RESEARCH INTO SMALL AREA ESTIMATION AT THE U.S. DEPARTMENT OF AGRICULTURE

Barry Ford, Doug Bond, Nancy Carter

One of the goals of the Statistical Reporting Service in the U.S. Department of Agricultural (USDA) is to develop a clear, rigorous statistical process for small area estimation of agricultural values. This paper is an overview of research which was conducted at USDA in two phases and written in two in-house reports [1,2] which give the technical details of the material. The first phase involved estimators which only relied on current survey information. The second phase involved estimators which were based on simple models that combined current and historical information.

Most of the small area estimation at USDA centers on counties and small collections of counties called districts. The discussion in this report focuses on counties because they are the "lowest" level of interest and because in the USDA research the results for districts were much the same as counties. Currently, USDA makes county estimates by allowing statisticians in their 44 field offices to gather information from any available source and to determine subjectively the level of each county estimate. Examples of sources of information are the U.S. Census of Agriculture run by the Bureau of the Census every five years, state censuses of agriculture, and economic and weather conditions in each county. However, USDA is interested in developing a formal statistical procedure to make county estimates — a procedure which would also use current information from its operational surveys although these surveys have been designed to establish state and/or national estimates.

The research on which this paper is based began in 1979 when the field office in North Carolina requested help on making county estimates. Until 1979 North Carolina had a state census of agriculture each year which determined the levels of agricultural variables in the 100 counties of North Carolina. Thus, a sound historical series had been established, and the field office wanted to continue the series by using the best statistical procedures possible.

PHASE ONE: Evaluation of Direct and Synthetic Estimators

In order to evaluate estimators which depend solely on current survey information, a state survey of agricultural acreage and production was expanded to a sample size which yielded approximately 18,000 respondents. The sample was stratified into four strata which classified farmers by an auxiliary variable giving the number of acres in the farming operation. The population within each stratum was ordered by the county in the mailing address, and the sample allocation for each stratum was selected systematically across the counties. The very large sample size guaranteed that at least two units were selected from each county in a stratum, and in those cases where nonresponse prevented the collection of data from less than two units, strata were collapsed.

Data from this survey was used to evaluate two county estimators. The first was the direct estimator. This estimator only used whatever sample units fell in county i to make estimates for that county:

$$D_i = \sum_{j=1}^{4} N_{ij} \bar{x}_{ij} .$$

where i refers to the county, j to the stratum, N_{ij} is the number of population units in county i and stratum j, and $\bar{x}_{ij}$ is the mean of the sampled units in county i and stratum j.

Although this estimator is mathematically unbiased and has "tremendous appeal to those individuals responsible for regional, state, and local planning" [5] it usually requires a large sample size to attain standard errors which are reasonably small.

The second estimator was the synthetic estimator. In each stratum j this estimator used $\bar{x}_{Gj}$ the sample mean of district G (the district which contained county i) as the stratum mean of county i:

$$S_i = \sum_{j=1}^{4} N_{ij} \bar{x}_{Gj} .$$

Thus, the synthetic estimator used the estimate from a large area in order to form estimates for a small area. Standard errors for both estimators are straightforward when the sample sizes in each county are assumed fixed.

Synthetic estimates have an intuitive appeal, are generally easy and inexpensive to obtain [5], and usually have a much smaller standard error than direct estimates. However, the synthetic estimator is also biased. In the USDA research the bias was the result of how much the district means differed from the county means. For most surveys it is difficult to estimate the mean square error -- the squared bias plus the squared standard error -- for each county. However, an estimate of the average mean square error across all counties is possible [3], and thus an estimate of the average squared bias across all counties is also possible.

Table 1 compares the direct and synthetic estimates with regard to mean square error (MSE) and its components -- variance and squared bias -- for seven agricultural variables. Except for the number of hogs, the direct estimator had a much smaller MSE than the synthetic estimator. Although the synthetic estimator has a smaller variance, i.e. making it a more stable estimator, it also had a larger bias.

The large sample size of approximately 18,000 farmers had a big impact on the results in Table 1. When the sample size is extremely large, the bias rather than the variance dominates the MSE. For smaller sample sizes, the bias will probably remain at the same level, but the variances of both estimators will increase.

The state sample sizes for which the direct and synthetic estimators would have had the same MSE were: all land in farm -- 5399, hogs -- 26,268, cattle -- 4158, corn -- 3654, tobacco -- 2040, soybeans -- 4153, and sorghum -- 8960. Across the seven variables the sample size averaged about 7500 (or about 75 sample units per county) for the MSE of the two variables to be equal. Without the hog variable the average would be about 5000 (or about 50 sample units per county).

For the sake of completeness, it should be mentioned that there is a composite estimator which combines the direct and synthetic estimators by weighting them according to the mean square errors [6]. When composite estimates were computed for this study, the sample size was so large that the composite estimators were almost exactly the same as the direct estimates. Thus, the composite estimator offered little improvement as it might for smaller sample sizes.

Although Table 1 shows that the direct estimator is better than the synthetic estimator in terms of MSE, both estimators had variances which were too large for the uses of USDA. These variances translated into coefficients of variation (CV) for the direct estimator which ranged from 0.14 to 1.80 and averaged about 0.42. Again the effect of the hog variable was great; most CV's were in the 0.14 to 0.24 range. CV's this high would make county estimates fluctuate so much from year to year that time trends and relationships among counties would be unrecognizable. Thus, USDA felt that county estimates needed further stabilization. Any increase in the sample size beyond 18,000 was impossible because of time and cost constraints -- in fact, North Carolina already planned to decrease its sample size to between 10,000 and 12,000 because of costs. Therefore, USDA planned a second phase to investigate the use of models to make county estimates.

PHASE TWO: Using Models to Combine Historical and Current Data to Make County Estimates

The purpose of this phase was to overcome the instability of the direct and synthetic estimators by creating a model-based procedure. This model used the historical data of "official" values for counties in North Carolina from 1972 to 1980 in order to measure relationships among the counties over time. From 1972 until 1978 North Carolina still had a state farm census, but after 1978 "official" values were based on information from current surveys, control data, opinions of crop conditions, etc. The official values after 1978 should not be considered the exact truth for any time period but should only be considered as a "best" guesses that USDA made by using the statistical, economic, meteorological, and historical information available at the time.

Phase Two concentrated on data from three major crops in North Carolina -- corn, soybeans, and tobacco. For corn and soybeans three variables were analyzed -- planted acres, harvested acres, and produced bushels. For tobacco, two variables were analyzed -- harvested acres and produced pounds.

In order that past values for counties accurately predict future values, relationships over time had to be discernible. Evaluation of data in North Carolina showed that although some gross time trends were evident using totals (e.g. bushels of corn produced, acres of tobacco planted), time trends became much more stable when totals were translated into percentages. For example, although the total acres of soybeans planted in a particular county fluctuated a great deal

from year to year, the county's percentage of the state total for planted acres of soybeans remained relatively stable from year to year. Using the historical data, simple linear regressions for each county were fitted where the independent variable was a time variable of the years 1972-1980 and the dependent variable was percentage of the state total in that county. Although some of the time trends were slightly curved, simple linear regressions were used as approximations to capture the general nature of the time trends.

The authors decided to change the aim of the research from the estimation of the county totals to the estimation of each county's percentage of the state total in order to take advantage of strong time trends in the percentages. This slight change in the aim emphasized that the estimation of the county values was an allocation process. Whatever estimates were determined for the state by using current surveys, etc. could be allocated to the counties by using the percentage estimates.

The percent of explained variation in the regressions varied from 26% to 45% of the total variation and averaged about 37%. The percent of explained variation was moderate not because of scatter in the data but because of the horizontal nature of many of the time trends. The result of fitting simple linear regressions in this type of situation was to model the data as an average percentage over time -- an average percentage which usually had a small standard error. Thus, although the percent of explained variation was moderate, the regression estimates usually had small standard errors -- a fact which is evident in later tables. The authors decided to continue using the regression model because it helped some variables and did no harm for others.

Besides the time trends, there were strong relationships among variables in the historical data, and these relationships proved useful for the estimation of county values. For example, the harvested acreage of corn was highly correlated to the planted acreage of corn -- explained variation was 83% of total variation. These relationships among variables were also made part of the modeling process.

For each crop in each county, a three-stage process was constructed to estimate county values. The first stage was to estimate the percentage of planted acres for county i by weighting together two component estimators:

$$p_i = u_{i1} p_{i1} + u_{i2} p_{i2}$$

where for county i: p_{i1} was the estimated percentage based on direct estimates from the current survey, p_{i2} was the estimated percentage from a simple linear regression on the percentages of planted acres over 1972-1980, and u_{i1}, u_{i2} were optimal weights whose formulas are discussed later. Each p_i was proportionately adjusted so that the total of the p_i's across all counties would equal 1.

The second stage was to estimate the percentage of harvested acres in county i by weighting three component estimators:

$$h_i = v_{i1} h_{i1} + v_{i2} h_{i2} + v_{i3} h_{i3}$$

where for county i: h_{i1} was the estimated percentage based on direct estimates from the current survey, h_{i2} was the estimated percentage from a simple linear regression on the percentages of harvested acres over 1972-1980, h_{i3} was the estimated percentage based on the historical relationship of planted acres to harvested acres over 1972-1980, and v_{i1}, v_{i2}, v_{i3} were optimal weights. The estimate h_{i3} required the estimation of p_i from the first stage. Each h_i was proportionately adjusted so that the total of the h_i's across all counties equaled 1.

The third stage was to estimate the percentage of produced bushels (or pounds) for county i by weighting three component estimates:

$$r_i = w_{i1} r_{i1} + w_{i2} r_{i2} + w_{i3} r_{i3}$$

where for county i: r_{i1} was the estimated percentage based on direct estimates from the current survey, r_{i2} was the estimated percentage from a simple linear regression on the percentages of production over 1972-1980, r_{i3} was the estimated percentage based on the historical relationship of harvested acres to production over 1972-1980, and w_{i1}, w_{i2}, w_{i3} were optimal weights. The estimate r_{i3} required the estimation of h_i from the second stage. As in the other two stages, each r_i was proportionately adjusted so that the total of the r_i's across all counties equaled 1.

The optimal weights for each stage were a combination of the standard errors of each estimated percentage. These weights were optimal in that they minimized the standard errors of p_i, h_i, and r_i.

Evaluation of the three-stage procedure involved using the historical data from 1972-1980 to make county estimates for 1981. Since there were no "true" values for comparisons with these estimates, they were compared against the official values for 1981. These comparisons reflected how well the formal statistical procedure approximated the more unstructured and subjective process of making official estimates.

Table 2 shows the absolute differences between estimated percentages and official percentages for the 100 counties in North Carolina. The absolute differences are averaged across all 100 counties to show general effects without showing all of the differences for each county. The absolute differences are shown both for the component estimators and for the combined estimator calculated by weighting the component estimators together. At the county level, the differences for the time model were much smaller than for the other components. The time model percentages and the combined percentages were very close to the official percentages -- in fact, the time trends were slightly closer to the official values. The first inclination would be to discard all other estimators and only use the time trends, but the authors felt that the current data should have some impact because of atypical crop seasons where the current data would be useful even if the errors were large.

Table 3 shows the standard errors of the component percentages and the combined percentages as averages across the 100 counties. The relatively low standard errors from the time models and historical relationships contrast with the high standard errors from the current survey. For planted acreages of corn and soybeans and for harvested acreages of tobacco, the standard errors show that the time model percentages dominate the combined estimator. For other crop variables, the time model percentages are less important, and the percentages based on historical relationships with other variables are more important. Thus, trends in the historical data were accounting for almost all of the stability in the combined percentages.

The fact that percentages based on the time models agree so closely with the official percentages has two possible interpretations: (1) the official percentages from the time models were close to the "truth" and therefore the percentages from the time models were closer to the "truth" than the other components, or (2) field statisticians had subjectively used time trends to set official values.

If the first interpretation is true, then the procedure combining the historical and current data is indeed a very accurate procedure. If the second interpretation is true, then the procedure is approximating in a formal way the process that statisticians use in making official estimates. The authors adopted the second interpretation because it was less stringent than the first interpretation

and because it was highly likely to have occurred. Validation that the estimates were close to "truth" must wait until a study where the "true" values are available.

However, the official values for tobacco in Tables 2 and 3 were based on accurate control data because of regulation of the tobacco industry. Thus, the official estimates for tobacco can be regarded as "true" values. Results for tobacco did show that estimates from combining historical and current data were closer to the truth than just using current survey data. However, the regulation that enabled USDA to have good control data also probably stabilized the trends and relationships in tobacco values over time, perhaps making tobacco an atypical crop. Thus, although the results for tobacco were encouraging, they should not be accepted as proof that the procedure would yield estimates close to "truth" for other crops.

CONCLUSIONS AND FUTURE RESEARCH

This research showed that for USDA data a three-stage procedure which combined historical and current data gave estimated percentages for counties which were stable and close to official values. Plans now call for comparisons between 1982 estimates from the procedure and values from the 1982 U.S. Census of Agriculture. USDA would like to refine the allocation procedure in several ways: 1) by investigating more sophisticated time series models, 2) by including weather information into the allocation procedure since weather data is probably a major determinant of crop production, and 3) by extending the estimation to variables of yield, livestock, and minor crops.

References

[1] Ford, Barry L. "The Development of County Estimates in North Carolina." Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. 20250. 1981.

[2] Ford, Barry L.; et al. "Combining Current and Historical Data to Make District and County Estimates In North Carolina." Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. 20250. 1983.

[3] Gonzalez, Maria E. "Use and Evaluation of Synthetic Estimates," American Statistical Association: Proceedings of the Social Statistics Section. 1982.

[4] Steinberg, J. (ed.) Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion. National Institute of Drug Abuse Research Monograph 24, U.S. Government Printing Office. Washington, D.C. 1979.

[5] ----- (Chapter in reference [4]). Levy, P.S. "Small Area Estimation Synthetic and Other Procedures."

[6] ---- (Chapter in reference [4]). Schaible, W.A. "A Composite Estimator for Small Area Statistics."

Table 1. Using four strata, a comparison of the relative values of the mean square error, MSE, and its components -- the variance, V, and the squared bias, B^2 -- for direct and synthetic county estimates. By definition, $MSE=V + B^2$. The values in this table are average values across the 100 counties in North Carolina and are in relative terms because they are divided by the average county estimate.

Variable	Relative MSE		Relative V		Relative B^2	
	Direct	Synthetic	Direct	Synthetic	Direct	Synthetic
All Land in Farm (acres)	0.02	0.05	0.02	<0.01	----	0.05
Hogs (number of head)	3.24	2.40	3.24	0.56	----	1.84
Cattle (number of head)	0.05	0.19	0.05	0.01	----	0.18
Corn Harvested (acres)	0.05	0.21	0.05	0.01	----	0.20
Tobacco Harvested (acres)	0.04	0.28	0.04	<0.01	----	0.28
Soybeans Harvested (acres)	0.06	0.26	0.06	0.01	----	0.25
Sorghum Harvested (acres)	1.25	2.41	1.25	0.15	----	2.26

Table 2. Absolute differences between estimated percentages and official percentages as averages across the 100 counties in North Carolina during the 1981 crop year. A '*' indicates that a particular component was not available or not used.

	Tobacco		Corn			Soybeans		
Source of Estimator	Harvested Acreage	Produced Pounds	Planted Acreage	Harvested Acreage	Produced Bushels	Planted Acreage	Harvested Acreage	Produced Bushels
Current Survey	0.25	0.25	0.35	0.30	0.30	0.40	0.32	0.34
Time Model	0.08	0.11	0.05	0.06	0.15	0.09	0.09	0.13
Historic Relation With Other Variable	*	0.13	*	0.08	0.18	*	0.11	0.12
Combined Estimate	0.08	0.13	0.08	0.08	0.16	0.11	0.11	0.12

Table 3. Standard errors for estimated percentages indicating each county's part of the state total. Standard errors are shown as averages across the 100 counties in North Carolina. A '*' indicates that a component was not used or not available.

	Tobacco		Corn			Soybeans		
Source of Estimator	Harvested Acreage	Produced Pounds	Planted Acreage	Harvested Acreage	Produced Bushels	Planted Acreage	Harvested Acreage	Produced Bushels
Current Survey	0.31	0.31	0.33	0.20	0.22	0.44	0.26	0.26
Time Model	0.06	0.09	0.05	0.05	0.11	0.07	0.06	0.10
Historic Relation With Other Variable	*	0.04	*	0.02	0.15	*	0.02	0.08
Combined Estimate	0.06	0.03	0.05	0.02	0.07	0.08	0.01	0.04