Stratum Weight Determination Using Shortest Path Algorithm

Susan L. King[1]

Abstract.—Forest Inventory and Analysis uses post-stratification to calculate resource estimates. Each county has a different stratification, and the stratification may differ depending on the number of panels of data available. A "5 by 5 sum" filter was passed over the reclassified forest/nonforest Multi-Resolution Landscape Characterization image used in Phase 1, generating an image in which each pixel represents the count of forested pixels inside a 5 by 5 window. The forested pixel count ranges from 0 to 25 or 26 classes. In the next step, the ground plots are overlaid on the class map generated by the 5 by 5 window. The objective is to find the break points in the 26 classes that minimize the difference in the number of acres/plot between strata while simultaneously maximizing the number of strata. These are conflicting goals. More strata imply larger deviances between the strata. Also, the stratum must have contiguous classes with at least four plots. This is a nonlinear integer programming problem. Because software is not readily available to solve a nonlinear integer programming problem, the problem was reformulated to finding the shortest path through the network. For each county, the optimal one, two, three, four, five, six, and seven strata are found, and various heuristics for determining the final solution are investigated and compared.

Introduction

The annual Forest Inventory and Analysis (FIA) sampling design is composed of three phases. Phase 1 uses satellite imagery to classify the land area in a State as forest or nonforest. Phase 2 is the traditional ground sample. An interpenetrating hexagonal grid is placed across a State with one ground plot per grid cell. Each hexagonal grid represents 5,937 acres. One-fifth of the ground plots spread uniformly across the State are visited yearly. Each year's plots/hexagonal grids are referred to as a panel. On a subset of the Phase 2 plots, additional variables are measured to determine forest health. This subset is the Phase 3 sample. This article focuses on finding an automated and efficient procedure for determining the optimal number of strata and, hence, the stratum weights for Phase 2 forest land estimates or "on the fly" resource estimates of user-defined polygons. The objective is to minimize the difference in the Phase 1-to-Phase 2 ratio between the strata (deviance) while simultaneously maximizing the number of strata. Each stratum must have a minimum of four ground plots. When the population is divided into as many homogenous strata as possible, the variance of the population estimates tends to be lower. As the number of strata increase, it becomes more difficult to find break points in which all the strata have approximately equal Phase 1-to-Phase 2 ratios.

Methods

Phase 1 and Phase 2 Cost Information

The satellite imagery used for the Phase 1 sample was a forest/nonforest map acquired from National Land Cover Data (formerly Multi-Resolution Land Characterization [MRLC]). This vegetation map was made by the U.S. Geological Survey Earth Resources Observation Systems (EROS) Data Center (Vogelmann *et al*. 2001) and is based on 1992 Landsat 7 Thematic Mapper data; other intermediate-scale spatial data were used as ancillary data. For the forest/nonforest call, the MRLC was reclassified so that the forest classes and woody wetland received a value of 1, and other pixels received a value of 0. A "5 by 5 sum" filter was passed over the reclassified forest/nonforest MRLC image, generating an image in which each pixel represents the count of forested pixels inside a 5 by 5 window. The forested pixel count ranges from 0 to 25, which creates 26 classes (bins). The Phase 2 plots were overlaid on the filtered image to obtain a

[1] Operations Research Analyst, U.S. Department of Agriculture, Forest Service, Northeastern Research Station, Newtown Square, PA, 19073. Phone: 610–557–4048; fax: 610–557–4250; e-mail: sking01@fs.fed.us.

forested class call for each plot. For the Phase 1 sample, both the total number of pixels in a polygon of interest, such as a county, and the number of pixels in each bin are known. This information is used to develop a cost function, which is not limited to a monetary function. A cost function also can be time, distance, or another measure to be optimized. Again, the objective is to break the 26 bins into strata that minimize the variance by equalizing the Phase 1-to-Phase 2 stratum acres/plot costs while simultaneously maximizing the number of strata. Each stratum must have contiguous bins and at least four ground plots. The bins must be contiguous so that similar forested and nonforested bins are grouped together.

To mathematically formulate the cost information, let the 26 bins be numbered from 1 to 26. There are ℓ strata, and b_1, b_2,…, $b_{\ell-1}$ are the break points between strata. The first bin (end point b_0) is always in stratum 1, and the last bin (end point b_ℓ) is always in stratum ℓ. The Phase 1 area for each county or polygon for bin i is:

$$a_i = \left(\frac{\text{number of pixels in bin i}}{\text{total number of pixels in county}} \right) \text{(total acres in county)} \quad (1)$$

On a per stratum basis for a county, the Phase 1 area for stratum j is:

$$e_j = \sum_{i=b_{j-1}}^{b_j} a_i \qquad \text{for } j=1,\ldots,\ell \quad (2)$$

The Phase 2 county or polygon sample size for each bin is:

$$s_i = \sum_{i=b_{j-1}}^{b_j} \text{number of ground plots in bin i} \qquad \text{for } j=1,\ldots,\ell \quad (3)$$

The Phase 1-to-Phase 2 ratio is the number of acres/plot, also known as the stratum weight. This stratum weight will be used as the "cost" for stratum j.

$$C_j = \frac{e_j}{s_j} \qquad \text{for } j=1,\ldots,\ell \quad (4)$$

Objective Function

Two objective functions are defined. The deviance objective function is the sum of the absolute value of the cost differences between a stratum and the adjacent lower stratum. This is expressed mathematically as:

$$\text{deviance} = \sum_{j=1}^{\ell-1} \left| (C_{j+1} - C_j) \right| \quad (5)$$

The smoothed deviance is the sum of the absolute value of the cost differences between a stratum and all the previous strata. This is expressed mathematically as:

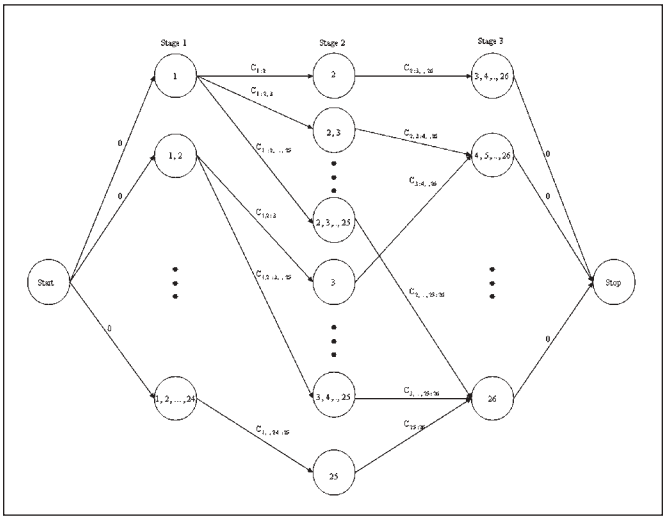
$$\text{smoothed deviance} = \sum_{j=1}^{\ell-1} \left| (C_{j+1} - C_j) \right| + \left| (C_{j+2} - C_j) \right| + \ldots + \left| (C_\ell - C_j) \right| \quad (6)$$

By including all possible pairs of strata, the smoothed deviance should better reduce the cost difference between strata over the deviance objective function. In each case, these objective functions tend to result in proportional allocation to the strata, which is the expectation from a systematic sample of plots.

Shortest Path

Mathematical optimization is a tool for finding the combination of decision variables and their values that minimize or maximize an objective function while simultaneously satisfying a set of constraints on the decision variables. In this article, the deviance or smoothed deviance function is the objective function, and the constraints are the contiguous bin requirement and the lower bound on the number of ground plots per stratum. One mathematical optimization procedure to optimally determine the best allocation of bins to a stratum is the shortest path algorithm. The problem is formulated as a feed-forward network (fig.1). A

Figure 1.—*Each stage in this three-stage network corresponds to a stratum. Bin combinations are located at the node, and the cost between bin combinations is located on the arcs.*



network has nodes joined by arcs, and, in this problem, the arcs are directed in only one direction. The objective is to traverse the network from the start node to the stop node with the least cost. Each arc has an associated cost, which could be time, money, distance, or another measure. In this case, cost is the deviance or smoothed deviance objective function. The stages in the network correspond to the strata. The network in figure 1 is a three-stage or three-strata network. The cost encountered from the start node to stage 1 is 0, and the cost encountered from stage 3 to the stop node is 0. The nodes in each stage correspond to the number of bins. In stage 1, all the nodes include the first bin, and, in stage 3, all the nodes contain bin 26. In stage 1, if bins 1 through 24 are selected, in the three-stratum case, bin 25 must be selected in stage 2, and bin 26 selected in stage 3. If only bin 1 is in the first stratum, many combinations exist for strata 2 and 3. If four plots are not in the bin combination at a node, the arc is assigned a large cost so that this path never will be selected. Paths containing these infeasible arcs are pruned from the network before solving for the shortest path through the network.

Many approaches and algorithms are available to solve a shortest path problem. Dijkstra's Algorithm (1959) is the classic method for computing the shortest path from a single source node to every other node in a weighted (stratum weight cost on an arc is a weight) network. This algorithm, a simple and consequently easily implemented algorithm for finding the shortest routes, is the most widely used in GIS software packages. Dijkstra's Algorithm is used for solving problems that require real-time solutions—for example, routing an ambulance to an accident site and from there to the nearest hospital. Its performance depends on the data structures (for example, heaps or priority queues) used to represent the network (Derekenaris *et al*. 2001). Improving the data structure efficiency of Dijkstra's Algorithm and other approaches to solving the shortest path problem are active areas of research. Nevertheless, shortest path algorithms are used routinely to solve large-scale problems and are available in most programming languages.

Almost any problem that can be formulated as a shortest path through a network also can be solved using dynamic programming, that is, the problem can be solved using recursive equations without special software. The drawback of dynamic programming is the "curse of dimensionality." As both the number of stages and nodes increase, so do the number of recursions. Information for the recursions must be stored in a lookup table. The feasibility of using dynamic programming for solving the stratum weight problem was not investigated.

Another mathematical optimization approach to optimally allocating bins to the strata is nonlinear integer programming. If bin i is assigned to stratum j, the decision variable is 1; otherwise, the decision variable is 0. In addition to the constraints requiring at least four ground plots and consecutive bins, constraints are added to ensure that each bin is assigned to only one stratum, and that each stratum has at least one bin. The objective function is either equation (5) or (6). The objective function is nonlinear because the denominator term, number of ground plots in a stratum, is an integer variable. Commercial software is not readily available to solve nonlinear integer programming problems, but is readily available to solve a shortest path problem.

Data

Three panels of annual inventory data from Pennsylvania were used to evaluate the procedures for determining the stratum weights for each county. From this information, estimates are calculated for the number of acres of nonforest and forest for each county and the State. The complete range of forested conditions is found in Pennsylvania, from heavily nonforested to heavily forested counties. Heavily forested or nonforested counties may require only one stratum, whereas counties with a mixture of forest conditions may require as many as seven strata.

Table 1 shows the possible stratifications from the shortest path algorithm for Mifflin County using the smoothed deviance objective function and three annual panels. The shortest path cost increases as the number of strata increases. The one-stratum solution starts at the first bin and stops at the last bin. The cost is 9,763 acres/plot. For the two-stratum solution, the first stratum has bins 1 through 25, and the second stratum has bin 26. The shortest path is the difference between the cost of 10,194 acres/plot and 9,417 acres/plot, or 776.7 acres/plot. The shortest path cost increases as the number of strata increases. This precludes building one network and allowing the algorithm to select the strata combination with the lowest cost. From the table, dividing the 26 bins into two groups of equal cost is easier than three groups of equal cost. Larger numbers of strata should have

Table 1.—*Possible stratifications for Mifflin County.*

Strata number	Start1	Stop1	Cost1	Start2	Stop2	Cost2	Start3	Stop3	Cost3	Start4	Stop4	Cost4	Shortest path
1	1	26	9,763										
2	1	25	9,417	26	26	10,194							776.76
3	1	1	5,116	2	24	11,204	25	26	11,005				12,174.29
4	1	1	5,116	2	6	6,471	7	22	12,244	23	26	12,149	27,059.84

lower sampling errors, however. From a mathematical point of view, stopping at bin 25 for the first of two strata makes sense, but is it wise from a biological perspective? Bins may have no or a sparse number of ground plots, and they are grouped with the strata that best balances the cost.

Because the shortest path cost directly increases as the number of strata increases, a single network cannot be built because the smaller strata solutions would prevail. Therefore, several heuristics were investigated for selecting the "optimal" number of strata. (Optimal is in quotation marks because the procedures are rules of thumb and not mathematically based procedures that develop necessary and sufficient conditions for optimality). One heuristic was to divide the shortest path by the number of difference pairs and graph the new cost versus the number of strata. The hypothesis was that the curve would decrease, reach a minimum, and then increase. The solution would be the number of strata at which the curve reached its minimum. The curves for each county did not follow the expected pattern, however; the cost per difference pair tended to increase with an increasing number of strata. A second heuristic ties the last bin in the first stratum and the first bin in the last stratum to NLCD imagery classification break points developed by Hoppus *et al*. (2001). Not all the counties could be classified with NLCD imagery break points because the requirements are too stringent or the county is essentially all forest or nonforest and the only appropriate stratification is one stratum. The final procedure is a series of relaxations on the NLCD imagery requirement. The procedure is as follows.

Imagery-Based Heuristic for Selecting the "Optimal" Number of Strata

Step 1. Create table 1 for each county. For each strata combination, calculate:

Cost range = largest cost of a strata – smallest cost of a strata

Next, sort by county and descending cost range. This sorting guarantees that the solution with the largest number of strata that meets the remaining criteria will be selected first.

Step 2. Separate the counties that can be stratified only by one stratum (group A) from the remaining counties. From the remaining counties, select the counties with a cost range of less than 6,000 acres/plot, a break point between the first and second stratum at bin 7 or lower, and the break point between the highest stratum and its adjacent lower stratum at bin 24 or higher. (These break points are the NLCD imagery classification break points.) From the counties that meet these criteria, select the solution with the largest number of strata (group B). Remove group B counties from the remaining data.

Step 3. Relax the standards on the remaining data (original data set minus groups A and B). From the remaining counties, select the counties with a cost range of less than 6,000 acres/plot, fewer than 12 bins in the first stratum, and the last stratum in bin 19 or higher. From the counties that meet these criteria, select the solution with the largest number of strata (group C). Remove group C counties from the remaining data.

Step 4. Relax the standards on the remaining data (original data set minus groups A, B, and C). From the remaining counties, select those counties with a cost range of less than 6,000 acres/plot. From the counties that meet these criteria, select the solution with the largest number of strata (group D). Remove group D counties from the remaining data.

Step 5. Select any remaining counties based on the smallest cost range. Place these counties in group E.

Step 6. Add groups A, B, C, D, and E to form the final solution.

Ratio Heuristic

Another heuristic is the ratio heuristic:

$$\text{Ratio}=\frac{\text{Total Land Area (acres) in State}}{\text{Total Number of Plots Measured}} \quad (7)$$

For the first panel, the ratio is approximately 30,000 acres/plot. For the second and third panel, the ratio is approximately 15,000 and 10,000 acres/plot, respectively. Exact numbers for the ratio heuristic could easily be calculated, but the approximations are used in this study.

To implement the ratio heuristic, apply Step 1 in the imagery-based heuristic. For Step 2, find the solution with the highest number of strata so that the cost range is less than the ratio in equation (7).

Results

Currently, a human expert performs the stratification for the annual inventory. From the number of plots in the county and their distribution in the 26 classes, the expert can estimate the number of strata. Using this information and a spreadsheet macro, the expert can visually examine the impact of different break points in the cost per stratum equation (4). The final solution is achieved when the expert believes that the cost cannot be further balanced among the strata. Table 2 shows the division of land between nonforest (0) and forest (1) for the three panels of annual inventory data in Pennsylvania.

The statistics in table 2 are calculated using FINSYS (Born and Barnard 1983), the computer program used by the Northeastern FIA unit to calculate sampling statistics. Because the data must be in a special format for FINSYS, "what if" questioning is difficult. As a result, the remaining statistics were calculated with a user-written SAS macro (SAS Institute 1999) and benchmarked against FINSYS. The results for the three-panel problem with a deviance objective function and a smoothed deviance objective function are shown in tables 3 and 4, respectively. According to the ratio rule, for three panels, the maximum cost range should be 10,000 acres/plot. For evaluation purposes, 6,000 acres/plot also was considered. The sampling

Table 2.—*Human expert's result for three-panel stratification in Pennsylvania.*

Forest land	Area (acres)	Mean area (%)	Sampling error (%)
0	12,030,500	41.9	1.2
1	16,652,100	58.1	0.9

Table 3.—*Three-panel stratification using optimization and deviance objective function.*

Procedure	Forest land	Area (acres)	Mean area (%)	Sampling error (%)
Imagery-based	0	11,767,610	41.03	1.179
	1	16,915,021	58.97	0.820
Cost range < 6,000 acres/plot	0	11,794,962	41.12	1.164
	1	16,887,670	58.88	0.813
Cost range < 10,000 acres/plot	0	11,769,360	41.03	1.172
	1	16,913,272	58.97	0.816

Table 4.—*Three-panel stratification using optimization and smoothed deviance objective function.*

Procedure	Forest land	Area (acres)	Mean area (%)	Sampling error (%)
Imagery-based	0	11,760,770	41.00	1.175
	1	16,921,862	59.00	0.817
Cost range < 6,000 acres/plot	0	11,794,956	41.12	1.169
	1	16,887,675	58.88	0.817
Cost range < 10,000 acres/plot	0	11,736,291	40.92	1.165

errors are lower for the optimization procedures. The mean area percentages differ by approximately 1 percent between the expert and the optimization. In the optimization groups, the imagery-based procedure had slightly higher sampling errors. For the optimization procedure, the lowest sampling errors were for the smoothed deviance objective function and the ratio decision rule using a cost range of less than 10,000 acres/plot. Consequently, in the remaining statistics, the imagery-based decision rule is not further investigated, and only the smoothed deviance objective function is investigated.

Table 5 shows the results for two-panel combinations. Panels 1 and 2 are shaded light gray, panels 1 and 3 are shaded dark gray, and panels 2 and 3 have a white background. For two panels, each plot is worth approximately 15,000 acres. A decision rule of cost range of less than 10,000 acres/plot is for comparison. The sampling errors are close. The cost range of less than 10,000 acres/plot decision rule has slightly lower sampling errors in two of the three cases. In the first and second panel combination, the lowest sampling error for the forest land is for the cost range of less than 10,000 acres/plot decision rule, and the lowest sampling error for nonforest is the cost range of less than 15,000 acres/plot decision rule.

Table 6 presents the results for the single panel. Panels 1 and 2 are shaded light gray, panels 1 and 3 are shaded dark gray, and panels 2 and 3 have a white background. The ratio decision rule would be to accept the stratification with the largest number

Table 5.—*Two-panel stratification using optimization and smoothed deviance objective function.*

Procedure	Forest land	Area (acres)	Mean area (%)	Sampling error (%)
Cost range < 10,000 acres/plot	0	11,824,972	41.23	1.457
	1	16,857,659	58.77	1.022
Cost range < 15,000 acres/plot	0	11,916,395	41.55	1.443
	1	16,766,236	58.45	1.025
Cost range < 10,000 acres/plot	0	11,824,972	41.23	1.457
	1	16,857,659	58.77	1.022
Cost range < 15,000 acres/plot	0	11,810,641	41.18	1.498
	1	16,871,991	58.82	1.048
Cost range < 10,000 acres/plot	0	11,633,254	40.56	1.457
	1	17,049,378	59.44	0.994
Cost range < 15,000 acres/plot	0	11,660,065	40.65	1.495
	1	17,024,023	59.35	1.024

Table 6.—*One-panel stratification using optimization and smoothed deviance objective function.*

Procedure	Forest land	Area (acres)	Mean area (%)	Sampling error (%)
Difference < 15,000 acres/plot	0	12,105,840	42.21	2.367
	1	16,576,792	57.79	1.728
Difference < 30,000 acres/plot	0	12,215,132	42.59	2.289
	1	16,467,500	57.79	1.698
Difference < 15,000 acres/plot	0	11,841,438	41.28	2.396
	1	16,841,194	58.72	1.685
Difference < 30,000 acres/plot	0	11,632,008	40.55	2.344
	1	17,050,624	49.45	1.599
Difference < 15,000 acres/plot	0	11,872,889	41.39	2.363
	1	16,809,742	58.61	1.669
Difference < 30,000 acres/plot	0	11,871,364	41.39	2.338
	1	16,811,268	58.61	1.651

of strata so that the cost range between the stratum with the largest and smallest cost is less than 30,000 acres/plot. These results are contrasted with those obtained from using 15,000 acres/plot. The sampling errors are larger with fewer panels. The sampling errors are lower for the difference of less than 30,000 acres/plot for all three panels.

Conclusions

The procedure described in this article is an automated approach to determining the stratification for a county, State, or other polygon. Using this procedure achieved lower sampling errors than with the current human expert procedure. By formulating the problem as a shortest path through the network, fast and efficient computational procedures are available to provide a real-time solution. Actual solution time depends on the number of arcs in the network. Arcs increase with the number of polygons to be simultaneously processed and the number of strata. Pruning of infeasible paths before optimization and a fast computer processor reduce the solution time. Different shortest path algorithms affect the solution speed.

From the shortest path formulation, the optimal one-, two-, three-, four-, five-, six-, and seven-strata solutions are found. To find the "optimal" solution, several heuristics were investigated. The ratio heuristic is easily implemented and provided the smallest sampling errors.

Literature Cited

Born, J.D.; Barnard, J.E. 1983. FINSYS-2: subsystem table-2 and output-2. Gen. Tech. Rep. NE-84. Broomall, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experimental Station. 68 p.

Derekenaris, G.; Garofalakis, J.; Makris, A.; *et al.* 2001. Integrating GIS, GPS and GSM technologies for the effective management of ambulances. Computers, Environment, and Urban Systems. 25: 267–278.

Dijkstra, E. 1959. A note on two problems in connection with graphs. Numerische Mathematik. 1: 269–271.

Hoppus, M.; Arner, S.; Lister, A. 2001. Stratifying FIA ground plots using a 3-year old MRLC forest cover map and current TM delivered variables selected by "decision tree" classification. In: Reams, G.A.; Mc Roberts, R.E.; Van Deusen, P.C., eds. Proceedings, 2nd annual forest inventory and analysis symposium; 2000 October 17–18; Salt Lake City, UT. Gen. Tech. Rep. SRS-47. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station: 19–24.

SAS Institute. 1999. SAS macro language: reference, version 8. Cary, NC: SAS Institute. 324 p.

Vogelmann, J.E.; Howard, S.M.; Yang, L.; *et al*. 2001. Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper data and ancillary data sources. Photogrammetric Engineering & Remote Sensing. 67(3): 650–661.