A Simple Small Area Procedure to Incorporate Two Years of Survey Data with an Application to Estimation of County-Level Cash Rental Rates

Emily Berg* William Cecere† Malay Ghosh‡

Abstract

Information on local cash rental rates is useful for farmers (in determining rental agreements) and policy-makers (in evaluating programs such as the Conservation Reserve Program). The National Agricultural Statistics Service (NASS) conducts an annual survey to obtain estimates of average cash rental rates at the county level. Because sample sizes for many counties are too small to support reliable direct estimators, NASS is investigating model-based procedures. Properties of the data from the Cash Rent Survey lead to several challenges in model development. For example, estimated variances are correlated with estimated means, and the sample distributions have outliers. Also, models must be developed for a wide range of situations. We describe a model-based estimation procedure that incorporates auxiliary information and historical survey data, and present preliminary results.

Key Words: Small area estimation, agricultural survey, EBLUP

1. Introduction

Estimates of cash rental rates serve an important role in US agriculture. Producers landowners, lenders, and appraisers use the estimates as one source of information when they negotiate cash rental rates for specific situations (Dhuyvetter and Kastens, 2010). Policy makers use the estimates for guidance in the administration of state and federal programs. For example, county-level cash rental rates are of interest to the Farm Service Agency because of their implicatons for the Conservation Reserve Program, a voluntary program that provides farmers with monetary incentives to preserve (instead of farm) their land.

Because of the importance of cash rental rates to US agriculture, the 2008 Farm Bill requires the National Agricultural Statistics Service (NASS) to conduct an annual cash rent survey. The objective of NASS's Cash Rent Survey is to obtain estimates of average cash rental rates for counties with at least 20,000 acres of cropland or pastureland. The cash rent data are collected for three different categories: irrigated cropland, nonirrigated cropland, and permanent pasture. Because of differences between the characteristics of the three types of land uses, NASS estimates cash rental rates for the three groups separately.

NASS is investigating model-based alternatives to direct estimation because estimated variances of the direct survey estimators are large for some counties. The main objective of a model-based approach is to obtain efficient estimators of county-level cash rental rates for the three land uses. A reasonable mean squared error estimator is also desired. Computational simplicity is important because of the relatively short time frame for processing the data. In addition to the data from the annual Cash Rent Survey, auxiliary data related to yields, soil productivity, and the value of agricultural products are available.

The model-based procedure that we propose uses survey data for two years and available auxiliary information to obtain estimates of county-level cash rental rates. The general approach, described in Section 4, is to specify two separate univariate models:

*Iowa State University, Ames, IA 50011

†National Agricultural Statistics Service, Fairfax, VA, 22030

‡University of Florida, Gainesville, FL, 32611

one for the average of the two years and one for the difference between the two years. Predictors for the two individual years are obtained by adding or subtracting the predictor of the average to one half of the predictor of the difference. An estimator of the mean squared error is obtained under an assumption that the predictor of the average and the predictor of the difference are uncorrelated.

We illustrate the proposed methodology using the data for 2010 and 2011 for South Dakota. We describe properties of the survey data and the auxiliary information in Section 2 and Section 3, respectively. In Section 4, we describe the proposed model-based procedure in detail, and we explain how we approach the main challenges that we encountered when implementing this procedure. We present results in Section 5, and we conclude in Section 6 with a summary and areas for future work.

2. NASS Cash Rent Survey

The NASS Cash Rent Survey uses a stratified design. The sample size is approximately 224,000 agricultural operations. The questionnaire consists of two main sections. First, a sampled operation is asked to report whether or not any land was rented for cash. If land was rented for cash, the operation is then asked to report the acres rented and the cash rent per acre or total dollars paid.

NASS computes direct estimators as weighted sums of the sampled units. The weight is the product of the inverse of the selection probability and the inverse of the response rate. Figure 1 shows the direct estimators for South Dakota counties, with 2011 direct estimators on the vertical axis and the 2010 direct estimators on the horizontal axis. The three colors distinguish the three different land uses. The sample correlations between the direct estimators for the two years for nonirrigated cropland, irrigated cropland, and permanent pasture (given in the legend of the plot), are 0.99, 0.66, and 0.95, respectively. The relatively high correlations suggest that we may be able to improve the estimator for the current year by taking advantage of the information in the data from the previous year.

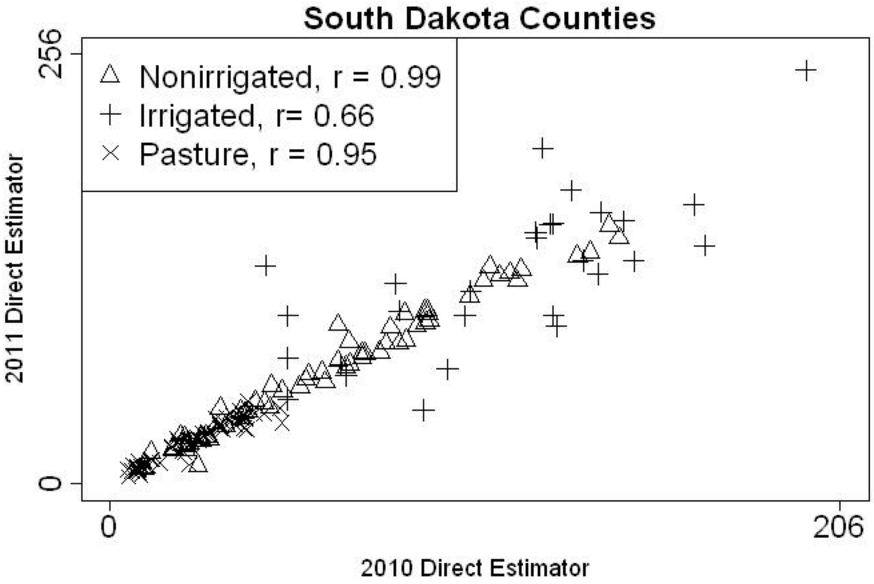

Figure 1: Direct estimators of county-level average cash rental rates for South Dakota for three different land uses.

In addition to a direct estimator of the cash rental rate, NASS computes a direct estimator of the design variance, or the sampling variance, using a jackknife procedure. In Figure 2, the estimated coefficients of variation (CVs) based on the direct estimators are plotted on the vertical axis with the realized sample sizes on the horizontal axis. The largest estimated CV is approximately 50%, and none of the estimated CVs is below 4.5%. The realized sample sizes range from 2 for irrigated cropland to almost 80 for nonirrigated cropland. As we expect, the estimated CVs decrease as the sample size increases. Another property of the estimated CVs is that the variability among the estimated CVs tends to decrease as the realized sample size increases. We discuss our use of generalized variance functions to obtain a variance estimator with a smaller variance than the direct variance estimator in Section 5.

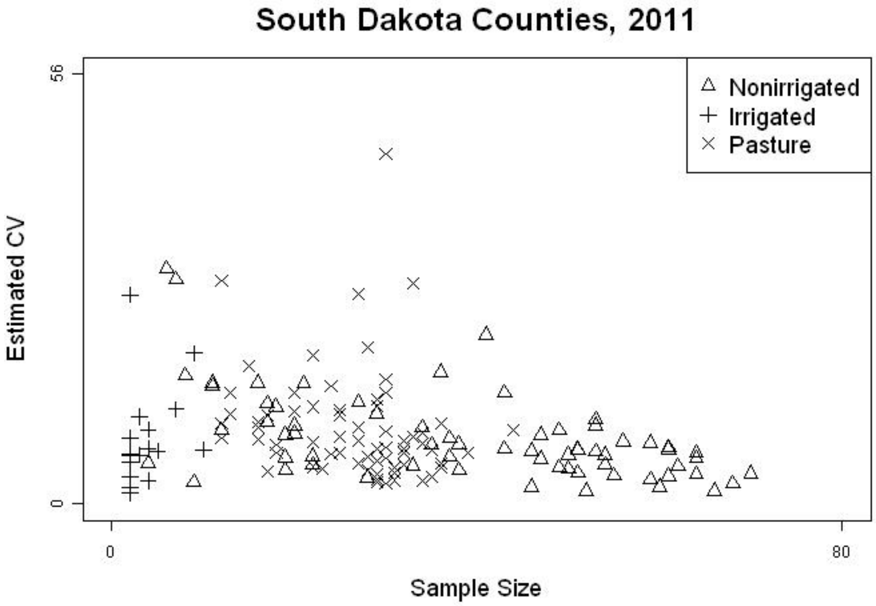

Figure 2: Estimated CVs (y-axis) and realized sample sizes (x-axis) for 2011 in South Dakota counties.

3. Auxiliary data

The 2007 Census of Agriculture is one source of auxiliary information. The 2007 Census provides an estimate of the total dollar value of agricultural production (TVP) in a county. We use the county-level TVP from the 2007 Census as a potential covariate.

The second set of potential covariates consists of four yield indexes, which are based on NASS's published county yields from 2004 through 2009. For states where yields are split by irrigated and nonirrigated cropland, two separate indexes are computed: one based on yields for irrigated cropland and a second based on yields for nonirrigated cropland. A total yield index, which includes yields from irrigated and nonirrigated cropland, is also computed for all states. We excluded crops involving hay from the irrigated, nonirrigated, and total yield indexes and computed a separate hay yield index. Appendix 1 describes the construction of the yield indexes in detail.

The third set of covariates are three National Commodity Crop Productivity Indexes (NCCPI-corn, NCCPI-wheat, NCCPI-cotton), which were developed by the Natural Resources Conservation Service (NRCS). The indexes reflect the quality of the soil for growing nonirrigated crops. The NCCPI-corn, NCCPI-wheat, and NCCPI-cotton are associated with climate conditions best suited for growing corn, wheat and cotton, respectively, on non-irrigated land. The indexes are originally measured at the "map-unit" level, which is a geographic region smaller than a county. County-level indexes are weighted averages of the map-unit indexes, where the weights reflect the acres associated with a map unit.

Table 1 contains the correlations between the average of the 2010 and 2011 direct estimators for South Dakota and the potential covariates. For each covariate, the correlation for irrigated cropland is smaller than the correlation for the other two land uses. For nonirrigated cropland, the total yield index has the largest correlation with the average of the direct estimators, and for irrigated cropland, the hay yield index has the largest correlation. (Both correlations are roughly 0.92.) For irrigated cropland, the total yield index has the largest correlation with the average of the direct estimators, and the correlation is 0.66. We do not compute separate irrigated and nonirrigated yield indexes for South Dakota. The soil productivity index associated with climate conditions suitable for growing cotton is irrelevant for South Dakota because South Dakota is not a cotton-producing state.

	Total Value	Yield Indexes		Soil Productivity	
Land Use	of Production	Total	Hay	Corn	Wheat
Nonirrigated	0.656	0.926	0.875	0.885	0.240
Pasture	0.683	0.864	0.919	0.847	0.349
Irrigated	0.316	0.658	0.603	0.639	0.160

Table 1: Correlations between direct estimators of average cash rental rates and potential covariates for South Dakota counties.

4. A Model Based Approach

To motivate the model-based approach, we discuss two relationships between the average cash rental rate (level) and the change in the cash rental rate over time. First, the cash rental rate for a single year can be written as the sum of the average for two years and one half of the difference between the two years. Letting $\theta_{i,t}$ be the true cash rental rate for county i and year t,

$$(\theta_{i,t-1}, \theta_{i,t}) = (\theta_i, \theta_i) + (-0.5\Delta_i, 0.5\Delta_i), \tag{1}$$

where $(\theta_i, \Delta_i) = (0.5(\theta_{i,t} + \theta_{i,t-1}), \theta_{i,t} - \theta_{i,t-1})$. The expression for $\theta_{i,t}$ in (1) shows that one way to obtain a predictor for time t is to add a predictor of the average for two years to one-half of a predictor of the difference between the two years.

To describe the second relationship between level and change, let $\hat{y}_{i,t}$ be the the direct estimator for county i in year t. If the variance of the direct estimator for time t is equal to the variance of the direct estimator for time $t-1$, then the direct estimator of the mean and the direct estimator of the difference are uncorrelated. In Figure 3, the realized sample size for 2010 is plotted against the realized sample size for 2011. The correlations between the realized sample sizes for the two years are 0.88, 0.94, and 0.79 for irrigated cropland, nonirrigated cropland, and permanent pasture, respectively. The high correlations between the sample sizes for the two years suggest that an assumption that the sampling errors in the direct estimators for the two years are approximately uncorrelated is reasonable.

These two relationships between level and change motivate us to specify separate univariate area-level models for the average and the difference. Under an assumption that

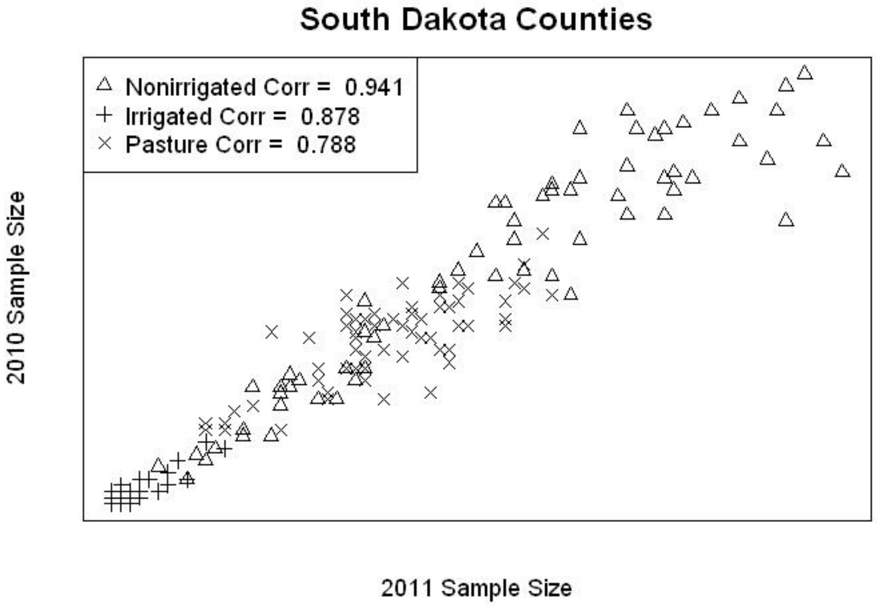

Figure 3: Realized sample sizes for 2010 (y-axis) and 2011 (x-axis) for three land uses in South Dakota counties.

the predictors of the average and the difference are approximately uncorrelated, an estimator of the mean squared error of the predictor for time t can be obtained by adding an estimator of the varaince of the average to one-fourth of an estimator of the variance of the difference. Details are presented below.

4.1 Area Level Models for Averages and Differences

We specify separate models for each state and land use (irrigated, nonirrigated, and pasture), so we omit subscripts for state and land use in the model specification. Letting $\hat{y}_i = 0.5(\hat{y}_{i,t} + \hat{y}_{i,t-1})$, a univariate area-level model for the average is,

$$\begin{aligned} \hat{y}_i &= \theta_i + e_i \\ \theta_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \end{aligned} \tag{2}$$

where $(u_i, e_i)' \sim \{\mathbf{0}, \text{diag}(\sigma_u^2, \sigma_{ei,avg}^2)\}$. Likewise, an area-level model for the difference is

$$\begin{aligned} \hat{d}_i &= \Delta_i + \eta_i \\ \Delta_i &= \boldsymbol{z}_i'\boldsymbol{\beta}_d + v_i \end{aligned} \tag{3}$$

where $(v_i, \eta_i) \sim \{\mathbf{0}, \text{diag}(\sigma_v^2, \sigma_{\eta i,diff}^2)\}$ and $\hat{d}_i$ is a direct estimator of the difference between the cash rental rates for county i in years t and $t-1$. In (2) and (3), $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are vectors of covariates, σ_u^2 and σ_v^2 represent variability between counties unexplained by the covariates, and $\sigma_{ei,avg}^2$ and $\sigma_{\eta i,diff}^2$ are estimates of the variances of the sampling errors in the direct estimators. The estimates of $\sigma_{ei,avg}^2$ and $\sigma_{\eta i,diff}^2$ are based on a generalized variance function. (See Section 5.) We treat the estimates of the variances of the sampling errors as fixed constants for estimation, which is standard practice in small area estimation. See, for example, Fay and Herriot (1979) and Rao (2003).

4.2 Predictors and MSE Estimators

The EBLUPs of the average and the difference are, respectively,

$$\hat{\theta}_i = \hat{\gamma}_{i,avg}\hat{y}_i + (1 - \hat{\gamma}_{i,avg})\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}, \tag{4}$$

and

$$\hat{\Delta}_i = \hat{\gamma}_{i,diff}\hat{d}_i + (1 - \hat{\gamma}_{i,diff})\boldsymbol{z}_i'\hat{\boldsymbol{\beta}}_d, \tag{5}$$

where $\hat{\gamma}_{i,avg} = (\hat{\sigma}_u^2 + \hat{\sigma}_{ei,avg}^2)^{-1}\hat{\sigma}_u^2$, and $\hat{\gamma}_{i,diff} = (\hat{\sigma}_v^2 + \hat{\sigma}_{\eta i,diff}^2)^{-1}\hat{\sigma}_v^2$. We use (27) - (32) of Wang and Fuller (2008) to estimate $\boldsymbol{\beta}$, $\boldsymbol{\beta}_d$, σ_u^2, and σ_v^2. Estimators of the MSEs of the average and the difference are, respectively,

$$M\hat{S}E_{i,avg} = \hat{\gamma}_{i,avg}\hat{\sigma}_{ei,avg}^2 + (1 - \hat{\gamma}_{i,avg})^2\boldsymbol{x}_i'\hat{\boldsymbol{V}}\{\hat{\boldsymbol{\beta}}\}\boldsymbol{x}_i, \tag{6}$$

and

$$M\hat{S}E_{i,diff} = \hat{\gamma}_{i,diff}\hat{\sigma}_{\eta i,diff}^2 + (1 - \hat{\gamma}_{i,diff})^2\boldsymbol{z}_i'\hat{\boldsymbol{V}}\{\hat{\boldsymbol{\beta}}_d\}\boldsymbol{z}_i, \tag{7}$$

where

$$\hat{\boldsymbol{V}}\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{i=1}^{m}\boldsymbol{x}_i(\hat{\sigma}_u^2 + \hat{\sigma}_{ei,avg}^2)^{-1}\boldsymbol{x}_i'\right)^{-1} \tag{8}$$

and

$$\hat{\boldsymbol{V}}\{\hat{\boldsymbol{\beta}}_d\} = \left(\sum_{i=1}^{m}\boldsymbol{z}_i(\hat{\sigma}_v^2 + \hat{\sigma}_{\eta i,diff}^2)^{-1}\boldsymbol{z}_i'\right)^{-1}. \tag{9}$$

Predictors of the mean cash rental rates for the two time points are,

$$(\hat{\theta}_{i,t-1}, \hat{\theta}_{i,t}) = (\hat{\theta}_i, \hat{\theta}_i) + (0.5\hat{\Delta}_i, -0.5\hat{\Delta}_i). \tag{10}$$

Under an assumption that $\hat{\theta}_i$ and $\hat{\Delta}_i$ are uncorrelated, an estimator of the MSE of the predictor of $\theta_{i,t}$ is

$$M\hat{S}E_{i,t} = M\hat{S}E_{i,avg} + 0.25M\hat{S}E_{i,diff}. \tag{11}$$

The MSE estimator (11) does not account for variability in the estimators of variances.

4.3 Benchmarked Predictors

In addition to the county estimates, NASS publishes estimates of state-level cash rental rates for the three land uses (irrigated, nonirrigated, and pasture). The published state-level estimates are based on data from NASS's June Area Survey in addition to data from the Cash Rent Survey. The state-level estimates are published before the county-level estimates are complete. Therefore, we impose the benchmarking restriction,

$$\sum_{i=1}^{m}\hat{w}_{i,t}\hat{y}_{i,t} = \hat{y}_{t,pub}, \tag{12}$$

where $\hat{y}_{t,pub}$ is the published estimate of the state-level cash rental rate,

$$\hat{w}_{i,t} = \hat{a}_{i,t}(\sum_{i=1}^{m}\hat{a}_{i,t})^{-1},$$

and $\hat{a}_{i,t}$ is the direct estimator of the acres rented in county i and year t. In addition to county and state estimates, NASS publishes estimates for agricultural statistics districts, groups of spatially contiguous counties. We use the two-stage benchmarking procedure of Ghosh and Steorts (2012) to obtain county-level and district-level estimates that satisfy the restriction (12).

4.4 Covariates

One set of challenges stems from the nature of the covariates described in Section 3. The availability of the covariates differs across states and counties. For example, we do not have sufficient data for an irrigated yield index in South Dakota, the TVP is unavailable for some counties in Texas, and the NCCPI is missing for some counties in Colorado. Also, several of the covariates are highly correlated. The correlation between the hay yield index and the total yield index for South Dakota is 0.89. Another problem associated with using multiple covariates in a linear model is that estimates of the average cash rental rates can be negative.

As an initial solution to these challenges, we combined the multiple covariates into a single index. The single index is a weighted average of the covariates, where the weights are larger for covariates that are more highly correlated with the average cash rental rate. All covariates are positive, and the single index is defined to be strictly positive. The model for the average cash rental rate is the simple linear model,

$$\theta_i = \beta_0 + \beta_1 x_i + u_i$$

where x_i is the covariate index. The estimate of β_0 is restricted to be nonnegative. A complete definition of the index is given in Appendix 2.

For nonirrigated cropland, the R^2 from a simple linear regression of the $\hat{y}_i$ on the four covariates in Table 1 is 0.89. The corresponding R^2 from a regression with an intercept and the covariate index is 0.83. Because the R^2 does not account for the degrees of freedom, we consider the reduction in the R^2 due to the use of the single index instead of the individual covariates to be small relative to the gain in simplicity and the guarantee of a reasonable predicted value. Similar results for other states suggest that little information is lost by using the single index instead of the individual covariates.

4.5 Outliers

A second challenge that arose when implementing the general approach of Sections 4.1 and 4.2 is that residuals have extreme values relative to a normal distribution. In Figure 4, standardized residuals from an initial model that was fit to the direct estimators of the differences are plotted against the theoretical quantiles of a normal distribution. The plot shows that the residuals have a heavy-tailed distribution. We used the method described in Appendix 3 to reduce the effects of extreme values on the estimators of the parameters.

4.6 Sampling Variances

Estimation of sampling variances is a third challenge. As discussed in Section 2 (see Figure 2), the direct estimators of the sampling variances may have large variances. In addition, the direct estimators of the sampling standard deviations are correlated with the direct estimators of the cash rental rates. For South Dakota, the correlation between the 2011 direct estimators of cash rental rates for nonirrigated cropland and the direct estimators of the sampling standard deviations is 0.80. A high correlation between the direct estimators of the means and the direct estimators of the variances has been documented to lead to a bias in the estimators of the regression coefficients. See Carroll and Ruppert (1988) and references cited there.

We specify a model for the variances to reduce the correlation between the estimated means and the estimated variances and to reduce the variances of the estimated variances. Our variance model assumes that the direct estimators of the standard deviations

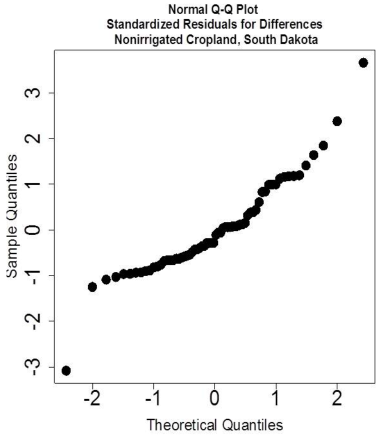

Figure 4: Normal quantile-quantile plot of residuals in an initial model fit to the direct estimators of the differences.

are linear in the covariate index. We set the intercept in the linear model equal to zero if necessary to avoid negative estimates of standard deviations.

5. Results

Table 2 shows the distributions of the ratios of the estimated MSEs of the model-based predictors to the estimates of the variances of the sampling errors for South Dakota. The true MSE ratios may be somewhat larger than the estimated ratios due to the effects of benchmarking and estimating variances. Nonetheless, the MSE ratios are uniformly smaller than one, suggesting that the use of historical data and covariates is effective in improving the quality of the predictors.

Land use	Min	25%	Median	Mean	75%	Max
Nonirrigated	0.37	0.58	0.63	0.62	0.67	0.77
Pasture	0.41	0.55	0.58	0.59	0.62	0.90
Irrigated	0.26	0.38	0.41	0.46	0.56	0.78

Table 2: Distribution of ratios of estimated MSEs of model-based predictors to estimated sampling varainces for nonirrigated cropland in South Dakota.

6. Concluding Remarks

We developed a small area procedure that incorporates survey data for two years and auxiliary data. The procedure is based on two separate univariate area-level models: one for the average of the two years and one for the difference between the two years. Predictors of the two time means are linear combinations of the predictors of the averages and the differences. A simple approximation for the mean squared error is obtained using an assumption that estimators of averages and differences are approximately uncorrelated. Initial results for the 2011 data for South Dakota support the proposed procedures. The estimated MSEs of the predictors for South Dakota are uniformly smaller than the estimated sampling variances. The computational simplicity of the proposed procedure is important for production because estimates must be processed in a relatively short time frame.

Outliers, correlated covariates, and estimated sampling variances create challenges in the implementation of this general procedure. We developed simple methods to deal with these challenges. More innovative solutions, such as explicit models for outliers and nonlinear models for variances, provide areas for future innovations. Other potential areas for future work include incorporating more than two years of data, examining spatial structure of the model residuals, refining the MSE estimator to account for estimated variances, and evaluating the goodness of fit of the model.

REFERENCES

Carroll, R. and Ruppert, D. (1988), *Transformation and Weighting in Regression*, CRC Press.

Dhuyvetter and Kastens (2010), “2009 Kansas Land Values and Cash Rents,” http://www.agmanager.info/farmmgt/land/county/CountyValues&Rents(Sep2009)-Revised(Oct2011).pdf.

Ghosh and Steorts (2012), ”Two-stage Bayesian benchmarking as applied to small area estimation,” Submitted.

Appendix 1: Yield Covariates

The published crop yields for a state fall into at most four categories: irrigated, nonirrigated, total for crop, and hay. The hay yield index incorporates the commodities, “Hay all (Dry),” “Hay Alfalfa (Dry),” and “Hay Other (Dry).” For a subset of states, NASS publishes separate estimates of yields for irrigated and nonirrigated cropland. For these states, we define two separate indexes, one based on yield estimates for irrigated cropland and the other based on yield estimates for nonirrigated cropland. For all states, the “total for crop” category includes estimates of yields for both irrigated and nonirrigated cropland. We define a “total for crop” yield index based on estimates of crop yields that include both irrigated and nonirrigated cropland. The commodities included in the hay yield index are excluded from the irrigated, nonirrigated, and total for crop yield indexes. We use the published yields for the four categories (nonirrigated, irrigated, total for crop, and hay) to construct at most four yield indexes for each state.

Because we construct a separate yield covariate for groups defined by combinations of states and the four categories (irrigated, nonirrigated, total for crop, and hay), we do not use subscripts for states or categories in the definition of the yield covariate. Let $i = 1, \ldots, m$ denote the counties in a state, and let $s = t-5, \ldots, t-1$ denote the years. For example, if we are estimating cash rental rates for 2010 and 2011, then $t - 5 = 2005$, and $t - 1 = 2010$. Let $c = 1, \ldots, C$ index the crops with a published yield estimate for at least one county in the state in at least one of the years. Let

$$h_c = \sum_{t=1}^{T} \sum_{i=1}^{m} h_{ict} \tag{13}$$

where h_{ict} is the harvested acres published for county i and year t for commodity c.

Let $h_{[1]}, \ldots, h_{[C]}$ be the published estimates of the harvested acres at the state level for the C commodities listed in increasing order, and let $S_k = \sum_{j=0}^{k} h_{[C-j]}$. The number of commodities needed to cover 95% of the harvested acres in the state is,

$$\ell = \min\{k : S_k (S_{[C]})^{-1} > 0.95\}. \tag{14}$$

The crops used to construct the yield covariate for a particular state and category are the commodities associated with the order statistics $C - \ell, \ldots, C$.

For $C-\ell, \ldots, C$, let z_{ics} be the published yield for county i, year s, and commodity c. The z_{ics} may be missing for some i or s. Let $\bar{z}_{ic}$ be the average of the $\{z_{ics} : s =$

$t-5, \ldots, t-1,$ and z_{ics} is not missing$\}$. If z_{ics} is not missing for at least four of the years, then the largest and smallest yields are omitted from the average.

Let $t_{ic} = s_c^{-1}(\bar{z}_{ic} - \bar{z}_c)$, where $\bar{z}_c$ and s_c are the average and standard deviation, respectively, of the $\bar{z}_{ic}$ across the counties where $\bar{z}_{ic}$ is not missing. Let

$$t_i = \max\{t_{ic} : c = C - \ell, \ldots, C \text{ and } t_{ic} \text{ is not missing}\}. \tag{15}$$

The yield covariate for county i for a particular state and category is $\tilde{t}_i$, where

$$\tilde{t}_i = t_i - \min[\min\{t_i : i = 1, \ldots, m\}, 0]. \tag{16}$$

We subtract the minimum of the t_i if the minimum of the t_i is negative to create a yield index with positive values.

We use one of the four possible yield indexes as the yield covariate according to the following rules. If we are estimating cash rental rates for (non)irrigated cropland and the state publishes separate estimates of yields for irrigated and nonirrigated cropland, then we use the (non)irrigated yield index. If we are estimating cash rental rates for irrigated or nonirrigated cropland and the state does not publish separate estimates for irrigated and nonirrigated crop yields, then we use the total for crop yield index. If we are estimating cash rental rates for permanent pasture, and the state has enough hay yields to compute a hay yield index, then we use the hay yield index. If the we are estimating cash rental rates for permanent pasture and the state does not have enough hay yield data to construct a hay yield index, then we use the total for crop yield index.

Appendix 2: Construction of the Single Index to Use as the Covariate

The possible covariates for a state are combined to form a single covariate index. The use of a single index avoids potential problems associated with automated selection procedures and negative estimates. The covariate index is a weighted average of the available covariates for a state with higher weights assigned to covariates that are more highly correlated with the estimates of the average cash rental rates. The covariate index is defined to be strictly positive and have a range similar to the range of the average cash rental rates so that we can set the intercept to be zero to guarantee positive predicted values if necessary.

Each of our covariates is missing for at least one county where an estimate of a cash rental rate is desired. If a covariate is missing for county i, we impute the average of the covariate values across the counties that are in the same agricultural statistics district as county i. If the covariate is missing for all counties in the agricultural statistics district, then we impute the average value of the covariate across the counties in a state.

We define a separate index for each state. Let x_{di}, $d = 1, \ldots, D$ be the set of initial covariates, and assume that $x_{di} > 0$. Let

$$(\bar{y}, S_y^2) = \left(m_1^{-1} \sum_{i=1}^{m_1} \hat{y}_i, (m_1 - 1)^{-1} \sum_{i=1}^{m_1} (\hat{y}_i - \bar{y})^2 \right), \tag{17}$$

where m_1 is the number of counties where the direct estimator of the average cash rental rate is not missing. Similarly, let

$$(\bar{x}_d, S_{x,d}^2) = \left(m_2^{-1} \sum_{i=1}^{m_2} x_{di}, (m_2 - 1)^{-1} \sum_{i=1}^{m_2} (x_{di} - \bar{x}_d)^2 \right) \tag{18}$$

where m_2 is the number of counties where x_{di} is not missing. Define

$$x_{di}^* = \bar{y} + \left(\min\left\{ \frac{\bar{y}}{\bar{x}_d}, \frac{S_y}{S_{x,d}} \right\} \right) (x_{di} - \bar{x}_d). \tag{19}$$

To see that $x^*_{di} > 0$, note that if $\bar{y}S_{x,d} < S_y\bar{x}_d$, then $x^*_{di} = \bar{x}_d^{-1}\bar{y}$. If $\bar{y}S_{x,d} > S_y\bar{x}_d$, then $\bar{y}S_{x,d} + S_y x_{di} > S_y\bar{x}_d$ (by the assumption that $x_{di} > 0$). Subtracting $S_y\bar{x}_d$ from both sides and dividing through by S_x shows that $x^*_{di} > 0$.

Let ρ_{dy} be the sample correlation between x^*_{di} and $\hat{y}_i$. The covariate index is,

$$x_i = \frac{\sum_{d=1}^{D} w_d x^*_{di}}{\sum_{d=1}^{D} w_d}, \tag{20}$$

where

$$w_d = \max\{\rho_{dy}, 0.1\}. \tag{21}$$

Appendix 3: Outliers in Estimates of Differences

Because of the large amount of data to be processed, we use a simple method to adjust for outliers. Let $\tilde{d}_i^{(0)}$ be an initial direct estimator of the difference for county i. For $t = 0, 1, \ldots$, let $m_{yd}^{(t)}$ and $S_{yd}^{(t)}$ be the median and sample standard deviation, respectively, of $\tilde{d}_i^{(t)}$ $(i = 1, \ldots, m)$. If $\tilde{d}_i^{(t)} > m_{yd}^{(t)} + 2.33S_{yd}^{(t)}$, then set $\tilde{d}_i^{(t+1)} = m_{yd}^{(t)} + 2.33S_{yd}^{(t)}$. If $\tilde{d}_i^{(t)} < m_{yd}^{(t)} - 2.33S_{yd}^{(t)}$, then set $\tilde{d}_i^{(t+1)} = m_{yd}^{(t)} - 2.33S_{yd}^{(t)}$. Otherwise, set $\tilde{d}_i^{(t+1)} = \tilde{d}_i^{(t)}$. Let $\hat{d}_i = \tilde{d}_i^{(4)}$, the result of four steps of the iteration.