

Two-phase Estimation by Imputation

F. Jay Breidt, Anita McVey and Wayne A. Fuller
Iowa State University, Ames, U.S.A.

SUMMARY

In the National Resources Inventory conducted by the U.S. Natural Resources Conservation Service in cooperation with Iowa State University, data are collected at two levels. The primary sampling unit 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 segment. Detailed data are collected at the points. In the 1992 inventory, the segment data were used to impute point data for land uses occurring in the segment but not observed at a point in that segment. The goal was to create a tabulation data set of sampled and imputed points, which would contain the information in the segment data. The imputation procedure is described and small area estimates constructed with the imputed data are compared with two-phase estimates using segment 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.

Key words : Resource inventory, Survey estimation, Small area estimation.

1. Introduction

The Iowa State Statistical Laboratory cooperates with the U.S. Natural Resources Conservation Service (formerly 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 Natural Resources Conservation Service; Iowa State University has responsibility for the sample design and for estimation. The survey is currently called the National Resources Inventory (NRI).

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 sample of points selected within the PSU according to a design which ensures good spatial dispersion. 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 being 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, defined on the basis of soil and land 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 contains 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.

Data at the PSU level on the acres of farmsteads, small water, small built-up, and large urban are collected for each PSU within the state for 1982, 1987, and 1992. Briefly, a farmstead is an area used by farm operators for buildings, storage, livestock, etc.; small water is a small water body less than 40 acres in size or a stream less than one eighth mile wide; small built-up is an area of nonfarm dwellings and (or) other urban activity less than 10 acres, and large urban is an urbanized area larger than 10 acres. More detailed data are collected on the points. We shall concentrate on estimation for points that are classified as urban (built-up or large 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. A random subsample of the NRI points falls on urban land, and so both point-level and PSU-level data are available for a subsample of the PSU's. This is an example of a two-phase sample. Estimation for two-phase samples is discussed in texts such as Cochran [1], Särndal, Swensson and Wretman [2], and Wolter [3].

Different methods can be used to combine the information from the first phase sample and the subsample to estimate the cover. If a regression two-phase estimator is used, the information in the PSU data is combined with the information on the points that is available in a subsample of the PSU's through a regression equation. One method of implementing a two-phase regression estimator is to create regression weights for each of the points such that the sum of the weights for the points applied to the characteristic (acres) is equal to the estimate constructed with the PSU data.

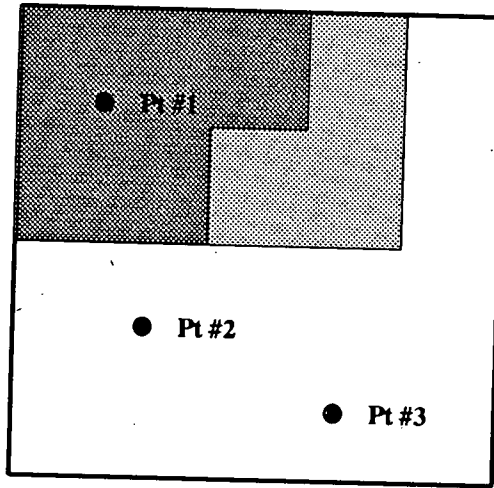
Not all PSU's that contain urban land have points falling on urban land. Thus, there will be few urban points in some small MLRAC's. This is particularly important for estimation of changes in land use. For example, in Arizona about 20 MLRAC's contained PSU's that showed an increase in urban land from 1987 to 1992 while only 13 MLRAC's contained a point that changed from nonurban to urban between 1987 and 1992. To reduce the variability in the ordinary two-phase small area estimator of small area characteristics, an alternative estimation scheme was developed. In the estimation data set created under our procedure, pseudo points are created for each PSU that contains urban land. Section 2 describes the pseudo point generation procedure. In Section 3, the pseudo point estimators are compared to two-phase regression estimators for small area characteristics in Missouri. Section 4 concludes with discussion of a possible variance estimation procedure.


2. Pseudo Point Generation


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 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 increase 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

Imputation Example 160 Acre PSU



 40 acres of Large Urban in 1982

 20 more acres of Large Urban in 1987 and 1992

Point	LU : Large Urban		CROP : Cropland		Acres
	1982	1987	1992		
1	LU	LU	LU		40
2	CROP	CROP	CROP		50
3	CROP	CROP	CROP		50
4*	CROP	LU	LU		20

* pseudo point

Figure 1. Schematic diagram illustrating pseudo point imputation procedure for a PSU in the 1992 NRI. The PSU data indicate an increase in urban acres from 1982 to 1987 that is not reflected in the point data. A pseudo point changing from nonurban in 1982 to urban in 1987 is generated to carry the PSU change information. See Section 2 for details.

the land use requirements of the change, the weight is divided equally among 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 sampled and pseudo points contain the information in the PSU data and form the tabulation data set used in estimation for the NRI.

Figure 1 illustrates the pseudo point imputation procedure for a hypothetical 160-acre PSU on which measurements have been obtained in 1982, 1987, and 1992. The PSU data indicate that there were forty acres of large urban in 1982 and sixty acres of large urban in 1987 and 1992. One point falls on an area that is large urban in all three years, 1982, 1987, and 1992, but no sampled point reflects the shift from nonurban to urban between 1982 and 1987. In order to represent what happened in this PSU, the imputation procedure creates one additional point, a pseudo point, which is large urban in 1987 and 1992 and is not large urban in 1982. The pseudo point is assigned a weight equal to 20 acres divided by the probability of selection. The data on acreage is provided by the PSU data but the remaining data elements, such as earth cover, must be obtained from other sources.

Data for the pseudo points are imputed using data from real points. 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. To obtain the 1982 coveruse and associated characteristics for the pseudo point in the example of Figure 1, the imputation procedure selects at random one of the real points in the PSU that was not large urban in 1982, and assigns the 1982 characteristics of the selected point to the created pseudo point. In this example, suitable donors are point 2 and point 3.

The second source of data used in the imputation 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. In the example of Figure 1, the 1987 and 1992 data might come from the urban point within the PSU. In many cases, however, it would be necessary to obtain the real urban point from outside the PSU. The urban point is selected via a kind of hierarchical hot-deck imputation procedure, in which the donor point is drawn at random from a set of "nearby" potential donors. The imputation procedure begins with the smallest set of potential donors, consisting of points within the same MLRAC. This class is searched for a suitable donor: first, with a limit set on the number of times a donor can be used; again, if necessary,

with the limit increased; and again, if necessary, with no limit. If no suitable donor is found, the imputation class is enlarged from points within the same MLRAC to points within the same county, and searched in the manner described above. The procedure is repeated, if necessary, with the imputation class enlarged to the entire state.

3. 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 coverage 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, N denotes coveruses other than U in 1982 and 1987, and N denotes coveruses other than U or S in 1992. The estimates based upon PSU data (first phase estimates) were constructed for six geographic subdivisions of Missouri. The pseudo point generation procedure described in Section 2 was used to create a weighted point data set. Then the weights on the real point 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

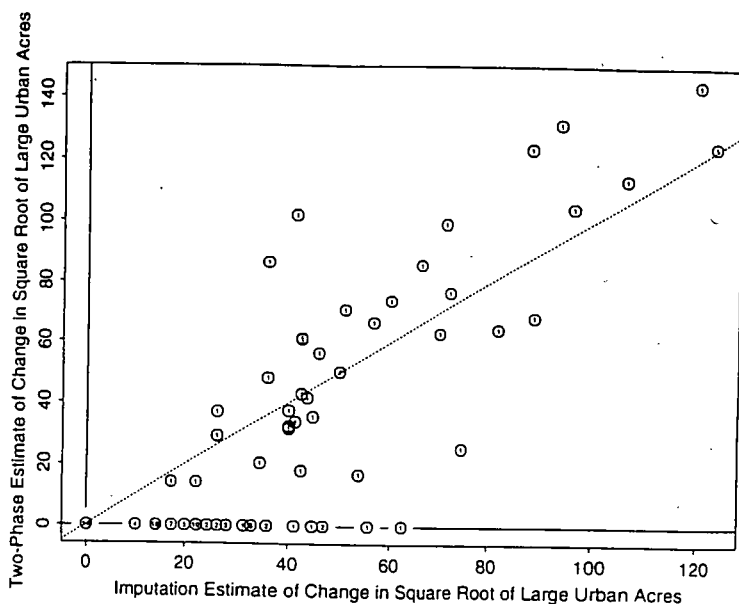


Figure 2. Two-phase estimates of change in large urban acres versus imputation estimates for the 115 counties in Missouri. Square root transforms have been applied to both axes. The circled number is the number of counties represented at the point. The dotted reference line has unit slope and zero intercept.

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 one expects the point generation procedure to give estimates more representative of the PSU data than the standard two-phase estimation scheme.

Figure 2 compares the two-phase and imputation estimates of change in large urban acres from 1982 to 1992 for the 115 counties of Missouri. (Because of the highly skewed nature of these estimates, square root transforms were applied to both axes.) Seventy-eight of the two-phase estimates of change in large urban acres are zero, while only fourteen of the imputation estimates are zero. Because some PSU's contain an increase in large 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. Figure 3 shows 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 (96/115 compared to 29/115).

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 ranges of the two phase estimates (vertical axes in Figures 2 and 3) are larger than the ranges of the imputation estimates (horizontal axes in Figures 2 and 3), some counties are inappropriately being assigned large estimates of change when the two-phase procedure is used. In particular, note the outliers in the upper right and lower left corners of Figure 3.

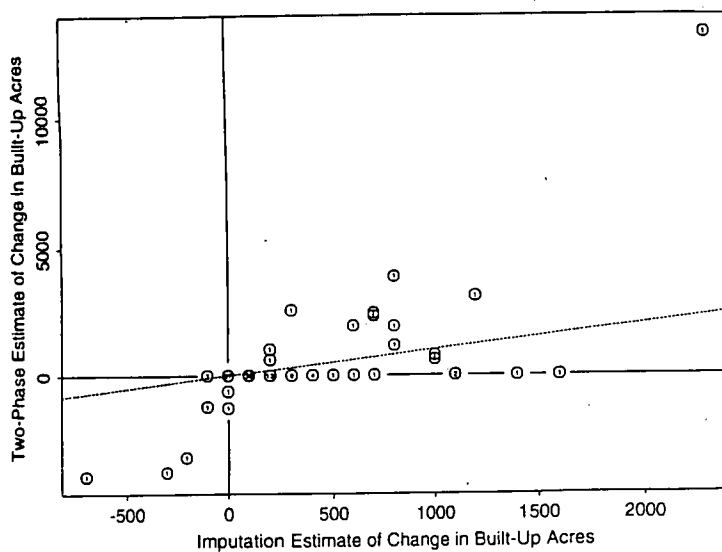


Figure 3. Two-phase estimates of change in small built-up acres versus imputation estimates for the 115 counties in Missouri. The circled number is the number of counties represented at the point. The dotted reference line has unit slope and zero intercept.

The two-phase estimation scheme is a consistent estimation procedure for any characteristic. The imputation estimator is consistent under the assumption that donors are selected such that the expected value of donors is equal to the expected value of the recipients. Because donors were selected on the basis of geographical location, the imputation scheme is not guaranteed to be unbiased for the point data items.

Table 1. Estimates constructed by pseudo point imputation 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 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	14383	1208	29016	44606

Tables 1 and 2 show the estimated total acres associated with change in land coveruse from 1982 to 1987 for the pseudo point imputation procedure and the standard two-phase estimation scheme, respectively. Similarly, Tables 3 and 4 show the changes in land uses from 1982 to 1992. The tables contain three combined coveruse groups. The coveruse groups are : cropland, large urban and other land (including forestland, rangeland, federal land, and small built-up). The 1992 marginals, the 1982 large urban to 1987 large urban diagonal, the 1982 large urban to 1992 large urban diagonal, and the 1982

Table 3. Estimates constructed by pseudo point imputation 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 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

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 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

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

Table 5. Difference of ratios tests of equivalence

Land Use			Difference	S.E.	T
1982	1987	1992			
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*

* Significantly different from zero at level 0.05.

where the sum is over the sample, i is geographic subdivision, j is the observation within subdivision, w_{ij} is the weight after the pseudo point imputation procedure,

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

δ_{ij} = 1 if the point is a real point and is zero otherwise,

and r_i is the ratio of the sum of weights in the total sample to the sum of the weights for the real points.

Thus,

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

A relatively simple estimator of the variance is possible in this case because we are interested in the difference and the variance can be calculated based on the first phase units. 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. It would be possible to develop an alternative imputation scheme in which the probability that a point is selected as a donor point is based upon a probability estimated from the real points.

Table 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

Table 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

significant differences in this table suggesting that there is little bias in the procedure used to select donor points.

4. Discussion

The pseudo point generation procedure described here, which produces points which carry PSU-level urban acreage information, is just one step in an extremely complex estimation strategy for the NRI. Other steps include further pseudo point generation (for other types of PSU data and for data collected in a county census), missing data imputation, and ranking ratio weight adjustments. Variance estimation in this context is difficult. An approximate variance estimation procedure which uses a replication idea is outlined. For simplicity, we focus only on pseudo point generation and weight adjustments rather than the entire estimation procedure. In the estimation procedure, states are processed independently. For variance estimation, the sampling strata are split or collapsed as necessary to create two PSUs per stratum. A half-sample and its complement are formed by randomly selecting one PSU per stratum. For each half-sample, the pseudo point generation procedure is completed using only the data for that half-sample. The estimation weights for each half-sample

are created such that the weights for each half-sample sum to one half of the external control totals.

Assume that the pseudo point generation procedure selects donors only from within the county (this assumption will at times be violated). The set of all points (including pseudos) within a half-sample and a county can then be considered a cluster of correlated observations that is independent of the other half-sample. The sample can then be treated as a sample of two clusters per stratum, where counties are the strata. Classical sampling theory can be used to produce approximate standard errors of estimated quantities. Theoretical and empirical properties of these standard errors are the subject of ongoing research.

ACKNOWLEDGEMENTS

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

REFERENCES

- [1] Cochran, W.G., 1977. *Sampling Techniques*, 3rd ed. John Wiley and Sons, New York.
- [2] Särndal, C.E., Swensson, B. and Wretman, J., 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- [3] Wolter, K.M., 1985. *Introduction to Variance Estimation*. Springer-Verlag, New York.