OPTIMUM STRATIFICATION FOR CLAM SURVEYS

Ronaldo Iachan and Christopher T. Gledhill, Iowa State University

Summary. The problem of finding stratum boundary points that minimize the variance of the survey estimates is reviewed. The practical case when stratification is on auxiliary variables is examined and applied to bottom trawl surveys of shellfish.

Key words: optimum stratification points, stratum boundaries, clam surveys.

1. Introduction

Given the number L of strata to be formed, the problem of finding optimum points of stratification (in the sense of minimum variance) is not difficult when the stratification variable is identical to the survey variable. It grows in complexity, however, in the more realistic situation when auxiliary variables only are available for stratification. Also, possibly conflicting objectives may arise in stratification in multipurpose surveys.

The population mean $\bar{Y} = \sum_h W_h\bar{Y}_h$ of the survey variable is unbiasedly estimated by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h\bar{y}_h ,$$

where $\bar{Y}_h(\bar{y}_h)$ is the h-th stratum (sample) mean, $W_h = N_h/N$ the h-th stratum weight and N_h its size, $N = \sum_h N_h$. If n_h units are selected at random (with replacement) in the h-th stratum and independently in different strata, the variance of the estimate $\bar{y}_{st}$ is then

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2S_h^2/n_h , \qquad (1.1)$$

where S_h^2 is the h-th stratum variance.

Let $n = \sum_{h=1}^{L} n_h$ be the total sample size. When the sample from each stratum is taken with proportional allocation, i.e., $n_h = n\cdot W_h$, (1.1) becomes

$$V_p(\bar{y}_{st}) = \sum_{h=1}^{L} W_hS_h^2/n . \qquad (1.2)$$

Minimizing (1.1) subject to a cost function of the form

$$C = C_0 + \sum_{h=1}^{L} C_hn_h$$

(C_0 the overhead cost and C_h the cost per unit in the h-th stratum) leads to n_h proportional to $W_hS_h/\sqrt{C_h}$. This is known as optimum allocation. The case $C_h = k$ (h = 1,...,L) gives the so-called Neyman allocation, i.e., $n_h = n\cdot W_hS_h/(\sum_{n=1}^{L} W_hS_h)$, under which the variance (1.1) is reduced to (approximately)

$$V_N(\bar{y}_{st}) = (\sum W_hS_h)^2/n . \qquad (1.3)$$

The problem of choosing stratum boundary points that minimize either variance (1.2) or (1.3) is briefly reviewed in Section 2. Further references are found in Kpedekpo (1973). In Section 3, the results reviewed are applied to stratification in bottom trawl surveys. Alternative stratifications on the auxiliary variables are compared, and the practical problem of mapping the strata is approached.

2. Optimum Stratification Points

2.1 Stratifying on the Estimation Variable

The problem of finding optimum boundary points was first approached by Dalenius (1950) in case the stratification variable is identical to the survey variable y, a < y < b. Minimizing (1.2) with respect to b_h (h = 1,...,L-1) gives

$$b_h = \frac{\bar{Y}_h + \bar{Y}_{h+1}}{2} , \quad (h = 1,\dots,L-1). \qquad (2.1)$$

For the case of Neyman allocation, the points $b_1,\dots,b_{L-1}$ that minimize (1.3) satisfy

$$\frac{(b_h - \bar{Y}_h)^2 + S_h^2}{S_h} = \frac{(b_h - \bar{Y}_{h+1})^2 + S_{h+1}^2}{S_{h+1}} ,$$

$$(h = 1,\dots,L-1). \qquad (2.2)$$

For convenience, let $b_0 = a$, $b_L = b$.

Solving (2.2) exactly is now feasible in general since $\bar{Y}_h$ and S_h^2 are now known. Mahalanobis (1952) suggests making

$$W_h\bar{Y}_h = \text{constant} \quad (h = 1,\dots,L) \qquad (2.3)$$

to obtain a "nearly optimum solution." Dalenius and Hodges (1957) propose approximations, for large L, to the solution points $b_1,\dots,b_{L-1}$, obtained by the so called "cum $\sqrt{f}$ " rule, where the population is now assumed infinite with density function f(y), a < y < b. The rule consists of making

$$\int_{b_{h-1}}^{b_h} \sqrt{f(t)} = \text{constant}, \quad (h = 1,\dots,L-1). \qquad (2.4)$$

Note that in this context,

$$W_h = \int_{b_{h-1}}^{b_h} f(t)\,dt ,$$

$$\bar{Y}_h = \frac{1}{W_h}\int_{b_{h-1}}^{b_h} t\,f(t)\,dt ,$$

$$S_h^2 = \frac{1}{W_h}\int_{b_{h-1}}^{b_h} t^2\,f(t)\,dt - \bar{Y}_h^2 .$$

Another approximation useful in univariate stratification is provided by Ekman (1959). Under suitable conditions on the density f(y), a < y < b, the b_h's that satisfy

$$(b_h - b_{h-1})W_h = C_n, \quad (h = 1,\dots,L) \qquad (2.5)$$

are shown to approximately attain optimum stratification with Neyman allocation. Here, C_n is a constant that depends on the sample size n.

Cochran (1961) compares empirically the rules due to Mahalanobis, Dalenius and Hodges, and Ekman. An additional method is suggested by Sethi (1963) who finds the optimum points of stratification for the normal and a set of chi-square distributions using proportional, equal, and Neyman allocations. The optimum points differ little under equal and Neyman allocations, being well approximated by the cum $\sqrt{f}$ rule.

2.2 Stratifying on the Auxiliary Variable

Dalenius and Gurney (1952), assume that the estimation variable y and the stratification variable x are related by

$$y = \psi(x) + \varepsilon \qquad (2.6)$$

where $E(\varepsilon) = 0$, $E(\varepsilon^2) = S_\varepsilon^2$, ε and $\psi(x)$ being uncorrelated for all x. If $\psi(x) = \alpha + \beta x$ then, under proportional allocation, the optimum set of points $B_x = \{b_{x,1},\ldots,b_{x,L-1}\}$ consists of

$$b_{y,h} = \alpha + \beta c_{x,h}, \quad (h = 1,\ldots,L-1) \qquad (2.7)$$

$c_{x,h} = \frac{1}{2}(\bar{X}_h + \bar{X}_{h+1})$ being the optimum points for x. Under Neyman allocation, the optimum stratification points satisfy

$$\frac{\beta^2(b_{x,h} - \bar{X}_h)^2 + (\beta^2 S_{x,h}^2 + 2S_\varepsilon^2)}{\sqrt{\beta^2 S_{x,h}^2 + S_\varepsilon^2}} = \frac{\beta^2(b_{x,h+1} - \bar{X}_{h+1})^2 + (\beta^2 S_{x,h+1}^2 + 2S_\varepsilon^2)}{\sqrt{\beta^2 S_{x,h+1}^2 + S_\varepsilon^2}}$$

$$(h = 1,\ldots,L-1) \qquad (2.8)$$

where $\bar{X}_h$ and $S_{x,h}^2$ denote the mean and the variance, respectively, for the variable x in stratum h. Note that these are approximately the optimum points for x if S_ε is small compared to $|\beta|S_{x,h}$, all h, or equivalently, if the correlation between x and y is high (in absolute value) within every stratum. Cochran (1977, p. 131) remarks that failure to use the optimum points for x should not be very harmful, however, if this correlation is only moderate.

Taga (1967) demonstrates the existence (under proportional allocation) of an optimum stratification for the estimation variable y based on the auxiliary variable x, and exhibits a method by which it can be obtained asymptotically. We shall be concerned with approximations to this optimum in our empirical study. Previous empirical investigation by Hess et al. (1966) had a high correlation ($\rho_{x,y} \doteq .9$) present here. Some approaches to overcome this difficulty shall be suggested.

3. Empirical Study

3.1 Population and Sampling Design

The National Marine Fisheries Service (NMFS) has designed a sampling scheme for shellfish (surf clam and ocean quahog) along the East Coast from Cape Hatteras to Nova Scotia, covering a sea surface of approximately 75,000 square nautical miles. The whole area has been divided into depth zones ranging from 5 to 200 fathoms (9 to 365 meters) and each zone is divided into strata according to geographical location. The strata may be combined to represent regions, which are studied separately.

Each stratum is first divided into areas of 5' latitude by 10' longitude. These areas are further divided into 10 units 2½' latitude by 2' longitude. Let N_h be the number of such units in the h-th stratum and $N = \sum_h N_h$ the total over the region. A predetermined number of sampling units, called stations (or tow locations), is selected at random within each stratum before each cruise. Presently, stratum weights are based on their areas and allocation is (roughly) proportionate, so that we may assume $n_h \propto W_h = N_h/N$ is the number of sampling units in the sample from stratum h. Proportional allocation is not only convenient but also an efficient compromise solution in multipurpose surveys. As the two survey variates, namely relative abundance of clams and quahogs, have quite distinct distributions in the region, we shall confine our study to proportional allocation. The stratifying variable x (depth), however, is efficient for both survey variates: clams are more abundant closer to shore (up to 40 meters deep) while ocean quahogs are found in water deeper than 30 meters.

The NMFS would like to improve the estimates of clam relative abundance in terms of both quantity and weight. Also, it is desirable to obtain an approximately optimum stratification for the fishing catch based on depth and geography (latitude and longitude). The purpose of this study is to compare various re-stratifications of each region, making use of the 1978-80 clam survey data from the original strata briefly described in Table 1.

3.2 Approximately Optimum Strata

Consider the (geographical) coordinates u (latitude) and v (longitude) and suppose optimum stratification has been performed on depth $x = g(u,v)$. If the optimum strata $P_i = (b_{i-1}, b_i)$ are obtained, $i = 1,\ldots,L$, the question of whether the stratification $g^{-1}(P_i)$ is optimum in the plane, or at least nearly so, then naturally arises.

Ignoring the second order term in a Taylor series expansion of $g\colon R^P \to R$ with continuous second derivatives, we may write

$$g(\underline{s}) \doteq g(\underline{a}) + g'(\underline{a})\cdot(\underline{s} - \underline{a}) \qquad (3.1)$$

and hence

$$V[g(\hat{\theta})] = E[g(\hat{\theta}) - g(\theta)]^2 \doteq [g'(\hat{\theta})]^2 \cdot V(\theta).$$

Therefore, minimization of $V[g(\hat{\theta})]$ is equivalent to that of $V(\hat{\theta})$ iff g is linear. In particular, an optimal stratification for $x = g(u,v)$ does <u>not</u> automatically carry over to the plane coordinates. In view of (3.1), however, the associated stratification in the plane should be nearly optimal if g is smooth.

In our example, depth $x = g(u,v)$ is a smooth function of latitude (u) and longitude (v). The (least square) linear fit

$$x \doteq 2282.68 - 12.85u - 23.61v \qquad (3.2)$$

has $R^2 = .54$, all parameters being significantly nonzero (via t-tests). A reasonable approximation to an optimum stratification may thus be based on depth contours, namely

$$\{(u,v):\ b_{i-1} < g(u,v) \leq b_i\} \qquad (3.3)$$

It is argued in Section 2.2 that, if x and y are at least moderately correlated then approximately optimal stratifications on x (depth) should be also nearly optimal for the survey variable (catch) $y = \psi(x)$, and more nearly so if ψ is linear (cf. 2.7)). Thus, the addition of variables u,v to the regression $y = \psi(x)$ should yield no significant improvement over the use of x alone.

Under proportional allocation, the optimum boundary points for stratification on x can be found with an iterative procedure suggested by Dalenius (1950). We start with arbitrary $b'_1,\ldots,b'_{L-1}$. If

$$\frac{1}{2}\,(\bar{x}'_h + \bar{x}'_{h+1}) < b'_h \qquad (h = 1,\ldots,L-1)$$

then a smaller value of b'_h should be used at the next iteration and conversely, until (2.1) is approximately achieved. Using the frequency distribution of the variable x, this procedure leads to the boundary points given in Table 2 for the separate regions. An approximately optimal stratification may be thus obtained by substituting the b_i $(i = 1,\ldots,6)$ levels into contours (3.3) or into a linearization such as (3.2).

3.3 Alternative Stratifications

The approximately optimum strata given in the last paragraph are not satisfactory to the extent that the linear fit $y = \psi_1(x)$ is far from perfect. An investigation into other functional relationships $y = \psi(x)$ may lead to more efficient stratifications by increasing the correlation between Y and $X' = \psi^{-1}(Y)$, say. We shall examine how the boundary points are affected by the choice both of a rule and of a function ψ. In particular, fits of $\psi_2(x) = ae^{-bx}$ and of $\psi_3(x) = ax^{-b}$ have been performed:

$$\begin{aligned} \psi_2(x) &= 125.2e^{-.086x} \quad (R^2 \doteq .30) \\ \psi_3(x) &= 17{,}500x^{-2.25} \quad (R^2 \doteq .20) \end{aligned} \qquad (3.4)$$

The use of transformations to obtain alternative stratifications is illustrated in Table 3 for the two functions in (3.4) as well as a linear function (in the last three columns), combined with different rules. The number of strata has been reduced to the L = 5 in Northern New Jersey to simplify the computation and presentation of the results. In order to apply Sethi's rule, a chi-square distribution χ^2_r was fitted, with r = 1 d.f. for λy and r = 30 d.f. for γx, r, γ and λ being determined from the first two moments ($\lambda = 37$, $\gamma = .94$). An iterative procedure for rule (2.1) was introduced in Section 3.2. An algorithm to find the stratification points according to the cum $\sqrt{f}$, Mahalanobis' and Ekman's rules simultaneously is available from the senior author.

3.4 Comparison of Stratification Methods and Discussion

Under proportional allocation, in view of (1.2), a comparison of stratification methods reduces to one between the value of $\sum_h W_h S_h^2$. The weights W_h for the strata defined in Table 3 can be computed from the distribution of the variates Y and X, and S_h^2 can be found from that of Y alone, available for our known population. We illustrate by comparing the variances for the two simple stratifications given in the first two columns of Table 3. Such variances are presented in Table 5, whereas other characteristics of the formed strata are given in Table 4. A full comparison of the current and optimal stratifications is provided in Table 6.

When stratification is carried out on the estimation variate Y itself, Mahalanobis' method assigns heavy weight to the lower stratum which exhibits a very low variability. Therefore, this method is highly efficient. The two middle methods, however, present a large contribution of the higher, variable stratum, and hence perform very poorly.

In the more realistic situation where X is the stratifying variate, all methods show approximately the same (low) efficiency, with a slight advantage to the cum $\sqrt{f}$ rule. Still, all four stratification methods result (with proportional allocation) substantially more precise than the one currently used. For instance, making use of the same 1981 NMFS data from Northern New Jersey, stratum variances (S_h^2) were estimated for the five study strata, namely strata 21, 25, 88, 89 and 90. The first two sample variances equal 36,051 and 357, respectively, and the remaining three are zero, so that with sample allocation 18, 9, 10, 10 and 2 units, respectively, we estimate $nV_p(\bar{y}_{st})$ by $\frac{18}{49}(36{,}051) + \frac{9}{49}(357) \doteq 13{,}309$. This is much larger than the values in Table 5.

Our findings suggest that an efficient stratification can be based on contours of depth, in case the sea is approximately planar. As expected, the strata are strips that are approximately parallel to the coast, i.e., normal to the gradient of the survey variate (relative abundance). Two other ways of finding the contours should be pointed out. The first one is using the exact ocean depth contours already mapped by the NMFS. In this case, the strata would not necessarily be strips. The second one is stratifying on the variable catch at a previous survey, and periodic redefinition of strata.

References

COCHRAN, W. G. (1961). "Comparison of Methods for Determining Stratum Boundaries." Bull. Int. Statist. Inst., 38, 345-358.

COCHRAN, W. G. (1977). Sampling Techniques. J. Wiley and Sons, New York.

DALENIUS, T. (1950). "The Problem of Optimum Stratification." Skand. Akt. Tidskrift, 203-213.

DALENIUS, T. and GURNEY, M. (1951). "The Problem of Optimum Stratification." Skand. Akt. Tidskrift, 133-148.

DALENIUS, T. and HODGES, J. L., Jr. (1957). "The Choice of Stratification Points." Skand. Akt. Tidskrift, 198-203.

EKMAN, G. (1959). "An Approximation Useful in Univariate Stratification." Ann. Math. Statist., 30, 219-229.

HESS, I., SETHI, V. K. and BALAKRISHNAN, T. R. (1966). "Stratification: A Practical Investigation." Jour. Amer. Statist. Assoc., 61, 74-90.

Acknowledgments: We are grateful to H. T. David at Iowa State University for his helpful and insightful suggestions. The cooperation of the NMFS staff is also gladly acknowledged.

KPEDEKPO, G. M. K. (1973). "Recent Advances on Some Aspects of Stratified Sample Design. A Review of the Literature." Metrika, 20, 54-64.

MAHALANOBIS, P. C. (1952). "Some Aspects of the Design of Sample Surveys." Sankhya, 12, Part 1 & 2, 1-7.

SETHI, V. K. (1963). "A Note on Optimum Stratification of Populations for Estimating the Population Means." Aust. J. Statist., 5, 20-33.

TAGA, Y. (1967). "On Optimum Stratification for the Objective Variable Based on Concomitant Variables Using Prior Information." Ann. Inst. Stat. Math., 19, 101-129.

Table 1. List of strata and the number of sample units in each stratum by management area †

NNJ		SNJ		Delmarva		VA-NC	
Str #	#units	Str #	#units	Str #	#units	Str #	#units
21	234	17	62	9	278	1	201
22	228	18	52	10	111	2	161
23	159	19	81	11	66	3	107
24	313	20	154	12	167	4	89
25	43	87	64	13	127	5	47
26	70			14	84	6	22
27	112			15	98	7	14
28	158			16	200	8	12
88	55			82	27	80	48
89	35			83	44	81	81
90	14			84	77		
				85	65		
				86	52		

† NNJ = Northern New Jersey, SNJ = Southern New Jersey, Delmarva = Delaware, Maryland and Virginia, VA - NC = Virginia and North Carolina.

Table 2. Stratum boundaries (b_i) formed by the cum $\sqrt{f}$ rule (A) and Dalenius rule (B) by management area. †

Boundary point	MANAGEMENT AREAS							
	NNJ		SNJ		Delmarva		VA-NC	
	A	B	A	B	A	B	A	B
b_0	9	9	9	9	9	9	9	9
b_1	21	21	15	15	17	17	17	17
b_2	30	30	24	25	27	28	27	28
b_3	39	39	38	38	36	37	36	37
b_4	54	54	48	50	50	50	50	50
b_L	110	110	110	110	110	110	110	110

† NNJ = Northern New Jersey, SNJ = Southern New Jersey, Delmarva = Delware, Maryland and Virginia, VA-NC = Virginia, North Carolina management areas (L = 5, stratification variable is depth in meters).

Table 3. Boundary Points (L=5)

Variable Rule	Y (catch)	X (depth)	ψ_1^{-1} (Y)	ψ_2^{-1} (Y)	ψ_3^{-1} (Y)
(a) Optimum	6.5	21.5	39.4	34.1	33.5
	23.6	30.0	33.8	19.3	18.9
	79.2	37.4	15.6	5.3	11.0
	226.2	44.0	0*	0*	6.9
(b) CUM $\sqrt{f}$	3.0	20.2	40.5	43.0	47.2
	10.2	25.3	38.2	28.9	27.4
	23.3	32.6	33.9	19.4	19.0
	81.0	37.9	15.1	5.0	10.9
(c) Sethi's	4.8	22.0	39.9	37.6	38.3
	11.4	28.0	37.8	27.6	26.1
	21.6	33.0	34.4	20.3	19.6
	36.0	40.0	29.7	14.4	15.6
(d) Mahalanobis	39.6	22.7	28.5	13.3	15.0
	114.5	28.8	4.12	1.0	9.4
	117.4	35.0	0*	0*	7.7
	326.0	40.0	0*	0*	5.9

*Zero values stand for negative transformed values.

Table 4. Stratification comparison - stratum sizes and stratum variances

On Y

Stratum Rule		h=1	h=2	h=3	h=4	h=5
(a)	n_h	27	20	9	7	3
	S_h^2	2.87	14.89	134.62	588.92	2,811.55
(b)	n_h	23	13	11	9	10
	S_h^2	1.10	5.31	9.60	134.62	10,244.96
(c)	n_h	26	12	9	6	13
	S_h^2	2.00	2.97	7.78	132.92	11,099.21
(d)	n_h	53	6	4	1	2
	S_h^2	94.13	775.14	513.00	0.00	36.00

On X

Stratum Rule		h=1	h=2	h=3	h=4	h=5
(a)	n_h	15	22	15	10	4
	S_h^2	14,517	3,555	532	32	0.
(b)	n_h	14	14	10	14	14
	S_h^2	15,199	5,105	172	570	28
(c)	n_h	17	16	9	15	9
	S_h^2	13,634	4,555	198	543	32
(d)	n_h	17	16	14	10	9
	S_h^2	13,634	4,555	600	12	32

Table 5. Comparison of different stratifications - approximate (normalized) sampling variances $n\ V_p(\bar{y}_{st}) = \sum_h W_h s_h^2$

Stratification Variable Rule	Y (catch)	X (depth)
(a) optimum	200.67	4,610.
(b) cum $\sqrt{f}$	1,573.65	4,470.
(c) Sethi's	2,200.68	4,771.
(d) Mahalanobis'	178.24	4,749.

Table 6. Stratum means and variances under present and optimal (Dalenius) stratification for clam cruises 278 (1978) and 080 (1980).

		CRUISE 278				CRUISE 080				
		STRATIFICATION				STRATIFICATION				
		present		optimal			present		optimal	
AREA	(n)	$\bar{y}$	$V(\bar{y})$	$\bar{y}$	$V(\bar{y})$	(n)	$\bar{y}$	$V(\bar{y})$	$\bar{y}$	$V(\bar{y})$
				SURF CLAMS						
NNJ	(90)	0.89	0.0540	0.51	0.0212	(62)	26.60	34.8537	30.70	49.5650
SNJ	(36)	6.74	2.4939	3.09	0.4745	(19)	51.66	1640.5300	68.00	3547.3400
Del.	(73)	6.42	1.8787	4.61	0.9431	(68)	68.56	1217.3008	67.99	1626.8900
VA-NC	(42)	2.04	0.3709	1.95	0.6512	(15)	28.56	778.7810	16.49	268.7900