Comparison of Model Based Methods for County Level Estimation of Crop Yields

Michael E. Bellow

USDA/NASS
3251 Old Lee Hwy., Rm. 305, Fairfax, VA 22030
mbellow@nass.usda.gov

1. Introduction

The National Agricultural Statistics Service (NASS) has been publishing estimates of crop, livestock and other commodities at the county level since 1917. The primary source of data for agricultural commodity estimation has always been surveys of farmers, ranchers and agribusinesses who voluntarily provide requested information on a confidential basis. Since surveys designed and conducted at the national and state levels are seldom adequate for obtaining reliable county estimates, NASS has made extensive use of ancillary data sources such as list sampling frame control data, previous year estimates, earth observing satellite data and Census of Agriculture data. When estimating area planted or harvested to a crop, the availability of reliable administrative data has been very helpful. The number of acres planted to a typical crop seldom varies dramatically from year to year, so crop area estimation is typically a straightforward and repeatable process. On the other hand, accurate estimation of county crop yield, i.e., the ratio of production to harvested area, has always been more challenging due to several factors: 1) a general lack of dependable administrative data, 2) the fact that yields can vary significantly from one year to the next due to climatological factors, and 3) lack of adequate survey data. County yield estimates are scrutinized heavily by crop insurance firms and other data users. Since county estimates have traditionally been computed based on a non-probability sample of farms with no nonresponse follow-up and differential sampling/response rates for small vs. large farms, standard small area estimation methods based on known selection probabilities could not be used. In recent years, there has been considerable interest in the potential of model based small area estimation techniques, a direction that NASS has pursued in its research program.

Three features of a county yield estimation method deemed important to NASS are: 1) repeatability of the estimation process, 2) ability to estimate variance accurately, and 3) capability to produce estimates for counties where no survey data is available. NASS has traditionally used a simple ratio estimator to derive county yields. This estimator is computed as the ratio between the county estimates of production and harvested area for a given crop. The ratio estimator can produce inconsistent yields due to fluctuations in harvested area from year to year. Furthermore, there is no usage of data from any other county in deriving a given county's estimate, nor is there any provision for obtaining an estimate in the absence of survey data for a county.

Stasny et al. (1995), working under NASS's former cooperative agreement with the Ohio State University, developed a Bayesian county yield estimation algorithm based on the idea that due to weather, cropping practices and other factors, the crop yields of counties in close geographic proximity tend to be more similar than those of counties further apart. This procedure, which will be referred to as the Stasny-Goel method, assumes a mixed effects model with farms as the sample units, farm size (reduced to two or three size groups based on total land operated) as the fixed effect, and county location as the random effect. The county effect is assumed to have a multivariate normal distribution whose mean vector is a constant multiple of the previous year's county yield estimates and whose variance-covariance matrix incorporates spatial correlation among neighboring counties. The algorithm attempts to fit the mixed effects model using the EM algorithm, which runs until relative group and log-likelihood distances reach a preset limit. The county estimates are computed as weighted averages of individual farm level estimates, with the weights derived from size group membership data from the most recent Census of Agriculture. The estimates can be adjusted to agree with published state totals. One drawback of this method is the fact that when very few reports are available for a commodity, convergence to a solution may not be achieved after a reasonable number of iterations.

Griffith (1999) proposed an alternative spatial county yield estimation method that predicts yield values using the published number of farms producing the crop of interest. Box-Cox and Box-Tidwell transformations are employed in conjunction with an autoregressive specification so as to optimize agreement with model assumptions. Estimates for counties with missing survey data are computed by an imputation routine that utilizes the spatial correlation among neighboring counties as well as previous year county level data. Griffith's method is more computationally intensive than the Stasny-Goel procedure as it

involves extensive nonlinear estimation. As with the other method, county estimates can be made to agree with state totals and there are situations where convergence may be difficult to achieve.

Both the Stasny-Goel and Griffith algorithms are programmed as SAS macros. A comprehensive study aimed at comparing their performance and utility, with the goal of eventually selecting one or the other for incorporation into NASS's operational county estimation systems, has been planned and begun. The research will eventually involve a variety of major and minor crops in ten geographically dispersed states: New York, Tennessee, Mississippi, Florida, Ohio, Michigan, North Dakota, Oklahoma, Colorado and Washington. The methods will be compared for two or more estimation cycles (years) in order to assess performance over time. A regression based simulation approach developed for the study is described in Section 3. Early results obtained for oats and barley in North Dakota are discussed in Section 4. Although issues related to algorithm convergence and adjustment for missing data are not covered in this paper, they will be explored in the near future.

2. Post-Stratification Size Groups

The Stasny-Goel county estimation method requires that post-stratification size groups be defined, with yield estimates computed for each size group in all counties. NASS sample survey data for the current year are post-stratified by county and farm size based on farm level data on total acres operated obtained from the most recent Census of Agriculture. The Census is conducted by NASS every five years (for years ending in '2' or '7'). The percentages of Census farm acres by size group in a particular county are used by the algorithm as weights to derive the county yield estimates. Due to the possibility of significantly altered agricultural conditions since the most recent Census year, these weights can be updated based on ratios between survey based estimates of total land operated and number of farms for the current and Census years. The size groups are also used in the simulation procedure described in Section 3.

Stasny et al. (1995) recommended that two or three size groups be defined such that each group had approximately the same number of farms (grouping method 1). Another possibility is grouping method 2, where all groups have roughly equal total land operated (summed over farms). In the next section, the two grouping methods will be compared, but only in terms of generating valid simulated data sets. Table 1 shows the definitions developed under methods 1 and 2 with three size groups for North Dakota, the state used for the estimator comparisons discussed later in this paper.

Table 1: Post-Stratification Size Group Definitions for North Dakota; TL = total land operated (acres)

Size Group	Grouping Method 1			Grouping Method 2		
	Definition	Census No. Farms	Census Total Land	Definition	Census No. Farms	Census Total Land
1	TL<500	8,773	1,829,042	TL<1700	18,661	12,098,859
2	500≤TL<1500	8,644	8,298,019	1700≤TL<3300	5,292	12,240,612
3	TL≥1500	8,655	26,118,472	TL≥3300	2,119	11,906,062

3. Simulation Methodology

In order to evaluate the Stasny-Goel and Griffith county yield estimation methods as well as the standard ratio estimator for a variety of crops in geographically dispersed states, a simulation approach was developed that attempts to emulate NASS survey data for yield as accurately as possible. The methodology is basically that used by Crouse (2000) in his study of the Stasny-Goel method for crops in Michigan, but with a few modifications. The approach was applied to oats and barley yield estimation using data from the state of North Dakota for the year 2002; results will be discussed in Section 4. Oats and barley, two crops grown primarily in the northern great plains region of the United States, are harvested throughout North Dakota. The NASS data sources required to conduct a simulation study are current year survey data (from the December Agricultural Survey in the case of North Dakota oats and barley), published county yield estimates for the previous year (2001), and data on number of farms and land in farms from the most recent Census of Agriculture (1997).

The objective was to generate a simulated population of yield values from which 'true' population parameters could be derived for later comparison with estimates computed over subsets of the population. For each crop of interest, a multiple regression analysis was performed with the survey yield response values being the dependent variable. Four independent variables were used: the official NASS county yield estimate for 2002, the weighted average neighbor yield (WANY), and two indicator variables pertaining to membership in post-stratification size groups. The WANY for a given county is computed as the weighted average of official yield estimates of all neighbors of that county, with the weight assigned to each neighbor county being the ratio of harvested acreage (official estimate) for that county to the total harvested acreage of all the neighbor

counties. This variable was added in an effort to increase the spatial correlation of the simulated data so as to better reflect real survey data.

The regression equation used to generate replications of simulated data has the following general form:

$$y_{ij}^{(r)} = \alpha + \beta_y Y_i + \beta_z Z_i + \beta_{s1}\Psi_{1j} + \beta_{s2}\Psi_{2j} + \varepsilon_{ij}^{(r)} \quad (1)$$

where:

$y_{ij}^{(r)}$ = simulated yield value for replication r, county i, survey record j

α, β's = regression parameters

Y_i = NASS official crop yield estimate for county i

Z_i = weighted average neighbor yield for county i

Ψ_{gj} = 1 if record j is in size group g (g=1,2)
= 0 otherwise

$\varepsilon_{ij}^{(r)}$ = random error for replication r, county i, survey record j

There was no need to include a size group indicator variable for group 3 since whether or not a given record belongs to that group can be determined from the indicator variables for groups 1 and 2. The random error term was assigned a normal distribution with mean zero and variance equal to the sample variance of survey yield response values.

A large number of replicated survey data sets (10,000) was generated in order to ensure that the 'true' population parameters computed from these simulated records would agree with the model. From this population, a sample of 500 data sets was randomly selected. The three county estimation methods were then applied to each of the sample data sets. For each county, the sample based estimates for a given method were averaged and compared with the 'true' population values.

The regression equation (1) occasionally produced negative yields, which were rounded up to zero. The rounding process induces a minor bias into the simulated data, so the intercept term needed to be adjusted. A pilot population of 10,000 simulated data sets was generated for this purpose. The adjustment term was selected so that the statewide crop yield averaged over the simulated data sets equaled the official state yield estimates for 2002. Those statewide estimates were 46 bushels per acre for barley and 44 for oats. The actual set of 10,000 simulated data sets used in the estimator comparison was generated via a different random number seed than the one used to create the pilot population. For internal consistency purposes, the same seed was used for both crops.

Table 2 shows the regression parameters obtained after applying equation (1) to 2002 North Dakota survey data for oats and barley, respectively. For each crop, regression was run for both post-stratification grouping methods discussed in Section 2. The intercept terms (α) shown do not reflect the subsequent adjustment discussed earlier. The R^2 values are the regression coefficients of determination, which for purposes of generating simulated data need not be especially high. The table shows that the choice of grouping method had a negligible effect on R^2 for both crops. For this reason, grouping method 1 (the 'equal number of farms' criterion recommended by Stasny) was used exclusively in the later analysis.

An important aspect of the simulation process is ensuring that the simulated data sets accurately reflect the spatial correlation present in real survey data. Moran's I, a statistic that measures spatial correlation (Moran, 1950), was computed for the original survey data sets and all 500 simulated data sets generated using grouping method 1 for each crop. Table 3 shows the average, minimum and maximum values of Moran's I for the simulated and survey data sets. For both crops, the survey value of Moran's I is within the range of the simulated values and not that far from the average simulated value. Thus the simulated data sets can be regarded as having an acceptable degree of spatial correlation.

Table 2: Regression Parameter Estimates Used to Generate Simulated Data Sets

Crop	Parameter	Estimate	
		Grouping Method 1	Grouping Method 2
Oats	α	1.79	3.39
	β_y	0.94	0.93
	β_z	0.064	0.087
	β_{s1}	-12.99	-9.42
	β_{s2}	-6.73	-2.42
	R^2	0.27	0.27
Barley	α	-2.03	-1.61
	β_y	0.73	0.73
	β_z	0.29	0.32
	β_{s1}	-2.94	-5.89
	β_{s2}	-4.42	-2.45
	R^2	0.34	0.34

Table 3: Spatial Correlation for Simulated and Survey Data Sets (Grouping Method 1 Used)

Crop	Data Set	Statistic	Moran's I
Oats	Simulated	Average	0.46
		Minimum	0.22
		Maximum	0.64
	Survey	Value	0.53
Barley	Simulated	Average	0.67
		Minimum	0.51
		Maximum	0.78
	Survey	Value	0.73

4. Results for North Dakota Crops

This section discusses results of the simulation study done for oats and barley in North Dakota for the year 2002. Counties having fewer than three positive harvested acreage records for a given crop were excluded from the comparison study for that crop since NASS data disclosure rules prohibit publication of estimates for such counties. This rule led to the removal of four of North Dakota's 53 counties (Adams, Sioux, Slope and Steele) for oats and two (Adams and Sioux) for barley. The Stasny-Goel, Griffith and ratio estimation methods were applied to each of the 500 simulated data sets for both crops. Convergence was achieved in an acceptable number of iterations in all cases. The simulated Stasny-Goel and Griffith county yield estimates were adjusted to agree with state level totals.

For both oats and barley, the three estimators were ranked according to five efficiency measures based on statistics computed over all 500 replicated data sets. Tables 4 and 5 show the number of counties as well as percent of counties for which estimates were ranked first, second and third according to the measures. Average ranks over counties are also shown. The method with lowest average rank for a given measure can be said to have 'done best' for that criterion. The five measures are absolute bias, variance, mean square error, lower tail (5^{th} percentile) proximity and upper tail (95^{th} percentile) proximity. Absolute bias was computed as the average value over simulations of the absolute differences between the estimates produced by a given method and the population 'true' county yields. Variance was computed as the sample variance of simulated county yield estimates, while mean square error was calculated by averaging the squared deviations between estimates and 'true' county yields. The last two measures were included to assess outlier properties of the estimators, i.e., the tendency to produce 'out of bounds' yield estimates. Lower tail proximity is the absolute difference between the 5^{th} percentile of simulated yield estimates and the 'true' county yield, while upper tail proximity is defined similarly using the 95^{th} percentile.

From the average rankings shown in the tables, the Stasny-Goel method was best among the three county estimation methods in all five efficiency categories for oats, and in four of the categories for barley (although only slightly better than the Griffith method in terms of lower tail proximity). The only exception was variance for barley, where the Griffith procedure was best in 88 percent of the counties. Both model based methods clearly outperformed the ratio estimator, which had the highest average rank in all cases.

Table 4: Estimator Rankings for Oats

Measure	Rank	Estimation Method					
		Stasny-Goel		Griffith		Ratio	
		Count	Percent	Count	Percent	Count	Percent
Absolute Bias	1	21	43	21	43	7	14
	2	22	45	12	24	15	31
	3	6	12	16	33	27	55
	Average	1.69		1.9		2.41	
Variance	1	29	59	20	41	0	0
	2	20	41	27	55	2	4
	3	0	0	2	4	47	96
	Average	1.41		1.63		2.96	
Mean Square Error	1	25	51	19	39	5	10
	2	19	39	15	31	15	31
	3	5	10	15	31	29	59
	Average	1.59		1.92		2.49	
Lower Tail (5th Percentile) Proximity	1	27	55	21	43	1	2
	2	20	41	22	45	7	14
	3	2	4	6	12	41	84
	Average	1.49		1.69		2.82	
Upper Tail (95th Percentile) Proximity	1	26	53	19	39	4	8
	2	20	41	20	41	9	18
	3	3	6	10	20	36	73
	Average	1.53		1.82		2.65	

Table 5: Estimator Rankings for Barley

Measure	Rank	Estimation Method					
		Stasny-Goel		Griffith		Ratio	
		Count	Percent	Count	Percent	Count	Percent
Absolute Bias	1	27	53	18	35	6	12
	2	22	43	9	18	20	39
	3	2	4	24	47	25	49
	Average	1.51		2.12		2.37	
Variance	1	6	12	45	88	0	0
	2	45	88	5	10	1	2
	3	0	0	1	2	50	98
	Average	1.88		1.14		2.98	
Mean Square Error	1	23	45	23	45	5	10
	2	26	51	5	10	20	39
	3	2	4	23	45	26	51
	Average	1.59		2.0		2.41	
Lower Tail (5th Percentile) Proximity	1	23	45	26	51	2	4
	2	27	53	17	33	7	14
	3	1	2	8	16	42	82
	Average	1.57		1.65		2.78	
Upper Tail (95th Percentile) Proximity	1	22	43	22	43	7	14
	2	26	51	15	29	10	20
	3	3	6	14	27	34	67
	Average	1.63		1.84		2.53	

Table 6: Results of Rank Sum Tests on Absolute Bias (α= .05)

Crop	Comparison	Favorable Results			Neither Method Better
		Stasny-Goel	Griffith	Ratio	
Oats	Stasny-Goel vs. Griffith	26	21	_	2
	Stasny-Goel vs. Ratio	32	_	10	7
	Griffith vs. Ratio	_	29	18	2
Barley	Stasny-Goel vs. Griffith	31	16	_	4
	Stasny-Goel vs. Ratio	41	_	7	3
	Griffith vs. Ratio	_	24	25	2

A more rigorous analysis of absolute bias properties was performed. For each county in the study and each pair of estimators, one-sided Wilcoxon rank sum location tests were run on the absolute differences between simulated estimates and population 'true' county yields in order to assess whether one method in the pair produced estimates with significantly lower absolute bias than the other. Results of tests performed at the five percent significance level are summarized in Table 6, which shows the number of counties for which rank sum test results favored one method or the other as well as the number of counties where neither method was favored (null hypothesis of equal absolute bias accepted). Combining the results obtained for both crops, the Stasny-Goel method was favored over the Griffith method in 57 percent of tests, compared with 37 percent favorable results for the Griffith and results favoring neither method six percent of the time. The ratio estimator was significantly worse than the Stasny-Goel estimator in 73 percent of tests and significantly worse than the Griffith estimator in 53 percent of tests.

Table 7 provides insight into the actual (not absolute) bias and coefficient of variation (CV) of the estimators. The differences between county yield estimates and population 'truth' values were averaged over simulations on a county basis to estimate bias for each method. The median, minimum and maximum (over counties) of these bias estimates were then computed. Similar computations were done on the estimated county level CV (standard deviation divided by mean simulated estimate) values. The median bias estimates for all three methods were reasonably close to zero with both crops, so there do not appear to be underestimation or overestimation tendencies. For both oats and barley, the maximum CV was much lower for the Stasny-Goel procedure than for the other two methods, suggesting that this method produces the most consistent estimates.

Table 7: Bias and CV Statistics

Measure	Crop	Estimation Method								
		Stasny-Goel			Griffith			Ratio		
		Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
Bias	Oats	-8.6	0.3	23.9	-12.9	1.3	18.9	-1.4	0.03	0.92
	Barley	-4.4	-0.03	17.0	-15.5	-0.07	19.1	-0.64	-0.03	0.55
CV (%)	Oats	6.7	9.8	15.8	4.6	9.8	36.3	10.2	19.4	76.6
	Barley	4.2	7.4	15.2	1.4	4.7	26.3	5.7	11.7	70.4

5. Concluding Remarks

Two model based county crop yield estimation methods that exploit spatial correlation among neighboring counties and make use of previous year survey data have been proposed as potential improvements to NASS's current operational methodology. The Stasny-Goel, Griffith and standard ratio estimators were compared using simulated NASS survey data for oats and barley in North Dakota. A regression based methodology for simulation was developed for this purpose. The estimators were evaluated with regard to bias, variance, mean square error and outlier measures. The results favored the Stasny-Goel method in almost all cases. Both model based methods were clearly superior to the ratio method.

Similar studies will be done in nine other geographically dispersed states for a number of crops, with the goal of eventually selecting one method and adopting it for use in NASS's county estimation program. There is also the possibility that a hybrid estimation procedure could be developed that incorporates favorable features of both the Stasny-Goel and Griffith methods. The estimator comparisons will be done over more than one survey cycle for each state in the project area. Two key issues not covered in this paper are algorithm convergence and adjustment for missing data, both of which will be explored in future phases of the research effort.

References

Crouse, C. (2000), "Evaluation of the Use of Spatial Modeling to Improve County Yield Estimation", Research Report No RDD-00-05, National Agricultural Statistics Service, U.S. Department of Agriculture.

Griffith, D. (1999), "A Methodology for Small Area Estimation with Special Reference to a One-Number Agricultural Census and Confidentiality: Results for Selected Major Crops and States", Research Report No. RD-99-04, National Agricultural Statistics Service, U.S. Department of Agriculture.

Moran, P. (1950), "Notes on Continuous Stochastic Phenomenon", Biometrika, 37, 17-23.

Stasny, E., Goel, P, Cooley, C. and Bohn, L. (1995), "Modeling County-Level Crop Yield with Spatial Correlations Among Neighboring Counties", Technical Report No. 570, Department of Statistics, The Ohio State University.