Estimation Procedures for the National Inventory of Soil and Water Conservation Needs

August, 1957 J. E. Dowd and J. C. Cassady * BU-88-M

In order to outline in detail the estimation procedures to be used in the Conservation Needs Inventory we will distinguish between the three following situations in which we might find ourselves with respect to available data.

1. No previous mapping and measurement has been carried out in the county and county estimates are to be derived from the sample data alone.

2. Some previous mapping and measurement has been carried out in the county and conforms to present soils mapping standards so that it can be used in conjunction with the sample data to form estimates of the county totals. In this case we can further distinguish as to the manner in which the previous mapping has been carried out:

   (a) Mapping and measurement has been completed for large contiguous blocks of land as a portion of a complete mapping program for that county.

   (b) Mapping and measurement have been carried out on individual and perhaps relatively isolated farms as a part of farm planning activities of the SCS or other activities.

   (c) In many counties a combination of mapping described in (a) and (b) will occur.

3. The county has been completely mapped and measured, the soils mapping being deemed satisfactory for the purposes of the Conservation Needs

---

*Biometrics Unit, Cornell University

study. It is desired to measure changes in land use which have occurred up to present and to relate changes in land use to various soils grouping, e.g. land capability units.

In none of the cases as outlined in section 2 can the land areas upon which the mapping and measurement has been carried out be considered as having been selected in a random fashion from the county as a whole. In case 2(a) whole sections of a county are liable to be as yet completely unmapped while other sections are completely mapped and measured. In case 2(b) there is probably a general tendency to map soils on that land which is, or may become, of agricultural importance. Farmers who request that their farms be mapped likely form a sample with land and conservation practices superior to that found in the county as a whole. Thus in either of these cases or in the combination mentioned in 2(c) there are liable to be unknown biases present in the data which would preclude their use as estimates taken alone.

We will deal first with the simpler case mentioned in (1). It will be recalled that the sample scheme (described in BU-86-M) is a stratified random sample of land areas, each county (or in many cases, land resource area) independently stratified. Though two sampling units were drawn from each stratum, only one will be mapped and measured at present.

If X represents a particular soil separation being measured, an estimate $\hat{X}$ of the county total of X is given as follows:

$$\hat{X} = \sum_{i=1}^{k} N_i X_i$$

where X_i is the amount of soil separation X found in the selected sampling unit in the i^{th} stratum, N_i is the number of 100 acre units in the i^{th} stratum, and k is the number of strata into which the county is divided.

Since stratification is a somewhat inexact process of division especially at county or L.R.A. boundaries and where some bodies of water and certain ownership classes are to be excluded from within some of the strata, the total for all strata $N = 100 \sum_{i=1}^{k} N_i$ will not equal the total land area of the county as given in the 1954 Census of Agriculture. To adjust for this discrepancy we will multiply the estimate $\hat{X}$ by the ratio N_0/N where N_0 is the total land area of the county given by the 1954 Census of Agriculture after appropriate land and water areas have been excluded.

Thus, we define a new estimate of X which expands to the approximate total as:

$$X' = \frac{N_0}{N} \sum_{i=1}^{k} N_i X_i$$

If it is desired to compute an estimate of the variance of $\hat{X}$ from the sample this can be done by grouping strata (usually contiguous pairs) which are similar with respect to the soil separation in question. The estimated variance is then calculated as a weighted sum of squared differences between the X values of the paired strata.

Thus the estimated variance of $\hat{X}$ is given by:

$$s_{\hat{X}}^2 = \sum_{g=1}^{G} \frac{N_g(N_g - n_g)}{n_g} s_{X_g}^2 \quad ,$$

where

$$s_{X_g}^2 = \frac{1}{L_g - 1} \sum_{h=1}^{L_g} \left(X_{gh} - \frac{N_{gh}}{N_g} X_g\right)^2 ,$$

G = number of groups of strata in the county,

N_{gh} = size of h^{th} stratum in the g^{th} group,

N_g = size of the g^{th} group = $\sum_{h=1}^{L_g} N_{gh}$,

L_g = number of strata in the g^{th} group,

n_g = number of sampling units in the g^{th} group,

X_{gh} = sample total for X from the h^{th} stratum, in the g^{th} group,

X_g = sample total for X in the g^{th} group = $\sum_{g=1}^{L_g} X_{gh}$.

The estimate of the variance of X' is then simply given as:

$$s_{X'}^2 = \frac{N_0^2}{\bar{y}^2} s_{\hat{X}}^2 .$$

Of the three situations that present themselves in case 2 the first situation outlined 2(a) is probably the easiest to handle in using a combination of previous data in conjunction with the sample estimates. Where soils mapping has been carried out in large contiguous blocks it is usually possible to approximate the total area of a single block by a grouping of strata. Thus it is possible to regard the results from the block mapping as representing complete measurement of certain groups of strata. Suppose of the k strata in the county we may regard k_1 of these, by virtue of their approximate coincidence with the previously mapped blocks, as being completely measured. The (uncorrected) estimate of X from a sample of n_i units from the i^{th} strata (i=1,•••,k) is given by $\hat{X} = \sum_{i=1}^{k} \frac{N_i}{n_i} X_i$. For the k_1 strata completely mapped and measured we have $n_i = N_i$ and for the remaining $k - k_1$ strata we have $n_i = 1$. Thus the estimate reduces to:

$$\hat{X} = \sum_{i=1}^{k_1} X_i + \sum_{i=k_1+1}^{k} N_i X_i$$

where $\sum_{i=1}^{k_1} X_i$ is just the total from those large blocks of land completely mapped and measured and $\sum_{i=k_1+1}^{k} N_i X_i$ is the sample total for the remaining $k-k_1$ strata.

The variance of this estimate then reduces to:

$$s_X^2 = \sum_{g=1}^{G'} N_g \frac{(N_g - n_g)}{n_g} s_{X_g}^2$$

where all terms are as previously defined except that G' = the number of groups of strata in the $k-k_1$ remaining strata.

There is no variance associated with the term $\sum_{i=1}^{k_1} X_i$ since measurement has been carried out on all sampling units within each of these k_1 strata.

Where the mapping is not complete in blocks, but is scattered throughout the county as described in 2(b) above, the question arises as to the advisability of using information tabulated from this sort of survey, even though it represents a large portion of the county. Further examination of this situation may yield some useful information.

The mean square error from expansion of such tabulated information to estimates of county totals would be $E(\hat{X}_B - X)^2$ where $\hat{X}_B$ is the biased estimate and X is the population characteristic that is being estimated. If we assume the model to be

$$\hat{X}_B = X + \epsilon_B + B \quad ,$$

then

$$E(\hat{X}_B - X)^2 = E(\epsilon_B + B)^2$$

$$M.S.E(\hat{X}_B) = \sigma_B^2 + B^2$$

where ϵ_B is the random error associated with the estimate X_B from the sample of size n_B.

B is the bias associated with the estimate $\hat{X}_B$ from the non-random sample of size n_B.

If we assume the model for the random estimate to be

$$\hat{X}_R = X + \epsilon_R ,$$

then $$E(\hat{X}_R - X)^2 = E(\epsilon_R)^2 = \sigma_R^2 .$$

where ϵ_R is the random error associated with the estimate $\hat{X}_R$ from a sample of size n_R.

If we were to use some linear combination of totals estimated from the 2% random sample and from the large non-random survey for our final estimate, so that $\hat{X}_F = \lambda_R \hat{X}_R + \lambda_B \hat{X}_B$, how could we choose λ's so as to minimize the effect of both bias and variance?

$$\hat{X}_F - X = \lambda_R(\hat{X}_R - X) + \lambda_B(\hat{X}_B - X) \text{ , if } \lambda_R + \lambda_B = 1 .$$

$$E[(\hat{X}_F - X)^2] = E[\lambda_R^2(\hat{X}_R - X)^2 + \lambda_B^2(\hat{X}_B - X)^2 + 2\lambda_R\lambda_B(\hat{X}_R - X)(\hat{X}_B - X)] .$$

$$M.S.E.(\hat{X}_F) = \lambda_R^2\sigma_R^2 + \lambda_B^2(\sigma_B^2 + B^2) + 2\lambda_R\lambda_B E(\hat{X}_R - X)(\hat{X}_B - X) .$$

Since $\lambda_B = 1 - \lambda_R$,

$$M.S.E.(\hat{X}_F) = \lambda_R^2\sigma_R^2 + (1-\lambda_R)^2(\sigma_B^2 + B^2) + 2\lambda_R(1-\lambda_R)Cov.(\hat{X}_R, \hat{X}_B) .$$

Differentiating with respect to λ_R,

$$\frac{dMSE}{d\lambda_R} = 2\lambda_R[\sigma_R^2+\sigma_B^2-2\ \mathrm{Cov.}(\hat{X}_R,\hat{X}_B)]+2[-\sigma_B^2+\mathrm{Cov.}(\hat{X}_R,\hat{X}_B)-B^2].$$

Setting this equal to zero to minimize MSE, and solving,

$$\lambda_R = \frac{\sigma_B^2+B^2-\mathrm{Cov.}(\hat{X}_R,\hat{X}_B)}{\sigma_R^2+\sigma_B^2+B^2-2\ \mathrm{Cov.}(\hat{X}_R,\hat{X}_B)} .$$

$$\lambda_B = \frac{\sigma_R^2-\mathrm{Cov.}(\hat{X}_R,\hat{X}_B)}{\sigma_R^2+\sigma_B^2+B^2-2\ \mathrm{Cov.}(\hat{X}_R,\hat{X}_B)} .$$

Thus, when $\sigma_B^2 = \sigma_R^2$, and $B^2 = 0$, $\lambda_R = \lambda_B$; and the larger the B^2 term, the larger λ_R. Also, if $B^2 = 0$, but $\sigma_B^2 \neq \sigma_R^2$, λ_R decreases as σ_R^2 increases. If $\sigma_R^2 > \sigma_B^2 + B^2$, $\lambda_B > \lambda_R$.

In order to examine this relationship further, we must make some assumptions about σ_R^2 and σ_B^2. If N is the number of units in the population, n_R the number in the random sample and n_B the number in the biased sample, then let us assume

that $$\sigma_R^2 = \frac{N-n_R}{N}\ \frac{S^2}{n_R}$$

and $$\sigma_B^2 = \frac{N-n_B}{N}\ \frac{S^2}{n_B} .$$

Then $$\sigma_R^2/\sigma_B^2 = \frac{N-n_R}{n_R} \cdot \frac{n_B}{N-n_B} .$$

Consequantly, the ratio of the variances depends upon the ratio of the sizes of the samples. The following table shows values computed for λ_R.

The size of the non-random sample is varied from 10% to 80% and the bias, from 0 to $4\sigma_R^2$.

This table illustrates the dominance of the bias in affecting the values of λ_R. However, one other item is worthy of attention: even though we assume a bias of $4\sigma_R^2$ in the non-random sample, apparently we can still use that information to improve our estimates, by weighting the random estimate by approximately 4/5 and the non-random by approximately 1/5.

Values of λ_R

Size of Samples	Size of Bias (B^2)					
	0	$\frac{1}{2}\sigma_R^2$	σ_R^2	$2\sigma_R^2$	$3\sigma_R^2$	$4\sigma_R^2$
2% - 10%	.16	.41	.54	.69	.76	.79
2% - 20%	.08	.37	.52	.68	.75	.78
2% - 30%	.05	.35	.51	.67	.75	.78
2% - 40%	.03	.34	.51	.67	.75	.78
2% - 50%	.02	.34	.50	.67	.75	.78
2% - 60%	.01	.34	.50	.67	.75	.78
2% - 70%	.01	.34	.50	.67	.75	.78
2% - 80%	.005	.34	.50	.67	.75	.78

The unanswered question, of course, is how to determine or attempt to estimate B^2. This will vary for each characteristic to be estimated. Attempts are being made to verify empirically some assumptions which relate the size and direction of the bias to certain groupings of soil separations (land capability units). These results will be reported at a later date.

In those counties for which a previously completed survey exists and for which the soils mapping completed is usable, as outlined in 3 above,

land use will be checked on the 2% sample to determine whether there are any trends in land use changes. Three estimates will be available for land use: the estimate for the population at time A, the estimate from the 2% sample at time A, and the estimate from the 2% sample at time B (present time). It is required to determine whether the sample estimate determined at time B differs from the population estimate at time A.

Let y_i be the acreage of land use F (say) on unit i at time B.

Let x_i be the acreage of land use F (say) on unit i at time A.

Let X be the acreage of land use F in the population at time A.

With simple stratified sampling the ratio estimate for Y would be

$$\hat{Y}_R = (\frac{y}{x})X$$

where y and x are the sample totals of y_i and x_i, i.e. $y = \sum_i N_i y_i$, $x = \sum_i N_i x_i$. A combined estimate for the variance of $\hat{Y}_R$ is

$$s^2_{\hat{Y}_R} = (\sum_i \frac{N_i^2}{n_i} S_{y_i}^2) + \hat{R}^2(\sum_i \frac{N_i^2}{n_i} S_{x_i}^2) - 2\hat{R}\sum_i(\frac{N_i^2}{n_i}\hat{\rho}_i S_{y_i} S_{x_i})$$

where $S_{y_i}^2$ is an estimate of the variance of y in the i^{th} stratum

$$= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$S_{x_i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$\hat{\rho}_i = \frac{\sum_{i=1}^{n_i} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)}{\sqrt{\Sigma(x_{ij} - \bar{x}_i)^2 \Sigma(y_{ij} - \bar{y}_i)^2}}$$

$$\hat{R} = \frac{y}{x}$$

Where there is some correlation between previous land use and present land use, gains in the precision with which land use is estimated can be expected.

An approximate test of significance for changes in land use is given by:

$$t = \frac{\hat{Y}_R - X}{s_{\hat{Y}_R}}$$

where t is assumed to have Student's t distribution with $\sum_{i=1}^{K} (n_i - 1) = n - K$. For (n-K) large, say greater than 50, normal tables may be used.