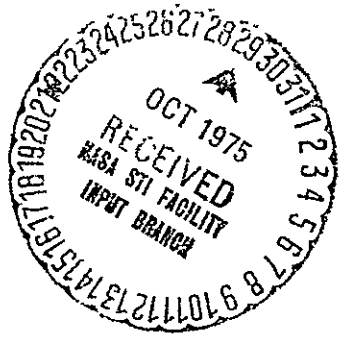
ANNUAL REPORT

Statistical Theory and Methodology for Remote Sensing Data Analysis With Special Emphasis on LACIE

Patrick L. Odell
Principal Investigator

June 1, 1974 - May 31, 1975



UTD

THE UNIVERSITY OF TEXAS AT DALLAS

Dallas, Texas

STATISTICAL THEORY AND METHODOLOGY FOR REMOTE SENSING DATA ANALYSIS WITH SPECIAL EMPHASIS ON LACIE

ANNUAL REPORT

ACKNOWLEDGMENTS

Research work contained in this annual report was carried out for NASA Johnson Space Center, Houston, Texas, under Contract NAS9-13512 to the University of Texas at Dallas, Richardson, Texas, for the period June 1, 1974, to May 31, 1975. This was accomplished in collaboration with Dr. A. H. Kvanli, UT-Dallas research scientist, Mr. Cary Simpson and Flo Marks, UT-Dallas graduate students, and the following consultants to UT-Dallas: Dr. T. L. Boullion and Dr. B. S. Duran of the Texas Tech University Statistics Faculty; Dr. W. A. Coberly of the University of Tulsa; Dr. J. P. Basu, research scientist UT-Dallas now with Lockheed Electronics Company, Inc.; Dr. H. L. Gray of the Southern Methodist University Statistics Department and D. D. McElroy, graduate student at SMU. Dr. Charles Peters and Dr. Jack Tubbs, National Research Council Fellows at NASA Johnson Space Center, co-authored two papers with Dr. Coberly (although they were not funded through this contract).

Patrick L. Odell
Principal Investigator

PREFACE

One of the major portions of the analytical procedure used by NASA-JSC in determining crop acreage is the method used for estimating crop proportions. One of the major tasks of this contract was to evaluate a set of five potential crop proportion estimators. Several studies of these proportion estimators are contained in this annual report including an empirical comparison of the different estimators using actual data and also an empirical study on the sensitivity (robustness) of one class of these estimators, referred to as the class of mixture estimators.

A concern when constructing a practical procedure for estimating crop production is the problem of encountering missing data, primarily due to cloud cover on one or more passes of the earth observation satellite. The effect of missing data upon the crop classification procedures is discussed in detail including a simulation of this missing data effect.

A basic discussion of the potential methods of crop acreage (proportion) estimation is contained in Paper 1. A literature study on these various estimators is contained in the report along with a complete description of each estimator including their known properties, bias, and MSE (mean square error). Paper 2 contains an empirical comparison of these proportion estimators using actual ERTS-A four channel multispectral scanner data taken over a 14 square mile test area site in Hill County, Montana. The report contains the results of several experiments including (for each experiment) the MSE's for each proportion estimator.

One such proportion estimator, the Odell-Chhikara (O-C) estimator, cannot be evaluated as previously proposed in earlier reports if the confusion matrix used to estimate the population proportions is singular. Paper 3 removes this problem by demonstrating two different methods for evaluating the proportion estimate utilizing matrix pseudoinverses and a modified Simplex procedure. The MSE for this estimator is also derived.

Paper 4 is a simulated sensitivity study of one class of these proportion estimators, the mixture estimators. Graphs and MSE's are included for each of the mixture proportion estimators for each simulated experiment. Paper 5 contains a technique for determining one such mixture proportion estimator, the maximum likelihood mixture estimator. The solution to the normal equations for determining the maximum likelihood estimates is an iterative one requiring initial estimates. This report compares the number of iterations necessary for convergence using different initial estimates on a set of actual ERTS data.

As mentioned previously, a major problem in any workable crop production estimation procedure is the problem of missing data. One such question that must be answered is "what happens to the quality of the proposed classification schemes when one encounters cloud cover on a pass of the satellite during one of the biological phases of the crops in some particular region?" In fact, this raises the general question of how to classify a region when one encounters missing data. For example, one could classify using only the complete data that exists or possibly estimate the missing values in some optimum fashion and proceed as if the estimated values were actual values. These and other methods of treating missing data are discussed at length in Paper 6 which includes

a comparison of the different methods using different sets of simulated data. Assuming a multivariate normal for the distribution of the multispectral scanner measurements, Paper 7 develops expressions for the estimators of the mean vector and covariance matrix using both complete and incomplete data. The report also contains a maximum likelihood scheme for classifying an observation vector into one of two multivariate populations with unknown means but known covariance matrices. Finally, Paper 8 considers the a priori probabilities of encountering missing data and calculates (and plots) the various probabilities of misclassification for the two population case.

The final problem addressed in this report is the problem of taking yield data (bushels per acre) gathered at several yield stations and extrapolating these values over some specified large region. For example, one might have such yield data at ten locations (stations) in the state of Kansas and wish to derive a yield estimate at each grid point (for some pre-determined grid) lying within the state. Paper 9 examines ten such methods ranging from merely using the sample mean to estimate the yield at <u>every</u> grid point to much more sophisticated techniques requiring extensive computer programming. The report also compares these ten extrapolation procedures using an empirical study with five years of wheat data from North Dakota covering the years 1962-1966. The authors also wrote a Fortran contour mapping program that plots the extrapolated results of any one method as a contour map of the entire region of interest. A <u>contour</u> is defined to be a line (or area) of roughly constant yield (i.e. lying within some specified interval). The effects of using one extrapolation method over another can

readily be determined by comparing the corresponding contour maps.

The final report, Paper 10, is a description of the Fortran computer programs written by the University of Texas at Dallas in support of some of the research activities carried out in these reports. The descriptions are very brief and are not meant to be program users' guides but instead are intended to acquaint the reader with their basic intent and input requirements. Further program documentation and computer card decks are available upon request.

TABLE OF CONTENTS

CONCERNING SEVERAL METHODS FOR ESTIMATING CROP ACREAGES USING REMOTE SENSING DATA

by

P. L. Odell [1]
J. P. Basu [2]

[1] The University of Texas at Dallas.

[2] Lockheed Electronics, Inc., Houston, Texas
(formerly at The University of Texas).

1. Introduction

In many areas of application of statistical data analysis unlabelled observations (observations of unknown classification) are available from several homogeneous populations constituting a single heterogeneous population and on the basis of these observations along with varying amount of information regarding these homogeneous subpopulations one has to estimate the proportions and sometimes the number and distribution parameters of these subpopulations. For example, in crop acreage estimation problem unlabelled observations, sometimes along with some labelled observations and perhaps some information about the distribution of individual crop populations, from several crops are available and on the basis of such information one has to estimate the acreage of a particular crop of interest or of all crops as proportion or proportions of total crop acreage.

The general statistical problem therein can be formulated as follows. Let

$$\mathcal{F} = \{F(x,\theta) : \theta \in \Omega \subset E_m\} \tag{1}$$

be a family of p-dimensional distribution functions, θ denoting a vector of m parameters belonging to a subset Ω of the m-dimensional Euclidean space E_m, and $G(\theta)$ any probability distribution function on Ω. Then the distribution function

$$H(x) = \int_{\Omega} F(x,\theta)\, dG(\theta) \tag{2}$$

is called a mixture on $\mathcal{F}$ with mixing distribution $G(\theta)$. When $\Omega = \{\theta_1, \theta_2, .., \theta_n\}$ is a finite set, then a probability distribution on Ω can be described by a finite set $G = \{\alpha_1\ \alpha_2, .., \alpha_N\}$, where

$$\alpha_i = P(\theta = \theta_i), \tag{3}$$

and a mixture H(x) can be obtained as

$$H(x) = \sum_{i=1}^{N} \alpha_i F(x, \theta_i) \ , \ \Sigma\alpha_i = 1, \tag{4}$$

a convex combination of a finite number of distinct elements of $\mathcal{F}$. The function H(x) is now called a finite mixture and the set G a finite mixing distribution. The probabilities α_i (i=1,.., N) given by (3) can also be interpreted as prior probabilities of $F(x,\theta_i)$'s. The class H of all finite mixtures on $\mathcal{F}$ is said to be identifiable (Teicher, 1960, 1961; Yakowitz, 1969), or, equivalently, the class $\mathcal{F}$ of finite mixing distributions are said to be identifiable (Chandra, 1969) if

$$\sum_{i=1}^{N} \alpha_i (F(x,\theta_i) = \sum_{j=1}^{N'} \alpha_j' F(x,\theta_j')$$

implies that $N = N'$ and for each i, $1 \leq i \leq N$, there is some j, $1 \leq j \leq N$, such that $\alpha_i = \alpha_j'$ and $\theta_i = \theta_j'$. A sample from H(x) is a set of observations $\{x_1,.., x_k,..\}$ on random vectors whose distribution functions $F(x,\theta)$ constitutes H(x) according to (2) or (4). Given a sample $x_1,.., x_k$ from H(x) and the parametric family of distribution functions to which the component distribution function of H(x) defined by (4) belong, the identification problem in its most general form is the problem of determining the number N, the mixing distribution G consisting of the proportions $\alpha_1,.., \alpha_N$ and the parameter vector $\theta_1,.., \theta_N$.

When the parametric family of distribution functions of the homogeneous subpopulation of crops is known, the crop acreage estimation problem can be viewed as a special case of identification problem. It is well known (Robbins, 1964) that the proportions $\alpha_1,.., \alpha_N$, the parameter vectors $\theta_1,.., \theta_N$ and the number N of subpopulations are estimable if, and only if, the class H of finite mixtures of distribution functions of subpopulations are identifiable. When the parametric family of the distribution functions of the subpopulations are

not known the acreage estimation problem then consists of estimating the proportions $\alpha_1, \ldots, \alpha_N$ and sometimes the number N of subpopulations. In this case also, estimation is possible if, and only if, the unknown subpopulation distribution functions define a class H of identifiable finite mixtures. Teicher (1960, 1961, 1963) and Yakowitz (1969) have proved that normal (multivariate and univariate), binomial and Poisson families of distribution functions are some of the families of distribution function, the class of finite mixtures on which are identifiable. In remote sensing data analysis, the mixtures are always assumed to be identifiable. So, the proportions of the subpopulations are always estimable.

When the number N of subpopulations in a mixture is unknown, Yakowitz (1969) has suggested a method of estimation for G and N using Levy distance between two distribution functions. But computation of Levy distance between multivariate distribution functions including even multivariate normal distribution functions is extremely difficult, practically impossible. Therefore, Yakowitz's method of estimation, the only available method of estimation in this case, is of no practical use.

In deriving all other estimators, so far mentioned in literature, the number N of subpopulations has been assumed to be known. Assuming N to be known, we have described some of the available methods of estimation for the mixing distribution G. Using the concept of statistically equivalent blocks (Tukey, 1947; Wilks, 1962) we have proposed some new estimators for the mixing distribution of a mixture of multivariate distributions. We have excluded graphical or semigraphical techniques (Harding, 1949, Cassie, 1954, Bliscke, 1964 and Bhattacharya, 1966) used in case of mixtures of univariate distributions, because, as Day (1969) has observed, these are difficult to extend to higher dimensions and appear to have poor sampling properties.

I No Training Sample Available

2. Moment Estimators

The moment estimators were first introduced by Karl Pearson (1894) in an attempt to estimate for a mixture of two univariate normal populations the means, variances and proportions by equating the first five moments with their sample values. Solving these five equations in the five unknown parameters leads to a ninth degree polynomial equation having at least one real root, eac real root giving a set of estimates for the parameters. Pearson proposed that the set having its sixth moment nearest the sample sixth moment be used as the final estimate.

Rao (1948) simplified the method of solution assuming the two univariate normal distributions in the mixture to have equal variances. Assume

$$H(x) = \alpha F_1(x) + (1-\alpha) F_2(x), \tag{5}$$

where F_1, F_2 are univariate normal distributions with mean μ_1, μ_2 and common variance σ^2. Let s_2, s_3 and s_4 be the second, third and fourth sample moments about the mean, and s_1 the first sample moment about the origin for a sample obtained from H. Equating these to the corresponding population moments introduces bias in the estimating equations for α. Hence, equating the first four k-statistics of Fisher to their expected values, which are the cumulants of H, avoids the bias. The first four k-statistics are given in terms of the sample moments by:

$$k_1 = s_1$$

$$k_2 = \frac{n}{n-1} s_2$$

$$k_3 = \frac{n^2}{(n-1)(n-2)} s_3$$

$$k_4 = \frac{n^2}{(n-1)(n-2)(n-3)} \{(n+1)s_4 - 3(n-1)s_2^2\} .$$

Equating these to the corresponding cumulants yields the estimating equations:

$$k_1 = \alpha \mu_1 + (1-\alpha)\mu_2$$

$$k_2 = \sigma^2 + \alpha d_1^2 + (1-\alpha)\, d_2^2$$

$$k_3 = \alpha d_1^3 + (1-\alpha)\, d_2^3$$

$$k_4 = \alpha d_1^4 + (1-\alpha)\, d_2^4 - 3[\alpha d_1^2 + (1-\alpha)\, d_2^2]^2$$

where $d_1 = \mu_1 - k_1$ and $d_2 = \mu_2 - k_1$.

Letting $x = d_1 d_2$, the value of x is obtained as the negative root of the cubic

$$x^3 + \frac{1}{2} k_4 x + \frac{1}{2} k_3^2 = 0.$$

If x is the required root, then d_1 is given by the negative root of the quadratic

$$d_1^2 + \frac{k_3}{x} d_1 + x = 0$$

and d_2 by $-(\frac{k_3}{x}) - d_1$. The estimates $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\alpha}$ and $\hat{\sigma}^2$ are given by

$$\hat{\mu}_1 = k_1 + d_1 \qquad \hat{\mu}_2 = k_1 + d_2$$

$$\hat{\alpha} = \frac{d_2}{(d_2 - d_1)} \qquad \hat{\sigma}^2 = k_2 + x.$$

The fundamental cubic equation

$$x^3 + \frac{1}{2} k_4 x + \frac{1}{2} k_3^2 = 0$$

has a single negative root greater than $-k_2$. Since the coefficient of x^2 is absent it is readily obtained.

The expressions for the standard errors of $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\alpha}$ and $\hat{\sigma}^2$ are very complicated, but it appears that the estimate of $\hat{\alpha}$ will have the highest

percentage of error whereas the estimates of $\ddot{\mu}_1$, $\ddot{\mu}_2$ and $\hat{\sigma}^2$ will be fairly reliable in large samples.

Using multivariate analogue of k-statistics Day (1969) has extended Rao's method to mixtures of two multivariate normals with common covariance matrix. In the p-variate case, with $p^2/2 + 5p/2 + 1$ parameters to estimate, there are p first moments, $p(p+1)/2$ second moments, $p + p(p-1) + p(p-1)(p-2)/6$ third moments and $p + 3\,p(p-1)/2 + p(p-1)(p-2)/2 + p(p-1)(p-2)(p-3)/24$ fourth moments. When $p>1$, since not all third or fourth moments are functionally independent, a choice has to be made as to which third and fourth moments, or functions of these moments, to use to obtain moment estimates.

When the covariance matrics of the two multivariate normal distributions forming the mixture are unequal, Martin (1936) observed that the multivariate analogue of Pearson's moment equations failed to provide any useful estimate.

Kabir (1968) considered moment estimates of the mixing distribution for a finite (more than two component) mixture on exponential families of distributions. It is believed that these estimates are of little practical use.

3. Maximum Likelihood Estimators

The maximum likelihood method has also been used in determining estimates of $\alpha_1,\alpha_2,\ldots,\alpha_N$ and $\theta_1,\theta_2,\ldots,\theta_N$ in the mixture $H(x) = \sum_{i=1}^{N} \alpha_i F_i(x)$. The procedure is to determine values of $\alpha_1,\alpha_2,\ldots\alpha_N$ and $\theta_1,\theta_2,\ldots,\theta_N$, that maximize the log of the likelihood function

$$L = \ell n \prod_{j=1}^{n} \left(\sum_{i=1}^{N} \alpha_i F_i(x_j;\theta_i) \right)$$

There are (N-1) α's and 2N θ's to estimate in the case when $F_1,\ldots,F_N$ are univariate normal distributions with $\theta_i = (\mu_i,\sigma_i^2)$. The estimates can be

obtained by solving the system

$$\frac{\partial L}{\partial \underline{\alpha}} (\underline{\alpha}, \theta_1, \ldots, \theta_N) = 0 , \tag{6}$$

$$\frac{\partial L}{\partial \underline{\theta}} (\underline{\alpha}, \theta_1, \ldots, \theta_N) = 0 .$$

This system is nonlinear, but can be solved by known iterative techniques such as the Newton-Raphson or gradient method. Hasselblad (1966) has considered the maximum likelihood approach, in the case of grouped data. The asymptotic variances for the estimates of the parameters were calculated and plotted. The asymptotic variances can be obtained from the diagonal elements of $-H^{-1}$ where H is the matrix

$$H = \begin{bmatrix} E\left(\frac{\partial^2 L}{\partial \underline{\alpha}^2}\right) & E\left(\frac{\partial^2 L}{\partial \underline{\alpha} \partial \underline{\theta}}\right) \\ N-1 \text{x} K-1 & N-1 \text{x} 2N \\ E\left(\frac{\partial^2 L}{\partial \underline{\alpha} \partial \underline{\theta}}\right) & E\left(\frac{\partial^2 L}{\partial \underline{\theta}^2}\right) \\ 2N \text{x} N-1 & 2N \text{x} 2N \end{bmatrix} \tag{7}$$

and $$\frac{\partial^2 L}{\partial \underline{\alpha}^2} = \left(\frac{\partial^2 L}{\partial \alpha_i \partial \alpha_j}\right) , \quad \frac{\partial^2 L}{\partial \underline{\theta}^2} = \left(\frac{\partial^2 L}{\partial \alpha_1 \partial \mu_1} \cdots \frac{\partial^2 L}{\partial \alpha_{N-1} \partial \sigma_N^2}\right)$$

$$\frac{\partial^2 L}{\partial \underline{\theta}^2} = \left(\frac{\partial^2 L}{\partial \mu_1 \partial \sigma_1^2} , \ldots , \frac{\partial^2 L}{\partial \mu_N \partial \sigma_N^2}\right)$$

The expected values in (7) are evaluated at the estimated parameter values. It is very difficult, if not impossible, to determine the asymptotic variances of the estimates analytically. The variances can, however, be obtained numerically.

Hasselblad (1969) considered finite mixtures of Poisson, Binomial and exponential distribution in a later study by means of maximum likelihood method.

Day (1969) considered the method of maximum likelihood in estimating the parameters of the mixture

$$H(x) = \alpha F_1(x) + (1-\alpha)\, F_2(x),$$

where F_1 and F_2 are multivariate normal distributions with pxl vector of means μ_1 and μ_2, respectively and common covariance matrix Σ. In this case the log likelihood function for a sample of size n is

$$L(\mu_1, \mu_2, \alpha, \Sigma) = \ln[(2\pi)^{-np/2}|\Sigma|^{-n/2} \cdot \prod_{i=1}^{n} \{\alpha \exp[-\tfrac{1}{2}(X_i-\mu_1)^T\Sigma^{-1}(X_i-\mu_1)] + (1-\alpha)\exp[-\tfrac{1}{2}(X_i-\mu_2)^T\Sigma^{-1}(X_i-\mu_2)]\}] \quad (8)$$

Taking logarithm and derivatives of L with respect to α, μ_1, μ_2 and Σ, we obtain

$$\sum_{i=1}^{n} \frac{e_{1i} - e_{2i}}{\hat{\alpha} e_{1i} + (1-\hat{\alpha}) e_{2i}} = 0$$

$$\sum_{i=1}^{n} \frac{(X_i - \hat{\mu}_1)\, \hat{\alpha} e_{1i}}{\hat{\alpha} e_{1i} + (1-\hat{\alpha}) e_{2i}} = 0$$

$$\sum_{i=1}^{n} \frac{(X_i - \hat{\mu}_2)(1-\hat{\alpha}) e_{2i}}{\hat{\alpha} e_{1i} + (1-\hat{\alpha}) e_{2i}}$$

$$-n\hat{\Sigma} + \sum_{i=1}^{n} \{(X_i-\hat{\mu}_1)(X_i-\hat{\mu}_1)^T \hat{\alpha} e_{1i} + (X_i-\hat{\mu}_2)^T (1-\hat{\alpha}) e_{2i}\} \cdot \{\hat{\alpha} e_{1i} + (1-\hat{\alpha}) e_{2i}\}^{-1} = 0$$

where

$$e_{ji} = e^{-\frac{1}{2}(X_i-\hat{\mu}_j)^T \hat{\Sigma}^{-1} (X_i-\hat{\mu}_j)}$$

If $P(k|X_j)$ denotes the probability that observation X_j arises from population k, then

$$\hat{P}(1|X_j) = \hat{\alpha}\, e_{ij} / \{\hat{\alpha}\, e_{1j} + (1-\hat{\alpha})\, e_{2j}\}$$

$$\hat{P}(2|X_j) = 1 - \hat{P}(1|X_j)$$

The m.ℓ. equations can then be written in the form

$$\frac{1}{\hat{\alpha}} \sum_{j=1}^{n} \hat{P}(1|X_j) = \frac{1}{1-\hat{\alpha}} \sum_{j=1}^{n} \hat{P}(2|X_j)$$

$$\hat{\mu}_1 = \{\sum_{j=1}^{n} X_j \hat{P}(1|X_j)\} / \sum_{j=1}^{n} \hat{P}(1|X_j) \tag{9}$$

$$\hat{\mu}_2 = \{\sum_{j=1}^{n} X_j \hat{P}(2|X_j)\} / \sum_{j=1}^{n} \hat{P}(2|X_j)$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{j=1}^{n} \{(X_j-\hat{\mu}_1)(X_j-\hat{\mu}_1)^T \hat{P}(1|X_j) + (X_j-\hat{\mu}_2)(X_j-\hat{\mu}_2)^T \hat{P}(2|X_j)\}$$

The mean μ_m, and covariance matrix, Σ_m, of the mixture are given by

$$\mu_m = \alpha\mu_1 + (1-\alpha)\mu_2, \quad \Sigma_m = \Sigma + \alpha(1-\alpha)(\mu_1-\mu_2)(\mu_1-\mu_2)'$$

Thus from the $\frac{1}{2}p^2 + \frac{5}{2}p + 1$ equations in (9), the m.ℓ. estimates of μ_m and Σ_m are given by the set of $\frac{1}{2}p^2 + \frac{3}{2}p$ equations

$$\hat{\mu}_m = \frac{1}{n} \sum_{j=1}^{n} X_j = \bar{X} \quad , \tag{10}$$

$$\hat{\Sigma}_m = \frac{1}{n} \sum_{j=1}^{n} (X_j-\bar{X})(X_j-\bar{X})'$$

Now

$$P(1|X) = [1 + e^{\underline{a}^T x + b}]^{-1}$$

where

$$\underline{a} = \Sigma^{-1}(\mu_2-\mu_1)$$

$$b = \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2) + \ell n(\frac{1-\alpha}{\alpha})$$

The likelihood funtion can be written in terms of μ_m, Σ_m, $\underline{a}$, and b, and equations (9) are transformed into the $\frac{1}{2}p^2 + \frac{3}{2}p$ equations of (4) and the p + 1 equations

$$\hat{\underline{a}} = \{\hat{\Sigma}_m - \hat{\alpha}(1-\hat{\alpha})(\hat{\mu}_1-\hat{\mu}_2)(\hat{\mu}_1-\hat{\mu}_2)^T\}^{-1} (\hat{\mu}_2-\hat{\mu}_1)$$

$$\hat{b} = \frac{1}{2}(\hat{\mu}_1-\hat{\mu}_2)\{\hat{\Sigma}_m-\hat{\alpha}(1-\hat{\alpha})(\hat{\mu}_1-\hat{\mu}_2)(\hat{\mu}_1-\hat{\mu}_2)^T\}^{-1}(\hat{\mu}_1+\hat{\mu}_2)+ n(\frac{1-\hat{\alpha}}{\hat{\alpha}})$$

which can be written as

$$\hat{\underline{a}} = \frac{\hat{\Sigma}_m^{-1} (\hat{\mu}_1-\hat{\mu}_2)}{1-\hat{\alpha}(1-\hat{\alpha})(\hat{\mu}_1-\hat{\mu}_2)^T\Sigma_m^{-1}(\hat{\mu}_1-\hat{\mu}_2)}$$

$$\hat{b} = -\frac{1}{2}\ \hat{\underline{a}}^T(\hat{\mu}_1-\hat{\mu}_2) + \ell n(\frac{1-\hat{\alpha}}{\hat{\alpha}}). \qquad (11)$$

The inversion of Σ_m rather than Σ might be preferable since $\hat{\Sigma}$ is more apt to be ill-conditioned. Now from equations (9), $\hat{\alpha}$, $\hat{\mu}_1$ and $\hat{\mu}_2$ are functions of the X's, $\hat{\underline{a}}$ and $\hat{b}$; and $\hat{\Sigma}_m$ is given from (10). Equations (11) then form a set of equations of the type

$$\hat{\underline{a}} = \phi_1(\hat{\underline{a}},\hat{b};X_1,\ldots,X_n), \qquad \hat{b} = \phi_2(\hat{\underline{a}},\hat{b};X_1,\ldots,X_n)$$

These equations can be solved by the usual methods; however, the solution may be somewhat laborious. The equations may yield several local maxima for the likelihood function. Thus the likelihood function should be examined at all local maxima so as to choose the solution yielding the overall maximum.

According to Day (1969) the m.ℓ. technique in the case of a mixture of 2 multivariate normals is computationally feasible for $p \leq 10$.

When the covariance matrices are not equal, the m.l. technique breaks down, since each sample point generates a singularity in the likelihood function.

4. Minimum Chi-Square Estimators

Let H(x) be a mixture of N p-variate normal distributions $F_1, \ldots, F_N$, so that

$$H(x) = \sum_{i=1}^{N} \alpha_i F_i(x), \tag{12}$$

$C_1, C_2, \ldots, C_k$ be k p-dimensional cubes of equal volume containing all n sample points from H(x), $H(C_i)$ be the probability of an observation falling in the ith cell C_i and n_i be the frequency of sample points in C_i. Then estimates of the parameters of the mixture, namely α_i's and the parameters of the F_i's, can be obtained by minimizing chi-square,

$$\chi^2 = \sum_{j=1}^{k} [\{n_j - nH(C_j)\}^2 / nH(C_j)], \tag{13}$$

or similar criteria, such as modified χ^2, Hellinger distance and Kullback-Leibber separator (see Rao, 1965, p 289).

In one dimension the minimum chi-square estimates can be obtained without great difficulty, and as one would expect, they behave well. In higher dimensions, however, the computation becomes prohibitive as it involves the evaluation of p-dimensional normal integrals over a series of cells.

Hasselblad (1966) used this approach in case of univariate mixtures when the data are grouped such that the length h of each class interval is relatively small compared with the variances $\sigma_1^2, \ldots, \sigma_N^2$. Assuming h to be 1, he approximated $F_i(C_j)$ by the density $f_i(x_j)$ of F_i evaluated at the mid point x_j of the jth cell C_j and obtained his estimates by minimizing χ^2 given by (13) with $H(C_j)$ replaced by $\Sigma\, \alpha_i f_i(x_j)$, j=1, ,k. Hasselblad then used his minimum chi-square estimates as an approximation to the maximum likelihood estimates.

In higher dimension also the computation difficulties involved in the evaluation of cell probabilities can be minimized by approximating the cell probabilities by $h^p \Sigma\, \alpha_i f_i(x_j)$, where $f_i(x_j)$ denote the value of the density function of F_i at the center x_j of the cell C_j, provided the length h of each cell is taken to be sufficiently small compared to $|\Sigma_i|^{1/p}$, the pth root of the determinant of the covariance matrix of F_i. However, it can be expected that use of such approximate values of cell probabilities in the chi-square criterion (13) may introduce serious bias in the estimates.

5. Least Square Estimator

A technique, which can be termed "least square" has been proposed by Choi (1969) and Choi and Bulgren (1968) for estimating parameters of a mixture (4) of univariate distribution functions. The proposed estimator based on a sample $X^{(n)}=(X_1,\ldots,X_n)$, denoted by $G(X^{(n)})=G_n=(\hat{\alpha}_1,\ldots,\hat{\alpha}_N;\ \hat{\theta}_1,\ldots,\hat{\theta}_N)$ is any $G(X^{(n)})$ which minimizes $S_N(G)$ given by

$$
\begin{aligned}
S_N(G) &= \int [H(x) - H_n(x)]^2 \, dH_n(x) \\
&= \int [\sum_{j=1}^{N} \alpha_j F(x;\theta_j) - H_n(x)]^2 \, dH_n(x) \\
&= \frac{1}{n} \sum_{i=1}^{n} [\sum_{j=1}^{N} \alpha_j F(X_{(i)},\theta_j) - \frac{i}{n}]^2,
\end{aligned}
$$

where $H_n(x)$ is the empirical distribution function of H(x), and $X_{(i)}$ denotes the ith order statistic of $X^{(n)} = (X_1,X_2,\ldots,X_n)$. To illustrate the procedure, let m = 1, that is, each F_i is indexed by a scalar θ_i. The procedure is to solve the system

$$\frac{1}{2}\dot{S}_n(g) = \begin{bmatrix} U_1(G) \\ U_2(G) \end{bmatrix} = 0$$

$$U_1(G) = \begin{bmatrix} \frac{\partial S_n(G)}{\partial \alpha_1} \\ \vdots \\ \frac{\partial S_n(G)}{\partial \alpha_N} \end{bmatrix} \qquad U_2(G) = \begin{bmatrix} \frac{\partial S_n(G)}{\partial \theta_1} \\ \vdots \\ \frac{\partial S_n(G)}{\partial \theta_N} \end{bmatrix}$$

Some properties of the estimator $\hat{G}_n = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_N, \hat{\theta}_1, \ldots, \hat{\theta}_N)$ are (under certain assumptions given in Choi [6]):

(i) If $G_0 = (\alpha_1^0, \alpha_2^0, \ldots, \alpha_N^0, \theta_1^0, \theta_2^0, \ldots, \theta_N^0)$ denotes the true values of the parameters then $\hat{G}_n \to G_0$ with probability one.

(ii) With probability one, there exists a neighborhood of G_0 such that, for all but finite n, $\hat{G}_n$ is the unique solution of $\dot{S}_n = 0$.

(iii) G_n has an asymptotic multivariate normal distribution

6. Statistically Equivalent Blocks

Let $X_1, \ldots, X_n$ be n observations on a pxl random vector X with distribution function $F(x)$, $h_1(x)$, $h_2(x), \ldots, h_n(x)$ be n functions of x, not necessarily different, such that the distribution of $h_i(x)$, i=1,.., n, is a continuous distribution function. Then for all i,

$$P[h_i(X_\alpha) = h_i(X_\beta),\ \alpha \neq \beta] = 0.$$

Also let $k_1, k_2, \ldots, k_n$ be a permutation of 1, 2,.., n and $X^{(k_1)}$ be defined as that X_α for which $h_{k_1}(X_\alpha)$ is the k_1 th order statistic among $h_{k_1}(X_1), \ldots, h_{k_1}(X_n)$. Then the cut

$$h_{k_1}(x) = h_{k_1}(X^{(k_1)})$$

defines two blocks

$$B_{1\cdots k_1} = \{x : h_{k_1}(x) \leqslant h_{k_1}(X^{(k_1)})\}$$

and

$$B_{k_1+1\ldots n+1} = \{x : h_{k_1}(x) > h_{k_1}(X^{(k_1)})\}.$$

The procedure is continued. Let $0 < k_2 < k_1$. Then the function $h_{k_2}(x)$ is used to order k-1 X_α's in $B_{1\ldots k_1}$ and $X^{(k_2)}$ is defined as that X_α for which $h_{k_2}(X_\alpha)$ is the k_2th order statistic among k-1 $h_{k_2}(X_\beta)$, $X_\beta \varepsilon\ B_{1\cdots k_1}$. Let

$$B_{1\cdots k_2} = B_{1\cdots k_1} \cap \{x : h_{k_2}(x) \leqslant h_{k_2}(X^{(k_2)})\}$$

$$B_{k_2+1\cdots k_1} = B_{1\cdots k_1} \cap \{x : h_{k_2}(x) > h_{k_2}(X^{(k_2)})\}.$$

If $k_1 < k_2$, rank the n-k, X_α's in $B_{k_1\cdots n+1}$ according to $h_{k_2}(x)$ and let $X^{(k_2)}$) be the (k_2-k_1)th in the ranking. Then

$$B_{k_1+1\cdots k_2} = B_{k_1+1\cdots n+1} \cap \{x : h_{k_2}(x) \leqslant h_{k_2}(X^{(k_2)})\}$$

and

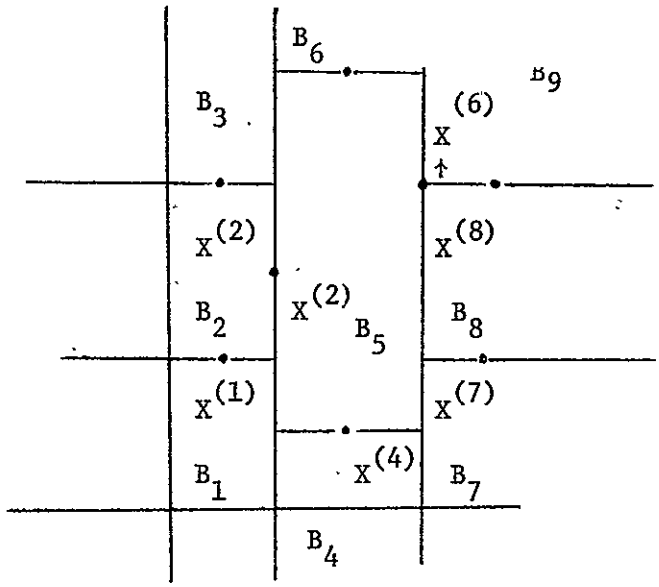
$$B_{k_2+1\cdots n+1} = B_{k_1+1\cdots n+1} \cap \{x : h_{k_2}(x) > h_{k_2}(X^{(k_2)})\}.$$

At the end of the mth stage there will be m+1 blocks: $B_{\cdots j_1}$, $B_{j_1+1\cdots j_2}$, $\cdots$, $B_{j_m+1\cdots n+1}$, where $j_1,\ldots, j_m$ are $k_1,\ldots, k_m$ arranged in ascending order. The function $h_{k_{m+1}}(x)$ is then used to order X_α's in the block having k_{m+1} as one of its indices and $X^{(k_{m+1})}$ is defined to be the X_α in this block such that $k_{m+1}-1$ X_α's are either in lower-ranking blocks (blocks with indices less than k_{m+1}) or ranked lower in this block by $h_{km+1}(x)$. This block is replaced by its intersection with $\{x : h_{k_{m+1}}(x) \leqslant h_{k_{m+1}}(X^{k_{m+1}})\}$ and its intersection with $\{x : h_{k_{m+1}}(x) > h_{k_{m+1}}(X^{(k_{m+1})})\}$. The procedure is continued until after n stages there are n+1 blocks $B_1,\ldots, B_{n+1}$. The blocks $B_1,\ldots, B_{n+1}$ are called statistically equivalent blocks (Tukey , 1947). The procedure has been described by Fraser (1957, section 4.3) and by Wilks (1962, section 8.7).

Example. To illustrate this method let

$$X_1 = \begin{bmatrix}1\\2\end{bmatrix},\ X_2 = \begin{bmatrix}3\\1\end{bmatrix},\ X_3 = \begin{bmatrix}5\\2\end{bmatrix},\ X_4 = \begin{bmatrix}2\\3\end{bmatrix}$$
$$X_5 = \begin{bmatrix}1\\4\end{bmatrix},\ X_6 = \begin{bmatrix}3\\5\end{bmatrix},\ X_7 = \begin{bmatrix}4\\4\end{bmatrix},\ X_8 = \begin{bmatrix}5\\4\end{bmatrix}$$

be 8 observations. Let $(k_1, \cdots, k_8) = (3,6,1,2,4,5,7,8)$ and let $h_3(x) = h_6(x) = x_1$, the first coordinate, and $h_k(x) = x_2$, the second coordinate, for $k \neq 3,6$. Then $X^{(3)}$ is the observation vector with the third smallest value of the first coordinate and $X^{(3)} = X_4$. Next, $X^{(6)}$ is the observation among the latter five which has the (6-3)th smallest value of the first coordinate. So, $X^{(6)} = X_7$. Of the two observations, X_1 and X_5 having the value of first coordinate less than the value of first coordinate of $X^{(3)}$, X_1 has the smallest second coordinate and X_5 the second smallest. So, $X^{(1)} = X_1$ and $X^{(2)} = X_5$. The process is continued until $X^{(k_8)} = X^{(8)}$ is obtained. The blocks $B_1, \cdots, B_8$ are shown in the following figure.



Extension of Least Square Method

Using the concept of statistically equivalent blocks, the least square method of Choi can now be extended very easily. Now we have,

$$S_N(G) = \Sigma_{i=1}^{n} \left[\sum_{j=1}^{m} \alpha_j F(X^{(i)}, \theta_j) - 1/n \right]^2, \qquad (15)$$

where $$F(X^{(i)}, \theta_j) = \sum_{k=1}^{i} F_j(B_k), \; F_j = F(x, \theta_j).$$

However, the evaluation as $F_j(B_k)$, the probability content of B_k under $F_j(x)$, may be very difficult.

7. Comparison of Estimation Procedures

In sections 2-6 we have discussed some of the most interesting procedures for estimating the mixing distributions of mixtures of univariate and multivariate identifiable distributions in the absence of training samples from the component distributions. The performances of these estimators are usually compared on the basis of bias and mean square error. But the sampling distributions of these estimators are extremely difficult, rather impossible, to obtain analytically. Robertson and Fryer (1970) obtained some approximate expression for the bias of moment estimators of a special type. Hasselblad (1966) was able to obtain graphs of variances of the estimates of the proportions (mixing distributions) against population distances in case of mixtures of univariate normal distributions. On the basis of Monte Carlo studies Day (1969) has tabulated the mean and variances of the moment estimators and m.l. estimators of the mixing distribution $(\alpha, 1-\alpha)$ of a mixture of two normal distributions with equal variance or covariance matrix. Choi and Bulgren (1968) have also tabulated the mean square error of the least square estimates of the mixing distribution of mixtures of univariate normal distribution.

II Training Sample Available

In the following section we discuss some procedures for estimating the mixing distribution $G(\alpha_1, \ldots, \alpha_N)$ of mixtures $H(x)$, given by (4), when we have training samples from the component distributions $F_1,.,F_N$. For the sake of simplicity we shall restrict our attention to mixtures of two populations. Throughout our discussions in the following sections we shall assume that $X_1,.,X_{n_1}$ is a sample from the first population Π_1, $Y_1, ., Y_{n_2}$ a sample from Π_2 and $Z_1,.,Z_n$ a sample from Π, a mixture of Π_1 and Π_2. The distribution $H(x)$ of the mixture is given by

$$H(x) = \alpha_1 F_1(x) + \alpha_2 F_2(x), \quad \alpha_1 + \alpha_2 = 1.$$

8. Confusion Matrix Estimators

A detailed account of this estimation method can be found in the papers by Odell and Chhikara (1974). In this procedure unlabelled sample from Π is classified by an arbitrarily preassigned classifier C. Let e_1 and e_2 be the expected proportion of the sample classified under Π, and Π_2 respectively and as a result of using the classifier let $\phi_1 = P(\Pi_1|\Pi_2)$ and $\phi_2 = P(\Pi_2|\Pi_1)$ be the respective probabilities of misclassifying an element from Π_2 into Π_1 and vice versa. Then the true proportions α_1 of Π_1 and $\alpha_2(=1-\alpha_1)$ of Π_2 are given by

$$\begin{aligned} e_1 &= (1-\phi_2)\alpha_1 + \phi_1(1-\alpha_1), \quad e_2 = 1-e_1 \\ \alpha_1 &= (e_1-\phi_1)/(1-\phi_1-\phi_2), \quad \alpha_2 = 1-\alpha_1. \end{aligned} \tag{16}$$

The matrix

$$P = \begin{bmatrix} 1-\phi_2 & \phi_1 \\ \phi_2 & 1-\phi_1 \end{bmatrix}$$

has been called confusion matrix. The misclassification probabilities are estimated from the training samples and the expected proportions from the unlabelled sample. There we have

$$\hat{\alpha} = (\hat{e}_1 - \phi_1)/(1-\phi_1-\phi_2), \tag{17}$$

if ϕ_1 and ϕ_2 are known and

$$\hat{\hat{\alpha}} = (\hat{e}_1 - \hat{\phi}_1)/(1-\hat{\phi}_1-\hat{\phi}_2), \tag{18}$$

if ϕ_1 and ϕ_2 are unknown, $\hat{\theta}$ denoting the estimate of the quantity θ.

$$\text{Var } \hat{\alpha} = \text{Var } (\hat{e}_1)/(1-\phi_1-\phi_2)^2 \tag{19}$$

Some approximate expression or the mean square error or α can be found in Odell and ~~Chhikara (1974).~~ Kvanli (1975),

9. Least Square Estimators

Let the component distributions F_1 and F_2 of the mixture $H(x)$ be known. Then in case of univariate populations, minimizing

$$S_2(G) = \sum_{i=1}^{n} [\sum_{j=1}^{2} \alpha_i F_i (Z_{(i)}) - i/n]^2,$$

where $Z_{(i)}$ is the i-th order statistic of $Z_1, ., Z_n$, we obtain

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} [F_1(Z_{(i)}) - F_2 (Z_{(i)})][F_2 (Z_{(i)}) - i/n]}{\sum_{i=1}^{n} [F_1(Z_{(i)}) - F_2(Z_{(i)})]^2} \tag{20}$$

and

$$\text{Var } (\hat{\alpha}_1) = \frac{\sum_{i=1}^{n} |F_1(Z_i) - F_2(Z_i)|^2 H(Z_i)[1-H(Z_i)]}{n \left(\sum_{i=1}^{n} |F_1(Z_i) - F_2(Z_i)|^2 \right)^2}. \tag{21}$$

In the multivariate case, one way to obtain an estimate of α, and an expression for its variance may be to replace $F_j(Z_{(i)})$, j=1,2, in (20) and (21) by $\Sigma_{k=1}^{i} F_j(B_k)$, where $B_1,\ldots, B_{n+1}$ are the statistically equivalent blocks determined by any prechoosen ordering functions $h_1(x)$, ., $h_n(x)$ and the sample Z_1, ., Z_n. Since evaluation of $F_j(B_k)$ may be extremely difficult, the least square method may not work very well in case of multivariate data.

In crop acreage estimation problem we are interested in estimating the mixing distribution only. So, the following estimation procedure may be of interest. Let $F_{j1}(x_1)$, $F_{j2}(x_2)$, ., $F_{jp}(x_p)$ be the marginals of the distribution function $F_j(x)$, j=1,2 and $H_1(x_1)$, ., $H_p(x_p)$ that of H(x). Then we have

$$\alpha_1 F_{11}(x_1) + (1-\alpha_1)F_{21}(x_1) = H_1(x_1)$$
$$\alpha_1 F_{12}(x_2) + (1-\alpha_1)F_{22}(x_2) = H_2(x_1)$$
$$\vdots$$
$$\alpha_1 F_{1p}(x_p) + (1-\alpha_1)F_{2p}(x_p) = H_p(x_p)$$

An estimate of α_1 can be obtained by minimizing

$$Q(G) = \sum_{k=1}^{p} \sum_{i=1}^{n} [\sum_{j=1}^{2} \alpha_j F_{jk}(Z_{k(i)}) - i/n]^2, \qquad (22)$$

where $Z_{k(i)}$ is the i th order statistic among Z_{1k}, ., Z_{nk}, the k th components of Z_1, ., Z_n. Thus we obtain

$$\hat{\alpha}_1 = \frac{\Sigma_{k=1}^{p} \Sigma_{i=1}^{n} [F_{1k}(Z_{k(i)}) - F_{2k}(Z_{k(i)})][F_2(Z_{k(i)}) - i/n]}{\Sigma_{k=1}^{p} \Sigma_{i=1}^{n} [F_{1k}(Z_{ik}) - F_{2k}(Z_{ik})]^2} \qquad (23)$$

and

$$\text{Var}(\hat{\alpha}_1) = \frac{\Sigma_{k=1}^{p} \Sigma_{i=1}^{n} |F_{1k}(Z_{ik}) - F_{2k}(Z_{ik})|^2 H(Z_{ik})[1-H(Z_{ik})]}{np\,(\Sigma_{k=1}^{p} \Sigma_{i=1}^{n} |F_{1k}(Z_{ik}) - F_{2k}(Z_{ik})|^2)^2} \qquad (24)$$

When $F_j(x)$'s are not completely known, but are only known to belong to certain parametric family of distribution functions and we have training samples from both $F_1(x)$ and $F_2(x)$, a fairly good estimate of F_1 and F_2 can be obtained by plugging in the estimates of their parameters in their functional forms. An estimate $\hat{\hat{\alpha}}_i$ of α_i can now be obtained from (20) or (23) by replacing $F_j(x)$'s or $F_{jk}(x_k)$'s by their estimates. The sampling distribution of α_i is extremely difficult to obtain analytically. So, we have been unable to give any expression for bias and mean square error of α.

10. Moment Estimator

Let the p-variate population Π be a mixture of N p-variate populations Π_1, ., Π_m such that

$$H(x) = \alpha_1 F_1(x) + \cdots + \alpha_N F_N(x), \ \alpha_1 + . + \alpha_N = 1, \qquad (25)$$

where $H(x)$ is the distribution function of Π and $F_i(x)$ that of Π_i and α_i's are proportions of Π_i in Π. Then

$$\int g(x)\, dH(x) = \sum_{i=1}^{N} \alpha_i \int g(x)\, dF_i(x), \qquad (26)$$

where $g(x)$ is any monotone function of x such that all the integrals involved in (26) exist. When α_i's are unknown, but $H(x)$ and $P_i(x)$'s are known, then α_i's can be estimated from the following set of N=1 equations

$$\int g_k(x)\, dH(x) = \sum_{i=1}^{N} \alpha_i \int g_k(x)\, dF_i(x), \ i=1,\cdot\cdot,N-1, \qquad (27)$$

where $g_1(x),.., g_{N-1}(x)$ are (N-1) linearly independent monotone functions such that integrals on both sides of (27) exist. When $F_i(x)$'s and/or $H(x)$ are also unknown, it has been proposed by Hartley (1974) that for estimating the α_i's the moments of $g_k(x)$'s in the system of equation (27) should be replaced by the corresponding sample moment, provided one has N sample from $F_i(x)$'s and a

sample from H(x). Let $Y^1, .., Y^n$ be a sample from the p-variate population H(x) and $X_i^1, .., X_i^{n_i}$ a sample from the p-variate population $F_i(x)$, i=1,.., N; also let $U_1, V_1, .., V_N$ be mx1 vectors defined as follows.

$$u_j = Y_j = \sum_{i=1}^{n} Y_{ij}/n \text{ and } \nu_{ij} = \sum_{k=1}^{n_i} X_{ij}^k / n_i, \ 1 \leq j \leq p,$$

$$u_{j+p} = \sum_{k=1}^{n} Y_1^k Y_j^k / n, \ \nu_{i,j+p} = \sum_{k=1}^{n_i} X_{i_1}^k X_{ij}^k / n_i, \ i=1,.., N; \ 1 \leq j \leq p$$

$$u_{j+2p} = \sum_{k=1}^{n} Y_2^k Y_j^k / n, \ \nu_{i,j+2p} = \sum_{k=1}^{n_i} X_{i2}^k X_{ij}^k / n_i, \ i=1,.., N; \ 2 \leq j \leq p$$

$$u_{j+2p+(p-1)} = \sum_{k=1}^{n} Y_3^k Y_j^k / n, \ \nu_{i,j+2p+(p-1)} = \sum_{k=1}^{n_i} X_{i3}^k X_{ij}^k / n_i, i=1, ,N; \ 3 \leq j \leq p$$

:

$$u_m = \sum_{k=1}^{n} (Y_p^k)^2 / n, \ \nu_{i,N} = \sum_{k=1}^{n_i} (X_{ip}^k)^2 / n_i, \ i=1,.., N; \ m=p+\frac{p(p+1)}{2}$$

Then estimates of α_i's can be obtained as a set of values $\alpha_1, .., \alpha_N$ for which

$$Q = || U - \sum_{i=1}^{N} \alpha_i V_i ||$$

is a minimum, $|| \cdot ||$ denoting some suitable norm. Apprehending difficulty in minimizing Q based on l_2-norm, Hartley (1974) has suggested that l_1 - or weighted l_1- norm should be used as the norm. Thus, when l_1- norm is used,

$$Q = \sum_{j=1}^{m} | u_j - \sum_{i=1}^{N} \alpha_i \nu_{ij} |$$

or when weighted l_1- norm is used,

$$Q = \sum_{j=1}^{m} w_j | u_j - \sum_{i=1}^{N} \alpha_i \nu_{ij} | ,$$

where w_j's are proportional to the inverse of sample standard derivation of u_j. We shall refer to these estimates as <u>moment estimates</u> with <u>minimum weighted l_1-norm deviation</u>. As can be expected, these estimates are very sensitive to the choice of weights.

Feiveson (1974) used this method to estimate proportions of 9 crop classes. He found appreciable change in the estimates due to

(i) change in weights beyond 3rd decimal place,

(ii) translation of the entire data.

The estimates are not expected to be translation invariant. It is believed that the relative sample sizes $n_1/n, .., n_m/n$ will also influence the estimates. This method of estimation has also been found to do better when the number of classes are smaller (say 2 or 3).

References

Anderson, T. W., "Some nonparametric multivariate procedures based on statistically equivalent blocks," in Multivariate Analysis-I, Ed. P. R. Krishnaiah, Academic Press, New York, 1966.

Boes, D. C., "On the estimation of mixing distributions," Ann. Math. Stat., Vol. 37, (1966) pp. 177-188.

_____ ."Minimax unbiased estimator of mixing distribution for finite mixtures," Sankhya, A. (1967), pp. 417-420.

Bhattacharya, C. G., "A simple method of resolution of a distribution into Guassian components," Biometrics, 23 (1967), pp. 115-135.

Cassie, R. M., "Some uses of probability paper in the analysis of size frequency distributions," Austral. J. of Marine and Freshwater Res., 5, (1954) 513-522.

Chandra, S. "On mixtures of probability distributions," Master's Thesis, Univ. of Chicago, 1969.

Choi, K. and Bulgren, W. G., "An estimation procedure for mixture distribution," J. Roy. Stat. Soc., B 30 (1968) pp. 444.

Choi, K., "Estimators for the parameters of a finite mixture of distributions," Ann. of Inst. of Stat. Math., Vol. 21, (1969) pp. 107-116.

Choi, K., "Empirical Bayes procedure for (pattern) classification with stochastic learning," Ann. of Inst. of Stat. Math., Vol 21, (1969) pp. 117-115.

Day, N. E., "Estimating the components of a mixture of normal distributions," Biometrika, Vol. 56, No. 3, (1969), pp. 463-474.

Feiveson, A., Personal communication, August, 1974.

Fraser, D. A. S., Nonparametric Methods in Statistics, John Wiley & Sons, New York, 1957.

Harding, J. P., "The use of probability paper for the graphical analysis of polymodal frequency distributions," J. of the Marine Biological Assoc., 28, (1949), pp. 141-153.

Hartley, H. O., "The estimation of acreages from satellite data," Technical Report, NASA-JSC, 1974.

Hasselblad, V., "Estimation of parameters for a mixture of normal distributions," Technometrics, Vol. 8, No. 3 (1966), pp. 431-446.

_____ "Estimation of finite mixtures of distributions from the exponential family," J. of American Stat. Assoc. (1969), pp. 1459-1471.

Kabir, A. B. M. L., "Estimation of parameters of a finite mixture distribution," J. Roy. Stat. Soc., B 30 (1968), pp. 470-482.

Odell, P. L. and Chhikara, R., "Estimation of a large area crop acreage inventory using remote sensing technology," Tech. Report, NASA-JSC., 1974.

Pearson, K., "Contributions to the mathematical theory of evolution," Phil. Trans. R. Soc. A, 185 (1894), pp. 71-110.

Rao, C. R., Advanced Statistical Methods in Biometric Research, John Wiley & Sons, New York, 1952.

_____ Linear Statistical Inference and Its Application, John Wiley & Sons, New York, 1965.

Robbins, H., "The empirical Bayes approach to statistical decision problems," Ann. Math. Stat. 35 (1964), pp. 1-20.

Robertson, C. A. and Fryer, J. G., "The bias and accuracy of moment estimators," Biometrika, 57 (1970), pp. 57-65.

Teicher, H., "On the mixture of distributions," Ann. Math. Stat. (1960), pp. 55-73.

_____ "Identifiability of mixtures," Ann. Math. Stat. (1961), pp. 244-248.

_____ "Identifiability of finite mixtures," Ann. Math. Stat. (1963), pp. 1265-1269.

Wilks, S. S., Mathematical Statistics, John Wiley and Sons, New York, 1962.

Yakowitz, S., "A consistent estimator for the identification of finite mixtures," Ann. Math. Stat. (1969) Vol. 40, No. 5, pp. 1728-1735.

AN EMPIRICAL COMPARISON OF FIVE PROPORTION ESTIMATORS

by

W. A. Coberly[1/] and P. L. Odell[2/]

[1/] Department of Mathematical Sciences, University of Tulsa

[2/] The University of Texas at Dallas

AN EMPIRICAL COMPARISON OF FIVE PROPORTION ESTIMATORS

W. A. Coberly[1] and P. L. Odell[2]

I. Introduction: Let $\pi_1,\ldots, \pi_m$ be m pattern classes having probability density functions $f_1,\ldots, f_m$ respectively. Let $\alpha = (\alpha_1,\ldots, \alpha_m)^T$ be the proportion vector defining the mixture density

$$f(x) = \sum_{k=1}^{m} \alpha_k f_k(x), \tag{1}$$

where $\alpha_k = 1$ and $\alpha_k > 0$ $k=1,\ldots, m$. The measurement vector x is n - dimensional. Consider the following problem: Given a random sample $X_1,\ldots, X_N$ from the mixture distribution, find an estimate for α. The density functions $f_1,\ldots, f_k$ are assumed to be known. A short discussion of each of the five estimators considered in this paper follows:

Classification Estimator (CL) Define

$$x_k(x) = \begin{cases} 1 & \text{if } f_k(x) > f_j(x) \quad j \neq k \\ 0 & \text{otherwise} \end{cases}$$

for $k=1,\ldots, m$. Now the CL estimate is given by

$$\hat{\alpha}_k = \frac{1}{N} \sum_{i=1}^{N} x_k(X_i) \quad k=1,\ldots, m. \tag{2}$$

Simply, $\hat{\alpha}_k$ is the proportion of the sample which would be classified into class π_k by the maximum - likelihood classifier.

---

[1]University of Tulsa

[2]University of Texas at Dallas

Odell-Chhikara Estimator (OC) Let P be the confusion matrix defined by the maximum - likelihood classifier. That is, $P = (p_{ij})$ where

$$p_{ij} = Pr\{\chi_i(X) = 1 \mid X \in \pi_j\}$$

for $i, j=1,\ldots, m$. Simply, p_{ij} is the probability of classifying an observation from π_j into π_i. In [2] it is shown the $E[\hat{\alpha}] = P\alpha$, where $\hat{\alpha}$ is the CL estimator defined in (2). This relation suggests an estimate of α given by the least squares solution of the matrix equation

$$\hat{\alpha} = P\alpha \tag{3}$$

constrained by $\Sigma\alpha_i = 1$ and $\alpha_i > 0$. Denote this estimate by $\hat{\beta}$. The properties of $\hat{\beta}$ are discussed at length in [1].

Maximum - Likelihood Estimator (ML). For the sample $X_1,\ldots, X_N$, the log - likelihood function with the added constraint $\Sigma\alpha_i = 1$ gives the objective function

$$\psi(\alpha) = \log \prod_{i=1}^{N} f(X_i) - \lambda\left(\sum_{k=1}^{m}\alpha_k - 1\right)$$

$$= \sum_{i=1}^{N} \log \sum_{k=1}^{m}\alpha_k f_k(X_i) - \lambda\left(\sum_{k=1}^{m}\alpha_k - 1\right).$$

Now $dL/d\alpha = 0$ implies that

$$\alpha_k = \frac{1}{N}\sum_{i=1}^{N} \alpha_k f_k(X_i) \Big/ \sum_{j=1}^{m}\alpha_j f_j(X_i) \tag{4}$$

for $k=1,\ldots, m$. Let G be a vector valued function whose components are defined by the RHS of (4). Then the ML estimate of α, if it exists, must be a solution to the fixed point equation

$$\alpha = G(\alpha). \tag{5}$$

The numerical behavior of this optimization problem is discussed further in [3].

Mixture Estimate (MX) The equation relating the mixture distribution function and the component distribution functions corresponding to (1) is

$$F(x) = \sum_{k=1}^{m} \alpha_k F_k(x).$$

Denote the corresponding jth marginal dist
and $F_k^{(j)}$ respectively, $j=1,\ldots, n$. Consider the following system of equations.

$$F^{(j)}(x_{ij}) = \sum_{k=1}^{m} \alpha_k F_k^{(j)}(x_{ij}) \tag{6}$$

$j=1,\ldots, n$, $i=1,\ldots, s$. In this experiment the functions $F^{(j)}$ were estimated from the sample $X_1,\ldots, X_N$ and x_{ij} was chosen to be the $100i/(s+1)$ percentile of $\hat{F}^{(j)}$. Define a vector Y and a matrix A by

$$y_p = \hat{F}^{(j)}(x_{ij})$$

$$a_{p_k} = F_k^{(j)}(x_{ij})$$

for $p = (j-1)s+i$, $j=1,\ldots, n$, $i=1,\ldots, s$. Then the mixture estimate MX is defined to be the least squares solutions of the linear system

$$Y = A\alpha \tag{7}$$

constrained by $\Sigma\alpha_i = 1$, $\alpha_i > 0$. (It is assumed that for each j the percentiles x_{ij} $i=1,\ldots, s$ are unique. If not, the redundant equation is deleted.)

Moment Estimator (MO) Let μ_{oj} and μ_{oj}^k denote the jth component of the mean vector for the mixture density and the kth component density $k=1,\ldots, m$. Likewise let μ_{ij} and μ_{ij}^k denote the ijth element of the noncentral dispersion matrix of the mixture density and the kth component density. From (1)

$$\mu_{ij} = \sum_{k=1}^{m} \alpha_k \mu_{ij}^k$$

for $i=0,\ldots, n$, $j=1,\ldots, n$. Estimate the raw moments μ_{ij} from the sample $X_1,\ldots, X_N$ and denote these by $\hat{\mu}_{ij}$. Now define a vector Y and a matrix A by

$$y_j = \hat{\mu}_{oj},\ a_{jk} = \mu_{oj}^k$$

for $j=1,\ldots, n$ and

$$y_p = \hat{\mu}_{ij},\ a_{pk} = \mu_{ij}^k$$

where $p = n+(i+1)i/2+j$ and $i=1,\ldots, n$; $j=1,\ldots, i$ and $k = 1,\ldots, m$. Then the moment estimate MO is defined to be the least squares solution of the linear system

(8) $WY - WA\alpha$

constrained by $\Sigma\alpha_i = 1$ and $\alpha_i > 0$. Here W is a diagonal weighting matrix defined in this experiment by

$$W_{pp} = \begin{cases} (\hat{\mu}_{pp})^{-1/2}, & p=1,\ldots, n \\ (\hat{\mu}_{ii}\,\hat{\mu}_{jj} + \hat{\mu}_{ij}^{2})^{-1/2}, & \\ \quad \text{for } p = n + (i+1)i/2+j, & i=1,\ldots, n,\ j=1,\ldots, i. \end{cases}$$

A comment is now in order about the assumption that the component densities f_k are known. In this experiment the densities are estimated by sampling labeled or ground truth data prior to the estimation of the proportion vector α based on the mixture sample $X_1,\ldots, X_N$. Hence the density estimates and the proportion estimates are found independently. This distinguishes the problem posed here from the general mixture problem in which the component density functions and the proportions are estimated simultaneously from the mixture sample. That is, no labeled data is assumed available.

II. Experimental Procedure The data base used for this experiment consisted of ERTS-A 4-channel multispectral scanner data, taken on May 5, May 23, June 11, and June 29, 1973, over a 14 square mile test site in Hill County (N), Montana. Only the data from the June 11 pass was used in these results. A ground truth map and summary dated May 5, were used in conjunction with the Earth Resources Interactive Processing

System (ERIPS) at NASA/JSC to identify and tag pixels, which fell into recognized homogeneous fields of one of five general classes: Wheat, fallow, barley, grass and stubble. Of a total of 8400 pixels in the test site, 2600 were labeled accordingly. The training and test samples were determined as follows. A random sample of 30% of the labeled data was selected for the training data set and (after replacement) another 30% random sample was selected for the test data set. For the training set, each pixel selected was grouped according to class tag and statistics and marginal histograms for the total 16 channel data set were computed and saved for each class. The test data was then classified using the training statistics (June 11 pass only) and the confusion matrix estimate was based on those results. This procedure was repeated 30 times. The five proportion estimates described in the previous section were then computed for each of the 30 trials, first using the 2600 labeled pixels as the mixture sample (Experiment I) and second using the total 8400 pixels (Experiment II). The results of each of the 30 trials for both experiments are exhibited in Appendices A and B.

Five major computer programs were required. CLASS computes the CL and ML estimates. This program reads each set of training statistics, classifies the mixture sample and computes the CL estimate. Furthermore, as the values of the density functions $f_1(X_1),\ldots, f_m(X_i)$ (m=s) are computed to make the classification decision, they are written to temporary storage for use in computing the ML estimate. A simple iteration method was used to solve this fixed point equation $\alpha = G(\alpha)$.

The initial guess, denoted by α°, was taken to be CL estimate just computed, then successive approximations were found by the following rule.

$$\alpha^n = G(\alpha^{n-1}),\ n=1,\ 2,\ldots$$

until $\| \alpha^n - \alpha^{n-1} \|$ was "sufficiently small." In this experiment ten iterations were sufficient to achieve 3 or 4 place convergence. The CONF program reads the training statistics and classifies a random sample of the labeled data in order to estimate the confusion matrix. The ODE program reads the CL estimates and the confusion matrix estimates and finds the constrained least squares solution of (3). The MIXTUR program reads the mixture and component histograms, sets up the linear system given in (6) and finds the constrained least squares solution. The MOMENT program reads the mixture and component statistics, converts them to the noncentral moments, sets up the linear system given in (B) and finds the constrained least square solution. The last three programs require subroutines PREP and QUADPR which set up the quadratic objective function and solve the resulting quadratic programming problem, respectively. The program QUADPR was adapted from [2]. All programs were coded in Fortran and are listed in Appendix C.

Table 1. Summary of Experiment I.
(Labeled Data, 2600 Pixels)

		CLASS	ODELL	MLE	MIX	MCM	GT
MEAN	WH	.282384	.297041	.300439	.294353	.274 49	.371900
	FA	.265500	.274428	.296764	.312696	.235257	.286200
	BA	.186089	.174306	.176688	.183300	.197059	.115400
	GR	.101679	.086066	.085825	.074200	.075168	.079200
	ST	.164346	.168155	.140178	.135448	.217765	.147300
VAR	WH	.000038	.000174	.000083	.000201	.000459	
	FA	.000445	.001385	.000408	.002216	.003027	
	BA	.000031	.000190	.000084	.000123	.000288	
	GR	.000080	.000051	.000060	.000165	.000463	
	ST	.000344	.001829	.000383	.002376	.004145	
TOTAL VAR		.000937	.003628	.001017	.005020	.008383	
MSE		.015172	.013322	.010086	.016572	.032066	

Table 2. Summary of Experiment II
(Total Data Set, 8400 Pixels)

		CLASS	ODELL	MLE	MIX	MCM	GT
MEAN	WH	.252733	.262044	.272700	.226467	.084333	.294000*
	FA	.207300	.183526	.185133	.353019	.023415	.249000
	BA	.167633	.154997	.151200	.218460	.325809	.124000
	GR	.186466	.187198	.189166	.158955	.184282	.138000
	ST	.186066	.212231	.201700	.043098	.382159	.196000
VAR	WH	.000095	.000263	.000310	.000232	.000949	
	FA	.000269	.007430	.000472	.002872	.002114	
	BA	.000063	.000299	.000206	.000135	.000610	
	GR	.000261	.000363	.000529	.001042	.000900	
	ST	.000141	.009569	.000832	.002975	.003402	
TOTAL VAR		.000829	.017924	.002350	.007257	.007974	
MSE		.008622	.057982	.010273	.055378	.180347	

III. Results and Conclusions In Experiment I only the labeled subset of the Hill County data was used. In Table 1. the mean and variance of the components of the proportion vector estimates are tabulated, averaged over the 30 trials. TOTAL VAR is the sum of the variances and MSE is the mean square error about the true proportion vector given in the column GT. It should be noted that one wheat field, accounting for 6.4% of the labeled data was consistantly described as barley, apparently by all five estimators. Hence the discrepancy between the ground truth and the five estimates with respect to wheat and barley.

If however, the 6.4% is added to the Barley Class, then the MSE in Table 1. would read MSE .002858 .004393 .001286 .018915 .006146 which reverses the ordering of (MO, MX) and (CL, OD). The ML estimator remains the best. Otherwise all five estimates apparently will be relatively stable with CL and ML having the lowest variance and MSE with MO apparently having a significantly larger variance and MSE than the others. A trial by trial summary is given in Appendix A.

In Experiment II the total Hill County data set was used. In Table 2 the GT column is not precise since many small classes listed in the ground truth summary were arbitrarily consolidated into the five class model. Conclusions drawn from the MSE should reflect this uncertainty. The CL and ML estimates are again lowest in variance and MSE followed by the OC and MX estimates. However, in this experiment the MO estimate is very bad. Apparently the moments of the total data set deviated sufficiently from the five class model to cause the distorted results. In Experiment I when only the labeled data was used the model more accurately reflected

the sample and the MO estimate performed reasonably well. If further studies find this lack of robustness to be a consistant problem, then the moment estimate should be eliminated from consideration in the application.

Based upon the results of these two experiments the following ordering of the five estimates is suggested:

$$(CL, ML) > (OC, MX) > MO$$

Several operational considerations should also be noted however. The MX and MO require design decisions prior to implementations such as choice of the percentiles and weights used in the construction of the linear systems in (6) and (8) respectively. In an automated system these additional parameters might be considered as a nuisance. The OC estimator requires allocation of some labeled data in order to estimate the confusion matrix. An alternative would be to use Monte Carlo methods on the model described in [1] to obtain numerical approximations of the confusion matrix given the estimates of the component densities. The CL, OC and ML estimators require classification of the total data set and hence require much more computational time than the MX and MO estimates.

In conclusion, on the basis of this study, the maximum likelihood estimator would appear to be the best of the five with respect to MSE, numerical stability, and adaptability in an automated system.

REFERENCES

[1] Chhikara, R. S. and Odell, P. L. (1974), "Acreage Estimates for Crops Using Remote Sensing Data," Technical Report for NASA.

[2] Kuester, J. L. and Mize, J. H., "Optimization Techniques with Fortran," pp. 106-119, McGraw-Hill.

[3] Peters, Charles, Personal communications.

ON SOLVING FOR THE PROBABILITY VECTOR p IN EQUATION Ap = e WHERE THE COLUMNS OF A AND e ARE PROBABILITY VECTORS

by

P. L. Odell*

A. H. Kvanli

*University of Texas at Dallas

I. INTRODUCTION

Consider m different populations $\pi_1, \pi_2, \ldots, \pi_m$ and a random sample of n values from these populations. Let n_i be the number of points classified into π_i, i = 1, ..., m, using a classification algorithm. If n (i|j) is the number of data points classified into π_i but actually belonging to π_j, then

$$n_i = n(i|1) + n(i|2) + \ldots + n(i|m) \tag{1.1}$$

Also, $$\frac{n_i}{n} = \sum_{j=1}^{m} \frac{n(i|j)}{n} \qquad i = 1, 2, \ldots, m$$

are the observed proportions for the sample data under the classification algorithm used. The observed proportion n_i/n is a biased estimate of p_i, where p_i = the actual proportion of the sample that is in π_i. n_i/n is an unbiased estimate of e_i, where

$$e_i = E\left(n_1/n\right)$$

$$= \sum_{j=1}^{m} E\left(\frac{n(i|j)}{n}\right)$$

$$= \sum_{j=1}^{m} p_j \cdot P(i|j) \tag{1.2}$$

and where $P(i|j)$ denotes the probability of classifying a data point from π_j into π_i under the classification algorithm.

If $P(i|j) = 0$ for $i \neq j$ or $n(i|j) = n(j|i)$ for all i,j then $e_i = p_i$ for $i = 1, 2, \cdots, m$, i.e. the sample proportions provide unbiased estimates of the actual proportions. In general, a classification algorithm will be subject to error and these two proportions will not be the same.

Denoting the observed proportion n_i/n by $\hat{e}_i$, then by (1.2) it follows that

$$\begin{aligned} e &= E(\hat{e}) \\ &= Pp \end{aligned} \tag{1.3}$$

where

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix} \qquad p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix}$$

and

$$P = \begin{bmatrix} P(1|1) & P(1|2) & \cdots & P(1|m) \\ P(2|1) & P(2|2) & \cdots & P(2|m) \\ \vdots & & & \\ P(m|1) & P(m|2) & \cdots & P(m|m) \end{bmatrix}$$

Considering the case where P is unknown, th

$$\hat{P}p = e \tag{1.4}$$

and an estimate of $\hat{p}$ is

$$\hat{p} = \hat{P}^{-1} \hat{e} \qquad (1.5)$$

where $\hat{P}$ is an estimate of P, $\hat{P}$ non-singular.

$\hat{p}$ is clearly a biased estimate of p, where both bias and mean square error of $\hat{p}$ depend on the performance of the classification algorithm as well as the degree to which the sample represents the population.

Consider the case for m = 2. If N sample values are used to estimate $P(1|1)$, $P(2|1)$ and N sample values are used to estimate $P(1|2)$, $P(2|2)$ then

$$\hat{P} \begin{pmatrix} \frac{x}{N} & 1-\frac{y}{N} \\ 1-\frac{x}{N} & \frac{y}{N} \end{pmatrix}$$

where

x = the number correctly classified into π_1

y = the number correctly classified into π_2.

Now, $\hat{P}$ is singular iff x+y = N. Since X and Y are independently distributed as binomial variables, then $\hat{P}$ is singular with positive probability. Consequently, the pseudoinverse of $\hat{P}$ must be employed for these situations.

Definition 1.1 A matrix A^+ is called a pseudoinverse of A if it satisfies the following conditions

(i) $AA^+A = A$

(ii) $A^+AA^+ = A^+$

(iii) $(AA^+)^T = AA^+$

(iv) $(A^+A)^T = A^+A$.

Theorem 1.1 Each matrix has one and only one pseudoinverse. [1].

II. DEFINITION OF PROBLEM

The problem is to

$$\text{Min } || \hat{P}p - \hat{e} || \tag{2.1}$$

subject to $J^T p = 1$

$p \geq 0$

where $J^T = (1, 1, 1, \ldots, 1)$.*

(2.1) is equivalent to solving

$$\text{Max } (\hat{e}^T \hat{P}p - \frac{1}{2} p\hat{P}^T \hat{P}p)$$

$$\text{subject to } J^T p = 1 \tag{2.2}.$$

$$p \geq 0$$

Another approach would be to solve

$$\text{Min } || \begin{pmatrix} \hat{P} \\ J^T \end{pmatrix} p - \begin{pmatrix} \hat{e} \\ 1 \end{pmatrix} ||$$

subject to $p \geq 0$.

This approach may not produce a solution vector $\hat{p}$, which sums to (exactly). For this reason (2.1) will be solved instead.

Two methods for solving (2.2) will be discussed. The first method (Iterative Search Method) is due to D. L. Nelson [3], and the second method uses a modified version of the Simplex procedure.

III. Iterative Search Method (Method I)

Define $C = (I - \frac{1}{m} JJ^T)\ (\hat{P}^T \hat{P})\ (I - \frac{1}{m} JJ^T)$

and $H = (I - \frac{1}{m} JJ^T)\ \hat{P}^T\ (\hat{e} - \frac{1}{m}\ \hat{P}\ J)$

$$\text{Let } p^*_y = \frac{1}{m} J + C^+ \hat{P}^T (\hat{e} - \frac{1}{m} \hat{P} J) + (I - \frac{1}{m} JJ^T - C^+C)\ y \qquad (3.1)$$

where y is any mx1 (real) vector.

Theorem 3.1 If $p^*_y \geq 0$ for some y, say y_o, then $p^*_{y_o}$ is an optimal solution to (2.2). If $p^*_y < 0$ for all vectors (i.e. $p^*_y \geq 0$ is infeasible), then let $\{ p_{n_i}, i = 1, 2, \ldots, k\}$ be the negative components of p^*_y. Then there exists a vector $\hat{p}$ which satisfies (2.2) such that at least one of the $p_{n_i} = 0$.

To determine if $p_y^* \geq o$ has a feasible solution use the Phase I procedure of the Simplex method. This procedure will determine if there is a feasible solution to the matrix inequality

$$A\,y \leq b \qquad (3.2)$$

where $A = C^{+}C + \frac{1}{m} JJ^{T} - I$

$$b = \frac{1}{m} J + C^{+} \hat{P}^{T} \left(\hat{e} - \frac{1}{m} \hat{P}\, J\right)$$

Using any linear programming routine, solve

$$\text{Min } \phi = \sum_{1}^{m} (v_i^{+} + v_i^{-})$$

$$\text{subject to } (A \; -A \; I \; I \; -I) \begin{pmatrix} Y^{+} \\ Y^{-} \\ S_{+} \\ V^{+} \\ V^{-} \end{pmatrix} = b$$

where S = slack variables

$$V^{+} = \begin{pmatrix} v_1^{+} \\ \vdots \\ v_m^{+} \end{pmatrix}, \; y^{+} = \begin{pmatrix} y_1^{+} \\ \vdots \\ y_m^{+} \end{pmatrix}$$

$$V^{-} = \begin{pmatrix} v_1^{-} \\ \vdots \\ v_m^{-} \end{pmatrix}, \; y^{-} = \begin{pmatrix} y_1^{-} \\ \vdots \\ y_m^{-} \end{pmatrix}$$

Consequently, (3.2) has a solution iff the sum of the artificial ariables equal zero, i.e. $\phi = 0$. If this is the case, the y vector atisfying (3.2) is given by

$$y = y^{+} - y^{-}$$

For the case where no feasible p_y^* exists, then by Theorem 3.1, a solution to (2.2) exists of the form (3.1), where at least one of the components is equal to zero. For such a situation, two cases are considered.

Case A: $C^+ CH = H$

Step 1

Choose any real vector y, say y_o.

Let $S_o = \{ i \mid p^*_{y_{oi}} \geq 0 \}$

and $T_o = \{ i \mid p^*_{y_{oi}} < 0 \} \neq \phi$,

where $p^*_{y_{oi}}$ = the i-th component of $p^*_{y_o}$

For each $i \in T_o$, set $p_i = 0$ and see if a feasible solution exists in the (m-1) - dimensional space which remains. This is equivalent to removing the i-th column of $\hat{P}$ and solving for p_y^* as before.

Let $S_1 = \{i \mid i \in T_o$ and $p_i = 0$ provided a feasible solution$\}$

and $T_1 = \{i \mid i \in T_o$ and $p_i = 0$ provided an infeasible solution$\}$.

If $T_1 = \phi$, go to Step 4. Otherwise continue to Step 2.

Step 2

Consider all pairs of components (i, j) such that $i \in T_1$ and $j \notin S_i$ and set $p_i = p_j = 0$. Solve the remaining (m-2) - dimensional space, if possible. This is equivalent to removing the i-th and j-th columns of $\hat{P}$ and again solving for p_y^* as before.

Let $S_2 = \{ (i, j) \mid p_i = p_j = 0$ provided a feasible solution$\}$

and $T_2 = \{ (i, j) \mid p_i = p_j = 0$ provided an infeasible solution$\}$.

If $T_2 = \phi$, go to Step 4. Otherwise continue to Step 3.

Step 3

Do the same for all triples (i, j, k) such that at least one element $\in T_2$ and such that (i, j), (i, k), (j, k) $\notin S_2$.

Next consider 4- tuples, etc., until all tuples have been considered or $T_i = \phi$ for some i.

Note: In Steps 2, 3, at no time should a set of components be set equal to zero when some proper subset (set equal to zero) brought a solution. For such a situation, the objective junction in (2.2) will not be improved [3].

Step 4

After a finite number of steps, the optimum solution can be derived by determining which vector obtained in Steps 2, 3 maximizes the expression in (2.2).

Case B: $C^{+} CH \neq H$

Step 1´

Attempt to find a solution by setting $p_j = 0$ for $j = 1, 2, \ldots, m,$ one at a time.

Let $S_1 = \{i \mid p_i = 0$ provided a feasible solution$\}$

and $T_1 = \{i \mid p_i = 0$ provided an infeasible solution$\}$

If $T_1 = \phi$, go to Step 4´. Otherwise go to Step 2´.

Steps 2´, 3´, 4´ are exactly the same as Steps 2, 3, 4 (respectively) for Case A.

IV. Simplex Method (Method II)

Let $f = \hat{e}^T \hat{P}$ and $D = \hat{P}^T \hat{P}$.

Thus (2.2) can be written

$$\text{Max } p_o = f\, p - \frac{1}{2} p^T D\, p$$

subject to $J^T p = 1,\ p \geq 0.$

Let $g(p) = J^T p - 1$.

Imposing the Kuhn-Tucker conditions for this problem (see [2]), there must exist a set of p_j and λ such that, for each P_j,

$$\left[\frac{\partial p_o}{\partial P_j} + \lambda \frac{\partial g(p)}{\partial P_j}\right] \leq 0 \quad j = 1, \ldots, m \tag{4.1}$$

λ , unrestricted in sigr

Thus (since D is symmetric)

$$f^T - Dp + J \leq 0.$$

Introduce a set of slack variables $S = (s_1, s_2, \ldots, s_m)^T$ such that

$$f^T - Dp + \lambda J + S = 0 \quad . \tag{4.2}$$

As a consequence of (4.2) (see [2]), it follows that $p_j s_j$ must equal zero in the optimum solution.

A starting feasible solution is $p_1 = 1$, $p_2 = p_3 = \ldots = p_m = 0$, and so Phase I of the Simplex method can be omitted. The second step is to introduce a set of non-negative artificial variables v_j. These are subtracted from the equation set (4.2) and the

objective function is defined as the sum of these artificial variables.

Letting $V = (v_1, v_2, \ldots, v_m)^T$, then the problem is

$$\text{Max} \quad -J^T V \tag{4.3}$$

$$\text{subject to (a)} \quad J^T p = 1$$

$$\text{(b)} \quad f^T - Dp + \lambda_1 J - \lambda_2 J + S - V = 0$$

$$\text{(c)} \quad p_j s_j = 0$$

$$\text{(d)} \quad p \geq 0 .$$

(4.3) can be expressed in matrix notation as

$$\text{Max} \quad -J^T V \tag{4.4}$$

$$\text{subject to} \quad \begin{bmatrix} D & -J & J & -I & I \\ J^T & 0 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} p \\ \lambda_1 \\ \lambda_2 \\ S \\ V \end{pmatrix} = \begin{pmatrix} f^T \\ 1 \end{pmatrix}$$

$$p_j \, s_j = 0$$

$$\lambda_1 \geq 0, \ \lambda_2 \geq 0, \ p \geq 0, \ S \geq 0, \ V \geq 0.$$

First, (4.4) will be put in canonical form with starting basis $p_1, v_1, v_2, \ldots, v_m$.

The constraints in (4.4) can be written

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} & -1 & 1 & -1 & 0 & \cdots 0 & 1 & 0 & \cdots 0 \\ d_{21} & d_{22} & & d_{2m} & -1 & 1 & 0 & -1 & \cdots 0 & 0 & 1 & \cdots 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & & \vdots & & \\ d_{m1} & d_{m2} & & d_{mm} & -1 & 1 & 0 & 0 & \cdots -1 & 0 & 0 & \cdots 1 \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots 0 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \\ 1 \end{bmatrix}$$

The first step is to remove $d_{11}, \ldots, d_{m1}$. Thus row $(m + 1)$ is multiplied by $- d_{i1}$ and added to row i providing

$$\begin{bmatrix} 0 & d'_{12} & \cdots & d'_{1m} & -1 & 1 & -1 & 0 & \cdots 0 & 1 & 0 & \cdots 0 \\ 0 & d'_{22} & & d'_{2m} & -1 & 1 & 0 & -1 & 0 & 0 & 1 & \cdots 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & d'_{m2} & & d'_{nn} & -1 & 1 & 0 & 0 & \cdots -1 & 0 & 0 & \cdots 1 \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 & 0 & \cdots 0 & 0 & 0 & \cdots 0 \end{bmatrix} = \begin{bmatrix} f'_1 \\ f'_2 \\ \vdots \\ f'_m \\ 1 \end{bmatrix}$$

$$\text{where} \quad d'_{ij} = d_{ij} - d_{i1}$$

$$f'_i = f_i - d_{i1}$$

The next step is to multiply each row by -1 for which $f'_i < 0$, except for column $(2m + 2 + i)$. This remains a positive 1 . For sake of discussion, it is assumed that $f'_i \geq 0$ for $i = 1, \ldots, m$. Writing

the objective function as the first row, the simplex tableau becomes

$-P_o$	p_1	p_2	$\cdots$	p_m	λ_1	λ_2	s_1	s_2	$\cdots$	s_m	v_1	v_2	$\cdots$	v_m	
1	0	0	...	0	0	0	0	0	...	0	-1	-1	...	-1	0
0	0	d'_{12}	...	d'_{1m}	-1	1	-1	0	...	0	1	0	...	0	f'_1
0	0	d'_{22}		d'_{2m}	-1	1	0	-1	...	0	0	1	...	0	f'_2
$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
0	0	d'_{m2}		d'_{mm}	-1	1	0	0		-1	0	0		1	f'_m
0	1	1	...	1	0	0	0	0	...	0	0	0	...	0	1

The final step is to remove the artificial variables $v_1, \ldots, v_m$ from the objective function. This is accomplished by adding the sum of rows 2 through (m + 1) to row 1 providing the tableau

$-P_o$	p_1	p_2	$\cdots$	p_m	λ_1	λ_2	s_1	s_2	$\cdots$	s_m	v_1	v_2	$\cdots$	v_m	
1	0	θ_2	...	θ_m	-m	m	-1	-1	...	-1	0	0	...	0	γ
0	0	d'_{12}	...	d'_{1m}	-1	1	-1	0	...	0	1	0	...	0	f'_1
0	0	d'_{22}		d'_{2m}	-1	1	0	-1	...	0	0	1	...	0	f'_2
$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
0	0	d'_{m2}	...	d'_{mm}	-1	1	0	0		-1	0	0	...	1	f'_m
0	1	1	...	1	0	0	0	0	...	0	0	0	...	0	1

$$\text{where } \theta_i = \sum_{j=1}^{m} d'_{ji} \qquad i = 2, \ldots, m$$

$$\gamma = \sum_{i=1}^{m} f'_i$$

Note that the first row may differ slightly if previously any of the rows were multiplied by -1. Columns 1, 2m + 3, ..., 3m + 2 will always contain zero. The simplex method can now be applied to the above tableau with the restriction that s_j and p_j cannot be in the basis at the same time. This algorithm can be easily coded on a digital computer.

V. Examples

A. $m = 3$, rank $(\hat{P}) = 3$

$$\hat{P} = \begin{pmatrix} .8 & .3 & .2 \\ .2 & .6 & .1 \\ 0 & .1 & .7 \end{pmatrix} \quad , \quad \hat{e} = \begin{pmatrix} .6 \\ .1 \\ .3 \end{pmatrix}$$

(i) Method I

$$p^*_y = \begin{pmatrix} .69 \\ -.14 \\ .45 \end{pmatrix} + (0)y \qquad \text{(Infeasible).}$$

The optimal solution is found for $p_2 = 0$, where

$$p^*_y = \begin{pmatrix} .605 \\ .395 \end{pmatrix} + (0)y \text{ , i.e. } p_1 = .605,\ p_3 = .395.$$

(ii) Simplex Method

The optimal solution is given by

$$p_1 = .605$$
$$p_3 = .395$$
$$\lambda_2 = .018$$
$$s_2 = .040$$

All other variables = 0.

Here it can be seen that Nelson's procedure is rather inefficient since it was necessary to examine all solutions for which $p_2 = 0$. For each method, the norm in (2.1) $= .0056 > 0$.

B. $m = 6$, rank $(\hat{P}) = 6$

$$\hat{P} = \begin{bmatrix} .71 & .04 & .13 & .01 & .03 & .05 \\ .02 & .82 & .06 & .11 & .02 & .01 \\ .11 & .01 & .66 & .03 & .07 & .02 \\ .03 & .05 & .04 & .75 & .01 & .01 \\ .06 & .03 & .08 & .06 & .85 & .03 \\ .07 & .05 & .03 & .04 & .02 & .88 \end{bmatrix} \quad \hat{e} = \begin{bmatrix} .10 \\ .18 \\ .22 \\ .28 \\ .12 \\ .10 \end{bmatrix}$$

(i) Method I

$$p^*_y = \begin{bmatrix} .065 \\ .148 \\ .294 \\ .343 \\ .077 \\ .073 \end{bmatrix} + (0)y$$

Since each $p_i > 0$, then the above is an optimal solution.

(ii) Simplex method

The optimal solution is

$$\hat{p}_1 = .065$$

$$\hat{p}_2 = .148$$

$$\hat{p}_3 = .294$$

$$\hat{p}_4 = .343$$

$$\hat{p}_5 = .077$$

$$\hat{p}_6 = .073$$

$$\lambda_1 = 0.0$$

All other variables = 0.

The norm in (2.1) in each case = 0.0

C. $m = 6$, rank $(\hat{P}) = 4$

$$\hat{P} = \begin{bmatrix} .71 & .246 & .13 & .040 & .03 & .05 \\ .02 & .052 & .06 & .015 & .02 & .01 \\ .11 & .550 & .66 & .045 & .07 & .02 \\ .03 & .038 & .04 & .010 & .01 & .01 \\ .06 & .076 & .08 & .440 & .85 & .03 \\ .07 & .038 & .03 & .450 & .02 & .88 \end{bmatrix}, \quad \hat{e} = \begin{bmatrix} .10 \\ .18 \\ .22 \\ .28 \\ .12 \\ .10 \end{bmatrix}$$

(i) Method I

$$p_y^* = \begin{bmatrix} .1248 \\ .2410 \\ .2701 \\ .1214 \\ .1233 \\ .1194 \end{bmatrix} + \begin{bmatrix} .024 & -.119 & .095 & 0 & 0 & 0 \\ -.119 & .595 & -.476 & 0 & 0 & 0 \\ .095 & -.476 & .381 & 0 & 0 & 0 \\ 0 & 0 & 0 & .667 & -.333 & -.333 \\ 0 & 0 & 0 & -.333 & .167 & .167 \\ 0 & 0 & 0 & -.333 & .167 & .167 \end{bmatrix} y$$

Thus for any y, the corresponding p_y^{**} is an optimal solution to (2.2).

(ii) Simplex Method

The optimal solution is given by

$$\hat{p}_1 = .173$$

$$\hat{p}_3 = .463$$

$$\hat{p}_4 = .360$$

$$\hat{p}_5 = .004$$

$$\lambda_1 = .084$$

$$s_6 = 0$$

$$v_{19} = 0$$

All other variables = 0.

Using a Phase I procedure, it can be shown that for

$$y = \begin{bmatrix} 0 \\ 0 \\ .505 \\ 0 \\ -.716 \\ 0 \end{bmatrix}$$

the result using Method I reduces to the above results using the Simplex method.

Using either method, the value of the norm in (2.1) is .1242.

VI. Mean Square Error of p_y^*

First consider the possible values for $\hat{P}$. If n_1 sample values are used to estimate the first column, n_2 to estimate the second column, etc., then

$$\hat{P} = \begin{bmatrix} \frac{n(1|1)}{n_1} & \cdots & \frac{n(1|m)}{n_m} \\ \vdots & & \vdots \\ \frac{n(m|1)}{n_1} & & \frac{n(m|m)}{n_m} \end{bmatrix} \qquad (6.1)$$

Where $n(i|j)$ is the number of data elements from population π_j that are classified into population π_i. Assuming that the m samples (one for each column) are independently taken, then column j has a multinomial distribution with probability

$$\frac{n_j!}{n(1|j)! \cdots n(m|j)!} \; p_{1j}^{n(1|j)} \; p_{2j}^{n(2|j)} \cdots p_{mj}^{n(m|j)}$$

Where p_{ij} = (i, j) -th element of P

$= p(i|j)$

Theorem 5.1

If $\hat{P}$ is defined as in (6.1), then there are $Z = \prod_{i=1}^{m} \binom{n_i + m - 1}{n_i}$ possible values of $\hat{P}$.

Proof: The proof will be constructed using a generating function.

Consider the expression

$$(1 + x + x^2 + \ldots + x^{n_i})^m \qquad (6.2)$$

Consider (6.2) as the product of m terms of the form $(1 + x + x^2 + \ldots + x^{n_i})$. Consequently the coefficient of x^{n_i} in the expanded expression in (6.2) will be the total number of ways of constructing the i-th comumn of $\hat{P}$. This is generally called the number of ordered partitions of n_i. Now,

$$(1 + x + x^2 + \ldots + x^{n_i})^m = \frac{(1-x^{n_i+1})^m}{(1-x)^m} \tag{6.3}$$

Also, $$\frac{1}{(1-x)^m} = \sum_{i=0}^{\infty} \binom{m-1+i}{i} x^i, \qquad |x|<$$

Thus the right hand side of (6.3) can be written

$$(1-x^{n_i+1})^m \sum_{i=0}^{\infty} \binom{m-1+i}{i} x^i. \tag{6.4}$$

The coefficient of $x^{\hat{}i}$ in (6.4) is clearly

$$z_i = \binom{n_i+m-1}{n_i}.$$

Thus the total number of possible values for $\hat{P}$ is

$$z = \prod_{i=1}^{m} \binom{n_i+m-1}{n_i}.$$

By the previous discussion, for each value of the matrix $\hat{P}$, say $\hat{P}_j$ $(1 \leq j \leq z)$, there is a corresponding probability of obtaining this $\hat{P}_j$, say q_j. Thus

$$q_j = \prod_{i=1}^{m} k_i \, p_{1i}^{n(1|i)} p_{2i}^{n(2|i)} \cdots p_{mi}^{n(m|i)}$$

where $$k_i = \frac{n_i!}{n(1|i)! \ \ldots \ n(m|i)!}$$

$$n_i = n(1|i) + n(2|i) + \ldots + n(m|i),$$

$$1 = p_{1i} + p_{2i} + \ldots + p_{mi}.$$

If p_y^* as given by (3.1) is ≥ 0 for some y, say y_o, then for some $y = y_1$ (possibly y_o), $p_{y_1}^* \geq 0$ (hence optimal by Nelson's procedure) and is also equal to the solution obtained by using the Simplex Method. If there is no y such that $p_y^* \geq 0$, then by the previous discussion, there does exist a solution which minimizes the norm in (2.1) such that at least one of the components equals zero. Setting one of the components, say p_k, equal to zero and resolving the problem amounts to deleting the k-th column from $\hat{P}$ and solving for p_y^* in (3.1), i.e. $\hat{P}$ in (3.1) is replaced by $\hat{P}$ with the k-th column removed. Thus an optimal solution of the form (3.1) always exists where it is understood that $\hat{P}$ represents the original $\hat{P}$ matrix with a certain number of columns, say r, removed where r = the number of components in p_y^* set equal to zero. Also, m is replaced by $m-r$. For the sake of discussion, in deriving the MSE for p_y^*, it will be assumed that $r = 0$.

Since $\hat{P}$ is a discrete random variable with z possible values, then the expected value of any function of $\hat{P}$ is easily derived, e.g.

$$E(\hat{P}) = \sum_{i=1}^{z} q_i \hat{P}_i$$

$$E(\hat{P}^T\hat{P}) = \sum_{i=1}^{z} q_i \hat{P}_i^T\hat{P}_i$$

$$E(C^+C) = \sum_{i=1}^{z} q_i C_i^+ C_i$$

where $$C_i = (I - \frac{1}{m} JJ^T)\, \hat{P}_i^T\, \hat{P}_i\, (I - \frac{1}{m} JJ^T)$$

We will assume here that $\hat{P}$ and $\hat{e}$ are obtained from independent samples. Furthermore, $\hat{e}$ is obtained from a sample of size n, where $n\hat{e}$ has a multinomial distribution, hence it follows that

$$\text{Cov}\,(e_i, e_j) = \begin{cases} \dfrac{e_i(1-e_i)}{n}, & i=j \\ \dfrac{-e_i e_j}{n}, & i\neq j \end{cases}$$

Consequently, as with $\hat{P}$, $\hat{e}$ has $s = \binom{n+m-1}{n}$ possible values (say $\hat{e}(1), \hat{e}(2), \ldots, \hat{e}(s)$) with corresponding probabilities q_i' $(i=1, \ldots, s)$

where $$q_i' = \frac{n!}{k_1!k_2! \cdots k_m!}\, e_1^{k_1}\, e_2^{k_2} \cdots e_m^{k_m}$$

and $$\hat{e}(i) = \begin{bmatrix} k_1/n \\ k_2/n \\ \vdots \\ k_m/n \end{bmatrix}$$

Thus it is possible to determine the expected value of any function of $\hat{P}$ and $\hat{e}$, e.g.

$$E(\hat{P}\ \hat{e}\ \hat{e}^T\ \hat{P}^T) = \sum_{i=1}^{s} \sum_{j=1}^{z} q_i' q_j\ \hat{P}_j\ \hat{e}(i)\ \hat{e}(i)^T\ \hat{P}_j^T .$$

Consequently,

$$\mu_p = E(p_y^*) = \frac{1}{m} J + (I - \frac{1}{m} JJ^T)\ y + E_1 + E_2 + E_3$$

where $E_1 = E(C^+ \hat{P}^T \hat{e})$

$$= E(C^+ \hat{P}^T)\ E(\hat{e})$$

$$= [\sum_{i=1}^{z} q_i\ C_i^+ \hat{P}_i^T]\ e$$

$$E_2 = -\frac{1}{m} E(C^+ \hat{P}^T \hat{P})\ J$$

$$= -\frac{1}{m} [\sum_{i=1}^{z} q_i\ C_i^+ \hat{P}_i^T \hat{P}_i]$$

$$E_3 = -E(C^+C)\ y$$

$$- [\sum_{i=1}^{z} q_i\ C_i^+ C_i]\ y$$

Also,

$$MSE\ (p_y^*) = E\ (p_y^* - p)(p_y^* - p)^T$$

$$= E(p_y^*\ p_y^{*T} - p\ p_y^{*T} - p_y^*\ p^T + pp^T)$$

Now, $E(p_y^* p_y^{*T}) = \{[\frac{1}{m} J + C^+ \hat{P}^T(\hat{e} - \frac{1}{m} \hat{P}J) \cdot$

$+ (I - \frac{1}{m} JJ^T - C^+C)y] \times [\frac{1}{m} J^T + (\hat{e}^T - \frac{1}{m} J^T \hat{P}^T)\hat{P}C^+ + y^T(I - \frac{1}{m} JJ^T - CC^+)]\}$

$p_y^* p_y^{*T}$ can take on $z \cdot s$ possible values considering all possible values for $\hat{P}$ and $\hat{e}$.

Thus,

$$E(p_y^* p_y^{*T}) = \sum_{i=1}^{s} \sum_{j=1}^{z} q_i' q_j \Psi_{ij} = W$$

where $\Psi_{ij} = [\frac{1}{m} J + C_j^+ \hat{P}_j^T (\hat{e}(i) - \frac{1}{m} \hat{P}_j J) + (I - \frac{1}{m} JJ^T - C_j^+ C_j)y]$

$\times [\frac{1}{m} J^T + (\hat{e}(i) - \frac{1}{m} J^T \hat{P}_j^T) \hat{P}_j C_j^+ + y^T (I - \frac{1}{m} JJ^T - C_j C_j^+)]$.

Thus $MSE(p_y^*) = W - p\,\mu_p^T - \mu_p p^T + pp^T$

and $MSE(p^*_{yi}) = w_{ii} + p_i^2 - 2\,\mu_{pi}\, p_i$

where (i) p^*_{yi} = i-th component of p_y^*

(ii) p_i = i-th component of p

(iii) μ_{pi} = i-th component of μ_p.

To illustrate the derivation of $MSE(p_y^*)$ consider the case for $m=2$, $n_1=2$, $n_2=2$, $n=3$, i.e.

$$\hat{P} = \begin{pmatrix} \frac{x}{2} & 1 - \frac{y}{2} \\ 1 - \frac{x}{2} & \frac{y}{2} \end{pmatrix} \qquad x, y = 0, 1, 2$$

$$\hat{e} = \begin{pmatrix} \frac{k}{3} \\ 1 - \frac{k}{3} \end{pmatrix} \qquad k = 0, 1, 2, 3.$$

Referring to the previous discussion, the number of possible values of $\hat{P}$ is $z = \binom{3}{2}\binom{3}{2} = 9$, and the number of possible values of $\hat{e}$ is $s = \binom{4}{3} = 4$. Assuming that an optimal solution is obtained using the entire $\hat{P}$ matrix (for same vector y), then

$$\mu_p = E(p_y^*) = \frac{1}{2} J + (I - \frac{1}{2} JJ^T) Y + \sum_{i=1}^{3} E_i$$

where, for example,

$$E_1 = \sum_{i=1}^{9} q_i C_i^+ P_i^T e$$

and $P_1 = \begin{pmatrix} 2 & 2 \\ 0 & 0 \end{pmatrix}$, $P_2 = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$, ..., $P_9 = \begin{pmatrix} 0 & 0 \\ 2 & 2 \end{pmatrix}$

$$q_1 = \binom{2}{2} p^2(1|1)\,[1 - p(1|1)]^0 \cdot \binom{2}{2} p^0(2|2)\,[1-p(2|2)]^2$$
$$\vdots$$
$$q_9 = \binom{2}{0} p^0(1|1)\,[1 - p(1|1)]^2 \cdot \binom{2}{0} p^2(2|2)]^0.$$

Thus, for example, $C_2^+ = [(I - \frac{1}{2} JJ^T)\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}(I - \frac{1}{2} JJ^T)]^+$.

So, $p_y^* p_y^{*T}$ can assume $z \cdot s. = 36$ possible values, and the $MSE(p_y^*)$ given by

$$MSE(p_y^*) = \sum_{i=1}^{4} \sum_{j=1}^{9} q_i' q_j \Psi_{ij} - p\mu_p^T - \mu_p p^T + pp^T$$

where, for example, $q_1' = \binom{3}{0} e_1^0 (1-e_1)^3$

$$q_2 = \binom{2}{2} p^2(1|1)\,[1-p(1|1)]^0 \cdot \binom{2}{1} p(2|2)\,[1-p(2|2)]$$

$$\Psi_{12} = \{\frac{1}{2} J + C_2^{+} \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix} [\begin{pmatrix} 0 \\ 3 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} J] + [I - \frac{1}{2} JJ^T - C_2^{+}C_2]y\}$$

$$x\{\frac{1}{2} J^T + [(0\ 3) - \frac{1}{2} J^T \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}] \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} C_2^{+} + y^T[I - \frac{1}{2} JJ^T - C_2C_2^{+}]\}$$

VII. CONCLUSION

It has been shown that a vector $p=p_y^*$ exists which minimizes $|| \hat{P}_p - \hat{e} ||$ subject to $J^T p=1$ and $p \geq 0$. Using the Iterative Search Technique (Nelson's procedure), the form of the solution was derived and was given by (3.1). Two methods of determining the solution vector were derived and illustrated by several examples. The modified Simplex Method appears to provide the most "efficient" solution to the problem and can be easily coded for a digital computer. The two methods provide identical (and unique) results when $\hat{P}$ is of full rank. When $\hat{P}$ is less than full rank, there will exist no unique solution; however, for some vector y, the Simplex result is identical to the result using Nelson's procedure. The MSE for p_y^* was derived in Section VI. Note that the expression for MSE (p_y^*) also provides the MSE for the solution vector provided by the Simplex Method since, for some vector y, the two methods produce identical results.

REFERENCES

[1] Boullion, T. L. and Odell, P. L. (1971), Generalized Inverse Matrices. New York: Wiley Interscience.

[2] Gottfried, B. S. and Weisman, J. (1973), Introduction to Optimization Theory. Englewood Cliffs, New Jersey: Prentice Hall, Inc.

[3] Nelson, D. L. (1969)"Quadratic Programming Techniques Using Matrix Pseudoinverses", Unpublished Ph.D. Dissertation, Texas Tech University.

AN EMPIRICAL SENSITIVITY STUDY OF MIXTURE PROPORTION ESTIMATORS

by

J. D. Tubbs [1/] and W. A. Coberly [2/]

1/ NASA/NRS Research Associate, Johnson Space Center, Houston, Texas

2/ Department of Mathematical Sciences, University of Tulsa

AN EMPIRICAL SENSITIVITY STUDY OF MIXTURE PROPORTION ESTIMATORS

J. D. Tubbs* and W. A. Coberly**

0. ABSTRACT

The sensitivity of several proposed estimators of the mixture proportions $\alpha = (\alpha_1,\ldots,\alpha_m)^T$ defining the normal mixture density $p(x;\alpha) = \sum_{k=1}^{m} \alpha_k p_k(x)$ are investigated when the component densities p_k are subjected to changes in location. The particular deviations studied are motivated by an application of this model to crop acreage assessment using satellite multispectral sensor data.

1. INTRODUCTION

A current problem in the NASA Earth Resources Program is the application of mixture proportion estimators to crop acreage assessment using multispectral sensor data from satellite platforms. The Earth Resources Technology Satellite (ERTS)

---

* NASA/NRC Research Associate, Mission Planning and Analysis Division, Johnson Space Center, Houston, Texas 77058.

** Department of Mathematical Sciences, University of Tulsa, Tulsa, Oklahoma 74104. This work was supported by NASA contract NAS-9-13512 under the University of Texas at Dallas.

records reflected (or radiated) energy in four spectral wave bands corresponding to square 80 meter plots on the ground from an altitude of approximately 500 nautical miles. In an agricultural region training areas (labeled data), consisting of known crops, are used to model or estimate the probability density functions of the existing crops. Then these density functions are used to construct a mixture model of a nearby recognition area in which no labeled data exists. That is, the mixture density of the recognition is given by

$$p(x;\alpha) = \sum_{k=1}^{m} \alpha_k p_k(x) \qquad (1)$$

where p_k is the component density and α_k the proportion of the k^{th} crop in the recognition area. The proportion vector $\alpha = (\alpha_1,\ldots,\alpha_m)^T$ must satisfy the constraints $\sum_{k=1}^{m} \alpha_k = 1$ and $\alpha_k \geq 0$ for $k=1,\ldots,m$ where m is the number of crops. In this application α_k is interpreted as the acreage proportion of crop k in the recognition area. It is assumed that the component densities p_k and the number of components m of the mixture are completely specified and α is the only unknown parameter.

Several estimators of α have been proposed [2,4,5,6] for this model. (See [1,3,7] for a discussion of the mixture estimation problem in a more general setting, i.e. when the densities p_k are not completely specified.) The following four estimators were chosen for this study.

i) Classification (CLASS) - α_k is the proportion of points in the recognition area which are classified into the

k^{th} class by a maximum likelihood classifier defined by the training densities p_k.

ii) Maximum Likelihood (MLE) - α is the vector which maximizes the likelihood function defined in (1).

iii) Moment (MOM) - α is the vector which fits the mixture moments with the empirical moments of the unlabeled data in the least squares sense.

iv) Minimum Chi-square (MIX) - α is the vector which fits the percentiles of the marginal mixture distribution functions with the corresponding empirical percentiles of the unlabeled data set in the least squares sense.

A detailed description of each estimator is given in the appendix. The estimator proposed in [5], which modifies the classification estimate by using the knowledge of the confusion matrix associated with the classification rule, was not included in this study since its behavior would exactly parallel that of the classification estimator if the confusion matrix were based only on the training area data. If additional information was known in the recognition area, then it is felt that this method would be worth considering.

In this application it is common for the components densities of the crops in the recognition area to deviate from the model which was based on the training area, even though the two areas are close geographically. For example, if the planting times of one crop, say wheat, were different in the two areas, other crops remaining fixed, then the difference in

maturity would cause a shift in the wheat distribution and the model would be inaccurate. If the data acquisition times were different, say by a day, then different sun angle and atmospheric conditions would cause a shift in all crop distributions. This study was undertaken in order to determine how each of the above estimators behave under such deviations of the model. It is assumed that no labeled data exists in the recognition area, otherwise the crop distributions would be adjusted to reflect any errors in the model.

2. SIMULATION STUDY

2.1 Description of the Simulation

The aim of this simulation study was to evaluate the effectiveness of the proportion estimators when the data in the recognition area has been "shifted" from the training area. The simulation was made as simple as possible in order to remove as many extraneous factors (sampling error, correlated variates, etc.) as possible. The study may be outlined as follows:

(a) Generate a random sample $X_1^{(i)} \ldots X_{n_i}^{(i)}$ from population π_i, where $X^{(i)} \sim MVN(\mu_i, \sigma I)$ for $i=1,2,3$ when n_i, μ_i and σ are known fixed parameters.

(b) Train the proportion estimators using the training segment.

(c) Create a new data set (recognition segment)

$Y_1 \ldots Y_N$ where $Y = X^{(i)} + dv_i$ when $N = \sum_{i=1}^{3} n_i$, d is a known scalar and v_i is a fixed direction vector for population π_i.

(d) Determine the proportion estimates in the new recognition segment and evaluate the estimators using the sum of squared error deviation from the ground truth.

In this study $\mu_1 = (30,30)^T$, $\mu_2 = (40,40)^T$, $\mu_3 = (40,20)^T$ and $\sigma = 7$. Using these parameters we defined two experiments, (1) $n_1 = 180$, $n_2 = 75$, $n_3 = 45$, (2) $n_1 = n_2 = n_3 = 100$. For both experiments we defined the following three sets of direction vectors:

(1) $v_1 = (1,0)^T$, $v_2 = (0,0)^T$, $v_3 = (0,0)^T$ and $d = 3,6,9$.

(2) $v_1 = (1,1)^T$, $v_2 = (0,0)^T$, $v_3 = (0,0)^T$ and $d = 3,6,9$.

(3) $v_1 = v_2 = v_3 = (1,0)^T$ and $d = 3,6,9$.

Each case was repeated three times for each experiment using an independent data set. See Fig. 1

2.2 Results

The results of the simulation have been summarized in Figures 1-6 and Tables 1-2. The values found in both the figures and tables represent the average mean-squared error and crop proportions, taken over the three data sets. In Figures 1-6 the mean-squared error is plotted against the distance the recognition area was shifted. Tables 1-2 give

FIG 1

CASE 1 EXPERIMENT

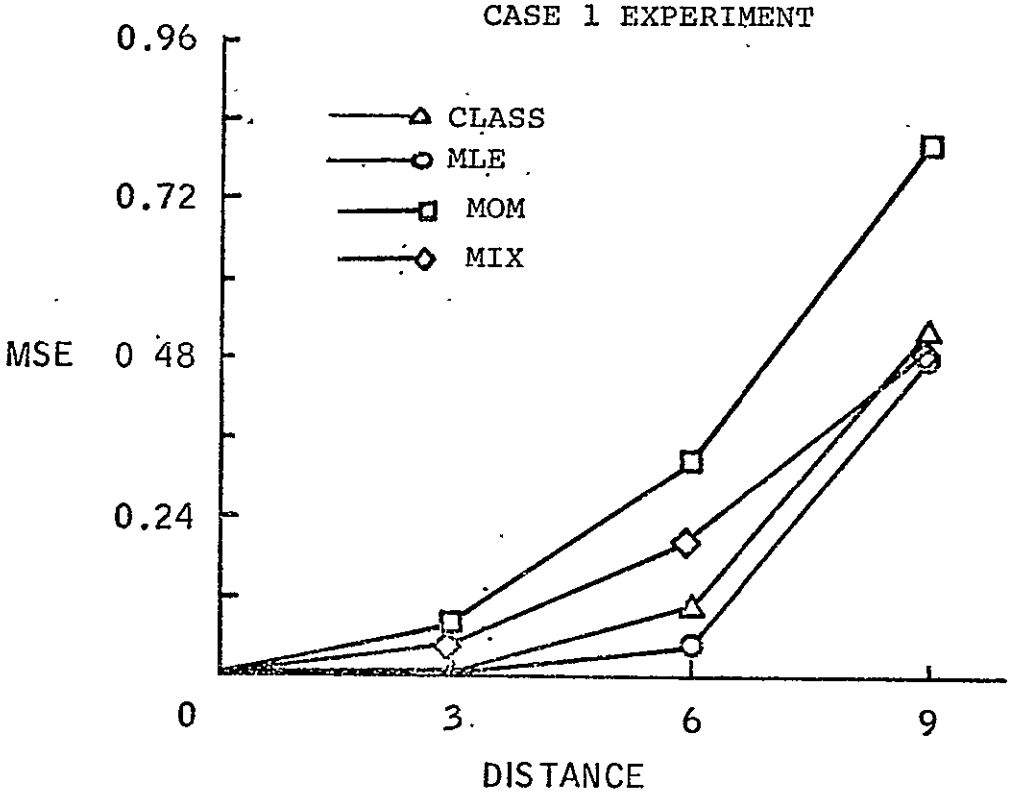


FIG 2

CASE 2 EXPERIMENT 1

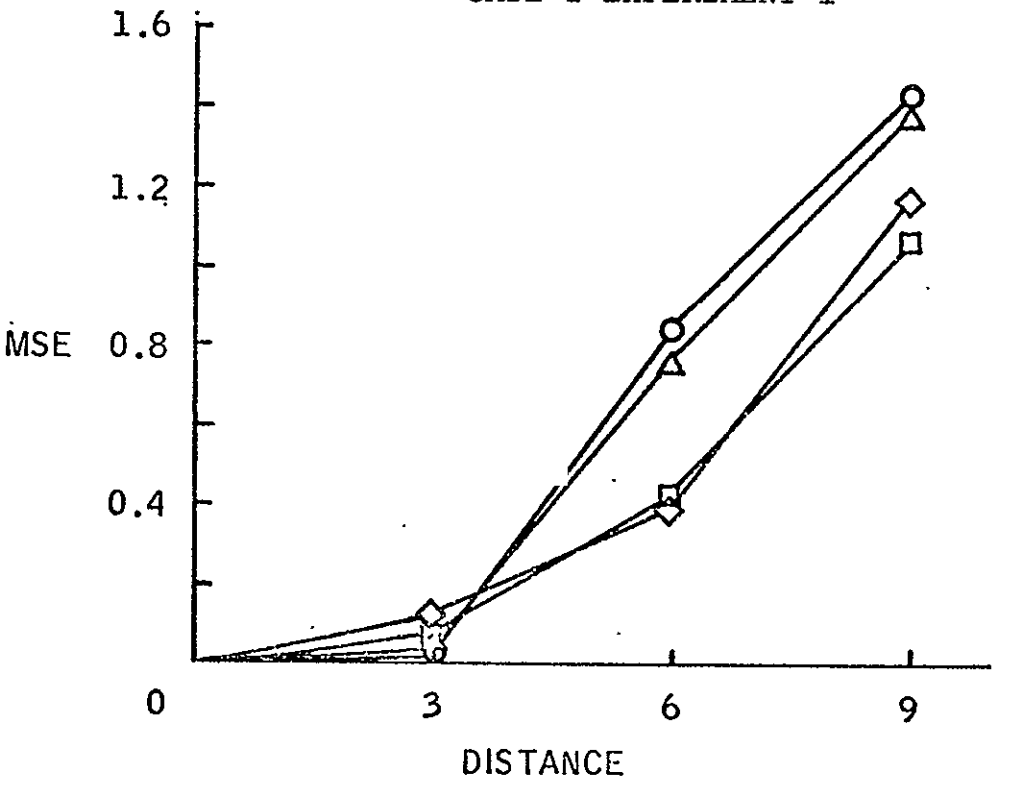


FIG 3

CASE 3 EXPERIMENT 1

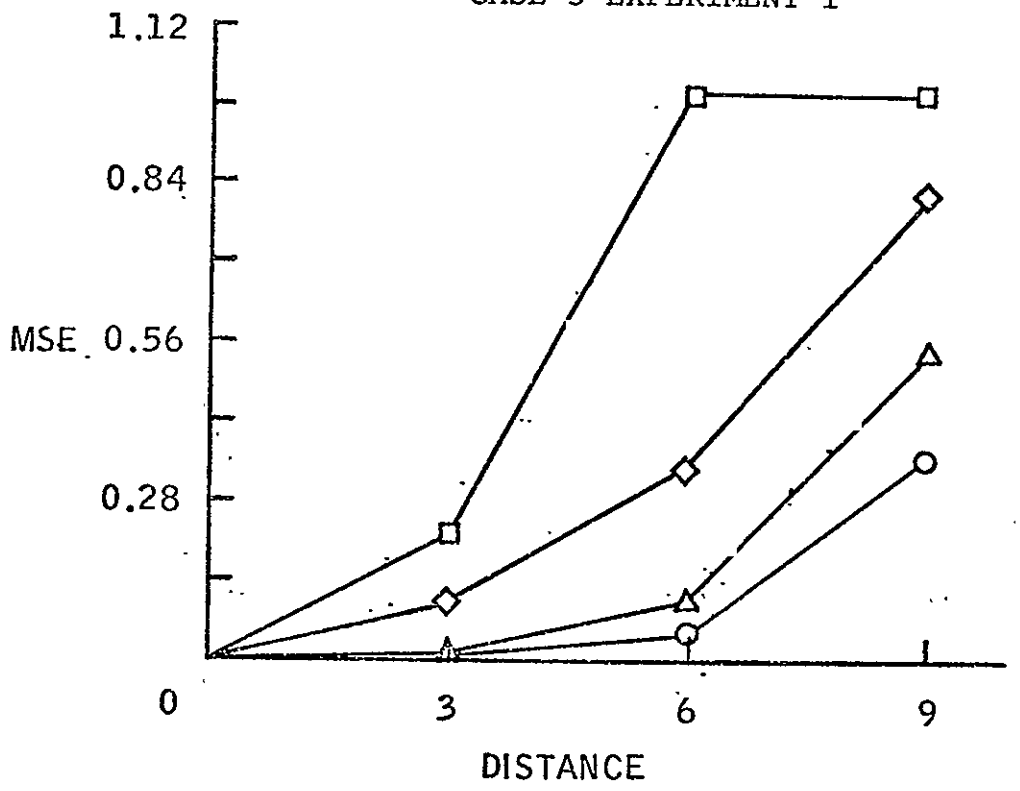


FIG 4

CASE 1 EXPERIMENT 2

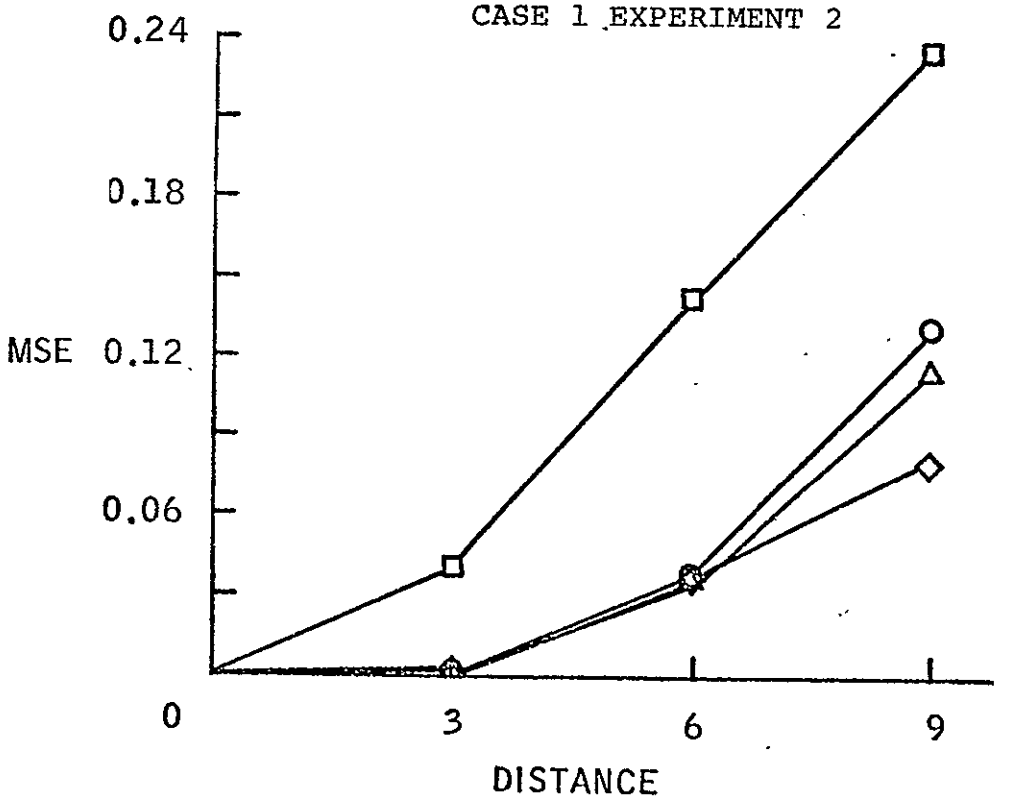


FIG 5

CASE 2 EXPERIMENT 2

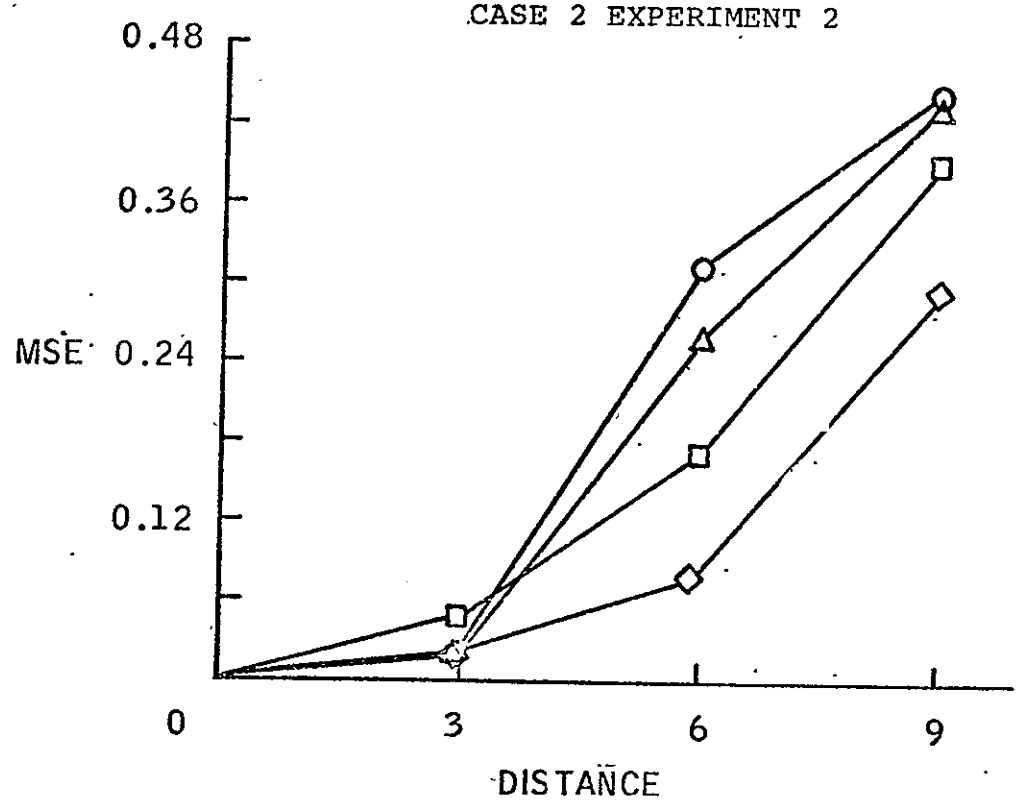


FIG 6

CASE 3 EXPERIMENT 2

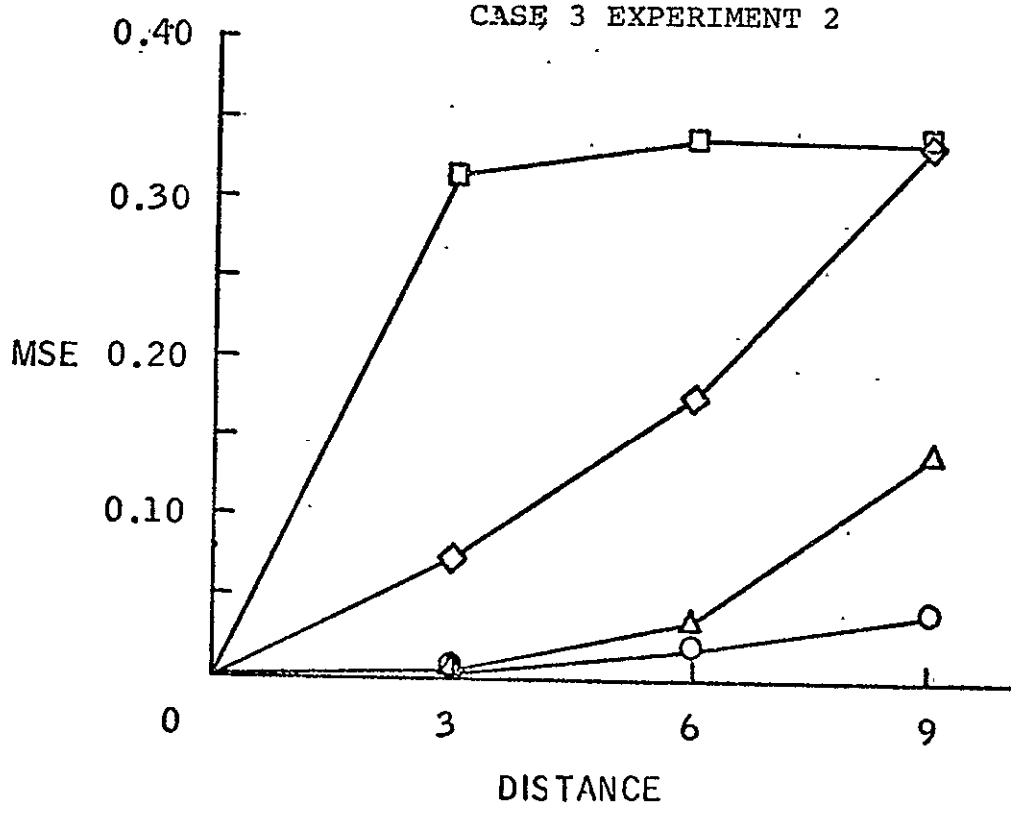


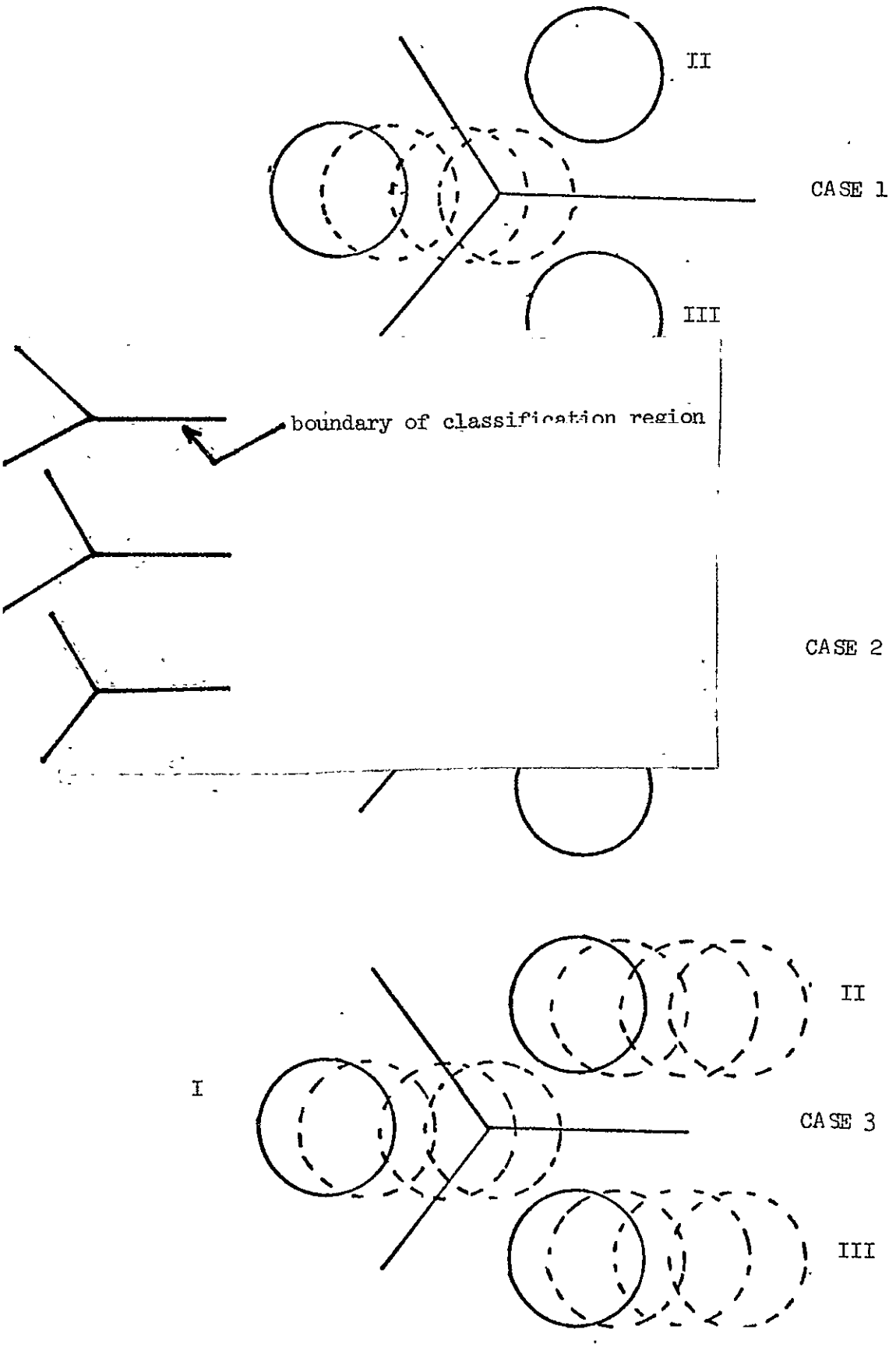
TABLE I

CASE	DIST.	POP.	CLASS	MLE	MOM	MIX	ANS.
1.1	0.	I	.5956	.6036	.6027	.6000	.6000
		II	.2533	.2491	.2431	.2500	.2500
		III	.1511	.1472	.1494	.1500	.1500
	3.	I	.5600	.5773	.4368	.4640	
		II	.2711	.2630	.3216	.3155	
		III	.1689	.1598	.2416	.2205	
	6.	I	.4122	.4560	.2565	.3536	
		II	.3422	.3267	.4035	.3663	
		III	.2456	.2173	.3400	.2801	
	9.	I	.1811	.1954	.0677	.2052	
		II	.4300	.4493	.4886	.4305	
		III	.3889	.3553	.4437	.3643	
1.2	3.	I	.4878	.5188	.4412	.4525	
		II	.3622	.3342	.4094	.4474	
		III	.1500	.1471	.1494	.1002	
	6.	I	.1644	.1364	.2626	.3034	
		II	.6878	.7153	.5877	.5676	
		III	.1478	.1483	.1497	.1291	
	9.	I	.0122	.0005	.0727	.0661	
		II	.8411	.8500	.7723	.7639	
		III	.1467	.1494	.1500	.1700	
1.3	3.	I	.5567	.5745	.3023	.4140	
		II	.2711	.2626	.3825	.3370	
		III	.1722	.1628	.3153	.2488	
	6.	I	.4089	.4651	.0072	.2707	
		II	.3422	.3125	.5171	.4039	
		III	.2489	.2224	.4757	.3254	
	9.	I	.1778	.2635	.0000	.0953	
		II	.4300	.3939	.5350	.4815	
		III	.3900	.3426	.4650	.4232	

TABLE II

CASE	DIST.	POP.	CLASS	MLE	MOM	MIX	ANS.
2.1	0.	I	.3322	.3357	.3260	.3333	.3333
		II	.3333	.3324	.3390	.3333	.3333
		III	.3344	.3320	.3350	.3333	.3333
	3.	I	.3133	.3148	.2307	.2824	
		II	.3433	.3458	.3827	.3588	
		III	.3433	.3394	.3866	.3588	
	6.	I	.2300	.2272	.1304	.2223	
		II	.3900	.3945	.4283	.3898	
		III	.3800	.3782	.4413	.3879	
	9.	I	.1017	.0867	.0070	.1407	
		II	.4417	.4625	.4797	.4342	
		III	.4567	.4510	.5133	.4251	
2.2	3.	I	.2656	.2684	.2335	.2731	
		II	.4000	.3996	.4317	.4144	
		III	.3344	.3320	.3349	.3126	
	6.	I	.0756	.0532	.1342	.1955	
		II	.5900	.6137	.5310	.4819	
		III	.3344	.3331	.3348	.3226	
	9.	I	.0022	.0008	.0296	.0628	
		II	.6644	.6674	.6326	.6074	
		III	.3333	.3319	.3341	.3297	
2.3	3.	I	.3122	.3133	.0173	.1769	
		II	.3444	.3485	.4787	.4168	
		III	.3433	.3381	.5039	.4063	
	6.	I	.2289	.2549	.0000	.0913	
		II	.3911	.3790	.5009	.4610	
		III	.3800	.3661	.4991	.4476	
	9.	I	.1056	.2124	.0000	.0000	
		II	.4400	.3886	.5079	.5081	
		III	.4544	.3991	.4872	.4870	

FIG. 7


II
CASE 1
III
boundary of classification region
CASE 2
I
II
CASE 3
III

the crop proportions found using the proportion estimators at the various distances.

3. CONCLUSIONS

Since the training model consists of component classes in which the interclass confusion is small, all estimators do well when the model is exact (d=0). Further the behavior of the estimators is similar under both experiments. That is, the true proportion vector does not seem to be a definite factor in the sensitivity analysis. Now consider each case separately. In case 1 the CLASS, MLE, and MIX estimators perform similarly with the first two better for d = 3,6. The MOM estimator is uniformly the most sensitive to the model deviations. In case 2 when the first class is shifted toward the second the likelihood based estimators (CLASS and MLE) are worse as the 1st class crosses the classification boundary, and the MOM and MIX perform similarly. In case 3 the MOM and MIX estimators are more sensitive than the CLASS and MLE.

In conclusion, based on the types of deviations considered here, it appears that the ordering of the four estimators, according to the degree of sensitivity, which is suggested by this experiment would be (CLASS, MLE) $\geq$ MIX $\geq$ MOM. It is also apparent however that the particular type of shift deviation from the model may give a different ordering. Therefore, if the suspected deviation is known to be of one particular type

or direction, then a ~~specialized~~ specified experiment should be ~~ran~~ run to investigate the sensitivity under this alternative.

4. APPENDIX

4.1 Classification (CLASS)

Define

$$\chi_k(X) = \begin{cases} 1 & \text{if } p_k(X) \geq p_j(X) \quad \neq \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1,\ldots,m$. Now the CLASS estimate is given by

$$\hat{\alpha}_k = \frac{1}{N} \sum_{j=1}^{N} \chi_k(X_j) \qquad k = 1,\ldots,m,$$

where $X_1,\ldots,X_N$ is the sample of unlabeled data. $\hat{\alpha}_k$ is simply the proportion of the sample which is classified into the k^{th} class by the maximum-likelihood classifier.

4.2 Maximum-Likelihood (MLE)

A necessary condition for $\hat{\alpha}$ to be a maximum-likelihood estimator of α is that

$$\hat{\alpha}_k = \frac{1}{N} \sum_{j=1}^{N} \hat{\alpha}_k p_k(X_j)/p(X_j;\hat{\alpha})$$

for $k = 1,\ldots,m$. See [6] for a more detailed discussion of this estimator.

4.3 Moment (MOM)

For each class k let μ_j^k, $j = 1,\ldots,n+\frac{n(n+1)}{2}$ denote the first and second order noncentral moments. (Here n is the dimension of the multivariate observations and μ_j^k is the mean of the j^{th} variable for $j = 1,\ldots,n$ and one of the distinct

elements of the noncentral second moment matrix for $j = n+1, \ldots, n+\frac{n(n+1)}{2}$.) Let χ_j be the corresponding sample moment from the unlabeled data for $j = 1,\ldots, n+\frac{n(n+1)}{2}$. Now construct the system of equations

$$w_j \bar{\chi}_j = w_j \sum_{k=1}^{m} \alpha_k \mu_j^k \qquad j = 1,\ldots,n+\frac{n(n+1)}{2}$$

where the weights w_j are proportional to the variance of the sample moment $\bar{\chi}_j$. Now the MOM estimator is defined to be the constrained least squares solution of this linear system.

4.4 Minimum Chi-square (MIX)

For each class k let F_j^k be the j^{th} marginal distribution function and let $\hat{F}_j$ be the j^{th} marginal empirical distribution function of the unlabeled data for $j = 1,\ldots,n$. Let $\tilde{\chi}_{ij}$ be the $100\frac{i}{s+1}$ percentile of $\hat{F}_j$ for $i = 1\ldots,s$. Now construct

$$\hat{F}_j(\tilde{\chi}_{ij}) = \sum_{k=1}^{m} \alpha_k F_j^k(\tilde{\chi}_{ij})$$

for $i = 1,\ldots,s$ and $j = 1,\ldots,n$. (In this experiment $s = 9$ and $n = 4$). The MIX estimator is defined to be the constrained least squares solution to this linear system.

4.5 Odell-Chhikara (O-C)

Let $P = (p_{ij})$ be the confusion matrix defined by the maximum likelihood classification procedure where

$$p_{ij} = Pr[\chi_i(x) = 1 | x \in \pi_j].$$

Since $E(\hat{\alpha}) = P\alpha$, where $\hat{\alpha}$ is the estimator defined by CLASS, the O-C estimator is given by

$$\hat{\hat{\alpha}} = \hat{P}^{-1}\hat{\alpha}$$

when $\hat{P}$ is an estimate of the confusion matrix.

REFERENCES

[1] Day, N. E., Estimating the components of a mixture of normal distributions, Biometrika, 56, 3, p. 463 (1969).

[2] Hartley, H. O., The estimation of acreages from satellite data, Tech. Report, NASA-JSC, 1974.

[3] Hasselblad, V., Estimation of parameters for a mixture of normal distributions, Technometrics, 8, No. 3, p. 431 (1966).

[4] Odell, P. L. and Basu, J. P., Concerning several methods for estimating crop acreages using remote sensing data, Tech. Report, NASA-JSC, 1974.

[5] Odell, P. L. and Chhikara, R., Estimation of a large area crop acreage inventory using remote sensing technology, Tech. Report, NASA-JSC, 1974.

[6] Peters, C. and Coberly, W. A., The numerical evaluation of the maximum-likelihood estimate of mixture proportions, Tech Report, NASA-JSC, 1975.

[7] Wolfe, J. H., Pattern classification by multivariate mixture analysis, Multivariate Behavioral Research, 5, p. 329, (1970).

THE NUMERICAL EVALUATION OF THE MAXIMUM-LIKELIHOOD ESTIMATE OF MIXTURE PROPORTIONS

by

Charles Peters+ and William A. Coberly++

+ NASA/NRC Research Associate, Johnson Space Center, Houston, Texas 77058

++ Division of Mathematical Sciences, University of Tulsa, Tulsa, Oklahoma 74104.
This work was begun while the author was a NASA/NRC Research Associate at the Johnson Space Center and completed under support of NASA contract NAS 9-13512.

ABSTRACT

Let $p_1, \ldots, p_m$ be multivariate density functions and let $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ be a proportion vector defining the mixture density $p(x,\alpha) = \Sigma_{i=1}^{m} \alpha_i p_i(x)$, where $\Sigma_{i=1}^{m} \alpha_i = 1$ and $\alpha_i \geq 0$ for $i = 1, \ldots, m$. This paper investigates the problem of evaluation of the maximum-likelihood estimate for α. An acreage estimation application is presented using remotely sensed data.

1. Introduction

Let $\pi_1, \ldots, \pi_m$ be m pattern classes having distinct multivariate probability density functions $p_1, \ldots, p_m$ respectively. Let $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ be the proportion vector defining the mixture density

$$p(x; \alpha) = \sum_{k=1}^{m} \alpha_k p_k(x) \tag{1}$$

where $\sum_{k=1}^{m} \alpha_k = 1$ and $\alpha_k \geq 0$ $k = 1, \ldots, m$. Let $\mathcal{X} = \{X_1, \ldots, X_N\}$ be a random sample drawn from the mixture density p. A necessary condition for a vector

$$\beta \in S = \{\gamma \in R^m : \Sigma \gamma_i = 1;\ \gamma_i \geq 0,\ i = 1, \ldots, m\}$$

to be the maximum-likelihood estimator (MLE) of α in (1) is well known (see [2,p.192]). However, the problem is usually considered in the context of the general mixture problem, where the component density functions and the proportion vector are simultaneously estimated from the mixture sample X. In this paper the component densities are assumed to be completely specified. This is an appropriate mathematical model even when labeled data, independent of $\mathcal{X}$, is used to estimate the component densities prior to drawing the mixture of unlabeled sample. The following is included for completeness.

Theorem 1 A necessary condition for $\beta \in S$ to be a MLE of α in (1) is

$$\beta_k = \frac{1}{N} \sum_{i=1}^{N} \beta_k p_k(X_i)/p(X_i; \beta) \tag{2}$$

for $k = 1, \ldots, m$. That is, the MLE must be a solution of the fixed point equation

$$\beta = G(\beta), \tag{3}$$

where G is a vector valued function defined component-wise by the RHS of (2).

Proof: The log-likelihood function

$$l(\beta) = \sum_{j=1}^{N} \log \sum_{i=1}^{m} \beta_i p_i(X_j)$$

is concave and the constraint set S is compact and convex. Hence, $\beta \in S$ maximizes

l on S if and only if $\nabla l\ (\beta)(\gamma - \beta) \leq 0$ for all $\gamma \in S$. That is, if and only if

$$\frac{\partial l}{\partial \beta_k}(\beta) \leq \nabla l(\beta)\,\beta \qquad k = 1, \ldots, m.$$

It is easily verified that $\nabla l(\beta)\,\beta = N$ and thus β maximizes l on S if and only if

$$(4) \qquad \frac{1}{N}\frac{\partial l}{\partial \beta_k}(\beta) = \frac{1}{N}\sum_{j=1}^{N}\frac{p_k(X_j)}{p(X_j;\beta)} \leq 1$$

$k = 1, \ldots, m$, with equality whenever $\beta_k > 0$. Multiplying both sides of these inequalities by the corresponding β_k yields the required necessary condition.

2. An Iterative Method for Obtaining the MLE.

The fixed point equation (3) suggests that a possible method of obtaining the MLE, say α^*, is to iterate G and form the sequence of successive approximations $\beta^n = G^n(\beta)$, where β is the initial guess of α^*. The purpose of this section is to provide partial theoretical justification for its use.

Let S^0 denote the relative interior of the constraint set S. We will assume that the $N \times m$ matrix $P = (p_i(X_j))$ has rank m so that the Hessian of the log-likelihood function

$$(5) \qquad H(\beta) = -\sum_{j=1}^{N}\begin{pmatrix} p_1(X_j)^2 & \cdots & p_1(X_j)p_m(X_j) \\ \vdots & & \\ p_m(X_j)\,p_1(X_j) & \cdots & p_m(X_j)^2 \end{pmatrix} \Big/ p(X_j,\beta)^2$$

is negative definite for all $\beta \in S$ and consequently the MLE α^* is unique. Moreover, with this assumption, both S and S^0 are invariant under G. In practice, N is very much larger than m so that the assumption that P has rank m is almost always satisfied.

Note that G has fixed points other than α^*. In particular, each vertex of S is fixed by G. The next theorem shows that G is unstable near these extraneous solutions

of the likelihood equations.

Theorem 2 If $\beta \in S^0$ and $\hat{\beta} = \lim_{n \to \infty} G^n(\beta)$, then $\hat{\beta} = \alpha^*$.

Proof: If $\hat{\beta}$ is not a MLE, then since $\hat{\beta} = G(\hat{\beta})$ it follows from (4) that $\frac{1}{N} \frac{\partial l}{\partial \beta_k} (\hat{\beta}) > 1$ for some k such that $\hat{\beta}_k = 0$. Since $\hat{\beta} = \lim \beta^n$, for large n we have

$$\frac{1}{N} \frac{\partial l}{\partial \beta_k} (\beta^n) > 1$$

and $\beta_k^n > 0$, since $\beta \in S^0$. Therefore,

$$\beta_k^{n+1} = \beta_k^n \cdot \frac{1}{N} \frac{\partial l}{\partial \beta_k} (\beta^n) > \beta_k^n$$

for sufficiently large n. It follows that β_k^n cannot converge to $0 = \hat{\beta}_k$, a contradiction.

The next theorem shows that under certain restrictions, G is a local contraction at the MLE α^*. That is, for some norm on R^m there exists $\epsilon > 0$ and k, $0 \leq k < 1$. Such that

$$\| \alpha^* - G(\beta) \| \leq k \| \alpha^* - \beta \|$$

whenever $\| \alpha^* - \beta \| < \epsilon$. Thus if β is sufficiently near α^*, the sequence $\beta^n = G^n(\beta)$ converges geometrically to α^*. The proofs of Theorems 3 and 4 are given in the Appendix.

Theorem 3 If the rank of P is m and the MLE α^* is in S^0, then G is a local contraction at α^*.

For the two class case, we have the following somewhat stronger result.

Theorem 4 If $m = 2$ and $\beta \in S^0$ then $G^n(\beta)$ converges to α^*.

3. Choosing a Starting Vector:

It is clear that any starting vector β should be in S^0. If no other information is available, then a naive starting vector would be $\beta = \frac{1}{m} (1, \ldots, 1)^T$. If the com-

component densities are widely separated, then the probability of error P_e of the maximum-likelihood classifier is small and the proportion of observations classified into each component class would be approximately the MLE. That is, let

$$\chi_k(x) = \begin{cases} 1 & \text{if } p_k(x) > p_j(x) \text{ for } j \neq k \\ 0 & \text{otherwise} \end{cases}$$

and

$$\beta_k = \frac{1}{N} \Sigma_{j=1}^{N} \chi_k(X_j) \ , \ k = 1, \ldots, m.$$

Even if P_e is not negligable, β should serve as a very good starting vector providing $\beta \in S^0$.

4. An Example:

In this section, the previous results are applied to the problem of estimating crop acreage from satellite data. The data used was taken by the multispectral scanner aboard the NASA launched, Earth Resources Technology Satellite (ERTS). This sensor records reflected (or radiated) energy in four spectral wave bonds corresponding to square 80 meter plots on the ground from an altitude of approximately 500 nautical miles. Five crops were identified and their density functions were estimated from labeled data using a multivariate normal model. In Case I, all five crops were represented by at least 10% of the total mixture (unlabeled) sample of 1000 points. In Case II, the fourth crop was retained in the model, but deleted from the mixture sample.

The seven starting vectors which were tried in each case are listed in Table I, along with the MLE in each case. All starting points converged to the MLE in both cases under the iterative procedure. The first two starting vectors are those proposed in section III.

The last five were chosen near vertices of S to illustrate the convergence rate when a "poor" choice of starting point is used.

		CROP			
MLE	1	2	3	4	5
I	.2937060	.3616114	.1581606	.1000086	.0865135
II	.3224670	.4115446	.1745424	.0032386	.0882074

	β	1	2	3	4	5
1	I	.278000	.278000	.170000	.106000	.168000
	II	.304444	.306667	.186667	.023333	.178889
2		.20	.20	.20	.20	.20
3		.01	.01	.01	.01	.96
4		.01	.01	.01	.96	.01
5		.01	.01	.96	.01	.01
6		.01	.96	.01	.01	.01
7		.96	.01	.01	.01	.01

TABLE 1 Starting Vectors

In Table 2 the number of iterations required for 2,3, and 4 place accuracy (using the sup-norm)are noted for each starting vector. That is, the table entry is the number of iterations n for which

$$\max_{j=1,\ldots,5} | \alpha_j^* - G_j^n(\beta) | \leq .5 \times 10^{-k} \quad \text{for } k = 2,3,4 .$$

β	CASE I			CASE II		
	k = 2	3	4	2	3	4
1	20	37	55	18	33	49
2	22	40	58	20	36	52
3	31	49	66	29	44	60
4	21	38	56	19	35	51
5	22	40	58	21	36	52
6	29	48	66	26	42	58
7	22	40	58	20	36	52

TABLE 2. Iterations needed for k place accuracy using starting vector β

For these two examples the iteration procedure appears to be very stable with only starting vectors 3 and 6 requiring more than 3 iterations more than the best starting vector.

C,2

APPENDIX

Proof of Theorem 3: The Frechet derivative $G'(\beta)$ of G is represented by the mxm matrix $(\frac{\partial G_k}{\partial \beta_j}(\beta))$. From (2) it follows after a brief calculation that

$$G'(\beta) = \text{diag}\left(\frac{1}{N}\ \frac{\partial I}{\partial \beta_k}(\beta)\right) + \frac{1}{N}\ \text{diag}(\beta_k)\ H(\beta) \tag{6}$$

where $H(\beta)$ is the Hessian of I given in (5). Since $G'(\beta)$ is a continuous function of β, it follows from the mean value theorem that G is a local contraction at α^*, if, with respect to some norm on R^m.

$$\| G'(\alpha^*) \| < 1$$

where $\| G'(\alpha^*) \| = \sup_{\|\beta\|=1} \|G'(\alpha^*)\ \beta\|$. By [2,p.46] this is true if and only if the spectral radius $\rho(G'(\alpha^*))$ is less than 1. Since α^* is the MLE and $\alpha^* \in S^0$, it follows from (4) and (6) that

$$G'(\alpha^*) = I + \frac{1}{N}\ \text{diag}(\alpha_k^*)\ H(\alpha^*).$$

Since the eigenvalues of $\text{diag}\ (\alpha_k^*)\ H(\alpha^*)$ are negative, $\rho(G'(\alpha^*)) < 1$ if and only if $\rho(\text{diag}(\alpha_k^*) H(\alpha^*)) < 2N$. The entries in the matrix $\text{diag}\ (\alpha_k^*) H(\alpha^*)$ are all ≤ 0 and it is easily verified that

$$\text{diag}(\alpha_k^*) H(\alpha^*)\ \alpha^* = -N\alpha^*.$$

It follows by Frobenius' Theorem, [5,p.49], that the spectral radius of $\text{diag}(\alpha_k^*) H(\alpha^*)$ is N. Hence $\rho(G'(\alpha^*)) < 1$ and the proof is complete.

Proof of Theorem 4: We will show that whenever $\beta \in S^0$ and $\beta \neq \alpha^*$, then

$$|G_1(\beta) - \alpha_1^*| < |\beta_1 - \alpha_1^*| \tag{7}$$

from which the theorem follows. From the concavity of I we have

$$0 < \nabla l(\beta)(\alpha^* - \beta) = (\alpha_1^* - \beta_1)\left(\frac{\partial l}{\partial \beta_1}(\beta) - \frac{\partial l}{\partial \beta_2}(\beta)\right).$$

Hence, $\beta_1 < \alpha_1^*$ if and only if $\frac{\partial l}{\partial \beta_1}(\beta) > \frac{\partial l}{\partial \beta_2}(\beta)$. Since $\beta_1 \frac{\partial l}{\partial \beta_1} + \beta_2 \frac{\partial l}{\partial \beta_2} = N$, $\beta_1 < \alpha_1^*$ if and only if $\frac{\partial l}{\partial \beta_1}(\beta) > N$. Similarly $\beta_1 > \alpha_1^*$ if and only if $\frac{\partial l}{\partial \beta_1}(\beta) < N$.

Assuming that $\beta_1 < \alpha_1^*$,

$$G_1(\beta) = \frac{\beta_1}{N} \frac{\partial l}{\partial \beta_1}(\beta) > \beta_1$$

and (7)' follows if $G_1(\beta) < \alpha_1^*$; i.e., if $\frac{\partial l}{\partial \beta_1}(G(\beta)) > N$.

But,

$$\frac{\partial l}{\partial \beta_1}(G(\beta)) = \sum_{j=1}^{N} \frac{p_1(X_j)}{G_1(\beta)p_1(X_j) + G_2(\beta)p_2(X_j)}$$

$$= \sum_{j=1}^{N} \frac{p_1(X_j)}{\beta_1 p_1(X_j)\frac{1}{N}\frac{\partial l}{\partial \beta_1} + \beta_2 p_2(X_j)\frac{1}{N}\frac{\partial l}{\partial \beta_2}}$$

$$> \sum_{j=1}^{N} \frac{p_1(X_j)}{(\beta_1 p_1(X_j) + \beta_2 p_2(X_j))\frac{1}{N}\frac{\partial l}{\partial \beta_1}}$$

$$= N.$$

Therefore, $G^n(\beta)$ converges to α^*.

REFERENCES

[1] Day, N.E., "Estimating the components of a mixture of normal distributions," Biometrika, 56, 3, p. 463 (1969).

[2] Duda, R. O. and Hart, P. E., Pattern Classification and Scene Analysis, Wiley, 1973.

[3] Hasselblad, V., "Estimation of parameters for a mixture of normal distributions," Technometrics, 8, No.3, p. 431 (August, 1966).

[4] Hasselblad, V., "Estimation of finite mixtures of distributions from the exponential family," J. Amer. Statist. Assoc., 64, p. 1459, (1969).

[5] Householder, A.S., The Theory of Matrices in Numerical Analysis, Blaisdell, 1965.

[6] Kale, B. K., "On the solution of likelihood equations by iteration processes. The multiparametric case," Biometrika, 49, 3, p. 470, (1962).

[7] Wolfe, J. H., "Pattern clustering by multivariate mixture analysis," Multivariate Behavioral Research, 5, pp 329-350 (July 1970).

SOME RESULTS ON RANDOMLY MISSING DATA IN DISCRIMINANT ANALYSIS

by

T. L. Boullion*

*Mathematics Department, Texas Tech University

Some Results on Randomly Missing Data in Discriminant Analysis

by T.L. Boullion

Several alternatives are available for handling randomly missing data from multispectral scanner measurement acquired over an agricultural area. One such procedure is described in [1]. In this report, other methods are investigated and compared using the asymtotic unconditional probability of correct classification.

The five methods considered are (1) Use only the complete observation vectors. All vectors with any missing components are discarded. The linear discriminant function is calculated in the usual manner with only the complete observation vectors. (2) Use all sample values in computing means and variances for each variable, but only complete pairs to compute covariances. Then the linear discriminant function is calculated with the sample covariance matrix formed from these statistics. (3) Compute means based on all available values and substitute these mean values for the missing values. Calculate the discriminant function from this completed set of vectors. (4) Use the estimate from the regression equations, regressing the missing values on those which are not missing, to complete the vectors with missing components. The linear discriminant function is

calculated from this set of completed vectors. (5) Use the first principal component from each of the two sample matrices to estimate the missing values as follows: Let X_1 $(p \times m_1)$ and X_2 $(p \times m_2)$ be the data matrices whose columns are the observation vectors on two crops of interest. Replace X_i by Y_i, $i = 1,2$ where $y_{jk} = (x_{jk} - \bar{x}_j)/\sqrt{s_{jj}}$ for known values and $y_{jk} = 0$ for missing values. $\bar{x}_j$ is the sample mean of the j^{th} variate and s_{jj} is the pooled variance of the j^{th} variate. The coefficients of the first principal component of Y_i are then obtained, say $q_1^{(i)}$ where $q^{(i)} = (q_1^{(i)},\ldots,q_p^{(i)})^T$ is the eigenvector of unit length associated with the largest eigenvalue of $Y_i Y_i^T$. Missing values in Y_i are replaced by the coordinate of the nearest point on the first principal component. That is, for each data matrix, Y_{jk} is replaced by $a_K q_j$, where $a_K = \sum_{j=1}^{p} Y_{jk} q_j$. After all missing values are estimated, Y_1 and Y_2 are transformed back to their original units X_1 and X_2, and the usual discriminant function calculated.

To compare the methods, the actual probability of correct classification using the usual discriminant function based on a single pair of samples is used. The expression is

$$p = \frac{1}{2}\,\Phi\left\{\frac{[\mu^{(1)} - \frac{1}{2}(\bar{X}^{(1)}+\bar{X}^{(2)})]^T S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})}{\sqrt{(\bar{X}^{(1)}-\bar{X}^{(2)})^T S^{-1}\Sigma S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})}}\right\} +$$

$$\frac{1}{2}\,\Phi\left\{\frac{-[\mu^{(2)}-\frac{1}{2}(\bar{X}^{(1)}+\bar{X}^{(2)})]^T S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})}{\sqrt{(\bar{X}^{(1)}-\bar{X}^{(2)})^T S^{-1}\Sigma S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})}}\right\} \qquad (1)$$

In the above $\mu^{(i)}$, i = 1-2 are the population means, Σ is the population covariance matrix, $\bar{X}^{(i)}$ is the vector estimate of $\mu^{(i)}$ from the particular missing value method used, and S is the pooled estimate of Σ from the particular missing value method.

The asymptotic limits of $\bar{X}^{(1)}$, $\bar{X}^{(2)}$ and S for each missing value method were obtained and substituted into (1) to obtain numerical values for the asymptotic probability of correct classification which equals E_p. The variances were taken equal to one without any loss of generality and all correlation coefficients were taken equal; hence $\Sigma = R = (\rho)$.

Methods (1) and (2) attain the maximum probability of correct classification, hence only methods (3), (4) and (5) were compared. The Mahalanobis distance $\Delta^2 = \mu^T R^{-1} \mu$ between the two populations was taken to be 4, and the proportion of missing values was taken to be m = .2. The correlation ρ was restricted to being greater than $-1/(p-1)$ so that the equicorrelation matrix would be positive definite.

Partial results are given in table I.

TABLE I

Asymptotic Probability of Correct Classification, E_p

p	method	ρ						
		.8	.5	.2	0	-.1	-.2	-.4
		(a) $\mu = (d,0,\ldots,0)^T$						
2	(3)	.8330	.8397	.8411	.8413			

	(4)	.8399	.8400	.8411	.8413	----	----	----
	(5)	.8413	.8405	.8388	.8370	----	----	----
3	(3)	.8370	.8410	.8411	.8413	----	.8406	.8299
	(4)	.8410	.8409	.8411	.8413	----	.8407	.8388
	(5)	.8413	.8410	.8400	.8381	----	.8325	.8022
4	(3)	.8387	.8404	.8411	.8413	----	----	----
	(4)	.8412	.8411	.8412	.8413	----	----	----
	(5)	.8413	.8412	.8405	.8388	----	----	----
8	(3)	.8406	.8410	.8412	.8413	.8376	----	----
	(4)	.8413	.8413	.8412	.8413	.8374	----	----
	(5)	.8413	.8413	.8411	.8400	.8292	----	----

(b) $\mu = (d,d,0)^T, (d,d,0,0)^T, (d,d,d,d,0,0,0,0)^T$ for p = 3,4 and 8, respectively.

3	(3)	.8300	.8390	.8410	.8413	----	.8408	.8359
	(4)	.8404	.8402	.8410	.8413	----	.8410	.8405
	(5)	.8413	.8406	.8390	.8372	----	.8332	.8158
4	(3)	.8348	.8394	.8409	.8413	----	.8391	----
	(4)	.8410	.8407	.8410	.8413	----	.8400	----
	(5)	.8413	.8409	.8397	.8373	----	.8265	----
8	(3)	.8370	.8396	.8406	.8413	.8374	----	----
	(4)	.8412	.8409	.8408	.8413	.8392	----	----
	(5)	.8413	.8411	.8404	.8374	.8233	----	----

$\Delta = 2$, $p_1 = .8413$, $m = .2$, $R = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & \ldots & \rho & 1 \end{pmatrix}$

$p_1 = .8413$ is the highest attainable probability of correct classification. For $\mu = (d,\ldots,d)^T$, $Ep = p_1$ for all entries.

Though the asymtotic performance among the missing value methods have slight variations, the differences of Ep are not substantial. Thus, it is concluded that the treatment of missing values in discriminant analysis must be studied for small and medium size samples.

With this in mind, a pilot study was conducted, comparing methods (1) and (4) in the two population discrimination problem.

This missing data problem is such that there are numerous variables which affect the outcome. This study focuses on two of these: 1) The percent of complete versus incomplete vectors and 2) The discriminatory power in the mean vector relative to the particular component or components which are missing. To gain insight into the effect of these factors several simplifying conventions were adopted. First of all, only two class patterns were specified at any time. The first class of vectors consisted of those which were complete and the second class consisted of vectors with one particular component missing. The second convention was that two simple but strong covariance structures were used throughout in order that the strongest statement with respect to the use of method (1) and (4) could be obtained. The first covariance structure exhibited near perfect correlation between the component which was to be missing in the second class pattern and one of the other components; all other covariances were zero. The first covariance structure for $p = 4$ might be represented as

$$\Sigma = \begin{pmatrix} 1 & & & \\ 1^{-} & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The second covariance structure had the correlation between the component to be missing in the second class pattern and all other components to be $1/\sqrt{p-1}^{-}$ (i.e. the multiple correlation coefficient is arbitrarily close to one); all other covariances are zero. This structure might be represented by

$$\Sigma = \begin{pmatrix} 1 & & & \\ 1/\sqrt{3}^{-} & 1 & & \\ 1/\sqrt{3}^{-} & 0 & 1 & \\ 1/\sqrt{3}^{-} & 0 & 0 & 1 \end{pmatrix}$$

Thus, viewing the missing data problem in terms of filling in the missing components with their regression estimates, given the components of the vector which are not missing, the above convention adopted in this study will provide for the maximum increase in information brought about by the inclusion of the partial data vectors. The mean vector for group 1 either had all components equal and nonzero or otherwise the mean associated with the particular component in the second class pattern which was to be missing was nonzero, and all other means were zero. The mean vector associated with group 2 was always the zero vector.

The actual simulation methodology will now be described. It

was arbitrarily determined that $p = 4$. Mean vectors and a covariance matrix were specified and training samples of sizes $n_1 = n_2 = 100$ were generated from each of two groups such that a specified percentage of the vectors were complete and the rest had a particular conponent missing. The well known Bayes linear classification rule was determined using only the complete vectors in the training samples and the method (4) was employed using all the data. Next, samples of size $N_1 = N_2 = 100$ were generated from the two groups and the vectors classified by both rules into one of the two groups and the number of misclassifications were counted. The entire procedure was then repeated until 25 simulations had been performed. It was determined that 25 simulations per particular set of mean vectors, common covariance matrix, and training sample design was sufficient to ascertain the relative performances of the two rules. The actual simulation results are summarized below.

For a fixed set of parameters μ_1, μ_2, Σ, $n_1=n_2=100$ and n_{ij} $(i=1,2; j=1,2)$, $N_1=N_2=100$ 25 simulations were performed

		Method (1)				Method (4)			
	$n_{i2}/n_i \times 100\%$	$\hat{P}(1/2)$	$\hat{\sigma}^2$	$\hat{P}(2/1)$	$\hat{\sigma}^2$	$\hat{P}(1/2)$	$\hat{\sigma}^2$	$\hat{P}(2/1)$	$\hat{\sigma}^2$
A) $\mu_1=[.5,.5,.5,.5]$ $\mu_2=0$	25%	.354	.00296	.360	.00361	.353	.00186	.360	.00352
$\Sigma=\begin{bmatrix} 1.0 & & & \\ 1.0^- & 1.0 & & \\ 0 & 0 & 1.0 & \\ 0 & 0 & 0 & 1.0 \end{bmatrix}$	50%	.333	.00291	.347	.00343	.328	.00241	.346	.00268
	75%	.331	.00329	.372	.00441	.333	.00250	.353	.00296
$D^2=.750 \Rightarrow P(1/2)=P(2/1)\cong.33^-$	90%	.369	.01160	.369	.00816	.346	.00550	.352	.00270
B) $\mu_1=[.1,0,0,0]$, $\mu_2=0$	25%	.133	.00104	.136	.00212	.134	.00110	.139	.00214
Σ same as above	50%	.137	.00159	.135	.00187	.132	.00146	.137	.00197
$D^2=5.00 \Rightarrow P(1/2)=P(2/1)\cong.13^+$	75%	.122	.00205	.139	.00232	.118	.00168	.136	.00187
	90%	.132	.00399	.159	.00384	.123	.00337	.156	.00402
C) $\mu_1=[.05,.05,.05,.05]$, $\mu_2=0$	25%	.315	.00302	.316	.00237	.316	.00298	.314	.00242
$\Sigma=\begin{bmatrix} 1.0 & & & \\ 1/\sqrt{3} & 1.0 & & \\ 1/\sqrt{3} & 0 & 1.0 & \\ 1/\sqrt{3} & 0 & 0 & 1.0 \end{bmatrix}$	50%	.299	.00228	.312	.00220	.296	.00247	.306	.00270
	75%	.295	.00431	.335	.00351	.284	.00508	.326	.00423
$D^2=1.11 \Rightarrow P(1/2)=P(2/1)\cong.30^+$	90%	.330	.00942	.349	.00934	.289	.00909	.301	.00731
D) $\mu_1=[.05,0,0,0]$ $\mu_2=0$	25%	.242	.00348	.232	.00210	.248	.00274	.222	.00212
Σ same as above	50%	.253	.00222	.237	.00252	.245	.00194	.242	.00248
$D^2=2.06 \Rightarrow P(1/2)=P(2/1)\cong.23^+$	75%	.276	.00344	.244	.00335	.271	.00394	.238	.00329
	90%	.299	.00842	.250	.00718	.265	.00827	.228	.00460

Summary

It seems logical that the inclusion of additional information into a classification rule might reduce the probability of misclassification; however, it must be remembered that given a set of parameters (i.e. mean vectors and covariance matrices) and the Bayes linear discriminant rule, the probability of misclassification is fixed and that the classical Bayes rule, regardless of sample size, can be used to provide unbiased estimates of such. Hence, one should expect the increase in performance resulting from the imput of additional information (i.e. data, whether it be complete or incomplete) to manifest itself in the variances of the associated estimates of the probability of misclassification. However, the estimated variances of the estimates of the probabilities of misclassification as computed in this simulation study do not in general tend to support this claim with respect to the inclusion of incomplete data. Even with the exceptional covariance structures specified in this study, the only evidence, and it is extremely weak, found in support of smaller variances associated with method (4) is found when the percentage of incomplete vectors is very high (i.e. 90%) and when the population mean vectors differ in every component. Further interpretation of the above results is left to the reader. It is felt that the common and known covariance matrix is at least partially responsible for the results obtained in this study since the estimation involved in the development of the two rules is only with respect to estimating the mean vectors associated with the two groups.

References

1. Boullion, T.L., Duran, B.S., Odell, P.L., "Estimation and Classification with Incomplete Data"

2. Chan, Linda S., and Dunn, Olive Jean, "A Note on the Asymptotic Aspect of the Treatment of Missing Values in Discriminant Analysis."

ESTIMATION AND CLASSIFICATION WITH INCOMPLETE DATA

by

T. L. Boullion[1], B. S. Duran[1], and P. L. Odell[2]

[1] Mathematics Department, Texas Tech University

[2] The University of Texas at Dallas

ESTIMATION AND CLASSIFICATION WITH INCOMPLETE DATA

Thomas L. Boullion, Benjamin S. Duran and Patrick L. Odell

1. Introduction and Summary

In this paper we consider the problem of missing data from multispectral scanner measurements acquired over an agricultural area. Frequently, due to partial cloud cover and other factors, not all elements of the multivariate observation vectors are meaningful on every occasion. Since these occur randomly, this information is taken into account to estimate parameters using the incomplete as well as the complete data vectors.

Assuming a multivariate normal for the distribution of the multispectral scanner measurements we develop expressions for the estimators of the mean vector and covariance matrix based on all the data. Also, expressions are given showing the gain in precision which may be obtained by using incomplete as well as the complete data.

Since this phenomenon of missing data is present in the training samples as well as for vector observations to be classified later, we develop a maximum likelihood scheme for classifying an observation vector in one of two multivariate populations with unknown means, but known covariance matrices.

2. Description of the Model

Assume a training sample of size N is taken from a p-variate normal distribution, but some of the p-vectors of observations have randomly occurring missing entries. In remote sensing applications, each observation vector X is of the form $X^T = [X_1^T, X_2^T, X_3^T, X_4^T]$, where $X_i(4\times1)$ represents a multispectral scanner measurement taken at time i. (For instance, each X_i

could be an observation on an agricultural unit at various growth stages for a crop of economic interest, such as wheat.) In this case, missing data occurs whenever a complete subvector X_i is missing, or can be identified to be cloud cover over the unit. Thus, there are 2^4-1 (in general, there are 2^p-1) possible sets of partial data vectors. Let R_1 be the set consisting of all complete data vectors. Let R_i denote the i-th set of partial data vectors arranged in an arbitrary order starting with R_1. To illustrate, let R_2 be the set of all vectors with X_4 missing, R_3 be the set of all vectors with X_3 and X_4 missing, etc.

Let $X_{i\alpha}(\alpha=1,\ldots,m_i)$ denote the $(p_i\times 1)$ observation vectors corresponding to $R_i(i=1,2,\ldots,2^p-1)$, and $\bar{X}_i = \frac{1}{m_i}\sum_{\alpha=1}^{m_i} X_{i\alpha}$, where $\bar{X}_i$ does not exist if $m_i=0$.

This leads us to assume the following statistical model:

$X_{1\alpha}$ is distributed as a p-variate normal with means μ and covariance Σ; and $X_{i\alpha}$ $(i>1)$ is distributed as a p_i-variate normal with mean μ_i and covariance Σ_i, where $\mu_i = D_i\mu$ and $\Sigma_i = D_i\Sigma D_i^T$, where D_i is a matrix of ones and zeros indicating which observations are missing. For instance, if R_2 is the set of all vectors with X_4 missing, then

$$\mu_2 = \begin{pmatrix}\mu_{21}\\ \mu_{22}\\ \mu_{23}\end{pmatrix} = \begin{pmatrix}I & 0 & 0 & 0\\ 0 & I & 0 & 0\\ 0 & 0 & I & 0\end{pmatrix}\begin{pmatrix}\mu_{21}\\ \mu_{22}\\ \mu_{23}\\ \mu_{24}\end{pmatrix} = D_2\mu\ ,$$

$$\text{and } \Sigma_2 = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ \Sigma_{41} & \Sigma_{42} & \Sigma_{43} & \Sigma_{44} \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix}$$

3. The Likelihood Equations When Σ is Known

Having the training sample properly categorized in the sets R_i $(i=1,2,\ldots,k)$ (note $k = 2^p-1$ if none of the R_i are empty), $R_1 = \{X_{11}, X_{12}, \ldots, X_{1m_1}\}$, $R_2 = \{X_{21}, X_{22}, \ldots, X_{2m_2}\}$, etc., the likelihood functions of the sample can be written as:

$$L = L_1, L_2, \ldots, L_k, \quad \text{where}$$

$$L_i = \frac{1}{(2\pi)^{\frac{p_i m_i}{2}} |\Sigma_i|^{\frac{m_i}{2}}} \; e^{-\frac{1}{2} \sum_{\alpha=1}^{m_i} (X_{i\alpha}-\mu_i)^T \Sigma_i^{-1} (X_{i\alpha}-\mu_i)}$$

The logarithm of the likelihood function is thus, $\log L = \sum_{i=1}^{k} \log L_i$, where

$$\log L_i = -\frac{1}{2} p_i m_i \log(2\pi) - \frac{m_i}{2} \log|\Sigma_i| - \frac{1}{2} \operatorname{tr}(\Sigma_i^{-1} M_i)$$

where the matrix M_i is given by $M_i = \sum_{\alpha=1}^{m_i} (X_i - \mu_i)(X_{i\alpha}-\mu_i)^T = m_i(\hat{\Sigma}_i + H_i)$. The matrices $\hat{\Sigma}_i$ and H_i are given by $\hat{\Sigma}_i = \frac{1}{m_i} \sum_{\alpha=1}^{m_i} (X_{i\alpha}-\hat{\mu}_i)(X_{i\alpha}-\hat{\mu}_i)^T$ and

$H_i = (\mu_i-\hat{\mu}_i)(\mu_i-\hat{\mu}_i)^T$, where $\hat{\mu}_i = \frac{1}{m_i} \sum_{\alpha=1}^{m_i} X_i$. Thus, $\log L_i$ can be written as

$$\log L_i = -\frac{1}{2} p_i m_i \log(2\pi) - \frac{m_i}{2} \log|\Sigma_i| - \frac{m_i}{2} \operatorname{tr}(\Sigma_i^{-1}[\hat{\Sigma}_i + H_i]) \quad \text{and}$$

$$\frac{\partial \log L_i}{\partial \mu_i} = -\frac{m_i}{2} \operatorname{tr}\left(\Sigma_i^{-1} \frac{\partial H_i}{\partial \mu_i}\right) = -\frac{m_i}{2}\left(\Sigma_i^{-1}\, 2(\mu_i - \hat{\mu}_i)\right) = -m_i \Sigma_i^{-1}(\mu_i - \hat{\mu}_i).$$

Since $\frac{\partial \log L_i}{\partial \mu} = D_i^T \left(\frac{\partial \log L_i}{\partial \mu_i}\right)$, we have $\frac{\partial \log L_i}{\partial \mu} = -\sum_{i=1}^{k} m_i D_i^T \Sigma_i^{-1}(D_i \mu - \hat{\mu}_i)$.

Differentiating $\log L_i$ a second time with respect to μ_i yields

$\frac{\partial^2 \log L_i}{\partial \mu_i^T \partial \mu_i} = -m_i \Sigma_i^{-1}$. But, the negative expected value of this is the portion of the information matrix corresponding to μ_i, say W_{μ_i}, which is equal to $m_i \Sigma_i^{-1}$. Similarly, $\frac{\partial^2 \log L}{\partial \mu^T \partial \mu} = -\sum_{i=1}^{k} m_i D_i^T \Sigma_i^{-1} D_i$ and thus the total information matrix for μ is $W_\mu = \sum_{i=1}^{k} D_i^T \Sigma_i^{-1} D_i$. This enables us to write

$$\frac{\partial \log L}{\partial \mu} = -W_\mu \mu + \sum_{i=1}^{k} D_i^T W_{\mu_i} \hat{\mu}_i , \qquad (1)$$

which upon setting equal to zero and solving for μ yields the maximum likelihood estimator $\hat{\mu}$ for μ (for Σ known) as

$$\hat{\mu} = W_\mu^{-1} \sum_{i=1}^{k} D_i^T W_{\mu_i} \hat{\mu}_i .$$

The matrix W_μ^{-1} is of interest since it is the asymptotic covariance matrix of $\hat{\mu}$. Also, if $k=2$ the estimator $\hat{\mu}$ becomes

$$\tilde{\mu} = W_\mu^{-1} [W_{\mu_1} \hat{\mu}_1 + D_2^T W_{\mu_2} \hat{\mu}_2] \qquad \text{or}$$

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} \hat{\mu}_{11} - \frac{n_2}{N}(\hat{\mu}_{11} - \hat{\mu}_{21}) \\ \vdots \\ \hat{\mu}_{12} - \frac{n_2}{N}\frac{\sigma_{12}}{\sigma_{11}}(\hat{\mu}_{11} - \hat{\mu}_{21}) \end{pmatrix}$$

where $\hat{\mu}_{ij}$ is the sample mean of the j-th component of μ computed from vectors in R_i. The large sample covariance matrix is given by W_{μ}^{-1} where $W_{\mu} = W_{\mu_1} + D_2^T W_{\mu_2} D_2$. It can be easily verified that W_{μ}^{-1} can be written in the form:

$$W_{\mu}^{-1} = \left[I - W_{\mu_1}^{-1} D_2^T (D_2 W_{\mu_1}^{-1} D_2^T + W_{\mu_2}^{-1})^{-1} D_2 \right] W_{\mu_1}^{-1}$$

The second term represents the gain in precision which may be expected if both groups of data are used as opposed to only the complete observation vectors to estimate μ. Investigating the variances of each component of $\hat{\mu}$ we obtain the following:

$\text{var}(\tilde{\mu}_i) = \sigma_{ii}/N$ if the i-th component of the data vectors is not missing in R_2, and $\text{var}(\tilde{\mu}_i) = (1 - \frac{m_2}{N} R_i^2)\, \sigma_{ii}/m_1$ otherwise. R_i^2 is the multiple correlation coefficient obtained when regressing the i-th variable on the variables corresponding to R_2.

4. The Likelihood Equations When Σ is Unknown

It is convenient to display the elements of Σ as a column vector whose elements are ordered as the columns of Σ. To illustrate, let

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{pmatrix}, \quad \text{then } \sigma^T = (\sigma_{11}, \sigma_{12}, \sigma_{22}, \sigma_{13}, \sigma_{23}, \sigma_{33}).$$ Thus, σ is a

vector of length $\frac{1}{2}p(p+1)$ defined by $\sigma_{\cdot} = (\sigma_{ij};\ 1\leq i\leq j = 1,2,\ldots,p)$. The vector σ_i of length $\frac{1}{2}p_i(p_i+1)$ will represent the corresponding column array of Σ_i, i=1,2,...,k. To relate σ_i to σ we introduce the matrix C_i so that $\sigma_i = C\sigma$. For Σ given above, and $\Sigma_i = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ & \sigma_{33} \end{pmatrix}$ we have

$$C_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Letting σ_{irs} denote the element in the (r,s) position of Σ_i and Σ_{irs} denote the matrix $\frac{\partial \Sigma_i}{\partial \sigma_{irs}}$ we have:

$$\frac{\partial \log|\Sigma_i|}{\partial \sigma_{irs}} = \mathrm{tr}(\Sigma_i^{-1}\Sigma_{irs})$$

and

$$\frac{\partial\, \Sigma_i^{-1}}{\partial \sigma_{irs}} = -\,\Sigma_i^{-1}\ \Sigma_{irs}\ \Sigma_i^{-1}$$

Using these results we obtain

$$\frac{\partial \log L_i}{\partial \sigma_{irs}} = -\frac{1}{2}m_i\ \mathrm{tr}(\Sigma_i^{-1}\Sigma_{irs}) + \frac{1}{2}\mathrm{tr}(\Sigma_i^{-1}\ \Sigma_{irs}\ \Sigma_i^{-1}M_i)\ , \qquad (2)$$

and

$$\frac{\partial^2 \log L_i}{\partial \sigma_{itu}\partial \sigma_{irs}} = \frac{1}{2}m_i\ \mathrm{tr}(\Sigma_i^{-1}\Sigma_{itu}\Sigma_i^{-1}\Sigma_{irs}) - \frac{1}{2}\mathrm{tr}(\Sigma_i^{-1}\Sigma_{itu}\Sigma_i^{-1}\Sigma_{irs}\Sigma_i^{-1}M_i)$$

$$-\frac{1}{2}\mathrm{tr}(\Sigma_i^{-1}\Sigma_{irs}\ \Sigma_i^{-1}\Sigma_{itu}\Sigma_i^{-1}M_i)\ .$$

Recalling that $E(M_i) = m_i\Sigma_i$ we find that the negative expected value of the above expression is given by

$$\frac{1}{2} m_i \operatorname{tr}(\Sigma_i^{-1}\Sigma_{irs}\Sigma_i^{-1}\Sigma_{itu}) \qquad (3).$$

Hence, the portion of the information matrix of L_i corresponding to σ_i is the square matrix W_{σ_i} of dimension $\frac{1}{2} p_i(p_i+1)$ whose elements are given by (3).

In order to express (2) in a more tractable form, we note that

$$\Sigma_i = \sum_{r,s,=1}^{p_i} \sigma_{irs}\Sigma_{irs} \text{ so that}$$

$$\operatorname{tr}(\Sigma_i^{-1}\Sigma_{irs}) = \operatorname{tr}(\Sigma_i^{-1}\Sigma_{irs}\Sigma_i^{-1} \sum_{t,u=1}^{p_i} \Sigma_{itu}\Sigma_{itu}) = \sum_{t,u=1}^{p_i} \operatorname{tr}(\Sigma_i^{-1}\Sigma_{irs}\Sigma_i^{-1}\Sigma_{itu})\sigma_{itu}.$$

It can be confirmed that this is just the rs-th component of $-W_{\sigma_i}\sigma_i$. Similarly, the second term on the right side of (2) is just the rs-th component of the vector $W_{\sigma_i}(\hat{\sigma}_i+h_i)$, where $\hat{\sigma}_i$ and h_i are the vector forms of Σ_i and H_i (ordered by columns).

Hence, we have

$$\frac{\partial \log L_i}{\partial \sigma_i} = - W_{\sigma_i}[\sigma_i-(\hat{\sigma}_i+h_i)] \quad \text{and since}$$

$$\frac{\partial \log L_i}{\partial \sigma} = C_i^T \frac{\log L_i}{\partial \sigma_i} \qquad \text{we obtain}$$

$$\frac{\partial \log L}{\partial \sigma} = - \sum_{i=1}^{k} C_i^T W_{\sigma_i}[X_i\sigma-(\hat{\sigma}_i+h_i)] = -W_\sigma\sigma + \sum_{i=1}^{k} C_i^T W_{\sigma_i}(\hat{\sigma}_i+h_i) \qquad (4)$$

where W is the total information matrix for σ given by $W_\sigma = \sum_{i=1}^{k} C_i^T W_{\sigma_i} C_i$

It has been tacitly assumed in the above development that all elements of μ are estimable. This is not the case for all elements of σ; hence, in (4) the vector σ should be interpreted as the vector of estimable parameters of Σ.

In order to obtain estimates for μ and σ from (1) and (4), we could use the method of steepest ascent to determine a stationary point for the likelihood function, since we have the expressions for the gradient vectors $\frac{\partial \log L}{\partial \mu}$ and $\frac{\partial \log L}{\partial \sigma}$. An alternative is to equate (1) and (4) to zero, obtaining the likelihood equations

$$W_{\mu} \cdot \mu = \sum_{i=1}^{k} D_i^T W_{\mu_i} \hat{\mu}_i \tag{5}$$

and

$$W_{\sigma} \cdot \sigma = \sum_{i=1}^{k} C_i^T W_{\sigma_i} (\hat{\sigma}_i + h_i) \ . \tag{6}$$

Estimates for W_{μ_i}, W_{σ_i} and h_i can be obtained from initial estimates of the parameters. Then solving the likelihood equations for μ and σ by multiplying by W_{μ}^{-1} and W_{σ}^{-1}, respectively, yields new estimates for μ and σ. Using these estimates to estimate W_{μ_i}, W_{σ_i} and h_i, the process can be repeated. This process should converge rapidly, and at termination, we have estimates of W_{μ}^{-1} and W_{σ}^{-1}, the large sample covariance matrices of $\tilde{\mu}$ and $\tilde{\sigma}$, respectively.

Consider the case k=2, that is, we have a set of m_1 complete observation vectors and a set of m_2 incomplete vectors. The large sample covariance matrix is W_{σ}^{-1} where $W_{\sigma} = W_{\sigma_1} + C_2^T W_{\sigma_2} C_2$. Hence W_{σ}^{-1} can be written as

$$W_{\sigma}^{-1} = [I - W_{\sigma_1}^{-1} C_2^T (C_2 W_{\sigma_1}^{-1} C_2^T + W_{\sigma_2}^{-1})^{-1} C_2] W_{\sigma_1}^{-1} \ .$$

The second term represents the gain in precision which may be expected if both groups of data are used as opposed to only the complete observation vectors to estimate Σ.

Considering the variances of each component of $\tilde{\sigma}$ we obtain the following:

$\text{var}(\tilde{\sigma}_{ii}) = 2\,\sigma_{ii}^2/N$ if the i-th component of the data vectors is not missing in R_2,

and $\text{var}(\sigma_{ii}) = (1-\frac{n_2}{N} R_i^4)\, 2\,\sigma_{ii}^2/n$, otherwise.

Although the likelihood equations (5), (6) in general must be solved numerically as indicated, there are certain special cases in which they may be solved analytically. This is true for any situation with only two groups of data where one group consists of the complete observation vectors, and for nested incomplete vector observations. By nested we mean that group R_1 consists of $m_1 > p$ complete vectors and it is possible to label the remaining groups so that Σ_{i+1} is a principal submatrix of Σ_i for $i=1,2,\ldots,k-1$.

Consider the likelihood equations (5), (6) for the nested case. They can be solved sequentially as follows: Consider equations (5) and (6) with $k=2$ and replace the elements of W_{μ_i} and W_{μ_2} in (5) by their estimates using $\hat{\sigma}_1$ only. The solution of (5) yields $\tilde{\mu}$ which is maximum likelihood for μ if $k=2$. Substituting $\tilde{\mu}$ for μ in h_i and using $\hat{\sigma}_1$ to estimate $W_{\hat{\sigma}_1}$ and $W_{\hat{\sigma}_2}$ in (6) enables us to solve for σ, say $\tilde{\sigma}$, which is maximum likelihood for σ if $k=2$. We next consider $k=3$ and repeat the above process with $\tilde{\sigma}$ as initial estimate rather than $\tilde{\sigma}_1$. The resulting estimates are maximum likelihood for $k=3$. This is continued until all k groups are exhausted and the final solutions are maximum likelihood.

For the remote sensing application, it is expected that $k=2$ will be the most frequently occurring situation since a portion of the subvectors representing data for a particular pass at a point in time over a certain region will be missing due to partial cloud cover.

5. Derivation of Maximum Likelihood Classifier

We now consider the problem of classifying an observation vector in one of two multivariate normal populations when dealing with incomplete data vectors. It will be assumed that sufficient training data are available to estimate the covariance matrices very accurately. Hence, we assume the covariance matrices are known and unequal. These will be denoted by Σ_1 and Σ_2 with corresponding submatrices Σ_{1j} and Σ_{2j}. Let X_0 denote an observation vector to be classified into either $N(\mu_1,\Sigma_1)$ or $N(\mu_2,\Sigma_2)$. Letting $X_{i\alpha}$ denote the samples from $N(\mu_1,\Sigma_1)$ and $Y_{i\alpha}$ from $N(\mu_2,\Sigma_2)$, the logarithm of the likelihood function L_i of all the observations can be expressed as

$$\log L_i = \log K_i - \frac{1}{2}\left[\sum_{\alpha=1}^{k}\left\{\sum_{\alpha=1}^{m_j}(X_{j\alpha}-\mu_{1j})^T\Sigma_{1j}^{-1}(X_{j\alpha}-\mu_{1j}) + \sum_{=1}^{n_j}(Y_{j\alpha}-\mu_{2j})^T\Sigma_{2j}^{-1}(Y_{j\alpha}-\mu_{2j})\right\}\right.$$

$$\left. + (X_0-\mu_{i0})^T\,\Sigma_{i0}^{-1}\,(X_0-\mu_{i0})\right]$$

where $K_i = \prod_{j=1}^{k'}(2\pi)^{-\frac{P_j(m_j+n_j)-P_0}{2}}|\Sigma_{ij}|^{-\frac{m_j}{2}}|\Sigma_{2j}|^{-\frac{n_j}{2}}|\Sigma_{i0}|^{-\frac{1}{2}}$, and Σ_{i0} is the covariance matrix corresponding to X_0 of size p_0. Likewise, μ_{i0} is the mean vector corresponding to X_0 for i=1,2. Note that

$$\log L_i = \log K_i + \sum_{j=1}^{k}(\log L_{1j} + \log L_{2j}) + (X_0-\mu_{i0})^T\Sigma_{i0}^{-1}\,(X_0-\mu_{i0})$$

Hence,

$$\frac{\partial \log L_1}{\partial \mu_1} = \sum_{j=1}^{k} D_j^T\,\frac{\partial \log L_{1j}}{\partial \mu_{1j}} + \Sigma_{10}^{-1}\,(X_0 - \mu_{10})$$

$$\frac{\partial \log L_{1j}}{\partial \mu_{1j}} = \sum_{\alpha=1}^{j} \Sigma_{1j}^{-1}(X_{j\alpha}-\mu_{1j}) = m_j\Sigma_{1j}^{-1}\bar{X}_j - m_j\Sigma_{1j}^{-1}D_j^T\mu_1 \text{ , where } \Sigma_{10}^{-1}(X_0-\mu_{10})$$

is to be multiplied on the left by the appropriate D_j^T if $p_0<p$. This will be denoted by writing $D_{10}^T\ \Sigma_{10}^{-1}(X_0-\mu_{10})$. Thus, we have

$$\frac{\partial \log L_1}{\partial \mu_1} = \sum_{j=1}^{k} m_j D_j^T(\Sigma_{1j}^{-1}\bar{X}_j) + D_{10}^T\Sigma_{10}^{-1}X_0 - \left[\sum_{j=1}^{k} m_j(D_j^T\Sigma_{1j}^{-1}D_j) + D_{10}^T\Sigma_{10}^{-1}D_{10}\right]\mu_1.$$

Similarly,

$$\frac{\partial \log L_1}{\partial \mu_2} = \sum_{j=1}^{k} n_j D_j^T\Sigma_{2j}^{-1}\bar{Y}_j - \left[\sum_{j=1}^{k} n_j D_j^T\Sigma_{2j}^{-1}D_j\right]\mu_2$$

For ease of presentation we introduce the following definitions:

$$Q_{11} = \sum_{j=1}^{k} m_j(D_j^T\Sigma_{1j}^{-1}D_j) + D_{10}^T\Sigma_{10}^{-1}D_{10}, \qquad q_{11} = \sum_{j=1}^{k} m_j D_j^T(\Sigma_{1j}^{-1}\bar{X}_j) + D_{10}^T\Sigma_{10}^{-1}X_0$$

$$Q_{12} = \sum_{j=1}^{k} m_j D_j^T\Sigma_{1j}^{-1}D_j \qquad q_{12} = \sum_{j=1}^{k} m_j D_j^T\Sigma_{1j}^{-1}\bar{X}_j$$

$$Q_{22} = \sum_{j=1}^{k} n_j(D_j^T\Sigma_{2j}^{-1}D_j) + D_{20}^T\Sigma_{20}^{-1}D_{20}, \qquad q_{22} = \sum_{j=1}^{k} m_j D_j^T\Sigma_{2j}^{-1}\bar{Y}_j + D_{20}^T\Sigma_{20}^{-1}X_0$$

$$Q_{21} = \sum_{j=1}^{k} n_j D_j^T\Sigma_{2j}^{-1}D_j \qquad q_{21} = \sum_{j=1}^{k} n_j D_j^T\Sigma_{2j}^{-1}\bar{Y}_j$$

Setting $\frac{\partial \log L_1}{\partial \mu_1} = 0$, $\frac{\partial \log L_1}{\partial \mu_2} = 0$ yields $Q_{11}\mu_1 = q_{11}$ and $Q_{21}\mu_2 = q_{21}$.

Similarly $\frac{\partial \log L_2}{\partial \mu_1} = 0$, $\frac{\partial \log L_2}{\partial \mu_2} = 0$ yields $Q_{12}\mu_1 = q_{12}$ and $Q_{22}\mu_2 = q_{22}$.

Thus, the maximum likelihood estimates for μ_1,μ_2 under the hypothesis that X_0 is from $N(D_{10}^T\mu_1,\ D_{10}^T\Sigma_1 D_{10})$ is given by:

$$ {}_1\hat{\mu}_1 = Q_{11}^{-1}q_{11}, \quad {}_1\hat{\mu}_2 = Q_{21}^{-1}q_{21} . $$

Likewise, if X_0 is from $N(D_{20}^T\ {}_2, D_{20}^T\Sigma_2 D_{20})$, the maximum likelihood estimates for μ_1,μ_2 are:

$$ {}_2\hat{\mu}_1 = Q_{12}^{-1}q_{12} \text{ and } {}_2\hat{\mu}_2 = Q_{22}^{-1}q_{22} . $$

Letting L_i denote the maximum of $L_i (i=1,2)$ under variation of μ_1,μ_2 we get

$$ \hat{L}_1 = \prod_{j=1}^{k} (2\pi)^{-\frac{P_j(m_j+n_j)}{2}} |\Sigma_{1j}|^{-\frac{m_j}{2}} |\Sigma_{2j}|^{-\frac{n_j}{2}} |\Sigma_{10}|^{-\frac{1}{2}} (2\pi)^{-\frac{P_0}{2}} $$

$$ \exp\left\{ -\frac{1}{2} \sum_{j=1}^{k} \left[\sum_{\alpha=1}^{m_j} (X_{j\alpha} - {}_1\hat{\mu}_{1j})^T \Sigma_{1j}^{-1} (X_{j\alpha} - {}_1\hat{\mu}_{1j}) + \sum_{\alpha=1}^{n_j} (Y_{j\alpha} - {}_1\hat{\mu}_{2j})^T \Sigma_{2j}^{-1} \right.\right. $$

$$ \left.\left. (Y_{j\alpha} - {}_1\hat{\mu}_{2j}) \right] + (X_0 - \hat{\mu}_{10})^T \Sigma_{10}^{-1} (X_0 - \hat{\mu}_{10}) \right\} , $$

$$ \hat{L}_2 = \prod_{j=1}^{k} (2\pi)^{-\frac{P_j(m_j+n_j)}{2}} |\Sigma_{1j}|^{-\frac{m_j}{2}} |\Sigma_{2j}|^{-\frac{n_j}{2}} |\Sigma_{20}|^{-\frac{1}{2}} (2\pi)^{-P_0/2} $$

$$ \exp\left\{ -\frac{1}{2} \sum_{j=1}^{k} \left[\sum_{\alpha=1}^{m_j} (X_{j\alpha} - {}_2\hat{\mu}_{1j})^T \Sigma_{1j}^{-1} (X_{j\alpha} - {}_2\hat{\mu}_{1j}) + \sum_{\alpha=1}^{n_j} (Y_{j\alpha} - {}_2\hat{\mu}_{2j})^T \Sigma_{2j}^{-1} \right.\right. $$

$$ \left.\left. (Y_{j\alpha} - {}_2\hat{\mu}_{2j}) \right] + (X_0 - \hat{\mu}_{20})^T \Sigma_{20}^{-1} (X_0 - \hat{\mu}_{20}) \right\} $$

Forming $-2 \log \frac{\hat{L}_1}{\hat{L}_2}$ and simplifying, we get

$$-2 \log \frac{\hat{L}_1}{\hat{L}_2} = \log \frac{|\Sigma_{10}|}{|\Sigma_{20}|} + (X_0 - {}_1\hat{\mu}_{10})^T \Sigma_{10}^{-1} (X_0 - {}_1\hat{\mu}_{10}) - (X_0 - {}_2\hat{\mu}_{20})^T \Sigma_{20}^{-1} (X_0 - {}_2\mu_{20})$$

$$+ q_{12}^T Q_{12}^{-1} q_{12} - q_{21}^T Q_{21}^{-1} q_{21} - q_{11}^T Q_{11}^{-1} (D_0^T \Sigma_{10}^{-1}) Q_{11}^{-1} q_{11} + q_{22}^T Q_{22}^{-1} (D_{20}^T \Sigma_{20}^{-1} D_{20}) Q_{22}^{-1} q_{22}$$

$$-q_{11}^T Q_{11}^{-1} q_{11} + 2 q_{11}^T Q_{11}^{-1} D_{10}^T \Sigma_{10}^{-1} X_0 + q_{22}^T Q_{22}^{-1} q_{22} - 2 q_{22}^T Q_{22}^{-1} D_{20}^T \Sigma_{20}^{-1} X_0 \quad \text{if } X_0 \text{ is of}$$

size $p_0 < p$.

If $L(Z) = -2 \log \frac{\hat{L}_1}{\hat{L}_2} = L(\bar{X}_1^T, \bar{X}_2^T, \ldots, \bar{X}_K^T, X_0, \bar{Y}_1^T, \ldots, \bar{Y}_K^T)$, then classify X_0 in $N(\mu_1, \Sigma_1)$ if $L(Z) < 0$ and in $N(\mu_2, \Sigma_2)$ if $L(X) > 0$.

The above expression simplifies whenever X_0 is $p \times 1$. In that case,

$$-2 \log \frac{\hat{L}_1}{\hat{L}_2} = \log \frac{|\Sigma_1|}{|\Sigma_2|} + X_0^T (\Sigma_1^{-1} - \Sigma_2^{-1}) X_0 + q_{12}^T Q_{12}^{-1} q_{12} + q_{22}^T Q_{22}^{-1} q_{22} - q_{21}^T Q_{21}^{-1} q_{21}$$

$$- q_{11}^T Q_{11}^{-1} q_{11}$$

References

[1] Hartley, H. O. and Hocking, R. R., "The Analysis of Incomplete Data," Biometrics, Vol. 27, No. 4, (1971), pp. 783-823.

[2] Hocking, R. R. and Smith, Wm. B., "Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations," JASA, (1968), pp. 159-173.

[3] _____, "Optimum Incomplete Multinormal Samples," Technometrics, Vol. 14, #2, (1972), pp. 299-307.

[4] Srivastava, J. N. and Zaatar, M.K., "On the Maximum Likelihood Classification Rule for Incomplete Multivariate Samples and Its Admissibility," J. of Multivariate Analysis, Vol. 2 (1972) pp. 115-126.

[5] Trawinski, I. M. and Bargmann, R. E., "Maximum Likelihood Estimation with Incomplete Multivariate Date," Annals of Math Statist., (1964) PP. 647-657.

PROBABILITY OF MISCLASSIFICATION WITH MISSING DATA

by

D. D. McElroy[1] and H. L. Gray[2]

[1] Graduate Student, Statistics Department, Southern Methodist University

[2] Statistics Department, Southern Methodist University

PROBABILITY OF MISCLASSIFICATION WITH MISSING DATA

1. Introduction

In the problem of discriminant analysis the probability of misclassification is of significant importance. The difficulty in obtaining explicit expressions for the measure of probability of misclassification has prompted various developments of statistical methodology of classification. The problem of misclassification is further complicated by the problem of missing data. This report investigates the relationship between the probability of missing data and the probability of misclassification. Since divergence is a widely used measure of distance, it will be investigated as a criteria upon which to base a discriminant function.

2. Definition of the Problem

$$\text{Let } X = \underset{np\times 1}{\begin{bmatrix} X_1 \\ n\times 1 \\ X_2 \\ n\times 1 \\ \vdots \\ X_p \\ n\times 1 \end{bmatrix}} \sim MVN_{np}(\mu, \underset{np\times np}{\Sigma}) \qquad (1)$$

where X consists of p subvectors, each of length n. The data will be lost in such a way that an entire subvector will be lost. Let q_i be the probability that the i^{th} subvector is missing, i=1,2,...,p.

As a mechanism for identifying the outcome of missing data define I as a random vector of 0's and 1's, with 0 indicating a missing subvector. I will be of length p, i.e. for some j, I_j' = (0 1 1 ... 1) indicates the first subvector, X_1, is missing. That is, we use the notation I_j to denote a value of I. There will be 2^p possible outcomes of I which are mutually exclusive and independent. The probability of outcome I_j above can be calculated:

$$P[I_j' = (0\ 1\ 1\ 1\ \ldots\ 1)] = q_1(1-q_2)(1-q_3)\ \ldots\ (1-q_p) = \gamma_j. \tag{2}$$

Let $\pi_1, \pi_2, \ldots \pi_m$ be the m populations and $p_1, p_2, \ldots p_m$ be the corresponding a priori probabilities. Let $P[i|j]$ be the probability of classifying an observation from the j^{th} population into the i^{th} population and $C(i|j)$ = the associated cost. For the remainder of this paper the cost $C(i|j) = 1$ if $i \neq j$ and $C(i|i) = 0$.

If $f_i(x)$ represents the density function of the i^{th} population, <u>divergence</u>, Δ, is defined by

$$\Delta = \int_{-\infty}^{\infty} [f_i(x) - f_j(x)]\ \ln [f_i(x)/f_j(x)]\ dx\ . \tag{3}$$

Letting $f_i(x)$, i=1,2,...,m be the multivariate normal density function with mean vector μ_i and variance-covariance matrix Σ_i, i=1,2,...m, we get divergence

$$\Delta = \frac{1}{2}\ \{tr(\Sigma_i-\Sigma_j)(\Sigma_j^{-1}-\Sigma_i^{-1}) + (\Sigma_i^{-1}+\Sigma_j^{-1})(\mu_i-\mu_j)(\mu_i-\mu_j)'\} \tag{4}$$

3. Equal Variance-Covariance Matrix Case

Anderson [1] considers the case for equal variance-covariance matrices and multivariate normal distribution. Divergence then reduces to

$$\Delta = (\mu_i-\mu_j)'\ \Sigma^{-1}\ (\mu_i-\mu_j) \tag{5}$$

In the case of two populations and a priori probabilities p_1 and p_2, the probability of drawing a sample from population π_2 and misclassifying it into π_1 is $P(1|2)p_2$ and the probability of drawing a sample from population π_1 and misclassifying it into π_2 is $P(2|1)p_1$. Thus the total probability of misclassification is

$$P[\text{misclassification}] = p_2\, P(1|2) + p_1\, P(2|1) \tag{6}$$

Anderson shows the discriminant function which minimizes the P[misclassification] is the usual Bayes discriminant function:

$$U = X'\Sigma^{-1}[\mu^{(1)} - \mu^{(2)}] - \frac{1}{2}[\mu^{(1)} + \mu^{(2)}]' \,. \tag{7}$$

The best regions of classification are given by, classify into π_1 if

$$x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})'\, \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \geq \ln p_2/p_1 \,. \tag{8}$$

Otherwise classify into π_2.

Anderson finds the distribution function for (7) and using this distribution obtains the probability of misclassification as follows (still with $\Sigma_1 = \Sigma_2 = \Sigma$):

$$P(1|2) = \int_{\ln p_2/p_1 + \frac{\sqrt{\Delta}}{2}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}y^2}\, dy \tag{9}$$

and

$$P(2|1) = \int_{-\infty}^{\ln p_2/p_1 - \frac{\sqrt{\Delta}}{2}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}y^2}\, dy \,. \tag{10}$$

Both probabilities are monotonically decreasing functions of Δ. Therefore, the total probability of misclassification is given by:

$$P[\text{misclassification}] = p_2\, P(1|2) + p_1\, P(2|1) \tag{11}$$

which is a monotonically decreasing function of divergence.

4. The Bayes Discriminant Function with Missing Data

We wish to incorporate the probability of missing subvectors into the problem of the probability of misclassification.

For the remainder of this report we will for convenience restrict ourselves to the case of two populations, π_1 and π_2, with a priori probabilities p_1 and p_2. The extension to n populations is obvious. We will consider $\Sigma_1 = \Sigma_2 = \Sigma$ and take n=4 and p=4. Of course the development which follows is the same for any n and p but we have selected these values from the physical model for definiteness. Thus

$$X = \underset{16x1}{\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}} \sim \underset{16x1}{MVN}\ (\ \mu\ ,\ \underset{16x16}{\Sigma}\ ) \tag{12}$$

There are $2^4 = 16$ ways in which the subvectors can be lost and 16 mutually exclusive and independent outcomes of I with associated probabilities, γ_j, j=1,2,... 16. For example, if $I_j' = (1,0,1,0)$ we have

$$P[I_j' = (1,0,1,0)] = (1-q_1)\ q_2\ (1-q_3)\ q_4 = \gamma_j. \tag{13}$$

(Clearly $\sum_{j=1}^{16} \gamma_j = 1$).

Definition 1:

Define $D(U,I_j)$ as the usual Bayes discrimination function, but based only on the available data as reflected in outcome I_j.

In calculating the probability of misclassification using $D(U,I_j)$ the probability of misclassification is based on the parameter

$$\alpha_j = (\mu_j^{(1)} - \mu_j^{(2)})'\ \Sigma_j^{-1}\ (\mu_j^{(1)} - \mu_j^{(2)}) \tag{14}$$

where the subscript indicates that the parameter is caluclated only from the

subsets of μ and Σ associated with the available data in outcome I_j. Only in the outcome $I_1 = (1, 1, 1, 1)$, where no data is missing does $\alpha_1 = \Delta$.

The distribution of the Bayes discrimination function has been altered by including the probability of missing data. It is now the function of two random variables. Thus, the probabilities of misclassification and outcome I_j are given by:

$$P[1|2,I_j]P[I_j] = \left[\int_{\ln p_2/p_1 + \frac{\sqrt{\alpha_j}}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \right] \gamma_j \tag{15}$$

$$P[2|1,I_j]P[I_j] = \left[\int_{-\infty}^{\ln p_2/p_1 - \frac{\sqrt{\alpha_j}}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \right] \gamma_j . \tag{16}$$

The total probability of misclassification and outcome I_j is:

$$P[\text{misclassification and } I_j] = \left[p_2 \int_{\ln p_2/p_1 + \frac{\sqrt{\alpha_j}}{2}}^{\infty} e^{-\frac{1}{2}y^2} dy + p_1 \int_{-\infty}^{\ln p_2/p_1 - \frac{\sqrt{\alpha_j}}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \right] \gamma_j \tag{17}$$

We can then define an "expected probability of misclassification," which is actually just the probability of misclassification in the problem as we have now modeled it. We have simply used the term "expected

misclassification" to emphasize that we have accounted for the possibility of missing data in our calculations, thus

$$P[\text{misclassification}] = E_I\, P[\text{misclassification}]$$

$$= \sum_{j=1}^{16} \gamma_j \left[p_2 \int_{\ell n\, p_2/p_1 + \frac{\sqrt{\alpha_j}}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dy + p_1 \int_{-\infty}^{\ell n\, p_2/p_1 - \frac{\sqrt{\alpha_j}}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dy \right] \quad (18)$$

Figures A-1 through A-15 demonstrate the behavior of (18) with different γ_j, i.e. different $q_1 = q_2 = q_3 = q_4 = q$.

Definition 2:

In the case where all the data is missing but a region must be classified, we can classify the region by making use of the a priori probabilities. That is, we simply sample from a random device with those a priori probabilities and make the classification. For example, in the two population case we take a "coin" which has probability p_1 of heads and $p_2 = 1-p_1$ of tails. Then we flip the coin and classify the region accordingly. Thus we have once more

$$\begin{aligned} P[\text{misclassification}] &= P[1|2]\, p_2 + P[2|1]\, p_1 \\ &= p_1p_2 + p_2p_1 = 2p_1p_2 \\ &= 2p_1(1-p_1) \end{aligned} \quad (19)$$

Thus, if we denote outcome (0, 0, 0, 0) by I_{16}, we have

$$E_I\left[P[\text{misclassification}]\right] = \sum_{j=1}^{15} \gamma_j [p_2P(1|2) + p_1\, P(1|2)] + 2\gamma_{16}\, p_1(1-p_1)$$

$$= \sum_{j=1}^{15} \gamma_j [p_2P(1|2) + p_1P(2|1)] + 2\, q_1q_2q_3q_4\, p_1(1-p_1) \quad (20)$$

One should note that if $p_1=p_2$ (20) is equivalent to taking $\alpha_{16} = 0$ which is certainly reasonable.

By definition (and by a realistic approach to the physical problem) the data obtained will be considered "essential data" in the sense that any data lost will increase the probability of misclassification, i.e.

$$P[i|j, I_k] > P[i|j, I_\ell] \tag{21}$$

if I_k represents an outcome where more data is lost than in outcome I_ℓ, i.e. every element in I_k is contained in I_ℓ. Since the outcome of I determines the parameter in the calculation of the probability of misclassification, this imposes the condition in (21) above that $\alpha_k < \alpha_\ell$. This makes the definition more suitable mathematically, so we will make this our formal definition.

Definition 3. Data will be said to be essential data if every element of I_k is in I_ℓ implies $\alpha_k \leq \alpha_\ell$.

One such case would be where $\Sigma^{-1} = \text{diag}\left[\frac{1}{\sigma_{ii}}\right]$ $i = 1,2,..,16$.

Theorem 1

Let $q = (q_1,q_2,q_3,q_4)$ and consider two cases. Case 1: at least one of the q_i is greater than 0. Case 2: $q_i=0, \forall i$. This implies no missing data, i.e. the only outcome possible for I is $I_1 = (1,1,1,1)$. Then

E[P[misclassification; I_j] in case 1] $\geq$ E[P[misclassification; I_j] in case 2].

Proof:

Since Case 2 is trivially true we consider Case 1 where we let

$$C = \int_{\ell n\, p_2/p_1 + \frac{\sqrt{\Delta}}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \qquad (22)$$

Now in case 1, if there exists at least one q_i greater than zero, γ_1 becomes $0 < \gamma_1 < 1$ and there exists at least one other γ_j greater than 0. Hence

$$E_I[P(1|2)] = \gamma_1 \int_{\ell n\, p_2/p_1 + \frac{\sqrt{\Delta}}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy + \ldots$$

$$+ \gamma_j \int_{\ell n\, p_2/p_1 + \frac{\sqrt{\alpha_j}}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy + \ldots + \gamma_{16} \int_{\ell n\, p_2/p_1 + \frac{\sqrt{\alpha_{16}}}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy. \qquad (23)$$

But by definition 3, each $\alpha_{k\neq 1} < \Delta$ and thus each integral of equation (23) (other than the first integral) is greater than C and equation (23) can be written

$$E_I[P(1|2)] > \gamma_1 C + \gamma_2 C + \ldots + \gamma_j C + \ldots + \gamma_{16} C \qquad (24)$$

$$= \sum_{i=1}^{16} \gamma_i C$$

$$= C \qquad \text{since } \sum_{i=1}^{16} \gamma_i = 1 .$$

In a similar fashion, we can show

$E_I[P(2|1)$ for at least one $q_i > 0] > E[P(2|1, I_j)\ q_i = 0, \forall i]$,

and finally

P[misclassification; if at least one $q_i > 0$] > P[misclassification; $q_i = 0, \forall q_i$]

5. P[misclassification] as a function of divergence

Unfortunately, P[misclassification] = E_I[P[misclassification]] is not in general a decreasing function of divergence, as shown in the following example.

Example: Let the divergence between populations $\pi_1, \pi_2, \pi_3, \pi_4$ be the following (without loss of generality we can assume equal a priori probabilities).

Let $\Delta_1(\pi_1, \pi_2, \Sigma_1 = \Sigma_2) = 16$ $\Delta_2(\pi_2, \pi_3, \Sigma_2 = \Sigma_3) = 36$ in the following way:

$$\Delta_1 = (\mu^{(1)} - \mu^{(2)})' \Sigma_1^{-1} (\mu^{(1)} - \mu^{(2)}).$$

$$= \underset{1\times16}{(1,1,\ldots 1)} \begin{bmatrix} 3.99 & & & & & & & & & & & & & & & \\ & 3.99 & & & & & & & & & & & & & & \\ & & 3.99 & & & & & & & & & & & & & \\ & & & 3.99 & & & & & & & & & & & & \\ & & & & .003 & & & & & & & & & & & \\ & & & & & .003 & & & & & & & & & & \\ & & & & & & \cdot & & & & & & & & & \\ & & & & & & & \cdot & & & & & & & & \\ \phi & & & & & & & & \cdot & & & & & & & \\ & & & & & & & & & .003 & & & & & & \\ & & & & & & & & & & .007 & & & & & \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ \\ \\ 1 \end{bmatrix}$$

16×16 $\qquad$ 16×1

$$= 4(3.99) + 11(.003) + 1(.007) = 16$$

$$\Delta_2 = (\mu^{(3)}-\mu^{(4)}) \cdot \Sigma_2^{-1} (\mu^{(3)}-\mu^{(4)})$$

$$= \underset{1\times16}{[1,1,\ldots\ 1]} \begin{bmatrix} 2.75 & & & & & & & & & & & & & & & \\ & 2.75 & & & & & & & & & & & & & & \\ & & 2.75 & & & & & & & & & & & & & \\ & & & 2.75 & & & & & & & & & & & & \\ & & & & 2 & & & & & & & & & & & \\ & & & & & 2 & & & & & & & & & & \\ & & & & & & \cdot & & & & & & & & & \\ & & & & & & & \cdot & & & & & & & & \\ & & & & & & & & \cdot & & & & & & & \\ & & & & & & & & & 2 & & & & & & \\ & & & & & & & & & & 3 & & & & & \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ \\ \\ \\ \\ \\ 1 \end{bmatrix}$$

16×16 $\qquad$ 16×1

$$= 4(2.75) + 11(2) + 1(3) = 36 .$$

Now let $q_1 = .5$ and $q_2 = q_3 = q_4 = 0$.

Then $\gamma_1 = .5$ and $\gamma_2 = .5$ and all the remaining terms, $\gamma_i = 0$. The only two outcomes of I that are possible are $I_1' = (1\ 1\ 1\ 1)$ and $I_2' = (0\ 1\ 1\ 1)$. Thus

$$E[P(1|2,\Delta_1) = .5(.0228) + .5(.4602) = .2415$$

and

$$E[P(1|2,\Delta_2) = .5(.0013) + .5(.0065) = .0039 .$$

So, even though $\Delta_2 > \Delta_1$

$$E[P(1|2,\Delta_2)] < E[P(1|2,\Delta_1)].$$

6. Equal q_i, i=1,2,3,4

While it is probably true that for equal Σ and $q = q_1=q_2=q_3=q_4$, that P[misclassification] is an increasing function of q, it may be hard to show analytically.

Consider the two dimensional case and equal a priori as a function of $q: g(q) = P[\text{misclassification}]$

$$= (1-q)^2 C_1 + q(1-q) C_2 + q(1-q) C_3 + q^2 C_4 \qquad 0 \leq q \leq 1$$

$$\text{where by def. 3} \quad C_1 > C_2, C_3 \quad \text{and} \quad C_2, C_3 > C_4 \,. \tag{25}$$

Now $g(0) = C_1$ and $g(1) = C_4$ so we know $0 < g(0) < g(1)$. Note that this is true for any dimensional case.

However, in general to prove there are no critical values for n dimensions would involve proving there are no roots for an n-1 degree polynomial for $0 \leq q \leq 1$. This may be very difficult to show analytically.

The question also arises with equal q_i whether or not $P[\text{misclassification}] = E_I[P[\text{misclassification}]]$ is a monotonically decreasing function of divergence in the equal covariance case, even though it has been shown not to be true for unequal q. Again this is probably true but would be difficult to prove analytically.

7. Unequal Covariance Matrices Case

It would be desirable to extend the above theory to encompass the case of unequal covariance matrices. However, Chang[4] points out that divergence is neither uniquely nor monotonically related to Bayes' classification errors. Since the relationship of the new model which incorporates the probability of missing data and divergence with equal covariance matrices is still uncertain, further investigation into the properties and possibly other discriminant functions should be investigated.

8. Plots

Appendix A contains several plots of probabilities of misclassification for equal q_i's and $\Sigma = I$ under different a priori assumptions.

References

[1] Anderson, T. W. (1958), An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, New York

[2] _____ (1957), "Maximum Likelihood Estimates for a Multivariate Normal Distribution when Some Observations are Missing," JASA, Volume 52.

[3] Boullion, T. L. (1974), "Some Results of Randomly Missing Data in Discriminant Analysis," Technical Report for NASA.

[4] Chang, C. Y. (1971), "Divergence and Probability of Misclassification," Technical Report for NASA prepared by Lockheed Electronics Company Inc.

[5] Chhikara, R. S. (1975). "Effect of Photo-Misinterpretation of Training Fields Upon Classification Performance," Lockheed Technical Report No. 4353, prepared by Lockheed Electronics Company for NASA under contract NAS 9-12200

[6] _____, Odell, P.L. (1974), "On Designing Simulation Models for Evaluating Discriminant Analysis Routines," Computers and Mathematics with Applications, Volume I.

[7] Van Ness, J. and Simpson, C. (1975), "On the Effects of Dimension in Discriminant Analysis," Master's Thesis at The University of Texas at Dallas (accepted by Technometrics)

Appendix A. Plots of Probability of Misclassification

Figures A-1 through A-15 contain plots of the probability of misclassificaion. P(1 GIVEN 2) denotes $P(1|2)$ and is given by

$$P(1|2) = \sum_{1}^{16} \gamma_j [1 - \Phi(A_j)]$$

where (1) Φ = standard normal distribution function

(2) $A_j = \ell n\ p_2/p_1 + \sqrt{\alpha_j}/2$

P(2 GIVEN 1) denotes $P(2|1)$ and is given by

$$P(2|1) = \sum_{1}^{16} \gamma_j \Phi(B_j)$$

where $B_j = \ell n\ p_2/p_1 - \sqrt{\alpha_j}/2$

The total probability of misclassification is

$$P[\text{misclassification}] = p_2\ P(1|2) + p_1\ P(2|1) .$$

For each plot $\Sigma = I$. Let

$$d_i = i^{th} \text{ component of } \mu^{(1)} - \mu^{(2)} .$$

The values for the d_i's for each of the plots is given in Table A.1 .

i =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Figures A-1 thru A-9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Figures A-10 thru A-12	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
Figures A-13 thru A-15	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4

Table A.1

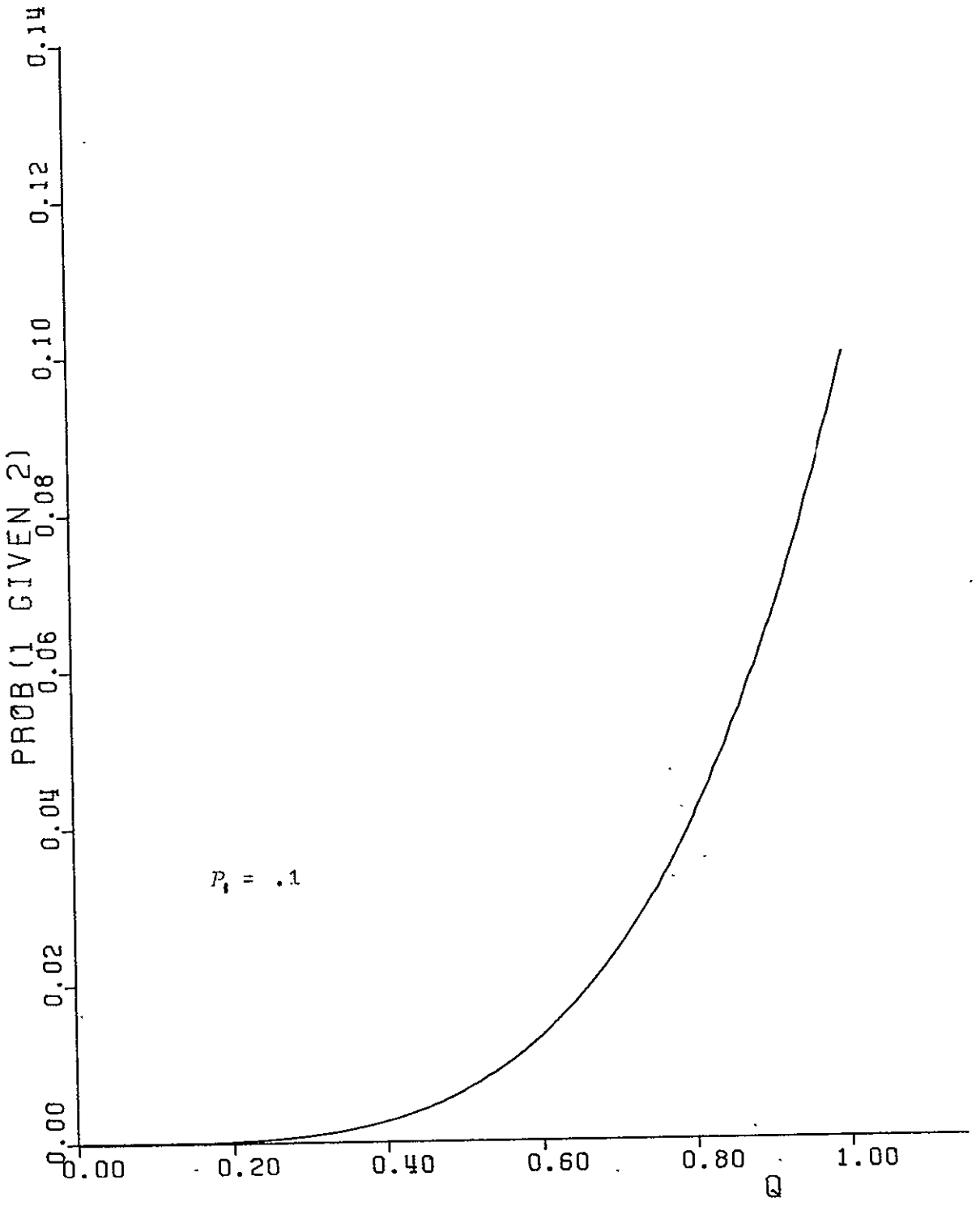
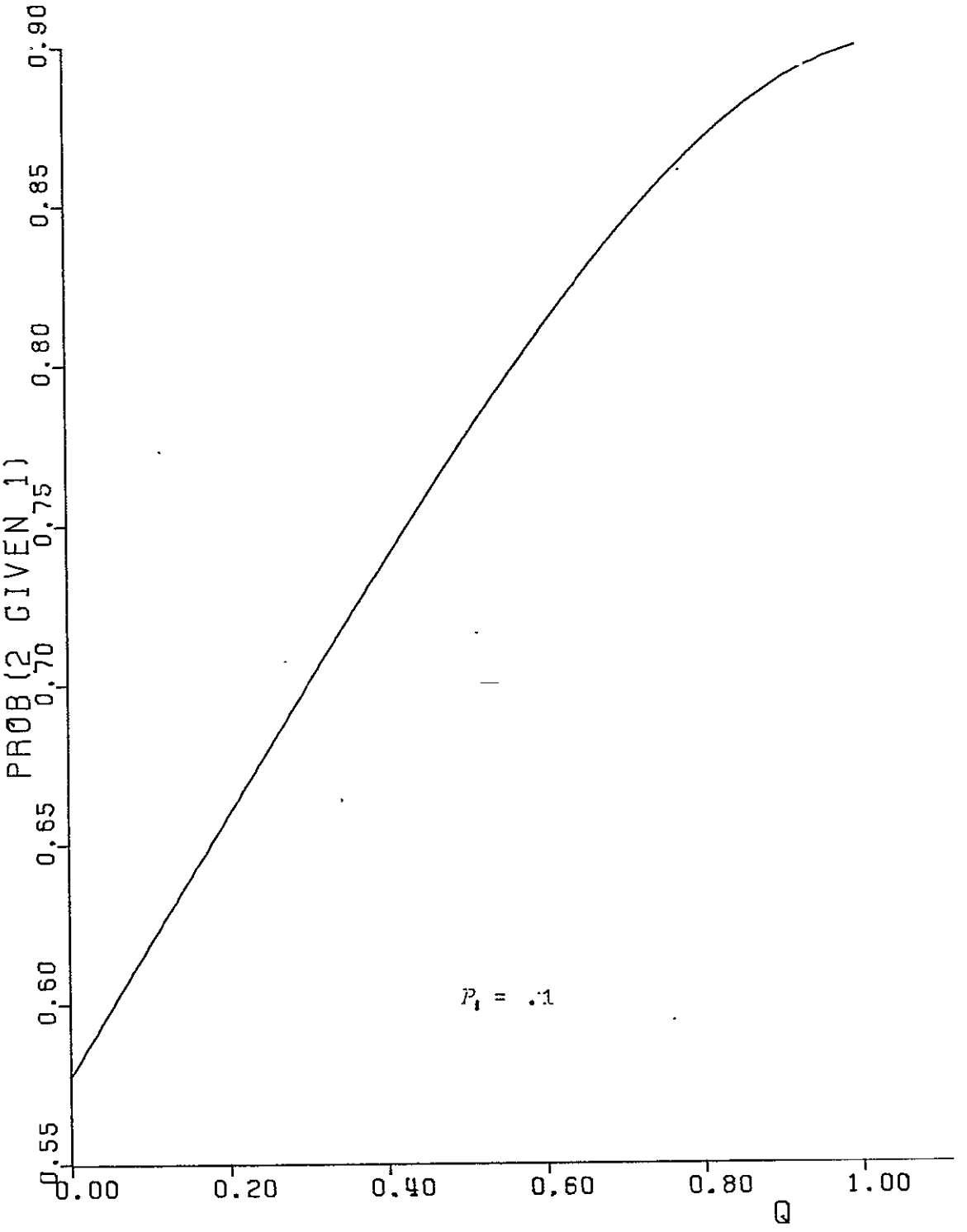

PROB(1 GIVEN 2)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .1

FIGURE A-1


PROB(2 GIVEN 1)
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .1

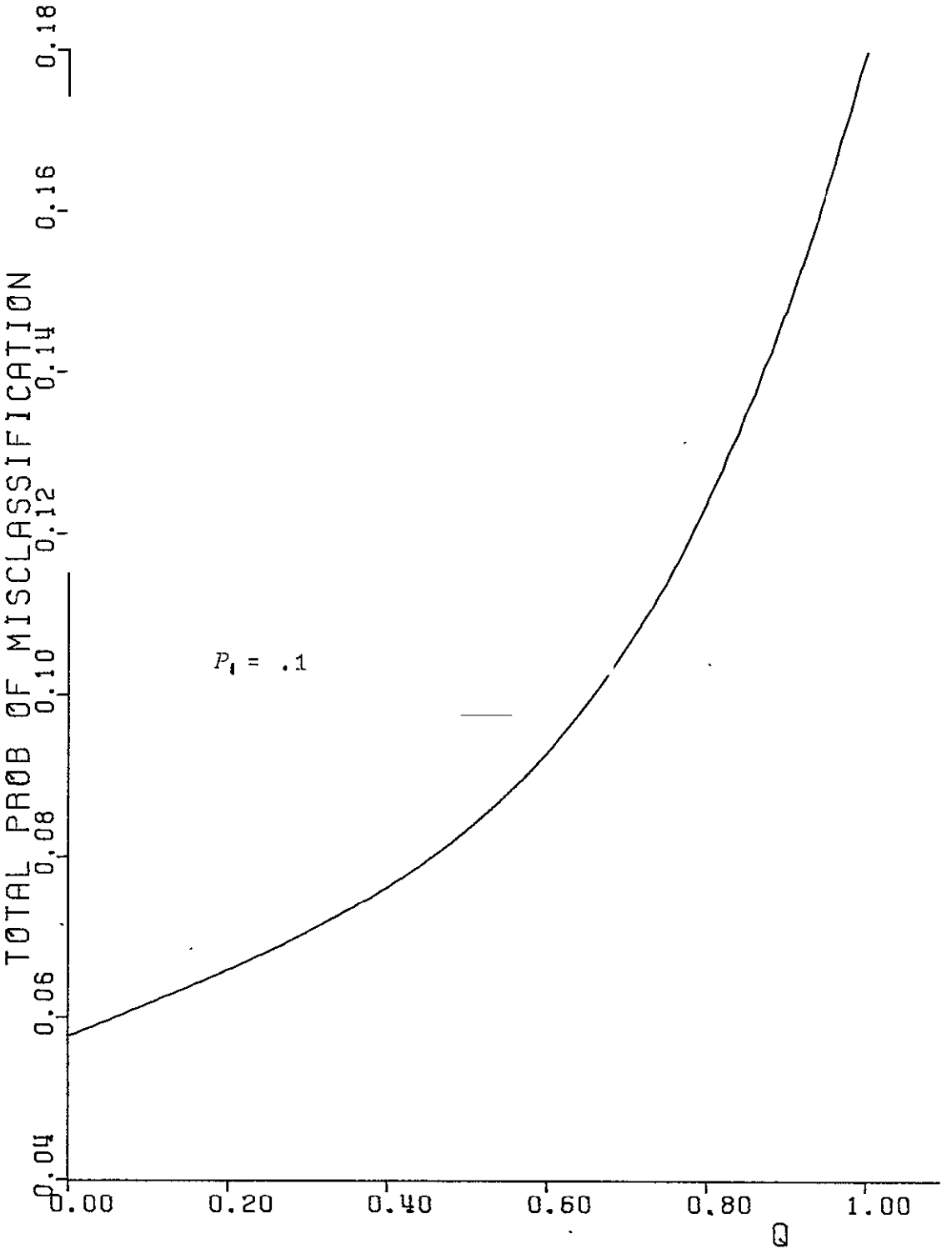

TOTAL PROB OF MISCLASSIFICATION
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
P_1 = .1
0.00
0.20
0.40
0.60
0.80
1.00
Q

FIGURE A-3

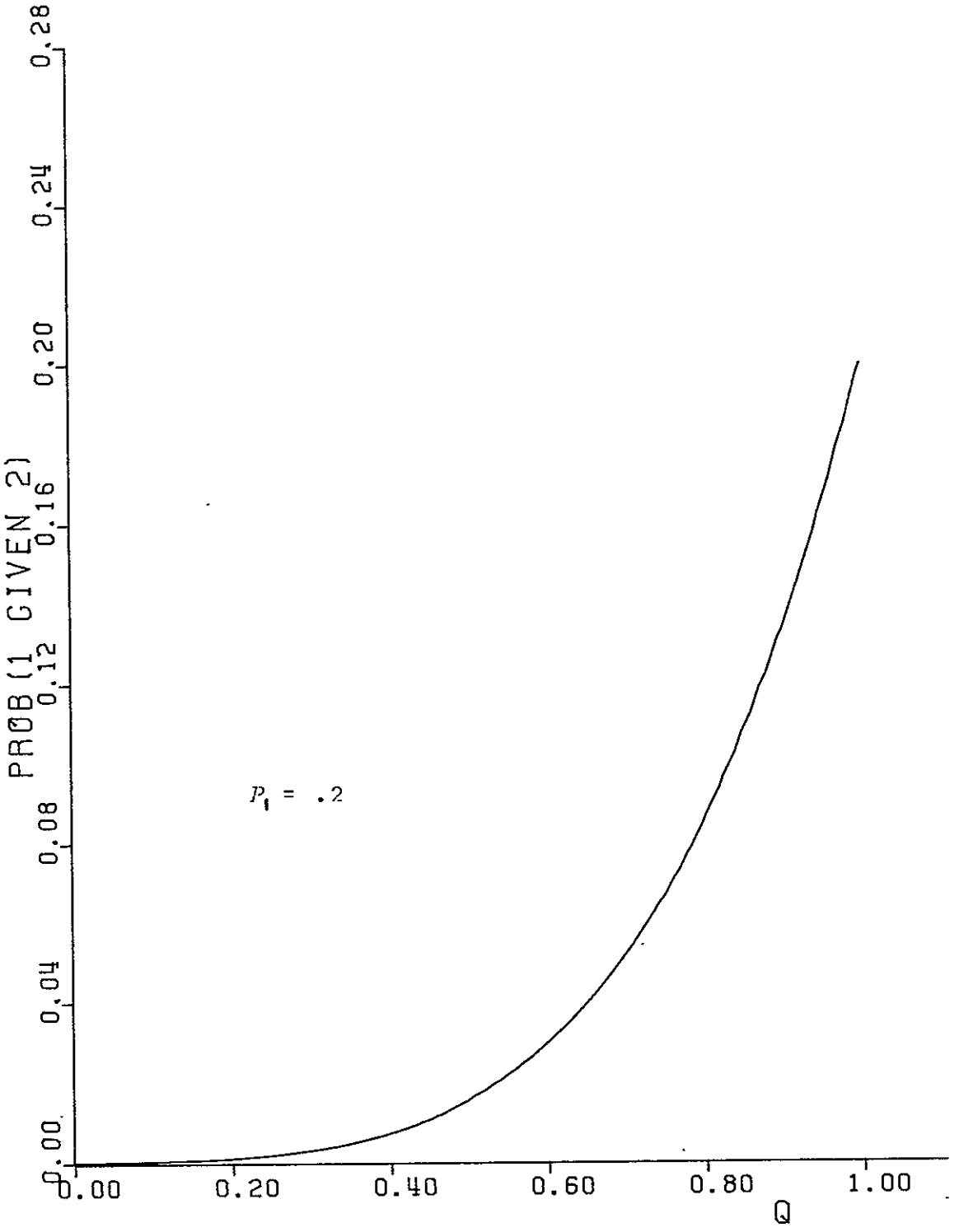
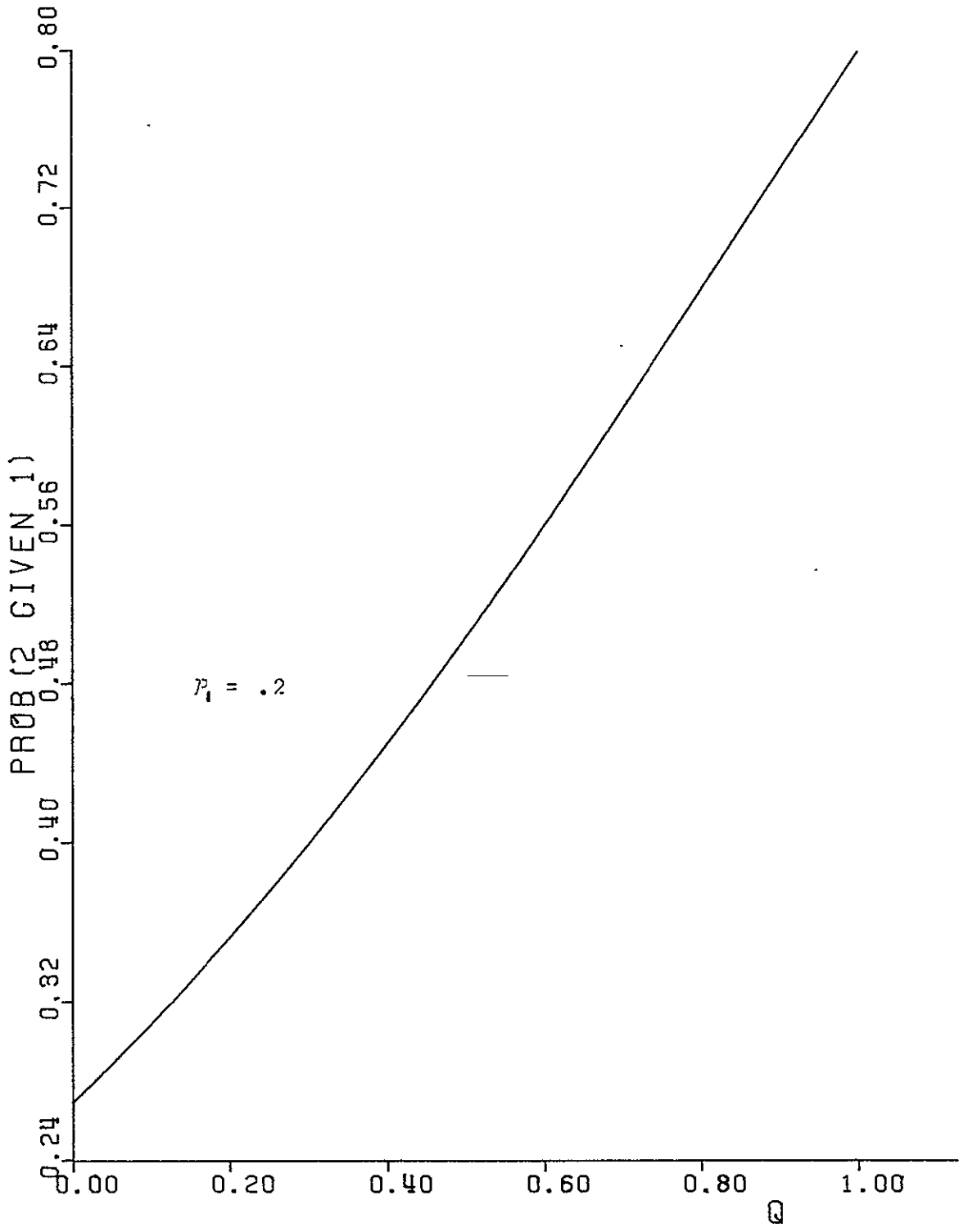

PROB(1 GIVEN 2)
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .2

FIGURE A-4


PROB(2 GIVEN 1)
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.00
0.20
0.40
0.60
0.80
1.00
Q
p_1 = .2

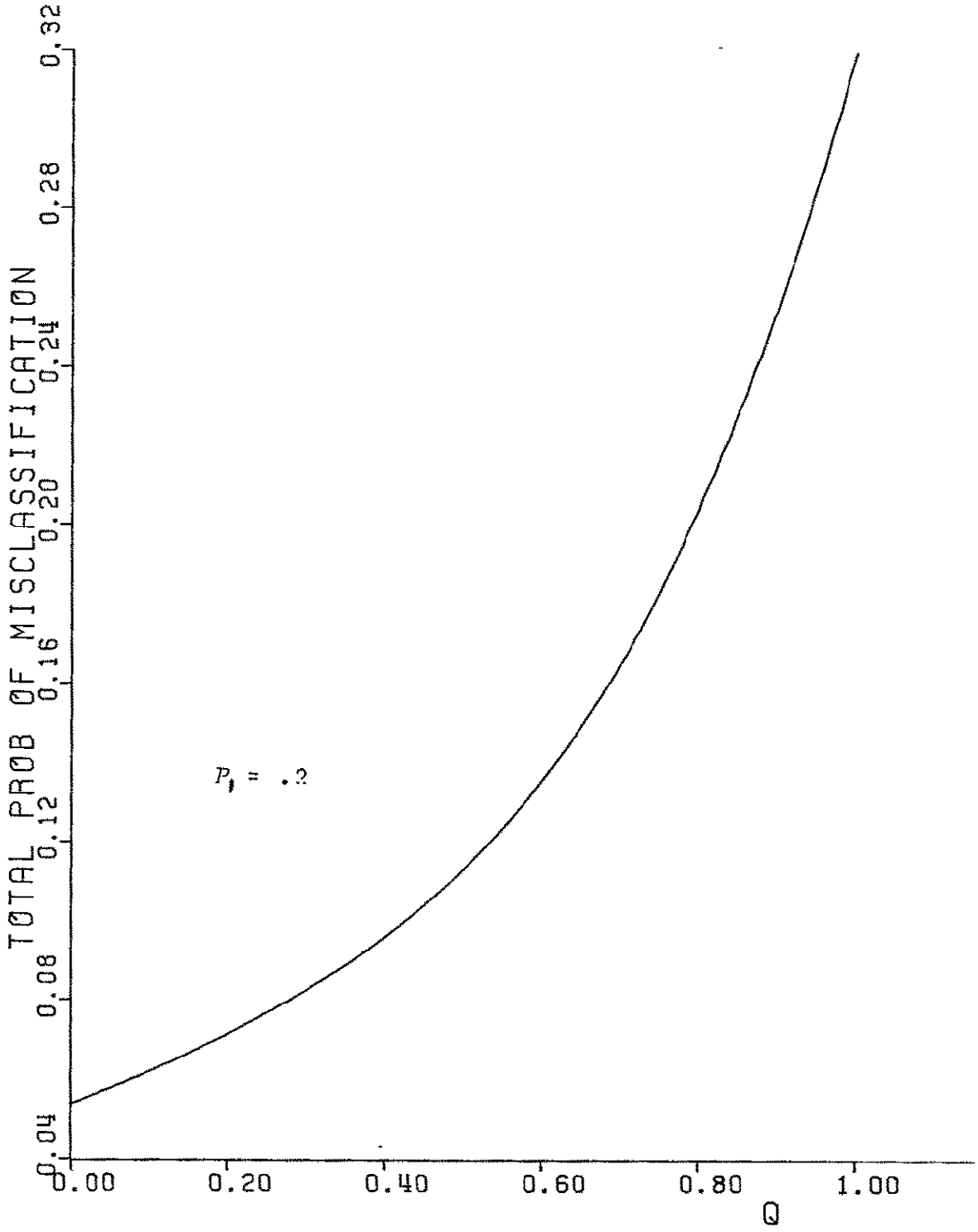

TOTAL PROB OF MISCLASSIFICATION
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.32
P_1 = .2
0.00
0.20
0.40
0.60
0.80
1.00
Q

FIGURE A-6

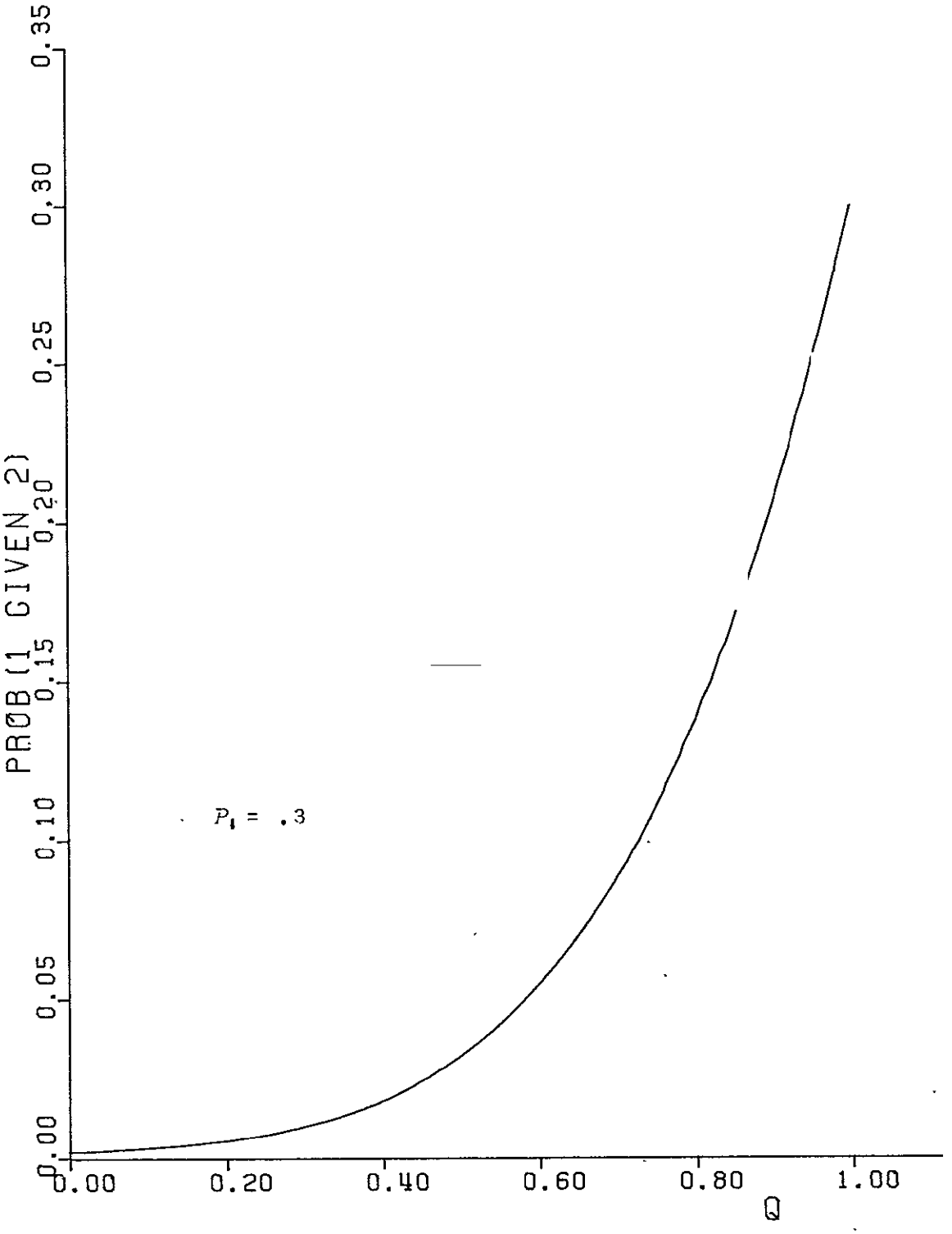

PROB(1 GIVEN 2)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
P_1 = .3
0.00
0.20
0.40
0.60
0.80
1.00
Q

FIGURE A-7

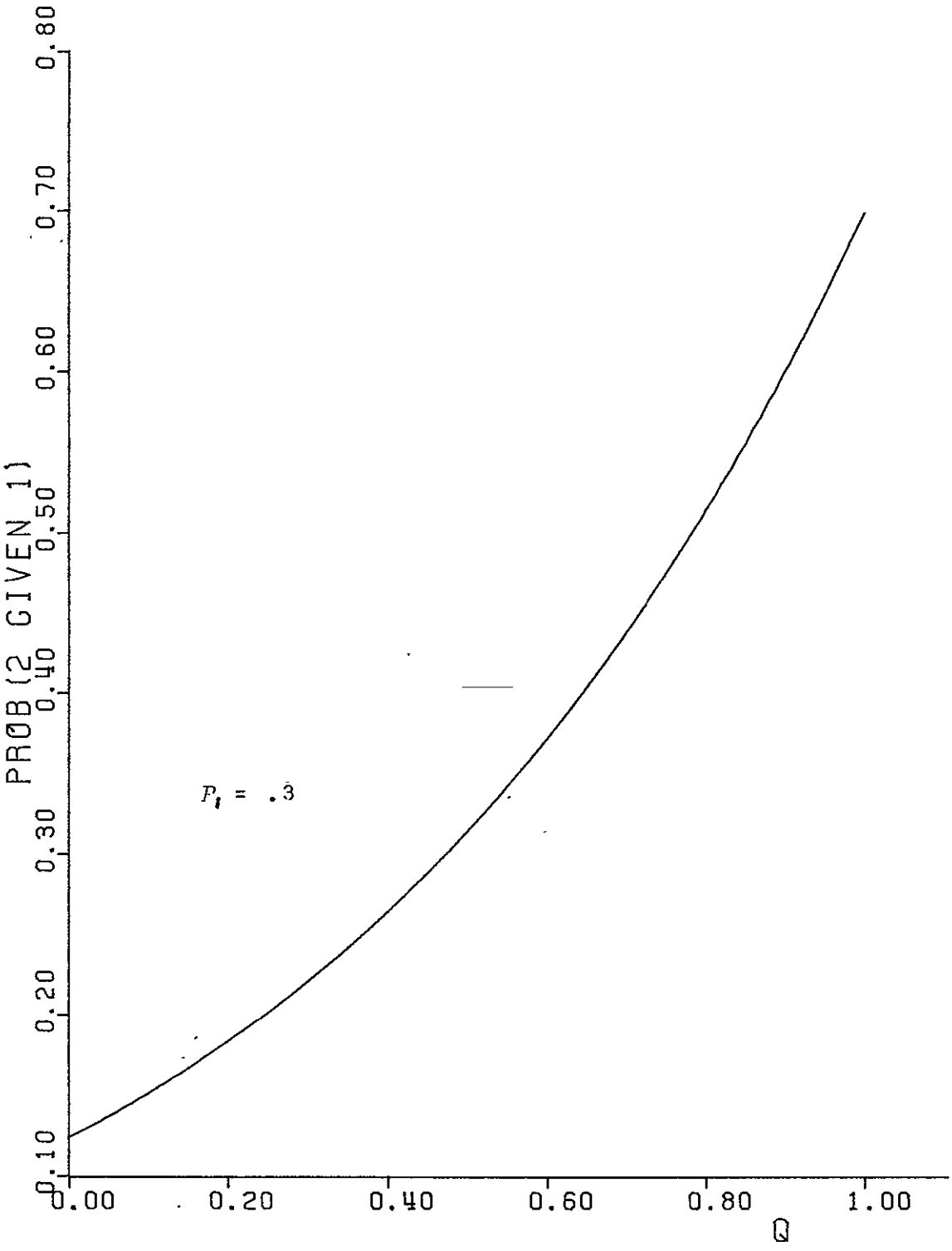

PROB(2 GIVEN 1)
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .3

FIGURE A-8

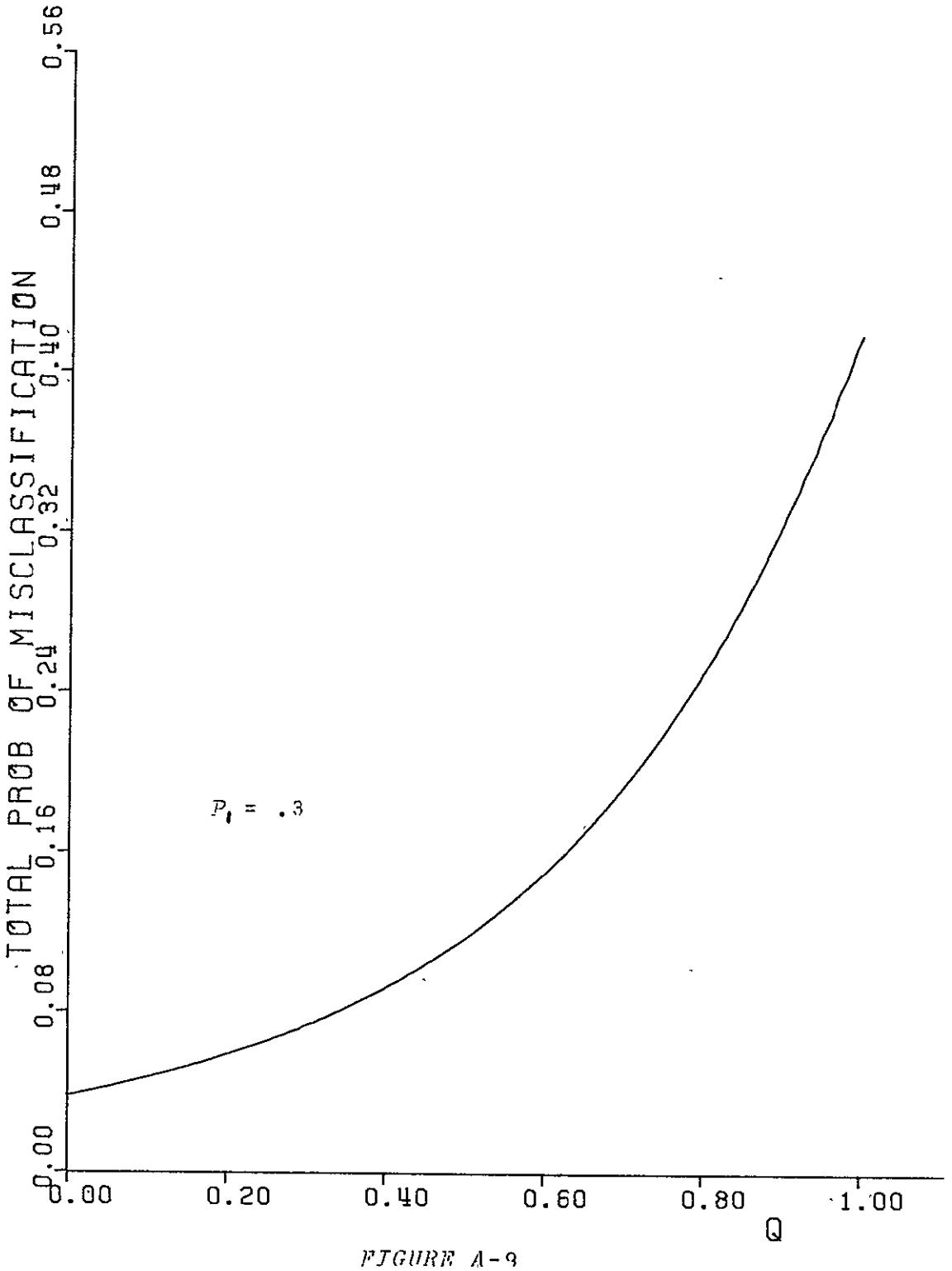

TOTAL PROB OF MISCLASSIFICATION
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .3

FIGURE A-9

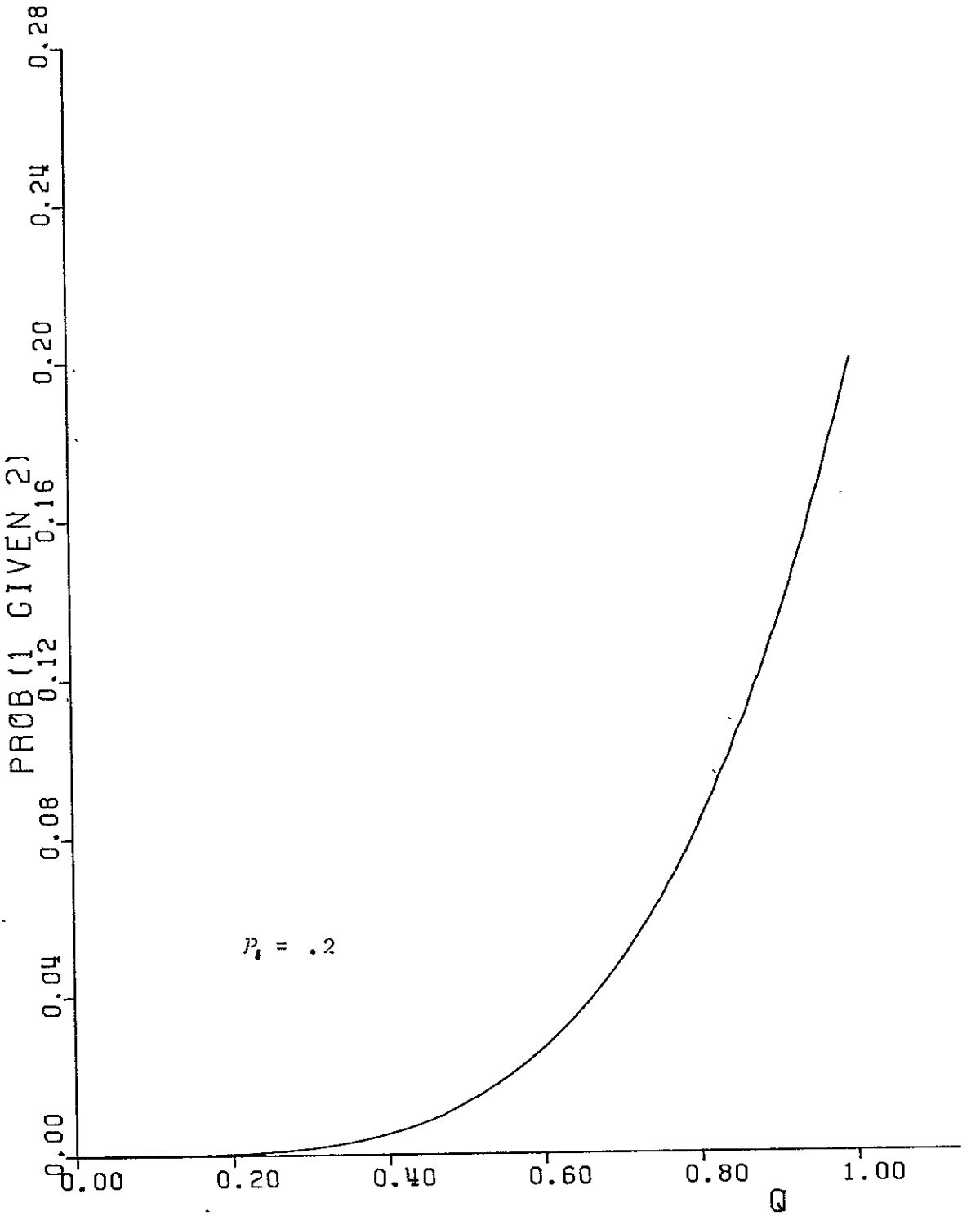

PROB(1 GIVEN 2)
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .2

FIGURE A-10

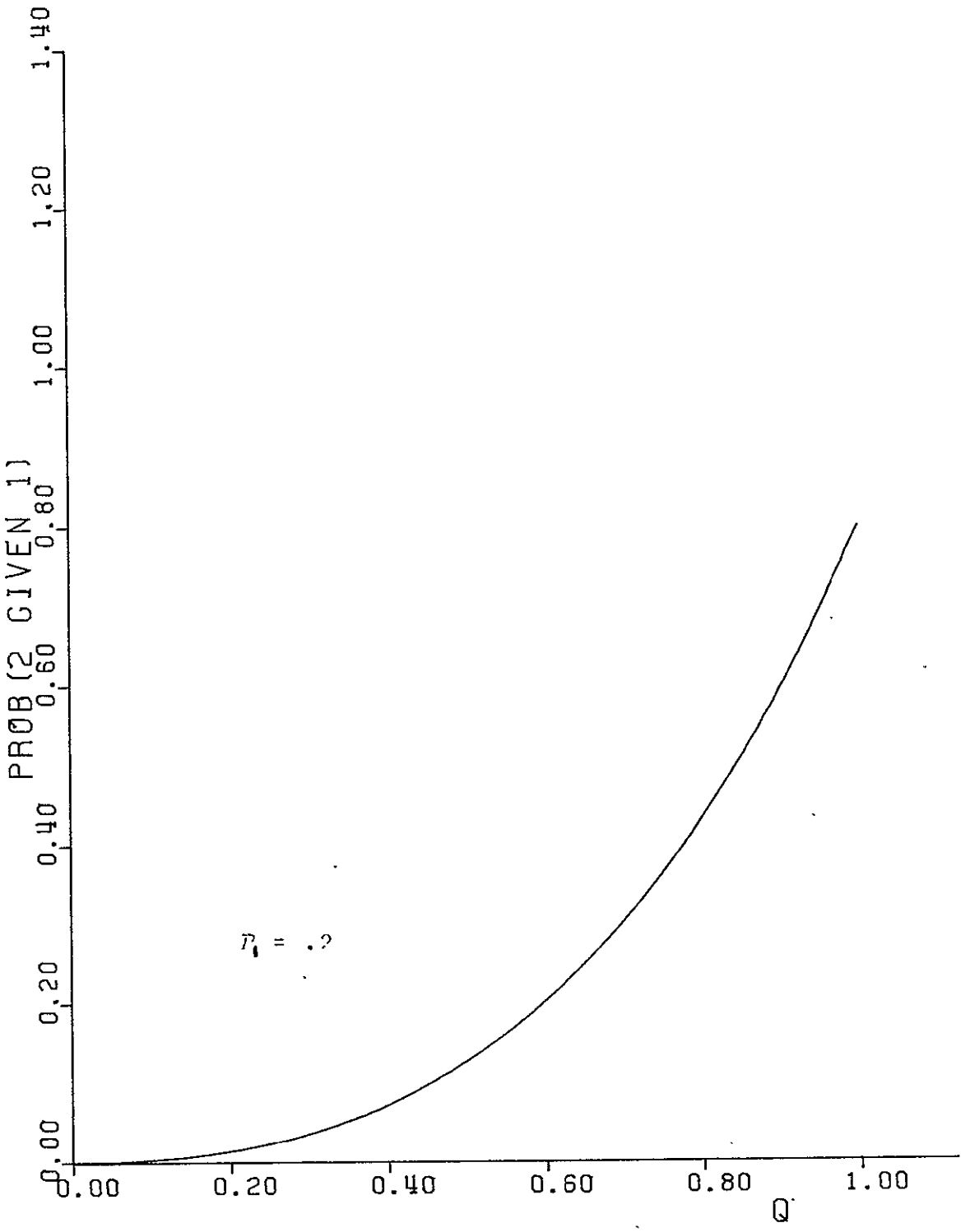

PROB(2 GIVEN 1)
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .?

FIGURE A-11

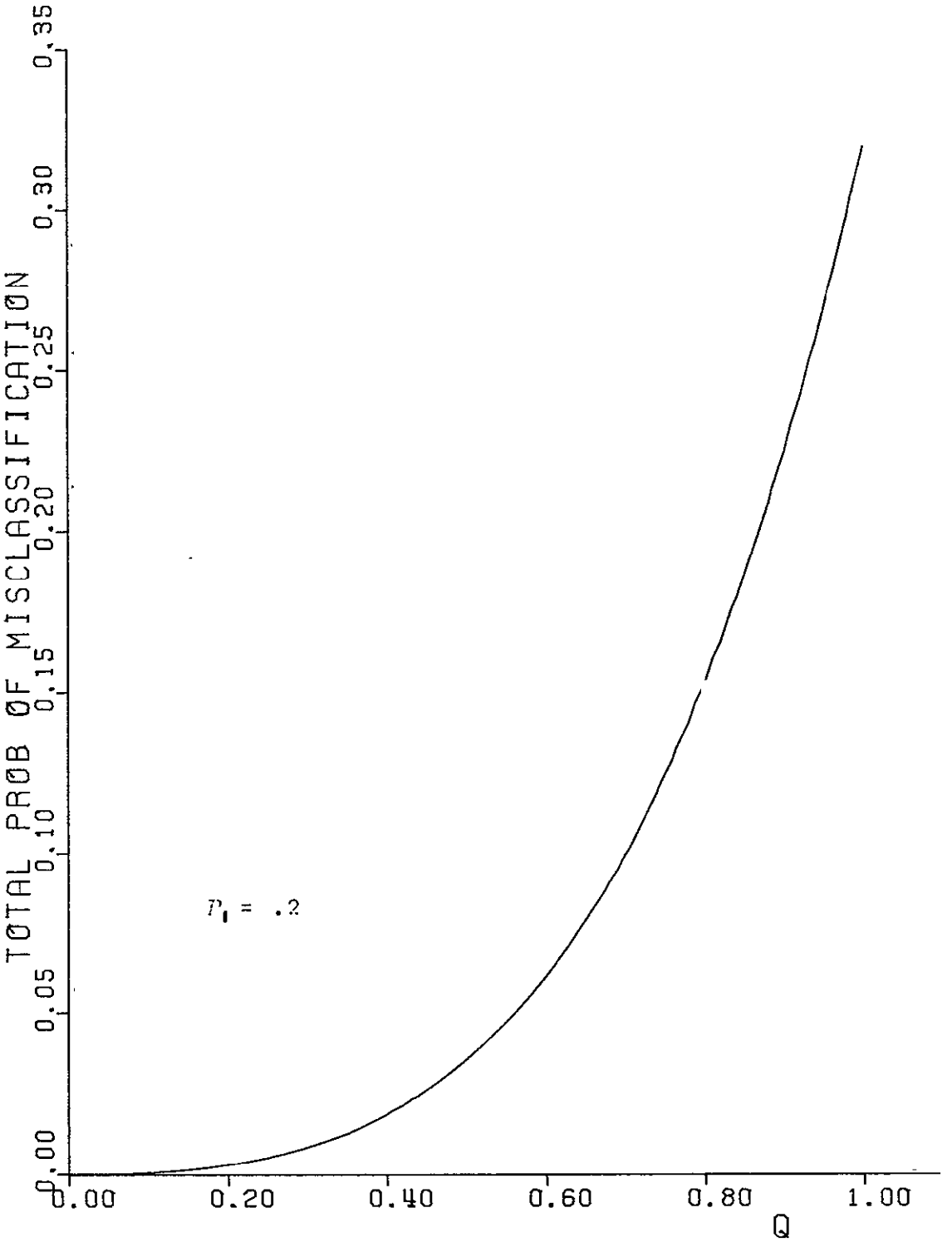

TOTAL PROB OF MISCLASSIFICATION
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .2

FIGURE A-12

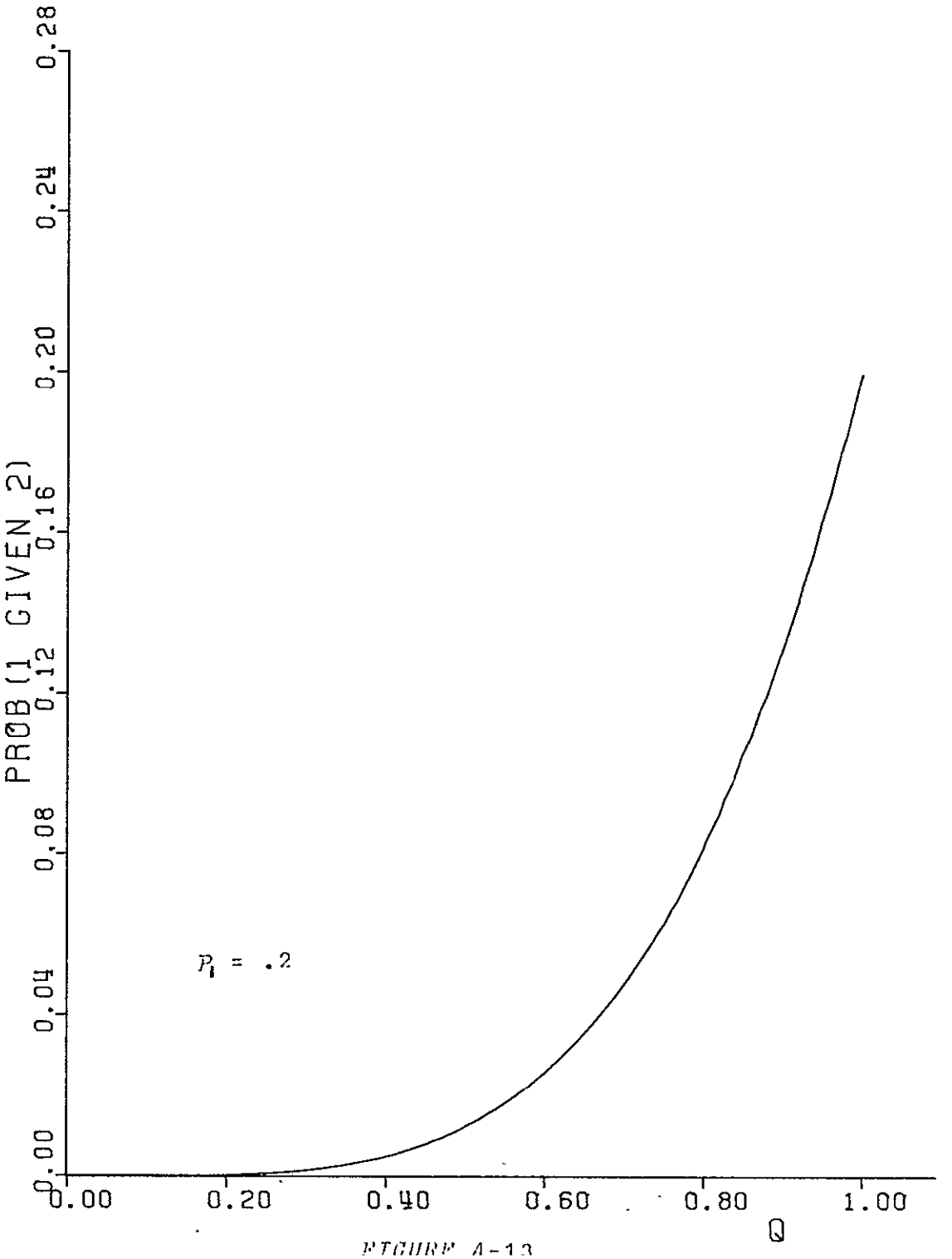

PROB(1 GIVEN 2)
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .2

FIGURE A-13

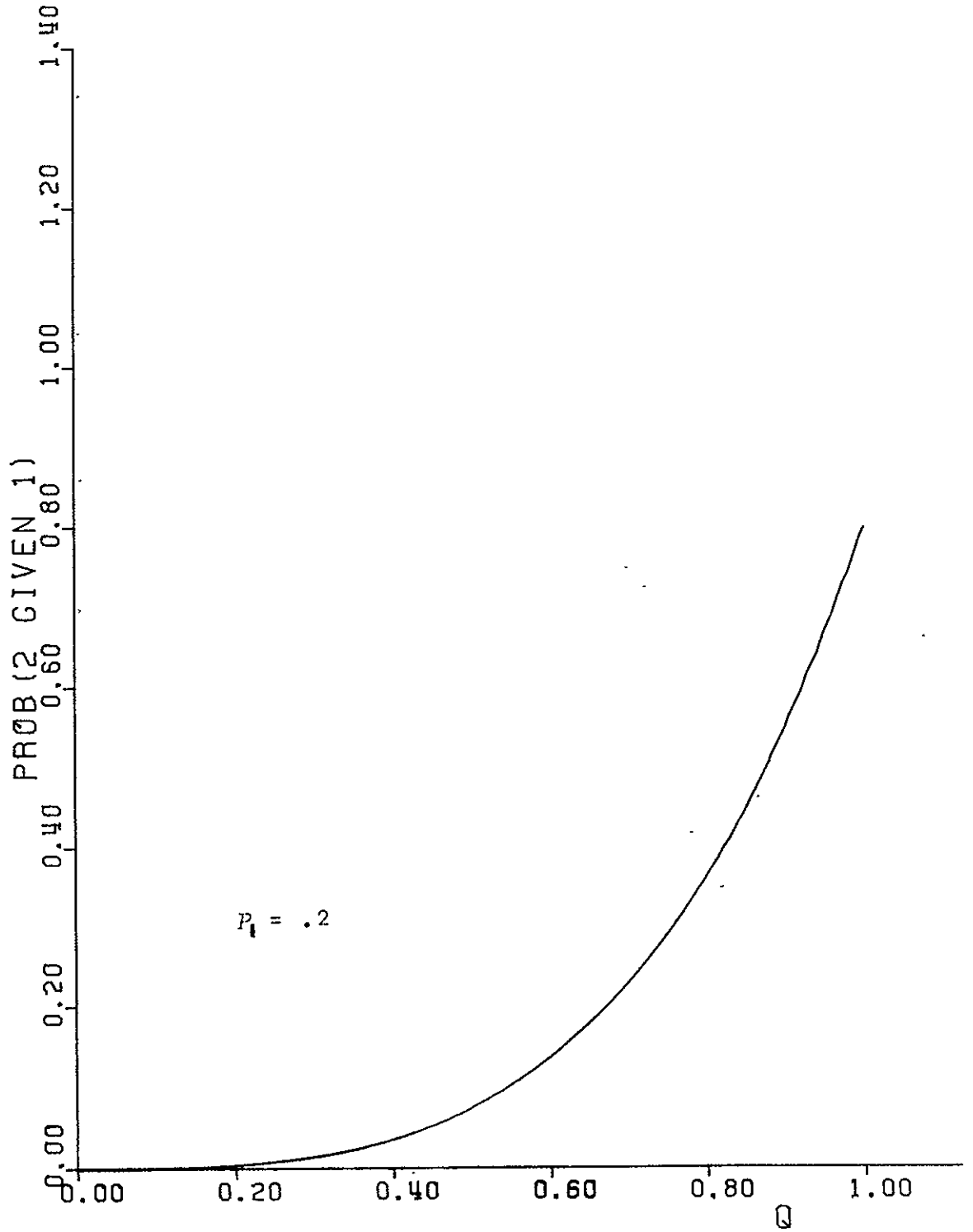

PROB(2 GIVEN 1)
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
0.00
0.20
0.40
0.60
0.80
1.00
Q
P_1 = .2

FIGURE A-14

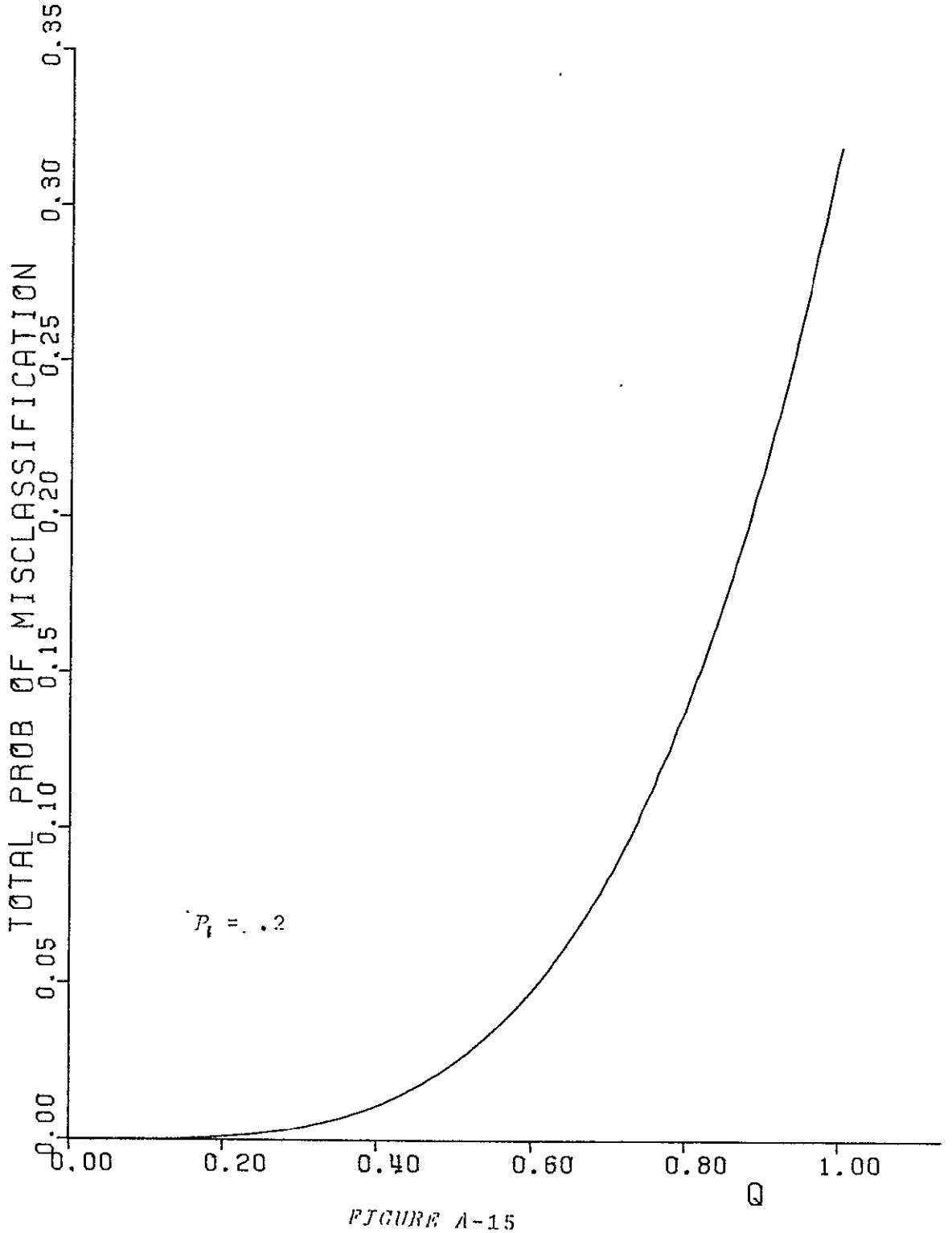

TOTAL PROB OF MISCLASSIFICATION
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
P_1 = .2
0.00
0.20
0.40
0.60
0.80
1.00
Q

FIGURE A-15

EXTRAPOLATION PROCEDURES FOR IRREGULARLY SPACED SPARSE DATA - A REVIEW AND COMPARISON

by

P. L. Odell*

A. H. Kvanli*

C. Simpson*

*University of Texas at Dallas

TABLE OF CONTENTS

1.0 Introduction

Much work has been done in recent years concerning the problem of extrapolating a set of irregularly spaced data to obtain a surface (possibly continuous but not necessarily) covering a large region of interest. The example used in this paper to illustrate the various extrapolation procedures will be an agricultural one but the methods can be used in other fields such as biology, meteorology, and city planning.

Suppose that we are given yield data (bushels per acre) for a particular agricultural crop such as wheat at several irregularly spaced locations within a region R. Designate these data as (x_1,y_1,z_1), ... , (x_N,y_N,z_N) where z_i = the yield at point (x_i,y_i). See Figure 1.

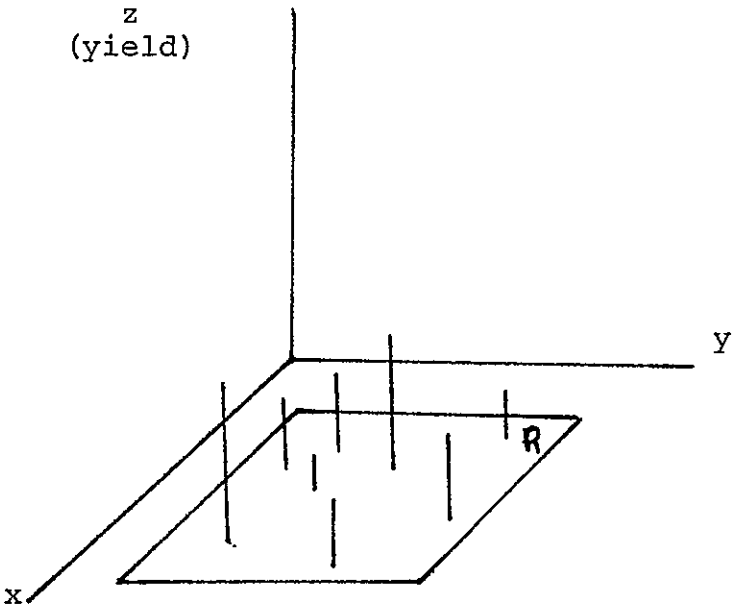


Figure 1

Consider these yield data as being particular points on a surface, henceforth referred to as a yield surface. The problem then is to devise a method which extrapolates this set of data over the entire region R, producing a yield estimate for each point of R. The term "extrapolate" as used in this paper may actually mean "interpolate" if the point under consideration lies within the convex hull defined by the data points. All of the proposed procedures for solving this problem will assume that a grid (coordinate system) has been constructed over R with each data point assigned to the nearest grid point. The problem then reduces to extrapolating the input data to each grid point lying within R.

A number of techniques ([1],[4],[8]) exist for interpolating within the convex hull of a set of data points, where the convex hull is the convex region of minimum area containing all of the data. The methods examined in this paper were selected because of their ability to extrapolate outside the convex hull of a sparse data set. One method (referred to as Method 3) is to fit the data points with a least squares linear surface of the form

$$\hat{z}_i = c_o + c_1 x_i + c_2 y_i \quad , \quad i = 1,2,\ldots,N.$$

McLain [3] expands this approach to include quadratic terms and (more importantly) uses a weighted least squares regression to determine the c_i's where the weights reflect the distance of the data points from (x_i, y_i). Consequently, a new set of regression coefficients is derived for each extrapolated point on the yield

surface. This method is referred to as Method 10 and will be discussed further in Section 2.9.

Shepard [9] has derived a surface generation technique which produces a surface that is continuous and passes through the data points. The method uses a weighted average of the data points to estimate the surface height at a given point where the weights reflect the relative distance and direction of the data points. This method is referred to as Method 9 in this paper and will be described in more detail in Section 2.8.

The methods described in subsequent sections were chosen because they seemed to be likely candidates for extrapolating sparse data. Because of the assumption of sparse data, the degree of sophistication the extrapolating procedure should have is questionable. One could argue that with an extremely sparse data set of yield data that using the sample mean to estimate the entire yield surface would be a "safe and reasonable" thing to do. On the other hand, one should possibly use a more sophisticated technique to glean as much information as possible from the set of data, although it becomes increasingly difficult to substantiate this sophistication from your sample data. Consequently, the authors present and compare ten methods ranging from the sample mean to the highly sophisticated procedure proposed by Shepard. It is hoped that the reader will apply each of these procedures to his own particular application to determine which of them most accurately models his situation.

In Section 3.0, the authors compare these methods using five years of wheat yield data from North Dakota. For this study, five years of data existed for seven yield test stations along with yield data for approximately 45 check points (cities). A Fortran contour mapping program was written and used on several of the methods to compare the final yield surfaces to each other and to a contour map of the full set of available data (i.e., the 52 yield data points). The latter map was assumed to accurately model the actual wheat yield distribution of North Dakota and was used as a check of the various extrapolation procedures.

Section 4.0 contains the conclusions of this study along with a table containing the desirable and undesirable characteristics of each model.

2.0 Proposed Methods of Solution

As mentioned previously, these methods will vary considerably in their complexity beginning with a very simple technique of yield extrapolation (Method 1) to the more sophisticated procedures of Shepard and McLain (Methods 9 and 10).

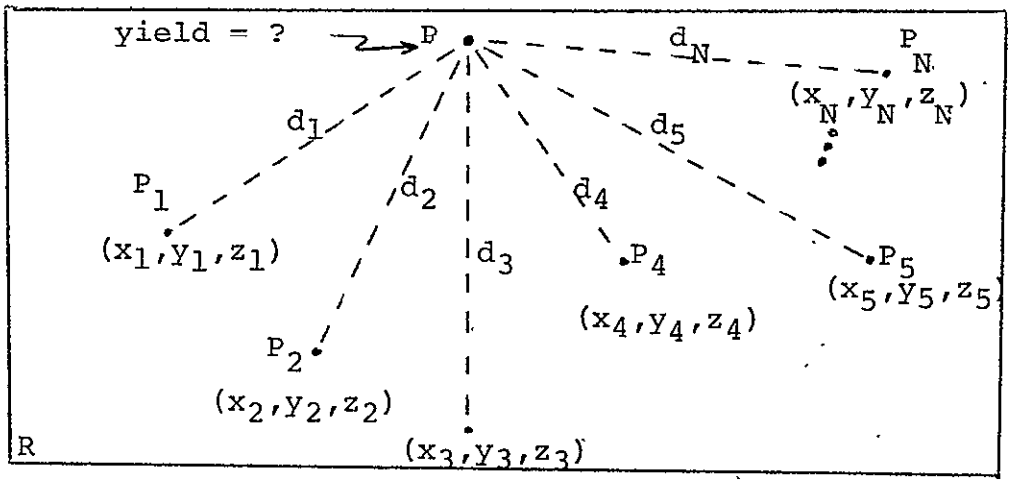


Figure 2

2.1 Method 1 - Composite Average

If one wishes to ignore the relative location of the yield data points, then a reasonable estimator for the yield at any point in R would be the sample mean, i.e.

$$\hat{z}_{P,1} = \text{yield estimate at P using Method 1}$$
$$= \frac{1}{N} \sum_{1}^{N} z_i = \bar{z}$$

where z_i = yield at P_i.

This method would be appropriate if the user ignores location because of the extreme sparcity of the data or he feels that the yield values are randomly distributed about R. This estimator provides a continuous surface which is well behaved near the boundary of R. However, it does not reflect any variations within the region, in particular in small neighborhoods about each data point. One rationale behind using $\bar{z}$ is that while it may serve as a relatively poor local (point) estimator it should provide an accurate global estimate of the <u>total</u> yield for the region, R.

2.2 Method 2 - Nearest Neighbor (NN)

This method (another conservative estimator) estimates the yield at P by the yield of that data point which is <u>nearest</u> P, i.e.

$$\hat{z}_{P,2} = \text{yield estimate at P using Method 2}$$
$$= z_k,$$

where (1) k is such that $d_k = \text{Min}_{1 \leq j \leq N} \{d_j\}$

(2) d_j = Euclidean distance from P to P_j.

This estimator provides better estimates in small neighborhoods about each data point and also is well behaved near the boundaries of R. However, the resulting yield surface is discontinuous with

the maximum yield estimate being the maximum z_i. For each point P, this estimator completely ignores the yield information of the N-1 furthest data points.

2.3 Method 3 - Least Squares Linear Surface (LS)

The Method 3 estimator fits the yield data with a plane that provides a least squares fit to the set of yield data. So

$$\hat{z}_{P,3} = c_o + c_1 x_P + c_2 y_P$$

where (1) coordinates of P are (x_P, y_P)

(2) c_o, c_1, c_2 are given by

$$c_1 = \frac{1}{D}\,(s_{yy}s_{xz} - s_{xy}s_{yz})$$

$$c_2 = \frac{1}{D}\,(s_{xx}s_{yz} - s_{xy}s_{xz})$$

$$c_o = \frac{1}{N}\,(s_z - c_1 s_x - c_z s_y)$$

and (3) $$s_{xx} = \sum_1^N (x_i - \bar{x})^2$$

$$s_{yy} = \sum_1^N (y_i - \bar{y})^2$$

$$s_{xy} = \sum_1^N (x_i - \bar{x})(y_i - \bar{y})$$

$$s_{xz} = \sum_1^N (x_i - \bar{x}) z_i$$

$$s_{yz} = \sum_1^N (y_i - \bar{y}) z_i$$

$$s_z = \sum_1^N z_i, \quad s_y = \sum_1^N y_i, \quad s_x = \sum_1^N x_i$$

$$D = s_{xx}s_{yy} - s_{xy}s_{xy}$$

This estimator provides intuitively "good" yield estimates in the convex hull (see Section 2.4) of the data points. The resulting surface is continuous and does reflect trends in the data (e.g. the

yield values increase from North to South). The major problem with this technique, as with any linear regression technique, is that it may be extremely ill behaved outside the range of the input data points, i.e. near the boundary of R. If the yield surface (a plane) is extremely inclined, one can encounter negative or unrealistically large yield estimates.

2.4 Method 4 - LS/NN

This method attempts to provide more conservative yield estimates for points outside the convex hull defined by the input points. The convex hull of a set of data points is defined to be that convex region of minimum area containing all of the points. The authors have developed an algorithm (see Appendix A) and computer program (Fortran) to determine the convex hull of an input set of two-dimensional data points. Figure 3 illustrates this for a set of six data points.

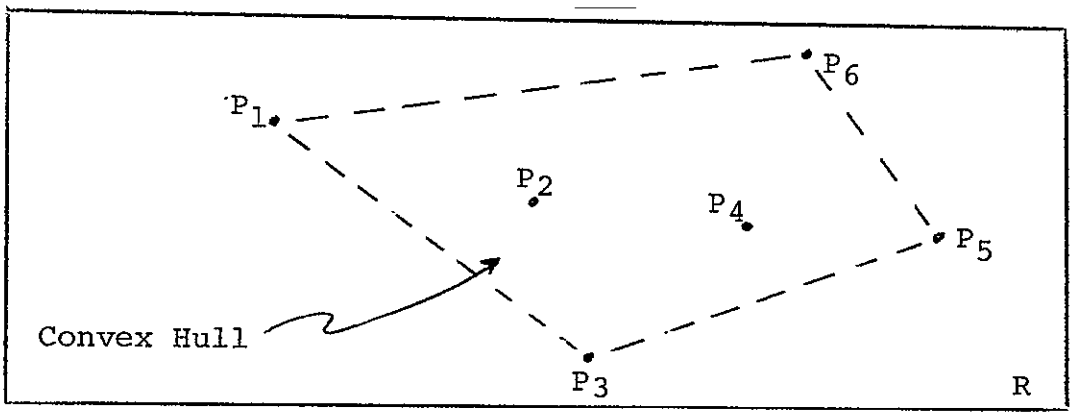


Figure 3

The LS/NN estimator uses the LS estimator (Method 3) inside the convex hull and the nearest neighbor estimator (Method 2) outside this region, i.e.

$$\hat{z}_{P,4} = \begin{cases} \hat{z}_{P,3} & \text{if } P \in \text{convex hull of data points} \\ \hat{z}_{P,2} & \text{otherwise} \end{cases}$$

The resulting surface although possibly better behaved near the boundary of R will be discontinuous at the boundary of the convex hull. However, for data that varies considerably, this more conservative estimator may be much more desirable than the overall regression estimator.

2.5 Method 5,6 - Average Linkage

The average linkage estimator predicts yield by a using a weighted average of the data points. Thus

$$\hat{z}_{P,j} = \left(\sum_{i=1}^{N} g(d_i) z_i\right) / \left(\sum_{i=1}^{N} g(d_i)\right)$$

where (1) d_i = distance from P to P_i

(2) $g(d_i)$ is a monotonically decreasing function of d_i such that $g(d_i) \longrightarrow 0$ as $d_i \longrightarrow \infty$.

For j=5, $g(d_i) = 1/d_i$ and for j=6, $g(d_i) = 1/d_i^2$. The Method 6 estimator will reduce the effect of distant data points more so than Method 5, i.e. the effect of the i-th data point on the yield estimate at P "dies out" faster using Method 6. If one wishes to further decrease the effect of distant data points, a possible candidate would be $g(d_i) = e^{-d_i}$ or $e^{-d_i^2}$.

The resulting yield surface using either method will be continuous and is well behaved near the boundary of R. Moreover, it includes the effect of each data point yet is completely dominated by the yield at P_k for any P in a small neighborhood about P_k. The resulting surface is conservative, however, since the maximum yield value will occur at one of the data points. The average linkage method is also unable to reflect any directional trends in the data.

2.6 Method 7 - Average Linkage with Directional Correlation

The Method 7 estimator attempts to improve the average linkage estimator by incorporating significant directional trends that are present in the data.

$$\hat{z}_{P,7} = (1-w_1-w_2)\ z_{P,6} + w_1\hat{z}_{P,7,x} + w_2\hat{z}_{P,7,y}$$

where (1) $$w_1 = \begin{cases} w \times [\delta_x/(\delta_x + \delta_y)] & \text{if } \delta_x \text{ or } \delta_y = 1 \\ 0 & \text{if } \delta_x = \delta_y = 0 \end{cases}$$

(2) $$w_2 = \begin{cases} w \times [\delta_y/(\delta_x + \delta_y)] & \text{if } \delta_x \text{ or } \delta_y = 1 \\ 0 & \text{if } \delta_x = \delta_y = 0 \end{cases}$$

(3) $$w = w_{max} \times [\delta_x |\hat{\rho}_{x,z}| + \delta_y |\hat{\rho}_{y,z}|] \times (\delta_x + \delta_y)^{-1}$$

(4) $$\delta_x = \begin{cases} 1 & \text{if } |\hat{\rho}_{x,z}| \geq \rho_{min} \\ 0 & \text{if } |\hat{\rho}_{x,z}| < \rho_{min} \end{cases}$$

(5) $$\delta_y = \begin{cases} 1 & \text{if } |\hat{\rho}_{y,z}| \geq \rho_{min} \\ 0 & \text{if } |\hat{\rho}_{y,z}| < \rho_{min} \end{cases}$$

(6) $\hat{\rho}_{x,z}$ = sample correlation between x_i's and z_i's

(7) $\hat{\rho}_{y,z}$ = sample correlation between y_i's and z_i's

(8) $\hat{z}_{P,7,x}$ = yield estimate at P obtained by regressing yield on the x-coordinate, i.e. $\hat{z}_{P,7,x} = A_x + B_x \cdot x_P$ for regression coefficients A_x, B_x

(9) $\hat{z}_{P,7,y}$ = yield estimate at P obtained by regressing yield on the y-coordinate, i.e. $\hat{z}_{P,7,y} = A_y + B_y \cdot y_P$ for regression coefficients A_y, B_y.

(10) w_{max}, ρ_{min} are predetermined constants.

w_{max} essentially measures the maximum importance that the experimenter wishes to give to the directional trend estimates and ρ_{min} is the minimum <u>significant</u> directional correlation that is allowed. Consider the network of data in Figure 3.

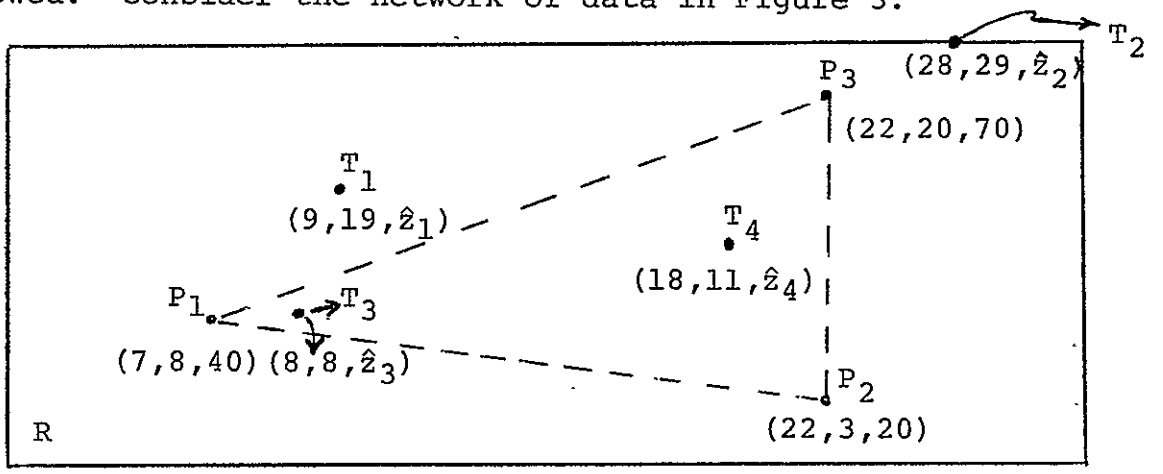


Figure 3

Here we have yield sample data of 40,20, and 70 at P_1, P_2, P_3 respectively. Setting $\rho_{min} = .8$, and computing directional correlations, then $\hat{\rho}_{x,z} = .114$ (not significant) and $\hat{\rho}_{y,z} = .992$ (significant). Regressing yield on the y coordinates produces $\hat{z}_{P,7,y} = 13.78 + 2.86\ y_P$. Letting $w_{max} = .8$, then $w = .8 \times .992 = .794$, and

$$\hat{z}_{P,7} = .206\ \hat{z}_{P,6} + .794\ \hat{z}_{P,7,y}$$
$$= .206\ \hat{z}_{P,6} + .794 \times (13.78 + 2.86\ y_P)$$

Thus:
$$\hat{z}_{t_1,7} = .206\ (49.36) + 10.941 + 2.271 \cdot (19)$$
$$= 64.26$$
$$\hat{z}_{t_2,7} = .206\ (61.61) + 10.941 + 2.271 \cdot (29)$$
$$= 89.49$$
$$\hat{z}_{t_3,7} = .206\ (40.00) + 10.941 + 2.271 \cdot (8)$$
$$= 37.35$$

$$\hat{z}_{t_4,7} = .206\ (42.08) + 10.941 + 2.271\ (11)$$
$$= 44.59$$

The choice of w_{max} is, of course, very subjective. If the experimenter feels that apparant directional data trends should be heavily considered, then a large value of w_{max} would be appropriate. If one is uncertain about such trends, then a lesser value of w_{max} (.3 - .7) should be used, and if one wishes to totally disregard any apparant trends, then the average linkage estimator should be used. It should be pointed out that $\hat{z}_{P,6}$ (the average linkage estimator) can be replaced by any of the remaining methods in this paper. Whichever method one uses in place of $\hat{z}_{P,6}$, the result of $\hat{z}_{P,7}$ will be to incorporate the effect of directional correlations into this yield estimator. Another question to consider in choosing w_{max} is "how representative are the data points of the entire region, R?" If the points are, say, clustered in a small portion of R, then one may be wary of directional predictors and may wish to use a more conservative estimator by choosing a small value for w_{max}. If, however, the points are distributed evenly about the entire region then this would suggest using a larger value for w_{max}.

The resulting yield surface will be continuous, but could, in some cases, be ill behaved near the boundary of R. This can be controlled somewhat by the choice of w_{max}.

2.7 Method 8 - Objective Analysis

This method is a result of a meteorological approach to sparse data extrapolation. Sasaki ([5], [6], [7]) and Wagner [10]

have proposed such a method referred to as an objective approach to analyzing data fields. This method also utilizes a low pass filter to compensate for the sparcity of data and to suppress the high-frequency noise contained in the initial data. This procedure can be broken down into the three following steps:

Step 1: Initialization

The region of interest, R, is fitted with a grid and, as in all the previous methods, the object will be to provide a yield estimate at each such grid point. The initialization process begins by assigning the N sample values to their nearest grid point and setting the remaining z (yield estimate) values equal to zero. Finally, R is augmented by a set of boundary grid values, all set equal to zero (see Figure 4). Call the new region R´.

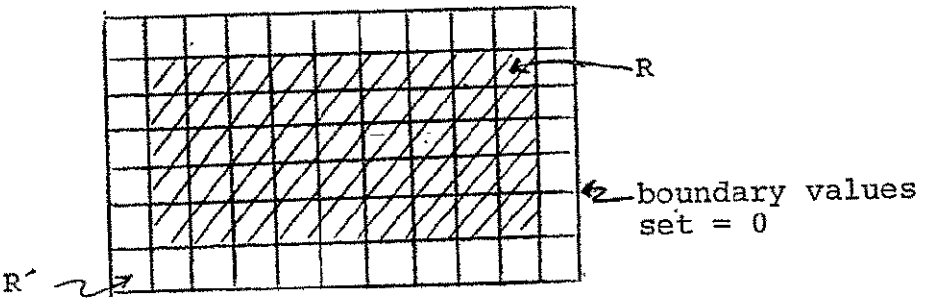


Figure 4

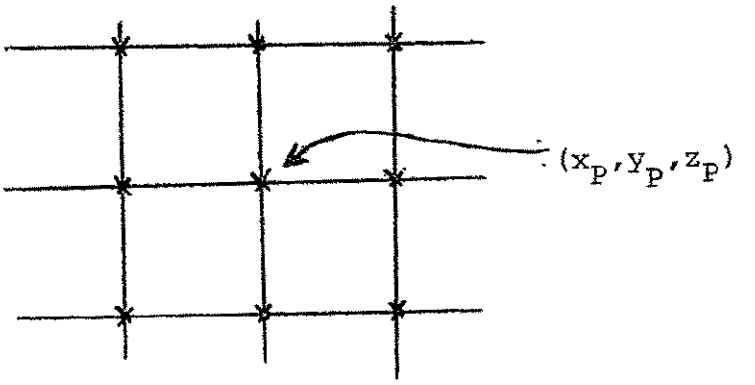


Figure 5

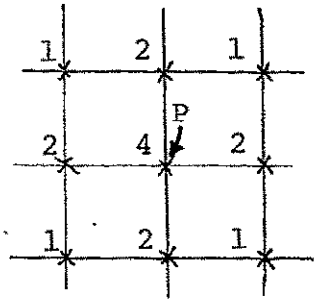
Referring to Figure 5, consider an arbitrary grid point P. The initialization process is actually a sequential procedure which continues until all grid points have been assigned a non-negative yield estimate. At each step, the yield estimate at point P is z_P, where

$$z_P = \begin{cases} \text{existing } z_P & \text{if } z_P > 0 \\ \text{average of the 9 points in Figure 5} & \text{if at least one of these points is} > 0 \\ \text{unspecified} & \text{if all 9 points} = 0 \end{cases}$$

Thus on the first pass, z_P = the existing z_P only if point P is one of the original N data points. After point P is finally assigned a non-zero initial estimate, it remains unchanged on succeeding steps of the initialization procedure. This procedure continues until <u>all</u> of the grid points have been assigned on non-zero initial yield estimate (i.e. no z_P is unspecified).

Step 2: Smoothing

As a first attempt to filter out some of the high-frequency components of the initialization process, a smoothing process (filter) is applied.



Figure

Considering, for the moment, a one dimensional situation, this smoothing filter can be written as

$$(1 - S)\, f_j + \frac{S}{2}\,(f_{j+1} + f_{j-1}) \qquad (2.7.1)$$

for an arbitrary function f_j and constant S. For S = 1/2, (2.7.1) reduces to $f_j' = 1/4\ (f_{j+1} + 2f_j + f_{j-1})$.

Referring to Figure 6 for the two-dimensional case, the smoothing algorithm replaces z_p by its "smoothed" value

$$z_p' = \sum_{i=1}^{9} w_i z_i/16 \qquad (2.7.2)$$

where (1) the index runs over the 9 points in Figure 6

(2) the weights (w_i) are the numbers to the left of each point in Figure 6.

For points on the boundary (but not the corner) of R (not R'), the nine point weighted average of (2.7.2) is replaced by the corresponding six point average where the weights are illustrated in Figure 7. Finally, using Figure 8, the smoothed values for the corner points of R can be determined using a four point weighted average. Note that for this case, the denominator of (2.7.2) is replaced by 8.0.

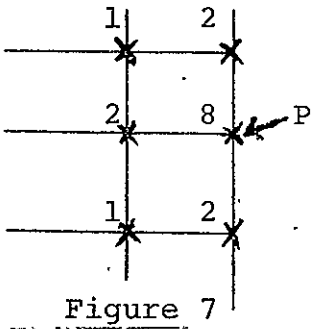


Figure 7

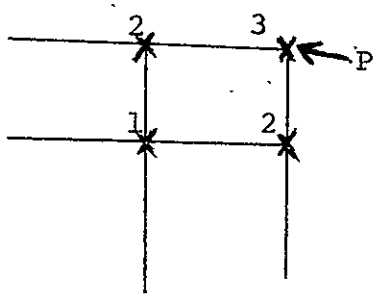


Figure 8

Step 3: Extrapolation Using Observational and Low Pass Filtering Constraints

Briefly, this step attempts to define a surface (field) of data which does not differ a great deal from the surface generated by Step 2 yet, at the same time, contains a minimum of high-frequency modes. The problem is essentially one belonging to the

Calculus of Variations where the purpose is to minimize I, given in (2.7.3).

$$I = \Sigma_R \left[\alpha(z - z')^2 + \alpha_1(\nabla_i z)^2\right] \quad (2.7.3)$$

where (1) z' represents the final surface generated by Step 2

(2) z is the desired yield estimate after Step 3

(3) $\nabla_i z$ is a difference operator applied to the yield estimates which controls their rate of change across the grid points of R

(4) α and α_1 are predetermined weighting factors.

Thus Step 3 determines a yield surface which resembles that of Step 2 but minimizes rapid changes in yield across the grid points of R. The Method 8 estimator is the outcome of Step 3 with the addition of another term in the summation in (2.7.3), namely another low pass filter term of higher order:

$$I' = \Sigma_R \left[\alpha(z-z')^2 + \alpha_1 (\nabla z)^2 + \alpha_2 (\nabla^2 z)^2\right] \quad (2.7.4)$$

For a more in depth and concise description of Sasaki's approach and difference operators, see Haltiner [2, Chapter 14.5].

The Step 3 algorithm proposed by Wagner [10], minimizes I' by determining the Euler-Lagrange equations corresponding to (2.7.4) and then solving these equations by a numerical method known as relaxation. This method is an iterative solution which continues until the yield at each grid point of R changes by an amount less than some predetermined amount. It is similar to the procedures of Steps 1 and 2 in that a weighted average (using α, α_1, and α_2) is computed for each grid point of R and compared to the existing yield value. The yield estimate at the grid point is then adjusted towards (not see equal to) this new estimate and the procedure

continues to the next grid point.

The choice of weighting factors in (2.7.4) is of course very subjective. In Wagner's thesis, he often uses weights of $\alpha = 100$, $\alpha_1 = \alpha_2 = 1$, i.e. placing a very high weight on fitting the Step 2 surface (referred to as the "observed" values by Wagner).

In summary, the resulting surface using Method 8 will be a "smooth" surface which will be well behaved throughout R but will not exactly fit the original set of N data points. For a set of data of low variance and using a large α value (e.g. $\alpha = 100$) we would expect the final yield surface to strongly resemble the "smoothed" surface generated by Step 2. For this reason, there appears to be a serious lack of attention paid to this smoothing procedure and, in fact, Wagner essentially describes his low-pass filtering approach beginning with a complete grid of smoothed data.

2.8 Method 9 - Modified Linkage

This approach, proposed by Shepard [9], is an extension of the average linkage estimators (Methods 5,6). Shepard attempts to improve the straight weighted average estimator by (1) selecting only nearby data points to be used in estimating the yield at some grid point in R, (2) including the effect of the directions between this grid point and the data points used in the estimation, and (3) correcting for the zero gradients at the N data points on the yield surface generated by Methods 5 and 6.

To correct for the zero gradient at data point P_i, small increments were added to nearby data points so that the generated yield surface would have "desired" partial deviatives at P_i. Since

we are assuming a sparse data set (e.g. 1% grid occupancy) the authors did not feel that this modification would improve the accuracy of the linkage estimators. Later experiments showed this to be true. Consequently, only the nearby point rule and the direction correction were used in the Method 9 estimator.

Step 1: Selecting Nearby Points

Denote the Method 6 estimator by $f_1(P) = \frac{\sum_{1}^{N} w_i z_i}{\sum w_i}$

where $w_i = 1/d_i^2$, $i = 1,2,\ldots,N$.

Since this estimator practically removes the effect of distant data points and we are using a sparse data set, using only nearby data values may not provide a significant improvement in estimation. The resulting suface should however contain less noise.

The first step is to define a circle of radius R, which when constructed arbritrarily about R, will contain (on an average) a predetermined number, n_1, of data points. Thus

$$\pi r^2 = n_1 \cdot (A/N)$$

where A = approximate total area of R. The choice of n_1 will primarily be a function of the sparcity of the data. For each grid point P, a collection C_P' of data points near P and a new radius r' will be defined. Let $C_P = \{P_i \mid d_i \leq r\}$ where P_i = i-th data point and d_i = distance from P to P_i. Also let $n(C_P)$ = the number of points in C_P. Next, order the N data points by increasing distance from P, i.e. $d_{i_1} \leq d_{i_2} \leq \ldots \leq d_{i_N}$, and let

$$C_P^k = \{P_{i_1}, P_{i_2}, \ldots, P_{i_k}\} \text{ for } k \leq N.$$

Correspondingly, define

$$r'(C_P^k) = d_{i_{k+1}}$$

Suppose that each yield estimate should use at least n_{min} data points but no more than n_{max}. Then define

$$C_P' = \begin{cases} C_P^{n_{min}} & \text{if } 0 \leq n(C_P) \leq n_{min} \\ C_P & \text{if } n_{min} < n(C_P) \leq n_{max} \\ C_P^{n_{max}} & \text{if } n_{max} < n(C_P) \end{cases}$$

and $$r_P' = \begin{cases} r'(C_P^{n_{min}}) & \text{if } 0 \leq n(C_P) \leq n_{min} \\ r & \text{if } n_{min} < n(C_P) \leq n_{max} \\ r'(C_P^{n_{max}}) & \text{if } n_{max} < n(C_P). \end{cases}$$

Finally, a new weighting function $s(d_i)$ is defined which is continuously differentiable for $d_i > 0$ and such that $s(d_i) = 0$ for $d_i > r'$. This is given by

$$s_i = s(d_i) = \begin{cases} 1/d_i & \text{if } 0 < d_i \leq \frac{r'}{3} \\ \frac{27}{4r'}\left(\frac{d_i}{r'} - 1\right)^2 & \text{if } \frac{r'}{3} < d_i < r' \\ 0 & \text{if } r' < d_i. \end{cases}$$

The new yield approximation equation is given by

$$f_2(P) = \left[\sum_{P_i \varepsilon C_P'} s_i^2 z_i\right] \Big/ \sum_{P_i \varepsilon C_P'} s_i^2 \qquad (2.8.1)$$

where $f_2(P) = z_k$ if $d_k = 0$ for any $P_k \varepsilon C_P$.

Step 2: Correcting for Direction

Consider Figure 9 (a), (b)

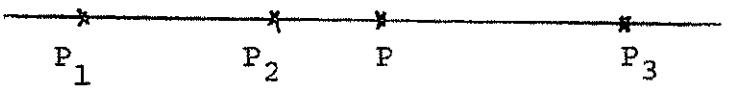


Figure 9(a)

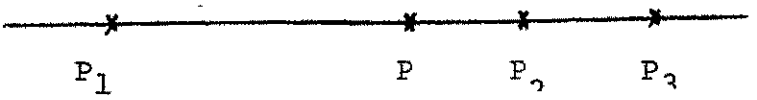


Figure 9(b)

Intuitively, it would seem the yield value at P in Figure 9(a) should be closer to the yield at P_3 than in Figure 9(b) due to the intervening effect of P_2 in (b). A weighted average based strictly on distances would provide identical yield estimates at P for both configurations. Thus a new directional weighting term is derived which represents this shadowing effect and is denoted by

$$t_i = \left[\sum_{P_j \varepsilon C_P'} s_j \left|1 - \cos\left(P_i P P_j\right)\right|\right] \Big/ \sum_{P_j \varepsilon C_P'} s_j .$$

The distance weighting factor s_j is included in the definition of t_i since data points near P should have a larger shadowing effect than distant data points. Note that $0 \leq t_i < 2$ for $i = 1,2,\ldots,N$.

The final weighted average incorporates this directional correlation and is given by (2.8.2).

$$\hat{z}_{P,9} = f_3(P) = \left[\sum_{P_i \varepsilon C_P'} w_i' z_i\right] \Big/ \sum_{P_i \varepsilon C_P'} w_i' \qquad (2.8.2)$$

where (1) $w_i' = s_i^2 \times (1 + t_i)$

and (2) $f_3(P) = z_k$ if $d_k = 0$ for some $P_k \in C_P'$.

The resulting yield surface using Method 9 will be continuously differentiable and will pass through the N data values. For extremely sparse data situations, we might expect little improvement over the average linkage estimators.

2.9 Method 10 - Modified Least Squares

This method, proposed by McLain [3], is an attempt to improve the linear regression estimator (Method 3) by (1) performing a weighted linear regression where the weights are functions of the distances d_i (2) deriving a different set of regression coefficients for each grid point P and (3) introducing x^2, y^2, xy terms into the regression model. For extremely sparse data sets one might expect to gain very little by adding the quadratic directional effects into the model (as later experiments demonstrated). However, the procedure of requiring those data points close to P to carry more weight than distant points seemed very desirable. Consequently, for each point P, this method determines a polynomial of the form

$$f(x_P, y_P) = c_{00} + c_{10}x_P + c_{01}y_P + c_{20}x_P^2 + c_{11}x_Py_P + c_{02}y_P^2.$$

The coefficients are chosen to minimize

$$Q = \sum_{i=1}^{N} [f(x_i, y_i) - z_i]^2 \cdot w(d_i) \qquad (2.9.1)$$

for some weighting function $w(d_i)$, e.g. $w(d_i) = 1/d_i$. The solution to (2.9.1) can be readily obtained by solving the six linear equations

$$\frac{\partial Q}{\partial C_{rs}} = 0$$

for the six coefficients c_{00}, c_{01}, c_{10}, c_{11}, c_{20}, c_{02}.

Having determined these coefficients, then the Method 10 estimator is given by

$$\hat{z}_{P,10} = c_{00} + c_{10} x_P + c_{01} y_P + c_{20} x_P^2 + c_{11} x_P y_P + c_{02} y_P^2$$

The weighting function $w(d_i)$ is again subjective. McLain discusses such functions ranging from $w(d_i) = 1/d_i$ (slow die-out) to $w(d_i) = \exp(-\alpha d_i^2) / (d_i^2 + \varepsilon)$ for suitable constants α and ε (fast die-out). McLain recommends the latter weighting function although it should be pointed out that he is assuming a grid containing between a hundred and a thousand data points. The authors in working with sparse data sets (see Section 3.0) had more success in using less drastic weighting functions such as $w(d_i) = 1/d_i$ or $1/d_i^2$.

This method will produce a surface which should better reflect yield differences across the entire region R and will pass through the N sample data values. By using a different regression equation for each grid point, the problem of ill behavior near the boundary of R (encountered in Method 3) is avoided. Overall, we would expect a better mean square error using Method 10 over Method 3.

3.0 An Empirical Study

3.1 Data Description

The authors used five years of yield data (bushels/acre) for wheat in North Dakota to test the ten extrapolation procedures*. The intent of the test was to see how well these methods extrapolated across the entire state from data acquired at seven yield stations located at Jamestown, Minot, Grafton, Bismark, Dickinson, Fargo, and

*obtained through NASA contract NAS9-13512

Williston. The data existed for the years 1962 - 1966 at these seven stations along with yield data at approximately forty five additional check points (cities) (Figure 3.1). The methods could then be checked for accuracy by applying each to the seven yield observations and comparing to the actual yield data at the forty five check points. The yield data for these seven stations is contained in Table 3.1.

	1962	1963	1964	1965	1966
Bismark	31.3	15.9	20.5	22.4	19.1
Dickinson	26.1	22.1	20.4	22.2	23.1
Fargo	24.1	24.0	24.2	28.8	24.7
Grafton	28.8	29.3	29.8	30.7	26.0
Jamestown	27.6	17.1	23.7	22.5	22.7
Minot	34.8	26.7	27.4	28.2	29.0
Williston	28.5	26.3	23.1	22.0	20.3

Table 3.1 Yield Data (bu/acre) for North Dakota Yield Stations

3.2 Method 9 Results

In the description of this method, three parameters were defined, namely:

n_1 = the average number of data points contained in a circle of radius r

n_{min} = minimum number of data points to be used in the weighted sum

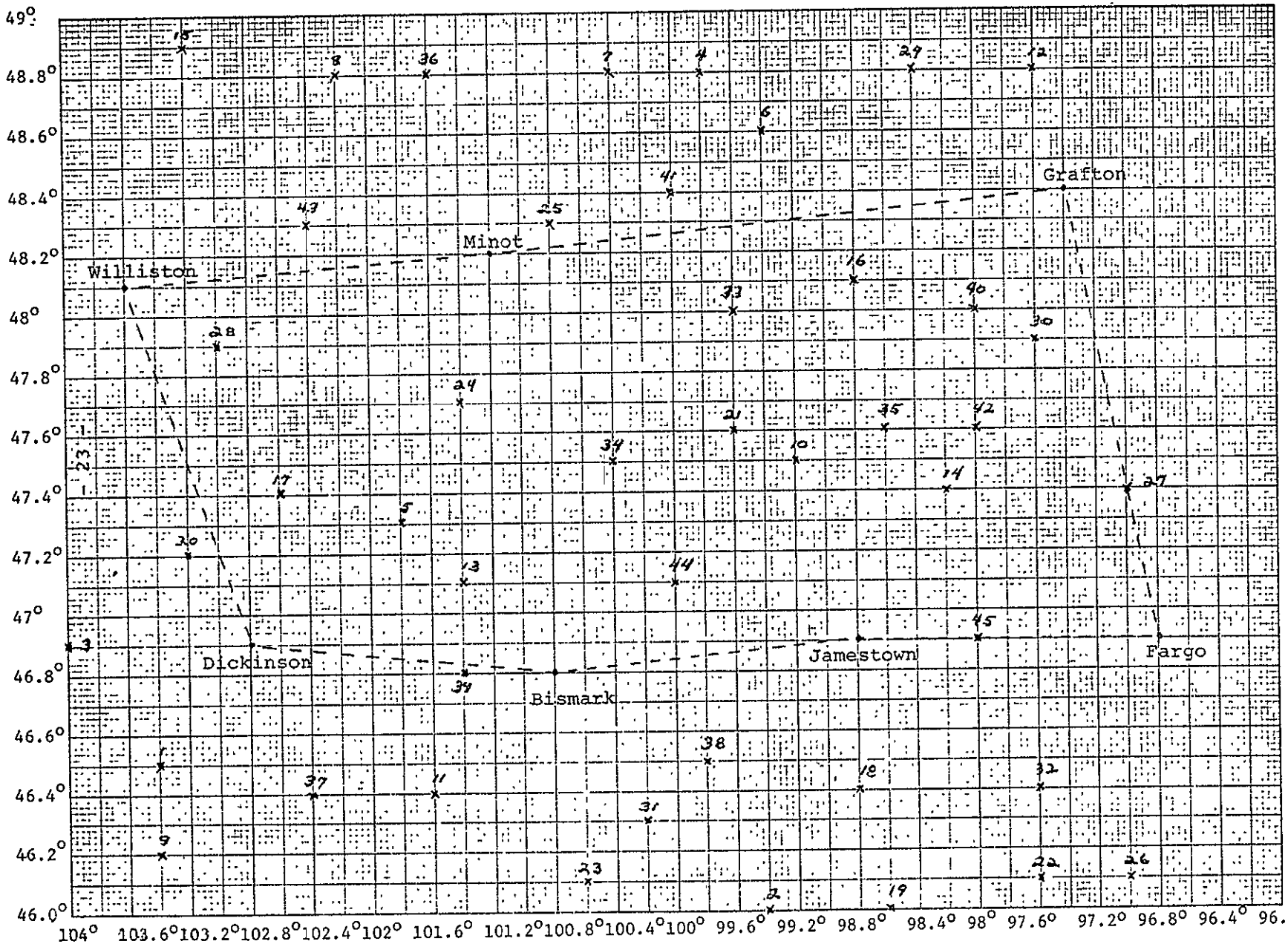


Figure 3.1 North Dakota Stations

n_{max} = maximum number of data points to be used in the weighted sum.

n_{max} was set equal to 7 with n_{min} and n_1 ranging from 2 to 7. The results of these computer runs are contained in Table 3.2. In four of the five years we obtained a good improvement over the average linkage estimates by using Shepard's modifications. The adjustment to the weights for direction (shadowing) provided a slight improvement in three of the five years and was actually the best of the three in two of the years. The value of n_1 made no difference since any value from 2 to 7 provided the same MSE. The ideal value for n_{min} unfortunately varied considerably and there appears to be no one "optimum" value to use in subsequent applications.

Year	Method 6		Using Nearest Points		Correcting for Direction	
1962	MSE = 11.20	*	n_1 = 2-7 n_{min} = 6,7 MSE = 12.10		n_1 = 2-7 n_{min} = 6,7 MSE = 11.85	
1963	MSE = 7.92		n_1 = 2-7 n_{min} = 6,7 MSE = 7.55		n_1 = 2-7 n_{min} = 6,7 MSE = 7.52	*
1964	MSE = 12.06		n_1 = 2-7 n_{min} = 3 MSE = 10.23		n_1 = 2-7 n_{min} = 3 MSE = 10.16	*
1965	MSE = 7.61		n_1 = 2-7 n_{min} = 4 MSE = 5.42	*	n_1 = 2-7 n_{min} = 3 MSE = 5.46	
1966	MSE = 8.17		n_1 = 2-7 n_{min} = 6,7 MSE = 7.81	*	n_1 = 2-7 n_{min} = 5 MSE = 7.98	

*Best for year

Table 3.2 Best MSE's using Method 9

3.3 Method 10 Results

Four models were considered in evaluating the Method 10 estimator, namely

Model I: $\hat{z}_P = A$

Model II: $\hat{z}_P = A + B\, x_P + C\, y_P$

Model III: $\hat{z}_P = A + B\, x_P + C\, y_P + D\, x_P y_P$

Model IV: $\hat{z}_P = A + B\, x_P + C\, y_P + D\, x_P y_P + E\, x_P^2 + F\, y_P^2$.

Five weighting factors were considered in deriving the regression coefficients; 1.0, $(1/d_i)$, $(1/d_i)^2$, exp $(-.002\, d_i^2)$, and exp $(-\alpha d_i^2)$ where α = (average distance between data points)$^{-1}$ and d_i = distance from P to data point P_i. For the five sets of data, the Model III estimator was found to provide the minimum MSE using a weighting factor of $1/d_i$. This will henceforth be referred to as the Method 10 estimator.

3.4 Summary of Results

Appendix B contains the results of Methods 1-8 and lists the actual and approximated results for each of the available check points. Table 3.3 summarizes the results of this experiment and contains the MSE's for each method and for each year. The MSE's for Method 9 are the smallest of the three MSE's as discussed in Section 3.2.

Method	Description	1962	1963	1964	1965	1966
1	Composite Average	13.84	16.53	21.78	16.55	12.84
2	Nearest Neighbor	14.96	14.33	10.69	7.53	9.59
3	Least Squares Linear Surface	9.14**	14.73	9.00**	5.95	7.00**
4	LS/NN	11.46	9.57	10.17***	6.92	10.49
5	Average Linkage (1/d)	12.00	10.36	15.56	10.98	9.88
6	Average Linkage $(1/d)^2$	11.20***	7.92***	12.06	7.61	8.17
7	Average Linkage with Coordinate Correlation	11.20***	9.74	12.06	7.61	8.17
8	Meteorological Objective Filtering	11.65	7.41*	10.21	5.88***	8.59
9	Modified Linkage	11.20***	7.52**	10.16	5.42*	7.81***
10	Weighted Least Squares A+Bx+Cy+Dxy with wt. = 1/d	7.77*	12.40	8.86*	5.63**	6.68*

Table 3.3 MSE's for Methods 1-10

* = smallest ** = 2nd smallest *** = 3rd smallest

3.5 Contour Maps

The authors wrote a Fortran contour mapping program to further analyze the results of the ten extrapolation procedures. Appendix C contains 12 such maps. Figures C-1 through C-5 applied Method 10 to the full set of 52 (or whatever number was available) yield values to obtain the most accurate representation of the North Dakota yield data for that year. These figures can be compared to Figures C-6 through C-10 which are the results of applying the best extrapolation

procedure (Methods 8,9 or 10) to the set of seven data points as discussed in Section 3.1. Figure C-7 is a contour map after 50 iterations of Step 3 using Method 8. Figure C-11 demonstrates Method 8 applied to the 1963 data immediately after Step 2 (the smoothing procedure). This figure illustrates an early point that due to the nature of the data, the basic contours are determined after the smoothing process. Finally, Figure C-12 is a contour map of Method 3 (one of the more consistently good estimators) applied to the 1963 data.

4.0 Conclusions

Based on the results of Section 3.4, the superior method for extrapolating North Dakota yield data appears to be Method 10 (Weighted Linear Regression). It provided the minimum MSE in three of the five years. Not far behind are Method 8 (Objective Analysis with Low-Pass Filtering Constraints), Method 3 (Least Squares Linear Surface), and Method 9 (Modified Average Linkage). The Method 3 estimator provided very consistently good MSE's.

The main point to be made here is that it does appear that the more sophisticated methods (8,9,10) appear to work very well on these sets of extremely sparse data. It should be pointed out however that there was not a lot of variation in the yearly yield data.

Table 4.1 contains the main features (good and bad) of each of the ten methods.

Method	Advantages	Disadvantages
1 Composite Average	Provides "reasonable" estimates over entire region. Yield surface is continuous and well behaved on boundaries.	Does not reflect variations within the region (particularly in small neighborhoods about each data point).
2 Nearest Neighbor	Provides good estimate in small neighborhoods about each data point and is well behaved on boundaries.	Surface defined by yield estimates is discontinuous. Maximum yield estimate is the maximum data value. Estimate completely ignores the influence of the other N-1 points. Rather poor at picking up directional trends.
3 Least Squares Plane	Provides good estimate in convex hull. Extremely good at picking up directional trends. Yield surface is continuous.	May provide unusually large or small (possibly negative) estimates on boundary of R if slope of plane is large.
4 LS/NN	Provides good estimate in convex hull and reflects directional trends in this region. Well behaved on boundaries.	Yield surface is discontinuous. Ignores influence of other N-1 data points outside convex hull.
5 Average Linkage (1/d)	More accurately models the influence of all N data points, yet in a small neighborhood about data point k, the yield estimate is completely dominated by z_k. Yield surface is continuous and well behaved.	Maximum yield estimate occurs at that data point having the largest yield. Does not reflect directional trends in the data. Yield surface has zero gradients at data points.

Method	Advantages	Disadvantages
6 Average Linkage $(1/d)^2$	Same as Method 5 - in addition the yield estimates appear to more accurately reflect the relative distances from each data point.	Same as Method 5
7 Linkage with Directional Correlation	Same as Method 6 - in addition reflects significant directional correlation	Could be ill behaved near the boundary if the regression estimators have large slope.
8 Objective Filtering	Provides a "smooth" well behaved surface over R. Considers the effect of all N data points.	Does not exactly fit data points. Severely affected by smoothing procedure.
9 Modified Linkage	Continuously differentiable. Passes through data points. Reflects relative distances and directions.	Same as Method 5
10 Weighted Regression	Passes through data points. Adjusts model for each point to be estimated.	Computationally time consuming.

Table 4.1

Appendix A: A Procedure for Determining the Convex Hull of Two-Dimensional Data

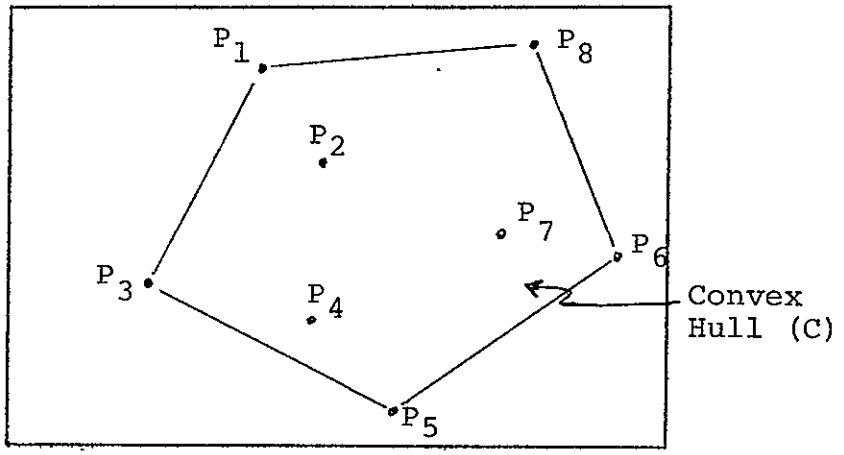


Figure A.1

The convex hull (C) of a set of two-dimensional data is the smallest (in area) convex region containing all of the data points. See Figure A.1. Given a set of data points $P_1,\ldots,P_N$ with coordinates $(x_1,y_1),\ldots,(x_N,y_N)$, this procedure defines P_i to be

(a) an interior point of C if P_i lies within any triangle formed by three other data points (P_2,P_4,P_7 in Figure A.1)

or (b) a boundary point of C if P_i lies in no such triangle (P_1,P_3,P_5,P_6,P_8 in Figure A.1).

The procedure will be to determine the set of boundary points of the input data. Once these points have been determined, it is a relatively easy task to connect this set of points to form the convex hull.

Thus, it remains to determine if P_i lies in any triangle formed by three other data points.

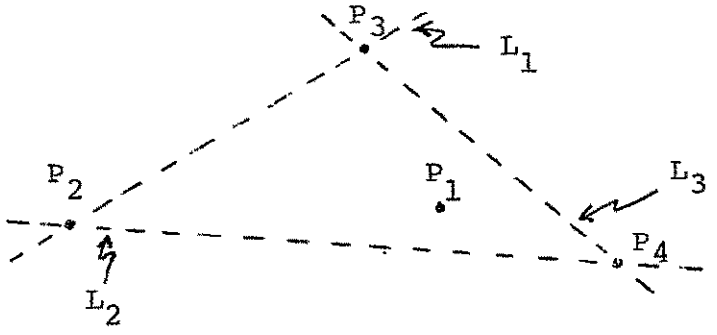


Figure A.2

Referring to Figure A.2, P_1 will lie within the triangle formed by P_2,P_3, and P_4 if

(1) P_1 and P_4 lie on the same side of L_1

(2) P_1 and P_3 lie on the same side of L_2

(3) P_1 and P_2 lie on the same side of L_3.

Care must be taken to insure that P_2,P_3, and P_4 are not colinear. If P_1 passes the above three tests for all sets of points P_i,P_j, and P_k $(i,j,k \neq 1)$, then P_1 will be classified as an interior point. The Fortran programs required to accomplish this are relatively straight forward and consist of

(1) a routine to determine a matrix containing the possible set of subscripts for triangles containing data point P_i. For example, for N = 5 and data point P_3 this matrix is

$$\begin{bmatrix} 1 & 2 & 4 \\ 1 & 2 & 5 \\ 1 & 4 & 5 \\ 2 & 4 & 5 \end{bmatrix}$$

(2) a routine to determine if P_i lies within the triangle formed by P_j,P_k, and P_l.

The output from these computer programs will be the set of boundary points which define the convex hull of the original data set.

When using the Method 4 yield estimation procedure, the above technique can be used to determine if a particular grid point P lies within the convex hull of the N sample points (in which case one would use the linear regression estimator). The point P will lie in the convex hull if it lies within (or on) any triangle formed by any three data points and this can be determined using the previously defined computer programs. For N = 5, the matrix of possible coefficients becomes

$$\begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 4 \\ 1 & 3 & 5 \\ 1 & 4 & 5 \\ 2 & 3 & 4 \\ 2 & 3 & 5 \\ 2 & 4 & 5 \\ 3 & 4 & 5 \end{bmatrix}$$

Appendix B

Methods 1-8 Results

NORTH DAKOTA SIMULATION - 1962

FIGURE B-1

NO.	NAME	ACTUAL	ESTIMATIONS METH1	DIFF	METH2	DIFF	METH3	DIFF	METH4	DIFF
1	AMID	20.20	28.74	-8.5	26.10	-5.9	27.80	-7.6	26.10	-5.9
2	ASHL	23.40	28.74	-5.3	27.60	-4.2	25.28	-1.9	27.60	-4.2
3	BEAC	27.30	28.74	-1.4	26.10	1.2	28.89	-1.6	26.10	1.2
4	BELC	30.10	28.74	1.4	34.80	-4.7	31.53	-1.4	34.80	-4.7
5	BEUL	32.10	28.74	3.4	26.10	6.0	28.97	3.1	28.97	3.1
6	BISB	29.00	28.74	0.3	34.80	-5.8	30.95	-1.9	34.80	-5.8
7	BOTT	32.80	28.74	4.1	34.80	-2.0	31.74	1.1	34.80	-2.0
8	BOWM	22.60	28.74	-6.1	26.10	-3.5	27.15	-4.5	26.10	-3.5
9	CARR	31.50	28.74	2.8	27.60	3.9	28.48	3.0	28.48	3.0
10	CARS	25.00	28.74	-3.7	31.30	-6.3	26.94	-1.9	31.30	-6.3
11	CAVA	25.60	28.74	-3.1	28.80	-3.2	30.74	-5.1	28.80	-3.2
12	CENT	26.10	28.74	-2.6	31.30	-5.2	28.39	-2.3	28.39	-2.3
13	COOP	32.60	28.74	3.9	27.60	5.0	27.90	4.7	27.90	4.7
14	CROS	28.20	28.74	-0.5	28.50	-0.3	32.96	-4.8	28.50	-0.3
15	DEVI	28.00	28.74	-0.7	28.80	-0.8	29.64	-1.6	29.64	-1.6
16	DUNN	29.00	28.74	0.3	26.10	2.9	29.48	-0.5	29.48	-0.5
17	EDGE	24.30	28.74	-4.4	27.60	-3.3	25.94	-1.6	27.60	-3.3
18	ELLE	24.80	28.74	-3.9	27.60	-2.8	24.99	-0.2	27.60	-2.8
19	FAIR	27.00	28.74	-1.7	26.10	0.9	29.26	-2.3	26.10	0.9
20	FESS	31.20	28.74	2.5	27.60	3.6	28.84	2.4	28.84	2.4
21	FORM	21.00	28.74	-7.7	24.10	-3.1	24.85	-3.9	24.10	-3.1
22	FORT	31.00	28.74	2.3	31.30	-0.3	25.93	5.1	31.30	-0.3
23	GARR	31.80	28.74	3.1	34.80	-3.0	29.70	2.1	29.70	2.1
24	GRAN	29.90	28.74	1.2	34.80	-4.9	30.79	-0.9	34.80	-4.9
25	HANK	17.60	28.74	-11.1	24.10	-6.5	24.64	-7.0	24.10	-6.5
26	HILL	29.00	28.74	0.3	24.10	4.9	27.47	1.5	27.47	1.5
27	KEEN	29.80	28.74	1.1	28.50	1.3	30.71	-0.9	30.71	-0.9
28	LANG	27.80	28.74	-0.9	28.80	-1.0	31.03	-3.2	28.80	-1.0
29	LARI	31.30	28.74	2.6	28.80	2.5	28.78	2.5	28.78	2.5
30	LINT	27.40	28.74	-1.3	31.30	-3.9	26.22	1.2	31.30	-3.9
31	LISB	22.00	28.74	-6.7	24.10	-2.1	25.51	-3.5	24.10	-2.1
32	MADD	30.40	28.74	1.7	34.80	-4.4	29.71	0.7	29.71	0.7
33	MCCL	28.30	28.74	-0.4	31.30	-3.0	28.91	-0.6	28.91	-0.6
34	MCHE	28.50	28.74	-0.2	27.60	0.9	28.48	0.0	28.48	0.0
35	MOHA	30.90	28.74	2.2	34.80	-3.9	32.17	-1.3	34.80	-3.9
36	MOTT	25.00	28.74	-3.7	26.10	-1.1	27.23	-2.2	26.10	-1.1
37	NAPO	25.40	28.74	-3.3	31.30	-5.9	26.51	-1.1	31.30	-5.9
38	NEW	31.00	28.74	2.3	31.30	-0.3	27.74	3.3	31.30	-0.3
39	PETE	30.10	28.74	1.4	28.80	1.3	29.14	1.0	29.14	1.0
40	RUGB	29.60	28.74	0.9	34.80	-5.2	30.73	-1.1	34.80	-5.2
41	SHAR	34.70	28.74	6.0	27.60	7.1	28.27	6.4	28.27	6.4
42	STAN	30.40	28.74	1.7	34.80	-4.4	31.37	-1.0	34.80	-4.4
43	TUTT	26.70	28.74	-2.0	31.30	-4.6	27.89	-1.2	27.89	-1.2
44	VALL	27.90	28.74	-0.8	27.60	0.3	26.74	1.2	26.74	1.2
AVERAGE ABS. DIFF.				2.85		3.35		2.42		2.78
MSE				13.84		14.96		9.14		11.46

NORTH DAKOTA SIMULATION - 1962

FIGURE B-2

NO.	NAME	ACTUAL	ESTIMATIONS METH5	DIFF	METH6	DIFF	METH7	DIFF	METH8	DIFF
1	AMID	20.20	28.25	-8.1	27.23	-7.0	27.23	-7.0	26.20	-6.0
2	ASHL	23.40	28.60	-5.2	28.57	-5.2	28.57	-5.2	31.20	-7.8
3	BEAC	27.30	28.40	-1.1	27.53	-0.2	27.53	-0.2	26.10	1.2
4	BELC	30.10	29.73	0.4	30.88	-0.8	30.88	-0.8	33.80	-3.7
5	BEUL	32.10	29.24	2.9	29.32	2.8	29.32	2.8	30.10	2.0
6	BISB	29.00	29.53	-0.5	30.33	-1.3	30.33	-1.3	32.00	-3.0
7	BOTT	32.80	30.12	2.7	31.82	1.0	31.82	1.0	34.50	-1.7
8	BOWM	22.60	28.49	-5.9	27.81	-5.2	27.81	-5.2	26.10	-3.5
9	CARR	31.50	28.85	2.6	28.77	2.7	28.77	2.7	29.60	1.9
10	CARS	25.00	29.05	-4.1	29.43	-4.4	29.43	-4.4	30.20	-5.2
11	CAVA	25.60	28.84	-3.2	28.83	-3.2	28.83	-3.2	28.90	-3.3
12	CENT	26.10	29.45	-3.3	30.04	-3.9	30.04	-3.9	30.40	-4.3
13	COOP	32.60	28.20	4.4	27.67	4.9	27.67	4.9	27.50	5.1
14	CROS	28.20	29.47	-1.3	29.81	-1.6	29.81	-1.6	29.60	-1.4
15	DEVI	28.00	29.04	-1.0	29.21	-1.2	29.21	-1.2	30.00	-2.0
16	DUNN	29.00	28.82	0.2	28.24	0.8	28.24	0.8	28.70	0.3
17	EDGE	24.30	28.28	-4.0	27.90	-3.6	27.90	-3.6	27.80	-3.5
18	ELLE	24.80	28.29	-3.5	27.91	-3.1	27.91	-3.1	27.60	-2.8
19	FAIR	27.00	28.24	-1.2	27.14	-0.1	27.14	-0.1	26.90	0.1
20	FESS	31.20	29.21	2.0	29.55	1.7	29.55	1.7	30.60	0.6
21	FORM	21.00	27.85	-6.8	26.92	-5.9	26.92	-5.9	25.10	-4.1
22	FORT	31.00	29.01	2.0	29.53	1.5	29.53	1.5	31.30	-0.3
23	GARR	31.80	30.20	1.6	31.77	0.0	31.77	0.0	32.10	-0.3
24	GRAN	29.90	31.81	-1.9	34.17	-4.3	34.17	-4.3	34.30	-4.4
25	HANK	17.60	27.64	-10.0	26.40	-8.8	26.40	-8.8	24.10	-6.5
26	HILL	29.00	27.44	1.6	26.03	3.0	26.03	3.0	26.00	3.0
27	KEEN	29.80	29.17	0.6	28.99	0.8	28.99	0.8	29.50	0.3
28	LANG	27.80	29.04	-1.2	29.13	-1.3	29.13	-1.3	29.20	-1.4
29	LARI	31.30	28.46	2.8	28.34	3.0	28.34	3.0	28.00	3.3
30	LINT	27.40	29.07	-1.7	29.65	-2.3	29.65	-2.3	30.90	-3.5
31	LISB	22.00	27.58	-5.6	26.44	-4.4	26.44	-4.4	25.40	-3.4
32	MADD	30.40	29.55	0.8	30.44	-0.0	30.44	-0.0	31.60	-1.2
33	MCCL	28.30	29.63	-1.3	30.46	-2.2	30.46	-2.2	31.60	-3.3
34	MCHE	28.50	28.60	-0.1	28.35	0.2	28.35	0.2	28.60	-0.1
35	MOHA	30.90	30.33	0.6	32.16	-1.3	32.16	-1.3	34.40	-3.5
36	MOTT	25.00	28.51	-3.5	27.84	-2.8	27.84	-2.8	27.90	-2.9
37	NAPO	25.40	28.93	-3.5	29.25	-3.8	29.25	-3.8	30.00	-4.6
38	NEW	31.00	29.54	1.5	30.39	0.6	30.39	0.6	30.20	0.8
39	PETE	30.10	28.65	1.4	28.60	1.5	28.60	1.5	28.60	1.5
40	RUGB	29.60	30.11	-0.5	31.80	-2.2	31.80	-2.2	33.10	-3.5
41	SHAR	34.70	28.30	6.4	27.85	6.8	27.85	6.8	27.70	7.0
42	STAN	30.40	30.02	0.4	30.99	-0.6	30.99	-0.6	31.80	-1.4
43	TUTT	26.70	29.23	-2.5	29.69	-3.0	29.69	-3.0	30.30	-3.6
44	VALL	27.90	27.71	0.2	27.05	0.9	27.05	0.9	26.40	1.5
AVERAGE ABS. DIFF.				2.64		2.64		2.64		2.84
MSE			12.00		11.20		11.20		11.65	

NORTH DAKOTA SIMULATION - 1963

FIGURE B-3

NO.	NAME	ACTUAL	ESTIMATIONS METH1	DIFF	METH2	DIFF	METH3	DIFF	METH4	DIFF
1	AMID	20.10	23.06	-3.0	22.10	-2.0	16.54	3.6	22.10	-2.0
2	ASHL	10.70	23.06	-12.4	17.10	-6.4	14.66	-4.0	17.10	-6.4
3	BEAC	20.20	23.06	-2.9	22.10	-1.9	18.74	1.5	22.10	-1.9
4	BELC	25.60	23.06	2.5	26.70	-1.1	31.10	-5.5	26.70	-1.1
5	BEUL	19.10	23.06	-4.0	22.10	-3.0	21.70	-2.6	21.70	-2.6
6	BISB	26.50	23.06	3.4	26.70	-0.2	30.03	-3.5	26.70	-0.2
7	BOTT	24.80	23.06	1.7	26.70	-1.9	30.94	-6.1	26.70	-1.9
8	BOWB	25.20	23.06	2.1	26.70	-1.5	30.45	-5.3	26.70	-1.5
9	BOWM	20.20	23.06	-2.9	22.10	-1.9	14.77	5.4	22.10	-1.9
10	CARR	23.70	23.06	0.6	17.10	6.6	23.58	0.1	23.58	0.1
11	CARS	16.50	23.06	-6.6	15.90	0.6	16.43	0.1	15.90	0.6
12	CAVA	22.00	23.06	-1.1	29.30	-7.3	31.69	-9.7	29.30	-7.3
13	CENT	16.00	23.06	-7.1	15.90	0.1	20.62	-4.6	20.62	-4.6
14	COOP	21.10	23.06	-2.0	17.10	4.0	23.26	-2.2	23.26	-2.2
15	CROS	25.40	23.06	2.3	26.30	-0.9	30.77	-5.4	26.30	-0.9
16	DEVI	27.20	23.06	4.1	29.30	-2.1	27.23	-0.0	27.23	-0.0
17	DUNN	21.00	23.06	-2.1	22.10	-1.1	22.07	-1.1	22.07	-1.1
18	EDGE	16.70	23.06	-6.4	17.10	-0.4	17.19	-0.5	17.10	-0.4
19	FAIR	20.50	23.06	-2.6	22.10	-1.6	20.73	-0.2	22.10	-1.6
20	FESS	22.50	23.06	-0.6	17.10	5.4	24.06	-1.6	24.06	-1.6
21	FORM	19.00	23.06	-4.1	24.00	-5.0	15.74	3.3	24.00	-5.0
22	FORT	15.90	23.06	-7.2	15.90	0.0	14.93	1.0	15.90	0.0
23	GARR	23.80	23.06	0.7	26.70	-2.9	24.17	-0.4	24.17	-0.4
24	GRAN	19.50	23.06	-3.6	26.70	-7.2	27.88	-8.4	26.70	-7.2
25	HANK	20.40	23.06	-2.7	24.00	-3.6	15.90	4.5	24.00	-3.6
26	HILL	26.40	23.06	3.3	24.00	2.4	23.58	2.8	23.58	2.8
27	KEEN	21.50	23.06	-1.6	26.30	-4.8	24.92	-3.4	24.92	-3.4
28	LANG	26.50	23.06	3.4	29.30	-2.8	31.48	-5.0	29.30	-2.8
29	LARI	26.10	23.06	3.0	29.30	-3.2	26.38	-0.3	26.38	-0.3
30	LISB	17.90	23.06	-5.2	24.00	-6.1	17.51	0.4	24.00	-6.1
31	MADD	24.00	23.06	0.9	26.70	-2.7	26.43	-2.4	26.43	-2.4
32	MCCL	20.10	23.06	-3.0	15.90	4.2	23.26	-3.2	23.26	-3.2
33	MCHE	18.70	23.06	-4.4	17.10	1.6	24.33	-5.6	24.33	-5.6
34	MOHA	26.20	23.06	3.1	26.70	-0.5	30.61	-4.4	26.70	-0.5
35	MOTT	20.00	23.06	-3.1	22.10	-2.1	16.22	3.8	22.10	-2.1
36	NEW	18.10	23.06	-5.0	15.90	2.2	18.85	-0.8	15.90	2.2
37	PETE	27.60	23.06	4.5	29.30	-1.7	26.86	0.7	26.86	0.7
38	RUGB	23.10	23.06	0.0	26.70	-3.6	28.68	-5.6	26.70	-3.6
39	SHAR	28.20	23.06	5.1	17.10	11.1	24.50	3.7	24.50	3.7
40	STAN	26.00	23.06	2.9	26.70	-0.7	27.44	-1.4	26.70	-0.7
41	VALL	20.10	23.06	-3.0	17.10	3.0	20.36	-0.3	20.36	-0.3
AVERAGE ABS. DIFF.				3.41		2.96		3.03		2.35
MSE				16.53		14.33		14.73		9.57

NORTH DAKOTA SIMULATION - 1963

FIGURE B-4

NO.	NAME	ACTUAL	ESTIMATIONS METH5	DIFF	METH6	DIFF	METH7	DIFF	METH8	DIFF
1	AMID	20.10	22.24	-2.1	21.94	-1.8	18.92	1.2	22.00	-1.9
2	ASHL	10.70	21.22	-10.5	19.55	-8.8	16.18	-5.5	16.00	-5.3
3	BEAC	20.20	22.66	-2.5	22.47	-2.3	20.67	-0.5	22.10	-1.9
4	BELC	25.60	24.01	1.6	25.05	0.5	29.00	-3.4	27.00	-1.4
5	BEUL	19.10	22.10	-3.0	21.39	-2.3	21.90	-2.8	21.50	-2.4
6	BISB	26.50	24.01	2.5	25.11	1.4	28.23	-1.7	27.70	-1.2
7	BOTT	24.80	24.06	0.7	25.19	-0.4	29.04	-4.2	26.80	-2.0
8	BOWB	25.20	24.10	1.1	25.20	-0.0	29.04	-3.8	26.60	-1.4
9	BOWM	20.20	22.18	-2.0	21.72	-1.5	17.67	2.5	22.10	-1.9
10	CARR	23.70	21.92	1.8	20.59	3.1	22.43	1.3	21.40	2.3
11	CARS	16.50	20.94	-4.4	19.13	-2.6	17.63	-1.1	17.20	-0.7
12	CAVA	22.00	25.56	-3.6	28.13	-6.1	29.98	-8.0	29.30	-7.3
13	CENT	16.00	21.00	-5.0	18.99	-3.0	20.34	-4.3	19.40	-3.4
14	COOP	21.10	22.13	-1.0	21.02	0.1	22.18	-1.1	22.20	-1.1
15	CROS	25.40	24.05	1.4	25.11	0.3	29.41	-4.0	26.40	-1.0
16	DEVI	27.20	23.74	3.5	24.76	2.4	26.14	1.1	26.90	0.3
17	DUNN	21.00	22.77	-1.8	22.68	-1.7	22.71	-1.7	23.30	-2.3
18	EDGE	16.70	20.83	-4.1	18.88	-2.2	17.54	-0.8	17.40	-0.7
19	FAIR	20.50	22.65	-2.2	22.41	-1.9	21.84	-1.3	22.70	-2.2
20	FESS	22.50	22.23	0.3	21.38	1.1	23.08	-0.6	22.10	0.4
21	FORM	19.00	21.99	-3.0	21.34	-2.3	17.15	1.8	22.10	-3.1
22	FORT	15.90	20.91	-5.0	18.88	-3.0	16.36	-0.5	15.90	0.0
23	GARR	23.80	23.20	0.6	23.95	-0.1	24.30	-0.5	23.90	-0.1
24	GRAN	19.50	24.79	-5.3	26.28	-6.8	27.42	-7.9	26.50	-7.0
25	HANK	20.40	22.33	-1.9	22.14	-1.7	17.41	3.0	24.00	-3.6
26	HILL	26.40	23.21	3.2	23.54	2.9	22.99	3.4	24.60	1.8
27	KEEN	21.50	24.16	-2.7	25.40	-3.9	25.56	-4.1	25.70	-4.2
28	LANG	26.50	24.72	1.8	26.83	-0.3	29.57	-3.1	29.10	-2.6
29	LARI	26.10	24.34	1.8	26.24	-0.1	25.83	0.3	27.00	-0.9
30	LISB	17.90	21.99	-4.1	21.51	-3.6	18.39	-0.5	21.40	-3.5
31	MADD	24.00	23.24	0.8	23.62	0.4	25.38	-1.4	25.50	-1.5
32	MCCL	20.10	22.06	-2.0	21.21	-1.1	22.63	-2.5	21.40	-1.3
33	MCHE	18.70	22.48	-3.8	21.75	-3.1	23.20	-4.5	23.20	-4.5
34	MOHA	26.20	24.17	2.0	25.39	0.8	29.10	-2.9	26.70	-0.5
35	MOTT	20.00	21.73	-1.7	21.12	-1.1	18.26	1.7	20.00	0.0
36	NEW	18.10	20.12	-2.0	17.63	0.5	18.72	-0.6	17.80	0.3
37	PETE	27.60	24.38	3.2	26.35	1.3	26.26	1.3	27.20	0.4
38	RUGB	23.10	23.90	-0.8	24.98	-1.9	27.40	-4.3	26.80	-3.7
39	SHAR	28.20	22.95	5.3	22.91	5.3	23.57	4.6	24.30	3.9
40	STAN	26.00	24.19	1.8	25.36	0.6	27.12	-1.1	26.30	-0.3
41	VALL	20.10	21.30	-1.2	19.76	0.3	19.80	0.3	20.10	0.0
AVERAGE ABS. DIFF.				2.65		2.07		2.47		2.06
MSE			10.36		7.92		9.74		7.41	

NORTH DAKOTA SIMULATION - 1964

FIGURE B-5

NO.	NAME	ACTUAL	ESTIMATIONS METH1	DIFF	METH2	DIFF	METH3	DIFF	METH4	DIFF
1	AMID	19.90	24.16	-4.3	20.40	-0.5	17.90	2.0	20.40	-0.5
2	BELC	26.90	24.16	2.7	27.40	-0.5	29.75	-2.8	27.40	-0.5
3	BEUL	20.20	24.16	-4.0	20.40	-0.2	22.29	-2.1	22.29	-2.1
4	BISB	28.40	24.16	4.2	27.40	1.0	29.28	-0.9	27.40	1.0
5	BOTT	28.70	24.16	4.5	27.40	1.3	29.28	-0.6	27.40	1.3
6	BOWB	24.70	24.16	0.5	27.40	-2.7	27.88	-3.2	27.40	-2.7
7	BOWM	17.40	24.16	-6.8	20.40	-3.0	16.72	0.7	20.40	-3.0
8	CARR	28.00	24.16	3.8	23.70	4.3	25.11	2.9	25.11	2.9
9	CARS	16.00	24.16	-8.2	20.50	-4.5	18.91	-2.9	20.50	-4.5
10	CAVA	25.80	24.16	1.6	29.80	-4.0	31.47	-5.7	29.80	-4.0
11	CENT	19.30	24.16	-4.9	20.50	-1.2	21.82	-2.5	21.82	-2.5
12	COOP	27.30	24.16	3.1	23.70	3.6	25.49	1.8	25.49	1.8
13	CROS	22.80	24.16	-1.4	23.10	-0.3	27.49	-4.7	23.10	-0.3
14	DUNN	17.60	24.16	-6.6	20.40	-2.8	22.06	-4.5	22.06	-4.5
15	EDGE	19.50	24.16	-4.7	23.70	-4.2	21.09	-1.6	23.70	-4.2
16	FAIR	16.20	24.16	-8.0	20.40	-4.2	20.80	-4.6	20.40	-4.2
17	FESS	28.70	24.16	4.5	23.70	5.0	25.19	3.5	25.19	3.5
18	FORM	19.10	24.16	-5.1	24.20	-5.1	20.85	-1.7	24.20	-5.1
19	GARR	25.70	24.16	1.5	27.40	-1.7	24.18	1.5	24.18	1.5
20	GRAN	23.10	24.16	-1.1	27.40	-4.3	27.00	-3.9	27.40	-4.3
21	HANK	18.80	24.16	-5.4	24.20	-5.4	21.32	-2.5	24.20	-5.4
22	HILL	27.80	24.16	3.6	24.20	3.6	26.43	1.4	26.43	1.4
23	KEEN	18.30	24.16	-5.9	23.10	-4.8	23.71	-5.4	23.71	-5.4
24	LANG	28.90	24.16	4.7	29.80	-0.9	30.84	-1.9	29.80	-0.9
25	LISB	19.40	24.16	-4.8	24.20	-4.8	22.03	-2.6	24.20	-4.8
26	MADD	28.10	24.16	3.9	27.40	0.7	26.76	1.3	26.76	1.3
27	MCCL	22.00	24.16	-2.2	20.50	1.5	24.17	-2.2	24.17	-2.2
28	MCHE	25.90	24.16	1.7	23.70	2.2	25.97	-0.1	25.97	-0.1
29	MOHA	29.30	24.16	5.1	27.40	1.9	28.35	1.0	27.40	1.9
30	MOTT	19.10	24.16	-5.1	20.40	-1.3	18.28	0.8	20.40	-1.3
31	NAPO	15.70	24.16	-8.5	20.50	-4.8	20.71	-5.0	20.50	-4.8
32	NEW	19.20	24.16	-5.0	20.50	-1.3	20.64	-1.4	20.50	-1.3
33	RUGB	24.20	24.16	0.0	27.40	-3.2	28.02	-3.8	27.40	-3.2
34	SHAR	29.80	24.16	5.6	23.70	6.1	26.44	3.4	26.44	3.4
35	STAN	25.70	24.16	1.5	27.40	-1.7	25.75	-0.1	27.40	-1.7
36	TUTT	16.70	24.16	-7.5	20.50	-3.8	22.91	-6.2	22.91	-6.2
37	VALL	23.70	24.16	-0.5	23.70	0.0	23.68	0.0	23.68	0.0
AVERAGE ABS. DIFF.				4.12		2.77		2.52		2.69
MSE				21.78		10.69		9.00		10.17

C-3

NORTH DAKOTA SIMULATION - 1964

FIGURE P-6

NO.	NAME	ACTUAL	ESTIMATIONS METH5	DIFF	METH6	DIFF	METH7	DIFF	METH8	DIFF
1	AMID	19.90	22.54	-2.6	21.21	-1.3	21.21	-1.3	20.40	-0.5
2	BELC	26.90	25.09	1.8	26.06	0.8	26.06	0.8	27.70	-0.8
3	BEUL	20.20	23.13	-2.9	22.31	-2.1	22.31	-2.1	22.40	-2.2
4	BISB	28.40	25.20	3.2	26.27	2.1	26.27	2.1	28.40	0.0
5	BOTT	28.70	25.02	3.7	26.01	2.7	26.01	2.7	27.50	1.2
6	BOWB	24.70	24.51	0.2	25.08	-0.4	25.08	-0.4	26.80	-2.1
7	BOWM	17.40	22.82	-5.4	21.64	-4.2	21.64	-4.2	20.40	-3.0
8	CARR	28.00	24.25	3.8	24.17	3.8	24.17	3.8	25.30	2.7
9	CARS	16.00	22.65	-6.6	21.45	-5.5	21.45	-5.5	20.50	-4.5
10	CAVA	25.80	26.61	-0.8	28.84	-3.0	28.84	-3.0	29.80	-4.0
11	CENT	19.30	22.78	-3.5	21.62	-2.3	21.62	-2.3	21.60	-2.3
12	COOP	27.30	24.48	2.8	24.49	2.8	24.49	2.8	25.50	1.8
13	CROS	22.80	24.10	-1.3	24.04	-1.2	24.04	-1.2	23.80	-1.0
14	DUNN	17.60	23.01	-5.4	22.07	-4.5	22.07	-4.5	22.20	-4.6
15	EDGE	19.50	23.75	-4.2	23.54	-4.0	23.54	-4.0	23.30	-3.8
16	FAIR	16.20	22.52	-6.3	21.24	-5.0	21.24	-5.0	21.00	-4.8
17	FESS	28.70	24.25	4.4	24.25	4.4	24.25	4.4	25.40	3.3
18	FORM	19.10	24.02	-4.9	23.93	-4.8	23.93	-4.8	24.10	-5.0
19	GARR	25.70	24.23	1.5	24.88	0.8	24.88	0.8	24.90	0.8
20	GRAN	23.10	25.68	-2.6	27.01	-3.9	27.01	-3.9	27.30	-4.2
21	HANK	18.80	24.13	-5.3	24.09	-5.3	24.09	-5.3	24.20	-5.4
22	HILL	27.80	24.74	3.1	24.84	3.0	24.84	3.0	25.70	2.1
23	KEEN	18.30	23.58	-5.3	23.30	-5.0	23.30	-5.0	23.60	-5.3
24	LANG	28.90	25.87	3.0	27.77	1.1	27.77	1.1	29.60	-0.7
25	LISB	19.40	24.08	-4.7	24.02	-4.6	24.02	-4.6	24.00	-4.6
26	MADD	28.10	24.77	3.3	25.38	2.7	25.38	2.7	27.20	0.9
27	MCCL	22.00	23.89	-1.9	23.66	-1.7	23.66	-1.7	24.30	-2.3
28	MCHE	25.90	24.61	1.3	24.82	1.1	24.82	1.1	26.30	-0.4
29	MOHA	29.30	24.82	4.5	25.77	3.5	25.77	3.5	27.20	2.1
30	MOTT	19.10	22.56	-3.5	21.30	-2.2	21.30	-2.2	20.40	-1.3
31	NAPO	15.70	23.23	-7.5	22.48	-6.8	22.48	-6.8	21.60	-5.9
32	NEW	19.20	22.33	-3.1	21.05	-1.9	21.05	-1.9	20.60	-1.4
33	RUGB	24.20	25.06	-0.9	26.06	-1.9	26.06	-1.9	27.70	-3.5
34	SHAR	29.80	24.86	4.9	25.34	4.5	25.34	4.5	26.60	3.2
35	STAN	25.70	24.32	1.4	24.74	1.0	24.74	1.0	25.30	0.4
36	TUTT	16.70	23.34	-6.6	22.50	-5.8	22.50	-5.8	22.70	-6.0
37	VALL	23.70	24.11	-0.4	23.95	-0.2	23.95	-0.2	24.00	-0.3
AVERAGE ABS. DIFF.				3.48		3.03		3.03		2.66
MSE				15.56		12.06		12.06		10.21

NORTH DAKOTA SIMULATION - 1965

FIGURE B-7

NO.	NAME	ACTUAL	ESTIMATIONS METH1	DIFF	METH2	DIFF	METH3	DIFF	METH4	DIFF
1	AMID	21.90	25.26	-3.4	22.20	-0.3	18.93	3.0	22.20	-0.3
2	ASHL	18.60	25.26	-6.7	22.50	-3.9	21.79	-3.2	22.50	-3.9
3	BEAC	22.30	25.26	-3.0	22.20	0.1	19.47	2.8	22.20	0.1
4	BELC	26.70	25.26	1.4	28.20	-1.5	29.69	-3.0	28.20	-1.5
5	BEUL	24.50	25.26	-0.8	22.20	2.3	23.05	1.4	23.05	1.4
6	BISB	28.90	25.26	3.6	28.20	0.7	29.53	-0.6	28.20	0.7
7	BOTT	29.20	25.26	3.9	28.20	1.0	29.04	0.2	28.20	1.0
8	BOWB	30.00	25.26	4.7	28.20	1.8	27.09	2.9	28.20	1.8
9	BOWM	18.80	25.26	-6.5	22.20	-3.4	18.04	0.8	22.20	-3.4
10	CARR	23.30	25.26	-2.0	22.50	0.8	26.47	-3.2	26.47	-3.2
11	CARS	20.50	25.26	-4.8	22.40	-1.9	20.59	-0.1	22.40	-1.9
12	CAVA	29.90	25.26	4.6	30.70	-0.8	32.08	-2.2	30.70	-0.8
13	CENT	21.50	25.26	-3.8	22.40	-0.9	22.89	-1.4	22.89	-1.4
14	COOP	28.20	25.26	2.9	22.50	5.7	27.26	0.9	27.26	0.9
15	CROS	27.20	25.26	1.9	22.00	5.2	26.30	0.9	22.00	5.2
16	DEVI	31.30	25.26	6.0	30.70	0.6	28.69	2.6	28.69	2.6
17	DUNN	22.50	25.26	-2.8	22.20	0.3	22.48	0.0	22.48	0.0
18	EDGE	24.60	25.26	-0.7	22.50	2.1	23.63	1.0	22.50	2.1
19	ELLE	20.90	25.26	-4.4	22.50	-1.6	22.65	-1.8	22.50	-1.6
20	FAIR	22.00	25.26	-3.3	22.20	-0.2	21.23	0.8	22.20	-0.2
21	FESS	24.40	25.26	-0.9	22.50	1.9	26.33	-1.9	26.33	-1.9
22	FORM	25.80	25.26	0.5	28.80	-3.0	24.04	1.8	28.80	-3.0
23	FORT	20.20	25.26	-5.1	22.40	-2.2	20.78	-0.6	22.40	-2.2
24	GARR	28.70	25.26	3.4	28.20	0.5	24.68	4.0	24.68	4.0
25	GRAN	24.30	25.26	-1.0	28.20	-3.9	27.12	-2.8	28.20	-3.9
26	HANK	25.70	25.26	0.4	28.80	-3.1	24.69	1.0	28.80	-3.1
27	HILL	32.20	25.26	6.9	28.80	3.4	28.56	3.6	28.56	3.6
28	KEEN	23.10	25.26	-2.2	22.00	1.1	23.54	-0.4	23.54	-0.4
29	LANG	33.30	25.26	8.0	30.70	2.6	31.21	2.1	30.70	2.6
30	LARI	30.90	25.26	5.6	30.70	0.2	29.40	1.5	29.40	1.5
31	LINT	22.20	25.26	-3.1	22.40	-0.2	21.81	0.4	22.40	-0.2
32	LISB	24.60	25.26	-0.7	28.80	-4.2	24.93	-0.3	28.80	-4.2
33	MADD	25.70	25.26	0.4	28.20	-2.5	27.53	-1.8	27.53	-1.8
34	MCCL	22.50	25.26	-2.8	22.40	0.1	25.17	-2.7	25.17	-2.7
35	MCHE	22.70	25.26	-2.6	22.50	0.2	27.42	-4.7	27.42	-4.7
36	MOHA	30.10	25.26	4.8	28.20	1.9	27.74	2.4	28.20	1.9
37	MOTT	23.00	25.26	-2.3	22.20	0.8	19.72	3.3	22.20	0.8
38	NAPO	19.60	25.26	-5.7	22.40	-2.8	22.84	-3.2	22.40	-2.8
39	NEW	22.10	25.26	-3.2	22.40	-0.3	22.00	0.1	22.40	-0.3
40	PETE	34.40	25.26	9.1	30.70	3.7	29.26	5.1	29.26	5.1
41	RUGB	23.80	25.26	-1.5	28.20	-4.4	28.28	-4.5	28.20	-4.4
42	SHAR	31.20	25.26	5.9	22.50	8.7	28.07	3.1	28.07	3.1
43	STAN	27.40	25.26	2.1	28.20	-0.8	25.38	2.0	28.20	-0.8
44	TUTT	20.20	25.26	-5.1	22.40	-2.2	24.41	-4.2	24.41	-4.2
45	VALL	26.30	25.26	1.0	22.50	3.8	25.99	0.3	25.99	0.3
AVERAGE ABS. DIFF.				3.45		2.08		2.02		2.17
MSE				16.55		7.53		5.95		6.92

NORTH DAKOTA SIMULATION - 1965

FIGURE B-8

NO.	NAME	ACTUAL	ESTIMATIONS METH5	DIFF	METH6	DIFF	METH7	DIFF	METH8	DIFF
1	AMID	21.90	23.74	-1.8	22.71	-0.8	22.71	-0.8	22.20	-0.3
2	ASHL	18.60	24.63	-6.0	23.95	-5.4	23.95	-5.4	22.40	-3.8
3	BEAC	22.30	23.82	-1.5	22.83	-0.5	22.83	-0.5	22.20	0.1
4	BELC	26.70	26.04	0.7	26.91	-0.2	26.91	-0.2	28.50	-1.8
5	BEUL	24.50	24.21	0.3	23.56	0.9	23.56	0.9	23.70	0.8
6	BISB	28.90	26.19	2.7	27.17	1.7	27.17	1.7	29.20	-0.3
7	BOTT	29.20	25.93	3.3	26.80	2.4	26.80	2.4	28.30	0.9
8	BOWB	30.00	25.20	4.8	25.50	4.5	25.50	4.5	27.30	2.7
9	BOWM	18.80	23.98	-5.2	23.01	-4.2	23.01	-4.2	22.20	-3.4
10	CARR	23.30	25.13	-1.8	24.63	-1.3	24.63	-1.3	25.40	-2.1
11	CARS	20.50	23.90	-3.4	22.97	-2.5	22.97	-2.5	22.30	-1.8
12	CAVA	29.90	27.67	2.2	29.81	0.1	29.81	0.1	30.70	-0.8
13	CENT	21.50	23.99	-2.5	23.12	-1.6	23.12	-1.6	23.20	-1.7
14	COOP	28.20	25.55	2.7	25.26	2.9	25.26	2.9	26.30	1.9
15	CROS	27.20	24.64	2.6	24.03	3.2	24.03	3.2	23.00	4.2
16	DEVI	31.30	26.27	5.0	27.32	4.0	27.32	4.0	28.90	2.4
17	DUNN	22.50	23.98	-1.5	23.16	-0.7	23.16	-0.7	23.10	-0.6
18	EDGE	24.60	24.60	-0.0	23.72	0.9	23.72	0.9	22.90	1.7
19	ELLE	20.90	24.93	-4.0	24.49	-3.6	24.49	-3.6	22.60	-1.7
20	FAIR	22.00	23.61	-1.6	22.63	-0.6	22.63	-0.6	22.40	-0.4
21	FESS	24.40	25.19	-0.8	24.95	-0.6	24.95	-0.6	25.70	-1.3
22	FORM	25.80	25.49	0.3	25.73	0.1	25.73	0.1	27.10	-1.3
23	FORT	20.20	24.23	-4.0	23.34	-3.1	23.34	-3.1	22.40	-2.2
24	GARR	28.70	25.15	3.6	25.76	2.9	25.76	2.9	25.80	2.9
25	GRAN	24.30	26.53	-2.2	27.81	-3.5	27.81	-3.5	28.10	-3.8
26	HANK	25.70	25.81	-0.1	26.49	-0.8	26.49	-0.8	28.80	-3.1
27	HILL	32.20	26.62	5.6	27.67	4.5	27.67	4.5	28.60	3.6
28	KEEN	23.10	23.88	-0.8	22.85	0.2	22.85	0.2	23.20	-0.1
29	LANG	33.30	26.91	6.4	28.74	4.6	28.74	4.6	30.50	2.8
30	LARI	30.90	26.99	3.9	28.67	2.2	28.67	2.2	29.40	1.5
31	LINT	22.20	24.19	-2.0	23.26	-1.1	23.26	-1.1	22.40	-0.2
32	LISB	24.60	25.67	-1.1	26.08	-1.5	26.08	-1.5	26.50	-1.9
33	MADD	25.70	25.74	-0.0	26.23	-0.5	26.23	-0.5	27.80	-2.1
34	MCCL	22.50	24.89	-2.4	24.64	-2.1	24.64	-2.1	25.10	-2.6
35	MCHE	22.70	25.61	-2.9	25.53	-2.8	25.53	-2.8	26.70	-4.0
36	MOHA	30.10	25.60	4.5	26.41	3.7	26.41	3.7	27.90	2.2
37	MOTT	23.00	23.80	-0.8	22.82	0.2	22.82	0.2	22.20	0.8
38	NAPO	19.60	24.22	-4.6	23.29	-3.7	23.29	-3.7	22.40	-2.8
39	NEW	22.10	23.67	-1.6	22.73	-0.6	22.73	-0.6	22.50	-0.4
40	PETE	34.40	26.93	7.5	28.65	5.7	28.65	5.7	29.40	5.0
41	RUGB	23.80	26.00	-2.2	26.89	-3.1	26.89	-3.1	28.50	-4.7
42	SHAR	31.20	26.06	5.1	26.51	4.7	26.51	4.7	27.70	3.5
43	STAN	27.40	24.89	2.5	24.95	2.5	24.95	2.5	25.30	2.1
44	TUTT	20.20	24.32	-4.1	23.41	-3.2	23.41	-3.2	23.20	-3.0
45	VALL	26.30	25.20	1.1	24.62	1.7	24.62	1.7	25.20	1.1
AVERAGE ABS. DIFF.				2.75		2.26		2.26		2.05
MSE				10.98		7.61		7.61		5.88

NORTH DAKOTA SIMULATION - 1966

FIGURE B-9

NO.	NAME	ACTUAL	ESTIMATIONS METH1	DIFF	METH2	DIFF	METH3	DIFF	METH4	DIFF
1	AMID	25.60	23.56	2.0	23.10	2.5	19.52	6.1	23.10	2.5
2	ASHL	16.40	23.56	-7.2	22.70	-6.3	20.42	-4.0	22.70	-6.3
3	BEAC	22.80	23.56	-0.8	23.10	-0.3	20.18	2.6	23.10	-0.3
4	BELC	25.00	23.56	1.4	29.00	-4.0	27.05	-2.1	29.00	-4.0
5	BEUL	22.50	23.56	-1.1	23.10	-0.6	22.32	0.2	22.32	0.2
6	BISB	26.90	23.56	3.3	29.00	-2.1	26.78	0.1	29.00	-2.1
7	BOTT	28.00	23.56	4.4	29.00	-1.0	26.74	1.3	29.00	-1.0
8	BOWB	25.70	23.56	2.1	29.00	-3.3	25.78	-0.1	29.00	-3.3
9	BOWM	16.40	23.56	-7.2	23.10	-6.7	18.78	-2.4	23.10	-6.7
10	CARR	23.60	23.56	0.0	22.70	0.9	24.19	-0.6	24.19	-0.6
11	CARS	21.50	23.56	-2.1	19.10	2.4	20.23	1.3	19.10	2.4
12	CAVA	24.30	23.56	0.7	26.00	-1.7	28.22	-3.9	26.00	-1.7
13	CENT	17.10	23.56	-6.5	19.10	-2.0	22.05	-4.9	22.05	-4.9
14	COOP	24.10	23.56	0.5	22.70	1.4	24.48	-0.4	24.48	-0.4
15	CROS	24.60	23.56	1.0	20.30	4.3	25.50	-0.9	20.30	4.3
16	DEVI	27.20	23.56	3.6	26.00	1.2	25.87	1.3	25.87	1.3
17	DUNN	20.50	23.56	-3.1	23.10	-2.6	22.14	-1.6	22.14	-1.6
18	EDGE	21.20	23.56	-2.4	22.70	-1.5	21.71	-0.5	22.70	-1.5
19	ELLE	18.10	23.56	-5.5	22.70	-4.6	20.84	-2.7	22.70	-4.6
20	FAIR	20.50	23.56	-3.1	23.10	-2.6	21.34	-0.8	23.10	-2.6
21	FESS	21.30	23.56	-2.3	22.70	-1.4	24.22	-2.9	24.22	-2.9
22	FORM	21.90	23.56	-1.7	24.70	-2.8	21.62	0.3	24.70	-2.8
23	FORT	17.50	23.56	-6.1	19.10	-1.6	20.02	-2.5	19.10	-1.6
24	GARR	24.00	23.56	0.4	29.00	-5.0	23.52	0.5	23.52	0.5
25	GRAN	22.20	23.56	-1.4	29.00	-6.8	25.30	-3.1	29.00	-6.8
26	HANK	21.10	23.56	-2.5	24.70	-3.6	21.93	-0.8	24.70	-3.6
27	HILL	25.00	23.56	1.4	24.70	0.3	25.11	-0.1	25.11	-0.1
28	KEEN	21.10	23.56	-2.5	20.30	0.8	23.16	-2.1	23.16	-2.1
29	LANG	28.40	23.56	4.8	26.00	2.4	27.80	0.6	26.00	2.4
30	LARI	24.10	23.56	0.5	26.00	-1.9	26.02	-1.9	26.02	-1.9
31	LINT	18.50	23.56	-5.1	19.10	-0.6	20.73	-2.2	19.10	-0.6
32	LISB	21.30	23.56	-2.3	24.70	-3.4	22.35	-1.0	24.70	-3.4
33	MADD	24.50	23.56	0.9	29.00	-4.5	25.20	-0.7	25.20	-0.7
34	MCCL	17.50	23.56	-6.1	19.10	-1.6	23.56	-6.1	23.56	-6.1
35	MCHE	21.80	23.56	-1.8	22.70	-0.9	24.75	-3.0	24.75	-3.0
36	MOHA	29.50	23.56	5.9	29.00	0.5	26.10	3.4	29.00	0.5
37	MOTT	24.00	23.56	0.4	23.10	0.9	19.80	4.2	23.10	0.9
38	NAPO	18.60	23.56	-5.0	19.10	-0.5	21.43	-2.8	19.10	-0.5
39	NEW	18.50	23.56	-5.1	19.10	-0.6	21.31	-2.8	19.10	-0.6
40	PETE	28.40	23.56	4.8	26.00	2.4	26.05	2.3	26.05	2.3
41	RUGB	22.00	23.56	-1.6	29.00	-7.0	25.97	-4.0	29.00	-7.0
42	SHAR	25.60	23.56	2.0	22.70	2.9	25.07	0.5	25.07	0.5
43	STAN	24.50	23.56	0.9	29.00	-4.5	24.45	0.0	29.00	-4.5
44	TUTT	17.00	23.56	-6.6	19.10	-2.1	22.79	-5.8	22.79	-5.8
45	VALL	24.70	23.56	1.1	22.70	2.0	23.36	1.3	23.36	1.3
AVERAGE ABS. DIFF.				2.91		2.51		2.07		2.55
MSE				12.84		9.59		7.00		10.49

FIGURE B-10

NORTH DAKOTA SIMULATION - 1966

NO.	NAME	ACTUAL	ESTIMATIONS METH5	DIFF	METH6	DIFF	METH7	DIFF	METH8	DIFF
1	AMID	25.60	22.96	2.6	22.81	2.8	22.81	2.8	23.00	2.6
2	ASHL	16.40	22.98	-6.6	22.40	-6.0	22.40	-6.0	19.30	-2.9
3	BEAC	22.80	22.96	-0.2	22.75	0.1	22.75	0.1	23.10	-0.3
4	BELC	25.00	24.49	0.5	25.63	-0.6	25.63	-0.6	28.50	-3.5
5	BEUL	22.50	23.07	-0.6	22.70	-0.2	22.70	-0.2	23.00	-0.5
6	BISB	26.90	24.45	2.5	25.41	1.5	25.41	1.5	27.60	-0.7
7	BOTT	28.00	24.66	3.3	26.19	1.8	26.19	1.8	28.90	-0.9
8	BOWB	25.70	24.09	1.6	24.92	0.8	24.92	0.8	27.70	-2.0
9	BOWM	16.40	22.95	-6.6	22.66	-6.3	22.66	-6.3	23.10	-6.7
10	CARR	23.60	23.56	0.0	23.37	0.2	23.37	0.2	24.20	-0.6
11	CARS	21.50	22.40	-0.9	21.29	0.2	21.29	0.2	19.90	1.6
12	CAVA	24.30	24.76	-0.5	25.68	-1.4	25.68	-1.4	26.00	-1.7
13	CENT	17.10	22.47	-5.4	21.23	-4.1	21.23	-4.1	21.70	-4.6
14	COOP	24.10	23.70	0.4	23.61	0.5	23.61	0.5	24.10	0.0
15	CROS	24.60	23.37	1.2	22.95	1.7	22.95	1.7	21.80	2.8
16	DEVI	27.20	24.24	3.0	24.86	2.3	24.86	2.3	26.20	1.0
17	DUNN	20.50	23.10	-2.6	22.89	-2.4	22.89	-2.4	23.00	-2.5
18	EDGE	21.20	23.11	-1.9	22.76	-1.6	22.76	-1.6	22.40	-1.2
19	ELLE	18.10	23.22	-5.1	22.92	-4.8	22.92	-4.8	22.70	-4.6
20	FAIR	20.50	22.98	-2.5	22.86	-2.4	22.86	-2.4	22.90	-2.4
21	FESS	21.30	23.66	-2.4	23.66	-2.4	23.66	-2.4	24.80	-3.5
22	FORM	21.90	23.52	-1.6	23.55	-1.6	23.55	-1.6	24.10	-2.2
23	FORT	17.50	22.56	-5.1	21.42	-3.9	21.42	-3.9	19.10	-1.6
24	GARR	24.00	24.16	-0.2	25.39	-1.4	25.39	-1.4	25.60	-1.6
25	GRAN	22.20	26.08	-3.9	28.34	-6.1	28.34	-6.1	28.70	-6.5
26	HANK	21.10	23.66	-2.6	23.86	-2.8	23.86	-2.8	24.70	-3.6
27	HILL	25.00	24.08	0.9	24.42	0.6	24.42	0.6	24.80	0.2
28	KEEN	21.10	22.68	-1.6	21.57	-0.5	21.57	-0.5	22.30	-1.2
29	LANG	28.40	24.52	3.9	25.39	3.0	25.39	3.0	26.20	2.2
30	LARI	24.10	24.39	-0.3	25.12	-1.0	25.12	-1.0	25.40	-1.3
31	LINT	18.50	22.47	-4.0	21.24	-2.7	21.24	-2.7	19.50	-1.0
32	LISB	21.30	23.60	-2.3	23.73	-2.4	23.73	-2.4	23.90	-2.6
33	MADD	24.50	24.21	0.3	24.97	-0.5	24.97	-0.5	26.70	-2.2
34	MCCL	17.50	23.48	-6.0	23.36	-5.9	23.36	-5.9	24.40	-6.9
35	MCHE	21.80	23.79	-2.0	23.80	-2.0	23.80	-2.0	24.60	-2.8
36	MOHA	29.50	24.57	4.9	26.13	3.4	26.13	3.4	28.60	0.9
37	MOTT	24.00	22.78	1.2	22.43	1.6	22.43	1.6	21.70	2.3
38	NAPO	18.60	22.58	-4.0	21.61	-3.0	21.61	-3.0	20.40	-1.8
39	NEW	18.50	21.89	-3.4	20.29	-1.8	20.29	-1.8	20.40	-1.9
40	PETE	28.40	24.40	4.0	25.17	3.2	25.17	3.2	25.60	2.8
41	RUGB	22.00	24.69	-2.7	26.22	-4.2	26.22	-4.2	28.20	-6.2
42	SHAR	25.60	23.95	1.7	24.15	1.4	24.15	1.4	24.70	0.9
43	STAN	24.50	23.88	0.6	24.35	0.2	24.35	0.2	25.00	-0.5
44	TUTT	17.00	22.70	-5.7	21.65	-4.6	21.65	-4.6	21.80	-4.8
45	VALL	24.70	23.49	1.2	23.31	1.4	23.31	1.4	23.60	1.1
AVERAGE ABS. DIFF.				2.54		2.29		2.29		2.35
MSE				9.88		8.17		8.17		8.59

Appendix C. Contour Maps

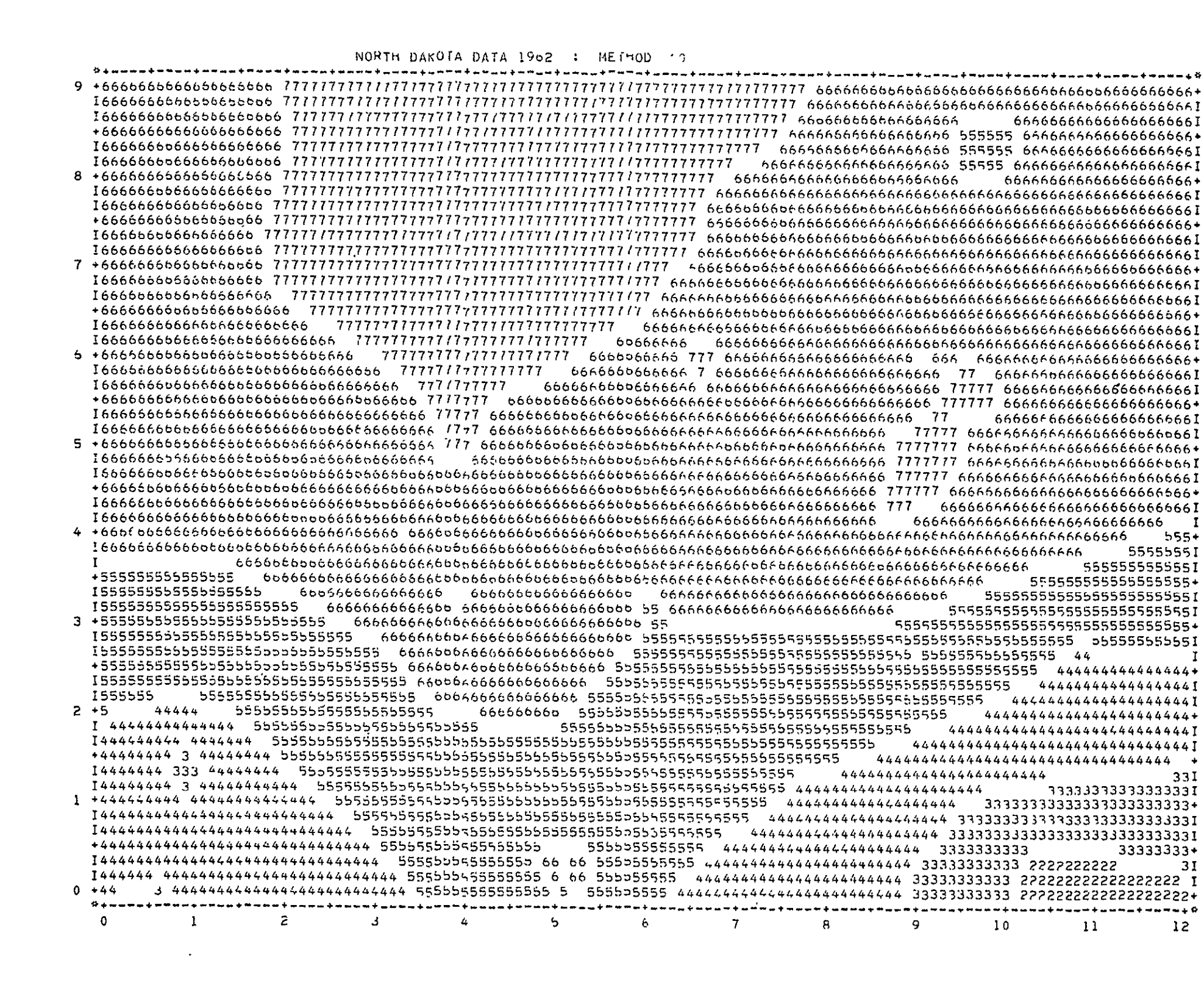
Table C.1 contains a description of the contour maps contained in the appendix and Table C.2 contains the ranges used for each symbol in the maps.

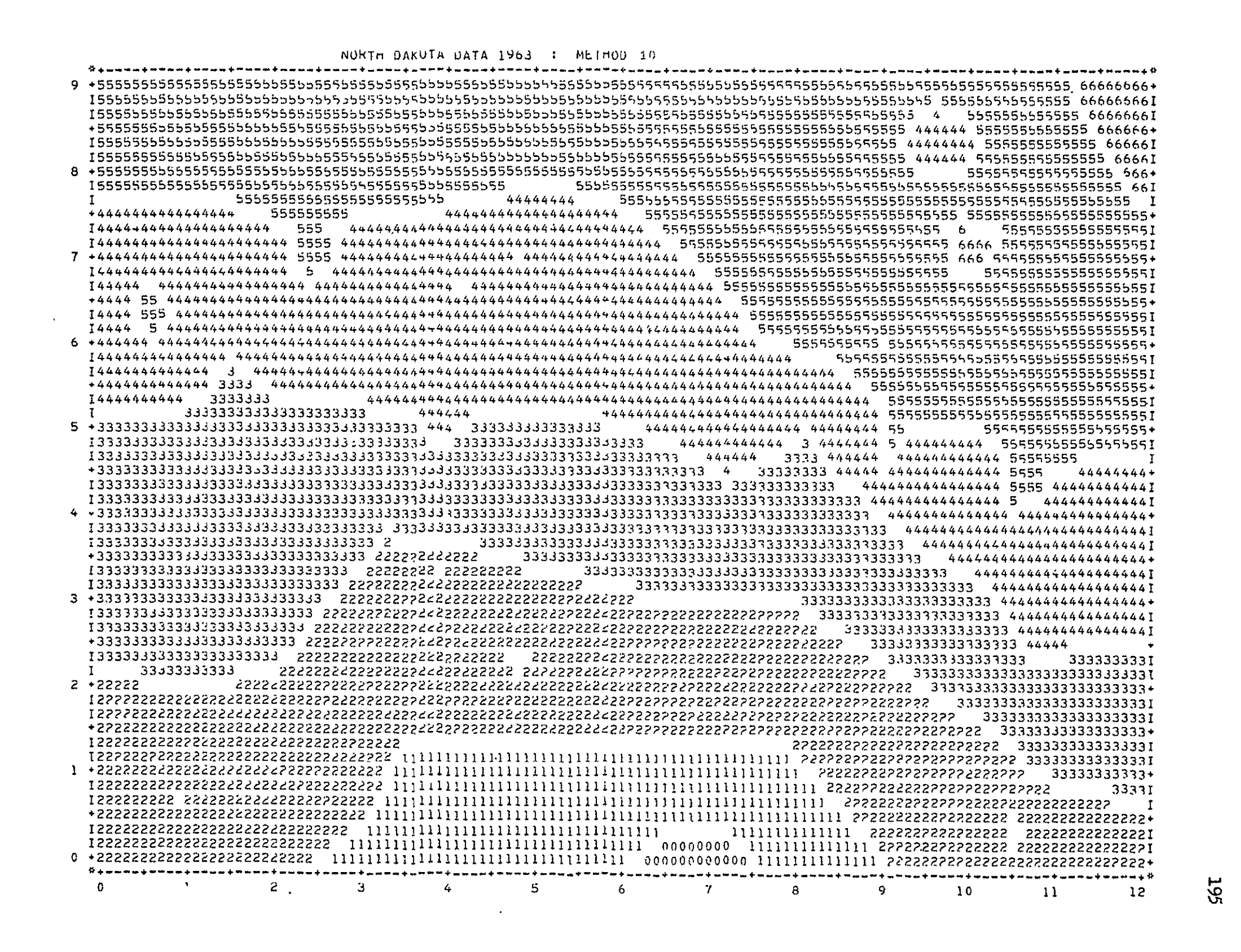
Figure	Description
C-1	Method 10 applied to 51 data points, 1962
C-2	Same, 1963 (48 points)
C-3	Same, 1964 (44 points)
C-4	Same, 1965 (52 points)
C-5	Same, 1966 (52 points)
C-6	Method 10 applied to 7 data points, 1962
C-7	Method 8 applied to 7 data points, 1963
C-8	Method 10 applied to 7 data points, 1964
C-9	Method 9 applied to 7 data points, 1965
C-10	Method 10 applied to 7 data points, 1966
C-11	Method 8 (7 point data) after Step 2, 1963
C-12	Method 3 applied to 7 data points, 1963

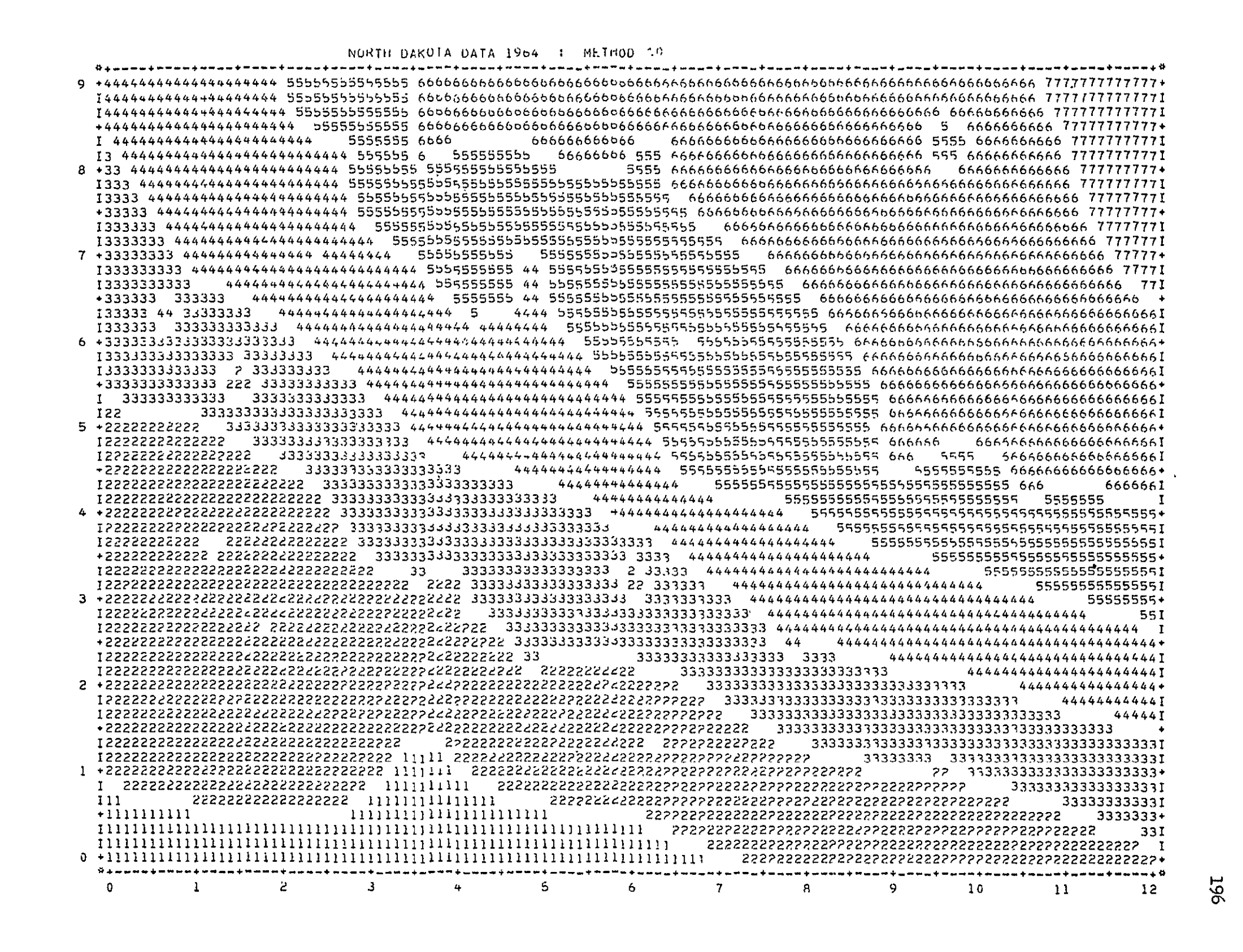
Table C.1 Description of Contour Maps

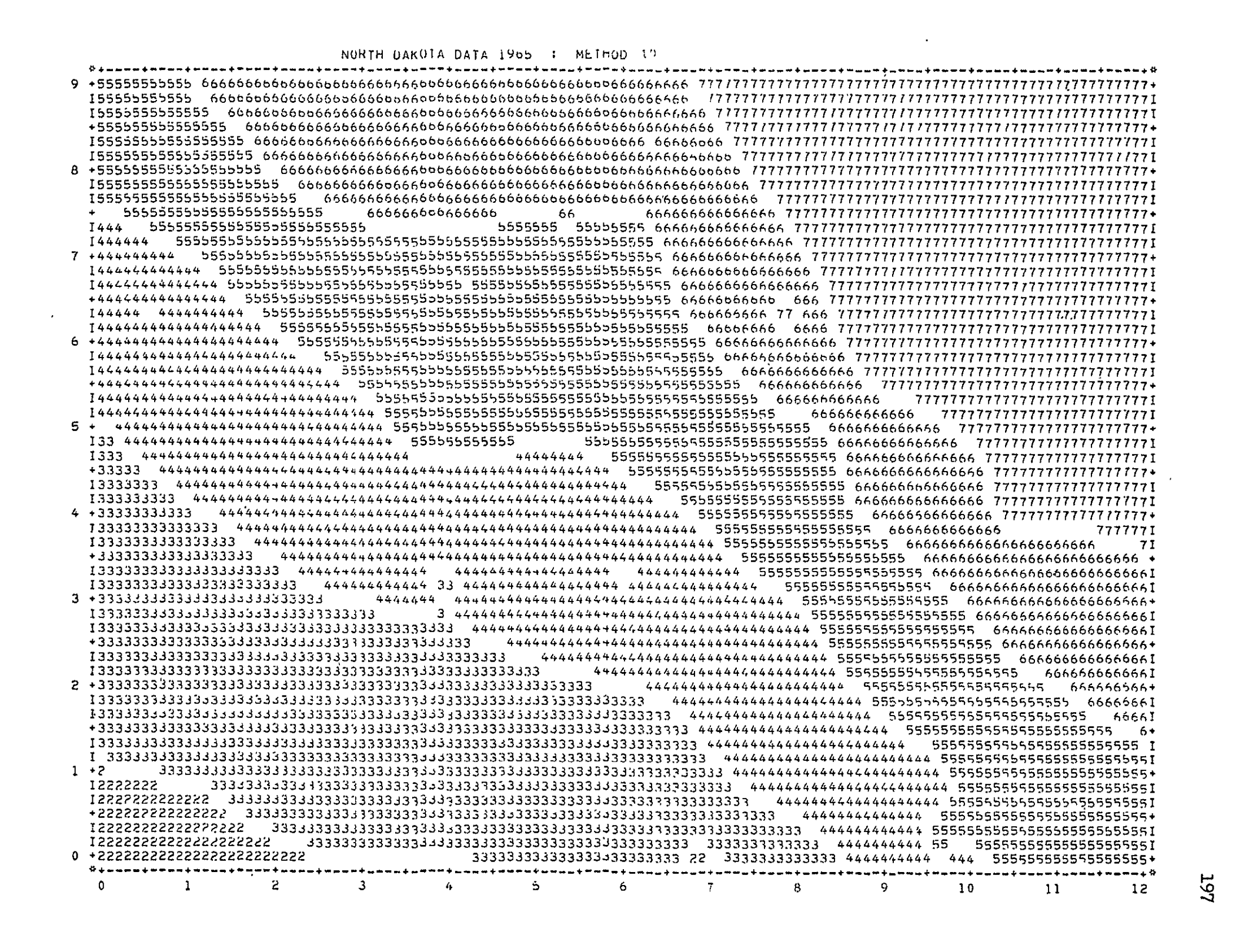
0 = less than 15
1 = 15.0 - 17.5
2 = 17.5 - 20.0
3 = 20.0 - 22.5
4 = 22.5 - 25.0
5 = 25.0 - 27.5
6 = 27.5 - 30.0
7 = greater than 30

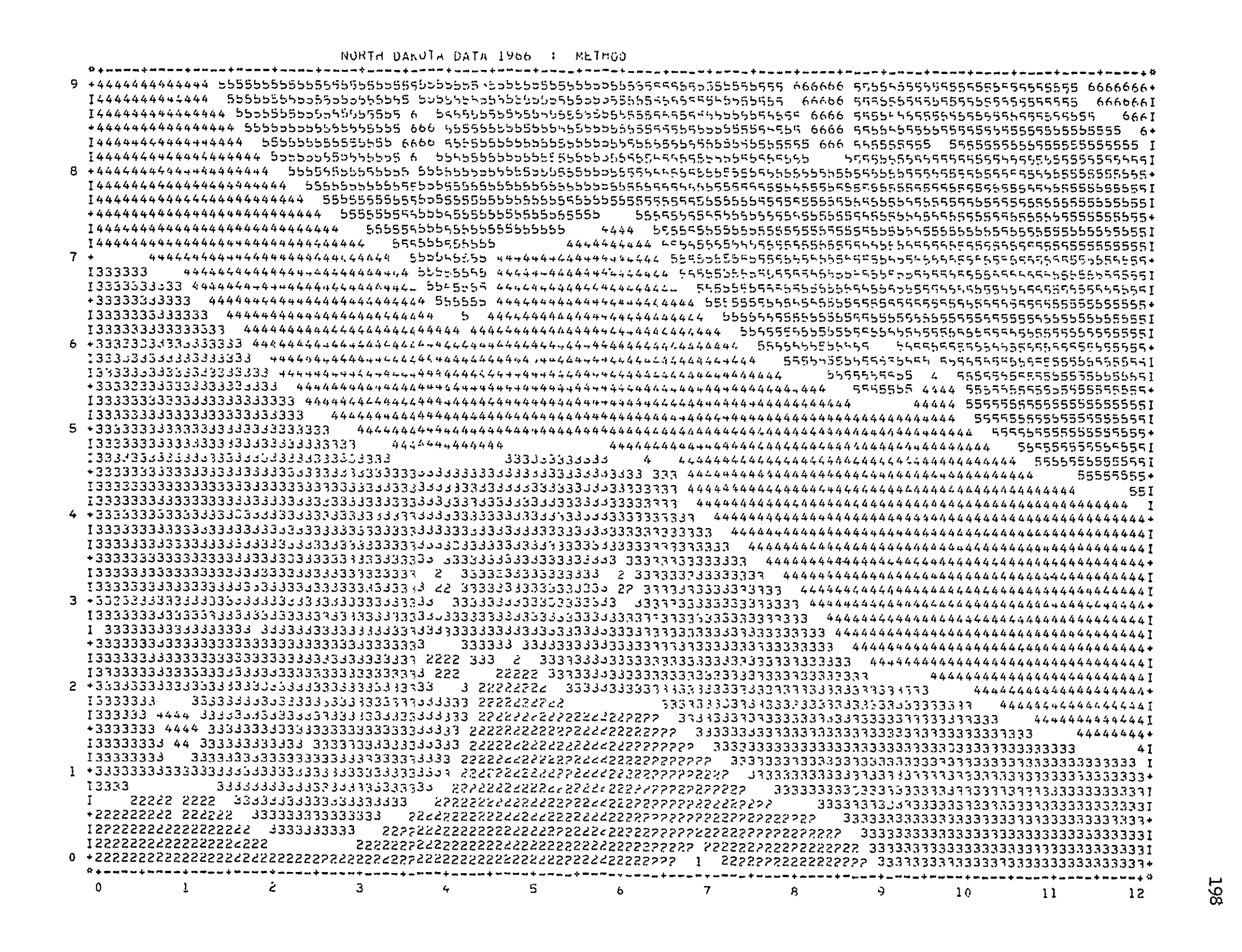
Table C.2 Ranges (bu./acre) for map symbols

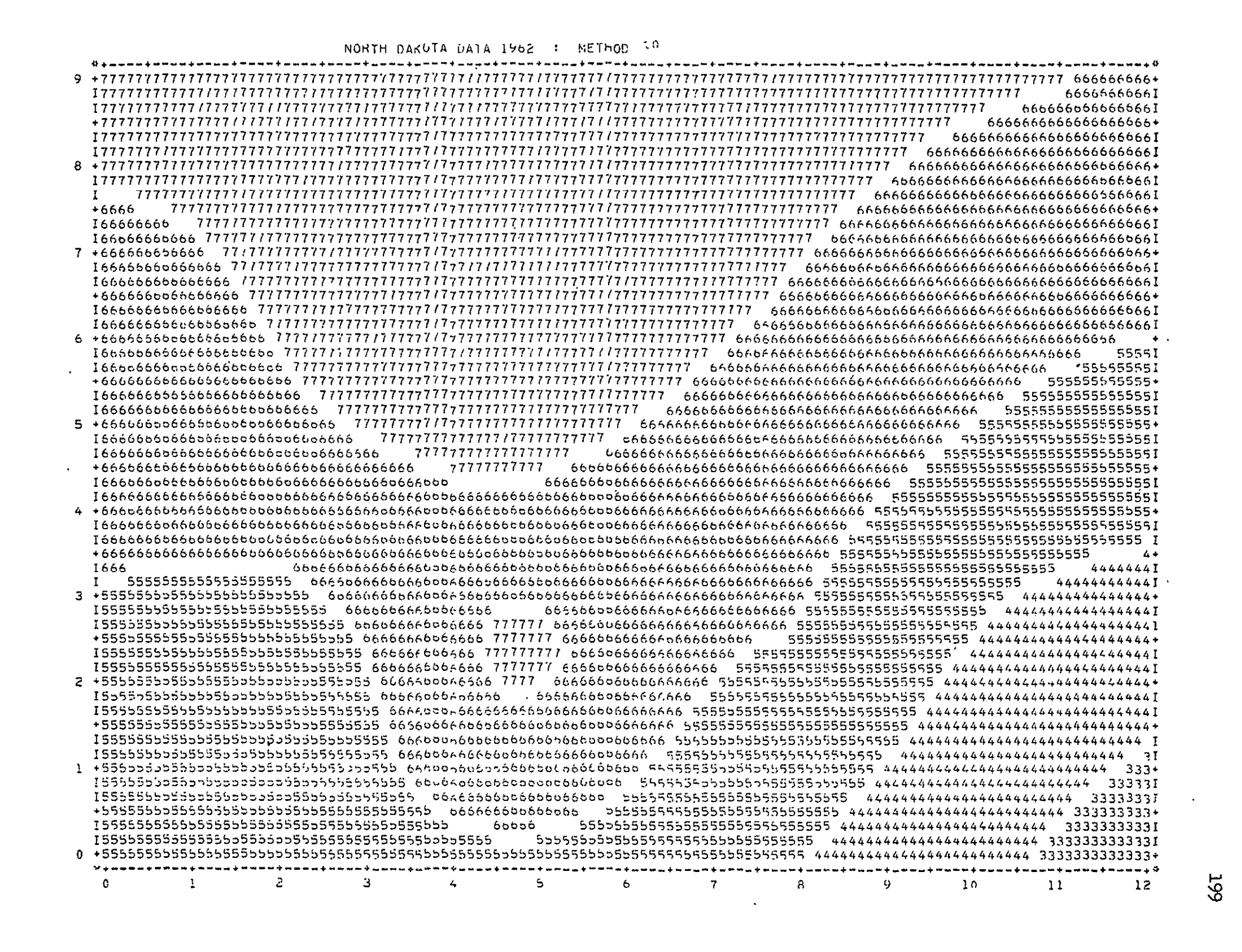

NORTH DAKOTA DATA 1962 : METHOD 10


NORTH DAKOTA DATA 1963 : METHOD 10


NORTH DAKOTA DATA 1964 : METHOD 10
0 1 2 3 4 5 6 7 8 9 10 11 12


NORTH DAKOTA DATA 1965 : METHOD 1)
0 1 2 3 4 5 6 7 8 9 10 11 12


NORTH DAKOTA DATA 1966 : METHOD
0 1 2 3 4 5 6 7 8 9 10 11 12


NORTH DAKOTA DATA 1962 : METHOD 10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
11
12

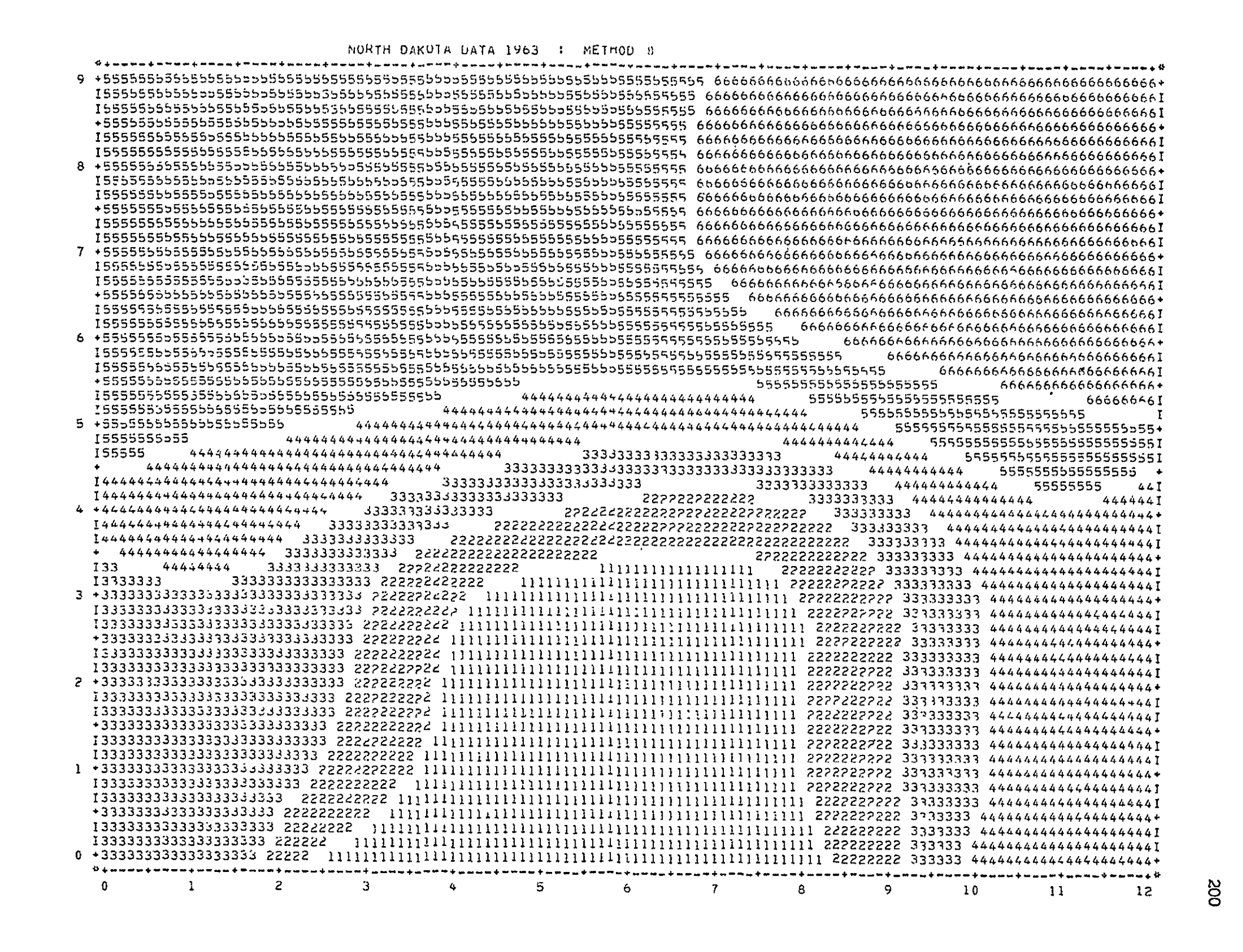
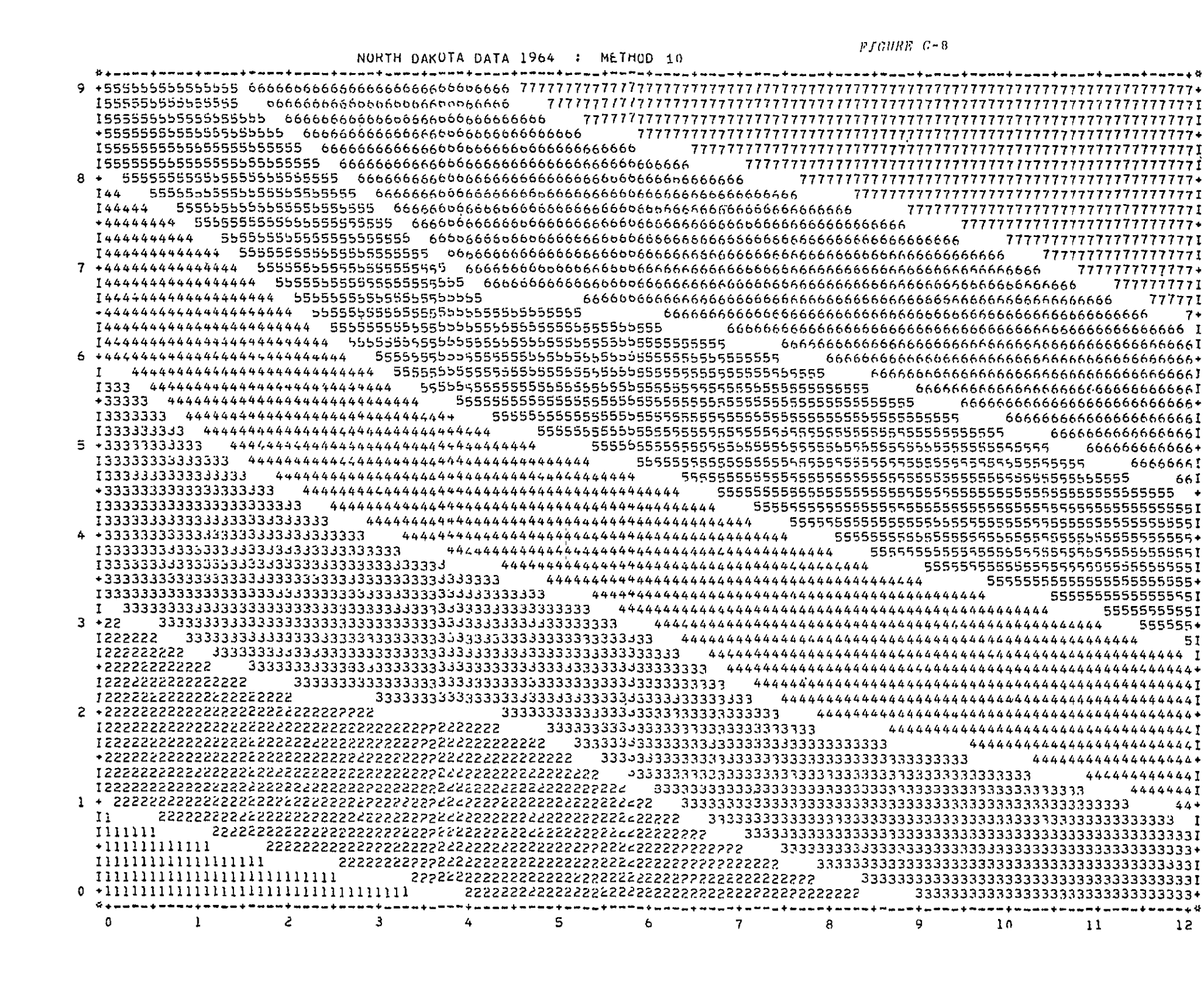
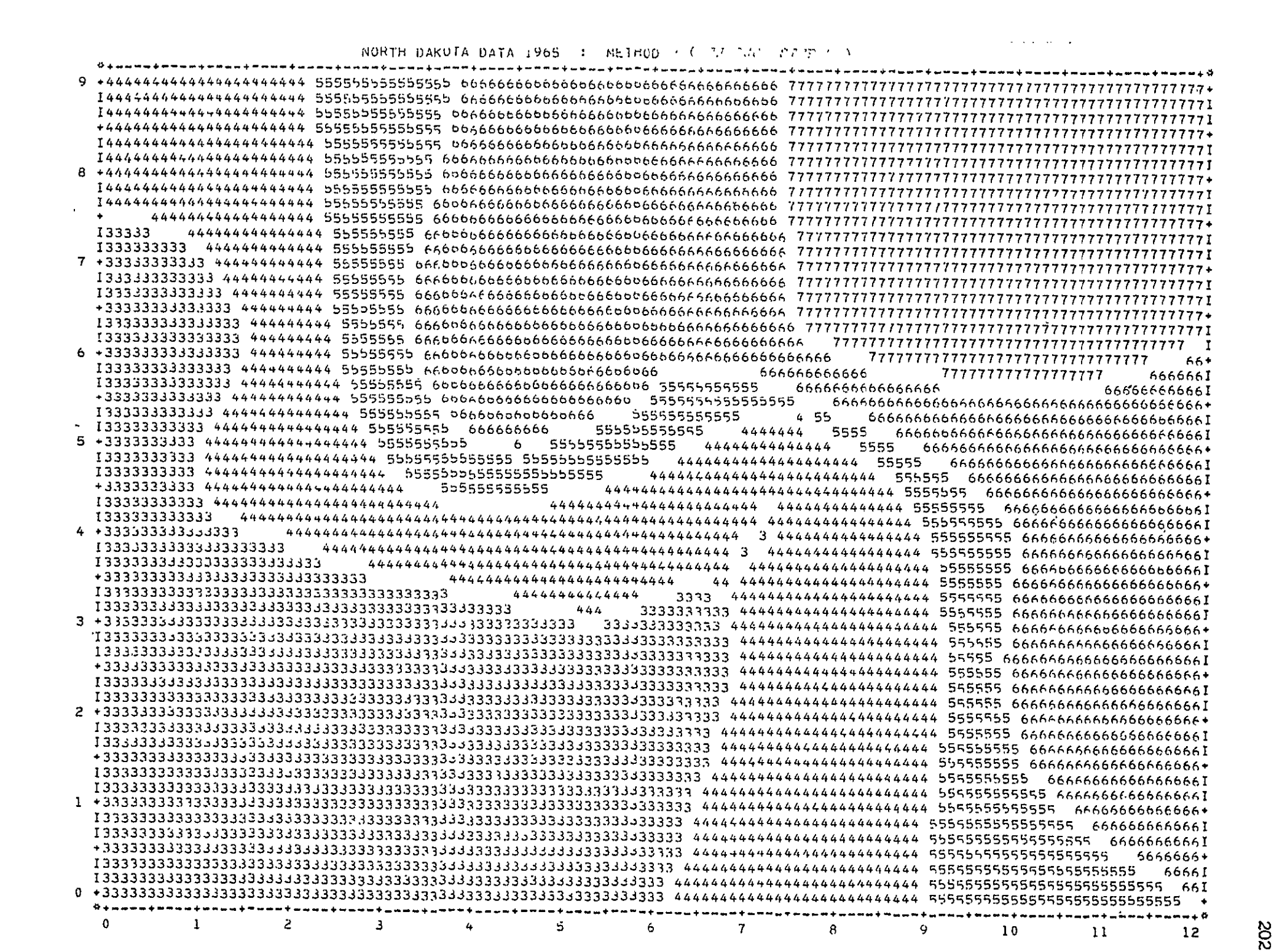
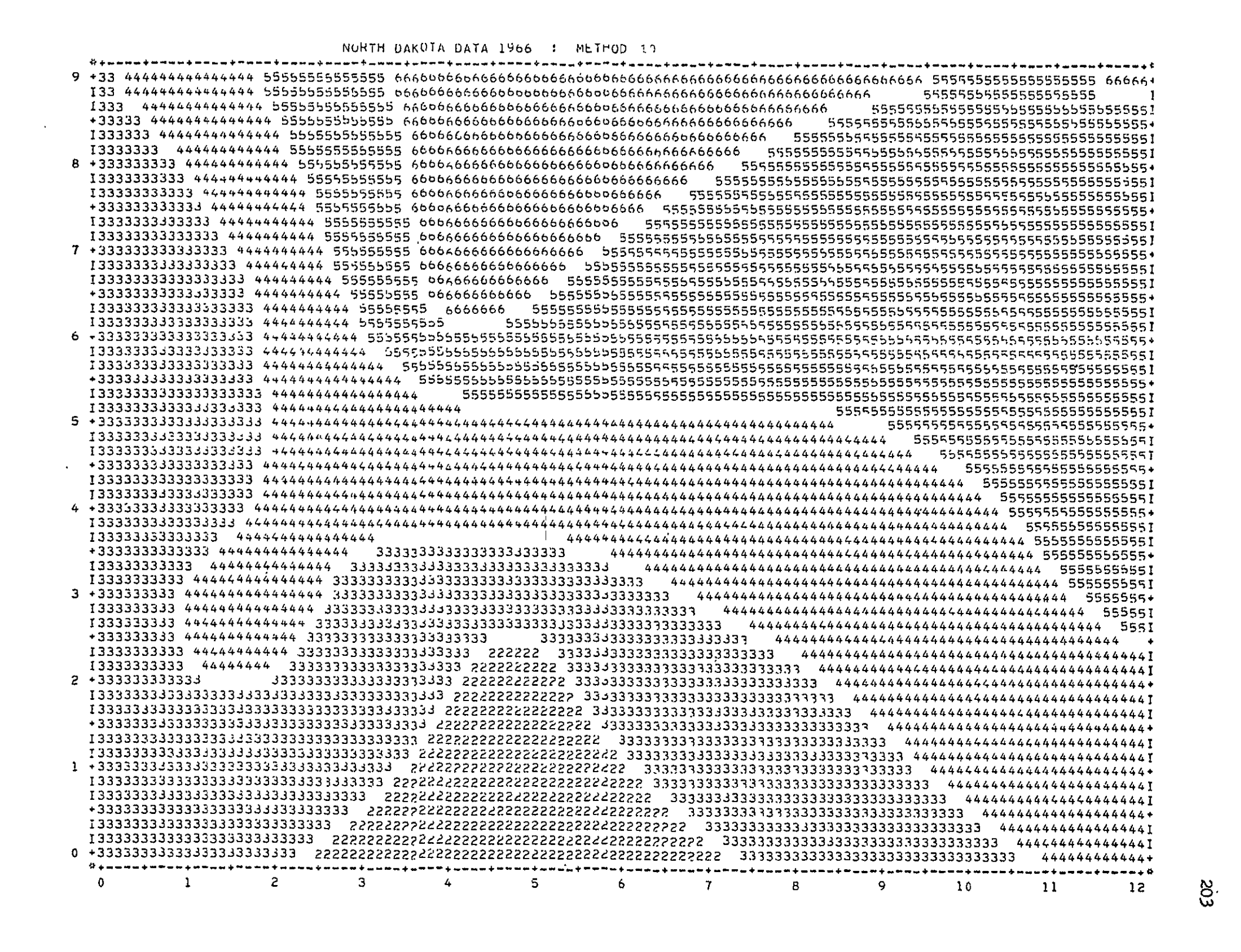

NORTH DAKOTA DATA 1963 : METHOD 8
0 1 2 3 4 5 6 7 8 9 10 11 12

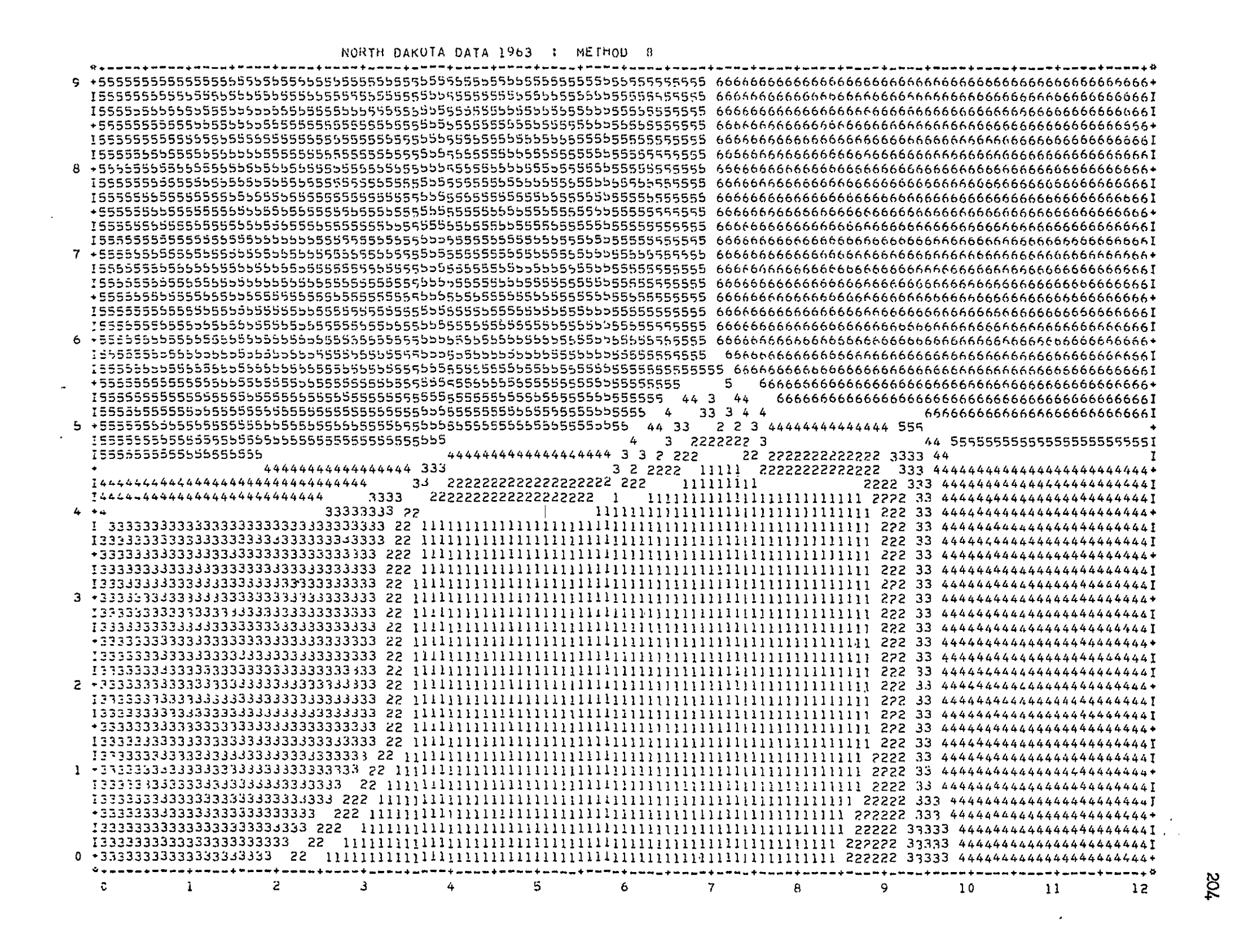
FIGURE C-8


NORTH DAKOTA DATA 1964 : METHOD 10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10 11 12



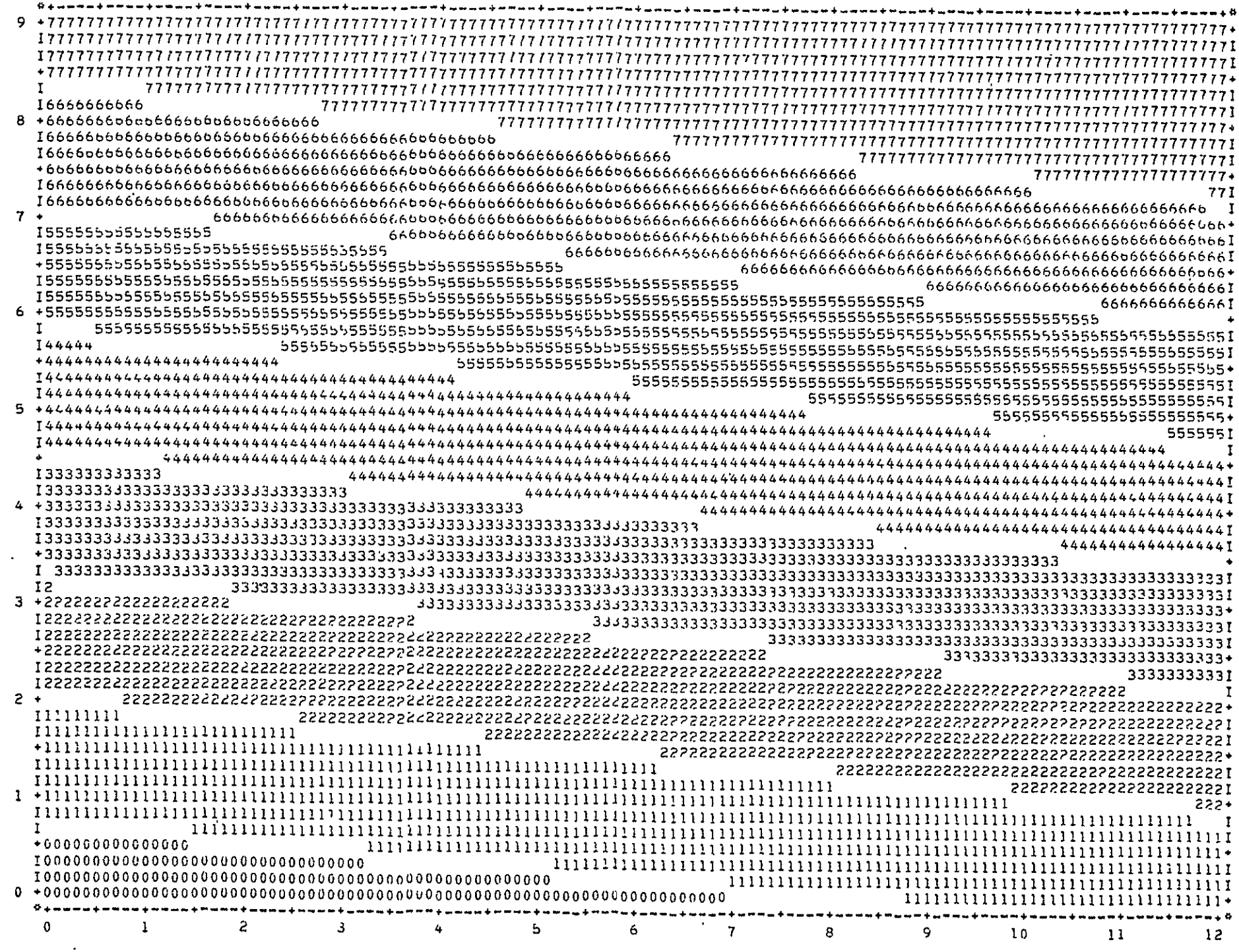

NORTH DAKOTA DATA 1966 : METHOD 10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
11
12

NORTH DAKOTA DATA 1963 : METHOD 8



0 1 2 3 4 5 6 7 8 9 10 11 12

NORTH DAKOTA DATA 1963 : METHOD 3



BIBLIOGRAPHY

[1] Bengtsson, B.E. and Nordbeck, S. (1964), "Construction of Isarithms and Isarithmic Maps by Computers", BIT, Vol. 4.

[2] Haltiner, G.J. (1971), Numerical Weather Predictions, John Wiley & Sons, New York.

[3] McLain, D.H. (1972), "Drawing Contours from Arbitrary Data Points", The Computer Journal, Volume 17, Number 4.

[4] Pitteway, M.L.V. (1967), "Algorithm for Drawing Ellipses or Hyperbolae with a Digital Plotter", The Computer Journal, Volume 10.

[5] Sasaki, Y. (1958), "An Objective Analysis Based on the Variational Method", J. Meteor. Soc. Japan, 36.

[6] ——————(1969), "Proposed Inclusion of Time Variation Terms, Observational and Theoretical, in Numerical Variational Objective Analysis", J. Meteor. Soc. Japan, 47.

[7] ——————(1970), "Some Basic Formalisms in Numerical Variational Analysis", Monthly Weather Review, 98.

[8] Schmidt, A.J. (1966), "SYMAP: A User's Manual", Tri-County Regional Planning Commission, Lansing, Michigan.

[9] Shepard, D. (1965), "A Two-Dimensional Interpolation Function for Irregularly Spaced Data", Proc. 23rd Nat. Conf. ACM.

[10] Wagner, K.K. (1971), "Variational Analysis Using Observational and Low-Pass Filtering Constraints", Unpublished Master's Thesis, Dept. of Meteorology, University of Oklahoma.

COMPUTER PROGRAM DESCRIPTIONS

by

A. H. Kvanli*

*The University of Texas at Dallas

COMPUTER PROGRAM DESCRIPTIONS

A number of Fortran Computer Programs were written during the completion of this contract. Rather than include all of the corresponding listings (approximately 50 pages), this section contains a list of the programs written by The University of Texas at Dallas along with the corresponding input and output.* Computer decks for any of these programs can be obtained upon request.

1.0 OBJECT

Purpose: Perform variational analysis using low-pass filtering constraints as outlined by Wagner [7].

Input: (1) Coordinates and values of data points.

(2) Coordinates of check points.

Output: (1) A complete grid of extrapolated yield values after the completion of Wagner's iterative procedure.

(2) A separate list of the yield values at the specified check points.

2.0 METH09

Purpose: Determine MSE's after Steps 1 and 2 using Method 9 as proposed by Shepard [6]. Different values of the various parameters are considered.

*The descriptions are brief. For a more detailed program description of any of these routines, contact The University of Texas at Dallas.

Input: (1) Coordinates and values of the data points.

(2) Parameter ranges to be considered.

(3) Coordinates and actual values for check points.

Output: For each set of parameter values, the output consists of the MSE's and average absolute deviations for (1) Method 6 (2) Method 9 after Step 1 (3) Method 9 after Step 2.

3.0 METHLO

Purpose: Determine MSE's for each of the four models using Method 9 as proposed by McLain [2] and discussed in [5]. For each model, four weighting factors are considered.

Input: (1) Coordinates and values of the data points.

(2) Coordinates and actual values for check points.

Output: For each weighting factor, a one page listing is obtained giving the actual vs. approximated value using each of the four models at each check point. Also determined are the MSE's and average absolute deviation for each model.

4.0 EXTRAP

Purpose: Calculate and display the results of applying Methods 1 through 8 to a set of yield data as described in [5].

Input: (1) Number of check points and data points.

(2) Check point coordinates.

(3) Data point coordinates and values.

(4) Actual values for check points.

(5) Method 8 (OBJECT) values for check points.

Output: Actual vs. estimated yield values for each check point and each method. Included also are the MSE's and average absolute deviation for each method.

4.1 REGION (part of EXTRAP)

Purpose: Determine if each of a set of input points lies within the convex hull of a set of data points (input).

Input: (1) Coordinates of data points.

(2) Coordinates of points to be checked.

Output: A vector of zero's and one's for each check point where a one indicates that the point lies within the convex hull of the data.

4.2 COMBIN (part of EXTRAP)

Purpose: Determine the matrix of n objects using 3 at a time.

Input: n

Output: Corresponding matrix, e.g. for n = 4, the output would be

$$\begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 1 & 3 & 4 \\ 2 & 3 & 4 \end{bmatrix}$$

5.0 CONMAP

Purpose: Plot a contour map for a grid of data.

Input: (1) The ranges for each of the 9 symbols in the map, e.g. 1 = 15-17.5, 2 = 17.5-20, etc.

(2) The value at each grid point.

Output: A contour map.

6.0 QUADPM

Purpose: Calculate the proportion vector p for the equation Pp = e as discussed in [4]. This routine uses a modified simplex method to determine p.

Input: Matrices P and e

Output: Value of p

7.0 NELSPM

Purpose: Same as 6.0 except this routine uses the procedure outlined by Nelson [3].

Input: Matrices P and e

Output: Value of p given by

$$p = a + By$$

where (1) y is an arbitrary vector

(2) a is an output vector

(3) B is an output matrix

7.1 PSEUDO (part of NELSPM)

Purpose: Determine the pseudoinverse of a matrix, A.

Input: A

Output: Pseudoinverse of A = A^{+}

8.0 MISCLS

Purpose: To plot probabilities of misclassification as discussed in [1].

Input: (1) a priori probability of $\pi_1 = p_1$

(2) Values for $\mu^{(1)} - \mu^{(2)}$

Output: CALCOMP plots of

(1) P (1|2)

(2) P (2|1)

(3) Total probability of misclassification

REFERENCES

[1] McElroy, D. D. and Gray, H. L. (1975) "Probability of Misclassification with Missing Data," Technical Report for NASA.

[2] McLain, D. H. (1972) "Drawing Contours from Arbitrary Data Points," The Computer Journal, Volume 17, Number 4.

[3] Nelson, D. L. (1969) "Quadratic Programming Techniques Using Matrix Pseudoinverses," Unpublished Ph.D. Dissertation, Texas Tech University.

[4] Odell, P. L. and Kvanli, A. H. (1975) "On Solving for the Probability Vector p in the Equation Ap = e where the Columns of A and e are Probability Vectors," Technical Report for NASA.

[5] Odell, P. L., Kvanli, A. H., and Simpson, C. (1975) "Extrapolation Procedures for Irregularly Spaced Sparse Data — A Review and Comparison," Technical Report for NASA.

[6] Shepard, D. (1965) "A Two-Dimensional Interpolation Function for Irregularly Spaced Data," Proc. 23rd Nat. Conf. ACM.

[7] Wagner, K. K. (1971) "Variational Analysis Using Observational and Low-Pass Filtering Constraints," Unpublished Masters Thesis, Department of Meteorology, University of Oklahoma.