha,\beta} = \overline{y}_{1-\alpha,1-\beta}$, that is, the estimator $\overline{y}_{\alpha,\beta}$ is invariant under a point reflection through the point $(\alpha,\beta) = (1/2, 1/2)$. In the point of symmetry $(\alpha,\beta) = (1/2, 1/2)$, the estimator reduces to the sample mean; that is, we have $\overline{y}_{1/2,1/2} = \overline{y}$. In fact, on the whole line $\beta = 1/2$ our proposed estimator reduces to the sample mean estimator, that is, $\overline{y}_{\alpha,1/2} = \overline{y}$. Similarly, we get $\overline{y}_{1,0} = \overline{y}_{0,1} = (\overline{xy})/\overline{X} = \overline{y}_P$ (product estimator) and $\overline{y}_{0,0} = \overline{y}_{1,1} = (\overline{y}\overline{X})/\overline{x} = \overline{y}_R$ (ratio estimator). Its simplicity (essentially just using convex combinations and/or a ratio of convex combinations) and that all three traditional estimators (sample mean, product, and ratio estimators) can be obtained from it by choosing appropriate parameters are the reasons why we study the estimator in (1.4) and compare it to the three traditional estimators.

However, in the outlook, Section 6.5, we also compare this estimator to more sophisticated estimators for the application in the groundwater data considered here.

2. First-Degree Approximation to the Bias

In order to derive the bias of $\overline{y}_{\alpha,\beta}$ up to $O(1/n)$, we set

$$e_1 = \frac{\overline{y} - \overline{Y}}{\overline{Y}}, \qquad e_2 = \frac{\overline{x} - \overline{X}}{\overline{X}}. \tag{2.1}$$

Thus, we have $\overline{y} = \overline{Y}(1 + e_1)$ and $\overline{x} = \overline{X}(1 + e_2)$, and the relative estimators are given by

$$\widehat{y} = \frac{\overline{y}}{\overline{Y}} = (1 + e_1), \qquad \widehat{x} = \frac{\overline{x}}{\overline{X}} = (1 + e_2). \tag{2.2}$$

Thus, the expectation value of the e_i's is

$$\mathbb{E}(e_i) = 0 \quad \text{for } i = 1, 2, \tag{2.3}$$

and under a simple random sample without replacement, the relative variances are

$$\begin{aligned} \mathbb{V}_{\text{rel}}(\overline{y}) &= \frac{\mathbb{V}(\overline{y})}{\overline{Y}^2} = \mathbb{E}\left(e_1^2\right) = \mathbb{V}(e_1) = \frac{1-f}{n}\left(\frac{S_Y}{\overline{Y}}\right)^2, \\ \mathbb{V}_{\text{rel}}(\overline{x}) &= \frac{\mathbb{V}(\overline{x})}{\overline{X}^2} = \mathbb{E}\left(e_2^2\right) = \mathbb{V}(e_2) = \frac{1-f}{n}\left(\frac{S_X}{\overline{X}}\right)^2, \end{aligned} \tag{2.4}$$

where $f = n/N$ is the *sampling fraction*. Also, we have

$$\mathbb{E}(e_1 e_2) = \frac{1-f}{n}\rho C_Y C_X, \tag{2.5}$$

see [2, 5, 6]. Furthermore, we note that $\mathbb{E}(e_1^2 e_2^2) = O(1/n^2)$, and $\mathbb{E}(e_1^i e_2^j) = 0$ when $(i + j)$ is an odd integer.

Now reexpressing (1.4) in terms of e_i's and by substituting $\overline{x}$ and $\overline{y}$, we have

$$\overline{y}_{\alpha,\beta} = \alpha\left[\frac{1 + e_2 - \beta e_2}{1 + \beta e_2}\right]\overline{Y}(1 + e_1) + (1 - \alpha)\left[\frac{1 + \beta e_2}{1 + e_2 - \beta e_2}\right]\overline{Y}(1 + e_1). \tag{2.6}$$

In the following, we assume that $|e_2| < \min\{1/|\beta|, 1/|1 - \beta|\}$, and therefore we can expand $(1 + \beta e_2)^{-1}$ and $(1 + (1 - \beta)e_2)^{-1}$ as a series in powers of e_2. (We note that $\min\{1/|\beta|, 1/|1 - \beta|\}$ attains its maximal value 2 at $\beta = 1/2$.) We get up to $O(e_2^3)$

$$\overline{y}_{\alpha,\beta} = (1 + e_1)\overline{Y} \cdot \left[1 - (1 - 2\alpha)(1 - 2\beta)e_2 + (1 - \alpha - \beta)(1 - 2\beta)e_2^2 + O\left(e_2^3\right)\right]. \tag{2.7}$$

We assume that the sample is large enough to make $|e_2|$ so small that contributions from powers of e_2 of degree higher than two are negligible; compare [6]. By retaining powers up to e_2^2, we get

$$\overline{y}_{\alpha,\beta} - \overline{Y} \approx \overline{Y}\left\{e_1 - (1+e_1)\left[(1-2\alpha)(1-2\beta)e_2 - (1-\alpha-\beta)(1-2\beta)e_2^2\right]\right\}. \tag{2.8}$$

Taking expectations on both sides of (2.8) and substituting $C = \rho(C_Y/C_X)$, we obtain the bias of $\overline{y}_{\alpha,\beta}$ to order $O(n^{-1})$ as

$$\begin{aligned}\mathbb{B}\left(\overline{y}_{\alpha,\beta}\right) &= \mathbb{E}\left(\overline{y}_{\alpha,\beta} - \overline{Y}\right) \\ &\approx \frac{1-f}{n}(1-2\beta)\left[(1-\alpha-\beta) - (1-2\alpha)\rho\frac{C_Y}{C_X}\right]C_X^2\overline{Y} \\ &= \frac{1-f}{n}(1-2\beta)\left[1-\alpha-\beta-(1-2\alpha)C\right]C_X^2\overline{Y}.\end{aligned} \tag{2.9}$$

Equating (2.9) to zero, we obtain

$$\beta = \frac{1}{2} \quad \text{or} \quad \beta = 1-\alpha-C+2\alpha C. \tag{2.10}$$

The proposed ratio-product-ratio estimator $\overline{y}_{\alpha,\beta}$, substituted with the values of β from (2.10), becomes an (approximately) unbiased estimator for the population mean $\overline{Y}$. In the three-dimensional parameter space $(\alpha, \beta, C) \in \mathbb{R}^3$, these unbiased estimators lie on a plane (in the case $\beta = 1/2$) and on a saddle-shaped surface, see Figure 1(a). Furthermore, as the sample size n approaches the population size N, the bias of $\overline{y}_{\alpha,\beta}$ tends to zero, since the factor $(1-f)/n$ clearly tends to zero.

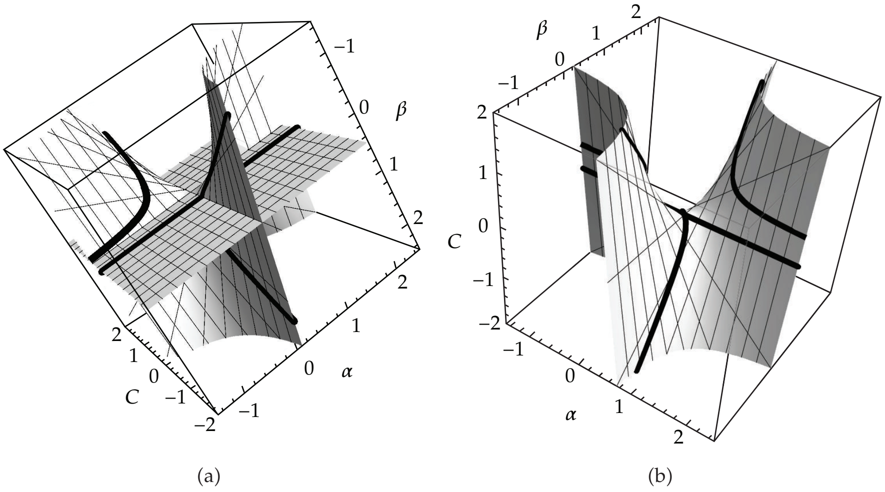
3. Mean Squared Error of $\overline{y}_{\alpha,\beta}$

We calculate the mean squared error of $\overline{y}_{\alpha,\beta}$ up to order $O(n^{-1})$ by squaring (2.8), retaining terms up to squares in e_1 and e_2, and then taking the expectation. This yields the first-degree approximation of the MSE

$$\mathrm{MSE}_1\left(\overline{y}_{\alpha,\beta}\right) = \frac{1-f}{n}\overline{Y}^2\left\{C_Y^2 + C_X^2(1-2\alpha)(1-2\beta)\left[(1-2\alpha)(1-2\beta)-2C\right]\right\}. \tag{3.1}$$

Taking the gradient $\nabla = (\partial/\partial\alpha, \partial/\partial\beta)$ of (3.1), we get

$$\nabla\mathrm{MSE}_1\left(\overline{y}_{\alpha,\beta}\right) = 4\,\frac{1-f}{n}\overline{Y}^2C_X^2\,\left[(1-2\alpha)(1-2\beta)-C\right](1-2\beta, 1-2\alpha). \tag{3.2}$$



(a) (b)

Figure 1: Surface of “bias-free estimators” defined by (2.9) in the parameter space $(\alpha, \beta, C) \in \mathbb{R}^3$. (b) Surface of “AOE parameters” defined by (3.4). The points of intersection of the two surfaces (see Section 5) are drawn as black curves.

Setting (3.2) to zero to obtain the critical points, we obtain the following solutions:

$$\alpha = \frac{1}{2}, \qquad \beta = \frac{1}{2} \tag{3.3}$$

or

$$C = (1 - 2\alpha)(1 - 2\beta). \tag{3.4}$$

One can check that the critical point in (3.3) is a saddle point unless $C = 0$, in which case we get a local minimum. However, the critical points determined by (3.4) are always local minima; for a given C, (3.4) is the equation of a hyperbola symmetric through $(\alpha, \beta) = (1/2, 1/2)$. Thus, in the three-dimensional parameter space $(\alpha, \beta, C) \in \mathbb{R}^3$, the estimators with minimal MSE (or better, minimal first approximation to the MSE; see calculation in (3.6)) lie on a saddle-shaped surface, see Figure 1(b).

We now calculate the minimal value of the MSE. Substituting (3.3) into the estimator $\overline{y}_{\alpha,\beta}$ yields the unbiased estimator $\overline{y}$ (sample mean) of the population mean $\overline{Y}$. Thus, we arrive at the mean squared error of the sample mean:

$$\text{MSE}\left(\overline{y}_{1/2,1/2}\right) = \text{MSE}(\overline{y}) = \frac{1-f}{n}\overline{Y}^2 C_Y^2 = \frac{1-f}{n} S_Y^2. \tag{3.5}$$

By substituting (3.4) into the estimator, an *asymptotically optimum estimator* (AOE) $\overline{y}_{\alpha,\beta}^{(o)}$ is found. For the first-degree approximation of the MSE, we find (independent of α and β)

$$\text{MSE}_1\left(\overline{y}_{\alpha,\beta}^{(o)}\right) = \frac{1-f}{n}\overline{Y}^2\left(C_Y^2 - C^2 C_X^2\right) = \frac{1-f}{n} S_Y^2\left(1 - \rho^2\right), \tag{3.6}$$

that is, the same minimal mean squared error as found in [2, 5–7]. In fact, Srivastava [3, 4] showed that this is the minimal possible mean squared error up to first degree of approximation for a large class of estimators to which the estimator (1.4) also belongs, for example, for estimators of the form $\overline{y}_g = \overline{y} \cdot g(\overline{x}/\overline{X})$ where g is a C^2-function with $g(1) = 1$. (In [8] it was shown that incorporating sample and population variances of the auxiliary variable might yield an estimator that has a lower MSE than $((1-f)/n)S_Y^2(1-\rho^2)$ especially when the relationship between the study variate y and the auxiliary variate x is markedly nonlinear.) Thus, whatever value C has, we are always able to select an AOE $\overline{y}_{\alpha,\beta}^{(o)}$ from the two-parameter family in (1.4).

4. Comparison of MSEs and Choice of Parameters

Here we compare $\mathrm{MSE}_1(\overline{y}_{\alpha,\beta})$ in (3.1) with the MSE of the product, ratio, and sample mean estimators, respectively. It is known (see [2, 5]) that

$$\mathrm{MSE}(\overline{y}) = \mathbb{V}(\overline{y}) = \frac{1-f}{n}\overline{Y}^2 C_Y^2, \tag{4.1}$$

$$\mathrm{MSE}_1(\overline{y}_R) = \frac{1-f}{n}\overline{Y}^2\left\{C_Y^2 + C_X^2(1-2C)\right\}, \tag{4.2}$$

$$\mathrm{MSE}_1(\overline{y}_P) = \frac{1-f}{n}\overline{Y}^2\left\{C_Y^2 + C_X^2(1+2C)\right\}. \tag{4.3}$$

4.1. *Comparing the MSE of the Product Estimator to Our Proposed Estimator*

From [2, 5–7], we know that, for $C < -1/2$, the product estimator is preferred to the sample mean and ratio estimators. Therefore, we seek a range of α and β values where our proposed estimator $\overline{y}_{\alpha,\beta}$ has smaller MSE than the product estimator.

From (4.3) and (3.1), the following expression can be verified:

$$\mathrm{MSE}_1(\overline{y}_P) - \mathrm{MSE}_1\left(\overline{y}_{\alpha,\beta}\right) = 4\frac{1-f}{n}\overline{Y}^2 C_X^2\left[1+2\alpha\beta-\alpha-\beta\right]\left[C-(2\alpha\beta-\alpha-\beta)\right], \tag{4.4}$$

which is positive if

$$\left[1+2\alpha\beta-\alpha-\beta\right]\left[C-(2\alpha\beta-\alpha-\beta)\right] > 0. \tag{4.5}$$

We obtain the following two cases:

(i) $C > 2\alpha\beta-\alpha-\beta > -1$ (if both factors in (4.5) are positive) or

(ii) $C < 2\alpha\beta-\alpha-\beta < -1$ (if both factors in (4.5) are negative).

Noting that we are only interested in the case $C < -1/2$, we get from (i)

$$-\frac{1}{2} > C > 2\alpha\beta-\alpha-\beta > -1. \tag{4.6}$$

We note that this implies $-1 < C < -1/2$, and the range for α and β where these inequalities hold are explicitly given by the following two cases.

(i) If $\beta < 1/2$, then $(\beta + C)/(2\beta - 1) < \alpha < (\beta - 1)/(2\beta - 1)$.

(ii) If $\beta > 1/2$, then $(\beta - 1)/(2\beta - 1) < \alpha < (\beta + C)/(2\beta - 1)$.

For any given C, we again note that the two regions determined here are symmetric through $(\alpha, \beta) = (1/2, 1/2)$. We also note that the parameters (α, β) which give an AOE (see (3.4)), which for a fixed C lie on a hyperbola, are contained in these regions.

In case (ii), where $C < -1$ (and therefore automatically $C < -1/2$), the following range for α and β can be found.

(i) If $\beta < 1/2$, then $(\beta - 1)/(2\beta - 1) < \alpha < (\beta + C)/(2\beta - 1)$.

(ii) If $\beta > 1/2$, then $(\beta + C)/(2\beta - 1) < \alpha < (\beta - 1)/(2\beta - 1)$.

The same remark as in the previous case applies. Furthermore, note that, for $C = -1$, the product estimator attains the same minimal MSE as our proposed estimator $\overline{y}_{\alpha,\beta}$ on the hyperbola given by (3.6). In Figure 2(a) we show the region in parameter space $(\alpha, \beta, C) \in \mathbb{R}^3$ calculated here and in the next two sections where the proposed estimator works better than the three traditional estimators.

4.2. Comparing the MSE of the Ratio Estimator to Our Proposed Estimator

For $C > 1/2$, the ratio estimator is used instead of the sample mean or product estimator; compare [2, 5–7]. As a result, we are concerned with a range of plausible values for α and β, where $\overline{y}_{\alpha,\beta}$ works better than the ratio estimator.

Taking the difference of (4.2) and (3.1), we have

$$\mathrm{MSE}_1(\overline{y}_R) - \mathrm{MSE}_1\left(\overline{y}_{\alpha,\beta}\right) = 4\frac{1-f}{n}\overline{Y}^2 C_X^2 [2\alpha\beta - \alpha - \beta][C - 1 - (2\alpha\beta - \alpha - \beta)] \tag{4.7}$$

which is positive if

$$[2\alpha\beta - \alpha - \beta][C - 1 - (2\alpha\beta - \alpha - \beta)] > 0. \tag{4.8}$$

Therefore,

(i) $C - 1 > 2\alpha\beta - \alpha - \beta > 0$ or

(ii) $C - 1 < 2\alpha\beta - \alpha - \beta < 0$.

Hence, from solution (i), where $C > 1$, we have the following.

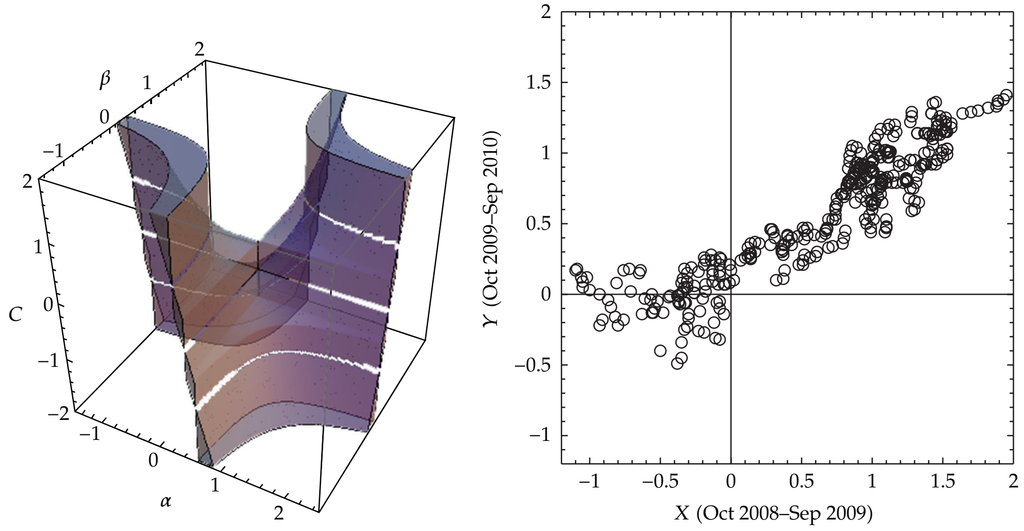
(i) If $\beta < 1/2$, then $(\beta + C - 1)/(2\beta - 1) < \alpha < \beta/(2\beta - 1)$.

(ii) If $\beta > 1/2$, then $\beta/(2\beta - 1) < \alpha < (\beta + C - 1)/(2\beta - 1)$.

Also, from solution (ii), where $1/2 < C < 1$, we obtain the following.

(i) If $\beta < 1/2$, then $\beta/(2\beta - 1) < \alpha < (\beta + C - 1)/(2\beta - 1)$.

(ii) If $\beta > 1/2$, then $(\beta + C - 1)/(2\beta - 1) < \alpha < \beta/(2\beta - 1)$.



(a)

(b)

Figure 2: (a) Region of the parameter space $(\alpha, \beta, C) \in \mathbb{R}^3$ where our proposed estimator $\overline{y}_{\alpha,\beta}$ has lower MSE than the traditional estimators. (b) Scatterplot of study and auxiliary variables for the groundwater data studied in Section 6.

4.3. Comparing the MSE of the Sample Mean to Our Proposed Estimator

Finally, we compare the $\text{MSE}(\overline{y})$ to our proposed estimator, $\text{MSE}(\overline{y}_{\alpha,\beta})$. From [2, 5–7], we know that sample mean estimator is preferred for $-1/2 \le C \le 1/2$.

Taking the difference of (4.1) and (3.1), we get

$$\text{MSE}(\overline{y}) - \text{MSE}_1\left(\overline{y}_{\alpha,\beta}\right) = \frac{1-f}{n}\overline{Y}^2 C_X^2 (1-2\alpha)(1-2\beta)\{2C - (1-2\alpha)(1-2\beta)\} \tag{4.9}$$

which is positive if

$$(1-2\alpha)(1-2\beta)\{2C - (1-2\alpha)(1-2\beta)\} > 0. \tag{4.10}$$

Therefore, either

(i) $\alpha > 1/2$, $\beta > 1/2$ and $C > 1/2(1-2\alpha)(1-2\beta)$,

(ii) $\alpha < 1/2$, $\beta > 1/2$ and $C < 1/2(1-2\alpha)(1-2\beta)$,

(iii) $\alpha > 1/2$, $\beta < 1/2$ and $C < 1/2(1-2\alpha)(1-2\beta)$, or

(iv) $\alpha < 1/2$, $\beta < 1/2$ and $C > 1/2(1-2\alpha)(1-2\beta)$.

Combining these with the condition $-1/2 \le C \le 1/2$, we get the following explicit ranges.

(i) If $0 < C \le 1/2$ and $\beta > 1/2$, then $1/2 < \alpha < (2\beta + 2C - 1)/2(2\beta - 1)$ (from (i)).

(ii) If $0 < C \le 1/2$ and $\beta < 1/2$, then $(2\beta + 2C - 1)/2(2\beta - 1) < \alpha < 1/2$ (from (iv)).

(iii) If $-1/2 \le C < 0$ and $\beta > 1/2$, then $(2\beta + 2C - 1)/2(2\beta - 1) < \alpha < 1/2$ (from (ii)).

(iv) If $-1/2 \le C < 0$ and $\beta < 1/2$, then $1/2 < \alpha < (2\beta + 2C - 1)/2(2\beta - 1)$ (from (iii)).

We note that the case $C = 0$ implies $r = 0$, and thus the sample mean estimator is the estimator with minimal MSE (and, as already noted, $\overline{y} = \overline{y}_{1/2,1/2}$).

In Figure 2(a) we show the region in parameter space $(\alpha, \beta, C) \in \mathbb{R}^3$ where the proposed estimator $\overline{y}_{\alpha,\beta}$ works better than the three traditional estimators. Note that the surface of "AOE parameters" in Figure 1(b) is a subset of this region, except for the values $C = 0$, $C = -1$, and $C = +1$ for which our proposed estimator only works as well as the sample mean, product, and ratio estimator, respectively. (We also remark that the points $(1, 1, 1)$ and $(0, 0, 1)$ (note that $\overline{y}_{1,1} = \overline{y}_{0,0} = \overline{y}_R$), $(0, 1, -1)$ and $(1, 0, -1)$ (note that $\overline{y}_{1,0} = \overline{y}_{0,1} = \overline{y}_R$) as well as the line $(\alpha, 1/2, 0)$ (note that $\overline{y}_{\alpha,1/2} = \overline{y}$) belong to the surface of "AOE parameters" in Figure 1(b).)

5. Unbiased AOE

Combining (2.10) and (3.4), we can calculate the parameters α and β where our proposed estimator becomes—at least up to first approximation—an unbiased AOE. We obtain a line with (recall that on this line our estimator always reduces to the sample mean estimator)

$$\beta = \frac{1}{2}, \qquad C = 0 \tag{5.1}$$

or a "curve" $(\alpha^*(C), \beta^*(C), C) \in \mathbb{R}^3$ in the parameter space with

$$\alpha^*(C) = \frac{1}{2}\left(1 \pm \sqrt{\frac{C}{2C-1}}\right), \qquad \beta^*(C) = \frac{1}{2}\left(1 \pm \sqrt{C(2C-1)}\right). \tag{5.2}$$

We note that the parametric "curve" in (5.2) is only defined for $C \leq 0$ or $C > 1/2$—in fact, this parametric "curve" is three hyperbolas. The surface of "bias-free estimator parameters" in Figure 1(a) and the surface of "AOE parameters" in Figure 1(b) only intersect in these three hyperbolas and the line $\beta = 1/2$ and $C = 0$. In the region $0 < C \leq 1/2$ of the parameter space $(\alpha, \beta, C) \in \mathbb{R}^3$, we have the common situation where minimising MSE comes with a trade-off in bias. The curves of intersection are included in Figure 1. Explicitly, our proposed estimator using the values in (5.2) is given by

$$\begin{aligned} \overline{y}^*(C) &= \overline{y}_{\alpha^*(C),\beta^*(C)} \\ &= \frac{2(C+1)\overline{X}^2 - 2(C-1)\overline{x}^2 + (2C^2 - C - 1)\left(\overline{X} - \overline{x}\right)^2}{4\overline{X}\overline{x} - (2C^2 - C - 1)\left(\overline{X} - \overline{x}\right)^2}\,\overline{y}. \end{aligned} \tag{5.3}$$

At first it might seem surprising that this estimator $\overline{y}^*(C)$ is also defined in the region $0 < C \leq 1/2$. (The denominator vanishes if $C = \left(1 \pm \sqrt{9 + 32\overline{X}\overline{x}/(\overline{X} - \overline{x})^2}\right)/4$.) However, one can also let the parameters (α, β) in the definition of our proposed estimator $\overline{y}_{\alpha,\beta}$ in (1.4) be complex numbers—but such that we still get a real estimator. One can check that $\alpha^*(C)$ and $\beta^*(C)$ in (5.2) for $0 < C < 1/2$ have this property.

Furthermore, we can check that the first degree of approximation of the bias and MSE of $\overline{y}^*(C)$ are given by

$$\mathbb{B}_1(\overline{y}^*(C)) = 0, \qquad \text{MSE}_1(\overline{y}^*(C)) = \frac{1-f}{n} S_Y^2 \left(1-\rho^2\right) \tag{5.4}$$

(compare (2.9) and (3.6)). Thus, the estimator $\overline{y}^*(C)$ of (5.3) is an unbiased AOE.

One might also ask whether inside $0 < C \leq 1/2$ there is a choice of real parameters $(\alpha, \beta) \in \mathbb{R}^2$ such that we get an AOE with small bias. Using (3.4) in (2.9), we get the first-degree approximation of the bias of an AOE

$$\mathbb{B}_1\left(\overline{y}^{(o)}_{\alpha,\beta}\right) = \frac{1-f}{n} C_X^2 \overline{Y}\ \frac{1}{2}\left[C(1-2C) + (1-2\beta)^2\right]. \tag{5.5}$$

From this expression (and the constraint (3.4)) it is clear that the bias can only be made zero if $C \leq 0$ or $C \geq 1/2$. Otherwise, there is always a positive contribution coming from the term $C(1-2C)$ that does not vanish no matter what we choose for $\beta \in \mathbb{R}$. In fact, it looks as if the choice $\beta = 1/2$ always yields the least possible bias; however two remarks are in order here. Firstly, given (3.4) and unless $C = 0$, we can only let β be close to $1/2$ and choose α accordingly (the absolute value $|\alpha|$ is then large). Secondly, we already noted that $\overline{y}_{\alpha,1/2} = \overline{y}$, and the MSE for the sample mean estimator is $((1-f)/n)S_Y^2$, not $((1-f)/n)S_Y^2(1-\rho^2)$ as for an AOE. We have arrived here at a point where the first-degree approximation to bias and MSE breaks down. To find a choice of real parameters for given C with minimal MSE and least bias, higher degrees of approximation would have to be considered.

6. Application and Conclusion

Using data taken from the Department of the Interior, United States Geology Survey [9], site number 02290829501 (located in Florida), a comparison of our proposed estimator $\overline{y}_{\alpha,\beta}$ to the traditional estimators was carried out. The study variables (denoted by Y) are taken to be the maximum daily values (in feet) of groundwater at the site for the period from October 2009 to September 2010. The auxiliary variables (denoted X) are taken as the maximum daily values (in feet) of groundwater for the period from October 2008 to September 2009. Our goal is to estimate the true average maximum daily groundwater $\overline{Y}$ for the period from October 2009 to September 2010.

The questions we ask are as the follows. How many units of groundwater must be taken from the population Y to estimate the population mean $\overline{Y}$ within $d = 10\%$ at a 90% confidence level ($\alpha = 0.10$)? And how well do the estimators perform given this data set with auxiliary information for the calculated sample size n?

Using the entire data set, we calculate the following parameters: $\overline{Y} \approx 0.5832$, $\overline{X} \approx 0.6277$, $S_Y \approx 0.4480$, $S_X \approx 0.7222$, $\rho \approx 0.9125$, $C_Y \approx 0.7681$, $C_X \approx 1.1504$, and $C \approx 0.6092$. A scatterplot of the data set is shown in Figure 2(b), which adds emphasis to the positive measure of association between the study variable Y and the auxiliary variable X.

One should note that the value of $C \approx 0.6092$ lies in the interval $(1/2, 1)$, so we choose values of α and β from Section 4.2 (resp., from Section 5). Indeed, we use (5.2) and choose $\beta = \beta^*(0.6092) \approx 0.3176$ and $\alpha = \alpha^*(0.6092) \approx -0.3349$. Note that $\beta = 0.3176$ yields

$(-0.8704) < \alpha < 0.200573$ in Section 4.2. Using the notation of Section 5, we also note that $\overline{y}_{-0.3349,0.3176} = \overline{y}^*(0.6092)$.

6.1. Calculating the Sample Size n

To estimate the population mean amount of groundwater recorded for the state of Florida from October 2009 to September 2010, a sample of size n is drawn from the population of size $N = 365$ according to the simple random sampling without replacement, see [10]. A first approximation to this sample size needed is the (infinite population) value

$$n_0 = \frac{Z_{\alpha/2}^2 \sigma^2}{d^2}, \tag{6.1}$$

where d is the chosen margin of error from the estimate of $\overline{Y}$ and $Z_{\alpha/2}$ is a standard normal variable with tail probability of $\alpha/2$. Accounting for the finite population size N, we obtain the sample size

$$n = \frac{1}{1/n_0 + 1/N}. \tag{6.2}$$

In general, the true value of σ^2 is unknown but can be estimated using its consistent estimator s^2. However, in our case σ^2 is calculated from the population and is given as $S_Y^2 \approx 0.2006$. Therefore, with $\alpha = 0.10$, that is, $Z_{0.05} \approx 1.6449$, and $d = 10\%$ of $\overline{Y}$ (i.e., $d \approx 0.0583$), the sample size can be calculated as follows: we have $n_0 \approx ((1.6449)^2 \cdot 0.2006)/(0.0583)^2 \approx 159.59$ and rounding up gives $n_0 = 160$; so, we get $n \approx 1/(1/160 + 1/365) \approx 111.23$ and thus take $n = 112$.

6.2. Relative Efficiencies

Table 1 shows the relative efficiencies of the traditional estimators (sample mean $\overline{y}$, ratio $\overline{y}_R$ and product $\overline{y}_P$ estimators) and our proposed two-parameter ratio-product-ratio estimator $\overline{y}_{\alpha,\beta}$ for the parameters $(\alpha, \beta) = (-0.3349, 0.3176)$. We note that, with this choice of parameters, the estimator is an (unbiased) AOE, namely, $\overline{y}_{-0.3349,0.3176} = \overline{y}^*(0.6092)$. The table shows that our two-parameter ratio-product-ratio estimator dominates the traditional estimators in the sense that it has the highest efficiency.

We can also observe that, in the computation of the relative efficiency, the specification of the sample size n is not important since the finite population correction factor $((1-f)/n)$ is canceled out (however, this would not be the case for higher degrees of approximation).

6.3. Constructing a 90% Confidence Interval for $\overline{Y}$ Using $\overline{y}_{\alpha,\beta}$

Constructing a 90% confidence interval, the following formulation can be used (similar formulae hold for all estimators discussed here), see [10]:

$$\left(\overline{y}^*(0.6092) \pm Z_{0.05}\sqrt{\frac{S_Y^2}{n}} \cdot \sqrt{\frac{N-n}{N-1}}\right). \tag{6.3}$$

The factor $\sqrt{(N-n)/(N-1)}$ is the *finite population correction*.

Table 1: Relative efficiencies comparisons.

$\text{MSE}(\overline{y})/\text{MSE}(\overline{y})$	$\text{MSE}(\overline{y})/\text{MSE}_1(\overline{y}_R)$	$\text{MSE}(\overline{y})/\text{MSE}_1(\overline{y}_P)$	$\text{MSE}(\overline{y})/\text{MSE}_1(\overline{y}^*(0.6092))$
100%	196.11%	16.73%	597.28%

Table 2: Comparison of the estimators according to the absolute deviation from the population mean $\overline{Y}$ (in 10 000 simulations).

Deviation from population mean							Counts
$\lvert\overline{y}^*(0.6092)-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_R-\overline{Y}\rvert$	$<$	$\lvert\overline{y}-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_P-\overline{Y}\rvert$	3 173
$\lvert\overline{y}_R-\overline{Y}\rvert$	$<$	$\lvert\overline{y}^*(0.6092)-\overline{Y}\rvert$	$<$	$\lvert\overline{y}-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_P-\overline{Y}\rvert$	2 978
$\lvert\overline{y}^*(0.6092)-\overline{Y}\rvert$	$<$	$\lvert\overline{y}-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_R-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_P-\overline{Y}\rvert$	1 528
$\lvert\overline{y}-\overline{Y}\rvert$	$<$	$\lvert\overline{y}^*(0.6092)-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_R-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_P-\overline{Y}\rvert$	972
$\lvert\overline{y}_P-\overline{Y}\rvert$	$<$	$\lvert\overline{y}-\overline{Y}\rvert$	$<$	$\lvert\overline{y}^*(0.6092)-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_R-\overline{Y}\rvert$	766
$\lvert\overline{y}-\overline{Y}\rvert$	$<$	$\lvert\overline{y}^*(0.6092)-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_P-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_R-\overline{Y}\rvert$	308
$\lvert\overline{y}-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_P-\overline{Y}\rvert$	$<$	$\lvert\overline{y}^*(0.6092)-\overline{Y}\rvert$	$<$	$\lvert\overline{y}_R-\overline{Y}\rvert$	275

Of course, by the choice of the sample size $n = 112$, we get a margin of error given by approximately $0.1 \cdot \overline{Y} \approx 0.0583$; more precisely, the calculation using the above formula yields $(\overline{y}^*(0.6092) \pm 0.0580)$.

6.4. Comparison of Estimators

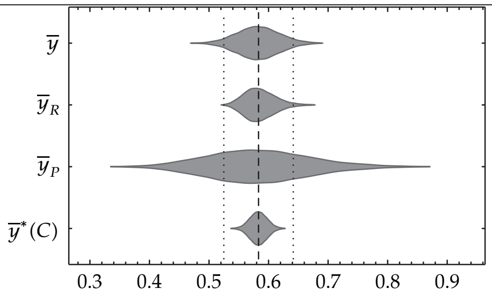
To compare the proposed estimator with the traditional ones, we selected 10 000 times a sample of size $n = 112$ and calculated the estimators from it. We note that there are $\binom{365}{112} \approx 2.5 \cdot 10^{96}$ possibilities to choose 112 data points out of a total 365 without replacement.

In Table 2 we show the relative position of the estimators with respect to the population mean $\overline{Y}$. In the 10 000 simulations, our proposed estimator outperformed the traditional estimators on 4 701 occasions. The ratio estimator, the suggested estimator for this value of C by [1], performs better than our proposed estimator 2 978 times (in these cases it is actually the best of the studied estimators; note that the ratio estimator is the worst 1 349 times).

In Table 3, we compare the estimators by looking at the following criteria. The *coverage probability* is the proportion of the 90% confidence interval covering the population mean $\overline{Y}$; as expected, the usual mean sample estimator yields around 90%, while the ratio estimator and our proposed estimator yield much higher values—in this simulation, all intervals calculated from our proposed estimator cover $\overline{Y}$. For those 90% confidence intervals that do not cover $\overline{Y}$, we check whether they lie to the left (*negative bias*) or to the right (*positive bias*) of $\overline{Y}$. We also state the statistical information *lower* and *upper quartile* and *median* that we get from the 10 000 simulations; we also show violin plots for the estimators (the dashed line indicates the value $\overline{Y}$; the dotted lines indicate the 90% confidence interval) to get a visual confirmation of the numbers just presented. In the violin plot, we see that the values obtained by our proposed estimators yield a narrow normal distribution around the true value (skewness is 0.0046; kurtosis is 2.9926), while the product estimator gives a spread-out distribution and

Table 3: Comparison of the estimators in 10 000 simulations. See text for details.

Estimator	Coverage	Neg. bias	Pos. bias	Lo. quart.	Median	Up. quart.
$\overline{y}$	89.86%	5.14%	5.00%	0.5583	0.5820	0.6062
$\overline{y}_R$	97.01%	0.12%	2.87%	0.5670	0.5831	0.6017
$\overline{y}_P$	49.59%	25.44%	24.97%	0.5240	0.5809	0.6410
$\overline{y}^*(0.6092)$	100.00%	0.00%	0.00%	0.5732	0.5829	0.5928



Estimator	MSE	MSE($\overline{y}$)/MSE(est)
$\overline{y}$	0.00127	100%
$\overline{y}_R$	0.00068	186.68%
$\overline{y}_P$	0.00755	16.79%
$\overline{y}^*(0.6092)$	0.00021	598.93%

the (traditionally preferred) ratio estimator gives a skewed distribution (skewness is 0.5230; kurtosis is 3.4253). Finally, we compare the values of the MSEs; the experimental values obtained agree with the theoretical values listed in Table 1.

We infer that our proposed estimator is more efficient and robust than the traditional sample mean, ratio, and product estimators.

6.5. Outlook

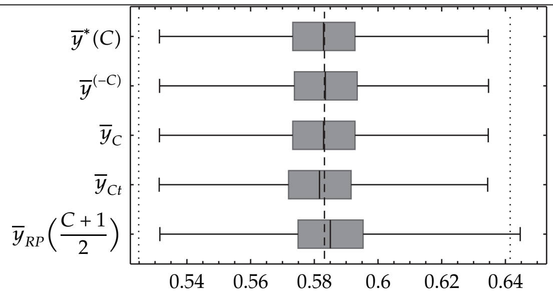
Several authors have proposed efficient estimators using auxiliary information. For example, Srivastava [11] and Reddy [12] consider a generalisation to the product and ratio estimator given by $\overline{y}^{(k)} = \overline{y}(\overline{x}/\overline{X})^k$; Reddy [12] also introduces the estimator $\overline{y}_k = \overline{y}\ \overline{X}/(\overline{X} + k\ (\overline{x} - \overline{X}))$; in Sahai and Ray [2] the estimator $\overline{y}_{kt} = \overline{y}\ (2 - (\overline{x}/\overline{X})^k)$ (where "t" stands for "transformed") is considered; Singh and Espejo [6] introduce a certain class of "ratio-product" estimators having the form $\overline{y}_{RP}(k) = \overline{y}(k \cdot (\overline{X}/\overline{x}) + (1-k) \cdot (\overline{x}/\overline{X}))$. Choosing appropriate parameters k for these estimators and calculating the first-degree approximation of the MSE, one can show that

$$\begin{aligned} \mathrm{MSE}_1\left(\overline{y}^{(-C)}\right) &= \mathrm{MSE}_1\left(\overline{y}_C\right) = \mathrm{MSE}_1\left(\overline{y}_{Ct}\right) = \mathrm{MSE}_1\left(\overline{y}_{RP}\left(\frac{C+1}{2}\right)\right) \\ &= \frac{1-f}{n} S_Y^2\left(1-\rho^2\right). \end{aligned} \tag{6.4}$$

Thus, these estimators and our proposed estimator (see (3.6)) are equally efficient up to the first degree of approximation, having the minimal possible MSE for this type of estimators [3, 4] (i.e., estimators which are given by a product of $\overline{y}$ and a function of $\overline{x}/\overline{X}$). Indeed, all

Table 4: Comparison of AOEs in 10 000 simulations. See text for details.

Estimator	Coverage	Neg. bias	Pos. bias	Lo. quart.	Median	Up. quart.
$\overline{y}^*(0.6092)$	100%	0%	0%	0.5732	0.5829	0.5928
$\overline{y}^{(-0.6092)}$	100%	0%	0%	0.5738	0.5835	0.5935
$\overline{y}_{0.6092}$	100%	0%	0%	0.5732	0.5829	0.5928
$\overline{y}_{0.6092t}$	100%	0%	0%	0.5719	0.5816	0.5915
$\overline{y}_{RP}(0.8046)$	99.99%	0%	0.01%	0.5749	0.5850	0.5953



Estimator	MSE	MSE($\overline{y}$)/MSE(est)
$\overline{y}^*(0.6092)$	0.000212	598.93%
$\overline{y}^{(-0.6092)}$	0.000214	593.91%
$\overline{y}_{0.6092}$	0.000212	598.92%
$\overline{y}_{0.6092t}$	0.000215	590.51%
$\overline{y}_{RP}(0.8046)$	0.000228	555.09%

these estimators give similar results as our proposed estimator in the above application, see Table 4. Comparing the first degree of approximation of the bias (doing calculations as in Section 2) reveals why our unbiased AOE $\overline{y}^*(C)$ and Reddy's $\overline{y}_C$ behave similarly—they are both unbiased AOEs:

$$\begin{gathered}
\mathbb{B}(\overline{y}) = \mathbb{B}_1(\overline{y}^*(C)) = \mathbb{B}_1(\overline{y}_C) = 0, \qquad \mathbb{B}_1(\overline{y}_R) = \frac{1-f}{n}(1-C)C_X^2\overline{y}, \\
\mathbb{B}_1(\overline{y}_P) = \frac{1-f}{n}CC_X^2\overline{y}, \qquad \mathbb{B}_1\left(\overline{y}^{(-C)}\right) = \frac{1-f}{n}\frac{C(1-C)}{2}C_X^2\overline{y}, \qquad (6.5) \\
\mathbb{B}_1(\overline{y}_{Ct}) = \frac{1-f}{n}\frac{C(1-3C)}{2}C_X^2\overline{y}, \qquad \mathbb{B}_1\left(\overline{y}_{RP}\left(\frac{C+1}{2}\right)\right) = \frac{1-f}{n}\frac{(1+2C)(1-C)}{2}C_X^2\overline{y}.
\end{gathered}$$

(With $C = 0.6092$, only $\overline{y}_{Ct}$ is negatively biased, compare the quartiles and the box plot in Table 4).

For our proposed estimator $\overline{y}_{\alpha,\beta}$ in (1.4) (which contains the three traditional estimators, namely, sample mean, product, and ratio estimators), we are able to use the two parameters α and β to obtain an estimator $\overline{y}^*(C)$ in (5.3) that is up to first degree of approximation both unbiased and has minimal MSE. While the idea behind creating $\overline{y}_{\alpha,\beta}$ is simple, the form of the unbiased AOE $\overline{y}^*(C)$ derived from it is not—and the above list shows that there are many AOEs, but they are not necessarily unbiased.

A thorough comparison of estimators using auxiliary information (e.g., the one in (1.4) and the ones mentioned above) involving higher degrees of approximation of MSE and bias as well as accompanying simulations might be desirable, for example, to find the estimator that behaves well if the parameter C is unknown in advance (in which case it may be replace with its consistent estimate, $\widehat{C} = r \cdot (\widehat{C}_Y / \widehat{C}_X)$, where r is the *sample Pearson correlation coefficient* and $\widehat{C}_Y$ and $\widehat{C}_X$ are the estimates of the coefficients of variation of Y and X, resp.). (Recall that our analysis in Section 5 shows that the first-degree approximation to MSE and bias for values of the parameter C close to zero breaks down.)

References

[1] M. N. Murthy, "Product method of estimation," *Sankhyā A*, vol. 26, pp. 69–74, 1964.

[2] A. Sahai and S. K. Ray, "An efficient estimator using auxiliary information," *Metrika*, vol. 27, no. 4, pp. 271–275, 1980.

[3] S. K. Srivastava, "A generalized estimator for the mean of a finite population using multi-auxiliary information," *Journal of the American Statistical Association*, vol. 66, no. 334, pp. 404–407, 1971.

[4] S. K. Srivastava, "A class of estimators using auxiliary information in sample surveys," *The Canadian Journal of Statistics*, vol. 8, no. 2, pp. 253–254, 1980.

[5] S. K. Ray and A. Sahai, "Efficient families of ratio and product-type estimators," *Biometrika*, vol. 67, no. 1, pp. 211–215, 1980.

[6] H. P. Singh and M. R. Espejo, "On linear regression and ratio-product estimation of a finite population mean," *Journal of the Royal Statistical Society D*, vol. 52, no. 1, pp. 59–67, 2003.

[7] A. Sahai and A. Sahai, "On efficient use of auxiliary information," *Journal of Statistical Planning and Inference*, vol. 12, no. 2, pp. 203–212, 1985.

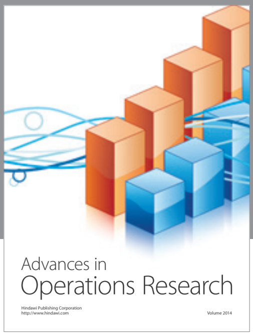
[8] S. K. Srivastava and H. S. Jhajj, "A class of estimators of the population mean in survey sampling using auxiliary information," *Biometrika*, vol. 68, no. 1, pp. 341–343, 1981.

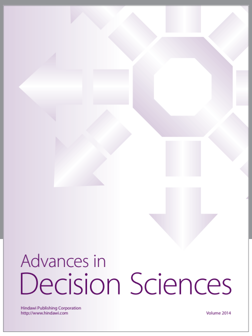
[9] United States Geology Survey, Water resources of the United States—annual water data report, 2011, http://wdr.water.usgs.gov/.

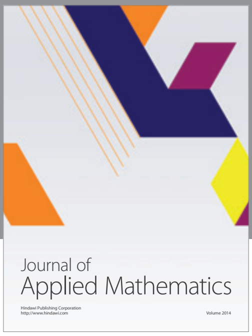
[10] W. G. Cochran, *Sampling Techniques*, John Wiley & Sons, New York, NY, USA, 3rd edition, 1977.

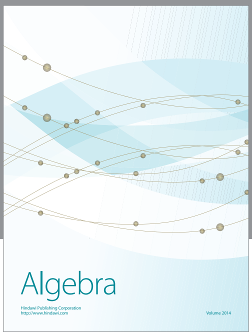
[11] S. K. Srivastava, "An estimator using auxiliary information in sample surveys," *Calcutta Statistical Association Bulletin*, vol. 16, no. 62-63, pp. 121–132, 1967.

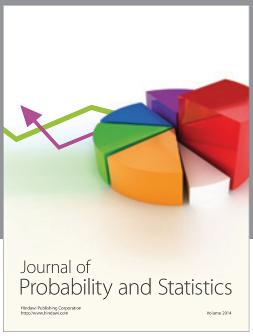
[12] V. N. Reddy, "On ratio and product methods of estimation," *Sankhyā B*, vol. 35, no. 3, pp. 307–316, 1973.


Advances in
Operations Research

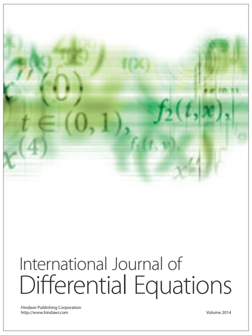

Advances in
Decision Sciences

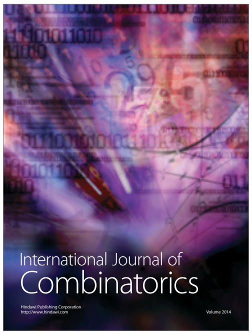

Journal of
Applied Mathematics

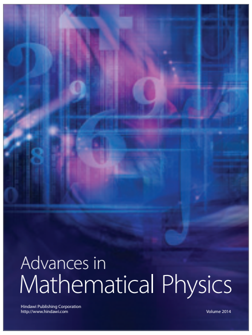

Algebra

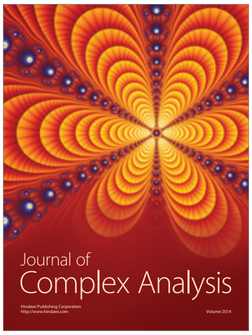

Journal of
Probability and Statistics

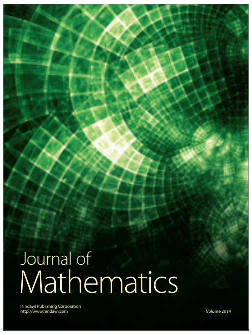
The Scientific
World Journal

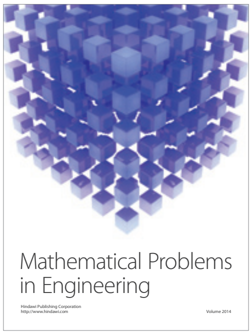

International Journal of
Differential Equations

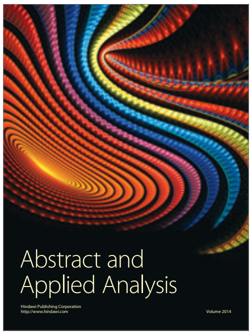

International Journal of
Combinatorics

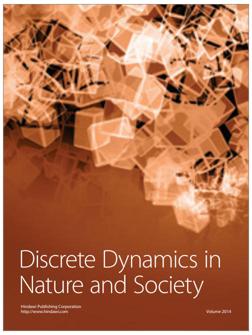

Advances in
Mathematical Physics

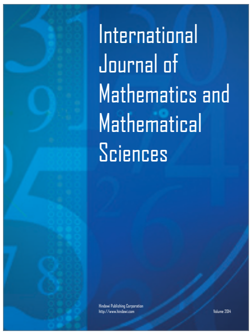

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

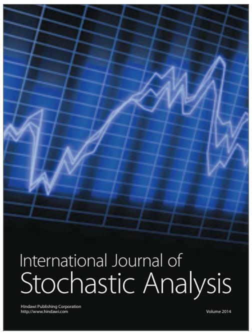

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization