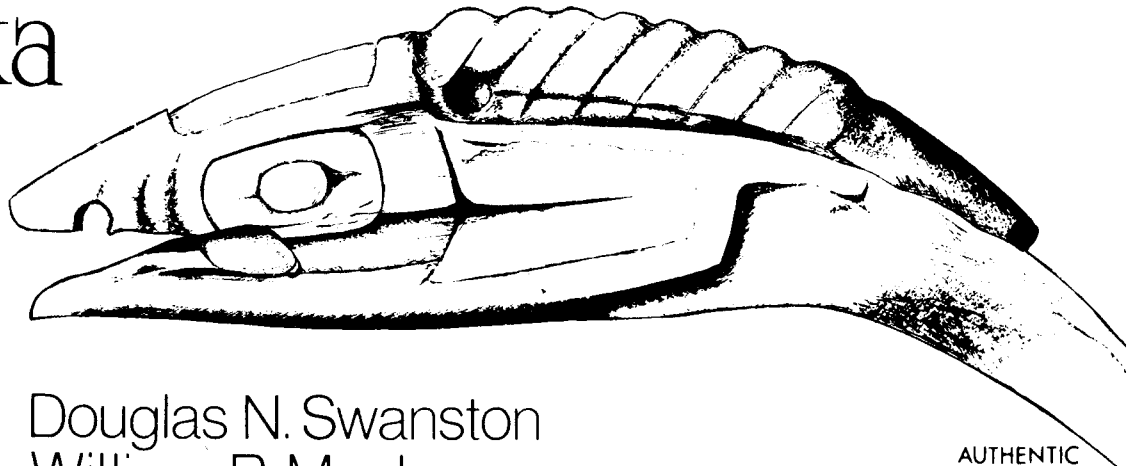
USDA FOREST SERVICE RESEARCH PAPER PNW-227
1977

# A Quantitative Geomorphic Approach to Predicting Productivity of Pink and Chum Salmon Streams in Southeast Alaska



AUTHENTIC
TLINGIT INDIAN
FISH CHARM

Douglas N. Swanston
William R. Meehan
James A. McNutt

PACIFIC NORTHWEST FOREST AND RANGE EXPERIMENT STATION

## Introduction

Salmon streams in southeast Alaska vary in amount of salmon produced per usable area. In many cases, the factors that limit production in some streams are readily discernible: overfishing, barriers to fish passage (falls, logjams), unfavorable streambed conditions (sedimentation, unstable gravel beds), excessive gradient, and highly variable streamflow. In some cases, however, streams that appear to have favorable conditions for salmon production do not support runs in proportion to their apparent potential. Questions then arise: Do geomorphic factors regulate production of salmon in these streams? Are these factors, or groups of factors, discernible directly as gross watershed characteristics, or are they more subtly related to local variations in the physical, chemical, or biological makeup of individual streams? Can the productivity of a particular stream be predicted from simple measurements of watershed characteristics on aerial photographs and topographic maps, or is a detailed investigation of streams and monitoring of fish populations necessary? Answers to these questions will have considerable bearing on the management of anadromous fisheries as the potential for damage from man's activities increases and as emphasis changes from maintenance of natural stocks of salmon to increased quality and productivity of watersheds.

## Factor Identification

Thompson and Hunt (1930) stressed the importance of the drainage basin as a whole, not just the stream, in their investigations of the basic nature of stream productivity. Slack (1955) reinforced this concept in his studies of stream productivity factors, demonstrating that the biological productivity of a stream is directly related to the physical environment of the watershed, which controls drainage pattern, flow rates, gravel size and shape, channel gradient, and general stability characteristics. Statistical analysis of quantitative geomorphic parameters of individual watersheds can help identify these factors.

Quantitative geomorphic techniques developed by Horton (1932, 1945) and Strahler (1952, 1953, 1954) provide a convenient method for obtaining numerical data on gross basin characteristics, given limited funding and difficulty of access and sampling of test streams. Measurement of physical parameters based on basin and channel geometry, obtainable from aerial photographs and topographic maps, provides correlation units such as drainage size and shape, stream density, and pattern, number, and length of tributaries. These in turn provide an estimate of stage of watershed development, probable basin discharge, extent of bedrock control of drainage,

impact of unstable slopes, and extent of channel suitable for spawning. Such techniques have been used successfully to analyze relationships between erosion, climate, surface properties, and geomorphology (Melton 1957, Maxwell 1960, Dissmeyer 1967). In 1973, Ziemer used quantitative geomorphic techniques to relate drainage basin and channel configuration to changes in production of pink salmon on Montague Island, Prince William Sound, Alaska, after large vertical tectonic adjustments resulting from the "Good Friday Earthquake" of 1964. Using five drainage system factors, he showed a correlation between drainage system geometry and freshwater production factors for pink salmon, with escapement as his indicator of production. He assumed that (1) the number of spawners using a stream is a sound measure of fish production in that stream, (2) escapement counts were consistently made from year to year and stream to stream, and (3) the impact of the fishery was consistent between stocks and years. He realized the problems involved by making these assumptions, but he had no other tools available.

Several quantitative methods were considered in an attempt to assign numbers to various degrees of salmon production. Enough time could not be spent on each stream to obtain even rough estimates of the standing crop of juveniles, egg or preemergent fry densities, or some other biological measure of production. Escapement counts are available for most southeast Alaska salmon streams for many years back. These counts are summarized in a set of catalogs that describe the physical characteristics of streams as well as the number of salmon that have returned to their spawning grounds. (Catalogs can be seen at the Alaska Department of Fish and Game, Juneau). However, salmon escapement data are not necessarily a reliable index of production of a given stream. Escapement is only one portion of the total run returning to a stream--the portion that has survived the onslaught of the fishery and has successfully completed the upstream migration to the spawning grounds. Intensity of the fishing effort as well as success of fishing is not necessarily the same for different streams. Consequently, the total return (catch plus escapement) to two streams may be similar. If fishing mortality, however, has accounted for two-thirds of the total return to one stream and one-third of the total return to another stream, escapement to the second stream may be twice as great. Other factors also may produce differential survival between stocks of fish. The ocean feeding area of one population may promote better growth and/or survival than another. The migration routes of one stock may subject that run to greater predation than the route or timing of another run. Aerial and ground surveys of escapements are often conducted at different stages of a given run in different years, by various observers, under different light conditions, etc. The main point is that escapement to a stream, although it may help in *qualitatively* describing the general level of production of the stream, does not necessarily indicate the biomass of salmon

that were or could be produced in that stream. A better *quantitative* indication of a stream's fish production would be the average number of smolts (seaward migrant juveniles) produced by a known number of spawning females over several years. Obtaining this type of information for many streams is costly, time consuming and, as a result, generally not done.

For this study, streams were categorized as either good producers or poor producers of pink and chum salmon. These categories were based on interviews and correspondence with district fishery management biologists throughout southeast Alaska and on the many years of escapement data (aerial and ground surveys by several agencies) summarized in the stream catalogs. Production was not based on escapement figures alone. Streams which were thought to be fair producers were not selected, so that only *very good* production and *very poor* production of pink and chum salmon were considered. Poor producers were further defined as streams to which no known causes for poor runs could be attributed; that is, they were streams which were accessible to migrating fish throughout most of their length (not blocked by falls, logjams, etc.), they appeared to have sufficient high quality water and gravels, they were not regularly "robbed" by illegal fishing, and they had not historically supported good pink and chum salmon runs. This subjective selection of good producers and poor producers may be criticized as not being statistically valid since the streams were not a randomly selected sample of all the available streams in southeast Alaska--one person's idea of "good" or "poor" may differ greatly from that of the next. However, we felt that this type of selection was justified since we specified only very good or very poor streams and since this is the type of selection process that may be necessary for the resource manager to use when he does not have the time and funds to obtain more quantitative estimates of production.

## Data Accumulation

A total of 78 watersheds were categorized as either exceptionally good producers or exceptionally poor producers based on the preceding criteria. These watersheds were scattered throughout southeast Alaska. They ranged in size from a minimum of 5.2 $km^2$ to a maximum of 422.2 $km^2$ (fig. 1); 22 were classified as poor and 56 as good.

To identify general similarities or differences between good producers and poor producers, we selected 21 independent variables for interbasin correlation purposes. These variables are listed in table 1. Of these, 19 were continuous--that is, they appeared at varying levels in every basin and could be simply measured on aerial photographs or 15-minute quadrangle maps. The other two were discontinuous; they classed each watershed according to whether it was underlain predominantly by igneous bedrock or metasedimentary bedrock. Of the 19 continuous variables, 14 ($X_1$-$X_2$; $X_5$-$X_8$; $X_{11}$-$X_{16}$; $X_{18}$-$X_{19}$) were

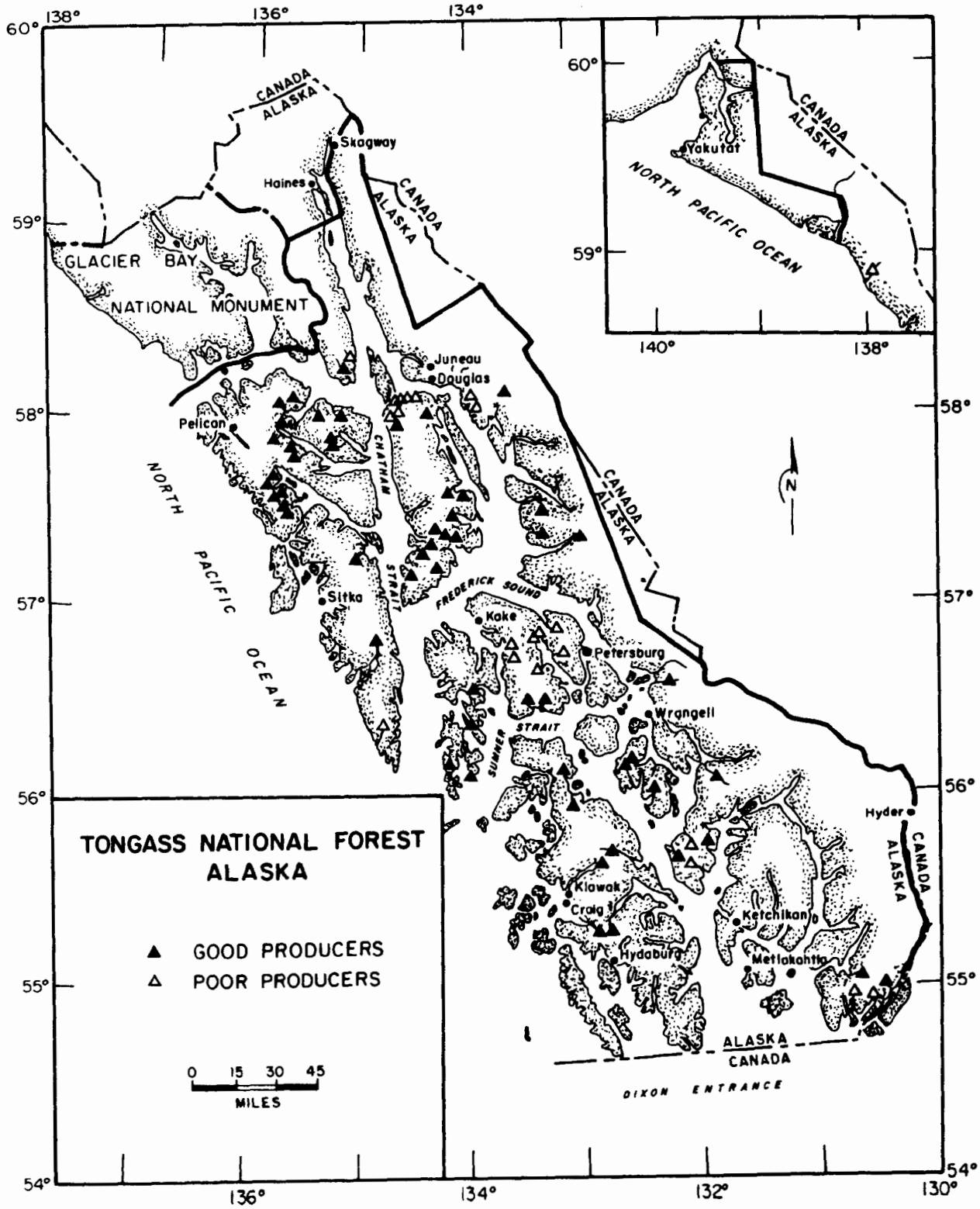



*Figure 1.--Map showing location of sample watersheds.*

Table 1--*List of quantitative geomorphic variables used for interbasin correlation purposes*

| Symbol | Variable | Unit of measurement |
|---|---|---|
| $X_1$ | Area of drainage basin | $km^2$ |
| $X_2$ | Mean valley side slope | degrees |
| $X_3$ | Basin area with slope above critical angle (34°) | percent |
| $X_4$ | Avalanche index (number of avalanches in watershed) | number |
| $X_5$ | Drainage density | dimensionless |
| $X_6$ | Bifurcation ratio | dimensionless |
| $X_7$ | Total length of channel segments | km |
| $X_8$ | Gradient of stream channel | degrees |
| $X_9$ | Length of stream with acceptable spawning gradient (<12 percent) | km |
| $X_{10}$ | Obstructions in main channel | number |
| $X_{11}$ | Basin perimeter | km |
| $X_{12}$ | Basin relief | m |
| $X_{13}$ | Channel frequency | dimensionless |
| $X_{14}$ | Relative relief | dimensionless |
| $X_{15}$ | Compactness coefficient | dimensionless |
| $X_{16}$ | Form factor | dimensionless |
| $X_{17}$ | Lakes in stream system | number |
| $X_{18}$ | Length ratio | dimensionless |
| $X_{19}$ | Basin orientation | degrees |
| $X_{20}$ | Predominatly sedimentary/ metamorphic rock (>50 percent) | 1 |
| $X_{21}$ | Predominantly igneous rock (>50 percent) | 2 |

standard quantitative geomorphic variables that provide a measure of basin or channel geometry, discharge characteristics, or stage of watershed development. Five ($X_3$, $X_4$, $X_9$, $X_{10}$, $X_{17}$) were developed for this study to provide a direct measure of basin stability and productivity.

## Analysis

The purposes of the data analysis were: (1) to test each variable for significance in differentiating between good producers and poor producers and (2) to construct a discriminant model that would give a decisionmaker or researcher the opportunity to classify salmon streams as either *very poor* producers or *very good* producers. Analysis dictated a search for a "best possible" model (where "best" is determined by trade-offs between statistical accuracy, data collection feasibility, and model application costs) that would provide more potential gain than cost to the user and that could be applied to any watershed for classifying it as a good producer or poor producer. Figure 2 shows such a model. Note that this approach requires equal dispersions for both groups but does not require equal sample sizes (Cooley and Lohnes 1971).

Data analysis was handled in four stages. Stage 1 involved computer-assisted evaluation of each explanatory variable property. This evaluation required examination of sample statistics, histograms, scatter diagrams, and correlation coefficients for each variable. Stage 2 centered on basic linear regression analysis of the 21 independent variables on salmon productivity. Stage 3 involved development of a family of discriminant functions from which a subset could be selected for further evaluation. Stage 4 included comparative analysis of selected key discriminant models. All analytical work was accomplished with the aid of special data processing facilities and software available through Oregon State University, Corvallis.

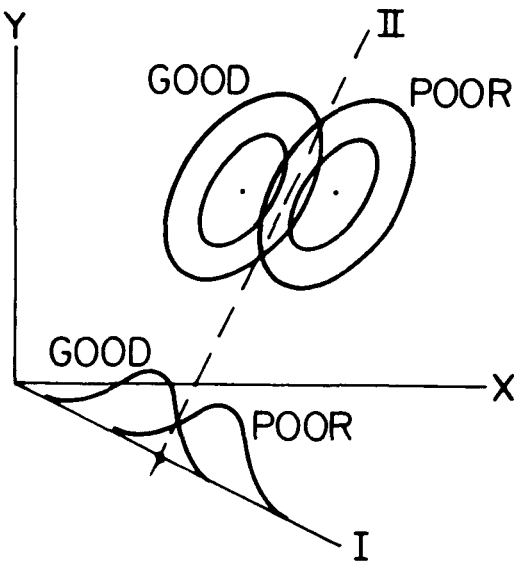



*Figure 2.--Geometric interpretation of discriminant analysis (after Cooley and Lohnes 1971, used with permission): (a) problem for two groups and two variates ($\underline{x}$ and $\underline{y}$); (b) line I $\perp$ line II; (c) elipse sets called centours for centile contours; (d) overlap of group good or poor is smaller than for any other set; (e) picture outer elipses as including 90 percent of each group and inner elipses 75 percent of each group.*

Stage 1 provided familiarity with the explanatory variables and all relevant interrelationships. Stage 2 provided a tool for evaluating possible model structures and behavioral characteristics. The basic approach was a "modified-backstep" regression analysis. All explanatory variables were regressed on a salmon productivity dummy variable. The least significant variables were dropped one at a time in each "backstep." At each juncture, the t-values of previously dropped variables were scanned. Any dropped variable which had a t-value that climbed back to a value of ±2.0 was reentered into the specified model. This approach allows for development of more significant models than does "stepwise regression" (Draper and Smith 1968). The reason for using regression modeling prior to using discriminant modeling is that a two-class linear, discriminant function is algebraically equivalent to a regression model. Model stability, structure, and order of variable importance (significance) are more easily examined and evaluated in a regression model than in a discriminant model; for example, for presence and impact of multicolinearity.

Stage 2 resulted in several significant models and a battery of test results that examined the reliability of six model assumptions (Kmenta 1971): (1) Error term is normally distributed, (2) expected value of the error term is zero, (3) variance of the error term is a constant, (4) error terms are not correlated in time and/or space, (5) each explanatory variable is nonstochastic, and (6) no explanatory variable has an exact linear relationship with any other explanatory variable. Examination of histograms and selected scatter diagrams of the residuals as well as of covariance matrices and correlation matrices did not indicate that any of these assumptions was violated significantly for the models considered (Draper and Smith 1968, Kmenta 1971).

Stage 3 produced a variety of discriminant models for later evaluation. The approach used in forming these models was the "modified-backstep" procedure that simply began with a fully specified model (all 21 explanatory variables) and dropped variables, one by one, in the same order as determined for the regression modeling procedure. For practical purposes, the largest model considered was a 12-variable model which included all explanatory variables with regression values so that: $-1.0 \geq t \geq +1.0$.[1/] The smallest model considered was a 5-variable model with values for all inclusive variables: $-2.0 \geq t \geq +2.0$. In all models considered, each discriminant function produced means of the good groups and poor groups that were significantly different at the $\alpha = 0.025$ level. Models with 8 or fewer variables produced significantly different means at the $\alpha = 0.01$ level (largest characteristic root test, Morrison 1967).

---

[1/] Note: For the t-test for sample sizes of n = 20, anything at or near 2.00 is significant at $\alpha = 0.05$. With n = 78, the t = 2.00 is significant at $\alpha = 0.025$, whereas the $\alpha = 0.05$ has a t-value = 1.67 and the $\alpha = 0.10$ one = 1.29 (Brownlee 1965).

Stage 4 was the most thorough and rewarding stage. Three key models were identified immediately for detailed comparative analysis. They were: 5-, 8-, and 12-variable models with the general characteristics shown in table 2.

Any process providing for final model selection is necessarily arbitrary and subjective. Because the explanatory variables had inclusive t-values near the $\alpha = 0.05$ level and the function discriminated between good and poor means at the $\alpha = 0.01$ level, the eight-variable model was chosen for reporting analysis. The extra costs of collecting data and manipulating a larger equation were also considered part of the trade-off in model selection. We accepted a middle ground between data costs and statistical accuracy.

The selected 8-variable discriminant function is:

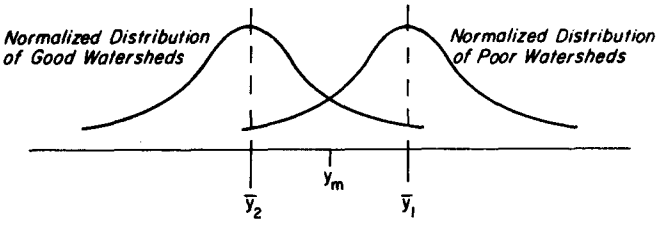
$$f(x) = -0.002787\ X_1 - 0.602159\ X_3 - 0.019247\ X_4 - 0.029875\ X_6 + 0.002146\ X_7 + 0.000332\ X_{12} - 0.023515\ X_{18} + 0.000660\ X_{19}.$$

For discussion purposes, let $f_1(x) = y_1$ express functional output for the poor group and $f_2(x) = y_2$ for the good group. Figure 3 underscores the generalized multivariate normality (multinormal) concept assumed related to this application of the discriminant analysis; it illustrates a tailored version of the discrimination provided by this model between the good groups and poor groups.

Table 2--*Discriminant model characteristics*

| Model | Explanatory variables | Significant delineating characteristics (s) |
|---|---|---|
| 5-variable[1/] | $X_1$, $X_3$, $X_4$, $X_{12}$, $X_{19}$ | Means significant @ $\alpha = 0.01$<br>All inclusive t-values:<br>$-2.0 \geq t \geq +2.0$ |
| 8-variable | $X_1$, $X_3$, $X_4$, $X_6$, $X_7$, $X_{12}$, $X_{18}$, $X_{19}$ | Means significant @ $\alpha = 0.01$<br>All inclusive t-values:<br>$-1.5 \geq t \geq +2.0$ |
| 12-variable | $X_1$, $X_2$, $X_3$, $X_4$, $X_6$, $X_7$, $X_{11}$, $X_{12}$, $X_{14}$, $X_{16}$, $X_{18}$, $X_{19}$ | Means significant @ $\alpha = 0.025$<br>All inclusive t-values:<br>$-1.0 \geq t \geq +1.0$ |

[1/] Smaller models were not considered due to very high degree of significance of removing variables beyond the five listed.




$\bar{y}_1 = 0.152$

$y_m = (\bar{y}_1 - \bar{y}_2) / 2 = 0.070$

$\bar{y}_2 = -0.012$

$\sigma_1^2 = \sigma_2^2 = 0.013$

$\bar{y}_1$ significantly different from $\bar{y}_2$ @ the $\alpha = 0.01$ level

*Figure 3.--Normalized distributions of good and poor watersheds for eight-variable discriminant model.*

## Application

The next step was development, application, and evaluation of classification rules. All classification rules depend directly on any "a priori" probability information on occurrence of both groups $y_1$ and $y_2$. One approach is to assume no such knowledge exists, hence probability ($P_1$) of $y_1$ is equal to probability ($P_2$) of $y_2$. And, because $P_1 + P_2 = 1.0$, both $P_1$ and $P_2$ are equal to 0.50. A second approach is to assume the available relative frequencies approximate the $P_1$ levels. Hence: $P_1 = 22/78 \simeq 0.30$ and $P_2 = 56/78 \simeq 0.70$. The problem with the second approach is that it assumes that the sampling techniques were truly random. When sampling is not random, relative frequencies are not used to approximate the $P_1$ "a priori" probabilities (Morrison 1967, Cooley and Lohnes 1971). For this problem, because of the state of the existing data base (highly local and ease of access oriented) and the constraints of accepting only truly poor or truly good producing streams, the technique was not random. This in no way affects any of the analytical procedures for discriminant functions (Morrison 1967); however, it precludes the use of relative frequencies as "a priori" probabilities. For classification analysis, $P_1 = P_2 = 0.50$ was the "a priori" estimate used. If and when future research indicates different levels of $P_1$ and $P_2$, such modifications can be entered in the analysis. The appendix illustrates just how $P_1$ and $P_2$ enter into the calculation of the posterior probabilities that dictate classification.

In general, where $P(y_1)$ and $P(y_2)$ represent posterior probabilities of $y_1$ and $y_2$, a classification rule is:

When $L_{1,2}P(y_2) \geq L_{2,1}P(y_1)$, classify as $y_2$; otherwise as $y_1$. Here, $L_{2,1}$ represents the expected loss for misclassifying a poor ($y_1$) as a good ($y_2$), and $L_{1,2}$ the expected loss for misclassifying a good ($y_2$) as a poor ($y_1$). The equal sign is relevant and can classify the function value as either a good producer or a poor producer. When $L_{1,2} = L_{2,1}$, this reduces to: Classify as $y_2$ when $P(y_2) \geq P(y_1)$; otherwise as $y_1$. For analysis purposes, we assumed $L_{1,2} = L_{2,1}$, and when $P_1 = P_2$ then $L_{1,2}P_1 = L_{2,1}P_2$. This simplifies the decision rule (Morrison 1967): When $f(x) \geq y_m$, classify as $y_1$; otherwise as $y_2$. Refer to figure 4 for this rule (Rule I). Application of Rule I to the 78 watersheds in the study yielded the results shown in table 3.

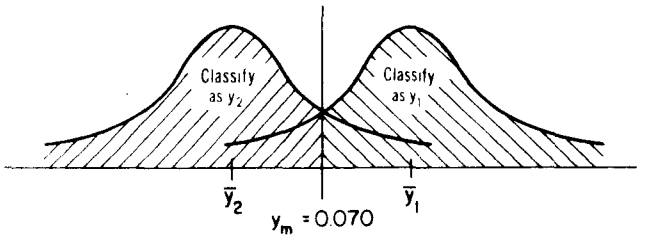



*Figure 4.--Classification Rule I.*

A modification to Rule I provides for selecting a band bracketing either side of $y_m$ so that any f(x) values falling within the band are not classified. Figure 5 illustrates this concept.

Application of Rule II to the 78 watersheds, where the lower limit ($E_\ell$) is 0.05, the upper limit ($E_u$) is 0.09, and $y_m$ is 0.07, yielded the results

Table 3--*Classification of 78 watershed by Rule I*

| Classified by discriminant function and Rule I | Preanalysis classification | | |
|---|---|---|---|
| | Good | Poor | Total |
| | Number of watersheds | | |
| Good | 41 | 5 | 46 |
| Poor | 15 | 17 | 32 |
| Total | 56 | 22 | 78 |

Here, 20 out of 78 watersheds were misclassified, abou 26 percent. If $L_{2,1}$, $L_{1,2}$, $P_1$ and $P_2$ are different from assumed levels, this would be altered by applying the techniques shown in the appendix. Under assumptions used here, application of the selected discriminant function yields much better results than assumptions of 50-50 possibilities (e.g., "a priori" probability = 0.50).

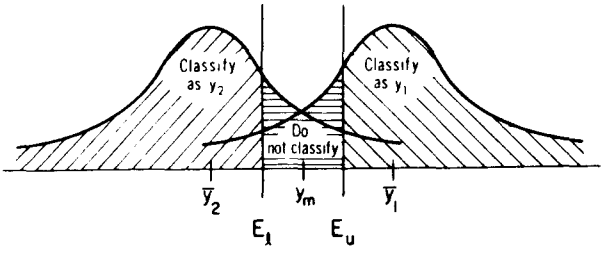



*Figure 5.--Classification Rule II.*

shown in table 4; where 13 watersheds out of 78 are misclassified (about 17 percent). Also, 13 watersheds out of 78 are not classified; so of the 65 classified, Rule II misclassifies 13 watersheds or 20 percent.

Actually, Rule I misclassifies 5 out of 22 poor (23 percent); Rule II, 2 out of 18 (11 percent). Rule I misclassifies 15 out of 56 good (27 percent); Rule II, 11 out of 47 (23 percent). Rule II is better than Rule I if we assume that the decisionmaker does not wish to classify the f(x) values clustered closely about $y_m$. Tables 5 and 6 illustrate the percentages of

Table 4--*Classification of 78 watersheds by Rule II*

| Classified by discriminant function and Rule II | Preanalysis classification | | |
|---|---|---|---|
| | Good | Poor | Total |
| | Number of watersheds | | |
| Good | 36 | 2 | 38 |
| Neutral | 9 | 4 | 13 |
| Poor | 11 | 16 | 27 |
| Total | 56 | 22 | 78 |

Table 5--*Generalized application of Rule I by percentages*

| Classification by normalized curves | Posterior percentage of total group | |
|---|---|---|
| | Good | Poor |
| Good | 77 | 23 |
| Poor | 23 | 77 |
| Total | 100 | 100 |

Table 6--*Generalized application of Rule II by percentages*

| Classification by normalized curves | Posterior percentage of total group | |
|---|---|---|
| | Good | Poor |
| Good | 71 | 18 |
| Neutral | 11 | 11 |
| Poor | 18 | 71 |
| Total | 100 | 100 |

the normalized curves for Rules I and II that generally fall into the categories specified.

## Discussion

The results of the data analysis provide us with a linear equation yielding at least a qualitative estimate of productivity of pink and chum salmon for southeast Alaska watersheds. The method is simple, flexible, much more accurate than assuming 50-50 probabilities, and responsive to the demands of the decision-maker. The land manager is provided with an analytical tool that can be used in solving land use problems. For example, a decisionmaker faced with a problem of allocating funds for protection or improvement of pink and chum salmon streams could use the discriminant function to classify very poor watersheds and very good watersheds. Then, based on managerial priorities for protection or enhancement of pink and chum salmon habitat commensurate with other resource values, he could determine appropriate allocation of funds.

Additional areas of managerial application involve land use decisions which may have an impact on salmon production. A land manager would benefit from knowing which watersheds are good producers of pink and chum salmon and which are poor producers. He could then take steps to minimize impact in watersheds with high production.

For the scientist and researcher, this tool provides a direct means of choosing an exceptionally good or an exceptionally poor salmon-producing watershed for more refined analysis of factors affecting productivity. It can direct the researcher to watersheds that have a higher probability of defining variables most likely to influence the level of pink and chum salmon production.

The important point is that, in general, the discriminant function developed from this research is a flexible tool that has potential dual utility: in land management decisions, a classification into poor or good categories aids the decisionmaking process; in research applications, it assists in detailed variable cause and effect analysis.

## *Literature Cited*


Brownlee, K. A.
1965. Statistical theory and methodology in science and engineering. p. 110-132, 560-561. John Wiley and Sons, New York.

Cooley, William W., and Paul R. Lohnes.
1971. Multivariate data analysis. p. 243-286. John Wiley and Sons, New York.

Dissmeyer, George E.
1967. Sheet erosion evaluation techniques used in the California north coastal river basin survey. 14 p., illus. USDA For. Serv. Interim Rep., Mass Erosion Conf., Berkeley, Calif.

Draper, N. R., and H. Smith.
1968. Applied regression analysis. p. 1-101. John Wiley and Sons, New York.

Horton, R. E.
1932. Drainage basin characteristics. Am. Geophys. Union Trans. 13:350-361.

Horton, R. E.
1945. Erosional development of streams and their drainage basins, a hydrophysical approach to quantitative morphology. Geol. Soc. Am. Bull. 56:275-370.

Kmenta, Jan.
1971. Elements of econometrics. p. 78-193. The MacMillan Company, New York.

Maxwell, J. C.
1960. Quantitative geomorphology of the San Dimas Experimental Forest, California. Tech. Rep. 19. ONR Proj. NR 389-042, Contr. N6 OHR 271-30: Nonr. 266(50). Off. Nav. Res., 95 p.

Melton, Mark A.
1957. An analysis of the relations among elements of climate, surface properties and geomorphology. Tech. Rep. 11. ONR Proj. NR 389-042. Contr. N6 OHR 271-30: Nonr. 266(50). Off. Nav. Res., 102 p

Morrison, Donald F.
1967. Multivariate statistical methods. p. 117-132. McGraw-Hill Book Co., New York.

Slack, Keith V.
1955. A study of the factors affecting stream productivity by the comparative method. Contrib. 501, Invest. Indiana Lakes and Streams 4(1):3-47.

Strahler, A. N.
1952. Hypsometric (area-altitude) analysis of erosional topography. Geol. Soc. Am. Bull. 63:1117-1142.

Strahler, A. N.
1953. Revision of Horton's quantitative factors in erosional terrain. (Abstr.) Hydrol. Sect., Am. Geophys. Union, Washington, D. C.

Strahler, A. N.
1954. Quantitative geomorphology of erosional landscapes. 19th Int. Geol. Congr., 1952, sect. 13, part 3:341-354, illus.

Thompson, D. H., and F. D. Hunt.
1930. The fishes of Champaign County: A study of the distribution and abundance of fishes in small streams. Bull. Ill. Nat. Hist. Surv. 19:5-101.

Wetherill, Barrie G.
1967. Elementary statistical methods. p. 146-161. Chapman and Hall, London.

Ziemer, G. L.
1973. Quantitative geomorphology of drainage basins related to fish production. State Alaska Dep. Fish and Game, Inf. Leafl. No. 162, 26 p.

# Appendix

*Posterior Probabilities and Classification Comments*

A. Posterior probabilities:

Two basic equations are solved simultaneously to obtain posterior probabilities:

1. $P(y_1) + P(y_2) = 1.0$.

2. $P(y_i) = C_b P_i f_i(y)$; $i = 1,2$.

Where: y is the value of the discriminant function for a watershed;

$P(y_1)$ is the posterior probability of the value y being classified in the poor group.

$P(y_2)$ is the posterior probability of the value y being classified in the good group.

$P_i$ is the "a priori" probability for the respective poor ($P_1$) and good ($P_2$) groups.

$C_b$ is a constant to be determined.

$f_i(y)$ is the value of the function on y:

$$f_i(y) = \frac{1}{\sqrt{2\pi}\ \sigma_i}\left[e^{-(1/2\sigma^2)(y-\bar{y}_i)^2}\right].$$

$\pi$ is the value 3.1416.

$\sigma_i$ is the standard deviation of group i discriminant values.

$\bar{y}_i$ is the mean of group i discriminant values.

e is natural log base e; value of 2.7183.

For this study two simplifications were made:

1. For the 8-variable model discriminant function,

a. $\sigma_1^2 = s_1^2$ and $\sigma_2^2 = s_2^2$.

s represents significant delineating characteristics (see table 2).

b. $s_1^2 = 0.0112$ and $s_2^2 = 0.0139$.

c. Assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and use:

$$s^2 = [(n-1)s_1^2 + (m-1)s_2^2]/(m + n-2).$$

m and n represent respective sample sizes for groups 1 and 2.

d. $s^2 = 0.013$; where $s^2 \doteq \sigma^2$.

e. Hence, $\sigma_1^2 = \sigma_2^2 = s^2$ (Wetherill 1967).

2. The "a priori" probabilities are assumed equal:

$$P_1 = P_2 = 0.50.$$

This allows reduction of the $f_i(y)$ function to only its variable portion:

$$f_i(y) = e^{-(1/2\sigma^2)\ (y-\bar{y}_i)^2}.$$

Here, $\frac{1}{\sqrt{2\pi}\ \sigma_i}$ will always be constant; therefore, it is included as part of the to-be-determined constant:

$C = \frac{1}{\sqrt{2\pi}\ \sigma_i} C_b$, and $C = 1/[\Sigma f_i(y)]$.

The posterior probabilities can be calculated now using this information. For:

$\bar{y}_1 = 0.1520$ (mean poor group).

$\bar{y}_2 = -0.0120$ (mean good group).

$\sigma^2 = s^2 = 0.0130$.

| y | $(y-\bar{y}_1)$ | $(y-\bar{y}_1)^2$ | $(y-\bar{y}_2)$ | $(y-\bar{y}_2)^2$ | $\sigma^2$ | $\frac{-1}{2\sigma^2}$ | $f_1(y)$ | $f_2(y)$ | $\Sigma f_i(y)$ | $C = \frac{1}{\Sigma f_i(y)}$ | $P(y_1)$ | $P(y_2)$ | $\Sigma P(y_i)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.040 | −0.112 | 0.0125 | 0.052 | 0.0027 | 0.013 | −38.46 | 0.6183 | 0.9014 | 1.520 | 0.658 | 0.4068 | 0.5932 | 1.0 |
| 0.050 | −0.102 | 0.0104 | 0.062 | 0.0038 | 0.013 | −38.46 | 0.6703 | 0.8640 | 1.534 | 0.649 | 0.4350 | 0.5650 | 1.0 |
| 0.060 | −0.092 | 0.0085 | 0.072 | 0.0052 | 0.013 | −38.46 | 0.7211 | 0.8187 | 1.540 | 0.650 | 0.4687 | 0.5313 | 1.0 |
| 0.070 | −0.082 | 0.0067 | 0.082 | 0.0067 | 0.013 | −38.46 | 0.7728 | 0.7728 | 1.550 | 0.645 | 0.5000 | 0.5000 | 1.0 |
| 0.080 | −0.072 | 0.0052 | 0.092 | 0.0085 | 0.013 | −38.46 | 0.8187 | 0.7211 | 1.540 | 0.650 | 0.5322 | 0.4678 | 1.0 |
| 0.090 | −0.062 | 0.0038 | 0.102 | 0.0104 | 0.013 | −38.46 | 0.8640 | 0.6703 | 1.534 | 0.649 | 0.5607 | 0.4393 | 1.0 |
| 0.100 | −0.052 | 0.0027 | 0.112 | 0.0125 | 0.013 | −38.46 | 0.9014 | 0.6183 | 1.520 | 0.658 | 0.5932 | 0.4068 | 1.0 |

For: $f(x) = -0.002787x_1 - 0.602159x_3 - 0.019247x_4 - 0.029875x_6 + 0.002146x_7 + 0.000332x_{12} - 0.023515x_{18} + 0.000660x_{19}$, $(f(x) = y)$.

When the "a priori" probabilities can be determined to be something other than equal (e.g., $P_1 = P_2 = 0.50$), the calculation of posterior probabilities is still straightforward. The analyst simply uses a value other than 0.50 in $P(y_i) = CP_i f_i(x)$ for the $P_i$ term. The theory and analysis remain unaltered.

B. Classification comments:

Classification into good and poor groups cannot rely alone on values of the discriminant function and respective posterior probabilities. The basic guideline is: $L_{2,1}\ P(y_1) \leq L_{1,2}\ P(y_2)$. Here $L_{2,1}$ is the cost of misclassifying a poor group as good; and $L_{1,2}$, the cost of misclassifying a good group as poor. When this inequality holds, classify the value for y as a member of the good group; otherwise as a member of the poor group. For $L_{2,1} = L_{1,2}$ the analyst can use just the posterior probabilities: $P(y_1) \leq P(y_2)$ implies classification as good, otherwise as poor.

The important point here is that the decisionmaker or researcher (user of analysis results) must determine his costs of wrong classification before the method can be applied. Exact costs need not be determined. Simple cost ratios (R) will suffice: $L_{2,1}/L_{1,2} = R_{2,1}/R_{1,2}$. Then use $R_{2,1}P(y_1) \leq R_{1,2}P(y_2)$ as the guideline.



☆GPO 799-359