An Evaluation of Categorical Data Analysis Methodology For County Estimates in North Carolina

by

Nancy J. Carter;

Assistant Professor,

California State University, Chico

95929-0525

and

Douglas C. Bond

Statistical Reporting Service;

U.S. Department of Agriculture;

Washington, D.C. 20250;

April, 1985.

ABSTRACT

Agricultural estimates at the county level are very important. This report details an evaluation of three different Categorical Data Analysis (CDA) estimators which were used to derive county-level estimates of agriculture in North Carolina. These three estimators Case 1 (Full Association Structure), Case 2 (Partial Association Structure), and Case 3 (Iterative Proportional Fitting) were each evaluated using the 1978 North Carolina Census of Agriculture data and the 1981, 1982, and 1983 A & P survey data. The Case 1 estimator seemed better than the other two estimators when evaluated using % Absolute Relative Differences. This report describes the analyses done and makes recommendations for further research in this area.

Keywords: Association Structure, Allocation Structure, % Absolute Relative Difference, Iterative Proportional Fitting.

ACKNOWLEDGEMENTS

This report would not have been possible without the assistance of several people:

Kay Schenk of the CSU, Chico Computer Center provided the expertise in computer programming. She worked long hours on writing, debugging, and running the many programs needed for this project. Her assistance and patience were both great.

Andrea Bowman of the Chico State Mathematics Department typed this report in a short time period. She worked very quickly and, as usual, very competently.

Barry Ford, Doug Kleweno, Bob Tortora, and Fred Vogel of USDA all contributed to the success of this project. They were instrumental in originating the research agreement and in continued assistance as the project proceded.

Of course, without the major financial support of the USDA/SRS, this project would not have been possible.

CONTENTS

SUMMARY

This research was intended to evaluate three different Categorical Data Analysis (CDA) estimators to see if any of them seemed appropriate for use in determining county-level agricultural estimates in North Carolina. The three estimators were derived and described by Noel Purcell (Purcell, 1979) in his Ph.D. dissertation. Purcell's application was to frequency data. This research was done to see if these techniques transfered to agricultural estimates of harvested acreage.

All three estimators use an associaton structure and an allocation structure. The association structure data came from the 1978 Census of Agriculture in North Carolina. The allocation structure data came from the North Carolina A & P surveys done in 1981, 1982, and 1983 (one allocation structure for each year). The Case 1 estimator used a full association structure, the Case 2 estimator used a partial association structure, and the Case 3 estimator used a full association structure and current state level agricultural estimates in an iterative fitting procedure.

The estimators were evaluated using % Absolute Relative Differences (% ARD's) and a correlation analysis. Both evaluation procedures incorporate SRS county estimates. The evaluation indicated the Case 1 and Case 3 estimators were far superior to the Case 2 estimator. The Case 1 and Case 3 estimates differed little in their evaluations with the Case 1 estimator perhaps slightly better than that of Case 3 due to the fact that Case 1

estimates are easier to derive than Case 3 estimates. Furhter research topics include incorporating NOL adjustments, reducing the magnitude of the % ARD's, adding an adjustment for the "unknown" stratum and expanding the time span of the research to include more years.

Introduction

Agricultural estimates at the county-level have been of interest for many years. They have generally been derived from population censuses, special surveys, or by using some nonprobability-based technique.

The need for improved methodology for setting county-level estimates stems from the fact that censuses and special surveys are usually very expensive. As a result, in many states, data for county-level estimates are collected from nonprobability surveys and the estimates are constructed by hand computation. Frequently, there is no sound statistical basis for the estimation techniques employed. For example, instead of using a probability-based approach, a bookkeeping type of method may be used, with the primary aim of this procedure being to avoid wide deviations from previous year estimates which were themselves the product of a similar procedure. As a result, there is usually no way to measure the precision of the estimates. Even in those states that have a large probability-based survey and computerized summary system, the process may be tedious and subjective. It is possible, given the methods and small sample sizes currently used, that the precision and accuracy of a number of county estimates are not good.

In recent years, the problem of deriving small area (such as county-level) estimates from survey data has been receiving increased attention. A number of new methods for estimation have been developed and evaluated by research statisticians in demography and health statistics. Noel Purcell

in his 1979 Ph.D. dissertation (Purcell, 1979) used a categorical data analysis (CDA) approach to try to develop efficient estimators for small domains. The evaluation of this CDA approach to SRS county-level agricultural estimates was the subject of the research reported in this paper. First, the CDA method will be explained and the estimators introduced. Next an evaluation of the methodology will be given. Finally, the results will be summarized and recommendations given for future work.

Description of Categorical Data Analysis for Small Area Estimation

The most extensive study of CDA for small area estimation is presented in Purcell's thesis (1979). A summary of his work can be found in a paper by Purcell and Kish (1980). Purcell's notation will be used in the following discussion and report.

The CDA approach to county-level agricultural estimation was evaluated on data gathered on harvested acreage in North Carolina for certain crops and land uses. Data have been collected in North Carolina for several years in a multiple frame, stratified, probability A & P survey designed to gather information from every county. A paper has been published by Ford (1981) on using these data to derive direct, synthetic, and composite estimates. Also, Ford, Bond, and Carter (1983) published a paper on further research using these data in a model that includes historical trends in acreage and production since 1972. Hence, a substantial amount of information has been gathered and evaluated (for other purposes) for North Carolina. In addition, a relatively recent census of agriculture (1978) was done in North Carolina.

This, along with the other information just mentioned, made North Carolina a good state for evaluation of CDA estimation.

The CDA estimation approach requires two data structures: an association structure and an allocation structure. The association structure consists of data that are broken into categories of the variable of interest, crosstabulated by associated variables and small areas. These data are normally obtained at some previous time, usually from a census. The allocation structure consists of data, again broken into categories of the variable of interest and crosstabulated by associated variables; but accumulated over small areas. These data are usually obtained from a current large scale survey. The allocation structure may include additional current information, such as data at the small area level accumulated over the categories of the variable of interest and of the associated variables.

For this research, the goal was to estimate the number of harvested acres for certain crops and land uses for each of the 100 counties in North Carolina. The association structure consisted of the 1200 cells of the crosstabulation of the categories i ($i=1,\ldots,20$) of the variable of interest, certain crops and land uses, by the categories g ($g=1,\ldots,6$) of the associated variable, farm size, by the counties, subscripted by h. The number of acres in each cell is denoted by N_{hig}. The allocation structure for a given year consists of a crosstabulation of crops and land uses by farm size, at the state level, obtained from the A&P survey for the particular year. Each cell of the allocation structure has a count $m_{.ig}$, where the dot denotes summation over a subscript. The allocation

structure may include additional information on current accurate county-level data on total farmland.

To estimate X_{hi}, the number of harvested acres for the twenty categories of crops and land uses, the association structure is adjusted in such a way that all interactions of variables are preserved, except those that are respecified by the allocation structure (the crops and land use by farm size margin). Then the adjusted association structure counts, denoted by X_{hig}, are summed over the associated variable (farm size) to obtain the county by crop and land use margin, whose cells, X_{hi}, are the desired estimates.

There are a number of ways to adjust the association structure, depending on the amount of information available in the association and allocation structures. Three cases were investigated by this research project.

Case 1 A full association structure was used which consisted of the 1978 North Carolina Census of Agriculture data (figure 1). The allocation structure for a given year consisted of estimates for the same categories of figure 1, at the state level. These estimates, as mentioned previously, came from the A&P survey for the particular year. By any of three methods - minimizing a weighted sum of squares, maximum likelihood, or minimizing a discriminant information criterion - the following estimate is obtained for the adjusted association structure counts:

$$x_{hig} = \frac{N_{hig}}{N_{.ig}} m_{.ig}$$

Recall that N_{hig} = number of harvested acres from the 1978 census for a particular county, crop, and farm size; $N_{.ig}$ = total # of harvested

acres in North Carolina, based on the 1978 Census, for a particular crop and farm size; and $m_{.ig}$ = total # of harvested acres in North Carolina for particular crop and farm size, based on the appropriate year's A&P survey.

Then the estimator of X_{hi}, the number of harvested acres for a particular crop or land use, at the county level is

$$X_{hi} = \sum_g X_{hig} = \sum_g \frac{N_{hig}}{N_{.ig}} m_{.ig} .$$

Case 2 Only an incomplete association structure is availalbe for this case. The association structure is a dummy structure, where total farmland data, crosstabulated by county and total farmland stratum (the hg margin of figure 1), are substituted at each level i (crops and land uses) of figure 1. These data came from the 1978 Census of Agriculture. Thus, for our problem, the 1200 cells of the association matrix were assumed to have counts $N_{h.g}$ where $N_{h.g}$ = total harvested acres for county h and farm size g. The allocation structure is the same as in Case 1 - state-level A&P estimates for all the ig categories of figure 1.

The estimator for the adjusted association structure is formed for this case as:

$$X_{hig} = \frac{N_{h.g}}{N_{..g}} m_{.ig} .$$

Hence, the county-level estimator is

$$X_{hi} = \sum_g \frac{N_{h.g}}{N_{..g}} m_{.ig} .$$

Case 3 All of the information of case 1 is available for this case. The association structure is the same as in case 1, where Census of Agriculture data was used, and the allocation structure includes the information in the allocation structure of cases 1 and 2. In addition, the allocation structure contains current accurate county-level data on total farmland. This came from the A&P survey for the current year. Estimators for the adjusted association structure are constructed using the method of iterative proportional fitting (Deming and Stephen, 1940). See Appendix A for a description of this method. Then as in cases 1 and 2, the resulting estimator X_{hig} is summed over the associated variable to obtain the county-level estimator:

$$X_{hi} = \sum_{g} X_{hig} .$$

This last case was of great interest since it utilized the most information of the three cases, and because it was the most accurate method in Purcell's application.

Evaluation of County Estimates

The three estimators described in the last section were determined and evaluated using the 1978 Census of Agruculture and the 1981, 1982, and 1983 A&P data. The results of the evaluation are given in tables and charts that follow. However, before examining these results, there are two peculiarities of the association and allocation structures which must be addressed.

The association structure was constructed using the 1978 Census of Agriculture data. The Census was not broken-down into exactly the same categories as those given in figure 1. Therefore, it was necessary to compute the numbers for the "Cropland Harvested Other Uses Acres" entries in the association matrix. The Census does have a category titled "Harvested Cropland." To get an estimate for "Cropland Harvested Other Uses Acres" for a particular county and farm size, the harvested acres for each of the crop categories in figure 1 were subtracted from the "Harvested Cropland" entry for that particular county and farm size. Occasionally, the number which resulted from this subtraction was negative. Double-cropping is the reason these negative numbers occurred. Farmers plant a crop, harvest it, and then plant a second crop on the same land. When the farm operator reports total harvested acres, he doesn't "double" the farm size. As a result, when the entries for harvested acres are summed over the individual crops reported , the total for this summation is larger than the total that results by summing over the entries reported for "total harvested acres on the farm." In order to correct for this double-cropping phenomenon and eliminate the negative numbers, the estimates for "Cropland Harvested Other Uses Acres" were determined in the following manner. When a positive number resulted from subtracting the harvested acres for the crop categories of figure 1 from the "Harvested Cropland" number, this positive number was used as the estimator for the "Cropland Harvested Other Uses Acres" entry. When this same subtraction resulted in a negative number, the entries for harvested acres for all of the crops

in the Census, for the particular county and farm size being considered, were recomputed by the following method. The entry for harvested acres for each crop was weighted so that the "new" estimates would sum up to the number reported for "Harvested Cropland." That is, if A = the number on the "Harvested Cropland" acres line in the Census and B = total number of harvested acres computed by summing over the harvested acres for all the crops in the Census, then each entry in the Census for harvested acres was multiplied by A/B. In other words, the harvested acres by crop were weighted so that they would sum up to the number reported for total harvested acres. The subtraction described earlier was then done again and this time a positive number resulted. This number was used as the estimator for the "Cropland Harvested Other Uses Acres" entry. This adjustment of the association structure only affected the estimators in Cases 1 and 3. For Case 2, the association structure entries are all given by $N_{h.g}$ and no adjustment was necessary. The $N_{h.g}$ came from summing up the Census figures over the crops and land use categories for each county and farm size.

The allocation structure also required some adjustment when computing the Case 3 estimators. In order to do the Iterative Proportional Fitting (IPF) procedure, it is necessary that $\sum_{h} m_{h..} = \sum_{i,g} m_{.ig}$. That is, the IPF procedure will not converge and satisfy the IPF criterion if the state total of harvested acreage derived by summing over the total harvested acres for each county, does not equal the state total derived by summing state totals for crop and farm size combinations. Therefore when working with Case 3, the entries in the allocation structure were adjusted by multiplying them

by the weight C/D where $C = \sum_{h} m_{h..}$ and $D = \sum_{i,g} m_{.ig}$. That is, the entries were forced to sum to C. The C total was thought to be perhaps more accurate than the D total. It was felt that the operators may be more accurate in reporting their total land acreage than in reporting breakdowns by crops and land uses. This adjustment was very small for all three years. The weights were all very close to one (specifically, .98122 for 1981, 1.0111 for 1982, and.97415 for 1983). Despite the fact that the allocation structure only had to be adjusted for Case 3, the estimators for Cases 1 and 2 were also computed using the adjusted allocation structure. Thus, all three cases were evaluating the same data structures. In addition, for comparative purposes, estimators for Cases 1 and 2 were computed using the unadjusted (i.e., unweighted) allocation structure. The evaluation of the results of using this unadjusted allocation structure are given in Appendix B. Comments will be made about these results after the evaluation of the estimators computed using the adjusted allocation structure.

An evaluation technique was needed in order to compare the three types of estimators. Methods of estimation of the variance-covariance matrix of the CDA estimates are given by Purcell. However, all of these methods are rather complex, and Purcell did not actually compute any variance-covariance matrices in his evaluation of CDA estimates. He points out that the variances of these estimates depend mainly on the variances of the allocation structure estimates, which can be controlled with sample design. The bias is therefore a more important source of error, and Purcell gives methods for estimating its size. The bias measure used for evaluation

purposes in this report is the percentage absolute relative difference (% ARD). For this type of evaluation, the CDA county estimates are compared with values from a current census, or from independent sources, by computing the %ARD:

$$\%ARD = \frac{|x_{hi} - X_{hi}|}{X_{hi}} \times 100 \quad ,$$

where x_{hi} = CDA county estimate, and X_{hi} = the "true" county-level value.

The %ARD formed the major part of the evaluation of the three estimators. The three estimators were compared with respect to the mean, median, and standard deviation of this measure across the 100 counties for seven different crops.

In addition, the Pearson product-moment correlation coefficients were computed to examine the degree of the relationship between he three different estimates and their respective "true" values.

As just stated, both the % ARD and Pearson correlation coefficients require the "true" county-level value before they can be computed. The official SRS county estimates were used as the "true" county-level values. The problem with using these estimates in the evaluation is that virtually none have check data for them. Only cotton, tobacco and peanuts can be verified by ASCS figures. Cotton is of no use because none was planted in North Carolina in 1981, 1982 and 1983. The tobacco figures would only be helpful for 1981 and 1982 since the 1983 crop appears to have been "estimated" exactly by the SRS estimates. Peanuts were planted in less than half of the

100 counties in all three study years. Therefore, despite the fact that the SRS estimates may be subjective and possibly inaccurate, with no measure of reliability, the CDA estimates were evaluated in comparison with these figures.

The main results of this evaluation are summarized in Tables 1 through 4 and Figures 2 through 15. A large quantity of results are presented and therefore, a short summary of the principal findings of the evaluation is given to aid in guiding the reader through these results. This summary is followed by an indepth discussion of the specific results with respect to each evaluation technique used.

Summary of principal findings: The first and foremost finding of the results was that the Case 1 and Case 3 estimators were clearly superior to the Case 2 estimator. However, the Case 1 and Case 3 estimators were very close in overall performance with Case 1 estimates perhaps slightly superior for most crops in the evaluation. There was a general deterioration over time for all three cases due to some breakdown over time in the validity of the assumptions concerning the stability of the preserved association structure. Sometimes the Case 3 estimates deteriorated less over time than the Case 1 estimate but generally their performance was similar with Case 1 even better over time for a few crops. Figures 9 through 15 show that the standard deviations of the % ARD's for the Case 1 and Case 3 estimates were quite close in magnitude.

Another feature of the results was the wide degree of difference in the performances of the three different estimators for the different crops.

For example, harvested acres of corn was more accurately estimated than harvested acres of oats. Generally, the more harvested acres a crop had, the better the estimates were.

Even though there was a difference in the level of accuracy of the estimates for the different crops, the relative accuracy of the three estimators, with respect to each other, is generally the same for all crops. Case 1 and Case 3 estimates outperform Case 2 and the Case 1 estimates frequently do better than the Case 3 estimates.

Analysis of the % ARD's: Tables 1 - 3 and Figures 2 through 15 summarize the results of the analysis of the % ARD's. When Tables 1 and 2 are compared, it is seen that in most cases the mean % ARD across counties is larger than the median % ARD. For example, the mean % ARD for the Case 1 estimator for soybeans is 34.41 while the median % ARD is 30.55. This indicates a noteworthy degree of positive skewness in the size of the % ARD's across counties. The exceptions to this "rule" are sorghums and oats for the Case 1 and Case 3 estimators which seemed to have negatively skewed % ARD's. This skewness seemed to suggest that the median was a more representative measure of "average" bias for the county estimates than was the mean. As a result, the following discussion concentrates on the evaluation of the median % ARD's, however, the results of the analysis of the mean % ARD's is practically identical. Since the median % ARD's for peanuts were zero over all estimators and years, the mean % ARD's are used for evaluation on peanuts.

The medians of the % ARD's are displayed in Table 2 and it is apparent that the accuracy of the different estimators varied considerably. In

addition, within each estimator there were differences over both crops and years. These differences are seen more clearly in Figures 2 through 8.

The Case 2 estimator (the synthetic estimates) performed quite poorly in relation to the other two estimators. The median % ARD's were higher for the Case 2 estimates than for the Case 1 and Case 3 estimates for all crops except for sorghums and oats. However on examining the standard deviations of the % ARD's for these two crops, it is seen that the Case 2 estimates had huge standard deviations when compared to the standard deviations of the % ARD's for the other two estimates. Certainly an estimator that results in a lower "average" bias across counties but has larger variability for the bias estimates is not necessarily a better estimator. Table 3 contains the standard deviations of the % ARD's across counties. Figures 9 through 15 more clearly compare these figures for the Case 1 and Case 3 estimators. The Case 2 estimates had standard deviations so large that they were not included on the figures. Generally, they demonstrate that the Case 1 and Case 3 estimates had similar degrees of variability.

Correlation Analysis: It was felt that the degree of linear relationship between the estimates and their true values might be of importance. This relationship was measured using Pearson product-moment correlation coefficients. The results of the correlation analysis are given in Table 4. There are two clear findings in these results. One conclusion is that, for all crops, the use of the full association structure (Cases 1 and 3) leads to far superior correlations than the corresponding figures for the Case 2 estimator which has an incomplete association structure. Secondly, there does not seem to

be any real difference between the Case 1 and Case 3 estimators as measured by the correlations.

The conclusion that the Case 2 estimator is inferior to both the Case 1 and Case 3 estimators is not surprising. It merely underscores the intuitive belief that the more knowledge there is at the start (a full association structure), the better the resulting estimates will be in the end. This result agrees with that of Purcell's result in his example. The fact that the correlations display great stability over time also agrees with Purcell's results.

Conclusions and Recommendations

On examining the results of both the % ARD and correlation analyses, it seems clear the Case 1 and Case 3 estimators consistently and significantly outperform the Case 2 estimator. As previously stated, this conclusion agrees with the intuitive belief that a full association structure should be superior to a partial association structure. Unfortunately, there is no pattern in the results which demonstrates superiority between the Case 1 and Case 3 estimators. Since the Case 1 estimator is easier to compute and requires less information, the logical conclusion is to recommend using Case 1 estimates when the necessary association and allocation structures are available. However, caution is warranted before such a recommendation is made. In Purcell's dissertation the Case 3 estimator was generally superior to the Case 1 estimator. A fact to note is that Purcell's example had a ten year analysis period whereas this report used data gathered over

at most five years, 1978 to 1983. The Case 1 and Case 3 estimators produced similar median % ARD's for the first four years of his evaluation but the Case 3 estimates were generally better with a more dramatic difference occurring toward the end of his ten year period. That is, the addition of current $m_{h..}$ information was more helpful as time went on and the census data in the association structure became more out of date. The results for the agricultural estimates in this report may or may not change over an extended time period. The frequency of the Census in North Carolina is therefore important. If the Censuses are five or six years apart, the Case 1 estimator would seem to be the one to use. A longer period of time between Censuses could mean the Case 3 estimator would be better in the later years.

There may be a problem with the Case 3 estimator which has not been discussed yet in this report. When doing the Case 3, IPF, procedure a convergence criterion is predetermined. The IPF procedure continues to run until this convergence criterion is met or until it is otherwise commanded to stop. The procedure is described in Appendix A. Briefly, the IPF procedure was to force either the sum of the estimates over all crops and farm sizes for a given county to be within .5 acre of the known $m_{h..}$ figures or to force the sum of the crop and farm size estimates over all counties to be within .5 acre of the known $m_{.ig}$ figures. Adjustments are made to all these figures at each iteration. This iterative procedure continues until the .5 acre criterion is satisfied. The IPF procedure converged for both the 1982 and 1983 data sets. The convergence occurred after approximately fifteen iterations. However, the 1981 data set failed to converge and the procedure was halted after 100 iterations. No real change in the iterative estimates occurred

after 15 iterations. The 1981 estimates were within 125 acres of the $m_{h..}$ figures and within 476 acres of the $m_{.ig}$ figures. The reason for this nonconvergence is not known. An interesting fact is that the magnitude of the numbers for 1982 and 1983 is very close while the 1981 numbers had a much larger range of values. This could mean the IPF procedure is "data dependent" and will not converge for some configurations of data. This is a problem which requires further investigation before the Case 3 estimation technique is routinely applied.

Another point which should be addressed is the size of the %ARD's in this report. The median % ARD's from Table 2 range from 26.37 to 85.23. In Purcell's example, the median ARD's ranged from .732 to 19.417. The standard deviations of the % ARD's in Purcell's thesis were from 1.483 to 20.628. Table 3 shows that the standard deviations of the % ARD's in this report ranged from 26.75 to 7752.12. From the definition of % ARD it is seen that the smaller the % ARD, the closer the estimates are to the "true" values. The magnitude of the numbers in this report is therefore of concern. The "large" % ARD's in this report means the CDA estimates differed in some cases from the SRS estimates by quite a bit. This difference could be due to a number of factors. First, the SRS estimates are themselves somewhat suspect in that they are not checked or verified as to accuracy. Second, the CDA procedures may not be working as well for agricultural acreage data as they did for the frequency counts in Purcell's thesis. Third, the CDA estimates may need further "adjustment" to make them "correct." The CDA county estimates did not have any adjustment for NOL (required due to incompleteness of the list frame). Also,

the CDA estimates had no adjustment for the "unknown" stratum composed of farm operators with farms of unknown size. These adjustments may improve the performance of the CDA estimates and reduce the % ARD's. Further research is called for to determine how best to make these adjustments and the effect on the % ARD's once the adjustments are made.

Generally speaking, the CDA approach to agricultural county estimation seems promising. It is a probability-based approach which uses past and current data to construct county-level estimates. The Case 1 estimator seems to be the best so far. Further investigation of the CDA approach seems called for to see if it can be routinely used to derive county-level agricultural estimates.

Table 1. Means of the Percentage Absolute Relative Differences of the Three Different Estimates

Estimator	Year	Crop						
		Corn-Grain	Soybeans	Tobacco	Peanuts	Sorghums	Oats	Barley
Case 1	1981	38.98	34.41	32.23	19.90	61.25	52.78	41.68
	1982	46.59	41.78	40.37	23.16	66.43	56.32	56.72
	1983	44.08	44.44	0.00	24.53	65.33	56.04	45.61
Case 2	1981	252.16	354.21	1210.29	1034.76	183.77	93.97	220.23
	1982	202.36	320.24	874.63	977.52	127.93	111.21	381.15
	1983	167.83	169.80	0.00	1072.07	134.60	99.64	292.28
Case 3	1981	50.52	38.33	37.50	21.80	66.24	50.62	43.38
	1982	50.13	42.76	41.13	23.20	67.17	55.48	49.55
	1983	46.16	45.27	0.00	25.44	67.77	56.11	48.34

Table 2. Medians of the Percentage Absolute Relative Differences of the Three Different Estimates

Estimator	Year	Crops						
		Corn-Grain	Soybeans	Tobacco	Peanuts	Sorghums	Oats	Barley
Case 1	1981	37.42	30.55	25.27	0.00	71.21	62.49	27.17
	1982	44.60	44.13	36.07	0.00	85.23	64.77	32.83
	1983	44.07	46.79	0.00	0.00	82.48	62.81	35.82
Case 2	1981	59.17	52.62	49.26	0.00	66.46	47.85	50.97
	1982	59.16	54.04	55.09	0.00	68.33	55.78	57.13
	1983	60.97	51.09	0.00	0.00	62.77	58.72	59.19
Case 3	1981	42.30	32.07	35.87	0.00	75.62	58.75	26.37
	1982	46.58	40.83	43.14	0.00	84.47	64.13	47.58
	1983	43.58	43.41	0.00	0.00	81.40	63.53	44.89

Table 3. Standard Deviations of the Percentage Absolute Relative Differences of the Three Different Estimates

Estimator	Year	Crop						
		Corn-Grain	Soybeans	Tobacco	Peanuts	Sorghums	Oats	Barley
Case 1	1981	29.74	28.26	30.33	35.79	42.43	36.76	42.64
	1982	28.76	28.96	28.65	36.95	39.12	36.73	53.10
	1983	27.36	28.75	0.00	37.59	39.87	33.38	40.00
Case 2	1981	754.81	2322.89	7752.12	4017.85	399.23	358.04	651.03
	1982	680.78	2236.90	5361.38	3369.39	267.20	369.19	1663.56
	1983	490.82	853.45	0.00	4052.12	284.57	273.40	1028.94
Case 3	1981	39.60	30.07	28.04	36.83	50.22	38.85	44.19
	1982	40.97	28.84	29.60	37.41	38.80	36.95	47.83
	1983	26.75	31.26	0.00	38.81	43.35	34.27	39.70

Table 4. Pearson Correlation Coefficients Between the Three Different Estimates and Their Respective SRS Values

Estimator	Year	Crop						
		Corn-Grain	Soybeans	Tobacco	Peanuts	Sorghums	Oats	Barley
Case 1	1981	.82	.98	.97	.99	.89	.69	.95
	1982	.83	.95	.98	1.00	.86	.74	.92
	1983	.82	.96	--	.99	.85	.84	.98
Case 2	1981	.77	.72	.76	.58	.21	.30	.16
	1982	.75	.72	.76	.54	.22	.33	.09
	1983	.75	.74	--	.55	.25	.47	.14
Case 3	1981	.74	.93	.93	.94	.89	.71	.91
	1982	.82	.87	.94	.97	.84	.76	.90
	1983	.80	.91	--	.99	.79	.82	.95

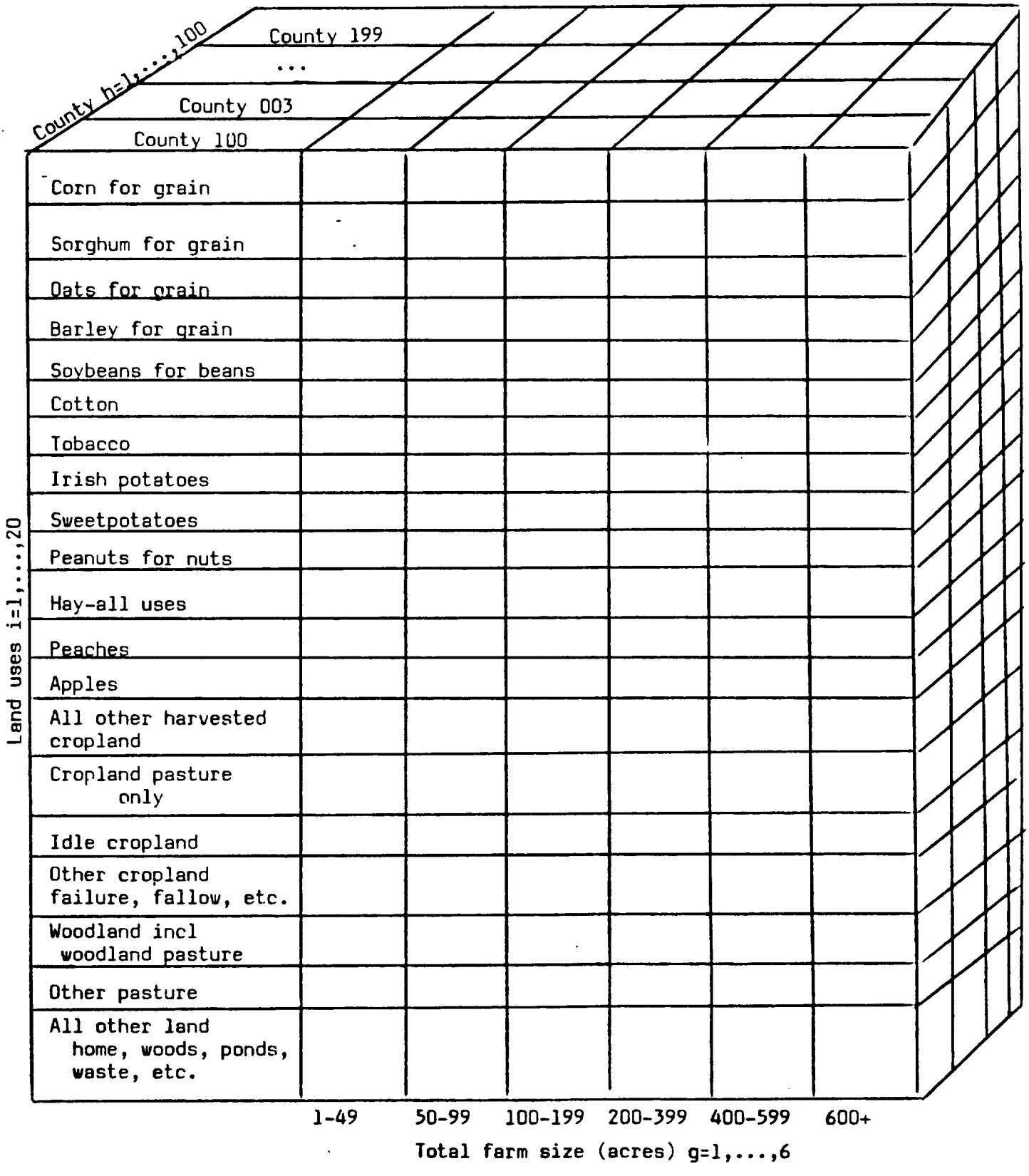

Figure 1 - Association structure for cases 1 and 3, based on Census of Agriculture.

Note: crop acreages are harvested acres.

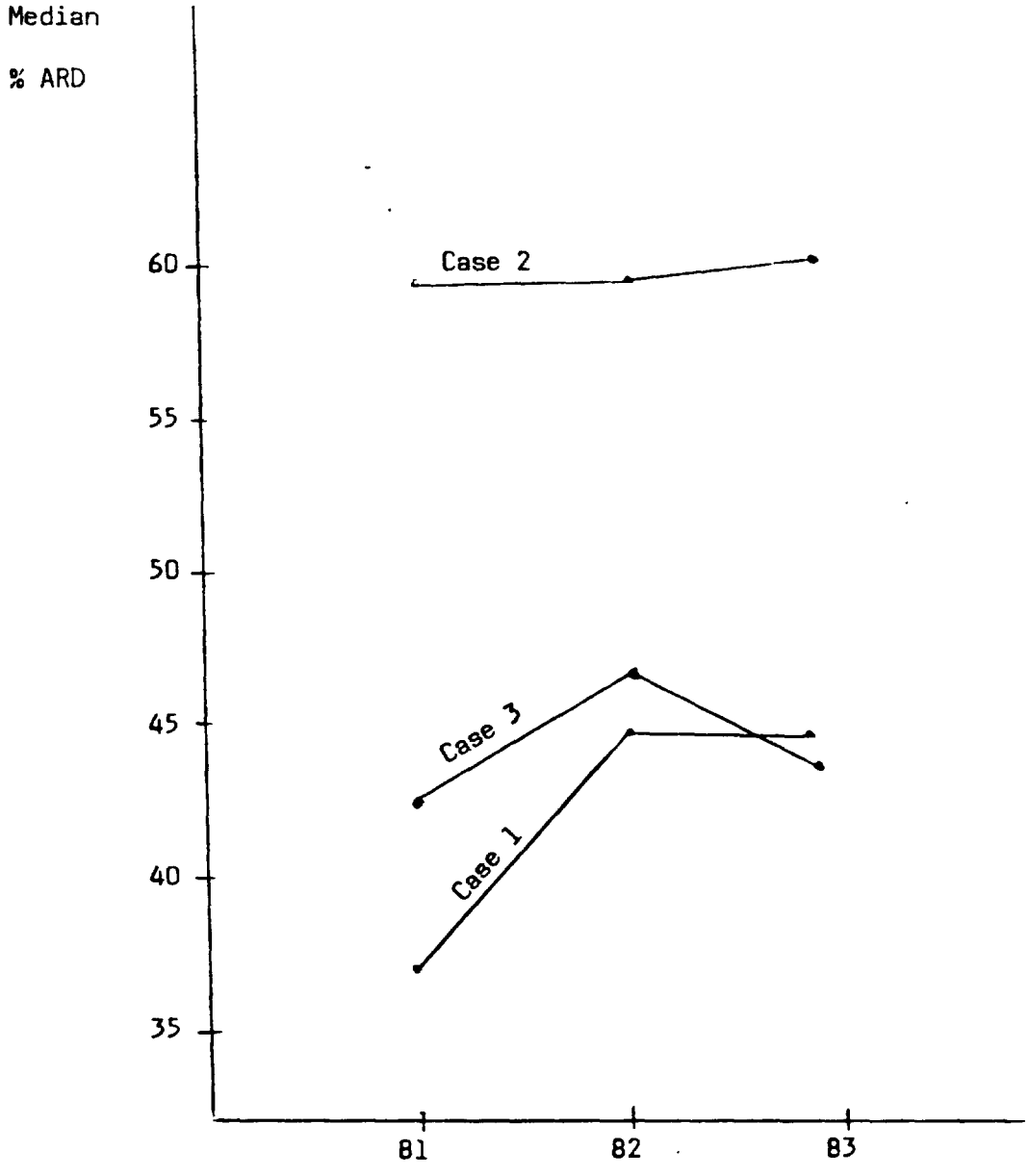

Figure 2. Median % ARD of the Three Different Estimates of Harvested Acres of Corn for Grain by Year

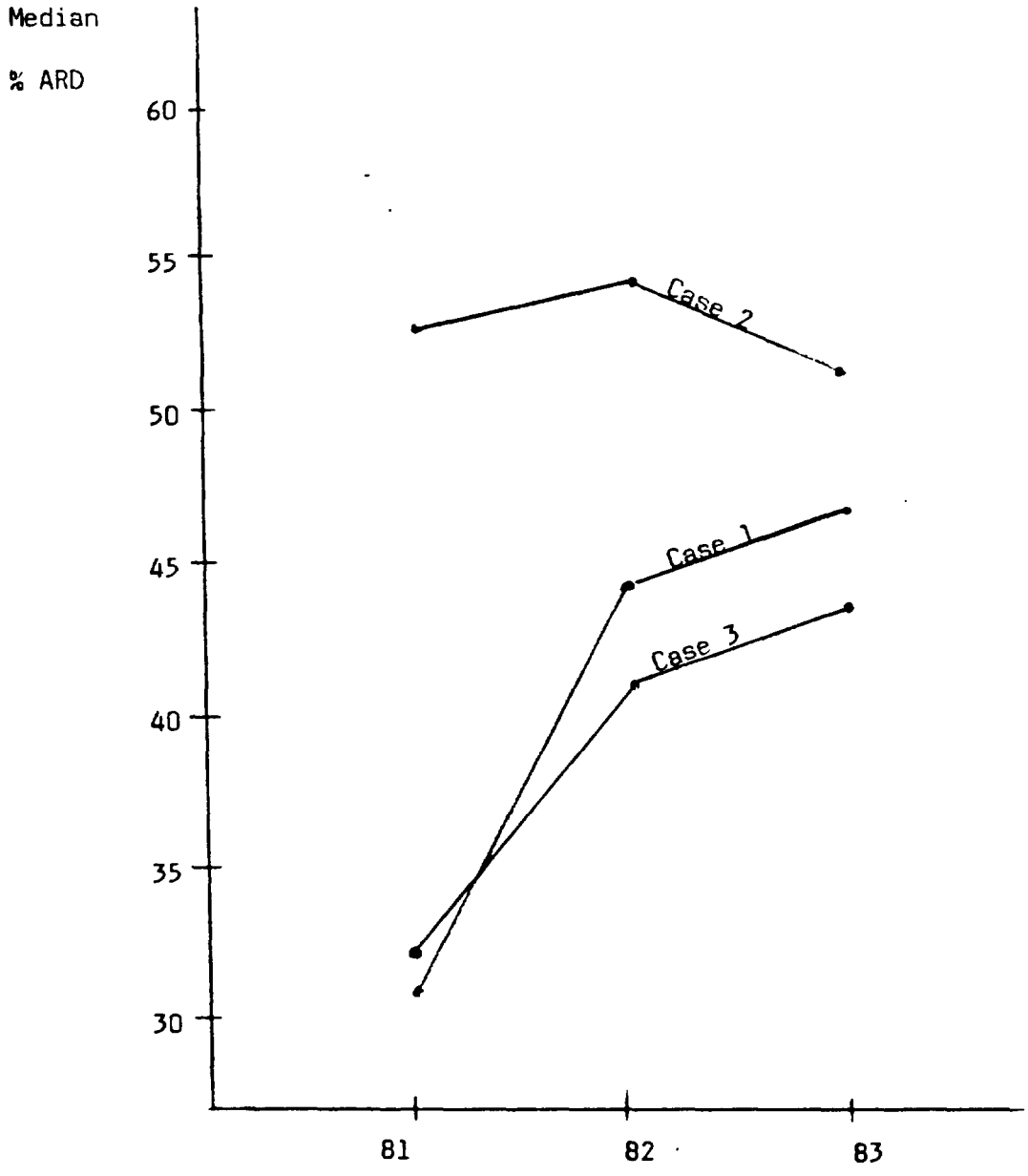

Figure 3. Median % ARD of the Three Different Estimates of Harvested Acres of Soybeans for Beans by Year

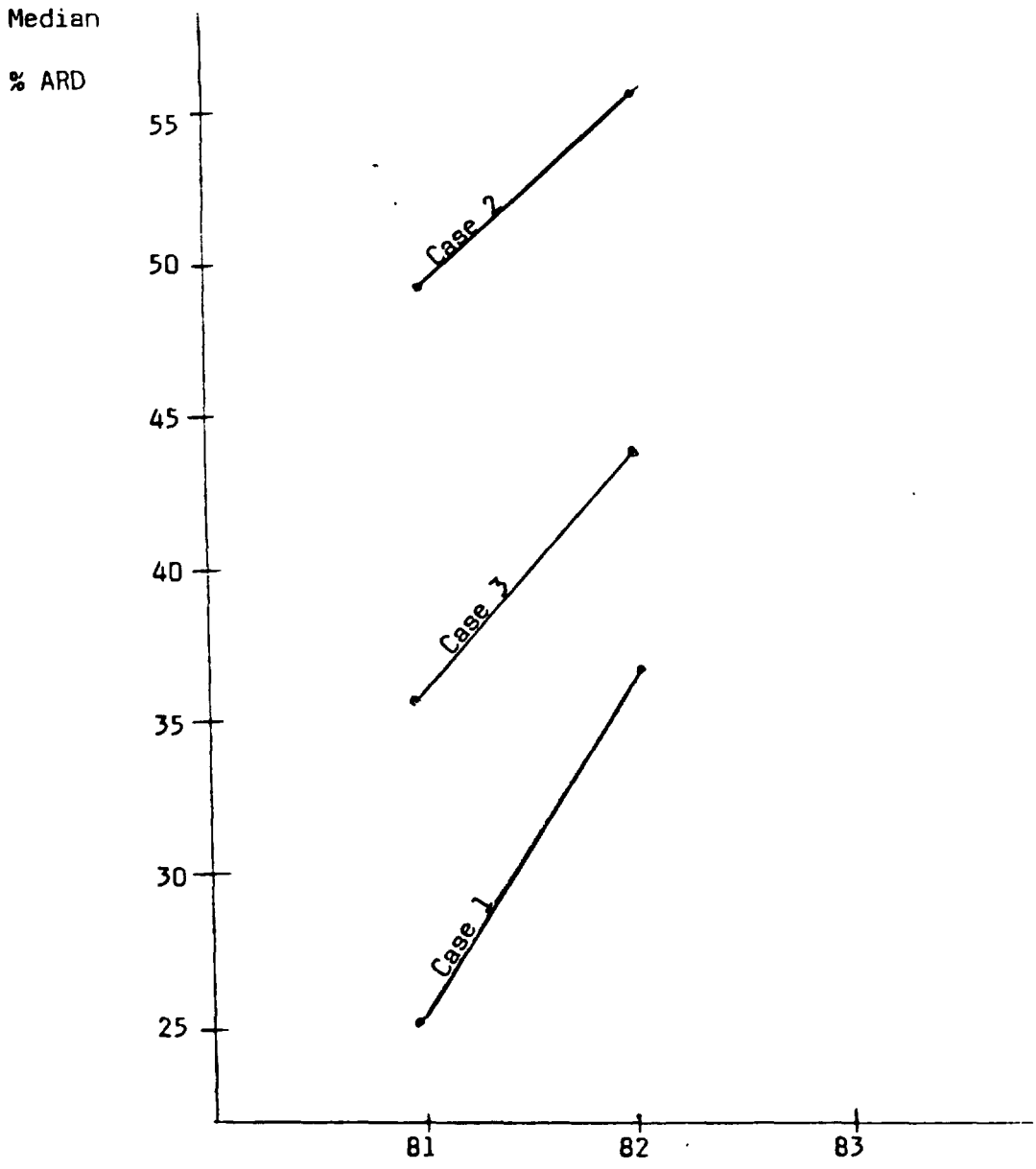

Figure 4. Median % ARD of the Three Different Estimates of Harvested Acres of Tobacco by Year

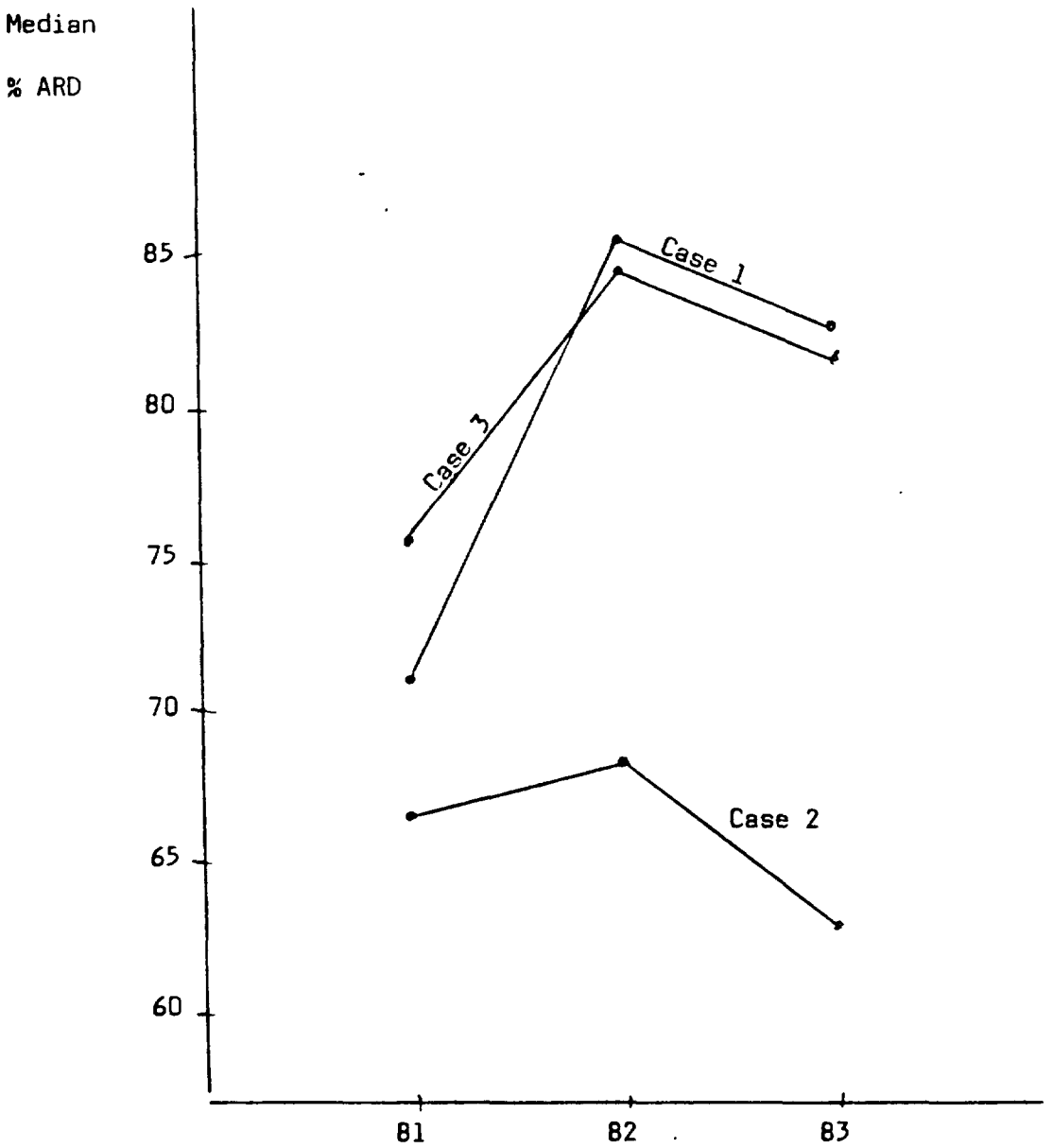

Figure 5. Median % ARD of the Three Different Estimates of Harvested Acres of Sorghams for Grain by Year

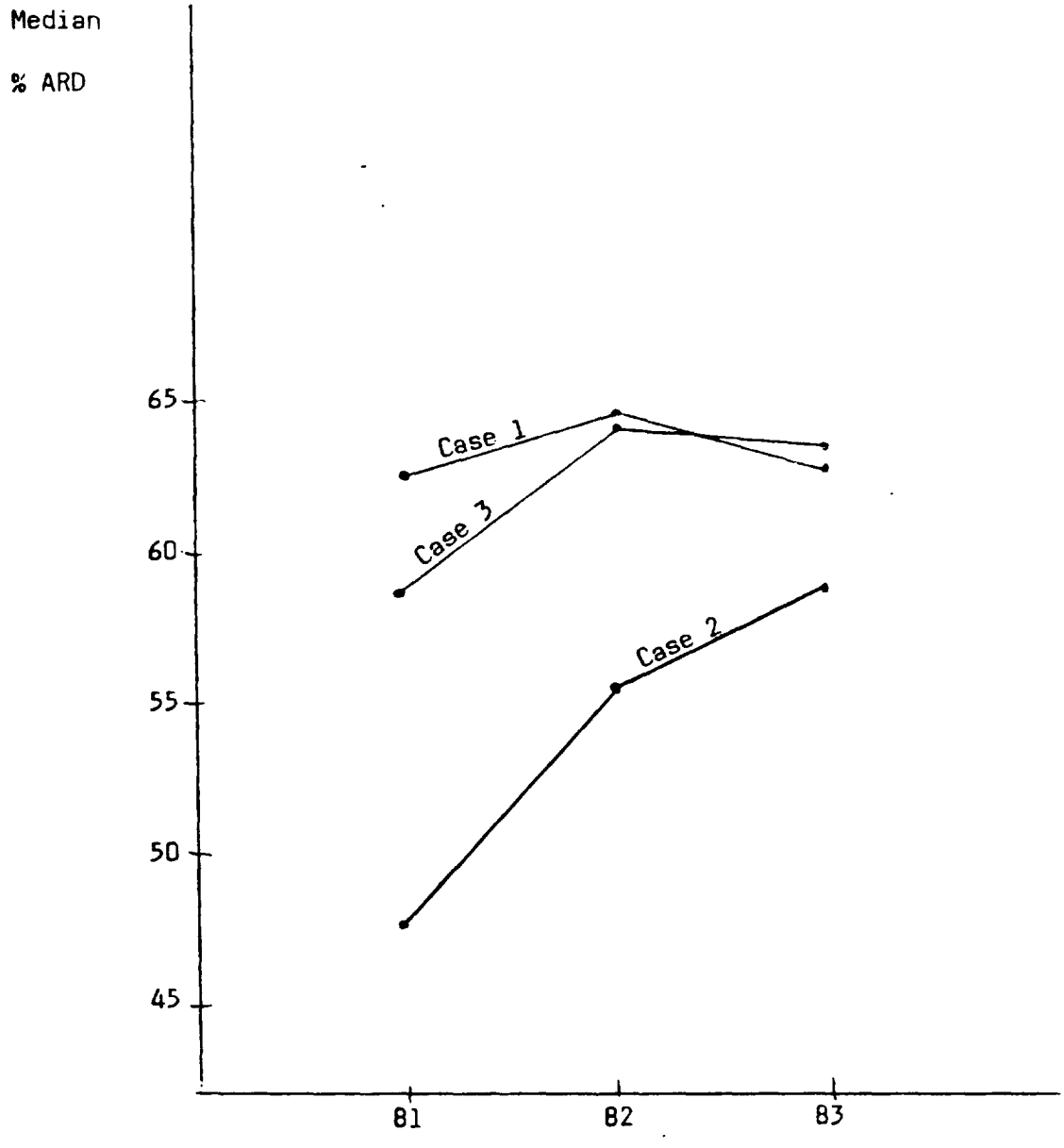

Figure 6. Median % ARD of the Three Different Estimates of Harvested Acres of Oats for Grain by Year

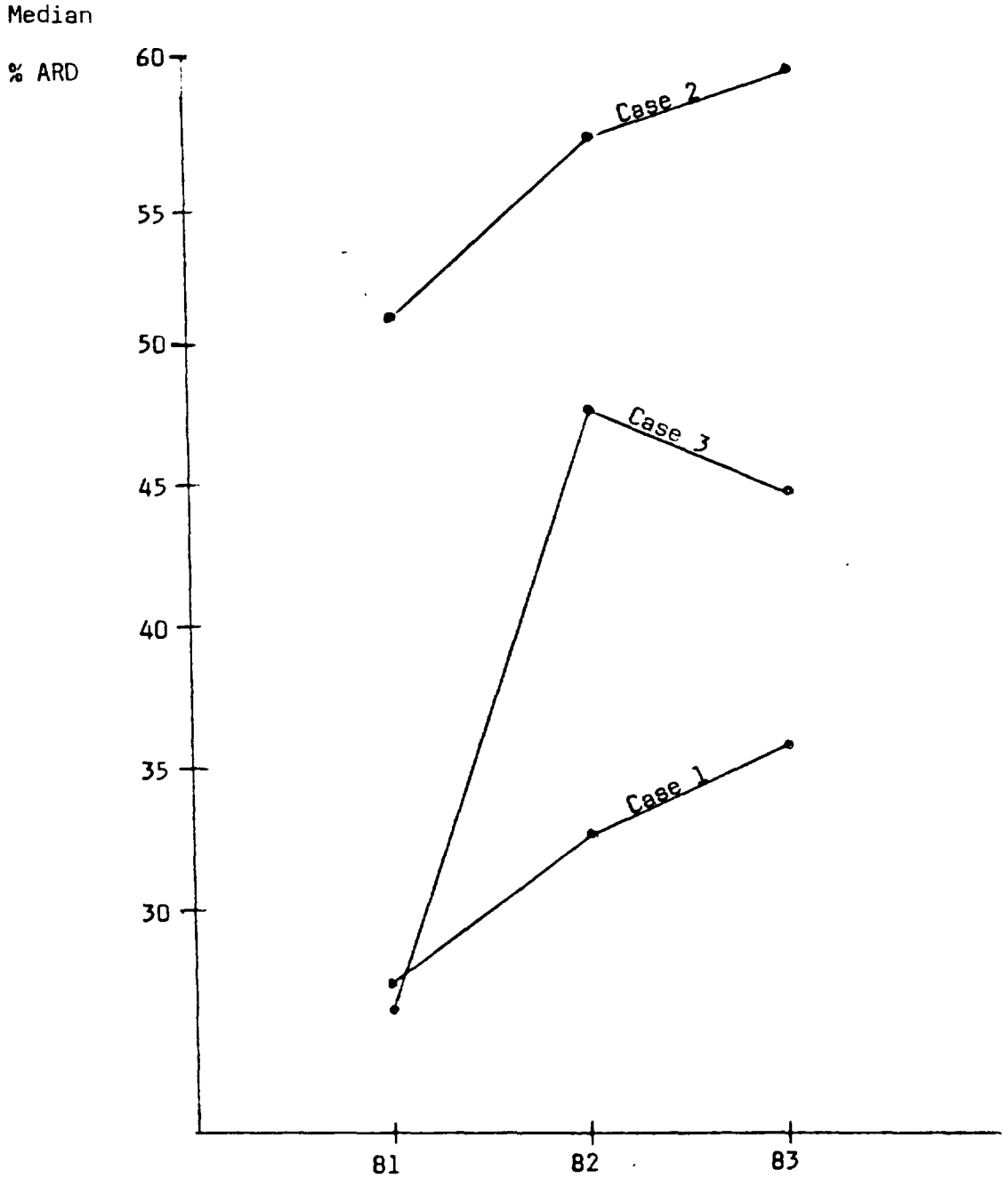

Figure 7. Median % ARD of the Three Different Estimates of Harvested Acres of Barley for Grain by Year

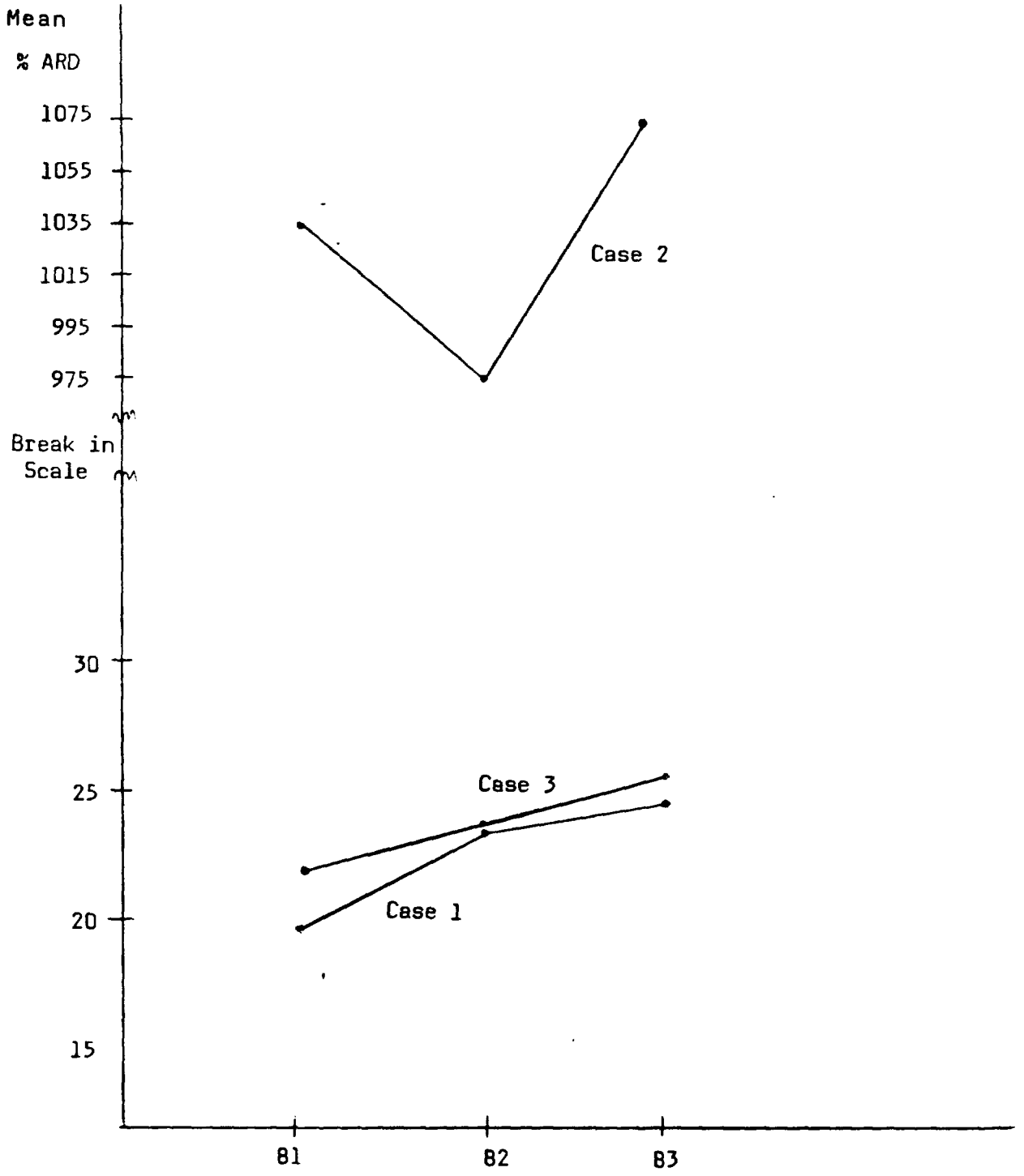

Figure 8. Mean % ARD of the Three Different Estimates of Harvested Acres of Peanuts for Nuts by Year

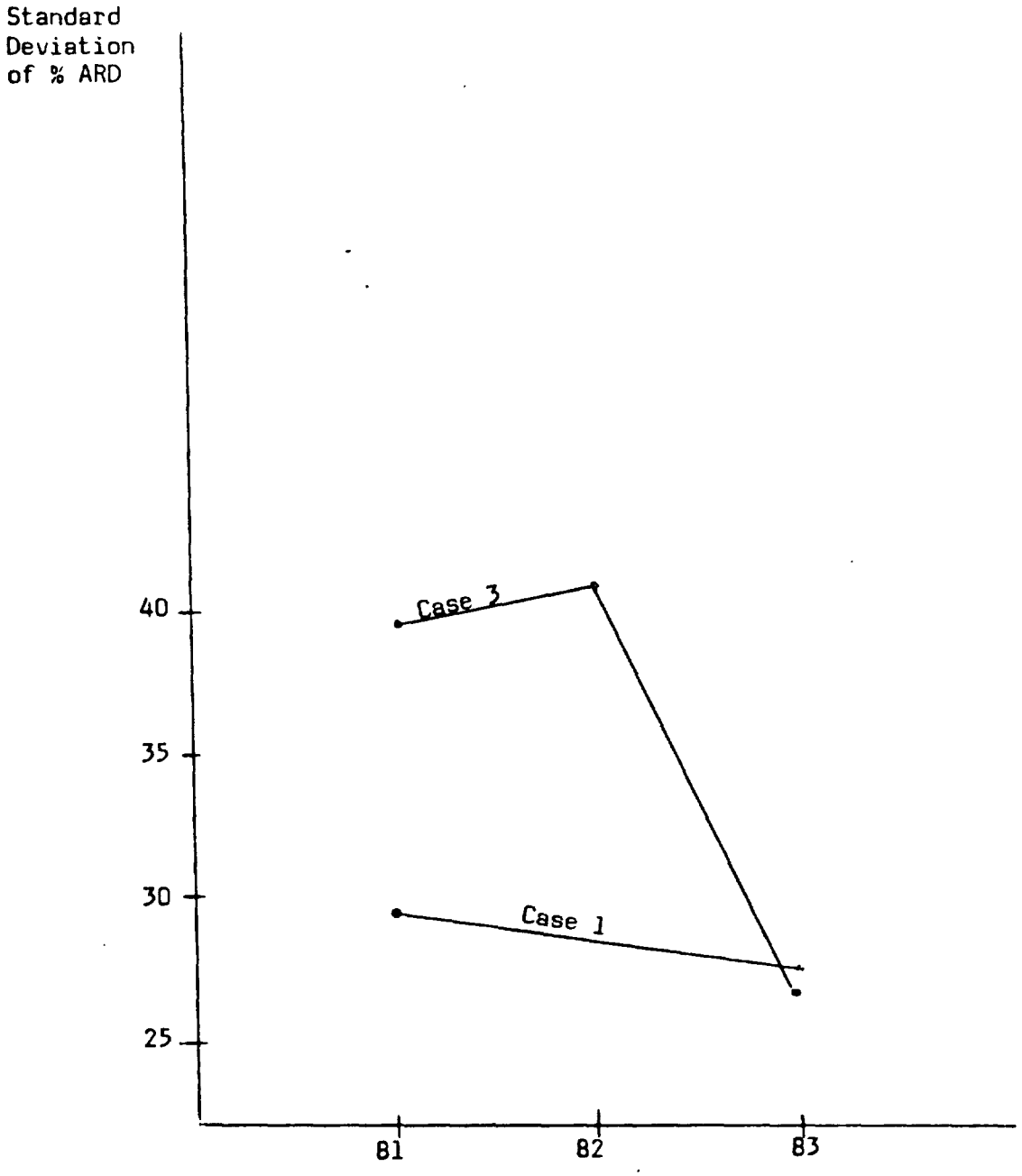

Figure 9. Standard Deviation of % ARD of Case 1 and Case 3 Estimates of Harvested Acres of Corn For Grain by Year

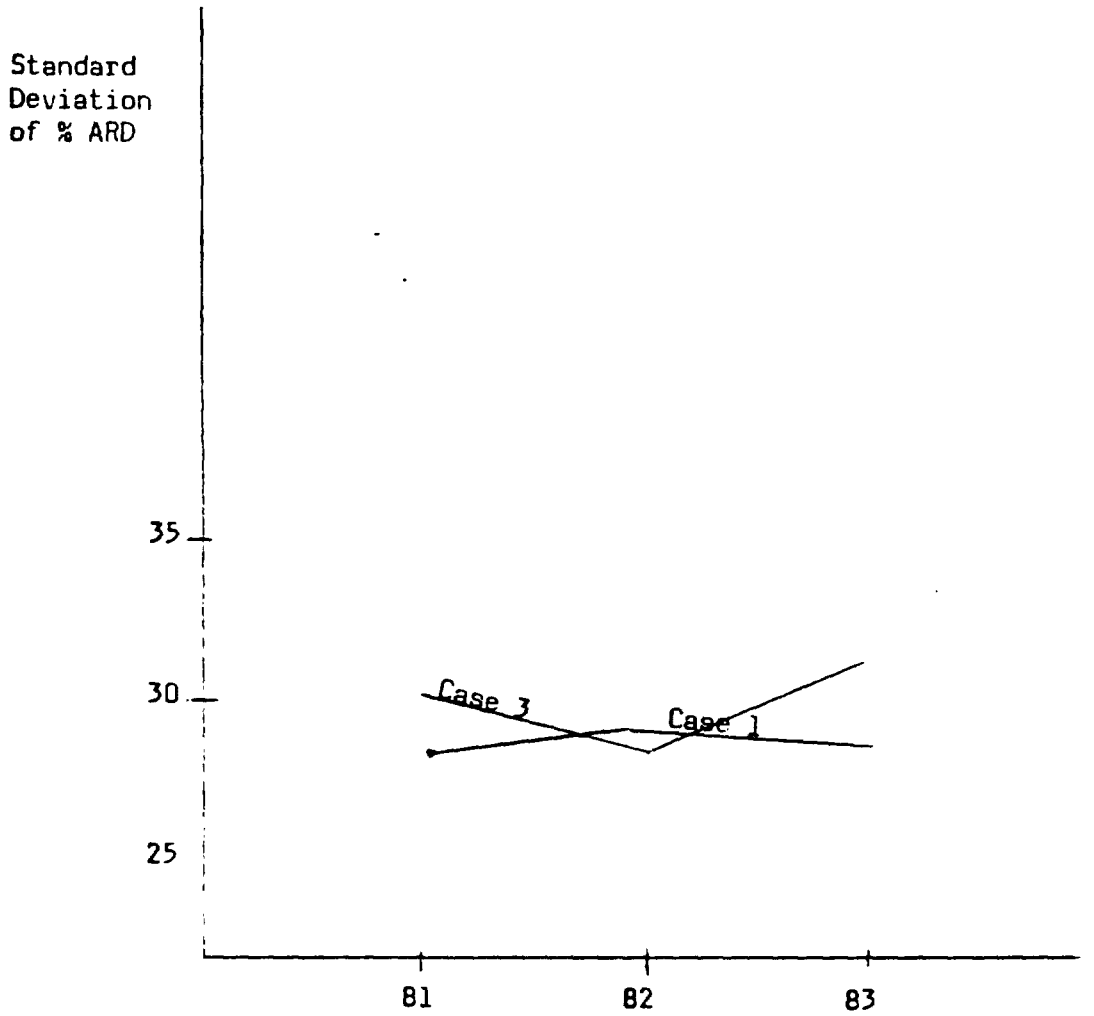

Figure 10. Standard Deviation of % ARD of Case 1 and Case 3 Estimates of Harvested Acres of Soybeans For Beans by Year

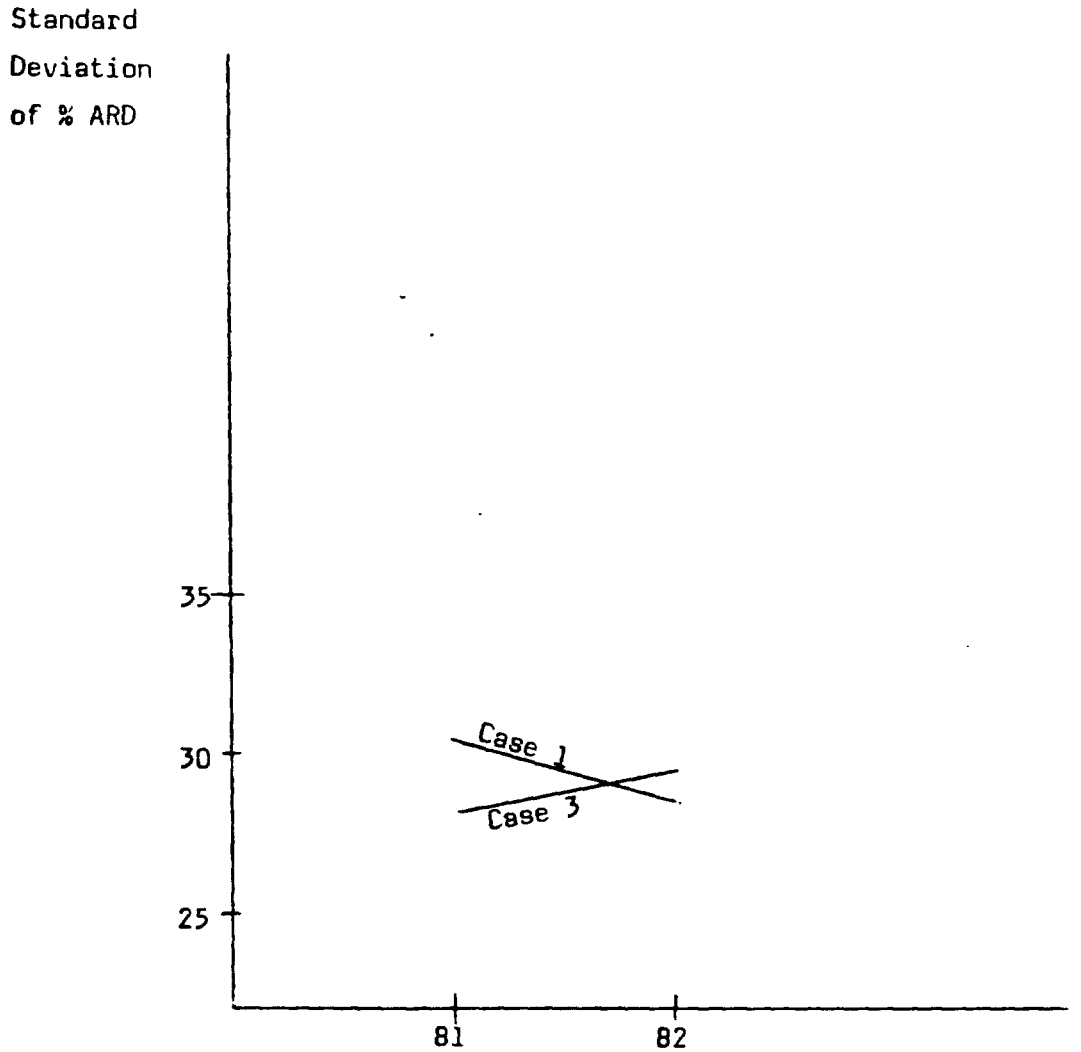

Figure 11. Standard Deviation of % ARD of Case 1 and Case 3 Estimates of Harvested Acres of Tobacco by Year

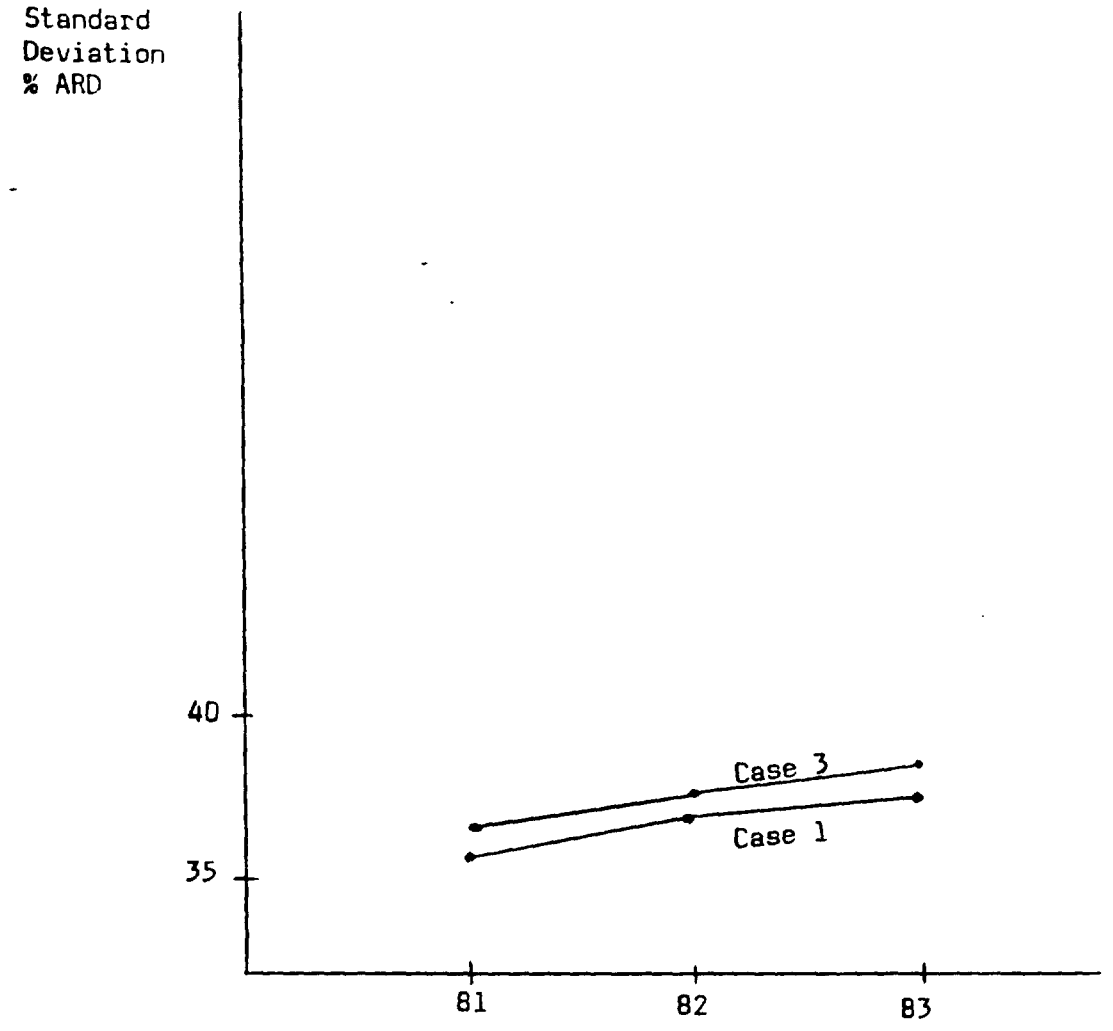

Figure 12. Standard Deviation of % ARD of Case 1 and Case 3 Estimates of Harvested Acres of Peanuts for Nuts by Year

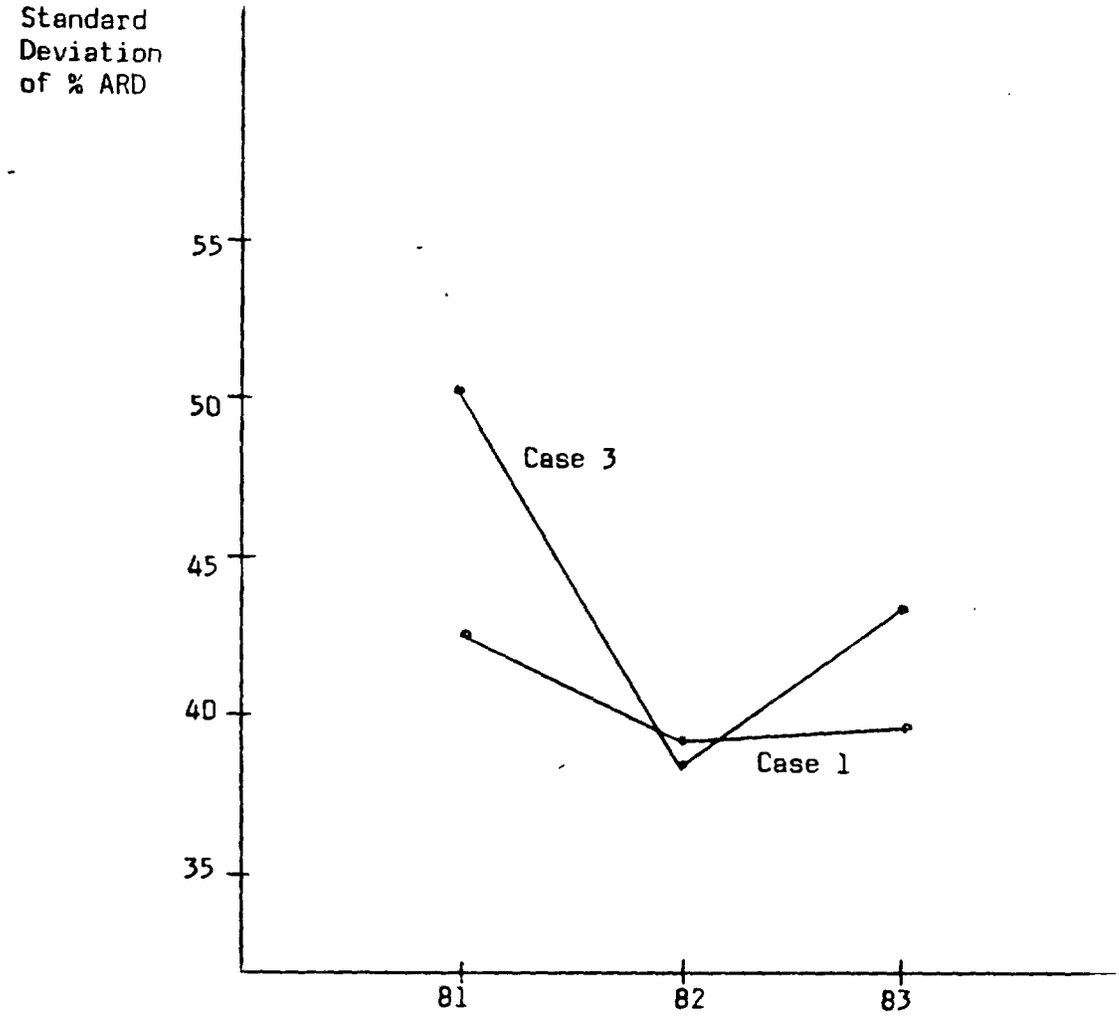

Figure 13. Standard Deviation of % ARD of Case 1 and Case 3 Estimates of Harvested Acres of Sorghums For Grain by Year

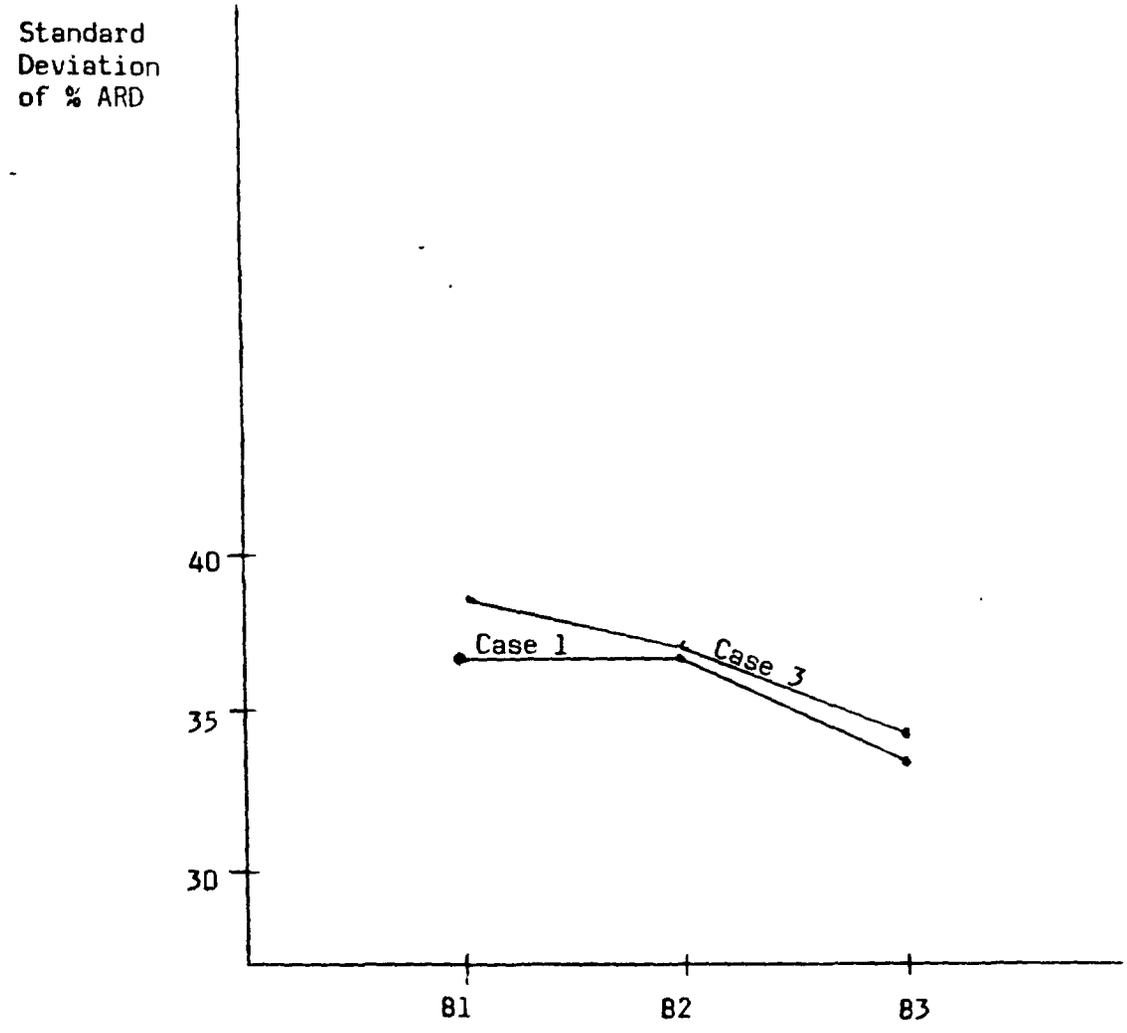

Figure 14. Standard Deviation of % ARD of Case 1 and Case 3 Estimates of Harvested Acres of Oats for Grain By Year

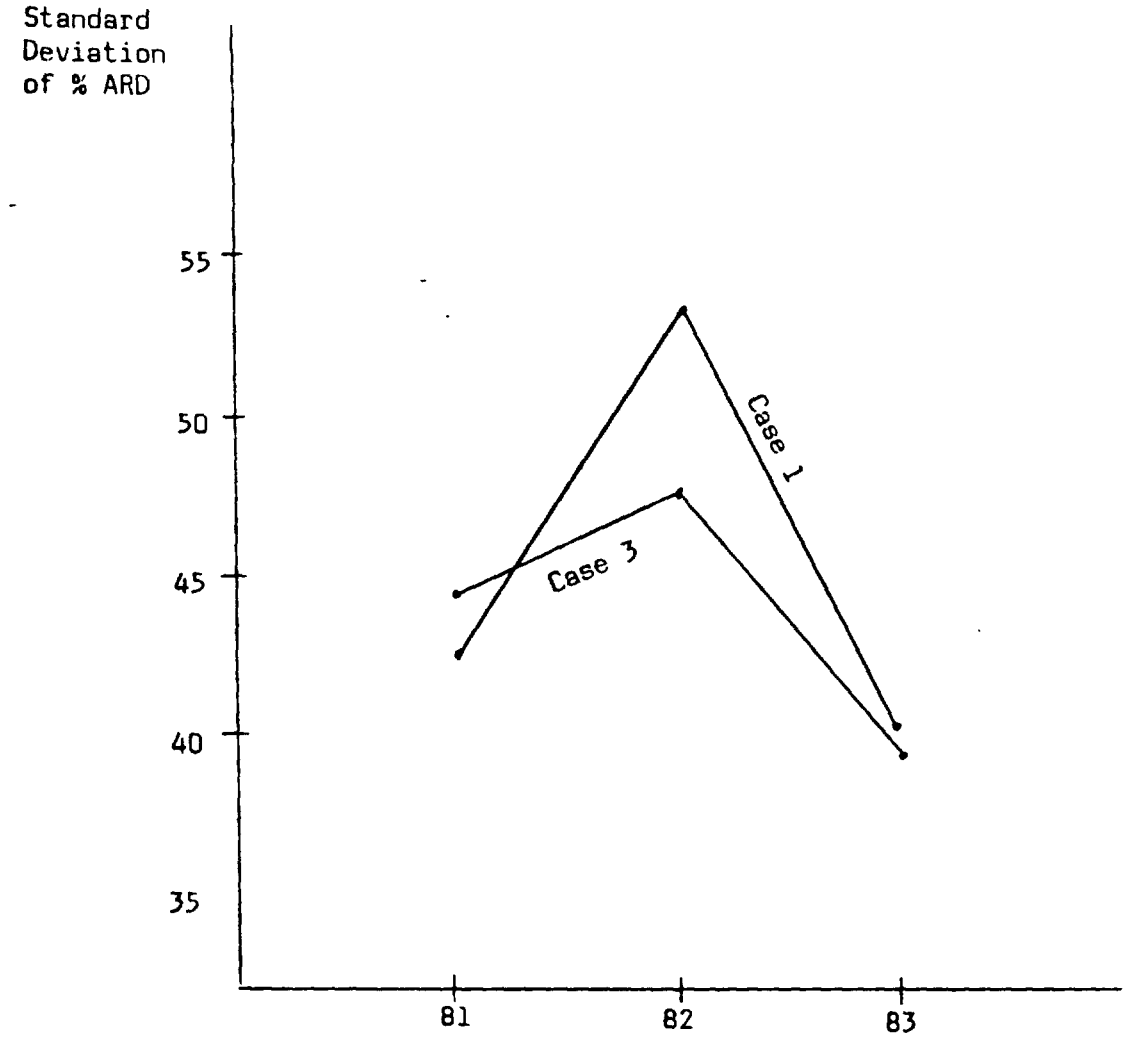

Figure 15. Standard Deviation of % ARD of Case 1 and Case 3 Estimates of Harvested Acres of Barley for Grain By Year

Biblingraphy

Deming, W.E. and F. F. Stephan. (1940). "On a Least Squares Adjustment of A Sampled Frequency Table When the Expected Marginal Totals Are Known," Annals of Mathematical Statistics, 11, pp. 427-444.

Ford, B.L. (1981). "The Development of County Estimates in North Carolina," Statistical Reporting Service, USDA, Washington, D.C.

Ford, B.L., D. Bond, and N.J. Carter. (1983). "Combining Historical and Current Data to Make District and County Estimates for North Carolina," Statistical Reporting Service, USDA, Washington, D.C.

Purcell, N.J. (1979). "Efficient Estimation for Small Domains: A Categorical Data Analysis Approach," unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Purcell, N.J. and L. Kish (1980). "Postcensal Estimates for Local Areas (or Domains)," International Statistical Review, 48, pp. 3-18.

Appendix A

This appendix contains a description of the Iterative Proportional Fitting (IPF) procedure that was used in deriving the Case 3 estimates. The notation and description follows that of Purcell (1979) and the interested reader is directed to pages 59-60 of his Ph.D. thesis.

Recall the IPF procedure used a full association structure, an allocation structure with state level estimates for $m_{.ig}$ from the current year A & P survey, and current accurate county-level data on total farmland, $m_{h..}$ (also from A&P survey). The initial step in the IPF procedure sets the starting values equal to the known past values, i.e.,

$$x_{hig}^{(0)} = N_{hig} .$$

These cell proportions are then adjusted to the first set of marginal constraints, specified by the allocation structure, $\sum_{h} x_{hig} = m_{.ig}$, then to the second set of marginal constraints, $\sum_{i,g} x_{hig} = m_{h..}$ in a cyclical iterative manner. An iteration cycle consists of two steps. At the k^{th} iteration we have

$$1x_{hig}^{(k)} = \frac{x_{hig}^{(k-1)}}{x_{.ig}^{(k-1)}} \; m_{.ig}$$

and

$$x_{hig}^{(k)} = \frac{1x_{hig}^{(k)}}{1x_{h..}^{(k)}} \; m_{h..} \quad ,$$

where $1x_{hig}^{(k)}$ are the estimates resulting from adjusting to the ig marginal constraints at the k^{th} iteration. The resulting estimates,

$x_{hig}^{(k)}$, are then used as inputs into the next cycle. This iteration process is continued until the convergence criterion is satisfied following an iteration cycle.

The convergence criterion was as follows. At the end of a cycle, the following were checked:

$$\left| \sum_{i,g} {}_1x_{hig}^{(k)} - m_{h..} \right| = \left| {}_1x_{h..}^{(k)} - m_{h..} \right| \leq .5$$

for all h; and

$$\left| \sum_{h} x_{hig}^{(k)} - m_{.ig} \right| = \left| x_{.ig}^{(k)} - m_{.ig} \right| \leq .5$$

for all ig combinations. If <u>either</u> of these sets of inequalities was satisfied, the IPF procedure was stopped and the estimates for the individual cells were the current $x_{hig}^{(k)}$ values. Therefore, the county-level estimates were:

$$x_{hi} = \sum_{g} x_{hig}^{(k)} .$$

Appendix B

This appendix contains the % ARD and correlation analyses results for the Case 1 and Case 2 estimates when the allocation structure was not "adjusted". Recall in order for the Case 3 (IPF) procedure to converge, it was necessary for the allocation structure data that $\Sigma m_{h} {}_{h..} = \Sigma m_{i,g} {}_{.ig}$. Since this equality was not satisfied with the original data, the data in the allocation structure were weighted by multiplying the $m_{.ig}$ figures by C/D where $C = \Sigma m_{h} {}_{h..}$ and $D = \Sigma m_{i,g} {}_{.ig}$. This adjustment forced the "adjusted" $m_{.ig}$ figures to sum to C. The actual weights used were .98122 for 1981, 1.0111 for 1982, and .97415 for 1983. These weights are all close to one and therefore the data actually only had a small adjustment. However, for purposes of comparison, the % ARD and correlation analyses were performed using estimates derived for Cases 1 and 2 with the original (unadjusted) allocation structure. The results are given in Tables 5 through 8. When Tables 5 through 8 are compared with Tables 1 through 4 it is seen the figures are nearly the same. Therefore, all relationships and conclusions based on Tables 1 through 4 will apply to the estimates evaluated in Tables 5 through 8.

Table 5. Means of the Percentage Absolute Relative Differences of the Three Different Estimates

Estimator	Year	Crop						
		Corn-Grain	Soybeans	Tobacco	Peanuts	Sorghums	Oats	Barley
Case 1	1981	39.02	33.59	31.61	20.11	61.28	52.75	41.91
	1982	46.93	42.27	40.82	23.29	66.47	56.40	46.73
	1983	43.23	43.39	0.00	24.31	65.32	55.44	44.93
Case 2	1981	261.39	360.80	1233.44	1054.80	187.47	95.09	244.51
	1982	200.18	316.97	865.23	966.69	126.63	110.35	376.98
	1983	172.19	173.67	0.00	1100.68	138.01	101.01	299.85

Allocation structure was not "adjusted" in deriving these results.

Table 6. Medians of the Percentage Absolute Relative Differences of the Three Different Estimates *

Estimator	Year	Crops						
		Corn-Grain	Soybeans	Tobacco	Peanuts	Sorghums	Oats	Barley
Case 1	1981	36.46	29.22	23.84	0.00	70.66	63.05	27.80
	1982	45.21	44.75	36.77	0.00	85.39	64.21	33.57
	1983	42.58	45.38	0.00	0.00	83.13	61.82	34.12
Case 2	1981	59.18	51.71	48.42	0.00	66.35	46.85	50.89
	1982	58.67	54.54	55.58	0.00	68.68	56.26	56.75
	1983	59.94	50.46	0.00	0.00	61.78	57.96	58.11

Allocation structure was not "adjusted" in deriving these results.

Table 7. Standard Deviations of the Percentage Absolute Relative Differences of the Three Different Estimates

Estimator	Year	Crop						
		Corn-Grain	Soybeans	Tobacco	Peanuts	Sorghums	Oats	Barley
Case 1	1981	29.91	28.32	30.31	35.78	42.79	36.93	42.90
	1982	28.59	28.87	28.63	36.98	39.13	36.61	52.77
	1983	27.43	28.88	0.00	0.00	33.13	31.82	34.12
Case 2	1981	769.69	2367.61	7900.70	4095.17	407.48	365.89	664.00
	1982	673.04	2212.17	5302.30	3332.12	263.92	364.89	1645.06
	1983	504.51	876.57	0.00	4160.18	292.99	281.33	1056.92

Allocation structure was not "adjusted" in deriving these estimates.

Table 8. Pearson Correlation Coefficients Between the Three Different Estimates and Their Respective SRS Values

Estimator	Year	Crop						
		Corn-Grain	Soybeans	Tobacco	Peanuts	Sorghums	Oats	Barley
Case 1	1981	.82	.98	.97	.99	.89	.69	.95
	1982	.83	.95	.98	1.00	.86	.74	.92
	1983	.82	.96	--	.99	.85	.84	.98
Case 2	1981	.77	.72	.76	.58	.21	.30	.16
	1982	.75	.72	.76	.54	.22	.33	.09
	1983	.75	.74	--	.55	.25	.47	.14

Allocation structure was not "adjusted" in deriving these results.