An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data

GEORGE E. BATTESE, RACHEL M. HARTER, and WAYNE A. FULLER*

Knowledge of the area under different crops is important to the U.S. Department of Agriculture. Sample surveys have been designed to estimate crop areas for large regions, such as crop-reporting districts, individual states, and the United States as a whole. Predicting crop areas for small areas such as counties has generally not been attempted, due to a lack of available data from farm surveys for these areas. The use of satellite data in association with farm-level survey observations has been the subject of considerable research in recent years. This article considers (a) data for 12 Iowa counties, obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture and (b) data obtained from land observatory satellites (LANDSAT) during the 1978 growing season. Emphasis is given to predicting the area under corn and soybeans in these counties. A linear regression model is specified for the relationship between the reported hectares of corn and soybeans within sample segments in the June Enumerative Survey and the corresponding satellite determination for areas under corn and soybeans. A nested-error model defines a correlation structure among reported crop hectares within the counties. Given this model, the mean hectares of the crop per segment in a county is defined as the conditional mean of reported hectares, given the satellite determinations and the realized (random) county effect. The mean hectares of the crop per segment is the sum of a fixed component, involving unknown parameters to be estimated and a random component to be predicted. Variance-component estimators in the nested-error model are defined, and the generalized least-squares estimators of the parameters of the linear model are obtained. Predictors of the mean crop hectares per segment are defined in terms of these estimators. An estimator of the variance of the error in the predictor is constructed, including terms arising from the estimation of the parameters of the model. Predictions of mean hectares of corn and soybeans per segment for the 12 Iowa counties are presented. Standard errors of the predictions are compared with those of competing predictors. The suggested predictor for the county mean crop area per segment has a standard error that is considerably less than that of the traditional survey regression predictor.

KEY WORDS: Small-area estimation; LANDSAT; June Enumerative Survey; Components of variance; Nested-error model.

1. INTRODUCTION

The U.S. Department of Agriculture (USDA) has been investigating the use of LANDSAT satellite data, both to improve its estimates of crop areas for crop-reporting districts and to develop estimates for individual counties. The methodology used in some of these studies was presented by Cárdenas, Blanchard, and Craig (1978), Hanuschak et al. (1979), and Sigman, Hanuschak, Craig, Cook, and Cárdenas (1978). Additional research was presented by Chhikara (1984).

The USDA is engaged in several interrelated types of research. Some research is directed toward transforming satellite information into good estimates of crop areas at the individual pixel and segment levels. The "segment" is the primary sampling unit, and a "pixel" (a term for "picture element") is the unit for which satellite information is recorded. Segments are about 250 hectares; a pixel is about .45 hectares. Other research is aimed at producing good estimators of total crop areas for both large and small geographical units. Studies by Hanuschak et al. (1979) and Hung and Fuller (1987) concentrated on obtaining good functions of the satellite data.

In this article we consider the prediction of areas under corn and soybeans for 12 counties in north-central Iowa, based on 1978 June Enumerative Survey and satellite data. The USDA Statistical Reporting Service field staff determined the area of corn and soybeans in the 37 segments of these 12 counties by interviewing farm operators. Data for more than one sample segment are available for several counties. Based on LANDSAT readings obtained during August and September 1978, USDA procedures were used to classify the crop cover for all pixels in the 12 counties. Table 1 presents (a) the number of segments in each county, (b) the number of hectares of corn and soybeans for each sample segment (as reported in the June Enumerative Survey), (c) the number of pixels classified as corn and soybeans for each sample segment, and (d) the county mean number of pixels per segment classified as corn and soybeans.

A preliminary analysis of the corn data indicated that the second segment in Hardin county deviated from other observations: The reported hectares of corn for the second segment were identical to that of the first segment. Therefore, all data for that (second) segment are deleted from our analyses. The soybean data are deleted for convenience, so the same number of observations is involved for both crops.

* George E. Battese is Senior Lecturer, Department of Econometrics, University of New England, Armidale, New South Wales 2351, Australia. Rachel M. Harter is Associate Research Director, A. C. Nielsen Company, Northbrook, IL 60062. Wayne A. Fuller is Distinguished Professor, Department of Statistics, Iowa State University, Ames, IA 50011. This research was partly supported by Research Agreement 58-319T-1-0054X with the Statistical Reporting Service of the U.S. Department of Agriculture, and Joint Statistical Agreement 82-6 with the U.S. Bureau of the Census. The authors thank Cheryl Auer and Stephen Miller for assistance in writing computer programs for the empirical analyses. Comments of the editors and referees resulted in numerous changes to earlier drafts of the article. A part of this research was conducted during the periods the first author was at Iowa State University, on study leaves from the University of New England.

Table 1. Survey and Satellite Data for Corn and Soybeans in 12 Iowa Counties

	No. of segments		Reported hectares		No. of pixels in sample segments		Mean number of pixels per segment*	
County	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	189.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humboldt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	29.46	340	87		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

* The mean number of pixels of a given crop per segment in a county is the total number of pixels classified as that crop, divided by the number of segments in that county.

Figures 1 and 2 plot the reported hectares of corn and soybeans for the remaining 36 segments against the number of pixels of corn and soybeans, respectively. Observations from segments within given counties are identified with different symbols and jointed by lines, so the county data are more clearly indicated. It is evident that there is a strong relationship between the reported hectares of corn and the number of pixels of corn, and between the reported hectares of soybeans and the number of pixels of soybeans. In addition, the plots indicate that observations for segments within counties tend to be closer together than observations for the whole sample.

Predictors of mean crop areas per segment in the sample counties are obtained under the assumption that a linear regression model defines the relationship between the survey and satellite data. The random errors of the model are assumed to be defined by the nested-error model, in which deviations within a county are correlated. Estimation of this model was discussed by Fuller and Battese (1973) and was suggested for small-area estimation by Battese and Fuller (1981) and Fuller and Battese (1981). Alternative approaches to small-area estimation were given by Fuller and Harter (1987). Fuller and Harter (1987) also presented additional details for the methodology in this article.

2. COMPONENTS-OF-VARIANCE MODEL

The reported crop hectares for corn (or soybeans) in sample segments within counties are expressed as a function of the satellite data for those sample segments, such that the reported crop hectares are positively correlated within given counties but uncorrelated from different counties. The model is

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_{ij}, \qquad (2.1)$$

where i is the subscript for county ($i = 1, 2, \ldots, T$,

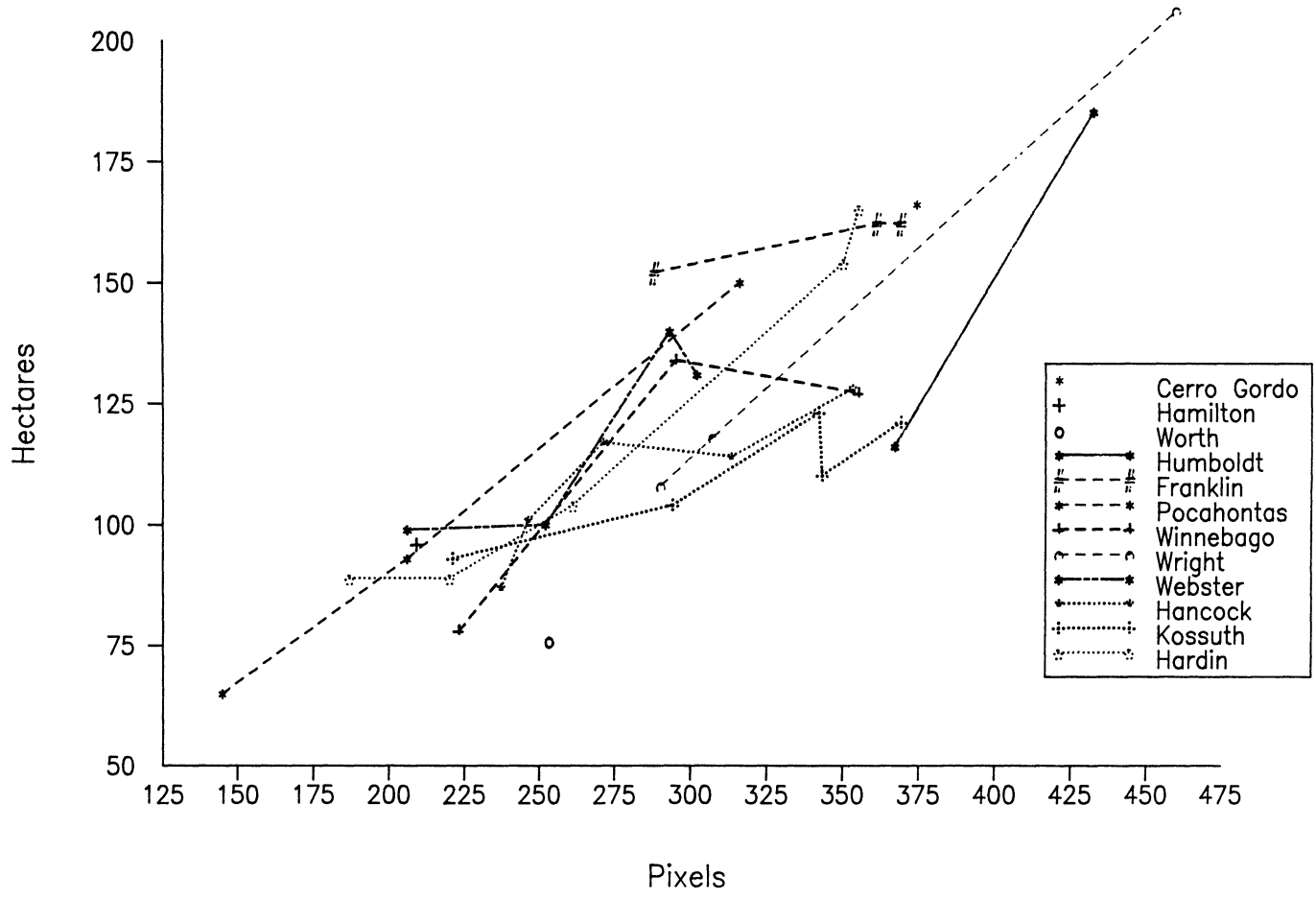

Figure 1. Plot of Corn Hectares Versus Corn Pixels by County.

where $T = 12$); j is the subscript for a segment within a given county ($j = 1, 2, \ldots, n_i$, where n_i is the number of sample segments in the ith county); y_{ij} is the number of hectares of corn (or soybeans) in the jth segment of the ith county, as reported in the June Enumerative Survey; x_{1ij} and x_{2ij} are the number of pixels classified as corn and soybeans, respectively, in the jth segment of the ith county; and β_0, β_1, and β_2 are unknown parameters.

The random error u_{ij}, associated with the reported crop area y_{ij}, is expressed as

$$u_{ij} = v_i + e_{ij}, \tag{2.2}$$

where v_i is the ith county effect and e_{ij} is the random effect associated with the jth sample segment within the ith county. The random errors, v_i ($i = 1, 2, \ldots, T$), are assumed to be iid $N(0, \sigma_v^2)$ random variables independent of the random errors, e_{ij} ($j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, T$), which are assumed to be iid $N(0, \sigma_e^2)$ random variables. These assumptions imply that the covariance structure of the random errors, u_{ij}, is given by

$$\begin{aligned} E(u_{ij}u_{pq}) &= \sigma_v^2 + \sigma_e^2, && i = p, j = q, \\ &= \sigma_v^2, && i = p, j \neq q, \\ &= 0, && i \neq p. \end{aligned} \tag{2.3}$$

This components-of-variance model is only one possible model for area effects associated with observations from similar geographic regions. Other correlation structures, where reported crop hectares for geographically closer segments have stronger correlation than those farther apart, were considered. Models were estimated where correlation was a function of distance between segments, but the distance effect was not statistically significant.

The components-of-variance model (2.1)–(2.2) does not explicitly define a correlation structure between reported hectares of corn and soybeans in sample segments within counties. The model can be expressed in a multivariate framework that considers the correlation between reported areas of corn and soybeans. Fuller and Harter (1987) covered the multivariate extension of the model (2.1)–(2.3). The extension did not improve the precision of estimation for our data, however, so we confine our attention to the univariate case.

The model for reported hectares of corn (or soybeans), defined by (2.1), was chosen after some preliminary investigations in which the reported hectares of corn (or soybeans) were defined in terms of quadratic functions of the numbers of pixels of corn and soybeans. In each case, however, the null hypothesis—that the coefficients of the nonlinear terms are 0—was not rejected at the 5% level. (Additional evaluation of the model is described in the discussion of empirical results.)

The sample mean of the reported hectares of corn (or soybeans) per segment in the ith county is denoted by

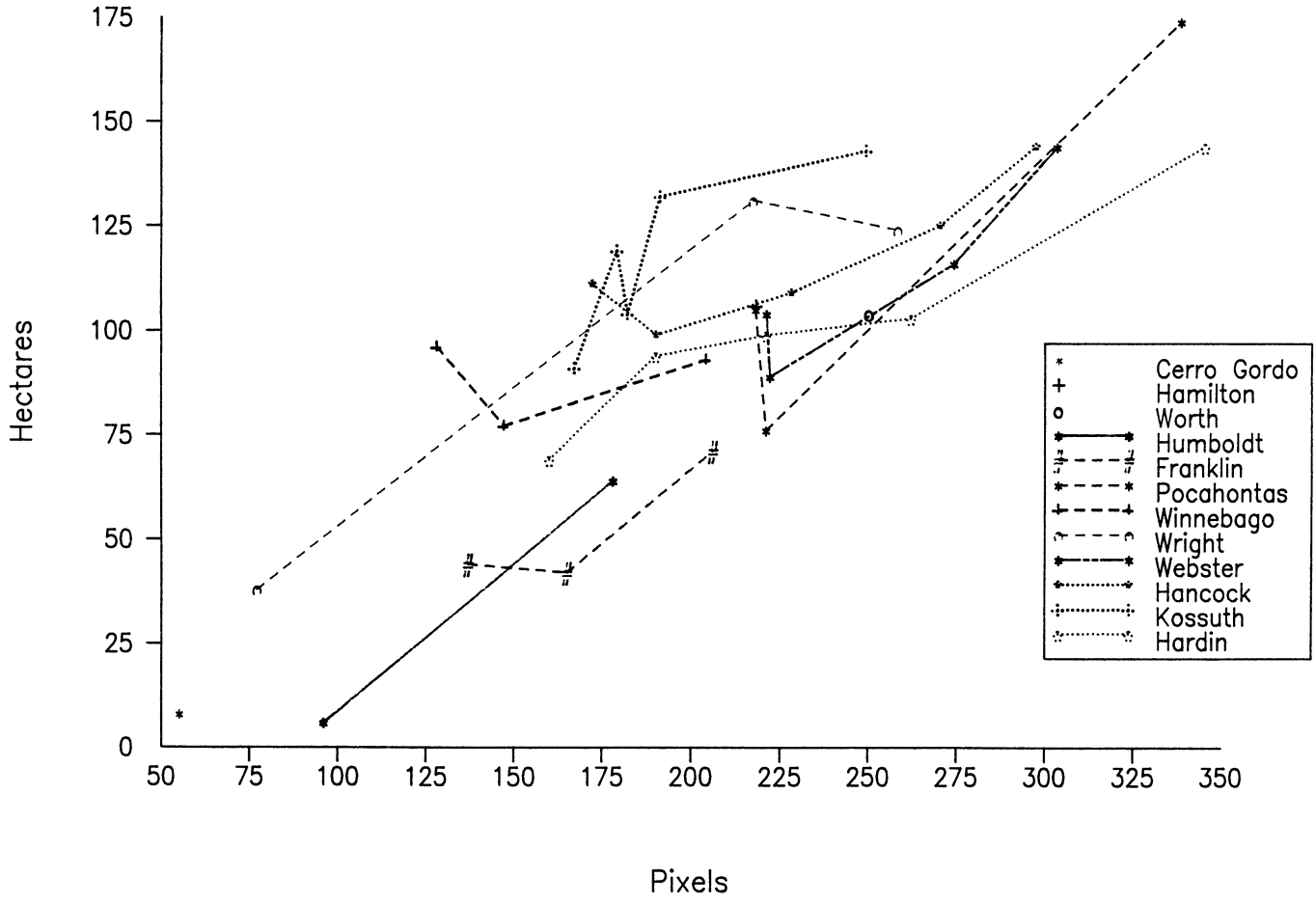

Figure 2. Plot of Soybean Hectares Versus Soybean Pixels by County.

$\bar{y}_{i.}$, where $\bar{y}_{i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$. The sample mean is expressed in terms of the parameters of the model (2.1)–(2.2) by

$$\bar{y}_{i.} = \beta_0 + \beta_1 \bar{x}_{1i.} + \beta_2 \bar{x}_{2i.} + v_i + \bar{e}_{i.}, \qquad (2.4)$$

where $\bar{x}_{1i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} x_{1ij}$ and $\bar{x}_{2i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} x_{2ij}$ are the sample mean numbers of pixels of corn and soybeans, respectively, in the n_i sample segments within county i, and $\bar{e}_{i.} \equiv n_i^{-1} \sum_{j=1}^{n_i} e_{ij}$ is the sample mean of the within-county effects for the sample segments in the ith county.

The population mean hectares of corn (or soybeans) per segment in the ith county is defined as the conditional mean of the hectares of corn (or soybeans) per segment, given the realized county effect v_i and the values of the satellite data. Under the assumptions of the model (2.1)–(2.2) this mean, denoted by y_i, is

$$y_i \equiv \beta_0 + \beta_1 \bar{x}_{1i(p)} + \beta_2 \bar{x}_{2i(p)} + v_i, \qquad (2.5)$$

where $\bar{x}_{1i(p)} \equiv N_i^{-1} \sum_{j=1}^{N_i} x_{1ij}$ and $\bar{x}_{2i(p)} \equiv N_i^{-1} \sum_{j=1}^{N_i} x_{2ij}$ are the population mean numbers of pixels classified as corn and soybeans per segment, respectively, in the ith county, and N_i is the total number of segments in the ith county. Because the number of pixels of corn and soybeans are available from the satellite classifications for all segments in the ith county, the population mean pixel values $\bar{x}_{1i(p)}$ and $\bar{x}_{2i(p)}$ are known. The prediction of the mean crop hectares per segment, defined by (2.5), is the focus of this article.

In a finite-population model, the mean hectares of corn (or soybeans) per segment in the ith county is $\bar{Y}_{i.} \equiv N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$, where Y_{ij} denotes the hectares of the crop in the jth segment in county i and the summation is over all segments in the population. The mean $\bar{Y}_{i.}$ is not equivalent to y_i [as defined in (2.5)], because the sum of the e_{ij}'s over the finite population of segments in county i is not identically 0. As shown in Section 3, however, the predictor for the mean y_i is an appropriate predictor for the finite-population mean $\bar{Y}_{i.}$ when the sampling rate is small.

Obtaining the mean crop hectares per segment in county i, y_i, involves predicting the sum of a known linear function of unknown parameters and an unobserved random variable, v_i. This is a special problem in predicting a linear combination of fixed effects and random effects (see Harville 1976, 1979; Henderson 1975; Kackar and Harville 1984; Peixoto and Harville 1986; Reinsel 1984, 1985). The theory for our parameter estimators and crop-area predictors is an extension of results in the articles cited previously, and is presented in more detail in Fuller and Harter (1987).

Before introducing the estimators and predictors, we present the components-of-variance model (2.1)–(2.3) in matrix notation. Let $\mathbf{Y}_i$ represent the column vector of the reported hectares of the given crop for the n_i sample segments in the ith county, $\mathbf{Y}_i \equiv (y_{i1}, y_{i2}, \ldots, y_{in_i})'$. Furthermore, let $\mathbf{Y}$ represent the column vector of the

reported hectares of the crop for the sample segments in the T counties, $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_T')'$. Thus model (2.1), expressed in matrix notation, is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{2.6}$$

where the row of $\mathbf{X}$ that corresponds to the element y_{ij} in $\mathbf{Y}$ is $\mathbf{x}_{ij} \equiv (1, x_{1ij}, x_{2ij})$ and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2)'$.

The covariance matrix for the random vector $\mathbf{u}$ in (2.6) is given by

$$E(\mathbf{u}\,\mathbf{u}') \equiv \mathbf{V} = \text{block diag}(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_T), \tag{2.7}$$

where

$$\mathbf{V}_i = \mathbf{J}_i\sigma_v^2 + \mathbf{I}_i\sigma_e^2, \tag{2.8}$$

with $\mathbf{J}_i$ the square matrix of order n_i with every element equal to 1 and $\mathbf{I}_i$ the identity matrix of order n_i.

The mean crop hectares per segment (2.5), expressed in matrix notation, is

$$y_i \equiv \bar{\mathbf{x}}_{i(p)}\boldsymbol{\beta} + v_i, \tag{2.9}$$

where $\bar{\mathbf{x}}_{i(p)} \equiv N_i^{-1}\sum_{j=1}^{N_i}\mathbf{x}_{ij} = (1, \bar{x}_{1i(p)}, \bar{x}_{2i(p)})$.

3. ESTIMATION AND PREDICTION

Basic to the prediction of the mean crop area (2.9) for the ith county is the prediction of the county effect, v_i. If the random errors u_{ij} $(j = 1, 2, \ldots, n_i)$ are known, then the best predictor of v_i is the conditional expectation of v_i, given the sample mean $\bar{u}_{i.}$, where $\bar{u}_{i.} \equiv n_i^{-1}\sum_{j=1}^{n_i}u_{ij}$. Under the assumptions of the model (2.1)–(2.2), the random variables v_i and $\bar{u}_{i.}$ have a bivariate normal distribution with zero mean vector and covariance matrix

$$\begin{pmatrix} \sigma_v^2 & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + n_i^{-1}\sigma_e^2 \end{pmatrix}.$$

The expectation of v_i, conditional on $\bar{u}_{i.}$, is

$$E(v_i|\bar{u}_{i.}) = \bar{u}_{i.}g_i, \tag{3.1}$$

where $g_i = m_i^{-1}\sigma_v^2$ and $m_i = (\sigma_v^2 + n_i^{-1}\sigma_e^2)$. The error variance in this best predictor is

$$\begin{aligned} \mathrm{E}\{(v_i - \bar{u}_{i.}g_i)^2\} &= \sigma_v^2(1 - g_i) \\ &= n_i^{-1}\sigma_e^2 - n_i^{-2}\sigma_e^2 m_i^{-1}\sigma_e^2. \end{aligned} \tag{3.2}$$

If the variances σ_v^2 and σ_e^2 are known, the β parameters of the model can be estimated by the generalized least-squares estimator

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{3.3}$$

Then a possible predictor for the ith county effect, v_i, is

$$\tilde{v}_i = \tilde{u}_{i.}g_i, \tag{3.4}$$

where $\tilde{u}_{i.} = n_i^{-1}\sum_{j=1}^{n_i}\tilde{u}_{ij}$ and $\tilde{u}_{ij} = y_{ij} - \mathbf{x}_{ij}\tilde{\boldsymbol{\beta}}$. The corresponding predictor $\tilde{y}_i$ for the county mean crop area per segment (2.9) is

$$\tilde{y}_i = \bar{\mathbf{x}}_{i(p)}\tilde{\boldsymbol{\beta}} + \tilde{v}_i. \tag{3.5}$$

This is the best linear unbiased predictor of y_i (see Harville 1985).

The variance of the error in the predictor (3.5) is

$$\mathrm{E}\{(\tilde{y}_i - y_i)^2\} = \sigma_v^2(1 - g_i) + \mathbf{c}_i\mathbf{V}(\tilde{\boldsymbol{\beta}})\mathbf{c}_i', \tag{3.6}$$

where $\mathbf{V}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, $\mathbf{c}_i = \bar{\mathbf{x}}_{i(p)} - g_i\bar{\mathbf{x}}_{i.}$, and $\bar{\mathbf{x}}_{i.} = n_i^{-1}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$. The variance of (3.6) is larger than that of (3.2) by the term associated with the estimation of $\boldsymbol{\beta}$.

A predictor for the finite population mean crop hectares per segment in the ith county [see the paragraph following (2.5)] is

$$N_i^{-1}\left[\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i}(\mathbf{x}_{ij}\tilde{\boldsymbol{\beta}} + \tilde{v}_i)\right].$$

In this predictor, the unobserved y_{ij} are replaced by the model predictions. It approaches the predictor (3.5) as the sampling rate decreases. Because the sampling rates are small in our application, we use the predictor (3.5).

This predictor is one of several that have been suggested for the small-area problem. Let a class of predictors of the county mean crop area y_i be defined by

$$\bar{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}} + (\bar{y}_{i.} - \bar{\mathbf{x}}_{i.}\hat{\boldsymbol{\beta}})\delta_i, \tag{3.7}$$

where δ_i is a nonnegative constant and $\hat{\boldsymbol{\beta}}$ is an estimator for $\boldsymbol{\beta}$.

For $\delta_i = g_i$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$, the predictor (3.7) is the best linear unbiased predictor. For $\delta_i = 0$, the predictor (3.7) is $\bar{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}}$; this is called the *regression synthetic predictor.* The term *synthetic* is used for predictors that are functions of $\bar{\mathbf{x}}_{i(p)}$, which may not be linear in $\bar{\mathbf{x}}_{i(p)}$. The predictor (3.7) when $\delta_i = 1$ is $\bar{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}} + (\bar{y}_{i.} - \bar{\mathbf{x}}_{i.}\hat{\boldsymbol{\beta}})$, which is equivalent to the *survey regression predictor* $\bar{y}_{i.} + (\bar{\mathbf{x}}_{i(p)} - \bar{\mathbf{x}}_{i.})\hat{\boldsymbol{\beta}}$. The survey regression predictor adjusts the sample survey mean $\bar{y}_{i.}$, using the difference between the population mean of the regressor vector $\bar{\mathbf{x}}_{i(p)}$ and the sample mean of the regressor values for the sample segments $\bar{\mathbf{x}}_{i.}$ in county i. The survey regression predictor, with an alternative form for the estimator $\hat{\boldsymbol{\beta}}$, was considered by Särndal (1984). Under the model in which $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$ the survey regression predictor is unbiased for y_i, conditional on the realized county effect v_i and the values of the satellite data.

The generalized least-squares estimator (3.3) and the predictor (3.5) for the county mean crop area are infeasible, because the variances σ_v^2 and σ_e^2 associated with the nested-error model are unknown. Harville (1977) reviewed a number of methods for estimating the variances for components-of-variance models. We obtain the fitting-of-constants estimator for σ_e^2, denoted by $\hat{\sigma}_e^2$, which is defined by the residual mean square for the regression model (2.1), with dummy variables for the counties. Alternatively, $\hat{\sigma}_e^2$ is expressed as

$$\hat{\sigma}_e^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}\left[\sum_{i=1}^{T}(n_i - 1) - 2\right]^{-1}, \tag{3.8}$$

where $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ is the residual sum of squares for the regression of the y deviations, $y_{ij} - \bar{y}_{i.}$, on the x deviations, $\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i.}$, for those counties with $n_i > 1$. Under the assumptions of model (2.1)–(2.3), the estimator $\hat{\sigma}_e^2$ is unbiased for σ_e^2 and is distributed as a multiple of a chi-squared random

variable. That is, $d_e\hat{\sigma}_e^2/\sigma_e^2$ has a chi-squared distribution with d_e df, where $d_e \equiv \sum_{i=1}^{T}(n_i - 1) - 2$.

An estimator for the variance of county effects, $\hat{\sigma}_v^2$, is obtained by considering the average of the ordinary least-squares residuals for county i,

$$\overset{\Delta}{u}_{i.} = \bar{y}_{i.} - \bar{\mathbf{x}}_{i.}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.9}$$

It is readily verified that

$$E(\overset{\Delta}{u}_{i.}^2) = b_i\sigma_v^2 + d_i\sigma_e^2, \tag{3.10}$$

where

$$b_i = 1 - 2n_i\bar{\mathbf{x}}_{i.}(\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}}_{i.}' + \bar{\mathbf{x}}_{i.}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{j=1}^{T} n_j^2\bar{\mathbf{x}}_{j.}'\bar{\mathbf{x}}_{j.}\right)(\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}}_{i.}'$$

and $d_i = n_i^{-1}[1 - n_i\bar{\mathbf{x}}_{i.}(\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}}_{i.}']$. Thus the weighted sum of squares of the average residuals for the counties,

$$\hat{m}_{..} \equiv \left(\sum_{i=1}^{T} n_i b_i\right)^{-1}\left(\sum_{i=1}^{T} n_i\overset{\Delta}{u}_{i.}^2\right), \tag{3.11}$$

has expectation

$$\mathrm{E}(\hat{m}_{..}) \equiv m_{..} = \sigma_v^2 + c\sigma_e^2, \tag{3.12}$$

where $c = (\sum_{i=1}^{T} n_i b_i)^{-1}(\sum_{i=1}^{T} n_i d_i)$. Under the assumptions of the model (2.1)–(2.2), the weighted sum of squares $\hat{m}_{..}$ is independent of $\hat{\sigma}_e^2$. Our estimator for σ_v^2 is

$$\hat{\sigma}_v^2 = \max\{\hat{m}_{..} - c\hat{\sigma}_e^2, 0\}. \tag{3.13}$$

An estimator of g_i is

$$\hat{g}_i = (\hat{\sigma}_v^2 + n_i^{-1}\hat{\sigma}_e^2)^{-1}\hat{\sigma}_v^2. \tag{3.14}$$

A feasible predictor for the mean crop area (2.9) in county i is

$$\hat{y}_i \equiv \bar{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}} + \hat{u}_{i.}\hat{g}_i, \tag{3.15}$$

where $\hat{\boldsymbol{\beta}}$ is the estimated generalized least-squares estimator for $\boldsymbol{\beta}$, obtained by replacing $\mathbf{V}$ of (3.3) with $\hat{\mathbf{V}}$, where $\hat{\mathbf{V}}$ is the estimator for the covariance matrix (2.7) obtained by using the estimators $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$, defined by (3.8) and (3.13), respectively; $\hat{u}_{i.} \equiv \bar{y}_{i.} - \bar{\mathbf{x}}_{i.}\hat{\boldsymbol{\beta}}$; and $\hat{g}_i$ is an alternative estimator to (3.14), which is defined in the Appendix. The estimator $\hat{g}_i$, which is approximately unbiased for g_i, was suggested by Fuller and Harter (1987).

An approximation for the variance of the prediction error, $y_i - \hat{y}_i$, and estimators for this variance, were given by Fuller and Harter (1987) for the multivariate case. An estimator for the variance of the prediction error is given in the Appendix. For more detail on the predictor, and the estimator for the variance of the error in the predictor, readers should consult Fuller and Harter (1987).

4. EMPIRICAL RESULTS

Estimates for the parameters of the model (2.1)–(2.2) are obtained by using a modification of the nested-error option of SUPER CARP (see Hidiroglou, Fuller, and Hickman 1980). The modification of SUPER CARP incorporates the alternative estimator for the variance σ_v^2, defined by (3.13). The variance components are first estimated, and then the estimated generalized least-squares estimators for the β parameters are obtained. The estimated parameters for corn are

$$\hat{y}_{ij} = \underset{(25)}{51} + \underset{(.050)}{.329}\,x_{1ij} - \underset{(.056)}{.134}\,x_{2ij},$$

$$\hat{\sigma}_e^2 = \underset{(45)}{150}, \qquad \hat{\sigma}_v^2 = \underset{(89)}{140.}$$

The estimated parameters for soybeans are

$$\hat{y}_{ij} = \underset{(29)}{-16} + \underset{(.058)}{.028}\,x_{1ij} + \underset{(.065)}{.494}\,x_{2ij},$$

$$\hat{\sigma}_e^2 = \underset{(59)}{195}, \qquad \hat{\sigma}_v^2 = \underset{(49)}{272.}$$

The value of the constant c, defined by (3.12), is .349.

The three estimates of the regression function are statistically significant in the corn function, but only the coefficient of soybeans pixels is significantly different from 0 for the soybean function. The estimated variances for within- and among-county variation in reported crop hectares are approximately equal for corn, but for soybeans the among-county variance is about 60% of the total of the two variances. The among-county variance is significant at the 10% level for corn and the 1% level for soybeans.

In our model (2.1)–(2.2) the errors are assumed to be normally distributed. The predictor of the mean crop areas for the counties retains desirable properties for nonnormal errors, but the estimated variances of the prediction errors can be seriously biased when the errors are not normally distributed. Normal probability plots are presented in Figures 3 and 4 for the transformed residuals, $\hat{u}_{ij}^*$, for the corn and soybean models, respectively, which are defined by

$$\hat{u}_{ij}^* = (y_{ij} - \hat{\alpha}_i\bar{y}_{i.}) - (\mathbf{x}_{ij} - \hat{\alpha}_i\bar{\mathbf{x}}_{i.})\hat{\boldsymbol{\beta}},$$

where $\hat{\alpha}_i \equiv 1 - [\hat{\sigma}_e^2/(\hat{\sigma}_e^2 + n_i\hat{\sigma}_v^2)]^{1/2}$. These transformed residuals are approximately uncorrelated with variances approximately equal to σ_e^2 (see e.g., Fuller and Battese 1973, p. 627). The Shapiro–Wilk W statistic for the transformed residuals had values of .985 and .957 for corn and soybeans, respectively. If the residuals were independent normal samples, then the probabilities of values less than those observed would be .921 and .299, respectively. The sample is small, but these analyses give no reason to reject the hypothesis that the errors in the model (2.2) are normally distributed.

Given the assumptions of the model (2.1)–(2.2), the parameters β_1 and β_2 can be estimated from the multiple regression of the within-county deviations $y_{ij} - \bar{y}_{i.}$ on the deviations $x_{1ij} - \bar{x}_{1i.}$ and $x_{2ij} - \bar{x}_{2i.}$ [see Eq. (3.8)]. The expectation of these estimators for β_1 and β_2, represented by $\hat{\boldsymbol{\beta}}_W$, is the same as the expectation of the generalized least-squares estimators of β_1 and β_2, represented by $\hat{\boldsymbol{\beta}}_G$. Hence the estimators $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_G$ can be used to construct a test of model (2.1). Let $\hat{\boldsymbol{\Sigma}}_W$ be the estimated covariance matrix of the within-county estimator $\hat{\boldsymbol{\beta}}_W$, and let $\hat{\boldsymbol{\Sigma}}_G$ be

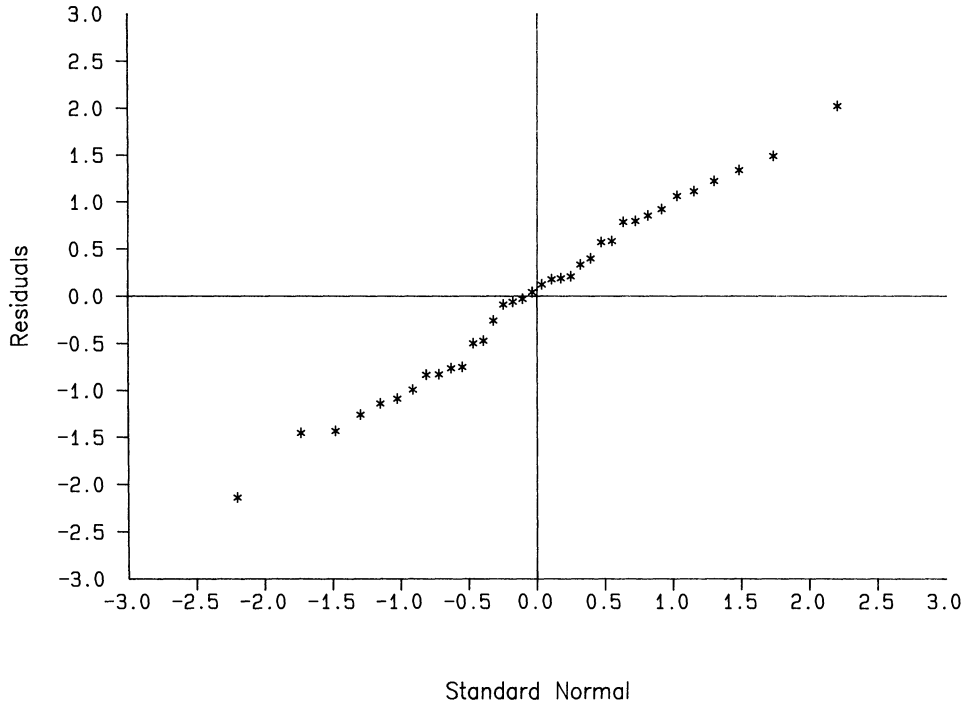

Figure 3. Full Normal Plot Residuals of the Transformed Corn Model.

the estimated covariance matrix of the generalized least-squares estimator $\hat{\boldsymbol{\beta}}_G$. Then the approximate distribution of the statistic

$$F = 2^{-1}(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_G)'(\hat{\boldsymbol{\Sigma}}_W - \hat{\boldsymbol{\Sigma}}_G)^{-1}(\hat{\boldsymbol{\beta}}_W - \hat{\boldsymbol{\beta}}_G)$$

is the F distribution with 2 and 22 df, under the null hypothesis that the slope parameters are the same within and among counties. This result follows from the fact that the estimated covariance between $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_G$ is $\hat{\boldsymbol{\Sigma}}_G$. The test statistic is .46 for corn and .60 for soybeans. Hence we accept the hypothesis that the parameters β_1 and β_2 are the same for within and among counties, as postulated in (2.1).

With the predictor (3.15) we obtain the predictions for the mean crop hectares per segment. Results are given in Tables 2 and 3 for corn and soybeans, respectively, along with the estimated standard errors for the best predictor (3.15), the survey regression predictor, and the sample

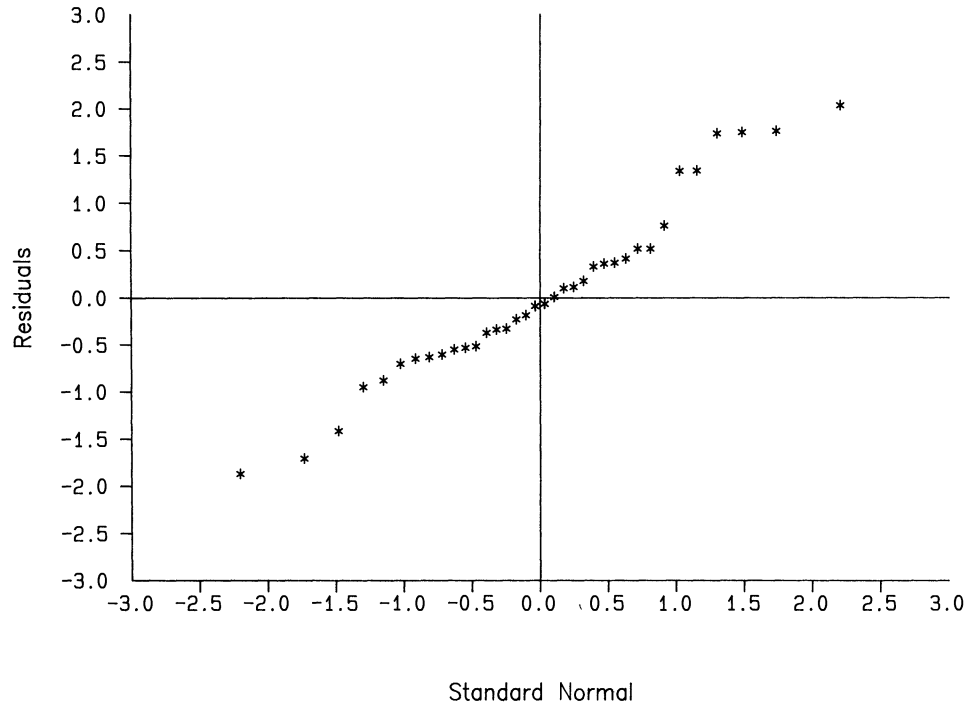

Figure 4. Full Normal Plot Residuals of the Transformed Soybean Model.

Table 2. Predicted Hectares of Corn With Standard Errors of Alternative Predictors

County	Sample segments	Predicted hectares	Standard errors: Best predictor	Standard errors: Survey regression predictor	Standard errors: Sample mean
Cerro Gordo	1	122.2	9.6	13.7	30.5
Hamilton	1	126.3	9.5	12.9	30.5
Worth	1	106.2	9.3	12.4	30.5
Humboldt	2	108.0	8.1	9.7	21.5
Franklin	3	145.0	6.5	7.1	17.6
Pocahontas	3	112.6	6.6	7.2	17.6
Winnebago	3	112.4	6.6	7.2	17.6
Wright	3	122.1	6.7	7.3	17.6
Webster	4	115.8	5.8	6.1	15.2
Hancock	5	124.3	5.3	5.7	13.6
Kossuth	5	106.3	5.2	5.5	13.6
Hardin	5	143.6	5.7	6.1	13.6

mean of the survey data. The estimated standard error of the sample mean is the square root of the within-county mean square, divided by the number of segments in the given county.

The differences between the predicted hectares of corn and soybeans and the corresponding sample means decrease (see Table 1) as the number of sample segments increases. This is because the standard errors of the sample means are larger for counties with small numbers of sample segments. The standard errors of the sample mean are considerably greater than those for the survey regression predictor. The ratio of the standard error of the best predictor to that for the survey regression predictor increases from about .77 to .97 as the number of sample segments increases from 1 to 5. When there are no more than 3 sample segments, the best predictor has a standard error considerably less than that for the survey regression predictor. The improvement in the precision of the predictor, obtained by increasing the number of sample segments in a county from 3 to 4 or 5, is modest.

5. COMMENTS

The survey regression predictor is unbiased for the 12 counties' mean crop area, and it has relatively small variance. Hence the survey regression predictor is adequate for the entire area. It then becomes desirable to modify the individual county predictors so that the properly weighted sum equals the unbiased survey regression predictor for the total area. A possible adjusted predictor for the ith county mean crop area involves adding to the best predictor a proportion of the weighted sum of the differences between the survey regression predictors and the corresponding best predictor for the counties involved. This predictor is defined by

$$\overset{*}{y}_i = \hat{y}_i + a_i \left[\sum_{j=1}^{T} W_j (\bar{y}_{i.} - \bar{\mathbf{x}}_{i.}\hat{\boldsymbol{\beta}})(1 - \hat{g}_i) \right],$$

where $a_i = [\Sigma_{j=1}^{T} W_j^2 \hat{V}(\hat{y}_j)]^{-1} W_i^2 \hat{V}(\hat{y}_i)$. $\hat{V}[\hat{y}_i]$ is the estimated variance of the prediction error for predictor (3.15), and W_j is the weight for the jth area used in constructing the predictor for the total area. It is clear that $\Sigma_{i=1}^{T} W_i \overset{*}{y}_i$ is equal to the unbiased survey regression predictor for the total area, $\Sigma_{i=1}^{T} W_i[\bar{y}_{i.} + (\bar{\mathbf{x}}_{i(p)} - \bar{\mathbf{x}}_{i.})\hat{\boldsymbol{\beta}}]$. The adjustment produces a very small increase in the variance of the small-area predictors under the components-of-variance model with unequal n_i and/or unequal W_i.

The nested-error regression model (with satellite data as auxiliary variables) offers a promising approach to predicting crop areas in small domains. The USDA has conducted exploratory analyses with the software developed for the univariate nested-error approach to predicting county crop areas. The procedure allows for the use of

Table 3. Predicted Hectares of Soybeans With Standard Errors of Alternative Predictors

County	Sample segments	Predicted hectares	Standard errors: Best predictor	Standard errors: Survey regression predictor	Standard errors: Sample mean
Cerro Gordo	1	77.8	12.0	15.6	29.1
Hamilton	1	94.8	11.8	14.8	29.1
Worth	1	86.9	11.5	14.2	29.1
Humboldt	2	79.7	9.7	11.1	20.6
Franklin	3	65.2	7.6	8.1	16.8
Pocahontas	3	113.8	7.7	8.2	16.8
Winnebago	3	98.5	7.7	8.3	16.8
Wright	3	112.8	7.8	8.4	16.8
Webster	4	109.6	6.7	7.0	14.6
Hancock	5	101.0	6.2	6.5	13.0
Kossuth	5	119.9	6.1	6.3	13.0
Hardin	5	74.9	6.6	6.9	13.0

supplementary information, such as estimates of variances from other areas and other years, in the estimation of variance components. Modification of the model to account for stratification, according to land use within counties, was investigated by both Walker and Sigman (1982) and Harter (1983).

APPENDIX: COMPUTATIONAL FORMULAS

The model parameters, county predictions, and standard errors in the empirical section were computed with an adaptation of the nested-error regression procedure of SUPER CARP (Hidiroglou, Fuller, and Hickman 1980). The modifications are based on the multivariate estimators suggested by Fuller and Harter (1987). Univariate forms of the estimators for this specific example follow.

The predictor for the county mean crop hectares per segment, defined by (3.15), is

$$\hat{y}_i = \bar{\mathbf{x}}_{i(p)}\hat{\boldsymbol{\beta}} + (\bar{y}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot}\hat{\boldsymbol{\beta}})\hat{g}_i, \qquad \text{(A.1)}$$

where $\hat{g}_i = 1 - \hat{h}_i$,

$$\hat{h}_i = [\hat{m}_i + \hat{k}_i + (n_i^{-1} - c)^2\hat{w}_i]^{-1}[n_i^{-1}\hat{\sigma}_e^2 + (n_i^{-1} - c)n_i^{-1}\hat{w}_i],$$

$$\hat{m}_i = \hat{m}_{\cdot\cdot} + (n_i^{-1} - c)\hat{\sigma}_e^2, \qquad \hat{w}_i = 2d_e^{-1}\hat{m}_i^{-1}\hat{\sigma}_e^4,$$

$$\hat{k}_i = 2\hat{\sigma}_e^2(\ddot{\sigma}_{ff} + n_i^{-1})^{-1}\left[\sum_{j=1}^{T} n_j b_j\right]^{-2}\left[\sum_{j=1}^{T} n_j^2 b_j(\ddot{\sigma}_{ff} + n_j^{-1})^2\right],$$

and $\ddot{\sigma}_{ff} = \max[0, (T-5)^{-1}(T-3)\hat{\sigma}_e^{-2}\hat{m}_{\cdot\cdot} - c]$. The constant b_i is defined after (3.10), c is defined after (3.12), and $d_e = 22$ for this application.

The variance of the error in the predictor (A.1) is estimated by

$$\hat{V}\{\hat{y}_i - y_i\} = n_i^{-1}\hat{\sigma}_e^2 - \hat{\phi}_i + \hat{\mathbf{c}}_i\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})\hat{\mathbf{c}}_i' + \hat{h}_i^2\hat{k}_i + d_e^{-1}\hat{r}_i^2\hat{\phi}_i + d_e^{-1}\hat{r}_i^2\hat{h}_i\hat{\sigma}_e^2, \qquad \text{(A.2)}$$

where $\hat{\mathbf{c}}_i = \bar{\mathbf{x}}_{i(p)} - \hat{g}_i\bar{\mathbf{x}}_{i\cdot}$, $\hat{\phi}_i = (d_e + 1)^{-1}d_e\ddot{\phi}_i - d_e^{-1}n_i^{-1}\hat{\sigma}_e^2\hat{h}_i$,

$$\ddot{\phi}_i = n_i^{-2}[\hat{\sigma}_e^2 + (n_i^{-1} - c)\hat{w}_i]^2[\hat{m}_i + \hat{k}_i + (n_i^{-1} - c)^2\hat{w}_i]^{-1},$$

and $\hat{r}_i = 1 - (1 - n_i c)\hat{h}_i$. The last three terms of (A.2) are nonnegative and arise from the estimation of the parameter $m_{\cdot\cdot}$, defined by (3.12), and the estimation of the variance, σ_e^2.

[*Received June 1984. Revised July 1987.*]

REFERENCES

Battese, G. E., and Fuller, W. A. (1981), "Prediction of County Crop Areas Using Survey and Satellite Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 500–505.

Cárdenas, M., Blanchard, M. M., and Craig, M. E. (1978), *On the Development of Small Area Estimators Using LANDSAT Data as Auxiliary Information,* Washington, DC: Economics, Statistics, and Cooperatives Service, USDA.

Chhikara, R. S. (ed.) (1984), *Communications in Statistics—Theory and Methods,* 13, No. 23 (special issue on crop surveys using satellite data).

Fuller, W. A., and Battese, G. E. (1973), "Transformations of Estimation of Linear Models With Nested-Error Structure," *Journal of the American Statistical Association,* 68, 626–632.

——— (1981), "Regression Estimation for Small Areas," in *Rural America in Passage: Statistics for Policy,* eds. D. M. Gilford, G. L. Nelson, and L. Ingram, Washington, DC: National Academy Press, pp. 572–586.

Fuller, W. A., and Harter, R. M. (1987), "The Multivariate Components of Variance Model for Small Area Estimation," in *Small Area Statistics: An International Symposium,* eds. R. Platek, J. N. K. Rao, C. E. Särndal, and M. P. Singh, New York: John Wiley, pp. 103–123.

Hanuschak, G., Sigman, R., Craig, M., Ozga, M., Luebbe, R., Cook, P., Kleweno, D., and Miller, C. (1979), "Obtaining Timely Crop Area Estimates Using Ground-Gathered and LANDSAT Data," Technical Bulletin 1609, Washington, DC: Economics, Statistics, and Cooperatives Service, USDA.

Harter, R. M. (1983), "Small Area Estimation Using Nested-Error Models and Auxiliary Data," unpublished Ph.D. thesis, Iowa State University, Dept. of Statistics.

Harville, D. A. (1976), "Extension of the Gauss–Markov Theorem to Include the Estimation of Random Effects," *The Annals of Statistics,* 4, 384–395.

——— (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association,* 72, 320–338.

——— (1979), "Some Useful Representations for Constrained Mixed-Model Estimation," *Journal of the American Statistical Association,* 74, 200–206.

——— (1985), "Decomposition of Prediction Error," *Journal of the American Statistical Association,* 80, 132–138.

Henderson, C. R. (1975), "Best Linear Unbiased Estimation and Prediction Under a Selection Model," *Biometrics,* 31, 423–447.

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980), *SUPER CARP* (6th ed.), Iowa State University, Survey Section, Statistical Laboratory.

Hung, H.-M., and Fuller, W. A. (1987), "Regression Estimation of Crop Acreages With Transformed Landsat Data as Auxiliary Variables," *Journal of Business & Economic Statistics,* 5, 475–482.

Kackar, R. N., and Harville, D. A. (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association,* 79, 853–862.

Peixoto, J. L., and Harville, D. A. (1986), "Comparisons of Alternative Predictors Under the Balanced One-Way Random Model," *Journal of the American Statistical Association,* 81, 431–436.

Reinsel, G. (1984), "Estimation and Prediction in a Multivariate Random Effects Generalized Linear Model," *Journal of the American Statistical Association,* 79, 406–414.

——— (1985), "Mean Squared Error Properties of Empirical Bayes Estimators in a Multivariate Random Effects General Linear Model," *Journal of the American Statistical Association,* 80, 642–650.

Särndal, C. E. (1984), "Design-Consistent Versus Model-Dependent Estimation for Small Domains," *Journal of the American Statistical Association,* 79, 624–631.

Sigman, R. S., Hanuschak, G. A., Craig, M. E., Cook, P. W., and Cárdenas, M. (1978), "The Use of Regression Estimation With LANDSAT and Probability Ground Sample Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 165–168.

Walker, G., and Sigman, R. (1982), "The Use of LANDSAT for County Estimates of Crop Areas: Evaluation of the Huddleston–Ray and the Battese–Fuller Estimators," SRS Staff Report AGES-820909, USDA, Washington, DC.