Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(1) 2012 65-74

County Level Estimation using Data from the U.S. National Resources Inventory

Pushpal K. Mukhopadhyay[1*], Tapabrata Maiti[2] and Wayne A. Fuller[3]

[1]*Research Statistician, SAS Institute Inc., 600 Research Drive, Cary, NC 27513*

[2]*Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824*

[3]*Department of Statistics and Department of Economics, Iowa State University, Ames, IA 50011*

Received 30 April 2011; Revised 12 September 2011; Accepted 12 September 2011

SUMMARY

We use a transformed Fay-Herriot model to estimate wind erosion at the county level in Iowa. A soil erodibility index is available from administrative records for each county in Iowa and is used to form the predictor. The response variable is the soil loss due to wind as recorded in the 2002 National Resources Inventory. We propose bias corrected, and calibrated small area predictors such that the weighted sum of the county predictors matches the state level direct estimate. The standard errors are estimated by using a parametric double-bootstrap method. The small area predictions have estimated coefficients of variation that average about three fifth of those of the direct estimates.

Keywords : Small area estimation, Wind erosion, Calibration, Double-Bootstrap, Soil erodibility index.

1. INTRODUCTION

Wind erosion is a severe problem in some mid-western states in the US and the resulting soil loss can lead to decreases in soil productivity. Although some work has been done by the Wind Erosion Research Unit (WERU) to estimate wind erosion at the national level, there has been little work to estimate wind erosion at the county level. We propose estimating wind erosion at the county level using National Resources Inventory (NRI) data. A transformed Fay-Herriot model is used to predict the county means. We include the design weight in our proposed model in such a way that the final estimates are calibrated with the state direct estimate. A parametric double bootstrap method is used to estimate the mean squared error (MSE) of the proposed estimator. The proposed model fits the data well and produces a mean squared error of prediction that is approximately one half of the design standard error for most counties.

Section 2 gives an introduction to the NRI survey. The problem of wind erosion and some previous results are discussed in Section 3. In Section 4, we present results from an exploratory study and in Section 5 calibration for estimated county means is discussed. Section 6 discusses the estimation of the MSE for the proposed estimators and Section 7 summarizes the results.

2. THE NRI SURVEY

The NRI is a longitudinal survey conducted by the US Department of Agriculture's (USDA) Natural Resources Conservation Service (NRCS) in cooperation with the Center for Survey Statistics and Methodology (CSSM) at Iowa State University. The survey is designed to assess conditions and trends for land cover, soil, water, and related natural resources on non-federal lands in the US. The NRI was conducted every 5 years during 1982-1997. The basic design of NRI surveys is

[*] *Corresponding author* : Pushpal K. Mukhopadhyay
E-mail address : pushpal.mukhopadhyay@sas.com

a stratified, two-stage area sample. The land area of most states in the US is divided according to the Public Land Survey (PLS) system and the PLS provides a convenient structure for NRI sample selection and for locating primary sampling units (PSUs) in the field. The primary sampling units are also called segments. A typical PSU in the midwest is a square, one half mile on a side. PSU sizes vary and are somewhat smaller in the East where the PLS does not apply. Three sample points are selected within each PSU according to a restricted randomization procedure, see Nusser and Goebel (1997). The 1997 NRI contains approximately 300,000 PSUs and over 800,000 sample points. Sampling rates generally range from 2% to 6% of the land area, though rates occasionally fall outside this 6% range. The sampling rate within a county is increased when larger sample sizes are needed for special studies or when heterogeneous patterns exist for soil types, land uses, major land resource areas, or hydrologic regions (Nusser and Goebel 1997). Since 2000, the original structure of the NRI has been replaced by a two-phase supplemented panel sampling design in which the 1997 NRI segments serve as a first phase and each year a partially overlapping panel is selected through a stratified sampling design as a second phase. The second phase sample includes approximately 42,000 "core" segments that are observed every year. An additional 30,000 segments are selected from the remaining 268,000 PSUs each year to form a supplemental sample. All points in every selected segment are part of the annual sample (Fuller 2003). Data are collected at two levels. Urban land, water, and roads are collected at the PSU level whereas soil properties, land use, and land cover are collected at the point level.

3. VARIABLE OF INTEREST

Wind erosion is a serious problem in many parts of the world. Areas susceptible to wind erosion include parts of North Africa; the Near East; Asia; the Siberian Plains; Australia; China; South America; and North America (Wind Erosion Research Unit (WERU)), http://www.weru.ksu.edu/new_weru/problem/problem.shtml).

An extensive dry spell during the 1930's produced dust storms and catastrophic soil damage. Still today wind erosion can severely damage agricultural land throughout the Great Plains (WERU). Wind erosion is responsible for about 40 percent of the total soil loss in the US (Hagen 1994), and the percent can increase in drought years (Hagen and Woodruff 1973). On approximately 74 million acres of land in the US, wind erosion is a dominant problem. The 1992 National Resources Inventory (NRI) survey estimated 2.2 tons per year soil loss due to wind erosion on non-federal rural lands in the US.

Although several studies address the issue of estimating soil loss at the national level, little effort has been made to provide estimates of soil loss at a lower level (e.g., county). Our objective is to produce estimates of wind erosion for counties in Iowa that are susceptible to wind erosion. There are 44 counties in Iowa where wind erosion is important and where wind erosion data are collected. Data were used from the 2002 NRI survey for those 44 counties. It is not practical to measure wind erosion directly for large areas. Therefore the wind erosion equation (WEQ) is used by several national agencies, including the NRCS, to estimate soil loss due to wind. Wind erosion is calculated as a function of several factors, including soil erodibility, climate, slope, conservation practices, and land cover (Woodruff and Siddoway 1965). We use WEQ02 to denote the wind erosion calculated for the year 2002. The measure of soil erodibility used in the WEQ is called the soil erodibility index and is denoted by IFact. A higher IFact value indicates greater susceptibility to wind erosion. IFact can be obtained from the NRCS soil survey database available through the NRCS Soil Data Mart (SDM) for each county in the US. Since IFact is a soil characteristic, it changes little over time. For this study, we use IFact from the soil science database as the predictor variable, and use WEQ02 as the response variable.

4. EXPLORATORY ANALYSIS

Let y_i be the survey weighted mean WEQ02 for county i, and let x_i be the mean IFact for the same county, where $i = 1, 2, ..., m$. We first considered an area level small area model in the original units.

The model in the original units (Model I) is

$$y_i = \beta_0 + \beta_1 x_i + u_i + e_i, \qquad (1)$$

where β_0 and β_1 are unknown parameters, e_i is the sampling error and u_i is the area specific random effect. We estimate the parameters under the assumptions

$e_i \overset{ind}{\sim} N(0, D_i)$, $u_i \overset{ind}{\sim} N(0, \sigma_u^2)$; the e_i and u_i are independent; and the D_i are known. An overall PSU variance $\left(\sigma_e^2\right)$ for the state was estimated as the pooled within county variance for the 44 counties. The design variance of the mean for county i with a sample size n_i is σ_e^2 / n_i. The empirical generalized least squares (EGLS) estimator of $\beta = (\beta_0, \beta_1)^T$ is

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} \mathbf{y}$$

$$= \left[\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T \left(\sigma_u^2 + D_i\right)^{-1}\right]^{-1} \left[\sum_{i=1}^{m} \mathbf{x}_i y_i \left(\sigma_u^2 + D_i\right)^{-1}\right] \quad (2)$$

where $\mathbf{y} = (y_1, \ldots, y_m)^T$, $\mathbf{x}_i = (1, x_i)^T$, $X = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_m^T)^T$, $V = \text{diag}\ \{\sigma_u^2 + D_i\}_{i=1}^m$, $\hat{V} = \text{diag}\ \{\hat{\sigma}_u^2 + D_i\}_{i=1}^m$ and $\hat{\sigma}_u^2$ is an estimator of σ_u^2. For a set of scalars $\{a_1, a_2, \ldots, a_m\}$, $\text{diag}\{a_i\}_{i=1}^m$ denotes a diagonal matrix with elements $a_1, a_2, \ldots, a_m$ and $\mathbf{A}^T$ denotes the transpose of a matrix $\mathbf{A}$.

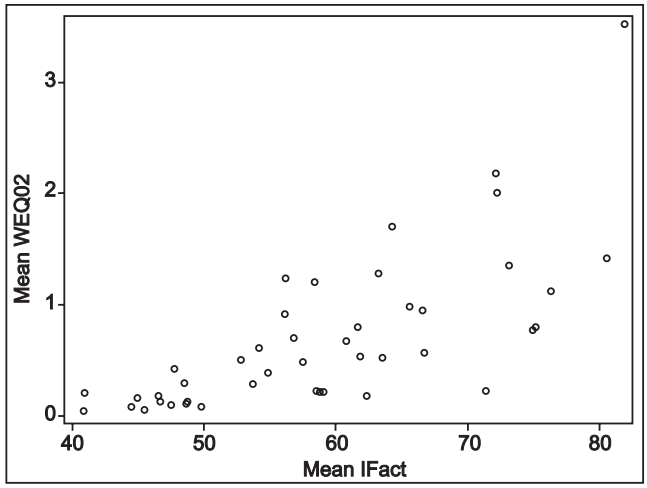

Fig. 1. Scatter plot of direct estimates for the mean WEQ02 and the mean IFact

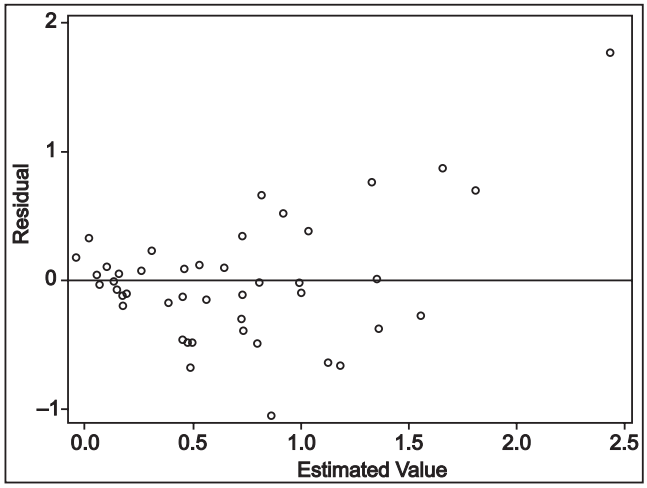

Fig. 2. Residuals and estimated values for the model in original scale

Fig. 1 is a scatter plot of the mean WEQ02 and the mean IFact, and Fig. 2 is the plot of the residuals against predicted values obtained by regressing mean WEQ02 on mean IFact. The residual for county i is $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the EGLS estimators given in equation (2). Nonlinearity and unequal residual variances are clear from the plot which leads us to consider a transformation of y_i. Several transformations were considered, and a cube root transformation of the mean WEQ02 fit the data reasonably well.

The cube root transformed model (Model II) is

$$(y_i)^{1/3} = \beta_0 + \beta_1 x_i + u_i + e_i^*, \quad (3)$$

where $u_i \overset{ind}{\sim} N(0, \sigma_u^2)$, $e_i^* \overset{ind}{\sim} N(0, D_i^*)$, u_i and e_i are independent, and D_i^* are treated as known. In the remaining analysis the cube root transformation is treated as known. The $D_i^* = (9 y_i^{4/3})^{-1} D_i$, where D_i is the estimated variance of y_i.

The regression parameters of model (3) were estimated using the EGLS (2) and the between area variance parameter was estimated using the REML. The GLIMMIX procedure in SAS is used to estimate the parameters of the model. Parameter estimates and their standard errors are given in Table 1. Both regression parameters are significant at 0.05 level. The estimate of σ_u^2 is about 1.5 of the standard error. The likelihood ratio statistic for testing $\sigma_u^2 = 0$ is $\chi_1^2 = 3.59$ with a p-value of 0.06. The average sample size is 15.68 and the variance of mean erosion is 0.0204 for a sample of size 16. Thus the estimated σ_u^2 is about 0.43 of the variance for an average sized sample.

Table 1. Parameter estimates and their standard errors for the cube root transformed small area model.

	β_0	$10\beta_1$	$10\sigma_u^2$
Estimates	0.804	0.199	0.088
Standard Errors	0.026	0.025	0.062

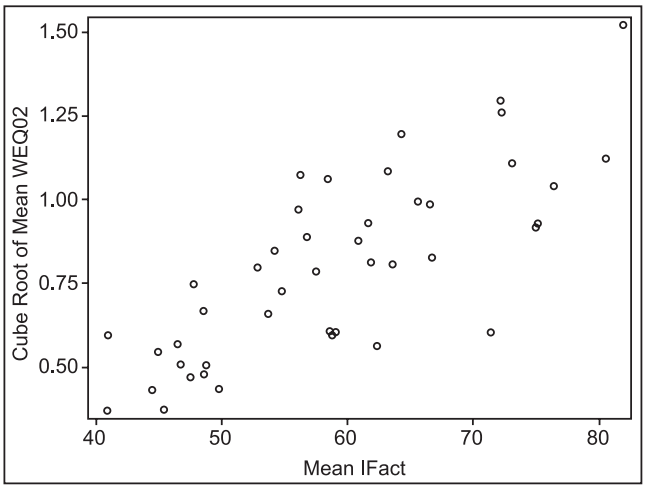

Fig. 3. Scatter plot of the cube root transformed mean WEQ02 and mean IFact

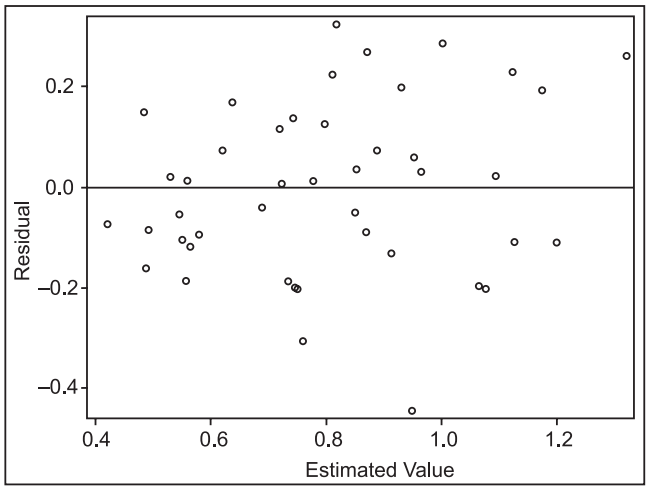

Fig. 4. Residuals and estimated values for the transformed model

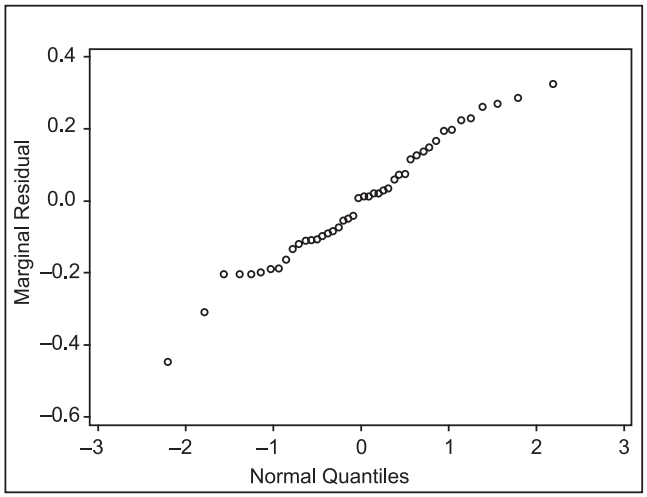

Fig. 5. Normal quantiles plot of residuals for the transformed model

Fig. 3 shows a scatter plot of the cube root of the mean WEQ02 and the mean IFact. The residual plot for the fitted cube root transformed model (3) in Fig. 4 supports the linear model. The normal quantiles plot for the residuals from the transformed model is given in Fig. 5. A test of normality for the residuals was performed by using the Kolmogorov-Smirnov test (Conover 1999, Section 6.2). The Kolmogorov-Smirnov D statistic is 0.08 with a *p*-value greater than 0.15. Overall, the transformed model provides a reasonable fit to the data.

The empirical best linear unbiased predictor (EBLUP) of $\mu_i = \mathbf{x}_i^T \beta + u_i$ for a linear model is given by

$$\hat{\mu}_i = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i)\mathbf{x}_i^T \hat{\beta} \quad (4)$$

$$= \mathbf{x}_i^T \hat{\beta} + \hat{\gamma}(y_i - \mathbf{x}_i^T \hat{\beta}), \quad (5)$$

where

$$\hat{\gamma}_i = (\hat{\sigma}_u^2 + D_i)^{-1}\hat{\sigma}_u^2, \quad (6)$$

Prasad and Rao (1990) expressed the MSE of the EBLUP (5) as

$$E[\hat{\mu}_i - \mu_i]^2 = g_{1i}(\sigma_u^2) + g_{2i}(\sigma_u^2) + g_{3i}(\sigma_u^2) \quad (7)$$

where

$$g_{1i}(\sigma_u^2) = (\sigma_u^2 + D_i)^{-1}\sigma_u^2 D_i \quad (8)$$

is the MSE when β and σ_u^2 are known,

$$g_{2i}(\sigma_u^2) = (1-\gamma_i)^2 \mathbf{x}_i^T \left[\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^T (\sigma_u^2 + D_i)^{-1}\right]^{-1} \mathbf{x}_i \quad (9)$$

is the effect of estimating $\hat{\beta}$, and

$$g_{3i}(\sigma_u^2) = D_i^2 (\sigma_u^2 + D_i)^{-3} \overline{V}(\hat{\sigma}^2) \quad (10)$$

is the effect of estimating σ_u^2, and $\overline{V}(\hat{\sigma}^2)$ is the asymptotic variance of $\hat{\sigma}_u^2$. The asymptotic variance of the residual maximum likelihood (REML) estimator of σ_u^2 is (Rao 2003)

$$\overline{V}(\hat{\sigma}^2) = 2\left[\sum_{i=1}^{m} (\sigma_u^2 + D_i)^{-2}\right]^{-1} \quad (11)$$

The MSE in (7) can be estimated by

$$\text{mse}[\hat{\mu}_i] = g_{1i}(\hat{\sigma}_u^2) + g_{2i}(\hat{\sigma}_u^2) + 2g_{3i}(\hat{\sigma}_u^2) \qquad (12)$$

where $\hat{\sigma}_u^2$ is the REML estimator of σ_u^2 (Prasad and Rao 1990).

Because a transformation of y_i is used, the standard small area predictions (5) and the estimation of MSE (12) need to be adjusted. For the transformation, the sampling model can be written as

$$y_i^{1/3} = \mu_i^* + e_i^*, \qquad (13)$$

where μ_i^* are the unobserved small area means in the transformed scale, $e_i^* \overset{ind}{\sim} N(0, D_i^*)$, and $i = 1, 2, \ldots, m$. The linking model can be written as

$$\mu_i^* = \beta_0 + \beta_1 x_i + u_i, \qquad (14)$$

where $u_i \overset{ind}{\sim} N(0, \sigma_u^2)$, and u_i and e_i are independent. Therefore, the small area parameter of interest in the original scale is defined as $\mu_i = (\mu_i^*)^3 = (\beta_0 + \beta_1 x_i + u_i)^3$. Although the design expectation of y_i is the same as the small area parameter of interest for the untransformed model, this is no longer valid for the transformed model. See Rao (2003), section 10.4 for details. Therefore a bias corrected small area predictor is necessary. Slud and Maiti (2006) proposed a bias corrected small area predictor for a small area parameter that is a smooth monotone function, h, of μ_i^*. The approximately bias corrected estimator is

$$\hat{\mu}_i = h(\hat{\theta}_i)\hat{E}[h(\theta_i)] / \hat{E}[h(\hat{\theta}_i)] \qquad (15)$$

where $\hat{E}(\cdot)$ is an estimator of $E(\cdot)$ that is obtained by substituting $(\hat{\beta}, \hat{\sigma}_u^2)$ for (β, σ_u^2), and $E(\cdot)$ is an asymptotically correct expression for the expectation. Following Slud and Maiti (2006), we propose,

$$\hat{\mu}_{i,BE} = \left\{\hat{\gamma}_i (y_i)^{1/3} + (1-\hat{\gamma}_i) x_i^T \hat{\beta}\right\}^3 \left\{3\hat{\sigma}_u^2 \hat{\gamma}_i + \left(x_i^T \hat{\beta}\right)^2\right\}^{-1} \left\{3\hat{\sigma}_u^2 + \left(x_i^T \hat{\beta}\right)^2\right\} \qquad (16)$$

for the cube root transformed model, where $\hat{\beta}$ is the EGLS estimate of β, $\hat{\gamma}_i = \left(\hat{\sigma}_u^2 + D_i^*\right)^{-1} \hat{\sigma}_u^2$, and $\hat{\sigma}_u^2$ is the REML estimator of σ_u^2.

5. A REGRESSION BASED CALIBRATED SMALL AREA ESTIMATOR

If a domain with acceptable direct estimates is divided into a number of small areas, then it is often desirable that the weighted sum of the small area predicted means be the same as the direct domain mean. For the NRI survey, it is desired that the design weighted predicted county means equal the state direct mean because the state estimate has been released. Small area estimates such that the weighted sum of the estimated county means equals the direct estimate of the state mean are said to be calibrated. The direct estimate of the state mean (SDE) is defined as SDE = $\sum_{i=1}^m (w_i y_i) / \sum_{i=1}^m (w_i)$ and the state level model based estimate (SME) is defined as SME = $\sum_{i=1}^m (w_i \hat{\mu}_i) \Big/ \sum_{i=1}^m (w_i)$, where the w_i is the survey weight associated with county i, y_i is the direct county mean, and $\hat{\mu}_i$ is a model-based small area prediction. We define the relative absolute calibration (RAC) error by RAC = |SME – SDE| / SDE. The RACs for the untransformed and transformed models are 1.2% and 7% respectively.

To see why calibration at the state level is not necessarily achieved through the proposed model based estimators, consider model I with σ_u^2 known. The normal equations for estimating $\beta = (\beta_0, \beta_1)^T$ for model (1) are given by

$$X^T V^{-1} X \tilde{\beta} = X^T V^{-1} \mathbf{y} \qquad (17)$$

where X is the design matrix for the model I, $V = \text{diag}\left\{(\sigma_u^2 + D_i)^{-1}\right\}_{i=1}^m$, and $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$. One of the normal equations for β can be written as

$$\sum_{i=1}^m \frac{y_i - x_i^T \tilde{\beta}}{\sigma_u^2 + D_i} = 0. \qquad (18)$$

But for $\hat{\mu}_i$ from (5), $\sum_{i=1}^m w_i \hat{\mu}_i$ can be written as,

$$\sum_{i=1}^m w_i \hat{\mu}_i = \sum_{i=1}^m w_i y_i - \sum_{i=1}^m w_i \frac{D_i}{\sigma_u^2 + D_i} (y_i - x_i^T \tilde{\beta}) \qquad (19)$$

If $w_i \propto D_i^{-1}$ the second term in (19) is zero by (18) and the calibration condition is achieved. Several approaches have been taken to obtain a calibrated

estimator. For example see Fuller (2009). Following Wang and Fuller (2002), we include an additional variable in the small area model. The proposed estimator will have the properties of EBLUP under the model and will be fully calibrated. For the untransformed model, we include $x_{2i} = D_i w_i$ as a covariate.

Model I(b):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_{2i} + u_i + e_i, \qquad (20)$$

where β_2 is a fixed parameter and β_0, β_1, u_i and e_i are defined in model I. As before, we assume $u_i \overset{ind}{\sim} \mathrm{N}(0, \sigma_u^2)$, $e_i \overset{ind}{\sim} \mathrm{N}(0, D_i)$ and u_i and e_i are mutually independent. Small area means, $\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, are estimated by

$$\hat{\mu}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)(\hat{\beta}_0 + \hat{\beta}_i x_{1i} + \hat{\beta}_2 x_{2i}), \qquad (21)$$

where $\hat{\beta}$ and $\hat{\gamma}_i$ are given in (2) and (6) respectively. From one of the normal equations for β when σ_u^2 is known,

$$\sum_{i=1}^{m} \frac{1}{\sigma_u^2 + D_i} \{D_i w_i y_i - D_i w_i \tilde{\beta}_0 - D_i w_i x_i \tilde{\beta}_1 - (D_i w_i)^2 \tilde{\beta}_2\} = 0 \qquad (22)$$

But using the predicted mean from model I(b),

$$\sum_{i=1}^{m} w_i \hat{\mu}_i = \sum_{i=1}^{m} w_i y_i - \sum_{i=1}^{m} \frac{1}{\sigma_u^2 + D_i} \{D_i w_i y_i - D_i w_i \tilde{\beta}_0 - D_i w_i x_i \tilde{\beta}_1 - (D_i w_i)^2 \tilde{\beta}_2\}. \qquad (23)$$

The second term on the right side of the equation in (23) vanishes by (22). Therefore we achieved calibration by including x_{2i} in model I. We can test the significance of calibration by conducting a significance test on β_2. If model I is correct then by including x_{2i} we have included an unnecessary variable. Note that when σ_u^2 is unknown, we replace σ_u^2 in equation (18), (19), (22), and (23) with a suitable estimator, $\hat{\sigma}_u^2$, and the overall calibration condition remains valid.

Proposition (5.1) gives a method for obtaining calibrated predictors that is appropriate for models where a function of y_i is linearly related to x_i.

Proposition 5.1. Let the small area model be

$$y_i = h(z_i),$$
$$z_i = \beta_0 + \beta_1 x_i + u_i + e_i, \qquad (24)$$

where β_0 and β_1 are regression parameters, e_i are sampling errors, u_i are area specific random effects, $h(\cdot)$ is a smooth function and $i = 1, 2,...,m$ indexes the m small areas. Assume $e_i \overset{ind}{\sim} \mathrm{N}(0, D_i)$, $u_i \overset{ind}{\sim} \mathrm{N}(0, \sigma_u^2)$, e_i and u_i are independent, the D_i are known, $\gamma_i = \left(\sigma_u^2 + D_i\right)^{-1} \sigma_u^2$, $\hat{\gamma}_i = \left(\hat{\sigma}_u^2 + D_i\right)^{-1} \hat{\sigma}_u^2, \hat{\sigma}_u^2$ is an estimator of σ_u^2, and w_i are the survey weights. Let, $\hat{z}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$,

$$\hat{\mu}_{p,i} = h[\hat{z}_i + \hat{\gamma}_i (z_i - \hat{z}_i)], \qquad (25)$$

$$\hat{\alpha} = \left(\sum_{i=1}^{m} \xi_i^2 v_i^{-1} \right)^{-1} \sum_{i=1}^{m} \xi_i v_i^{-1} (y_i - \hat{\mu}_{p,i}) \qquad (26)$$

and

$$\xi_i = w_i v_i, \qquad (27)$$

where $(\hat{\beta}_0, \hat{\beta}_1)$ is the EGLS estimator of (β_0, β_1), and v_i are constants. Then

$$\hat{\mu}_i = \hat{\mu}_{p,i} + \hat{\alpha} \xi_i \qquad (28)$$

is a fully calibrated estimator of μ_i in that $\sum_{i=1}^{m} w_i y_i = \sum_{i=1}^{m} w_i \hat{\mu}_i$.

Proof of Proposition 5.1. Note that by equation (26) for any v_i,

$$\sum_{i=1}^{m} w_i (y_i - \hat{\mu}_{p,i} - \hat{\alpha} w_i v_i) = 0, \qquad (29)$$

but from equation (28),

$$\sum_{i=1}^{m} w_i y_i - w_i \hat{\mu}_i = \sum_{i=1}^{m} w_i (y_i - \hat{\mu}_{p,i} - \hat{\alpha} \xi_i) \qquad (30)$$

$$= \sum_{i=1}^{m} w_i (y_i - \hat{\mu}_{p,i} - \hat{\alpha} w_i v_i) \qquad (31)$$

Therefore for any choice of v_i estimator (28) is fully calibrated.

We apply Proposition 5.1 to Model II to construct calibrated estimators of the small area means. We

regress the adjusted residuals $\tilde{q}_i = (y_i - \hat{z}_i^3)$ on $w_i(1-\hat{\gamma}_i)$ with $v_i = (1-\hat{\gamma}_i)$ to obtain $\hat{\alpha} = 3.14 \times 10^{-5}$. The calibrated predictor is

$$\hat{\mu}_{i,CE} = \left\{\hat{z}_i + \hat{\gamma}_i\left(z_i - \hat{z}_i\right)\right\}^3 + (3.14 \times 10^{-5}) w_i\left(1-\hat{\gamma}_i\right), \quad (32)$$

where the least squares standard error of $\hat{\alpha}$ is 3.32×10^{-5}.

Fig. 6 shows a plot of design unbiased county means and predicted means from the small area models.

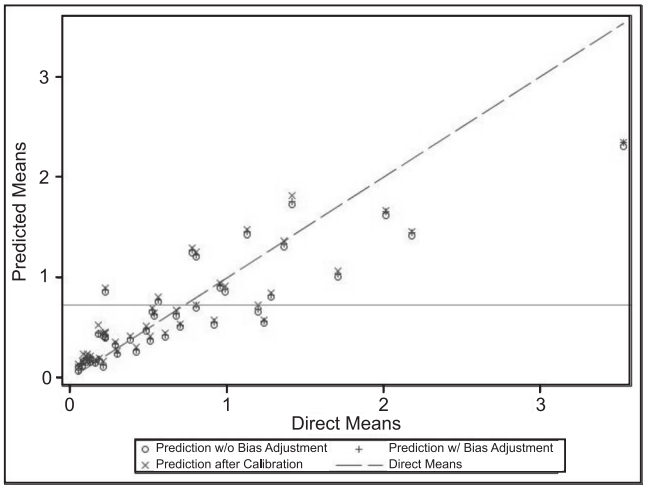

Fig. 6. Predicted small area means for the transformed estimator, bias corrected estimator, and calibrated estimator

Predicted values from the transformed and bias corrected estimators are very close for all counties. All calibrated predicted values are larger than the predicted means because $w_i(1-\hat{\gamma}_i)$ is always positive and the sum of the original predicted values is less than the weighted state total.

6. MSE ESTIMATION

An estimator of the MSE of prediction for the untransformed model is given in (12). However, the result is not easily extended to the transformed model or to the calibrated estimator that are required for this dataset. An alternative approach is to use a replication based estimator of the MSE. Hall and Maiti (2006b) proposed a parametric double-bootstrap, and Hall and Maiti (2006a) proposed a non-parametric double-bootstrap method to estimate the MSE. We computed the parametric double-bootstrap estimator for the transformed estimator ($\hat{\mu}_{i,TE}$), bias corrected estimator ($\hat{\mu}_{i,BE}$), and the calibrated estimator ($\hat{\mu}_{i,CE}$) for the transformed model. Given the normal probability plot of Fig. 5, a parametric bootstrap based on the normal model is a reasonable approach. The first bootstrap samples are obtained by sampling independently from $u_i^I \overset{ind}{\sim} \text{N}(0, \hat{\sigma}_u^2)$, and $e_i^I \overset{ind}{\sim} \text{N}(0, D_i^*)$ where $\hat{\beta}$ and $\hat{\sigma}_u^2$ are the estimated values of β and σ_u^2 from the transformed model and D_i^* are the design variances in the transformed scale. Compute the bootstrap estimate $(\hat{\beta}^I, \hat{\sigma}_u^{2I})$ of (β^I, σ_u^{2I}) by using data from the first bootstrap samples. The second bootstrap samples are obtained by sampling independently from $u_i^{II} \overset{ind}{\sim} \text{N}(0, \hat{\sigma}_u^{2I})$, and $e_i^{II} \overset{ind}{\sim} \text{N}(0, D_i^*)$. The estimation method can be described as follows. Let $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}_u^2$ denote the EGLS estimates of β_0, β_1, and σ_u^2, respectively, for the transformed model. Let $\mathbf{u} = (u_1, \ldots, u_m)^T$, and $\mathbf{e} = (e_1, \ldots, e_m)^T$.

- Repeat for $r = 1$ to R
  - Generate $\mathbf{u}^{I(r)} \overset{ind}{\sim} \text{N}(\mathbf{0}, \hat{\sigma}_u^2)$, and $\mathbf{e}^{I(r)} \overset{ind}{\sim} \text{N}(\mathbf{0}, \boldsymbol{D}^*)$ so that they are independent
  - Compute $\{\mathbf{y}^{1/3}\}^{I(r)} = X\hat{\beta} + \mathbf{u}^{I(r)} + \mathbf{e}^{I(r)}$
  - Compute the EGLS estimators $\hat{\beta}^{I(r)}$, and $\hat{\sigma}_u^{2I(r)}$; and the small area predictor $\hat{\mu}_i^{I(r)}$
  - Repeat for $s = 1$ to S
    - Generate $\mathbf{u}^{II(r,s)} \overset{ind}{\sim} \text{N}(0, \hat{\sigma}_u^{2I(r)})$, and $\mathbf{e}^{II(r,s)} \overset{ind}{\sim} \text{N}(\mathbf{0}, \boldsymbol{D}^*)$ so that they are independent
    - Compute $\{\mathbf{y}^{1/3}\}^{II(r,s)} = X\hat{\beta}^{I(r)} + \mathbf{u}^{II(r,s)} + \mathbf{e}^{II(r,s)}$
    - Compute the EGLS estimators $\hat{\beta}^{II(r,s)}$, and $\hat{\sigma}_u^{2II(r,s)}$; and the small area predictors $\hat{\mu}_i^{II(r,s)}$

- Compute $\text{MSE}_{2i}^{(r)}$

$$= S^{-1}\sum_{s=1}^{S}\left(\hat{\mu}_i^{II(r,s)} - \tilde{\mu}_i^{II(r,s)}\right)^2,$$

where $\tilde{\mu}_i^{II(r,s)} = x_i\hat{\beta}^{I(r)} + u_i^{II(r,s)}$

- Compute $v_{1i} = MSE_{1i}$

$$= R^{-1}\sum_{r=1}^{R}\left(\hat{\mu}_i^{I(r)} - \tilde{\mu}_i^{I(r)}\right)^2,$$

where $\tilde{\mu}_i^{I(r)} = x_i\hat{\beta} + u_i^{I(r)}$

- Compute $v_{2i} = \hat{\text{E}}\,(\text{MSE}_{2i}) = R^{-1}\sum_{r=1}^{R}\text{MSE}_{2i}^{(r)}$

Finally, the double-bootstrap estimator of the MSE of $\hat{\mu}_i$ is given by

$$\text{mse}(\hat{\mu}_i) = v_{1i} + m^{-1}\tan^{-1}\{m(v_{1i} - v_{2i})\},\ v_{1i} \geq v_{2i}$$

$$v_{1i}^2 \big/ [v_{1i} + m^{-1}\tan^{-1}\{m(v_{2i} - v_{1i})\}],\ v_{1i} < v_{2i}. \quad (33)$$

The two-case definition of the mse in (33) is used to ensure that the estimator is always positive. MSE of predictions for the transformed estimator, bias corrected estimator, and calibrated estimator are computed by using the double-bootstrap estimator (33) with $R = 220$, and $S = 100$. On average, the double-bootstrap MSEs are about 0.999 of the naive bootstrap MSEs.

The double-bootstrap MSEs have large variances compared to the estimated MSEs because of the relatively small R. We use the order one variance in the cube root scale and county IFact to smooth bootstrap MSEs in the cube root scale. Let $v_{1ii}^* = (\hat{\sigma}_u^2 + D_i^*)^{-1} D_i^*\hat{\sigma}_u^2$ be an estimate of the first order term in the variance of the predictor in the cube root scale. Let $v_{2ii}^* = 1 + 0.5n\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1}(x_i - \bar{x})^2$ be a crude approximation for a multiple of the order n^{-1} term in the variance of the prediction in the cube root scale. Let b_i^* be the double-bootstrap estimate of the MSE in the cube root scale, and b_i be the double-bootstrap estimate of the MSE in the original data scale for county i. The smooth estimate of the MSE in the cube root scale is

$$\hat{b}_i^* = v_{1ii}^* + \hat{\alpha}v_{2ii}^*, \quad (34)$$

where $\hat{\alpha}$ is the regression coefficient for the regression of $b_i^* - v_{1ii}^*$ on v_{2ij}^*. Let $c_i = \hat{b}_i^*\{3(x_i\hat{\beta})^2\}^2$ be the transformed estimate for the MSE in the original scale. The smooth bootstrap estimates for the MSEs in the original scale are the fitted values from the regression of b_i on c_i without an intercept.

7. SUMMARY OF RESULTS

Statistics for the original direct estimators, the predicted means using the transformed predictor, the bias corrected predictor, and the calibrated predictor are given in Table 2. The table has four pairs of rows. The first pair is for the direct estimates. The first row gives statistics for the direct estimator, where, for example, the median of the 44 direct estimates is 0.514. The second row gives statistics for the estimated coefficient of variation, denoted by $\widehat{\text{CV}}$. The $\widehat{\text{CV}}$ for the direct estimator are calculated using the back transformed D_i^*.

Table 2. Summary statistics for the predicted means.

		First Quartile	Median	Mean	Third Quartile	Range
Original	Estimated	0.196	0.514	0.696	0.970	3.477
Mean	$\widehat{\text{CV}}$	0.426	0.514	0.534	0.645	0.707
Simple	Predicted	0.216	0.488	0.637	0.859	2.229
Predictor (35)	$\widehat{\text{CV}}$	0.248	0.290	0.310	0.364	0.377
Bias Corrected	Predicted	0.232	0.509	0.658	0.881	2.256
Predictor (16)	$\widehat{\text{CV}}$	0.273	0.319	0.342	0.401	0.416
Calibrated	Predicted	0.254	0.532	0.683	0.906	2.228
Predictor (32)	$\widehat{\text{CV}}$	0.252	0.294	0.315	0.370	0.383

The estimates in the second pair of rows are obtained by constructing the EBLUPs in the cube root scale and then transforming back to the original scale. That predictor is,

$$\hat{\mu}_{i,TE} = \{\hat{\gamma}_i (y_i)^{1/3} + (1-\hat{\gamma}_i)\, x_i^T \hat{\beta}\}^3 \qquad (35)$$

For the transformed estimator, the mean of predicted values (0.637) in Table 2 is the simple average of predicted values for the 44 counties. The mean $\widehat{\text{CV}}$ (0.310) in Table 2 is the average $\widehat{\text{CV}}$ for the 44 counties.

The third and the fourth pairs of rows in Table 2 are for the bias corrected and the calibrated predictors, respectively. The bias correction increases the predictions by an average of about 0.02 which is about 3% of the average. Calibration increases the predictions by about 0.05 relative to the simple predictions. The average of model $\widehat{\text{CV}}$s for the calibrated estimator is 31.5%, and the average of model $\widehat{\text{CV}}$s for the bias corrected estimator is 34.2%. The average $\widehat{\text{CV}}$s for the direct estimates is 53.4%. The third quartile of $\widehat{\text{CV}}$s the model $\widehat{\text{CV}}$s for the calibrated estimator (37.0%) and the bias corrected estimator (40.1%) are also lower than the third quartile of the $\widehat{\text{CV}}$s for the direct estimates (64.5%). The model $\widehat{\text{CV}}$s for all three small area predictors are smaller than the $\widehat{\text{CV}}$s for the direct estimates in all counties.

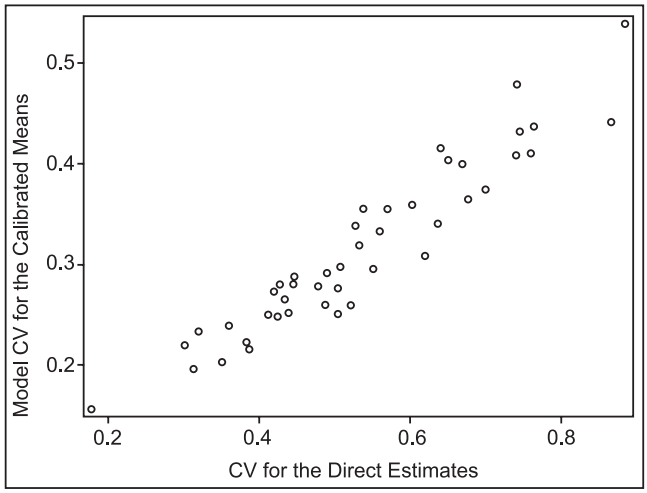

Fig. 7. Coefficients of variation for the small area predicted values and the direct estimates

Fig. 7 is a plot of $\widehat{\text{CV}}$s for the calibrated estimator plotted against the $\widehat{\text{CV}}$s for the direct estimator. The $\widehat{\text{CV}}$s for the model estimates are 50% to 88% of the corresponding $\widehat{\text{CV}}$s for the direct estimates. As is usual in small area estimation, the gain from model predictions are larger for areas whose direct estimators have large $\widehat{\text{CV}}$s.

ACKNOWLEDGEMENTS

This paper describes research and analyses conducted by the authors, and is released to inform interested parties and encourage discussion. Results and conclusions are the authors' and have not been endorsed by the NRCS. The research was supported in part by Cooperative Agreement No. 63-3A754-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University. The second author's research was partially supported by NSF Grant No. SES-0961649. The paper was in part prepared while Pushpal K. Mukhopadhyay was a graduate Research Assistant at the Department of Statistics, Iowa State University. The authors are grateful to the anonymous reviewer and the Associate Editor for their constructive suggestions, which led to a much improved paper.

REFERENCES

Conover, W.J. (1999). *Practical Nonparametric Statistics*. 3rd edition, John Wiley & Sons, New York.

Fuller, W.A. (2003). Sample selection for the 2000 NRI-2004 NRI surveys. Unpublished Manuscript, Center for Survey Statistics and Methodology.

Fuller, W.A. (2009). *Sampling Statistics*. John Wiley & Sons, Hoboken, New Jersey.

Hagen, L.J. (1994). Wind erosion in the United States. In : *Proceedings of Wind Erosion Symposium*, 25-32. Poznan, Poland.

Hagen, L.J. and Woodruff, N.P. (1973). Air pollution from dust storms in the great plains. *Atmospheric Environment*, **7**, 323-332.

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Ann. Statist.*, **34(4)**, 1733-1750.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *J. Roy. Statist. Soc.,* **B68(2)**, 221-238.

Nusser, S.M. and Goebel, J. (1997). The national resource inventory: A long-term multi-resource monitoring program. *Environ. Ecol. Statist.*, **4**, 181-204.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of the small area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology, Hoboken, New Jersey.

Slud, E.V. and Maiti, T. (2006). Mean-squared error estimation in transformed Fay-Herriot models. *J. Roy. Statist. Soc.,* **B68(2)**, 239-258.

Wang, J. and Fuller, W.A. (2002). Small area estimation under a restriction. In: *Proceedings of the Joint Statistical Meetings, volume* CD-ROM.

Woodruff, N.P. and Siddoway, F.H. (1965). A wind erosion equation. In: *SSSA Proceedings*, **29**, 602-608.