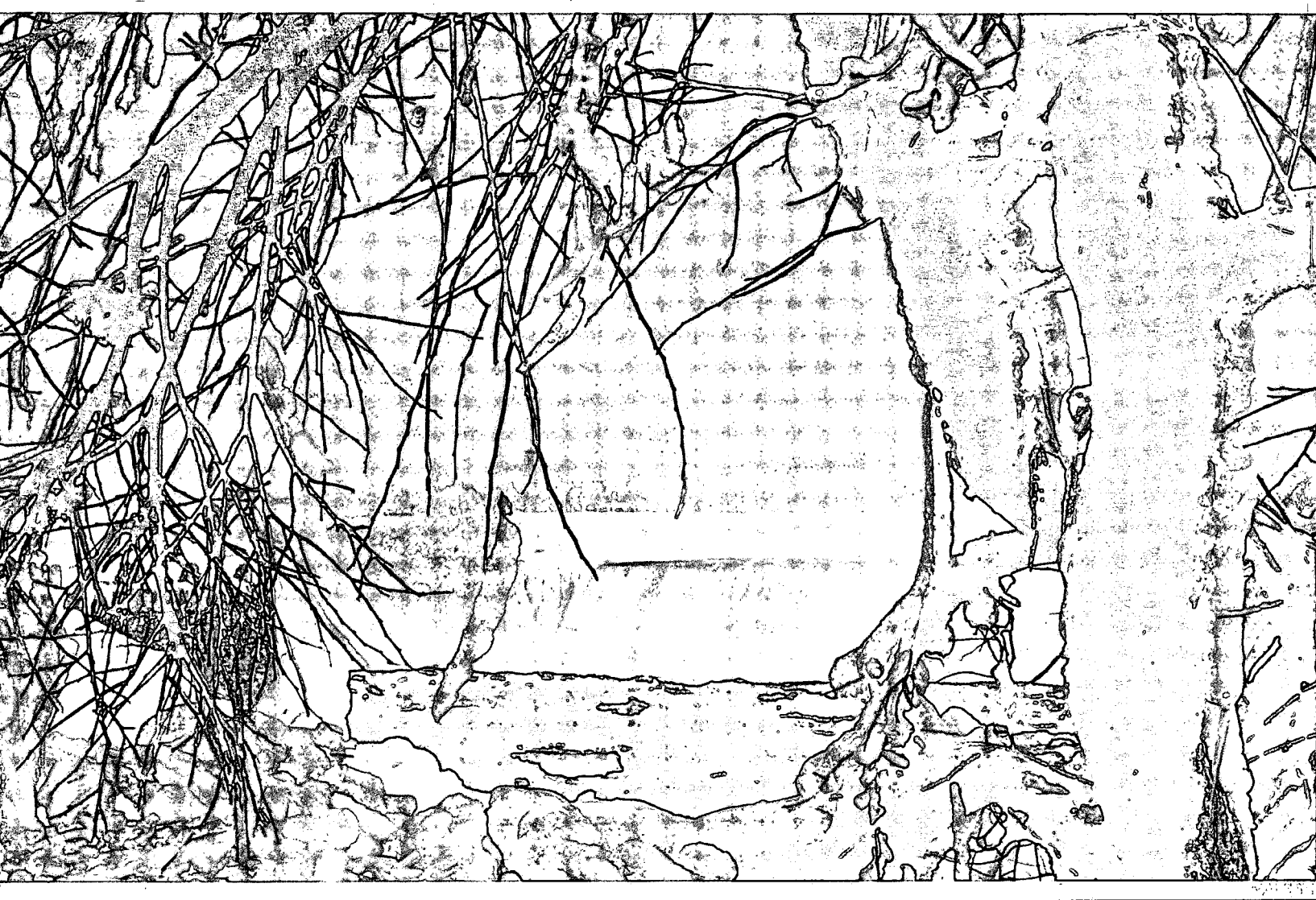
Environment Canada
Environnement Canada

Statistical Procedures for Classification of a Lake

A.H. El-Shaarawi and K.R. Shah



SCIENTIFIC SERIES NO. 86
(Résumé en français)

INLAND WATERS DIRECTORATE,
NATIONAL WATER RESEARCH INSTITUTE,
CANADA CENTRE FOR INLAND WATERS,
BURLINGTON, ONTARIO, 1978.

Environment Canada Environnement Canada

Statistical Procedures for Classification of a Lake

A.H. El–Shaarawi and K.R. Shah

SCIENTIFIC SERIES NO. 86
(Résumé en français)

INLAND WATERS DIRECTORATE,
NATIONAL WATER RESEARCH INSTITUTE,
CANADA CENTRE FOR INLAND WATERS,
BURLINGTON,ONTARIO, 1978.

Cat. No. En 36-502/86
ISBN 0-662-10039-5
Contract No. KL229-7-1040
THORN PRESS LIMITED

Contents

Tables

Illustrations

Abstract

Statistical classification procedures for univariate and multivariate limnological data are presented. A regression model in terms of additive temporal and spatial components is fitted to the data after a search for an appropriate transformation. When the spatial component is found to be significant, a hierarchical procedure is suggested to divide the lake into regions. The procedure is illustrated using the data on phytoplankton biomass from Lake Superior collected in 1973.

Résumé

Le présent rapport renferme des renseignements sur des méthodes statistiques de classification s'appliquant à des données limnologiques à une et plusieurs variables. Un modèle de régression en ce qui concerne les composantes temporelles et spatiales additives est appliqué aux données après la recherche d'une transformation appropriée. Dans les cas où la composante spatiale est significative, une méthode hiérarchique est proposée afin de diviser le lac en régions. La méthode est démontrée à l'aide de données sur la biomasse du phytoplancton du lac Supérieur, recueillies en 1973.

Statistical Procedures for Classification of a Lake

A.H. El–Shaarawi and K.R. Shah*

INTRODUCTION

In limnological investigations it is important to classify a given body of water into zones according to the values of a specific character or a set of characters. A statistical procedure in the framework of a regression model is presented here. Since a regression model does not always fit raw data (Taylor, 1961), it would be desirable to look for a transformation so that the assumptions of the standard regression model approximately hold for the transformed data. A procedure of Box and Cox (1964) is used to find a suitable transformation. An additive linear model with seasonal and spatial components is fitted to the transformed data. A hierarchical classification procedure using estimates of spatial effects is proposed here. A multivariate generalization is outlined briefly.

The univariate procedure is illustrated using the data on phytoplankton biomass from Lake Superior collected by the Canada Centre for Inland Waters (CCIW) in 1973. The plots of residuals indicate that the model is reasonable. The classification procedure divides the lake into three zones. The maximum biomass was found in mid-summer.

THE MODEL

Let Y_{ij} denote the observed measurement on the character of interest during the ith cruise at the jth sampling station, where $i = 1, 2, \ldots, \ell_1$; $j = 1,2, \ldots, \ell_2$. These measurements are assumed to be a realization of n independent random variables whose probability behaviour is described below. As is usually the case, technical difficulties prevented the collection of observations from each station during each cruise and hence n, the total number of observations, is less than $\ell_1\ell_2$. Box and Cox (1964) considered a family of transformations given by

$$z_{ij} = \begin{cases} (y_{ij}^{\lambda} - 1)/\lambda \ , \lambda \neq 0 \\ \ell n y_{ij} \qquad , \lambda = 0 \end{cases} \tag{1}$$

where the parameter λ defines a particular transformation and the random variable z_{ij} is defined for $y_{ij} > 0$. It is assumed that for a set of values of λ, say $\lambda_1 \leqslant \lambda \leqslant \lambda_2$, the random variable z_{ij} is approximately normally distributed with the mean

$$E(z_{ij}) = \mu + \alpha_i + \beta_j, \tag{2}$$

and the variance

$$\text{var}(z_{ij}) = \sigma^2, \tag{3}$$

where μ, α_i and β_j are unknown constants. According to the above it is assumed that the non-linear transformation on y_{ij} results in resolving the mean value of z_{ij} into three additive components: μ is the general mean, α_i is the effect due to the ith cruise, and β_j is the effect due to the jth sampling station. Hence, apart from λ, the transformation parameter, the problem is reduced to that of estimating the main effects in a non-orthogonal factorial experiment with only two factors. The first factor has ℓ_1 levels (the number of cruises) and the second has ℓ_2 levels (the number of stations). These formulations can be expressed in matrix notations as follows. Let $\underline{z}$ be the vector of transformed observations, then

$$E(\underline{z}) = 1_n\mu + M_1'\underline{\alpha} + M_2'\underline{\beta}, \tag{4}$$

and

$$\text{var}(\underline{z}) = I\sigma^2, \tag{5}$$

where $\underline{1}_n$ is a column vector of length n with each element unity, M_i', $i = 1, 2$, is a binary incidence matrix of order $(n \times \ell_i)$ with rank ℓ_i, $\underline{\alpha}$ is a column vector of length ℓ_1 and its ith element is α_i, $\underline{\beta}$ is a column vector of length ℓ_2 and its jth element is β_j, and I is a unit matrix. Setting $A = [\underline{1}_n : M_1' : M_2']$ and $\underline{\theta}' = [\mu, \underline{\alpha}', \underline{\beta}']$, where $\underline{\alpha}'$ is the transpose of $\underline{\alpha}$, Equation 4 can be written as

$$E(\underline{z}) = A\underline{\theta}. \tag{6}$$

Assuming that $\underline{z}$ has a multivariate normal distribution and $\underline{y}$ represents the original observational vector, the probability of obtaining $\underline{y}$ or the likelihood function for $\underline{\theta}$, λ and σ^2 in relation to the original vector of observations is

$$(2\pi)^{-n/2} \cdot \sigma^{-n} \cdot \exp\{-(1/2\sigma^2) [(\underline{z} - A\underline{\theta})'(\underline{z} - A\underline{\theta})]\} J(\lambda, \underline{y}), \tag{7}$$

*Department of Statistics, University of Waterloo, Waterloo, Ontario.

where

$$J(\lambda, \underline{y}) = \prod_{i,j} \left| \frac{dz_{ij}}{dy_{ij}} \right|.$$

Equation 7 represents the general model. The application of this model to a particular case requires the estimation of the unknown parameters and testing of different hypotheses about them. Moreover, it is necessary to check the suitability of the model to the particular case by analyzing the residuals.

ESTIMATION OF THE PARAMETERS

The method of maximum likelihood can be used to estimate the parameters $\underline{\theta}$, σ^2 and λ. This method may be applied in two steps. First, for a given λ, Equation 7, except for a constant factor, is the likelihood for a standard least squares problem. Hence the estimate $\hat{\underline{\theta}}(\lambda)$ of $\underline{\theta}$ is a solution of the normal equation (Plackett, 1960)

$$(A'A)\hat{\underline{\theta}}(\lambda) = A'\underline{z}, \tag{8}$$

and the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2(\lambda) = 1/n \left\{ [\underline{z} - A\hat{\underline{\theta}}(\lambda)]' [z - A\hat{\underline{\theta}}(\lambda)] \right\}. \tag{9}$$

Second, if we substitute for $\underline{\theta}$ and σ^2 in Equation 7 their estimates $\hat{\underline{\theta}}(\lambda)$ and $\hat{\sigma}^2(\lambda)$ and take the logarithm, then for a fixed λ, the maximized log likelihood, except for a constant, is

$$L_{max}^{(\lambda)} = -(n/2)\, \ell n\, (\hat{\sigma}^2_{(\lambda)}) + \ell n(J(\lambda, \underline{y})), \tag{10}$$

where, from Equation 1, we have

$$\ell n(J(\lambda, \underline{y})) = (\lambda - 1) \sum_{ij} \ell n(y_{ij}). \tag{11}$$

The value $\hat{\lambda}$ which maximizes Equation 10 is the maximum likelihood estimate for λ. This estimate is not available analytically, and numerical methods such as Newton-Raphson method can be used to obtain $\hat{\lambda}$, which is the root of

$$dL_{max}^{(\lambda)} / d\lambda = 0. \tag{12}$$

However, an approximate estimate for λ can be obtained graphically by plotting $L_{max}^{(\lambda)}$ against λ for a trial series of values, and $\hat{\lambda}$ may be read from the plot. Once $\hat{\lambda}$ is found, the maximum likelihood estimates for $\underline{\theta}$ and σ^2 can be calculated from Equations 8 and 9, respectively. The solution of the normal equation is rather cumbersome to obtain because of large dimensionality and the non-orthogonality of the matrix (A'A). An analytical solution that reduces substantially the size of the matrix that needs to be inverted is given in El-Shaarawi (1972).

HYPOTHESES TESTING

The regression and the residual sums of squares are given respectively by $REG = \hat{\underline{\theta}}'(\hat{\lambda})\, A'\, \underline{z}$ and $RES = \underline{z}'\underline{z} - REG$. REG gives the total variability explained by the model, while RES represents the unexplained variability or measures the chance variation if the model is correct. Under the assumptions given about the distribution of $\underline{z}$ and ignoring the random fluctuations in $\hat{\lambda}$, RES/σ^2 has a χ^2 distribution with $N = n - \ell_1 - \ell_2 + 1$ degrees of freedom. To test whether there are spatial differences, i.e. differences between stations, we test the null hypothesis

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_{\ell_2} = 0.$$

Under this hypothesis Equation 4 reduces to

$$E(\underline{z}) = \underline{1}_n \mu + M_1' \underline{\alpha}$$

$$= A_1 \underline{\theta}_1,$$

where $A_1 = [\underline{1}_n : M_1']$ and $\underline{\theta}_1' = [\mu : \underline{\alpha}']$. Let $\hat{\underline{\theta}}_1(\lambda)$ be the least squares estimate of $\underline{\theta}_1$. The regression sum of squares becomes $REG1 = \hat{\underline{\theta}}_1'(\lambda) A_1' \underline{z}$. The reduction in the regression sum of squares, which resulted from accepting H_0, is $RED1 = REG - REG1$. Under H_0 the statistic $RED1/\sigma^2$ has a χ^2 distribution with $(\ell_2 - 1)$ degrees of freedom. Since RES and RED1 are independently distributed, the statistic

$$F_1 = [RED1/(\ell_1 - 1)]/(RES/N)$$

has Fisher's F distribution with $(\ell_1 - 1)$ and N degrees of freedom (d.f.), and hence can be used for testing H_0. Similarly the statistic

$$F_2 = [RED2/(\ell_2 - 1)]/(RES/N)$$

can be used to test the differences between cruises by comparing its observed value with that for an appropriate F distribution, where RED2 is obtained by making obvious changes in the procedure for computing RED1.

Two methods can be used for making inferences about λ. The first method makes use of the fact that for large samples the statistic $-2\ell nR(\lambda)$ is distributed approximately as χ^2 with a single degree of freedom, where

$$R(\lambda) = \exp(L_{max}^{(\lambda)}) / \exp(L_{max}^{(\hat{\lambda})}).$$

This leads to a test of significance and the confidence intervals for λ in the usual way. The other method requires the plotting or tabulating of $R(\lambda)$ for different values of λ, and constructing the interval for λ with a specified degree of plausibility according to the method suggested by Kalbfleisch and Sprott (1970). Though the two methods are operationally similar, the first one is based on large sample theory and has probability interpretation.

To examine the adequacy of the suggested model we present the plots of residuals of z_{ij} for $\lambda = \hat{\lambda}$. More specifically, we plot $e_{ij}(\hat{\lambda}) = [z_{ij}(\hat{\lambda}) - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j] / \hat{\sigma}(\hat{\lambda})$ against $z_{ij}(\hat{\lambda})$. Random pattern is expected if the model is adequate. We also present Q-Q plots of residuals to examine if the transformed variables are approximately normally distributed (Wilk and Gnandesikan, 1968).

CONSTRUCTION OF ZONES

If the previous analysis showed that there are no significant differences between the sampling stations, the data then suggest that the lake can be regarded as a single homogeneous zone. On the other hand, if the analysis suggested the existence of real differences between the stations then the lake can be divided into more than one zone. This can be accomplished in the following manner. The sampling stations may be ranked according to the values of $\hat{\beta}_i$ (arranged in increasing or decreasing order). One may then combine the pair of stations which are "closest" in some sense into one group. One way of doing this is to rewrite the model, assigning a common β value for a pair of stations, and to compute the decrease in REG resulting from this change in the model. This may be computed for all possible pairs of stations and the pair giving the smallest decrease may be regarded as closest. However, for simplicity we combine into a group the pair (i, j) for which $\hat{\beta}_i - \hat{\beta}_j$ is minimum. The difference between this and the first method is due to the non-orthogonality in the data and is expected to be small. This process of grouping can be continued until all the stations are combined into a single group. This procedure determines a hierarchical classification of stations and may be graphically represented in the form of a "Dendrogram" (Hartigan, 1975) or classification tree.

This tree can be used to divide the lake into zones. This could either be done subjectively or by using an ad hoc statistical procedure such as the following one. Let Z_i denote the change in REG at the ith stage. Distribution of $NZ_i/ni\hat{\sigma}^2$ is not easily obtained. If this is regarded as an F with (i, N) d.f. the resulting level of significance will be higher than the nominal level. A large sample approximation to the distribution of $NZ_i/ni\hat{\sigma}^2$ may be obtained from its first few moments.

Another reasonable procedure would be to stop when Z_i/RED1 exceeds a predetermined number such as 0.05 or 0.1.

MULTIVARIATE METHODS

In most limnological investigations there are several measurements which should be taken into account. One approach would be to combine these into a single measurement, for example, by taking the linear combination of measurements that corresponds to the first principal component. If this linear combination is nearly the same for the different cruises, the univariate methods described earlier can be used.

Another approach is to start with analysis of dispersion for two-way non-orthogonal classification using the general methods given in Rao (1965). This analysis should be preceded by an analysis of transformation carried out separately on each character in a manner described earlier.

Assuming that the measurements on each of the p characters are available whenever a station was visited in a cruise, the multivariate model may be written as

$$\underset{n\times p}{Z} = \underset{n\times k}{A}\ \underset{k\times p}{\Theta} = \underset{n\times p}{E}$$

where the ith column of Z gives the measurements on the ith character, the ith column of Θ gives the $k(=\ell_1 + \ell_2 + 1)$ parameters for the ith character, and E is the matrix of errors. Rows of E are assumed to be independently and identically distributed, each having a p-variate normal distribution with zero mean and the covariance matrix Σ. Let $\hat{\Theta}$ denote the estimate of Θ which is obtained by solving $(A'A)\hat{\Theta} = A'Z$. As is well known (Rao, 1965), this amounts to obtaining the least squares estimates separately for each character. The estimate of Σ is given by

$$\hat{\Sigma} = (Z - A\hat{\Theta})'(Z - A\hat{\Theta})/(n - \ell_1 - \ell_2 + 1).$$

Let $\underline{\beta}_i$ denote the p x 1 vector of station effects for the p characters for the ith station. Initially, one may examine the null hypothesis $\underline{\beta}_1 = \underline{\beta}_2 = \cdots = \underline{\beta}_{\ell_1}$. This may be done by rewriting the model by incorporating the hypothesis and re-computing the $\hat{\Sigma}$ matrix for this model. The ratio of determinants of the $\hat{\Sigma}$ matrices under the model and under the hypothesis or some other appropriate function of these matrices may be used as a test statistic. When this null hypothesis is rejected, one groups the stations into zones. The following hierarchical method may be used for this. Let $a_{ij}\hat{\Sigma}$ denote the covariance matrix for $\hat{\beta}_i - \hat{\beta}_j$. It may be noted that a_{ij} can be obtained from the univariate

Table 1. Dates and Estimates of Cruise Effects

	Date of cruise					
	May 12 – 24	June 15 – 28	July 26 – Aug. 9	Sept. 4 – 18	Oct. 9 – 29	Nov. 13 – Dec. 3
Estimated cruise effect	– 0.62694	– 0.23075	0.57607	0.13560	0.18096	– 0.43871

methods described in the previous section and a_{ij} would all be equal if the data were orthogonal. The "distance" between the ith and the jth stations may be computed as $d(i, j) = (\hat{\beta}_i - \hat{\beta}_j)' \hat{\Sigma}^{-1} (\hat{\beta}_i - \hat{\beta}_j) / a_{ij}$. We first search for the pair for which d(i, j) is minimum. For the next stage we rewrite the model with a common $\underline{\beta}$ for these two stations but we retain the same $\hat{\Sigma}$ as for the initial model. The process can be continued as in the univariate case until all the stations are grouped into a single zone.

APPLICATIONS

The data discussed here were obtained during six cruises on Lake Superior in 1973. These cruises form a part of the surveillance program carried out by the CCIW at Burlington, Ontario. The data on phytoplankton biomass were obtained from 37 stations. The first row in Table 1 gives the dates of these cruises, while the pattern of stations is given in Figure 1. The analysis of transformation gave $\hat{\lambda} = 0.16$. The analysis of variance using this transformation was carried out. For the hypothesis of equality of station effects the observed value of the F statistic based on 36 and 145 d.f. was 5.73, which is significant at 1% level. For the hypothesis of equality of cruise effects the observed value based on 5 and 145 d.f. was 18.73, which is also significant at 1% level. It may be noted that the total number of observations is 187, which is 35 less than the number obtainable if the data were available from each station for each cruise. This justifies the analysis that takes into account the non-orthogonality present in the data.

The cruise effects, given in Table 1, increase steadily during the first three cruises and then decrease steadily. This indicates that for Lake Superior as a whole the maximum phytoplankton biomass level is reached around midsummer.

Since the differences between the stations are highly significant, we proceed with a more detailed investigation consisting of clustering of the stations into zones. At the nominal 5% level this gave three zones. Estimates of station effects are given in Table 2. Figure 1 gives the map of the lake divided into three zones formed by this procedure. The three zones are roughly the near-shore, offshore and the main lake, with the main lake being the biggest zone and having the lowest biomass values.

Table 2. Estimates of Station Effects

Station number	Estimated station effect	Station number	Estimated station effect
5	0.51927	121	0.00705
9	–0.13458	127	–0.50767
12	–0.20786	139	0.38967
16	–0.11585	140	–0.15646
17	0.39464	144	0.08511
31	–0.22390	157	–0.41255
36	–0.50676	164	0.39418
43	–0.03807	169	0.03561
50	0.20122	178	0.17004
62	–0.01974	183	0.18580
69	–0.55833	189	–0.28234
72	–0.45911	192	0.51406
80	–0.03184	196	0.24657
86	–0.54256	205	–0.05887
89	–0.16852	211	0.29031
95	–0.65649	214	0.50154
105	–0.69295	220	1.32953
106	–0.05593	221	0.92228
120	–0.46403		

Plots of residuals against the estimated values of z_{ij}'s are given in Figure 2 separately for each cruise. Figure 3 gives the Q-Q plots again separately for each cruise. Figure 4 gives both these plots for all the data. These plots appear to indicate that the model is adequate.

Figure 5 gives the relative (maximized) likelihood function for λ. The shape of this function is nearly normal. The maximum of this function is reached at $\lambda = 0.16$ (approximately). Likelihood intervals for λ can be constructed from this graph. In Figure 5 we have shown an interval consisting of values of λ for which the relative likelihood function exceeds 0.1.

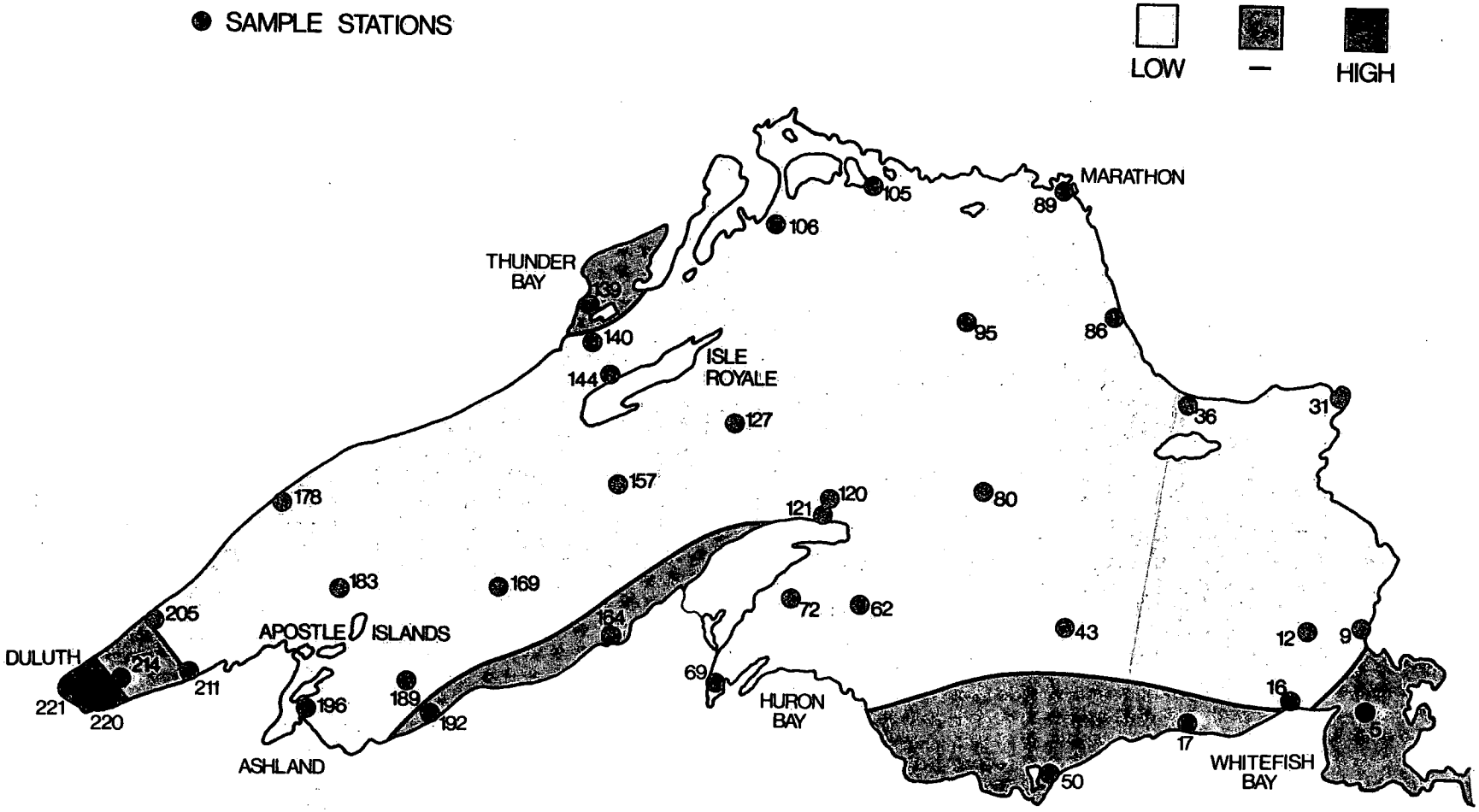


Figure 1. Statistical classification using total phytoplankton biomass as a parameter, 1973.

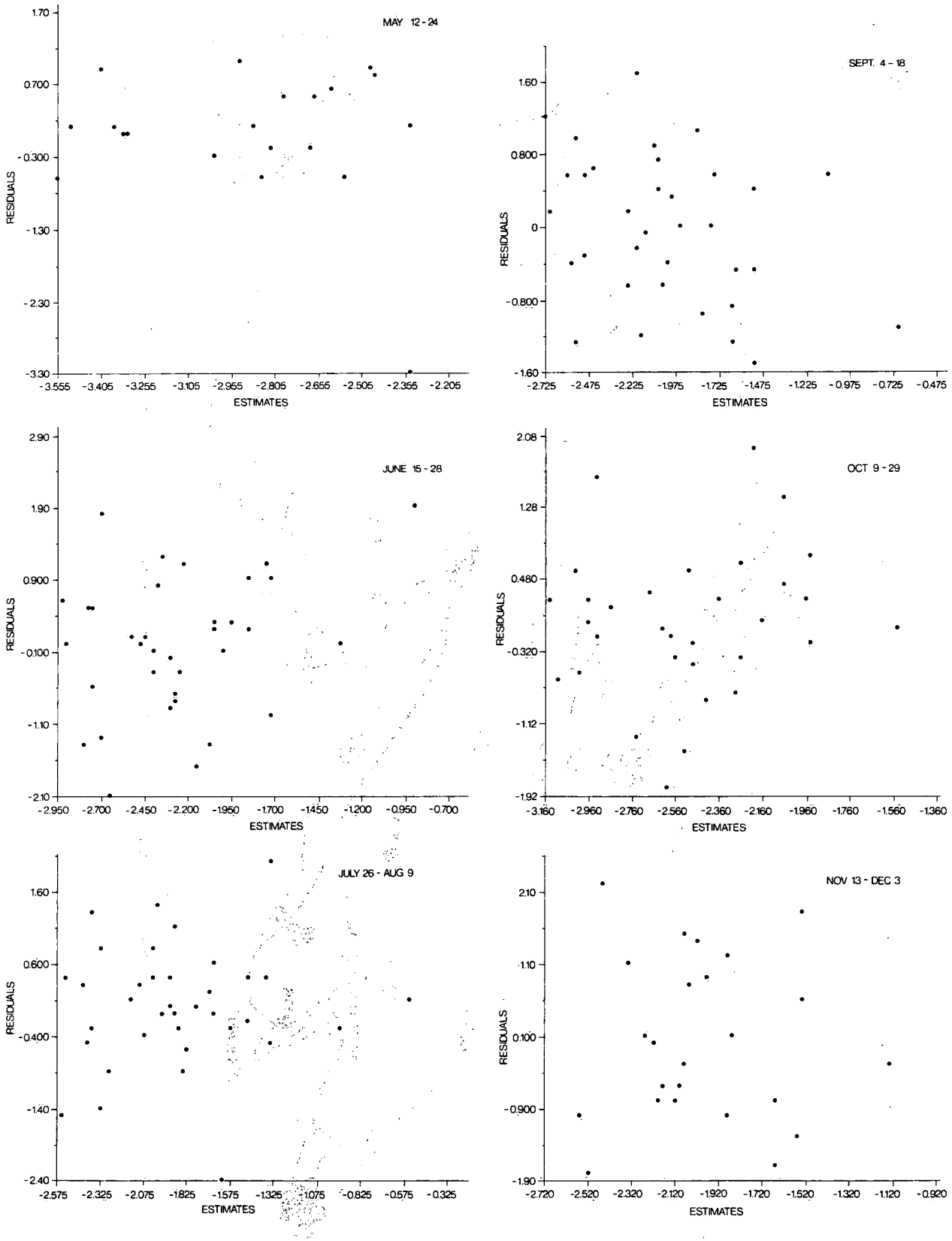


Figure 2. Residuals vs estimates of $E(z_{ij})$'s.

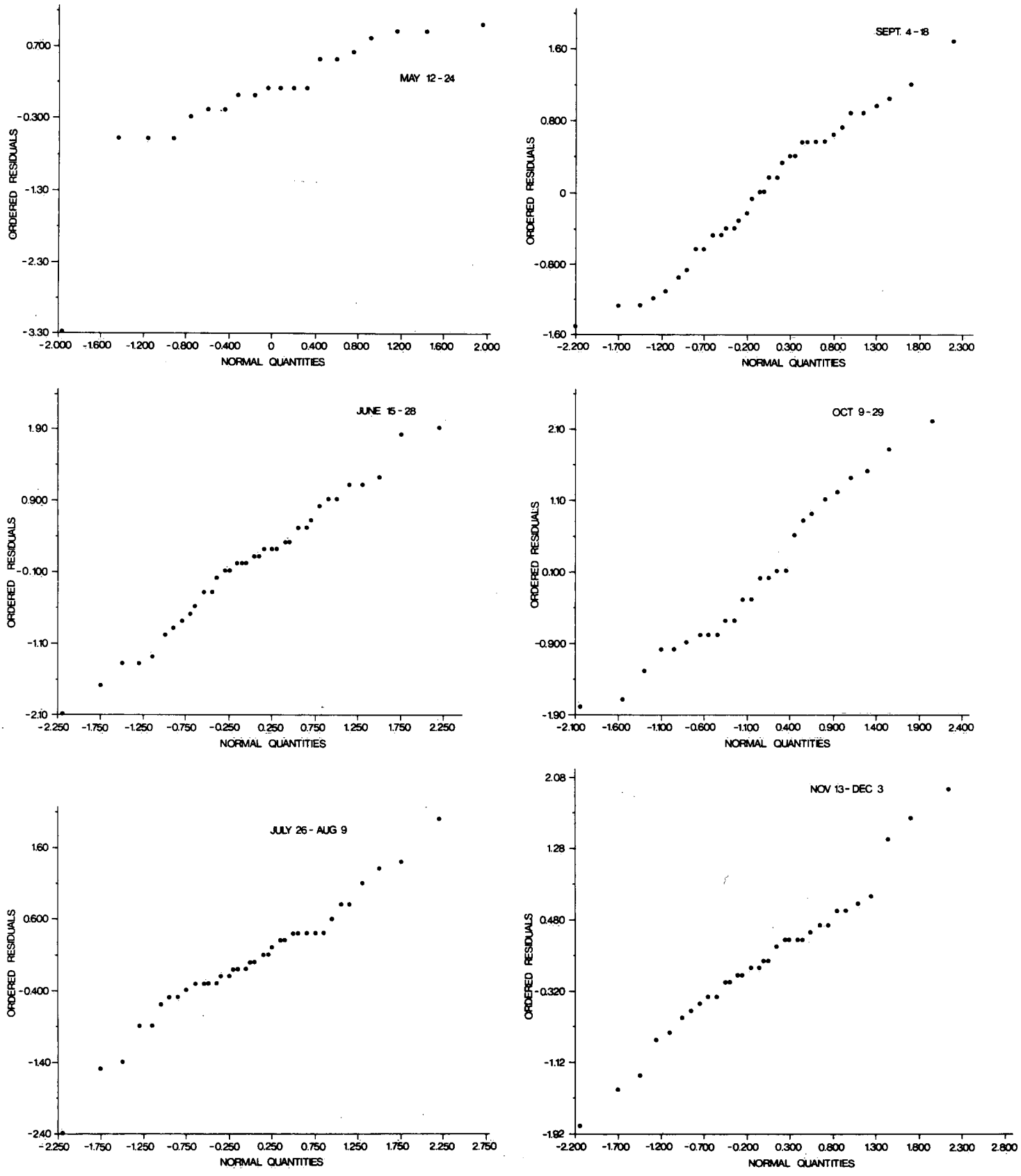


Figure 3. Q-Q plots for individual cruises.

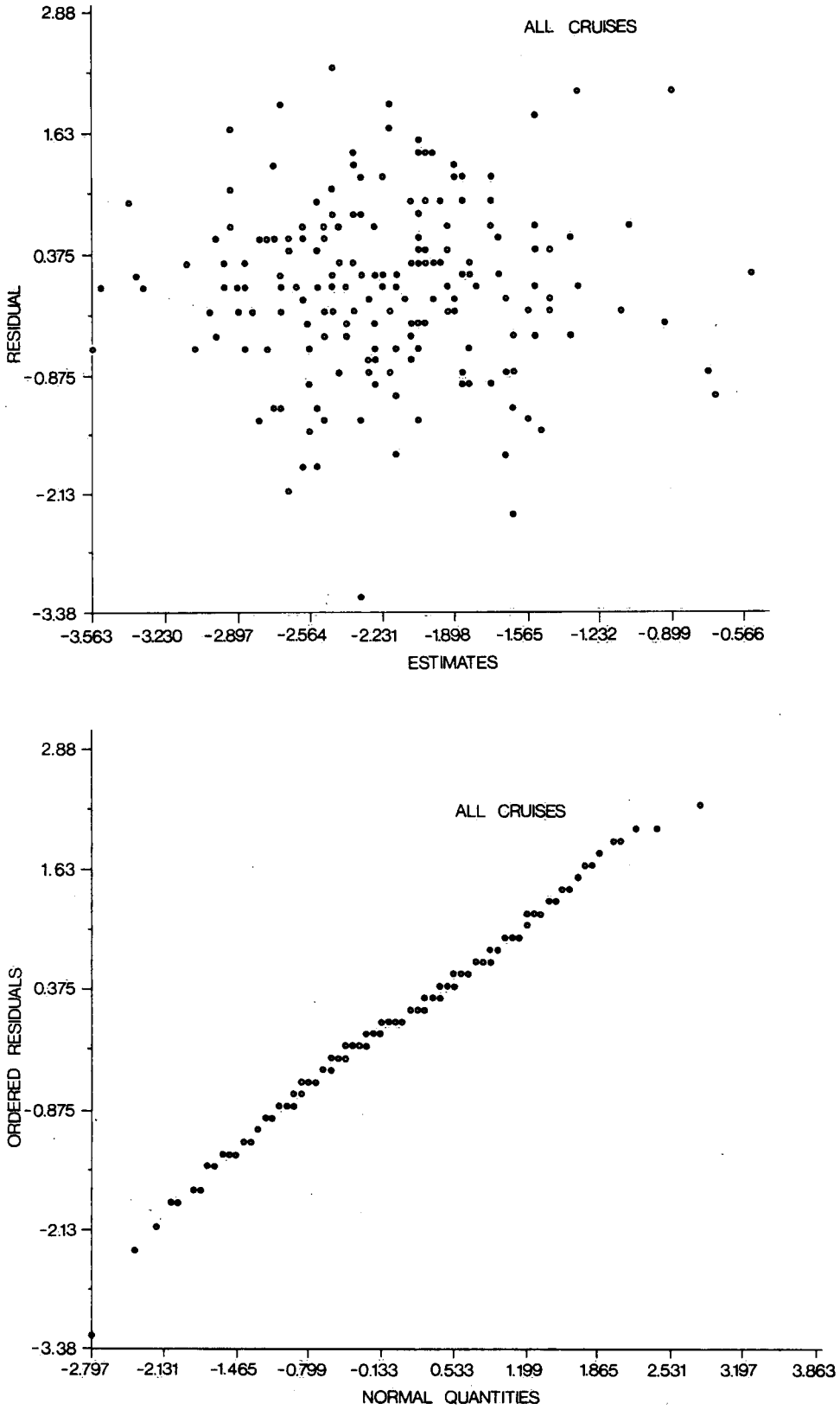


Figure 4. Q-Q plot and residuals vs estimates for the complete set.

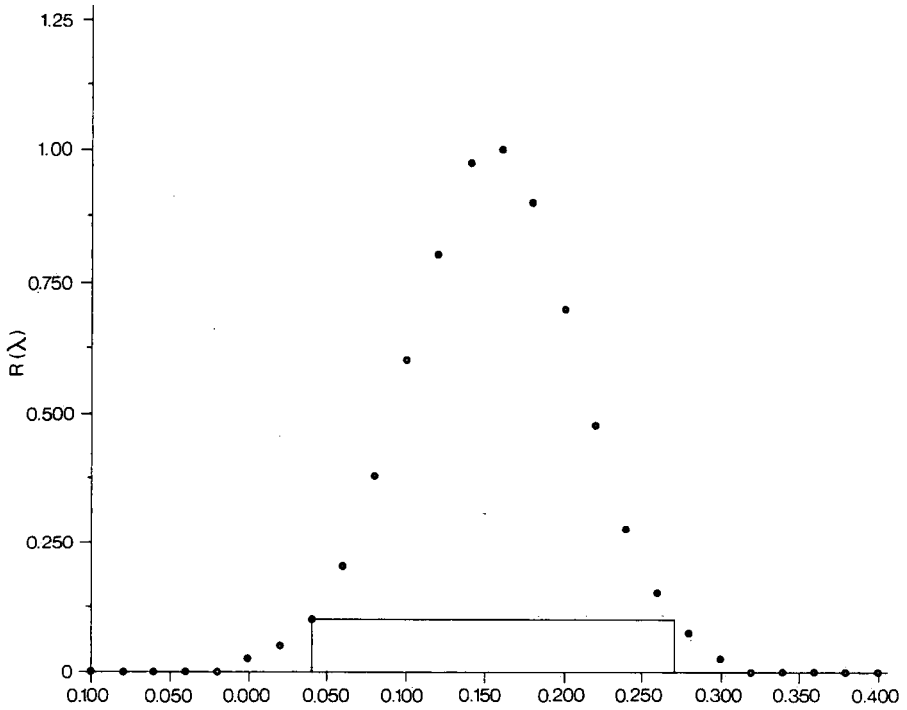


Figure 5. Plots of the relative likelihood.

ACKNOWLEDGMENT

The authors thank Mrs. A. Liu for carrying out the computations presented here.

REFERENCES

Box, G.E.P. and Cox, D.R., 1964. An analysis of transformations (with discussion). J.R. Stat. Soc., B, Vol. 26, 211-252.

El-Shaarawi, A.H., 1972. The statistical analysis of mortality rates. Ph.D. thesis, University of Waterloo, Waterloo, Ont.

Hartigan, J.A., 1975. Clustering algorithms. John Wiley & Sons, New York.

Kalbfleisch, J.D. and Sprott, D.A., 1970. Applications of likelihood methods to models involving a large number of parameters (with discussion). J.R. Stat. Soc., B, Vol. 32, 175-208.

Plackett, R.L., 1960. Principles of regression analysis. Oxford University Press.

Rao, C.R., 1965. Linear statistical inference and its applications. John Wiley and Sons, New York.

Taylor, L.R., 1961. Aggregation, variance and the mean. Nature (London), Vol. 189, 732-735.

Wilk, M.B. and Gnandesikan, R., 1968. Probability plotting methods for the analysis of data. Biometrika, Vol. 55, 1-17.