Survey Methodology, December 1991
Vol. 17, No. 2, pp. 211–225
Statistics Canada

County Estimates of Wheat Production

ELIZABETH A. STASNY, PREM K. GOEL
and DEBORAH J. RUMSEY[1]

ABSTRACT

Although farm surveys carried out by the USDA are used to estimate crop production at the state and national levels, small area estimates at the county level are more useful for local economic decision making. County estimates are also in demand by companies selling fertilizers, pesticides, crop insurance, and farm equipment. Individual states often conduct their own surveys to provide data for county estimates of farm production. Typically, these state surveys are not carried out using probability sampling methods. An additional complication is that states impose the constraint that the sum of county estimates of crop production for all counties in a state be equal to the USDA estimate for that state. Thus, standard small area estimation procedures are not directly applicable to this problem. In this paper, we consider using regression models for obtaining county estimates of wheat production in Kansas. We describe a simulation study comparing the resulting estimates to those obtained using two standard small area estimators: the synthetic and direct estimators. We also compare several strategies for scaling the initial estimates so that they agree with the USDA estimate of the state production total.

KEY WORDS: Non-probability sample; Regression; Simulation; Small area estimation.

1. INTRODUCTION

County estimates of farm production are more and more in demand by government agencies for use in local economic decision making and by companies selling fertilizers, pesticides, crop insurance, and farm equipment. The United States Department of Agriculture (USDA) is currently implementing a program to standardize and improve county estimates of farm production (Bass *et al.* 1989). County estimation programs in the past have been carried out individually within each state. Because of this there has been little consistency across states in data collection and estimation methods used to produce county estimates. The goal of the USDA program for county estimation is to provide a set of sampling and estimation procedures for the states so that county estimation programs across the United States may yield estimates of comparable quality.

The new USDA county estimation program encompasses every stage of the production of county estimates from the construction of sampling frames through the estimation itself. The research described here is concerned only with the estimation of bushels of wheat produced. We hope, however, that our methods may prove useful in other aspects of the county estimation program, for example in estimating acres planted and for crops other than wheat.

Although the county estimation procedures used in the past varied from state to state, some parts of the procedures were similar. A typical procedure involved obtaining initial estimates from the data available within each county. Then an expert would review the estimates, alter them in light of his personal knowledge of the farms in the sample, weather conditions, and other factors, and then note the implications of the adjustments on the estimated total production for the state. The expert might repeat this process for a number of iterations until the

[1] Elizabeth A. Stasny, Prem K. Goel and Deborah J. Rumsey, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, USA.

estimates within each county seemed reasonable and the resulting state production total agreed with the USDA state estimate. (The USDA state total is estimated based on a large probability sample and is thus thought to be a more accurate estimate than the total based on the county estimation procedure. For this reason, states typically constrain their county estimates to sum to the USDA estimate.)

Written documentation of the current county estimation procedures, as outlined above, typically is not available. Thus the assumptions and methods that the expert uses can not be inspected by others and it is practically impossible to study the procedures or replicate calculations. In addition, one cannot obtain variance estimates or use the procedures of one state in another state. New methods for county estimation must address these problems.

The data that we use in this research were collected in Kansas in 1987, before the new USDA county estimation sampling procedures were in use. Data from 1987 were used because the United States Agricultural Census was taken in that year and we may, therefore, use the Census data in our estimation procedure. Kansas data were chosen for use in this study because the county data collection program in Kansas was one of the more comprehensive programs in the United States. Nevertheless, the data used for county estimation in Kansas, as in most other states, were not collected from a probability sample of farms. Therefore, our estimation procedure must not require a probability sample of wheat farms. Such a procedure may also be useful under the new county estimation program since states still will not be required to choose probability samples of farms.

There is much recent research on small area estimation (see for example Platek *et al.* 1987). Standard small area estimation procedures, however, require known selection probabilities since the inverses of these probabilities are used to weight observations in standard estimators such as synthetic and direct estimators. (See for example Section 2 of Särndal and Hidiroglou 1989 for a discussion of standard small area estimators.)

The methods considered here must be different from the usual small area estimation techniques. First, the sample of farms available to produce county estimates is not typically a probability sample. Second, the county estimates must be constrained to sum to the USDA-produced state totals. Since most state agriculture departments currently do not have large computing facilities, an additional preliminary constraint on the estimation procedure is that computations must be simple enough to be performed on a personal computer. Thus, for our initial efforts, we prefer to avoid computationally intensive estimators such as those described by Fay and Herriot (1979). For these reasons, we consider a computationally simple estimator based on a regression model for producing county estimates of wheat production.

In Section 2 of this paper we describe the Kansas data bases used in this study. Section 3 presents the regression procedure for estimating wheat production while Section 4 describes several methods for scaling those estimates to the USDA state total. In Section 5 estimates from the regression models are obtained and compared to the published county estimates and to estimates produced using the synthetic estimator and the direct estimator. In Section 6 we present the results of a simulation study conducted to compare these same estimators. Section 7 gives conclusions and areas for future research.

2. KANSAS DATA

For the purpose of reporting farm production, all states are divided into nine or ten districts. Kansas is divided into nine districts such that each of the 105 counties in Kansas is completely contained within one of the districts. The locations of the districts and the number of counties within each district are as shown below:

District Number	District Location	Countries in District
1	Northwest	8
2	West Central	9
3	Southwest	14
4	North Central	11
5	Central	11
6	South Central	13
7	Northeast	11
8	East Central	14
9	Southeast	14

Two data bases which are used in the production of Kansas county estimates, the Planted Acres Data Base and the Small Grain Data Base, were available for our use in this research. Most of our work was done with 1987 data but we also verified our results with the 1988 data. The 1987 Planted Acres Data Base contains information on planted acreage for 37,094 farms throughout Kansas. (A farm is defined by USDA to be any place with annual sales of agricultural products of $1,000 or more.) Of these farms, the 22,300 that reported planting some wheat were used in the simulation study described in Section 5. The 1987 Small Grain Data Base contains production information for 5,802 farms which reported planting small grain crops. Of these, the 1,707 that reported planting some wheat were used in our study.

Records on the Planted Acres Data Base are a composite of Kansas farm data from a number of sources collected at a number of times. First a list of names and addresses of farms is created using data collected by county appraisers. This data may be replaced and/or corrected using data from the Quarterly Agricultural Surveys and from Monthly Farm Reports. The Quarterly Agricultural Surveys use stratified systematic samples of approximately 2,600 farms. The response rate is approximately 80%. The Monthly Farm Report is completed by about 3,000 farmers who have agreed to file the reports. The same farmer may complete monthly reports for many years. The most recent data for each item appears in the Planted Acres Data Base and the record for any one farm in any year may contain information from a number of sources.

The 1987 Small Grain Data Base contains information on acres planted, acres harvested, and bushels produced for farms responding to the Quarterly Agricultural Surveys and the Kansas Small Grain Survey. About 6,000 surveys were mailed to a random sample of farms for the 1987 Kansas Small Grain Survey; about 50% of the surveys were completed and returned.

In addition to the potential problem with nonresponse bias in the Small Grain Data Base, there is typically a problem with response bias. The production reported by farmers is often lower than the actual production. The non-standard sample, nonresponse bias, and response bias lead us to develop the county estimation procedure described in the following sections.

3. REGRESSION MODELING

We propose the development of a regression model for use in producing county estimates. The calculations for fitting a multiple regression model can be performed using a number of statistical packages available for personal computers. In addition, our proposed estimator allows for the fact that we do not have a probability sample of farms and will produce county estimates that sum to the desired state total.

The steps in our procedure are as follows:

1) Use multiple regression to model the relationship between farm production and some predictor variables using the non-probability sample of farms.
2) Assume that the regression relationship holds for the entire population of farms in the state, and estimate farm production for all farms in each county.
3) Adjust the estimates of farm production to sum to the USDA state total.

To describe the regression model we need the following notation. For $i = 1, 2, \ldots, I$ ($I = 105$ counties in Kansas) and $j = 1, 2, \ldots, n_i$ let

n_i = number of farms from i^{th} county in sample;

$n = \sum_{i=1}^{I} n_i$ = total sample size;

N_i = total number of farms from i^{th} county in population;

$N = \sum_{i=1}^{I} N_i$ = total number of farms in population;

Y_{ij} = wheat production of j^{th} farm in i^{th} county (in bushels);

$X_{ij} = (1\ X_{ij1}\ X_{ij2} \ldots X_{ijp})$ = vector of p predictors for j^{th} farm in i^{th} county.

It is important, as we will see later, to choose predictor variables for which county totals are known or for which very accurate estimates of the county totals are available. The predictor variables must also include information related to the probability that a farm is included in the sample, such as a measure of the size of a farm. This will allow us to use the regression model to adjust for the fact that the sample is not a probability sample.

We consider regression models of the form

$$Y_{ij} = f(X_{ij} \mid \beta) + \epsilon_{ij},$$

where $\beta = (\beta_0\ \beta_1\ \beta_2 \ldots \beta_p)$ is a vector of parameters and ϵ_{ij} is a random error term with variance σ^2. Let the fitted values, which will be obtained using data from the Small Grain Data Base, be denoted by

$$\hat{Y}_{ij} = f(X_{ij} \mid \hat{\beta}).$$

Then the county total for the i^{th} county may be estimated as follows:

$$\hat{Y}_{i+} = \sum_{j=1}^{N_i} \hat{Y}_{ij} = \sum_{j=1}^{N_i} f(X_{ij} \mid \hat{\beta}),$$

where a "+" in a subscript indicates summation over the corresponding subscript.

For a general form of $f(X_{ij} \mid \beta)$, it would be necessary to know the value of X_{ij} for all farms in the i^{th} county. It is, of course, not possible to have such extensive information. If, however, $f(X_{ij} \mid \beta)$ is a linear function, then we only need to know county totals of the predictor variables. This is the case since, for a linear regression equation,

$$\hat{Y}_{i+} = \sum_{j=1}^{N_i} \hat{Y}_{ij} = \sum_{j=1}^{N_i} [\hat{\beta}_0 + \hat{\beta}_1 X_{ij1} + \hat{\beta}_2 X_{ij2} + \ldots + \hat{\beta}_p X_{ijp}]$$

$$= \hat{\beta}_0 N_i + \hat{\beta}_1 X_{i+1} + \hat{\beta}_2 X_{i+2} + \ldots + \hat{\beta}_p X_{i+p},$$

where X_{i+k} is the total of the k^{th} predictor for the i^{th} county.

The $\hat{Y}_{i+}$ will be reasonable county estimates if the regression model describes the relationship between the predictor variables and production for all farms in each county as well as for the farms in the data base. These county estimates, however, will not necessarily sum to the USDA state total for production. Methods for resolving this problem will be considered in Section 4.

In addition to providing county estimates of farm production, the linear regression model proposed above also permits us to obtain variance estimates. This is easiest to see if we write the county estimates in terms of matrices. Let

$\boldsymbol{X}$ = $n \times (p + 1)$ matrix of actual data with rows being the X_{ij} defined above;

$\boldsymbol{Z}$ = (unknown) $N \times (p + 1)$ matrix of predictor variables for all farms in the state;

$\hat{\boldsymbol{Y}}$ = (unknown) $N \times 1$ vector of estimates of wheat production for all N farms in state;

$\boldsymbol{B}_i$ = $N \times 1$ column vector with elements b_{ij}

$$\text{where } b_{ij} = \begin{cases} 1 \text{ if the } j^{\text{th}} \text{ farm is in the } i^{\text{th}} \text{ county} \\ 0 \text{ otherwise} \end{cases}$$

$\boldsymbol{A}$ = $[\boldsymbol{B}_1 \boldsymbol{B}_2 \boldsymbol{B}_3 \ldots \boldsymbol{B}_I]_{N \times I}$.

The estimation procedure described above does not provide $\hat{\boldsymbol{Y}}$ but instead provides a vector of county estimates $\hat{\boldsymbol{Y}}_{i+} = \boldsymbol{A}^T \hat{\boldsymbol{Y}}$, where "$T$" indicates the transpose of a matrix.

The variance for the county estimates is thus

$$\text{Var}(\hat{\boldsymbol{Y}}_{i+}) = \text{Var}(\boldsymbol{A}^T \hat{\boldsymbol{Y}}) = \boldsymbol{A}^T \text{Var}(\hat{\boldsymbol{Y}}) \boldsymbol{A} = \boldsymbol{A}^T \text{Var}(\boldsymbol{Z}\hat{\beta}) \boldsymbol{A} = \sigma^2 \boldsymbol{A}^T \boldsymbol{Z} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{Z}^T \boldsymbol{A}.$$

Although $\boldsymbol{Z}$ itself is unknown, the product $\boldsymbol{A}^T \boldsymbol{Z}$ is a known matrix containing only the numbers of farms in a county, N_i, and the county totals, X_{i+k}, for the predictor variables. Thus, if we use the regression mean square error (mse) as an estimate of σ^2, we may obtain estimates of the variances of the county estimates. Variance estimates for county estimates have not previously been available.

The estimator based on a regression model as described in this section meets the requirements for a computationally simple estimator from a non-probability sample. In the following section we consider methods to adjust the estimates to sum to the USDA state totals for farm production.

4. SCALING ESTIMATES TO SUM TO STATE TOTAL

Let Y be the USDA's estimated total wheat production for Kansas. In general, $\sum_{i=1}^{I} \hat{Y}_{i+} \neq Y$. Thus, we define new estimates

$$\tilde{Y}_{i+} = c_i \hat{Y}_{i+},$$

where the c_i are constants such that $\sum_{i=1}^{I} \tilde{Y}_{i+} = \sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$. An important question is how to choose the c_i. Current methods used for county estimation take $c_i = c$ (at the district level) and thus adjust all estimates by a common proportion. Instead, one could choose the c_i to minimize the sum of the squared differences or relative differences between the $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$. Values of c_i and $\tilde{Y}_{i+}$ for three criterion for choosing c_i are given below.

1) **Choose $c_i = c$**

If c_i is taken to be a constant, then it is easy to show that

$$c_i = c = Y \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}$$

and

$$\tilde{Y}_{i+} = Y\left(\hat{Y}_{i+} \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}\right).$$

2) **Choose c_i to minimize the sum of squared differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$**

To choose c_i to minimize the sum of the squared differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$ subject to $\sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$, we must minimize $\sum_{i=1}^{I} (\tilde{Y}_{i+} - \hat{Y}_{i+})^2 = \sum_{i=1}^{I} (c_i \hat{Y}_{i+} - \hat{Y}_{i+})^2$ with respect to c_i using a Lagrange multiplier to impose the desired constraint. Doing this, we find that

$$c_i = 1 + \left[\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \hat{Y}_{i+}^2 \sum_{i=1}^{I} (1/\hat{Y}_{i+})\right]$$

and

$$\tilde{Y}_{i+} = \hat{Y}_{i+} + \left[\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \hat{Y}_{i+} \sum_{i=1}^{I} (1/\hat{Y}_{i+})\right].$$

Note that the scaled estimates, $\tilde{Y}_{i+}$, are obtained by adjusting the original estimates, $\hat{Y}_{i+}$, by adding a factor which is a proportion of the difference between the USDA state total and the sum of the original county estimates. The proportion is based on the harmonic mean of

the original estimates. Although some of these scaled estimates could be negative in theory, this is not considered likely in practice because farmers often underreport the amount of production on their farms. If the total of the original estimates exceeds the USDA state total, then scaled estimates corresponding to counties with small original estimates may be negative.

3) Choose c_i to minimize the sum of squared relative differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$

To choose c_i to minimize the sum of the squared relative differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$ subject to $\sum_{i=1}^{I} c_i \hat{Y}_{i+} = Y$, we must minimize $\sum_{i=1}^{I} [(\tilde{Y}_{i+} - \hat{Y}_{i+})/\hat{Y}_{i+}]^2 = \sum_{i=1}^{I} (c_i - 1)^2$ with respect to c_i using a Lagrange multiplier to impose the desired constraint. Doing this, we find that

$$c_i = 1 + \left[\hat{Y}_{i+}\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}^2\right]$$

and

$$\tilde{Y}_{i+} = \hat{Y}_{i+} + \left[\hat{Y}_{i+}^2\left(Y - \sum_{i=1}^{I} \hat{Y}_{i+}\right) \Big/ \sum_{i=1}^{I} \hat{Y}_{i+}^2\right]$$

The scaled estimates, $\tilde{Y}_{i+}$, are again obtained by adjusting the original estimates, $\hat{Y}_{i+}$, by adding a factor which is a proportion of the difference between the USDA state total and the sum of the original county estimates. The proportion here is based on the squared values of the original estimates. As in method 2, these scaled estimates may be negative, although it is unlikely in practice.

Note that we have chosen to consider the difference $\tilde{Y}_{i+} - \hat{Y}_{i+}$ relative to $\hat{Y}_{i+}$ rather than to $\tilde{Y}_{i+}$. This choice was made because in the later case the estimator, $\tilde{Y}_{i+}$, does not have a closed-form solution. Thus, to meet the goal of developing computationally simple estimators, we chose to consider the difference $\tilde{Y}_{i+} - \hat{Y}_{i+}$ relative to $\hat{Y}_{i+}$.

In the following section we will consider the effects of these three scaling methods on the county estimates of wheat production.

5. COMPARISON OF ESTIMATES OF WHEAT PRODUCTION

We used a linear regression model, as described in Section 3, to model the relationship between wheat production (measured in total bushels produced) and some predictor variables for farms in the 1987 Small Grain Data Base. The possible predictor variables that we considered included: acres planted in wheat, acres of wheat harvested, a prediction of wheat production based on the 1986 county estimates, acres of irrigated wheat, acres of non-irrigated wheat, indicators of the district in which the farm is located, indicators of region of the state (east, central, west), and interaction terms.

The most important predictor variables for the regression model were acres planted in wheat and some indicator of the location of the farm within the state. The variable based on the previous year's county estimates did not seem to be a useful predictor for the amount of wheat produced on a farm in the current year. Because other possible predictor variables, such as irrigated acres, are not known as accurately at the county level, we decided that acres planted would be the single continuous predictor variable included in the model. Not all district indicators were needed in the regression model; that is, some districts were similar and could

Table 1
Regression Models Fitted to Actual Data

	Fitted Models	R^2	√mse
Model 1	Bushels = $-811 + 32(\text{Pla}) + 3{,}248I_1 + 3{,}088I_2 + 2{,}190I_3 + 2{,}526I_4 + 1{,}241I_5 - 562I_6 + 1{,}047I_7 + 399I_8$	85	5,945
Model 2	Bushels = $-281 + 28(\text{Pla}) + 138I_1 + 1{,}861I_2 + 2{,}328I_3 + 329I_4 - 359I_5 - 334I_6 - 42I_7 + 500I_8 + 11(\text{Pla})I_1 + 5(\text{Pla})I_2 + 3(\text{Pla})I_3 + 11(\text{Pla})I_4 + 9(\text{Pla})I_5 - 0.2(\text{Pla})I_6 + 15(\text{Pla})I_7 - 7(\text{Pla})I_8$	86	5,818

Note: Pla is planted acres, I_i is the indicator variable for the i^{th} district.

have been grouped together. We decided, however, to include all district indicators in the model since groupings of districts might change from year to year or might be different for crops other than wheat.

We chose to focus our study on two possible regression models: Model 1 contained acres planted in wheat and the district indicators while Model 2 contained these same variables and the interaction terms involving acres planted and the indicator variables. The models and measures of their fits are shown in Table 1. Although the root mean squared errors did not differ considerably for the two models, we felt that the difference might be magnified when the models were used to estimate farm production for the entire state. Thus, in the following, we obtain and compare estimates from both models.

To verify that these regression models are not simply a result of some unusual feature in the 1987 Kansas Small Grain Data Base, we used the same set of possible predictor variables and searched for reasonable regression models using the 1988 data. The fits of Models 1 and 2 to the 1988 data are similar to the 1987 fits and no other model appeared to be superior for fitting the 1988 data. The estimates for the parameter corresponding to acres of wheat planted were fairly similar in both 1987 and 1988, but the parameters corresponding to the indicator variables for districts showed considerable change. We believe that the indicator variables for districts are reflecting the effects of weather and different farming practices in different parts of the state. For example, irrigation is more commonly used in western and central Kansas than in eastern Kansas. Although farming practices are not likely to change dramatically from one year to the next, weather conditions may be quite different. Thus, it seems reasonable that the contribution of the district variable in predicting wheat production could change considerably from year to year.

Both models were used to obtain county estimates for all 105 counties in Kansas. In Table 2, the unscaled estimates and their standard errors under both Models 1 and 2 are given for nine counties, one county chosen at random from within each district so that the nine counties are spread over the entire state. An inspection of Table 2 suggests that the estimated standard error for Shawnee county is an anomaly. The variance of a county estimate depends on the number of farms in the county, the total acres planted in wheat in the county, and the number of farms sampled from the district in which the county lies. District 8, in which Shawnee county is located, had relatively few farms in the Small Grain Data Base. The county has a moderate number of farms growing wheat but these farms are small in terms of acres planted. These three factors together result in the rather large standard error for the estimates from Shawnee county.

Table 2
Regression Model Estimates for Nine Counties in Kansas

District	County	Estimated Bushels of Wheat Produced (in thousands of bushels)	
		Model 1 (no interaction terms)	Model 2 (with interaction terms)
1	Decatur	4,944 (180)	4,778 (179)
2	Trego	4,378 (174)	4,229 (188)
3	Hodgeman	4,808 (123)	4,908 (125)
4	Jewell	5,555 (275)	5,550 (269)
5	Marion	5,144 (313)	4,931 (315)
6	Comanche	2,615 (59)	2,480 (63)
7	Leavenworth	231 (53)	262 (61)
8	Shawnee	232 (106)	226 (104)
9	Butler	2,374 (331)	2,272 (338)

Note: Standard errors are given in parentheses below each estimate.

The estimates shown in Table 2 are reasonably similar to the published county estimates (Kansas Agricultural Statistics 1988). While it is encouraging that our estimates are not wildly different from those published by Kansas, there is no theoretical basis for using the Kansas estimates as a standard. Thus, we carried out a simulation study to help us evaluate our estimators. This study is described in the following section.

6. SIMULATION STUDY

6.1 The Estimators to be Compared

In the simulation study, we compared the estimates from our two regression models with those from two standard small area estimators: the synthetic and direct estimators. (See, for example, Section 2 of Särndal and Hidiroglou (1989) for a discussion of standard small area estimators, including the synthetic and direct estimators.) The synthetic estimates are obtained by allocating the state total for wheat production to the counties according to the proportion of total acres planted in wheat within each county. The direct estimates are obtained using only the sampled farms in a county to estimate wheat production for that county.

We expect the synthetic estimates to have a large amount of bias because counties in different parts of the state have different farming practices and different weather conditions, while the synthetic estimator treats each county as if it were representative of the entire state. The synthetic estimates, however, will have relatively small variances because they are obtained using all the data from the entire state.

Since the direct estimate for a county is based only on the sample data within that county, it will have a relatively large variance but it should have smaller bias than the synthetic estimate. At least one farm from a county must appear in the sample to make it possible to obtain an estimate for that county, and at least two farms are needed in the sample to make variance estimation possible. In the 1987 Kansas Small Grain Data Base, three counties had no wheat farms in the sample and three additional counties had only a single farm in the sample. Although we are comparing our regression model estimates to the synthetic and direct estimates, it should be noted that the latter two estimators require that the data be from a probability sample. This requirement is not met by the Kansas data.

Table 3

Numbers of Farms and Production Levels by District and Planted Acres

District		Planted Acres in Farm				
		0-99	100-249	250-499	500-999	≥1,000
1	M_i*	354	638	531	302	85
	m_i*	27	45	51	40	9
	bu/pa*	34.68	37.18	37.76	39.21	38.68
2	M_i	266	550	572	377	161
	m_i	27	49	47	55	33
	bu/pa	35.92	33.62	36.78	39.09	34.85
3	M_i	264	549	610	537	264
	m_i	31	80	76	98	61
	bu/pa	26.93	32.84	35.03	36.79	33.13
4	M_i	956	939	626	271	50
	m_i	62	37	23	21	7
	bu/pa	36.81	36.91	39.70	39.87	39.41
5	M_i	1,236	1,529	912	350	54
	m_i	92	93	51	26	3
	bu/pa	31.79	32.25	31.69	36.85	33.65
6	M_i	1,181	1,427	1,160	793	249
	m_i	96	96	81	55	20
	bu/pa	26.24	26.88	28.78	27.87	26.72
7	M_i	957	242	67	9	3
	m_i	62	5	2	0	0
	bu/pa	33.87	40.81**	40.81**	40.81**	40.81**
8	M_i	1,126	251	52	9	1
	m_i	56	11	2	0	0
	bu/pa	26.02	11.48**	11.48**	11.48**	11.48**
9	M_i	1,122	431	166	59	12
	m_i	47	19	7	3	1
	bu/pa	23.57	23.87	27.63**	27.63**	27.63**

* M_i is the number of farms on the Planted Acres Data Base, m_i is the number of farms in the Production Data Base, and bu/pa is the ratio of bushels produced to acres planted.
** Cells of this district were grouped to obtain bu/pa values.

6.2 The Simulated Population and Samples

We first simulated a population of wheat farms by generating production values for all 22,300 farms reporting acres planted in wheat on the Planted Acres Data Base. Because production rates appear to vary by district and size of farm (see Table 3), we generated bushels-per-planted-acres (bu/pa) from 37 different distributions. These distributions were based on the bu/pa data from the Small Grain Data Base. (Notice that in the eastern districts of Kansas, districts 7, 8 and 9, there were few or no sampled farms in several size-of-farm classifications. Those classifications were grouped as indicated in Table 3 for the purpose of simulating bu/pa values.) Histograms of the observed bu/pa from the Small Grain Data Base were generated by district and five sizes of farm: 0-99, 100-249, 250-499, 500-999, and 1000 or more acres of wheat planted. Since these histograms generally appeared mound-shaped, we chose to use normal distributions to model the distributions of the bu/pa. The means and variances of the normal distributions were taken to be the sample means and variances of bu/pa from the wheat farms in the Small Grain Data Base within the 37 district by size-of-farm classifications.

After the bu/pa values were generated from the appropriate normal distributions for each farm, the bushels of wheat produced were obtained by multiplying the simulated bu/pa by the reported acres planted in wheat for each farm. Ten samples were generated from the resulting simulated population. Since there was no sampling design to follow in creating these samples, we sampled each farm within the district by size-of-farm classifications with probabilities equal to the observed frequencies with which farms on the Planted Acres Data Base appeared in the Small Grain Data Base. That is, farms within classification C, say, were chosen to be in the sample with probability equal to

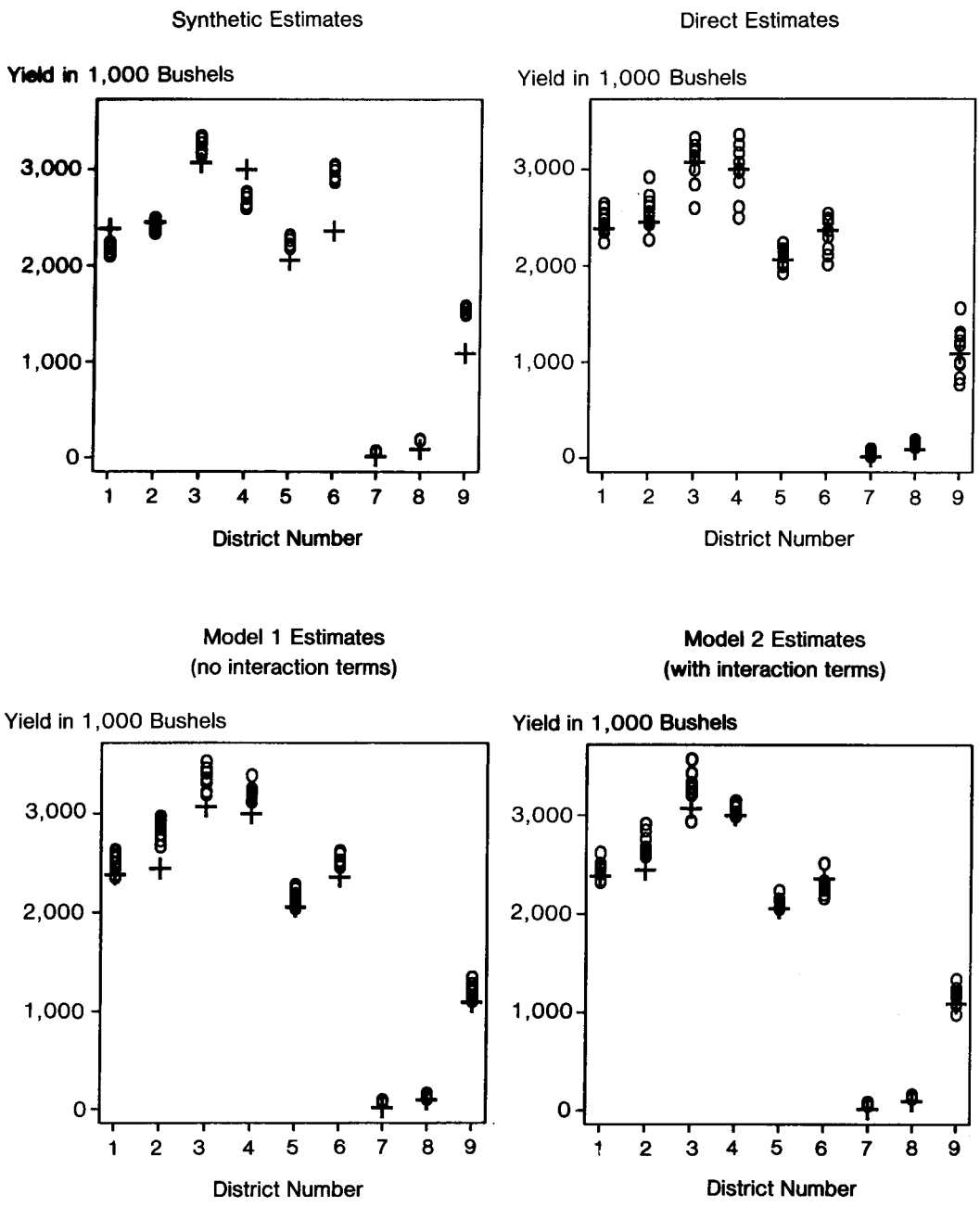
$$\frac{\text{Number of farms in classification } C \text{ in Small Grain Data Base}}{\text{Number of farms in classification } C \text{ in Planted Acres Data Base}}.$$

Our goal in using such a sampling scheme was to make the simulated samples as similar as possible to actual samples even though we do not know what the selection probabilities for the actual samples were.

6.3 Comparison of the Four County Estimators

We used the four county estimators (the two regression, the synthetic, and the direct) to obtain wheat production estimates for all 105 counties from each of the ten simulated samples. The resulting estimates were then compared to the "true" production values obtained for each county from the simulated population. This comparison allows us to evaluate the amounts of bias and variability in the estimates for each county. Figure 1 presents the values of all four estimates from each of the ten samples along with the true production values for the nine-randomly chosen counties, one from each district, which were previously mentioned in Section 5.

As expected, the synthetic estimates exhibit considerable bias. Indeed, only in district 2 does the range of estimates include the true population value. The ranges of the direct estimates are all larger than those of the synthetic estimates but those ranges do include the population values. The ranges of estimates from the regression models appear to be less than those of the direct estimates. For about half of the counties pictured in Figure 1, the estimates from Model 1 appear to exhibit some bias. The estimates under Model 2 seem to exhibit less bias. On the basis of this comparison of estimators we prefer Model 2, the regression model with the interaction terms.



Note: Estimates are for one county chosen at random from within each district.
o = estimate from one of the ten simulated samples,
+ = true value from simulated population.

Figure 1. Comparison of Estimators for Nine Counties

6.4 Comparison of the Scaling Methods

The same four sets of estimates for all counties from the ten sets of simulated samples were next scaled to agree with the state total from the simulated population using the three scaling methods described in Section 4. The resulting scaled estimates were compared to the true county production values for the simulated population. The comparison was made using the mean of the absolute value of relative error which is defined as follows:

$$(1/I) \sum_{i=1}^{I} \left| (\tilde{Y}_{i+} - Y_{i+})/Y_{i+} \right|.$$

Figure 2 shows the values for all ten samples of the mean over the 105 counties of absolute relative error. This error is given for all four estimators under no scaling and under each of the three methods of scaling.

From Figure 2A, we see that the scaling method which minimizes the sum of squared differences produces very poor final estimates; the average of absolute relative differences between the final estimates and the county production values for the simulated population is quite large compared to that of the other scaling methods. This large error results from the fact that the total wheat production in one county may be quite different from that in another county. Since the scaling procedure minimizes the squared differences between the original and the final estimates, a county with a very small original estimate may have a final estimate that is changed considerably relative to the original estimate. These large changes in estimates do not seem warranted; hence we drop this method of scaling from consideration.

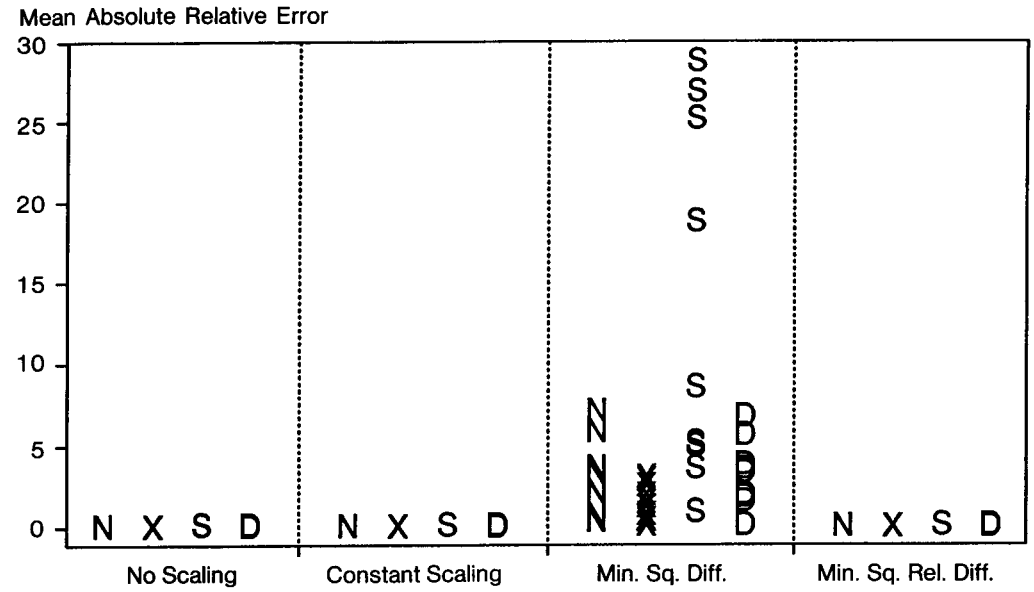
Figure 2B, a refinement of Figure 2A, provides a more detailed comparison of the four estimators under no scaling and under the two remaining scaling methods. We see from this figure that the error is generally smallest for the regression model with the interaction terms. This supports our choice of Model 2 in the previous subsection. In addition, Figure 2B suggests that there is little difference between the original unscaled estimates and the final estimates under either scaling method. In fact, the total of the original county estimates is not far from the simulated population total. Thus, the scaling constants, c_i, are all quite close to one. Since the two methods of scaling produce similar estimates, there is no reason to use the more difficult scaling method; the constant scaling method may be used.

7. CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

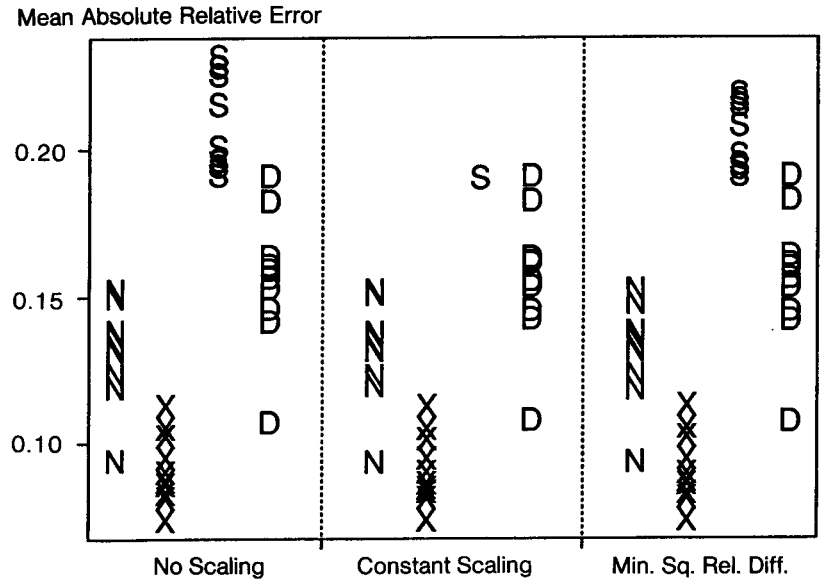
We have shown that a regression model may be used to obtain reasonable county estimates of wheat production. The model we selected used acres planted, district indicators, and interaction terms as predictor variables. The regression model does not require a probability sample of farms and it does permit the estimation of variances of the county estimates. The estimates based on the regression model may be scaled to agree with state total production using a constant scaling factor since the alternative scaling method did not produce markedly different county estimates.

Many areas for future research in county estimation of farm production remain. For example, the county estimates from our simulation study suggested that the inclusion or exclusion of large farms (1,000 or more acres of wheat planted) from the sample for a district could have a large effect on the estimates for counties in that district. This was particularly true for

A.



B.



Note: Data are from the simulated samples.
N = Model 1 Estimator (no interaction terms),
X = Model 2 Estimator (with interaction terms),
S = Synthetic Estimator,
D = Direct Estimator.

Figure 2. Comparison of Estimators and Scaling Methods

districts which had few of these larger farms. Since large farms most likely account for a sizable proportion of farm production, it might be worthwhile to handle large farms separately in a county estimation procedure. States might also consider altering their sampling plans so that the largest farms are included in the samples with certainty.

Additional work is needed to determine whether a regression model similar to that developed for wheat is appropriate for other crops as well. In particular, it would be useful to discover if such models can be used for rare crops where there is much less available data. We should also note that the similarity in the state total and the total of the county estimates, which was observed for the actual data as well as for the simulated samples, may be characteristic of wheat production but not of all crop production. Future research should consider whether other crops require a scaling method other than constant scaling.

We chose to begin our research on the county estimation problem by studying methods of estimating production. An additional problem for future research is the estimation of total acreage planted for various crops. In this research we used 1987 agricultural census data to provide the needed information on numbers of farms and acres planted in wheat within each county. The agricultural census, however, is taken only every five years. In the intermediate years, changes in numbers of farms and acres planted must be estimated from sample data. We expect such changes in census values to be small for major crops like wheat in Kansas, but we anticipate greater difficulty estimating these quantities for less common crops.

Finally, the requirement for a computationally simple estimator, which led us to propose an estimator based on a regression model, may no longer be necessary as state agricultural offices are being linked to a large, national computer system. Thus, in our future research on county estimates of farm production, we plan to consider more computationally intensive small-area estimators.

ACKNOWLEDGEMENTS

This research was supported in part by the United States Department of Agriculture under Cooperative Agreement No. 58-3AEU-9-80040. The authors take sole responsibility for the contents of this paper. The authors wish to thank Gary Keough and Leland Brown at USDA and Ronald Sadler, Melvin Perrott, Eldon Thiessen, and M. E. Johnson at the Kansas Department of Agriculture for their help on this project. We also thank the referee and associate editor for their helpful comments on an earlier version of this paper.

REFERENCES

BASS, J., GUINN, B., KLUGH, B., RUCKMAN, C., THORSON, J., and WALDROP, J. (1989). Report of the Task Group for Review and Recommendations on County Estimates. USDA National Agricultural Statistics Service, Washington, D.C.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

KANSAS AGRICULTURAL STATISTICS. (1988). Kansas Farm Facts, prepared by the Statistical Division of the Kansas Department of Agriculture in cooperation with the National Agricultural Statistics Service of the U. S. Department of Agriculture, Topeka, Kansas.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.E., and SINGH M.P. (Eds.) (1987). *Small Area Statistics*. New York: John Wiley & Sons.

SÄRNDAL, C.E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.