Composite Predictions of Yield for Agricultural Commodities

Timothy P. Keller
William H. Wigton

Sampling and Estimation Research Section
Research and Development Division
National Agricultural Statistics Service
United States Department of Agriculture
3251 Old Lee Highway, Room 305
Fairfax, Va. 22030-1504

1. Introduction.

The National Agricultural Statistics Service (NASS), as the primary fact-collecting and reporting agency of the U.S. Department of Agriculture, is responsible for the national program of timely, accurate, and relevant statistics. NASS reports cover most aspects of the agricultural sector including production of crops and livestock, prices paid and received, and stocks on hand. NASS runs hundreds of national, state and local surveys that include data published in more than 400 publications of interest to government policy makers, private sector planners and producers and educators and researchers. Most of these surveys have evolved to the point where their foundations are firmly based on probability sampling, employing both area and list sampling frames. However, as one might surmise, this complex structure sometimes leads to having multiple indications for the same parameter.

Combining several estimators of a common parameter into a single composite estimator, which is, in some sense, better than any of the constituent estimators, is the central problem of composite estimation. There are many agricultural applications of composite estimation. The application we discuss is the problem of optimally combining several predictions of yield for a given agricultural commodity, over a given geographic area, into a single composite yield prediction. Setting the yield estimates which are ultimately published entails the application of much agricultural expertise and subject matter knowledge on a vast body of information; the formal composite estimation problem is only a small part of this larger process. Nevertheless, having a rational, mathematically sound method of combining estimates provides a useful starting point for the subsequent application of expert knowledge. In the larger view of things, the proper role for mathematical statistics is not to provide a replacement for the operation of subjective expert judgment, but to aid that judgment by providing tools built on a rational, intellectually sound basis.

2. Brief Review of the Classical Theory of Composite Estimation

The key result in the classical theory of composite estimation is stated in the following proposition.

Theorem 1

If $T_1, \cdots, T_p$ are unbiased estimators of θ with covariance matrix Σ, then $w_1 + \cdots + w_p = 1$ implies that $w_1 T_1 + \cdots + w_p T_p$ is an unbiased estimator of θ. Among all such linear combinations of $T_1, \cdots, T_p$ the minimum variance estimator of θ is

$$(2.1) \qquad \hat{\theta} = \frac{e' \Sigma^{-1} T}{e' \Sigma^{-1} e},$$

where $e' = (1, \cdots, 1)$ and $T' = (T_1, \cdots, T_p)$. See, for instance, Raj (1968), for details.

The application of this result involves two key steps:

(1) The estimators available are often significantly biased; hence one 'corrects' these estimators for bias in some fashion.

(2) Using the bias-corrected estimators one computes an estimate of the covariance matrix , S, to obtain the so-called Graybill-Deal composite estimator:

$$(2.2) \qquad \tilde{\theta} = \frac{e' S^{-1} T}{e' S^{-1} e} \quad .$$

Note that (2.2) presents a composite estimator which is a linear combination of estimators for which the coefficients, as well as the constituent estimators, are random variables. Hence, as intuition suggests,

$$(2.3) \qquad \mathrm{Var}(\tilde{\theta}) > \mathrm{Var}(\hat{\theta})$$

(Note that as long as S and T are independent, the Graybill-Deal estimator is still unbiased.) This observation leads to one of the main objections to the application of Theorem 1 : for small sample sizes the variance estimator given by (2.2) may be larger than the the variance of at least one of the constituent estimators, defeating the purpose of computing a composite estimator! Keller and Olkin (2002) gave a simple expression for the relative efficiency of the Graybill-Deal composite estimator in the important case for which $T \sim N(\theta e, \Sigma)$, which demonstrates that this objection is of no consequence as long as the sample size on which the constituent estimators are based are of moderate size.

The central issues in applying Theorem 1 are the formulation of reasonable model assumptions on which to base estimates of biases and the covariance matrix.

3. Model Assumptions for the Application to Yield Prediction

Calculations are based on historical data for the constituent yield estimators and the final yield estimates published by NASS. Thus, for a given combination of commodity, geographic region, and month of prediction, one has for year i , a vector W_i of p constituent yield indications, and the published final yield for the corresponding combination of commodity, geographic area and year. Given this data, we make the following assumptions:

(1) The official published final yield for year i is essentially the true yield, θ_i, for the geographic region in question.

(2) Over a period of N previous years, the biases and covariance structure of the constituent indications under consideration, are essentially constant.

(3) Yield indications are independent, across years.

If Y_i denotes the vector of the p *bias-corrected* constituent yield estimates for year i, then the corresponding estimate, S, for the covariance matrix is :

$$(3.1) \qquad S = \frac{\sum_{i=1}^{N} (Y_i - \theta_i e)(Y_i - \theta_i e)'}{N}$$

Bias-Correction Techniques

Three methods of correcting the constituent indications for bias were investigated:

(1) Bias modeled as an additive constant over time.

An estimate of bias for a given combination of month, region and commodity is based on the average deviation of the yield estimator from the published final yield over past years.

(2) Bias modeled as a multiplicative constant over time.

This amounts to using the regression estimate of yield based on a simple linear regression of each estimator (for a given combination of month, region and commodity) on the published final yield, with the regression constrained to pass through the origin.

Generalizing this idea, we also considered

(3) Bias correction via simple linear regression of the yield estimators on the published final yields.

The performance of these three methods of bias correction, measured by the mean square error of the corresponding monthly yield predictions, were about the same overall. In some instances, bias correction via simple linear regression seems to have a modest advantage; but, as yet, a definitive conclusion is not possible.

Estimating the Covariance Matrix

The geographic region corresponding to the yield predictions was either a state, or the major producing region for a commodity comprising several states. Since the number of years of historical data available for a given state was sometimes only 10 or 12, it was suggested that the sample size for estimating the covariance matrix could be increased if one could reasonably assume that the covariance structure for the constituent estimators not only had *temporal* stability, as previously stated, but also a measure of *spatial* stability. I.e. the covariance matrices corresponding to a set of p yield estimators of the same type, were essentially the same for Illinois and Indiana. In practice, the mean square error of the resulting yield predictions were little better, if at all, than the yield predictions based on state by state estimates of the covariance matrix. It seems that whatever gains can be achieved through having a larger sample size, are counter balanced by the loss of precision in using a single estimate for somewhat different state level covariance matrices.

Incorporating Expert Knowledge into the Modeling Process

The issues of how to select historic data for estimating bias and covariance structure, and the issues relating to the development of a more realistic model of bias than any of the three rather simplistic models previously discussed, are issues that require subject matter knowledge, not merely knowledge of statistical methodology. It is our hope that as the users of these ideas, come to 'own' the ideas, their expert knowledge will be brought to bear on these issues.

4. Some Typical Results

NASS conducts two major surveys relating to yield. One survey simply asks a random sample of producers of the commodity in question to give the their best estimate of the yield they expect at harvest for their own fields. The other survey makes counts and measurements on plants from plots randomly located throughout the major producing region for the commodity in question, and makes a yield prediction based on a simple biological model of yield. Based on historical evidence the yield prediction based on the former survey generally has a negative bias, while the yield prediction based on the latter survey generally has a positive bias.

Security considerations preclude presenting actual yield data ; however, to give the reader an idea of the of typical performance of the composite yield prediction, we present a table of results which are the result of carefully transforming actual data in order to mask information. The columns labeled 'farmer reported yield ' and ' biological yield model ' are the yield predictions corresponding to the surveys just described, but corrected for bias using a

constant bias correction. The column labeled ' panel of experts' is the yield prediction produced by a group of commodity specialists, privy not just to the two survey results previously described, but a wealth of other information relating to the yield of the commodity in question. The column labeled 'true yield' is the estimate of final yield published by NASS after harvest of the commodity in question is complete.

Table 1

Predicted yield (weight per area) of commodity Z for state X in month Y.

year	farmer reported yield	biological yield model	composite estimate	panel of experts	true yield
1	88.0	87.5	87.8	89.5	87.8
2	82.5	80.0	81.5	82.5	87.3
3	83.0	86.5	84.2	85.8	85.3
4	73.5	79.0	75.3	76.3	76.8
5	79.0	84.5	81.3	83.3	78.3
6	82.0	83.5	82.5	83.8	89.0
7	83.0	79.8	81.8	85.0	82.5
8	80.8	84.0	81.8	81.3	84.0
9	81.0	83.0	81.7	81.8	82.3
10	79.0	79.0	79.0	81.0	80.8
11	64.0	76.0	68.3	67.5	68.3
12	80.5	83.8	81.6	83.0	83.0
13	83.0	87.0	84.4	85.0	85.0
14	81.5	78.5	80.4	82.0	81.8

Root Mean Square Errors :

Farmer reported yield : 3.06
Biological yield model: 3.92
Composite estimator: 2.68
Panel of experts: 2.58

As these sample results indicate, a composite estimate can provide a useful starting point for producing a yield prediction. Perhaps the most exciting aspect of this work is the generality of the ideas, which allow application in numerous instances besides the particular application discussed in this paper.

Bibliography :

[C] W. Cochran, Problems arising in the analysis of a series of similar experiments, *Journal of the Royal Statistical Society* **4** (1937)102-118

[Co] A. Cohen, Combining estimates of location, *Journal of the American Statistical Association* **71** (1976) 172-175

[G] F. Graybill, R. Deal , Combining unbiased estimators, *Biometrics* **15** (1959) 543-550

[Gr] E. Green, W. Strawderman, A James-Stein type estimator for combining unbiased and possibly biased estimators, *Journal of the American Statistical Association* **86** (1991) 1001-1006

[H] E. Houseman, Composite Estimation, Statistical Reporting Service Airlie Conference, April (1970)

[K1] T. Keller, I. Olkin, Combining Correlated Unbiased Estimators of the Mean of a Normal Distribution, *Stanford University Technical Reports*, No. 2002-5, (2002).

[K2] T. Keller, et al, Research on Composite Indications of Crop Yield, National Agricultural Statistics Service United States Department of Agriculture, (in process).

[Na] K. Nair, Variance and distribution of the Graybill-Deal estimator of the common mean of two normal populations, *The Annals of Statistics* **8** (1980) 212-216

[No] E. Norwood, K. Hinkelmann, Estimating the common mean of several normal populations, *The Annals of Statistics* **5** (1977) 1047-1050

[P] C. Perry, P Cook, M Holko, and S Wiyatt, Composite Estimators in the June Hog Series, *NASS USDA* 1989.

[R] D. Raj , *Sampling Theory* , McGraw-Hill (1968) 16 - 17

[W] W. Wigton and H. Robert Van der Vaart, A Note on Combining Estimates 1973, (unpublished)

[Z] S. Zacks, Unbiased estimation of the common mean of two normal distributions based on small samples of equal size, *Journal of the American Statistical Association* **61** (1966) 467-476