TWO-PHASE ESTIMATION BY IMPUTATION

Anita McVey, F. Jay Breidt, Wayne A. Fuller, Iowa State University
Wayne A. Fuller, 219 Snedecor Hall, Iowa State University, Ames, Iowa 50011

Key Words: Resource inventory, survey estimation, small area estimation

Abstract

In the National Resources Inventory (NRI) conducted by the Soil Conservation Service in cooperation with Iowa State University, data are collected at two levels. The primary sampling unit (PSU) is an area segment of land, often 160 acres in size. The secondary sampling unit is a point. Some data, such as urban and built-up area, are collected on the PSU. Detailed data are collected at the points. In the 1992 NRI, the PSU data were used to impute point data for land uses occurring in the PSU but not observed at a point in that PSU. The imputation procedure is described and small area estimates constructed with the imputed data are compared with two-phase estimates using PSU data as the first phase estimates. Analysis of data collected in Missouri indicates that the imputation procedure produces far fewer small area estimates of no change in urban acres than the standard two-phase estimation procedure. Tests of equivalence for the two procedures indicate that the imputation procedure is generally unbiased

I. Introduction

The Iowa State Statistical Laboratory cooperates with the U.S. Soil Conservation Service on a large survey of land use in the United States. The survey was conducted in 1958, 1967, 1975, 1977, 1982, 1987 and 1992. The survey collects data on soil characteristics, land use and land cover, potential for converting land not used for crops to cropland, soil and water erosion, and conservation practices. The data are collected by employees of the Soil Conservation Service. Iowa State University has responsibility for the sample design and for estimation.

The sample is a two-stage stratified sample of the nonfederal area of the 50 states and Puerto Rico. The first-stage sampling units are areas of land called segments or primary sampling units (PSU's). The PSU's vary in size from 40 acres to 640 acres. Data are collected for the entire PSU on items such as urban land and small water area. Detailed data on soil properties and land use are collected at a random sample of points within the PSU. A point within the PSU is the second-stage sampling unit. Generally, there are three points per PSU, but 40-acre PSU's contain two points and the samples in two states contain one point per PSU. Some data, such as total land area and area in roads, are collected on a census basis external to the sample survey.

In 1982, the sample contained about 350,000 PSU's and nearly one million points. The 1987 sample was composed of about 100,000 PSU's. The majority of the 1987 sample PSU's were a subsample of the 1982 PSU's. However, about 1,500 new PSU's were selected in areas of rapid urban growth. Data were collected on about 280,000 points in 1987. The sample for 1992 is the 1987 sample plus the majority of the 1982 sample not observed in 1987. About 290,000 PSU's and 800,000 points were observed in 1992. The design of the sample is a simple form of a panel survey in that the 1987 sample was nearly a subsample of the 1982 sample and the 1992 sample is nearly the 1982 sample.

The sample was designed to produce reasonable estimates for geographical units called Major Land Resource Areas. These areas are defined on the basis of soil and cover characteristics. There are about 180 Major Land Resource Areas in the study area. Also, the acreage estimates for any county were to be consistent with the total acreage of that county. There are about 3,100 counties in the sample. Because the sample must provide consistent acreage estimates for both counties and Major Land Resource Areas, the basic tabulation unit is the portion of a Major Land Resource Area within a county. There are 5,530 of these units, which we call MLRAC's.

In 1992, it was decided that longitudinal data analysis would be performed using data for the three years 1982, 1987, and 1992. Thus, the final data set will contain data for those three years for about 290,000 PSU's. The 1987 data on cover for PSU's not observed in 1987 was collected in 1992, primarily from aerial photography.

II. Point Generation

Data at the PSU level on the acres of land in farmsteads, small water bodies, built-up, and large urban are collected for each PSU within the state for 1982, 1987, and 1992. More detailed data are collected on the points. We shall concentrate on

estimation for points that are classified as urban for at least one of the three years.

Consider the estimation of a characteristic such as the cover on an urban point, where cover includes information on percent of a circular area around the point that is grass, percent that is trees, etc. If a regression two-phase estimator is used, the information in the PSU data is combined with the information on the points that are available in a subsample of the PSU's through a regression equation. One method of implementing a two-phase estimator is to create regression weights for each of the points such that the sum of the weights for the points applied to the PSU characteristic (acres) is equal to the estimate constructed with the PSU data.

Not all PSU's that contain urban land have points falling on urban land. Thus, there will be few urban points in some small MLRAC's. This is particularly important for estimation of changes in land use. For example, in Arizona about 20 MLRAC's contained PSU's that showed an increase in urban land from 1987 to 1992 while only 13 MLRAC's contained a point that changed from nonurban to urban between 1987 and 1992. To reduce variability in small area estimates, we developed an alternative estimation scheme rather than using the ordinary two-phase estimator. In the estimation data set created under our procedure, pseudo points are created for each PSU that contains urban land.

Changes in the PSU data over the three collection years are the key to determining what kind of points must be created in the PSU to ensure that all of the PSU data are represented in point form. The acres for a given land use in a PSU can remain constant, increase or decrease in each of the two intervals 1982-1987 and 1987-1992. For example, the acres for large urban might increase from 1982 to 1987 and then decrease from 1987 to 1992. Once the type of change has been identified for a land use in each of the intervals, the number and kind of points can be determined for that PSU. For example, if the acres in large urban increases from 1982 to 1987 for a given PSU, a point with a non-urban land use in 1982 and a large urban land use in 1987 will appear in the tabulation data set. Such a point might have been sampled. If not, a pseudo point will be created for that PSU. If a sampled point in the PSU exists and has the required land uses, it is assigned a weight equal to the acres associated with the change in PSU acres for that land use divided by the probability of selection. If more than one sampled point within the PSU meets the land use requirements of the change, the weight is divided equally between the sampled points. If no sampled point meets the land use requirements, a pseudo point is generated. A pseudo point represents a real change in a land use that is not observed at the point level. The pseudo point is assigned a weight equal to the change in PSU acres for that land use divided by the probability of selection.

The data for the pseudo points are imputed using data from real points. For example, if a PSU shows an increase in urban land from 1982 to 1987, a point that is not urban in 1982 and is urban in 1987 and 1992 is required. Two sources of data are used in the imputation. The first source is used to impute the land coveruse in the years for which coveruse is unknown. Coveruse is an exhaustive classification based on the use of the land and the characteristics of the land. Some of the coveruse categories are cropland, pastureland, rangeland, forestland, urban, and small built-up. If a PSU showed an increase in urban from 1982 to 1987, and if no point showed such a change, a point is created that has the urban coveruse in 1987 and a different coveruse in 1982. The coveruse and associated characteristics for 1982 are imputed by selecting one of the points in the PSU at random and assigning the 1982 characteristics of the selected point to the created pseudo point. The second source is an urban point selected from a PSU "near" the PSU under consideration. The selected urban point provides the data on urban characteristics for 1987 and 1992.

The sampled and pseudo points contain the information in the PSU data and form the tabulation data set used in estimation for the NRI data.

III. Comparison of Alternative Estimators

In this section, we compare estimates constructed for Missouri using a standard two-phase estimation scheme with those constructed using our point generation procedure. The two-phase estimator is the estimator using the first phase (PSU data) to define strata. Five types of points are identified in terms of the coveruse in each of the three years 1982, 1987, and 1992. The types are UUU, NUU, NNU, NNS, NNN, where U denotes urban, S denotes small built-up, and N denotes all remaining coveruses. The estimates based upon PSU data (first phase estimates) were constructed for six geographic subdivisions of Missouri. Then the weights on the real points in each of the types within subdivisions were ratio adjusted to give the PSU estimated acres. Thus, for a subdivision, the imputed data set and the set composed of real points will give the same estimate of urban acres for each of the three years 1982, 1987, and 1992. It is in the

small area estimates (e.g. county or MLRAC totals) where we expect the point generation procedure to give estimates more representative of the PSU data than the standard two-phase estimation scheme.

Figures 3.1 and 3.2 contain the distribution of estimated change in urban acres from 1982 to 1992 for the 115 counties of Missouri. Because some PSU's contain an increase in urban acres but no point showing this change, the use of real points in two-phase estimation produces a much higher fraction of counties with an estimated change of zero. Figures 3.3 and 3.4 show the distribution of estimated change in small built-up for the 115 counties. As with large urban, the two-phase estimator has a much higher fraction of zero estimates. In both the large urban and the small built-up figures, the total acres of change in urban land for the state is the same for the two estimation procedures. Since the two-phase distributions (Figures 3.2 and 3.4) are wider than the imputation distributions (Figures 3.1 and 3.3), some counties are inappropriately being assigned some large estimates of change when the two-phase procedure is used.

Tables 3.1 and 3.2 show the estimated total acres associated with change in land coveruse from 1982 to 1987 for the point generation procedure and

Figure 3.1

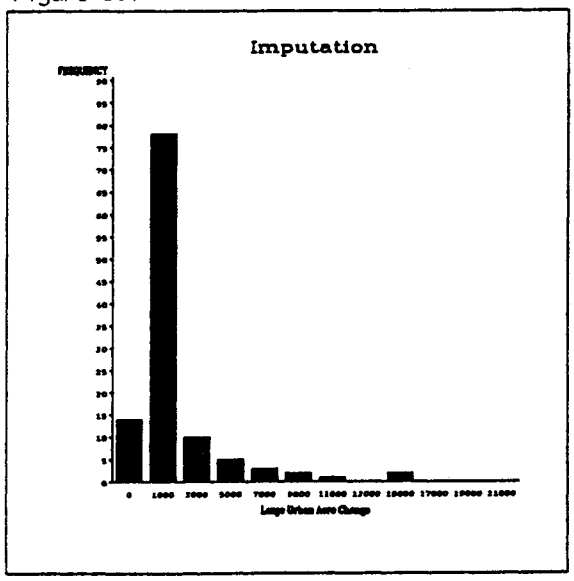

Figure 3.3

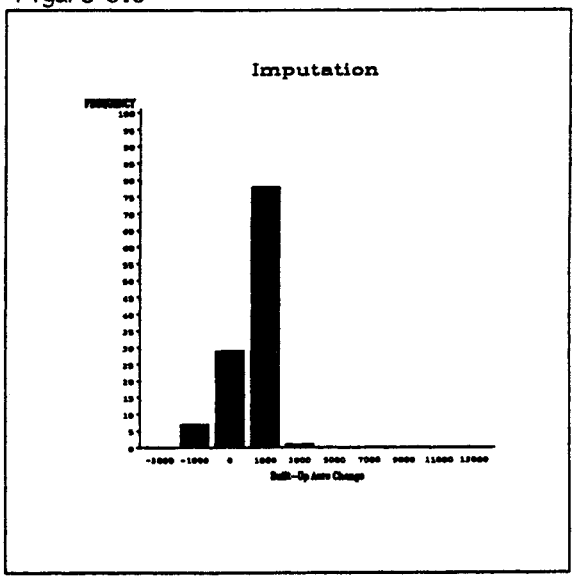

Figure 3.2

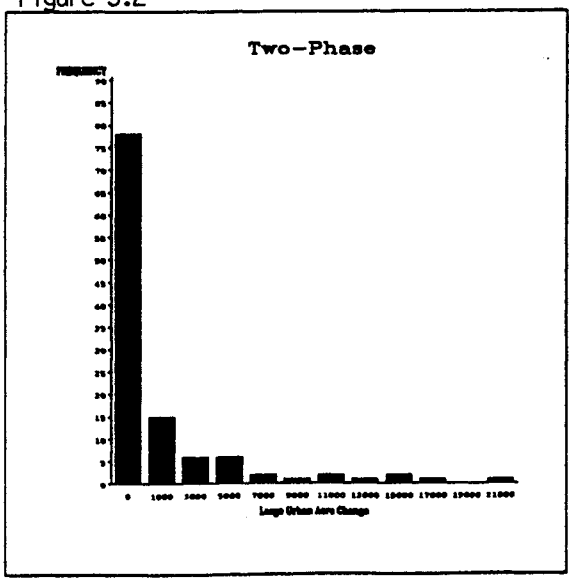

Figure 3.4

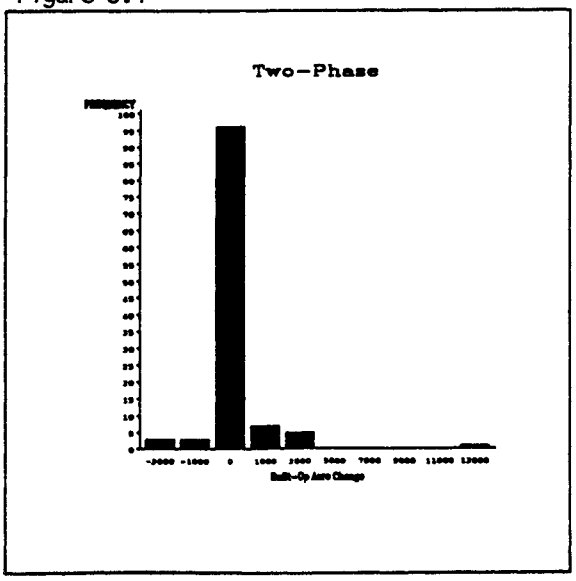

Table 3.1. Estimates constructed by Point Generation Procedure (in 1000's of acres)

1982 Land Use	1987 Land Use			
	Cropland	Large Urban	Other Land	Total
Cropland	13527	24	1448	14999
Large Urban	0	1117	0	1117
Other Land	858	67	27565	28490
Total	14385	1208	29013	44606

Table 3.2 Standard Two-Phase Estimates (in 1000's of acres)

1982 Land Use	1987 Land Use			
	Cropland	Large Urban	Other Land	Total
Cropland	13523	19	1442	14984
Large Urban	0	1117	0	1117
Other Land	859	72	27574	28505
Total	14382	1208	29016	44606

the standard two-phase estimation scheme, respectively. Similarly, Tables 3.3 and 3.4 show the changes in land uses from 1982 to 1992. The table contains three combined coveruse groups. The coveruse groups are: cropland, large urban and other land (including forestland, rangeland, federal land, and small built-up). The method of estimation used means that the 1992 marginals, the 1982 large urban to 1987 large urban diagonal, the 1982 large urban to 1992 large urban diagonal, and the 1982 cropland to 1992 cropland diagonal are the same for the two estimation procedures. Of interest is the land that is converted into urban, especially the land that is converted to urban from cropland. These tables show that the state estimates of land use changes constructed by the two estimation schemes are very similar.

Table 3.5 contains *t*-values for tests of equivalence of the two procedures for several characteristics. The test statistics were constructed on the difference of the two estimated ratios of the acres of a characteristic of interest to total acres. The estimate of the difference of the two ratios is

$$\frac{\sum\sum w_{ij}y_{ij}}{\sum\sum w_{ij}} - \frac{\sum\sum w_{ij}x_{ij}}{\sum\sum w_{ij}r_i\delta_{ij}},$$

where the sum is over the sample, i is geographic subdivision, j is the observation within subdivision, w_{ij} is the original sample weight,

y_{ij} = the sample observations (including real points and pseudo points)

x_{ij} = $r_i y_{ij}$ if the point is a real point

= 0 otherwise,

r_i is the ratio of the sum of weights in the total sample to the sum of the weights for the real points, and $\delta_{ij} = 1$ if the point is a real point and is zero otherwise. Thus,

$$r_i = \left(\sum_j w_{ij}\delta_{ij}\right)^{-1} \sum_j w_{ij} .$$

Table 3.3. Estimates Constructed by Point Generation Procedure (in 1000's of acres)

1982 Land Use	1992 Land Use			
	Cropland	Large Urban	Other Land	Total
Cropland	12151	49	2799	14999
Large Urban	0	1117	0	1117
Other Land	1196	142	27152	28490
Total	13347	1308	29951	44606

Table 3.4. Standard Two-Phase Estimates (in 1000's of acres)

1982 Land Use	1992 Land Use			
	Cropland	Large Urban	Other Land	Total
Cropland	12151	40	2793	14984
Large Urban	0	1117	0	1117
Other Land	1196	151	27158	28505
Total	13347	1308	29951	44606

Table 3.5. Difference of Ratios Tests of Equivalence

Land Use					
1982	1987	1992	Difference	S.E.	T
Prime	Large Urban		0.00001649	0.0001162	0.142
Prime		Large Urban	0.00001181	0.0002023	0.058
Cropland	Large Urban		0.00012085	0.0001333	0.907
Cropland		Large Urban	0.00020744	0.0002215	0.937
Built-Up	Large Urban		-0.00003279	0.0001062	-0.309
Built-Up		Large Urban	-0.00008859	0.0001251	-0.708
Cropland		Built-Up	0.00013858	0.0000488	2.839*

*Significant at 95% confidence level

Table 3.6. Difference of Ratios Tests of Equivalence

Cover	Difference	S.E.	T
Grass and crops	1.52848	1.08980	1.403
Trees and shrubs	-0.44861	0.79998	-0.561
Barren and artificial	-1.19577	1.52797	-0.783
Water	0.11590	0.08905	1.302

The estimated variance of the difference of the ratios was computed recognizing the stratified cluster nature of the design.The calculated *t*-values suggest that it is reasonable to conclude that the two procedures are estimating the same quantity, with the exception of the estimated shift of cropland to small built-up. The imputation procedure gives a larger estimated shift of cropland to small built-up. The imputation procedure randomly selects one of the points in the PSU to provide the previous coveruse for land shifting into small built-up. If certain land, such as forest land, has a higher probability of being shifted into small built-up, then the imputation procedure is biased.

Table 3.6 contains a comparison of earth cover estimated by the two procedures. Earth cover is an estimate of the fraction of the area covered by different types of cover (grass, shrubs, trees, hard surface such as concrete) when viewed from above. The source of the imputed data for this item is a real point of the required coveruse near the pseudo point.

There are no significant differences in this table suggesting that there is little bias in the procedure used to select donor points.

ACKNOWLEDGEMENTS

This research was partly supported by Cooperative Agreement 68-3A75-4-86 with the Soil Conservation Service.

REFERENCES

Cochran, W. G. (1977). *Sampling Techniques.* New York: John Wiley and Sons.

Fuller, W. A. (1990). Analysis of Repeated Surveys. *Survey Methodology*, 16, 167-180.