Integration of Multiple Sources of Data by the National Agricultural Statistics Service

Malay Ghosh
University of Florida

August 5, 2019

Outline

- Background
- To Model or not to Model
- Multiple Data Sources for Crops
- Multiple Data Sources for Cash Rents
- Bayesian Analysis of Cash Rent Data
- Final Remarks

Background

- The National Agricultural Statistics Service (NASS) is the primary data collection agency within the US Department of Agriculture (USDA).
- The agency conducts multiple surveys annually and prepares reports covering virtually every apect of agriculture,
- NASS produces county level estimates for crops (planted acres, harvested acres, production and yield by commodity).
- Also, they produce cash rental rates for irriggated cropland, non-irrigated cropland and permanent pastureland.
- Key users of these county level estimates include USDA's Farm Service Agency (FSA) and Risk Management Agency (RMA).
- They use these estimates as part of their processes for distributing farm subsidies and providing farm insurance respectively.

- USDA began regularly reporting estimates of acreage, yield and production of principal crops starting 1866.
- Currently, NASS is responsible for estimating acreage, production and yield for most commodities grown in US.
- National and State level estimates are among the primary economic indicators of the Office of Management and Budget (OMB).
- NASS has been publishing county-level crop and livestock inventories since 1917.
- These county level estimates, originally prepared with Federal funding, have now evolved into Federal-State cooperative funding agreements.
- Crop surveys are usually sponsored by states.
- NASS statisticians in states define samples and processes to be used in developing these estimates.

- Nass has also long estimated rental rates for farmland at the state level.
- Between 1950 and 1994, state level cash rents were estimated using a list survey of real estate appraisers.
- Beginning in 1994, state level cash rents were estimated primarily by the June Area Survey.
- This survey asks the farmers directly about their rental agreements.
- The 2000 farm bill mandated NASS to provide mean rental rates for all counties with at least 20,000 acres of cropland plus pasture.

- Cash rents usually reflect the value of the farmland as an input to agricultural production.
- These estimates are used by farm operators and land owners in negotiating rental agreements.
- Also used by bankers in making farm operating and ownership loans.
- Other users are real estate agents, analysts, financial advisors and many others.
- Information on cash rents helps building efficiency of farmland markets by providing participants in these markets with data on supply and demand.

- The need for disintegrated and accurate information on acreage, yield etc. was felt over time.
- In fact, this was required by FSA for several of their programs under the 2014 farm bill.
- In particular, this was required for their Conservation Reserve Program (CRP).
- Under this program, the Federal Govt. makes a payment to a farmer in exchange for removing environmentally sensitive land from agricultural production, and planting species to help improving environmental health and quality.
- Pre 2008, FSA administered the CRP using rental rates estimated by appointed FSA board of directors in each state.
- Beginning 2009, FSA used NASS statistical survey of county average rental rates for cropland and pastureland to establish rental rates to be used for CRP.

- NASS took steps to meet these administrative requirements by collecting data at the county level.
- First, the agency initiated the County Agricultural Production Survey (CAPS) in 2011 for a few states, and expanding this to 44 states in 2012.
- Today, CAPS provides survey data to estimate acreage and production of selected crops at the county level to be used for state and federal programs.
- One virtue was replacement of nonprobability samples by probability sampling.
- Today the NASS county crop estimate program is joinly administered by NASS, FSA and RMA.

- In the last decade NASS has responded to the increasing use of county level data by introducing large scale probability surveys.
- Often however, farmers have complained that the county averages do not meet their own farm experience.
- Indeed, a small variation in a county yield estimate may result in relatively large changes in subsidy payments.
- Substantive variability in the payment rates from one county to the next can also be driven by these yield differences.
- The monetary significance of a county yield estimate therefore focuses attention on how it is derived, and for an individual farmer, on how well it corresponds to a farmer's own yield outcomes.

- Another issue is the shrinkage in the number of counties for which survey-based estimates are available due to non-response.
- In the absence of NASS county estimates, FSA needs to resort to secondary sources of information which may not properly reflect a county's experience.
- Without a NASS county estimate or lack of sufficient data to report a credible RMA county yield, FSA often resorts to a NASS and/or RMA estimate for a neighboring county.
- This may result in yield determaination higher for some counties and lower for others compared to what it actually should be.

- NASS applies a publication standard to determine those counties for which the number of survey reponses is adequate to support estimation.
- The standard requires that there should be at least 30 valid responses for a crop in a county.
- If there are fewer, they collectively must account for at least 25% of acreage by crop.
- If this standard is not met, NASS does not publish the county's estimate.
- For the CAPS, the response rate was 62.8% in 2011 and 56.4% in 2016.
- For the Cash Rents Survey, the response rate was 75.9% in 2014, but fell to 69% in 2016.
- Reflecting this decline, NASS published data for 2,879 counties in 2,014 and 2,579 counties in 2016.

To Model or not to Model

- Traditionally government statistical agencies avoided the use of models and favored design-based approaches for perceived objectivity.
- More recently, there is a growing demand for esimates at finer levels of detail where direct survey estimates fail to meet the desired precision.
- Also there are methodological advances in what has now become known as small area estimation.
- Statistical agencies are now more open to use of models to achieve fusion of survey-based information with other data sources in pursuit of greater precision in estimation.
- Direct survey-based estimates are appropriate for large samples.
- But incorporation of additional information with the aid of statistical models is essential for small domains where sample is scarce.

- Like other statistical agencies, NASS has traditionally been involved in producing survey-based estimates.
- However, especially for crop estimation program, NASS has been fortunate to have high quality alternative sources of data.
- The NASS approach has been to provide input to ASB with these sources of information.
- In addition, it provides also a set of rules to ASB so that it can make some good estimates by integrating these multiple sources of data.
- Indeed the ASB does review all indications such as survey, administrative and/or model-based and prepares estimates for dissemination.
- A similar ASB process is used for county level estimates.
- The concern, however, is that such estimates are neither transparent nor reproducible.

- NASS requires consistency of its published estimates at all geographical levels: national, principal regions, states and Agricultural Statistics District (ASD) and counties.
- National, principal region and state estimates for acreage and production are based on surveys from the NASS list frame and June Area Survey.
- Estimates for crop acreage and production are first set by ASB for the total produced for the region containing the principal states in which the commodity is grown.
- State totals for that region must add to the regional total and are set second.
- Then the national estimate is derived from the principal region total and the total for the remaining states producing that commodity.
- No national estimate is published for cash rents, but data obtained from June Area Survey are used to publish state level estimates.

- NASS employs a "top-down" procedure for establishing its substate estimates.
- National and state estimates are published before ASD and county estimates, even though there may be ongoing data collection.
- To incorporate survey and auxiliary information into its crop estimates, NASS computes composites of available inputs for its acreage and production totals.
- The ASD-level composites are ratio-adjusted to the published state total, and then rounding rules are enforced to publish the official ASD statistics.
- Subsequently the county composite estimates are ratio-benchmarked to the rounded ASD totals.
- Rounding rules are then enforced again at the county level to produce the official county estimates.

Multiple Data Sources for Crops

- ASB considers multiple data sources when developing county-level crop estimates.
- NASS produces annual estimates for planted acres, harvested acres, production and yield (production divided by harvested acres).
- The main source of data used to develop these estimates is the County Agricultural Production Survey (CAPS).
- Over time , however, NASS has been able to publish estimates only for a fewer counties because of falling response rates.
- This makes it a necessity to use auxiliary data.
- These auxiliary data sources are administrative data from the FSA and the RMA and data observed from remote sensing tools.

- NASS utilizes several modeling approaches for planted acres, harvested acres and production.
- The first involves a weighted estimate prepared from (1) direct survey estimates, (2) administrative data, (3) remote sensing data and (4) estimates made from the past year.
- Each component is first ratio benchmarked so that the aggregate over counties equals the previously published state total.
- The currently used weights seem stable over time, although there are provisions to make adjustment if necessary.
- In the table that follows, I present the weights that are currently attached to (1) direct survey estimates, (2) planted indication times the direct survey ratio, (3) FSA estimates, (4) planted indication minus RMA failed, (5) remote sensing and (6) past data.
- * and ** correspond to availability and non-availability of remote sensing.

Table. Weights used to compute the composite indication.

Data source	(1)	(2)	(3)	(4)	(5)	(6)
Planted acres(*)	.15	0	.65	0	.20	0
Planted acres(**)	.20	0	.80	0	0	0
Harvested acres	.14	.70	0	.08	0	.08
Production (*)	.12	.78	0	0	.05	.05
Production (**)	.15	.80	0	0	0	.05

- The second approach is based on NASS Cropland Data Layer (CDL).
- This is a remote sensing product which provides pixels classified by land use (including a range of crops) as the basis for independent county level indications of planted acreage by crop.
- The currently used Landsat Satellite Data along with the June Area Survey Data are together used for finding estimates using a method of Battese, Harter and Fuller (1988).
- There is also a model-based approach proposed in Erilescu, Cruze and Nandram.

Cash Rents

- NASS is responsible for estimating average cash rental rates at the county level mandatedby the 2008 Farm Bill.
- A cash rental rate refers to the market value of land rented on a per acre basis for cash only.
- Estimates of cash rental rates are useful to farmers, economists, and policy makers.
- Farmers use the estimates for guidance in determining rental agreements.
- Because realized sample sizes at the county level are often too small to support reliable direct estimators, predictors based on models are inevitable.
- However, to maintain internal consistency, the county predictors must satisfy a benchmarking restriction, i.e. ratio benchmarked to the state estimators.

- Cash Rent Surveys use a stratified sample design.
- To define the stratification, nine groups are formed on the basis of the dollars rented that an operation reported on previous surveys and censuses.
- The strata are the intersections of the nine groups and agricultural statistics districts.
- The sampling fractions within strata are defined so that operations with higher dollars rented on previous surveys and censuses have greater probabilities of selection.
- For illustration, we consider 2009 and 2010 Cash Rent Surveys.
- The same sample was used for the 2009 and 2010 Cash Rent Surveys, which had a national sample size of approximately 224,000 operations.
- A unit may respond in only one year either because of nonresponse or because the operation only participated in a rental agreement in one of the two years.

$y_{ij10} - \bar{y}_{i10}$
50
0
-50

- 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.
- For the states considered in Berg, Cecere, and Ghosh (2014), the correlations between the direct estimates for the two years range from 0.20 to 0.99, where the correlation is across counties for a particular state.
- Because the emphasis is on unit level models, the focus is on relationships over time at the unit level.

- Berg, Cecere and Ghosh (2014) use area-level models to predict county-level cash rental rates for all states and for the three land use categories of non-irrigated cropland, irrigated cropland, and permanent pasture.
- For each combination of land use category and state, the method of Berg, Cecere and Ghosh (2014) uses data from two years.
- An assumption that the variances for the two years are the same motivates the Pitman-Morgan transformation, which converts the vector of observations for the two time points into an average and a difference.
- After separate univariate models are applied to the average and the difference, the predictor for each time point is obtained by adding or subtracting the predictor of the average to half of the predictor of the difference.
- The simplifying assumption of the same variability in two years is often questionable, especially now that the data collection is made every other year.

- In an effort to improve the precision of the estimators of average cash rental rates at the county level, auxiliary variables were desired that would explain both the variability among the county means as well as the variability among units within a county.
- Auxiliary information for modeling cash rental rates is available from several sources external to the Cash Rent Survey.
- The potential covariates divide into three broad categories, depending on whether the covariate relates principally to land quality, the commodity value sold, or other farm characteristics.
- The list below summarizes the three categories of covariates, indicates whether each covariate is recorded at the county level or the unit level, and specifies if the covariate is only available for a particular state.
- Unit-level covariates are only available for units in the Cash Rent Survey sample, while area level covariates are treated as population means.

- Land Quality: Four National Commodity Crop Productivity Indexes (NCCPIs) are county-level covariates available for all states.
- Three climate-specific indexes called NCCPI-corn, NCCPI-wheat, and NCCPI-cotton reflect the quality of the soil for growing non-irrigated crops in three different climate conditions.
- The fourth index, Max-NCCPI, is the maximum of the three climate-specific indexes.
- The indexes are originally constructed at the level of a "mapunit," an area that has relatively homogeneous soil properties.
- The county-level covariates are averages of the indexes across all mapunits in a county.

- Value of the Commodity Sold: Total value of production for a county based on the 2007 Census of Agriculture is available for all states.
- Expected sales for an operation (unit) recorded on the NASS list frame are available for all states at the unit-level.
- Other farm characteristics: Farm type is a unit level categorical covariate, available for all states. Farms are partitioned into 17 farm types on the NASS list frame. To define a covariate, the farm types are aggregated into two groups: (1) grains/oilseeds, and (2) other.
- Acres rented for non-irrigated cropland recorded on the NASS Cash Rent Survey are available at the unit level for all states.

- The correlation between the 2009 and 2010 cash rental rates suggests that using the information in the data from 2009 has the potential to improve the predictions for 2010.
- A bivariate hierarchical model for a state is specified as a way to incorporate the data for both years.
- Let $a_{ij,t}$ and $y_{ij,t}$ be the acres and dollars per acre, respectively, rented by operator j in county i and year t. $t = 09$ or $t = 10$.
- Let $\boldsymbol{x}_{ij,t}$ be the associated column vector of auxiliary variables with dimension p_t.
- For covariates that are constant across years and individuals, $\boldsymbol{x}_{ij,t} = \boldsymbol{x}_{i109}$.
- Let $w_{ij,t} = a_{ij,t} N_{g(ijt)} n^{-1}_{g(ijt)}$.
- $N_{g(ijt)}$ and $n_{g(ijt)}$ are the population size and number of respondents, respectively, in year t for the stratum g that contains unit (ij).

- To specify the model, the respondents are divided into three sets:
- Set 1 consists of units (ij) that report a non-irrigated cash rental rate in both 2009 and 2010.
- Set 2 consists of units (ij) that only report a non-irrigated cash rental rate in 2009.
- Set 3 consists of units (ij) that only report a non-irrigated cash rental rate in 2010.
- Observations in Set 1 follow a bivariate model:
  $y_{ij,09} = \boldsymbol{x}_{ij,09}^T \boldsymbol{\beta}_{09} + \nu_{i,09} + e_{ij,09}$
  $y_{ij,10} = \boldsymbol{x}_{ij,10}^T \boldsymbol{\beta}_{10} + \nu_{i,10} + e_{ij,10}$
- $(\nu_{i,09}, \nu_{i,10})^T \sim \mathrm{N}_2(\boldsymbol{0}, \boldsymbol{\Sigma}_\nu)$
- $(e_{ij,09}, e_{ij,10})^T \sim \mathrm{N}_2(\boldsymbol{0}, \boldsymbol{D}^{-0.5} \boldsymbol{\Sigma}_e \boldsymbol{D}^{-0.5})$.
- $\boldsymbol{D} = \mathrm{diag}(w_{ij,09}, w_{ij,10})$.
- $\boldsymbol{\Sigma}_e = \mathrm{diag}(\sigma^2_{e09}, \sigma^2_{e10})$.

- For units (ij) in set 2 or 3, we assume
  $y_{ij,t} = \boldsymbol{x}_{ij,t}^T \boldsymbol{\beta}_t + \nu_{i,t} + e_{ij,t}^*$,
  where $e_{ij,t}^* \sim \mathrm{N}(0, w_{ij,t}^{-1} \tau_{e,t}^2)$, $t = 09$ for set 2, and $t = 10$ for set 3.
- The model not only allows the variances for the unit-level errors to differ across time points but also allows the variances of unit-level errors for units that respond in both time points to differ from the variances for units that only respond in one time-point.
- The quantity to predict for 2010 is $\theta_{i,10} = \bar{\boldsymbol{x}}_{N_{i,10}}^T \boldsymbol{\beta}_{10} + \nu_{i,10}$.
- The variances of the unit-level errors, $e_{ij,t}$ and $e_{ij,t}^*$, are assumed to be inversely proportional to the weight, $w_{ij,t}$, for two reasons.
- First, incorporating the weights in the model aims to reduce bias that could arise if the design is informative for the model.
- The second reason to incorporate the weights is that the variances of residuals from preliminary analyses decrease as the acres increase.

- Diffuse proper priors are specified for the unknown regression coefficients and variances.
- Specifically, $\boldsymbol{\beta}_t \sim \mathrm{N}(\mathbf{0}, 10^6 \boldsymbol{I})$.
- $\tau^2(e,t) \sim \mathrm{IG}(0.001, 0.001)$.
- The covariance matrices, $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\Sigma}_\nu$ have inverse-Wishart prior distributions with shape parameter 0.01 and a diagonal scale matrix with diagonal elements 0.001.
- The analysis is based on one chain of length 20,000 for each of the three states, Iowa, Kansas and Texas, where the first 1000 iterations are discarded for burn-in.

- We used Gibbs sampling to obtain a Monte Carlo approximation to the posterior distribution.
- An analysis based on three MCMC chains, each with 20,000 iterations, indicated that 1000 iterations is sufficient for burn-in.
- By the choices of the likelihood and the priors, the full conditional distributions are all known distributions.
- We obtained first the posterior means and posteror variances of the county level casha rent estimates.
- Then we obtained two-stage benchmarked estimators following the method proposed in Ghosh and Steorts (2013).

- NASS obtains estimates of cash rental rates at the state level using data from the June Area Survey in addition to 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.
- To retain internal consistency, appropriately weighted sums of county estimates must equal the district estimates and appropriately weighted sums of district estimates must equal the previously published state estimate.
- We did fit the non-irrigated cropland cash rental rates reported on the 2009 and 2010 Cash Rent Surveys for Iowa, Kansas, and Texas.
- These three states were chosen to reflect a range of situations.

Table: Posterior distributions of correlations between 2009 and 2010.

	$Cor\{\nu_{i09}, \nu_{i10}\}$		$Cor\{e_{ij09}, e_{ij10}\}$	
State	Median	95% Credible Interval	Median	95% Credible Inter
Iowa	0.746	[0.611, 0.839]	0.570	[0.548, 0.592]
Kansas	0.919	[0.870, 0.950]	0.727	[0.701, 0.751]
Texas	0.884	[0.831, 0.921]	0.691	[0.667, 0.714]

- To demonstrate the gain in efficiency due to the use of the bivariate model relative to a univariate model, we compare the posterior mean squared errors of the predictors from the bivariate model to the posterior mean squared errors of the predictors from a corresponding univariate model.
- The assumptions of the univariate models are the same as the assumptions of the bivariate models except that the covariance parameters in Σ_{ee} and $\Sigma_{\nu\nu}$ are assumed to equal zero.
- To fit the univariate models, we use inverse-gamma prior distributions for σ^2_{et} and $\sigma^2_{\nu\ t}$ $(t = 09, 10)$.
- To compare the bivariate and univariate models, we define the relative posterior MSE (RelMSE) for county i by

  $$\text{RelMSE}_{i,10} = \frac{MSE^{BBench}_{i10}}{MSE^{UNIBench}_{i10}},$$

  where where MSE^{BBench}_{i10} is defined in and $MSE^{UNIBench}_{i10}$ is the posterior MSE based on the corresponding univariate model.

- The average relative MSEs for Iowa, Kansas, and Texas are 88.71%, 97.27%, and 88.65%, respectively, where the average relative mean squared error for a state is $D^{-1}\sum_{i=1}^{D}\text{RelMSE}_{i,10}$.
- The average relative mean squared error for a state is $D^{-1}\sum_{i=1}^{D}\text{RelMSE}_{i,10}$.
- Benchmarked estimators are used for both the bivariate and univariate models.
- Because of the significant correlations in the model errors for the two time points, the posterior MSE from a bivariate model is smaller than the posterior MSE from the corresponding univariate model, and the average relative efficiencies are less than one.

- To assess model fit, we use the posterior predictive p-value, which measures departures between the observed data and the model based on some summary statistics.
- We use only the elements observed in both 2009 and 2010 and we denote this set by A_i, $i = 1, \ldots, D$.
- We consider two summary statistics: the mean for each year and the multivariate skewness.
- The mean for year t is the mean of the observations in set 1 for year t and is defined by $\bar{y}_t = (\sum_{i=1}^{D} |A_i|)^{-1} \sum_{i=1}^{D} \sum_{j \in A_i} y_{ijt}$, where A_i denotes the elements in set 1 for county i.
- The multivariate skewness is defined by $\hat{\gamma}_{1,p} = (\sum_{i=1}^{D} |A_i|)^{-1} \sum_{i=1}^{D} \sum_{i'=1}^{D} \sum_{j \in A_i} \sum_{\ell \in A_i} m_{iji'\ell}^3$, where $m_{iji'\ell} = (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})' \boldsymbol{S}^{-1} (\boldsymbol{y}_{i'\ell} - \bar{\boldsymbol{y}})$, $\boldsymbol{y}_{ij} = (y_{ij,09}, y_{ij,10})'$, $\bar{\boldsymbol{y}} = (\bar{y}_{09}, \bar{y}_{10})'$, and $\boldsymbol{S} = (\sum_{i=1}^{D} |A_i| - 1)^{-1} \sum_{i=1}^{D} \sum_{j \in A_i} (\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}})'$.

Summary and Conclusion

- We use a bivariate HB model to obtain predictors of county-level cash rental rates for non-irrigated cropland in Iowa, Kansas, and Texas.
- The model incorporates auxiliary information related to land quality, commodity values, and farm characteristics.
- Significant correlations exist between the 2009 and 2010 model random effects at both the unit and county levels.
- As a consequence, using the information in the 2009 cash rent estimates reduces the posterior MSE relative to a univariate model.
- To incorporate unit-level covariates with unknown population means, we add a level to the hierarchical model that justifies adding a term to the posterior mean squared error to account for uncertainty in the unknown population means of the unit-level covariates.