A Multi-Phase Sample Design to Co-ordinate Surveys and Limit Response Burden

Phillip S. Kott, USDA-NASS and Matt Fetter, USDA-NASS
Matt Fetter, National Agricultural Statistics Service, 4818 South Bldg., Washington, D.C. 20250

Key Words: Nearly Unbiased, Subpopulation, Systematic PPS Sampling

Motivation

The Agricultural Resource Management Study (ARMS) was developed to provide mutually exclusive samples for multiple surveys . These surveys were used to collect cropping practices, cost of production (through three separate surveys), farm operator resources, and cattle health. Because all of these surveys were considered to be highly burdensome to the respondent, it was necessary to limit to one the number of surveys that any farm operation would receive. Samples were drawn individually by state, and the number of applicable surveys varied by state. Data were collected in three phases- a screening phase, a cropping practice and cost of production phase, and a economic phase. This paper will outline the derivation of the estimation weights for Nebraska in which there were five applicable survey modules.

Developing Estimation Weights for the ARMS Survey

Nebraska in 1996 has two cost of production (COP) commodities, corn for grain - hereafter called "grain corn" - and beef, and three cropping practices' "multi-crops" (MC's), corn, wheat, and soybeans. It also has animal health (NAHMS) cattle survey.

We will use the term *module* to refer to one of the following sets:

FOR (Farm Operator Resources) farm-level variables; NAHMS farm-level variables; grain corn COP farm-level variables; beef COP farm-level variables; combined FOR and COP farm-level variables; grain corn COP field-level variables (includes some pro-rated farm-level variables); beef COP enterprise-level variables; combined beef COP and NAHMS farm/enterprise-level variables; corn MC field-level variables; soybean MC field-level variables; wheat MC field-level variables; and combined corn COP and corn MC field-level variables.

The sample that applies to each module will have its own set of weights. There are two samples for the corn COP field-level module - a fall sample and a spring sample. The latter, which includes only fields on farms completing the spring economic questionnaire, is a subset of the former. The samples may not coincide due to nonresponse. Analogously, there will be fall and spring samples for the Beef COP enterprise-level module.

The Screening Sample

A stratified simple random sample of farms has been taken within each state for screening. Let i denote a farm in the Nebraska screening sample, and $n_{SCR(i)}/N_{SCR(i)}$ denote the sampling fraction of the screening stratum containing i. This unusual notation simplifies the exposition, which remains complicated enough. *For now, we will assume that each sampling unit - hereafter called a "farm" - is a single operation.* We relax this assumption in a later section.

The screening weight for farm i is

$$w_{SCR(i)} = \frac{\sum^{N_{SCR(i)}} t_g}{\sum^{u_{SCR(i)}} t_g}, \qquad (1)$$

where t_g denotes the control *total farm value of sales* for farm g, and $u_{SCR(i)}$ are the screener usables in the same screening stratum as i. All summations in this note are indexed by the farm variable g (i denotes the farm of interest; g a farm like i). The set over which the sum is taken is implied by the limit of summation on top of the "$\sum$."

Equation (1) assumes that we do not know whether a non-respondent is in business or what its Data Adjustment Factor (DAF) is. More complex and accurate weights can - and will - be derived when addition information is known about screener non-respondents. Moreover, an addition weight adjustment will occur in modules with a NOL component.

The FOR Sample

The usables from the screening sampled are subsampled for the FOR sample. Let $\pi_{FOR(i)} = n_{FOR(i)}/N_{FOR(i)}$ denote the sampling fraction of the FOR stratum containing i. In Nebraska, the FOR strata are virtually the same as the screening strata. The major difference is that farms with DAF=0 will be removed from the FOR strata.

The FOR weight for farm i is

$$w_{FOR(i)} = \frac{\sum_{N_{FOR(i)}} t_g w_{SCR(g)}}{\sum_{u_{FOR(i)}} t_g w_{SCR(g)} / \pi_{FOR(g)}} w_{SCR(i)} / \pi_{FOR(i)} \qquad (2)$$

$$= \frac{\sum_{N_{FOR(i)}} t_g w_{SCR(g)}}{\sum_{u_{FOR(i)}} t_g w_{SCR(g)}} w_{SCR(i)} .$$

The form is equation (2) will prove useful when we allow sampled farms to be multiple operations.

Other Farm Samples

Samples of farms for the NAHMS and the two COP surveys - beef then grain corn - are drawn sequentially using systematic "probability proportional to size (pps)" sampling. The term "pps" is in quotes because the measures of size described below have little to do with size.

To draw the NAHMS sample, all farms from the screener who are judged to be in-scope for NAHMS (i.e., are in business with DAF≠0 and have cattle) and have *not* been selected for the FOR survey are given the following measure of size:

$$m_{NAHMS(i)} = \min\{2,\ 1/(1-\pi_{FOR(i)})\},$$

where the bounding value 2 in the above equation may be increased upon further study and may be different in other states than Nebraska. We then sort available farms by their screener cattle inventories and draw a systematic pps sample with unit i receiving conditional (on it being available for sampling) selection probability:

$$\pi_{NAHMS(i)} = n_{NAHMS} m_{NAHMS(i)} / \sum^{N_K} m_{NAHMS(g)} .$$

This can be written more generally as

$$\pi_{K(i)} = n_K m_{K(i)} / \sum^{N_K} mK(g), \qquad (3)$$

where the subscript K equals NAHMS.

To draw the beef COP sample, all farms not yet selected and in-scope are given the measure of size:

$$m_{BFCOP(i)} = m_{NAHMS(i)} / (1-\pi_{NHAMS(i)}) ,$$

which leads to a conditional selection probability expressed by equation (3) with K equal to BFCOP.

Similarly, to draw a grain corn COP sample, all farms not yet selected and in-scope are given the measure of size:

$$m_{GCCOP(i)} = \min\{2, 1/(1-\pi_{FOR(i)})\}/[(1-\pi_{BFCOP(i)})(1-\pi_{NAHMS(i)})].$$

This measure of size leads to a conditional selection probability expressed by equation (3) with K equal to GCCOP. Note that $\pi_{BFCOP(i)} = 0$ if farm i was not eligible for the beef COP.

To draw the three multi-crop farm samples, we can first assign each in-scope farm the measure of size:

$$m_{K(i)} = \min\{2,1/(1-\pi_{FOR(i)})\}/[1-\pi_{GCCOP(i)})(1-\pi_{BFCOP(i)})(1-\pi_{NAHMS(i)})],$$

where K denotes either corn, soy beans, or wheat.

Each available farm is assigned a permanent random number prn(i) from 0 to 1. For the crop K sample, we calculate a(i) = prn(i) / $m_{K(i)}$ for each available farm having crop K and then select the n_K farms with the smallest a(i) values.

The conditional selection probability of farm i for the crop K sample has approximately the same form as equation (3) when all $\pi_{K(i)} < .7$. When that is not the case, conditional selection probabilities can be estimated via simulation (i.e., draw 100 multi-crop samples and compute the fraction of times farm i is selected for the crop K sample). A practical alternative is to simply use $m_{K(i)} = 1$ as the measure of size in the first place. We plan to do that in most states.

If an operation is selected for all three MC samples, we randomly delete it from one, given each crop an equal probability of being the one excluded one. This causes $\pi_{K(i)}$ in equation (3) to be multiplied by 2/3.

Farm Weights for the NAHMS and COP Modules

Let us call

$$w_{NAMHS(i)}^{sa} = w_{SCR(i)}/[(1-\pi_{FOR(i)})\pi_{NAHMS(i)}] \qquad (*)$$

the *screener-adjusted sampling weight* for a NAHMS farm i. Similarly, we call

$$w_{BFCOP(i)}^{sa} = w_{SCR(i)}/[(1-\pi_{FOR(i)})(1-\pi_{NAHMS(i)})\pi_{BFCOP(i)}] , \quad (*)$$

and

$$w_{CNCOP(i)}^{sa} = w_{SCR(i)} / [(1-\pi_{FOR(i)})(1-\pi_{NAHMS(i)})(1-\pi_{BFCOP(i)})\pi_{GCCOP(i)}] \ , \quad (*)$$

the screener-adjusted sampling weight for farm i in the beef and grain corn COP, respectively.

Before proceeding we need to focus on a variable of interest from the screener survey, either cattle inventory or grain corn acres as appropriate. Let x_i^O be the screener variable value of interest for farm i times the farm's DAF. Unfortunately, this value is known to be positive but is not reported for some screener farms. When that happens for farm i, we impute a value using the following formula:

$$x_i = \frac{\sum_{u_{SCR(i)}^*} x_g^O}{\sum_{u_{SCR(i)}^*} t_g} t_i \ , \quad (4)$$

where $u_{SCR(i)}^*$ are the usables in the same screener stratum as i that are in business with DAF≠0 and have positive and known x-values. For farms where x_i^O is known, we define x_i to be x_i^O.

The fully-adjusted weight for farm i in module K (NAHMS, BFCOP, or GCCOP) can be expressed as

$$w_{K(i)} = \frac{\sum_{S_{K(i)}} x_g w_{SCR(g)}}{\sum_{u_{K(i)}} x_g w_{K(g)}^{sa}} w_{K(i)}^{sa} \ , \quad (5)$$

where $S_{K(i)}$ is the set of farms in the screener sample that are also in the same post-stratum as farm i for module K, and $u_{K(i)}$ is the set of usables in module K in the sample post-stratum as i.

Equation (5) applies no matter what rule we use to define the post-strata for module K. A major purpose of post-strata is to serve as model groups for non-response adjustment. One way to define post-strata for this purpose is to divide the sample for module K into three post-strata based on the relative size of the farms' x-values; that is, create one post-stratum consisting of farms with the highest x-values, one containing farms with the lowest x-values, and one with the rest of the in-scope screener farm sample.

Multi-Operation Farms

When a sampled farm for any module has more than one operation, we do the following: select one in-scope operation at random allowing each operation an equal probability of selection. We can treat this as an additional stage of sampling. Let $\pi_{MOK(i)}$ be the conditional probability of selecting operation i given that the multi-operation farm containing i has been selected for module K (i.e., $\pi_{MOK(i)}$ = 1 / [the number of in-scope operations in i]).

This value can be multiplied by $n_{FOR(i)} / N_{FOR(i)}$ to determine $\pi_{FOR(i)}$, which in turn can be plugged into equation (2) to determine the FOR farm weight for i.

Similarly, $\pi_{MOK(i)}$ can be multiplied by the right-hand-side of equation (3) to determine a new $\pi_{K(i)}$. This leads to a new set of screener-adjusted sampling weights, $w_{K(g)}^{sa}$, where only the $\pi_{K(g)}$ (not π-values from previous phases) is multiplied by $\pi_{MOK(g)}$. The final farm operation weights for module K are then calculated with equation (5).

Some care may be needed to impute a missing value for x_i using equation (4) when i is an operation on a multi-operation farm. It may require that the farm total value of sales for a sampling unit be divided among its component operations.

Field Weights for Grain Corn COP and MC Modules

Let ij denote field j on farm operation i. If $F_{K(i)}$ is the number of grain corn fields on a farm operation sampled for the grain corn COP, then $1/F_{K(i)}$ is the field's conditional selection probability. Here K denotes grain corn, but we leave it in general form for future use.

Let z_i be the grain corn acres reported for all (of the operation selected from) farm i on the grain corn COP and z_{ij} be the grain corn acres on field j. The weights for the grain corn COP field-level module have the form:

$$w_{K(ji)} = \frac{\sum_{S_{K(i)}} x_g w_{SCR(g)}}{\sum_{u_{K(i)}} x_g w_{K(g)}^{sa}} \frac{\sum_{u_{K(i)}} w_{K(g)}^{sa} z_g}{\sum_{u_{Kf(i)}} w_{K(g)}^{sa} z_{gj} F_{K(g)}} w_{K(i)}^{sa} F_{K(i)} \ , \quad (6)$$

where $u_{Kf(i)}$ is the usable field sample in the same post-stratum as the fields in operation i (this assumes that the same post-stratum definitions are used for the farm and field samples). Note that $u_{Kf(i)}$ - and the corresponding $u_{K(i)}$- in equation (6) may be different for the fall sample

of grain corn fields and the spring sample of such fields because the latter is restricted to fields in farm operations for which economic data is also available.

Equation (6) can not be used to determine weights for the three MC modules because we do not have the appropriate z_g.-values. One approach appropriate for corn and soybeans is to set $z_g = x_g$, and let equation (6) collapses to

$$w_{K(ji)} = \frac{\sum_{S_{K(i)}} x_g w_{SCR(g)}}{\sum_{u_{Kf(i)}} w_{K(g)}^{sa} z_{gj} F_{K(g)}} \; w_{K(i)}^{sa} F_{K(i)} . \qquad (6')$$

Alternatively, for wheat (because of planted acres versus harvested acres issues), we can set $x_g = z_g = 1$ and $z_{gj} = 1/F_{K(g)}$ and collapse (6) to

$$w_{K(ji)} = \frac{\sum_{S_{K(i)}} w_{SCR(g)}}{\sum_{u_{Kf(i)}} w_{K(g)}^{sa}} \; w_{K(i)}^{sa} F_{K(i)} . \qquad (6'')$$

We define $w_{K(i)}^{sa}$ (K = corn, wheat, or soy beans) in (6') and (6") as

$$w_{K(i)}^{sa} = \qquad (*)$$

$$\frac{w_{SCR(i)}}{[(1-\pi_{FOR(i)})(1-\pi_{NAHMS(i)})(1-\pi_{BFCOP(i)})(1-\pi_{GCCOP(i)})\pi_{K(i)}]} ,$$

where $\pi_{K(i)}$ is scaled by 2/3 and/or the inverse of the number of operations in farm i as appropriate.

Farm Operation Weights for the Economic (Joint FOR/COP) Module

Let us now consider the farm (operation) weight for the economic module that combines data from the FOR and spring COP. Those operations in the usable FOR sample *without* either grain corn or at least 10 weaned calves on the screener get their FOR weight from equation (1); that is $w_{ECON(i)} = w_{FOR(i)}$.

For those spring usable farm operations in one of the two COP samples, compute:

$$w_{COP(i)}^{sa} = \qquad (*)$$

$w_{BFCOP(i)}^{sa}$ if farm i is eligible for beef COP but not grain corn COP,

$w_{CNCOP(i)}^{sa}$ if farm i is eligible for grain corn COP but not beef COP,

$w_{BFCOP(i)}^{sa}/2$ if farm i is eligible for both COP's and chosen for Beef COP,

$w_{CNCOP(i)}^{sa}/2$ if farm i is eligible for both COP's but chosen for grain corn,

and then use equation (5) with K=COP, $x_g = t_g$. We restrict $S^{COP(i)}$ to farm operations that are eligible for one of the two COPS's.

Let u_{CFOR} be the number of usable FOR operations that have either grain corn or at least 10 weaned calves on the screener. The economic weight for such operations has the form:

$$w_{ECON(i)} = \lambda w_{FOR(i)} , \qquad (7)$$

while the economic weight for a operation from one of the COP samples has the form:

$$w_{ECON(i)} = (1 - \lambda) w_{COP(i)} . \qquad (8)$$

The value of λ has yet to be determined. One obvious choice for λ is $u_{CFOR}/(u_{CFOR} + u_{COP})$. Other possibilities that do not require a count of usables may be more practical. The techique of setting a single λ for all operations in both the COP and FOR domains is called *composite weighting*.

Other Composite Weights

For the combined NAHMS and Beef COP module, all NAHMS farm operations with less than 10 weaned calves on the screener get a combined cattle weight equal to their NAHMS weight. Other operations in the NAHMS sample receive a weight equal to λ times their NAHMS weight, while operations in the beef COP receive a weight of $(1-\lambda)$ times their beef COP weight.

For the cropping practices module combining the field-level corn COP and MC samples, the weight for each usable non-grain field equals its MC weight:. Other fields in the MC sample recieve a weight equal to λ times their MC weight, while fields in the grain corn COP

receive a weight of $(1-\lambda)$ times their grain corn COP weight.

Other Issues

Surprises - If a farm operation selected for a COP or MC survey reports no quantity of the commodity during the fall survey in contrast to information provided on the screener, NASS will determine whether the operation is better treated as a nonrespondent or a valid zero for the module. *An operation with valid zeroes is treated as a usable* in the weight equations. Moreover, all fall sampled COP farm operations remain eligible for the spring economic module.

Vegetable Farms - Suppose a farm operation is selected for a module of the ARMS and is also selected for the Vegetable Chemical Use Survey (VCUS). Such an operation may be randomly assigned to one or the other survey. If that happens then the relevant value of $\pi_{K(i)}$ will be multiplied by 2.

Variance Estimation - When computing a delete-a-group jackknife (Kott 1997), the screening sample is first divided into 15 groups (treating a multi-operation farm as a single sample unit). Jackknife replicate r is defined as the screening sample *minus* the r'th group. A replicate-specific screener weight, $w_{SCR(i)[r]}$, is calculated thusly:

$$w_{SCR(i)[r]} = \frac{\sum_{N_{SCR(i)}} t_g}{\sum_{u_{SCR(i)[r]}} t_g} \qquad (1r)$$

when farm i is in jackknife replicate r; 0, otherwise; where $u_{SCR(i)[r]}$ is the number of screening sample usables in jackknife replicate r.

This has a ripple effect when calculating jackknife-replicate weights. In particular, the replicate version of equation (2) is

$$w_{FOR(i)[r]} = \frac{\sum_{N_{FOR(i)}} t_g w_{SCR(g)[r]}}{\sum_{u_{FOR(i)}} t_g w_{SCR(g)[r]} / \pi_{FOR(g)}} w_{SCR(i)[r]} / \pi_{FOR(i)} . \qquad (2r)$$

For the screener-adjusted replicate weights (the original equations are denoted by (*), we have

$$w_{NAMHS(i)[r]}^{sa} = \frac{w_{SCR(i)[r]}}{[(1-\pi_{FOR(i)})\pi_{NAHMS(i)}]},$$

$$w_{BFCOP(i)[r]}^{sa} = \frac{w_{SCR(i)[r]}}{[(1-\pi_{FOR(i)})(1-\pi_{NAHMS(i)})\pi_{BFCOP(i)}]},$$

$$w_{CNCOP(i)[r]}^{sa} = \frac{w_{SCR(i)[r]}}{[(1-\pi_{FOR(i)})(1-\pi_{NAHMS(i)})(1-\pi_{BFCOP(i)})\pi_{GCCOP(i)}]},$$

$$w_{K(i)[r]}^{sa} = \frac{w_{SCR(i)[r]}}{[(1-\pi_{FOR(i)})(1-\pi_{NAHMS(i)})(1-\pi_{BFCOP(i)})(1-\pi_{GCCOP(i)})\pi_{K(i)}]},$$

$w_{COP(i)[r]}^{sa}$ =

$w_{BFCOP(i)[r]}^{sa}$ if farm i is eligible for beef COP but not grain corn COP,

$w_{CNCOP(i)[r]}^{sa}$ if farm i is eligible for grain corn COP but not beef COP,

$w_{BFCOP(i)[r]}^{sa} / 2$ if farm i is eligible for both COP's and chosen for Beef COP,

$w_{CNCOP(i)[r]}^{sa} / 2$ if farm i is eligible for both COP's but chosen for grain corn.

The replicate version of equation (5) is

$$w_{K(i)[r]} = \frac{\sum_{S_{K(i)}} x_g w_{SCR(g)[r]}}{\sum_{u_{K(i)}} x_g w_{K(g)[r]}^{sa}} w_{K(i)[r]}^{sa} . \qquad (5r)$$

Similarly, we have

$$w_{K(ji)[r]} = \frac{\sum_{S_{K(i)}} x_g w_{SCR(g)[r]}}{\sum_{u_{K(i)}} x_g w_{K(g)[r]}^{sa}} \frac{\sum_{u_{K(i)}} w_{K(g)[r]}^{sa} z_g}{\sum_{u_{Kf(i)}} w_{K(g)[r]}^{sa} z_{gj} F_{K(g)}} w_{K(i)[r]}^{sa} F_{K(i)} , \qquad (6r)$$

$$w_{ECON(i)[r]} = \lambda w_{FOR(i)[r]}, \quad (7r)$$

and

$$w_{ECON(i)[r]} = (1 - \lambda) w_{COP(i)[r]}. \quad (8r)$$

The variants of equation (6) are handled analogously.

Observe that what are *not* changed in these equations are the conditional selection probabilities for farm i (or the operation selected within it) and field j, the determination of which stratum and post-stratum farm i is in, and the value of λ in composite estimation. For simplicity, we will also employ the same x_i (see equation (4)) when computing jackknife-replicate weights. This shortcut is not strictly speaking correct, but it will save a considerable amount of work and is not likely to have a meaningful effect on the results.

The NRI Sample - The composite estimation principles underlying equations (7) and (8) will be used to combine that portion of the NRI sample that can be aggregated with the MC sample for a particular crop. Farms in the NRI sample that have also been sampled for another ARMS module may not be enumerated in some instances. When that happens, they will be treated as NRI non-respondents. When the selected field in the NRI farm sample is also selected for the grain corn COP or a MC module that field will be enumerated for both modules using the COP or MC instrument. The value $F_{K(ij)}$ in equation (6) or (6') or (6") may be replaced by the relevant farm operation acres divided by the field acres - (fall farm operation grain corn acres for the COP, screener crop acres for the MC. (The same replacement is appropriate for MC fields samples that coincide with Objective Yield samples.)

Data Adjustment Factors - For the 1996 ARMS, it is most expedient to simply multiply all final weights (and corresponding jackknife-replicate weights) by the final data adjustment factor (DAF). In the future, we may scale $w_{SCR(i)}$ by the first-phase DAF for farm i and then the final weight for i by its final DAF divided by its first-phase DAF.

Non-Screening States - For states, unlike Nebraska, that have an FOR sample but no screening sample. We can treat the FOR sample of farms as the screening sample. In those states, t_g has been effectively set to 1 in equations (1) and (1r). Equations (2) and (2r) collapse to

$$w_{FOR(i)} = W_{SCR(i)} / \pi_{MOFOR(i)}$$

and

$$w_{FOR(i)[r]} = W_{SCR(i)[r]} / \pi_{MOFOR(i)},$$

where $\pi_{MOFOR(i)}$ is the probability of selecting a particular operation from sample farm i.

Results

The main objective of the ARMS design was to control overlap across several survey modules within each state rather than to increase precision of estimates obtained from the previous designs. In fact, we anticipate that CVS for many of the estimates will be somewhat higher than the same estimates produced from previous designs.

We feel that the 1996 ARMS design has performed well with resulting estimates falling into line with what was expected. Estimated CVS and the performance of the jackknife as a variance estimator for all the estimates produced under the ARMS design are still being reviewed. As of the date of this publication we unfortunately cannot give any final results with regard to CVS obtained by this design.

Reference

Kott, Phillip S. (1997), *Using the Delete-A-Group Variance Estimator in NASS Surveys,* National Agricultural Statistics Service Research Report, forthcoming.