OPTIMAL DESIGN OF A PLOT CLUSTER FOR MONITORING

Charles T. Scott
Research Forester
Northeastern Forest Experiment Station
U.S.D.A., Forest Service
Delaware, OH 43015 USA

SUMMARY

Traveling costs incurred during extensive forest surveys make cluster sampling cost-effective. Clusters are specified by the type of plots, plot size, number of plots, and the distance between plots within the cluster. A method to determine the optimal cluster design when different plot types are used for different forest resource attributes is described. The method requires cost and variance relationships in order to develop an optimal design. The cluster design developed for the Forest Health Monitoring Program of the United States is presented.

INTRODUCTION

Land managers and policy makers need accurate resource statistics to make informed decisions on the management of the forest resource. When developing the forest survey to collect data, the survey planner makes two key decisions: the sample unit selection rule and the sampling unit (cluster) design. The objective of this study was to develop a multiresource cluster design. This design was used for the Forest Health Monitoring Program of the United States -- a network of multiresource permanent clusters distributed across the Nation. Because the focus of this paper is on the cluster design, only simple random sampling of clusters was considered. Extensions to other selection rules can be made.

METHODS

A cluster sample can be composed of various plot types (subsampling units) used to measure a variety of resource attributes. Scott (1981) described the sampling unit characteristics as a combination of a specified number of plots arranged in a specific spatial pattern. Plot characteristics are:

1. type of elements sampled, such as trees, shrubs, soils, or streams;
2. element selection rule, such as equal probability versus probability proportional to size;
3. geometric type, such as point, line, area, and volume;
4. shape, such as circular or rectangular.

The cluster design can be determined by specifying for each plot type, k:

1. number of plots within the cluster, m_k,
2. plot size, z_k,
3. distance between plots, d_k, and their spatial arrangement.

Together with the total number of clusters to be sampled, n, these cluster design variables influence the survey variances and costs. The first step in specifying the cluster design is to determine which plot types are to be used.

Sampling Unit Design

Several methods of selecting overstory trees have been suggested, but the most commonly used are fixed-area plots and variable-radius plots (horizontal point samples). Many studies were conducted comparing the efficiencies of fixed-area plots versus variable-radius plots, such as Grosenbaugh and Stover 1957. In general, attributes that are related to tree frequency are more efficiently estimated using fixed-area plots, and attributes related to tree basal area are best estimated using variable-radius plots (Scott and Alegria 1990).

Since the 1970's, the emphasis in natural resource surveys has broadened to include the entire forest ecosystem, including all plant species, soils, wildlife habitat, and water resources. Fixed-area plots have also been used extensively for plant species in the understory (Lindsey et al. 1958). Line-intersect sampling has been used to estimate the amount of edge between various land cover types (Barnes and Scott 1983). Thus, a multiresource cluster design can include a variety of plot types, such as fixed-area, variable-radius, and line samples.

Once the plot types have been specified, the cluster design variables, m_k, d_k, and z_k, must be determined in a cost-effective manner. Most studies in forestry have focused on plot size, z_k, and on one or two attributes of interest (for example, Freese 1961). In these studies, the variance and costs were modeled as a function of the plot size. This approach was then extended to the study of clusters of plots for one or two attributes (for example, Arvanitis and O'Regan 1972). Some have explicitly modeled the distance between elements, d_k, as a design variable (for

example, Nyyssonen et al. 1971 and Ek et al. 1984).

The problem can be set up either to minimize variance subject to a fixed cost constraint or to minimize cost subject to fixed precision constraints. Arvanitis and O'Regan (1972) used a simulation approach in which designs were tested and compared. Another approach is to use linear or nonlinear programming techniques.

Variance Estimation

In one of the first attempts to relate a cluster design variable to variance, Smith (1938) developed a model of variance, V, as a power function of plot size. This relationship has been used widely and has performed well. Scott (1981) extended this model to all of the cluster design variables:

$$R_i(m_k,\bar{d}_k,z_k) = b_{0i}\, m_k^{b_{1i}}\, \bar{d}_k^{b_{2i}}\, z_k^{b_{3i}} \qquad (1)$$

where:

R_i = rel-variance between clusters of attribute $y_i = v(y_i)/\bar{y}^2$

b_{ji} = regression coefficients with j=0,1,2,3

$$\bar{d}_k = \begin{cases} \text{average paired distance between subunits of type } k \\ 1, \text{ if only one subunit } (m_k=1) \end{cases}$$

The rel-variance is convenient when the precision is specified as a percent of the mean, for example, ± 10 percent. The exponents of the cluster design variables are negative and theoretically should range from 0 to -1. Thus, as the cluster design variable increases, the rel-variance decreases.

Often two concentric plot types can collect data on the same attribute of interest, such as a poletimber plot of size, z_k, within a large sawtimber plot of size, $z_{k'}$. Both plot types contribute to the estimate of the number of trees. In this case, the model form chosen was:

$$R_i(m_k,\bar{d}_k,z_k) = b_{0i}\, m_k^{b_{1i}}\, \bar{d}_k^{b_{2i}}\, z_k^{b_{3i}}\, z_{k'}^{b_{4i}} \qquad (2)$$

Costs

As the cluster design variable increases, the cost also increases. The cost functions must relate the survey costs to the cluster design variables and to the number of clusters, n. They should be as accurate as possible and should reflect the actual field and perhaps office work involved, so long as the costs are related to the design variables. Scott (1981) introduced the effect of daily costs. The number of days, D, is the number of clusters, n, divided by the number of clusters per day, CPD. Daily costs include traveling from the office to the first cluster of the day and back from the last cluster, lodging and meals, paid breaks, and supervision. Other

types of costs include the travel time between clusters, which is a function of the plot density and the number of clusters per day. Costs that relate only to the number of clusters include locating and establishing the first point in the cluster, and returning from the cluster. Measurement costs are related to both plot size and type.

Cluster Configuration

The cluster configuration defines the way in which the cluster is laid out on the ground. In this study, the cluster design is a systematic arrangement of the plots. The configuration is a function of the number of plots of each type, and the distance and direction between plots. In order to keep the plots within the minimum sample area of 0.4 ha, the configuration used had one center plot surrounded by the remaining plots which were equally spaced at a fixed distance from the center, thus forming a regular polygon with one plot in the center. By specifying the distance, d_k, between the center and the remaining plots for this configuration, the total walking distance, d_{walk}, and the average paired distance between plots can be determined. Once the radius, d_k, is specified, the walking distance can be computed as walking from center to plot 2, then along the sides to the remaining plots:

$$d_{walk} = d_k + (m_{max}-1)\, d_k \sqrt{2(1-\cos(2\pi/m_{max}))} \qquad (3)$$

where:

$$m_{max} = \max(m_k-1); \quad \text{for } k=1,K$$

The average paired distance was computed as the average of all possible distances between pairs of plots. To avoid recalculating it for each situation, it was computed as a function of the number of plots and the radius, d_k:

$$\bar{d}_k = d_k f(m_k) \qquad (4)$$

where:

f(1) = undefined	f(6) = 1.3592	f(11) = 1.3298
f(2) = 1.0	f(7) = 1.3520	f(12) = 1.3258
f(3) = 1.3333	f(8) = 1.3453	f(13) = 1.3224
f(4) = 1.3660	f(9) = 1.3394	f(14) = 1.3194
f(5) = 1.3657	f(10) = 1.3343	f(15) = 1.3167

Optimization Problem

When various cost components are put into a single form and lower bounds are added to the design variables, the optimization problem can be stated as:

$$\text{minimize } C = D\left[C_{daily}+(CPD-1)C_{bet}+ C_{crew}\{T_{daily}+(CPD-1)T_{bet}\}\right] + n\, C_{crew}\left[T_{loc}+ T_{walk}d_{walk}+ m_{max}T_{est}+ \sum_{k=1}^{K} m_k T_k(z_k)\right] \quad (5)$$

Subject to:

$$R_i(m_k,\bar{d}_k,z_k)/n \le V_i \quad (i=1,I)$$

$$\frac{HPD-T_{daily}+T_{bet}}{T_{bet}+T_{plot}} \ge CPD \text{ (an integer)}$$

$$n \ge 2$$

$$m_k \ge 0 \quad (k=1,K)$$

$$d_k \ge 0 \quad (k=1,K)$$

$$z_k \ge 0 \quad (k=1,K)$$

where:

$T_k(z_k)$ = time to measure a plot of size z_k

V_i = desired precision (variance as a proportion of the mean)

HPD = number of working hours per day (e.g., 8)

$$T_{plot} = T_{loc}+ T_{walk}d_{walk}+ m_{max}T_{est} + \sum_{k=1}^{K} m_k T_k(z_k)$$

The cost and time variables are defined in Table 1.

Although this study focused on minimizing costs for fixed precision levels, the same variance and cost functions could be used to minimize some weighted function of the variances to achieve a fixed cost.

Solution Technique

The model can be treated as a large integer nonlinear programming problem. The particular solution technique used was the m-neighborhood method (Garfinkel and Nemhauser 1972, p. 330). By treating the distance, d_k, and plot size, z_k, as integers (for example, 5 m and 0.001 ha, respectively), this method searches all possible combinations of the design variables in the m-neighborhood, that is, within m increments of the starting value. The combination which minimizes the cost the most is used as the next starting point. This process is repeated until no

improvements are found. Other starting points can be tested until the best combination is determined. Although this method does not guarantee finding the global optimum, it usually improves upon any existing design.

DATA AND RESULTS

Estimating the cost and variance function parameters requires the collection of preferably 40 or more clusters. In addition to the resource data, crew times and other costs must be recorded. The cluster must be designed in such a way that several distances between plots can be evaluated. Also, the cluster should have more plots than would be expected in the final solution. Also, several plot sizes should be evaluated by recording tree locations on a large plot, or by collecting data on a series of sizes.

The first set of data used in this study was taken in 1968 and 1981 by Forest Inventory and Analysis (FIA), USDA Forest Service, on 61 remeasured plots in Hancock County, Maine. These plots were unique in that two cluster configurations were co-located: 1) a 1/10-ac (0.04 ha) poletimber plot and a 1/5-ac (0.08 ha) sawtimber plot; and 2) a cluster of ten 37.5 ft^2/ac (8.6 m^2/ha) horizontal point samples (Figure 1). Detailed time data on each observation were recorded by an additional crew member. In addition, data were collected on 20 new plots located in central New Jersey. Some new variables and less detailed time data were recorded, but the cluster configuration (Figure 2) provided data on a wider variety of configurations.

Cost Functions

Cost functions and times were developed using standard regression techniques (Table 1). The values presented are either simple averages or relate time to a cluster design variable. In general, the times were highly variable. Attempts to relate the times to the magnitude of the item recorded resulted in low correlations. However, the averages do prove useful in determining the optimal cluster design.

Rel-variance Functions

The rel-variance functions were developed using nonlinear regression on various subsets of the data. The subsets were selected to form various cluster configurations with different numbers of plots, distances between them, and plot sizes. For each of these combinations, the rel-variances across all clusters were computed. These rel-variances were then regressed against the cluster design variables (Table 2). In general, the rel-variances also proved to be highly variable and proved difficult to model. Often the parameter for distance would have the incorrect sign. This was a result of both its minimal affect on rel-variance at these

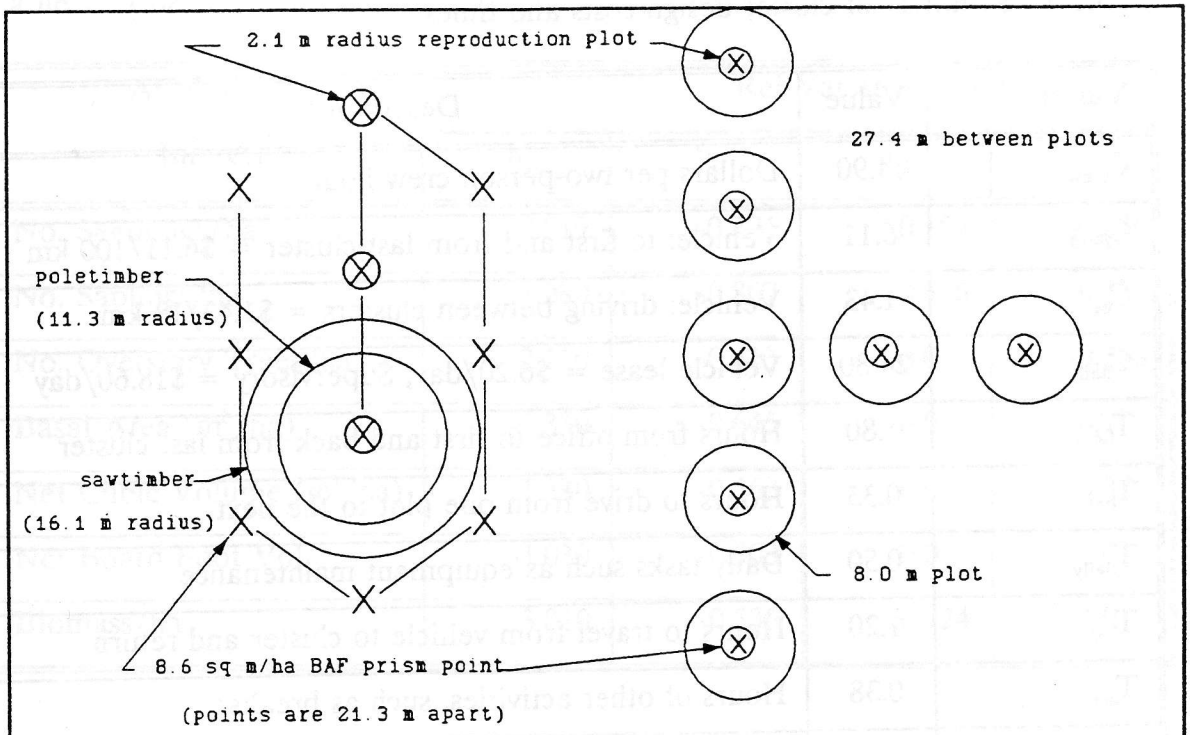

Figure 1. Maine cluster design

Figure 2. New Jersey cluster design

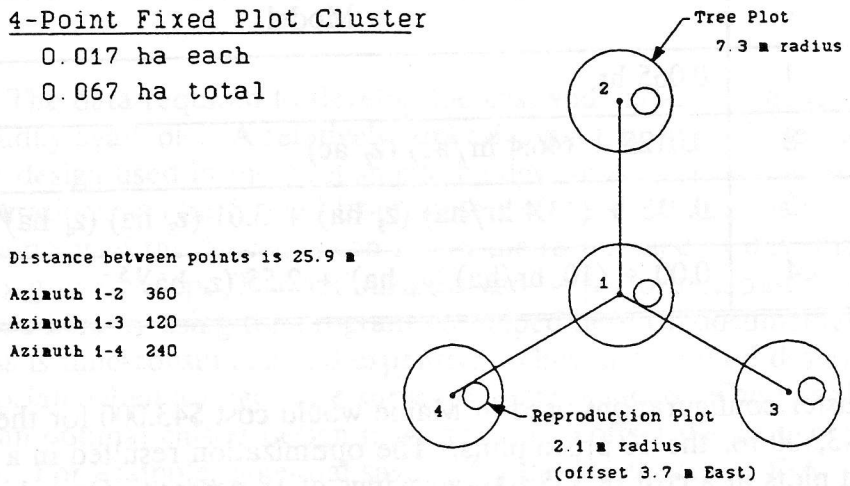

Figure 3. Optimal Forest Health Monitoring cluster design

short distances, and its correlation with the number of plots. Techniques such as ridge regression were not investigated.

Optimization Results

The optimum cluster design program was written in FORTRAN and was based on FIA's precision requirements which could be stated in terms of the maximum rel-variance allowed. To meet the precision requirements specified in this

Table 1. Summary of cluster design costs and times

Variable	Value	Description
C_{crew}	14.90	Dollars per two-person crew hour
C_{field}	6.11	Vehicle: to first and from last cluster = \$6.11/100 km
C_{bet}	1.43	Vehicle: driving between clusters = \$1.43/18 km
C_{daily}	24.80	Vehicle lease = \$6.20/day; Supervisory = \$18.60/day
T_{field}	1.80	Hours from office to first and back from last cluster
T_{bet}	0.33	Hours to drive from one plot to the next
T_{daily}	0.50	Daily tasks such as equipment maintenance
T_{loc}	1.20	Hours to travel from vehicle to cluster and return
T_{act}	0.38	Hours of other activities, such as breaks
T_{walk}	.0003	Hours per meter of walking
T_{est}	0.04	Hours to establish (monument) each plot center
Function	k	Model
Foliage	1	0.065 hr
Reproductn.	2	0.0125 + (46.4 hr/ac) (z_2 ac)
Poletimber	3	0.013 + (13.4 hr/ha) (z_3 ha) + 5.61 (z_3 ha) (z_4 ha)$^{0.5}$
Sawtimber	4	0.00 + (10. hr/ha) (z_4 ha) + 2.55 (z_4 ha)$^{1.5}$

study, the two cluster configurations used in Maine would cost \$43,000 for the fixed-area plots and \$43,500 for the 10 prism plots. The optimization resulted in a cluster of four fixed-area plots at a cost of \$35,400--a savings of 18 percent. One cluster could be observed each day, and the day was fully utilized. The optimum design is shown in Figure 3.

In 1990 in the six New England states, a network of Forest Health Monitoring clusters was established. The purpose of the monitoring is to characterize the forest conditions and potential stressors, to detect and quantify changes in conditions, and to correlate these changes with potential stressors. The cluster design chosen was based on the optimization performed for FIA. The design was modified by referencing each point back to the center and by offsetting the reproduction plot 3.7 m to the East to avoid trampling (Figure 3). The design is being used as the network of clusters expands throughout the United States in the coming years.

Table 2. Coefficients of rel-variance function (2) for seven attributes.

Attribute of Interest	Rel-Variance Coefficients			
	b_0	b_1	b_2	b_3
No. Seedlings/ha	0.019	-0.771	-0.154	-0.236
No. Saplings/ha	-1.080	-0.807	-0.326	-0.536
No. Overstory Trees/ha	-2.110	-0.413	0.0	-0.368
Basal Area (m^2/ha)	-3.350	-0.645	0.0	-0.772
Net Cubic Volume (m^3/ha)	-1.330	-0.563	0.0	-0.375
Net Board Foot Vol.	-1.030	-0.503	0.0	-0.425
Biomass/ha	-3.810	-0.535	-0.124	-1.0

DISCUSSION AND CONCLUSIONS

The data required to develop the cost and variance functions are generally not readily available. A relatively expensive pilot survey will be required. The cluster design used in the pilot should be developed so that a variety of cluster configurations can be formed from subsets of the data, such as in Figure 2. The development of the functions can follow the forms used in this study and need not take long to develop. The optimization can be performed based on the algorithms reported here or using the program developed here (undocumented). The whole process is time-consuming and expensive. Thus, the method described here is only appropriate when a large-scale survey is being planned. The benefits derived in using an optimal cluster design must more than offset the costs of developing the design. For example, potential savings for the Northeastern FIA are about $90,000 annually.

The cost functions used in this study may be applicable in other areas or can be modified. More detailed information on these and other variables is available from the author. The rel-variance functions proved difficult to model. Other model forms or regression techniques should be investigated. Also, other optimization methods could be used which take advantage of the fact that the distance and plot size variables are continuous. Another approach, which avoids the model development and optimization steps, is to use the cost and variance data directly. Find the subset of data which meets the precision requirements and minimizes cost. Although this cluster configuration may not be optimal, it avoids many of the analytical steps and model assumptions involved.

The optimization methods proposed in this study can be thought of in a very general way. The first step is to identify the design variables. In this situation they are: 1) the number of clusters, 2) the number of plots within a cluster, 3) the distance between plots, and 4) the plot size. All costs that are related to these variables must be quantified and put into a form similar to equation (5). The rel-variances of all attributes of interest should be modeled as a function of the design variables. The costs are then minimized subject to precision constraints. The final step is to perform a sensitivity analysis to determine the robustness of the design chosen. These steps can be applied to almost any survey sampling problem resulting in a very efficient design.

LITERATURE CITED

Arvanitis, L.G.; O'Regan, W.G., 1972: Cluster or satellite sampling in forestry: a Monte Carlo computer simulation study. Proceedings Third Conference IUFRO Advisory Group of Forest Statisticians, Jouy-en-Josas, France. pp. 191-205.

Barnes, R.B. and Scott, C.T., 1983: Edge sampling techniques from aerial photographs. Wildlife Society Bulletin, vol. 11, pp. 389-391.

Ek, A.R.; Scott, C.T.; Zeisler, T.R.; Benessallah, D., 1984: Cluster sampling: from theory to practice. Proceedings Inventorying Forest and Other Vegetation of the High Latitude and High Altitude Regions. Fairbanks, AK, Society of American Foresters SAF 84-11, pp. 78-81.

Freese, F., 1961: Relation of plot size to variability: an approximation. Journal of Forestry, vol. 59, p. 679.

Garfinkel, R.S. and Nemhauser, G.L., 1972: Integer Programming. John Wiley and Sons, New York. 427 p.

Grosenbaugh, L.R. and Stover, W.F., 1957: Point sampling compared with plot sampling in southeast Texas. Forest Science, vol. 3, no. 1, pp. 2-14.

Lindsey, A.A.; Barton, J.D.; Miles, S.R., 1958: Field efficiencies of forest sampling methods. Ecology vol. 39, pp. 428-444.

Nyyssonen, A.; Roiko-Jokela, P.; Kilkki, P., 1971: Studies of the number of sample plots of different sizes. Acta Foresta Fennica 75.2, 17 p.

Scott, C.T., 1981: Design of optimal two-stage multiresource surveys. Ph.D. Dissertation, University of Minnesota, St. Paul, 138 pp.

Scott, C.T. and Alegria, J., 1990: Fixed- versus variable-radius plots for change estimation. In LaBau, V.J. and Cunia, T., eds. Symposium on State of the Art Methodology of Forest Inventory, Workshop Proceedings. July 30 - August 5, 1989; Syracuse, NY: General Technical Report PNW-GTR-263. Portland, OR: USDA, Forest Service, Pacific Northwest Research Station. p. 126-132.

Smith, H.F., 1938: An empirical law describing heterogeneity in the yields of agricultural crops. Journal of Agricultural Science, vol. 28, no. 1, pp. 1-23.

18 November, 1991