MODEL – ASSISTED-CUM-DESIGN-BASED ADAPTIVE SAMPLING

Ismaila Adeleke[1], **Ray Okafor[2]**, Ebenezer Esan[2], Athanasius Opara[3]
[1]Actuarial Science and Insurance, University of Lagos, Lagos, Lagos, Nigeria
[2]Mathematics, University of Lagos, Lagos, Lagos, Nigeria
[3]Distance Learning Institute, University of Lagos, Lagos, Lagos, Nigeria

INTRODUCTION

Abstract

Adaptive sampling designs are designs in which additional units or sites for observation are selected depending on the interpretation of observations made during initial sampling. Additional sampling is driven by the observed results from an initial sample.

Two problems often crop up in adaptive sampling. One, it may not be feasible to sample according to a designated sampling plan. And two, the prescribed sampling plan may result in very small selection probabilities for some units thereby giving large weights to such units in estimation.

In order to ameliorate these problems, we propose a regression procedure that combines design and model-based techniques of estimation.

KEY WORDS: Model-based, adaptive sampling, inclusion probabilities

Adaptive sampling designs are designs in which additional units or sites for observation are selected depending on the interpretation of observations made during initial sampling. Additional sampling is driven by the observed results from an initial sample.

For population with complex features or characteristics that are rare, unevenly distributed, hidden, or hard to reach, conventional sampling designs such as simple random sampling lead to estimates with large variances and potential biases. With sufficient previous knowledge of the population, precision can be increased through such devices as stratification, systematic designs, and the use of auxiliary information in the design and estimation stages (Cochran 1977; Thompson and Seber 1996)

Most times, however, the uneven patterns in the populations cannot be predicted before sampling. For example, pattern of drugs use may change over time, epidemic(s) progress through cycles, neighborhoods may change their compositions, and economic changes occur; similarly, natural populations of animals or fish may change unpredictably in spatial pattern. For such populations, adaptive sampling strategies can be useful.

Adaptive cluster sampling design is implemented using the following basic elements:

(1) Selecting the initial probability-based sample, (2) specifying a rule or criterion for performing additional sampling, and (3) defining the neighborhood of a sampling unit (Chambers 2003).

A grid is placed over a geographical area of interest (target population) where each grid square is a potential (primary) sampling unit (Thompson , 1992). This is illustrated in (APPENDIX) Figure 1(a). Shaded areas on the figure indicate the area of interest; for instance, areas of elevated contaminant levels. This example has four regions of contamination. The 12 darkened rectangles in the figure represent a randomly selected set of 12 sampling units constituting the initial sample. Whenever a sampled unit is found to exhibit the characteristic of interest — that is, the unit intersects any part of the shaded areas — neighboring sampling units are also sampled using a consistent pattern. An example follow-up sampling pattern is shown in Figure 2, where the x's indicate the neighboring sampling units to be sampled. The follow-up sampling pattern is called the *neighborhood* of a sampling unit. The five grid units in the figure make up the neighborhood of the initially sampled unit. In Figure 1(a), four initial sampling units intersect the shaded areas. The units adjacent to these four initial units are sampled next, as shown in Figure 1(b). Some of these sampled adjacent units also intersect the shaded areas, so the units adjacent to these are sampled next, as shown in Figure 1(c). Figures 1(d) to (f) show subsequent sampling until no more sampled units intersect the shaded areas. Figure 3 shows the initial random sample and the final sample. Note that the final sample covers three of the four regions of contamination. If at least one of the initial units had intersected the fourth area, it would also have been covered by a cluster of observed units.

The final sample consists of *clusters* of selected (observed) units around the initial observed units. Each cluster is bounded by a set of observed units that do not exhibit the characteristic of interest. These are called *edge units.* A cluster without its edge units is called a *network.* Any observed unit, including an edge unit, that does not exhibit the characteristic of interest is a network of size one. Hence, the final sample can be partitioned into non-overlapping networks. These definitions are important in understanding the estimators for statistical parameters.

EARLY DEVELOPMENTS IN ADAPTIVE SAMPLING

A Horvitz-Thompson (HT) and Hansen-Hurwitz (HH)-type estimators of the mean and variance (of the sampled population) based on the final sample, as proposed by Thompson (1990) are typically used with adaptive cluster sampling. Dryver (1999) notes that the selection probabilities generally cannot be determined for all the units in the final sample and that is why a modified version of the HT-type and HH-type estimators using the Rao-Blakwell theorem was proposed (Dryver and Thompson 1999). Unfortunately, as noted by Dryver, "there is not one estimator which is uniformly better than another". But he noted that generally the HT-type estimator is more efficient than the HH-type estimator in the univariate case.

The usual unbiased estimators in adaptive cluster sampling are very simple but do not necessarily utilize all the information gathered. In the case where an initial sample is taken with replacement repeat selections can occur. The usual unbiased estimators do not take this into account. A more efficient estimator that utilizes this information of a repeat selection was discussed by Dryver (1999). Improvements have also been made in the case when an initial sample is taken without replacement. In particular, the values of edge units are utilized in the estimators only for edge units that were picked in the initial sample. Estimators that can incorporate this information can be obtained using the Rao-Blackwell method conditioning on the minimal sufficient statistic (Thompson 1997). These estimators can be computationally challenging. For computing the Rao-Blackwell estimators, Salehi (1998) derived expressions based on inclusion-exclusion formulas. Dryver (1999) derived a new easy-to-compute estimators of higher efficiency by taking the expected value of the usual estimators conditional on a sufficient statistic that is not minimally sufficient.

The improved unbiased estimator proposed was shown to be more efficient that the usual HT estimator.

THE PROBLEM AND WHY IT IS A PROBLEM

There are still many areas of adaptive sampling that deserve attention. A general problem with all design-based unbiased estimators (which is the area researched thus far) is that they are dependent upon the design being carried out properly. When the sampling is not carried out according to the design this can profoundly affect estimation of the parameter of interest. In addition, these sampling problems may be correlated with the parameters of interest. For example, a researcher may not have enough funding to sample the entire network if it is too large and many design-based unbiased adaptive sampling estimators require the entire network to be observed.

For the latter reasons it is important to study model based estimators in conjunction with an adaptive sampling design being used. All the estimators developed for adaptive sampling thus far are design-based. The design-based approach to survey inference has a number of strengths that makes it popular among its practitioners: it automatically takes into account features of the survey design, and it provides reliable inferences in large samples, without the need for strong modeling assumptions. On the other hand, it is essentially asymptotic, and yields limited guidance for small-sample adjustments. It lacks a theory for optimal estimation (Godambe 1995, 1986) and estimates from the approach are potentially inefficient.

The Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952) that was used to derive the estimators for adaptive sampling applies this idea of design based inference more generally. Consider inference about the population total:

$$Y = y_1 + y_2 + \cdot\cdot\cdot + y_n$$

and any sampling design with positive inclusion probabilities

$$\pi_i = E(I_i / y) > 0 \quad for\ units\ i,\ i = 1, 2, \ldots, N$$

the HT estimator is then

$$\hat{Y}_{HT} = \sum_{i=1}^{n} y_i / \pi_i = \sum_{i=1}^{N} I_i y_i / \pi_i \tag{1}$$

and is design unbiased for *Y*.

But (1) above has a major deficiency, for example, when an outlier in the sample has a low selection probability and hence receives a large weight. Basu's (1971) famous circus elephant example provides an amusing example.

Strengths and weaknesses of design-based inference

Little (2003) noted that the design-based approach to survey inference has a number of strengths that make it popular with practitioners. In addition to the fact that it automatically takes into account features of the survey design, it also provides reliable inferences in large samples, without the need for strong modeling assumptions. Though it is essentially asymptotic, and hence yields limited guidance for small-sample adjustments. Unlike models, which lead to efficient inferences based on likelihood or Bayesian principles, the design-based approach is not prescriptive for the choice of estimator. It lacks a theory for optimal estimation (Godambe 1955), and estimates from the approach are potentially inefficient.

The practical bent of survey samplers is illustrated by the fact that Basu (a Bayesian) made fun of the frequentist position by placing it in the domain of "mathematical statistics". On his part, Leslie Kish, an avid design-based advocate, similarly criticizes mathematical statisticians for focussing on i.i.d. models that fail to account for the complex sample design (Kish 1995, Section 9).

Little (1991, 1993, 2002) also noted that superpopulation models are not the basis for inference in the design-based approach, but they can be useful to motivate the choice of estimator; in particular many of the classical estimators for incorporating covariate information, such as the ratio estimator or the regression estimator (e.g. Cochran 1977), can be motivated as arising from linear superpopulation models (see also Elliot and Little 2000).

Model-based-cum-Design-based inference

In situations where the HT model is not reasonable (Little, 2003), a model-assisted modification is to predict the non-sampled values using a more suitable model as proposed below, and then apply the HT estimator to the residuals from that model. Specifically, the generalized regression estimator of $\hat{Y}$ takes the form:

$$\hat{Y} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{n} (y_i - \hat{y}_i)/\pi_i \qquad (2)$$

where $\hat{y}$ is the prediction from a linear regression model relating y to the covariates. The second term on the right side of (2) conveys it with the useful property of design consistency (Brewer 1979, Isaki and Fuller 1982), which means informally that the estimator converges to the population quantity being estimated as the sample size increases, in a manner that maintains the features of the sample design. Design-based statisticians usually weight cases by the design weights w_i when computing this regression, but the estimator (9) is also design consistent if the regression is variance weighted. For discussions of generalized regression estimator and alternatives, see for example Cassel, Särndal and Wretman (1977), Särndal, Swensson, and Wretman (1992).

Another general approach to design-based inference incorporate models by basing inference on "pseudo-likelihoods" that reflect survey design features (Binder, 1983;
Godambe and Thompson, 1986).

ANALYSIS

The analysis is conceived in the context of two phases:
"phase" 1: Outside the networks
"phase" 2: Inside the networks

We estimate separately in each "phase" and combine the estimates. This is done because estimation in "phase" 1 is fairly easy due to random selection of sample units there.

Model Assumptions

1. Initial observations are randomly selected. Subsequent observations in each network are not randomly selected. They are dependent on the initial selection in the network
2. There are Y_i , $i = 1, 2, 3, \ldots, n$ observations in the sample. There are $n^{|}$ observations in the networks and (n- $n^{|}$) observations outside the networks.

3. The population consists of N units made up of $N^{|}$ inside the networks and N- $N^{|}$ outside the networks
4. *y_i in “phase”1 is distributed as i.i.d $N(\mu, \sigma^2)$ while y_i from “phase” 2 is multinomial $N(\underline{\mu}, \sigma^2 A)$*

Analysis proper
In “phase”1
The procedure is to use ordinary least squares to obtain $\hat{y}_i$ since sampling is random. If there is evidence of heteroscedacity, then use generalized least squares.
Thus,

$$Y_{s1} = \sum_{i=1}^{N-N^1} \hat{y}_i + \sum_{i=1}^{n-n^1} (y_i - \hat{y}_i)/\pi_i \qquad (2)$$

In “phase”2
In this phase, there is the real possibility of spatial correlation among the y_i in the network, especially those close to each other and hence the need for a variance-covariance matrix.
Eventually,

$$Y_{s2} = \sum_{i=1}^{N^1} \hat{y}_i + \sum_{i=1}^{n^1} (y_i - \hat{y}_i)/\pi_i \qquad (3)$$

the estimate of the population total can be estimated eventually by,

$$\hat{T} = w_1 Y_{s1} + w_2 Y_{s2} \qquad (4)$$

where $w_1 = \dfrac{N_1}{N}$ *and* $w_2 = \dfrac{N_2}{N}$
N_1 – population size outside the network
N_2 – population size inside the network
N – population size.
Also, w_1 *and* w_2 can be obtained such that the variance of $\hat{T}$ would be minimized. In which case;

$$w_1 = \frac{\operatorname{var}(Y_{s2})}{\operatorname{var}(Y_{s1}) + \operatorname{var}(Y_{s2})} \quad \textit{and} \quad w_2 = \frac{\operatorname{var}(Y_{s1})}{\operatorname{var}(Y_{s1}) + \operatorname{var}(Y_{s2})}$$

with the estimate of the population variance given as

$$\operatorname{var}(\hat{T}) = w_1^2 \operatorname{var}(Y_{s1}) + w_2^2 \operatorname{var}(Y_{s2})$$

suppose Ys1 and Ys2 are correlated, w1 and w2 are estimated thus respectively,

$$w_1 = \frac{Cov(Y_{s1}, Y_{s2}) - \operatorname{var}(Y_{s2})}{2Cov(Y_{s1}, Y_{s2}) - \operatorname{var}(Y_{s1}) - \operatorname{var}(Y_{s2})} \quad \textit{and}$$

$$w_2 = \frac{Cov(Y_{s1}, Y_{s2}) - \operatorname{var}(Y_{s1})}{2Cov(Y_{s1}, Y_{s2}) - \operatorname{var}(Y_{s1}) - \operatorname{var}(Y_{s2})}$$

An illustrative example:
The following example serves to illustrate the computation of the new estimators for a given sample and shows, for a small population, the relative properties of the different types of estimators. The population consists of N=8 units. The initial sample is a simple random sample of n = 2 units. Neighboring (adjacent) units are added whenever the condition $y_i \geq 10$ is satisfied.
Table 1: The first line is the unit labels and the second line is their associated values while the third line is the associated covariates. The population consists of the eight units. The following line of the table are necessary components for calculating various estimators in adaptive cluster sampling (m_i being the number of units in network i, w_i represents the average value of a unit in that network which contains unit i, y_k^* being the sum of units in network i, and α_k, the inclusion probabilities) , with n =2 and a condition $y_i \geq 10$

units	=	1	2	3	4	5	6	7	8
y_i	=	2	15	16	14	9	8	1	4
x_i	=	3	31	32	29	17	17	1	9
m_i	=	1	2	2	2	1	1	1	1
w_i	=	2	15.5	15.5	15	9	8	1	4
Network # k	=	1	2	2	2	3	4	5	6
y_k^*	=	2	31	31	30	9	8	1	4
α_k	=	1/4	13/28	13/28	13/28	1/4	1/4	1/4	¼

Table 2: This table consists of all possible initial samples and a few possible associated estimates of population mean μ. In the first column number before semi-colon represent the initial sample and numbers after the semi-colon represent adaptively added units. **where** $\hat{\mu}_{xy}$ is the proposed estimator of the population mean μ, $\hat{\mu}_{HT}$ the Horvitz-Thompson estimator and $\hat{\mu}_{HT+}$, the improved estimator proposed by Dryver (1999)

The sample	$\hat{\mu}_{xy}$	$\hat{\mu}_{HT}$	$\hat{\mu}_{HT+}$
2,15; 16,14,9	13.153	13.1154	14.8654
2,16;15,14,9	13.153	13.1154	14.8654
2,14;16,9	7.893	9.0769	10.8269
2,9	5.5	5.5	5.5
2,8	5	5	5
2,1	1.5	1.5	1.5
2,4	3	3	3
15,16;14,9,2	14.393	15.5769	12.4103
15,14;16,9,2	14.393	15.5769	12.4103
15,9;16,14,8,2	16.653	16.6154	15.4487
15,8;16,14,9,2	16.153	16.1154	15.2821
15,1;16,14,9,2	12.653	12.6154	14.1154
15,4;16,14,9,2	14.153	14.1154	14.6154
16,14;15,2,9	15.743	15.3462	12.5128
16,9;15,2	13.243	12.8462	11.0962
16,8;15,2	12.743	12.3462	10.8462
16,1;15,2	9.243	8.8462	9.0962
16,4;15,2	10.743	10.3462	9.8462
14,9;16,15,2	16.653	16.6154	14.8654
14,8;16,15,2	16.153	16.1154	14.6154
14,1;16,15,2	12.653	12.6254	12.8654
14,4;16,15,2	14.153	14.1154	13.6254
9,8	8.5	8.5	8.5
9,1	5	5	5
9,4	6.5	6.5	6.5
8,1	4.5	4.5	4.5
8,4	6	6	6
MEAN	10.71941	8.625	8.625
BIAS	2.09441	0.0000	0.0000
MEAN SQUARE ERROR	28.13806	50.72843	48.79968

Table 3: This table consists of all possible initial samples and their variances from table 1. In the first column, numbers before the semi-colon represent the initial sample and numbers after the semi-colon represent adaptively added units. **where** $\hat{\mathrm{var}}(\hat{\mu}_{xy})$ is the variance of the proposed estimator and $\hat{\mathrm{var}}(\hat{\mu}_{HT+})$ is the improved estimator of the Horvitz-Thompson proposed by Dryver (1999)

The sample	$\hat{\mathrm{var}}(\hat{\mu}_{HT}$	$\hat{\mathrm{var}}(\hat{\mu}_{xy})$	$\hat{\mathrm{var}}(\hat{\mu}_{HT}$
2,15; 16,14,9	77.36428	60.47542	74.30178
2,16;15,14,9	77.36428	60.47542	74.30178
2,14;16,9	33.57701	2.319527	30.51451
2,9	24.5	24.5	24.5
2,8	18	18	18
2,1	0.5	0.5	0.5
2,4	2	2	2
15,16;14,9,2	87.953	9.278107	77.92522
15,14;16,9,2	87.953	9.278107	77.92522
15,9;16,14,8,2	84.73447	60.47542	83.37336
15,8;16,14,9,2	82.55658	60.47542	81.86214

15,1;16,14,9,2	77.81139	60.47542	75.56139
15,4;16,14,9,2	77.59504	60.47542	77.34504
16,14;15,2,9	82.03	2.319527	74.00222
16,9;15,2	42.6408	2.319527	39.5783
16,8;15,2	40.54926	2.319527	38.29926
16,1;15,2	36.40805	2.319527	36.34555
16,4;15,2	35.9331	2.319527	35.6831
14,9;16,15,2	84.73447	60.47542	82.48447
14,8;16,15,2	82.55658	60.47542	82.49408
14,1;16,15,2	77.81139	60.47542	75.56139
14,4;16,15,2	77.59504	60.47542	77.53254
9,8	0.5	0.5	0.5
9,1	32	32	32
9,4	12.5	12.5	12.5
8,1	24.5	24.5	24.5
8,4	8	8	8
MEAN = MSE	50.72843	28.13806	48.79968

Comment

As observed from tables 3 and 4 above, the proposed model is biased but gives more precise estimates as compared to the HT estimates.

Currently, the environmental management unit of the department of Chemistry, University of Lagos in collaboration with the department of Mathematics, University of Lagos is conducting a research on the presence and spread of heavy metals at Owode-Oniring area of Lagos, Nigeria. The adaptive sampling method is used in sample selection and the model-assisted-cum-design-based analysis shall be used amongst others to analyse the data from the sample.

References

Basu, D. (1971), "An essay on the logical foundations of survey sampling, Part 1," in *Foundations of Statistical Inference*, Toronto: Holt, Rinehart and Winston, pp. 203-242.

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Sample Surveys. *International Statistical Review* 51, 279-92.

Brewer, K. R. W. (1979), "A Class of Robust Sampling Designs for Large-Scale Surveys," *Journal of the American Statistical Association* 74, 911-915

Cassel, C-M, Särndal, C-E. and Wretman, J.H. (1977), *Foundations of Inference in Survey Sampling*, New York: Wiley.

Chambers, R.L. (2003), "Introduction to Part A," in *Analysis of Survey Data*, R.L. Chambers and C.J. Skinner, eds., New York: Wiley, pp.13-27.

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

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, London: Chapman Hall.

Dryver A.L., and Stephen Thompson (1999). "*Improving Unbiased Estimators in Adaptive Sampling* ". Journal of the American Statistical Association, 89, 727-731

Dryver A.L., (1999). Adaptive Sampling designs and Associated Estimators. PhD thesis, the Pennsylvania State University.

Elliott, M. R. and Little, R.J.A. (2000). Model-Based Alternatives to Trimming Survey Weights. *Journal of Official Statistics* 16, No. 3, 191-209.

Godambe, V.P. (1955), "A Unified Theory of Sampling from Finite Populations," *Journal of the Royal Statistical Society*, B 17, 269-278.

Godambe, V.P. and Thompson, M.E. (1986), "Parameters of Superpopulations and Survey Population: the ir Relationship and Estimation. *International Statistical Review* 54, 37-59.

Horvitz, D.G., and Thompson, D.J. (1952), "A Generalization of Sampling without
Replacement from a Finite Universe," *Journal of the American Statistical Association* 47, 663-685.

Isaki, C. T., and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model", *Journal of the American Statistical Association* 77, 89-96.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Kish, L. (1995). "The Hundred Years' Wars of Survey Sampling," *Statistics in Transition*, 2, 813-830. Reproduced as Chapter 1 of G. Kalton and S. Heeringa, (2003,eds.) *Leslie Kish: Selected Papers*, New York: Wiley.

Little, R.J.A. (1991), "Inference with Survey Weights," *Journal of Official Statistics* 7,
405-424.

Little, R.J.A. (1993), "Post-Stratification: a Modeler's Perspective," *Journal of the*
American Statistical Association 88, 1001-1012.

Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data,* 2nd edition,
New York: Wiley.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992), *Model Assisted Survey Sampling*, Springer Verlag: New York.

Salehi, M.M. (1998). Adaptive Cluster Sampling. PhD thesis, University of Auckland

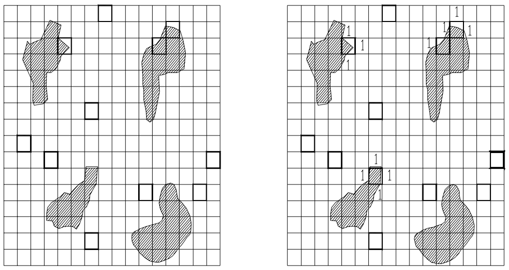
Thompson, S.K. (1990), Adaptive Cluster Sampling. Journal of the American Satistical Association 85, 1050-1059

Thompson, S.K. (1992), Sampling. New York:Wiley

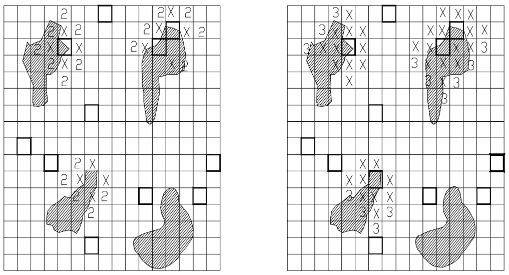
Thompson, S.K., Seber, G.A.F. (1996), Adaptive Sampling. New York: Wiley

Thompson, S.K. (1997), Adaptive Sampling in behavioural survey. In Harrison, L., and Hughes, A. eds., The Validity of Self-reported Drug Use: Improving the Accuracy of Survey Estimates. NIDA Research Monograph 167. Rockville, MD: National Institute of Drug Abuse, 263-319

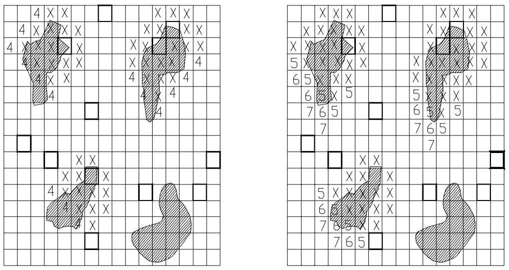
APPENDIX



(a) Initial sample (b) First batch of adjacent units



(c) Second batch of adjacent units (d) Third batch of adjacent units



(e) Fourth batch of adjacent units (f) Fifth, seventh and sixth batch of ad

Fig 1: Population Grid with Shaded Area of Interest, Initial Simple Random Sample, and Follow-up Sample

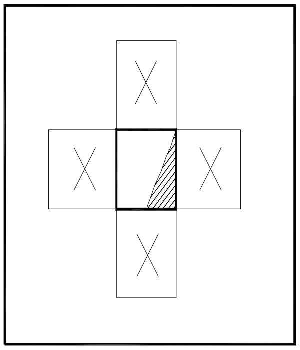
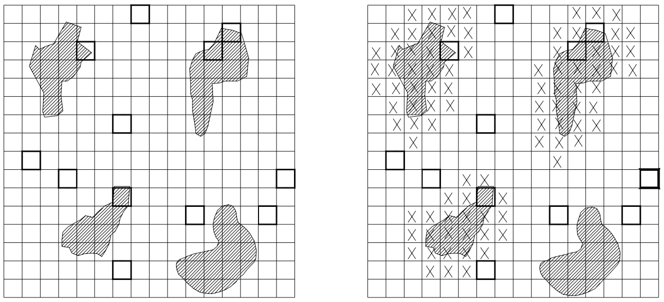


Fig. 2: Follow-up Sampling Pattern



Population Grid with Shaded Areas of Interest and Initial Sample Random Sample

Final Adaptive Sampling Results

X = Observed Sampling unit

Fig. 3: Population Grid with Shaded Areas of Interest, Initial Simple Random Sample, and Final Sample