United States Department of Agriculture

Statistical Reporting Service

Research Division

Staff Report AGES 811119

November, 1981

The Development of County Estimates in North Carolina

Barry L. Ford

The Development of County Estimates In North Carolina. By Barry L. Ford; Research Division; Statistical Reporting Service; U.S. Department of Agriculture; Washington, D.C. 20250; November, 1981. Staff Report AGES811119.

ABSTRACT

The purpose of this report is to describe and evaluate the survey design and county estimators for the 1979 Probability Crop and Livestock Survey (PCLS) in North Carolina. Using data from the 18,361 respondents to the 1979 PCLS, this study analyzed three county level estimators--the direct, synthetic, and composite estimators. The direct estimator had a smaller mean square error than the synthetic estimator. This outcome resulted from the large sample size of the PCLS and the bias of the synthetic estimator. For samples yielding less than 5000 respondents, analysis showed that for some variables the synthetic estimator would probably have a mean square error less than or equal to that of the direct estimator. For the data set in this study the composite estimates differed very little from the direct estimates.

CONTENTS

Summary

This study evaluated the survey design and county estimators for the 1979 Probability Crop and Livestock Survey (PCLS) in North Carolina. Results of the analysis showed that:

- o four strata were almost as efficient as ten
- o the allocation which was used was effective
- o the subsampling plan saved $10,000 over a comparative design involving no subsampling
- o an estimated 4 percent of the units on the list frame had incorrect county codes
- o operations with livestock in more than one county were estimated at only 0.1 percent of the population

Three county estimators were evaluated--direct, synthetic, and composite. The direct estimator was better than the synthetic estimator because it had a smaller mean square error. This outcome resulted from the large number of respondents (18,361) and the bias of the synthetic estimator. However, for samples with less than 5000 respondents, analysis indicated that for some variables the mean square error of the synthetic estimator may be less than that of the direct estimator. For the data set in this study the composite estimator, although more complicated to compute than the other two estimators, produced estimates and standard errors which were almost the same as the direct estimates.

Although the direct estimator was better for large samples than the synthetic estimator, coefficients of variation (CV) associated with it were still large. This study showed that county estimates for most variables had CV's in the range 0.14-0.24. With CV's this large county estimates can fluctuate so much from year to year that time trends are unrecognizable. Thus, county estimates still need much improvement. Possibilities for improving county estimates include combining information from other surveys, keeping part of the sample in from one year to the next, using census data to model relationships among the county values, and using historical data to model time trends in the county values.

THE DEVELOPMENT OF COUNTY ESTIMATES IN NORTH CAROLINA

This paper has six sections. The first section is an introduction which explains the circumstances leading up to this study and previous research done on the topic of small area estimation. The second and third sections describe and evaluate, respectively, the survey design which was used. The fourth and fifth sections describe and evaluate, respectively, the county estimators. The last section summarizes the results and makes recommendations for future work.

Introduction

In 1978 the state legislature in North Carolina discontinued the annual state farm census and appropriated funds for a probability survey to yield county estimates of crops and livestock. The North Carolina State Statistical Office (NCSSO) decided to adapt its state acreage and production surveys for this purpose. In the fall there were actually two surveys of acreage and production--an early fall survey collecting data on planted and harvested acreages and a late fall survey collecting data on harvested acreages, production, and prices. The plan was to redesign these two surveys on a probability basis that would allow for county estimates. The two redesigned surveys were collectively called the Probability Crop and Livestock Survey (PCLS).

The problem of constructing county or other small area estimates from survey data has been an important topic for many survey organizations throughout the history of survey sampling. Traditionally, large-scale data collection was used to solve the problem [11]. Occassionally during the 1950's and 1960's other methods were tried. The Radio Listening Survey described in Hansen, Hurwitz, and Madow [6] and a method used by Lillian Madow in a report for the Advertising Research Foundation [1] are two early examples. In 1968 in a report on disability in the United States, the National Center for Health Statistics first used "synthetic" estimates to make small area estimates [7]. During the 1970's a great deal of discussion and research went into small area estimates--most of the interest being devoted to the study of synthetic estimates.

An important result of this interest in synthetic estimation was the Workshop on Synthetic Estimates for Small Areas, which was cosponsored in 1978 by the National Institute on Drug Abuse and the National Center for Health Statistics. This workshop allowed the presentation of many important papers on both the theory and application of small area estimates. The workshop also served as a forum for discussion among private and government agencies interested in this topic. The papers and discussions from the workshop were published in a monograph in 1979 [8].

Three estimators for making small area estimates are discussed in this monograph. The first is the direct estimator. This estimator only uses whatever sample units fall in an area to make an estimate for that area. Although this estimator is mathematically unbiased and has "tremendous appeal to those individuals responsible for regional, state, and local planning" [9], it may be expensive or have large standard errors. The second is the synthetic estimator. This estimator uses subclass means from a large area, such as a state, and forms a small area estimate by calculating a sum in which each subclass mean is weighted by the proportion of the small area population which falls in that subclass. For example, to make a county estimate in North Carolina, a statistician might use the strata from the survey design as the subclasses.

The synthetic estimate for a particular county would then be a sum in which the estimated state mean for each stratum is weighted by the proportion of the county's population in the stratum. Use of this synthetic estimator assuming that for each stratum the county mean is equal to the state mean. Paul Levy has noted [9] that not only do synthetic estimates have an "intuitive appeal", but also they are "generally easy and inexpensive" to obtain. The third estimator, the composite estimator, combines the direct and synthetic estimators by weighting them according to the mean square errors involved [10]. The intention of this study was to investigate and compare these three estimators.

Description of the Survey Design

This section describes the survey design of the 1979 PCLS in detail because of the importance of the survey design to the county estimates. These details cover: 1) the timing of the survey, 2) types of information collected, 3) the stratification, 4) the allocation, 5) the initial selection of the sample, 6) the subselection strategy used on the sample, 7) the adjustments for missing data, and 8) a proposed nonoverlap estimate to measure incompleteness of the list frame.

The NCSSO collected data for the 1979 PCLS at two different times. Half the sample was surveyed in October, and the other half was surveyed in December. The October PCLS had a different questionnaire from the December PCLS. Copies of both questionnaires are in Appendix A. Although both questionnaires asked about harvested acreages, the October PCLS also asked about planted acreages while the December PCLS asked about amounts harvested, amounts sold or to be sold, and average price per unit. Questions on the total land in the farm and the number of livestock were common to both questionnaires. Also common to both questionnaires were several questions which checked for specific problems that the NCSSO and the Survey Research Section thought might be frequent. These potential problems were: 1) the farm operation was actually in a different county than specified by the county code on the sampling frame, and 2) the livestock on land operated by the farm are located in more than one county.

In short, data for 92 variables were collected--90 quantitative variables and 2 qualitative variables. There were 27 variables collected only on the October PCLS, 30 variables collected only on the December PCLS, and 35 variables collected on both occasions. Therefore, it was planned that the 57 variables collected on only one occasion would have approximately half the sample size of the 35 variables collected on both occasions.

The sampling frame was a list of 93,434 possible farm operators. Each sample unit on the list frame was coded by county and by crop reporting district. Approximately 43 percent of the list units had a measure of the total acres in the farm. For this part of the list the total acres was used as a control variable to stratify the frame into three strata. The rest of the list was put into an "unknown acres" stratum. Because the optimum number of strata was unknown, these four strata were divided into a total of ten substrata. By proportionally sampling the substrata within each stratum, this structure allowed an analysis of the efficiency of four strata vs. ten strata. The optimal boundaries for the four strata and the ten substrata were found by using the "cumulative $\sqrt{f}$" rule [2]. These boundaries are shown in Table 1.

Table 1. Population Sizes and Sample Sizes for the 1979 Probability Crop and Livestock Survey.

Stratum	Sub-Stratum	Stratum Boundaries*	Population Size	% of Total Population	Sample Size	Sampling Rate Within Stratum %
1	1	unknown	40,005	42.9	13,702	34
2	2	1-49	10,099	10.8	2,892	28
	3	50-99	20,459	21.9	5,856	28
	4	100-199	13,683	14.6	5,596	42
3	5	200-399	6,234	6.7	2,550	42
	6	400-599	1,475	1.6	604	42
	7	600-899	767	0.8	416	54
	8	900-2999	603	0.6	326	54
4	9	3000-8999	94	0.1	50	54
	10	9000 +	15	0.02	8	54
TOTAL			93,434	100 %	32,000	

* Stratum boundaries are based on the total number of acres in farm.

The total sample size of the surveys was 32,000--16,000 in October and 16,000 in December. This sample size was the maximum allowable under cost constraints and was the sample size which had previously been used for non-probabilit state acreage and production surveys. Of the 32,000 units, 13,702 units were allocated to stratum 1 by proportional allocation. The remaining 18,298 units in the sample were assigned to the other three strata using an optimum allocation [2] which took into account the population sizes and the variances of the control variable. Although the stratum variances for the control variable could have been obtained by summarizing the control values on the frame, the author actually made the conservative assumption of the uniform distribution in each stratum and estimated each stratum variance as 0.29 times the range of the control values [3]. This approximation was made to save time and trouble. As mentioned previously, proportional allocations were made to the substrata within each stratum. The resulting sample sizes are in Table 1.

To select the sample the NCSSO sorted all units in the frame by substratum, district, county, and identification number. Then six replicates were selected systematically within each substratum--three for October and three for December. The purpose of multiple replicates was to allow unbiased estimates of standard errors. The sorting and systematic selection guaranteed that each district and county was proportionally represented within each substratum.

Just before data collection began on the 1979 PCLS, the NCSSO was put under travel and cost restrictions. These restrictions resulted in a complex subsampling strategy used by the NCSSO. Figure 1 shows this strategy graphically. First, the NCSSO mailed questionnaires to the entire selected sample. Second, the

NCSSO separated those units which completed mail questionnaries from those units which were inaccessible by mail. Third, the NCSSO randomly selected half the units inaccessible by mail to receive telephone interviews. Fourth, within the group selected for telephone interviews, the NCSSO separated the units which completed telephone interviews from the units which were inaccessible by telephone. Fifth, the NCSSO randomly selected a third of the units inaccessible by telephone from the NCSSO to receive field interviews (where a field enumerator would try to contact a farm operator either personally or by telephone).

Weights were assigned to each unit with a completed questionnaire to reflect the subsampling strategy. Units completed by mail interview received a weight of "1", units completed by telephone interview from the NCSSO received a weight of "2", and units completed by field interview received a weight of "6". The units completed by field interview received a weight of "6" because they must represent themselves, the other two-thirds in the group which were inaccessible by telephone from the NCSSO, and that part of the "not selected for telephone interview from NCSSO" group which *would* have been telephone inaccessibles.

The effect of the subsampling strategy was a poststratification of the sample. Within each substratum there were three poststrata defined by whether a unit was enumerated by mail, telephone, or field interview. Thus, for state estimates the estimators and their standard errors had to reflect this subsampling structure. Formulas for estimators at the state level and their standard errors are in Appendix B.

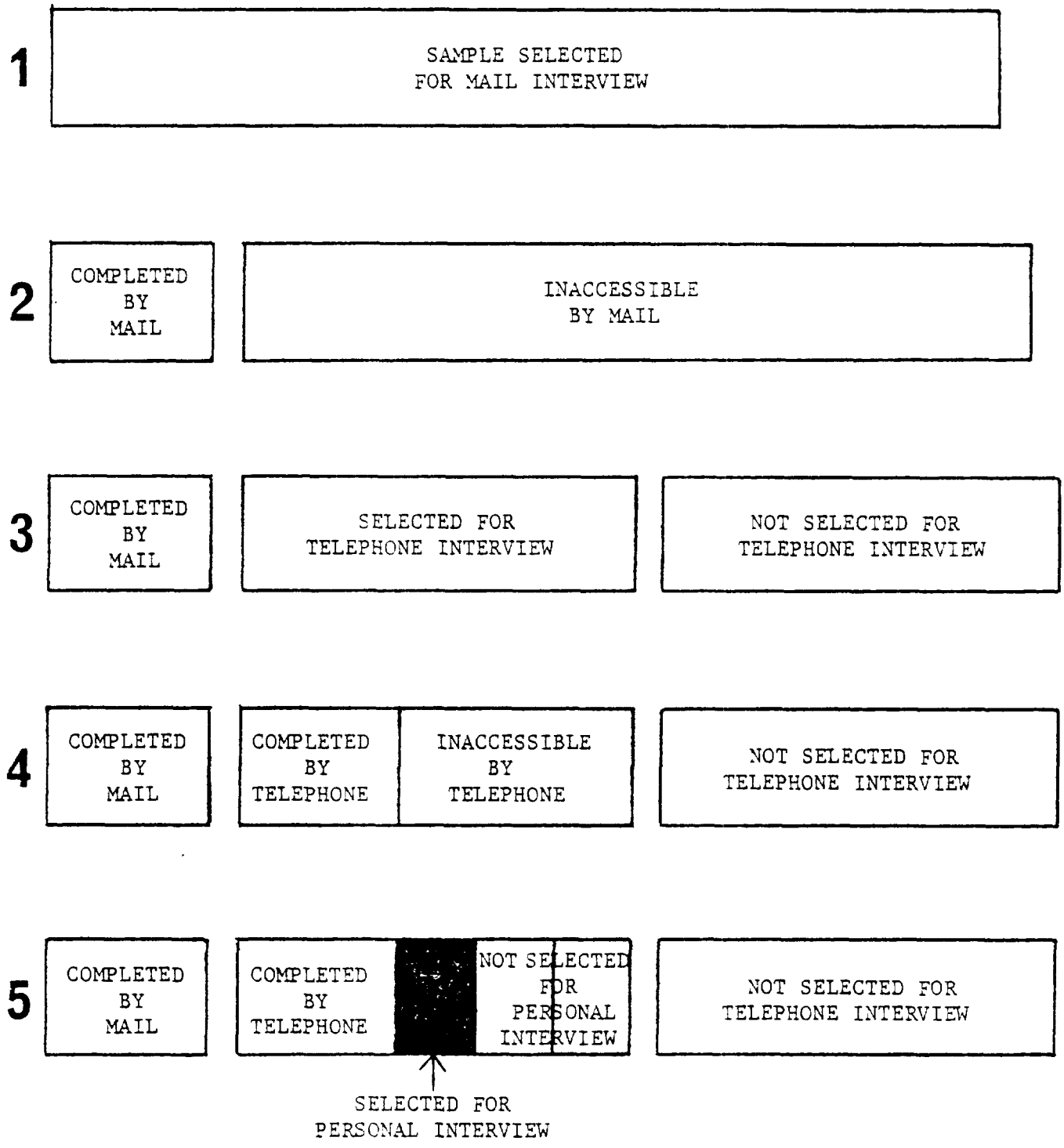
As in most surveys there were two types of missing data--missing units, i.e. refusals and inaccessibles, and missing items, i.e. for a particular unit most questions were answered but a few were missing. On the PCLS missing units were omitted from any estimation. This procedure assumes that the missing units are distributed in the same way as the units which reported information. Although this assumption is not strictly true, a more sophisticated approach to this problem should wait until the survey data can be analyzed for the effects of nonresponse--an analysis outside the scope of this study.

For the purposes of the PCLS a missing item was any value missing from a unit which had at least reported the total acres in the farm. Values were imputed for missing items by using ratio relationships computed from the data of the completed questionnaires. In general, imputations were done in a logical order which branched from the value for the total acres in the farm. For example, suppose questionnaire A reported the "total acres in the farm" but the "acres of oats harvested for grain" and the "amount of oats harvested for grain" were not reported. Then the procedure:

1: estimated the ratio "acres of oats for grain" ÷ "total farm acres" from the complete questionnaires
2: multiplied the ratio in step 1 by the "total farm acres" on questionnaire A to impute a value for the "acres of oats for grain"
3: estimated the ratio "amount of oats for grain" ÷ "acres of oats for grain" from the complete questionnaires
4: multiplied the ratio in step 3 by the "acres of oats for grain" imputed on questionnaire A in step 2 to impute a value for the "amount of oats for grain".

Thus, imputations were done by ratioing in a logical, but complex sequence.

Figure 1: Diagram showing the subselection of the initial sample.



There were two exceptions to this imputation procedure. First, missing values for livestock items were imputed by using the average of the nonzero values on the complete questionnaires. This procedure assumed that if a missing value should have been zero, then a "0" was hand edited on the questionnaire. Second, if an average price for a particular crop was missing on the December PCLS, then a weighted average price from the complete questionniares was imputed. The weight was the amount of the crop sold.

The intention of this study was to estimate standard errors which account for the imputations. However, with the effects of substratification, poststratification, and subsampling, the replicates divided the sample so finely that there were many cells with one or no observations. This problem was mainly in the substrata with larger farms and would have made it difficult to estimate the standard errors. Thus, the replicates were ignored, and the standard errors were estimated by analyzing the imputed data as if they were reported data. Research has shown that this procedure can lead to biases in the standard error estimates [4], but a better solution could not be found.

Incompleteness of the list frame used on the 1979 PCLS was not measured. However, plans have been made to measure the incompleteness of the 1980 PCLS by making a nonoverlap estimate using the sample segments from an area frame survey, the 1980 June Enumerative Survey (JES). The NCSSO plans to match the names of operators who reside within the segments against names on the list frame. All operators not on the list frame will be sent PCLS questionnaires to complete. Using JES expansion factors, the NCSSO can then make nonoverlap estimates. For the final state estimates, the NCSSO will add the nonoverlap estimates to the estimates for the list sample in order to have complete coverage of the population of farm operators.

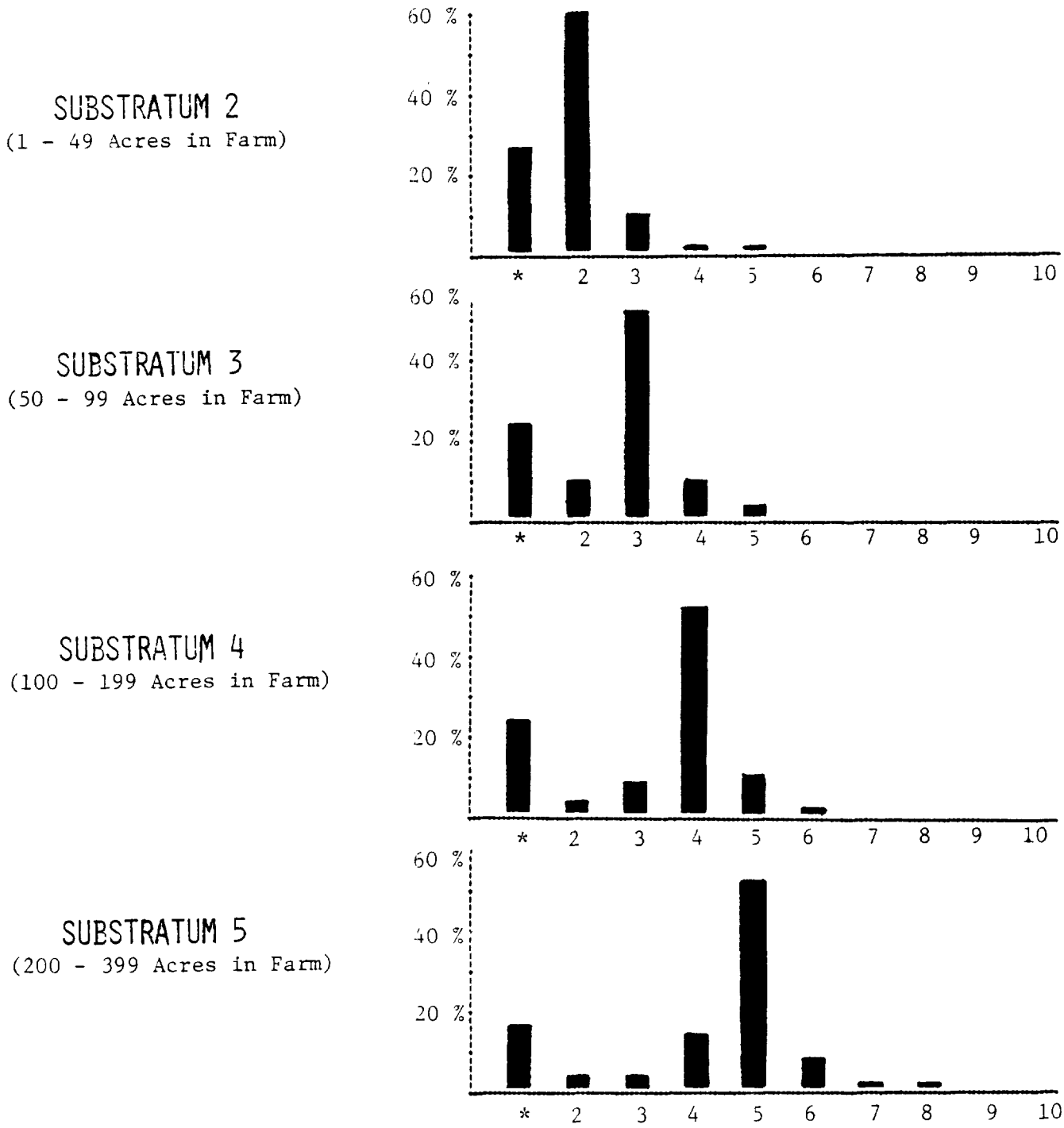
An important problem which was not directly addressed on the PCLS questionnaire was the complexities of partnership operations. Although partnership information was prorated if the respondent listed the partners on the questionnaire--the respondent was most likely to list partners under question 1 (see Appendix A)--, there were no questions directly asking the respondent to describe the structure of the operation. This omission was due to a matter of space on the questionnaire and should be corrected in the future.

Evaluation of the Survey Design

This section evaluates certain aspects of the survey design for the list sample. This evaluation covers: 1) accuracy of stratification, 2) size of the sample, 3) standard errors, 4) optimum allocation, 5) comparison of the efficiency of four strata vs. ten strata, 6) effects of geographic substratification, 7) number of farms with incorrect county codes on the frame, 8) number of farms with livestock located in more than one county, and 9) efficiency of the subsampling scheme.

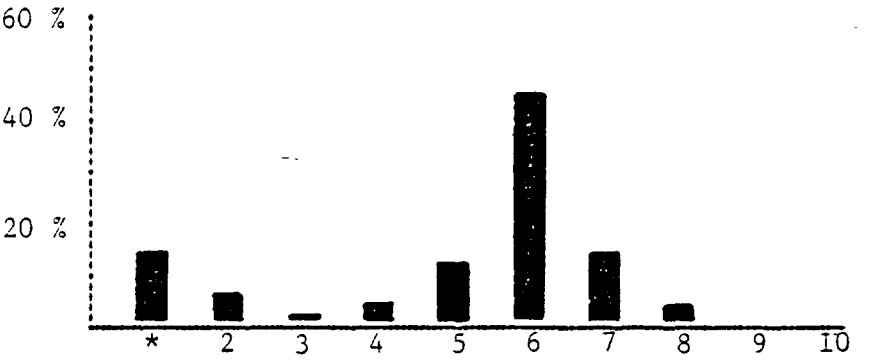
The NCSSO stratified the list frame using the variable "all land in the farm". This variable was also reported by farmers on the PCLS. For each substratum Figure 2 displays graphically the percentages of the farmers in their "true" substrata. Most substrata have a correct classification rate between 50 percent and 60 percent--a fairly accurate classification.

Figure 2. For each substratum vertical axes show the percentage of the sample and horizontal axes show the "true" substratum. A "*" for the "true" substratum indicates that the farm operation went out of business. Because substratum 1 only contains farms of unknown size, it is not shown.



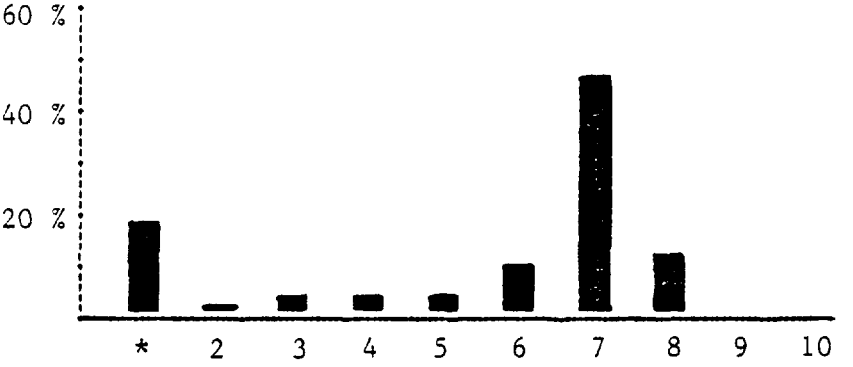
SUBSTRATUM 6

(400 - 599 Acres in Farm)



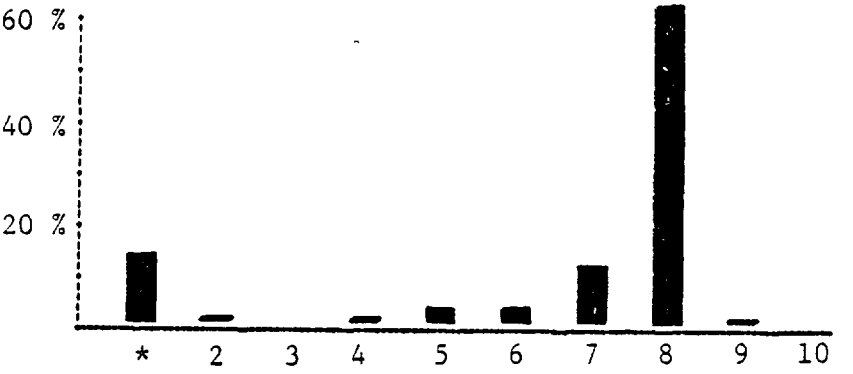
SUBSTRATUM 7

(600 - 899 Acres in Farm)



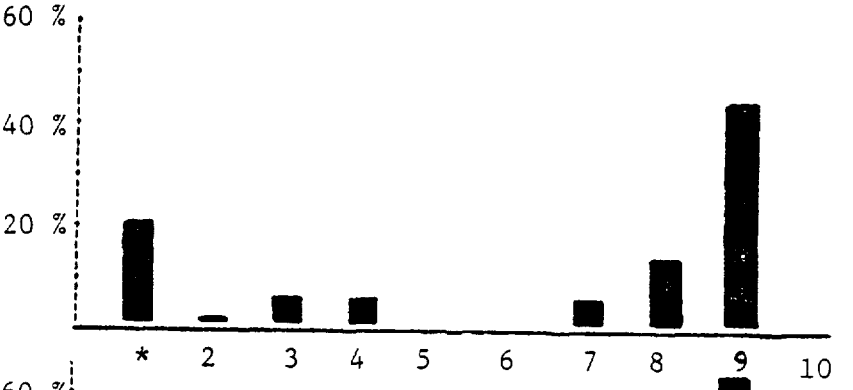
SUBSTRATUM 8

(900 - 2999 Acres in Farm)



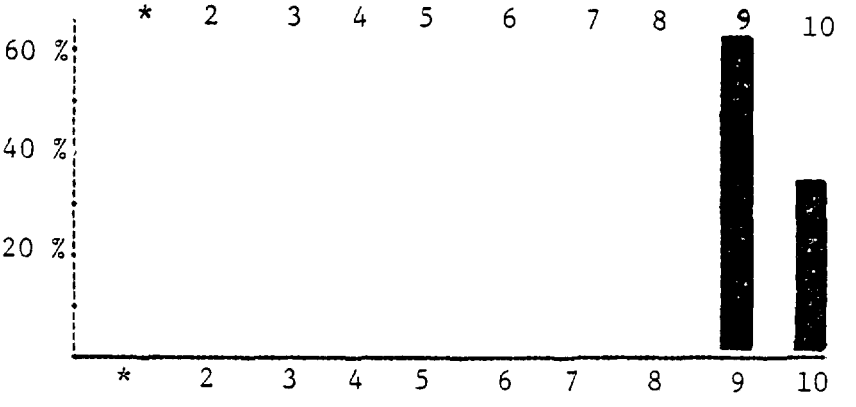
SUBSTRATUM 9

(3000 - 8999 Acres in Farm)



SUBSTRATUM 10

(9000 + Acres in Farm)



The initial sample size for both phases of the 1979 PCLS was 32,000. The effects of the subsampling scheme reduced the initial size to 19,499. Nonresponse further decreased the size to 18,361. Thus, the number of reporting units was 57 percent of the initial sample. The nonresponse was actually larger than the numbers above indicate ($\frac{19,499 - 18,361}{19,499} = 0.06$) because most nonresponse was on telephone or field interviews where the subsampling required weights of "2" and "6" respectively. When the weighting was taken into account, the nonresponse rate was 18 percent.

Of the 18,361 units which reported information 4488, or 24 percent, required imputation of one or more variables. The following variables were the most frequently imputed: amount of corn harvested for grain (imputed on 1610 units), amount of soybeans harvested for beans (imputed on 1134 units), amount of tobacco harvested (imputed on 969 units), and amount of hay harvested (imputed on 1107 units). The amount of imputation was slightly related to the timing of survey since 60 percent of the imputations were on the October PCLS and 40 percent were on the December PCLS. This study did not evaluate the effects of the imputations although this research should be done later.

For state estimates the coefficient of variation (CV) averaged across the 90 quantitative variables was 0.14. With such a large sample size an average CV of 0.14 seemed rather high, but many variables corresponded to rare items--for example, cotton and lespedeza. For common variables such as tobacco or soybeans, the CV's were often between 0.03 and 0.05. The CV for each of the 90 variables are in Appendix C.

Once the 1979 PCLS data had been collected, the optimum allocation was compared to the actual allocation. Table 2 shows this comparison for a sample size of 16,000--the size of the October or December PCLS. The actual allocations were not very close to the optimum, especially in stratum four where the sample size for optimum allocation is practically the entire population. However, given that all 16,000 units reported, the average CV under optimum allocation would have been 0.097 whereas the average CV under the actual allocation would have been 0.099. Thus, the gains in efficiency are negligible, but the actual allocation lessens the respondent burden for the large farms in stratum 4.

Table 2. A comparison of the actual and optimum allocations for the 1979 PCLS. Values represent average allocations over 90 quantitative variables.

Stratum	Population Size	Actual Allocation	Optimum Allocation
1	40,005	6,851	6,620
2	30,558	4,374	3,291
3	21,392	4,315	5,006
4	1,479	400	1,083
Total	93,434	16,000	16,000

As stated in Section 2, the sample was divided into ten substrata based on farm size in order to evaluate the statistical efficiency of four strata vs. ten strata. Given: 1) an optimum allocation, 2) a sample size of 16,000, 3) no missing data, and 4) no subsampling scheme, the average CV with ten substrata was 0.096. Compared to an average CV of 0.097 with four strata, the gain from ten substrata was trivial for state estimates.

Although the four strata were almost as efficient as the ten substrata when using farm size as the stratification variable, it was possible that further stratification based on geography would improve the efficiency of the estimators. To analyze this hypothesis, the average coefficients of variation are computed across all 90 quantitative variables using a substratification based on districts and a substratification based on counties. The results in Table 3 show that there were some modest gains in geographic substratification.

Table 3. Average coefficients of variation across the 90 quantitative variables collected on the 1979 PCLS.

	Average Coefficient of Variation
No Geographic Substratification	0.138
Substratification Using Districts	0.124
Substratification Using Counties	0.117*

*This value was calculated by ignoring the poststratification caused by the subsampling scheme.

The effect of the subsampling to obtain telephone and personal interviews complicated the estimators used on the PCLS and increased the standard errors. The efficiency of the subsampling scheme was evaluated by comparing costs of the 1979 PCLS to those of a hypothetical sample which had the same design except there was no subsampling. Analysis showed that the hypothetical sample only had to yield 11,000 respondents to achieve the same average CV. Allowing for a nonresponse rate of 18 percent would mean that approximately 13,000 units would need to be selected initially in order for the hypothetical sample to yield 11,000 respondents. The cost of the 1979 PCLS was approximately:

$$\begin{aligned}&(0.185)\ (\text{the number of mailed questionnaires})\\ &\quad + (\$1.36)\ (\text{the number of telephone interviews})\\ &\quad + (\$10.50)\ (\text{the number of field interviews})\end{aligned}$$

or:

$$(0.185)\ (32{,}000) + (\$1.36)\ (4514) + (\$10.50)\ (1054) \doteq \$23{,}216$$

Taking into account the number of telephone and field interviews that could be expected without subsampling, the comparable cost for the hypothetical sample would have been:

$$(0.185)\ (13{,}000) + (\$1.36)\ (3520) + (\$10.50)\ (2530) = \$33{,}757.$$

Thus, the subsampling plan saved approximately $10,000.

There were two check questions on the PCLS to estimate: 1) the percentage of units on the list frame which had been classified into the incorrect county and 2) the percentage of units which had livestock in multiple counties. Neither of these percentages--4 percent and 0.1 percent, respectively--were large. When a questionnaire did show classification of a unit into an incorrect county, the information from that unit was summarized in the correct county. The very few questionnaires which had livestock in multiple counties were simply summarized with the county from which they were selected.

Description of the County Estimators

This section describes the three small area estimators--the direct, synthetic, and combined--mentioned in the introduction. In this section they are applied to the specific problem of making county estimates from the PCLS.

The direct estimator used only those sample units which fell into a particular county to form an estimate for that county. Formulas used in this study for the direct estimator and its standard error are in Appendix B. Systematic sampling and a large sample size guaranteed that units from the sample would fall into every county. However, the substratification was ignored because many counties did not have at least 2 units sampled in every substratum. Also, in a few counties there were strata which did not have any units in the sample--particularly strata 3 and 4. For these few counties enough strata were collapsed to attain a sample size of at least 2. Except for the effects of collapsing, direct estimators are mathematically unbiased.

Poststratification was not possible when estimating standard errors for each county estimate because of the large number of counties which had poststrata with one or no units in the sample. Since statistical tests within each stratum showed no significant difference among the means from the poststrata, the poststratification was ignored in the calculation of standard errors. However, the weights which arose because of the poststratification were included in the estimation of totals and means. The formulas found in Appendix B reflect this procedure.

The second county estimator, the synthetic estimator, used the stratum mean of the district as the stratum mean of the county. On each stratum the total for each county was estimated by multiplying the mean for the district by the population size for the county. Formulas for the synthetic estimator and its standard error are also in Appendix B.

Although the synthetic estimator has a much smaller standard error than the direct estimator in most applications, the synthetic estimator is biased. In this study the amount of bias was a direct result of how much the county means differed from the district means. For most surveys it is difficult to estimate the mean square error--the squared bias plus the squared standard error--in each county. However, an estimate of the <u>average</u> mean square error across all counties is possible [5], and thus, an estimate of the <u>average</u> bias across all counties is also possible. The formula for the average mean square error is in Appendix B.

The third estimator is the composite estimator. This estimator combines the direct and synthetic estimators by weighting them and adding them together. Each weight is determined by the inverse of the mean square error associated with that estimator. Each weight is affected by the sample sizes involved [10]. As the sample size in a particular county increases, the weight for the direct estimator increases. This relationship follows intuition. With a small sample size in a county, the synthetic estimator is better, and with a large sample size in a county, the direct estimator is better. The formulas for the composite estimator are in Appendix B.

Evaluation of the County Estimators

This section compares the direct and synthetic estimators with regard to mean square error (MSE) and its components. The composite estimator is not directly evaluated because, as analysis shows, the composite estimates differed very little from the direct estimates. Part of the analysis shows the sample sizes for which the synthetic estimator is as efficient as the direct estimator. This section also includes a comparison of the direct and synthetic estimates with estimates from the 1978 U.S. Census of Agriculture and with "true" values in Robeson County, North Carolina.

To make the calculations for all 90 quantitative variables on the PCLS would have been extremely expensive. Therefore, seven representative variables were selected for evaluation of the county estimators. These are listed in Table 4.

A MSE can be calculated which is the average MSE across all counties in North Carolina although the MSE can not be calculated individually for each county. Of course, given that the direct estimator is unbiased, the MSE of the direct estimator is only composed of the variance. Table 4 gives a comparison of the synthetic and direct estimators with regard to MSE and its components--variance and squared bias--for the seven variables.

Table 4 shows that except for "Hogs" the direct estimator had a much smaller MSE than the synthetic estimator. Although the synthetic estimator had a smaller variance, i.e. it is a more stable estimator, it also had a larger bias. Table 4 shows the results when the four strata were used in the calculations, and Table 5 shows the results when the ten substrata were used. The results in Table 4 and 5 are much the same--providing evidence that the ten substrata did not cause an improvement in the county estimates.

The large sample size had a big impact on the results in Tables 4 and 5. When the sample size is 18,361 respondents, 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. Thus, the bias of the synthetic estimator becomes less important for smaller sample sizes.

Tables 4 and 5 also show that although the direct estimator is better in terms of MSE, it still has variances which are too large. These variances translate into CV's which range from 0.14 to 1.80 and average 0.42. Most variables are in the 0.14 to 0.24 range. Thus, the fluctuations in county estimates from year to year due to sampling would not measure the time trends in the true county values.

Table 4.

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	less than 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 for Grain (acres)	0.05	0.21	0.05	0.01	----	0.20
Tobacco Harvested (acres)	0.04	0.28	0.04	less than 0.01	----	0.28
Soybeans Harvested for Beans (bushels)	0.06	0.26	0.06	0.01	----	0.25
Sorghum Harvested for Grain (bushels)	1.25	2.41	1.25	0.15	----	2.26

Table 5.
Using ten substrata, 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 were 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	----	0.04
Hogs (number of head)	3.32	2.41	3.32	0.57	----	1.84
Cattle (number of head)	0.05	0.15	0.05	0.02	----	0.13
Corn Harvested for Grain (acres)	0.05	0.20	0.05	0.01	----	0.19
Tobacco Harvested (acres)	0.04	0.24	0.04	less than 0.01	----	0.24
Soybeans Harvested for Beans (bushels)	0.07	0.23	0.07	less than 0.01	----	0.23
Sorghum Harvested for Grain (bushels)	1.34	2.42	1.34	0.21	----	2.21

Table 6 shows values of n*, the number of respondents for which the direct and synthetic estimators have the same MSE. Appendix D gives the formulas and assumptions which were used to compute the results in Table 6. Obviously, the number of respondents must be much smaller for the synthetic estimator to be a reasonable alternative. Table 6 indicates that the average number of respondents for which the synthetic estimator has a mean square error less than or equal to the direct estimator is 7856. However, the effect of the "hogs" variable is great. Without this variable the average would be approximately 5000.

When estimates were computed using the composite estimator, there was very little difference between the composite estimates and the direct estimates. This result was caused by the fact that the weights of the direct estimates averaged about 0.8 and the weights of the synthetic estimates averaged about 0.2 because the mean square errors of the direct estimates were so much less than the mean square errors of the synthetic estimates. Thus, in this study the composite estimator did not improve over the direct estimator.

Table 6. Number of respondents below which the relative mean square error of the synthetic county estimator is less than the relative mean square error of the direct county estimator.

Variable	Number of Respondents
All Land in Farm (acres)	5,399
Hogs (number of head)	26,628
Cattle (number of head)	4,158
Corn Harvested for Grain (acres)	3,654
Tobacco Harvested (acres)	2,040
Soybeans Harvested for Beans (bushels)	4,153
Sorghum Harvested for Grain (bushels)	8,960
	Average = 7,856

The 1978 U.S. Census of Agriculture, carried out by the Bureau of the Census, provided an independent measure of the number of acres in farms. Although the PCLS data was from 1979, one year should not have caused a large change in this variable. The Census of Agriculture's total value for all 100 counties in North Carolina was 11,001,686 acres while the total of the direct estimate was 11,021,118 acres and the total of the synthetic estimate was 11,071,198 acres. All three totals were very close, especially if one remembers that the definitions and data collection techniques of the Census were different from those of the PCLS.

When the estimates for each county were compared, there were some large differences. Table 7 lists the Census values for each county and the direct and synthetic estimates relative to the Census values. Some discrepancies--for example, in Wake County and Nash County--were large and occurred in important agricultural counties. Many of the discrepancies were probably due to the sampling variability associated with the direct and synthetic estimators and reveal how unstable those county estimators are, even with 18,361 respondents. Over all counties the direct estimator was only slightly closer (in terms of absolute distance) than the synthetic estimator to the Census values.

Admittedly, the comparison of 1979 synthetic and direct estimators with 1978 Census data is fragile evidence. All surveys, including censuses, contain nonsampling errors which hinder efforts to determine which estimator is "best". Firmer evidence would require a "true" measurement of some of the variables.

During July, 1980 "true' values were collected in Robeson County--a major agricultural county in North Carolina--in connection with another research project. Four of the "true" values had corresponding direct and synthetic estimates on the 1980 PCLS. Table 8 shows the "true" values vs. the direct and synthetic estimates. The synthetic estimates were closer to the "true" values for three of the four variables and had a lower standard error for all variables. As theory predicted, the synthetic estimator showed a bias. It consistently underestimated the "true" values. Although it appeared that the direct estimator was overestimating the "true" value, this appearance had no theoretical foundation and was probably a random effect from sampling fluctuation.

Table 9 shows that Robeson County had means for acreage variables which were larger than the means of the crop reporting district. These differences led to the consistent underestimation of the synthetic estimates shown in Table 8.

Conclusions and Recommendations

The conclusions of this report in regard to the 1979 PCLS fell into three major areas--conclusions about the survey design, conclusions about the county estimators, and future research.

In the area of survey design, analysis showed that 4 strata were almost as efficient as 10 strata and that the stratification was fairly accurate. The actual allocation of the sample to the strata was acceptable when compared to the optimum allocation because the actual allocation minimized the standard errors of the estimates while also minimizing the respondent burden on large farms. Although complicating the estimators, the subsampling plan of the PCLS was economical--being approximately $10,000 less expensive than a sample design which had no subsampling. The problem of incorrect county codes proved to be a small problem as did the problem of livestock belonging to one farm and located in several counties. The NCSSO should plan to give consideration to the problem of partnerships because of the high possibility of duplicate reporting.

Table 7. A comparison between the 1978 U.S. Census of Agriculture and the 1979 Probability Crop and Livestock Survey (PCLS) with regard to county estimates of the total number of acres in farms.

County	Percentage of 1979 PCLS to 1978 U.S. Census of Agriculture	
	Direct Estimate	Synthetic Estimate
Alleghany	108	86
Ashe	69	77
Avery	72	137
Caldwell	63	58
Surry	91	78
Watauga	70	87
Wilkes	79	79
Yadkin	107	112
Bancombe	105	55
Burke	72	48
Cherokee	72	96
Clay	95	105
Graham	91	143
Haywood	70	85
Henderson	61	88
Jackson	62	84
McDowell	103	143
Macon	81	97
Madison	81	74
Mitchell	83	84
Polk	81	87
Rutherford	82	54
Swain	91	136
Transylvania	83	87
Yancey	73	106
Alamance	134	145
Caswell	92	86
Durham	121	144
Forsyth	97	136
Franklin	114	93
Granville	119	89
Guilford	121	138
Orange	70	61
Person	86	87
Rockingham	122	129
Stokes	73	94
Vance	104	84
Warren	80	87
Alexander	128	120
Catawba	76	76
Chatham	70	67
Davidson	81	78
Davie	79	57
Iredell	75	69
Lee	150	173
Randolph	90	115
Rowan	93	84
Wake	150	150
Anson	79	82
Cabarrus	88	88
Cleveland	104	107
Gaston	101	100
Lincoln	99	120
Mecklenburg	97	104
Montgomery	112	85
Moore	118	138
Richmond	69	86
Stanly	77	70
Union	84	66
Bertie	144	125
Camden	154	100
Chowan	128	160
Currituck	112	100
Dare	no farm acreage	
Edgecombe	75	47
Gates	133	122
Halifax	99	116
Hertford	67	104
Martin	111	114
Nash	202	180
Northhampton	60	86
Pasquotank	91	134
Perquimans	80	153
Tyrrell	72	69
Washington	139	105
Beaufort	141	128
Carteret	107	43
Craven	121	160
Green	122	103
Hyde	77	59
Johnston	85	123
Jones	117	113
Lenoin	151	153
Pamlico	114	56
Pitt	123	115
Wayne	123	127
Wilson	124	127
Bladen	61	68
Brunswick	66	90
Columbus	116	145
Cumberland	121	91
Duplin	67	70
Harnett	104	92
Hoke	82	54
New Hanover	101	211
Onslow	73	106
Penden	102	12
Robeson	120	90
Sampson	93	85
Scotland	66	60
Total	100	101

Table 8. Comparison of "true" acreage values for Robeson County, North Carolina to estimates from the 1980 PCLS.

Variable	"True" Value	Direct Estimate		Synthetic Estimate	
		Total	Relative Standard Error	Total	Relative Standard Error
Corn	90,842	100,190	19%	86,838	7%
Soybeans	115,154	157,350	24%	103,435	9%
Tobacco	24,142	20,050	14%	18,449	6%
Cotton	10,699	19,343	77%	9,500	35%

Table 9. A comparison of county mean vs. mean of the crop reporting district for Robeson County, North Carolina. Estimated means are from 1980 PCLS.

	Robeson County	Crop Reporting District Containing Robeson County
Corn	26	20
Soybeans	41	25
Tobacco	5	4
Cotton	8	3

Of the three county estimators evaluated in this study, the direct estimator was the best in that it had the smallest mean square error. This outcome resulted from the extremely large sample size and the bias of the synthetic estimator. Analysis indicated that the number of respondents for which the synthetic estimator has a mean square error equal to the direct estimator was approximately 8000. When one hog variable was omitted, this average dropped to about 5000. In this study the composite estimator yielded estimates which differed very little from the direct estimator. Thus, the composite estimator offered no improvement over the direct estimator.

Although in this study the direct estimator was better than the synthetic estimator, there were fairly large coefficients of variation for both estimators. This study shows that county estimates for most variables probably have CV's as high as 0.14-0.24. High CV's make county estimates fluctuate so much from year to year that time trends and relationships among counties are unrecognizable. Thus, county estimates still need much improvement. Increases in sample size are impractical because of time and cost constraints. Possibilities for improving county estimates include combining information from other surveys, retaining part of the sample from one year to the next, using census data to model relative relationships among county values, and using historical data to model time trends. These possibilities should be the subject of future research.

Bibliography

[1] Advertising Research Foundation. U.S. Television Households, by Region States, and County. New York. March, 1956.

[2] Cochran, W.G. Sampling Techniques (2nd ed.) New York. John Wiley and Sons, 1960.

[3] Deming, W.E. Sample Design In Business Research. New York. John Wiley and Sons, 1960.

[4] Ford, B.L., Kleweno, D.G., and Tortora, R.D. "A Simulation Study to Compare Procedures Which Impute for Missing Items on an ESS Hog Survey," Proceedings of the Section on Survey Research Methods of the American Statistical Association, 1981.

[5] Gonzalez, M.E. "Use and Evaluation of Synthetic Estimates," Proceedings of the Social Statistics Section of the American Statistical Association, 1973.

[6] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. Sample Survey Methods and Theory, Vol. I, New York, John Wiley and Sons, 1953.

[7] National Center for Health Statistics. Synthetic State Estimates of Disability. Public Health Service Publication # 1759. Washington, D.C. U.S. Government Printing Office, 1968.

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

[9] ----------Levy, P.S. "Small Area Estimation--Synthetic and Other Procedures."
[10] ----------Schaible, W.A. "A Composite Estimator for Small Area Statistics."
[11] ----------Steinberg, J. "Introduction."

Appendix A

Questionnaires for the October and December

Probability Crop and Livestock Survey*

*These questionnaires were originally on legal size pages.

October PCLS

FARM INFORMATION FOR 1979

(Data collected under provisions of N. C. General Statutes)

NORTH CAROLINA
Crop & Livestock
Reporting Service

P.O. Box 27767
1 West Edenton Street
Raleigh, N. C. 27611
Phone (919) 755-4394

Dear Reporter:

Results from this survey will help provide county crop and livestock totals for farmers and others needing detailed information. All individual reports will be kept confidential. Your timely response is important.

Thank you.

Dan C. Tucker

DAN C. TUCKER
Statistician In Charge

980

Please make corrections in name, address and zip code, if necessary

1. Is your farm operation known by a name other than that on the address label?
   NO_____ YES_____ Enter other name_____
2. In what county or counties is your farm operation located?_____

REPORT FOR YOUR 1979 FARM OPERATION

(Include all cropland, idle land, pastures, woodland, and land rented from others but exclude land rented to others)

ITEM	ACRES	TOTAL PRODUCTION HARVESTED AND TO BE HARVESTED
3. How many acres of all land are there in the farm(s) you operate?	995	
-1979 CROPS-		
4. Corn planted for all purposes	130	
5. Corn harvested and to be harvested for grain	133	136 Bu.
6. Corn cut for silage	139	142 Tons
7. Corn cut for fodder, pastured and hogged down (without husking)	145	
8. Corn abandoned (will not be harvested or pastured)	148	
9. Soybeans planted for all purposes	625	
0. Soybeans harvested and to be harvested for beans	628	631 Bu.
1. Soybeans cut for hay, used for silage, pasture only, plowed under or abandoned	634	
2. Tobacco harvested	666	667 Lb.
3. Peanuts planted for all purposes	420	
4. Peanuts harvested and to be harvested for nuts	423	426 L

. Sorghums and sorghum grains planted for all purposes (exclude crosses with Sudan)	570	
. Sorghums harvested and to be harvested for grain	573	576 Bu.
. Sorghums cut for silage	579	582 Tons
. Sorghums cut for fodder and hay or used only for pasture	594	
. Sorghums used for syrup, molasses or abandoned	597	
. Cotton planted	180	
. Cotton harvested and to be harvested	181	183 Bales
. Sweet potatoes planted	445	
. Sweet potatoes harvested and to be harvested	446	447 55 Lb. Bu.
. Irish potatoes planted	435	
. Irish potatoes harvested and to be harvested	438	441 Cwt.
. All hay cut	316	319 Tons
. Lespedeza harvested and to be harvested for seed	508	509 (Clean Seed) Lbs.
. Wheat planted for all purposes in fall 1978	680	
. Wheat harvested for grain	685	690 Bu.
. Wheat used for hay, silage, pasture only, plowed under or abandoned	710	
. Oats planted for all purposes in fall 1978 and spring 1979	385	
. Oats harvested for grain	388	391 Bu.
. Oats used for hay, silage or pasture only	418	
. Oats plowed under or abandoned	419	
. Barley planted for all purposes in fall 1978 and spring 1979	001	
. Barley harvested for grain	006	011 Bu.
. Barley used for hay, silage, pasture only, plowed under or abandoned	031	
. Rye planted for all purposes in fall 1978	470	
. Rye harvested for grain	473	476 Bu.
. Rye used for hay, silage, pasture only, plowed under or abandoned	488	

-FALL SOWN CROPS-

. Wheat planted for all purposes in fall 1979 for 1980 crop.....	735	
. Rye planted for all purposes in fall 1979 for 1980 crop.....	491	

. Number of livestock and poultry on all land in your farm operation(s) on December 1, 1979:

	NUMBER OF HEAD ON DEC. 1, 1979
a. *Hogs and pigs*.	975
b. *All cattle and calves (Beef & Dairy)*..... ..	976
c. *Milk cows*.....	977
d. *Beef cows*	978
e. *Chickens (Exclude commercial broilers)*...	979

. If you have no hogs now, do you plan to have any during the next 12 months? No _____ Yes _____ 981

. Are all your livestock and poultry located in one county? No _____ Yes _____

ported By _____ Telepnone No. ( ) _____

Area Code

December PCLS

FARM INFORMATION FOR 1979

(Data collected under provisions of N.C. General Statutes)

NORTH CAROLINA
Crop & Livestock
Reporting Service

P.O. Box 27767
1 West Edenton Street
Raleigh, N. C. 27611
Phone (919) 755-4394

Dear Reporter:

Results from this survey will help provide county crop and livestock totals for farmers and others needing detailed information. All individual reports will be kept confidential. Your timely response is important. Thank you.

980

Please make corrections in name, address and zip code, if necessary.

DAN C. TUCKER
Statistician In Charge

1. Is your farm operation known by a name other than that on the address label?

NO____ YES____ Enter other name_____

2. In what county or counties is your farm operation located?_____

REPORT FOR YOUR 1979 FARM OPERATION

(Include all cropland, idle land, pastures, woodland, and land rented from others but exclude land rented to others)

	ACRES
3. Total acres of land in the farm(s) you operate	995
a. Acres of cropland harvested	304
b. Acres of cropland idle (no crops saved or grazed). . .	305
c. Acres of improved and unimproved pasture.	306
d. Acres of forest land (include woodland pasture). . . .	307
e. Acres of all other land (swamp, waste, homesite, ponds, etc.).	308

(NOTE: *ITEMS a through e should equal the total acres in Item 3.*)

4. Number of livestock and poultry on all land in your farm operation(s) on January 1, 1980:

	NUMBER OF HEAD ON JAN. 1, 1980
a. Hogs and pigs	975
b. All cattle and calves (Beef & Dairy).	976
C. Milk cows	977
d. Beef cows	978
e. Chickens (Exclude commercial broilers).	979

5. If you have no hogs now, do you plan to have any during the next 12 months? NO_____ YES_____

6. Are all your livestock and poultry located in one county? NO_____ YES_____

981

7. CROPS HARVESTED FROM YOUR ENTIRE FARM OPERATION FOR THE 1979 SEASON:

CROPS	ACRES HARVESTED	AMOUNT HARVESTED	AMOUNT SOLD AND TO BE SOLD	AVERAGE PRICE YOU RECEIVED 1/ (Dollars)
Corn for Grain	133	136 Bu.	013 Bu.	111 $. Per Bu.
Wheat for Grain	685	690 Bu.	210 Bu.	101 $. Per Bu.
Oats for Grain	388	391 Bu.	093 Bu.	112 $. Per Bu.
Barley for Grain	006	011 Bu.	103 Bu.	113 $. Per Bu.
Rye for Grain	473	476 Bu.	033 Bu.	104 $. Per Bu.
Sorghum for Grain	573	576 Bu.	163 Bu.	114 $. Per Cwt.
Sweet Potatoes	446	447 55 lb. Bu.	453 Bu.	302 $. Per Bu.
All Hay	316	319 Tons	203 Tons	189 $. Per Ton
Irish Potatoes	438	441 Cwt.	443 Cwt.	301 $. Per Cwt.
Soybeans for Beans	628	631 Bu.		154 $. Per Bu.
Tobacco	666	667 Lbs.		668 $. Per Lb.
Lespedeza Seed	508	509 (Clean Seed) Lbs.		723 $. Per Cwt.
Peanuts	423	426 Lbs.		153 (Cents) . Per Lb.
Cotton Lint	181	183 Bales		121 . Per Lb.

Appendix B

Formulas of Estimators

1. Formulas for State Estimators and Standard Errors

State estimates for the PCLS used formulas based on a poststratification. This poststratification resulted in three poststrata within each substratum of the original design: 1) a poststratum representing data collection by mail, 2) a poststratum representing data collection by telephone from the NCSSO, and 3) a poststratum representing data collection by field enumerators. There were different weights associated with each poststratum because of the subsampling within each poststratum.

Let T be the total value for variable x. Then, if i represents an index for the ten substrata, and j represents an index for the three poststrata within each substratum:

$$T = \sum_{i=1}^{10} \sum_{j=1}^{3} N_{ij} \mu_{ij} \qquad (B.1)$$

where N_{ij} represents the number of units in the population belonging to group (i,j) and μ_{ij} represents the mean of this group. The reader should think of the population belonging to (i,j) as that part of the population which is in substratum i and would have received the j^{th} method of data collection *if the entire population had been enumerated.* If n_{ij} is the number of units which were sampled from substratum i and had data collected by method j, then μ_{ij} is estimated by:

$$\hat{\mu}_{ij} = \frac{\sum_{k=1}^{n_{ij}} x_{ijk}}{n_{ij}} \qquad (B.2)$$

i.e. the average of the n_{ij} units in the sample belonging to group (i.j). However, N_{ij} in (B.1) must also be estimated. Let N_i be the known number of units in the population belonging to substratum i, and let p_{ij} be the proportion of the original sample--i.e. the sample before subselection--in substratum i which would have been in poststratum j. Then, the obvious estimator is:

$$\hat{N}_{ij} = p_{ij} N_i \qquad (B.3)$$

Because of the subselection, p_{ij} must be estimated. Let w_j represent the inverse of the subsampling rate in poststratum j. For the PCLS, w_j had the same value in every substratum--w_1 equaled "1" because there was no subsampling to

obtain mail interviews, w_2 equaled "2" because only half of the original sample which would have received telephone interviews was included in the final sample, and w_3 equaled "6" because only one-sixth of the original sample which would have received field interviews was included in the final sample.

Let $n_i^* = \sum_{j=1}^{3} w_j n_{ij}$, i.e. the number of units in the original sample. Then, if p_{ij} is the proportion of the original sample in substratum i which would have been in poststratum j, p_{ij} is estimated by:

$$\hat{p}_{ij} = \frac{w_j n_{ij}}{n_i^*} \tag{B.4}$$

Substituting back from (B.2), (B.3), and (B.4) to (B.1), one obtains the estimator of T:

$$\hat{T} = \sum_{i=1}^{10} \sum_{j=1}^{3} N_i \left(\frac{w_j n_{ij}}{n_i^*} \right) \sum_{k=1}^{n_{ij}} \frac{x_{ijk}}{n_{ij}} \tag{B.5}$$

Formula (B.5) can also be expressed as:

$$\hat{T} = \sum_{i=1}^{10} \frac{N_i}{n_i^*} \sum_{j=1}^{3} \sum_{k=1}^{n_{ij}} w_j x_{ijk} \tag{B.6}$$

so that summarization can be thought of as weighting the data by the appropriate w_j and applying the expansion factor N_i / n_i^*.

Some of the logic for the estimator $\hat{p}_{ij}$ changes when the nonresponse is taken into account. Units which were refusals and inaccessibles were omitted from the estimation phase of the PCLS so these units were not used to estimate μ_{ij} or N_{ij} although they were part of the original sample. When nonresponse is taken into account, the reader should think of n_i^* as the size of the original sample in substratum i *which would have responded if interviewed.* Similarly, $w_j n_{ij}$ is that part of the original sample in substratum i which would have been in poststratum j and *would have responded if interviewed.* These considerations complicate the explanation of the estimator (B.4) and, therefore, have been stated after presenting the estimator under simpler conditions.

To make standard error estimates we will treat the poststrata as if they were strata in the original design and $\hat{N}_{ij}$ as a known value rather than an estimated value. The rationale for using poststratification in this manner has been documented*. Let $SE(\hat{T})$ represent the estimated standard error of $\hat{T}$:

$$SE(\hat{T}) = \left[\sum_{i=1}^{10} \sum_{j=1}^{3} (\hat{N}_{ij})^2 \left(1 - \frac{n_{ij}}{\hat{N}_{ij}}\right) \frac{s_{ij}^2}{n_{ij}} \right]^{1/2} \qquad (B.7)$$

where:

$$s_{ij}^2 = \sum_{k=1}^{n_{ij}} \frac{(x_{ijk} - \hat{\mu}_{ij})^2}{n_{ij} - 1} \qquad (B.8)$$

2. Formulas for County Estimators and Standard Errors

To calculate county estimates the substratification was ignored because analysis showed there was little increase in efficiency with ten substrata rather than four strata. The direct estimator simply uses those units in the sample which fell in a certain county. Therefore, the estimator of T_h, the total value of variable x in county h, is similar to formula (B.6) except that terms are given corresponding to county h:

$$\hat{T}_h = \sum_{i=1}^{4} \frac{N_{hi}}{n^*_{hi}} \sum_{j=1}^{3} \sum_{k=1}^{n_{hij}} w_j x_{hijk} \qquad (B.9)$$

To compute the standard error of $\hat{T}_h$, the poststratafication was ignored because it created many poststrata with one or no observations. Also, statistical tests showed no statistical difference in the means of the poststrata. Therefore, the standard error of $\hat{T}_h$ was estimated by collapsing the poststrata and substrata:

$$SE(\hat{T}_h) = \left[\sum_{i=1}^{4} N_{hi}^2 \left(1 - \frac{n_{hi}}{N_{hi}}\right) \frac{s_{hi}^2}{n_{hi}} \right]^{1/2} \qquad (B.10)$$

*Holt, D. and Smith, T.M.F. "Post Stratification", Journal of the Royal Statistical Society. Series A, Volume 142. 1979.

where:

$$n_{hi} = \sum_{j=1}^{3} n_{hij}$$

$$s^2_{hi} = \sum_{j=1}^{3} \sum_{k=1}^{n_{hij}} \frac{(x_{hijk} - \bar{x}^*_{hi})^2}{n_{hi} - 1}$$

$$\bar{x}^*_{hi} = \sum_{j=1}^{3} \sum_{k=1}^{n_{hij}} \frac{x_{jijk}}{n_{hi}}$$

The synthetic estimator for county h uses the stratum means for the district to which county h belongs. Suppose $\bar{x}_{Gi}$ is the mean of stratum i in district G. We find $\bar{x}_{Gi}$ by ignoring the county index and making a direct estimate for district G, i.e. we only use those units belonging to stratum i in district G. Let T_{Gi} be the total value of district G for stratum i. Then we make a direct estimate by:

$$\hat{T}_{Gi} = \frac{N_{Gi}}{n^*_{Gi}} \sum_{j=1}^{3} \sum_{=1}^{n_{Gij}} w_j x_{Gijk} \qquad \text{(B.11)}$$

where terms are defined as in (B.9) except they are in reference to district G rather than county h. Then;

$$\bar{x}_{Gi} = \hat{T}_{Gi} / N_{Gi} \qquad \text{(B.12)}$$

The estimated standard error of $\bar{x}_{Gi}$ is also found by ignoring the poststratification:

$$SE(\bar{x}_{Gi}) = \left[(1 - \frac{n_{Gi}}{N_{Gi}}) \frac{s^2_{Gi}}{n_{Gi}} \right]^{1/2} \qquad \text{(B.13)}$$

where:

$$n_{Gi} = \sum_{j=1}^{3} n_{Gij}$$

$$s^2_{Gi} = \sum_{j=1}^{3} \sum_{k=1}^{n_{Gij}} \left[\frac{x_{Gijk} - \bar{x}^*_{Gi}}{n_{Gi} - 1} \right]^2$$

$$\bar{x}^{*}_{Gi} = \sum_{j=1}^{3} \sum_{k=1}^{n_{Gij}} x_{Gijk} / n_{Gi} .$$

Now the synthetic estimator of T_h can be expressed as:

$$\hat{\hat{T}}_h = \sum_{i=1}^{4} N_{hi} \bar{x}_{Gi} \qquad (B.14)$$

and the estimated standard error of $\hat{T}_h$ as:

$$SE(\hat{\hat{T}}_h) = \left[\sum_{i=1}^{4} N_{hi}^2 \{SE(\bar{x}_{Gi})\}^2 \right]^{1/2} \qquad (B.15)$$

The composite estimator of a total for county h, $\tilde{T}_h$, is found by weighting the direct and synthetic estimators for county h:

$$\tilde{T}_h = \nu \hat{T}_h + (1-\nu) \hat{\hat{T}}_h \qquad (B.16)$$

The optimum vlaue of ν, in the sense of minimizing the mean square error of $\tilde{T}_h$, is determined by the mean square errors of $\hat{T}_h$ and $\hat{\hat{T}}_h$ and their covariance. Schaible* shows that under certain simplifying assumptions the optimum value of ν is:

$$\nu \approx \frac{1}{1 - r} \qquad (B.17)$$

where r equals the mean square error of $\hat{T}_h$ divided by the mean square error of $\hat{\hat{T}}_h$, i.e. $\dfrac{MSE(\hat{T}_h)}{MSE(\hat{\hat{T}}_h)}$.

*Schaible, W.L. "A Composite Estimator for Small Area Statistics," Synthetic Estimators for Small Areas: Statistical Workshop Papers and Discussions. National Institute of Drug Abuse. Research Monograph 24. Washington, D.C. U.S. Government Printing Office, 1979.

If ν is a constant, then the standard error of $\tilde{T}_h$ is*:

$$SE(\tilde{T}_h) = \left[\nu^2 \{SE(\hat{T}_h)\}^2 + (1-\nu)^2 \{SE(\hat{\hat{T}}_h)\}^2 + 2(1-\nu)\, Cov(\hat{T}_h, \hat{\hat{T}}_h) \right]^{1/2} \quad (B.18)$$

where $Cov(\cdot)$ represents a covariance term.

3. Formulas for the Average Mean Square Error

The direct estimator is considered mathematically unbiased for the analysis in this study. Thus, the average $MSE(\hat{T})$ over h is the average $SE(\hat{T}_h)^2$. For the synthetic estimator, Gonzalez** gives an estimate of the average $MSE(\hat{\hat{T}}_h)$:

$$\overline{MSE(\hat{\hat{T}}_h)} = \frac{1}{H} \sum_{h=1}^{H} (\hat{\hat{T}}_h - \hat{T}_h)^2 - \frac{1}{H} \sum_{h=1}^{H} \sum_{i=1}^{4} p_{hi}^2 \, (1-2f_{hi}) \, \sigma_{hi}^2 \quad (B.19)$$

where:
H = the total number of counties
p_{hi} = proportion of the units in the h^{th} county which are in the i^{th} stratum
f_{hi} = proportion of all units in i^{th} stratum which are in the h^{th} county
σ_{hi}^2 = the variance of the units in the i^{th} stratum of the h^{th} county

For the composite estimator, the average MSE over all counties is***:

$$\overline{MSE(T_h)} = \nu\{MSE(\hat{T}_h)\} + (1-\nu) \{MSE(\hat{\hat{T}}_h)\} - \nu(1-\nu) \{E(\hat{T}_h - \hat{\hat{T}}_h)^2\} \quad (B.20)$$

*Hogg, R.V., and Craig, A.T. Introduction to Mathematical Statistics. London The MacMillan Company, 1970.

**Gonzalez, M.E. "Use and Evaluation of Synthetic Estimates". American Statistical Association. Proceedings of the Social Statistics Section. 1973.

***Schaible, W.L. "A Composite Estimator for Small Area Statistics", Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion, National Institute of Drug Abuse. Research Monograph 24. Washington D.C. U.S. Government Printing Office, 1979.

Appendix C

Coefficients of Variation from the Probability Crop and Livestock Survey (PCLS)

The following estimates are the coefficients of variation at the state level for the 90 quantitative variables collected on the PCLS.

Variable	Coefficient of Variation
All Land in Farm (acres)	0.02
Harvested Cropland (acres)	0.03
Idle Cropland (acres)	0.05
Pasture (acres)	0.03
Forest Land (acres)	0.03
Other Land (acres)	0.04
Hogs (number of head)	0.11
Cattle (number of head)	0.03
Milk Cows (number of head)	0.07
Beef Cows (number of head)	0.03
Chickens (number of head)	0.10
Corn Planted (acres)	0.04
Corn Harvested for Grain (acres)	0.03
Corn Harvested for Grain (bushels)	0.03
Corn Harvested for Silage (acres)	0.09
Corn Harvested for Silage (tons)	0.09
Corn Harvested for Fodder (acres)	0.16
Corn Abandoned (acres)	0.16
Corn Sold (bushels)	0.05
Corn Sold (average price per bushel)	0.03
Soybeans Planted (acres)	0.04
Soybeans Harvested for Beans (acres)	0.03
Soybeans Harvested for Beans (bushels)	0.03
Soybeans Harvested for Other Reasons	0.08
Soybeans Sold (average price per bushel)	0.02
Tobacco Harvested (acres)	0.03
Tobacco Harvested (pounds)	0.03
Tobacco Sold (average price per pound)	0.02
Peanuts Planted (acres)	0.08
Peanuts Harvested (acres)	0.06
Peanuts Harvested (pounds)	0.06
Peanuts Sold (average price per pound)	0.06
Sorghum Planted (acres)	0.11
Sorghum Harvested for Grain (acres)	0.11
Sorghum Harvested for Grain (bushels)	0.13
Sorghum Harvested for Silage (acres)	0.20
Sorghum Harvested for Silage (tons)	0.28
Sorghum Harvested for Fodder (acres)	0.18
Sorghum for Syrup or Abandoned (acres)	0.41
Sorghum Sold (bushels)	0.29
Sorghum Sold (average price cwt.)	0.18

Cotton Planted (acres)	0.19
Cotton Harvested (acres)	0.20
Cotton Harvested (bales)	0.21
Cotton Sold (average price per pound)	0.21
Sweet Potatoes Planted (acres)	0.21
Sweet Potatoes Harvested (acres)	0.16
Sweet Potatoes Harvested (bushels)	0.16
Sweet Potatoes Sold (bushels)	0.31
Sweet Potatoes Sold for Fresh Market (average price per bushel)	0.12
Sweet Potatoes Sold for Processed Market (average price per bushel)	0.27
Irish Potatoes Planted (acres)	0.40
Irish Potatoes Harvested (acres)	0.30
Irish Potatoes Harvested (cwt.)	0.29
Irish Potatoes Sold (cwt.)	0.48
Irish Potatoes Sold (average price per cwt.)	0.16
Hay Harvested (acres)	0.03
Hay Harvested (tons)	0.04
Hay Sold (tons)	0.17
Hay Sold (average price per ton)	0.09
Lespedeza Harvested (acres)	0.18
Lespedeza Harvested (pounds)	0.18
Lespedeza Sold (price per cwt.)	0.29
Wheat Planted (acres)	0.07
Wheat Harvested for Grain (acres)	0.06
Wheat Harvested for Grain (bushels)	0.06
Wheat Harvested for Other Reasons (acres)	0.13
Wheat Sold (bushels)	0.16
Wheat Sold (average price per bushel)	0.06
Oats Planted (acres)	0.06
Oats Harvested for Grain (acres)	0.06
Oats Harvested for Grain (bushels)	0.07
Oats Harvested for Silage (acres)	0.14
Oats Abandoned (acres)	0.10
Oats Sold (bushels)	0.13
Oats Sold (average price per bushel)	0.09
Barley Planted (acres)	0.10
Barley Harvested for Grain (acres)	0.08
Barley Harvested for Grain (bushels)	0.09
Barley Harvested for Other Purposes (acres)	0.17
Barley Sold (bushels)	0.26
Barley Sold (average price per bushel)	0.13

Rye Planted (acres)	0.06
Rye Harvested for Grain (acres)	0.11
Rye Harvested for Grain (bushels)	0.11
Rye Harvested for Other Purposes (acres)	0.07
Rye Sold (bushels)	0.27
Rye Sold (average price per bushel)	0.15
Wheat Planted in Fall for Harvest in Next Year (acres)	0.07
Rye Planted in Fall for Harvest in Next Year (acres)	0.06

Appendix D

Formulas to Calculate the Sample Size at Which the Mean Square Errors Are Equal

In a county let n* be the sample size for which the relative mean square error of the direct estimator, M_D^*, is equal to the relative mean square error of the synthetic estimator, M_S^*:

$$M_D^* = M_S^* \quad . \tag{D.1}$$

Although it is desirable to derive n* for each county, we are only able to calculate M_S^* as an average mean square error over all counties. Thus, in the equations of Appendix D we use terms that reflect the average county.

Because the direct estimator is mathematically unbiased but the synthetic is not, M_D^* is equal to $(C_D^*)^2$, the squared coefficient of variation of the direct estimator, but M_S^* is equal to $(C_S^*)^2+(B_S^*)^2$, the squared coefficient of variation of the synthetic estimator plus a squared relative bias term. Substituting into (D.1), one has:

$$(C_D^*)^2 = (C_S^*)^2 + (B_S^*)^2 \quad . \tag{D.2}$$

Now, suppose at sample size n the squared coefficients of variation are C_D^2 and C_S^2. For the direct estimator the squared coefficient of variation changes inversely with a change in sample size:

$$(C_D^*)^2 = \frac{C_D^2}{(n^*/n)} \quad . \tag{D.3}$$

If one assumes that a change in the sample size at the county level, $\Delta = \frac{n^*}{n}$, also represents a change of Δ in the sample size at the district level (or whatever aggregate level is used for the synthetic estimator), then also:

$$(C_S^*)^2 = \frac{C_S^2}{\Delta} = \frac{C_S^2}{(n^*/n)} \quad . \tag{D.4}$$

Substituting (D.3) and (D.4) into (D.2), one obtains:

$$\frac{C_D^2}{(n^*/n)} = \frac{C_S^2}{(n^*/n)} + (B_S^*)^2 \quad . \tag{D.5}$$

Assuming B_S^* is nonzero and rearranging terms, we have:

$$n^* = n \left[\frac{C_D^2 - C_S^2}{B_S^*} \right] . \qquad (D.6)$$

Thus, n*, as an average county value, can be found from (D.6) by substituting in the sample size, coefficients of variation, and bias from the current sample. When n*, the average sample size in a county, is multiplied by the number of counties in the state, the product is the sample size for the state.

The important assumptions for the above derivations are:

1) the bias is not related to sample size,
2) although the sample size in a county and district may change, they remain in the same proportion,
3) aspects of the sample design such as stratification and allocation remain the same.