A COMPARISON OF PERIODIC SURVEY DESIGNS EMPLOYING MULTI-STAGE SAMPLING

V. M. Lesser and W. D. Kalsbeek, Dept. of Biostatistics, University of North Carolina
V.M. Lesser, CB 7400, Dept. of Biostatistics, UNC-CH, Chapel Hill, NC 27599-7400

Key Words: **Panel survey, multi-stage sampling, survey costs.**

ABSTRACT

Three types of panel survey designs are evaluated within the framework of two-stage sampling. Comparisons of these designs will be discussed with regard to precision, cost, and other issues that need to be considered in planning long-term surveys. To compare precision, the underlying variance of a simple estimator of mean difference is derived for each of the three designs. Research is continuing on the cost model development, which will be combined with the variance models to evaluate cost-efficiency. These results will contribute to determining the best design to monitor agroecosystem health within EPA's Environmental Monitoring Assessment Program.

1. INTRODUCTION

With increasing concern on the status of environmental condition, EPA established a program to monitor ecological status and trends to establish baseline environmental conditions against which future changes can be documented (Overton, et.al, 1990). The program intends to assess the status of a number of different ecological resources, including surface waters, wetlands, near coastal wetlands, forests, arid lands, and agroecosystems. Presently, an overall design strategy is under consideration to sample all resources.

The agroecosystem component of EMAP has initiated the work discussed in this research. Each EMAP component has the task of reviewing currently operational monitoring programs with similar objectives as the EMAP program. For example, the United States Department of Agriculture (USDA) and the National Agriculture Statistics Service (NASS) conduct annual national scale surveys to obtain current statistics on the Nation's agriculture. The objective of this research is to compare the sample design employed by USDA/NASS to the sample design proposed by EMAP.

Three design options will be examined in this research, which are illustrated in Figure 1. In this Figure, N represents the sample observed in the j-th year of the program illustrated in the superscript as N^j. The panel or replicate number, which will be explained in subsequent paragraphs, is represented in the subscript, as the i-th panel. Therefore, the sample observed in the j-th year for the i-th panel is denoted as N_i^j. Each design option provides an area frame covering the total area of the US.

The NASS design will be referred to as a mixed-longitudinal or rotating panel design as is illustrated in Figure 1 (A). Each year one of these panels is introduced into the survey and measured annually for five years. At the same time, a panel, which has been measured for five years, is rotated out of the survey. Approximately 20 percent of the panels are replaced annually. This results in a 4/5 overlap of panel elements from year to year.

The EMAP design will be referred to as a longitudinal design with interpenetrating replicates and is illustrated in Figure 1 (B). This design consists of 4 interpenetrating replicates, which are measured within a repeating cycle length of 4 years. Once the sampling segments have been selected, they will remain in the survey for the duration of the study. No new sampling segments are introduced in this design.

The ENASS design is similar to the NASS design. The difference in the designs is due to the number of repeated measurements on the same observational unit. Sampling segments are measured only at the first and last year of the NASS rotating cycle. Figure 1 (C) illustrates this sampling strategy.

The primary difference between these design options is the number of measurements collected over time on the sampling units. This results in a comparison of a longitudinal design with two types of a mixed-longitudinal design. The statistical efficiency of these designs has been discussed in the literature, but not within the context of multistage sampling (Berger, 1986; Duncan and Kalton, 1987). Researchers, such as Kish (1988), have discussed the necessity of consideration of the survey design in applying statistical methods to the data.

This paper compares the statistical efficiency of these mixed-longitudinal and longitudinal designs within the context of multi-stage sampling. A measure of precision and cost models are derived for each of the design options. This information will contribute to determine and recommend to EPA the best design, in terms of precision and cost, to evaluate the condition of our Nation's agriculture.

2. VARIANCE MODEL DEVELOPMENT

A. Estimator of Interest and General Assumptions.

The estimator of interest in this research of the population value, obtained from the sample, using j and i to denote two different years, is defined as:

$$ {}_{ij}\theta = {}_{j}\bar{\bar{y}} - {}_{i}\bar{\bar{y}} = \sum_{z}^{n} {}_{j}y_z/n - \sum_{z}^{n} {}_{i}y_z/n \qquad 2.1 $$

In order to compare the precision of the three design options discussed in this research, the underlying sampling variance of this estimator is derived. In this research, each of the design options assumes the following.

1. Two-stage design.
2. Equal size clusters.
3. Sampling is unrestricted random sampling at the first stage.
4. Sampling is simple random sampling at the second stage.
5. No stratification at either stage.
6. The same sample size is assumed for each year for each design option.
7. Each replicate (panel) is considered an independent sample.
8. Epsem design.

B. Derivation of the Underlying Variances.

In the following, each component of the two-stage variance will be derived separately for each design and then combined. Capital letters refer to population values, while small letters refer to sample values. A defines clusters, while B defines elements within clusters.

1. Derivation of $E_\alpha[\mathrm{Var}_2({}_{ij}\theta|\alpha)]$.

a. To obtain the variance at the element level, the covariance term needs to account for the number of overlapping samples. In this derivation q refers to the proportion of overlapping panels.

$$ \mathrm{Var}_2({}_{ij}\theta|\alpha) = \mathrm{Var}_2({}_{j}\bar{\bar{y}} - {}_{i}\bar{\bar{y}}|\alpha) $$

$$ = \mathrm{Var}_2({}_{j}\bar{\bar{y}}|\alpha) + \mathrm{Var}_2({}_{i}\bar{\bar{y}}|\alpha) $$

$$ - 2\,\mathrm{Cov}_2\,({}_{j}\bar{\bar{y}},\, {}_{i}\bar{\bar{y}}|\alpha\,) $$

$$ = \sum_{\alpha}^{a}\frac{(1-f_b)}{a^2b}{}_{j}S^2_{\alpha b} + \sum_{\alpha}^{a}\frac{(1-f_b)}{a^2b}{}_{i}S^2_{\alpha b} $$

$$ - 2\sum_{\alpha}^{qa}\frac{(1-f_b)}{a^2b}\,{}_{ij}S_{\alpha b} \qquad 2.2 $$

where, the variance at year i or j (denote as w) is defined as:

$$ {}_{w}S^2_{\alpha b} = \sum_{\beta}^{B}\frac{({}_{w}Y_{\alpha\beta} - {}_{w}\bar{Y}_{\alpha})^2}{(B\text{-}1)} \qquad 2.3 $$

and the covariance between year i and j is defined as:

$$ {}_{ij}S_{\alpha b} = \sum_{\beta}^{B}\frac{({}_{i}Y_{\alpha\beta} - {}_{i}\bar{Y}_{\alpha})({}_{j}Y_{\alpha\beta} - {}_{j}\bar{Y}_{\alpha})}{(B\text{-}1)} \qquad 2.4 $$

b. Since the second stage fractions are assumed uniform for all first stage units, taking expectation over all possible PSU samples of size a results in:

$$E_\alpha[Var_2\ ({}_{ij}\theta|\alpha)] = \frac{(1-f_b)}{ab}\ {}_jS_b^2 + \frac{(1-f_b)}{ab}\ {}_iS_b^2 - 2\ q\ \frac{(1-f_b)}{ab}\ {}_{ij}S_b \qquad 2.5$$

where, the variance at year i or j (denote as w) is defined as:

$$ {}_wS_b^2 = \sum_\alpha^A \sum_\beta^B \frac{({}_wY_{\alpha\beta} - {}_w\bar{Y}_\alpha)^2}{A(B-1)} \qquad 2.6$$

and the covariance between year i and j is defined as:

$$ {}_{ij}S_b = \sum_\alpha^A \sum_\beta^B \frac{({}_iY_{\alpha\beta} - {}_i\bar{Y}_\alpha)({}_jY_{\alpha\beta} - {}_j\bar{Y}_\alpha)}{A(B-1)} \qquad 2.7$$

The result for the covariance term in Equation 2.5 is an extension of the argument given by M. Hansen, W. Hurwitz, and W. Madow in Sample Survey Methods and Theory - Volume II (1953, Equation 8.9), which accounts for the overlapping samples. The proportion of overlap varies for each of the designs:

A. EMAP: q=1 for comparisons of multiples of 4 years apart, otherwise q=0.
B. ENASS: q=1/2 for a comparison of 4 years, otherwise q=0.
C. NASS: q=(5-m)/5, where m=j-i, for comparisons $\leq$ 4 years, otherwise q=0.

2. Derivation of $Var_\alpha[E_2({}_{ij}\theta|\alpha)]$.

a. The expectation at the element level is derived as follows.

$$E_2\ ({}_{ij}\theta|\alpha) = E_2\ ({}_j\bar{\bar{y}} - {}_i\bar{\bar{y}}\ |\ \alpha\ ) = E_2\ ({}_j\bar{\bar{y}}|\ \alpha) - E_2({}_i\bar{\bar{y}}\ |\ \alpha\ ) = \sum_\alpha^a \frac{{}_j\bar{Y}_\alpha}{a} - \sum_\alpha^a \frac{{}_i\bar{Y}_\alpha}{a} \qquad 2.8$$

The same sample size is assumed for each year for each design option. To account for the summation over the panel numbers in the NASS and ENASS design, and to maintain the same sample size across designs, each panel consists of the same number of clusters. This results in the expectation given in Equation 2.8 for all designs.

b. In order to obtain the variance of the above formula, the variance of each of the above terms and the covariance of these terms over all possible PSU samples is obtained. Similar to the overlap found at the element level, the covariance between these two terms incorporates a term to account for the amount of overlapping samples. The variance over all possible PSU samples results in:

$$Var_\alpha[E_2\ ({}_{ij}\theta|\alpha)] = \frac{{}_j\sigma_a^2}{a} + \frac{{}_i\sigma_a^2}{a} - 2\ q\ \frac{{}_{ij}\sigma_a}{a} \qquad 2.9$$

where, the variance at year i or j (denote as w) is defined as:

$$ {}_w\sigma_a^2 = \sum_\alpha^A \frac{({}_w\bar{Y}_\alpha - {}_w\bar{\bar{Y}})^2}{A} \qquad 2.10$$

and the covariance between year i and j is defined as:

$$ {}_{ij}\sigma_a = \sum_\alpha^A \frac{({}_i\bar{Y}_\alpha - {}_i\bar{\bar{Y}})\ ({}_j\bar{Y}_\alpha - {}_j\bar{\bar{Y}})}{A} \qquad 2.11$$

3. Combining these terms results in:

$$Var\ ({}_{ij}\theta) = \frac{(1-f_b)}{ab}\left\{ {}_jS_b^2 + {}_iS_b^2 - 2\ q\ {}_{ij}S_b \right\} + \frac{1}{a}\left\{ {}_j\sigma_a^2 + {}_i\sigma_a^2 - 2\ q\ {}_{ij}\sigma_a \right\} \qquad 2.12$$

The variance components and the value of q for each design are defined above.

C. Simplification of Underlying Variances

In order to simplify Equation 2.12 for further comparisons among design options, a series of assumptions are made in this section redefining the underlying variances. These include:

(1) N is assumed large relative to n, resulting in a small sampling fraction f=n/N. Therefore, the finite population factors in the second stage disappear. Also, for each year, the expectation of s_b^2 is assumed equal to σ_b^2, and, the expectation of s_b is assumed equal to σ_b.

(2) The covariance terms in Equation 2.12 can be written in terms of the temporal correlation. This assumption uses:

$$ {}_{ij}\sigma_b = {}_{i}\sigma_b \; {}_{j}\sigma_b \; {}_{ij}\rho_b $$

and

$$ {}_{ij}\sigma_a = {}_{i}\sigma_a \; {}_{j}\sigma_a \; {}_{ij}\rho_a $$

(3) The between and overall components of variance are written in terms of the overall variance and δ, where δ is a measure of homogeneity between second stage units within the first stage units (Kish, 1965). This assumption uses:

$$ \sigma^2 = \sigma_a^2 + \sigma_b^2 $$

$$ \delta = \frac{\sigma_a^2 - \sigma_b^2/(b-1)}{\sigma^2} $$

To obtain the following expressions which are incorporated into the variance formula:

$$ \sigma_a^2 = \frac{\sigma^2}{b}\Big\{1 + \delta(b-1)\Big\} $$

$$ \sigma_b^2 = \frac{b-1}{b}\Big\{\sigma^2(1-\delta)\Big\} $$

(4) Finally, these equations are simplified further when we assume ${}_{j}\sigma^2 = {}_{i}\sigma^2$ and n=ab.

$$ \text{Var}\,({}_{ij}\theta) = \qquad 2.13 $$

$$ \Big\{\frac{2\sigma^2\,(1 - q\;\rho^m)}{n}\Big\}\Big\{\frac{(b-1)(1-\delta)}{b} + [1+\delta(b-1)]\Big\} $$

The variance components and the value of q for each design have been defined above. Notice that the first component in each of these equations is a simple random variance of the difference of two means. The second component can be considered a design effect. For the special case when b=1, this component equals 1.

3. COST MODEL DEVELOPMENT AND COST EFFICIENCY

A cost model has been obtained for each design option accounting for fixed and variable costs. The costs are partitioned among frame, interview, printing, training, and other operational costs. Presently, optimum allocation computations for a two-stage sample design is underway. Using this allocation and for a special case when the number of secondary sampling units is one, cost efficiency will be evaluated among designs. Efficiency is denoted as the inverse of the square root of the variance model and cost is denoted as the cost model defined for each design. Cost efficiency is defined as efficiency per unit cost, or the ratio of efficiency to cost.

A few measures to compare cost efficiency for the designs are presently under consideration. These include the ratio of cost efficiency of one design to another, the relative ratio of cost efficiencies, and the difference of the cost efficiencies. It is expected that sample size, the degree of panel overlap, the degree of temporal correlation, and the degree of intracluster homogeneity will affect these results. Four years of data collected by USDA/NASS have been obtained to verify some assumptions made in the variance derivation and to obtain measures of the

variance components.

4. FINAL ASSESSMENT

Other potential sources of survey error will be evaluated relative to the three designs, such as coverage, nonresponse, and measurement error to determine the impact of these errors on each of the designs. This evaluation, along with the cost efficiency results, will be used to contribute to determining the best design strategy to monitor agriculture health for EPA's EMAP program.

5. REFERENCES

Berger, M.P.F. (1986) "A Comparison of Efficiencies of Longitudinal, Mixed Longitudinal, and Cross Sectional Designs", Journal of Educational Statistics, 11, 171-181.

Cochran, W.G. (1977) Sampling Techniques, 3rd Edition, New York: John Wiley and Sons.

Duncan, G.J., Kalton, G. (1987) "Issues of Design and Analysis of Surveys Across Time", International Statistical Review, 55, 97-117.

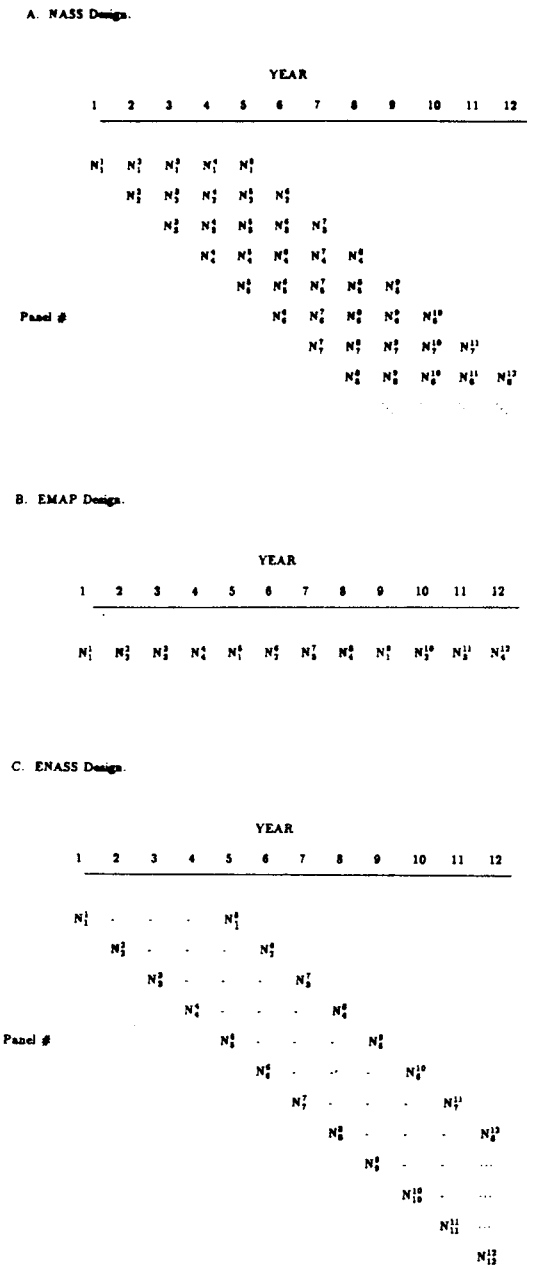
Hansen, M.H., Hurwitz, W.N., Madow, W.G. (1953) Sample Survey Methods and Theory Volume II, John Wiley and Sons, New York.

Kish, L. (1965) Survey Sampling, John Wiley and Sons, New York.

Kish, L. (1988) "Multipurpose Sample Designs", Survey Methodology, 14, 19 -32.

Overton, W.S., Stevens, D.L., Pereira, C.B., White, D., Olsen, T. (1990) "Design Report for EMAP, Environmental Monitoring and Assessment Program", March, 1990 Draft.

Figure 1. Three design options under consideration for the agroecosystem component of EMAP.[a]



[a] N represents the observed sample, the superscript represents year and the subscript represents panel number, denoting N_i^j.