A Problem in Small Area Estimation: Cash Rental Rates

William Cecere[1], Emily Berg[1], Malay Ghosh[2]
[1]National Agricultural Statistics Service USDA, [2]University of Florida

Abstract

The USDA National Agricultural Statistics Service (NASS) conducts an annual Cash Rents survey to estimate county level cash rental rates for three land use categories (non-irrigated cropland, irrigated cropland, and pasture). A cash rent is land rented on a per acre basis for cash only. Estimates of cash rental rates are useful to farmers in determining rental agreements, economists in studying research questions, and policy makers in computing payment rates for the Conservation Reserve Program. A major issue is that survey indications for counties are unstable due to small realized sample sizes. We investigate the use of mixed models to obtain reliable estimates of average cash rental rates at the county level.

Key Words: mixed models, small area estimation, empirical bayes, benchmarking

1. Introduction

The National Agricultural Statistics Service (NASS) conducts hundreds of surveys each year to obtain estimates related to diverse aspects of US agriculture. NASS's large scale surveys produce reliable estimates of agricultural characteristics at national and state levels. Estimation for small domains, such as counties, is more difficult due to small realized sample sizes.

The focus of our research is estimation of average cash rental rates for non-irrigated cropland, irrigated cropland, and pastureland at the county level. In a cash rental agreement, a tenant rents cropland or pastureland from a landowner in units of dollars per acre. Cash rental agreements differ from another common setup called share-rental agreements, which involve payments in terms of a share of the produced goods. The tenant in a cash rental agreement is typically responsible for all management decisions, acquires all of the produced goods, and assumes all risk in producing those goods.

NASS estimates of county-level cash rental rates serve many purposes. Farmers use the estimates of local cash rental rates for guidance in determining appropriate rental agreements (Dhuyvetter and Kastens, 2009). Agronomists use the estimates to study research questions related to the interplay between cash rental rates and other economic characteristics such as commodity prices and fuel costs (Woodard et al., 2010). NASS published county-level cash rent estimates have immediate implications for the Conservation Reserve Program, a policy that aims to protect natural resources by providing rental payments to agricultural landowners who choose to conserve their land. Because of the role of cash rental rates in the Conservation Reserve Program, the 2008 Farm Bill required NASS to conduct an annual survey of cash rental rates for nonirrigated cropland, irrigated cropland, and permanent pasture.

To satisfy the requirements of the 2008 Farm Bill, NASS implemented an annual Cash Rent Survey. A concern is that direct estimators of average county rental rates from the Cash Rent Surveys are often unstable due to small realized sample sizes. Ultimately, estimates are desired for the three land use categories (non-irrigated cropland, irrigated cropland, and pastureland) for counties with at least 20,000 acres of cropland or pastureland. Estimates are published at three levels of geographic detail: state, agricultural statistics district, and county. We investigate the use of mixed models (Rao, 2003) to stabilize the estimators of county-level cash rental rates. Model-based procedures were developed using the 2009 and 2010 Cash Rent Survey data, and the methods were later applied to the data from the 2011 Cash Rent Survey.

NASS estimates state-level cash rental rates using a combination of data from a national survey called the June Area Survey and the Cash Rent Survey. The state-level cash rent estimates are considered to be more reliable and are published before county level estimation from the Cash Rent Survey is complete. To maintain internal consistency, it is important that appropriately weighted sums of county predictors are equal to the previously published state estimates. We use the benchmarking procedure proposed by Ghosh and Steorts (2011) to ensure that the county predictors preserve the previously published state estimates.

2. Data for Modeling Cash Rental Rates

2.1 NASS Cash Rent Survey

The main source of data for estimating county-level cash rental rates is the annual NASS Cash Rent Survey. The first section of the questionnaire asks whether or not the operation owned, rented or leased from others, or rented or leased to others. The second section asks if the operation rented cropland or pastureland for cash. The operations that rented cropland or pastureland for cash are asked to report the acres rented and the cash rental rate or the total dollars rented for each of the three categories: non-irrigated cropland, irrigated cropland, or permanent pasture.

The samples for the 2009 and 2010 Cash Rent Surveys were stratified random samples, where the strata where defined according to the total dollars rented reported on previous surveys or the 2007 Census of Agriculture. The stratification changed for the 2011 Cash Rents Survey such that operations with greater acres previously rented had higher selection probabilities. Because the sampling frame used for previous surveys may not cover the whole population, an additional stratum was added for 2011.

A direct survey estimator for a particular land use category is a ratio of a weighted sum of the dollars rented to a weighted sum of acres rented. The weight associated with a respondent is the population size of the stratum containing the respondent divided by the number of responding units in that stratum.

2.2 Auxiliary Information

In an effort to improve the precision of the county-level cash rent estimates, auxiliary variables were desired that would explain the variability among the direct county estimates. Auxiliary information for modeling cash rental rates is available from several sources external to the Cash Rent Survey. After searching through many options of covariates, three were chosen based upon usability and correlation with the item of interest.

The first covariate is the National Commodity Crop Productivity Index (NCCPI), which is developed and maintained by the Natural Resource Conservation Service. The NCCPI consists of three different indexes called NCCPI-corn, NCCPI-cotton, and NCCPI-wheat, which reflect the quality of the soil for growing non-irrigated crops in three different climate conditions. The indexes are constructed at the level of a "map unit," a subset of a county. (User Guide National Commodity Crop Productivity Index (NCCPI). Version 1.0, 2008. (ftp://ftp-fc.sc.egov/usda.gov/NSCC/NCCPI_user_guide.pdf.) The county-level indexes used as covariates are weighted averages of the map-unit values, where the weights are the acres of cropland covered by a map unit. The NRCS also produces a summary index called max-NCCPI. The max-NCCPI is obtained by first taking the maximum of the three commodity-specific indexes for each map unit and then computing a weighted average of the maxima across map units in a particular county. The NCCPI provides four potential covariates: the three commodity specific indexes and the max-NCCPI.

Another measure of the quality of the land in a county is the realized crop yield. NASS publishes estimates of crop yields for a variety of crops in many counties. Defining a covariate based up NASS published yields is challenging because (1) no yields are published for many counties where an estimate of the cash rental rate is desired, and (2) for many states, the published yields for different counties are associated with different crops and different land usages. We attempt to define a yield covariate that overcomes these challenges.

We start by averaging the published yields across five years in an effort to smooth the published yields and reduce the variance. Using an average across years also helps reduce the number of counties with missing yields. Instead of defining a single covariate for each crop, we define a yield index based on multiple crops. This allows us to construct a single yield covariate for relatively diverse state where different crops are grown in different parts of the state. The published yields for a state fall into at most four categories: irrigated, non-irrigated, total for crop, and hay. We use the published yields for the four categories (non-irrigated, irrigated, total for crop, and hay) to construct at most four yield covariates for each state.

One consequence of having yield and NCCPI broken into four separate covariates each is that there is a tendency for high correlation between covariates of similar type. To account for this, we define a single index for both yield and NCCPI so that we can avoid problems resulting from multicollinearity. This single index is constructed by using the weighted average of the original covariates, where the weight is based off of the correlation between a given covariate and the county cash rental rate relative to the other covariates.

Our final source of auxiliary information is the Total Value of Production (TVP). The TVP is a county-level covariate obtained from the 2007 Census of Agriculture and is a measure of the value of goods produced in a county.

3. Cash Rent Modeling

The current NASS procedure for publishing county-level cash rent estimates involves having experts examine the current year survey data and previous year survey data to set estimates. They examine various survey-based indications such as the direct estimator for the current year and the ratio of the estimators for the current and previous year. Our main objective is to develop a model-based estimation procedure that uses historical data, auxiliary data, and data from neighboring counties in a way that is objective and provides a measure of mean squared error. We specify a separate model for each practice: non-irrigated cropland, irrigated cropland, and pasture. We assume that the state estimates for the current year are already determined and are available.

Initial analyses of the data from the 2009 and 2010 Cash Rent Surveys showed significant correlations between the cash rental rates for 2009 and 2010 at both the unit and county levels. The strong correlation suggests that incorporating information from the previous year may improve the estimate for the current year. One challenge in modeling is determining a method for incorporating data from the previous year that is both statistically defensible and computationally feasible.

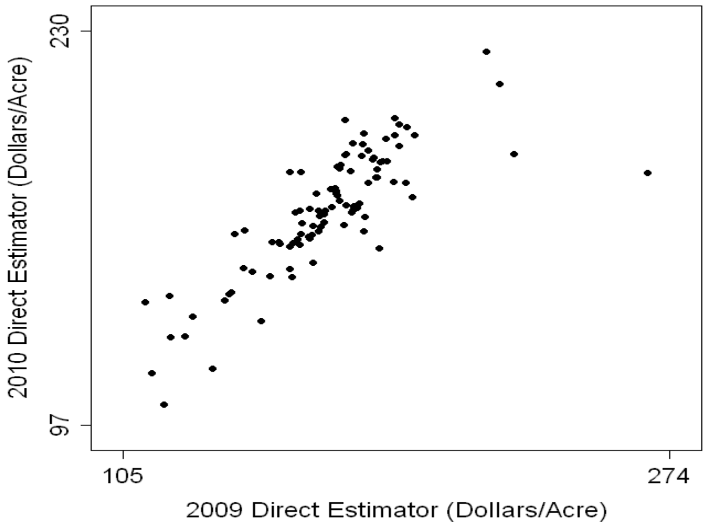

Figure 1: Direct estimates of cash rental rates for non-irrigated cropland in Iowa counties from the 2009 (x-axis) and 2010 (y-axis) Cash Rent Surveys

As an initial effort, we specified a univariate model for the unit-level data. One of the covariates in the unit-level model is the value of the cash rental rate reported in the previous year. This method is computationally feasible but lacks statistical validity. One important problem is that using the cash rent per acre from the previous survey as a covariate in a unit-level model treats the cash rent per acre from the previous year as a fixed value. It is more appropriate to treat the cash rent per acre from the previous year as a random variable. Conditioning on the previous year cash rental rate can lead to mean squared error estimators for predictions that are artificially small. Another undesirable property of the unit-level univariate model with the previous year cash rental rate as a covariate is that only units that report positive dollars and positive acres in both years can be used to estimate model parameters. Using only a subset of the data to estimate the model parameters can be inefficient relative to a procedure that uses all of the available data.

As a second attempt, we specified a unit-level bivariate model. The bivariate unit-level model overcomes the drawbacks associated with the univariate unit-level model: The bivariate model treats the previous year cash rental rate as a random variable, and all of the data from the two years are used to estimate the parameters of the bivariate models. One practical problem associated with the bivariate unit-level model is computation time. We used Gibbs sampling to fit the bivariate model, and the sampling procedure can take several hours to run for a given state and usage.

In this document, we propose a method based on a bivariate area-level model. The procedure incorporates the cash rent per acre from the previous year as a random variable (which overcomes the statistical problems of the univariate unit-level model), and the estimators of model parameters have closed form expressions (which overcomes the practical problems of the bivariate model). The procedure is based on two separate univariate area-level models. One of the univariate models is based on the average of the two time means. The second univariate model is a model for the difference of the two time means. The predictor for the average cash rent per acre for a single year is a sum of an estimate of the average (based on the first model) and an estimate of the change (based on the second model).

3.1 Area-Level Model

Let be the average cash rent per acre for county in year . Let $\hat{y}_{it}$ be the direct estimator or indication of , and assume that direct estimators are available for two consecutive time points. Note that we specify a separate model for each practice, so we do not include a subscript for the practice. Assume

$$\begin{aligned} \hat{y}_{it} &= \theta_{it} + e_{it} \\ \theta_{it} &= x_i^{'}\beta_t + \delta_i + \eta_{it} \end{aligned} \tag{1}$$

Where $\delta_i \sim (0, \sigma_\delta), (e_{i,t-1}, e_{it}) \sim [0, \sigma_{ei}^2, R_{ei}]$, and $(\eta_{i,t-1}, \eta_{it}) \sim [0, \sigma_\eta^2, R_\eta]$, where R_η and R_{ei} are 2x2 correlation matrices with correlation parameters ρ_η and ρ_{ei} respectively. We assume that estimates of σ_{ei}^2 and ρ_{ei} are available from the unit-level data.

Let $\hat{y}_i = 0.5(\hat{y}_{i,t-1} + \hat{y}_{it})$ and $\hat{d}_i = (\hat{y}_{it} - \hat{y}_{i,t-1})$. From (1),

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

where $\beta = 0.5(\beta_{t-1} + \beta_t), u_i = \delta_i + 0.5(\eta_{it-1} + \eta_{it})$, and $e_i = 0.5(e_{it-1} + e_{it})$. Similarly,

$$\begin{aligned} \hat{d}_i &= \Delta_i + \eta_i \\ \Delta_i &= x_i^{'}\beta_d + v_i \end{aligned} \tag{3}$$

Where $\beta_d = (\beta_t - \beta_{t-1}), v_i = \delta_{it} - \delta_{it-1}$, and $\eta_i = (e_{it} - e_{i,t-1})$. The equations in (2) specify a univariate model for the average of the direct estimators for the two time points, and (3) is a univariate model for the difference. Because the variances for the two time points are assumed to be equal in (1), $(\hat{y}_i, \theta_i)$ is uncorrelated with $(\hat{d}_i, \Delta_i)$. One can obtain predictors of θ_i and Δ_i along with associated MSE estimators using standard methodology for univariate area-level models. (Rao, 2003, Chapter 7.) If $\hat{\theta}_i$ and $\hat{\Delta}_i$ are unbiased estimates of θ_i and Δ_i, then

$$\hat{\theta}_{it} = \hat{\theta}_i + 0.5(\hat{\Delta}_i) \tag{4}$$

is an unbiased estimator predictor of θ_{it}. If, in addition, $\hat{\theta}_i - \theta_i$ and $\hat{\Delta}_i - \Delta_i$ are uncorrelated, then an estimate of the MSE of $\hat{\theta}_{it}$ is

$$M\hat{S}E(\hat{\theta}_{it}) = M\hat{S}E(\hat{\theta}_i) + 0.25M\hat{S}E(\hat{\Delta}_i), \tag{5}$$

where $M\hat{S}E(\hat{\theta}_i)$ and $M\hat{S}E(\hat{\Delta}_i)$ are estimates of the MSEs of $\hat{\theta}_i$ and $\hat{\Delta}_i$, respectively.

We assume that the models (2) and (3) hold for direct estimators of the average and difference, respectively. The predictor of the average for county is

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

where $\hat{\gamma}_{i,avg} = (\hat{\sigma}_u^2 + \sigma_{ei,avg}^2)^{-1}\hat{\sigma}_u^2$, and $\hat{\sigma}_u^2$ is an estimator of the variance of u_i of (2) and $\hat{\sigma}_{ei,avg}^2$ is an estimate of the sampling variance in the direct estimator of the average. The sampling variance is represented by the variance of e_i of (2). The predictor of the difference for county *i* is

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

Where $\hat{\gamma}_{i,diff} = (\hat{\sigma}_v^2 + \sigma_{\eta i,avg}^2)^{-1}\hat{\sigma}_v^2$ where $\hat{\sigma}_v^2$ and $\hat{\sigma}_{\eta i,avg}^2$ are estimates of the variances of v_i and η_i of (3). The procedure outlined in (27)-(32) of Wang and Fuller (2008) is used to obtain the estimates, $(\hat{\sigma}_u^2, \hat{\sigma}_v^2, \hat{\beta}, \hat{\beta}_d)$. The mean squared error estimator has the form in (5), the two components of the MSE ignore the variance of the variance estimators.

3.2 Estimation of the Difference

Recall that for an estimate of the average, we start with the simple average of the current and previous years $\bar{y}_i = 0.5(\bar{y}_{it} + \bar{y}_{it-1})$. The equivalent approach to constructing a direct estimator of the difference would be to use the simple difference between the two direct estimates $\bar{y}_{it} - \bar{y}_{it-1}$. In the cash rent application, some units respond in both time points. We define an estimator of the difference that treats units that respond in both time points distinctly from units that respond in only one time point as an effort to obtain an estimator of the difference with a smaller variance than $\hat{y}_{it} - \hat{y}_{it-1}$. To define the estimator of the difference, let index the units within county , and let be the number of units in county that report a cash rental rate in both time points. Let $\tilde{d}_{ij} = y_{ijt} - y_{ijt-1}$, and let d_{ij} be the result of applying a modification for outliers analogous to the method of Appendix B to the $\tilde{d}_{ij}$. An estimator of the difference based on only the observations that respond in both time points is

$$\bar{y}_{di} = (\sum_{j=1}^{n_i} d_{ij}a_{ij})(\sum_{j=1}^{n_i} a_{ij})^{-1},$$

where a_{ij} is the average of the acres rented by unit *(ij)* across years *t-1* and *t* .

The three distinct scenarios with which we can measure change from one year to the next are as follows:

(1) For a given county-usage combination, we have records that responded in both years, responded in the first year and not the second, and responded in the second year and not the first.

(2) For a given county-usage combination, we have no identical respondents in both years, respondents in the first year and not the second, and respondents in the second year and not the first.

(3) For a given county-usage combination, we have only records that responded to both years.

For scenario (2) in which we have no records responding to both years, we set $\hat{d}_i = \bar{y}_{it} - \bar{y}_{it-1}$ as our estimate of $\hat{y}_{di}$. For scenario (3) in which there are only records responding to both years, we use $\bar{y}_{di}$ as our estimate of $\hat{y}_{di}$.

For scenario (1) where there are both records from (2) and (3), we use a weighted combination of $\hat{d}_i$ and $\bar{y}_{di}$ where the weights reflect the relative size for records from scenarios (2) and (3) respectively. The formula and derivations of the estimate of $\hat{y}_{di}$ for scenario (1) can be found in Appendix A.

3.3 Estimation of Sampling Variances

In the discussion of the small area models in section 3.1, we assumed that we are starting with an estimate of the sampling variance of the direct estimators of the average and the difference. NASS methodologists compute a variance estimate of the direct estimator of average cash rental rate for each year using a jackknife procedure. The jackknife estimates of the variances are design unbiased, but can have large variances for areas with small sample sizes due in part to outliers. For several states in our study, the jackknife estimates of the variances are correlated with the direct estimates of the cash rental rates. In this section, we define a generalized variance function to obtain estimates of the sampling variances.

Using estimates of sampling variances based on a generalized variance function GVF (Wolter, 1976) instead of the direct jackknife estimates of the sampling variances is common practice in small area estimation for several reasons. First, direct estimators of variances (such as jackknife estimators) may have large variances for areas with small sample sizes. Smoothing the estimators of the variance through a generalized variances function can reduce the mean squared error of the estimator of the variance. Second, a direct estimator of a variance is equal to zero if the sample size for the area is less than two. Our definition of the generalized variance function gives a positive estimate of the variance for areas with a sample size of one. Second, correlations between direct estimators of variances and direct estimators of means can lead to biases in the estimators of the regression coefficients. Use of a generalized variance function can reduce the correlation between the direct estimator of the variance and the direct estimator of the mean and subsequently improve the quality of the predictions.

3.3.1 A Generalized Variance Function for the Direct Estimators of the Sampling Variances

Let $\hat{S}_{it}^2$ be the jackknife estimator of the variance of $\hat{y}_{it}$, the direct estimator of the average cash rental rate for county *i* in year *t*. Let n_{it} be the number of respondents in county *i* with positive dollars rented and positive acres rented in year *t*. Assume,

$$n_{it}^{0.5}\hat{S}_{it} = \theta_0 + (x_i - x_{..})'\theta_1 + \delta_{it} \qquad (6)$$

where x_i is the covariate in the model for the average, and

$x_{..} = n_{..}^{-1}(\sum_{i=1}^{m} x_i n_{i.})$, $n_{i.} = \sum_{s=t-1}^{t} n_{is}$, $n_{..} = \sum_{i=1}^{m} n_{i.}$ and $\delta_{it} \sim (0, n_{it}^{-1}\sigma_{\delta}^{2})$. Let $(\hat{\theta}_0, \hat{\Theta}_1')$ be the generalized least squares estimator of (θ_0, Θ_1') based on (6), and let

$$\hat{S}_{mi}(n_{it}) = \hat{S}_{mit} = n_{it}^{-0.5}[\hat{\theta}_0 + (1 - n_{it}^{-0.25})(x_{i.} - x_{..})'\hat{\Theta}_1] \qquad (7)$$

We use $\hat{S}_{mit}^{2}$ as an estimator of the variance of $\hat{y}_{it}$. A danger in (7) is that the predicted value for a standard deviation may be negative. For states presented here, all predicted values for standard deviations are positive. Another problem with modeling standard deviations instead of variances is that the variance estimators that result from squaring the standard deviations are not unbiased even if the estimator of the standard deviation is unbiased. Refinements to the generalized variance function in (7) are an area of current work.

To be able to get an estimate of variance across two years, an estimator of the correlation between the sampling errors is required. Let

$$\tilde{r}_{ijt} = \hat{a}_{i}^{-1}(d_{ijt} - \hat{y}_{it}a_{ijt}) \tag{8}$$

Where a_{ijt} and d_{ijt} are the acres and dollars, respectively, rented from operator j in county i and year t. Note that a correction is made for extreme values in the calculation of $\tilde{r}_{ijt}$ and is explained further in Appendix B. Let r_{ijt} be the result of the procedure in Appendix B. Let A be the set of *(ij)* that report positive dollars and positive acres in both time points, *t* and *t-1*. Let

$$\hat{\rho} = \frac{\sum_{(ij)\in A}(r_{ijt} - \bar{r}_{.t})(r_{ijt-1} - \bar{r}_{.t-1})}{\sqrt{\sum_{(ij)\in A}(r_{ijt} - \bar{r}_{.t})^{2}\sum_{(ij)\in A}(r_{ijt-1} - \bar{r}_{.t-1})^{2}}} \tag{9}$$

where $\bar{r}_{.t} = | A |^{-1} \sum_{(ij)\in A} r_{ijt}$ and |A| is the number of *(ij)* in the set A. Let

$$\hat{\Sigma}_{ei} = diag(S_{mit-1}, S_{mit})R_{i}diag(S_{mit-1}, S_{mit}) \tag{10}$$

where R_i is a 2x2 matrix of ones on the diagonal and $\hat{\rho}$ on the off-diagonals. Our estimates of sample variance for the average and difference are obtained using these results.

3.4 Two-stage Benchmarking

NASS obtains estimates of cash rental rates at the state level using data from a national survey that is conducted in June in addition to the Cash Rent Survey. The state estimates are published before the county-level data from the Cash Rent Survey are fully processed. NASS also establishes estimates of cash rental rates for agricultural statistics districts, groups of spatially contiguous counties within a state. To retain internal consistency, it is important that appropriately weighted sums of county estimates equal the district estimates and appropriately weighted sums of district estimates equal the previously published state estimate. The benchmarking restrictions for a time *t* are,

$$\sum_{i\in d_k} w_i \hat{\vartheta}_i = \hat{\lambda}_k \tag{11}$$

and

$$\sum_{k=1}^{D} \eta_k \hat{\lambda}_k = \theta_{pub} \tag{12}$$

where $w_i = (\sum_{i\in d_k} z_i)^{-1} z_i$, $\eta_k = (\sum_{k=1}^{D}\sum_{i\in d_k} z_i)^{-1}\sum_{i\in d} z_i$, z_i is a direct estimate of the acres rented in county *i*, d_k is a set of indexes for the counties in district *k*, $\hat{\lambda}_k$ is the final estimate of the average cash rental rate for district *k*, and $\hat{\theta}_{pub}$ is the published estimate of the state-level cash rent per acre. The index for the year is suppressed in (11) and (12) for simplicity. The direct estimators of the acres rented at the county and district levels are treated as fixed for our analysis.

We use the two-stage benchmarking procedure proposed by Ghosh and Steorts (2011) to define benchmarked estimates. The benchmarked estimates minimize the quadratic form,

$$g(c,d) = \sum_{k=1}^{D}\sum_{i\in d_k} \varsigma_i (\hat{\theta}_i^B - c_i)^2 + \sum_{k=1}^{D}\rho_k (\hat{\theta}_{k,w}^B - d_k)^2 \qquad (13)$$

subject to the constraints in (14) and (15), where $c = (c_1,..., c_m)$, $d = (d_1,..., d_k)$, $\hat{\vartheta}_{k,w} = \sum_{i\in d_k} w_i \hat{\vartheta}_i^B$, and ρ_k and ς_i are constants selected by the analyst. We choose $\varsigma_i = w_i$ and $\rho_k = \eta_k$, which gives the benchmarked estimates,

$$\hat{\theta}_i = \hat{\theta}_i^B + \hat{\lambda}_{k(i)} - \hat{\theta}_{k(i),w}^B \qquad (14)$$

and

$$\hat{\lambda}_{k(i)} = \hat{\theta}_{k(i),w}^B + \frac{(\theta_{pub} - \hat{\theta}_w^B)\eta_{k(i)}(1+\eta_{k(i)})^{-1}}{\sum_{i\in d_{i(i)}} \eta_{k(i)}^2 (1+\eta_{k(i)})^{-1}} \qquad (15)$$

for county *i* and district $k(i)$, respectively, where $k(i)$ is the district containing county *i*. In (15) , $\hat{\theta}_w^B = \sum_{k=1}^{D} \eta_k \hat{\theta}_{,k,w}^B$. Each of the benchmarked estimates in (14) and (15) is a sum of the predictor and an adjustment term. If the predictor for the state is larger (smaller) than the previously published state total, then the adjustment is negative (positive), and the benchmarked county and district estimates are smaller (larger) than the predictors. We ignore the effect of benchmarking on the MSE of the predictor. In a Bayesian setting, the mean squared error of the benchmarked predictor is the sum of the variance of $\hat{\theta}_i$ and the squared difference between $\hat{\theta}_i$ and the benchmarked predictor. (You et al., 2002)

4. Results

Recall that we are using models for three usages, nonirrigated and irrigated cropland, and permanent pasture. For each state/usage combination, we select one model for the average and one for the difference. For the purposes of simplicity, we will focus on five states in our analysis: Florida, Mississippi, Michigan, Iowa, and Kansas. These states reflect the diversity in agriculture necessary to represent a broad range of challenges with modeling cash rental rates. In this analysis we will focus on the estimates for the year 2011.

A common measure of whether the model improves upon the initial direct survey estimate is the measure the relative root mean squared error. We define relative root mean squared error as,

$$RRMSE = \frac{RMSE(\hat{\theta}_{it})}{RMSE(\hat{y}_{it})}$$

where $RMSE(\hat{y}_{it})$ is the estimated RMSE of the direct indication and $RMSE(\hat{\theta}_{it})$ is the estimated RMSE of the model estimate. Table 1 gives the medians of the distributions of RRMSEs for each state/usage combination. A RRMSE less than one from our definition occurs when the modeled estimate has a lower MSE than the direct survey estimate. For each state, the estimated MSEs of the model-based predictors are smaller than the estimated variances of the direct survey estimators for most of the counties in the state.

Table 1: Median Relative Root Mean Squared Errors

State	Nonirrigated	Irrigated	Pasture
Florida	0.84	0.63	0.70
Iowa	0.50	0.79	0.58
Kansas	0.66	0.62	0.71
Michigan	0.63	0.57	0.62
Mississippi	0.78	0.49	0.56

For nonirrigated cropland, Iowa showed the most improvement, with RRMSE less than 0.5 for half of the counties. For the irrigated cropland and pasture, we see a substantial drop in the MSE for the modeled estimates on the whole. The smallest improvement is for nonirrigated cropland in Florida, where the relationship between the covariates and the average cash rental rate is relatively weak.

Another measure in evaluation of the modeled estimates is whether coefficients of variation (CVs) are reasonable. Table 2 shows medians of county CVs across the states for each usage. Florida has the largest amount of variation relative to its estimates. No median CV is over 30%,

Table 2: Median of Estimated CVs for Model Predictors (%)

State	Nonirrigated	Irrigated	Pasture
Florida	13.54	26.39	21.65
Iowa	3.52	10.40	8.91
Kansas	6.97	11.03	5.10
Michigan	6.98	17.86	18.58
Mississippi	13.41	10.20	12.66

As mentioned in section 3, NASS uses a procedure where experts examine current and previous year survey data to set estimates. It is often desired to examine how modeled estimates compare with NASS published values as published values are viewed as a standard. Figure 2 illustrates this comparison for nonirrigated cropland with model predictions on the y-axis and published values on the x-axis. The five states are represented by color according to their state abbreviation. The model predictions and published values lie close to the 45 degree line, indicating consistency with our comparison standard.

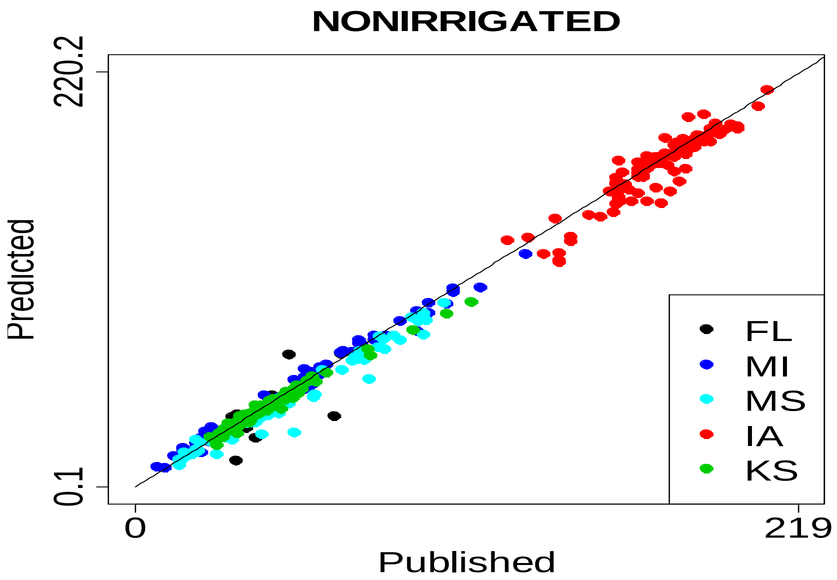

Figure 2: Plot of model predicted estimates vs. published values for 2011 nonirrigated cropland

5. Conclusions and Future Work

We use empirical Bayes methods to estimate county-level cash rental rates using previous year information and auxiliary data. We specify two separate univariate area-level models: one for the average of the two years and one for the difference between the two years. In this approach, the predictor for either time point is just a linear combinations of the estimators of the average and difference. We can obtain a MSE estimate by assuming no correlation between the estimators of average and difference. An advantage to this approach is that the processing time is quite small, which is important for use in a production environment.

In the states examined, the model-based procedure leads to an improvement in root mean squared error relative to the direct survey estimates. We also demonstrated relatively stable coefficients of variation for the three usages and a strong relationship with the published values of the same year.

Although these models demonstrate significant improvements, there are still several challenges. One is the potential for the rare case of a negative estimate to be produced. This is possible mainly for counties with relatively small survey indications and large variances due to the linear relationship with the covariates. A way to guarantee positive estimates that we are looking into is to make a single, positive covariate index with a positive correlation with the cash rental rate. Another potential for innovation is to combine contiguous states with small realized samples to pool from a larger area. We may also investigate the use of nonlinear models to this application.

Acknowledgements

The authors thank Wayne A. Fuller for his contributions to this research. Additional thanks are in order to Angie Considine, Sharyn Lavender, Curt Stock, Scott Shimmin, and Dan Beckler from the National and Rich Iovanna from the Farm Service Agency.

References

Battese, G.E., Harger, R.M., and Fuller, W.A. (1988), "An error-components model for prediction of county crop areas using survey and satellite data," *Journal of the American Statistical Association*, 83, 28 – 36.

Dhuyvetter, D., and Kastens, T. (2010). "Kansas Land Values and Cash Rents at the County Level." http://www/agmanager.info/farmmt/land/county/CountValues &Rents(Sep2010).pdf.

Ghosh, M. and Steorts, R. C., "Two Stage Bayesian Benchmarking as Applied to Small Area Estimation," (2011). To be submitted.

Wolter, K.(1976), *Introduction to Variance Estimation*, Springer series in statistics

Woodard, S., Paulson, N., Baylis, K., and Woddard, J. (2010). "Spatial Analysis of Illinois Agricultural Cash Rents," The Selected Works of Kathy Baylis. http://works.bepress.com/kathy_baylis/29.

Rao J.N.K.(2003): *Small Area Estimation*, New Work: John Wiley and Sons.

Wang, J., and Fuller, W.A. and Qu, Y. (2008), "Small area estimation under a restriction." *Survey Methodology*, 34, 29-36.

You, Y., Rao, J.N.K., Dick, P. (2002). "Benchmarking Hierarchicial Bayes Small Area Estimators with Applications in Census Undercoverage Estimation." *Proceedings of the Survey Methods Section*, Statistical Society of Canada.

Appendix A: Derivation for the Estimate of the Difference

Recall from section 3.2 that we are trying to get an estimate of $\hat{y}_{di}$ for scenario 1. Suppose

$$\begin{pmatrix} y_{ijt-1} \\ y_{ijt} \end{pmatrix} \sim \left(\begin{pmatrix} \mu_{it-1} \\ \mu_{it} \end{pmatrix}, \sigma_i^2 R_i \right)$$

Where R_i is a 2x2 correlation matrix with parameter ρ_i. Then, $V\{\bar{y}_{di}\} = 2(1-\rho_i)\sigma_i^2 n_{it-1,t}^{-1}$ and $V\{\hat{y}_{udi}\} = \sigma_i^2 n_{it-1,t-1}^{-1} + \sigma_i^2 n_{it,t}^{-1}$, where $\hat{y}_{udi} = \bar{y}_{dit,t} - \bar{y}_{dit-1,t-1}$, and $\bar{y}_{dis,s}$ is the simple average of the respondents with positive dollars and positive acres in years s but not in year t. A generalized least squares estimator of $\mu_{it} - \mu_{it-1}$ is

$$\hat{y}_{di,opt} = \alpha_i \bar{y}_{udi} + (1-\alpha_i)\bar{y}_{di} \tag{16}$$

where $\alpha_i = 2(1-\rho_i)n_{it-1,t}^{-1}[2(1-\rho_i)n_{it-1,t}^{-1} + n_{it-1,t-1}^{-1} + n_{it,t}^{-1}]^{-1}$. If $n_{it-1} = n_{it}$, then

$$\bar{d}_i = \frac{n_{i,bb}\hat{y}_{udi} + n_{it-1,t}\bar{y}_{di}}{n_{i,bb} + n_{it-1,t}} \tag{17}$$

where $\bar{y}_{di} = \bar{y}_{it} - \bar{y}_{it-1}$, and $\bar{y}_{it}$ is the simple average of y_{ijt} . If we solve for $\bar{y}_{udi}$ in (20), we obtain,

$$\bar{y}_{udi} = \frac{n_i}{n_{i,bb}}\bar{d}_i - \frac{n_{it-1,t}}{n_{i,bb}}\bar{y}_{di} \tag{18}$$

where $n_i = n_{i,bb} + n_{it-1,t}$. By substitution of the right hand side of (21) into (19),

$$\hat{y}_{di,opt} = \alpha_i \frac{n_i}{n_{i,bb}}\bar{d}_i + (1-\alpha_i \frac{n_i}{n_{i,bb}})\bar{y}_{di} \tag{19}$$

Finally, by replacing $\bar{d}_i$ in (22) with $\hat{d}_i$ we get our estimate of $\hat{y}_{di}$ used in scenario 1.

Appendix B: Modification to Extreme Taylor Deviates

Let $S_{r,t}$ be the sample standard deviation of $\tilde{r}_{ijt}$ defined in (11). Let med_i be the median of $\tilde{r}_{ijt}$. If $\tilde{r}_{ijt} > med_i + 3.3S_{rt}$, then set $\tilde{r}_{ijt} = med_i + 3.3S_{rt}$. If $\tilde{r}_{ijt} < med_i - 3.3S_{rt}$, then set $\tilde{r}_{ijt} = med_i - 3.3S_{rt}$. Otherwise, set $r_{ijt} = \tilde{r}_{ijt}$. We do not iterate the procedure for simplicity.