STATISTICAL ANALYSIS OF GEOCHEMICAL DATA ILLUSTRATED BY REFERENCE TO THE DALRADIAN OF N.E. SCOTLAND

by

MARK St. JOHN TURNER, B.Sc., M.Sc., DIC.

A thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

of the

UNIVERSITY OF LONDON

Applied Geochemistry Research Group

Royal School of Mines

Imperial College of Science and Technology

September 1986

ABSTRACT

The thesis undertakes the evaluation of several univariate and multivariate statistical techniques, some new to the analysis of geochemical data, which, in theory, have significant advantages over classical methods of analysis. This evaluation was achieved through an extensive comparative study of these techniques using a test set of stream sediment data from Norway (25 observations, 6 variables). The techniques were then applied to a large data set (503 observations, 24 variables) drawn from the BGS geochemical atlas database for the Moray-Buchan area of N.E. Scotland, with a view to investigating their value in a typical geochemical exploration context.

Robust statistics are equivalent to parametric statistics, but are designed to perform in a manner which is resistant to the distorting presence of outliers. Maximum likelihood, or M-estimates of standard univariate and multivariate parametric statistics perform well on geochemical data which display moderate to extreme positive skew.

The traditional logarithmic transformation, used to deskew or normalize frequency distributions, is inferior to the more flexible power transformation of Box and Cox

(1964). Parametric statistics generated using power transformed data are resistant to outliers, and are comparable with robust statistics generated using untransformed data. However, the process of data transformation causes the suppression of outliers, and the spread (or enhancement) of background observations, with the result that anomaly contrast is reduced. Thus, if anomaly detection is the aim of the analysis, any data transform should be avoided.

Trees are a recently developed graphical technique for the representation of multivariate data, based on a cluster analysis, and provide a useful exploratory means of viewing the data, and of making a preliminary determination of those samples which are likely to be of further interest.

Factor analysis proved superior to principal components analysis, and was used to identify and quantify variable associations present in the data. Robust estimation was incorporated into these techniques resulting in enhanced anomaly contrast.

Robust regression analysis is compared with iteratively reweighted least squares regression and proves to be theoretically and practically superior. A comparison of robust ridge estimation with ridge regression using an interactive iterative trimming procedure based on residuals proved that the former is more efficient and theoretically better

Acknowledgements

I should like to thank:

Dr. Richard Howarth

Dr. Martin Hale

Dr. Katherine Woodthorpe

Dr. Neil Coward

N.E.R.C.

I.M.M.

W.M.C.

List of Contents

Abstract

Acknowledgements

List of Contents

List of Figures

List of Tables

Glossary of Statistical Terms

CHAPTER ONE: INTRODUCTION

1.1 Aims of the research

1.2 Thesis layout

1.3 Previous work

CHAPTER TWO: GEOLOGY of the TEST DATA

2.1 The SIND data

2.1.1 The geology of the SIND data

2.2 The Moray-Buchan data

2.2.1 Sampling procedure

2.2.2 Regional geology

2.2.3 Depositional history

2.2.4 The Dalradian of the Buchan area, N.E. Scotland

2.2.5 Mineralization

2.2.6 The Aberfeldy Ba-Zn-Pb deposit

2.2.7 Exploration

CHAPTER THREE: UNIVARIATE STATISTICS and the ROLE of ROBUST ESTIMATORS

3.1 Introduction

3.2 Why robust statistics?

3.3 Robust estimators of location

3.3.1 L-estimators

3.3.2 M-estimators

3.3.3 Auxiliary scale estimators

3.3.4 Examples of location estimators

3.3.5 W-estimators

3.3.6 Summary

3.4 Robust estimators of scale

3.4.1 W-estimators of scale

3.4.2 Summary

3.5 Discussion

3.6 Robust estimation and the SIND data

3.7 Robust estimation and the Moray-Buchan data

3.8 Conclusions

CHAPTER FOUR: PROBABILITY PLOTS

4.1 Introduction

4.2 Univariate probability plots

4.3 Univariate probability plots and the SIND data

4.4 Univariate probability plots and the Moray-Buchan data

4.5 Multivariate probability plots

4.6 Multivariate probability plots and the SIND data

4.7 Multivariate probability plots and the Moray-Buchan data

4.8 Conclusions

CHAPTER FIVE: DATA TRANSFORMATIONS

5.1 Introduction

5.2 Data transformations in geochemistry

5.3 Power transformations

5.4 Program LAMT

5.5 The geometry of transformations

5.6 Data transformations and the SIND data

5.7 Data transformations and the Moray-Buchan data

5.8 Conclusions

CHAPTER SIX: COVARIANCE and CORRELATION

6.1 Introduction

6.2 Covariance and correlation matrices

6.3 Robust statistics

6.4 Robust correlation

6.4.1 Pairwise estimates

6.4.2 Multivariate estimates

6.4.2.1 Multivariate trimming

6.4.2.2 M-estimates

6.5 Discussion

6.6 Correlation and the SIND data

6.7 Correlation and the Moray-Buchan data

6.8 Conclusions

CHAPTER SEVEN: GRAPHICAL REPRESENTATION of MULTIVARIATE DATA

7.1 Introduction

7.2 Displays for multivariate data

7.3 Symbolic plots

7.4 Trees

7.5 Clustering

7.6 Faces

7.7 Trees, faces and the SIND data

7.7.1 SIND Trees

7.7.2 Chernoff Faces and the SIND data

7.8 Trees and the Moray-Buchan data

7.9 Conclusions

CHAPTER EIGHT: PRINCIPAL COMPONENT TECHNIQUES

8.1 Introduction

8.2 FA and PCA

8.3 Standardization and the similarity matrix

8.4 Factor analysis

8.5 Uses of FA and PCA

8.6 Summary of the differences between PCA and FA

8.7 Rotation

8.8 Component and factor scores

8.9 Robust PCA

8.10 PCA/FA and the SIND data

8.10.1 Data transformation

8.10.2 Robust estimation of the correlation matrix

8.10.3 Rotation

8.10.4 PCA or FA - practice

8.10.5 SIND data: conclusions

8.11 PCA/FA and the Moray-Buchan data

8.11.1 PCA - logten data

8.11.2 PCA - raw data

8.11.3 FA - logten data

8.11.4 FA - raw data

8.11.5 Moray-Buchan data: conclusions

CHAPTER NINE: MULTIPLE REGRESSION ANALYSIS

9.1 Introduction

9.2 The regression model

9.3 Standardization

9.4 Ridge regression

9.5 Robust regression

9.6 Regression and the SIND data

9.6.1 Robust versus non-robust regression

9.6.2 Transformation and regression

9.7 Robust ridge regression and the Moray-Buchan data

9.8 Conclusions

CHAPTER TEN: SUMMARY

10.1 Summary

10.2 Recommendations for further work

10.3 Conclusions

APPENDICES

A.1 Back-transformation of statistical parameters obtained from log-transformed and power-transformed data

A.2 Moray-Buchan anomalous samples detected using TREES analysis

A.3 Summary of Fortran programs used back pocket

Bibliography

List of Figures

2.1 SIND area

2.2 Moray-Buchan area showing study area and summary geology (in pocket in back cover)

2.3 Mineralization occurrences in the Scottish Dalradian

3.1 ρ and ψ functions for sample mean

3.2 ρ and ψ functions for sample median

3.3 Influence function for mean example

3.4 ρ and ψ functions for Huber

3.5 ρ and ψ functions for Hampel

3.6 ρ and ψ functions for Andrews

3.7 ρ and ψ functions for biweight

4.1 Four schematic Q-Q plots

4.2-3 Q-Q plots of SIND raw variables

4.4-9 Q-Q plots of MB raw variables

4.10-33 Point symbol maps of MB raw variables

4.34 Diagram showing relation between Mahalanobis distance and Euclidean distance

4.35 Q-Q plot $D_m^{2/3}$ SIND raw data

4.36 Q-Q plot robust $D_m^{2/3}$ SIND raw data

4.37 Q-Q plot $D_m^{2/3}$ MB raw data

4.38 Q-Q plot robust $D_m^{2/3}$ MB raw data

4.39 Q-Q plot 503 numbers generated randomly from a Normal distribution

4.40 Point symbol map robust $D_m^{2/3}$ MB raw data

5.1 Fe vs V $D_m^{2/3}$ contour plot

5.2-7 Q-Q plots of SIND raw, logten, boxcox variables

5.8-9 Q-Q plots non-robust and robust $D_m^{2/3}$ SIND raw, logten, boxcox data

5.10-33 Q-Q plots MB raw, logten, boxcox variables

5.34-35 Q-Q plots non-robust and robust $D_m^{2/3}$ MB raw, logten, boxcox data

5.36-37 Point symbol map robust $D_m^{2/3}$ MB logten, boxcox data

6.1 Six schematic scatter plots

6.2 Diagrams showing the effect of outliers on the correlation coefficient

6.3 correlation linkage diagram (non-robust) MB raw data

6.4 correlation linkage diagram (robust) MB raw data

6.5 correlation linkage diagram (non-robust) MB logten data

6.6 correlation linkage diagram (robust) MB logten data

6.7 correlation linkage diagram (non-robust) MB boxcox data

6.8 correlation linkage diagram (robust) MB boxcox data

7.1-5 Trees (non-robust) SIND raw data

7.6-10 Trees (robust) SIND raw data

7.11-15 Trees (non-robust) SIND logten data

7.16-20 Trees (robust) SIND logten data

7.21-25 Trees (non-robust) SIND boxcox data

7.26-30 Trees (robust) SIND boxcox data

7.31-32 Chernoff Faces emphasis on true anomalies SIND raw data

7.33-34 Chernoff Faces emphasis on false anomalies SIND raw data

7.35-36 Chernoff Faces SIND logten data

7.37-38 Chernoff Faces SIND boxcox data

7.39-74 Trees (non-robust) MB logten data

7.75-110 Trees (robust) MB raw data

8.1 Normalized variable vectors in object space

8.2 Eigenvectors and normalized variable vectors in object space

8.3 Eigenvectors in variate space

8.4 Eigenvectors in object space showing component scores

8.5 Diagram showing the effect of an outlier on eigenanalysis

8.6-7 Q-Q plots PCA scores (non-robust rotated) SIND raw data

8.8-9 Q-Q plots PCA scores (non-robust rotated) SIND logten data

8.10-11 Q-Q plots PCA scores (non-robust rotated) SIND boxcox data

8.12-13 Q-Q plots PCA scores (robust rotated) SIND raw data

8.14-15 Q-Q plots PCA scores (robust rotated) SIND logten data

8.16-17 Q-Q plots PCA scores (robust rotated) SIND boxcox data

8.18 Q-Q plots PC1 scores (non-robust non-rotated) SIND raw, logten, boxcox data

8.19 Q-Q plots PC2 scores (non-robust non-rotated) SIND raw, logten, boxcox data

8.20 Q-Q plots non-robust/robust Factor scores SIND raw data

8.21 Q-Q plots non-robust/robust Factor scores SIND logten data

8.22 Q-Q plots non-robust/robust Factor scores SIND boxcox data

8.23 Q-Q plots PC1 PC2 scores (non-robust rotated) MB logten data

8.24-25 Point symbol maps PC1 PC2 scores (non-robust rotated) MB logten data

8.26 Q-Q plots PC1 PC2 scores (robust rotated) MB raw data

8.27-28 Point symbol maps PC1 PC2 scores (robust rotated) MB raw data

8.29-30 Q-Q plots F1-F6 scores (non-robust rotated) MB logten data

8.31-37 Point symbol maps F1-F6 scores (non-robust rotated) MB logten data

8.38-39 Q-Q plots F1-F7 scores (robust rotated) MB raw data

8.40-46 Point symbol maps F1-F7 scores (robust rotated) MB raw data

9.1 Q-Q plots non-robust/robust regression residuals SIND raw data

9.2 Q-Q plots non-robust/robust regression residuals SIND logten data

9.3 Q-Q plots non-robust/robust regression residuals SIND boxcox data

9.4 Q-Q plots non-robust/robust Ba regression residuals MB raw data

9.5 Q-Q plots non-robust/robust Zn regression residuals MB raw data

9.6 Q-Q plots non-robust/robust Pb regression residuals MB raw data

9.7 Q-Q plots non-robust/robust Ba regression residuals MB logten data

9.8 Q-Q plots non-robust/robust Zn regression residuals MB logten data

9.9 Q-Q plots non-robust/robust Pb regression residuals MB logten data

9.10-11 Point symbol maps non-robust/robust Ba regression residuals MB raw data

9.12-13 Point symbol maps non-robust/robust Zn regression residuals MB raw data

9.14-15 Point symbol maps non-robust/robust Pb regression residuals MB raw data

9.16-17 Point symbol maps non-robust/robust Ba regression residuals MB logten data

9.18-19 Point symbol maps non-robust/robust Zn regression residuals MB logten data

9.20-21 Point symbol maps non-robust/robust Pb regression residuals MB logten data

List of Tables

2.1 Raw SIND data matrix

2.2 Stratigraphy of Dalradian Supergroup

3.1 Table of ψ and ω functions

3.2 Ordered SIND raw data

3.3 Absolute differences (x_i - Median) SIND raw data

3.4 Classical and robust location and scale estimates SIND raw data

3.5 Classical and robust location and scale estimates MB raw data

5.1 SIND logten data matrix

5.2 SIND boxcox data matrix

5.3 Chi-squared SIND raw, logten, boxcox data

5.4 Non-robust/robust location and scale estimates SIND raw, logten, boxcox data

5.5 Chi-squared MB raw, logten, boxcox data

5.6 Non-robust/robust location and scale estimates MB raw, logten, boxcox data

6.1 Breakdown points

6.2 Non-robust correlation matrices SIND raw, logten, boxcox data

6.3 Non-robust/robust correlation matrices SIND raw data

6.4 Non-robust/robust correlation matrices SIND logten data

6.5 Non-robust/robust correlation matrices SIND boxcox data

6.6 Robust correlation matrices SIND raw, logten, boxcox data

7.1 Agglomerative clustering recurrence formulae

8.1 Component loadings matrices (non-robust rotated) SIND raw, logten, boxcox data

8.2 Non-robust/robust component loadings matrices SIND raw data

8.3 Non-robust/robust component loadings matrices SIND logten data

8.4 Non-robust/robust component loadings matrices SIND boxcox data

8.5 Component loadings matrices (non-robust non-rotated) SIND raw, logten, boxcox data

8.6 Factor loadings matrices (non-robust rotated) SIND raw, logten, boxcox data

8.7 Non-robust/robust factor loadings matrices SIND raw data

8.8 Non-robust/robust factor loadings matrices SIND logten data

8.9 Non-robust/robust factor loadings matrices SIND boxcox data

Glossary of Statistical Terms

These definitions have been derived from: Kendall and Buckland's "Dictionary of Statistical Terms", 1982.

ADDITIVE MODEL

A model in which the factors influencing the response variable have an additive effect, e.g. a linear regression equation is additive.

ASYMMETRIC DISTRIBUTION

A distribution is not symmetrical, that is to say, for which there is no central value a such that f(x-a) = f(a-x), f(x) being the frequency function.

BIAS

Generally, an effect which deprives a statistical result of representativeness by systematically distorting it, as distinct from a random error which may distort on any one occasion, but balances out on the average.

BREAKDOWN BOUND

One of a number of similar measures used to compare the robustness of estimators in the presence of extreme

outliers. The bound gives an indication of the largest proportion of outliers that an estimator may be expected to tolerate before distortion and instability sets in.

CHI-SQUARED STATISTIC

This expression relates to a statistic of a particular kind which is, as a general rule, distributed more or less approximately in the χ^2 form. If a set of n values is distributed over k classes such that the observed frequency in the jth class is n_j and the theoretical (expected) frequency in that class is v_j, the statistic

$$\frac{\Sigma\ (\ n_j - v_j\ )^2}{v_j}$$

is called the χ^2 statistic, or simply the value of χ^2 , for the data. It is widely used to test agreement between observations, as represented by the n_j, and hypotheses as represented by the v_j.

CHI-SQUARED TEST

A test of significance based upon the chi-squared statistic. Such tests occur in many ways, the most prominant being:

1) An overall goodness-of-fit comparison of observed with hypothetical frequencies falling into specified classes.

2) Comparison of an observed with a hypothetical variance in Normal samples.

CLUSTER ANALYSIS

A general approach to multivariate problems in which the aim is to see whether the individuals fall into groups or clusters. There are several methods of procedure but they all depend on setting up a metric to define the "closeness" of individuals.

CORRELATION MATRIX

For a set of variates $x_1,\ldots,x_n$ with correlations between x_i and x_j denoted by r_{ij}, the correlation matrix is the square matrix of values (r_{ij}). The matrix is symmetrical since $r_{ij} = r_{ji}$. Unless otherwise specified the diagonal elements $r_{ii}(i = 1,\ldots,n)$ are unity.

COVARIANCE

The first product moment of two variates about their mean values (see Section 6.1).

COVARIANCE MATRIX

For n variates $x_1,\ldots,x_n$, for which the covariance of x_i and x_j is c_{ij}, the square matrix (c_{ij}) is called the covariance matrix. The diagonal terms are called the variances: var x_i. Alternative names are: dispersion matrix, variance-covariance matrix. If the variates are

standardized so as to have unit variances, the covariances become correlations and the matrix becomes the correlation matrix.

CIRCULAR CHART

A method of diagrammatic representation whereby the components of a single total can be shown as sectors of a circle. The angles of the sectors are proportional to the components of the total. Additional visual aid can be obtained with colour shading or cross-hatching.

DEGREES OF FREEDOM

The degrees of freedom of a set of observations is the number of values which could be assigned arbitrarily within the specification of a system. For example, in a sample of constant size n grouped into k intervals there are k-1 degrees of freedom because, if k-1 frequences are specified, the other is determined by the total size n; and in a contingency table of p rows and q columns with fixed marginal totals there are (p-1)(q-1) d.f.

DENDOGRAM

A diagram of a hierarchical kind exhibiting the relation of subsets of a structure; so called because the diagram branches like a tree.

DISCRIMINANT ANALYSIS

Given a set of multivariate observations on samples, known with certainty to come from two or more populations, the problem is to set up some rule which will allocate further individuals to the correct population of origin with minimal probability of misclassification. This problem and elaborations of it give rise to discrminant analysis.

DISTRIBUTION CURVE or CUMULATIVE DISTRIBUTION CURVE

The graph of cumulated frequency as ordinate against the variate value as abscissa, namely the graph of the distribution function.

DISTRIBUTION FUNCTION

The distribution function f(x) of a variate x is the total frequency of members with variate values less than of equal to x. As a general rule the total frequency is taken to be unity, in which case the distribution function is the proportion of members bearing values $< x$.

EFFICIENCY

The concept of efficiency in statistical estimation is an attempt to measure objectively the relative merits of several possible estimators. The criterion is that

of variance, an estimator being regarded as more 'efficient' than another if it has smaller variance; and if there exists an estimator with minimum variance v the efficiency of another estimator of variance v_1 is defined as the ratio v/v_1.

EIGENVALUE

The eigenvalue of a square matrix A is a value λ such that $| A - \lambda I | = 0$, where I is the identity matrix. For a p x p matrix there are, in general, p such eigenvalues. They are also known as latent roots or characteristic roots.

The corresponding row vectors U or column vectors V for which

$$UA = \lambda U \text{ or } AV = \lambda V$$

are called eigenvectors or characteristic vectors.

ESTIMATE

In the strict sense an estimate is the particular value yielded by an estimator in a given set of circumstances. The expression is, however, widely used to denote the rule by which such particular values are calculated. It seems preferable to use the word 'estimator' for the rule of procedure and 'estimate' for the values to which it leads in particular cases.

ESTIMATOR

An estimator is a rule or method of estimating a constant of a parent population. It is usually expressed as a function of sample values and hence is a variate whose distribution is of great importance in assessing the reliability of the estimate to which it leads.

EXPLANATORY VARIABLE

When a relation such as $y = f(x)$ is interprated in a casual sense, e.g. y is regarded as "caused" by x, the latter is sometimes called a cause variable and the former the effect variable. The cause variable is also known as the explanatory variable. The effect variable is also known as the response variable.

FISHER'S TRANSFORMATION

A transformation of the correlation coefficient r according to the formula

$$z = \tanh^{-1} r$$

The distribution of z for samples from a bivariate Normal population approaches normality of form much more rapidly than does that of r; and its variance, unlike that of r, is little affected by the population

correlation coefficient even for samples of moderate size. It is therefore useful in many problems involving estimation and tests significance.

FOURIER ANALYSIS

The theory of representing functions of a variable t as the sum of a series of sine and cosine terms of type:

$$a_j \cos (2 \pi_j / \lambda_j), \; j = 0, 1, \ldots, n.$$

The analysis is more general than harmonic analysis which considers series of terms such as $\cos (2 \pi_j / \lambda)$

where λ is some constant.

FREQUENCY CURVE

The graphical representation of a continuous frequency distribution, the variate being the abscissa and the frequency the ordinate. The frequency curve may be viewed as the limiting form of the frequency polygon as the number of observations tends to become infinitely large and the class intervals indefinitely small.

FREQUENCY FUNCTION

An expression giving the frequency of a variate value x as a function of x; or, for continuous variates, the frequency in an elemental range dx. Unless the contrary

is specified the total frequency is taken to be unity, so that the frequency function represents the proportion of variate values x. The frequency function is most conveniently regarded as the derivative of the distribution function.

HISTOGRAM

A univariate frequency diagram in which rectangles proportional in area to the class frequencies are erected on sections representing the corresponding class interval of variate.

INFLUENCE CURVE

The basis of an array of tools for assessing the influence on a given estimator of a contaminated observation.

INTERQUARTILE RANGE

The variate distance between the upper and lower quartiles. This range contains one half of the total frequency and provides a simple measure of dispersion which is useful in descriptive statistics.

KURTOSIS

A term used to describe, the extent to which a unimodal frequency curve is 'peaked'; that is to say, the extent of the relative steepness of ascent in the

neighbourhood of the mode. The term was introduced by Karl Pearson in 1906; and he proposed as a measure of kurtosis the moment ratio β_2 (= μ_4 / μ_2^2).

If a moment ratio is adapted as a measure of kurtosis, the value it assumes for a Normal distribution, namely 3, is taken as standard. Curves for which the ratio is less than, equal to or greater than 3 are known respectively as platykurtic, mesokurtic and leptokurtic.

LEAST SQUARES ESTIMATOR

An estimator obtained by the least squares method.

LEAST SQUARES METHOD

A technique of estimation by which the quantities under estimate are determined by minimising a certain quadratic form in the observation and those quantities.

In certain restricted cases it has demonstrably optimum properties:

1) where linear unbiased estimators with minimum variance are sought.

2) Where the model involves errors which are normally distributed and least squares estimation becomes equivalent to maximum likelihood estimation.

3) Where residuals are autocorrelated and the method provides estimates that are linear, unbiased, consistent and distribution free.

LINEAR MODEL

A model in which the equations connecting the variates or variables are in a linear form.

LOCATION PARAMETER

A parameter which 'locates' a frequency distribution in the sense of defining a central or typical value such as a mean, median, or mode.

LOGNORMAL DISTRIBUTION

If the logarithms of a set of variate values are distributed according to the Normal distribution the variate is said to have a logarithmic Normal distribution, or be distributed 'lognormally'.

MAIN EFFECT

An estimate of the effect of an experimental variable or treatment measured independently of other treatments which may form part of the experiment. Thus in a balanced experiment involving 3 factors A B C each at two levels, applied and not applied, the main effect due to A would be the average of the four effects where A was applied, and B and C were applied at each of the two levels, less the

average of the four effects where A was not applied.

MEAN or ARITHMETIC MEAN

The arithmetic mean of a set of values $x_1, x_2, \ldots, x_n$ is their sum divided by their number, namely:

$$1/n \; \Sigma \; x_i$$

It is often denoted by a bar, e.g. $\bar{x}$.

MEDIAN

The median is that value of the variate which divides the total frequency into two halves. As a partition value it may be defined for a continuous frequency distribution by the equation:

$$\int_{-\infty}^{M} f(x)dx = \int_{M}^{\infty} f(x)dx = 1/2$$

M being the median value. For a discontinuous variate of total frequency 2n+1 items, the median is the variate value of the (n+1)-th item: for 2n items it is customary to take the average of the n-th and (n+1)-th item.

M-ESTIMATE OF LOCATION

For an ordered sample $x_{(1)} \leq x_{(2)} \leq \ldots \; x_{(n)}$ from a symmetric distribution with scale parameter σ, the solution T of an equation of the form:

$$\frac{\Sigma\ \psi(x_{(i)} - T)}{s} = 0$$

where ψ is an odd function and s, an estimate of σ, is found independently.

MODEL

A model is a formalised expression of a theory or the causal situation which is regarded as having generated observed data. In statistical analysis the model is generally expressed in symbols, that is to say in a mathematical form, but diagrammatic models are also found.

MOMENTS

If x_{ij} is the i-th of n observations on variable j then the mean of x values is:

$$\bar{x} = 1/n\ \Sigma\ x_{ij}$$

The n-th moment (μ_n) of j is:

$$\mu_n = 1/n\ \Sigma\ (x_{ij} - \bar{x}_j)^n$$

The variance of j is given by:

$$s^2 = \mu_2 = 1/n\ \Sigma\ (x_{ij} - \bar{x}_j)^2$$

The standard deviation is given by:

$$s = \mu_2^{1/2}.$$

MONTE CARLO METHOD

A term which has been used with several different meanings:

1) To denote the approximate solution of distributional problems by sampling experiments.

2) To denote the solution of any mathematical problem by sampling methods; the procedure is to construct an artificial stochastic model of the mathematical process and then to perform sampling experiments upon it.

MULTICOLLINEARITY

In regression analysis, a situation in which there exists a linear relation connecting the explanatory variables.

NORMALIZING TRANSFORM

Tranformations of variables with the object of reducing the distribution to the Normal or very closely so. It may be noted that a transformation to achieve, say,

stabilization of variance may also yield a large measure of normalization. Usually, this expression denotes a transformation of a random variable so that its distribution approximates, or is exactly equal to the Normal (Gaussian) form. It also occurs in the sense of a transformation which reduces the distribution to some other standard form which is regarded as the norm.

OUTLIERS

In a sample of n observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are not from a different population, or that the sampling technique is at fault. Such values are called outliers.

PERCENTILES

The set of partition values which divide the total frequency into one hunderd equal parts.

POPULATION

In statistical usage the term population is applied to any finite or infinite collection of individuals.

PRINCIPAL COMPONENTS

If each member of a population bears the values of p variates 1,..., p it is, in general, possible to find a linear transformation to p new variates P_1,..., P_p which a)

are independent and b) account in turn for as much of the variation as possible in the sense that the variance of P_1 is a maximum among all linearly transformed variates; the variance of P_2 is a maximum among all linearly transformed variates orthogonal to P_1 and so on. Such variates are called principal components.

PROBABILITY DISTRIBUTION

A distributon giving the probability of a value x as a function of x. It is customary to use 'probability distribution' to denote the probability mass or probability density of either a discontinuous or continuous variate and some such expression as 'cumulative probability distribution' to denote the probability of values up to and inculding the argument x. From a frequency viewpoint the distinction is the same as beween 'frequency function' and 'distribution function'.

QUANTILE

The class of (n-1) partition values of a variate which divide the total frequency of a population or a sample into a given number of n equal proportions. For example, if n = 4 then the n-1 values are the quartiles although the central variate value is generaly termed the Median.

QUARTILE

There are three variate values which separate the total frequency of distribution into four equal parts. The central value is called the median and the other two th upper and lower quartiles respectively. They are a particular set of quantiles.

RESIDUAL

A general term denoting a quantity remaining after some other quantity has been subtracted. It occurs in a variety of particulr contexts. For example, if the true value of a variable is subtracted from an observed value then the difference may be called a residual; it is also frequently called an error. Similarly, if a mathematical model is fitted to data, the values by which the observations differ from the model values are called residuals.

RIDGE REGRESSION

A method proposed by Hoerl and Kennard (1970) to overcome shortcomings of standard least squares methods of estimating regression coefficients arising from the more or less close association btween variates. While the least squares estimates are minimum variance unbiased, they are not in general minimum mean square error (MSE). The ridge estimators are biased but with smaller MSE: the approach is essentially one of choosing a constant (which is added to

each diagonal element of the correlation matrix) to achieve a satisfactory balance between bias and variance.

ROBUSTNESS

Many test procedures involving probability levels depend for their exactitude on assumptions concerning the generating mechanism, e.g. that the parent variation is Normal (Gaussian). If the inferences are little affected by departure from those assumptions, e.g. if the significance points of a test vary little if the population departs quite substantially from the Normality, the test on the inferences are said to be robust. In a rather more general sense, a statistical procedure is described as robust if it is not very sensitive to departure from the assuptions on which it depends.

ROSE DIAGRAM

A form of diagrammatic representation for grouped angular data. Corresponding to each angular interval a sector is constructed with apex at the origin, radius proportional to the class frequency and arc subtending the interval.

SAMPLE

A part of a population, or a subset from a set of units, which is provided by some process or other, usually by

deliberate selection with the object of investigating the properties of the parent population or set.

SCALE PARAMETER

A parameter of a frequency distribution which is functionally related to the scale of the variable e.g. the standard derivation in a Normal distribution.

SKEW DISTRIBUTION

A distribution which is not symmetrical; a distribution for which a measure of skewness has some value other than zero.

SKEWNESS

An older and less preferable term for asymmetry, in relation to a frequency distribution; a measure of that asymmetry. The concept of asymmetry is easily defined, a measure of asymmetry less easily so. If a unimodal distribution has a longer tail extending towards lower values of the variate it is said to have negative skewness; in the contrary case, positive skewness.

Skewness may be measured by a moment value in the same fashion as kurtosis:

$$\text{Skewness} = \frac{(\mu_3)^2}{(\mu_2)^3}$$

STANDARD DEVIATION

The most widely used measure of dispersion or scale of a frquency distribution introduced by K Pearson (1893). It is equal to the positive square root of the variance.

STANDARD MEASURE

If x is a variate with mean μ and standard deviation σ the transformed variate $y = (x - \mu) / \sigma$ is said to be in standard measure or standardized. It has zero mean and unit variance.

STOCHASTIC

The adjective 'stochastic' implies the presence of a random variable; e.g. stochastic variation is variation in which at least one of the elements is a variate and a stochastic process is one wherein the system incorporates an element of randomness as opposed to a deterministic system.

SYMMETRIC DISTRIBUTION

A frequency distribution for which the variate values equidistant form a central value are equally frequent. In a symmetrical distribution all odd order moments about the mean are zero.

TRIMMING

A procedure suggested by Tukey (1962), for the rejection of observations, for removing equal numbers of lowest and highest observations in order to improve estimation of a location parameter from a simple random sample. The motive is to guard against outliers or a long tailed distribution.

UNBIASSED ESTIMATOR

An estimator, say t, which for all sample sizes, has its expected value equal to the parameter, say θ, under estimate, i.e. $E(t) = 0$. For example, the sample variance is not an unbiased estimator of the population variance σ^2 : since

$$E[1/n\ \Sigma(x_i - \bar{x})^2] = \sigma^2\ (n-1)/n.$$

On the other hand the statistic $[1/(n-1)]\ \Sigma(x_i - \bar{x})^2$ is an unbiassed estimator of the population variance since:

$$E[(1/(n-1))\ \Sigma(x_i - \bar{x})^2] = \sigma^2\ .$$

An estimator which is not unbiassed is called biassed. If it is unbiassed for all distributions, as in the case of the estimator for the variance given above, it is absolutely unbiassed. If it is unbiassed and linear in the observations it is an unbiassed linear estimator.

If the estimator tends to be unbiassed as the sample number increases it is asymptotically unbiassed. The first

estimator of the variance mentioned above, for example, is so.

VARIANCE

The variance is the second moment of a frequency distribution taken about the arithmetic mean as the origin. It is a quadratic mean in the sense that it is the mean of the squares of variations from the arithmetic mean (see entry under MOMENT).

CHAPTER ONE

INTRODUCTION

1.1 Aims of the Research

The research has two aims.

1) To evaluate some statistical techniques new to the analysis of geochemical data which, in theory, have significant advantages over classical methods of analysis.

2) To investigate the value of these techniques on a typical "exploration" geochemical database.

1.2 Thesis Layout

Each chapter describes the theoretical background to each statistical technique, followed by an extensive comparative analysis using a small set of test data. The lessons learnt from the comparative analysis are then

applied to a much larger and more complex set of data, typical of data sets generated in mineral reconnaissance surveys.

Chapter One contains the thesis outline, aims and a glossary of statistical terms.

Chapter Two contains a description of the geological setting of the locations from which the two sets of data used in the thesis are derived.

Chapter Three discusses the problem of the sensitivity of classical estimates of location and scale to the presence of outliers, and introduces the concept of robust estimation.

Chapter Four examines the role of probability plots as a visual tool for the analysis of both univariate and multivariate frequency distributions.

Chapter Four discusses the process of data transformation, the purpose of which is to obtain a frequency distribution that is near normal, and symmetric in shape, and well suited to the estimation of descriptive parameters. An alternative to the traditional logarithmic transformation is discussed and evaluated.

Chapter Six extends the discussion in Chapter Three to include the robust estimation of correlation coefficients and matrices.

Chapter Seven discusses the graphical representation of multivariate data and introduces "trees" - a new method, based on cluster analysis.

Chapter Eight describes and evaluates a new variation of Principal Components Analysis using a robust estimate of the variance-covariance matrix.

Chapter Nine describes a new variation to classical Multiple Linear Regression - that of Ridge Regression. Robust alternatives are described and evaluated.

Chapter Ten contains the final summary, conclusions and recommendations for further work.

1.3 Previous Work

Detailed references to particular techniques are given in the appropriate chapters. Those given here are more general, or span several techniques.

Davis (1973) presents an invaluable introductory text to statistics and data analysis for geologists. A more complete review of multivariate techniques is provided by Howarth and Sinding-Larsen (1983), in which they extensively analyse the "SIND" test set of data used in this thesis.

The above two references provide an excellent coverage of the classical approach to statistical analysis of geological data. Little or no mention is made of robust statistics, power transformation, graphical representation of multivariate data, or ridge regression. Howarth (1984),

in a review of recent developments in the statistical treatment of geochemical data, introduces these techniques in the context of an integrated program of exploratory data analysis. In his conclusion, Howarth comments that only extensive applications of these techniques will determine which are most useful, but a more robust approach to data analysis for the exploration geochemist cannot go amiss.

Hoaglin et al (1983) provide a general review of robust data analysis techniques from a purely statistical standpoint.

Mancey (1980) investigates the feasibility of some data analysis techniques when applied to the huge Wolfson geochemical atlas data set for England and Wales. His analysis shows that the data distributions tend to be positively skewed, but not necessarily log-normal. Of particular interest is his evaluation of the power transform, which he shows results in more normally distributed data than the traditional logarithmic transformation.

CHAPTER TWO

GEOLOGY OF TEST DATA AREAS

2.1 The SIND Data

Throughout the thesis the statistical techniques will be illustrated and compared using a set of real data drawn from a regional stream sediment geochemical survey undertaken by the Norwegian Geological Survey. The data consist of 25 samples, from a single, geologically homogeneous drainage basin, analysed for copper, lead, zinc, cadmium, iron and manganese (Table 2.1). The geographical location of these samples within the drainage basin is shown in Figure 2.1. The data set was selected because it represents a typical geochemical problem while involving a small enough number of samples to be conveniently manipulated. The samples from the drainage basin chosen had the highest Zn and Cu concentrations in the region. These could be attributable either to mineralization or to the results of element enhancement by Mn and Fe oxide scavenging.

Table 2.1

SIND RAW Data Matrix

(all values in ppm unless otherwise stated)

Sample No.	East	North	Zn	Fe (%,×100)	Mn	Cd (ppm,×10)	Cu	Pb
1	97	55	24	108	330	4	7	5
2	86	43	25	118	420	3	9	7
3	98	43	42	206	910	6	12	6
4	109	34	50	173	700	5	15	9
5	121	15	52	174	690	5	18	10
6	82	28	29	106	510	6	10	8
7	66	6	26	108	530	5	10	7
8	75	10	23	93	260	4	7	7
9	76	71	89	184	670	9	32	6
10	65	77	72	335	530	5	11	8
11	58	68	31	151	350	4	11	4
12	63	56	115	159	650	12	37	6
13	41	50	535	259	960	35	350	9
14	40	43	48	171	570	3	17	9
15	22	46	1010	271	1070	56	590	9
16	2	46	560	305	450	31	490	8
17	23	21	48	143	590	4	8	8
18	28	22	44	170	710	4	11	16
19	19	3	36	128	410	2	7	10
20	133	55	33	194	490	10	20	12
21	108	93	45	179	260	8	27	12
22	127	110	118	670	1930	12	33	17
23	140	120	274	610	5920	31	63	27
24	156	129	81	262	970	9	22	13
25	175	133	80	361	900	10	23	10

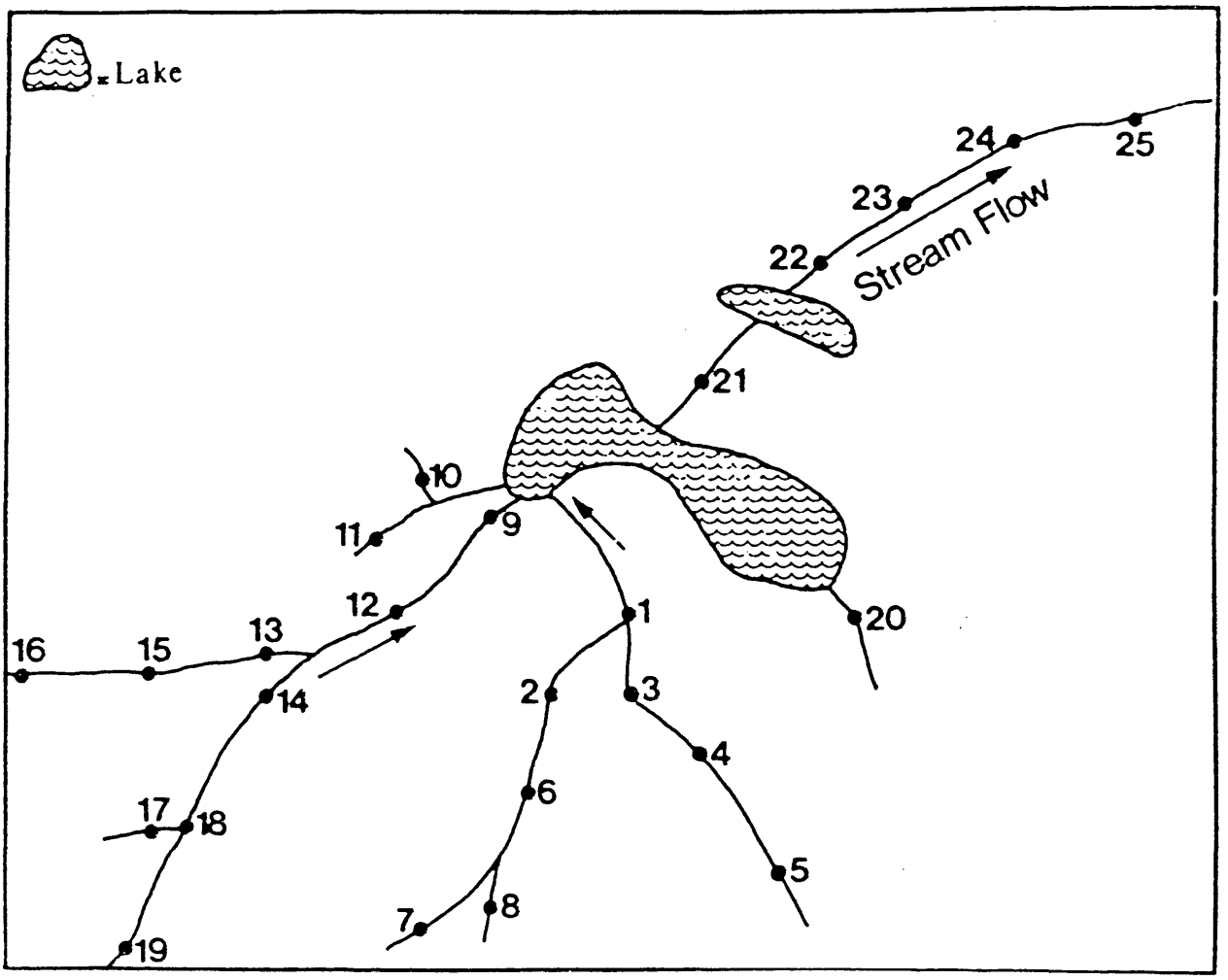


Figure 2.1
Sample location map for SIND data points.

The only information needed about the data is that there is a source of Zn-Cu-Cd mineralization upstream of sample number 16, that element scavenging by manganese and iron oxides is particularly dramatic downstream of the smaller of the two lakes and that the area is geologically homogeneous, i.e. no changes in element concentrations can be attributed to changes in lithology.

This data set has been similarly used as a test set of data by Howarth and Sinding-Larsen (1983) and by Garrett (1983). The data will be subsequently referred to as the SIND data.

2.1.1 The Geology of the SIND Area

The single drainage basin from which the SIND data are derived lies within the Central Norwegian Caledonides in the Grong district of Nord Trondelag. The major lithological unit of the Grong district is the Gjersvik Nappe, which consists of a thick sequence of greenstones overlain by sandy and conglomeritic flysch type rocks. The greenstones, chiefly of the island arc type, are mostly pillow lavas with associated lenses of sulphides.

Mineralization in the Grong district occurs in three main associations: volcanic, intrusive and sedimentary. The volcanic association is by far the most important. The

mineral occurrences are interpreted to be stratiform and syngenetic originating from volcanic exhalations and deposited as chemical sediments. A majority of the mineral occurrences are situated within the Gjersvik Group where they are generally closely associated with acidic, keratophyric tuffs and lavas, agglomerates and jasper or magnetitic chert (Kollung, 1979). The largest deposit in the Gjersvik Group is the Skorovas orebody which consists of approximately 10Mt of massive and disseminated pyritic ore with sphalerite, chalcopyrite and locally galena enrichments at the stratigraphic tops and margins of the ore lenses, giving approximate average grades of 1.3% Zn and 1.0% Cu with traces of Pb, As and Ag (Halls et al, 1977).

2.2 The Moray-Buchan Data

The "Moray-Buchan" data used in this thesis are derived from the regional geochemical stream sediment survey data generated by the Institute of Geological Sciences (IGS), now known as the British Geological Survey (BGS), for the Moray-Buchan Geochemical Atlas (in prep.).

In the late 1960's the IGS initiated a programme, subsequently sponsored by the Department of Industry, to prepare geochemical maps of the land area of Britain. The

principal purpose was to supplement available geological information and provide regional geochemical data for studies of crustal evolution and ore forming processes. The data were to be published in the form of geochemical atlases at a scale of 1:250000.

Following orientation surveys by the IGS in the Northern Highlands of Scotland, a region which was chosen because of the relative absence of industrial contamination, the first four Regional Geochemical Atlases, dealing with Shetland (1978), Orkney (1978), South Orkney and Caithness (1979) and Sutherland (1982) were published. The next to be published will be the Argyll Atlas and this will be followed by the Moray-Buchan Atlas.

All the geochemical atlases are based on stream sediment samples, being the best medium for systematic regional mapping (Plant and Moore, 1979).

2.2.1 Sampling Procedure

In Scotland the IGS have had to contend with several problems which adversely affect stream sediment sampling. These are related mainly to the coarse detrital nature of the alluvium, which makes the collection of representative samples difficult, and to the extensive areas of poorly drained peat and acid soils. These factors tend to obscure

the relationship between the trace element content of stream sediments and bedrock because some elements are coprecipitated with secondary hydrous Fe and Mn oxides.

The IGS have had to adapt traditional stream sediment sampling methods in order to develop a procedure which is practical, robust and reproducible. A sampling method is used whereby up to 50g (dry weight) of the -100 BSI (150 μm) mesh size material is collected directly in the field. At each site 10-15 kg of material is screened through an 8 BSI (2000 μm) mesh sieve, and then through a 100 BSI (150 μm) mesh sieve using the minimum of water. The sample is set aside for a fixed period of time at each site (c. 15 min) to allow the fine material to settle, and the resulting wet slurry is collected in a Kraft paper bag. This method eliminates the need for dry sieving which is a significant source of error.

A minimum sampling density of 1 sample per 1-2 km based on second or third order streams sampled immediately above confluences was selected for the standard procedure. The top few centimetres of sediment are removed before sample collection. This procedure reduces the effect of element coprecipitation with secondary Fe and Mn oxides since these precipitates tend to be concentrated in first order streams and in the top few centimetres of the sediment profile.

The sub-sample taken for analysis is small (15mg) and sub-sampling is therefore likely to be an important source of error. Although the -100 BSI (150 μm) mesh size

fraction employed for the programme is finer than is normally collected for geochemical exploration, it nevertheless consists of a heterogenous mixture and requires further sub-sampling and particle size reduction. Freeze drying has been shown to produce a dry, friable powder which is particularly suitable for homogenization and sub-sampling; sub-samples of 15-20g of freeze dried material are ground to a particle size of 95% -350 BSI (53 μm) mesh in agate planetary ball mills prior to analysis (Plant et al, 1984).

In order to satisfy the analytical requirements of the programme it was necessary to employ rapid, inexpensive multi-element methods and the IGS have found that d.c. arc emission spectrography is the most appropriate technique for the determination of all elements except U (Plant and Moore, 1979).

The 4289 stream sediment samples from the Moray-Buchan area were analysed for 33 elements: Ag, Al, As, B, Ba, Be, Bi, Ca, Co, Cr, Cu, Fe, Ga, K, La, Li, Mg, Mn, Mo, Nb, Ni, Pb, Rb, Se, Si, Sn, Sr, Ti, U, V, Y, Zn and Zr. In addition pH and conductivity measurements were recorded, and at some locations U in water was measured. The data also includes a recorded sample identifier, national grid coordinates and latitude and longitude for each sample. The major elements are quoted as their oxides: Al_2O_3, CaO, Fe_2O_3, K_2O, MgO, SiO_2 and TiO_2; the K/Rb ratio is also given.

The subarea chosen for study is shown in detail in Figure 2.2 (see also Figure 2.3 for regional setting) and includes all the Lower and Middle Dalradian and some of the Upper Dalradian outcrop in the Moray-Buchan region. The number of samples was reduced from 4289 to the 503 occurring in the study area. The following variables were then removed from the data because in the majority of samples they were below detection limit or undetermined: Ag, As, Bi, Mo, Nb, Se, Sn, U in waters, pH and conductivity. Gallium, the K/Rb ratio and SiO_2 were also omitted: Ga because geochemically it behaves similarly to Al_2O_3; K/Rb because it can be derived from two existing variables; SiO_2 because of closure problems (see Aitchison, 1981, 1984). All three variables also showed near-perfect Gaussian frequency distributions and were therefore unlikely to be of particular interest.

The final data set consists of 503 samples analysed for 24 elements.

2.2.2 Regional Geology

The Moray-Buchan area is situated in the N.E. part of the Scottish Grampian Highlands in rocks that form part of the Dalradian Supergroup, a succession of approximately

25km of marine metasedimentary rocks and metavolcanics that accumulated from ca. 700 to 510Ma in a complex N.E.-S.W. oriented ensialic basin at least 400km long and 150km wide.

The Dalradian Supergroup is of late Proterozoic to Cambro-Ordovician age and has been divided into three groups (Harris and Pitcher, 1975): the Lower (Appin) Group, the Middle (Argyll) Group and the Upper (Southern Highland) Group (see Table 2.2). The Appin Group rocks consist of relatively thin formations of limestone, pelite (dominantly carbonaceous), quartzite and characteristically striped and laminated pelite-psammite. At the base of the Appin Group interbanded quartzites and semipelites follow upwards from the Moine psammites without a break.

During the Ordovician (514 - 480Ma) Grampian Orogeny the Dalradian sedimentary pile underwent complex polyphase deformation, metamorphism and igneous intrusive activity.

The Dalradian Supergroup has recrystallized in two metamorphic facies series (Atherton, 1977). In the S.W. and Central Highlands recrystallization occurred in the Barrovian intermediate facies series with the index minerals being chlorite - biotite - garnet - (staurolite) - kyanite - sillimanite. The Dalradian of N.E. Scotland has recrystallized in the Buchan low pressure facies series cordierite - andalusite - sillimanite.

Table 2.2

Stratigraphy of the Dalradian of Scotland (after Harris and Pitcher, 1975)

Group	Subgroup	Kintyre-Arran	Loch Fyne	Loch Tay
Southern Highland				
Argyll	Tayvallich	Tayvallich Lmst	Loch Tay Lmst	Loch Tay Lmst
	Crinan	Crinan Grits (= Erins Qzite)	Ben Lui Schist / Erins Qzite	Ben Lui Schist
	Easdale	Craignish Phy (= Ardrishaig Phy) Easdale Slate	no outcrop	Farragon Ben Lawer Ben Eagach Carn Mairg
	Islay	Islay Qzites Dolomite beds Port Askaig Tillite		Killiecrankie Schichallion Dolomite beds Schichallion Boulder beds
Appin	Blair Atholl Ballachulish Lochaber			

2.2.3 Depositional History

The depositional history of the Dalradian succession is analysed in terms of several large scale, basin deepening and shallowing sequences (100's metres to kms thick) which record a progressive upward increase in tectonic instability and contemporaneous basic volcanism.

In considering the sedimentary evolution of the Dalradian rocks, Harris and Pitcher (1975) envisaged a restricted trough in which sediments of the Lochaber and Ballachulish Subgroups (Appin Group) were deposited. This restricted trough was eventually replaced by a progressively less restricted area in which the overlapping deposition of similar shallow water limestones, carbonaceous shales and orthoquartzites of the Blair Atholl Subgroup took place. The normally quiescent conditions of deposition of the Appin Group sediments were succeeded by higher energy conditions accompanying the dominantly clastic sedimentation of the Argyll Group, including the late Precambrian glacial episode. It was envisaged that Argyll Group sediments accreted southeastwards possibly with a migrating axis of maximum sedimentation coinciding with the leading edge of the sedimentary wedge; e.g. proximal turbiditic sandstones of the Crinan Grits in the Islay-Loch Awe region pass southeastwards into finer

grained more distal deposits (Ben Lui Schists). The provenance of the Argyll Group clastic sediments was thought of as being a northwestern continental landmass with northwesterly derived material subsequently being transported axially N.E.-S.W. In a change of provenance probably occurring in the Cambrian and coinciding with the outbreak of widespread basic volcanicity, a southern continent, of which the Ox Mountains may be a remnant, supplied the clasts of high grade metamorphic material that characterize the Southern Highland Group sandstones.

The tectonic framework during the evolution of this major, composite basin deeepening sequence was probably dominated by differential fault activity within a marginal basin.

2.2.4 The Dalradian of the Buchan Area, N.E. Scotland

The Dalradian rocks of the Buchan area are predominantly metasediments of the Argyll and Southern Highland Groups. Structurally the Buchan rocks lie in the upper (right way up) limb of the Tay Nappe, preserved in the area as a plunge depression, which has been modified by two N.E.-S.W. trending major folds known as the Boyndie Syncline and the Buchan Anticline (Fettes, 1970 and Hudson, 1980). On the structural timescale, minor recumbent isoclines (F1) with

penetrative axial schistocity (S1) are thought to be related to the Tay Nappe, whilst minor monoclinal folds (F3) commonly associated with crenulations in S1 and with crenulation cleavage (S3) are thought to be related to the major N.E.-S.W. folds. Second phase structures appear to be rare or absent at higher structural levels.

As mentioned above the metamorphic rocks of the Dalradian Supergroup can be divided into two major provinces - the Barrovian and the Buchan - on the basis of mineral assemblages in pelitic rocks. There is a major pressure difference between the two metamorphic provinces - the boundary being the kyanite-andalusite isograd. The outcrop pattern of the Buchan metamorphic zones is controlled by the Boyndie Syncline. The lowest grade (biotite zone) rocks occupy its core and the higher grade rocks outcrop on its limbs as two N.E.-S.W. trending sets of zones which show different zonal sequences (facies series) of mineral assemblages. In the western sequence, four prograde zones are recognized on the occurrence of the minerals biotite, cordierite, andalusite and staurolite. In the eastern sequence only three zones are developed (biotite, cordierite and andalusite). Both sequences pass into sillimanite bearing rocks at higher grade and the western sequence passes finally into the Barrovian kyanite zone to the west of the Newer Gabbro outcrops.

The metasediments of the Buchan area were intruded, prior to F3 movements, by a large, discontinuous,

sheet-like gabbroic body which has been dated at 501+17 Ma. A series of small, post metamorphic granites emplaced after the F3 event are representatives of a single intrusive event centred on 460 Ma. Around 400 Ma the final intrusive event occurred in the area with the emplacement of the "Late Caledonian" Granites.

The area is now in part overlain, with marked unconformity, by sediments of Old Red Sandstone age.

2.2.5 Mineralization

Of all the main rock divisions in Scotland, the Dalradian Supergroup contains the greatest number of significant base-metal showings, with the majority of these occurring in rocks of the Middle Dalradian Easdale Subgroup. Moreover, a notable areal concentration can be observed on the ground between Islay in the extreme S.W. and the Aberfeldy district in central Perthshire (see Figure 2.3). Many of the mineral showings in this highly mineralized belt are of vein type, such as those at Tyndrum Pb-Zn mine, but some, such as those at Creggans, McPhun's Cairn and Meall Mor, are by description stratiform, and others, for example Coille Bhraghad and Craignure, although often described as metasomatic replacements, are evidently stratabound (Smith et al, 1984). Except that at Tyndrum,

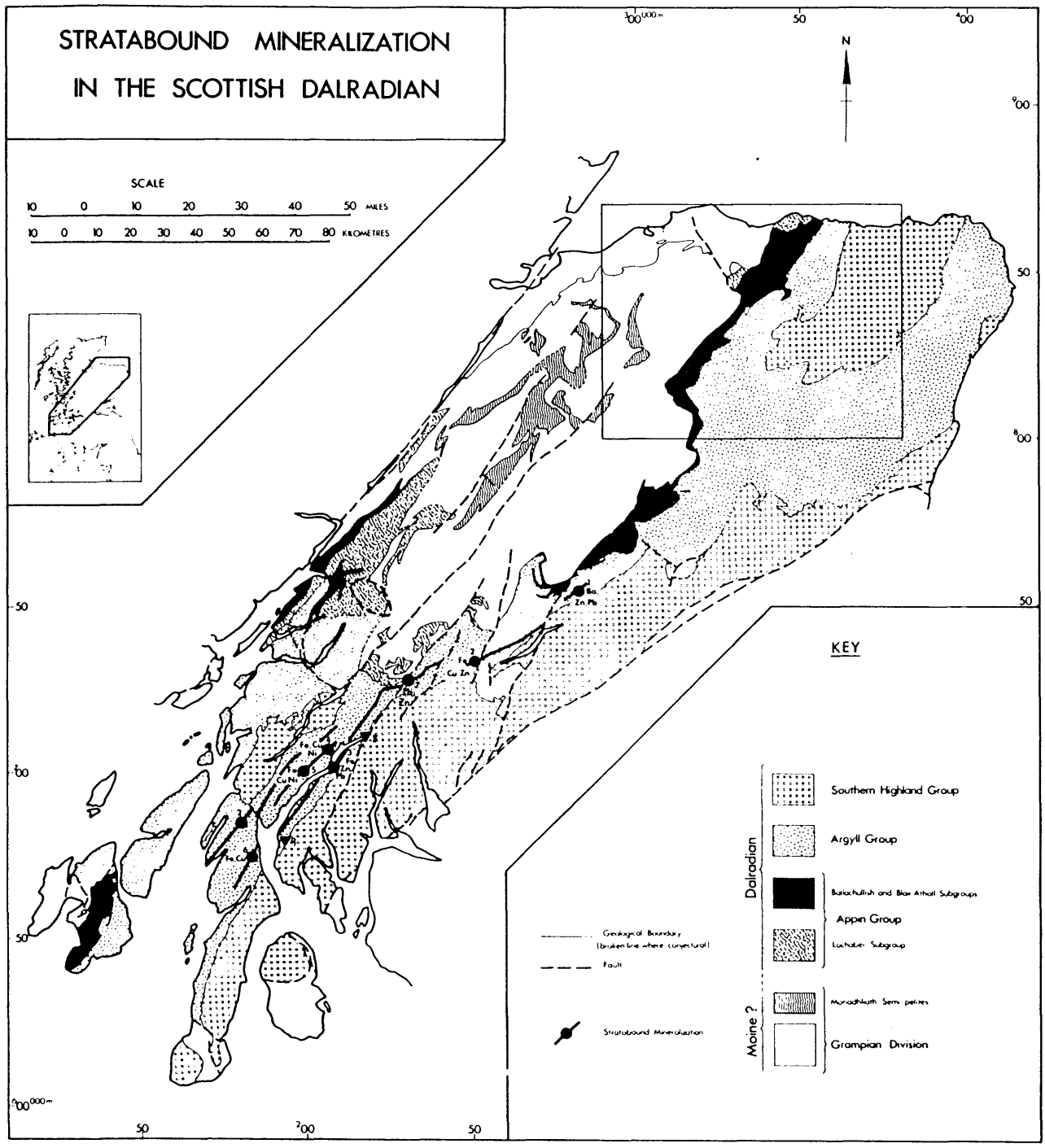


Figure 2.3

Map showing the occurrences of stratabound mineralization in the Dalradian of Scotland, plus extra locations mentioned in the text.

1. Aberfeldy; 2. Pyrite zone (Glen Lochay); 3. McPhun's Cairn and Creggans; 4. Coille Braghad; 5. Craignure; 6. Pyrite zone (Meall Mor); 7. Auchtertyre (Tyndrum); 8. Clachan Beag; 9. Kilfinan. [Adapted from Plant et al, 1984]

Inset shows Moray-Buchan study area - see Figure 2.2

all these occurrences are in either the Ardrishaig Phyllite or its lateral equivalent, the Erins Quartzite (see Loch Fyne section, Table 2.2). A further occurrence of possibly stratabound mineralization in the Loch Fyne area is the galena-sphalerite showing at Clachan Beag, which is contained in the Loch Tay Limestone of the Tayvallich Subgroup. In the Dalradian rocks of Banffshire the Lecht Fe-Mn deposit near Tomintoul was regarded as being fault controlled, but its orientation is such that it may well be stratabound. Analyses of lithiophorite and cryptomelane from the deposit show unusually high concentrations of Zn, Ba and Tl (Wilson et al, 1970).

In the 1950's the IGS recognized that there was a conspicuous development of pyrite with minor chalcopyrite closely following the stratigraphy for 190km from Knapdale (W. of Loch Fyne) to Glen Shee, reaching a cross strike width of at least 12km in the Loch Tay region where the horizon is repeated by open folds.

This "Dalradian Pyrite Horizon" is a concordant zone, averaging 180m in thickness, of disseminated pyrite and traces of chalcopyrite preferentially concentrated in calc-quartzites near the top of the Ben Lawers Schist, its lateral equivalent to the S.W. the Ardrishaig Phyllite and within the slightly younger Erins Quartzite. The pyrite content is generally less than 5%, but reaches 20% or more at Glen Lochay, Auchtertyre, Creggans and in the Meall Mor area. Base metals within the pyrite horizon are developed

in higher than average amounts and in more variety at Auchtertyre, 4km east of the Tyndrum Pb-Zn-Cu-Ag veins of Devonian-Carboniferous age. The disseminated sulphides comprise pyrite, sphalerite, pyrrhotite, chalcopyrite and in one section traces of molybdenite. They form up to 30% by volume of thin fine grained siliceous layers in a 200m succession of quartzites and quartz-mica schists near the top of the Ben Lawers Schist. Zinc assays a maximum of 0.4%, Cu 0.14% and Pb 160ppm together with traces of Au and Ag (Smith, 1977 and Smith et al, 1977a).

Geographically and stratigraphically related to the Pyrite Horizon are the following stratabound deposits: the Aberfeldy Ba-Zn-Pb deposit (see below); the Coille Bhraghad and Craignure Ni-Cu deposits; the McPhun's Cairn Zn-Pb mineralization (sub-economic); the Meall Mor Cu deposit; the Kilfinan Cu mineralization and the Clachan Beag Pb-Zn deposit. With the exception of the Clachan Beag deposit all the above deposits are thought to be stratabound occurrences of pre-metamorphic synsedimentary exhalative origin.

The mineralization occurs in generally steeply dipping metasediments on the N.W. margin of the Tay Nappe Flat Belt, which strikes N.E.-S.W. Any deposits at the same stratigraphic level to the S.E. will either have been eroded away (in the upper limb of the Tay Nappe) or will lie at depth in the lower inverted limb. The apparent absence of large stratabound deposits may therefore be due

to an accident of the structure and present erosion level (Willan, 1980).

The mineralization appears to be associated with changes in lithology at the top of the Ben Eagach Black Schist and Ben Lawers Schist/Ardrishaig Phyllite, which were in turn probably due to changes in basin configuration. The tectonic environment of the Dalradian Basin appears to have been an extensional one, possibly related to the commencement of the opening of the Iapetus Ocean or a marginal ocean basin to the S.E. Pulses of tectonic movement resulted in faulting, subsidence, a change in the pattern of sedimentation and volcanic and exhalative activity which produced the mineralization. If the extensional tectonics in the Dalradian Basin occurred in response to the same stress field which caused the commencement of ocean floor spreading to the S.E., then faulting within the Dalradian Basin would have been oriented N.E.-S.W. i.e. parallel to the "Caledonide" trend. Exhalative activity along such fractures would explain the lateral N.E.-S.W. persistence of such horizons as the Pyrite Horizon and the Green Beds.

Direct evidence of the heat source to power the mineralizing system is lacking. The mineralization is in a dominantly metasedimentary succession and there is no obvious association of the mineralization with igneous rocks. However, by Lower Cambrian time, the basin had filled with at least 6km of sediments derived from a

granitic/metamorphic terrain and would thus have had a high K, Th and U content and consequently a high potential for radiogenic heat production (Willan, 1980).

With the exception of the Aberfeldy Ba-Zn-Pb deposit (see below) the genesis of the Dalradian stratabound deposits mentioned above can be explained by a simple model (Russell et al, 1984) which invokes convectively circulating highly saline seawater which percolated down through growth faults initiated by extensional tectonics. In the thick, hot, Dalradian sedimentary pile the hot brine could leach major and trace elements and return to the sea-floor via syndepositional faults. Exhalative activity into the reducing environment of black muds in Ben Eagach times, with an abundant supply of bacteriogenically reduced sulphur, would lead to massive sulphide formation. The Dalradian Pyrite Horizon probably represents geographically widespread late stage exhalations into a Dalradian basin containing relatively low concentrations of biogenically reduced sulphur, insufficient to produce a large, economic, base metal deposit.

2.2.6 The Aberfeldy Ba-Zn-Pb Deposit

The Aberfeldy deposit is the most recent and significant discovery in the Grampian Highlands and was found by the

IGS as a result of the follow up of Ba, Zn and Pb anomalies detcted during regional stream sediment sampling. It is the first record of synsedimentary barium mineralization in the Caledonides and possibly the largest accumulation of barium silicates in the world.

The mineralized zone contains a number of lithologically unusual barium rich rocks including bedded barite rock, quartz-celsian rock, barian muscovite schist and fragmental rocks. A considerable volume of the zone, however, comprises graphitic, calcareous and dolomitic schists, which, although lithologically indistinguishable from rocks outside the zone, contain anomalous concentrations of barium. Carbonate is a volumetrically minor constituent but may contain appreciable amounts of lead and zinc. The zone varies in thickness from about 60m to 110m and extends intermittently over 7km of strike and through a proven vertical interval of 370m. The component rock types occur as bands and lenses that range in thickness from < 1cm to several metres. The banding everywhere appears to be concordant with the bedding plane foliation in the adjacent schists and is therefore most probably of sedimentary origin. Small scale interbanding of contrasting lithologies is common but, as yet, no systematic zoning has been identified (Coats et al, 1980).

In the Aberfeldy area the Ben Eagach Schist consists principally of black graphitic muscovite schist with impersistent bands of quartzite, non-graphitic mica schist

and limestone. However, close to its boundary with the Ben Lawers Schist it shows considerable lithological variation with, in addition to the suite of barian rocks, bands of calc-schist.

The mineralized strike length has been split into two sectors (Gallagher, 1984). In the western sector, known as the Foss deposit, barite thicknesses of 2.3-15.5m have been recorded at intervals over 1.8km. A minimum of 2Mt of barite ore has been inferred from drilling at Foss, but the recoverable amount could be smaller because of mining problems related to the complex folding.

In the eastern sector, referred to as the Ben Eagach deposit, an upper horizon with up to 10m of barite and an inferred strike length of 360m has been reported. In addition a lower mineralized horizon has been defined by systematic drilling over some 2km of strike. Barite thicknesses are up to 28m and inferred ore reserves are comparable to those at Foss.

The barite at the Aberfeldy deposits normally occurs as anhedral or polygonal crystals 1-2mm across and making up about 90% of the rock. Centimetre banding is picked out by enrichment in pyrite, sphalerite and galena or in magnetite, often accompanied by enrichment in quartz. Stratabound sphalerite and argentiferous galena are associated with the barite at Aberfeldy, more notably at the Ben Eagach deposit, so the potential for mining in the district is high. Large quantities of barian muscovite and

the barian feldspar, celsian, which occur in the mineralized horizons - although having no obvious industrial application at present - represent important guides to ore in exploration of the Middle Dalradian metallogenic belt (see below).

The presence of huge volumes of barium silicates associated with the Aberfeldy deposits poses special problems for the generation of an acceptable ore genesis model. Russell et al (1984) propose a model in which highly saline seawater was drawn several kilometres down growth faults where it gained heat from the ambient heat in the crust; the newly buoyant fluid returned convectively through active faults to the sea floor. The hydrothermal convection cells widened and deepened with time as the crust was cooled by the throughflow of seawater. Seawater in the downdraught was modified by interaction with lower Dalradian rocks, and eventually, with Moinian basement, both of which are relatively rich in barium. Deep weathering of the continental landmass to the N.W. of the Dalradian basin resulted in highly alkaline ground and river waters rich in potassium, silica and alumina which mixed with dense hypersaline seawater in evaporitic tidal flats and lagoons. The resulting dense brine carrying $Al(OH)_4^-$ and $SO_4^=$ sank into the Dalradian basin, filling fault bounded troughs such as the newly developed sub-basins at Aberfeldy into which acid hydrothermal solutions containing Ba, Mn, Fe, Zn, Pb, Ag and silica were

issuing. The concentration of barium in the hydrothermal solution was unusually high and although 10Mt of barite was precipitated, excess barium reacted with the aluminium and silicon complexes to produce the quartz-celsian rock as a sheath and apron to the barite deposit. Moreover, there was an excess of barium chloride which escaped beyond the confines of the sub-basins to form an aureole of Ba-aluminosilicate which reacted with the sedimentary illite during metamorphism to form barium rich mica.

2.2.7 Exploration

The Middle Dalradian of the Grampian Highlands of Scotland contains significant stratabound deposits of base-metal sulphides and barite which, with a few exceptions, may be classed as sediment hosted submarine exhalative deposits (Klau and Large, 1980; Large, 1980). Since only a small proportion of the 250km strike length of the Middle Dalradian has been prospected in detail, there is considerable potential for further exploration. The following are a few suggestions for exploration by Russell et al, (1983).

1) The Easdale Subgroup and the lowermost members of the

Crinan Subgroup are the only favourable "formations". Of these, the Ben Eagach Schist and its equivalents are the prime targets.

2) Since there was apparently insufficient seawater sulphate to satisfy the huge quantities of hydrothermal barium discharged into the Middle Dalradian sea, any search should concentrate on barium enrichment in the favourable horizons.

3) Abrupt changes of thickness and displacements in the Easdale Subgroup could signify coeval faulting, cross structures and depressions and hence favourable targets for mineralization.

4) It may be that the pods of serpentinite - talc - magnesite - chrome magnetite in the lowermost members of the Crinan Subgroup are the metamorphosed representatives of ultrabasic volcanic activity induced by extensional tectonics. If so they could be the stratigraphic hangingwalls to sulphide and (?) sulphate mineralization.

5) The Middle Dalradian of N.E. Scotland is regarded as particularly promising since it contains a few mineral showings - mostly veins, but including more enigmatic occurences such as the Lecht Fe-Mn deposit - a large

volume and variety of intrusive rocks (some with appreciable Cu mineralization) and extensive tracts of concealed ground.

6) Plant et al (1984) propose some geochemical indicators for mineral exploration in the Dalradian of Scotland. They suggest that B and Li and other elements of basic association can be used to indicate a sedimentary regime consistent with that of an extensional marginal basin. A linear zone of high Co and MgO levels which occurs stratigraphically above the mineralization over the lavas which mark the base of the Tayvallich Subgroup and the extension of the horizon to the N.E. is thought to record basic magmatism associated with the near surface development of rifting within the Dalradian basin.

The combination of such geochemical criteria as:

1) Pb > 40ppm, Ba > 1650ppm and Zn > 200ppm occurring within 20km of mineralization;

2) a zone of high Co and MgO stratigraphically above mineralization;

3) a region where B and Li levels are generally > 30ppm;

identifies all of the recorded occurrences of stratabound

mineralization in the Scottish Dalradian. Such criteria, particularly the identification of the zone of high Co-MgO, are considered by Plant et al (1984) to be of great value in identifying favourable areas for stratabound mineralization in the Caledonian province elsewhere.

CHAPTER THREE

CENTRAL TENDENCY AND DISPERSION AND THE ROLE OF ROBUST ESTIMATORS

3.1 Introduction

The technical term "robust" was coined only in 1953 (by G.E.P Box), and the subject of robust statistics only acquired recognition as a legitimate topic for investigation in the mid-1960's. The subject, however, was certainly not a new concept. Among the leading scientists of the late nineteenth century there were several practicing statisticians (e.g. the astronomer S.Newcomb, the astrophysicist A.Eddington, and the geophysicist H.Jeffreys) who had a perfectly clear, operational understanding of the idea; they knew the dangers of long-tailed error distributions, they proposed probability models for gross errors, and they even invented excellent robust alternatives to the standard estimates, which were rediscovered only recently. But for a long time theoretical statisticians tended to shun the subject as being inexact. The statistician P.J.Huber with his 1964 paper entitled "Robust Estimation of a Location Parameter" may have helped

to dispel such prejudices, and may be responsible for the respectibility with which the subject is now regarded. An excellent book by Andrews et al (1972) entitled "Robust Estimates of Location: Survey and Advances" was the product of a year-long seminar at Princeton University, U.S.A., with many leaders in the field participating. The first systematic book-length exposition of robust statistics was produced by P.J.Huber in 1981. Easier to understand for the non-statistician is "Understanding Robust and Exploratory Data Analysis" by Hoaglin et al (1983).

Robust statistics have only very recently begun to be used in the Earth Sciences. The problem of robust estimation of semivariograms in geostatistics has been addressed by Cressie and Hawkins (1980), and Krige and Magri (1982). The need to downweight multivariate outliers prior to regression analysis has been solved using robust techniques by Garrett (1983) and Garrett et al (in press).

3.2 Why Robust Statistics?

The classical statistical techniques are designed to be the best possible when stringent assumptions apply. These are:

1) Additivity;

2) Constancy of variance;

3) Normality.

During recent years, however, an enormous amount of effort has been devoted to robust techniques because it has become increasingly apparent that some of the most common classical statistical procedures (in particular those optimized for an underlying normal distribution) are excessively sensitive to seemingly minor deviations from the assumptions. Robustness in this context therefore signifies insensitivity to small deviations from the assumptions.

Primarily we are concerned with distributional robustness in which the shape of the true underlying distribution deviates slightly from the assumed model (usually the Gaussian model).

Two concepts that will be used many times in the thesis are those of location and scale. The location of a sample distribution is a measure of the centre of the main body of that distribution. Scale is a measure of the spread of observations about the location.

3.3 Robust Estimators of Location

The sample mean is the most common summary statistic of a batch of data. When the data come from a symmetric distribution, the mean gives an estimate of the location of the centre of that distribution. Other simple estimators such as the median also do this. How do we choose among them? We require estimation procedures (and estimators) that perform well for the conditions actually underlying the data. Theory gives us estimators that are best for particular, narrowly specified, conditions (such as the sample mean for the Gaussian distribution). In practice, however, we never know the underlying conditions precisely, if only because of possible disturbances in the data. Thus we seek estimators that perform well for a variety of underlying conditions - these we call <u>robust estimators</u>.

Two properties required of a robust estimator are resistance and robustness of efficiency. An estimator is resistant if it is relatively insensitive to at least two types of circumstance encountered in sampling.

1) A few large deviations in the data, often thought of as outliers, will tend to make the tails of the underlying distribution <u>heavier</u> than those of the Normal distribution. A heavy tailed distribution is one in which a larger proportion of the distribution is contained in the extremities (tails) compared to the equivalent part of a Normal distribution. In this situation the sample mean is no longer a good estimate

for the centre of the distribution because all the sample observations contribute equally to the value of the mean; estimators that attach less emphasis to extreme values perform better.

2) Many rather small deviations in the data, for example as a result of rounding or grouping, may adversely affect the sample median. Although the sample median is determined by ordering all the observations, only one or two of the central observations determine its numerical value. This value will respond sharply to the presence of small errors in the centremost observations.

An estimator has robustness of efficiency over a range of distributions if its variance is close to the minimum for each distribution. The ratio of variance of the best estimator for a given distribution to the variance of a given estimator provides a measure of the relative efficiency of the estimator. An estimator with high relative efficiency for a broad class of distributions has robustness of efficiency. Robustness guarentees that an estimator is good for a collection of distributions without necessarily being best for any particular one.

The most thoroughly investigated aspect of robust estimation has been estimating the centre (or location) of a symmetric distribution. Two reasons help to explain this emphasis.

1) When we are prepared to treat the data in a symmetric way as coming from a symmetric source, we know that the distribution's location is best described by its centre of symmetry.

2) If we have a sample from a symmetric distribution all the natural estimators of location (the mean, median, midmean, etc.) coincide.

When dealing with an underlying asymmetric distribution the natural estimators of location are likely to differ from one another. The data analyst must therefore decide on the appropriate quantity to estimate for the problem in hand. Consider, for example, a practical situation in which the distribution of element concentrations is positively skewed (to the right). The mean concentration is likely to overestimate the true background concentration, while the median concentration will probably be a better estimate.

So, unlike the symmetric case, in many asymmetrical situations we have no natural and consistent goal for location estimation.

Since a skewed distribution violates the third of the three basic assumptions that we impose on the statistical model - that of normality - it would seem sensible to transform the data prior to estimation so that this assumption holds. Data transformations will be fully

discussed in chapter 5, but for the purposes of this chapter, data distributions are assumed to be symmetric unless specified otherwise.

Two classes of estimators - L-estimators and M-estimators - have played a prominent role in recent research on robust estimation of location. L-estimators, which are linear combinations of the ordered observations, are discussed in Section 3.3.1. M-estimators (Section 3.3.2) generalize the idea of the maximum likelihood estimator of the location parameter in a specified distribution. Two other classes of estimators which are often mentioned in the statistical literature, but which are not discussed here, are R-estimators and Adaptive estimators. R-estimators are derived from rank tests, but do not perform as well as M-estimators, and are not as easy to work with. Adaptive estimators use the sample itself to choose the estimation procedure, for example some "adaptive trimmed means" determine the amount of trimming according to a sample estimate of the tail heaviness of the underlying distribution, whereas others attempt to minimize the estimated variance of the estimate.

3.3.1 L-estimators

The class of L-estimators is defined as linear combinations of the order statistics. The order statistics of a sample are the observations rearranged in order of

increasing magnitude. For example the order statistics of the sample (5,3,6,-2,4) are (-2,3,4,5,6).

Let x_1 , x_2 ,..., x_n be a sample of size n. The observations rearranged in ascending order of magnitude, denoted by $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, are called the order statistics of the sample, and $x_{(i)}$ is called the ith order statistic.

The class includes the arithmetic mean, median and trimmed means as special cases. For example, the mean of a sample of size n, $\Sigma x_i / n$, may be thought of as an L-estimator with all weights equal to 1/n. The sample median, the estimate of the midpoint of the set of ordered values, is an L-estimator with all but one or two of the coefficients equal to zero.

Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ be the order statistics of a sample of size n. Let $a_1, a_2, \ldots, a_n$ be real numbers, $0 \leq a_i \leq 1$, i=1,...,n, such that $\Sigma a_i = 1$

An L-estimator T with weights $a_1, a_2, \ldots, a_n$ is

$$T = \Sigma \; a_i x_{(i)}$$

Trimmed means, Midmeans and Medians

The arithmetic mean is a simple, well understood estimate of location. However, as we have seen, it is

highly non-robust, being very sensitive to extreme outliers. One simple way to make the arithmetic mean insensitive to extreme points is first to delete, or "trim" a proportion of these observations from each end of the ordered sample, and then to calculate the arithmetic mean of the remaining observations. These trimmed means form a family indexed by α, the proportion of the sample size removed from each end.

If α is a multiple of $1/n$, (n = sample size) an integral number of observations are deleted from each end and the trimmed mean is a simple average of the remaining observations. If α is not a multiple of $1/n$, $[\ \alpha n\ ]$ (where $[\ \alpha n\ ]$ denotes the greatest integer less than or equal to αn) observations are removed at each end, the largest and smallest remaining observations are given a weight $P = 1 + [\ \alpha n\ ] - \alpha n$. A weighted mean is then taken.

$$T_{(\alpha)} = \frac{px_{([\ \alpha n + 1])} + x_{([\ \alpha n + 2])} + \ldots + px_{(n - [\ \alpha n])}}{n(1 - 2\alpha)}$$

The 0% trimmed mean is the usual sample mean, the 25% trimmed mean is the midmean. The 50% trimmed mean is the median.

Trimean

Of the class of L-estimators, only the trimean cannot be viewed as an example of a trimmed mean. The trimean is motivated by the desire to include sample information farther from the centre by making use not only of the median, but also of a few additional order statistics - namely the upper and lower fourths of the sample.

Let F_l and F_u denote the lower and upper fourths of the sample (approximately the quartiles - see Glossary). The trimean, denoted by TRI, is

$$TRI = \frac{1}{4}(F_l + 2M + F_u) \quad \text{where } M = \text{Median}$$

From the collection of simple location estimators described above we want to find one or more that perform well for a variety of symmetric distributions. The variance of an estimator is used to measure its performance for a given distribution. To compare estimators across a collection of distributions, we use the concept of relative efficiency (introduced in Section 3.3). For a given estimator this is the ratio of the variance of the best trimmed mean for a given distribution to the variance of the specified estimator. We then choose as most robust, the estimator with maximum relative efficiency. A detailed discussion of the mathematics of this process may be found

in Hoaglin et al (1983) chapter 10.

One simple conclusion drawn from the detailed analysis undetaken by Hoaglin et al, is that the ideal amount of trimming is related directly to the heaviness of the tails of the distribution from which the data are sampled. The most efficient estimator for data from the Gaussian distribution is the sample mean (0% trimming). For slightly stretched tails the 10% trimmed mean is most efficient. For moderately stretched tails the midmean is most efficient. For heavy tailed distributions the median is the most efficient.

In most practical situations we only have a vague notion of the distribution generating the data. What is needed therefore, is not an estimator that is best for a specific situation, but one that is fairly good over a suitable range of situations. We want an estimator that has fairly high efficiency over a range of possible distributions.

When heavy tailed distributions are anticipated, slightly more than 25% trimming is preferred. When heavy tailed distributions are not expected, slightly less trimming is required, and the 20% trimmed mean is recommended. However, for a broad class of distributions, the midmean should provide an excellent estimate of location.

3.3.2 M-estimators

L-estimators, though simple to calculate, may sacrifice resistance by giving weight to extreme observations, or, conversely, they may sacrifice efficiency by placing too little weight on end observations that are not extreme. Modern research on robust methods offers even better performance if we can accept more complicated estimators of location. This Section introduces M-estimators, the class of estimators that offers the greatest advantages in performance, flexibility and convenience.

M-estimators minimize functions of the deviations of the observations from the estimate that are more general than the familiar sum of squared deviations, or the sum of absolute deviations. In this way the class of M-estimators includes the mean and median as special cases. Viewed in another way, M-estimators generalize the idea of the maximum likelihood estimator of the location parameter in a specified distribution.

A simple reformulation of M-estimators yields a weighted mean in which the weights depend on the data. The resulting M-estimators provide a straightforward way of modifying the familiar least squares method, particularly in regression problems, where they are the basis for the technique of iteratively reweighted least squares (discussed in Chapter

9). M-estimators minimize objective functions more general than the familiar sum of squared deviations associated with the sample mean. Instead of squaring the deviation of each observation x_i from the estimate t, we apply a function $\rho(x;t)$ and form the objective function by summing over the sample: $\Sigma\ \rho\ (x_i;t)$. Often $\rho(x;t)$ depends on x and t only through x-t, so that we could write $\rho(x-t)$; for the moment we retain the greater generality by treating x and t as separate arguments.

The M-estimate $T_n(x_1, \ldots, x_n)$ for the function ρ and the sample $x_1, \ldots, x_n$ is the value of t that minimizes the objective function $\Sigma\ \rho(x;t)$. When we know the derivative of ρ with respect to t, a function which we denote by ψ, we may find it more convenient to calculate T_n by finding the value of t that satisfies

$$\Sigma\ \psi(x_i;t) = 0$$

The most familiar M-estimate is the sample mean, the least squares estimate of location. For least squares estimation, ρ is the square of the residual:

$$\rho(x;t) = (x-t)^2$$

Thus we minimize $\Sigma(x_i - t)^2$. Differentiating the objective function ρ with respect to t and dropping the constant 2 leads to the equation

$$\Sigma(x_i - t) = 0$$

Because $t = \Sigma x_i / n$ solves this equation, T_n is indeed the sample mean. Figure 3.1 shows the functions ϱ and ψ for the sample mean.

A second example is the sample median. The objective function is the absolute value of the residual,

$$\varrho(x;t) = |x - t|$$

and the corresponding ψ function is

$$\psi(x;t) = \text{sgn}\ (x - t)$$

where

$$\text{sgn}\ (u) = \begin{cases} +1, & u > 0 \\ 0, & u = 0 \\ -1, & u < 0 \end{cases}$$

The expression $\Sigma\ \psi(x_i;t) = \Sigma\ \text{sgn}\ (x_i - t)$ counts each observation above t as +1 and each observation below t as -1, so that t = median yields a sum of zero. Figure 3.2 shows the ϱ and ψ functions for the sample median.

It is important that our estimator of location should remain unchanged when:

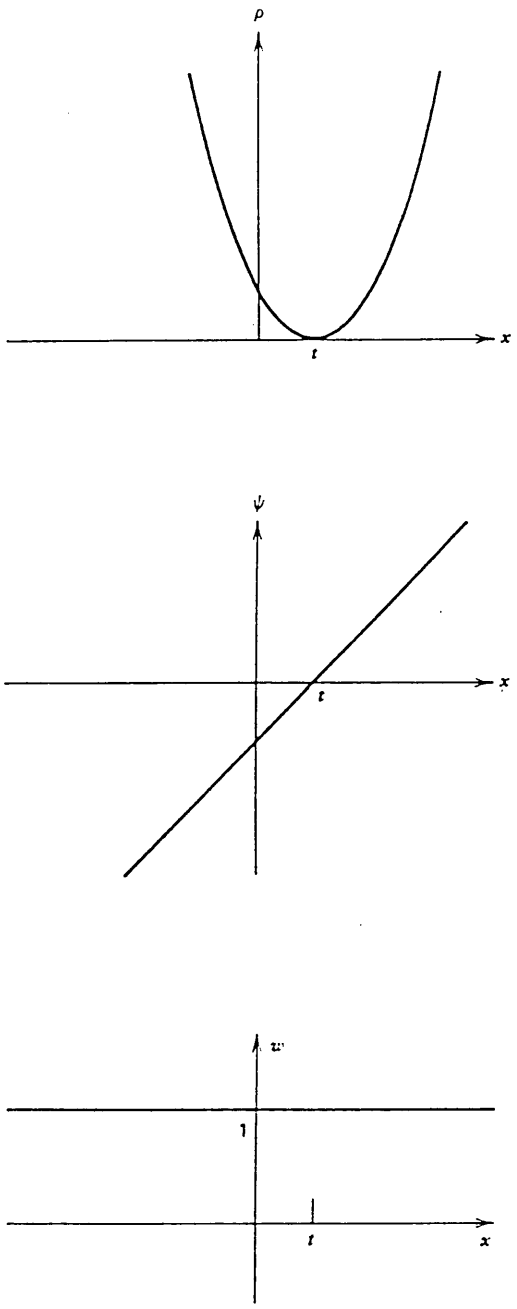


Figure 3.1

Objective function, ψ–function and weight function of the sample mean. (a) Objective function, $\rho(x;t)$. (b) ψ–function, $\psi(x;t)$. (c) Weight function, $\omega(x;t)$.

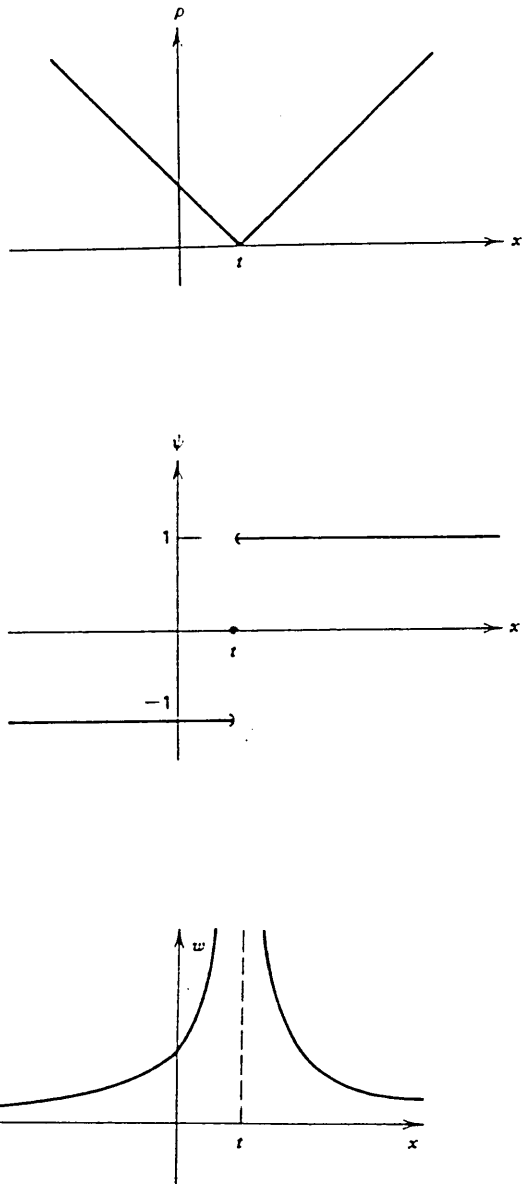


Figure 3.2

Objective function, ψ−function and weight function of the median. (a) Objective function, $\rho(x;t)$. (b) ψ−function, $\psi(x;t)$. (c) Weight function, $\omega(x;t)$.

1) the whole sample is shifted by an amount A; and

2) the whole sample is multiplied by a non-zero constant B.

Such estimators are said to be location and scale invariant. Most M-estimators, however, achieve this property only by incorporating some measure of the scale of the sample, denoted by s_n. Two exceptions are the mean and the median. Other M-estimators of location depend on the scale of the arguments of ρ and ψ. We therefore use an auxiliary estimator of scale s_n, which is a function of $x_1, \ldots, x_n$ and use this, together with a constant c, to rescale the measure of deviation $(x_i - t)$. We then have the centred and rescaled observations:

$$U_i = \frac{x_i - t}{cs_n} \qquad \text{Equation 3.1}$$

T_n is a value of t that minimizes $\Sigma\, \rho(U_i)$ or satisfies:

$$\Sigma\, \psi(U_i) = 0$$

In general, an M-estimator of location involves a fixed measure of scale s_n. The constant c in the definition of U_i

matches the basic scale of ρ or ψ to the scale of the data, measured in units of s_n. An inherent part of the definition of the M-estimator, c is known as the tuning constant because it can be chosen to fine-tune the estimator so that it has a specified efficiency at a chosen distribution (usually the Gaussian). For most M-estimators the ψ function is curved or is composed of several linear pieces, and some ψ functions are zero outside an interval. It is often convenient to take (-1,+1) as this interval. This may be achieved by setting the tuning constant c to an appropriate value. Thus an M-estimator of location is actually a family of estimators, all involving the same scale measure or auxiliary scale estimator; the value of the tuning constant determines the individual member of the family.

Influence Curve

Two properties required of a robust estimator are resistance and robustness of efficiency (see Section 3.3). We can study an estimator's influence curve to see how resistant that estimator will be. The influence curve of an estimator displays the effect on that estimator of changing the value of one observation from the sample.

For example, consider a sample of 11 observations. Let

10 of the observations have values such that they sum to zero. Let the 11th observation, x, move from large negative values to large positive values. Let us consider the sample mean as the estimator of location. Since 10 observations sum to zero, the grand total must equal x itself. Consequently the sample mean

$$\bar{x} = \Sigma x_i / n = x/n = x/11$$

Figure 3.3 shows the influence curve for this example.

It can be shown (Hoaglin et al, 1983, Chapter 11) that the influence curve is proportional to the ψ function. This means that although the size of the two curves may differ, the shape will not. The influence curve for the trivial example above (Figure 3.3) has the same shape as the ψ function for the sample mean (Figure 3.1b).

Breakdown

An estimator is resistant if it is altered to only a small extent by a small proportion of outliers. The estimator breaks down if the proportion becomes too large. The breakdown bound or breakdown point of an estimator is the largest possible fraction of the observations for which there is a bound on the change in the estimate when that

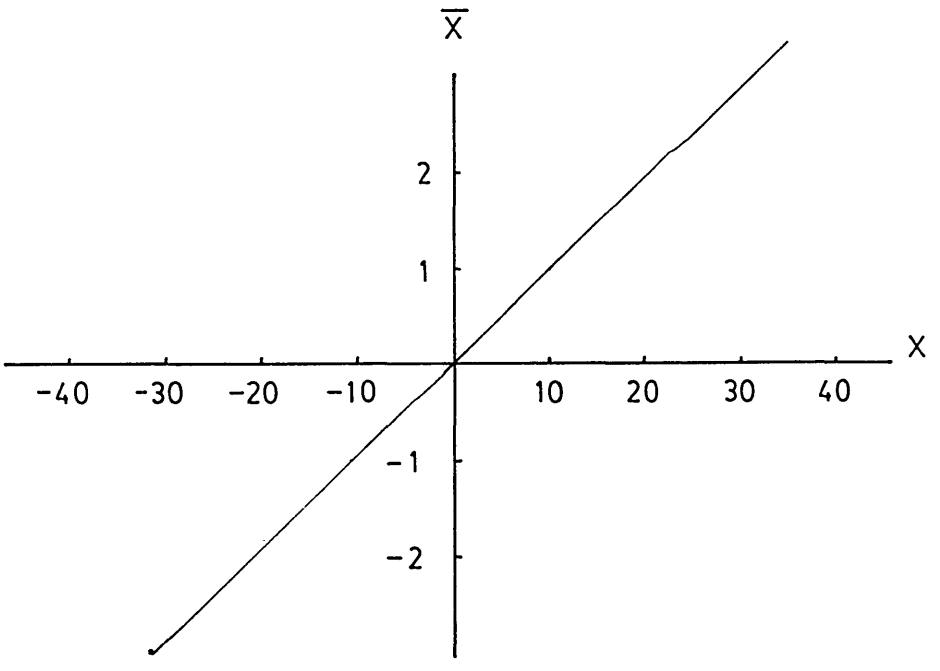


Figure 3.3 Influence curve of sample mean.

fraction of the sample is altered without restriction.

An estimator is resistant only if its breakdown bound is greater than zero. As a single observation is made larger, the mean increases without bound. Therefore the breakdown bound of the mean is zero. The breakdown bound of the median is the largest possible for a location estimator that treats observations on each side of the estimate symmetrically. It tends toward 50% for large sample sizes.

3.3.3 Auxiliary Scale Estimators.

Three common scale estimators are the standard deviation, the fourth spread and the median absolute deviation. The sample standard deviation has breakdown point zero; it is not resistant. In fact gross errors affect the standard deviation even more adversely than they affect the mean. The standard deviation is, therefore, not used as an auxiliary scale for M-estimators.

The most commonly used auxiliary scale estimators are:

1) The median absolute deviation (MAD),

$$MAD = med_i \, (\, | \, x_i - med_j \, (x_j) \, | \,).$$

2) The fourth spread (dF), is the difference between the

lower fourth and the upper fourth), or its close relative, the interquartile range.

The breakdown bound of the fourth spread, 25%, is smaller than that of the MAD, 50%. A high breakdown bound here indicates tolerance to sample asymmetry. For this reason the MAD is the preferred scale estimator despite it having slightly lower efficiencies than the dF for a range of distributions.

3.3.4 Examples of Location Estimators

Mean

The mean is the classical estimator of location. The objective function and ψ function are shown in Figure 3.1. It is the maximum likelihood estimator for the Gaussian distribution. The variance of the sample mean at the Gaussian distribution is the minimum. It is unbounded and therefore highly non-resistant.

Median

The median is the classical robust estimator. The objective function and ψ function are shown in Figure 3.2.

It has the maximum breakdown bound (close to 50%), but is sensitive to rounding and grouping errors of the middle observations.

The mean and median do not require auxiliary scale estimators. The estimators discussed below do require auxiliary scale estimators, aswell as associated tuning constants. Thus each example is a family of estimators.

Hubers

We have seen that the squared residual objective function (Figure 3.1) makes the estimator too sensitive to outliers. On the other hand, using the absolute residual makes the estimate too sensitive to the middle observations. A compromise is possible when we remember C.P.Winsor's principle: all distributions are Gaussian in the middle.

The objective function of a Huber estimator (Figure 3.4a) is, therefore, quadratic in the centre (like that for the sample mean), and linear in the tails (like that for the sample median). An auxiliary scale estimator is now essential in order to fix the point of transition from quadratic to linear.

The ψ function of a Huber estimator (Figure 3.4b) is linear in the centre and constant in the tails. The mean and median are extreme cases in Huber's family when k tends to ∞ and $k = 0$ respectively.

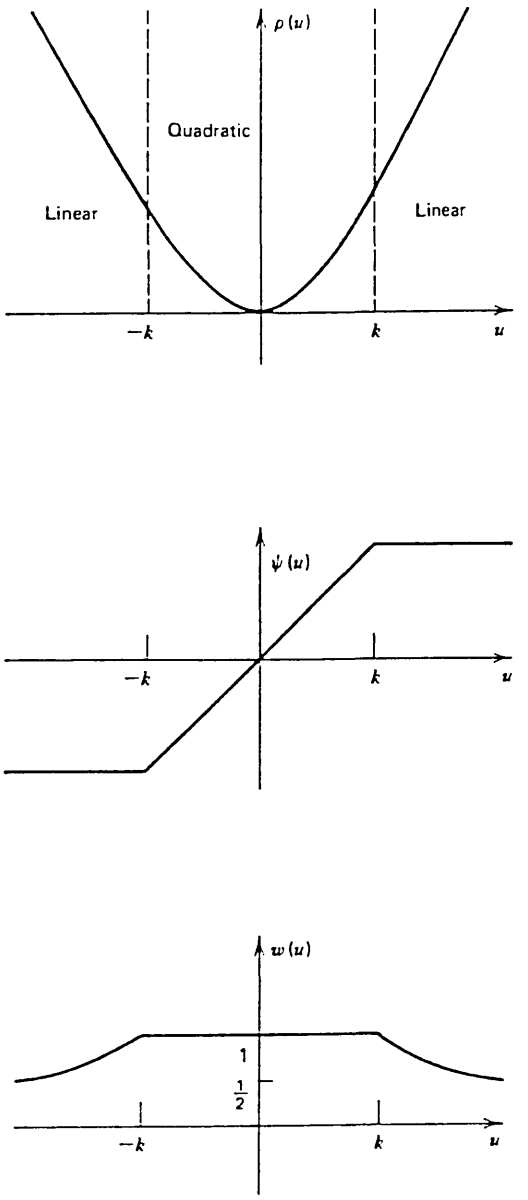


Figure 3.4

Objective function, ψ−function and weight function of Hubers estimators of location.
(a) Objective function, $\rho(u)$. (b) ψ−function, $\psi(u)$.
(c) Weight function, $\omega(u)$.

Rescending Estimators: Hampel's

The influence curve of Huber's estimator (Figure 3.4b) is constant for all observations beyond a certain point. An M-estimator can be made more resistant by having the ψ function, and hence the influence function, return to zero. Outliers have diminishing effects on a redescending estimator and beyond the rejection point they have no effect. The rejection point r is defined as the least distance from the location estimate beyond which observations do not contribute to the value of the estimate for a given scale estimate. The objective and ψ functions of the three part redescending M-estimator, sometimes called a Hampel, are shown in Figure 3.5.

Andrews' Wave

The abrupt changes in slope of the three part redescending estimators are intuitively unappealing because of the abrupt changes in the way data are used. Andrews' wave estimators, also called sine estimators, have a smooth redescending influence curve (Figure 3.6b). Andrews et al (1972) use a value of (2.1 π) for the tuning constant c.

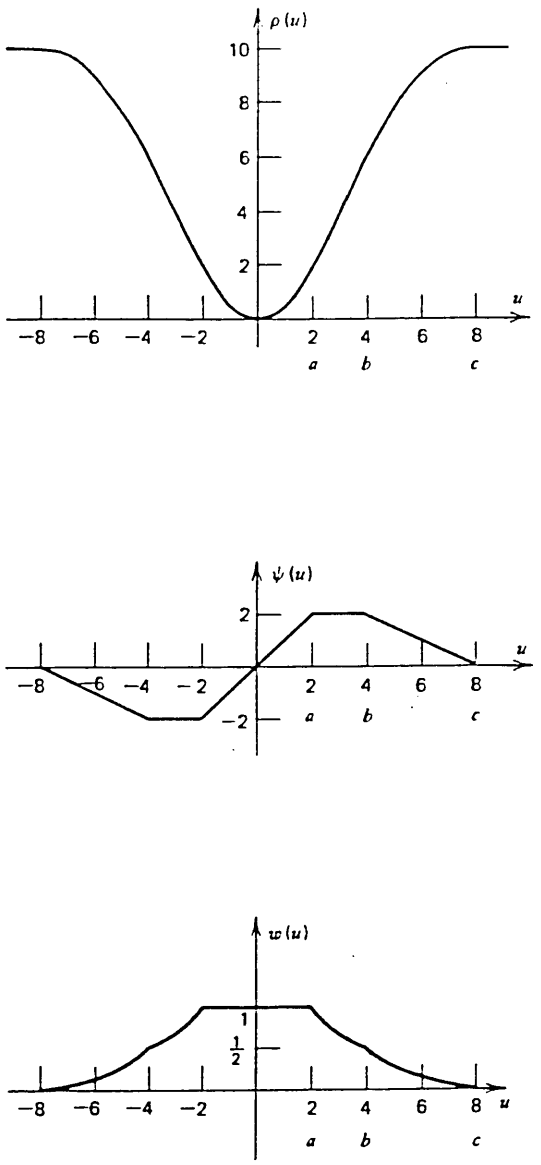


Figure 3.5

Objective function, ψ−function and weight function of the three−part redescending estimators of Hampel. Tuning constants are (a,b,c) = (2,4,8).
(a) Objective function, $\rho(u)$. (b) ψ−function, $\psi(u)$.
(c) Weight function, $\omega(u)$.

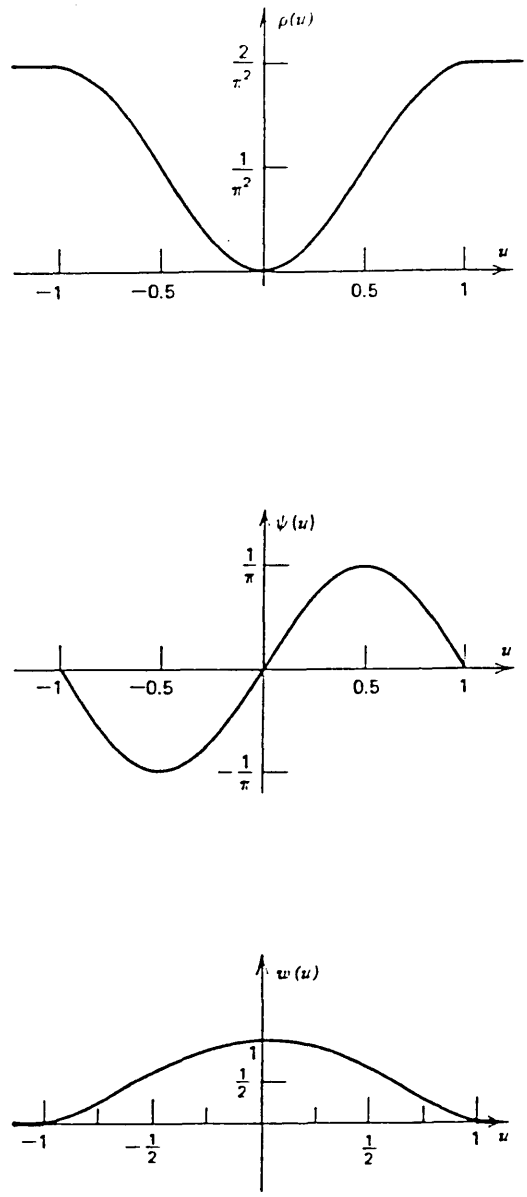


Figure 3.6

Objective function, ψ—function and weight function of Andrews wave estimators of location.
(a) Objective function, $\rho(u)$. (b) ψ—function, $\psi(u)$.
(c) Weight function, $\omega(u)$.

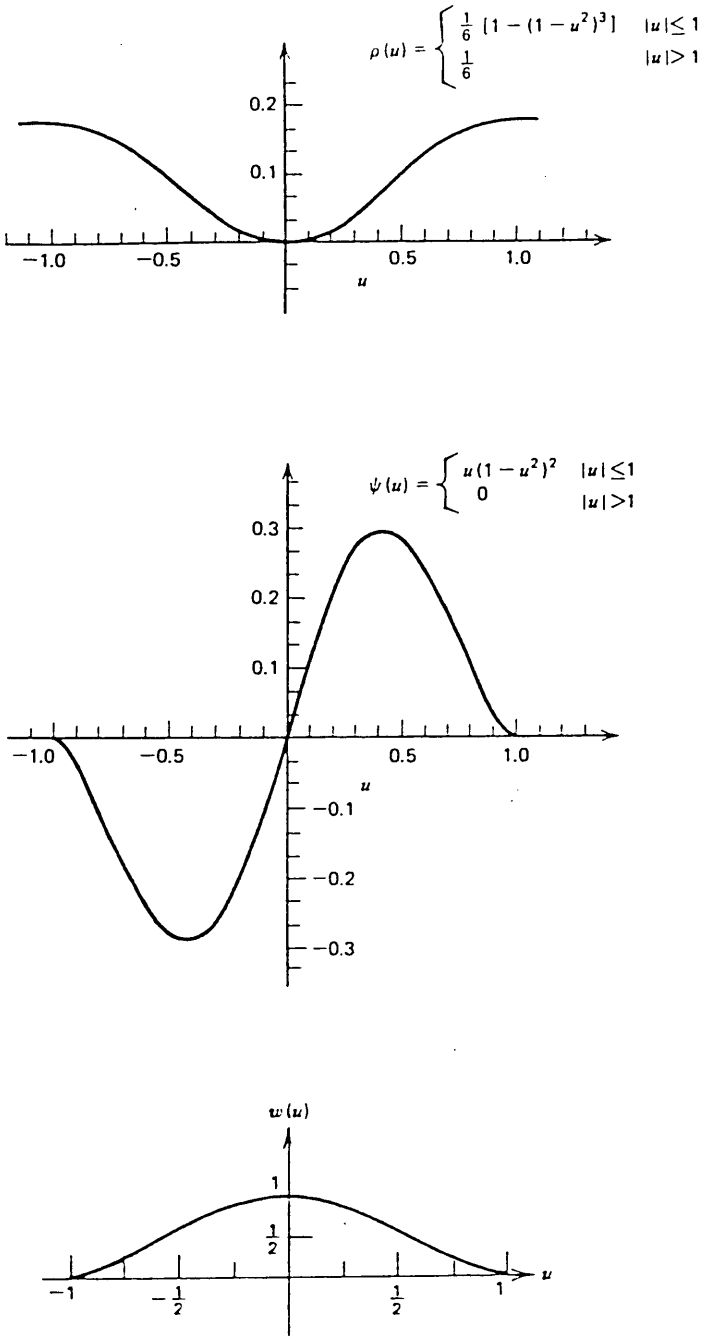


Figure 3.7

Objective function, ψ−function and weight function of the biweights. The vertical scale is 3 times the horizontal scale. (a) Objective function, $\rho(u)$. (b) ψ−function, $\psi(u)$. (c) Weight function, $\omega(u)$.

The Biweight

Tukey's biweight estimators are a popular robust family. Both Andrews' wave estimators and the biweights have smooth redescending ψ functions. For reasonable values of c, say $6 \leq c \leq 12$, the biweights have good robustness of efficiency and are resistant. See Figure 3.7 for the ρ and ψ functions.

A general principle of robust estimation of location is that the tuning constants should be smaller for heavier tailed distributions. For a given set of data as the value of the tuning constant, c, decreases the value of U_i (equation 3.1) will increase, and more observations will fall beyond the rejection point.

3.3.5 W-Estimators

An alternative form of M-estimation is called W-estimation. The M-estimate T_n is defined by

$$\frac{\Sigma \psi (x_i - T_n)}{cs_n} = 0$$

When we define w according to $Uw(U) = \psi(U)$ and substitute for $\psi(U)$, we get

$$\Sigma \left[\frac{(x_i - T_n)}{cs_n} \quad w \quad \frac{(x_i - T_n)}{cs_n} \right] = 0$$

or, rearranging

$$T_n = \frac{\Sigma\, x_i\, w\, [(x_i - T_n)/cs_n]}{\Sigma\, w\, [(x_i - T_n)/cs_n]} \qquad \text{Equation 3.2}$$

T_n is thus a weighted mean of the x_i. We call T_n, defined iteratively by equation 3.2, the W-estimate based on the weight function w.

A trivial example is the sample mean. No auxiliary scale is needed, and the ψ function is simply $\psi(U_i) = U_i$ with $U_i = x_i - T_n$. The weight function $w(x_i - T_n)$ is equal to one over the whole range of x_i and equation 3.2 becomes

$$T_n = 1/n\ \Sigma x_i$$

A direct algebraic solution of equation 3.2 is, however, rarely available. Instead we have to use an iterative solution. Let $T_n^{(k)}$ be the estimate at the kth iteration. Let

$$U_i^{(k)} = \frac{x_i - T_n^{(k)}}{cs_n}$$

Then the iteration formula is

$$T_n^{(k+1)} = \frac{\Sigma\, x_i\, w(U_i^{(k)})}{\Sigma\, w(U_i^{(k)})} \qquad \text{Equation 3.3}$$

Equation 3.3 is an example of a general procedure known as iteratively reweighted least squares (IRLS) estimation with weights $W_i^{(k)} = w(U_i^{(k)})$. The ordinary least squares (OLS) estimate T is the choice of t that minimizes $\Sigma\,(x_i-t)^2$, namely

$$T = 1/n \quad \Sigma\, x_i$$

The weighted least squares estimate T_w, based on fixed weights W_i, is the choice of t that minimizes $\Sigma\, W_i(x_i-t)^2$, namely

$$T_w = \frac{\Sigma\, W_i\, x_i}{\Sigma\, W_i}$$

Iteratively reweighted least squares estimation is a further development. Each weight depends on the residual at the preceding iteration (equation 3.3). Starting with a sensible estimate (for W-estimation of location, the sample

median is a good choice), we iterate until the sequence of estimates has converged to within our desired accuracy. The weights at each iteration correspond to the contribution of each observation to the estimate. For example, outliers may receive small or even zero weights. The weight functions for the six estimators of location discussed above are graphed in Figures 3.1 through 3.7.

The advantage of W-estimation lies in equation 3.2 which permits straightforward calculation. In most situations the W-estimate calculated from a sample will be the same as the corresponding M-estimate.

3.3.6 Summary

The careful user of the mean watches for pitfalls. When the data contain outliers, one approach is to remove the outliers and then to use the mean. A second approach is to use a robust estimator for that sample in place of the mean. A third approach is to apply a transformation to the data so that the adverse effect of the outliers on the shape of the distribution is reduced (see Chapter 4). Huber (1981) asks whether a two-step approach to robust estimation would suffice - i.e. first clean the data by applying some rule for outlier rejection, and then use classical estimation procedures on the remainder. His

answer is quite clearly that this approach will not suffice for the following reasons.

1) It is rarely possible to separate the two steps cleanly; for instance outliers may only be recognized by resorting to robust techniques.

2) Even if the original data conform to a normal distribution augmented with some gross errors, the cleaned data will not be normal (there will be statistical errors of both kinds, false rejections and false retentions). The situation is even worse when the original data derive from a genuine non-normal distribution. Therefore classical normal theory is not applicable to the cleaned data and classical estimators will tend to be distorted.

3) It is an empirical fact that the best rejection procedures do not quite reach the performance of the best robust procedures. The latter apparently are superior because they can make a smooth transition between full acceptance and full rejection of an observation.

Robust estimators provide a one-step procedure for limiting the influence of extreme observations. Huber

estimators are designed to be highly efficient for distributions close to the Gaussian. For heavy tailed distributions, however, very extreme observations may have to be manually removed from the data. Thus the need for data screening is reduced, but not eliminated when we use Huber estimators in place of the mean.

Redescending estimators totally remove the influence of very extreme observations. Manual rejection of outliers makes the Huber estimators analagous to redescending estimators. Either procedure will remove very extreme observations; the challenge is in the observations that we are not sure whether to keep in or to discard. The descending part of the ψ function of a redescending estimator resolves this difficulty in a precise way. An automatic procedure is vital when we must rapidly process a large number of samples.

The Hampel's ψ function has a piecewise linear shape with variable proportions for the ascending, level and descending parts. The Andrews wave estimator and the biweight have ψ functions of fixed shape which appeal because they are smooth and which resolve the problem of choosing the proportions. The biweight ψ function is preferred to the wave function because the redescending part extends relatively further for the biweight thus increasing the rejection point. This has important implications for the calculation of auxiliary scale estimators. Robust estimators of scale are weak for small

samples because, when observations in the tails of the sample are not used, too little of the sample remains. The resultant underestimation of scale is a particular problem for redescending estimators. The larger rejection point of the biweight ψ function gives the biweight estimators an advantage over the wave estimators.

3.4 Robust Estimators of Scale

Scale is a vague concept which can be made precise in many ways. We know from experience that some distributions and some samples are more spread out than others. To put this into numerical terms, we must adopt a specific measure of spread. This is why we encounter a variety of words describing the concept: spread, scale, dispersion, variability and width for example. For this thesis the term scale has been adopted.

The familiar sample standard deviation is defined by

$$s = \sqrt{\Sigma (x_i - \bar{x})^2 / (n-1)}$$

where $\bar{x}$ is the sample mean. For independent Gaussian observations, s^2 is the minimum variance unbiased estimator of the population variance σ^2 ; thus $s^2$is the most efficient unbiased estimator of σ^2 for Gaussian data. For

this reason, s has been the most commonly used measure of scale for a batch of data; but, like the sample mean, it lacks resistance and robustness.

Estimators that perform well under highly restrictive distributional assumptions (as the standard deviation does for the Gaussian situation) may cease to do so in the face of slight deviations from these assumptions. The standard deviation, for example, is a highly non-resistant estimator when the Gaussian distribution becomes contaminated. For this reason we hesitate to assume exact knowledge of the distribution from which the sample comes. Three alternative measures of scale are now introduced.

1) The mean absolute deviation from the sample median (AD) is given by

$$AD = 1/n \quad \Sigma \quad | x_i - M |$$

where M is the sample median;

2) The median absolute deviation from the sample median (MAD)

$$MAD = \text{median}_i \ (\ | x_i - M | \ ); \text{ and}$$

3) The fourth spread (dF) is the difference between the upper and lower fourths:

$$dF = F_u - F_l.$$

3.4.1 W-estimators of Scale

In the same way that the location estimator T_n can be iteratively defined using a weight function w (see Section 3.3.5) so can the scale estimator s_n.

$$s_n = \sqrt{\left[\frac{\Sigma\ W_i^2\ (x_i - T_n)^2}{(\ \Sigma\ W_i^2\ - 1)}\right]} \qquad \text{Equation 3.4}$$

Where $W_i = w(U_i) = w((x_i - T_n)/cs_n)$, s_n is thus a weighted standard deviation of the x_i.

To take a trivial example, if all the weights W_i are equal to one (in other words if we use the weight function of the sample mean: Figure 3.1) then equation 3.4 becomes

$$s_n = \sqrt{\left[\frac{\Sigma\ (x_i - T_n)^2}{(n-1)}\right]}$$

which is identical to the equation (above) for the sample standard deviation.

In the same way that the W-estimate of location is obtained iteratively, so too is the W-estimate of scale. Let $s_n^{(k)}$ be the scale estimate at the kth iteration. Let $T_n^{(k)}$ be the location estimate at the kth iteration. Let

$$U_i^{(k)} = \frac{x_i - T_n^{(k)}}{cs_n^{(k)}}$$

Then the iteration formula is

$$s_n^{(k+1)} = \sqrt{\left[\frac{\Sigma\, w(U_i^{(k)})^2 \; (x_i - T_n^{(k+1)})^2}{(\Sigma\, w(U_i^{(k)})^2 - 1)} \right]}$$

This process of W-estimation of scale is discussed further in Chapter 6.

3.4.2 Summary

Robust estimation of scale is not as straightforward as robust estimation of location even in a symmetric distribution, in part because different distributions do not share a readily agreed parameter to estimate (like centre of symmetry for location estimation).

The simple "classical" estimators, such as the standard

deviation and the mean deviation, rapidly lose efficiency when one moves away from the common but unrealistic assumption of Gaussian data and toward heavier tailed contamination or a heavy tailed distribution. Although they are not highly efficient for near-Gaussian data, such simple resistant estimators as the fourth spread and the median absolute deviation perform well for heavy tailed distributions and are adequate for exploratory data analysis work.

For good overall performance, however, the W-estimators of scale are best.

3.5 Discussion

From the applied viewpoint, M-estimators can be considered as a simple modification of classical estimators; the contribution of an observation to the statistic(s) of interest is given full unit weight if it is a reasonable observation, otherwise it's contribution is downweighted. M-estimators can be considered as classical estimators after a weight function has been applied to the data. The weight function (see Section 3.3.5) is given by

$$W_i = w(U_i) = \psi(U_i) / U_i$$

where, as we have seen, U_i is a measure of deviation reflecting the discrepancy of the i-th observation from the robust location estimate relative to a robust scale estimate:

$$U_i = \frac{x_i - T_n}{cs_n}$$

and ψ is an influence function, linear over the range of values of U_i corresponding to reasonable data, but bounded outside this range. Study of the Figures 3.1 through 3.7 will demonstrate the relationship between the ψ function and the weight function. To calculate robust location and scale estimates the data analyst must first choose the form the ψ function is to take (see Table 3.1). Initial estimates of T_n and s_n are the sample median and MAD (as discussed in Sections 3.3.5 and 3.4.1). Weights for all observations are calculated according to the w-functions in Table 3.1. New robust estimators of location (T_n^*) and scale (s_n^*) are then calculated according to the following equations:

$$T_n^* = \frac{\Sigma W_i x_i}{\Sigma W_i} \qquad \text{Equation 3.5}$$

and

Table 3.2

M-estimators of location

Estimator with Tuning Constant(s)	Figure	Objective Function ρ	ψ-Function	Weight Function ω	Range of u	Commonly used Denominator of u
Mean	3.1	$\frac{1}{2}u^2$	u	1	$-\infty$ to $+\infty$	none
Median	3.2	$\lvert u\rvert$	$\operatorname{sgn}(u)$	$\frac{\operatorname{sgn}(u)}{u}$	$-\infty$ to $+\infty$	none
Huber (k) for k	3.4	$\frac{1}{2}u^2$	u	1	$\lvert u\rvert \le k$	MAD
		$k\lvert u\rvert - \frac{1}{2}k^2$	$k\operatorname{sgn}(u)$	$\frac{k\operatorname{sgn}(u)}{u}$	$\lvert u\rvert > k$	
Redescending Estimators (Finite Rejection Point)						
Hampel (a,b,c) (three part redescending) $0<a\le b\le c$	3.5	$\frac{1}{2}u^2$	u	1	$\lvert u\rvert \le a$	MAD
		$a\lvert u\rvert - \frac{1}{2}a^2$	$a\operatorname{sgn}(u)$	$\frac{a\operatorname{sgn}(u)}{u}$	$a < \lvert u\rvert \le b$	
		$ab - \frac{1}{2}a^2 + (c-b)\frac{a}{2}\left[1-\left(\frac{c-\lvert u\rvert}{c-b}\right)^2\right]$	$a\frac{c-\lvert u\rvert}{c-b}\operatorname{sgn}(u)$	$a\frac{c-\lvert u\rvert}{c-b}\frac{\operatorname{sgn}(u)}{u}$	$b < \lvert u\rvert \le c$	
		$ab - \frac{1}{2}a^2 + (c-b)\frac{a}{2}$	0	0	$\lvert u\rvert > c$	
Andrews wave (c) $0<c$	3.6	$\frac{1}{\pi^2}(1-\cos\pi u)$	$\frac{1}{\pi}\sin\pi u$	$\frac{1}{\pi u}\sin\pi u$	$\lvert u\rvert \le 1$	c*MAD
		$\frac{2}{\pi^2}$	0	0	$\lvert u\rvert > 1$	
Tukeys biweight (c) $0<c$	3.7	$\frac{1}{6}\left[1-(1-u^2)^3\right]$	$u(1-u^2)^2$	$(1-u^2)^2$	$\lvert u\rvert \le 1$	c*MAD
		$\frac{1}{6}$	0	0	$\lvert u\rvert > 1$	

$$s_n^* = \frac{\Sigma W_i^2 (x_i - T_n^*)^2}{(\Sigma W_i^2 - 1)}$$

This sequence of operations iterates until the estimates converge to within a desired accuracy. Convergence is normally taken to occur when the absolute difference between the values of s_n for the last two iterations is less than 0.001.

3.6 Robust Estimation and the SIND Data

The purpose of this section is to introduce some of the estimators of location and scale that have been discussed above in the context of the SIND test data set. Table 2.1 shows the SIND data in its original, or, as it will be subsequently termed, its "raw" state. The number of observations or objects (n) is equal to 25, and the number of variables or dimensions (p) is equal to 6. Each row of data may be thought of as an object vector, while each column may be thought of as a variable vector. This terminology will be used in subsequent chapters.

Table 3.2 shows the SIND data ordered within each variable. This is the first step necessary for the calculation of L-estimators (see Section 3.3.1). For

Table 3.2

SIND RAW Data Ordered

Sample	Zn	Fe	Mn	Cd	Cu	Pb
1	23	93	260	2	7	4
2	24	106	260	3	7	5
3	25	108	330	3	7	6
4	26	108	350	4	8	6
5	29	118	410	4	9	6
6	31	128	420	4	10	7
7	33	143	450	4	10	7
8	36	151	490	4	11	7
9	42	159	510	5	11	8
10	44	170	530	5	11	8
11	45	171	530	5	12	8
12	48	173	570	5	15	8
13	48	174	590	6	17	9
14	50	179	650	6	18	9
15	52	184	670	8	20	9
16	72	194	690	9	22	9
17	80	206	700	9	23	10
18	81	259	710	10	27	10
19	89	262	900	10	32	10
20	115	271	910	12	33	12
21	118	305	960	12	37	12
22	274	335	970	31	63	13
23	535	361	1070	31	350	16
24	560	610	1930	35	490	17
25	1010	670	5920	56	590	27

example, the median for each variable is the observation that occurs exactly half way down each column, in this case the 13th observation, i.e. there are an equal number of observations greater than and less than this value. Table 3.3 is generated from Table 3.2 by calculating the absolute difference between each data point and the corresponding median. Table 3.3 may be used to generate several auxiliary scale estimators (see Sections 3.3.3 and 3.4) such as the MAD, which is the median of each column, and the AD, which is the average of each column.

Table 3.4 tabulates the various estimators of location and scale that have been discussed in previous sections. Included in the list of location estimators are the classical mean calculated both using all 25 samples, and retaining only 20 samples (sample numbers 13,15,16,22 and 23 having been omitted as being recognizably anomalous in a geochemical sense). Also included is the M-estimator generated using the iteratively reweighted technique discussed in Section 3.3.5.

As would be expected with data which display varying degrees of positive skewness, the classical mean overestimates the location of each variable distribution. As a progressively greater proportion of objects is trimmed from each end of the ordered data the estimate of location decreases. The percentage decrease is less for some variables than for others, for instance the estimates for Pb decrease by little more than 10% while those for Cu

Table 3.3

SIND RAW Data Absolute Deviation of Ordered Data from Median

$|X_{(i)}-M|$ M = Median

Sample No.	Zn	Fe	Mn	Cd	Cu	Pb
1	25	81	330	4	10	5
2	24	68	330	3	10	4
3	23	66	260	3	10	3
4	22	66	240	2	9	3
5	19	56	180	2	8	3
6	17	46	170	2	7	2
7	15	31	140	2	7	2
8	12	23	100	2	6	2
9	6	15	80	1	6	1
10	4	4	60	1	6	1
11	3	3	60	1	5	1
12	0	1	20	1	2	1
13	0	0	0	0	0	0
14	2	5	60	0	1	0
15	4	10	80	2	3	0
16	24	20	100	3	5	0
17	32	32	110	3	6	1
18	33	85	120	4	10	1
19	41	88	310	4	15	1
20	67	97	320	6	16	3
21	70	131	370	6	20	3
22	226	161	380	25	46	4
23	487	187	480	25	333	7
24	512	436	1340	29	473	8
25	962	496	5330	50	573	18

Table 3.4

Robust Estimators of Location and Scale SIND RAW Data

	Estimator	Zn	Fe	Mn	Cd	Cu	Pb
Location							
	Mean	139.6	225.5	871.2	11.32	73.60	9.72
	T(0.05)	102.71	208.71	669.0	9.58	49.72	9.18
	T(0.10)	79.65	196.22	635.5	8.5	28.37	8.95
	T(0.20)	57.73	188.27	621.33	6.8	18.13	8.73
	Midmean	55.16	185.9	612.2	6.6	17.48	8.62
	T(0.30)	53.95	181.5	604.0	6.5	16.8	8.65
	T(0.40)	48.60	176.2	602.0	6.0	16.4	8.60
	Median	48.0	174.0	590.0	6.0	17.0	9.0
	Trimean	57.5	187.5	630.0	6.75	19.12	9.0
	$^{1}T_{M}$	49.23	175.9	572.8	5.87	15.60	8.67
	$^{2}T_{20}$	49.7	176.15	572.5	5.90	15.7	8.7
Scale							
	$^{3}S_{25}$	231.2	144.4	1107.2	13.02	156.4	4.766
	$^{4}S_{20}$	25.25	71.44	206.6	2.97	8.7	2.96
	AD	105.2	88.32	438.8	7.24	63.48	2.96
	MAD	23.0	56.0	170.0	3.0	8.0	2.0
	d_F	70.0	131.0	470.0	7.0	22.5	4.0
	$^{5}S_{M}$	24.45	70.82	208.4	2.74	8.48	2.97

Notes

1) M-estimator using iteratively reweighted method on 25 samples.
2) Classical mean using 20 samples (13,15,16,22,23 omitted).
3) Classical standard deviation using 25 samples.
4) Classical standard deviation using 20 samples (13,15,16,22,23 omitted).
5) M-estimator using iteratively reweighted method on 25 samples.

decrease by over 75%. This suggests that the Pb distribution is nearer to a Gaussian distribution than any of the other variable distributions. The M-estimates of location were generated using a Hampel type redescending ψ function and should be the most robust location estimates available. If we take the M-estimates as "targets", we can see that, of the trimmed means, those with $\alpha = 0.4$ and $\alpha = 0.5$ (the median) are the closest to the target value in most cases. This means that only 5 samples or less are influencing the estimate, despite the fact that only 5 samples (numbers 13,15,16,22 and 23) are known to be anomalous. If these 5 anomalous samples are omitted from the data and the classical mean recalculated we have the estimator termed T_{20} in Table 3.4. This estimator is closer than any other to the M-estimator "target". The conclusions from this analysis as follows:

1) the L-estimators of location are inadequate in the presence of asymmetrical anomalies. This is due to the fact that objects are trimmed off both ends of the ordered data and, if the anomalous values are arranged asymmetrically, this will cause valuable background data to be omitted from the estimate along with the anomalous samples. The method is thus insensitive, and can be vastly improved by trimming only off the upper end of the ordered data.

2) The M-estimate (T_m) produces the same result automatically that is achieved by manual trimming of known anomalous samples. Calculation of the trimmed estimate T_{20} assumes that the analyst has some prior knowledge of the data, knowledge which is not always either available or easy to generate. The technique involves solving the problem before the analysis starts; this can only ultimately be self-defeating.

The scale estimates in Table 3.4 include the three L-estimators: the MAD, AD and the dF. Also included are the classical estimators using all 25 samples (s_{25}) and using only 20 samples (s_{20}). For the latter estimator the samples 13,15,16,22 and 23 were omitted in the same manner as for T_{20} . The M-estimator of scale (s_m) is also given. This is generated using the techniques discussed in Section 3.4.1. If we again take the M-estimator as "target" then we can say that the classical standard deviation (s_{25}), the AD and the dF all overestimate the scale of each distribution. The MAD, for an estimator that is simple to calculate, is not far from target, but tends to underestimate. Again it is not surprising that the manually trimmed estimate (s_{20}) is close to the M-estimate since the latter does automatically what is done manually in the former.

3.7 Robust Estimation and the Moray-Buchan Data

The analysis of the SIND data (section 3.6) revealed that L-estimators of location and scale are especially unsatisfactory in the presence of asymmetric outliers. Manual trimming of outliers from the data followed by classical estimation (a two step procedure) was found to give the same results as robust M-estimation (a one step procedure). With the SIND data it was known prior to the analysis which objects were anomalous and so it was easy to generate the trimmed estimates for comparison with the M-estimates. With the Moray-Buchan data such prior knowledge is not available and to obtain it without recourse to robust techniques would be difficult and inaccurate. As a result Table 3.5 gives only a comparison between classical estimates and M-estimates with the Median and MAD alone representing the L-estimators.

It is clear that both L-estimators generally underestimate location and scale (if we take T_m and s_m as the respective targets values). The ratio of s_m / s_d (where s_d is the classical standard deviation) for each variable gives a good indication of the deviation of that variable's frequency distribution from the Gaussian norm. If the frequency distribution is not significantly different from the Gaussian we can expect that s_m will not

Table 3.5

Robust Estimators of Location and Scale MB RAW Data

	LOCATION			SCALE			
	Mean	Median	T_m	Sd	MAD	S_m	S_m/Sd
Al	16.8	16.3	16.93	5.36	3.6	5.2	0.97
B	66.5	56.0	65.04	46.4	15.0	31.8	0.685
Ba	731.1	665.0	703.8	359.0	134.0	210.5	0.586
Be	2.4	2.1	2.25	2.57	1.1	1.47	0.572
Ca	1.69	0.97	1.44	1.84	0.62	1.28	0.696
Co	21.7	16.0	17.99	19.6	7.0	11.29	0.576
Cr	87.9	70.0	74.05	76.1	21.0	31.06	0.408
Cu	19.9	17.0	18.29	12.8	7.0	9.56	0.747
Fe	6.92	6.64	6.38	2.94	1.87	2.13	0.724
K	2.64	2.56	2.66	0.89	0.52	0.76	0.85
La	45.2	40.0	42.27	25.5	11.0	14.2	0.557
Li	37.6	37.2	38.09	13.9	10.0	12.5	0.899
Mg	2.63	2.02	2.60	1.91	0.75	1.9	0.995
Mn	2462.1	1380.0	1814.0	4697.0	634.0	1506.8	0.3208
Ni	37.6	29.0	30.76	42.0	11.0	16.27	0.387
Pb	34.0	26.0	28.79	35.2	10.0	17.86	0.507
Rb	123.0	118.0	121.3	52.5	28.0	39.29	0.748
Sr	174.7	157.0	166.2	73.6	28.0	50.5	0.686
Ti	1.39	1.14	1.27	0.80	0.27	0.51	0.644
U	4.07	3.21	3.35	7.18	0.57	1.04	0.1454
V	70.6	57.0	61.35	52.4	27.0	33.36	0.637
Y	37.8	34.0	35.0	17.5	6.0	7.96	0.455
Zn	127.9	110.0	113.6	96.7	29.0	46.66	0.482
Zr	986.4	870.0	985.5	561.0	269.0	493.44	0.879

be significantly different from s_d and the ratio will be equal to one. We can therefore conclude that the variables U, Mn, Ni and Cr are likely to show highly skewed frequency distributions, while Al and Mg will probably have frequency distributions close to Gaussian (see Chapter 4).

3.8 Conclusions

Classical estimates of location and scale have been clearly shown to be severely distorted in the presence of outliers. Since geochemical data tend to exhibit positive skew, we can say that these classical estimates will tend to overestimate the true values. From a practical point of view, this could lead to an artificially high background/anomaly threshold being placed on the data, with the result that a proportion of true anomalies could be missclassified as background. Mineral exploration effort these days often has to be directed towards the subtle anomaly, a situation which demands that data analysis techniques be sensitive enough to detect such anomalies. We can no longer afford the luxury of missclassifying true anomalies as background (or vica versa). Robust statistics enable data analysis techniques to be considerably more sensitive, accurate and precise in their anomaly detection process.

Although the robust estimators described in this chapter were developed to be resistant only for heavy tailed symmetric data distributions, the same estimators, as we have seen in the last two sections, are also resistant for highly asymmetric distributions for which theory would indicate possible breakdown.

CHAPTER FOUR

PROBABILITY PLOTS

4.1 Introduction

At the heart of probabilistic statistical analysis is the assumption that a set of data arises as a sample from a distribution in some class of probability distributions. The reasons for making distributional assumptions about data may be summarized as follows:

1) If we can describe a set of data as a sample from a certain theoretical distribution, say a Gaussian distribution, then we can describe the data in a compact manner. For example, in the Gaussian case, the data can be succinctly described by giving the mean and standard deviation and stating that the sample distribution of the data is well approximated by the Gaussian distribution.

2) Distributional assumptions lead to useful statistical procedures. For example, the assumption that data are generated by normal probability distributions leads to the analysis of variance and least squares theory.

3) Assumptions allow us to characterize the sampling distribution of statistics computed during the analysis and thereby make inferences and probabilistic statements about unknown aspects of the underlying distribution. For example, assuming the data are a sample from a Gaussian distribution allows us to use the t-distribution to form confidence intervals for the mean of the theoretical distribution.

4) Understanding the distribution of a set of data can sometimes shed light on the physical mechanisms involved in generating the data.

Analyses based on specific distributional assumptions about data are not valid if the assumptions are not met to a reasonable degree. This chapter looks at a commonly used graphical display for checking distributional assumptions about data.

4.2 Univariate Probability Plots

Let x_1 through x_n be raw data and let $x_{(1)}$ through $x_{(n)}$ be the corresponding order statistics of that data. Let F(x) be the cumulative distribution function of the theoretical distribution in question. The i-th quantile of F is taken to be

$$Q_i = F^{-1}((i-0.5)/n)$$

A theoretical quantile-quantile plot or Q-Q plot is obtained by plotting the quantiles of the ordered data $x_{(i)}$ against the corresponding quantiles of the theoretical distribution under consideration. Q-Q plots are often called probability plots. When the normal distribution is used the plot is commonly called a normal probability plot, or sometimes just probability plot. If, instead, the lognormal distribution was used the plot would be called a lognormal or log-probability plot. In the latter case it is easier in practice to log-transform the order statistics and plot these against the corresponding quantiles of the normal distribution (see Chapter 5).

The interpretation of probability plots is fairly straightforward. Consider a situation in which the

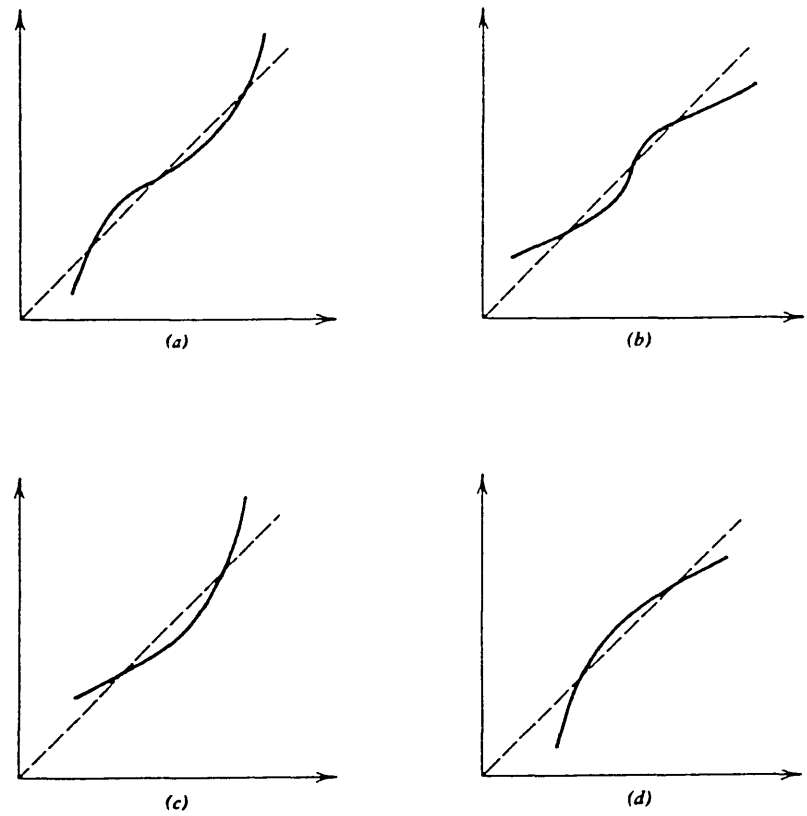


Figure 4.1

Hypothetical patterns in normal probability plots. Data on vertical axis versus Gaussian quantiles on horizontal axis. (a) Heavy-tailed. (b) Light-tailed. (c) Right-skewed. (d) Left-skewed.

theoretical distribution is a close approximation to the sample distribution. The quantiles of the data will then closely match the theoretical quantiles and the points on the plot will fall close to a straight line. If any large or systematic departures from the line occur, they may be interpreted as indicating lack of fit of the distribution to the data. Figure 4.1 shows, schematically, four types of systematic deviation from the straight line relationship.

For more details on these techniques refer to Chambers (1983).

4.3 Univariate Probability Plots and the SIND Data

Figures 4.2 and 4.3 show probability plots of each variable of the SIND raw data. This is an excellent way to obtain an initial impression of the behaviour of the data, something that is an essential precursor to further and more complex analysis. Time spent on this stage of data analysis will be more than repaid later on. A probability plot is a better way to view the distribution of a variable than the more conventional histogram for two reasons. The first is that histograms need bins or classes to be created in order to efficiently summarize the frequency distribution of the data - probability plots on the other hand keep the integrity of each object intact. The second

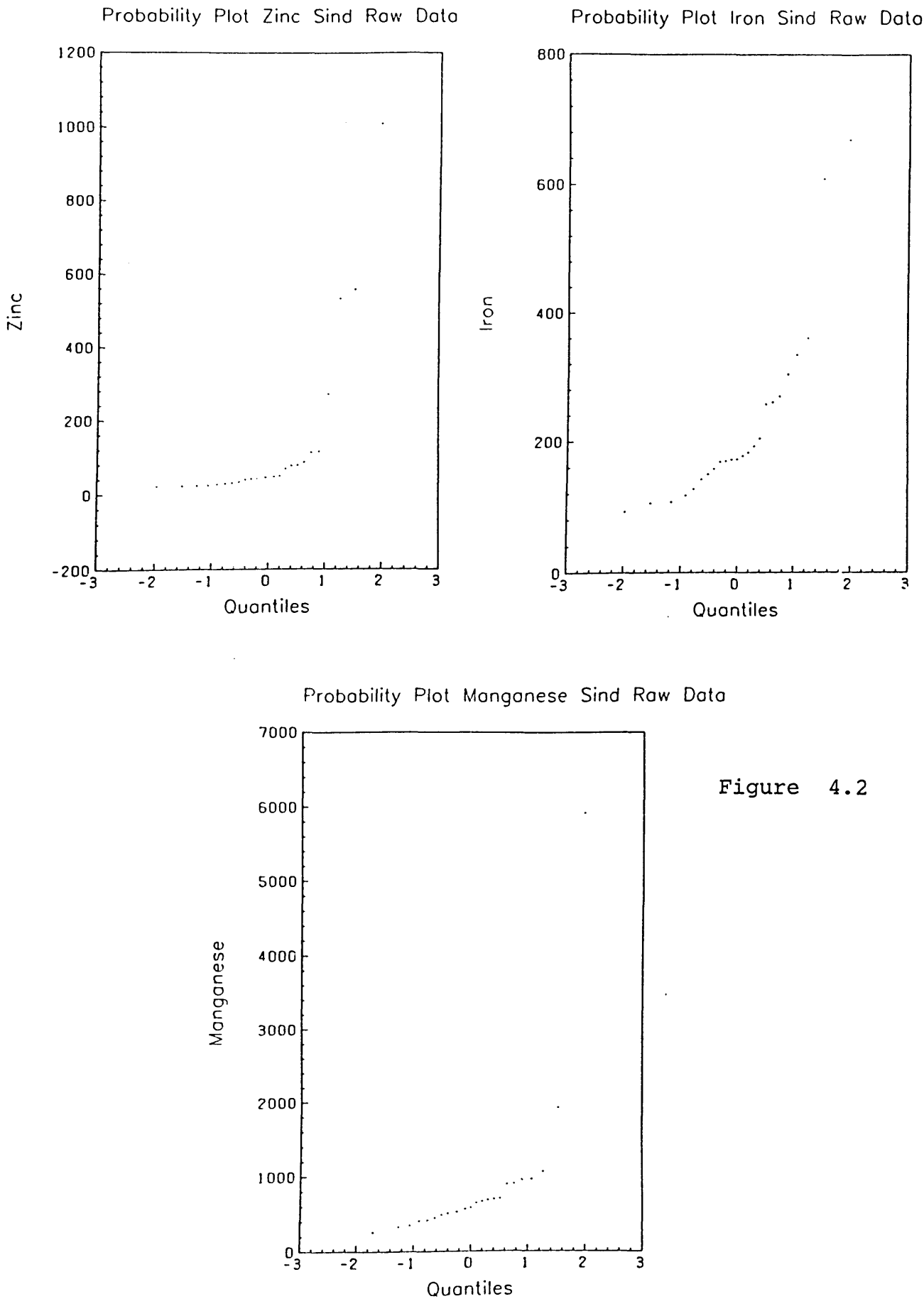
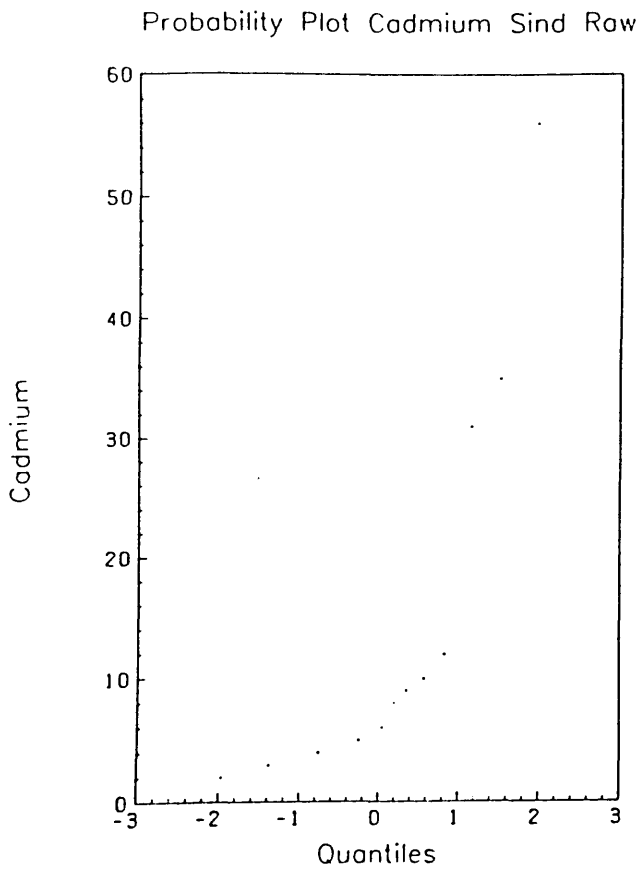
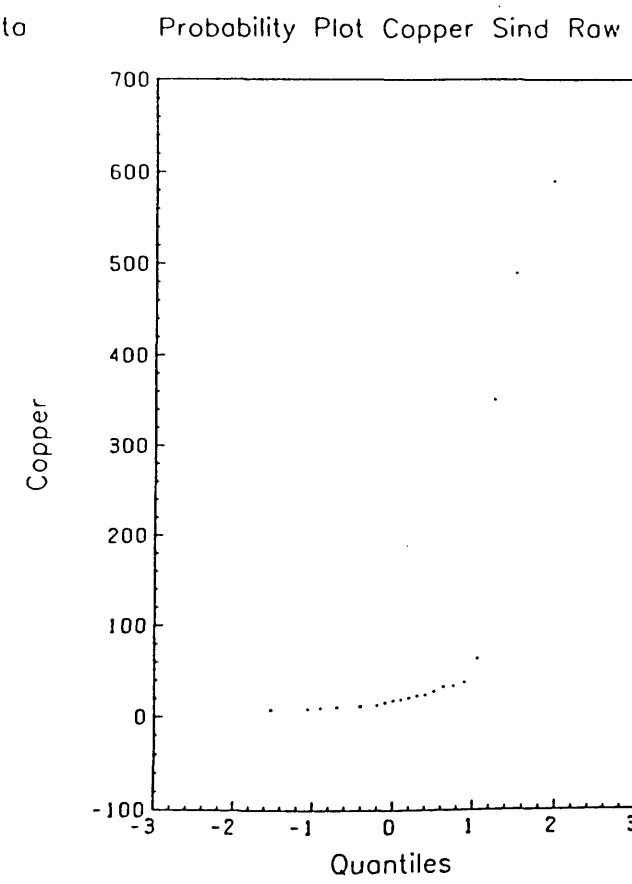

Probability Plot Zinc Sind Raw Data
Zinc
1200
1000
800
600
400
200
0
-200
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Iron Sind Raw Data
Iron
800
600
400
200
0
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Manganese Sind Raw Data
Manganese
7000
6000
5000
4000
3000
2000
1000
0
-3
-2
-1
0
1
2
3
Quantiles

Figure 4.2


Probability Plot Cadmium Sind Raw Data
Cadmium
60
50
40
30
20
10
0
-3
-2
-1
0
1
2
3
Quantiles


Probability Plot Copper Sind Raw Data
Copper
700
600
500
400
300
200
100
0
-100
-3
-2
-1
0
1
2
3
Quantiles

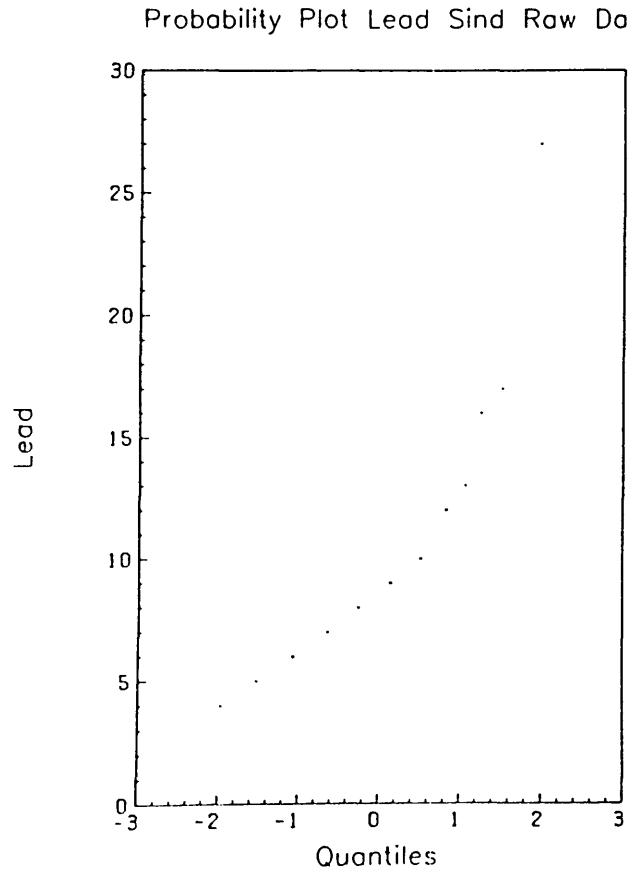

Probability Plot Lead Sind Raw Data
Lead
30
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
Quantiles

Figure 4.3

is that the assessment of the goodness-of-fit of the sample distribution to any theoretical distribution, be it normal, lognormal or whatever, is far easier using the appropriate probability plot than it is using a histogram.

Analysing Figures 4.2 and 4.3 we can come to the following conclusions:

1) The variables Zn, Cd and Cu have as anomalous samples the same four objects, numbers 15,16,13 and 23 (see Table 2.1). The only exception to this order is with Cd for which numbers 13 and 16 change places.

2) The variables Fe and Mn have only two anomalous objects, numbers 22 and 23 (this order is reversed for Mn).

3) The behaviour of Pb is unusual in that it has only one clearly defined anomalous sample (number 23).

4) All the probabilty plots conform to the hypothetical pattern for a right (positive) skewed distribution (see Figure 4.1c).

From these conclusions we may postulate that the variables Zn, Cd and Cu behave in a similar manner and may have similar sources - possibly indicative of base metal mineralization. This is confirmed from our knowledge of

the area (see Chapter 2). We may also postulate that Fe and Mn will vary together. Lead on the other hand appears to be unrelated to any of the other variables though its association with Fe and Mn is probably stronger than with Zn, Cd and Cu.

4.4 Univariate Probability Plots and the Moray-Buchan Data

Figures 4.4 through 4.9 show univariate probability plots of the 24 variables of the Moray-Buchan (MB) raw data. All distributions show some degree of positive skew, indicating the possible presence of a population or populations with anomalously large concentrations. The elements Li, K, Fe and Al have distributions which approximate closest to the Gaussian distribution, while U, Mn and Ni show extreme examples of positive skew.

After studying the frequency distributions of each variable it would be extremely useful to be able to see their spatial distribution on a map of the study area. Point symbol maps (Howarth, 1983) have been generated for each variable with classes based on the percentiles of the cumulative frequency distribution. The class intervals chosen were those concentrations corresponding to particular cumulative percentages of the observed variable distribution. The 20-, 60-, 90-, 95- and 99-th percentiles

Figure 4.4

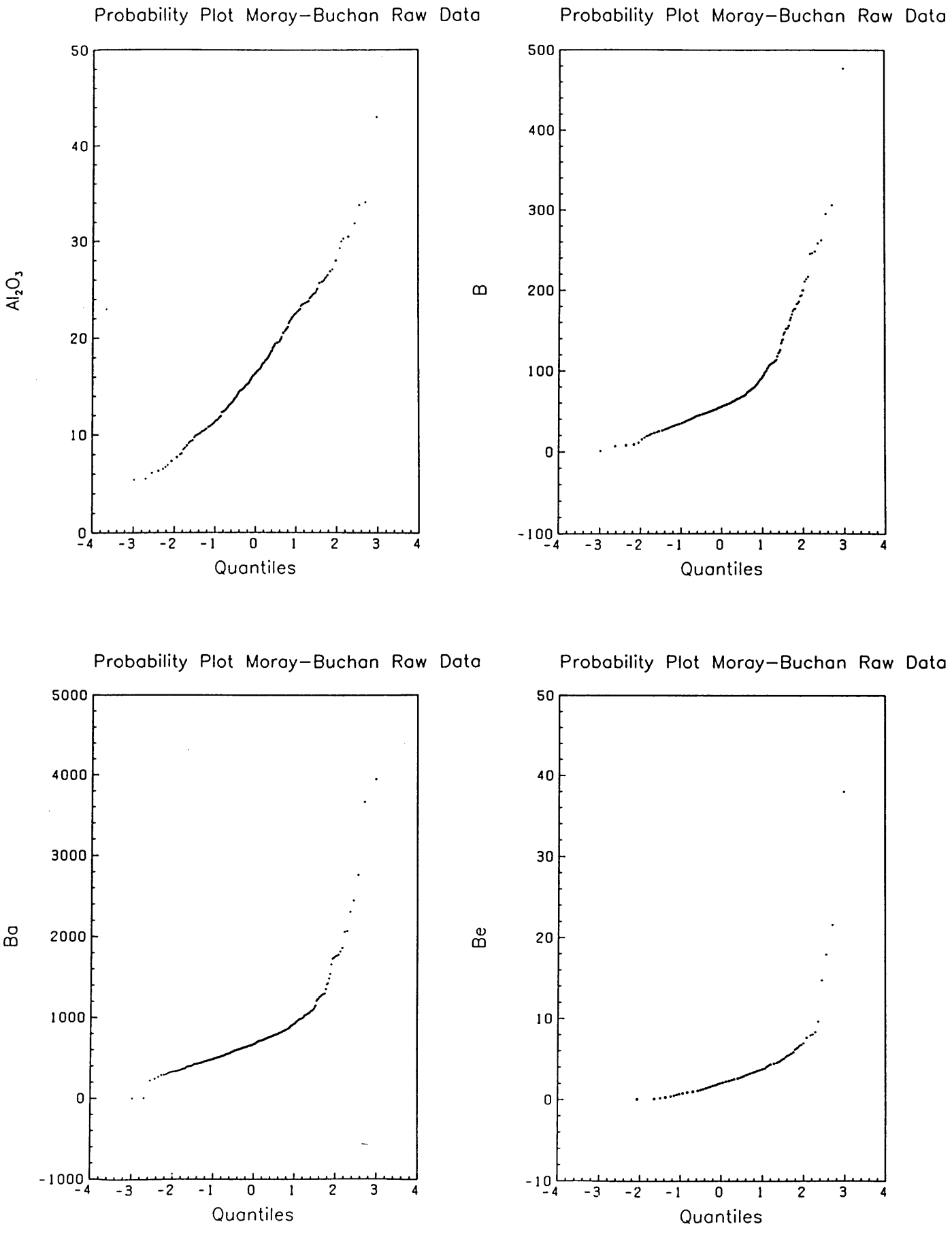
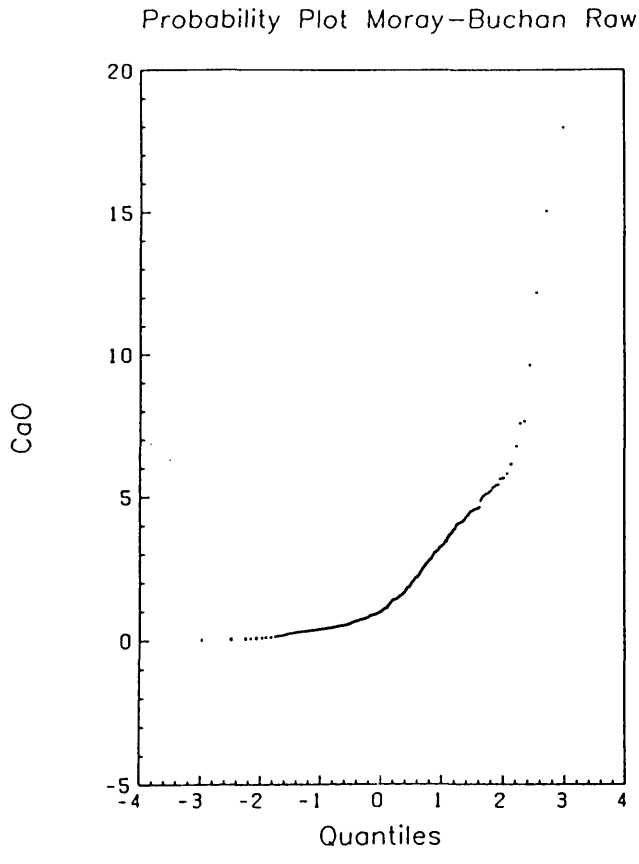
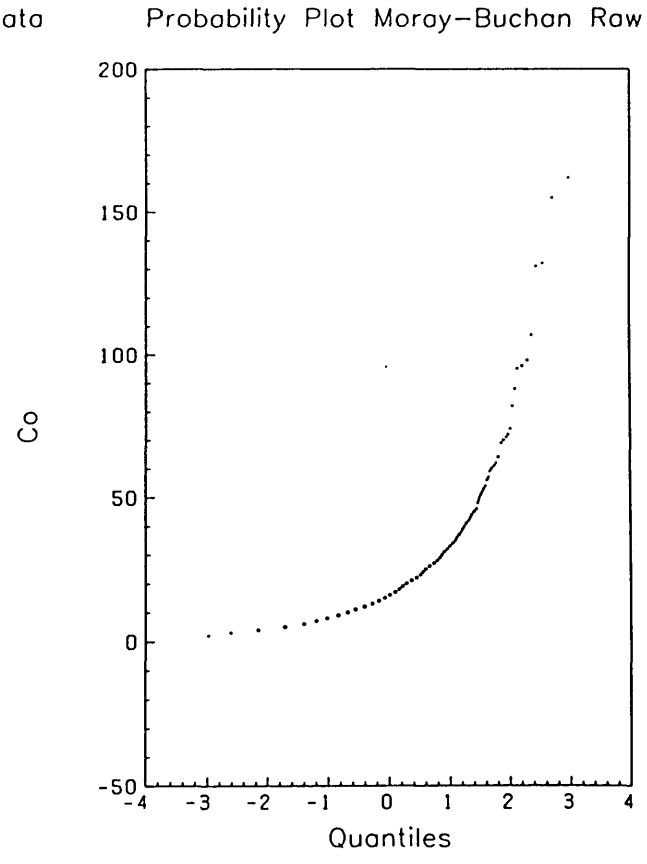
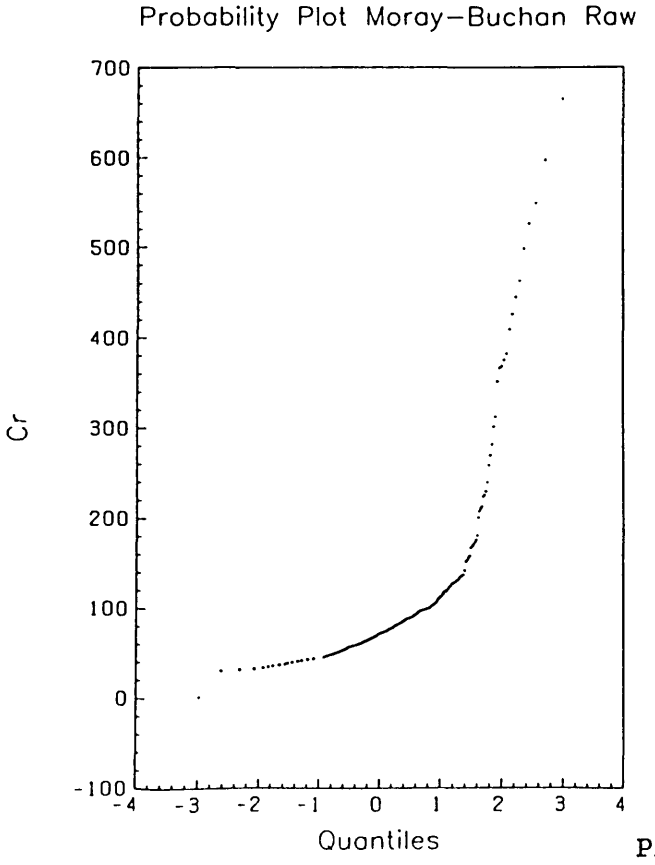

Probability Plot Moray–Buchan Raw Data
Al_2O_3
Quantiles
Probability Plot Moray–Buchan Raw Data
B
Quantiles
Probability Plot Moray–Buchan Raw Data
Ba
Quantiles
Probability Plot Moray–Buchan Raw Data
Be
Quantiles

Figure 4.5


Probability Plot Moray-Buchan Raw Data
CaO
20
15
10
5
0
-5
-4 -3 -2 -1 0 1 2 3 4
Quantiles


Probability Plot Moray-Buchan Raw Data
Co
200
150
100
50
0
-50
-4 -3 -2 -1 0 1 2 3 4
Quantiles


Probability Plot Moray-Buchan Raw Data
Cr
700
600
500
400
300
200
100
0
-100
-4 -3 -2 -1 0 1 2 3 4
Quantiles

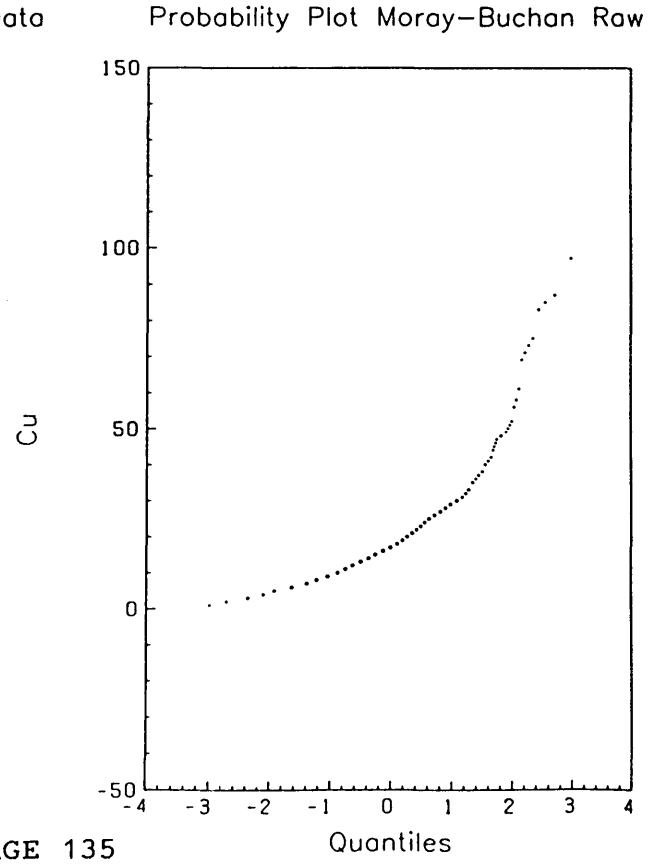

Probability Plot Moray-Buchan Raw Data
Cu
150
100
50
0
-50
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 4.6

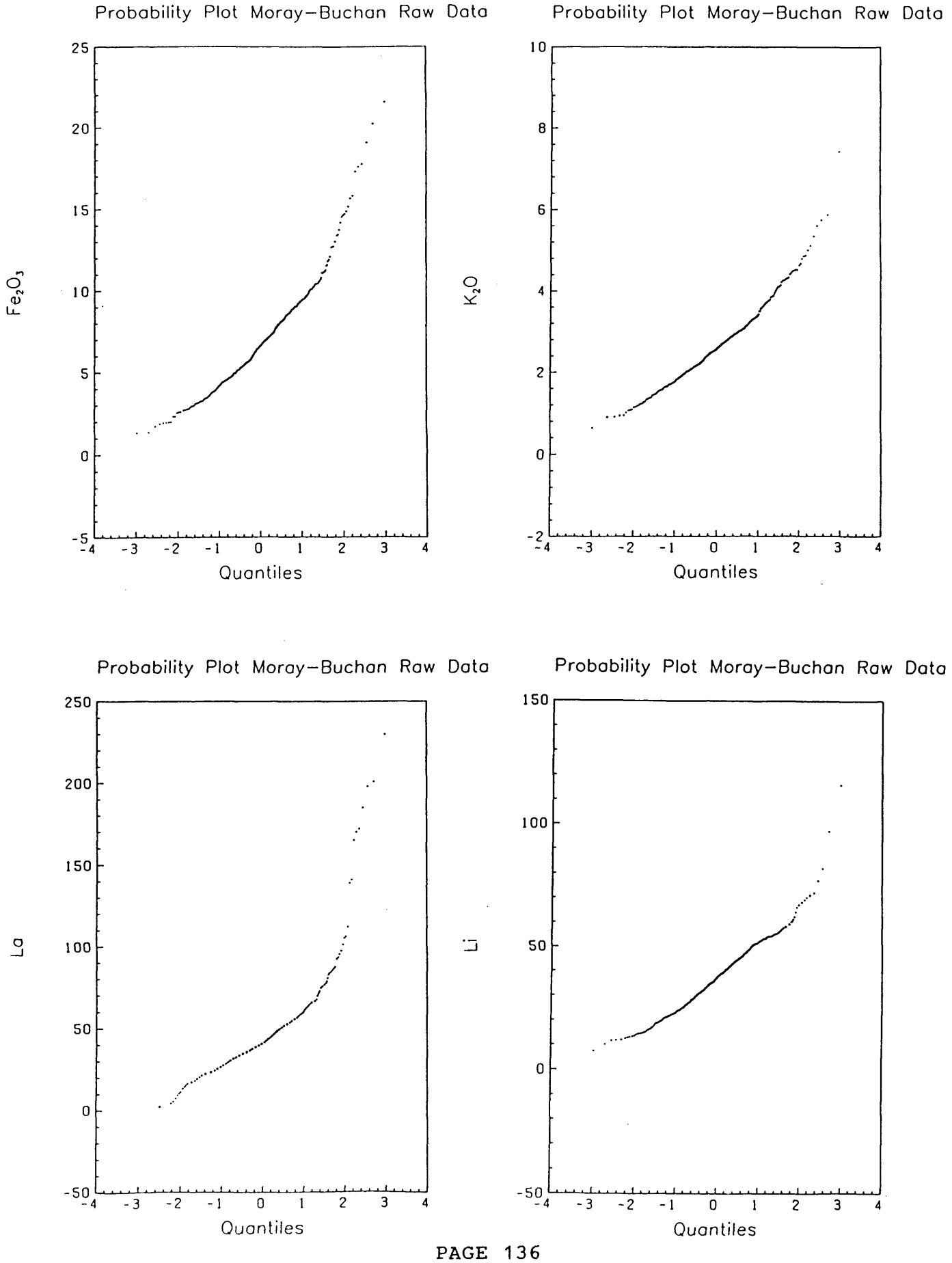
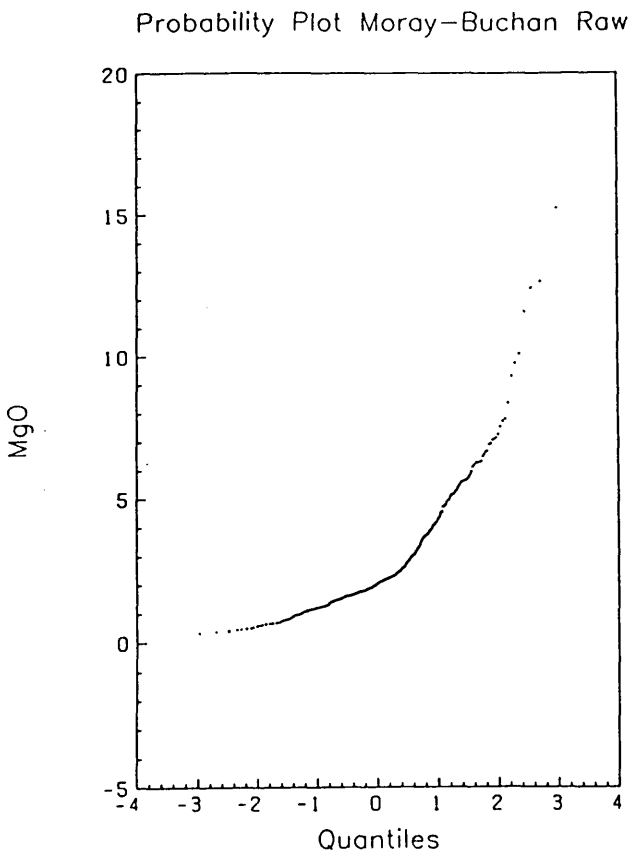
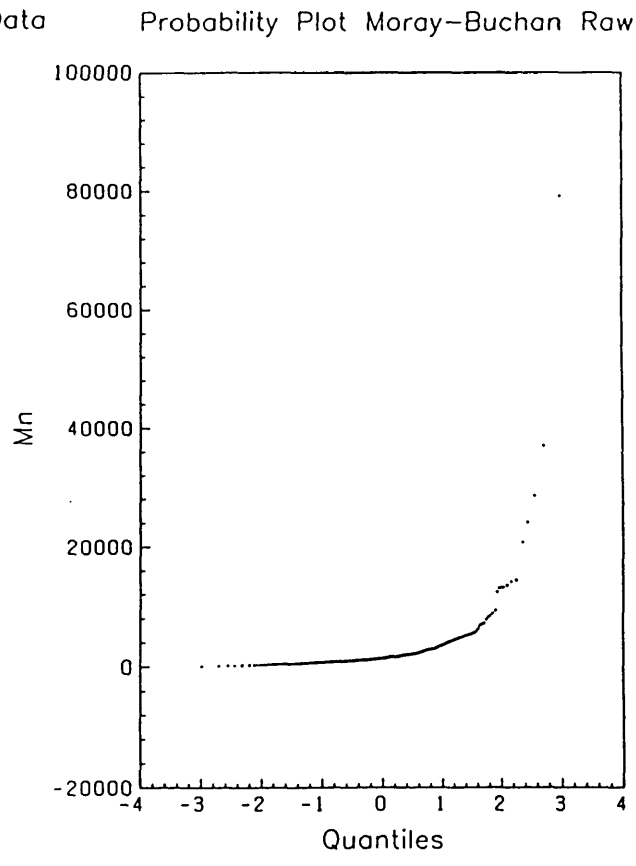
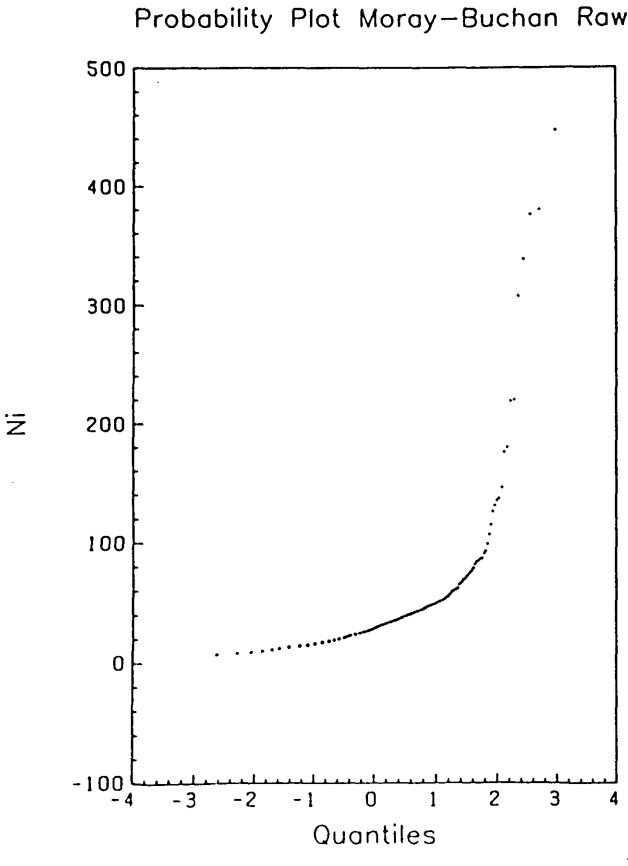

Probability Plot Moray–Buchan Raw Data
Fe2O3
Quantiles
Probability Plot Moray–Buchan Raw Data
K2O
Quantiles
Probability Plot Moray–Buchan Raw Data
La
Quantiles
Probability Plot Moray–Buchan Raw Data
Li
Quantiles

Figure 4.7


Probability Plot Moray-Buchan Raw Data
MgO
20
15
10
5
0
-5
-4 -3 -2 -1 0 1 2 3 4
Quantiles


Probability Plot Moray-Buchan Raw Data
Mn
100000
80000
60000
40000
20000
0
-20000
-4 -3 -2 -1 0 1 2 3 4
Quantiles


Probability Plot Moray-Buchan Raw Data
Ni
500
400
300
200
100
0
-100
-4 -3 -2 -1 0 1 2 3 4
Quantiles

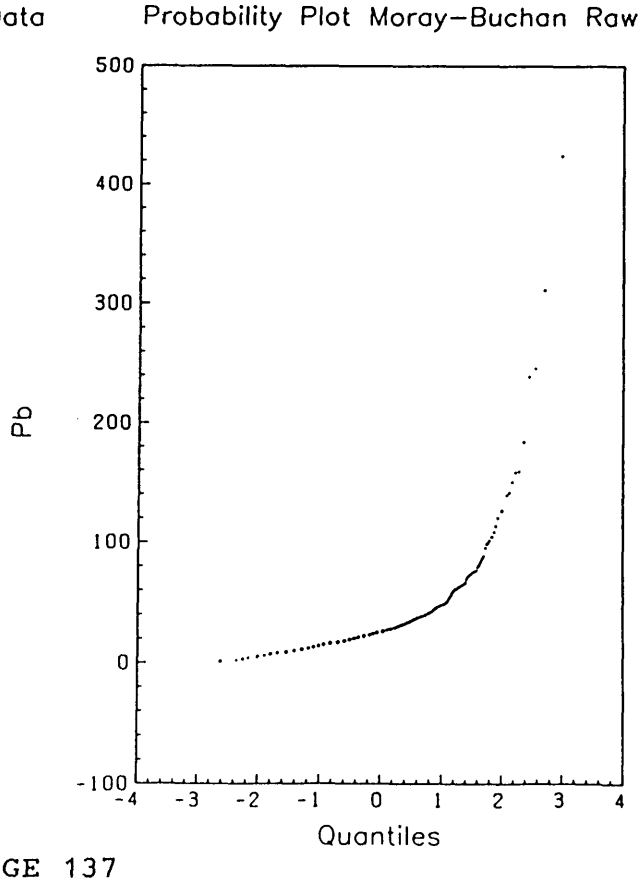

Probability Plot Moray-Buchan Raw Data
Pb
500
400
300
200
100
0
-100
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 4.8

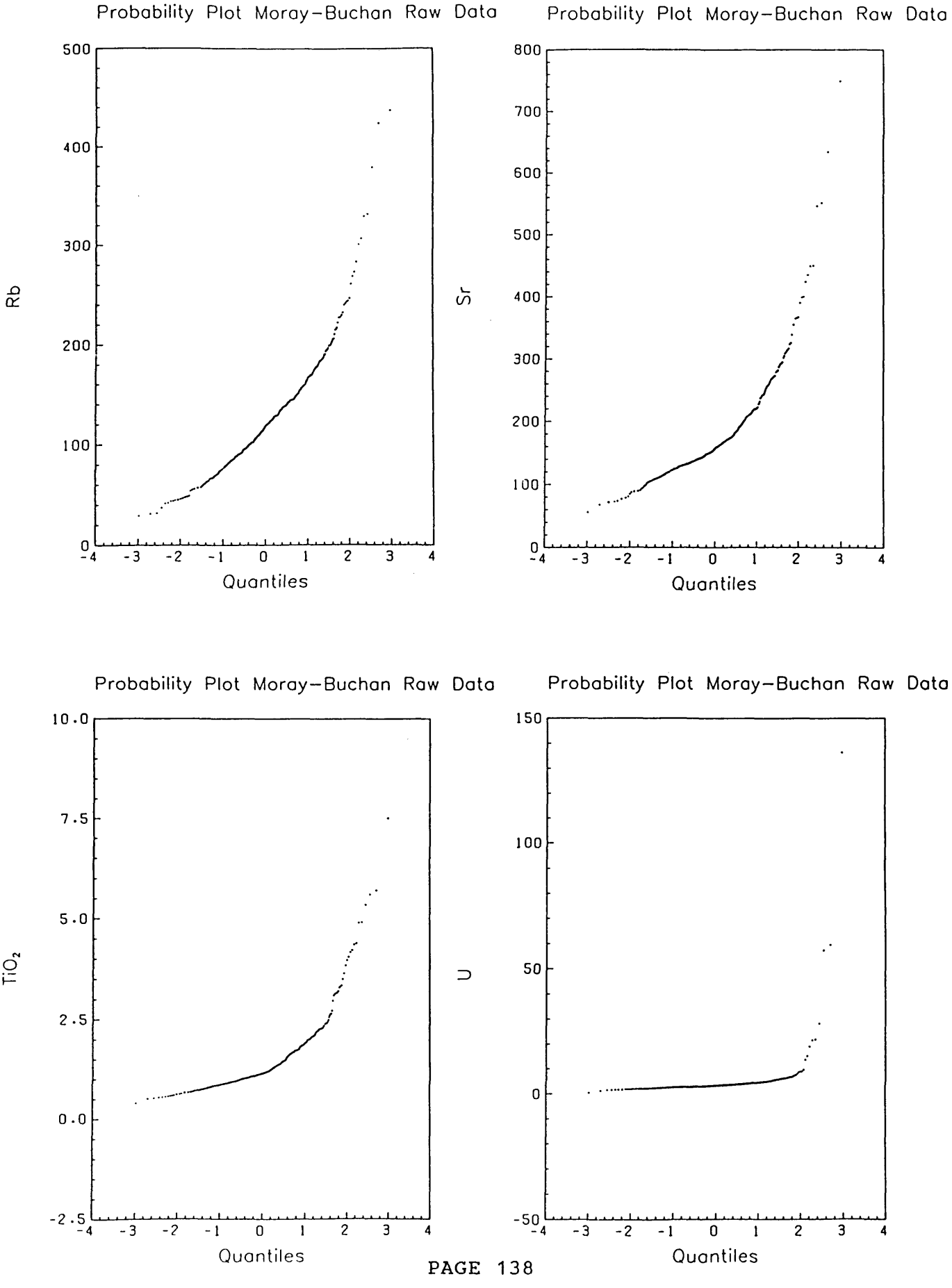
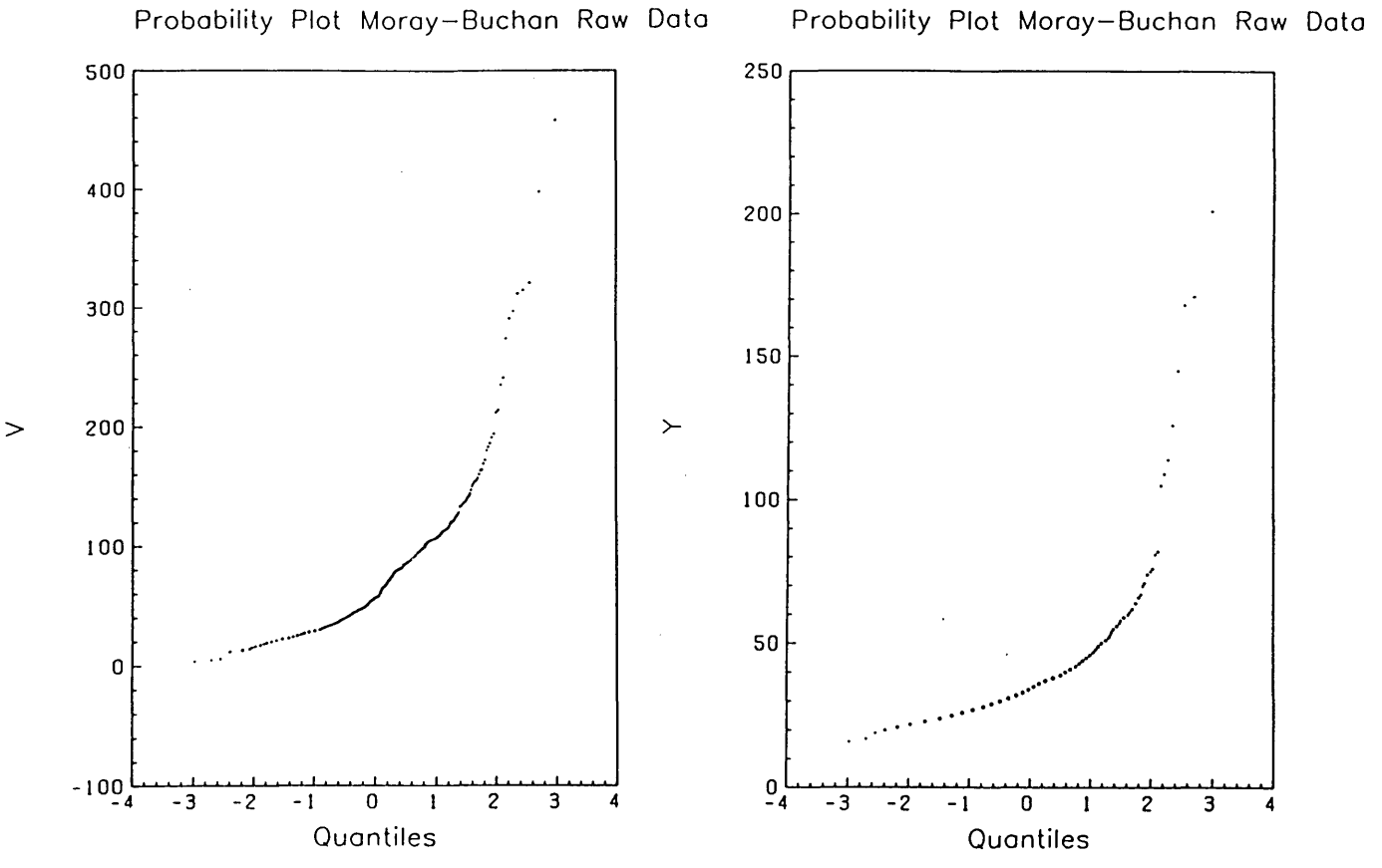
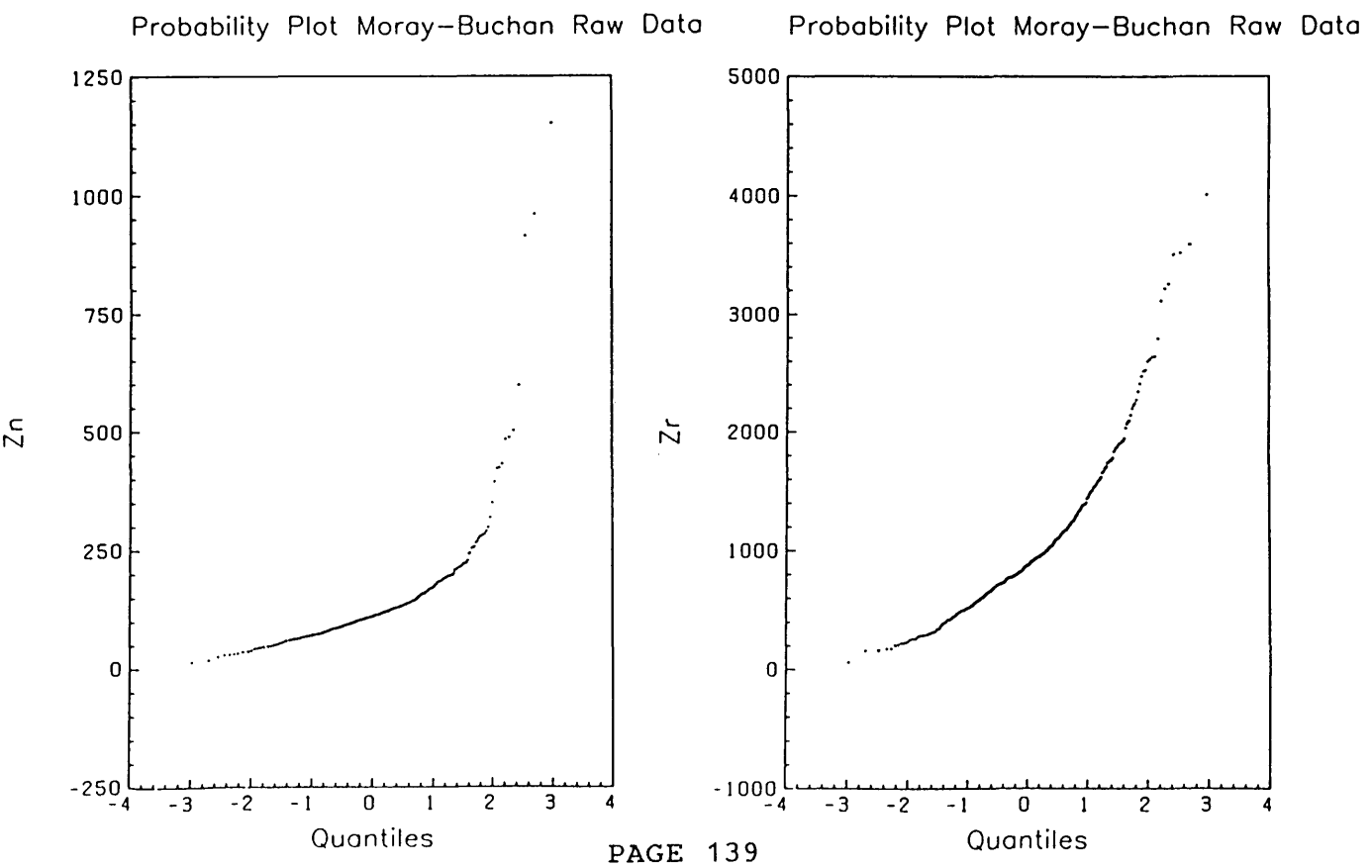

Probability Plot Moray-Buchan Raw Data
Rb
500
400
300
200
100
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Raw Data
Sr
800
700
600
500
400
300
200
100
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Raw Data
TiO2
10.0
7.5
5.0
2.5
0.0
-2.5
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Raw Data
U
150
100
50
0
-50
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 4.9


Probability Plot Moray-Buchan Raw Data
V
500
400
300
200
100
0
-100
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Raw Data
Y
250
200
150
100
50
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles


Probability Plot Moray-Buchan Raw Data
Zn
1250
1000
750
500
250
0
-250
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Raw Data
Zr
5000
4000
3000
2000
1000
0
-1000
-4 -3 -2 -1 0 1 2 3 4
Quantiles

were used to define six classes. The quantiles of the standardized normal distribution N(0,1) corresponding to the above percentiles are -0.84, 0.25, 1.28, 1.65 and 2.32. The symbols assigned to the six classes were chosen so that as the concentration level rises the visual presence of the symbol becomes increasingly dominant. The lowest concentration class (<20%) is represented by a small plus sign; the class 20%-60% is represented by a cross of the same size. The four subsequent classes are represented by open circles of increasing radius. The percentiles and symbols were chosen with the aim of drawing the eye towards those sites which display large concentrations and which may form part of an anomalous population. The 24 point symbol maps are reproduced in Figures 4.10 through 4.33 and the major conclusions from their analysis are presented below.

1) B, U, Y and Zr show marked enrichment in the Lower Dalradian. Coats et al (1981) comment that the elements Y, Zr, Be and B are characteristic of the Lower Dalradian, being probably related to the well-sorted shallow water marine nature of the sediments deposited in a shelf sea environment. The enrichment of U in the Lower Dalradian may be due to supply from the Cairngorm and Ben Rinnes granitic bodies. These elements show depletion in the Upper Dalradian.

Figure 4.10

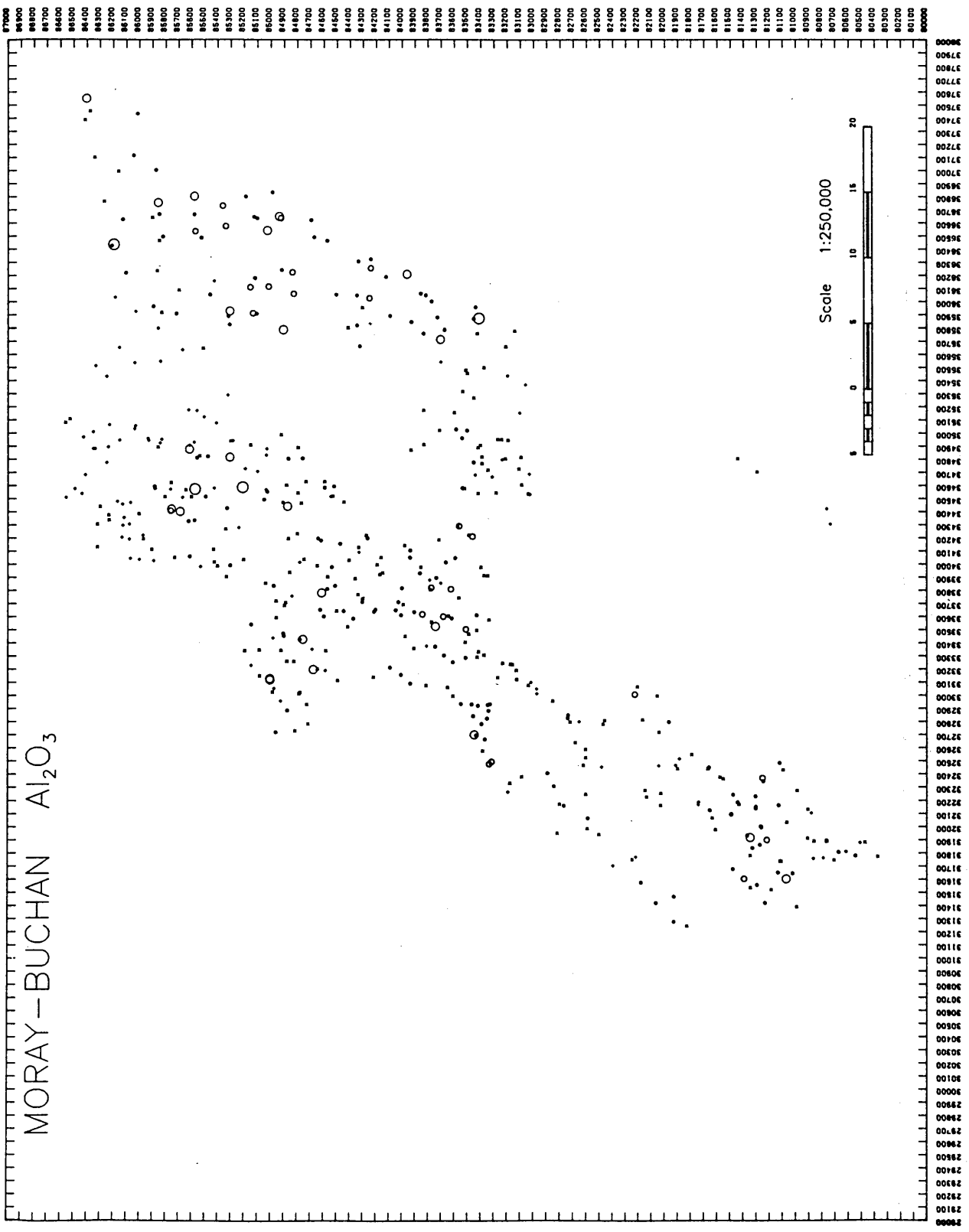


Figure 4.11

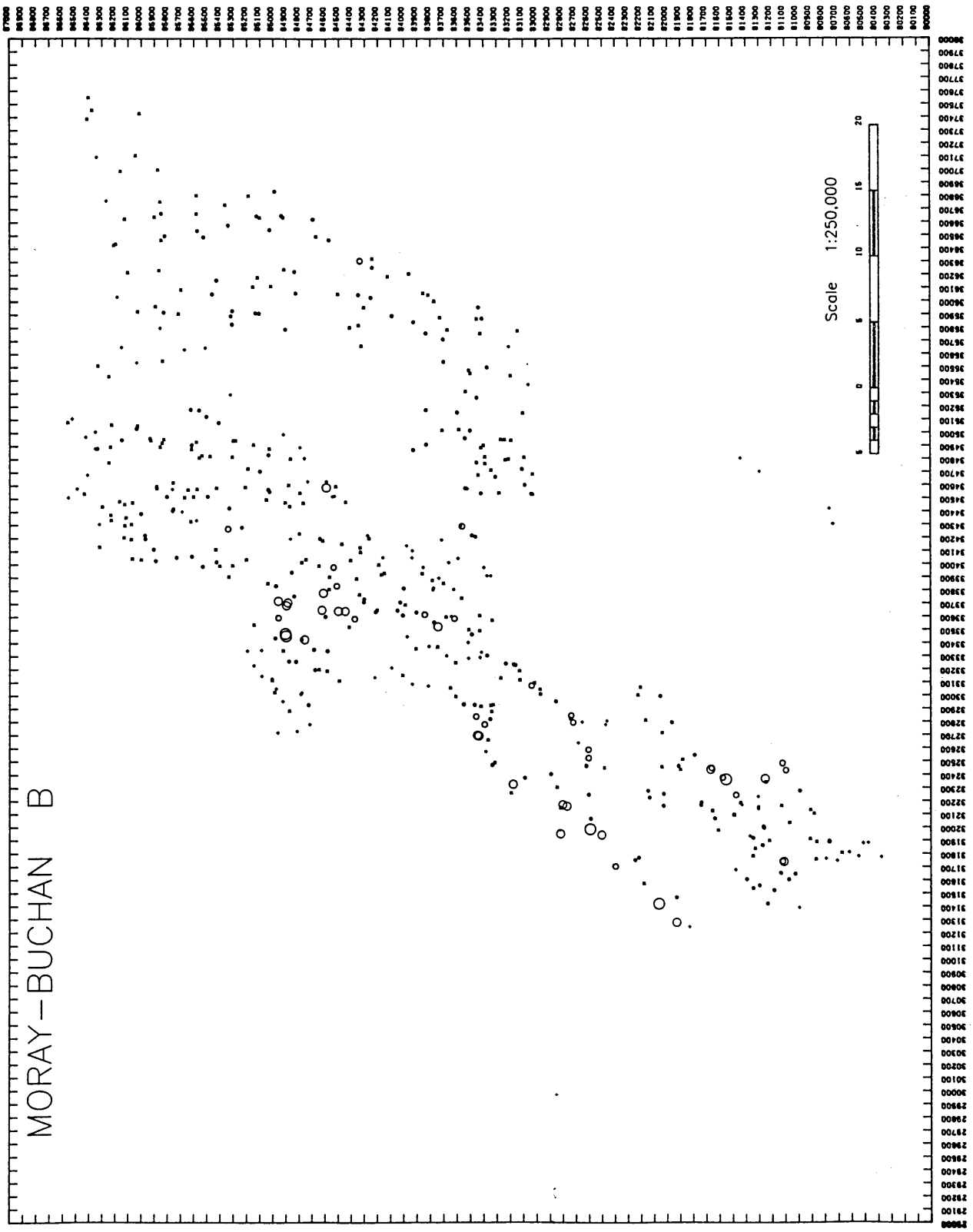


Figure 4.12

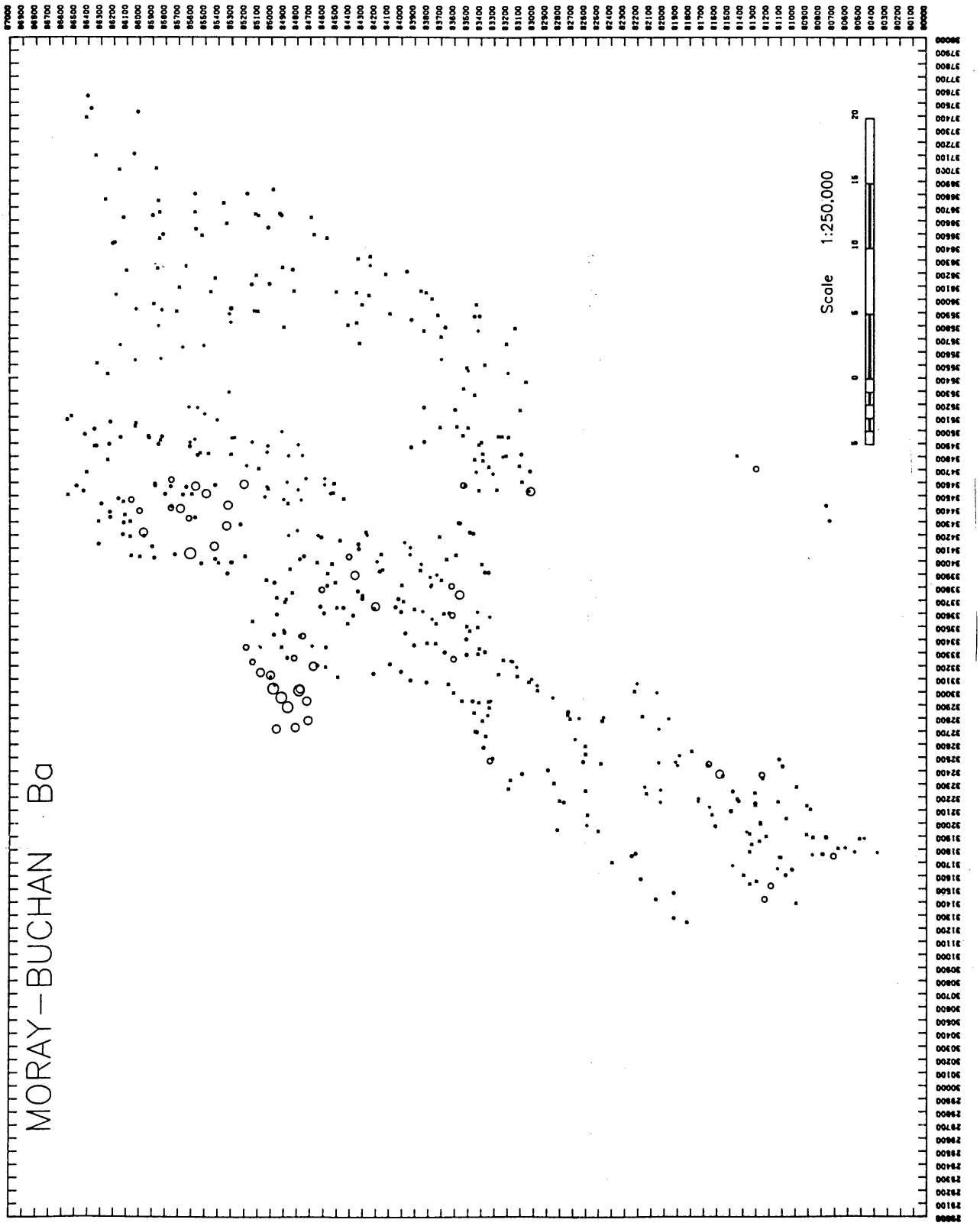


Figure 4.13

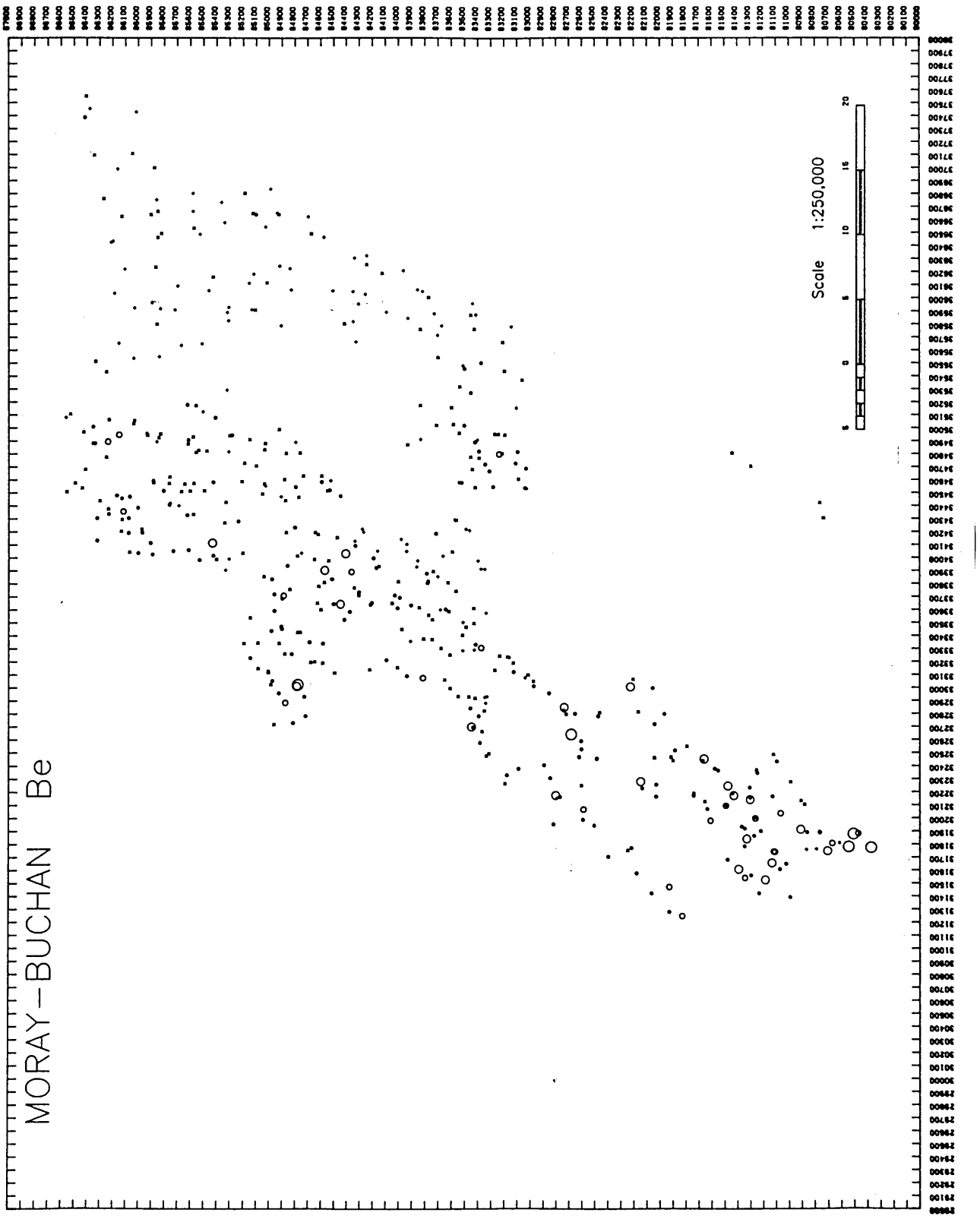


Figure 4.14

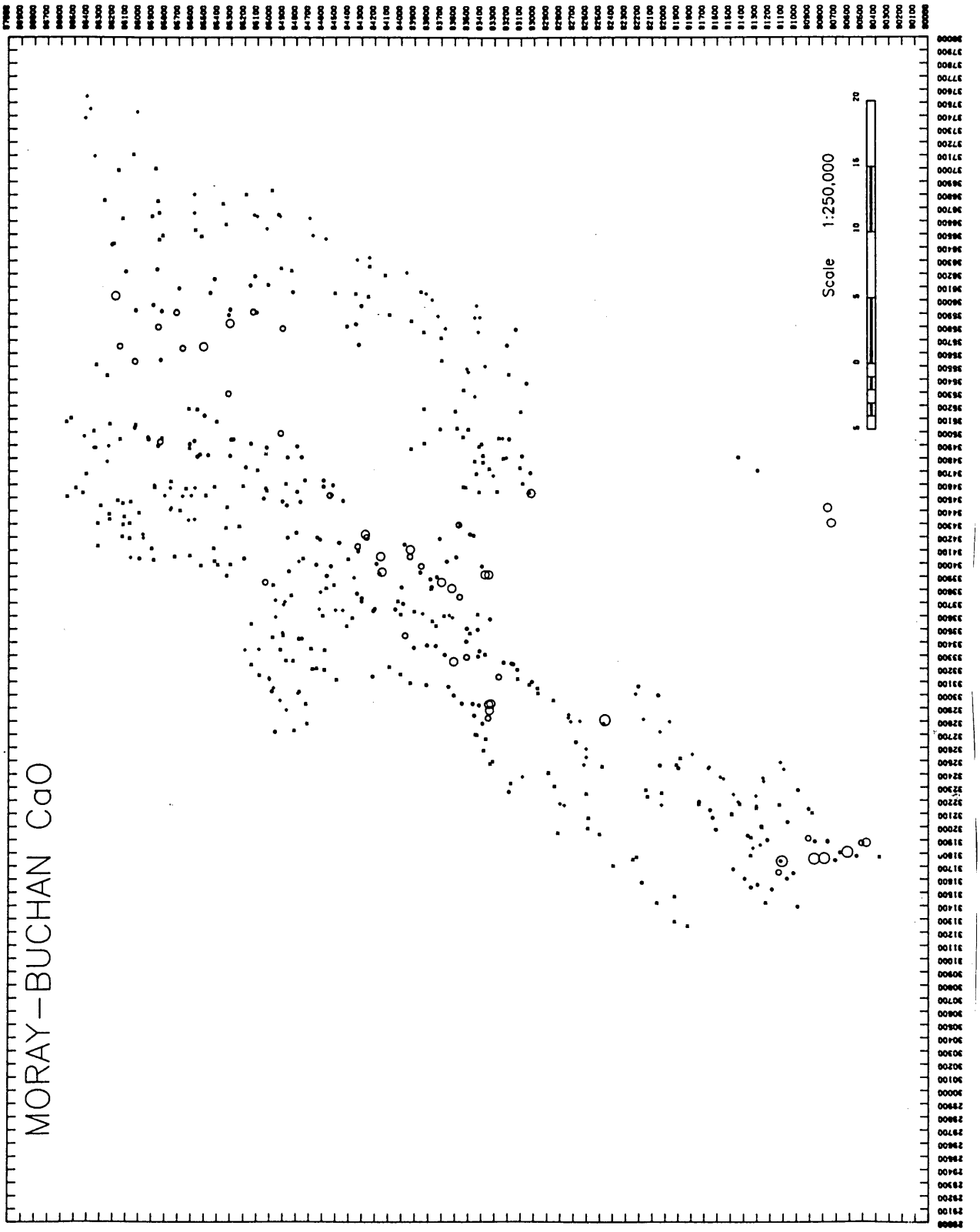


Figure 4.15

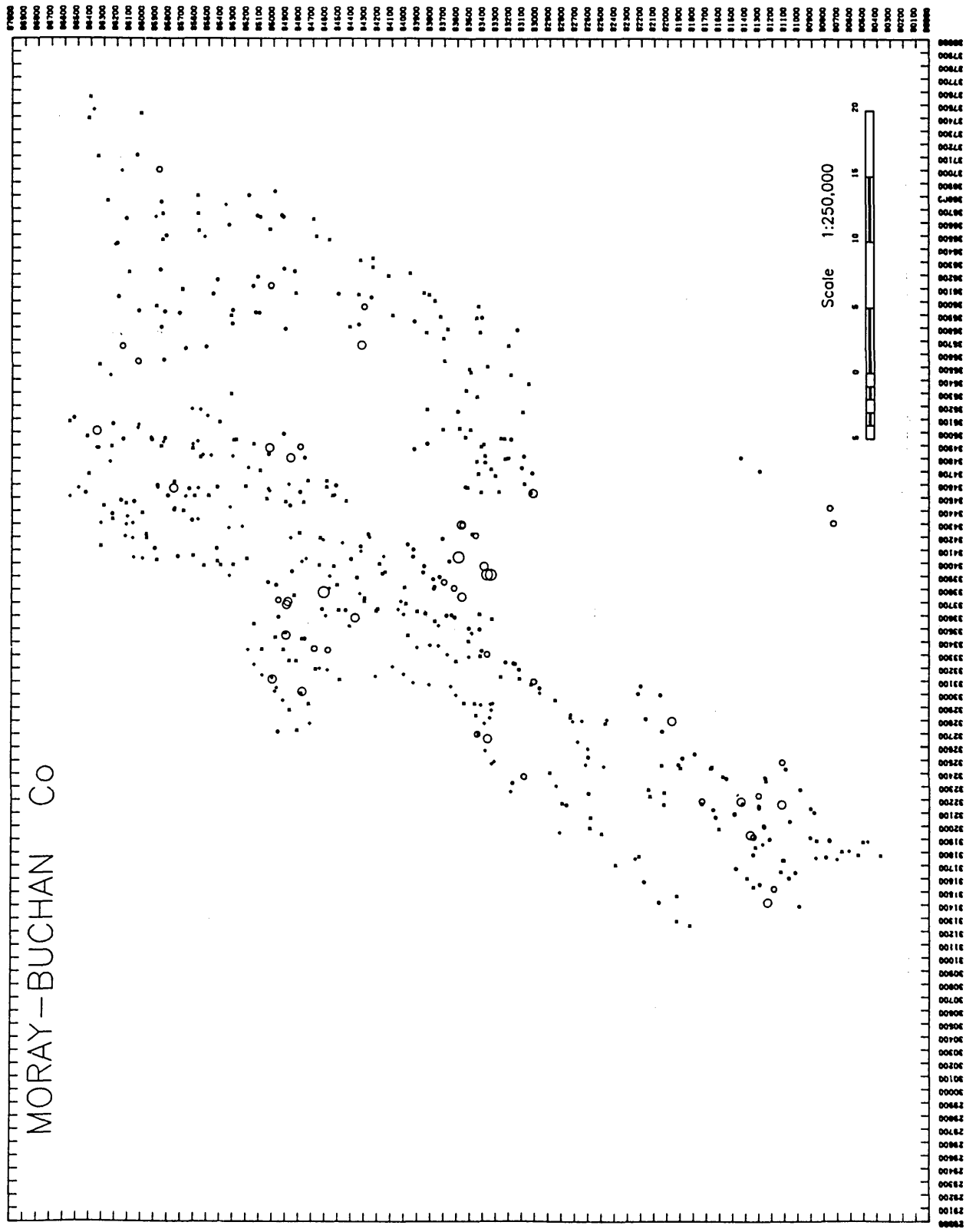


Figure 4.16

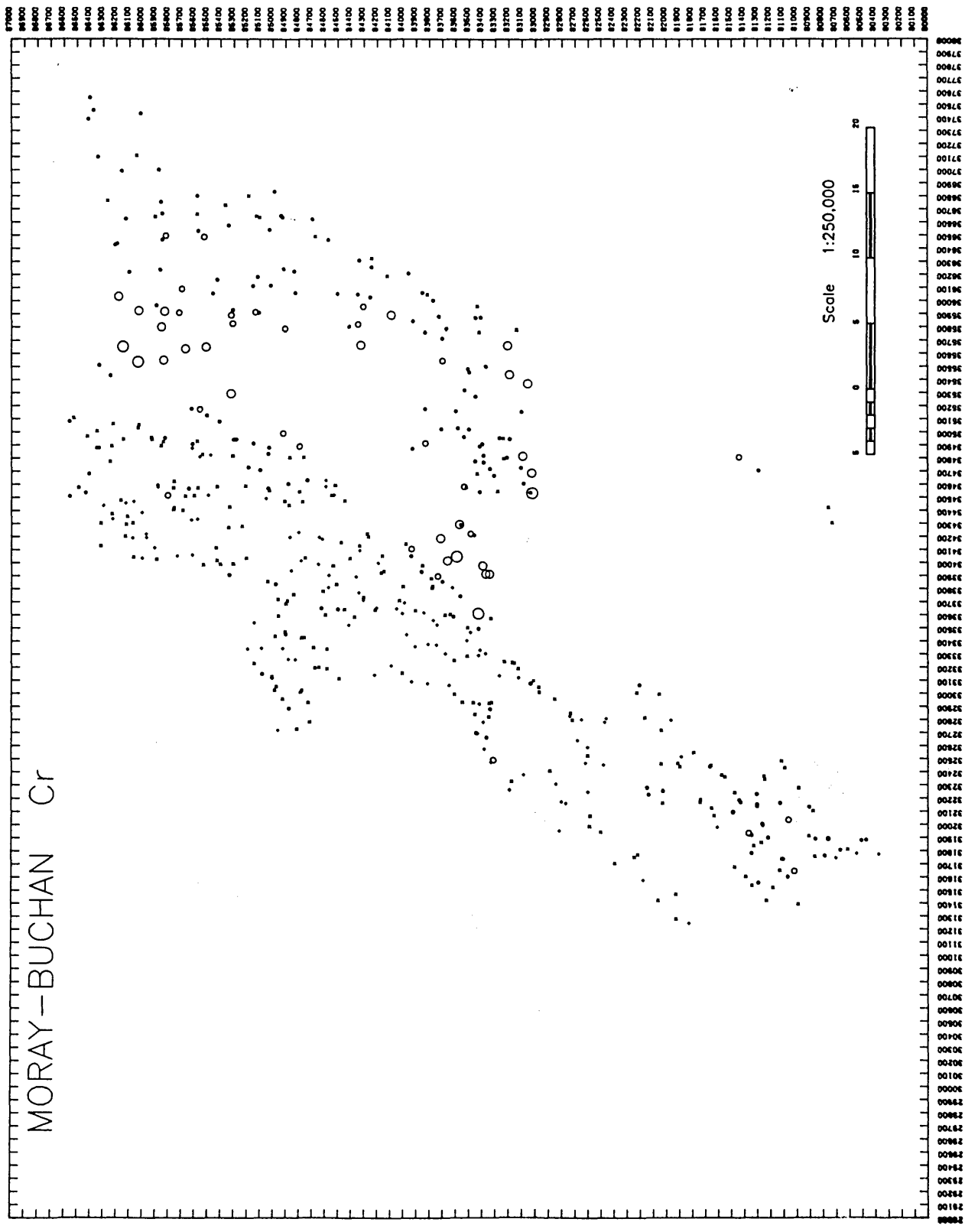


Figure 4.17

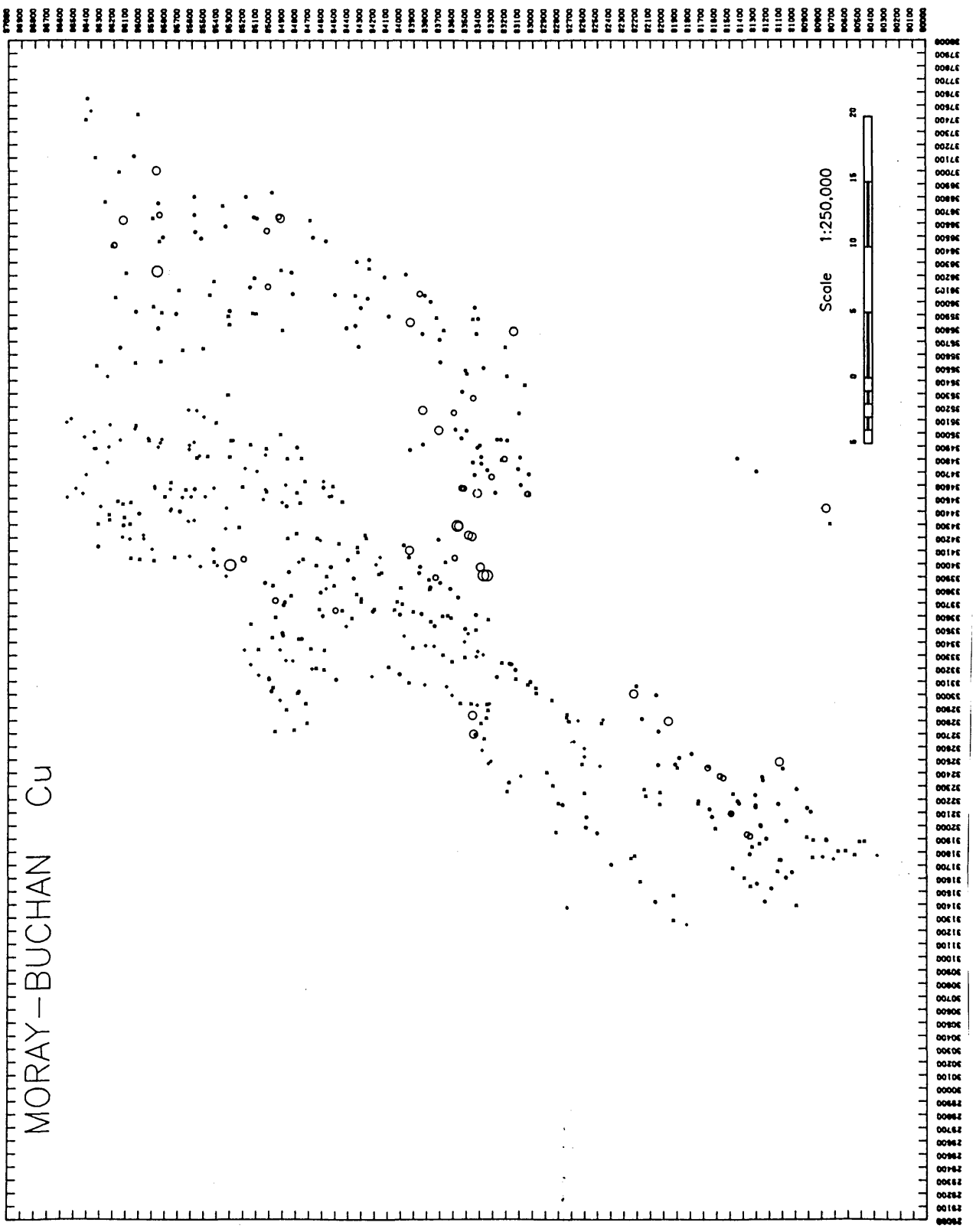


Figure 4.18

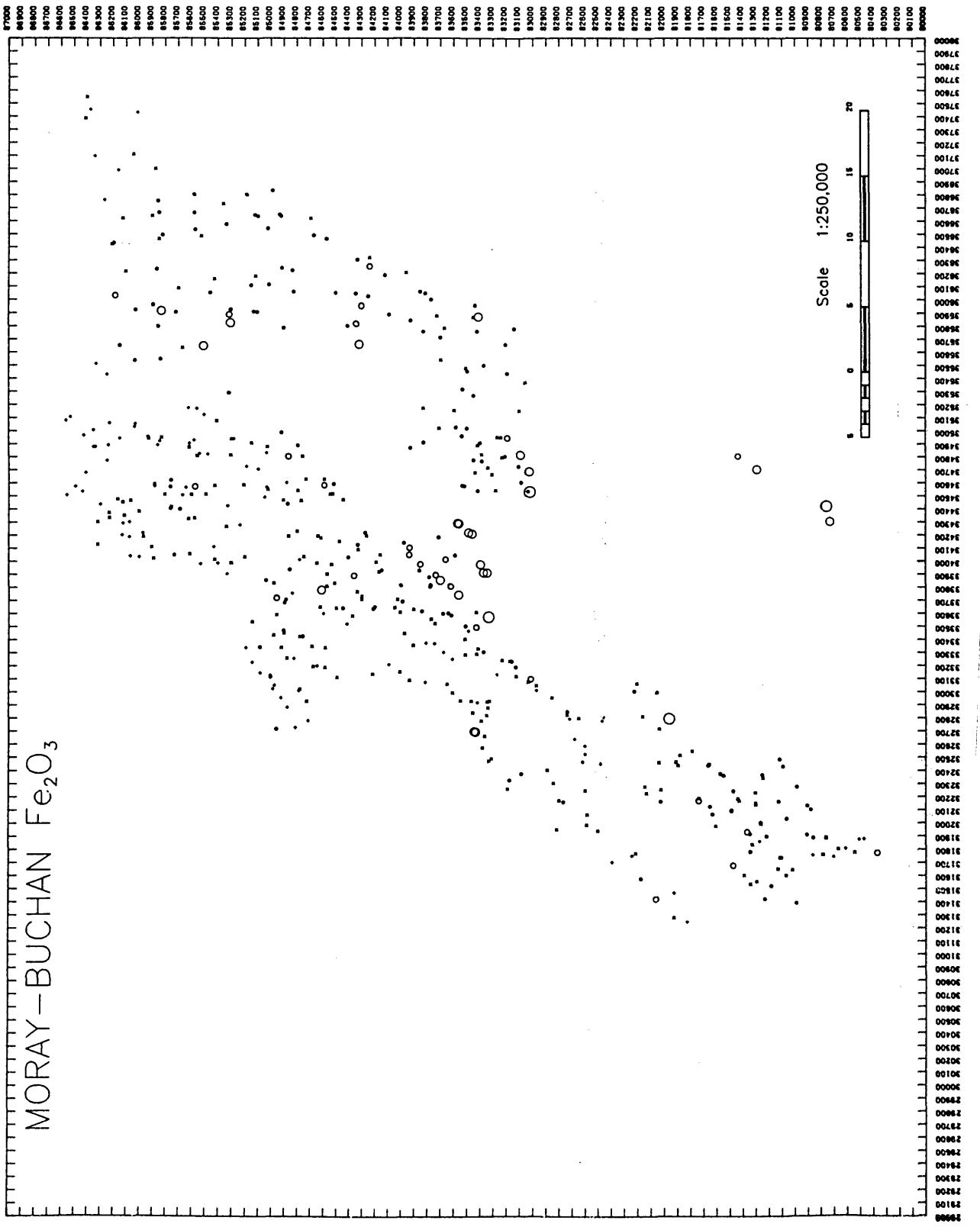


Figure 4.19

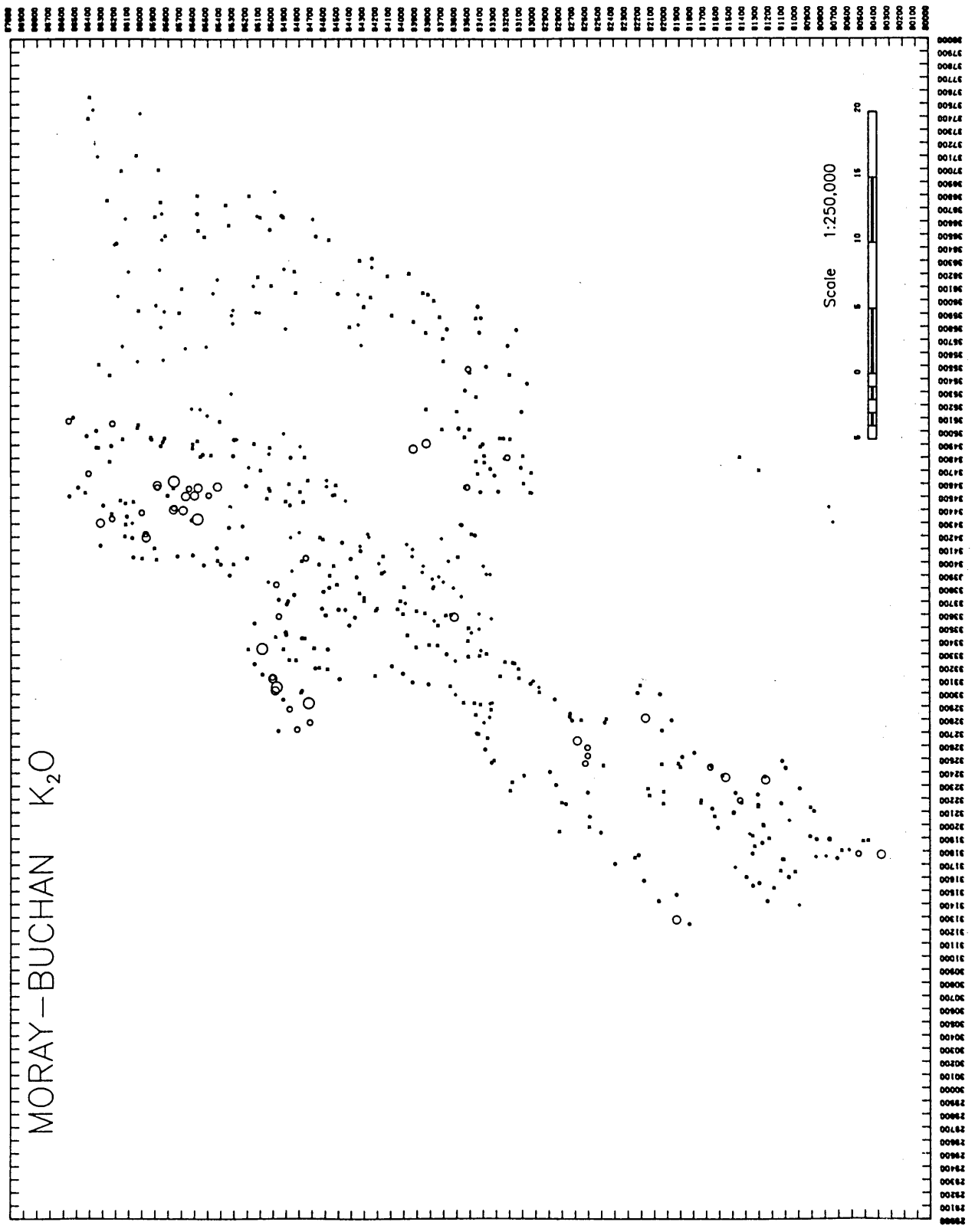


Figure 4.20

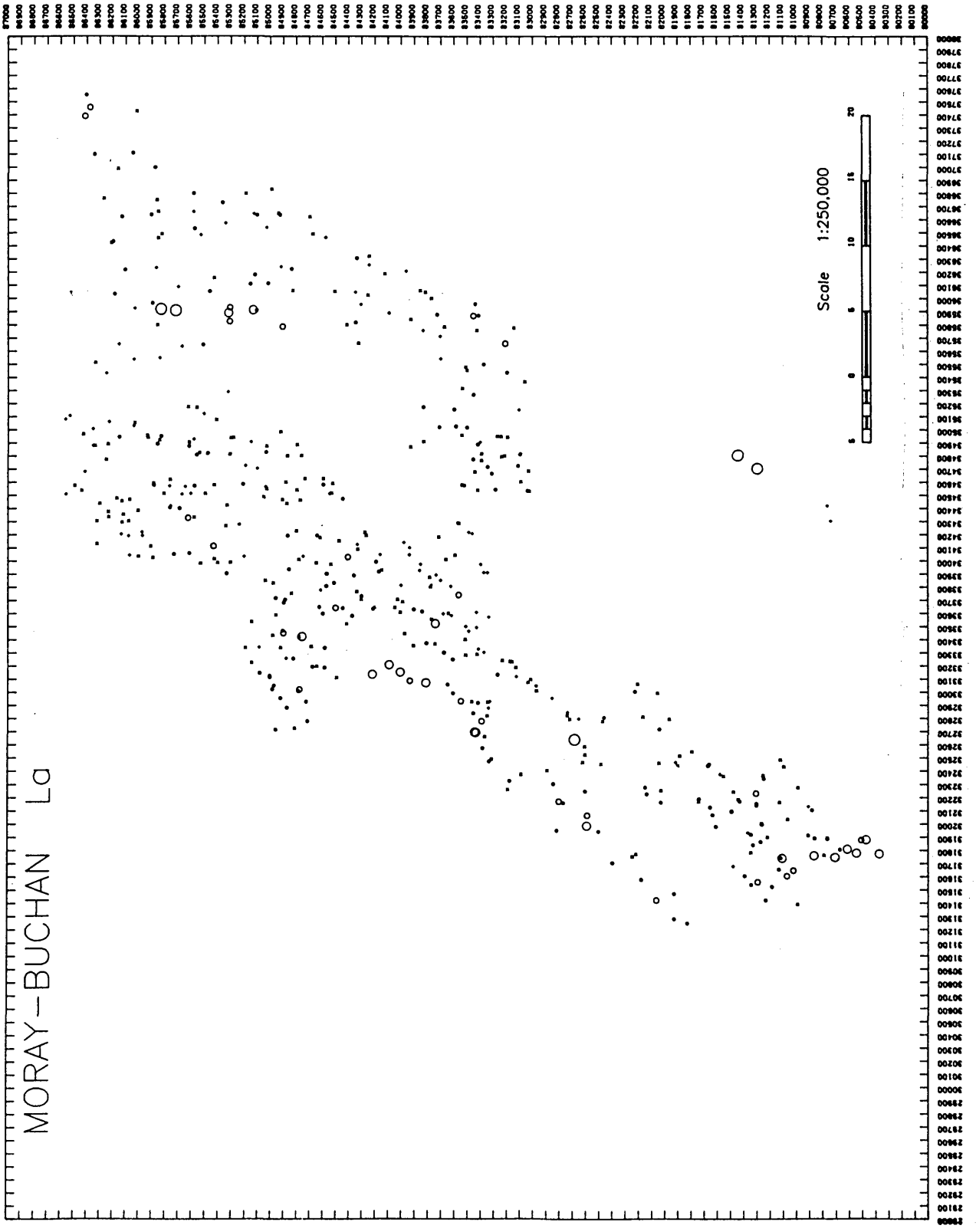


Figure 4.21

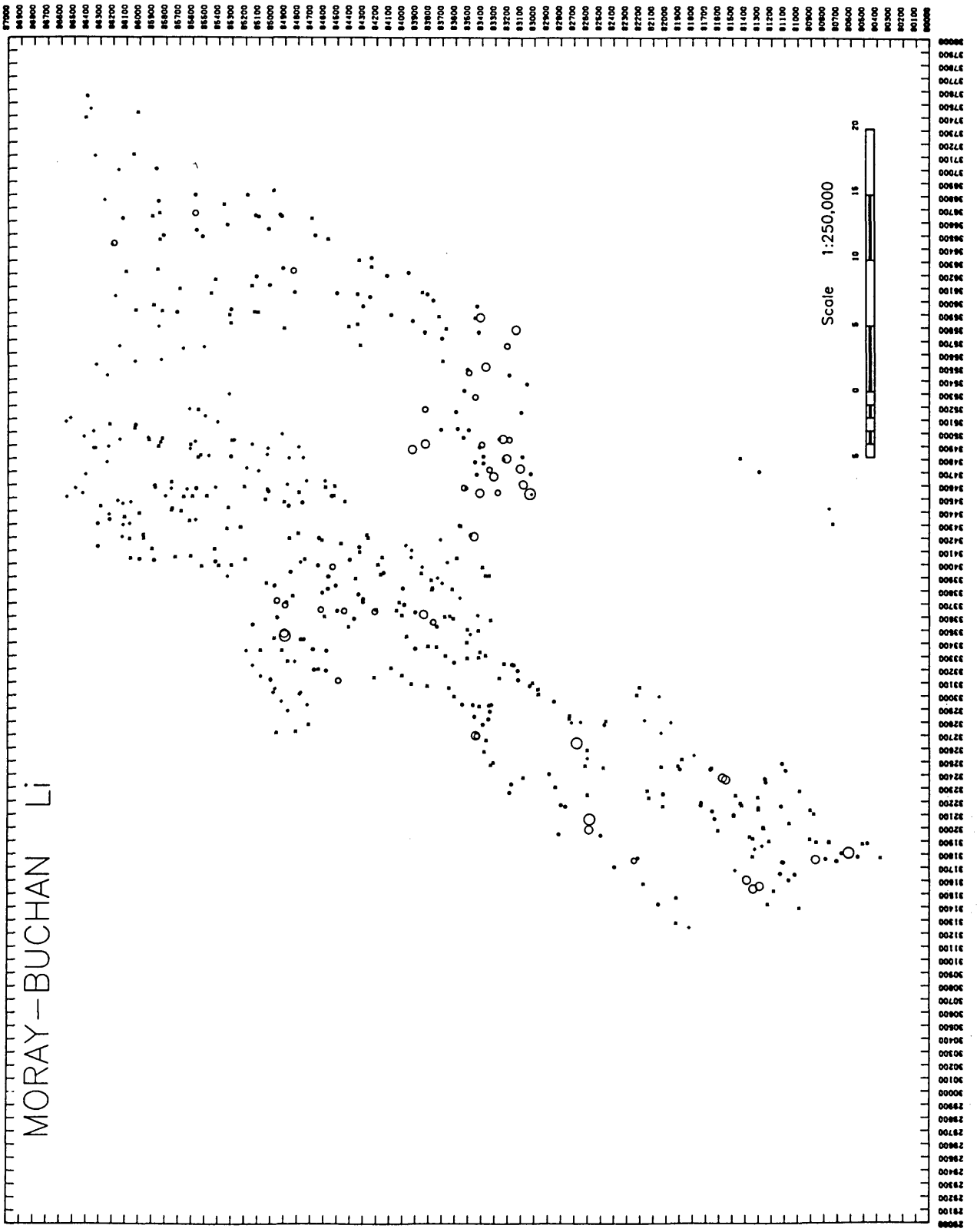


Figure 4.22

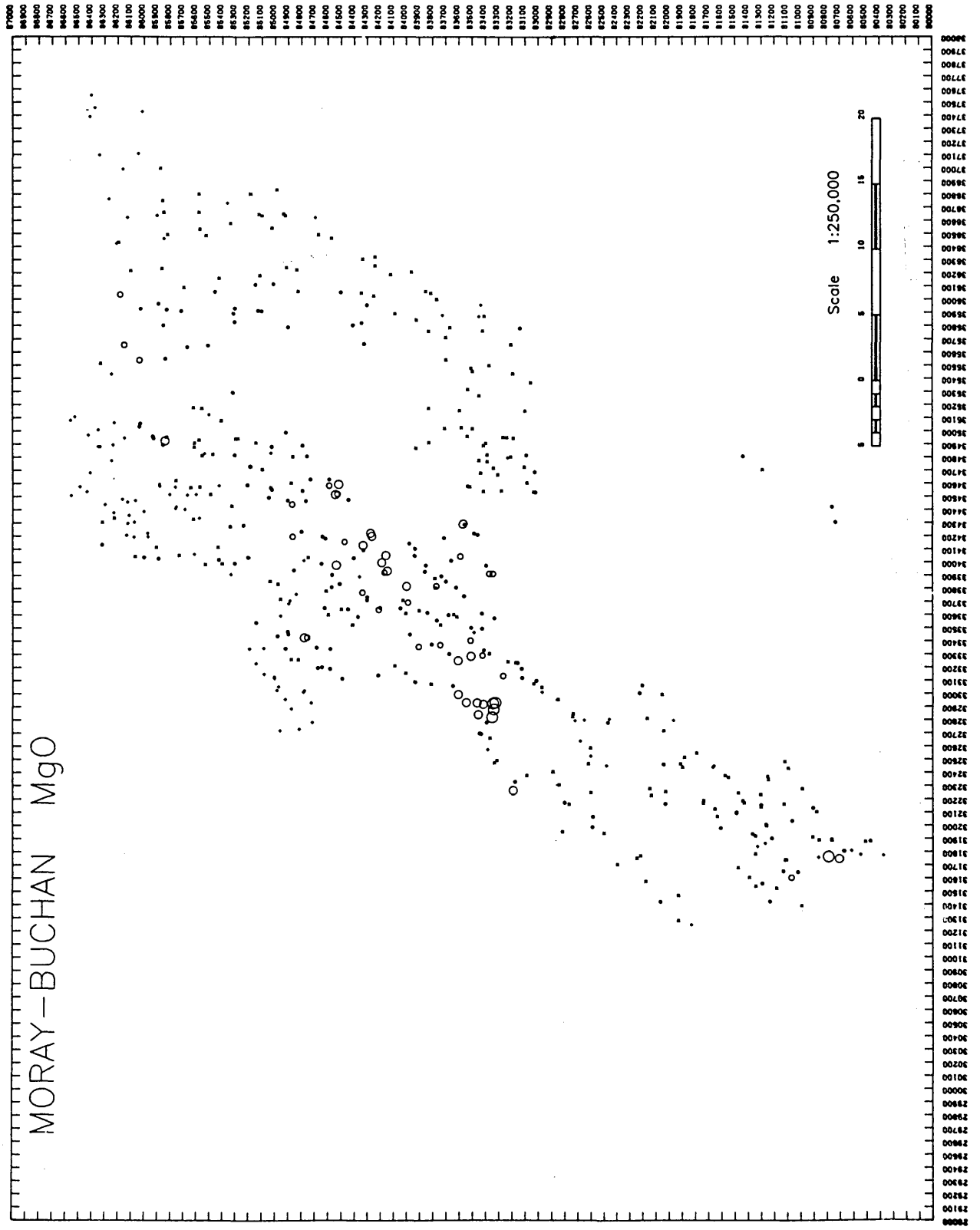


Figure 4.23

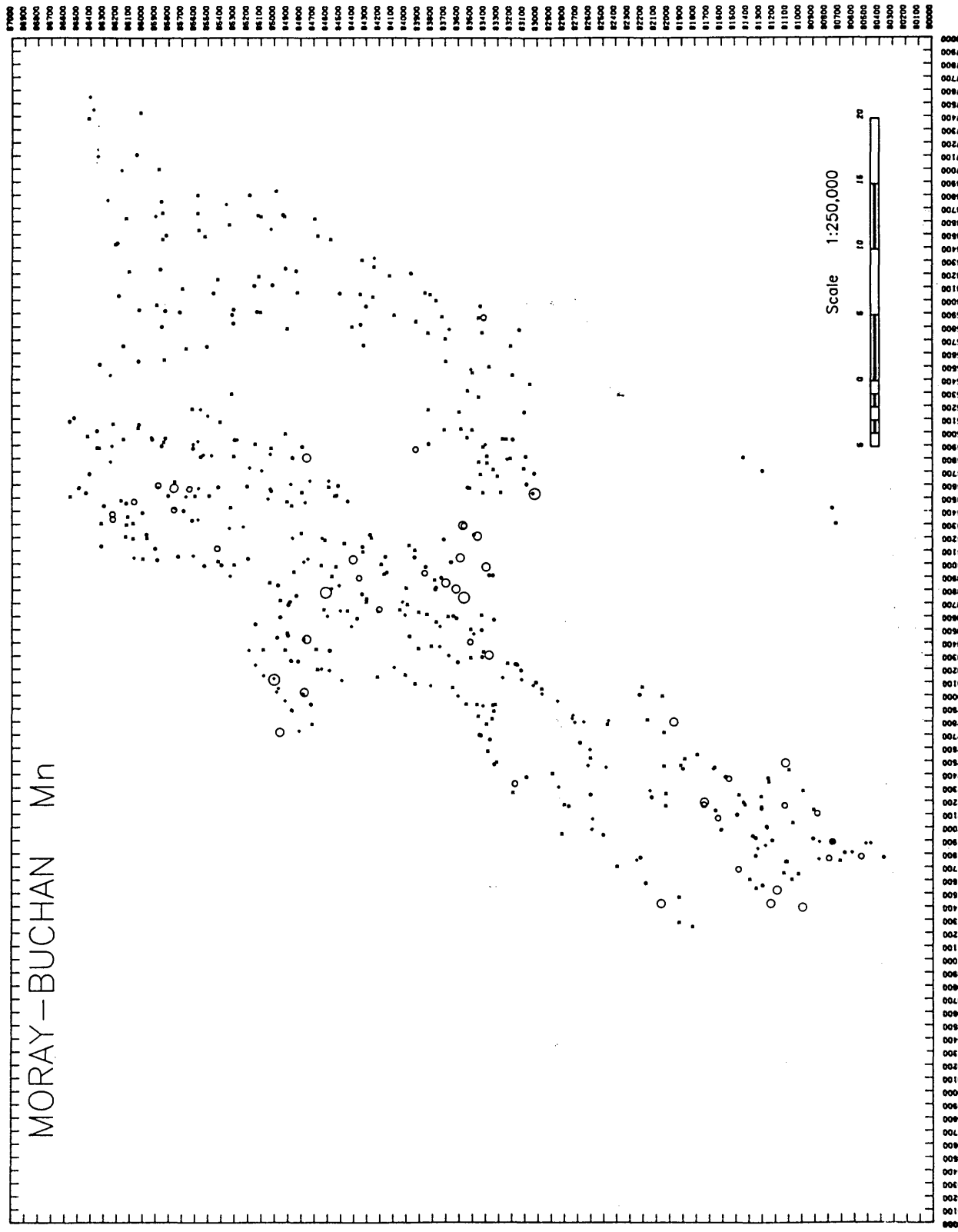


Figure 4.24

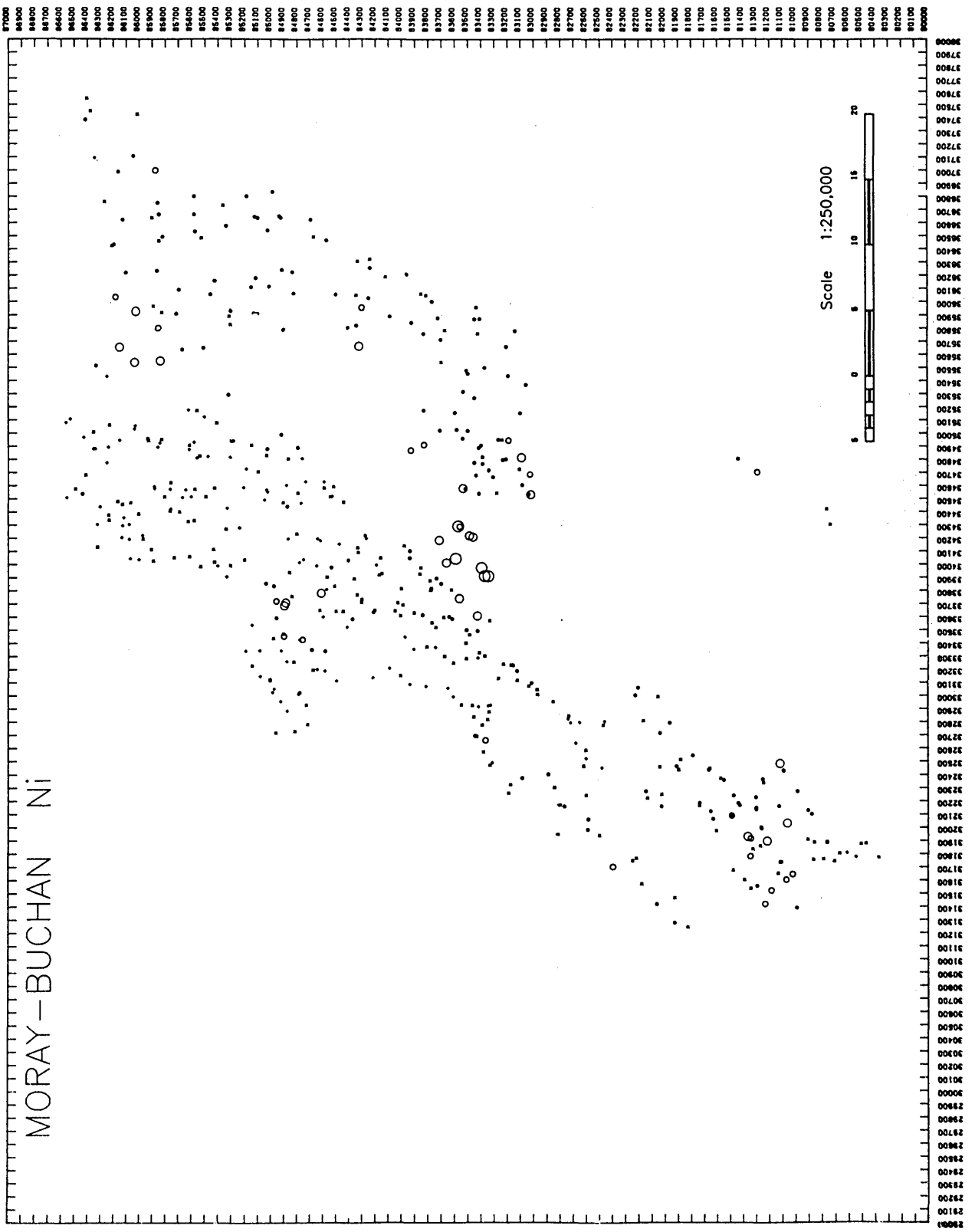


Figure 4.25

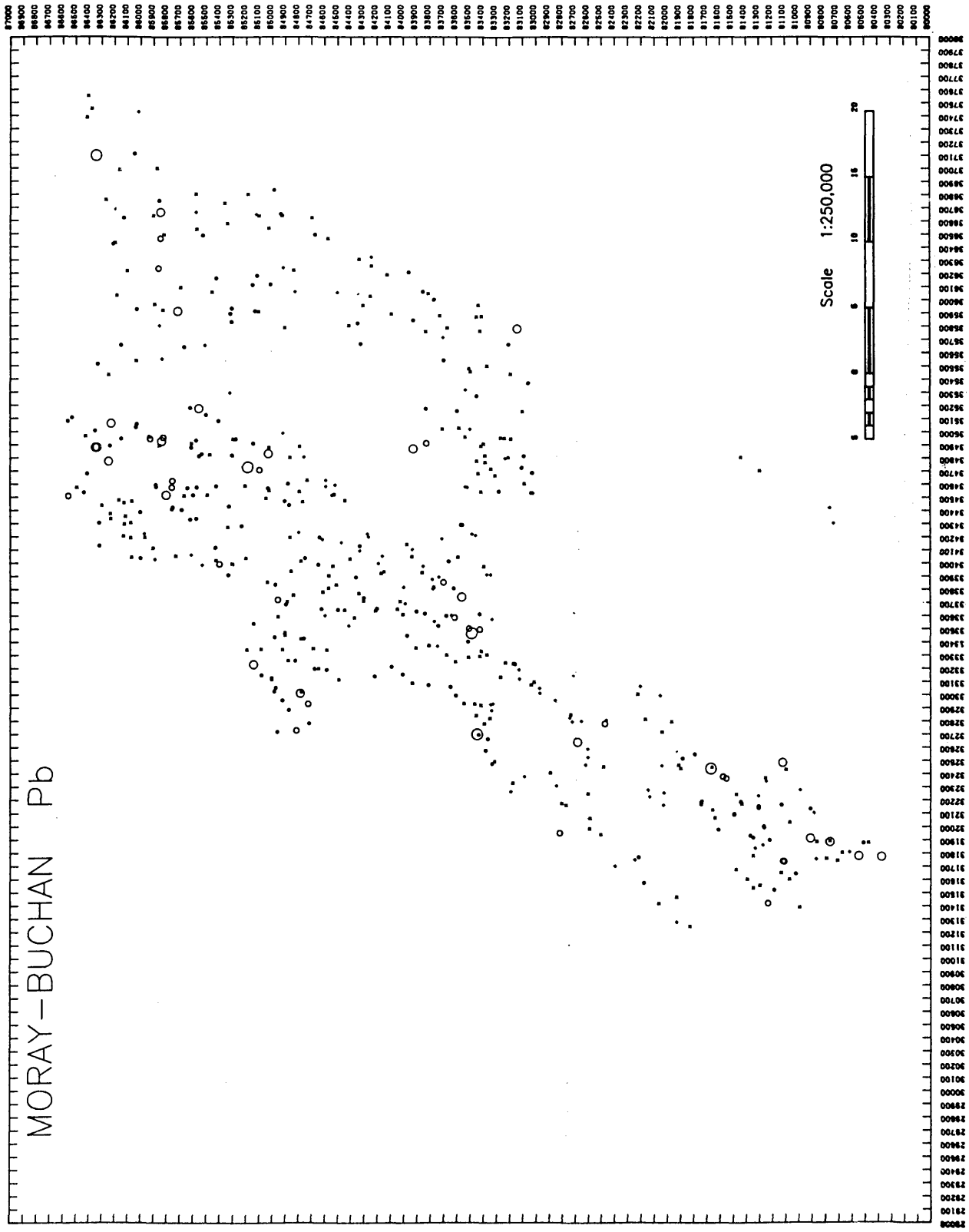


Figure 4.26

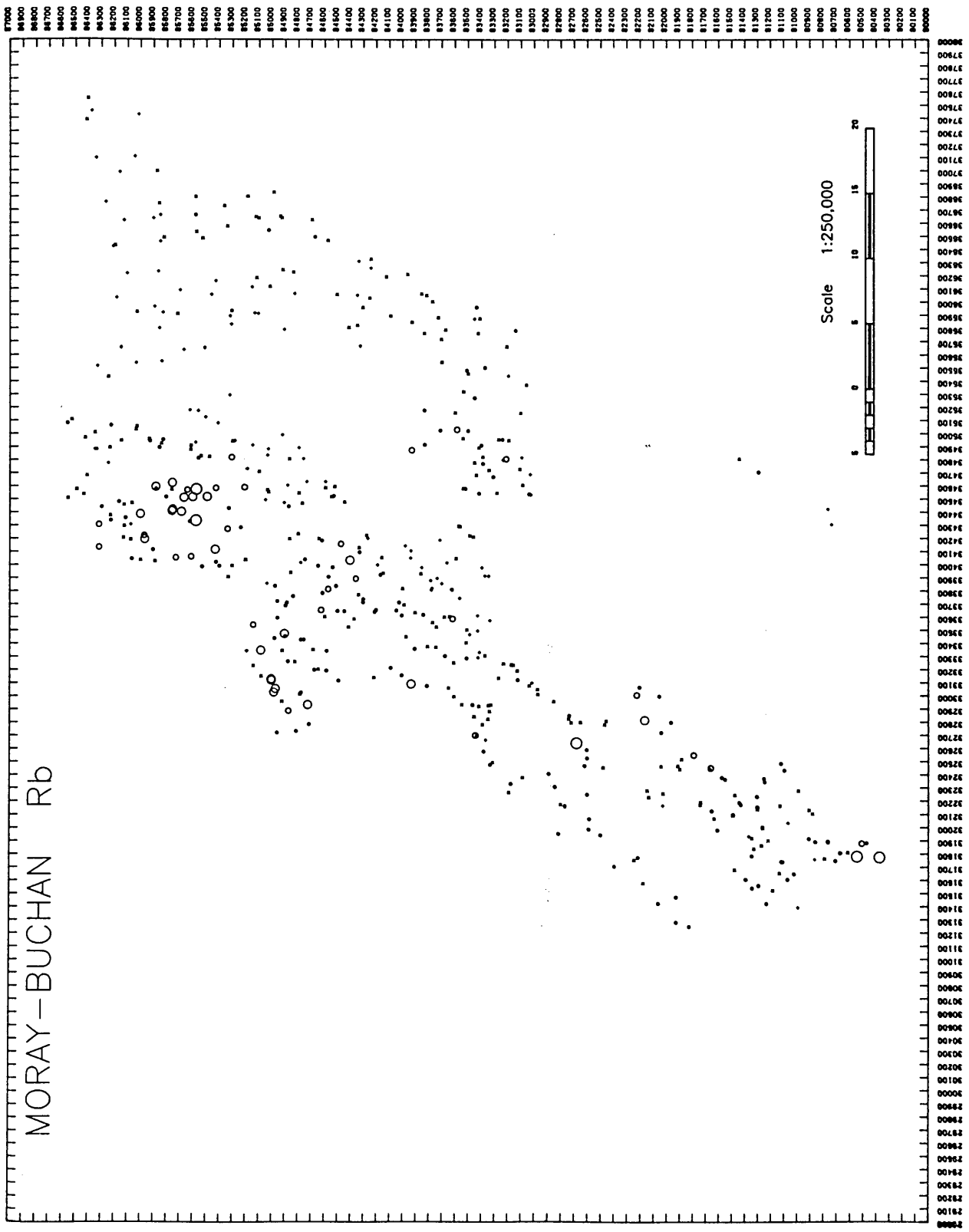


Figure 4.27

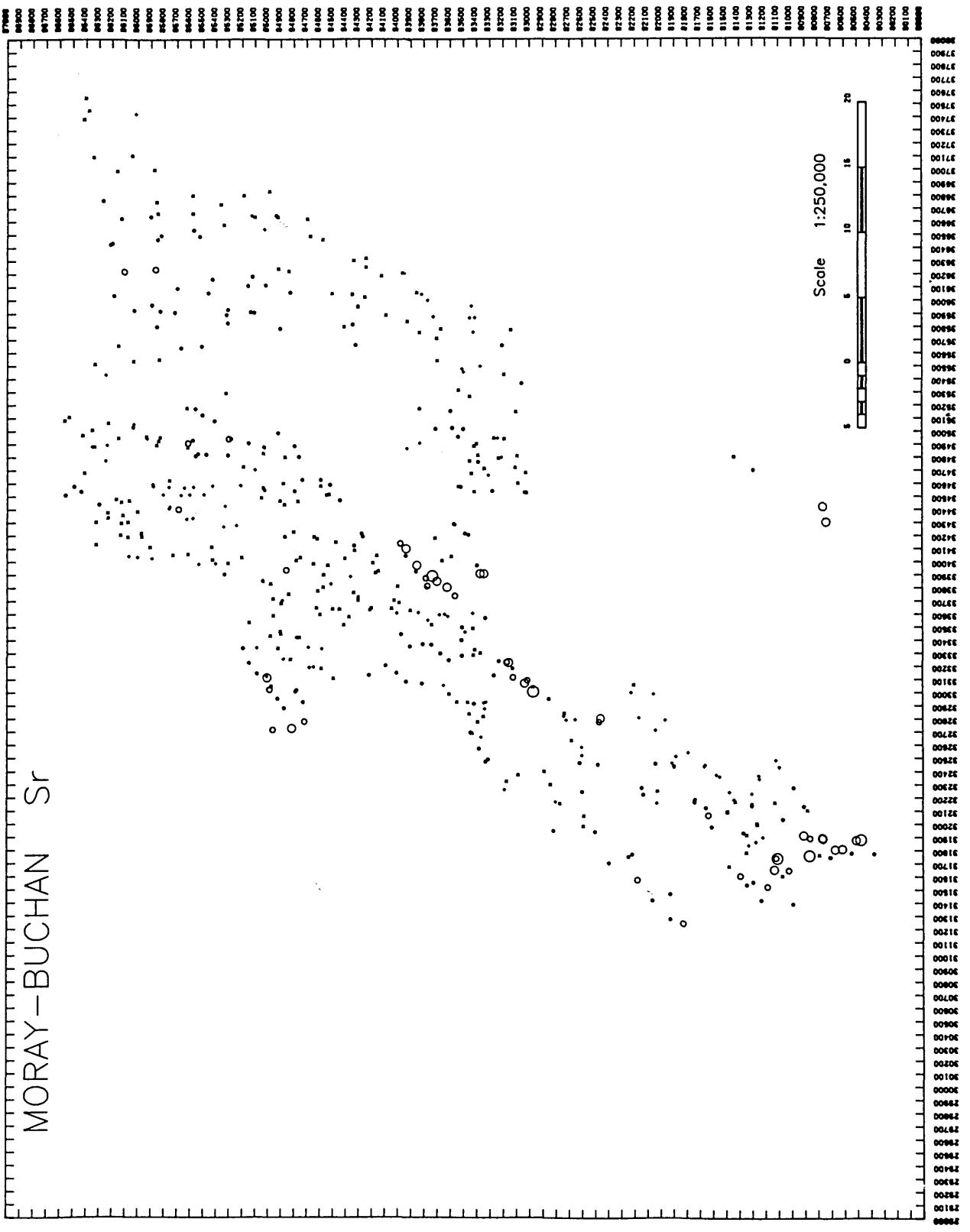


Figure 4.28

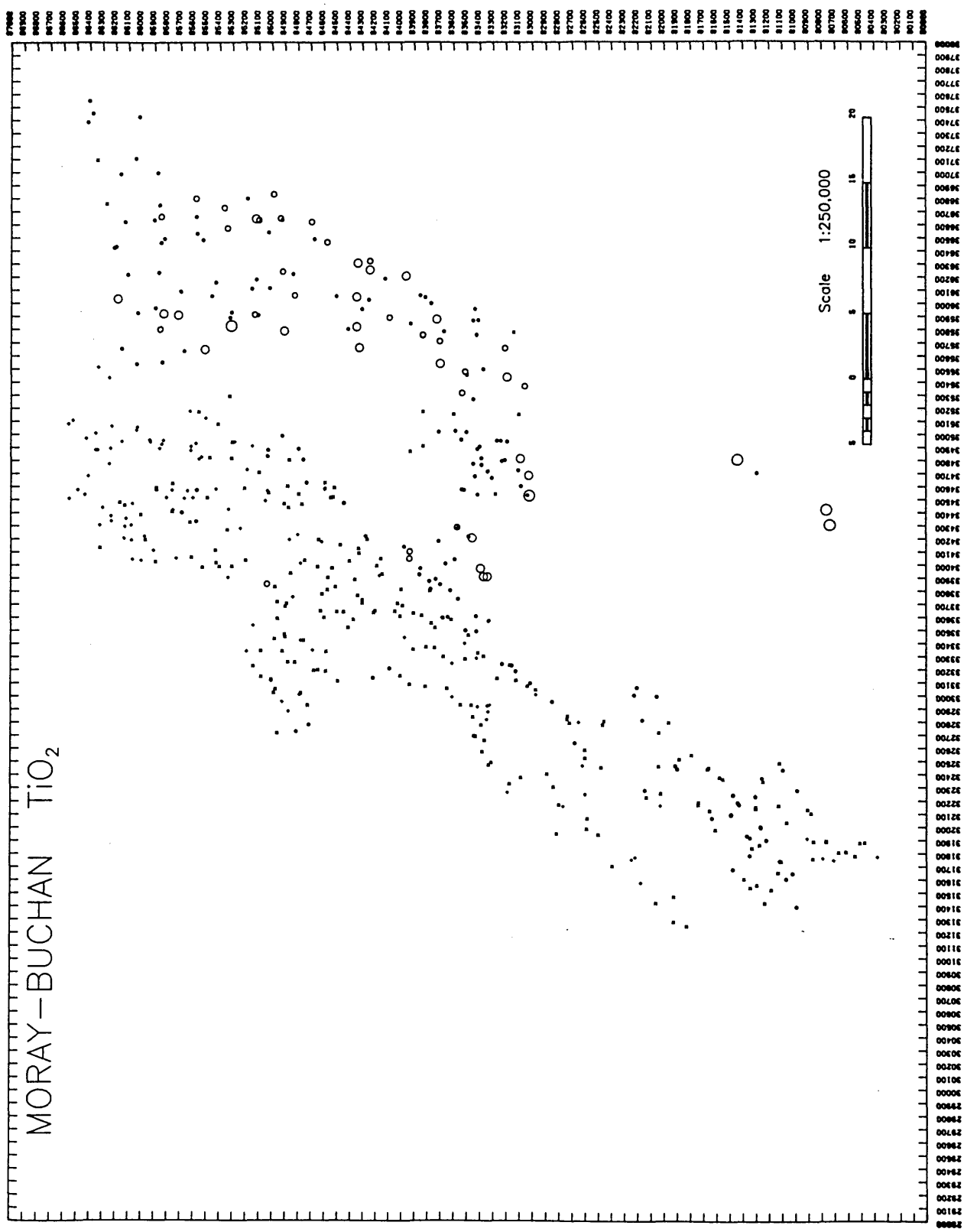


Figure 4.29

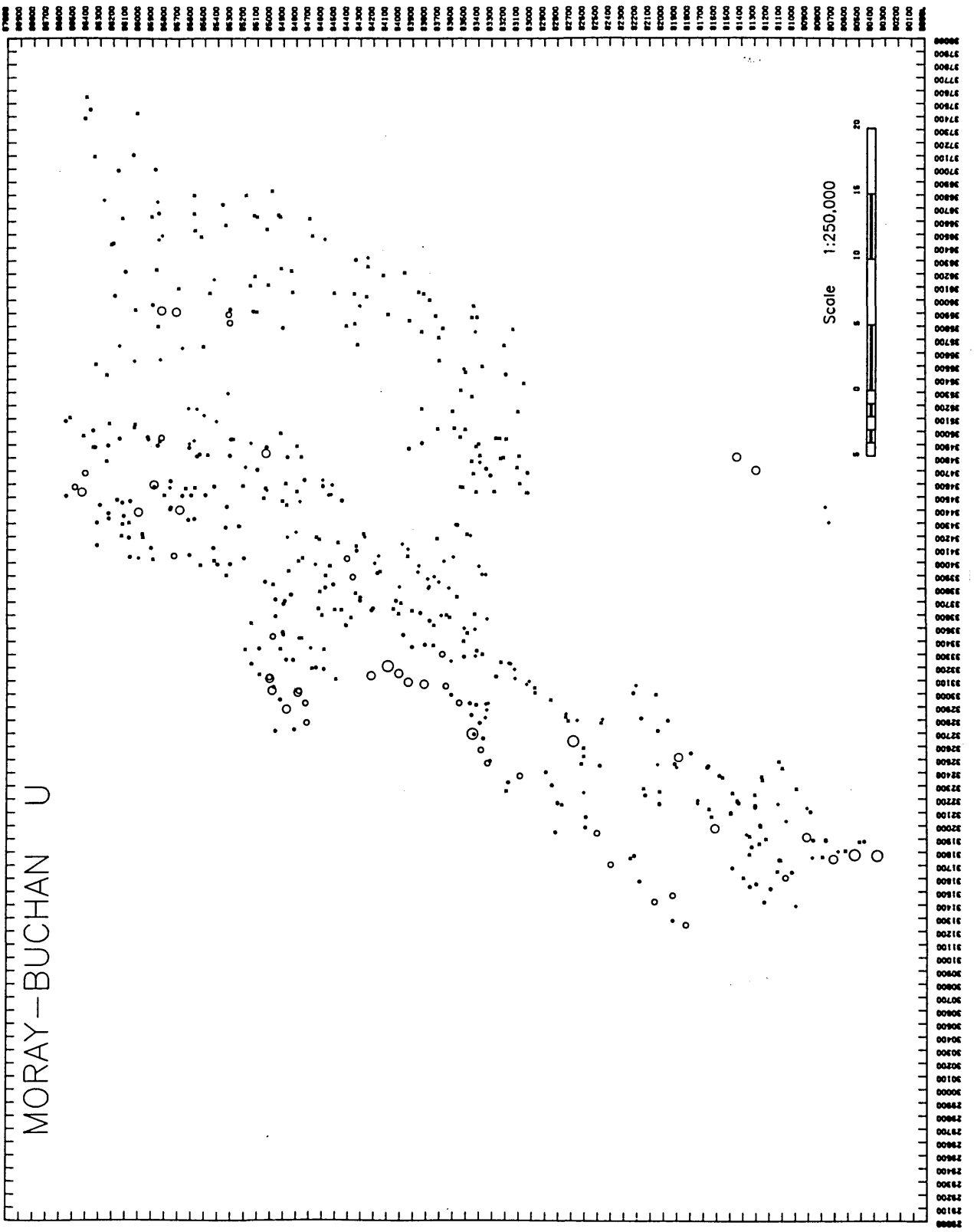


Figure 4.30

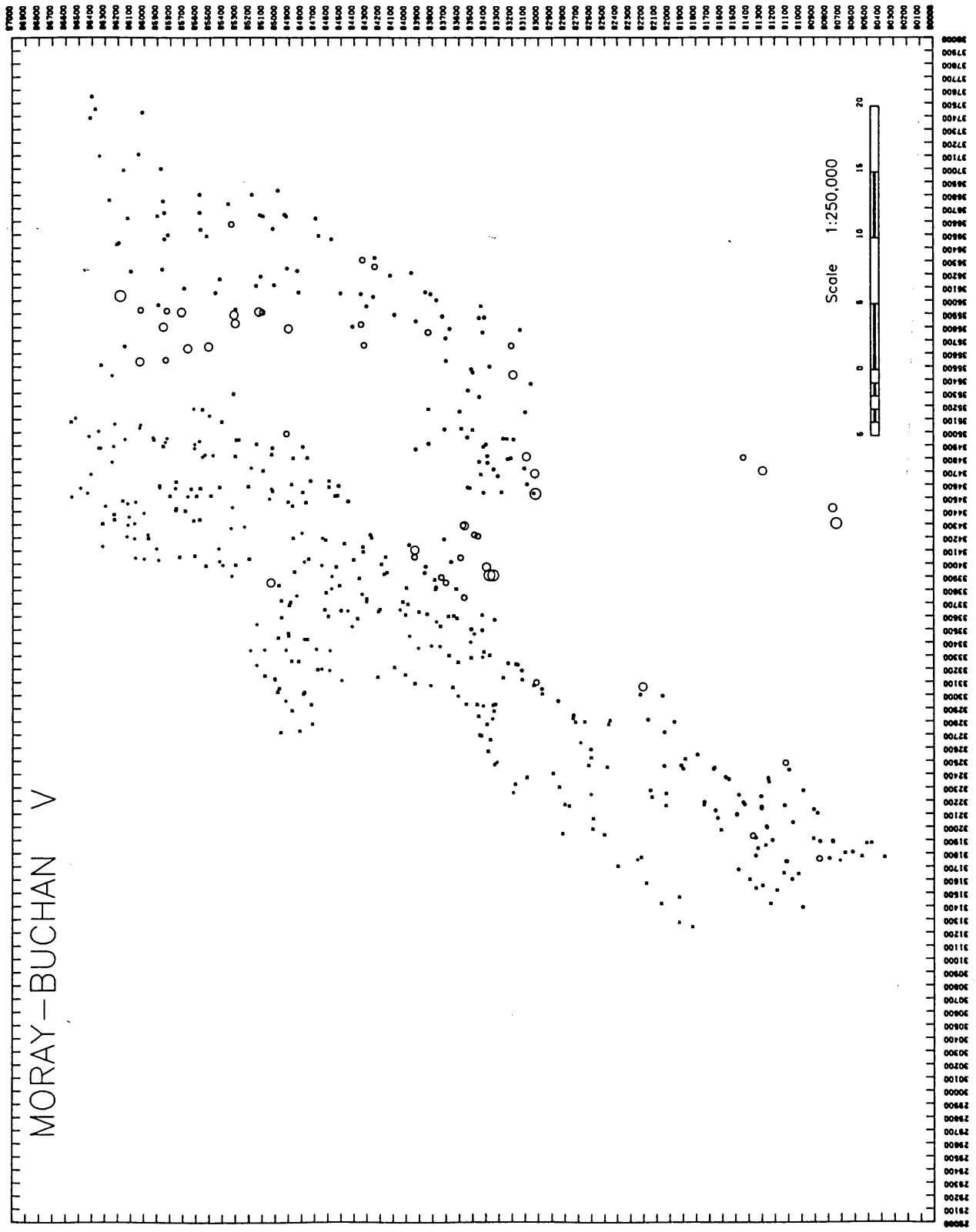


Figure 4.31

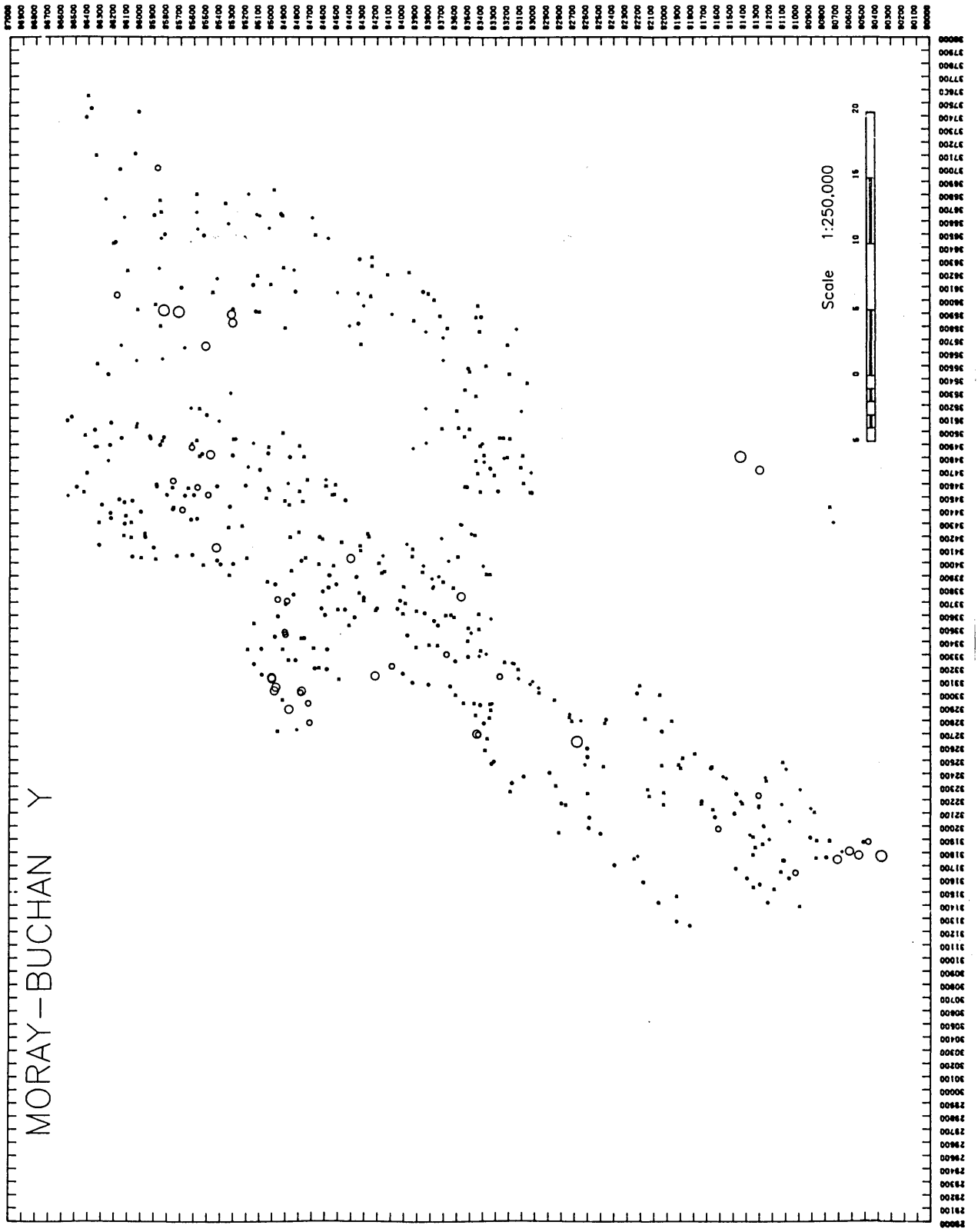


Figure 4.32

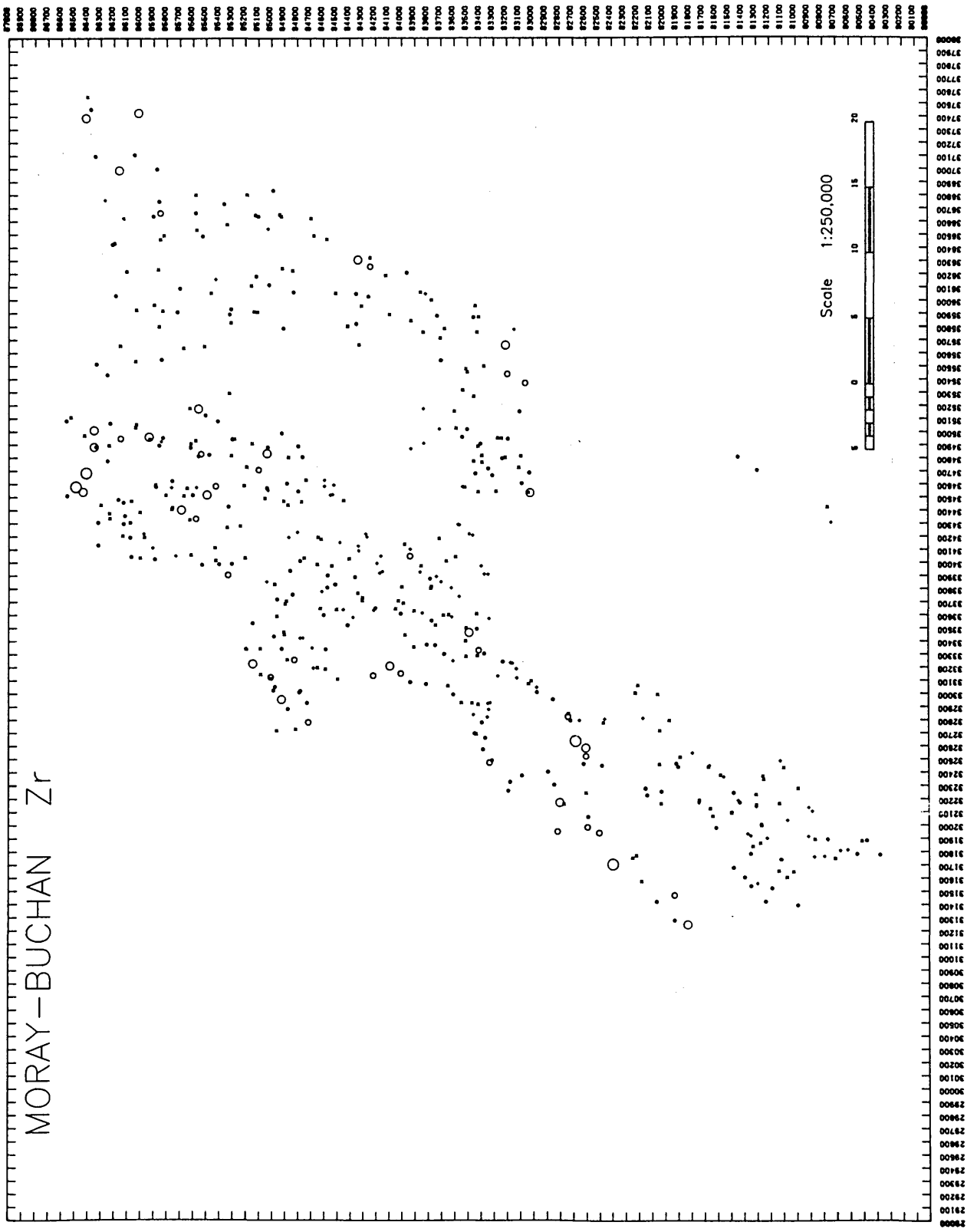
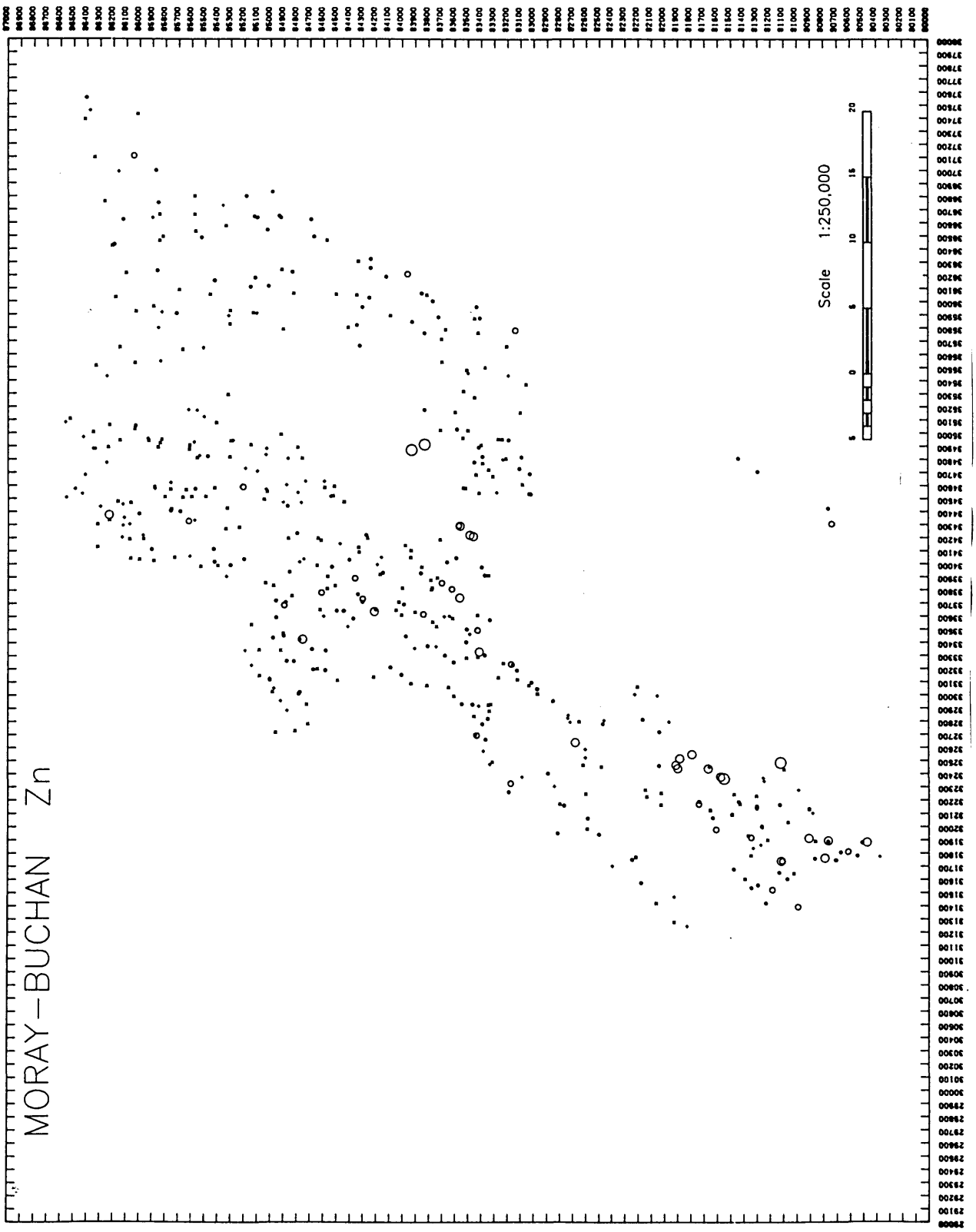


Figure 4.33



2) Cu, Fe, V and Ca show marked depletion in the Lower Dalradian. This may be due to the fact that basic volcanism commenced in the Middle Dalradian and reached its peak in the Upper Dalradian. This may also explain the enrichment of Li, Ti and V in the Upper Dalradian.

More localized concentrations of large values are seen in the following localities:

3) Ba, K and Rb show consistent large values in two localities. The first is on Lower Dalradian Lochaber quartzites in streams supplying the River Spey near Rothes. The second is on hills of Lower Dalradian Ballachulish mixed pelites with calcareous and quartzitic bands, about 7km north of Keith.

4) La, U, Y, Rb and Sr show marked enhancements near either the Cairngorms or the Ben Rinnes granite or both.

5) Co, Cr, Cu, Mn, Ti, Ni, V and Zn show a marked association with outcrop of basic or ultrabasic rocks.

6) B, Co and Ni show coincident anomalies over the Rosarie Forest about 7km NE of Dufftown, on Lower Dalradian Blair Atholl black schists.

7) Cu, Fe and Ni show anomalous values in a small locality about 11km SE of Dufftown, on undifferentiated Lower/Middle Dalradian pelites and black schists. The sample sites are close, to but upstream of, basic and ultrabasic outcrop.

4.5 Multivariate Probability Plots

In Chapter 3 we discussed estimates of the location and scale of a single univariate set of data. In many geological problems more than one variable is measured on each object. We conventionally arrange a set of multivariate data in the form of a matrix which has n rows, where n is the number of objects, and p columns, where p is the number of variables. For each variable an estimate of location and scale can be computed. Different variables measured on the same object usually tend to change together in some manner. Variables which have no relation to each other are said to be mutually independent; that is, an increase or decrease in one variable is not accompanied by a predictable change in the other variable. However, variables are usually not independent and so some measure of their mutual interaction is needed.

Computational procedures used to calculate the variance (s^2) of a single variable can be extended to a calculation

of a measure of mutual variability of a pair of variables. This measure, called the covariance, is the joint variation of two variables about their common mean. The sample <u>covariance</u> between two variables j and k from a data matrix X is given by:

$$COV_{jk} = \frac{\Sigma (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{(n-1)}$$

where $\bar{x}$ is the sample mean. It is possible to make some general conclusions about the behaviour of COV_{jk}. If the variables x_j and x_k are randomly scattered then $|COV_{jk}|$ tends to be small, while if they are independent then $COV_{jk} = 0$. If we confine ourselves to the situation in which the relationship between x_j and x_k is linear, then we can also say that if x_j and x_k vary antipathetically then $COV_{jk} < 0$ while if they vary sympathetically then $COV_{jk} > 0$. Although general conclusions such as these can be made about COV_{jk}, individual values are not too meaningful because, as with variances, they are dependent upon the units of measurement. Covariance may be used in matrix form to describe the structure of a data point cloud in p dimensional variate space. Thus the covariance matrix, denoted by the symbol S, is a symmetrical p x p matrix made up of variances along the leading diagonal and covariances in the off-diagonal positions.

There are two commonly used measures of the distance in

multivariate space between an object and the multivariate mean of the objects. The multivariate mean is simply a p dimensional vector containing the individual univariate means of the variables which make up the columns of the data matrix. The first measure is the Euclidean distance:

$$D^2 = (X_i - \bar{X})' (X_i - \bar{X})$$

where X_i is a p dimensional vector of the variable values on the i-th object, and $\bar{X}$ is the multivariate mean. Surfaces of equal Euclidean distance form circles (in two dimensions), spheres (in three dimensions), or hyperspheres (in more than three dimensions) about the multivariate mean (see Figure 4.34).

The second distance measure is the Mahalanobis squared distance:

$$D_m^2 = (X_i - \bar{X})' S^{-1} (X_i - \bar{X})$$

where S^{-1} is the inverse of the covariance matrix. The Mahalanobis squared distance has a significant advantage over the Euclidean distance since it takes into account the covariance structure of the data. By so doing the surfaces of equal D_m^2 about the multivariate mean are ellipsoidal rather than spherical in shape (see Figure 4.34), with the long axis of the ellipsoid being in the direction of maximum variance.

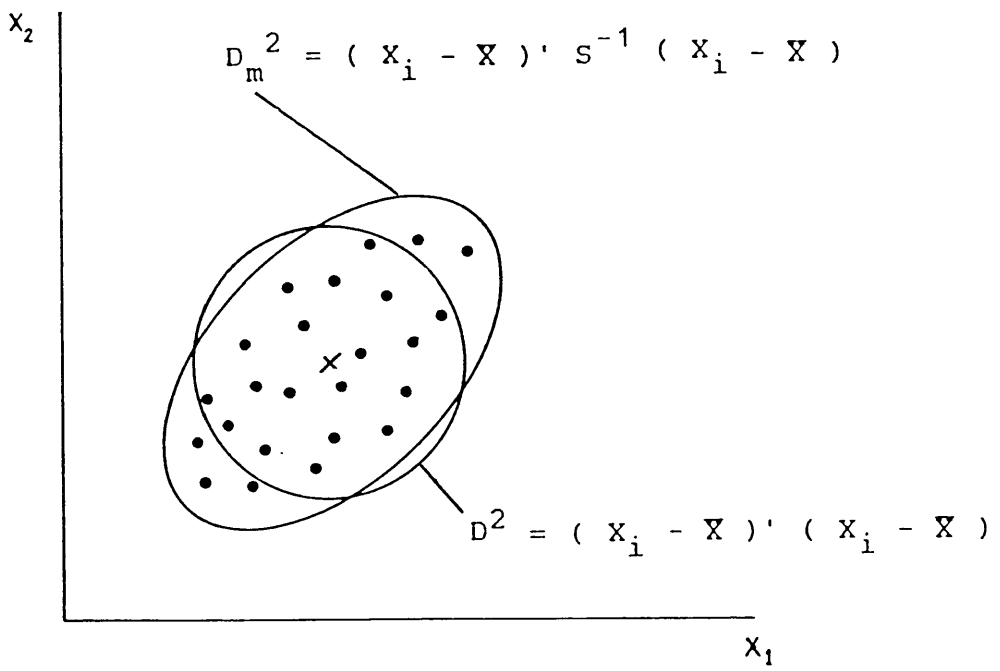


Figure 4.34

The ellipse represents a surface of constant Mahalanobis distance about the bivariate mean, indicating that the covariance structure of the data is taken into account. The circle represents a surface of constant Euclidean distance about the bivariate mean.

Mahalanobis squared distances have a chi-squared distribution, but it has been shown (Wilson and Hilferty, 1931; Healy, 1968) that the cube root of D_m^2 (i.e. $D_m^{2/3}$) closely approximates the Normal distribution. Healy (1968) therefore suggested an extension of probability plots of univariate data to the multivariate situation, by plotting $D_m^{2/3}$ for each observation against the order statistic for a Normal distribution. Normal probability plots of $D_m^{2/3}$ provide a method of checking for a multivariate normal distribution in a similar manner to univariate Q-Q plots, and detecting the presence of multivariate outliers, which may go unnoticed in inspection of univariate data (Campbell, 1979). The $D_m^{2/3}$ probability plot may also be used as a dimension reduction technique, since much of the information contained in a number of separate univariate QQ plots is summarized in one $D_m^{2/3}$ QQ plot. There will be some loss of detail but the information gained cannot be captured from the individual plots.

As we have seen from Chapter 3, an anomalous object (outlier) will influence both the mean and the variance of univariate data. In the same way an anomalous multivariate object vector will influence both the mean vector and covariance matrix of multivariate data. The tendency will usually be to deflate the correlations and possibly inflate the variances, and so to inflate the size of the associated concentration ellipsoid. This will in general have the effects of decreasing the Mahalanobis squared distance for

the anomalous object vector and distorting the rest of the plot (Campbell, 1979). The logical solution to this problem is to use robust estimates of the covariance matrix and the means in the calculation of the Mahalanobis squared distances (see Chapter 6 for a full explanation of how this is done).

4.6 Multivariate Probability Plots and the SIND Data

To illustrate the above discussion compare the two Q-Q plots in Figure 4.35. The first is the probability plot of the $D_m^{2/3}$ values for SIND raw data using classical estimators for the covariance matrix and mean vector. The second is a similar plot except that the $D_m^{2/3}$ values were calculated using robust estimators. It is clear from a comparison of the two plots that the presence of anomalous object vectors has caused a decrease in their $D_m^{2/3}$ values when these values are calculated in a non-robust fashion. In contrast the second figure shows how clearly these anomalies stand out when robust parameters are used. The robust parameters in this situation define the covariance structure and identify the multivariate mean vector of the data set from which the distorting influence of the anomalous object vectors have been removed. In this way

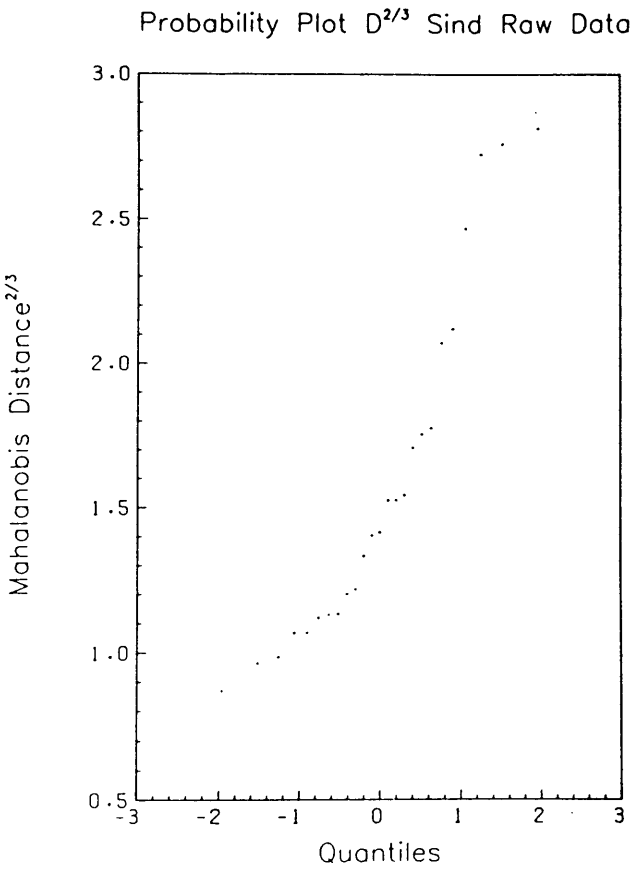
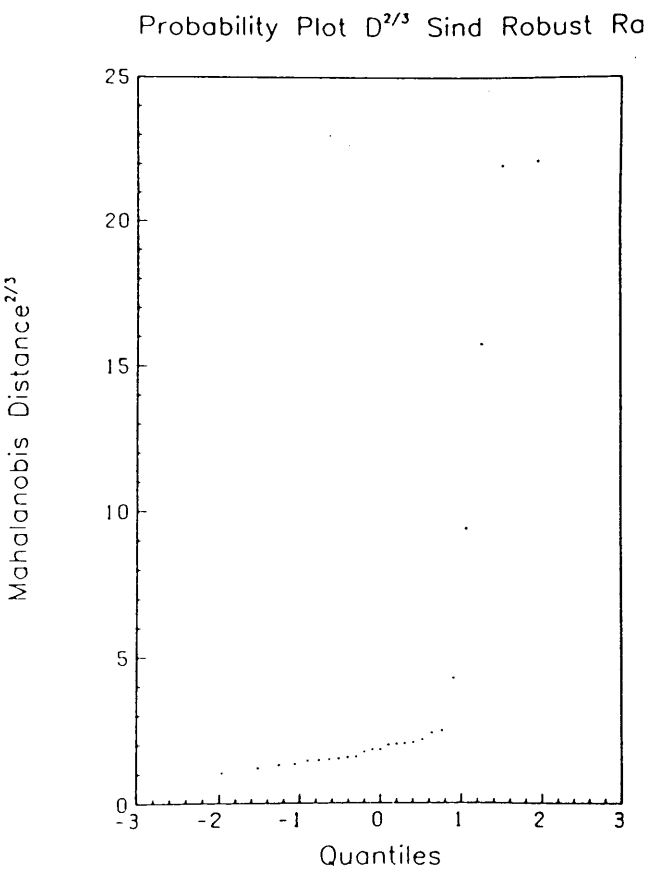

Probability Plot $D^{2/3}$ Sind Raw Data
Mahalanobis Distance$^{2/3}$
3.0
2.5
2.0
1.5
1.0
0.5
-3
-2
-1
0
1
2
3
Quantiles

Figure 4.35


Probability Plot $D^{2/3}$ Sind Robust Raw Data
Mahalanobis Distance$^{2/3}$
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
Quantiles

background relationships are accurately determined and the contrast between this background definition and the anomalous object vectors is maximized.

It may be of interest to note that the order of the anomalous object vectors in terms of decreasing $D_m^{2/3}$ (15, 16, 23, 13 and 22) is consistent with the univariate probability plots discussed in Section 4.3.

Mahalanobis distances play an important role in the robust estimation of multivariate location and scale, and in multivariate outlier detection (see Chapter 6).

4.7 Multivariate Probability Plots and the Moray-Buchan Data

Figure 4.36 shows, at the top, the probability plot of the $D_m^{2/3}$ values for MB raw data using classical estimators for the covariance matrix and mean vector. At the bottom of Figure 4.36 is a similar plot except that the $D_m^{2/3}$ values were calculated using robust estimators. One of the major goals of this thesis is to find techniques which maximize the definition between the background population and any superimposed anomalous populations. A phrase that will occur regularly in subsequent chapters is "anomaly contrast" which refers to the contrast between the background population and the anomalous object vectors

Probability Plot $Dm^{2/3}$ Moray–Buchan Raw Data

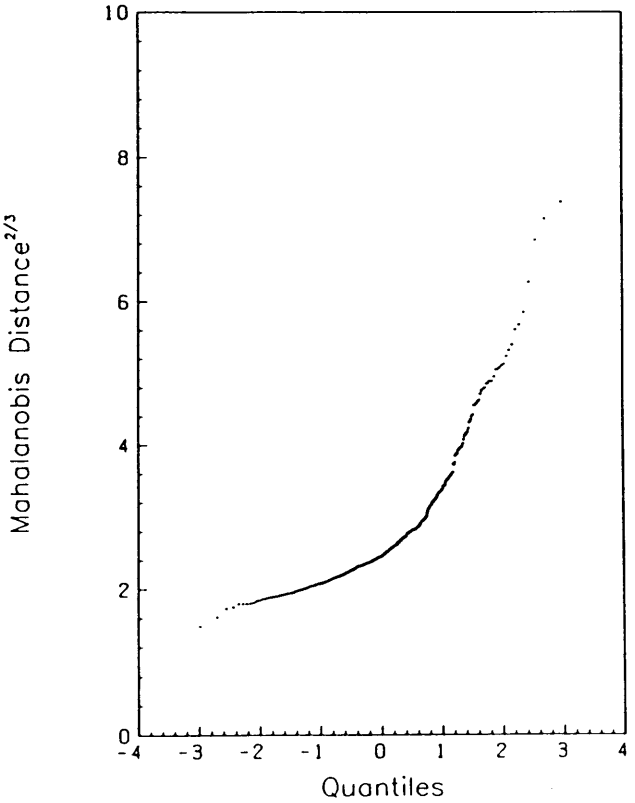
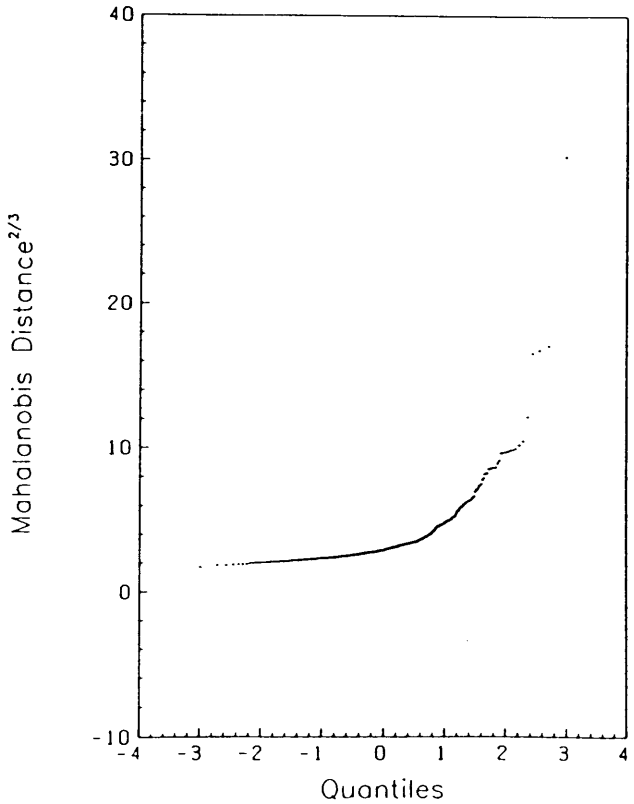


Figure 4.36

Probability Plot $Dm^{2/3}$ Moray–Buchan Robust Raw Data



whose identity we are trying to establish. A large anomaly contrast applies to a situation in which anomalous object vectors are easily distinguished from background, whereas a low anomaly contrast applies if the anomalous object vectors are merging with background. In the case of the MB raw data it is clear that the use of robust $D_m^{2/3}$ values results in a maximized anomaly contrast when compared with their non-robust equivalents. However, that does not make the plot any easier to interpret! There is a clear anomalous population above a threshold of $D_m^{2/3} = 10$ containing 7 samples, and a background population with $D_m^{2/3} < 3$. The latter value is approximately the point above which object vectors have non-unit weights. Between $D_m^{2/3} = 3$ and 10 the plot is relatively complex. Too much emphasis should not, however, be attached to minor changes of gradient. The reason for this is made clear by the following demonstration. Figure 4.37 shows a probability plot of 503 numbers generated at random from a standardized normal distribution N(0,1). Such a plot should, theoretically, be a perfect straight line but in practice this will only be attained as n tends towards infinity. With this consideration in mind, samples with $D_m^{2/3}$ values between 3 and 10 (Figure 4.36b) are probably derived from no more than a single population.

In the same way that the single element distributions can be portrayed spatially, so can the robust $D_m^{2/3}$ distribution. Figure 4.38 shows the point symbol map of

Figure 4.37

Probability Plot Random Nos. from std. Normal Distribution

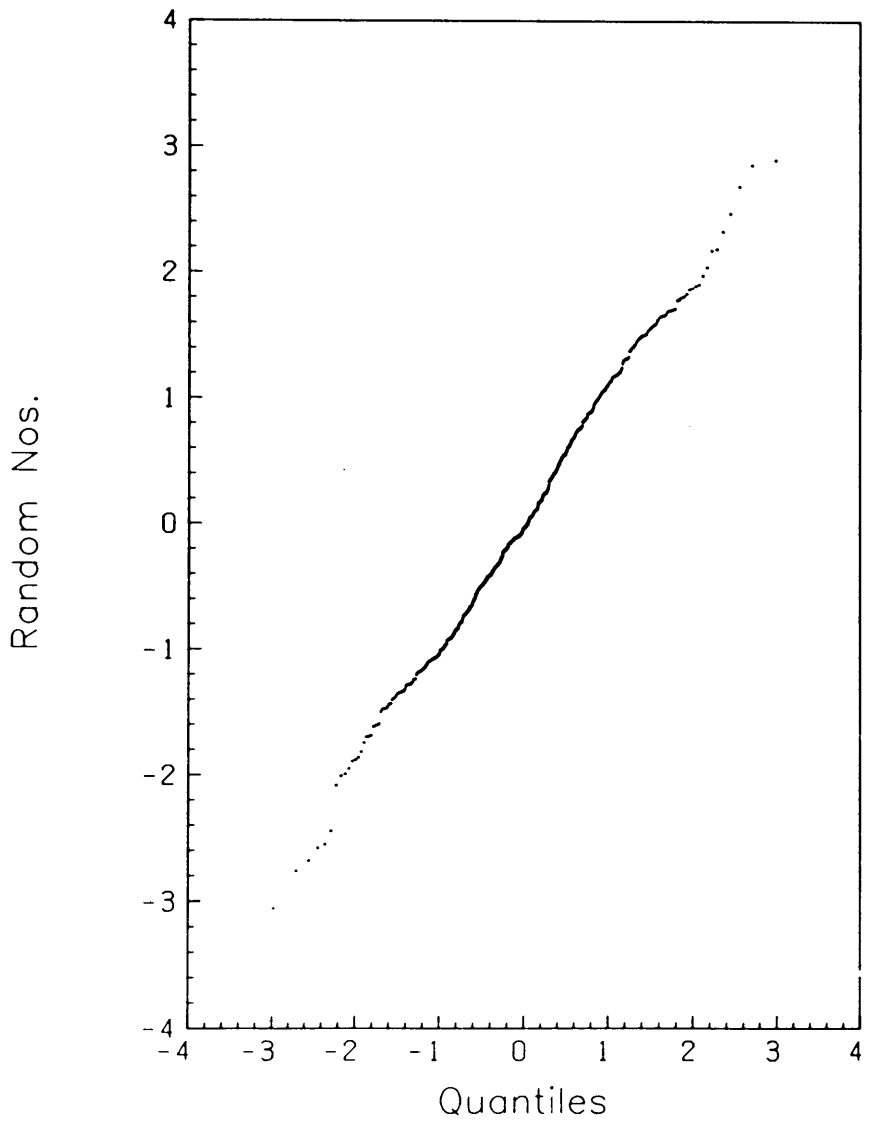
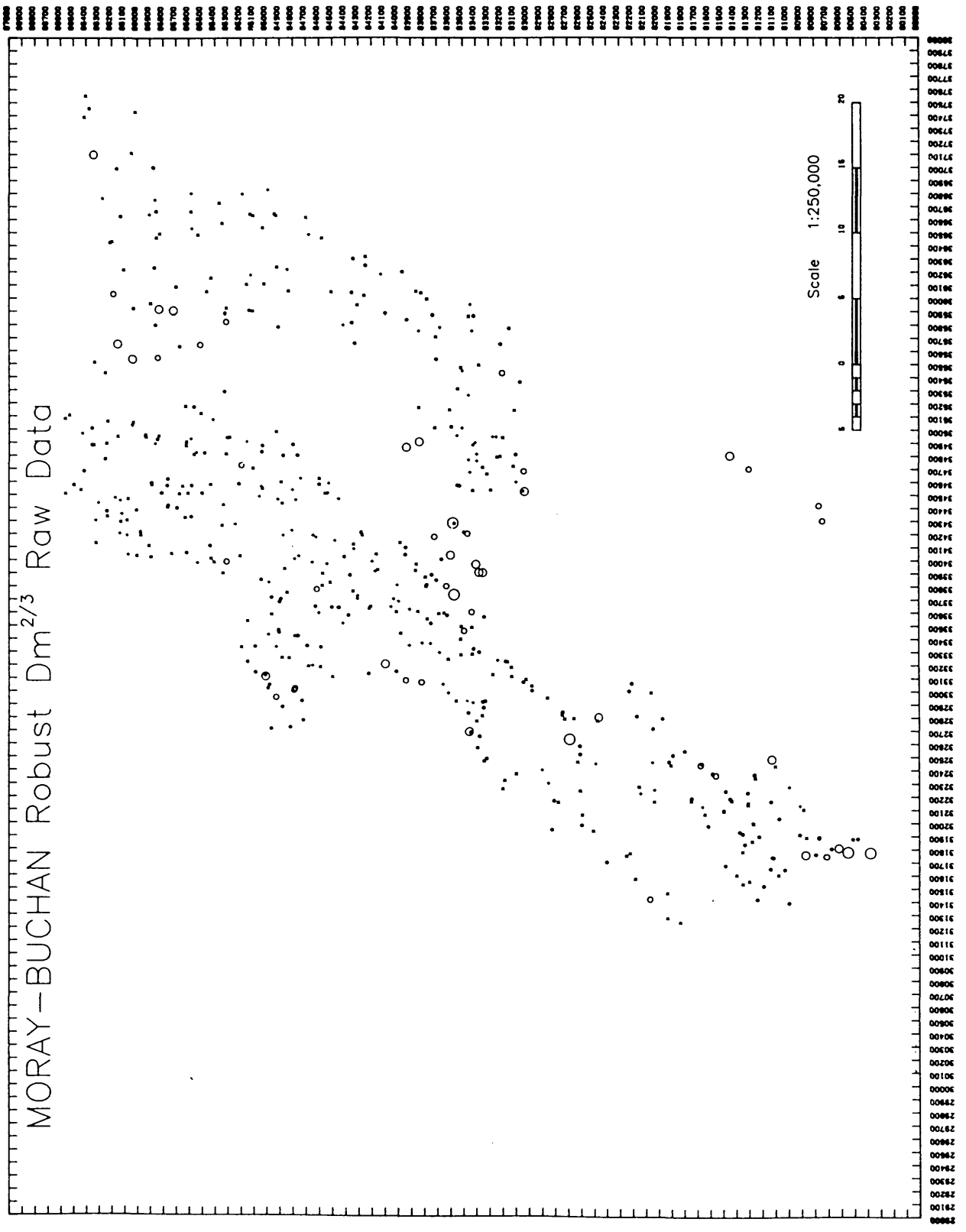


Figure 4.38



robust $D_m^{2/3}$ values using the same class intervals and symbols defined in section 4.4. The conclusions from the analysis of this map are as follows.

1) There is a group of samples with large Mahalanobis distances in the extreme south of the area close to the Cairngorm Granite. These large values occur in sites downstream from either granitic outcrop or metamorphosed limestone of undifferentiated Lower/Middle Dalradian age. The elements that contribute to these large $D_m^{2/3}$ values are U, La, Y, Rb and Be of probable igneous association, Ca and Sr of probable limestone association and Pb and Zn as additional contributers.

2) There is a second group of samples with large Mahalanobis distances in streams close to basic/ultrabasic outcrops approximately 11km SE of Dufftown. This area was previously identified in the analysis of the univariate point symbol maps as being high in elements of basic association, i.e. Co, Cr, Cu, Mn, Ti, Ni, V and Zn.

3) There are several isolated $D_m^{2/3}$ anomalies over Easdale Subgroup pelites in streams draining an ultrabasic body south of Portsoy.

4) Another group of samples with large Mahalanobis

distances occurs over Tayvallich/Crinan Subgroup mixed pelites in a locality about 5km NW of Aberchirder and south of Portsoy. The contributary elements are La, V, Y and Ti. The streams appear not to have any major basic or ultrabasic sources. A similar group of values occurs in the same Subgroup in the extreme SE of the study area.

5) The single largest $D_m^{2/3}$ value occurs in a stream draining undifferentiated Lower/Middle Dalradian quartzites near the boundary with Lower Dalradian black schists at (326000E 826000N). The contributary elements are Zr, Y, U, Li, La, K, Be, Zn and Pb.

6) The Lochaber quartzites around Rothes and the River Spey have generally elevated $D_m^{2/3}$ values.

4.8 Conclusions

The probability plot is an excellent graphical means of displaying the frequency distribution of either a single variable or a group of variables. The multivariate probability plot, based on the Mahalanobis distance, enables detection of anomalies which might not show up in

any individual univariate plot. These are termed minor anomalies and may be envisaged, in a simple bivariate scatter plot, as occurring to one side of the main point cloud, but not sufficiently far enough away to register as an outlier on either univariate plot. This analogy may be extended to more than two variables. In contrast, a major anomaly is one that occurs at the upper extremity of the point cloud, and will register as an anomaly in both univariate and multivariate probability plots.

The multivariate probability plot has the potential to detect subtle anomalies and, as mentioned in section 3.8, it is often these which prove most valuable. In order to fulfill the potential of the multivariate probability plot, it is necessary to generate robust mahalanobis distances. Robust estimation causes the contours of equal mahalanobis distance to contract, and redefine themselves about a new, robust, multivariate mean. This has the effect of revealing as anomalies, observations that had previously been defined as background. More credence may be placed on these anomalies than on those detected using a non-robust approach.

CHAPTER FIVE

DATA TRANSFORMATIONS

5.1 Introduction

The standard statistical techniques associated with the linear model have been developed with the following basic assumptions:

1) Additivity - that is to say the main effects combine linearly to "explain" the observations. For example, in linear regression the observations may be explained by the additive effect of the predictor variables.

2) Constancy of variance - that is the observations are assumed to have a constant variance about their varying means. Explicitly this means that the variance is independent of the expected value of the observations, and the sample size.

3) Normality - that is to say the observations are assumed to have a normal (Gaussian) distribution.

There are many cases in practice where these assumptions are known, or believed, not to hold. For instance the raw data may be known to have:

1) strong asymmetry in its distribution;

2) many outliers in one tail of the distribution;

3) batches at different levels (e.g. concentrations) with widely differing variances; and

4) large and systematic residuals from fitting a simple linear model to the data.

The data analyst then has the option of either developing new theory (e.g. robust statistics) or altering the data so that one or more of the assumptions are met. Transforming the data by applying a single mathematical function to all raw data values can result in one or more of the assumptions being tenable. The appropriate transformation must change the shape of the raw data distribution. Changes of origin or scale are linear transformations and leave shape unchanged. Stronger transformations such as the logarithmic and square root change shape.

Box and Cox (1964) have indicated methods of finding

transformations which simultaneously satisfy or nearly satisfy the three assumptions. This can usually be done so long as one does not press too hard for any one of them. The robustness of the usual linear model procedures means that only a little is lost if the assumptions are not exactly met. The family of transformations proposed by Box and Cox is described below (Section 5.3). For a more detailed account of data transformations and the philosophy behind the process see Hoyle (1973) and Hoaglin et al (1983).

5.2 Data Transformations in Geochemistry

The objective of all data interpretation in exploration geochemistry is the recognition of samples which, in terms of measured parameters, may be assigned to a population group associated with mineralization. Much of the early practice of exploration geochemistry using soils and stream sediments relied on large differences in element concentration between samples; this approach is adequate to detect near-surface and large mineralized areas which give rise to intensive and extensive halos of the elements associated with mineralization.

Attempts to extend the scope and usefulness of geochemistry in mineral exploration - to locate deeply buried deposits and to identify broad targets with

relatively few samples - has led to the increasing use of computer-assisted univariate and multivariate statistical data treatment to identify subtle anomalies. Whatever simple or complex data interpretation procedures are used, implicit in the designation of a sample as "anomalous" is the assumption that a homogeneous background distribution can be recognized. Because the theoretical distribution of elements in the earth's crust is in dispute, this is extraordinarily difficult in practice.

It is an empirical fact that the commonest form of frequency distribution of an element in samples from a seemingly homogeneous source is positively skewed. Ahrens (1954), studying trace element distributions from two igneous rock types in Canada, observed that when a logarithmic transformation was applied to the original positively skewed data the resultant distributional shape approximated that of the normal or Gaussian distribution. He therefore proposed a fundamental law concerning the nature of the distribution of the concentration of an element, namely: "the concentration of an element is lognormally distributed in a specific igneous rock". Despite considerable opposition to Ahrens' contention (Chayes, 1954; Vistelius, 1960; Shaw, 1961) it is common practice among exploration geochemists to calculate threshold values and to undertake various other statistical manipulations on log-transformed data. This practice has been contested by Govett et al (1975) among others, on the

grounds that:

1) there seems to be no satisfactory theoretical reason why an element distribution in a homogeneous parent population should be lognormal; and

2) in many situations where adequate control data are available, the samples in the tail of the distribution can be shown to belong to a different population.

Even in Ahrens' own data it can be shown (Govett et al, 1975) that some element distributions could as well obey a cube-root or square-root law as a logarithmic law. Even more plausible is the possibility that the distribution could belong to two dominant normal distributions which overlap.

Several authors have proposed the view that skewed distributions represent an overlapping mixture of two or more normal distributions. Vistelius (1960) suggested that the distributions of the concentrations of chemical elements in the earth's crust are divided into two groups. The first group of distributions is that of the fixed stage of the geochemical process; these distributions are normal. The second group of distributions is that of mixed products of many stages of the geochemical process; these distributions have large positive skew. In a similar vein, Smirnov (1963) suggested that sampling of heterogeneous

material and the presence of anomalous dispersion patterns "mask" the original normal distribution and cause skewed or even multimodal distributions.

Chapman (1976a) considers that the statistical distribution of data collected from exploration geochemical surveys do not follow an authentic lognormal distribution, but are frequently positively skewed and therefore mimic the lognormal distribution. He calls these distributions "pseudo-lognormal". Chapman postulates that the observed distribution of exploration geochemical data is a mixed distribution and is composed of many smaller aggregate populations which represent:

1) bedrock and other physiographic factors, and

2) mineralization.

These small populations are classed broadly as "background" and "anomalous" populations respectively. Each of these classes is composed of a number of populations and some of these populations are lognormally distributed.

It is important to note that trace element frequency distributions are usually much more positively skewed than those of major elements. The greater skewness, or asymmetry, of trace element distributions arises because most trace element observations have naturally low "background" concentrations, and therefore are more

susceptible to truncation by analytical detection limits. The converse may be true for major elements (e.g. Si and Ca) which can occur as significant proportions of a sample and can cause saturation of the analytical signal.

Govett et al (1975), Chapman (1976a) and Chapman (1976b) have put forward the view that attempts to make the assumption of normality more efficient by performing log-transformation on positively skewed data may be inappropriate. As Chapman (1976a) pointed out, the statistical distribution of an element is a mixed distribution, and the nature of the log-transformation enhances the background populations of the overall statistical distribution, and suppresses those populations which represent mineralization. This tends to lead to fusion of the two classes of populations. In this way transformation disrupts those relationships which statistical analysis attempts to investigate, and works directly against the goal of the analysis, i.e. investigation and delineation of associations caused by the presence of mineralization in the sample area. This suppression can be detrimental to the interpretation of principal component scores and regression residuals, for example. Variation due to mineralization may be concealed to the extent that areas of mineralization could remain undetected.

The above discussion raises the more philosophical question of whether any normalizing transformation is

really necessary. I propose to discuss this question in later sections (Sections 5.5 and 5.6).

5.3 Power Transformations

The lognormal distribution has for many years been the standard model for exploration geochemical data distributions because of their apparent similarity, and the fact that the skewness and kurtosis of the transformed data are less than for the original data. Indiscriminate application of the log-transformation could, however, have undesirable effects. For instance, an original slightly positively skewed distribution may be replaced by a negatively skewed distribution after log-transformation. Box and Cox (1964) suggested the use of a family of power transformations to improve normality that would ensure an optimum result for a particular set of data in that the non-normal aspects of the data would be minimized.

The family of transformations proposed by Box and Cox - subsequently referred to as the Boxcox (B-C) transform - takes the form

$$z = (x)^{\lambda} \qquad \text{Equation 5.1}$$

which include

$$z = \begin{cases} (x^{\lambda} - 1)/\lambda, & (\lambda \neq 0) \\ \ln(x), & (\lambda = 0) \end{cases} \quad x > 0$$

Equation 5.2

where z is the transformed observation, x the original observation and λ the tranforming power coefficient defined over all x. The value of λ is specific to a data variable and is calculated so as to minimize the departure of the transformed data from normality.

Equation 5.2 encompasses several familiar transforms for particular values of λ : reciprocal, $\lambda = -1$; natural logarithm, $\lambda = 0$; square root, $\lambda = 0.5$; and when $\lambda = 1$ the data are unchanged except in a linear manner which has no effect on distributional shape.

It is often necessary to back transform values such as means and variances obtained from transformed data into units consistent with untransformed data. For the mathematics of how this is done see Appendix A.1.

Several criteria have been proposed for determining the value of λ for a given data set (Howarth and Earle, 1979). These methods find the λ value which will minimize a

"normality" criterion. The criteria take the form of functions and may be one of the following:

1) $F1 = | \beta_1 |$

2) $F2 = W | \beta_1 | + | 3 - \beta_2 |$

3) $F3 = (-n/2)\ln(\sigma^2)+(\lambda -1) \quad \Sigma \ln(X)$

where β_1 is skewness, β_2 is the kurtosis of the transformed data. The first two functions are based on the values $\beta_1 = 0$ and $\beta_2 = 3$ for a normal or Gaussian distribution. It should be noted, however, that possession of zero skewness and a kurtosis of 3 does not necessarily imply that the data have a Gaussian distribution. W is a weighting function which may be one or two, to give two alternative forms of F2. Both σ^2 and ln(x) refer to the untransformed data. Howarth and Earle have evaluated the functions in detail and have published algorithms for each.The algorithms form the basis of an interactive FORTRAN program written by myself and referred to subsequently as LAMT.

The advantages of Boxcox transforming a data set are:

1) it will become as near normally distributed as possible;

2) the range of the data is reduced, making it easier to

display; and

3) over-transformation of the data is avoided - this is a common problem with log-transformation.

The disadvantages are:

1) the value of λ is dependent on the data distribution, so if this is changed by addition of new data a revised λ value must be calculated;

2) it is sensitive to the presence of outliers, as these have a large influence on skewness and kurtosis; it is, therefore, necessary to calculate λ using a data set trimmed of major outliers; and

3) a different λ value is usually required for each variable in a data set.

For a detailed analysis of the application of the Boxcox transform to large data sets refer to Mancey and Howarth (1980) and Mancey (1980).

5.4 Program LAMT

This program performs a power transformation of a set of multivariate data. The optimization criterion required (see above) must be specified prior to analysis. The first operation necessary is to trim the data of outliers prior to calculation of λ . Since it would be too time consuming to trim each variable of its outliers separately, this operation is done using a normal probability plot of $D_m^{2/3}$ (see Chapter 4) to select multivariate outliers. As suggested in Chapter 4 a robust calculation of $D_m^{2/3}$ is used (see Section 4.3). Trimming is done interactively at the computer terminal and may be iterated until a straight line relationship is achieved on the probability plot (see Section 4.2). Calculation of λ is now done for each variable in turn using only the trimmed data for that variable. The calculation requires:

1) an initial estimate for λ ;

2) transformation of the trimmed data (as in equation 5.2);

3) calculation of the optimization or "normality" criterion - calculations of skewness, kurtosis,

variance, etc are performed on the transformed trimmed data;

4) adjustment of the λ value; and

5) repetition of steps 2-4 until no significant change in the criterion is obtained.

The objective function is thus minimized and the optimum λ value obtained (see Howarth and Earle, 1979 for details).

5.5 The Geometry of Transformation

This section is presented to show the effect of transformation upon a bivariate set of data. A set of V and Fe concentration data for 612 marine surface sediment samples from throughout the S.W. Pacific was first Boxcox transformed and then both the raw and Boxcox data sets were standardized to zero mean and unit variance. Robust $D_m^{2/3}$ values were computed for each data set (using the FORTRAN program WTHGRP, see Chapter 6). Scatter plots of Fe against V were plotted for each data set and the $D_m^{2/3}$ values (of which there was one per sample) were able to be contoured on these plots. In addition the positions and sample numbers of all major outliers were marked on each plot.

When the two plots were overlain so that the axes matched it was found that the $D_m^{2/3}$ contours remained unchanged. These contours were, therefore, plotted (Figure 5.1), together with the change in position of the major outliers with an arrow pointed towards the Boxcox transformed location. The conclusions from this exercise are as follows:

1) The position of the $D_m^{2/3}$ contours does not change in bivariate space between raw and Boxcox situations.

2) Change in the $D_m^{2/3}$ value for individual samples is purely due to changes in their position in the bivariate space due to transformation. However, changes in $D_m^{2/3}$ are greater for samples off the bivariate regression line. Another way of viewing this is that samples with a large bivariate residual will change most in $D_m^{2/3}$. This illustrates and emphasises the fact (mentioned in Section 4.4) that Mahalanobis distance calculation takes into account the covariance structure of the data.

3) If there is no change in the shape of the distribution of a data set on transformation, there will be no change in $D_m^{2/3}$.

4) For a given movement in bivariate space on transformation, the change in $D_m^{2/3}$ will be a function

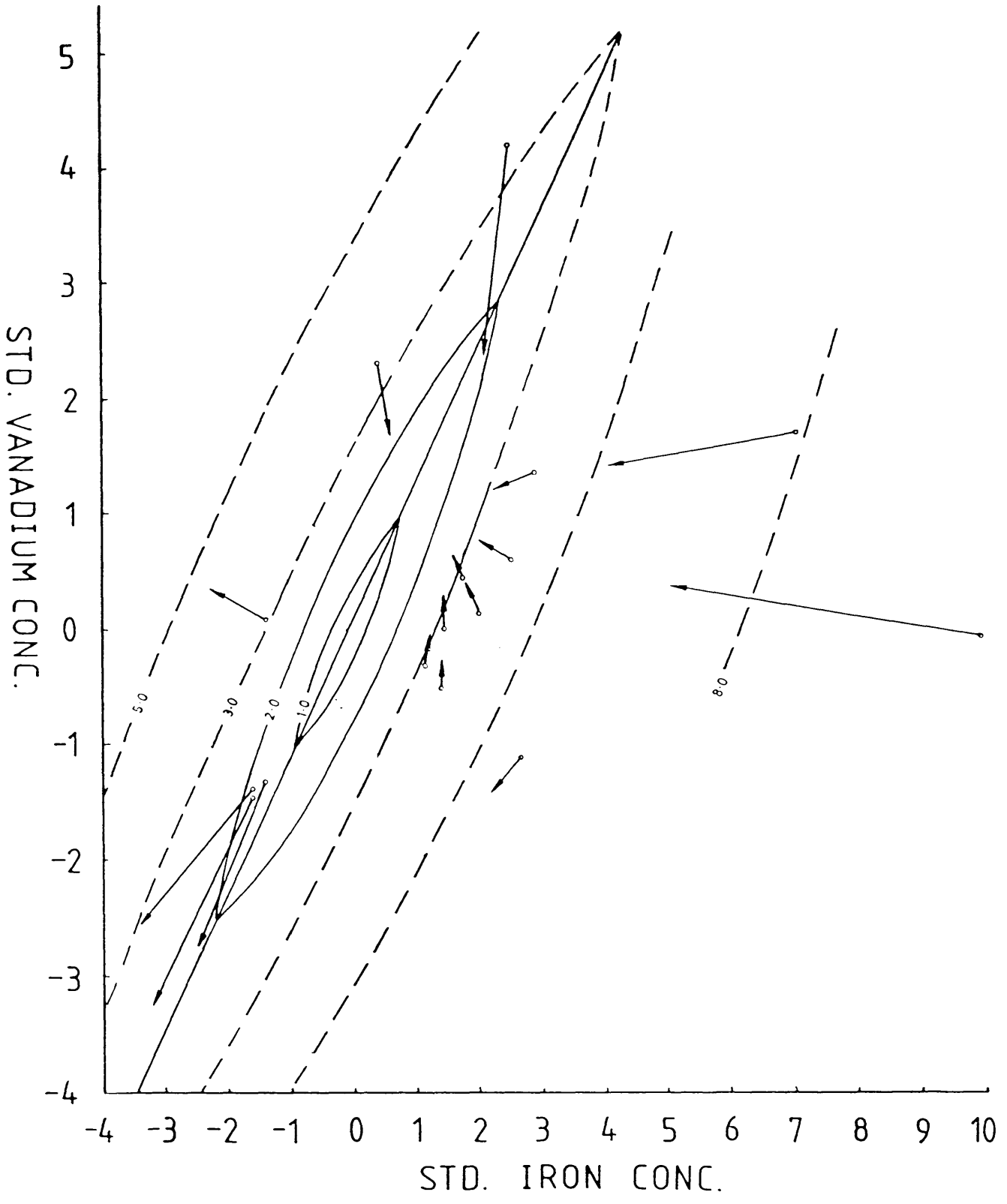


Figure 5.1

Showing the movement of outlying samples in standardized bivariate space from untransformed state → Boxcox transformed state. Contours are (Mahalanobis distance)$^{2/3}$ those in solid lines are well-defined, those in dashed lines are estimated.

of the perpendicular distance from the bivariate regression line.

5) Transformation will tend to suppress bivariate and univariate anomalies. Samples anomalous in only one variable will show most change in bivariate space in that variable, and will show negligible change in the other variable.

6) The background distribution in the raw data will be extended in bivariate space on transformation. Thus samples with low values in both variables will move in bivariate space similar distances to anomalously high samples.

7) This phenomenon will result in a separation of background samples, i.e. an extension of the background distribution, with the result that background relationships between samples will be enhanced.

8) The Euclidean distances between bivariate or univariate anomalies (in this case meaning anomalously high samples) decreases on transformation. The decrease is directly proportional to the size of the anomaly. This results in a decrease in contrast between anomalous samples.

9) Conversely, the extension of background samples on transformation is accompanied by an increase in Euclidean distances between samples which have background values in both variables. This results in an increase in contrast between background samples.

10) Summary.

a: Transformation is advantageous when looking at background relationships - the suppression of anomalies is good for this purpose.

b: Transformation is undesirable when attempting to fulfill the major task of geochemical exploration - that of trying to detect anomalous samples. For this task we need the maximum contrast obtainable between background and anomalies.

5.6 Data Transformations and the SIND Data

The purpose of this section is to analyse and discuss the effect of transformations on the SIND data and to attempt to answer the question posed earlier as to whether normalizing transformations are really necessary.

The first step in this analysis is obviously to generate a logarithmic (to base 10) transformed data set (termed "logten") and a power transformed data set (termed "boxcox"). These are tabulated in Tables 5.1 and 5.2 respectively and, with the raw data in Table 2.1, form the basis for this analysis. Let us first of all look at the univariate frequency distributions for each of these three data sets. As discussed in Section 4.3 we may do this best by studying probability plots - these are given in Figures 5.2 through 5.7 (note that Figures 4.2 through 4.3 are repeated here for convenience). For a detailed analysis of the raw data probability plots see Section 4.3. It is not necessary to give here a detailed analysis of the logten and boxcox plots since the main features are the same as for the raw data. The main differences between the raw data plots and the logten or boxcox plots are that the contrast between the anomalous objects and the background is reduced from the former to the latter, and the information presented about the background distribution is increased. These observations are consistent with the conclusions in Section 5.5.

The primary aim of data transformation is to achieve normality. Bearing in mind that a straight line relationship in a probability plot is indicative of a normal sample distribution, it is clear from the SIND probability plots that logten transformation represents a dramatic improvement in normality over raw data while the

Figure 5.2

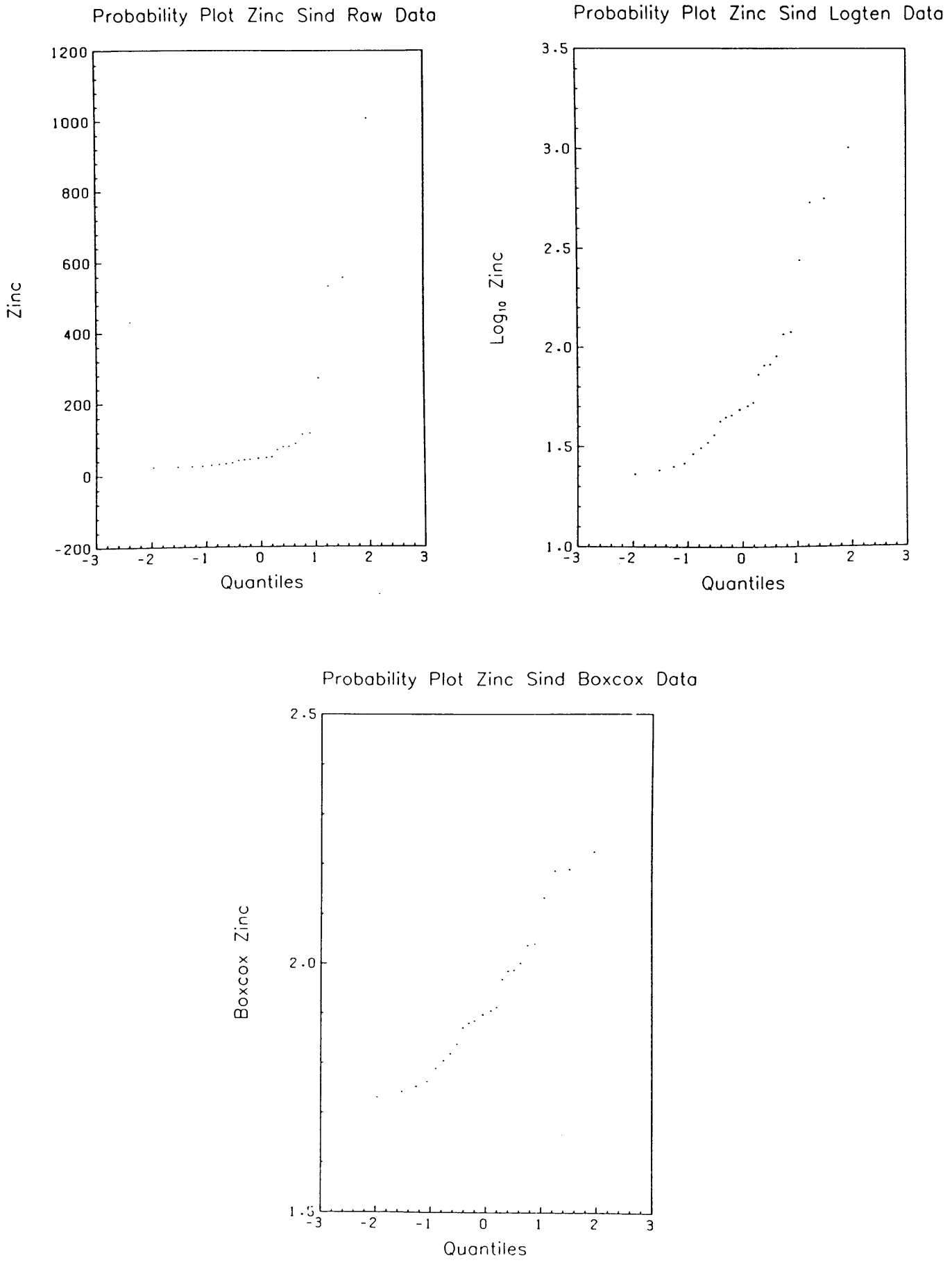
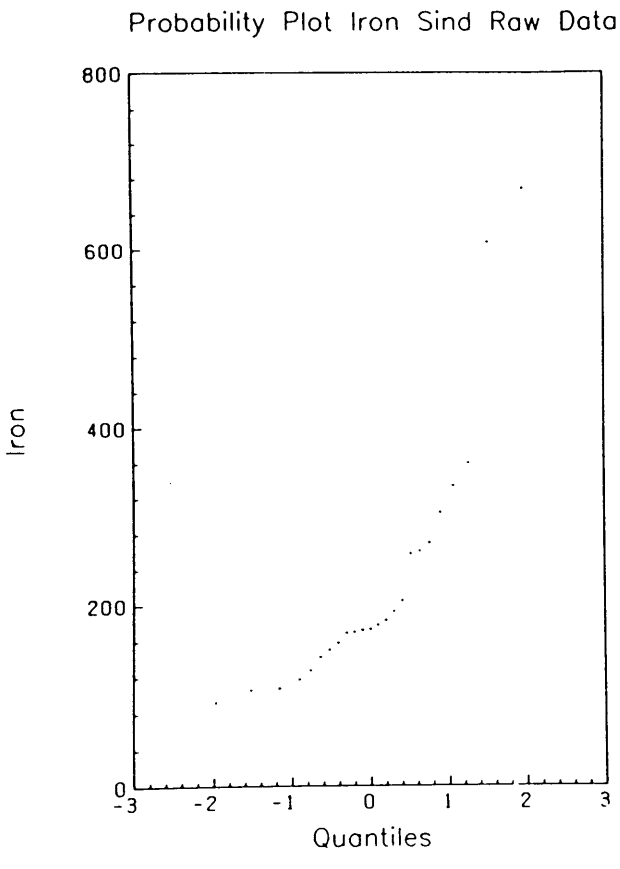
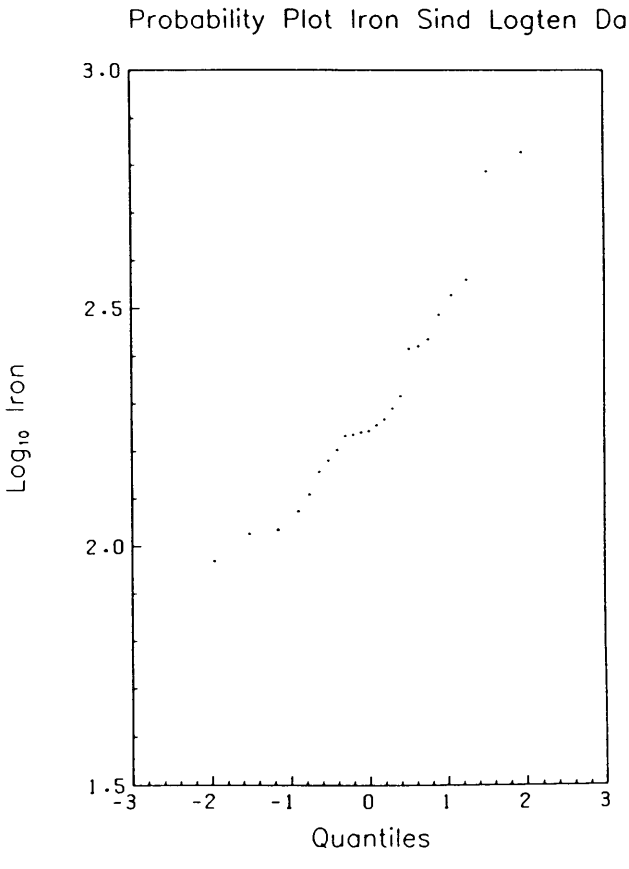

Probability Plot Zinc Sind Raw Data
Zinc
1200
1000
800
600
400
200
0
-200
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Zinc Sind Logten Data
Log10 Zinc
3.5
3.0
2.5
2.0
1.5
1.0
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Zinc Sind Boxcox Data
Boxcox Zinc
2.5
2.0
1.5
-3
-2
-1
0
1
2
3
Quantiles

Figure 5.3


Probability Plot Iron Sind Raw Data
800
600
400
200
0
Iron
-3
-2
-1
0
1
2
3
Quantiles


Probability Plot Iron Sind Logten Data
3.0
2.5
2.0
1.5
Log10 Iron
-3
-2
-1
0
1
2
3
Quantiles

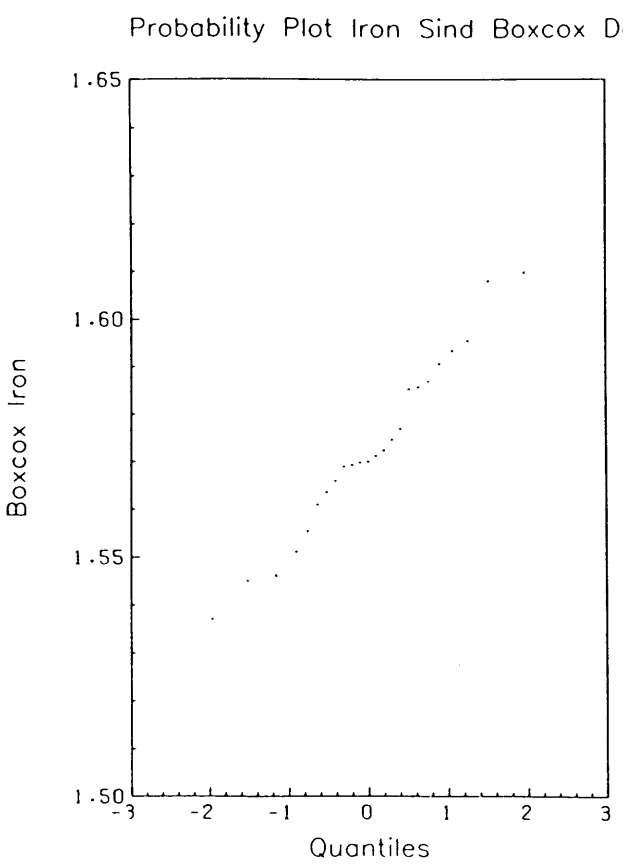
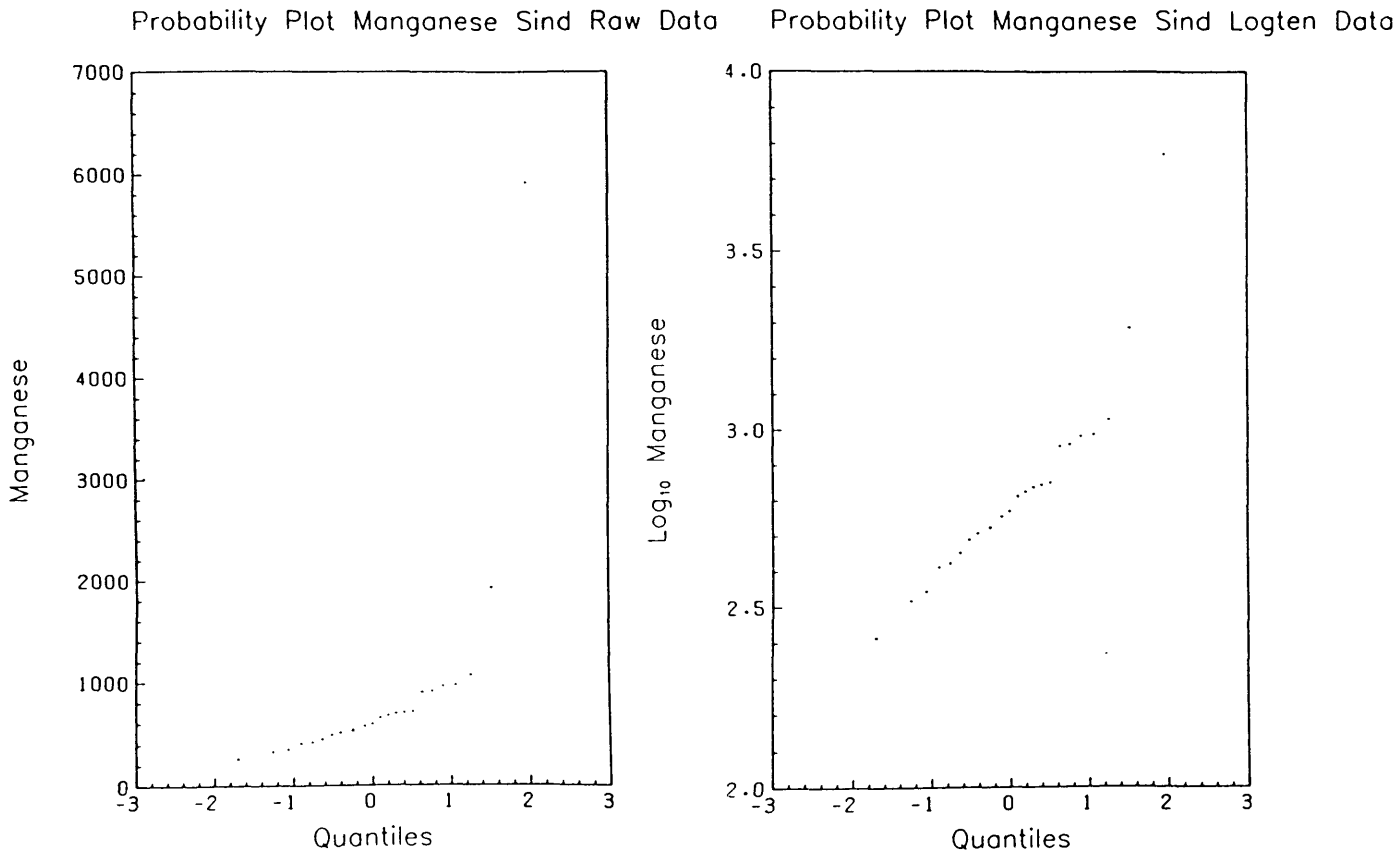

Probability Plot Iron Sind Boxcox Data
1.65
1.60
1.55
1.50
Boxcox Iron
-3
-2
-1
0
1
2
3
Quantiles

Figure 5.4


Probability Plot Manganese Sind Raw Data
Manganese
7000
6000
5000
4000
3000
2000
1000
0
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Manganese Sind Logten Data
Log10 Manganese
4.0
3.5
3.0
2.5
2.0
-3
-2
-1
0
1
2
3
Quantiles

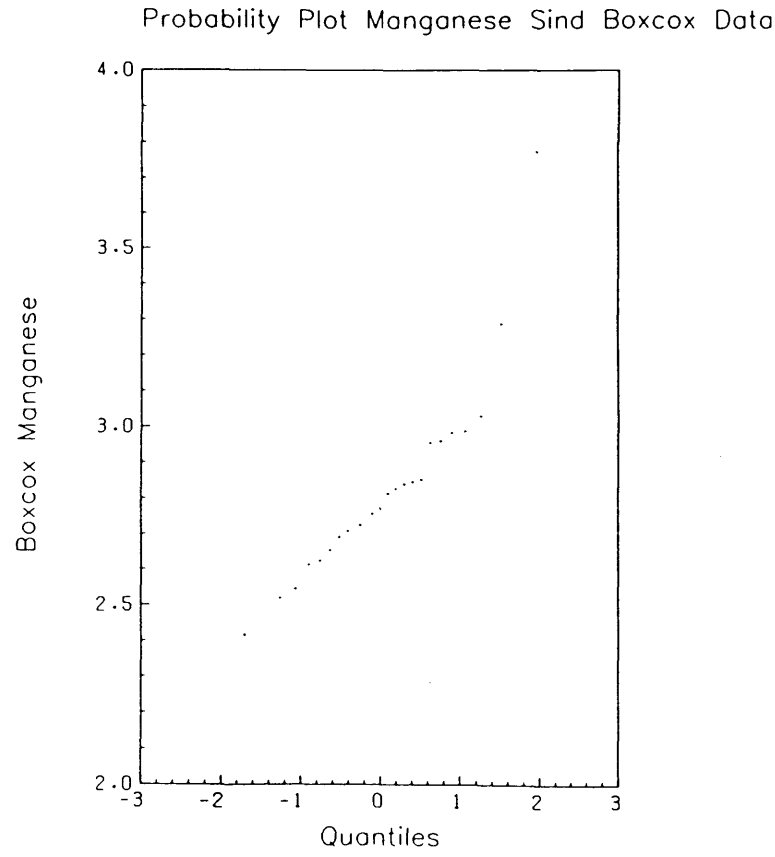
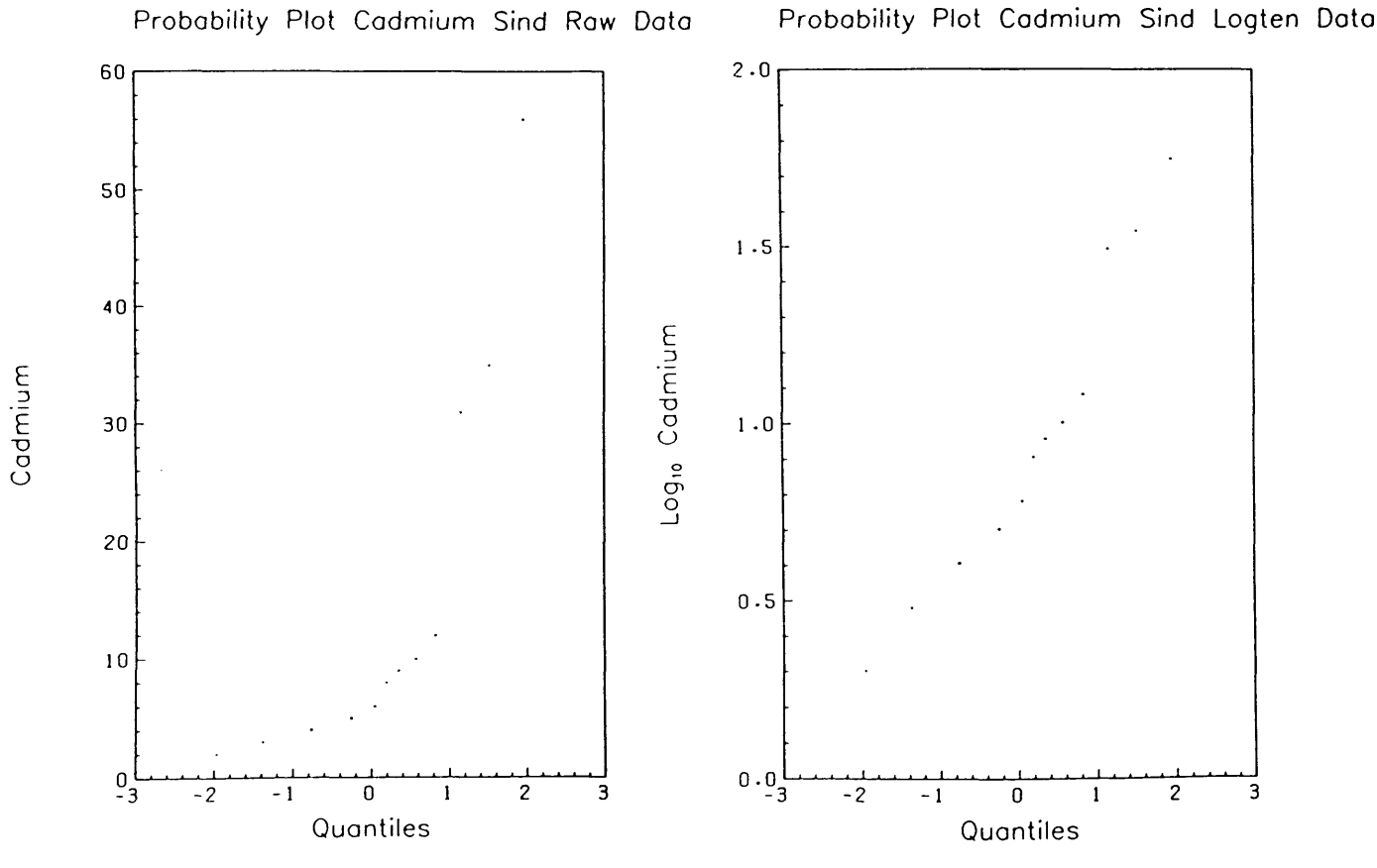

Probability Plot Manganese Sind Boxcox Data
Boxcox Manganese
4.0
3.5
3.0
2.5
2.0
-3
-2
-1
0
1
2
3
Quantiles

Figure 5.5


Probability Plot Cadmium Sind Raw Data
Cadmium
60
50
40
30
20
10
0
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Cadmium Sind Logten Data
Log10 Cadmium
2.0
1.5
1.0
0.5
0.0
-3
-2
-1
0
1
2
3
Quantiles

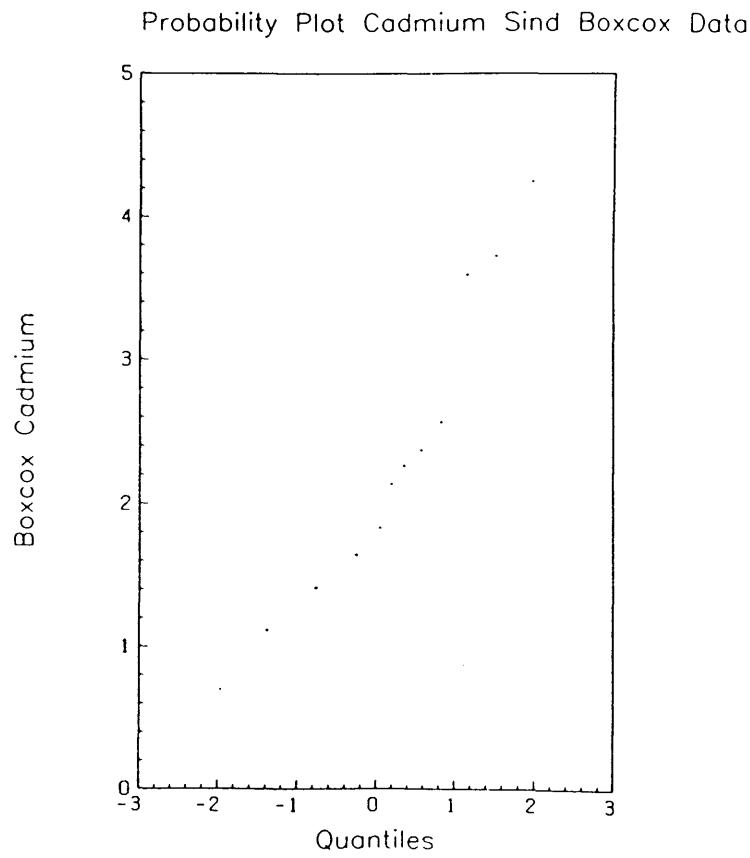

Probability Plot Cadmium Sind Boxcox Data
Boxcox Cadmium
5
4
3
2
1
0
-3
-2
-1
0
1
2
3
Quantiles

Figure 5.6

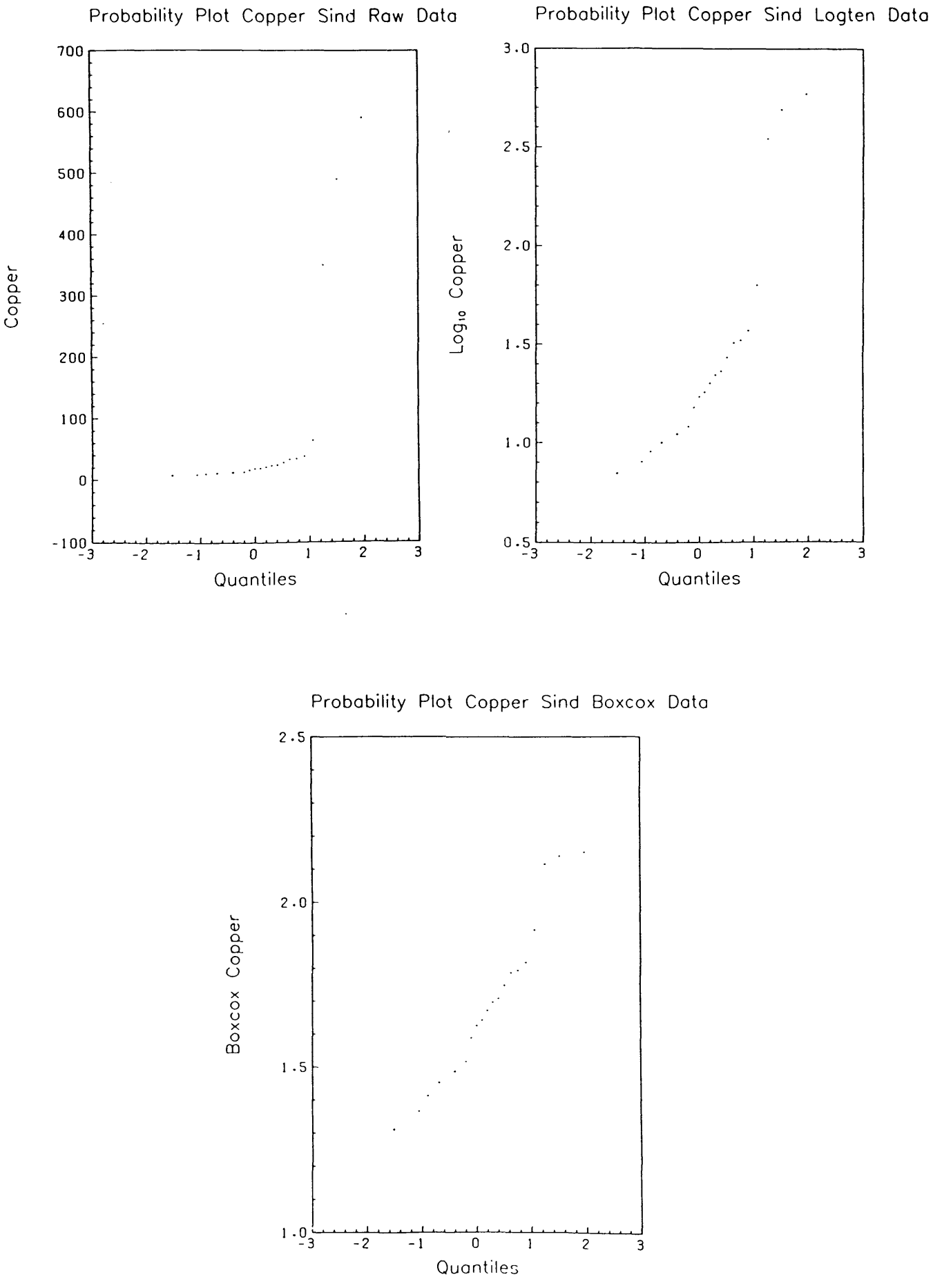
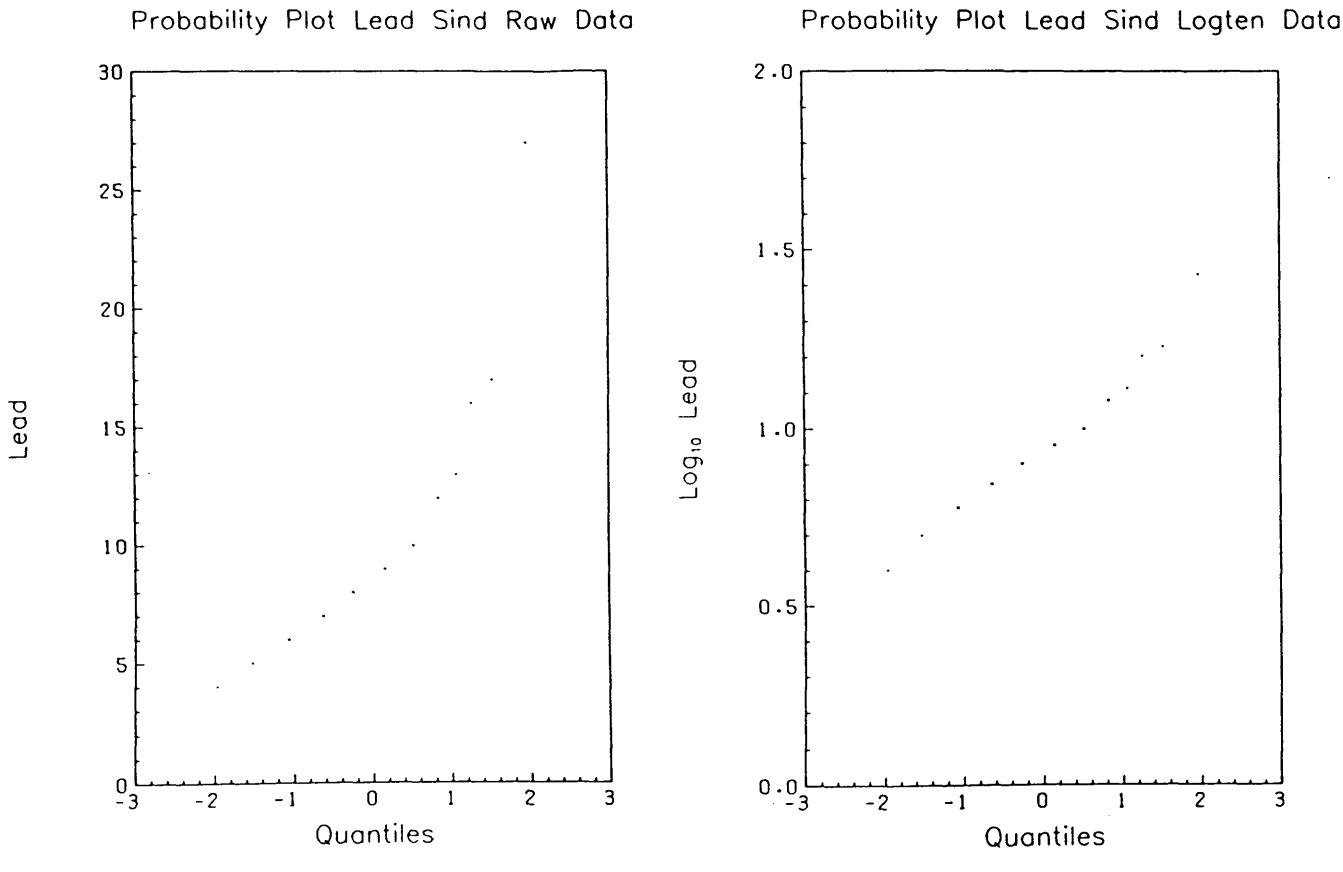

Probability Plot Copper Sind Raw Data
Copper
700
600
500
400
300
200
100
0
-100
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Copper Sind Logten Data
Log10 Copper
3.0
2.5
2.0
1.5
1.0
0.5
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Copper Sind Boxcox Data
Boxcox Copper
2.5
2.0
1.5
1.0
-3
-2
-1
0
1
2
3
Quantiles

Figure 5.7


Probability Plot Lead Sind Raw Data
Lead
30
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Lead Sind Logten Data
Log10 Lead
2.0
1.5
1.0
0.5
0.0
-3
-2
-1
0
1
2
3
Quantiles

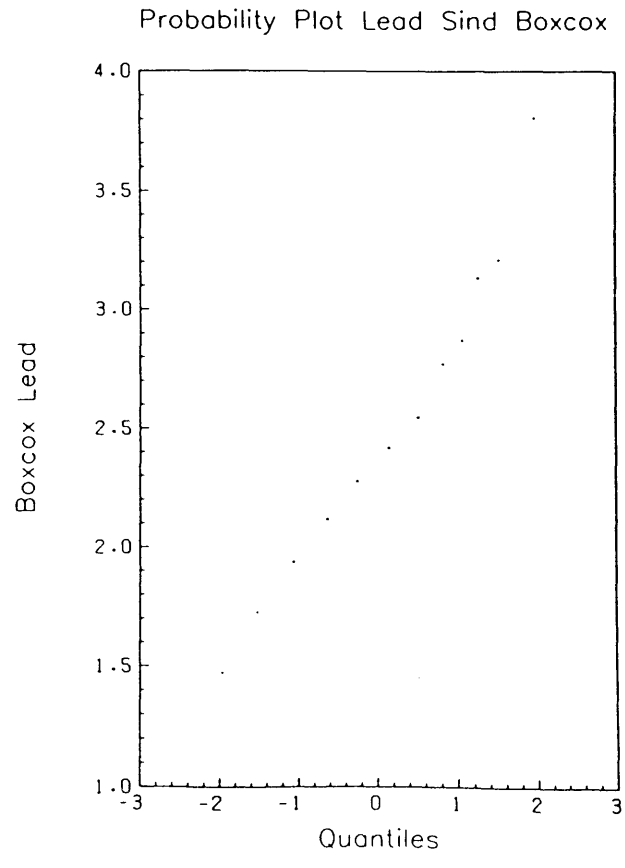

Probability Plot Lead Sind Boxcox Data
Boxcox Lead
4.0
3.5
3.0
2.5
2.0
1.5
1.0
-3
-2
-1
0
1
2
3
Quantiles

Table 5.1

SIND LOGTEN Data Matrix

Sample No.	East	North	Zn	Fe	Mn	Cd	Cu	Pb
1	97	55	1.3802	2.0334	2.5185	.6021	.8451	.6990
2	86	43	1.3979	2.0719	2.6232	.4771	.9542	.8451
3	98	43	1.6232	2.3139	2.9590	.7782	1.0792	.7782
4	109	34	1.6990	2.2380	2.8451	.6990	1.1761	.9542
5	121	15	1.7160	2.2405	2.8388	.6990	1.2553	1.0000
6	82	28	1.4624	2.0253	2.7076	.7782	1.0000	.9031
7	66	6	1.4150	2.0334	2.7243	.6990	1.0000	.8451
8	75	10	1.3617	1.9685	2.4150	.6021	.8451	.8451
9	76	71	1.9494	2.2648	2.8261	.9542	1.5051	.7782
10	65	77	1.8573	2.5250	2.7243	.6990	1.0414	.9031
11	58	68	1.4914	2.1790	2.5441	.6021	1.0414	.6021
12	63	56	2.0607	2.2014	2.8129	1.0792	1.5682	.7782
13	41	50	2.7284	2.4133	2.9823	1.5441	2.5441	.9542
14	40	43	1.6812	2.2330	2.7559	.4771	1.2304	.9542
15	22	46	3.0043	2.4330	3.0294	1.7482	2.7709	.9542
16	2	46	2.7482	2.4843	2.6532	1.4914	2.6902	.9031
17	23	21	1.6812	2.1553	2.7709	.6021	.9031	.9031
18	28	22	1.6435	2.2304	2.8513	.6021	1.0414	1.2041
19	19	3	1.5563	2.1072	2.6128	.3010	.8451	1.0000
20	133	55	1.5185	2.2878	2.6902	1.0000	1.3010	1.0792
21	108	93	1.6532	2.2529	2.4150	.9031	1.4314	1.0792
22	127	110	2.0719	2.8261	3.2856	1.0792	1.5185	1.2304
23	140	120	2.4378	2.7853	3.7723	1.4914	1.7993	1.4314
24	156	129	1.9085	2.4183	2.9868	.9542	1.3424	1.1139
25	175	133	1.9031	2.5575	2.9542	1.0000	1.3617	1.0000

Table 5.2

SIND BOXCOX Data Matrix

Sample No.	East	North	Zn	Fe	Mn	Cd	Cu	Pb
1	97	55	1.7417	1.5461	2.5185	1.4130	1.3114	1.7248
2	86	43	1.7521	1.5511	2.6232	1.1153	1.4132	2.1162
3	98	43	1.8704	1.5768	2.9590	1.8365	1.5169	1.9355
4	109	34	1.9046	1.5697	2.8451	1.6455	1.5888	2.4160
5	121	15	1.9119	1.5699	2.8388	1.6455	1.6425	2.5436
6	82	28	1.7887	1.5450	2.7076	1.8365	1.4527	2.2747
7	66	6	1.7620	1.5461	2.7243	1.6455	1.4527	2.1162
8	75	10	1.7306	1.5371	2.4150	1.4130	1.3114	2.1162
9	76	71	2.0012	1.5723	2.8261	2.2647	1.7867	1.9355
10	65	77	1.9684	1.5932	2.7243	1.6455	1.4869	2.2747
11	58	68	1.8043	1.5635	2.5441	1.4130	1.4869	1.4714
12	63	56	2.0372	1.5659	2.8129	2.5715	1.8177	1.9355
13	41	50	2.1868	1.5851	2.9823	3.7343	2.1141	2.4160
14	40	43	1.8968	1.5692	2.7559	1.1153	1.6261	2.4160
15	22	46	2.2251	1.5867	3.0294	4.2558	2.1503	2.4160
16	2	46	2.1899	1.5904	2.6532	3.6008	2.1383	2.2747
17	23	21	1.8968	1.5609	2.7709	1.4130	1.3669	2.2747
18	28	22	1.8797	1.5689	2.8513	1.4130	1.4869	3.1269
19	19	3	1.8379	1.5554	2.6128	.6998	1.3114	2.5436
20	133	55	1.8186	1.5745	2.6902	2.3768	1.6717	2.7671
21	108	93	1.8842	1.5711	2.4150	2.1398	1.7478	2.7671
22	127	110	2.0406	1.6095	3.2856	2.5715	1.7935	3.2038
23	140	120	2.1334	1.6077	3.7723	3.6008	1.9160	3.8043
24	156	129	1.9870	1.5855	2.9868	2.2647	1.6970	2.8663
25	175	133	1.9851	1.5953	2.9542	2.3768	1.7084	2.5436

boxcox transformation gives a slight improvement over logten. It is possible to quantify these observations by performing the usual chi-squared goodness-of-fit tests on each individual distribution. The chi-squared test provides a way to compare a sample (observed) distribution with a theoretical (expected) distribution. In our particular case we wish to evaluate the hypothesis that our sample distributions conform to a normal model at a certain level of significance. The results are given in Table 5.3. If the chi-squared value is greater than the corresponding critical value then we reject the hypothesis that the data are normally distributed with a 95% chance of being correct. It is possible to draw the following conclusions from these figures:

1) None of the raw data distributions conform to normality.

2) Logten transformed Zn, Fe, Cd, and Pb data are normally distributed at the specified level of significance. This is equivalent to stating that raw Zn, Fe, Cd and Pb data are lognormally distributed. Logten Mn and Cu have chi-squared values that exceed the critical level, due entirely to the presence of anomalous objects.

3) Boxcox transformation markedly improves the

Table 5.3

Chi-Squared Test Results for SIND RAW, LOGTEN and BOXCOX Data

		Zn	Fe	Mn	Cd	Cu	Pb
SIND RAW	Data						
	χ^2	93.35	43.43	65.25	48.35	53.56	39.3
Critical Value		15.51	15.51	15.51	15.51	14.07	14.07
α = 0.05							
SIND LOGTEN	Data						
	χ^2	11.55	9.91	32.63	7.76	17.90	11.71
Critical Value		15.51	14.07	16.92	15.51	14.07	16.92
α = 0.05							
SIND BOXCOX	Data						
	χ^2	3.65	4.15	32.63	7.76	3.98	9.42
Critical Value		14.07	14.07	16.92	15.51	14.07	16.92
α = 0.05							
Sample No. 23 Omitted from Mn Calculation							
	χ^2			3.55			
Critical Value				12.59			
α = 0.05							

conformity of Zn, Fe, Cd and Pb to normality and indeed results in the hypothesis for Cu being turned from rejection to acceptance. When undertaking boxcox transformation for Mn, no acceptable nonzero value for λ was found and so a logten transformation was used. The chi-squared value therefore remains the same for both logten Mn and boxcox Mn, and the hypothesis that these data sets conform to normality is rejected. However, if sample number 23 is omitted from the data the chi-squared value is dramatically reduced and the hypothesis is accepted.

We have established that the univariate distributions of the raw data are positively skewed to varying degrees, and that the boxcox transformation gives the closest approximation to distributional normality. It is now appropriate to undertake a comparative analysis of the non-robust and robust location and scale estimates for each of the three data sets.

For each of the three data sets (raw, logten and boxcox) two location and two scale estimates have been calculated. Classical estimators were used for the non-robust estimates and M-estimators were used for the robust estimates. With reference to Table 5.4 the following conclusions can now be made:

1) As would be expected for positively skewed data the

Table 5.4

Non-Robust and Robust Location and Scale Estimates for SIND RAW, LOGTEN and Boxcox data

		Zn	Fe	Mn	Cd	Cu	Pb
Location							
	Raw	139.6	225.5	871.2	11.32	73.60	9.72
	Raw Robust	49.23	175.9	572.8	5.87	15.60	8.67
	Logten	87.20	206.9	711.16	8.76	32.86	9.23
	Logten Robust	86.47	205.48	696.3	8.674	32.76	9.16
	Boxcox	83.79	213.39	711.16	10.26	27.83	9.66
	Boxcox Robust	82.85	212.38	690.92	10.10	27.55	9.53
Scale							
	Raw	231.2	144.4	1107.2	13.02	156.4	4.766
	Raw Robust	24.45	70.82	208.4	2.737	8.48	2.97
	Logten	67.73	71.9	319.68	5.306	33.17	2.534
	Logten Robust	66.89	69.9	294.74	5.203	33.18	2.43
	Boxcox	46.23	83.78	319.68	6.48	15.76	3.71
	Boxcox Robust	45.55	82.40	286.37	6.34	15.57	3.513

N.B. LOGTEN and BOXCOX estimates are given here in back-transformed units so that they may be directly compared with one another and with RAW data estimates.

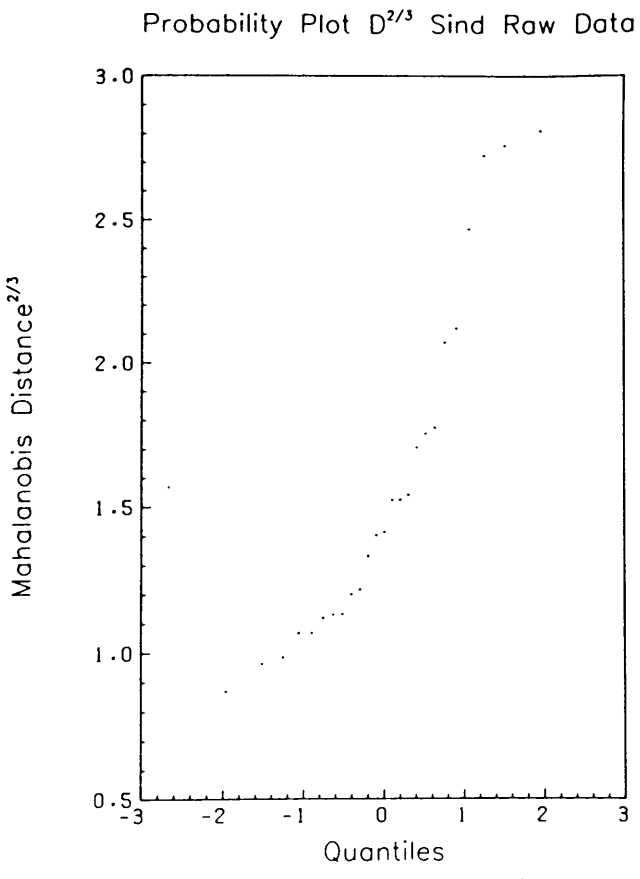
non-robust raw data estimates grossly overestimate location and scale. In contrast the differences between the non-robust and robust estimates for logten and boxcox data are very small. This is due to the fact that the univariate distributions of both these data sets closely approximate normality.

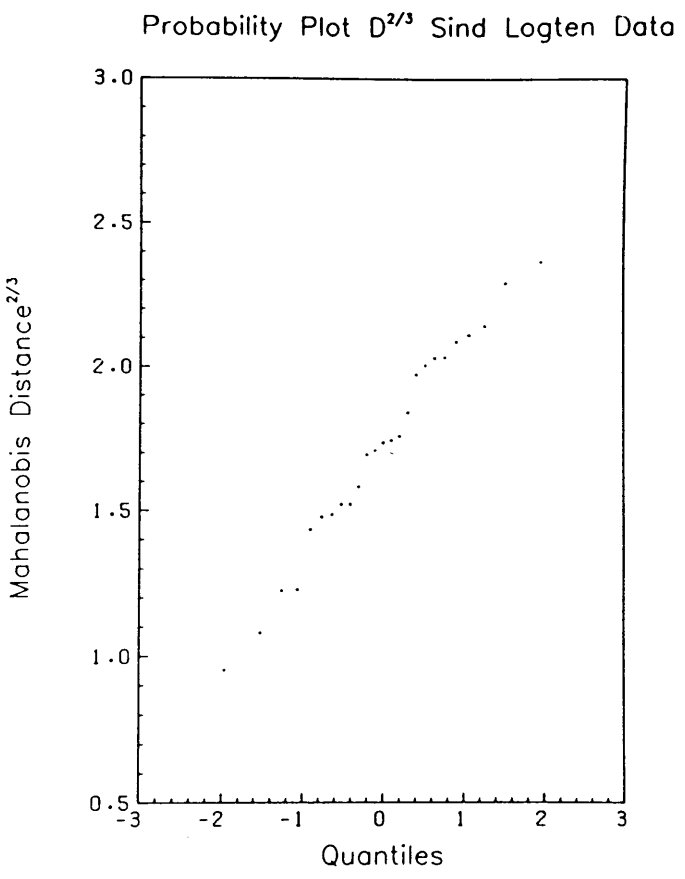
2) If we assume that the robust sample estimates for raw data are the closest available approximations to the true, unknown population location and scale parameters, then we can state that the logten and boxcox transformations tend to cause overestimation of location and scale.

Multivariate ($D_m^{2/3}$) probability plots (see Chapter 4) may be used to summarize the effects of transformation and also to show to what extent the use of robust parameters affects the multivariate distribution and anomaly contrast. Figures 5.8 through 5.9 show the non-robust and robust $D_m^{2/3}$ Q-Q plots for raw, logten and boxcox transformed SIND data. The following conclusions can be made from their analysis.

1) The two transformed data plots (see Figures 5.8 and 5.9) are very similar, demonstrating that both logten and boxcox transformations are equally effective at achieving multivariate distributional normality.

Figure 5.8


Probability Plot $D^{2/3}$ Sind Raw Data
Mahalanobis Distance$^{2/3}$
3.0
2.5
2.0
1.5
1.0
0.5
-3
-2
-1
0
1
2
3
Quantiles


Probability Plot $D^{2/3}$ Sind Logten Data
Mahalanobis Distance$^{2/3}$
3.0
2.5
2.0
1.5
1.0
0.5
-3
-2
-1
0
1
2
3
Quantiles

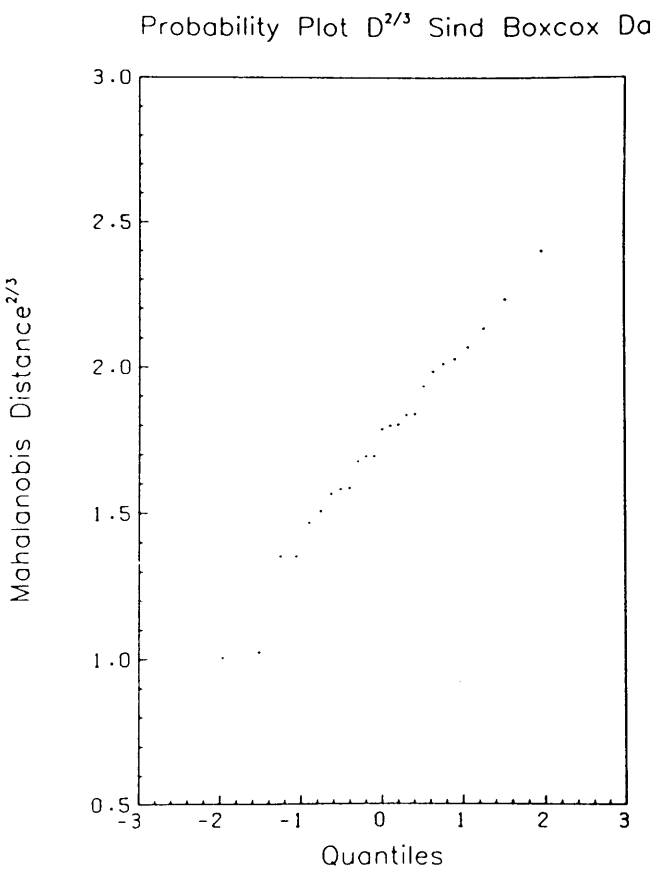
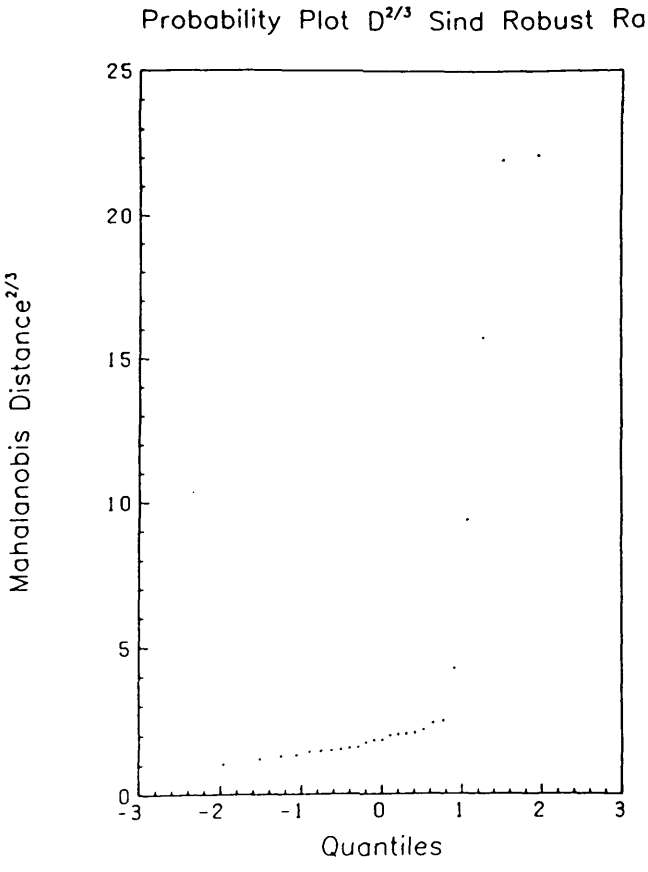
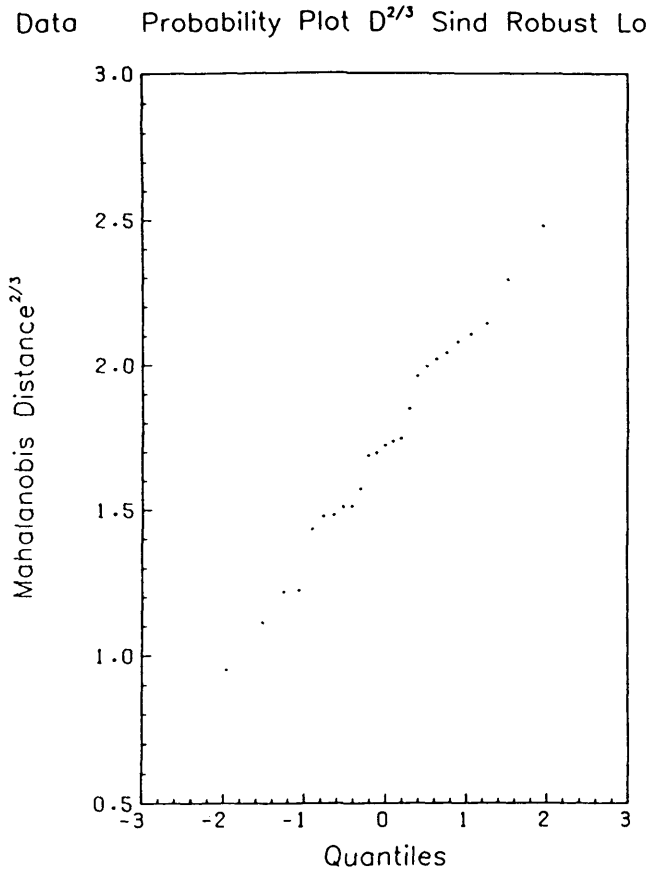
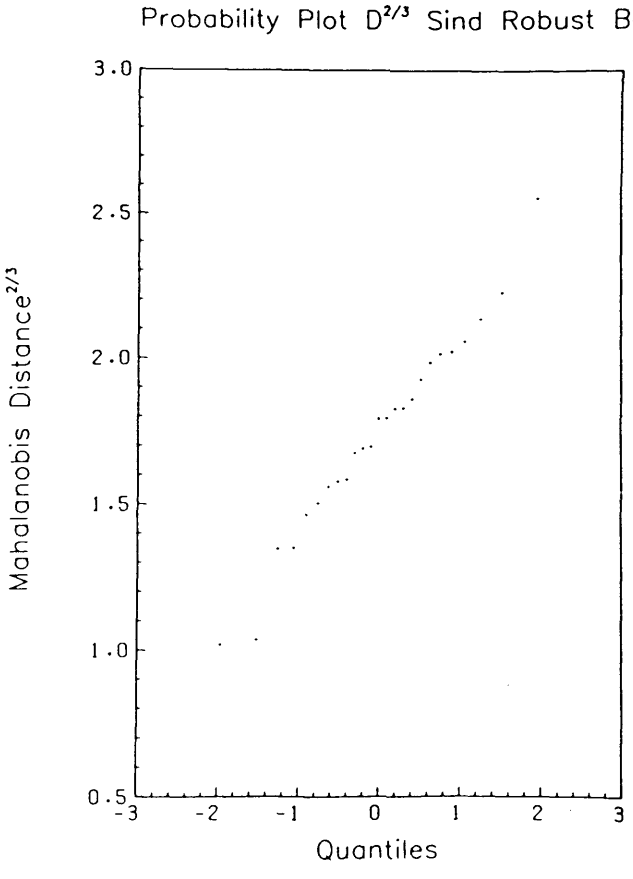

Probability Plot $D^{2/3}$ Sind Boxcox Data
Mahalanobis Distance$^{2/3}$
3.0
2.5
2.0
1.5
1.0
0.5
-3
-2
-1
0
1
2
3
Quantiles

Figure 5.9


Probability Plot $D^{2/3}$ Sind Robust Raw Data
Mahalanobis Distance$^{2/3}$
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
Quantiles


Probability Plot $D^{2/3}$ Sind Robust Logten Data
Mahalanobis Distance$^{2/3}$
3.0
2.5
2.0
1.5
1.0
0.5
-3
-2
-1
0
1
2
3
Quantiles


Probability Plot $D^{2/3}$ Sind Robust Boxcox Data
Mahalanobis Distance$^{2/3}$
3.0
2.5
2.0
1.5
1.0
0.5
-3
-2
-1
0
1
2
3
Quantiles

2) The use of robust parameters in the calculation of $D_m^{2/3}$ values has had extremely little effect on the Q-Q plots for logten and boxcox transformed data. This is because there is very little difference between the non-robust and robust parameters for the two data sets.

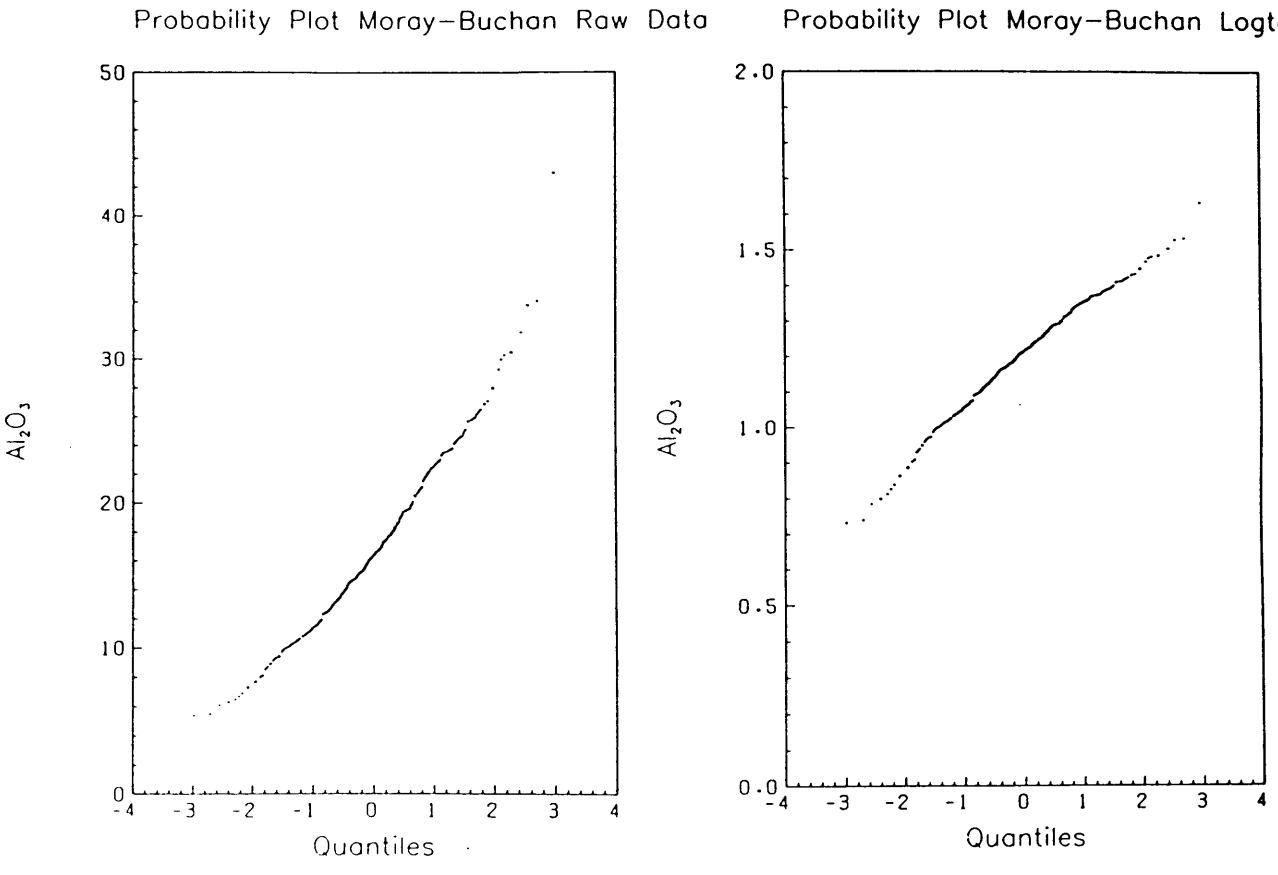
3) The non-robust $D_m^{2/3}$ Q-Q plot for raw data shows a slight positive skew. Due to overestimation by classical estimators of background location and scale parameters the contrast between the five known anomalies and the background object vectors is minimal.

4) Robust estimators used to generate robust $D_m^{2/3}$ values have been generated by effectively ignoring the five anomalous object vectors. This has the effect of dramatically reducing the length of the mean vector and making the covariance matrix reflect the background structure of the data. These factors combine to dramatically increase the contrast between the five anomalous object vectors and the background.

5.7 Data Transformations and the Moray-Buchan Data

As with the SIND data, the first step in the analysis of the MB data must be to generate a logten transformed data set and a boxcox transformed data set. A comparison of the univariate frequency distributions by means of Q-Q plots (Figures 5.10 through 5.33) demonstrates one of the principal arguments against logarithmic transformation. If the raw data frequency distribution shows a very strong positive skew the equivalent logten transformed distribution is likely to show a slight positive skew (see Ni, Sr, Ti and Y). Alternatively, if the raw distribution shows a slight positive skew then the equivalent logten distribution is likely to show a slight negative skew (see Al, K, Li and Zr). The logten transform is thus insensitive to the distribution of the data it purports to be normalizing. This is not true of the boxcox transform, which optimizes the normality of the transformed data. In the eight Q-Q plots specifically cited above the boxcox transformed distribution shows a perfect Gaussian form. Of the remaining 16 variables, 10 which show strong positive skew in Q-Q plots of raw data show close approximation to normality on logten and boxcox transformation. The remaining six variables show distinct improvements on transformation but demonstrate a tendency to

Figure 5.10


Probability Plot Moray-Buchan Raw Data
Al2O3
50
40
30
20
10
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Al2O3
2.0
1.5
1.0
0.5
0.0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

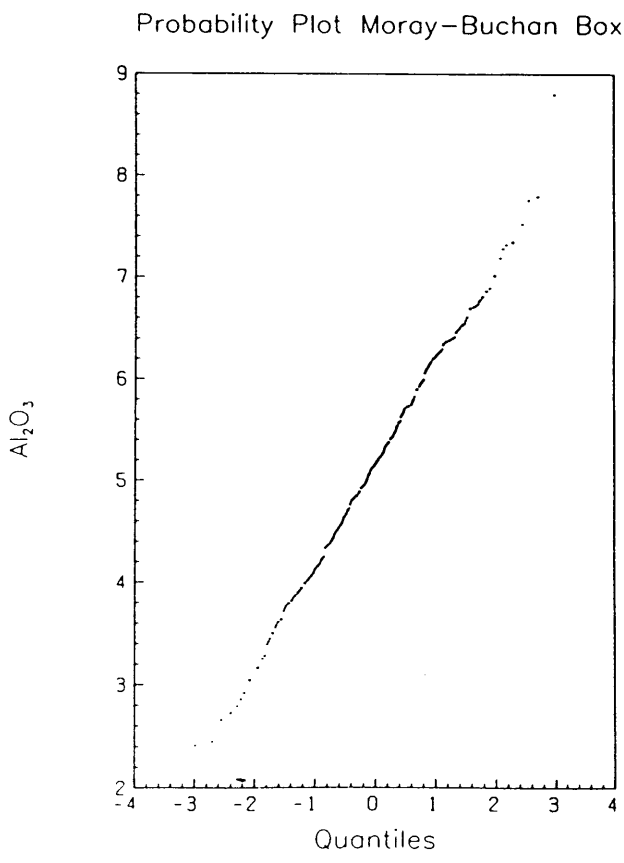

Probability Plot Moray-Buchan Boxcox Data
Al2O3
9
8
7
6
5
4
3
2
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.11

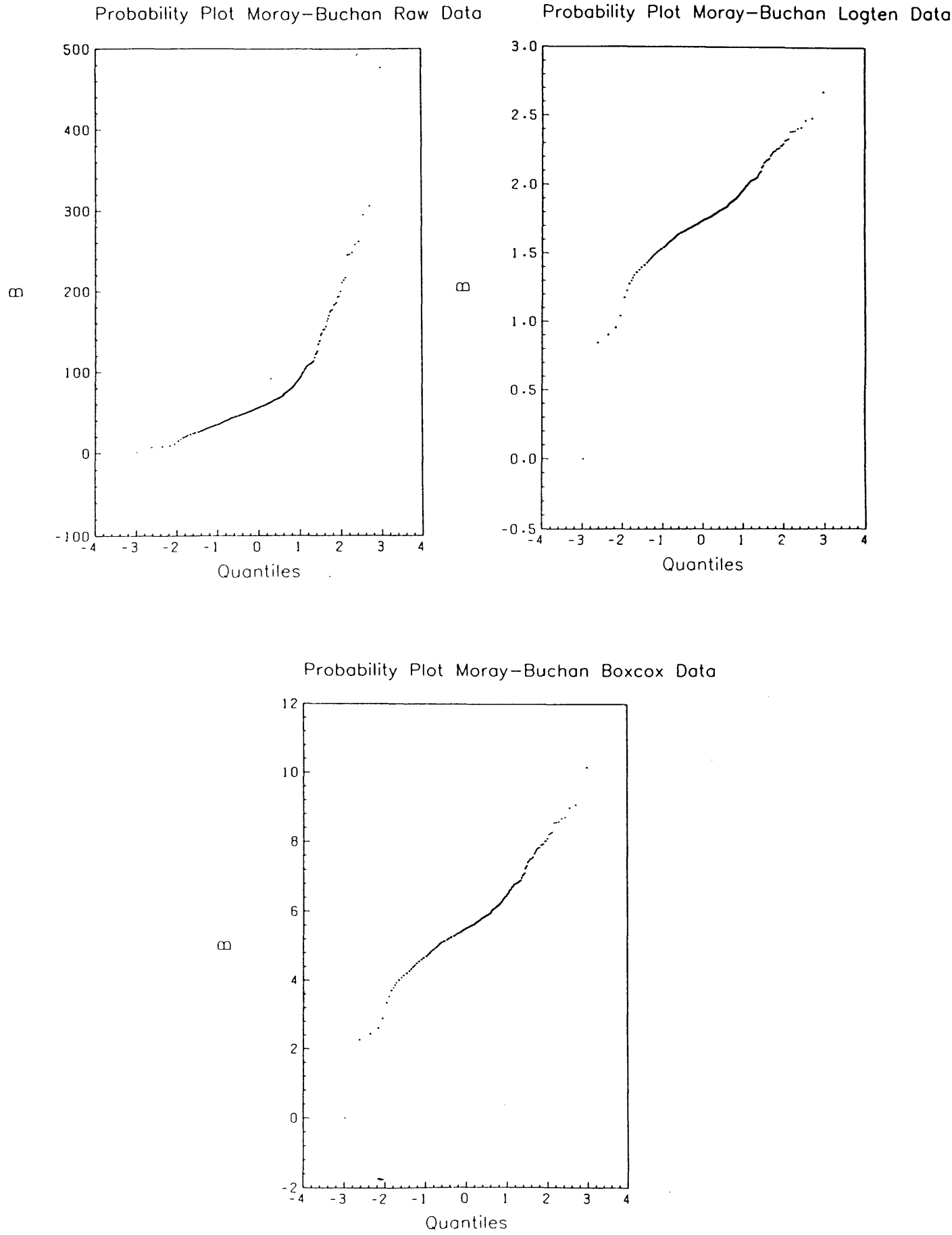
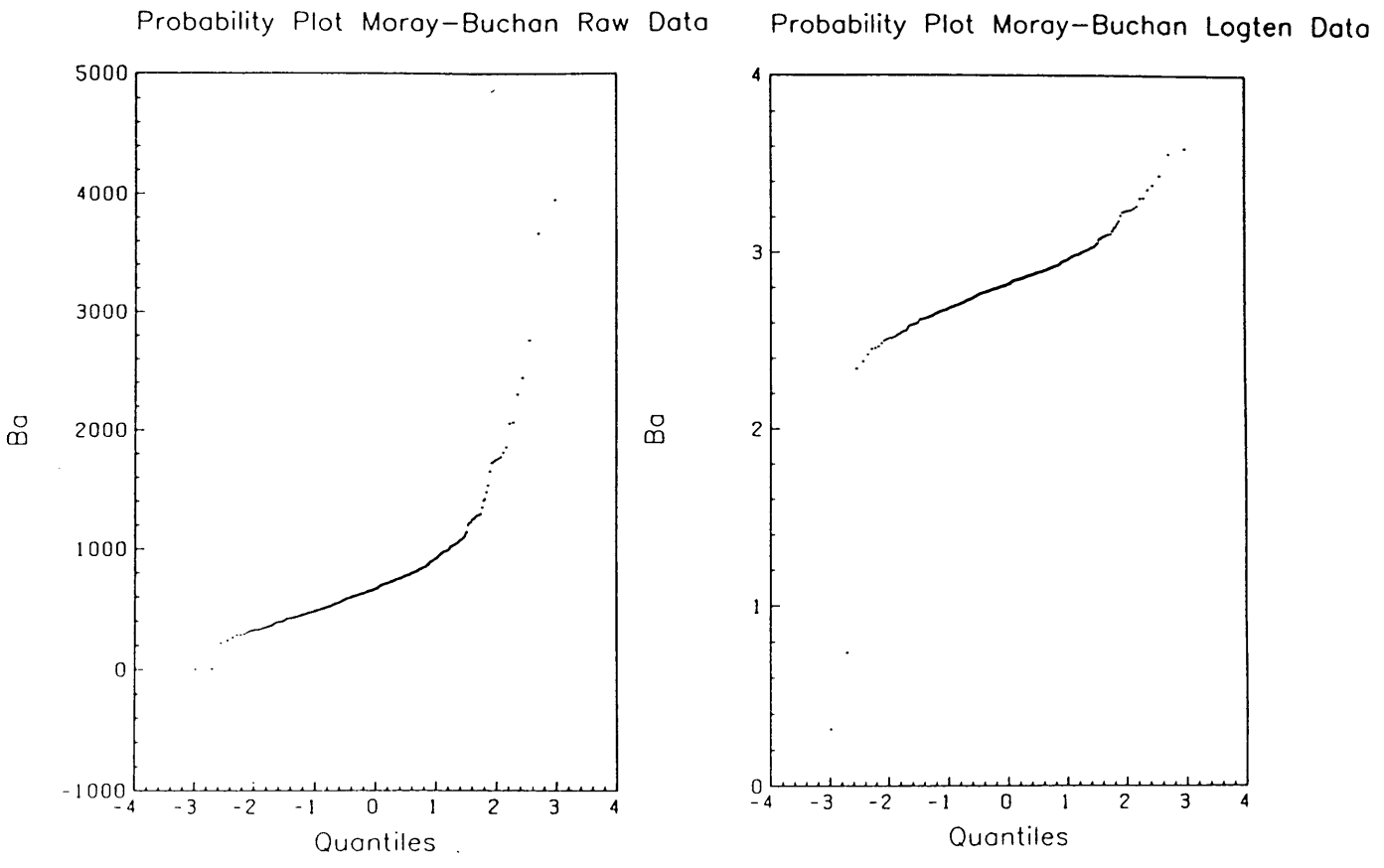

Probability Plot Moray-Buchan Raw Data
B
500
400
300
200
100
0
-100
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
B
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Boxcox Data
B
12
10
8
6
4
2
0
-2
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.12


Probability Plot Moray-Buchan Raw Data
Ba
5000
4000
3000
2000
1000
0
-1000
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Ba
4
3
2
1
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

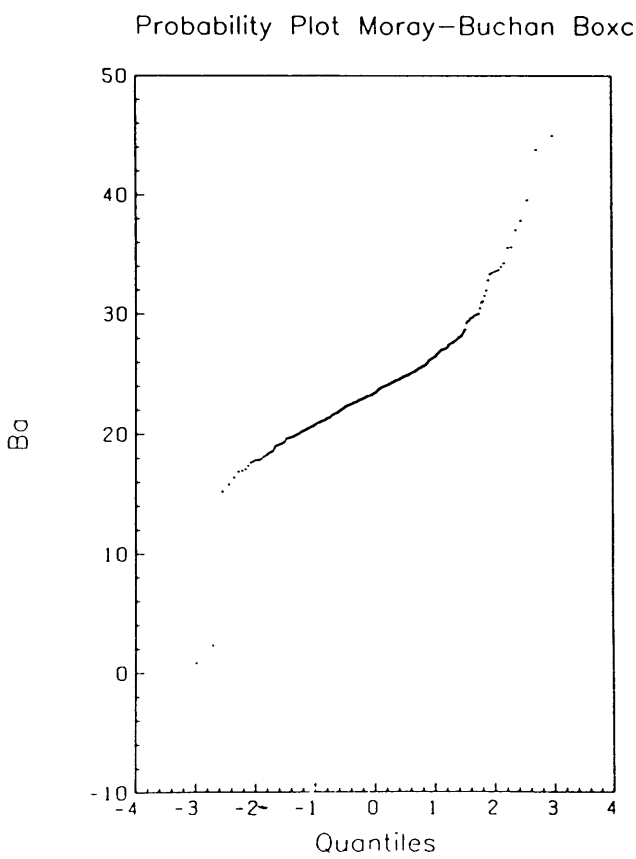

Probability Plot Moray-Buchan Boxcox Data
Ba
50
40
30
20
10
0
-10
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.13

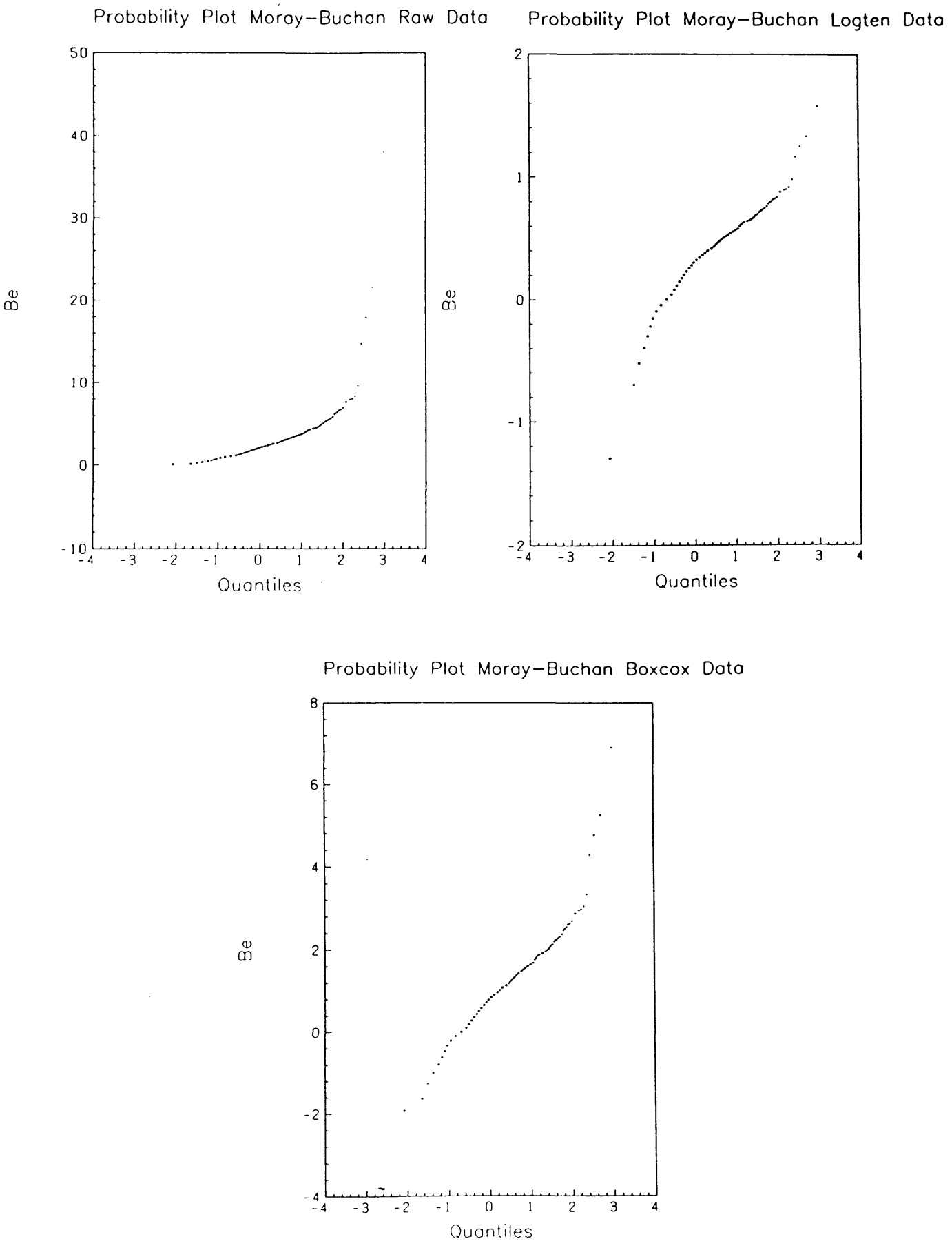
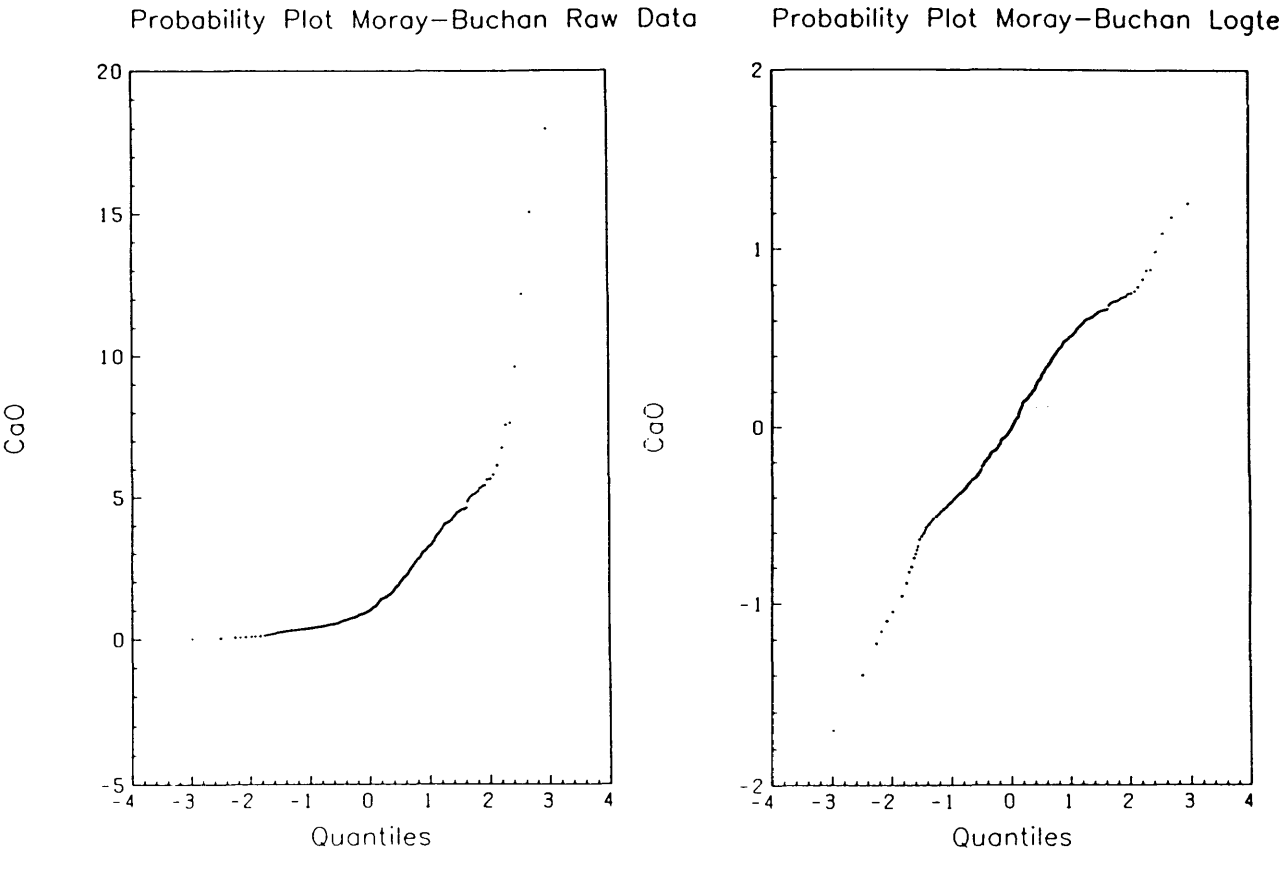

Probability Plot Moray-Buchan Raw Data
Be
Quantiles
Probability Plot Moray-Buchan Logten Data
Be
Quantiles
Probability Plot Moray-Buchan Boxcox Data
Be
Quantiles

Figure 5.14


Probability Plot Moray–Buchan Raw Data
CaO
Quantiles
Probability Plot Moray–Buchan Logten Data
CaO
Quantiles

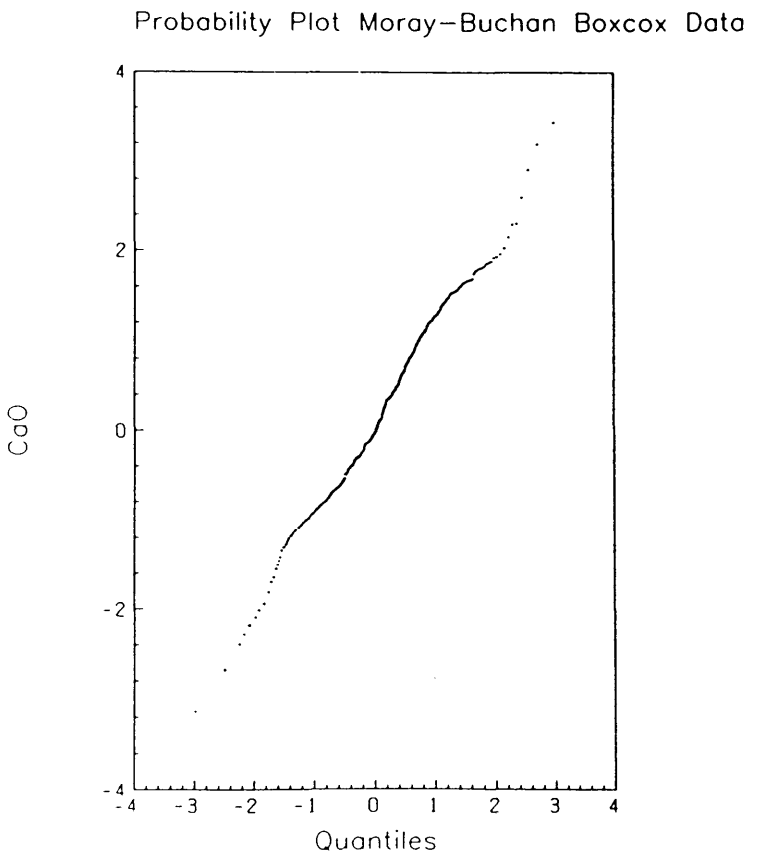
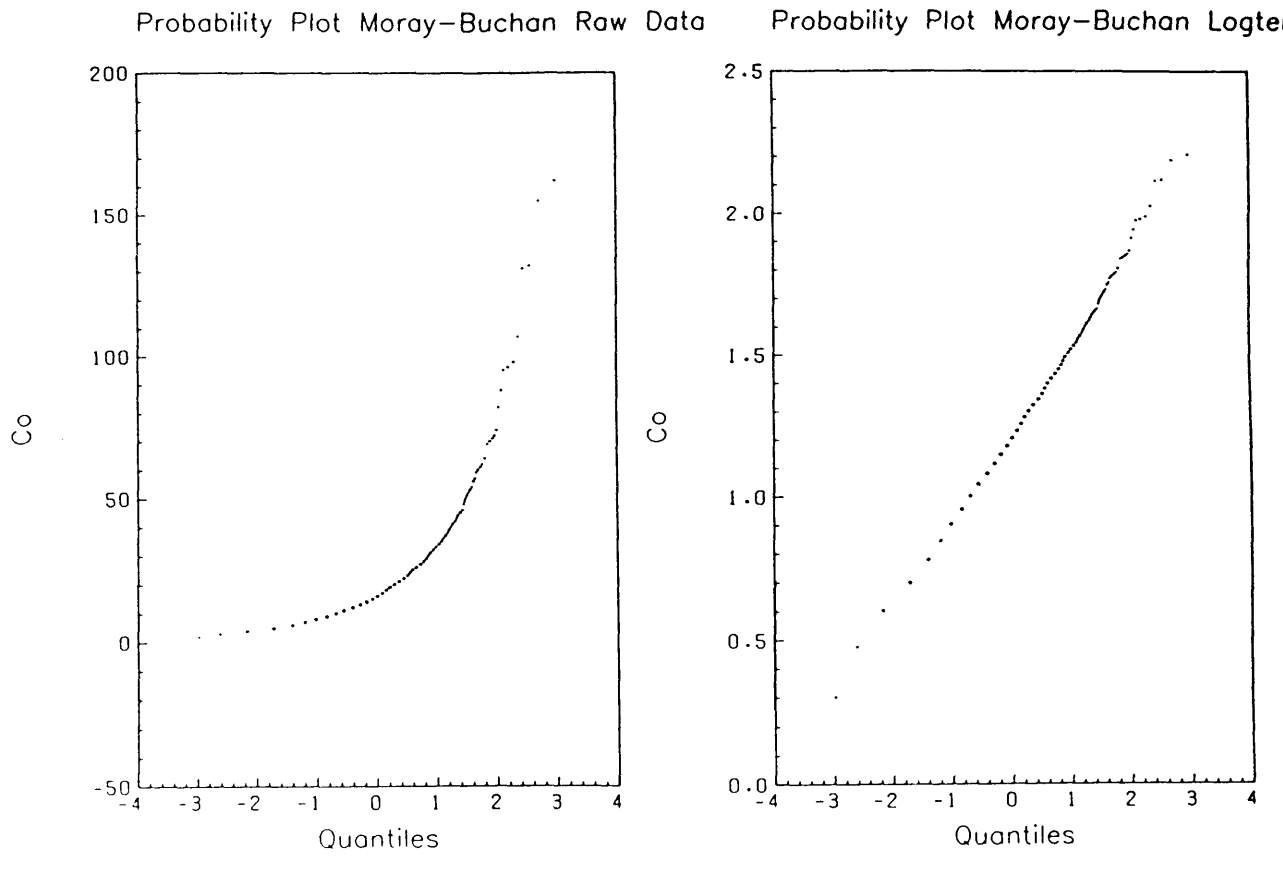

Probability Plot Moray–Buchan Boxcox Data
CaO
Quantiles

Figure 5.15


Probability Plot Moray-Buchan Raw Data
Co
200
150
100
50
0
-50
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Co
2.5
2.0
1.5
1.0
0.5
0.0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

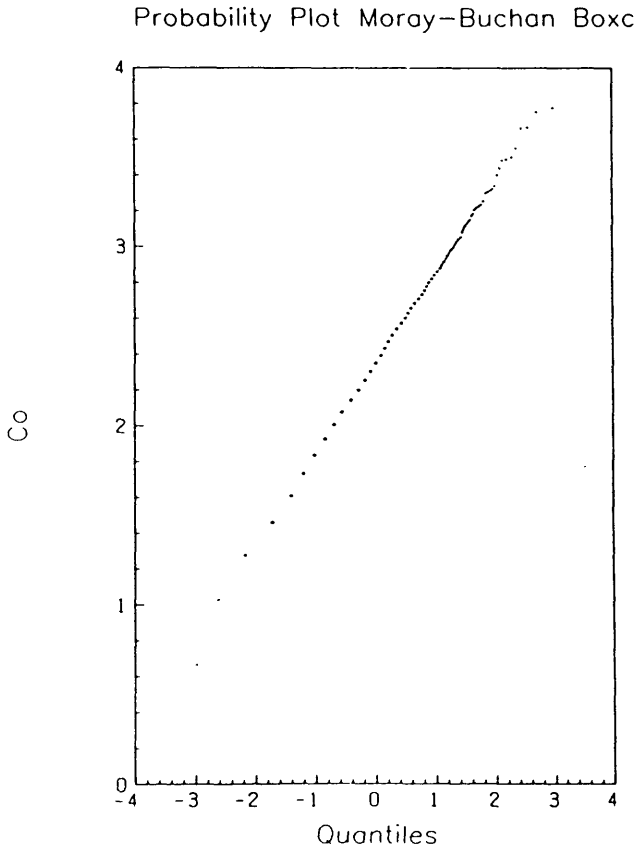
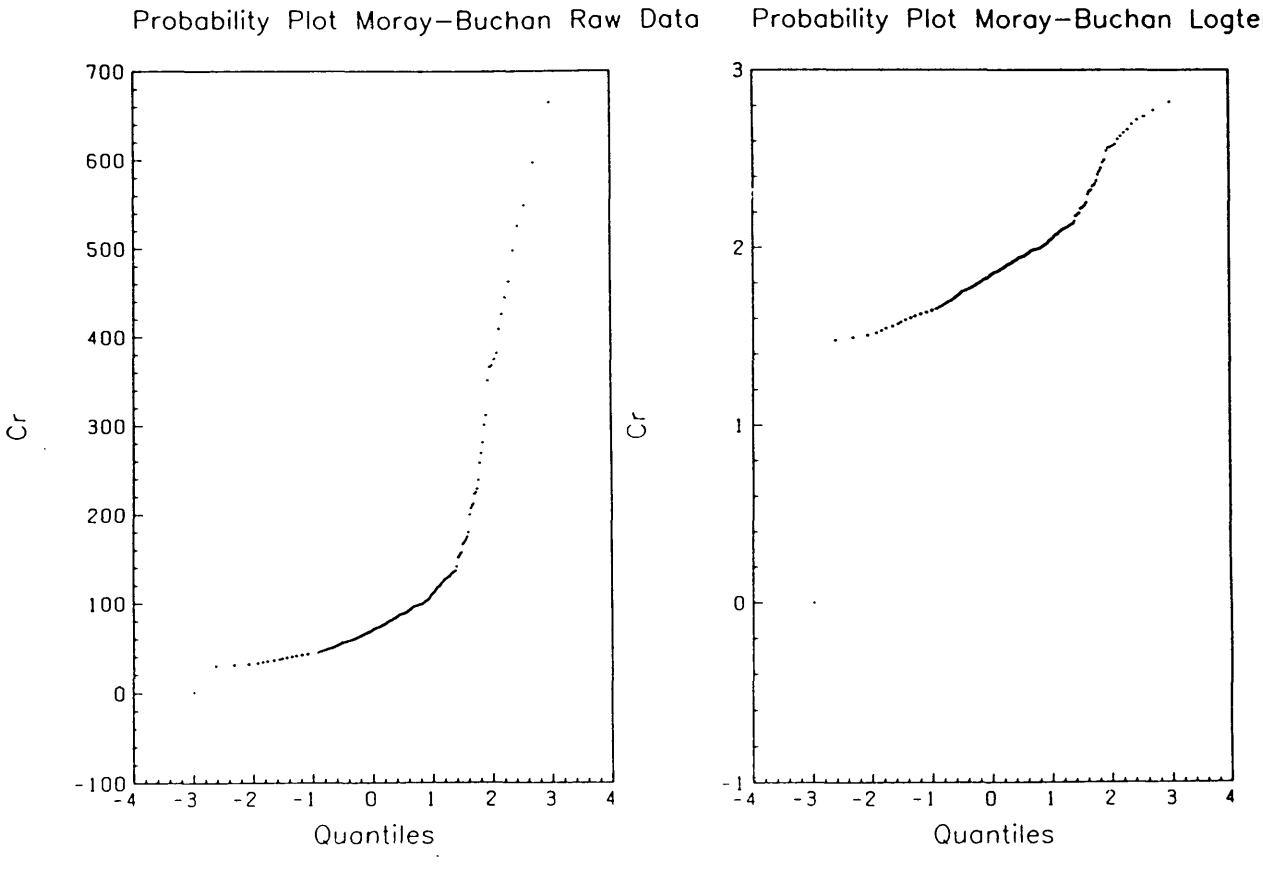

Probability Plot Moray-Buchan Boxcox Data
Co
4
3
2
1
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.16


Probability Plot Moray-Buchan Raw Data
Cr
700
600
500
400
300
200
100
0
-100
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Cr
3
2
1
0
-1
-4 -3 -2 -1 0 1 2 3 4
Quantiles

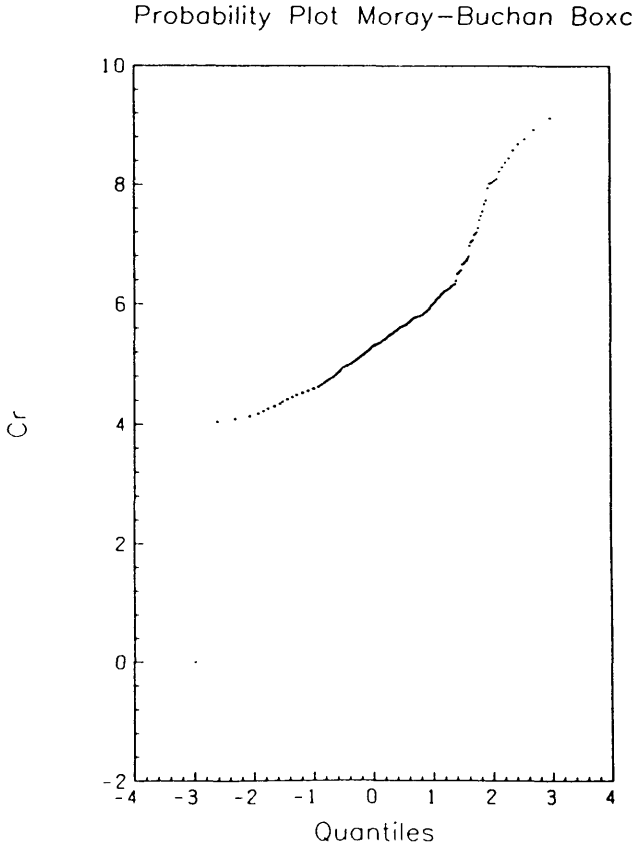

Probability Plot Moray-Buchan Boxcox Data
Cr
10
8
6
4
2
0
-2
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.17

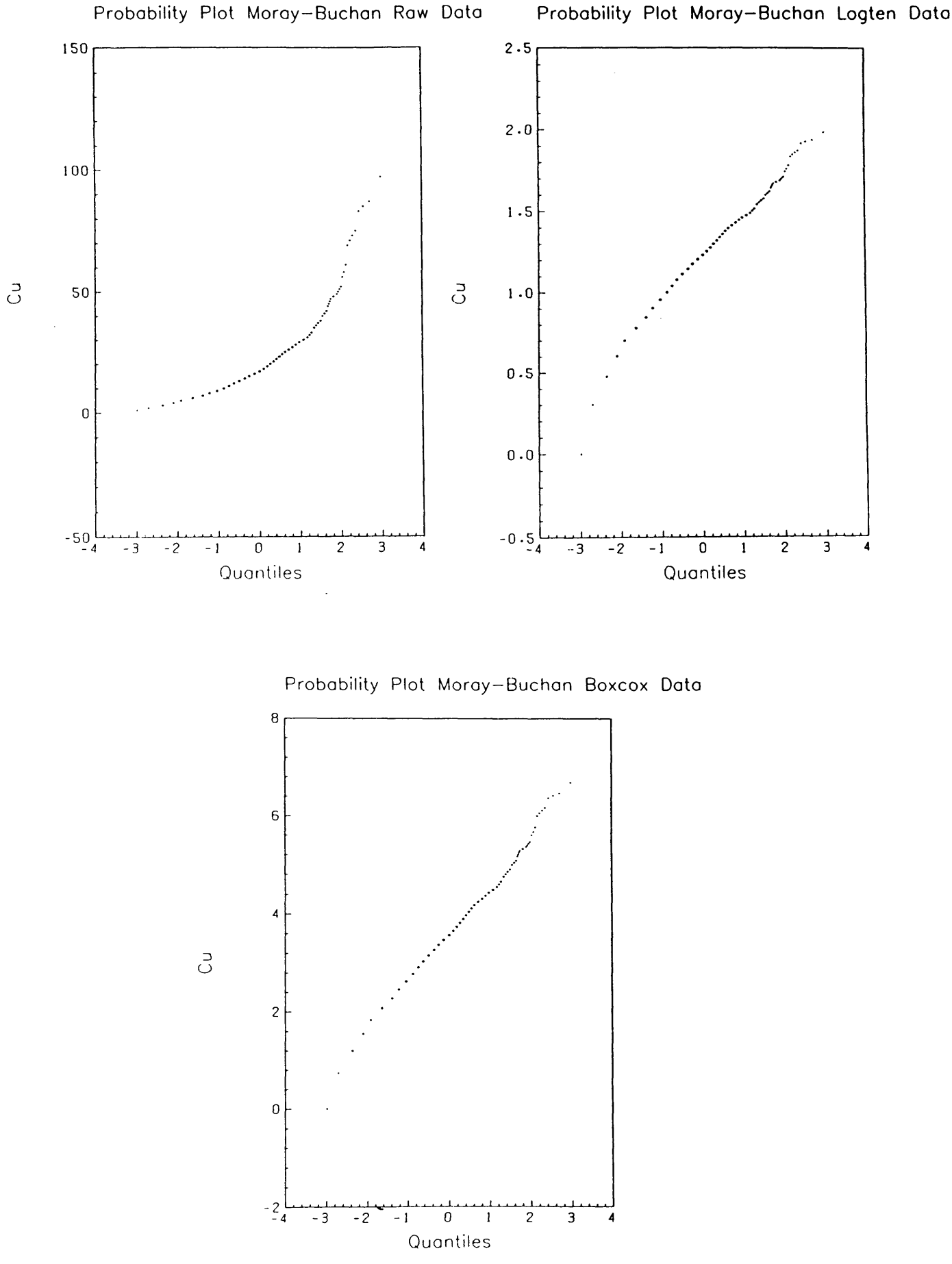

Probability Plot Moray-Buchan Raw Data
Cu
150
100
50
0
-50
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Cu
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Boxcox Data
Cu
8
6
4
2
0
-2
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.18

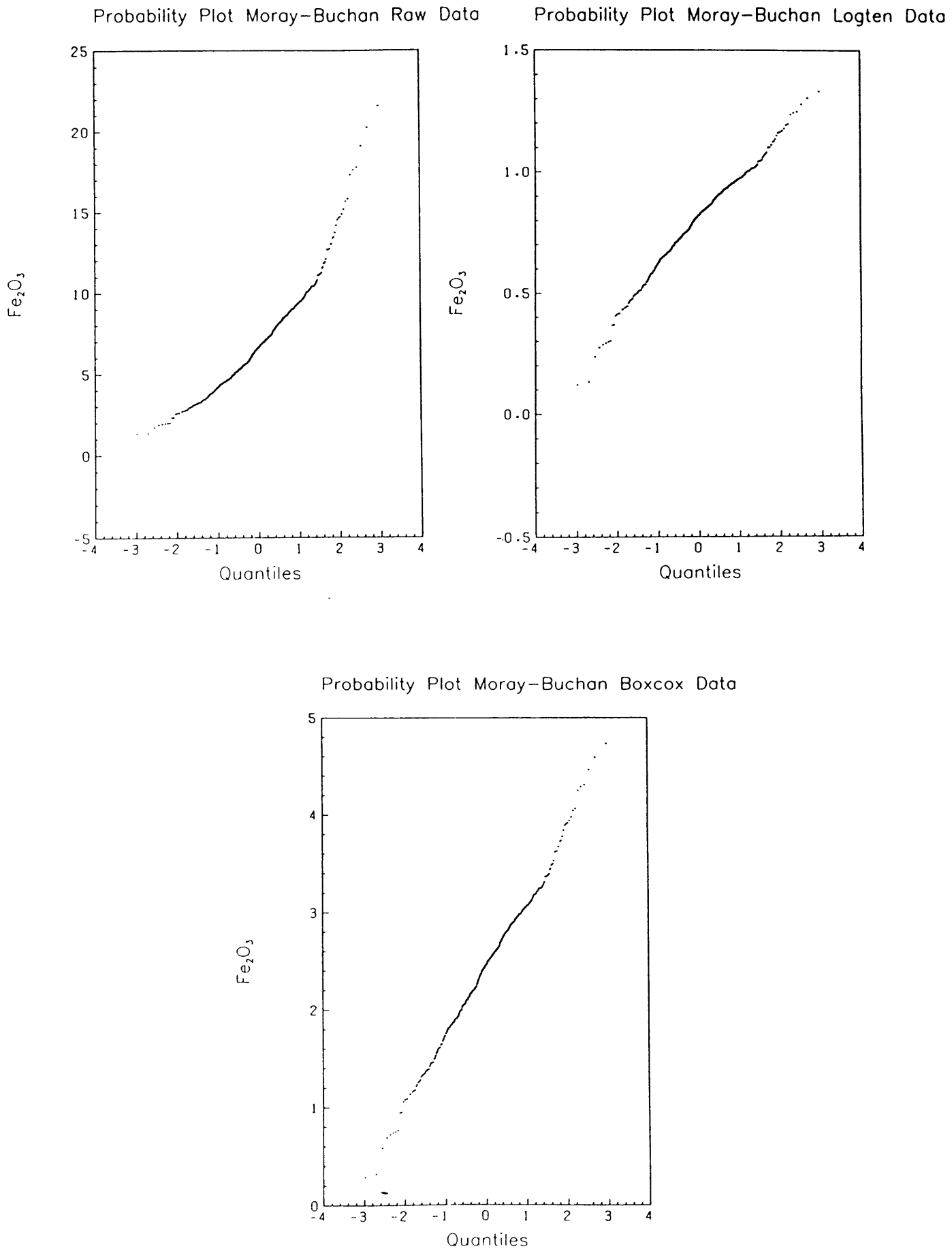
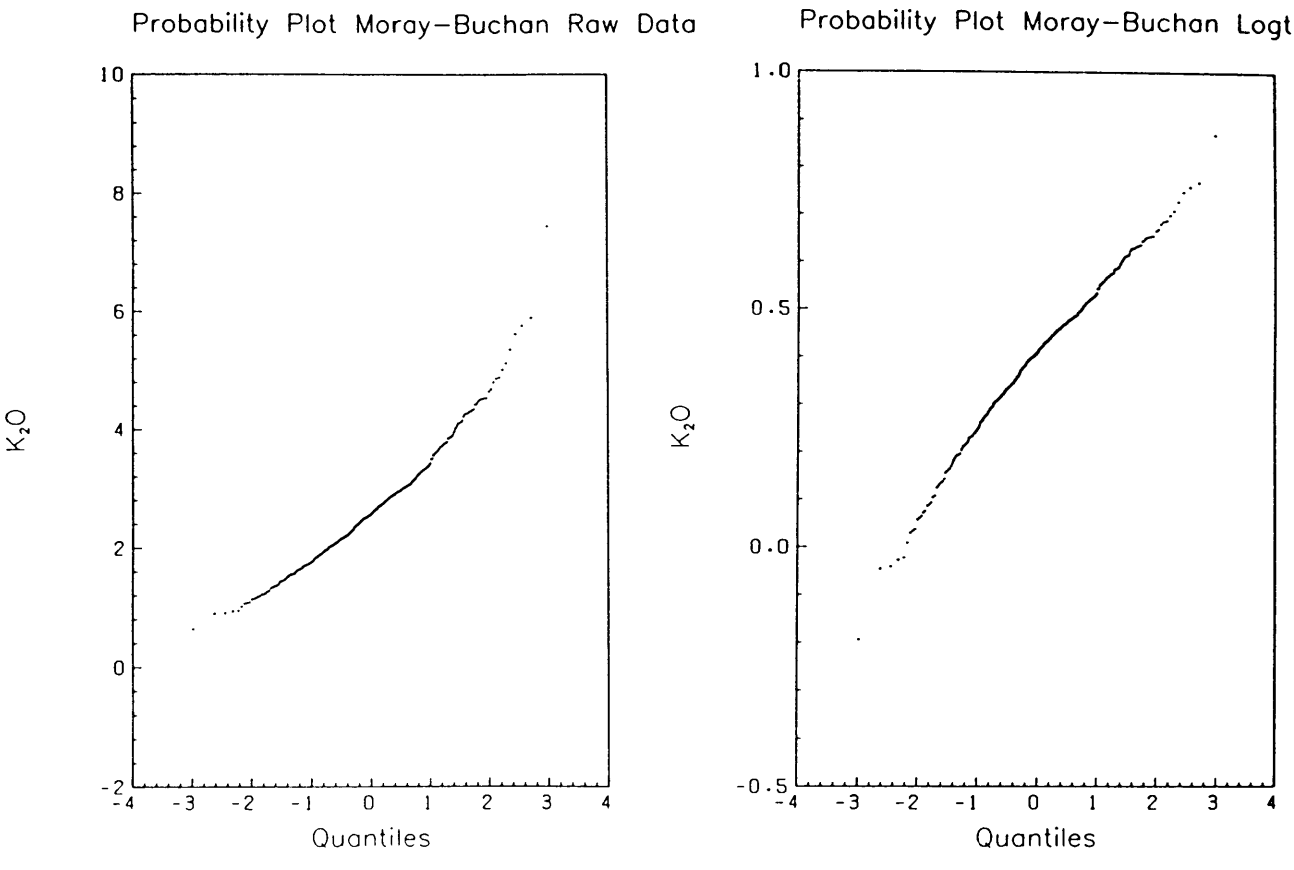

Probability Plot Moray-Buchan Raw Data
Fe2O3
Quantiles
Probability Plot Moray-Buchan Logten Data
Fe2O3
Quantiles
Probability Plot Moray-Buchan Boxcox Data
Fe2O3
Quantiles

Figure 5.19


Probability Plot Moray-Buchan Raw Data
Probability Plot Moray-Buchan Logten Data
K2O
Quantiles
10
8
6
4
2
0
-2
1.0
0.5
0.0
-0.5
-4
-3
-2
-1
0
1
2
3
4

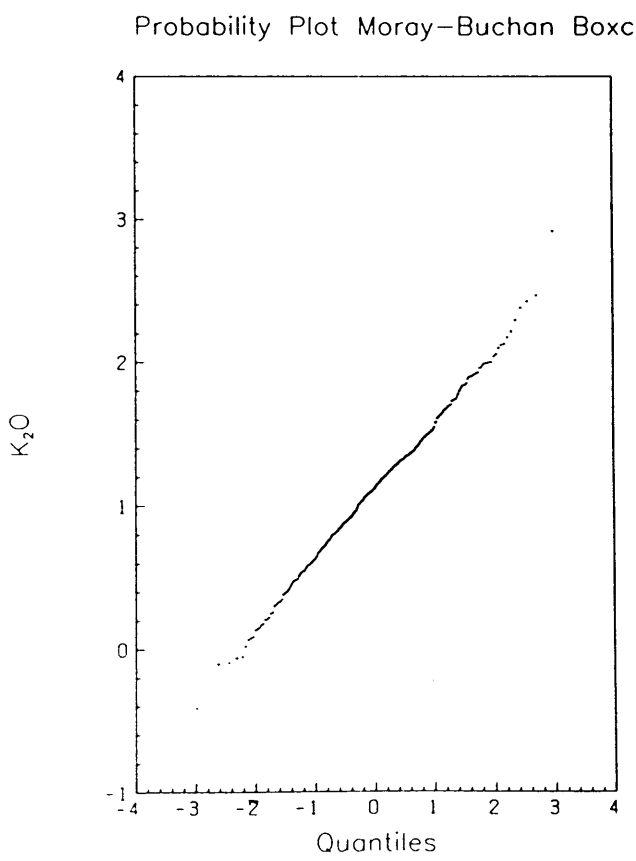

Probability Plot Moray-Buchan Boxcox Data
K2O
Quantiles
4
3
2
1
0
-1
-4
-3
-2
-1
0
1
2
3
4

Figure 5.20

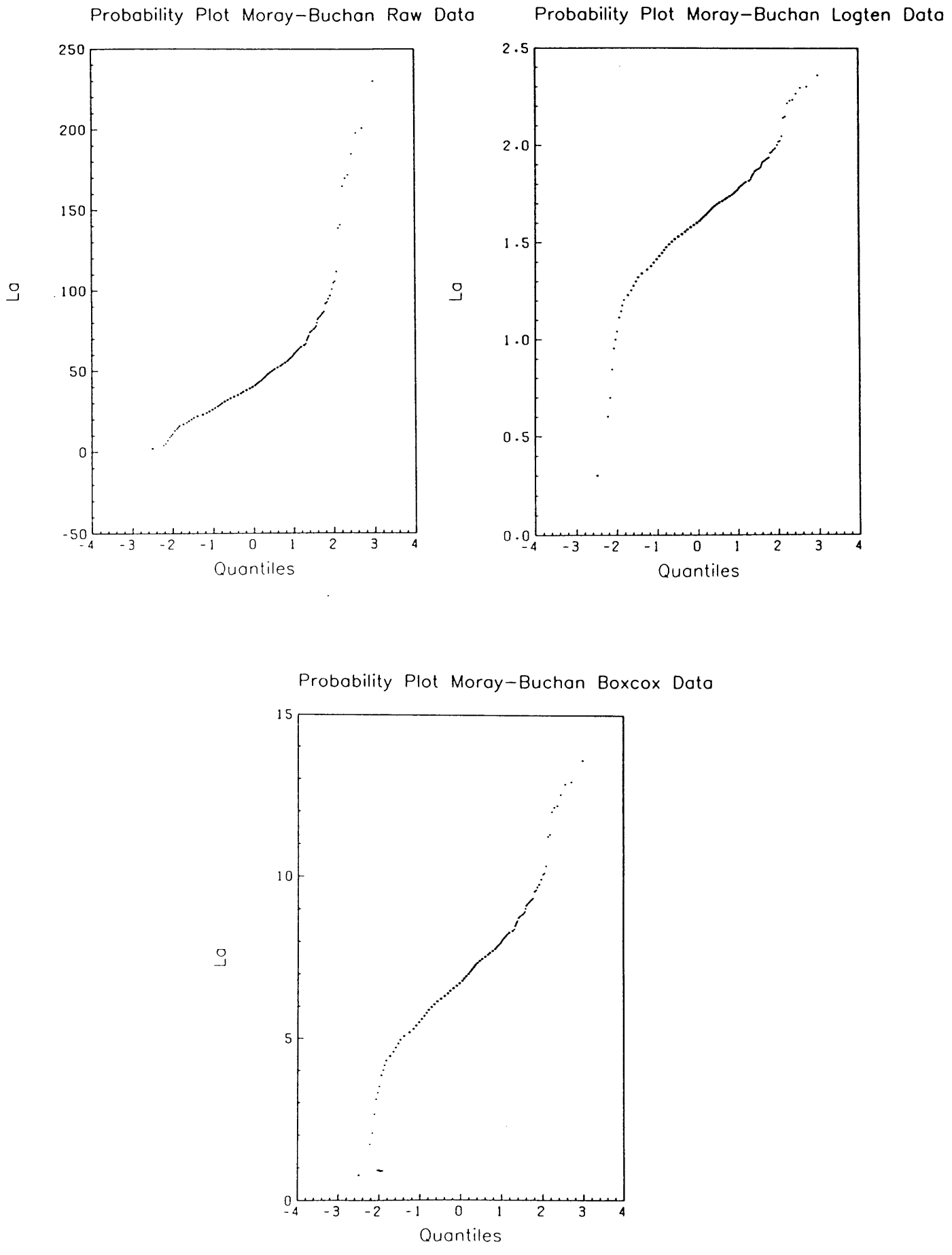
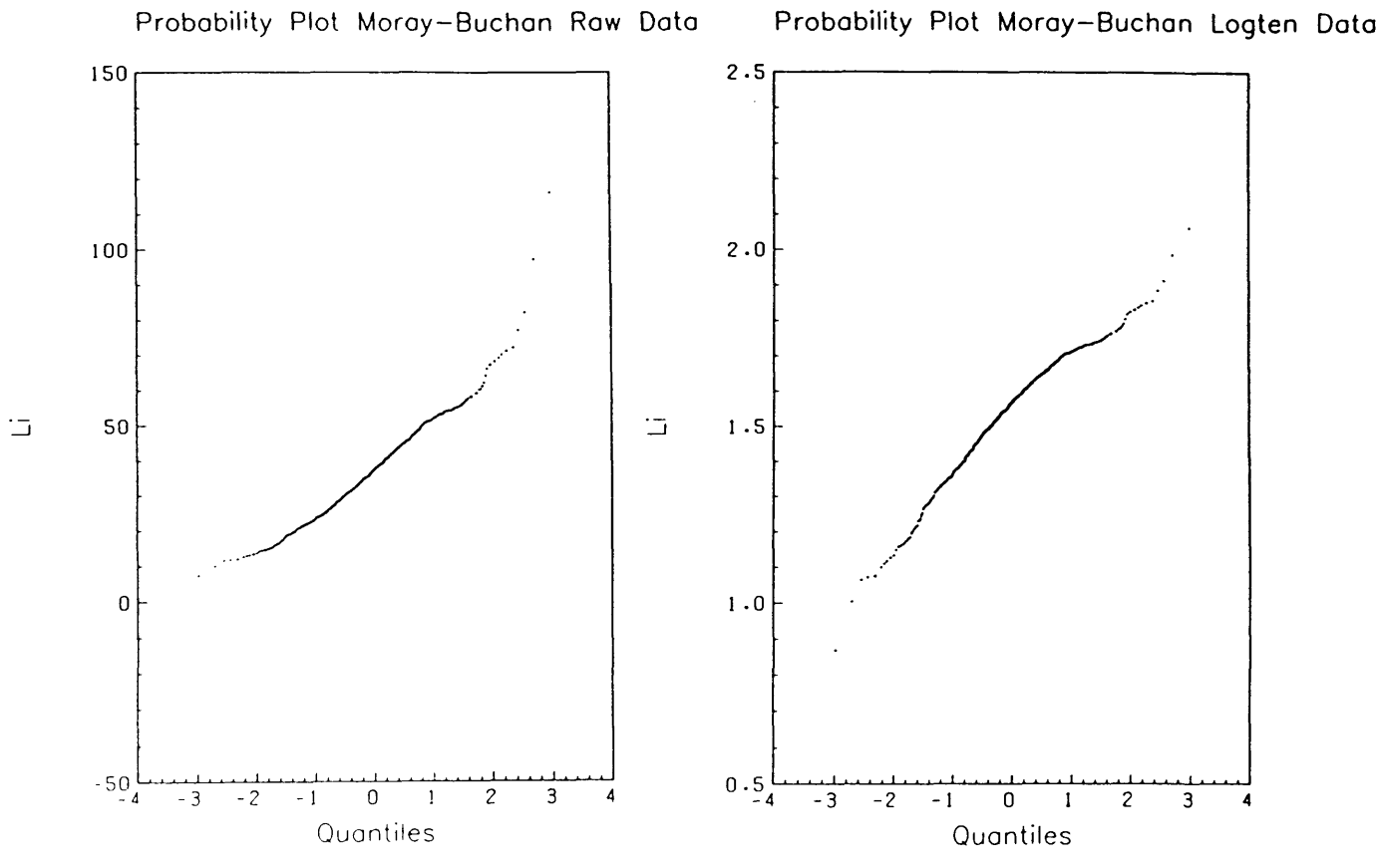

Probability Plot Moray-Buchan Raw Data
La
250
200
150
100
50
0
-50
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
La
2.5
2.0
1.5
1.0
0.5
0.0
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Boxcox Data
La
15
10
5
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.21


Probability Plot Moray-Buchan Raw Data
Li
150
100
50
0
-50
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Li
2.5
2.0
1.5
1.0
0.5
-4 -3 -2 -1 0 1 2 3 4
Quantiles

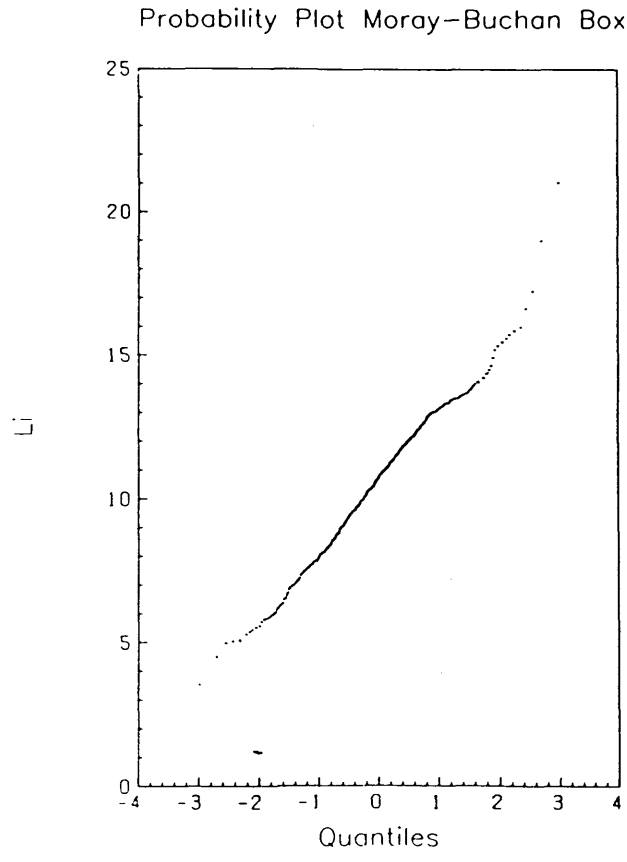
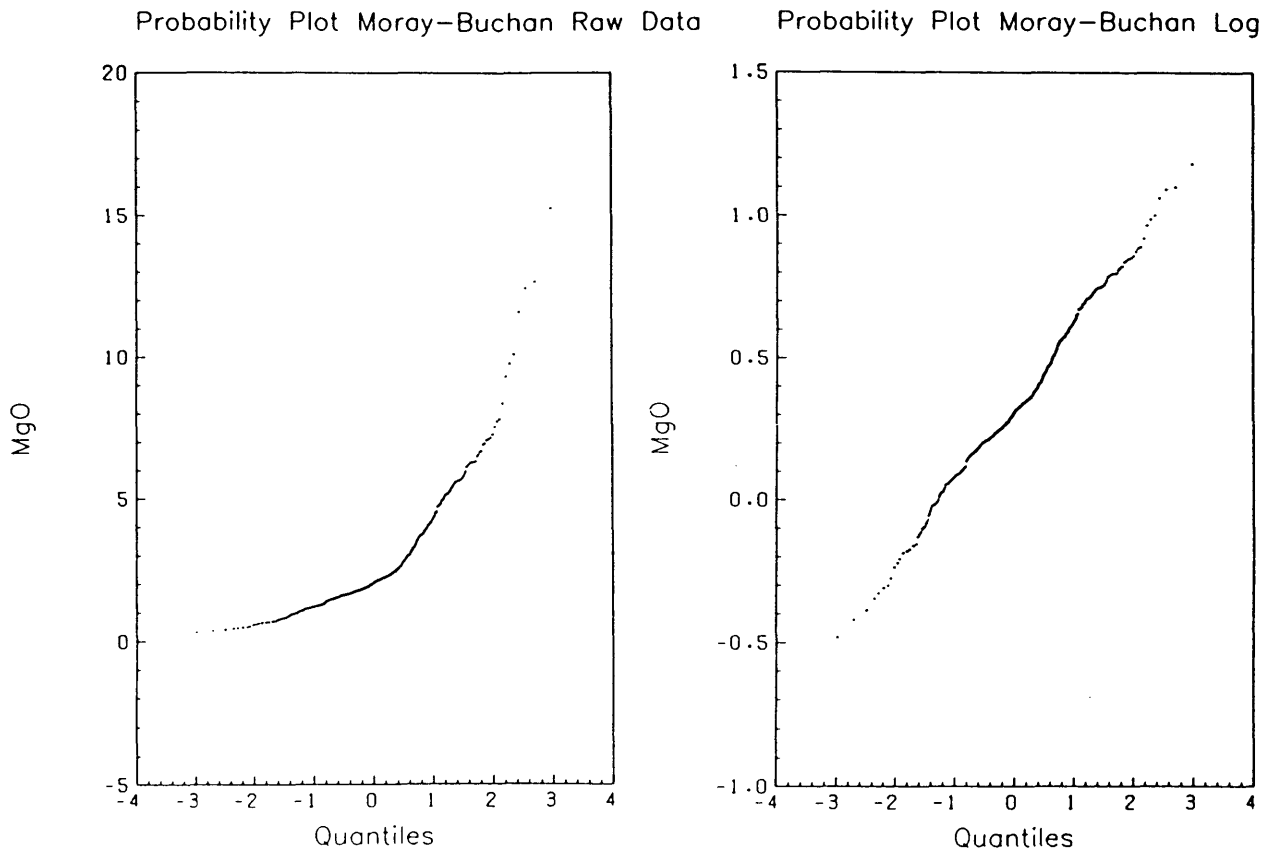

Probability Plot Moray-Buchan Boxcox Data
Li
25
20
15
10
5
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.22


Probability Plot Moray-Buchan Raw Data
MgO
20
15
10
5
0
-5
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
MgO
1.5
1.0
0.5
0.0
-0.5
-1.0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

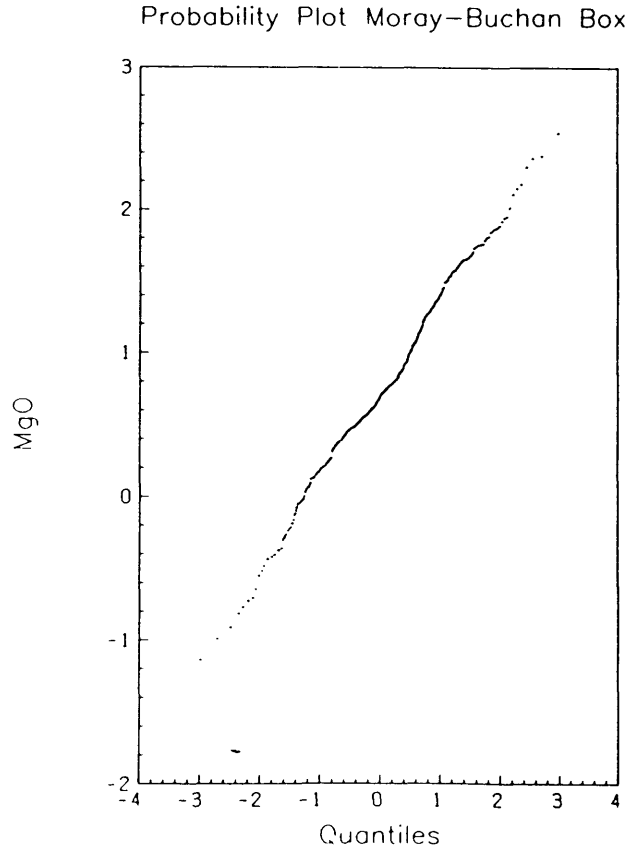
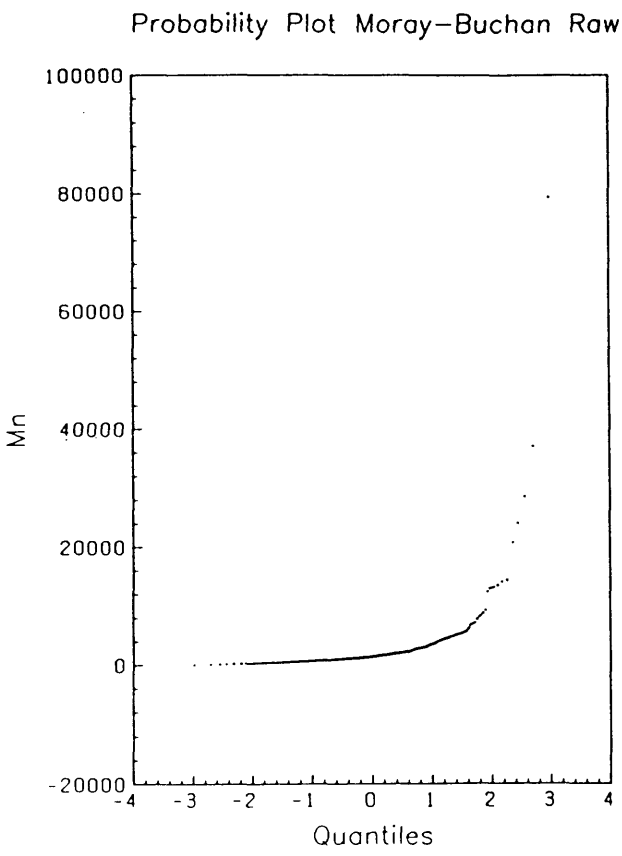
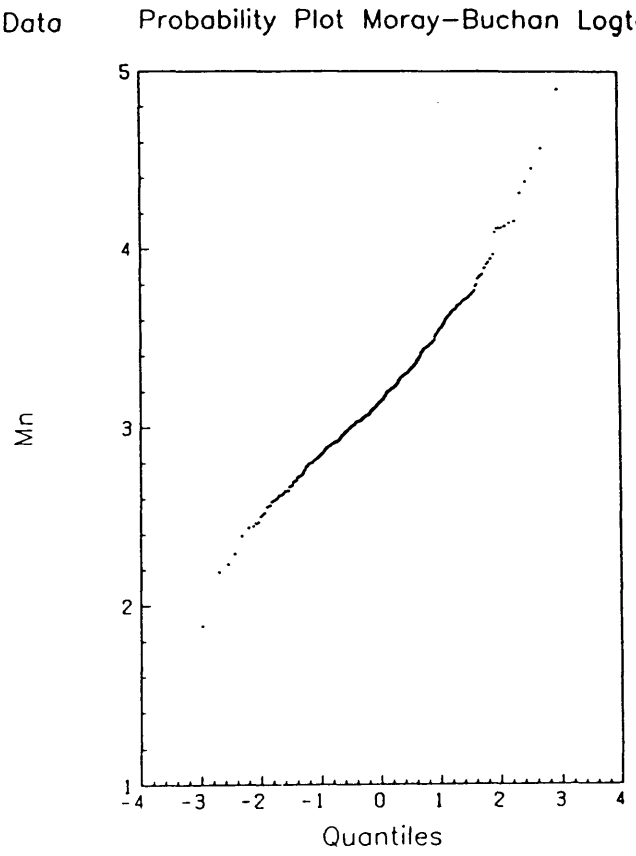

Probability Plot Moray-Buchan Boxcox Data
MgO
3
2
1
0
-1
-2
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.23


Probability Plot Moray–Buchan Raw Data
Mn
Quantiles
100000
80000
60000
40000
20000
0
-20000
-4 -3 -2 -1 0 1 2 3 4


Probability Plot Moray–Buchan Logten Data
Mn
Quantiles
5
4
3
2
1
-4 -3 -2 -1 0 1 2 3 4

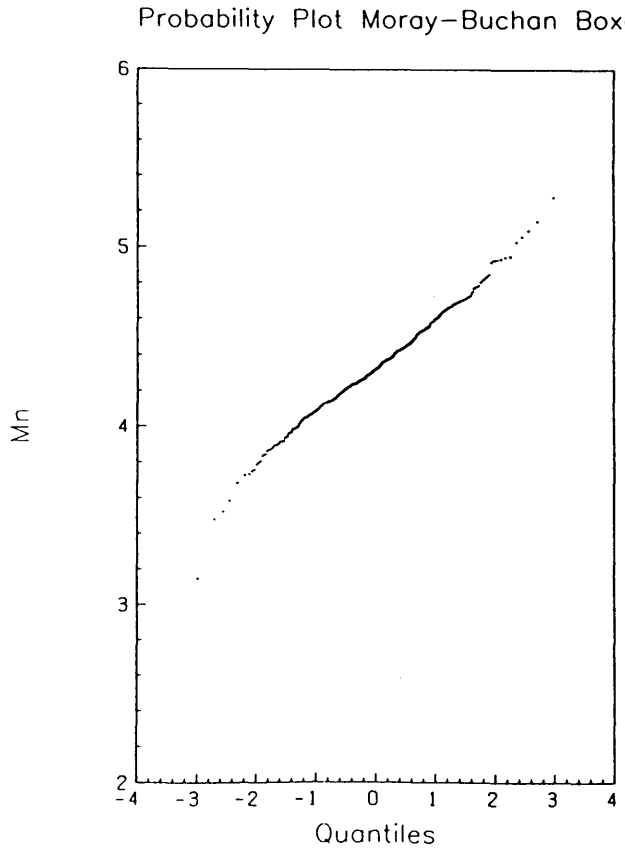

Probability Plot Moray–Buchan Boxcox Data
Mn
Quantiles
6
5
4
3
2
-4 -3 -2 -1 0 1 2 3 4

Figure 5.24

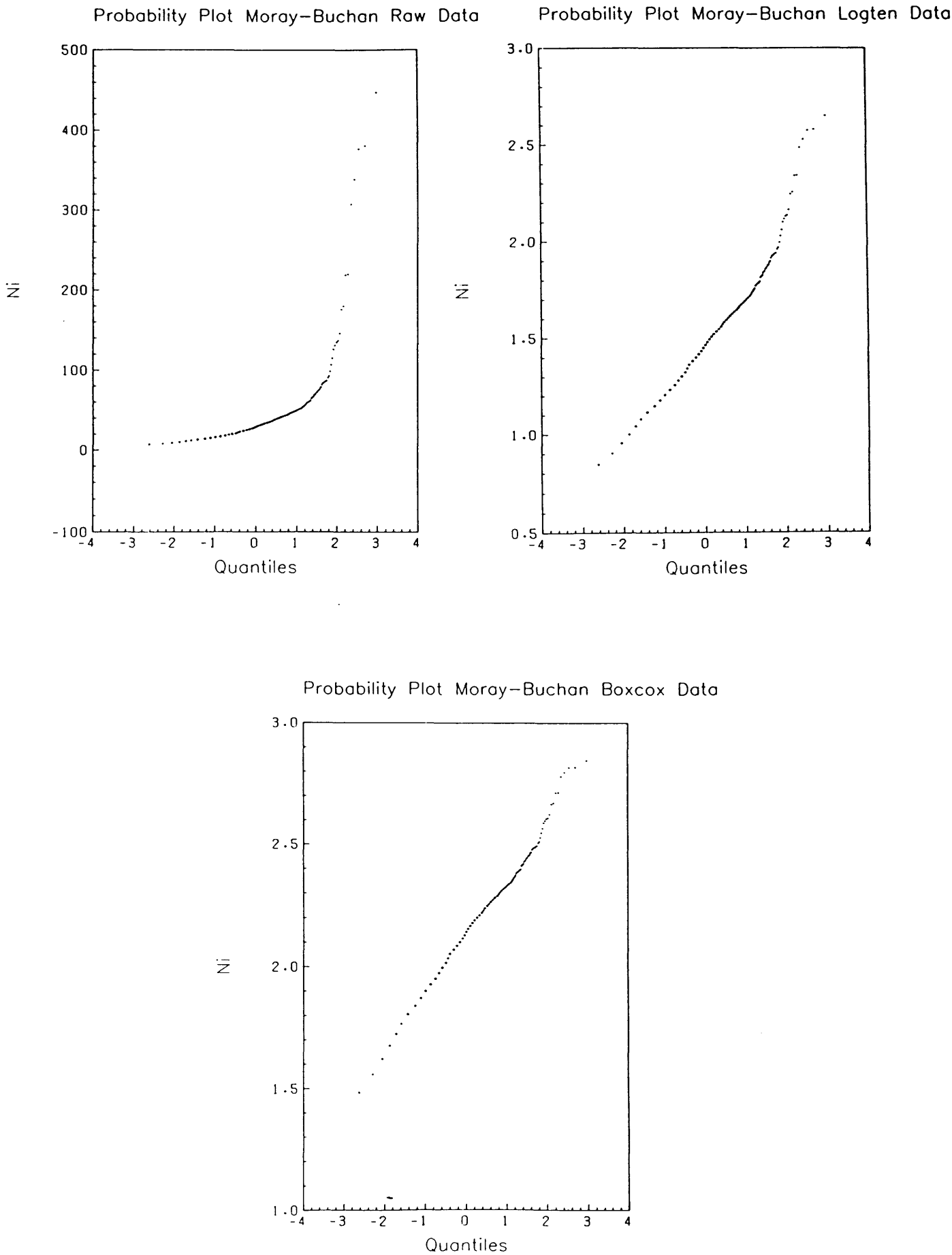

Probability Plot Moray-Buchan Raw Data
Probability Plot Moray-Buchan Logten Data
Probability Plot Moray-Buchan Boxcox Data
Ni
Quantiles
500
400
300
200
100
0
-100
3.0
2.5
2.0
1.5
1.0
0.5
-4
-3
-2
-1
0
1
2
3
4

Figure 5.25

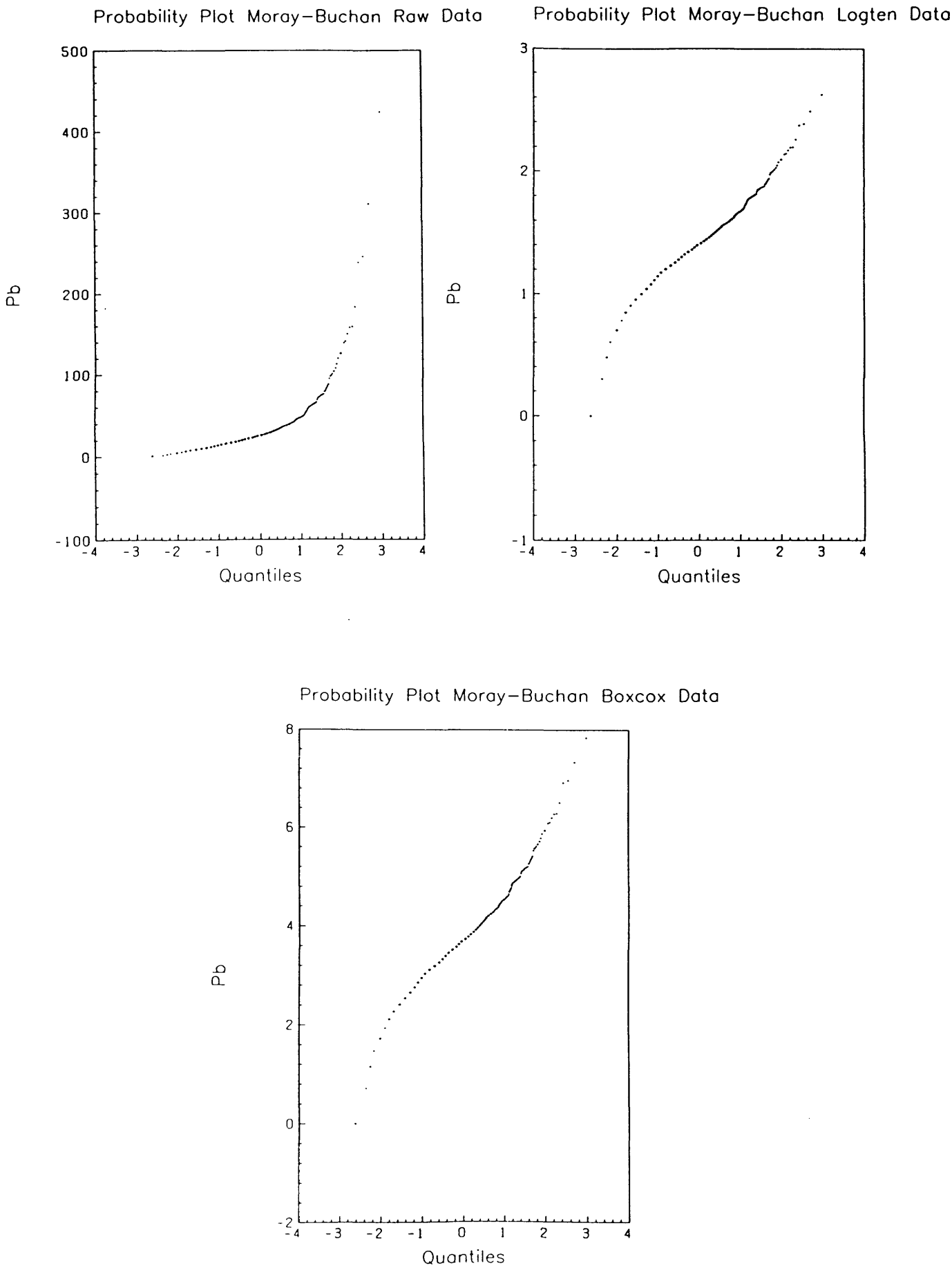
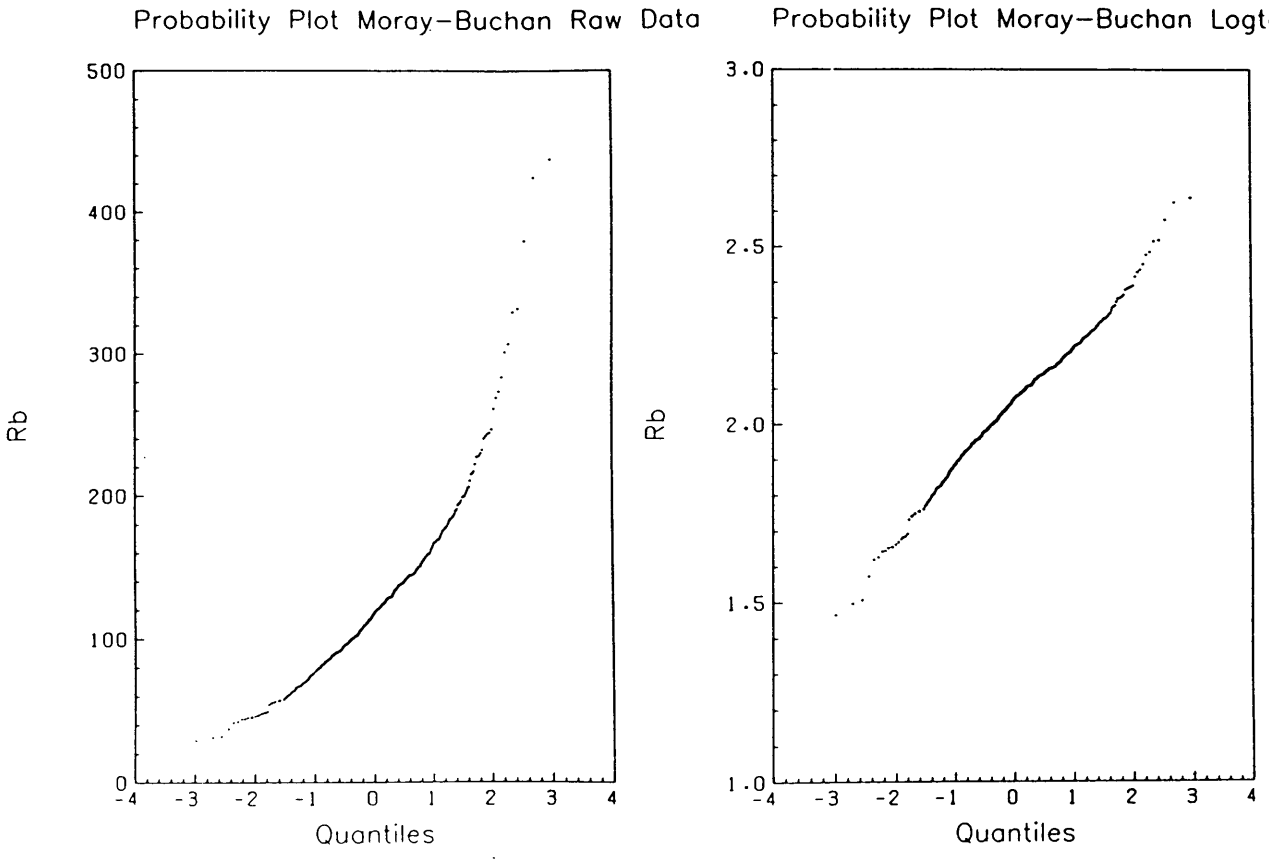

Probability Plot Moray–Buchan Raw Data
Pb
Quantiles
Probability Plot Moray–Buchan Logten Data
Pb
Quantiles
Probability Plot Moray–Buchan Boxcox Data
Pb
Quantiles

Figure 5.26


Probability Plot Moray-Buchan Raw Data
Rb
500
400
300
200
100
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Rb
3.0
2.5
2.0
1.5
1.0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

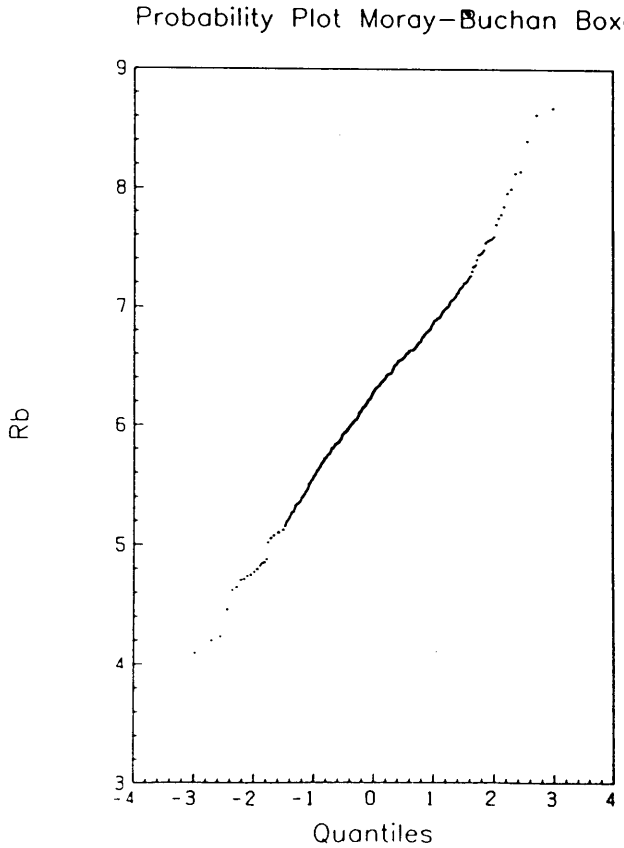

Probability Plot Moray-Buchan Boxcox Data
Rb
9
8
7
6
5
4
3
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.27

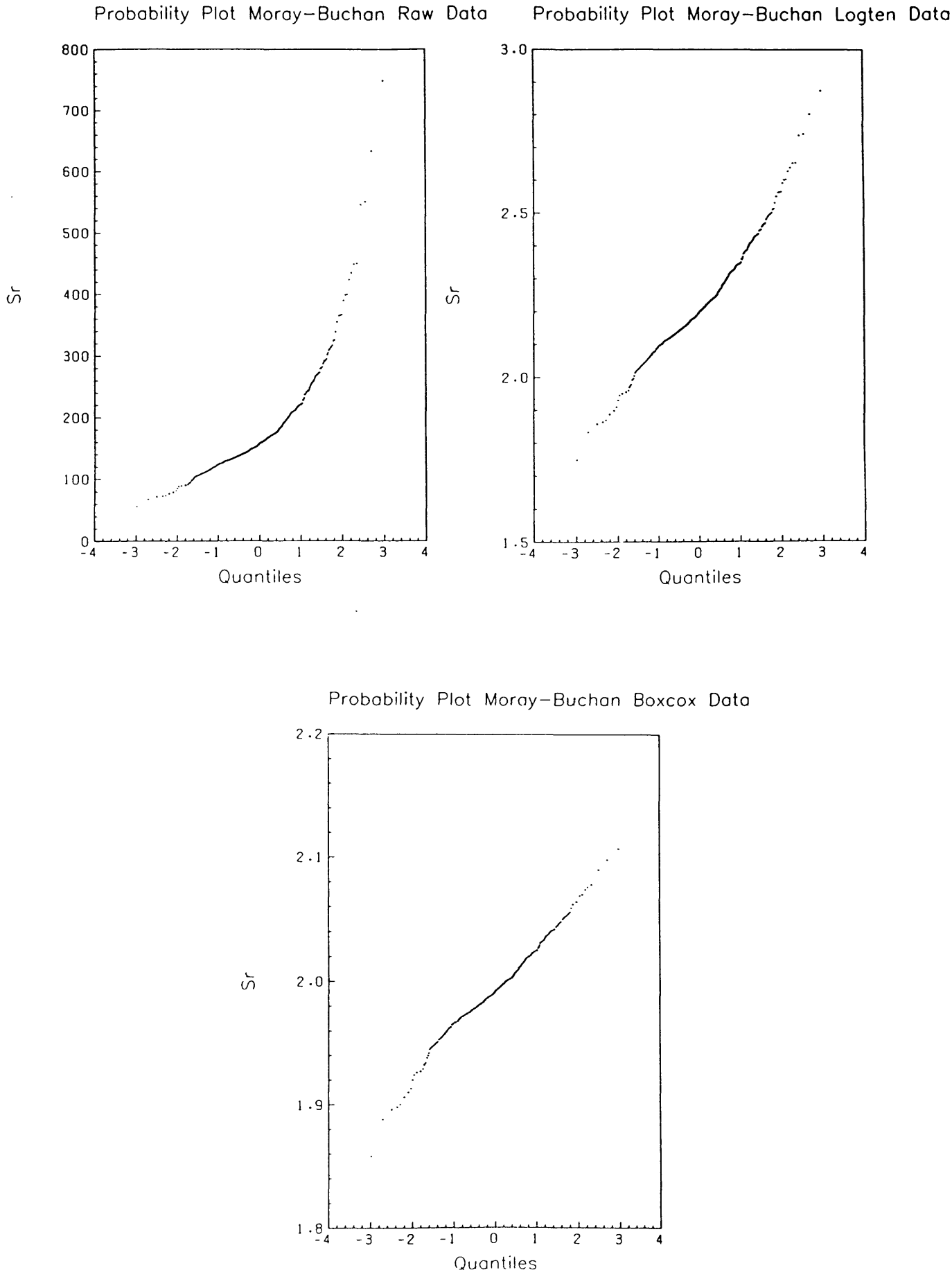
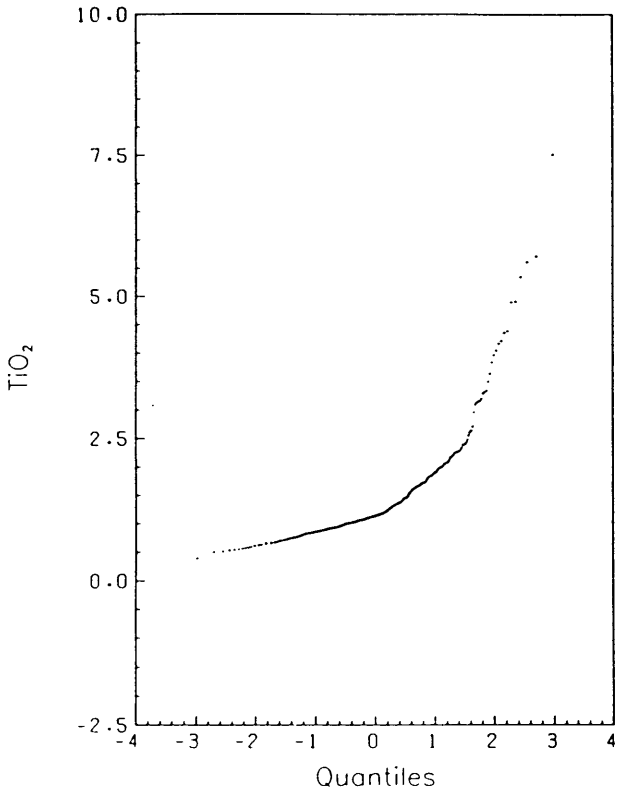
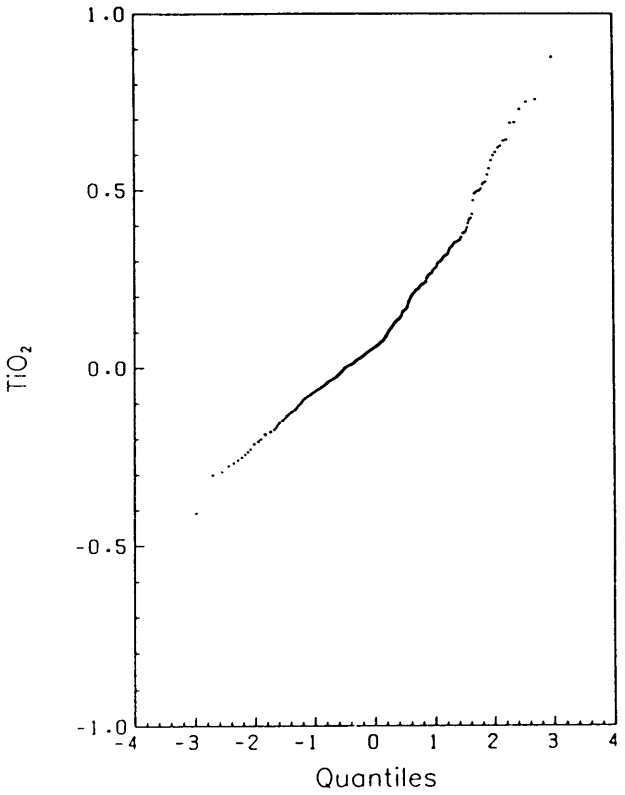

Probability Plot Moray-Buchan Raw Data
Sr
800
700
600
500
400
300
200
100
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Sr
3.0
2.5
2.0
1.5
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Boxcox Data
Sr
2.2
2.1
2.0
1.9
1.8
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.28

Probability Plot Moray–Buchan Raw Data



Probability Plot Moray–Buchan Logten Data



Probability Plot Moray--Buchan Boxcox Data

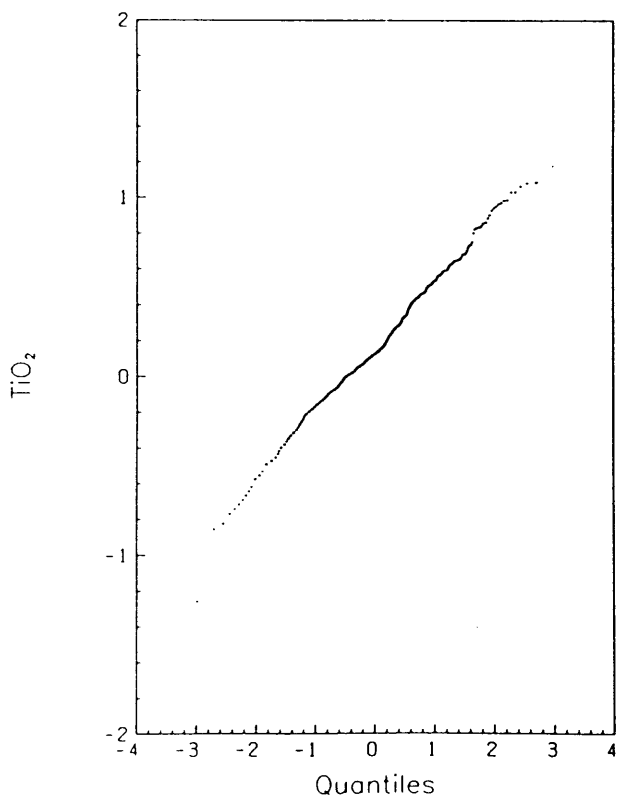
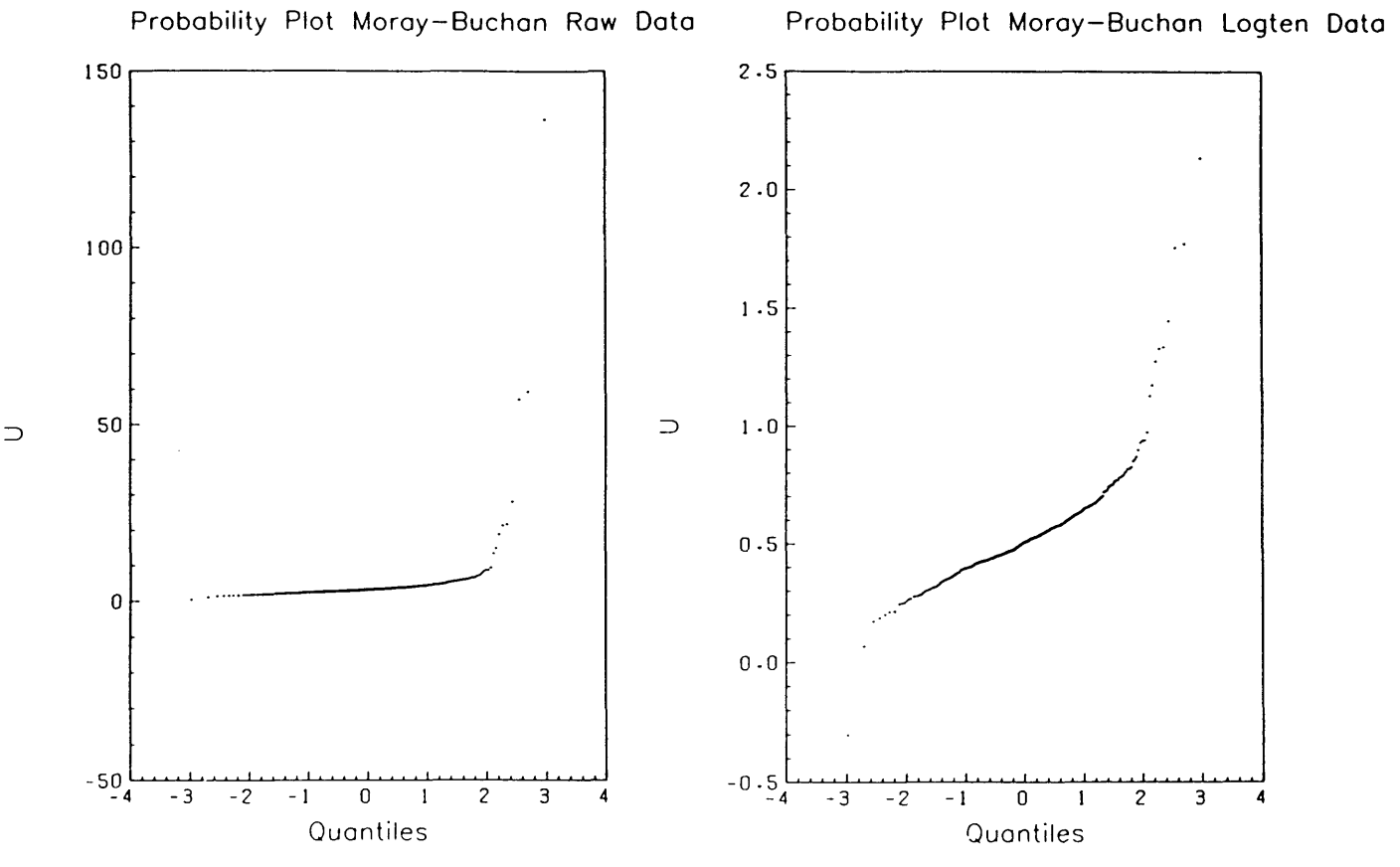


Figure 5.29


Probability Plot Moray-Buchan Raw Data
Probability Plot Moray-Buchan Logten Data
150
100
50
0
-50
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Quantiles

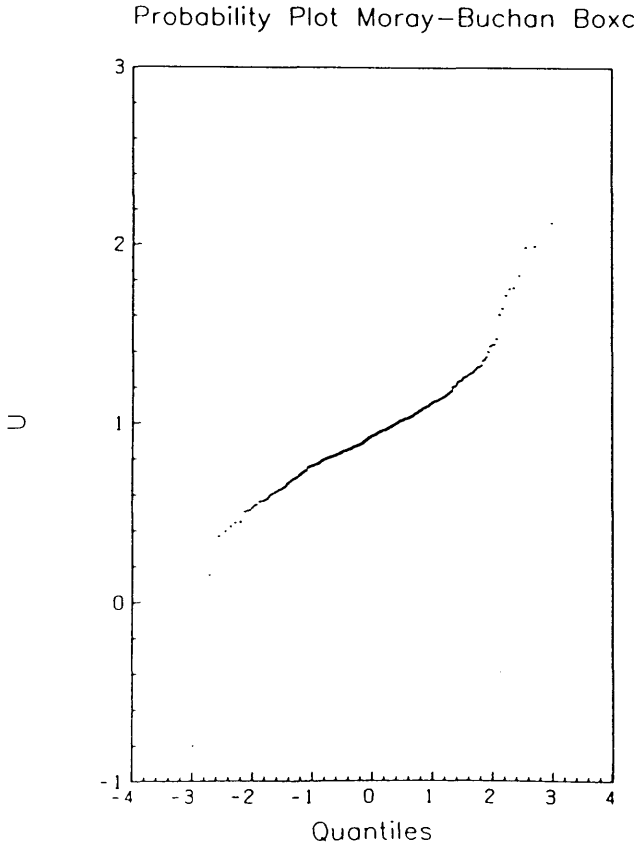

Probability Plot Moray-Buchan Boxcox Data
3
2
1
0
-1
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.30

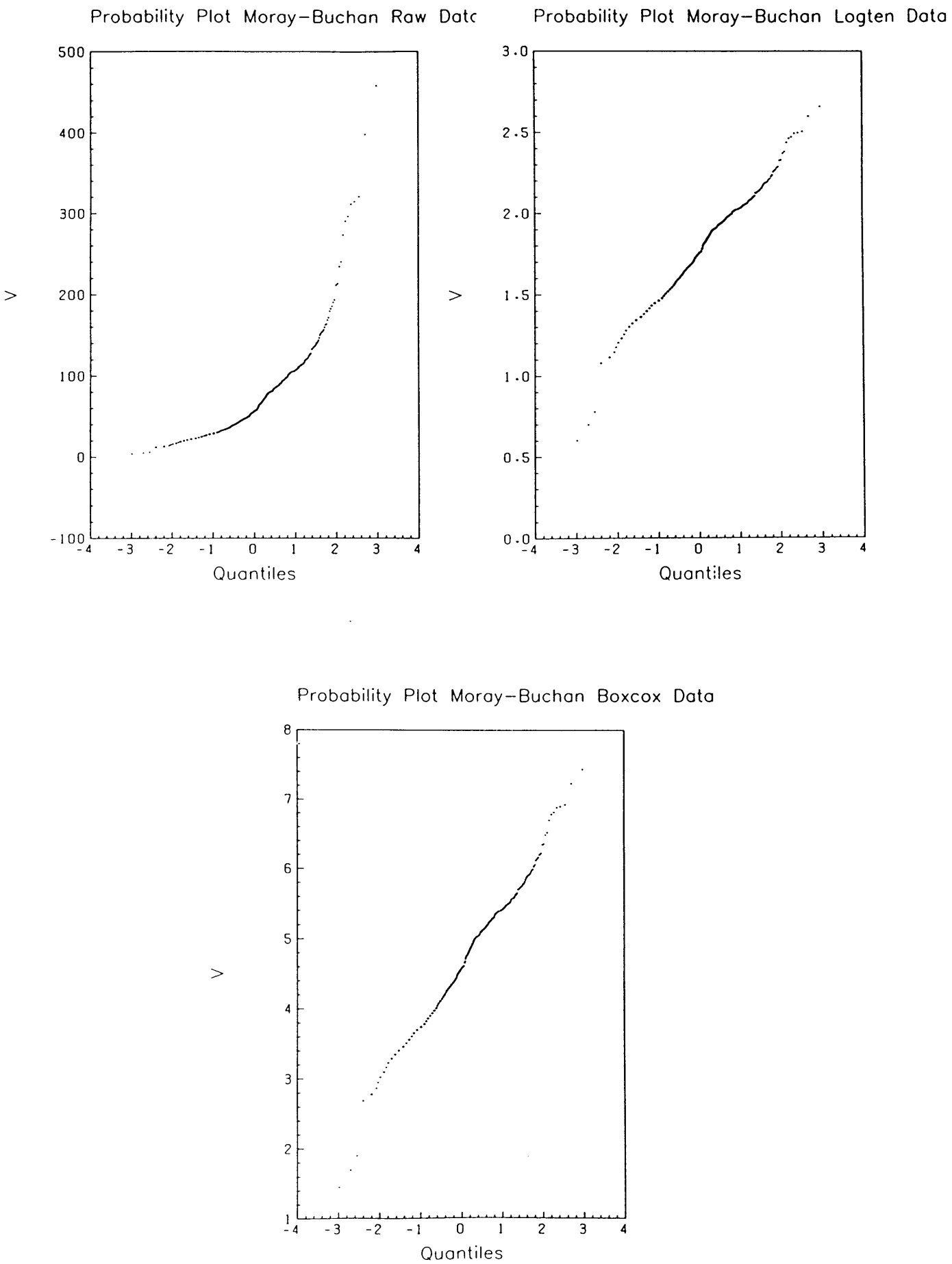
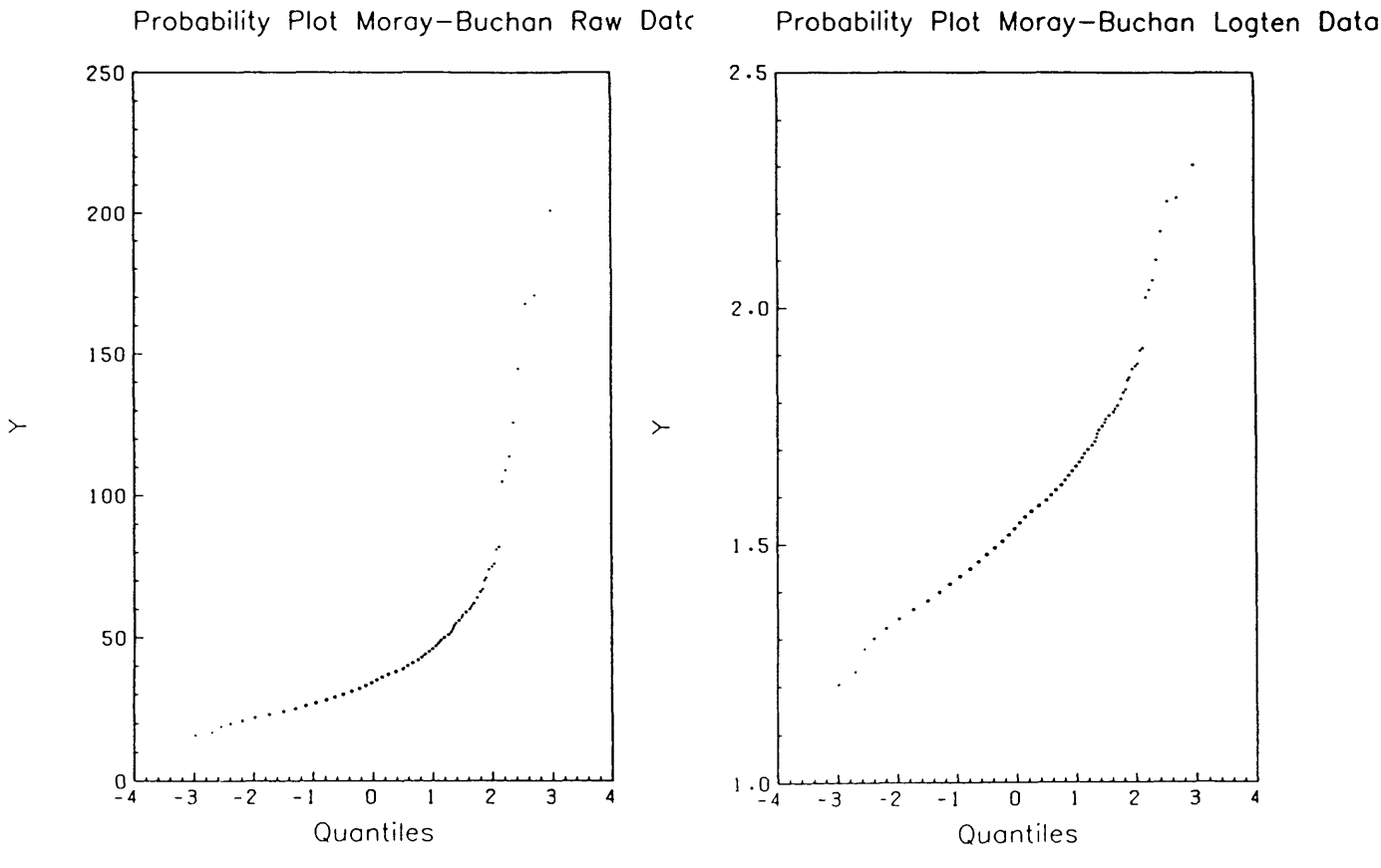

Probability Plot Moray-Buchan Raw Data
Probability Plot Moray-Buchan Logten Data
Probability Plot Moray-Buchan Boxcox Data
Y
Quantiles
500
400
300
200
100
0
-100
3.0
2.5
2.0
1.5
1.0
0.5
0.0
8
7
6
5
4
3
2
1
-4
-3
-2
-1
0
1
2
3
4

Figure 5.31


Probability Plot Moray-Buchan Raw Data
Y
250
200
150
100
50
0
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Y
2.5
2.0
1.5
1.0
-4 -3 -2 -1 0 1 2 3 4
Quantiles

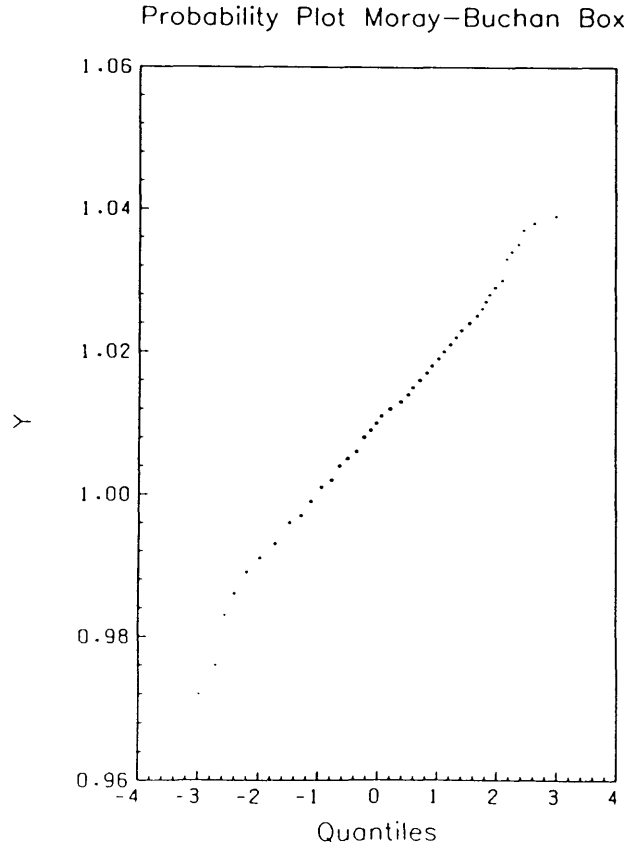

Probability Plot Moray-Buchan Boxcox Data
Y
1.06
1.04
1.02
1.00
0.98
0.96
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.32

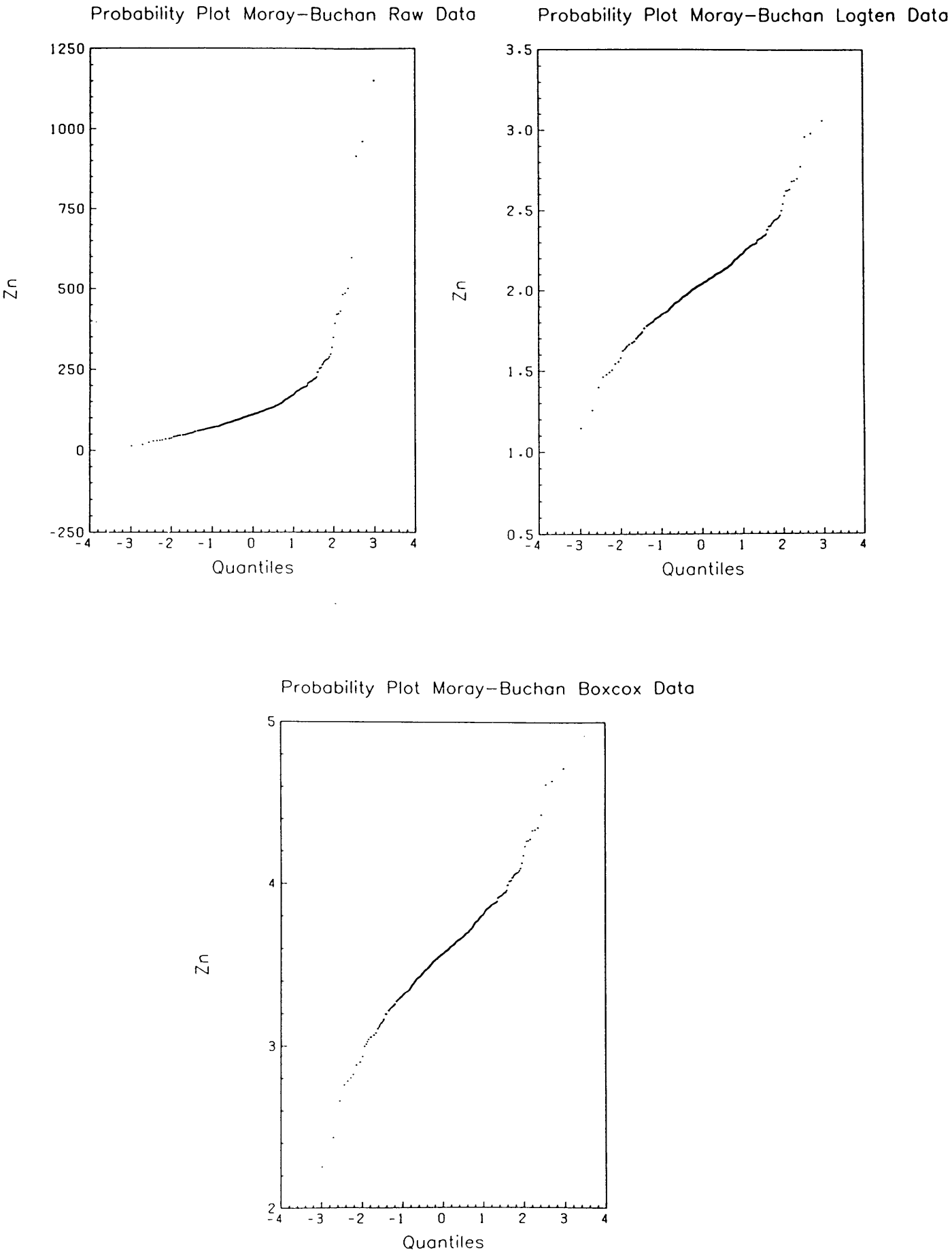
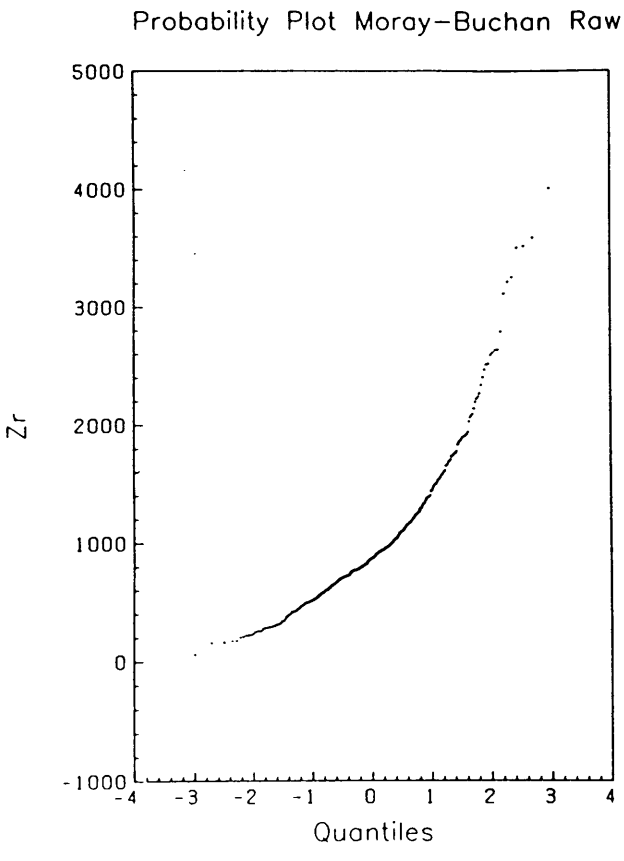
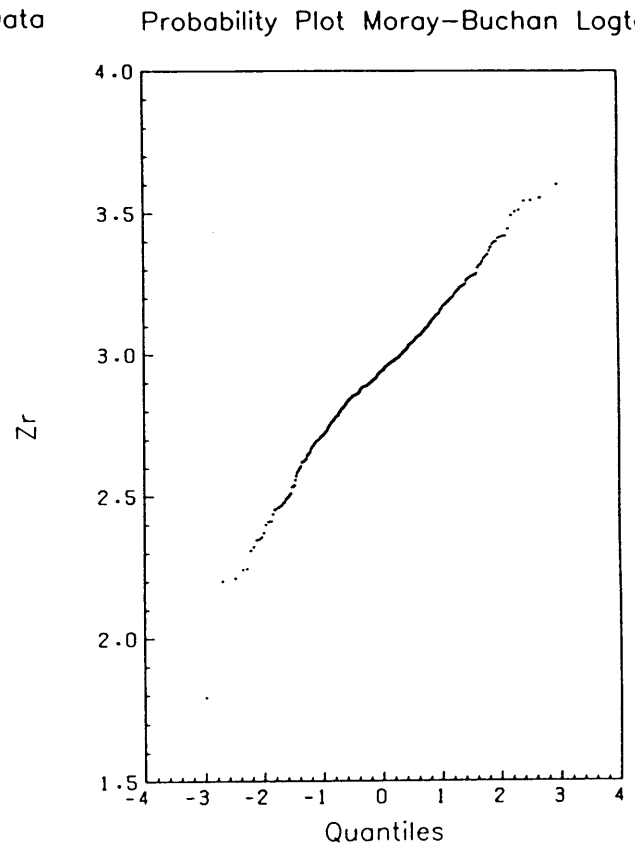
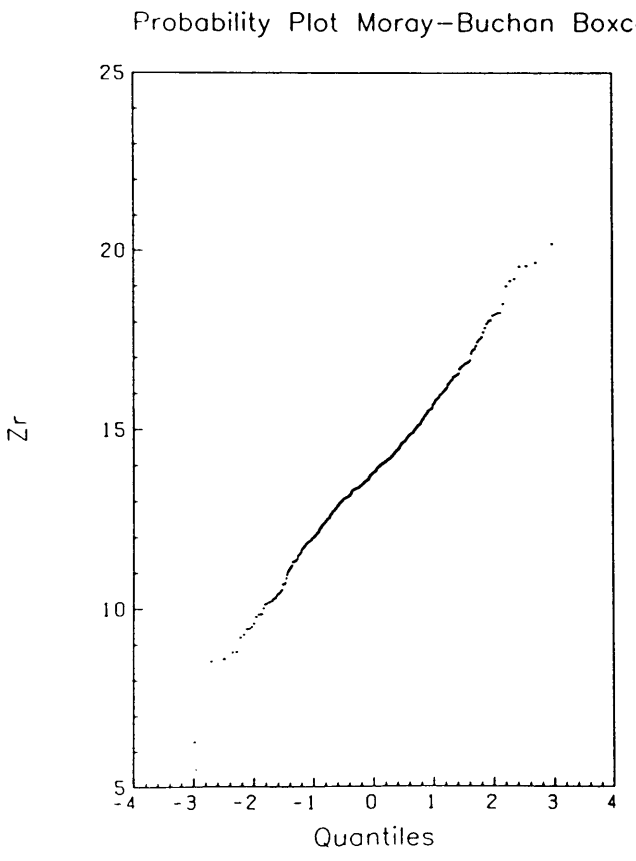

Probability Plot Moray-Buchan Raw Data
Zn
1250
1000
750
500
250
0
-250
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Logten Data
Zn
3.5
3.0
2.5
2.0
1.5
1.0
0.5
-4 -3 -2 -1 0 1 2 3 4
Quantiles
Probability Plot Moray-Buchan Boxcox Data
Zn
5
4
3
2
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 5.33


Probability Plot Moray-Buchan Raw Data
Zr
5000
4000
3000
2000
1000
0
-1000
-4 -3 -2 -1 0 1 2 3 4
Quantiles


Probability Plot Moray-Buchan Logten Data
Zr
4.0
3.5
3.0
2.5
2.0
1.5
-4 -3 -2 -1 0 1 2 3 4
Quantiles


Probability Plot Moray-Buchan Boxcox Data
Zr
25
20
15
10
5
-4 -3 -2 -1 0 1 2 3 4
Quantiles

heavy-tailedness (either symmetric or asymmetric).

We can quantify the analysis of univariate Q-Q plots by performing the chi-squared test on each variable for all three data sets (raw, logten and boxcox). The test results are given in Table 5.5. The following conclusions can be drawn from these figures.

1) Of the raw data distributions only Li conforms to normality at the specified level of significance. Of the remainder the frequency distributions of K and Al come closest to normality, while U and Mn clearly have the most non-normal distributions.

2) Logten transformed Al, Co, Cu, Fe, Pb, Rb, V, Zn and Zr data are normally distributed according to the chi-squared test. There appears to be some discrepancy between the form of the Q-Q plot and the chi-squared statistic for some variables. For instance, the variables Mn and Mg have logten Q-Q plots as straight as, say, Zn or Zr, but are judged by the chi-squared statistic not to be lognormally distributed.

The chi-squared test involves comparing the sample distribution to the standardized normal distribution N(0,1). Therefore, each variable must be <u>standardized</u> (by subtracting the mean form each observation and dividing by the standard deviation) prior to calculation of the test statistic. We have seen from Chapter 3 that

Table 5.5

Chi–Squared Test Moray–Buchan Data

5 D.F. Critical value = 15.09 at 1% level

	Raw	Logten	Boxcox
Al	18.55	14.88	4.58
B	238.96	21.16	40.51
Ba	153.46	77.12	46.07
Be	201.45	144.79	58.02
Ca	256.65	24.93	21.92
Co	270.05	12.71	2.77
Cr	310.73	52.35	67.46
Cu	83.08	9.43	5.05
Fe	31.29	14.24	8.47
K	19.61	16.98	2.05
La	116.66	49.52	30.89
Li	13.49	30.45	3.40
Mg	208.21	21.29	19.09
Mn	399.66	16.16	8.19
Ni	291.95	16.05	12.34
Pb	313.64	14.73	20.73
Rb	36.43	13.79	11.11
Sr	166.49	30.26	15.40
Ti	257.49	60.62	20.03
U	467.28	75.70	32.60
V	142.19	3.46	2.86
Y	208.02	28.36	6.79
Zn	240.42	15.09	9.92
Zr	97.30	11.41	8.09

the mean and standard deviation are highly sensitive to outliers and so it is possible that the presence of one anomalously high observation could increase both parameters, with the result that the sample distribution is offset in comparison with the theoretical distribution and is revealed as non-normal.

Note that on logten transformation Li becomes non-normally distributed. This demonstrates again the insensitivity of the logten transformation to the form of the original data distribution.

3) Boxcox transformation shows a general improvement over logten transformation with K, Li, Mn, Ni and Y being added to the list of normally distributed variables. However, Pb, accepted as lognormally distributed, is now considered to be non-normal. This is probably due to a slight increase in the mean through the influence of one or more anomalously high values, with the result that there are more observations with standardized values less than zero (mean) than there are observations greater than zero. The chi-squared statistic is thus increased. The variables B and Cr also show an increase in chi-squared from logten to boxcox transformation.

Having established that the boxcox tranformation gives significant advantages over the traditional logten

transformation in attaining distributional normality, we can now compare robust with non-robust location and scale estimators for each of the three data sets. Classical estimators were used for the non-robust estimates and M-estimators were used for the robust estimates. Referring to Table 5.6 the following conclusions can be made:

1) Robust location and scale estimates of variables whose frequency distributions show positive skew tend to be lower than the equivalent non-robust estimates. Conversely variables whose frequency distributions show negative skew tend to have their robust location estimates higher than their non-robust equivalents. Robust scale estimates, however, invariably show a decrease compared with non-robust estimates, no matter what form the frequency distribution takes.

2) Non-robust raw location and scale estimates are greater than the equivalent boxcox estimates which in turn are greater than logten estimates. The differences between raw, logten and boxcox location estimates are relatively small with, in the majority of cases, the boxcox estimate being closer to the raw estimate than to the logten estimate. Differences between raw, logten and boxcox scale estimates are easier to explain than the equivalent location estimates. For those variables whose raw frequency distributions show very strong

Table 5.6

Non-Robust and Robust Lc

Location

Raw

Raw Robust

Logten

Logten Robust

Boxcox

Boxcox Robust

Scale

Raw

Raw Robust

Logten

Logten Robust

Boxcox

Boxcox Robust

positive skew (U, Mn, Ni, Pb and Cr) the differences between non-robust raw scale estimates and the equivalent logten and boxcox estimates are large while the differences between the logten and boxcox scale estimates are small. However, for those variables whose raw frequency distributions come relatively close to normality (Li, K, Al, Fe, Rb and Cu) the differences between the raw, logten and boxcox scale estimates tend to be small with the boxcox estimates closer to the raw than to the logten equivalents.

Differences between the robust location and scale estimates for the three data sets are less clear cut than for the non-robust estimates. Robust logten location and scale estimates are clearly the lowest, but the robust raw estimates have decreased by a greater proportion (compared to their non-robust equivalents) than the robust boxcox estimates, with the result that raw estimates are no longer consistently greater than the boxcox.

A general conclusion from this analysis is that the more non-normal the frequency distribution of a variable the greater will be the relative change in its location and scale estimates when either that variable is transformed from its raw state, or robust estimates are used instead of the non-robust (classical) equivalents.

Multivariate ($D_m^{2/3}$) Q-Q plots may be used to summarize

the effects of transformations and the use of robust estimators. Figures 5.34 through 5.35 show the non-robust and robust $D_m^{2/3}$ Q-Q plots for raw, logten and boxcox transformed Moray-Buchan data. The following conclusions can be made from their analysis:

1) The logten and boxcox data $D_m^{2/3}$ Q-Q plots (see Figures 5.34 and 5.35) are very similar, both showing positive skewness, demonstrating that both transforms perform equally in terms of their effect on the multivariate distribution.

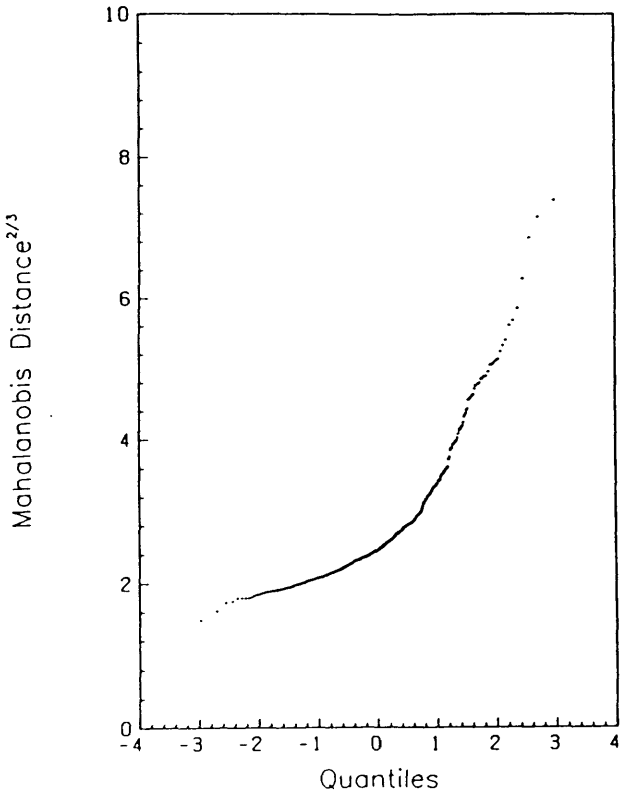
2) The use of robust parameters in the calculation of $D_m^{2/3}$ values has the effect of making the logten and boxcox Q-Q plots slightly more positively skewed - the effect is more noticable in the case of logten data.

3) The non-robust $D_m^{2/3}$ Q-Q plot for raw data shows a strong positive skew (see Figure 5.34) which is emphasized when robust parameters are used to calculate $D_m^{2/3}$ values (see Figure 5.35). See Chapter 4 for a full discussion of these two plots.

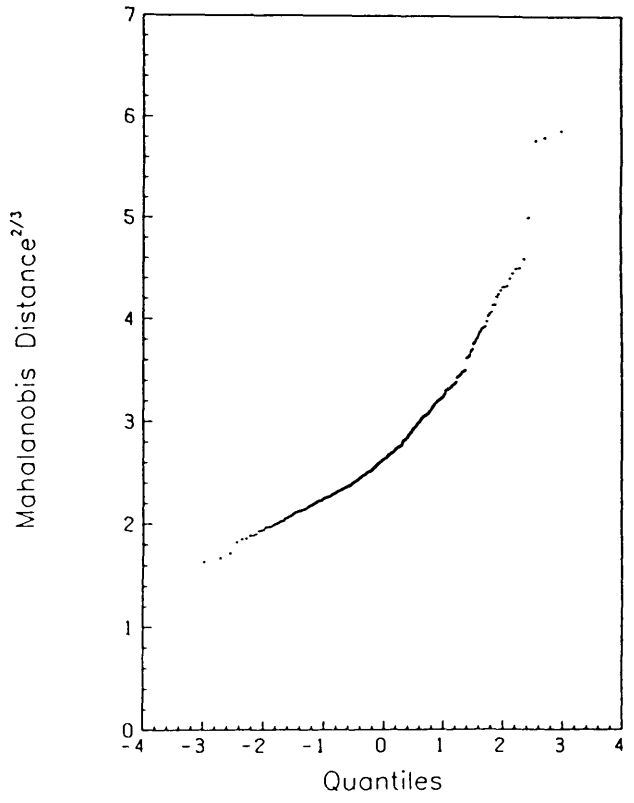
4) Compared with raw data both logten and boxcox transforms achieve a distinct improvement towards multivariate distributional normality.

Probability Plot $Dm^{2/3}$ Moray–Buchan Raw Data

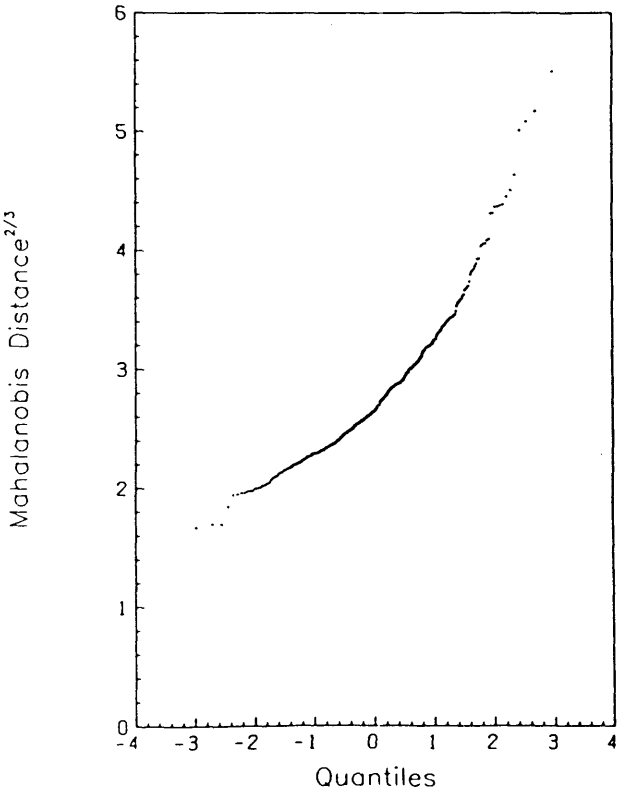
Figure 5.34



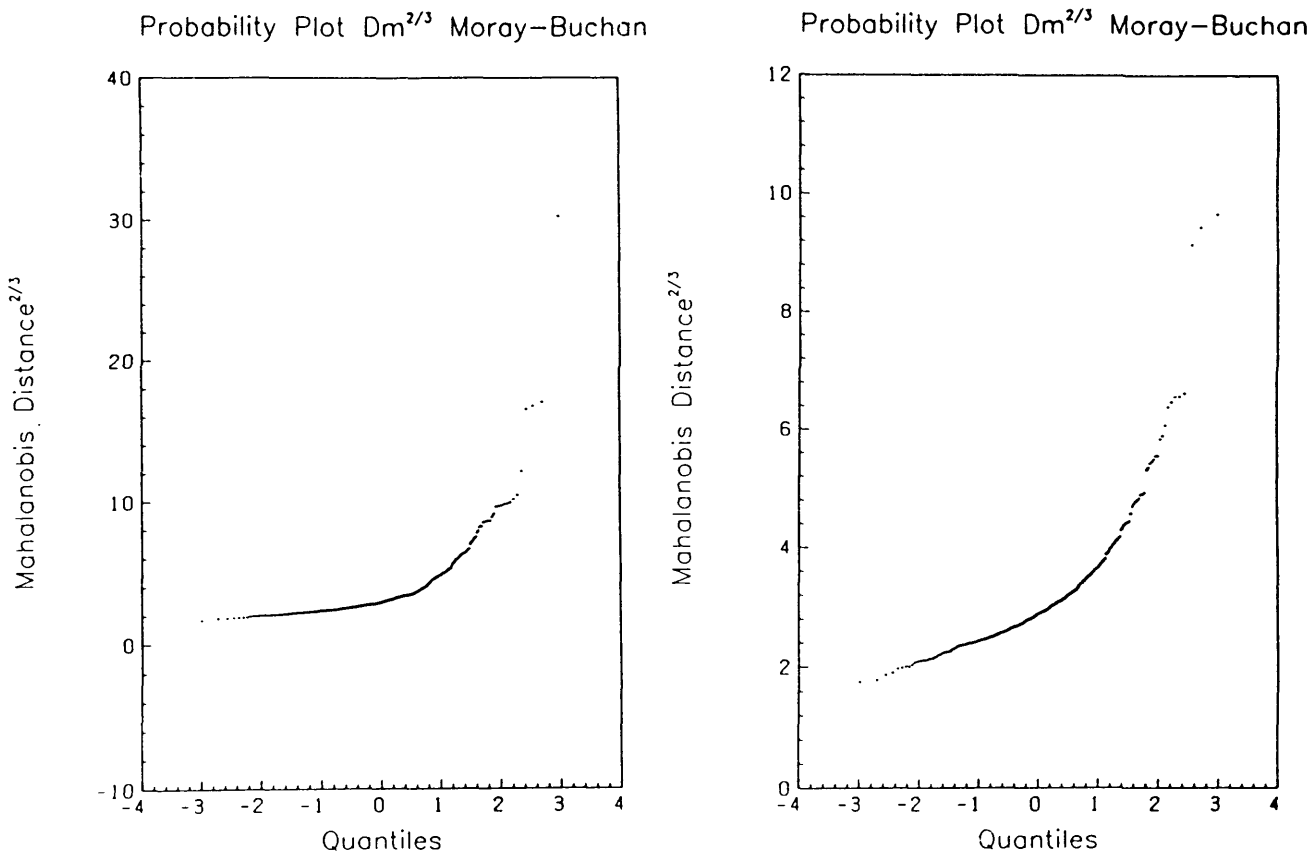
Probability Plot $Dm^{2/3}$ Moray–Buchan Logten Data

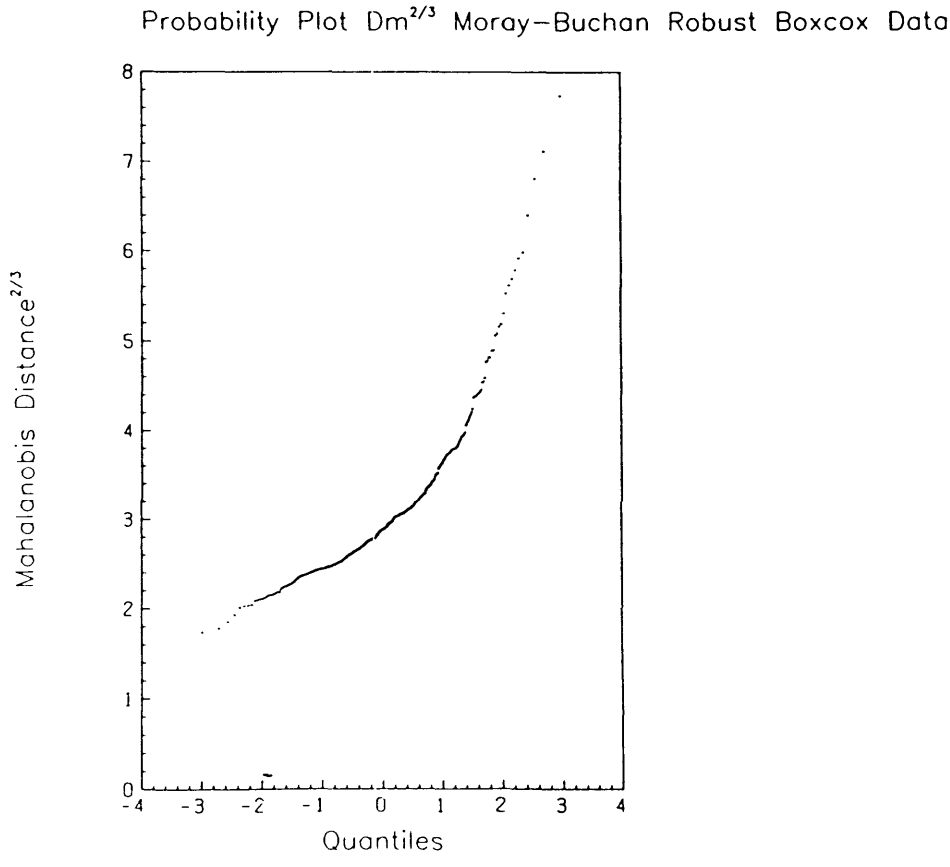


Probability Plot $Dm^{2/3}$ Moray–Buchan Boxcox Data



Robust Raw Data **Figure 5.35** Robust Logten Data





The spatial distribution of robust $D_m^{2/3}$ values for raw data (Figure 5.34) was fully interpreted in Chapter 4. Point symbol maps of robust $D_m^{2/3}$ values for logten and boxcox data are shown in Figures 5.36 and 5.37 respectively. The conclusions from the analysis of these maps and their comparison with Figure 4.38 are presented below. The point symbol maps of robust $D_m^{2/3}$ values for logten and boxcox data show similar areas of interest. They will therefore be discussed as one with any differences being noted.

1) There is a group of samples with large Mahalanobis distanceshigh in the extreme south of the area, close to the Cairngorm Granite. They occur in sites downstream from either granitic outcrop or metamorphosed limestone of undifferentiated Lower/Middle Dalradian age.

2) Another group of samples with large values occur in streams draining the Lower Dalradian Lochaber quartzites supplying the river Spey near Rothes. Robust $D_m^{2/3}$ values from the logten data (Figure 5.36) give less importance to these sites than robust $D_m^{2/3}$ values from the boxcox data. This area is much better emphasized using logten and boxcox data than with raw data.

Figure 5.36

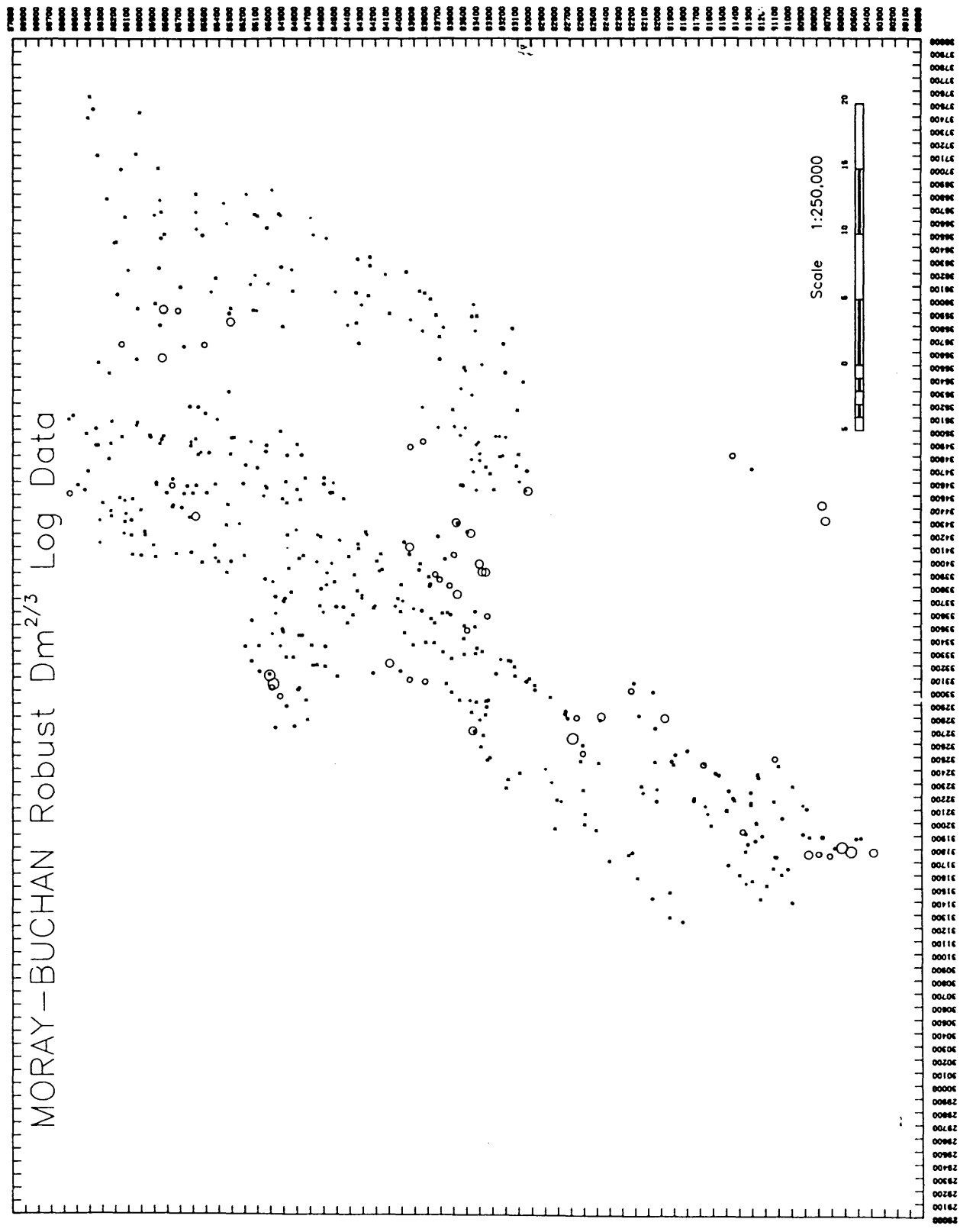
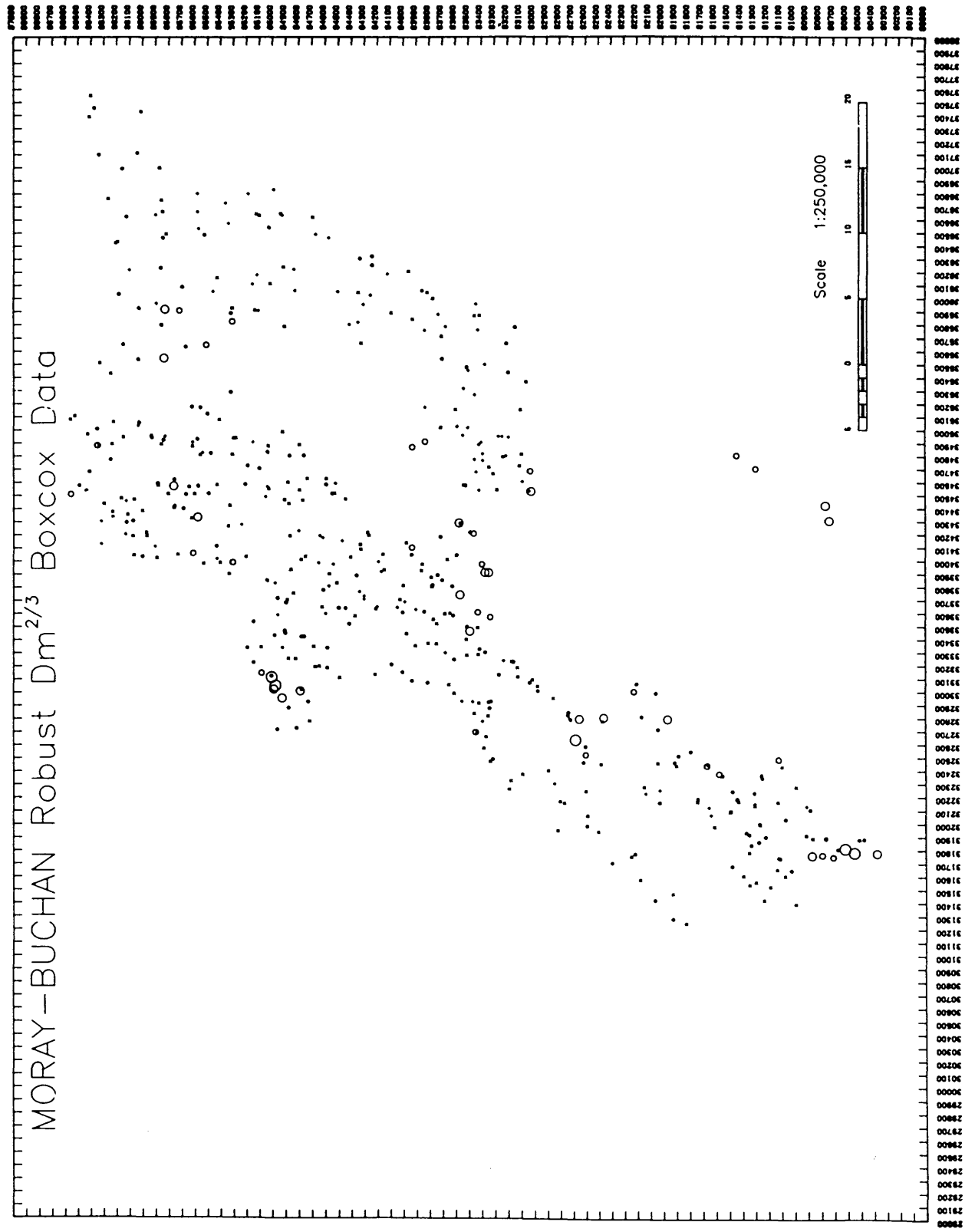


Figure 5.37



3) A group of samples with large Mahalanobis distances occur in streams close to basic/ultrabasic outcrops approximately 11kms SE of Dufftown.

4) A few large $D_m^{2/3}$ values occur over Tayvallich/Crinan Subgroup mixed pelites in the locality about 5km NW of Aberchirder and south of Portsoy. Another group of large values occurs in the same Subgroup in the extreme SE of the study area. The robust $D_m^{2/3}$ values for the raw data give more emphasis to these areas than do the equivalent values for either the boxcox or the logten data.

5) Another group of samples with large Mahalanobis distances occur just north and NE of a small granite outcrop about 10km NE of Tomintoul in streams draining undifferentiated Lower/Middle Dalradian pelites and quartzites.

4.8 Conclusions

Transformation of raw data has been shown to be efficient at achieving univariate and multivariate

distributional normality. However, this is at the expense of anomaly detection. Data transformation causes suppression of anomalies, emphasis of background effects and reduction of anomaly contrast. Whether transformation is required or not depends on the aims of the analysis. If the analyst wishes to examine background relationships and associations then transformation is necessary. If the analyst wishes to detect anomalies then transformation should be avoided, and robust estimators must be used.

As we shall see in later Chapters transformed data can be used to establish a background fit to the raw data. Then, when the background fit is removed from the raw data, the residuals show maximized anomaly contrast (i.e. residuals = data - fit).

CHAPTER SIX

COVARIANCE AND CORRELATION

6.1 Introduction

In Section 4.5 we introduced the concept of covariance as a measure of the interrelationship between two variables, and found that individual values are not particularly meaningful because they are dependent upon the units of measurement of the data.

In order to estimate the degree of interrelationship between variables in a manner not influenced by measurement units, the Pearson linear correlation coefficient (r) is used. This is defined as the ratio of the covariance of the two variables to the product of their standard deviations.

$$r_{jk} = \frac{COV_{jk}}{s_j . s_k} \qquad \text{Equation 6.2}$$

The correlation coefficient is unitless and, since

covariance may equal, but not exceed, the product of the standard deviations of its variables, its range is from +1 through -1. A correlation of +1 indicates a perfect direct relationship between two variables; a correlation of -1 indicates that one variable changes exactly inversely in relation to the other. Between the two extremes is a spectrum of less-than-perfect relationships, including zero which indicates the lack of any sort of linear relationship at all. Correlations may be visualized by plotting the two variables concerned against one another in what is termed a "scatter plot". Six scatter plots are given in Figure 6.1 together with their correlation coefficients. An interesting extreme is shown in Figure 6.1e. One variable is invariant, with the result that its standard deviation is zero. Attempts to calculate the correlation coefficient will encounter division by zero (see equation 6.2); the correlation is therefore undefined in this situation. In the scatter plot 6.1f there is an obvious interdependency between the two variables. However, the correlation coefficient will be zero. This is because the correlation coefficient is an expression of the linear relationship between two variables and is insensitive to non-linear relationships. If a non-linear trend is apparent from a scatter plot of two variables, a transformation is necessary to linearize their relationship (see Chapter 5). Alternatively a rank correlation coefficient may be calculated which is sensitive to non-linear relationships.

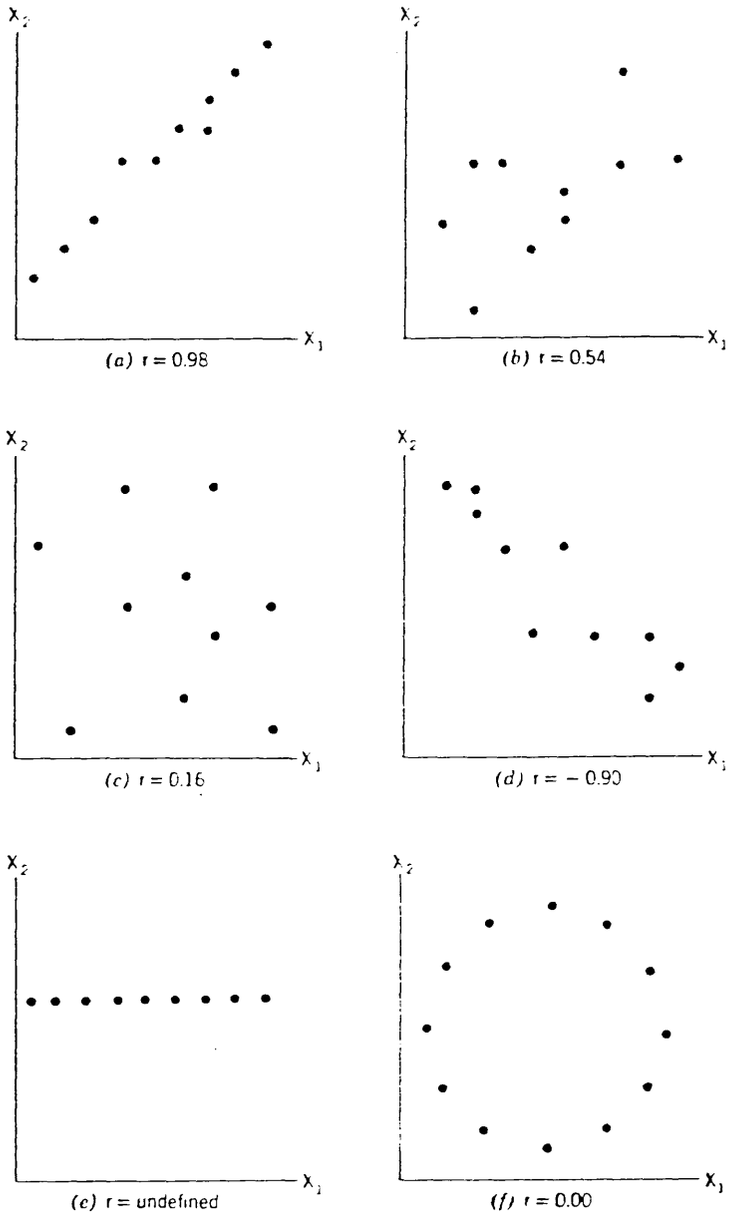


Figure 6.1

Scatter plots of two variables, showing different correlations between variables.

6.2 Covariance and Correlation Matrices

The statistical parameter of covariance or correlation may be used in matrix form to describe the structure of a data point cloud in p-dimensional variate space. Thus the covariance matrix is a symmetrical p x p matrix made up of variances along the leading diagonal and covariances in the off-diagonal positions. The correlation matrix is simply a scaled version of the covariance matrix with ones along the leading diagonal and off-diagonal values obtained using equation 6.2. The use of covariance and correlation matrices as summaries of dispersion and association is common practice in multivariate statistical procedures such as principal components analysis (PCA), cluster analysis and regression analysis. The emphasis and results presented in this chapter are on the correlation matrix rather than the covariance matrix for the following reasons:

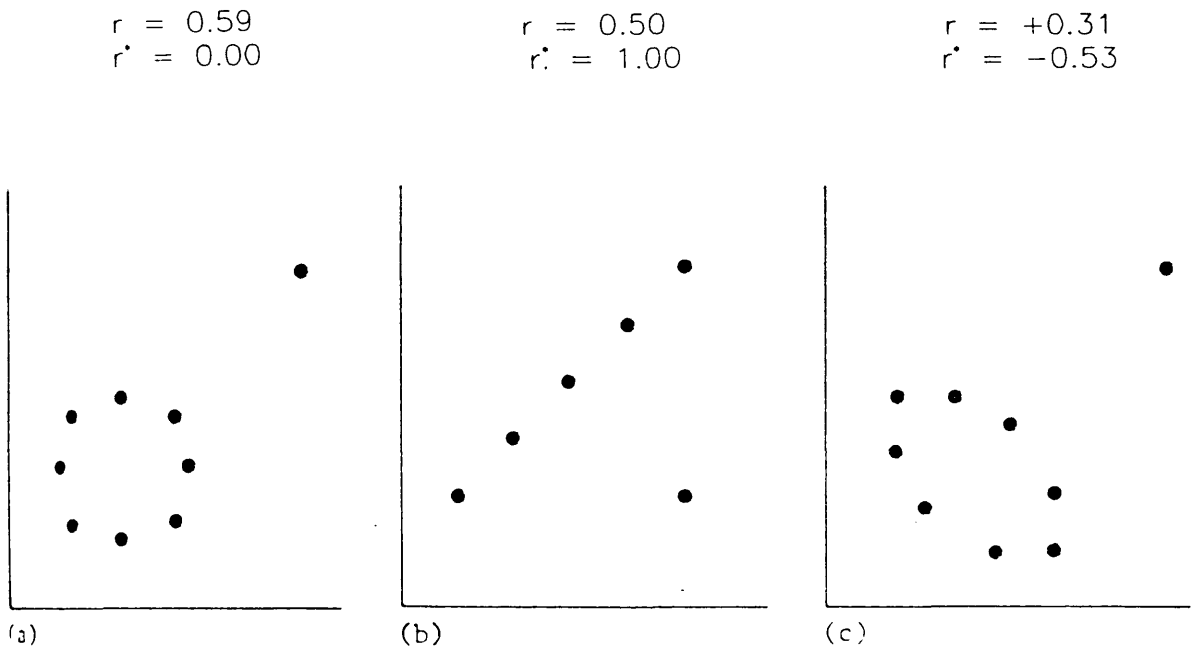
1) If variables in a data set are expressed in several different units of measurement, the elements of the covariance matrix will not be directly comparable. The elements of the correlation matrix, on the other hand, are unitless and therefore directly comparable one with another.

2) Since the elements in the covariance matrix may be very large in absolute value, the computational costs and errors involved in matrix manipulation may be excessive. The elements of the correlation matrix, however, are all between +1 and -1 and so costs and errors are kept to a minimum.

3) It is easy to relate the graphic scatter plot of two variables to its numerical correlation coefficient, whereas there is little such relation with the corresponding covariance.

6.3 Robust Statistics

We have seen in Chapter 3 how the presence of outliers in a univariate set of data can severely distort the classical estimators of location and scale (i.e. mean and variance). In a similar way the presence of outliers in bivariate space will distort the classical estimates of covariance and correlation. Figure 6.2 shows how the presence of outliers can inflate or deflate the value of the correlation coefficient, or even change its sign. Because classical statistical parameters are based on assumptions about the form of distributions, they are inherently sensitive to outliers. Robust statistics are



r = Correlation coefficient with the outlier.
r' = Correlation coefficient without the outlier.

Figure 6.2

Artificial examples showing how an outlier may (a) inflate or (b) deflate the estimate of the correlation coefficient, or even (c) change its sign.

designed to deal with these problems.

The object of robust statistics is to perform valid statistical inference even when faced with departures from the model assumptions. It can be considered as a modification of classical statistics such that small deviations from the assumptions will impair performance only slightly, and larger deviations will not result in catastrophic collapse. To date, robust procedures have dealt almost exclusively with departures from normality and major emphasis has been placed on the departure caused by the presence of symmetric outliers. This problem was formerly treated by outlier detection procedures in which anomalous values were identified visually or automatically and then classical techniques were applied after these observations were deleted.

Robust procedures are more powerful than rejection procedures in detecting multivariate outliers, especially minor outliers. These may be envisaged, in a simple bivariate scatter plot, as occurring to one side of the main point cloud, but not sufficiently far enough away to register as an outlier on either univariate plot (see Figure 6.2b). Robust procedures are more flexible in handling outliers. Unlike outlier rejection procedures in which outliers are completely removed from the analysis, robust procedures can partially incorporate information contained in outliers and achieve a better balance between robustness and loss of information. Robust procedures also

give analyses of the outliers themselves. This may be very helpful if the outliers have some special significance, e.g. indicating mineralization. Finally, robust procedures are more convenient to use and yield more consistent results than do most outlier rejection procedures, especially those involving visual inspection of the data, or graphics.

6.4 Robust Correlation

Proposed robust estimates of the correlation matrix can be divided into two groups: pairwise estimates and multivariate estimates.

6.4.1 Pairwise estimates

The approach is to estimate each off diagonal element of R^* (the robust correlation matrix) separately by a robust correlation coefficient r^*. In this "sums of squares and differences" (SSD) procedure (see Devlin et al, 1981) each

robust correlation coefficient is estimated from sums and differences of standardized data. In the study by Devlin et al standardization and the subsequent variance estimates were all calculated using 5% trimmed means and variances (other estimators which employ smoother downweighting of extreme objects may perform better). The basic equation is:

$$COV_{jk} = \frac{1}{4} [var(X_j + X_k) - var(X_j - X_k)]$$

where X_j and X_k are two random variables. If s_j^2 and s_k^2 are robust estimates of the variances of $(X_j + X_k)$ and $(X_j - X_k)$, then a robust estimate of the covariance between X_j and X_k is given by:

$$S_{jk} = \frac{1}{4} [s_j^2 - s_k^2]$$

The corresponding robust estimate of the correlation coefficient is:

$$r^*_{jk} = \left(\frac{S_{jk}}{S_{jj} \cdot S_{kk}} \right)^{1/2}$$

However, r^*_{jk} obtained in this manner does not necessarily lie within the range -1 through +1. To ensure that r^*_{jk} is in the admissable range, a modification is needed. Let Z_j and Z_k be the standardized form of X_j and X_k, and V^+ and V^- be robust estimates of the variances of $(Z_j + Z_k)$ and $(Z_j - Z_k)$. Then the robust estimate of the correlation coefficient is defined by:

$$r^*_{jk} = (V^+ - V^-)/(V^+ + V^-)$$

If the resulting $R^*(SSD)$ is not positive definite, i.e. if there are less than p nonzero eigenvalues (cf. Chapter 8 for definition), then it is "shrunk" until a positive definite matrix is achieved. The final result is denoted $R^*_+(SSD)$. An estimate of the covariance matrix may be obtained by rescaling the correlation matrix, using a diagonal matrix of trimmed standard deviations for instance.

6.4.2 Multivariate Estimates

Multivariate estimates fall into two groups: those

generated by multivariate trimming, and M-estimators. Both use the Mahalanobis distance as a means to identify outliers. We have seen from Chapter 4 that Mahalanobis distances play a basic role in multivariate robust estimation of location aswell as scale. Although we are primarily concerned in the present chapter with the robust estimation of the correlation matrix, the robust estimation of the mean vector is an important byproduct.

6.4.2.1 Multivariate Trimming

Multivariate trimming is an iterative procedure which involves the calculation at each step of Mahalanobis distances for all objects in the data matrix.

The Mahalanobis distance (D_m) is the distance of the multivariate object vector X_i from the current estimate of the mean vector $\bar{X}^*$ measured with relation to S^* the current estimate of the covariance matrix.

$$D_m = [(X_i - \bar{X}^*)' S^{*-1} (X_i - \bar{X}^*)]^{1/2} \qquad \text{Equation 6.3}$$

A specified percentage of the most extreme objects (having the largest D_m values) is temporarily set aside, and the remaining objects are used to compute the new estimates $\bar{X}^*$ and S^* in the same manner as one would the sample mean vector, $\bar{X}$, and S, the covariance matrix. To start the process, $\bar{X}^*$ and S^* are taken to be $\bar{X}$ and S. The S^* at every stage can be converted to R^* by rescaling. It continues with the iterative process which terminates as soon as the average absolute change in the Fisher Z transforms of the r^*_{jk} values between two successive iterations does not exceed 10^{-3} , or after the 25th step. The Fisher Z transform of a coefficient r^*_{jk} is given by:

$$Z_{jk} = 0.5 \log[(1 + r^*_{jk}) / (1 - r^*_{jk})$$

The final estimates are denoted $S^*(MVT)$ and $R^*(MVT)$.

6.4.2.2 M-estimates

Two other procedures proposed by Devlin et al (1981) are examples of M-estimators and are similar to the multivariate trimming approach. Instead of deleting some fraction of the data with large Mahalanobis distances, weights are assigned to the object vectors based on these

distances. In this section I shall discuss the principles of multivariate M-estimation and explain the specific procedure used in this thesis to generate M-estimates. This procedure was developed by Campbell (1979) and is implemented using a FORTRAN program called WTHGRP. This procedure is very similar to those of Devlin et al and so a full explanation of their techniques is not given here. I shall call the correlation matrix generated by M-estimation: $R^*_{(MLT)}$.

From the applied point of view, M-estimators can be considered as simple modifications of classical estimators; the contribution of an observation to the statistic of interest is given full unit weight if it is a reasonable observation, otherwise its contribution is downweighted. M-estimators can be considered as classical estimators computed after a weight function has been applied to the data. The weight function is given by

$$W_i = w(U_i) = \psi(U_i)/U_i \qquad \text{Equation 6.4}$$

where U_i is a measure of deviation reflecting the discrepancy of the ith object vector from the robust location estimate, relative to a robust scale estimate. The bounded influence function ψ is linear for the range of values of U_i corresponding to reasonable data, but bounded

outside this range. We have seen from Chapter 3 that for univariate M-estimation of location and scale:

$$U_i = \frac{X_i - T}{cS_n} \qquad \text{Equation 6.5}$$

and that ψ may have a variety of forms (see Figures 3.1 through 3.7). For robust M-estimation of multivariate location and scale, the appropriate measure of deviation turns out to be the Mahalanobis distance. The weight function then becomes:

$$W_i = w(D_m) = \psi(D_m)/D_m \qquad \text{Equation 6.6}$$

where D_m is given by equation 6.3. The equations used to define robust M-estimates of means and covariances are:

$$X^* = \frac{\Sigma W_i X_i}{\Sigma W_i} \qquad \text{Equation 6.7}$$

$$S^* = \frac{\Sigma\, W_i^2\, (X_i - \bar{X}^*)\,(X_i - \bar{X}^*)'}{(\Sigma\, W_i^2 - 1)} \qquad \text{Equation 6.8}$$

both $\bar{X}^*$ and S^* are obtained iteratively, as with the MVT procedure. The convergence criterion is the same as for MVT. The function ψ reflects the influence of an object on the statistic(s) of interest. The function is usually taken to be linear over the range of values of D_m corresponding to reasonable data, but bounded outside this range. Following the idea of the Hampel function (Figure 3.5), ψ is taken to redescend for object vectors with sufficiently large Mahalanobis distances, so that the influence and hence the weight of an extreme atypical object will be zero.

In the FORTRAN program WTHGRP the function ψ is taken to be

$$\begin{aligned}
\psi\,(D_m) &= D_m, && \text{if } D_m \leq D_0 \\
\psi\,(D_m) &= D_0, && \text{if } D_0 < D_m \leq D_1 \\
\psi\,(D_m) &= D_0 \exp[-0.5(D_m - D_1)^2 / b_2^2], && \text{if } D_m > D_1
\end{aligned}$$

Equation 6.9

The values of D_0 and D_1 are given by

$$D_j = [p(1+k^{1/2} b_j - k)^3]^{1/2} \; ; \quad k = 2/(9p) \; , \quad j = 0,1$$

values of $b_0 = 1.5$ and $b_1 = 2.0$ are suggested by Campbell (1979). When $b_2 = \infty$, the nondescending influence function proposed by Huber (see Figure 3.4) results. The rate of decrease of the redescending function is controlled by the parameter b_2. A value of $b_2 = 1.8$ is recommended by Campbell (1979). With this choice of coefficients the resultant robust estimates are within 2.5% of the usual estimates when the calculations are carried out on data generated from a multivariate Gaussian distribution.

The recommended procedure to determine robust $\bar{X}$ and S and the associated weights may be summarized as follows:

1) generate Mahalanobis distances based on initial estimates of means and covariances;

2) generate weights based on the nondescending ψ function ($b_2 = \infty$);

3) calculate means $\bar{X}^*$;

4) use means calculated in step 3 to generate robust covariance matrix S^*;

5) using these means and covariances, generate new Mahalanobis distances;

6) go to step 2 and iterate until convergence;

7) once the procedure converges with $b_2 = \infty$, the redescending ψ function with $b_2 = 1.8$ is introduced for up to a further 10 iterations.

There are three choices of initial estimates for use in step 1 that may be implemented in WTHGRP. The first is the usual sample means and covariances. The second and third choices involve robust M-estimation for each variable separately. For both these choices, the median and 0.74 of the interquartile range are used as robust initial estimates of location and scale and the univariate analogues of equations 6.3, 6.7 and 6.8 with $D_0 = 1.5$, $D_1 = 2.0$ and $b_2 = 1.8$ are carried out for up to 10 iterations. The second choice replaces the usual means and standard deviations by the resulting univariate M-estimates. The third choice recalculates the means and covariances using as weights the product of the weights given by the individual univariate M-estimates calculation.

6.5 Discussion

The following alternative estimators of the population correlation matrix have been discussed above:

R = classical sample correlation matrix;

R_+^*(SSD) = positive definite correlation matrix by element-wise bivariate method;

R^*(MVT) = multivariate trimmed; and

R^*(MLT) = maximum likelihood method.

A useful way to measure the robustness of an estimator is by the breakdown point (see Chapter 3). The breakdown point is the largest percentage of outliers which each estimator could tolerate and still give reasonable values. Outliers may be added in two ways - either symmetrically or asymmetrically - with differing results. The initial estimates of $\bar{X}$ and S are also critical to the robustness of the final estimator. Devlin et al (1981) have determined breakdown points using Monte Carlo simulation for all the above estimators, not only for symmetric and asymmetric outliers, but also for non-robust ($\bar{X}$ and S) and robust ($\bar{X}^*$ and S^*) initial estimates, where $\bar{X}^*$ equals the median vector and S^* in the univariate case would be the interquartile range or MAD (see Table 6.1).

Table 6.1

BREAKDOWN POINTS

% of Contamination Tolerated

	Type of Contamination	
Estimator	Symmetric	Asymmetric
R	0	0
R_+^* (SSD)	5	5
R^* (MVT)	10	4
R^* (MVT) with robust start	10	10
R^* (MLT)	15	2
R^* (MLT) with robust start	≥ 25	2

Using Monte Carlo simulation techniques Devlin et al arrived at certain conclusions about the estimation of correlation matrices. The sample correlation matrix R has a breakdown point of zero and is therefore unacceptable as either an estimate of the true population correlation matrix, or as input into robust multivariate statistics such as robust PCA. This is because the existence of only a few anomalous values in the sample will cause R to be distorted and unstable. R^*_+(SSD) is an adequate robust estimator but fails as a basis for estimating the eigenvalues and eigenvectors corresponding to near singularities in the data. R^*(MVT) has a major advantage over R^*(MLT) in that its breakdown point is the same as its trimming percentage and is independent of p - the number of variables. A robust start improves the breakdown point in the presence of asymmetric outliers. R^*(MVT) also tends to converge quickly - usually in two iterations. R^*(MLT) has a breakdown point which appears to be dependent upon p - usually no more than 1/(p+1). It also appears that breakdown is most likely to occur when the distribution of outliers is far from being symmetric, assuming the rest of the data are of this type. A robust start is also recommended. The main advantage of R^*(MLT) over other estimators is that it is an excellent method for obtaining robust estimates of the eigenvalues and eigenvectors. In this context it is also likely to be of use in robust extensions of other multivariate techniques such as cluster

analysis, regression analysis and discriminant analysis. One important advantage of M-estimation over multivariate trimming is that the latter technique trims a fixed, data-independent percentage of object vectors off the high end of the Mahalanobis distance distribution, irrespective of whether trimming is necessary or not. In contrast, the weights associated with the object vectors in M-estimation are calculated using data dependent parameters.

6.6 Correlation and the SIND Data - a Comparative Analysis

The comparative analysis has the following aims.

1) To examine the effects on correlation of data transformations;

2) To analyse the effects of robust compared with non-robust estimation on the correlation matrix for each of the three data sets (raw, logten and boxcox); and

3) To examine the effects on robust correlation of data transformations.

Table 6.2 gives the classical, non-robust correlation matrices for each of the three data sets superimposed on

Table 6.2

Non-Robust Correlation Matrices SIND RAW, LOGTEN and BOXCOX Data

		Zn	Fe	Mn	Cd	Cu	Pb
Zn	Raw	1.0					
	Logten	1.0					
	Boxcox	1.0					
Fe	Raw	0.296	1.0				
	Logten	0.664	1.0				
	Boxcox	0.789	1.0				
Mn	Raw	0.209	0.743	1.0			
	Logten	0.547	0.777	1.0			
	Boxcox	0.614	0.732	1.0			
Cd	Raw	0.967	0.428	0.404	1.0		
	Logten	0.907	0.643	0.556	1.0		
	Boxcox	0.871	0.652	0.556	1.0		
Cu	Raw	0.969	0.194	0.051	0.906	1.0	
	Logten	0.954	0.545	0.387	0.924	1.0	
	Boxcox	0.916	0.697	0.491	0.927	1.0	
Pb	Raw	0.096	0.723	0.830	0.267	0.026	1.0
	Logten	0.300	0.625	0.659	0.323	0.215	1.0
	Boxcox	0.367	0.586	0.669	0.325	0.316	1.0

each other so that any changes are readily seen. The following conclusions can be made on the effect that data transforms have on correlation.

1) The correlation coefficients for raw data are controlled by the anomalous objects. Recall from Chapter 2 that there are two separate processes operating in the SIND area. The first is that of mineralization which supplies Zn and Cu and accessory Cd to sites 13, 15 and 16. The second process is element enhancement by scavenging onto Fe and Mn oxides operating principally at sites 22 and 23. The remaining element, Pb, behaves in a largely independent fashion but if anything tends to relate more to Fe and Mn than to Zn, Cu and Cd. Correlation coefficients of raw data reflect these two processes. Correlations between the element groups (Zn, Cu and Cd) and (Fe, Mn and Pb) tend to be low, while correlations within the two groups tend to be high.

2) Logten transformation, by reducing the anomaly contrast and accentuating the spread of the background population, causes large changes in the correlation coefficients of between-group variables (see above). For instance the background correlation between Zn and Fe is strong positive. However, Zn-Cu mineralization has dramatically increased the Zn concentration in three

samples which were not unusually high in Fe. These two effects, working against each other, result in a low raw data correlation. Logten transformation causes little change in the positive background correlation between Zn and Fe, but has the effect of reducing the influence of mineralization. The resultant correlation coefficient (0.664) thus reflects the background relationship between the two variables rather than the anomalous.

Logten transformation has, conversely, little effect on the correlation coefficients of within group variables. Note that Cd and Pb behave slightly unusually, relative to the other variables, due to the weakness of their associations.

3) Boxcox transformation tends to accentuate the trend shown by logten transformation. Thus if logten transformation causes an increase in the correlation coefficient over raw data then the boxcox transform will tend to result in a further increase. This may be attributed to the fact that boxcox transformed variables tend to have frequency distributions closer to normal than logten transformed variables. This means that any anomalous effects (such as that of mineralization in the Zn-Fe example above) are further reduced and the resultant correlation coefficients are more representative of background relationships.

Table 6.3

Non-Robust and Robust Correlation Matrices SIND RAW Data

	Zn	Fe	Mn	Cd	Cu	Pb
Zn	1.0					
Robust	1.0					
Fe	0.296	1.0				
Robust	0.618	1.0				
Mn	0.209	0.743	1.0			
Robust	0.591	0.604	1.0			
Cd	0.967	0.428	0.404	1.0		
Robust	0.686	0.488	0.413	1.0		
Cu	0.969	0.194	0.051	0.906	1.0	
Robust	0.799	0.391	0.337	0.843	1.0	
Pb	0.096	0.723	0.830	0.267	0.026	1.0
Robust	0.099	0.306	0.289	0.156	0.146	1.0

Table 6.4

Non–Robust and Robust Correlation Matrices SIND LOGTEN Data

	Zn	Fe	Mn	Cd	Cu	Pb
Zn	1.0					
Robust	1.0					
Fe	0.664	1.0				
Robust	0.657	1.0				
Mn	0.547	0.777	1.0			
Robust	0.539	0.766	1.0			
Cd	0.907	0.643	0.556	1.0		
Robust	0.906	0.631	0.539	1.0		
Cu	0.954	0.545	0.387	0.924	1.0	
Robust	0.955	0.543	0.385	0.927	1.0	
Pb	0.300	0.625	0.659	0.323	0.214	1.0
Robust	0.277	0.603	0.628	0.293	0.202	1.0

Table 6.5

Non-Robust and Robust Correlation Matrices SIND BOXCOX Data

	Zn	Fe	Mn	Cd	Cu	Pb
Zn	1.0					
Robust	1.0					
Fe	0.789	1.0				
Robust	0.783	1.0				
Mn	0.614	0.732	1.0			
Robust	0.606	0.722	1.0			
Cd	0.870	0.652	0.566	1.0		
Robust	0.867	0.637	0.532	1.0		
Cu	0.916	0.699	0.491	0.927	1.0	
Robust	0.914	0.693	0.484	0.927	1.0	
Pb	0.367	0.586	0.669	0.324	0.316	1.0
Robust	0.338	0.560	0.624	0.283	0.294	1.0

Tables 6.3, 6.4 and 6.5 show the non-robust correlation matrix superimposed on the robust for each of the three data sets: raw, logten and boxcox respectively. We observe that:

1) Robust M-estimation on raw data results in five anomalous object vectors being assigned weights of zero. The object vectors are samples 13, 15, 16, 22 and 23 and, for the purposes of the calculation of robust univariate and multivariate statistics, may be regarded as omitted from the data. Removal of anomalies can have two effects. If the anomalous object vectors enhance the background relationship between two variables then their removal will tend to cause a reduction in the correlation coefficient. Conversely, if the anomalous vectors counteract the background relationship then their removal will tend to cause an increase in the correlation coefficient.

Of course, unusual situations are bound to occur. It may be that removal of anomalies has little effect on the correlation coefficient between two variables (e.g. Mn versus Fe, Fe versus Cd, Cd versus Mn), or that a low correlation may persist or be made worse by their removal (Pb versus Zn, Pb versus Cd, Pb versus Cu).

2) Robust M-estimation on logten transformed data (Table 6.4) results in only two object vectors (10 and 23) being assigned non-unit weights (0.97 and 0.76 respectively). Robust estimation on boxcox transformed data (Table 6.5) results in only one object vector (23) being given a non-unit weight (0.69). With such little change in the object vector weighting we expect and see little change in the robust compared to the non-robust correlation matrices for logten and boxcox transformed data. The only trend observed in Tables 6.4 and 6.5 is that in the vast majority of cases the robust coefficient is lower than the non-robust.

To examine the effects of data transformations on robust correlation matrices Table 6.6 is used, from which we find:

1) In all but one case (Mn versus Zn) the raw robust correlation coefficient is the smallest of the three.

2) There is, of course, no way of knowing which of the three robust matrices is closest to the true (unknown) correlation matrix. Since the logten and boxcox matrices are generated on the premise that data which has univariate and multivariate distributional normality is likely to produce the "best" matrix then we can state that the boxcox robust correlation matrix is preferred

Table 6.6

Robust Correlation Matrices SIND RAW, LOGTEN and BOXCOX Data

		Zn	Fe	Mn	Cd	Cu	Pb
Zn	Raw	1.0					
	Logten	1.0					
	Boxcox	1.0					
Fe	Raw	0.618	1.0				
	Logten	0.657	1.0				
	Boxcox	0.783	1.0				
Mn	Raw	0.591	0.604	1.0			
	Logten	0.539	0.766	1.0			
	Boxcox	0.606	0.722	1.0			
Cd	Raw	0.686	0.488	0.414	1.0		
	Logten	0.906	0.631	0.539	1.0		
	Boxcox	0.867	0.637	0.532	1.0		
Cu	Raw	0.799	0.392	0.337	0.843	1.0	
	Logten	0.955	0.543	0.385	0.927	1.0	
	Boxcox	0.914	0.693	0.484	0.927	1.0	
Pb	Raw	0.099	0.306	0.289	0.156	0.146	1.0
	Logten	0.277	0.603	0.628	0.293	0.202	1.0
	Boxcox	0.338	0.560	0.624	0.283	0.294	1.0

to the logten robust matrix.

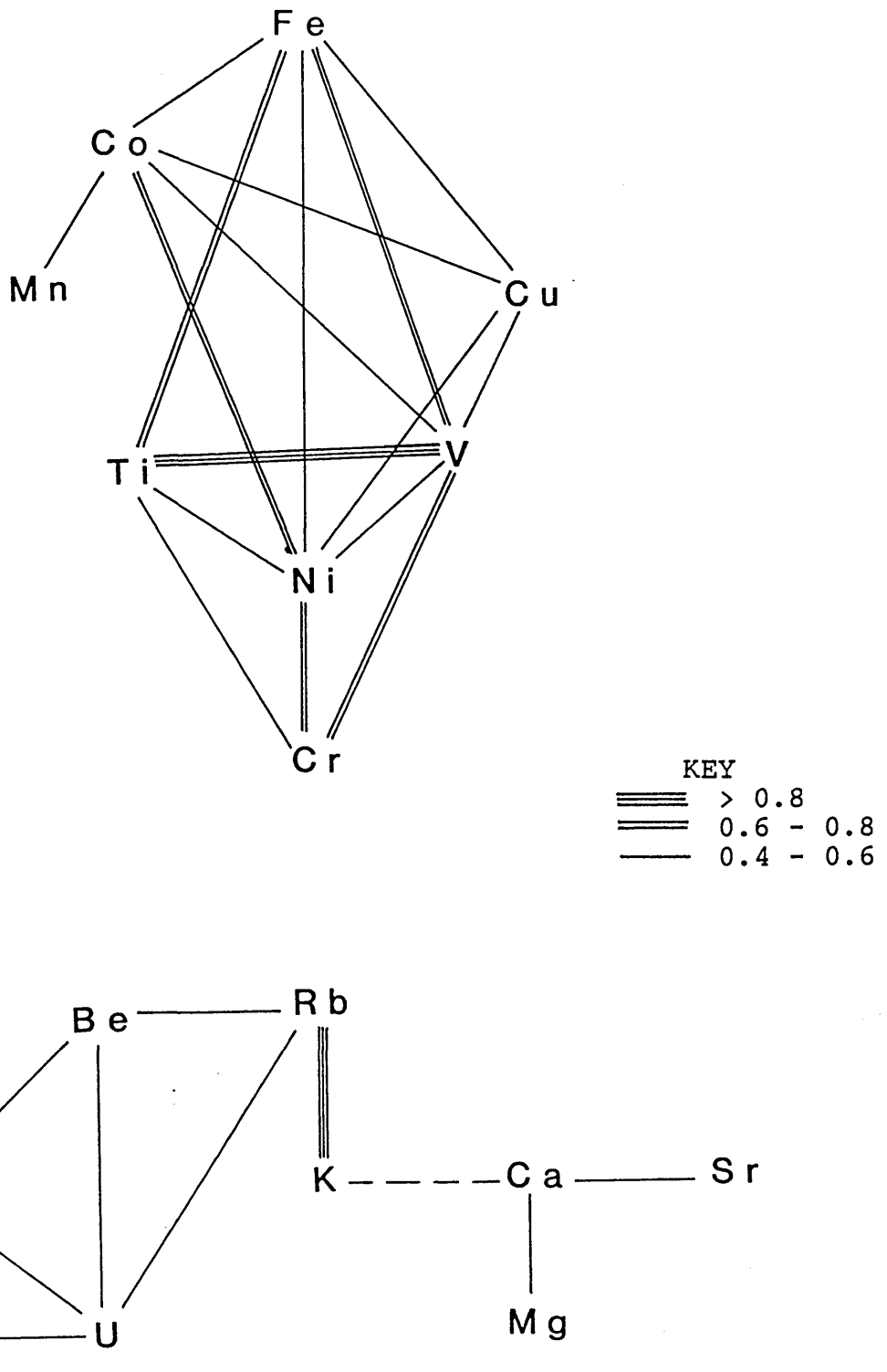
The close similarity between the robust correlation matrices for raw data and boxcox data suggest that, contrary to the conclusions of Devlin et al, 1981, robust techniques seem in practice to be resistant to highly asymmetric distributions (e.g.the SIND raw data) aswell as to heavy tailed symmetric distributions.

6.7 Correlation and the Moray-Buchan Data

The correlation matrices for the MB data are large and complex to interpret in tabular form so correlation linkage diagrams have been created for each matrix as a summary of the major associations. For each of the three data sets two correlation linkage diagrams have been produced, one for the non-robust matrix and one for the robust matrix (Figures 6.3 through 6.8).

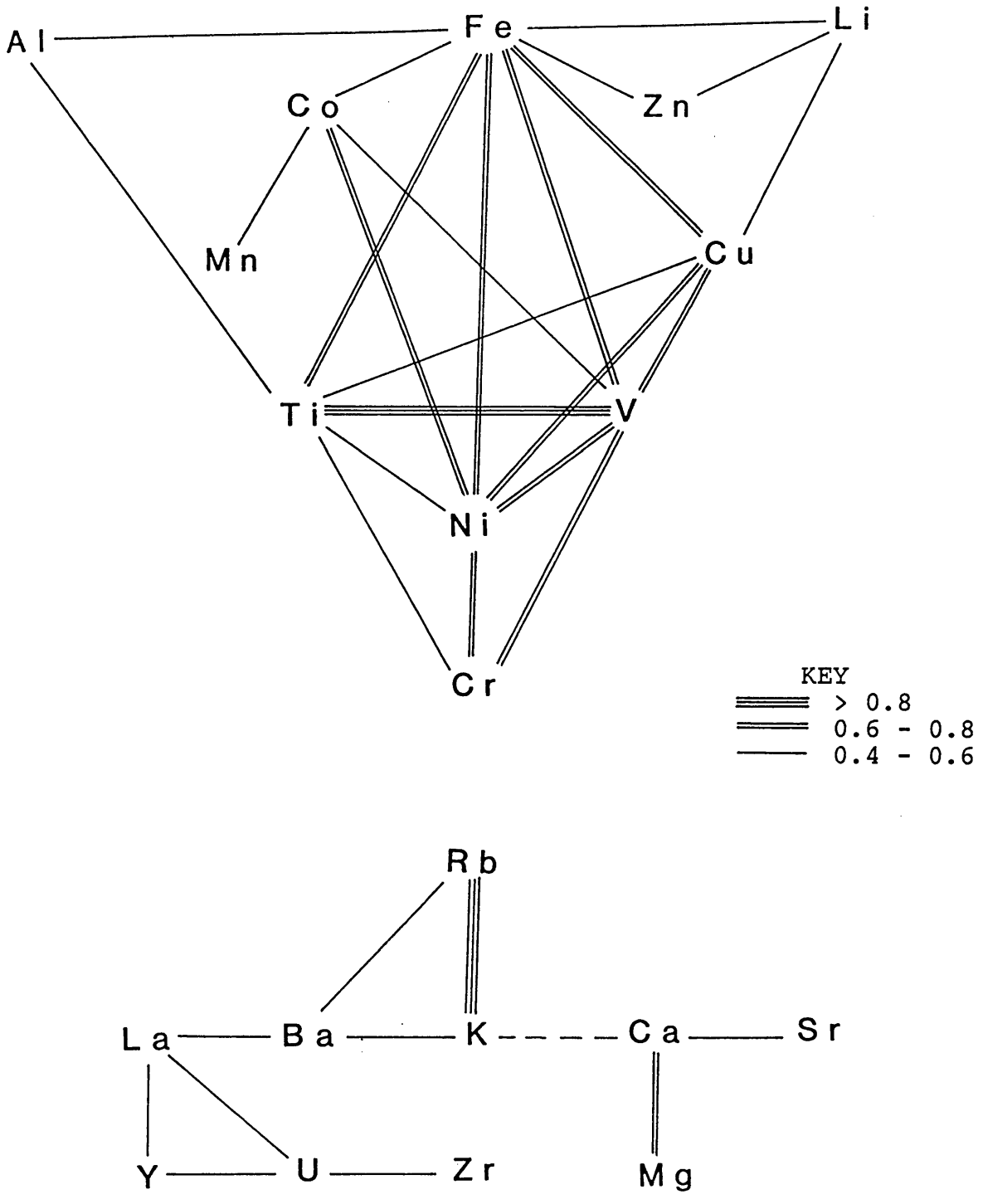
Two groups of variable associations are clearly seen on each of the six diagrams. The two groups are linked for logten and boxcox data, but by low negative correlations which serve to emphasise their separate identity. Group one is dominated by elements of mafic association (Co, Cr, Cu, Fe, Mn, Ni, Ti, V and Zn) with a subgroup consisting of Fe, Cu, Zn, Li and B. Group two may be split into two

Figure 6.3



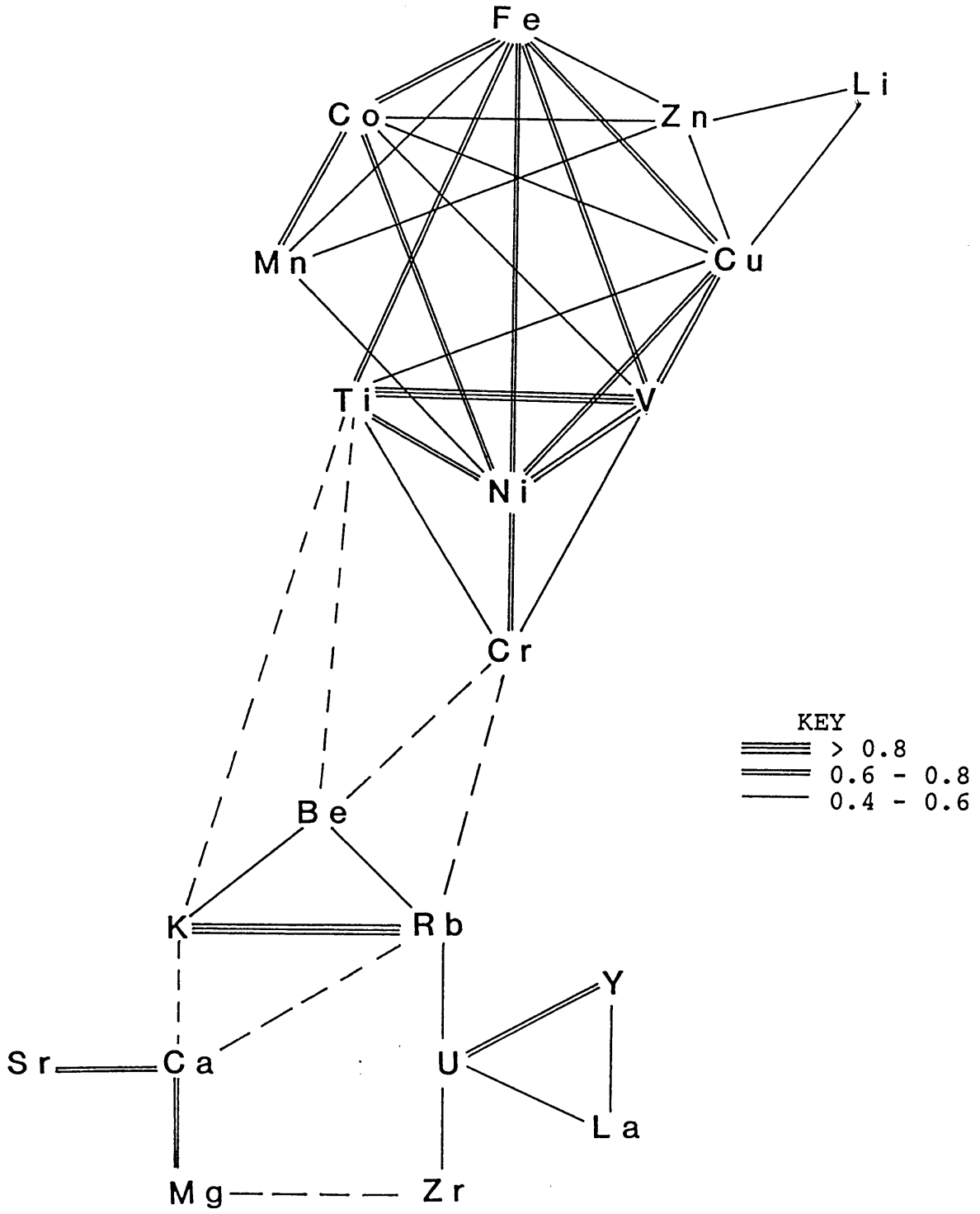
Moray-Buchan Raw data Correlation Linkage Diagram

Figure 6.4



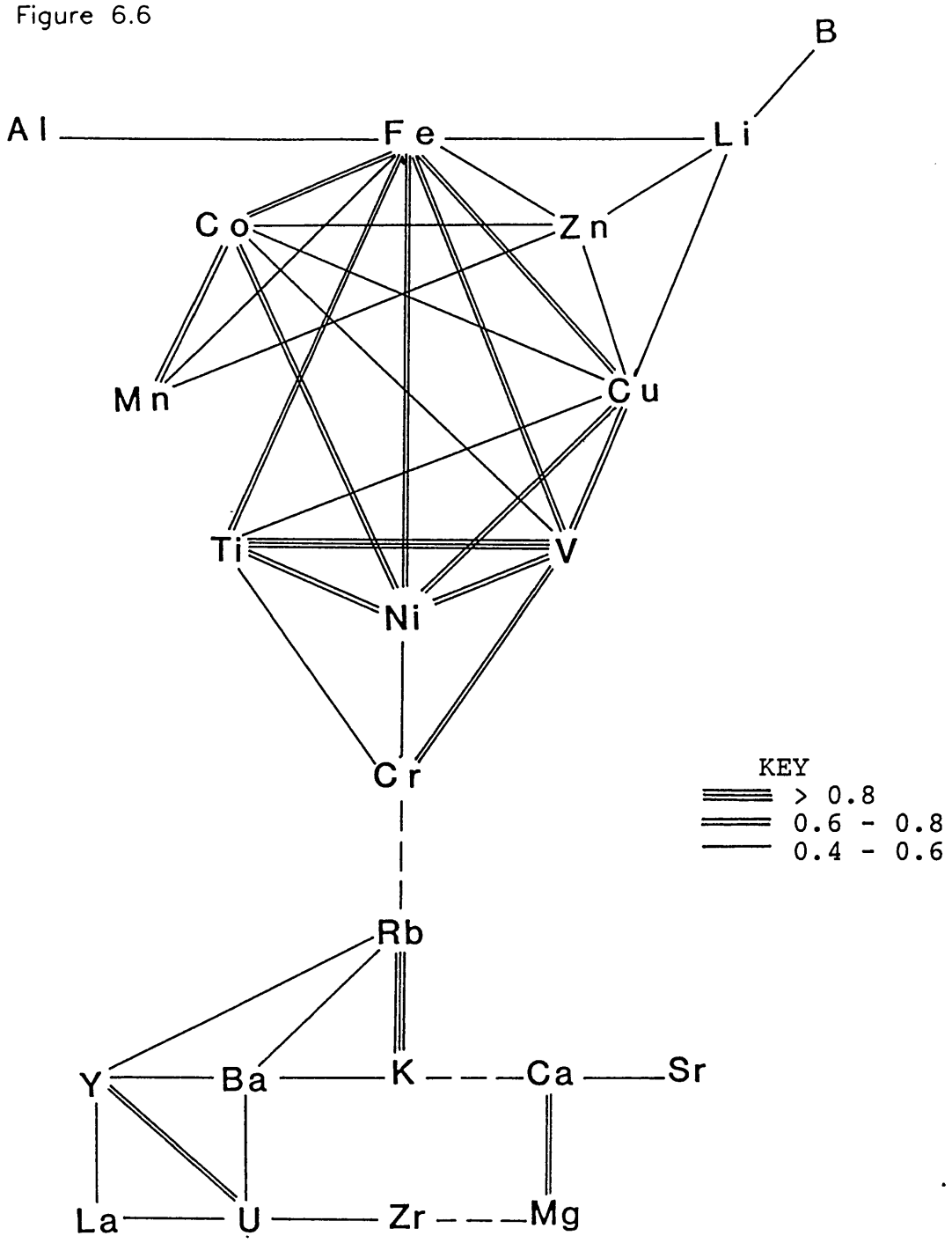
Moray-Buchan Raw data Robust Correlation Linkage Diagram

Figure 6.5



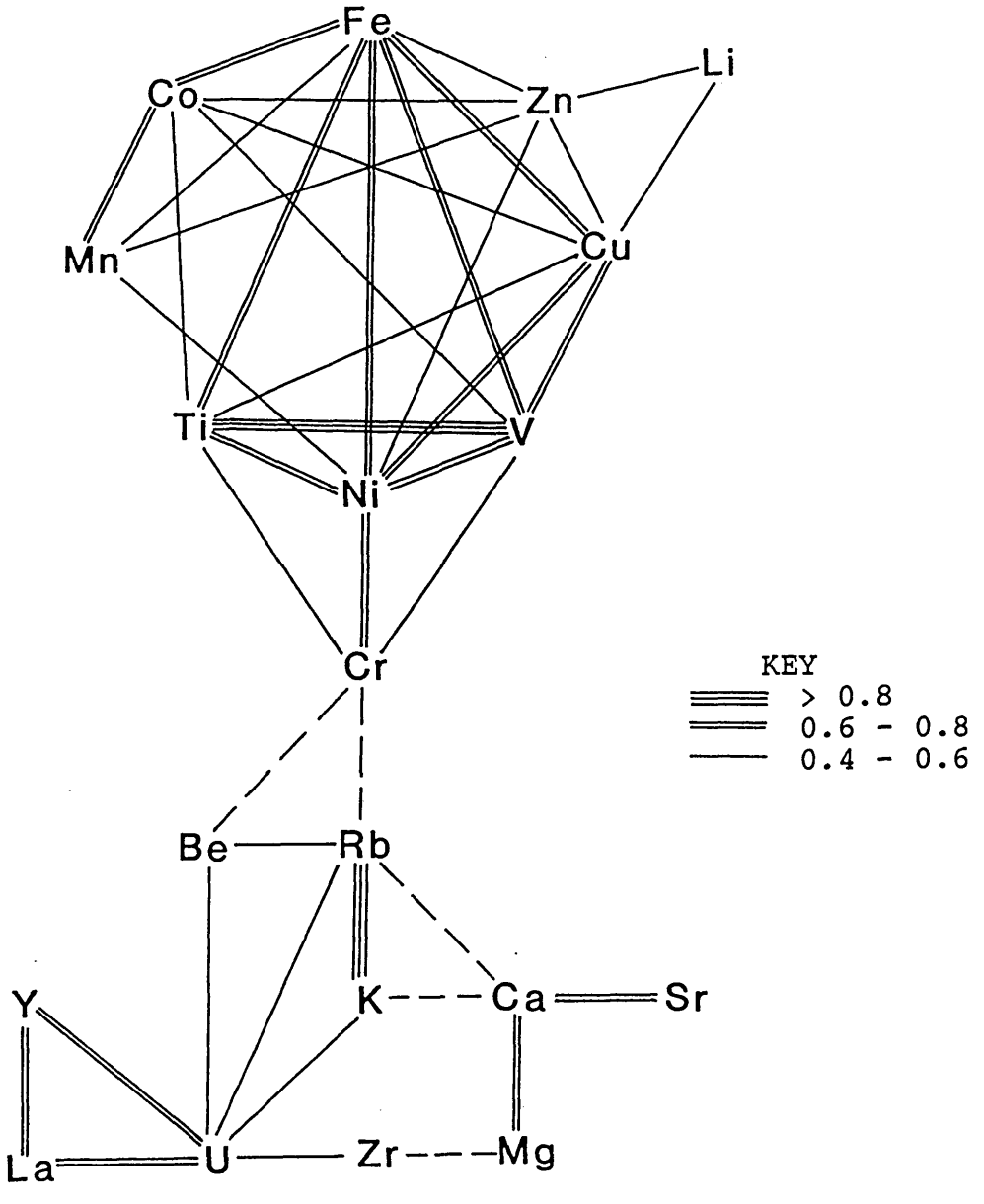
Moray-Buchan Logten data Correlation Linkage Diagram

Figure 6.6



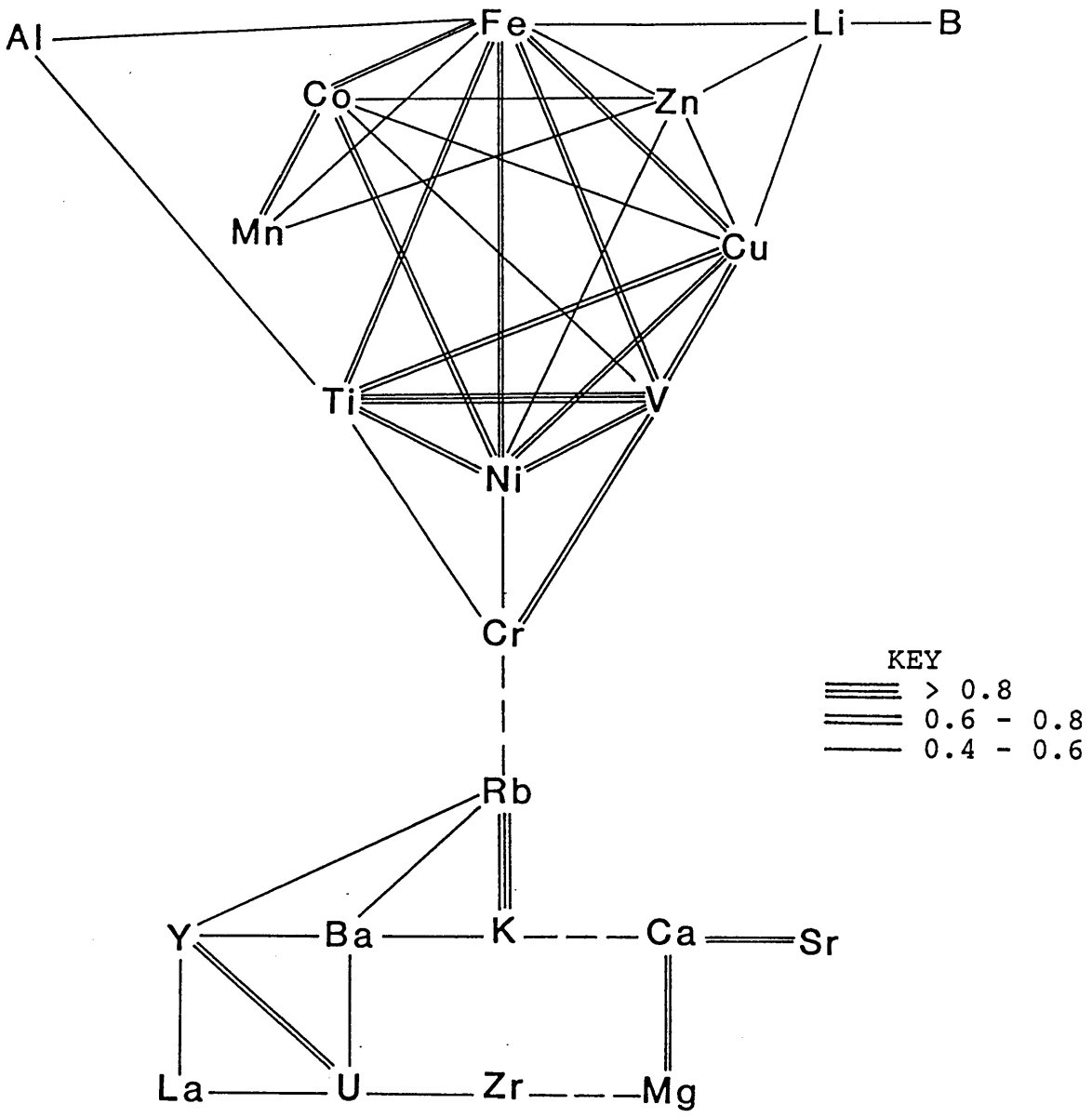
Moray-Buchan Logten data Robust Correlation Linkage Diagram

Figure 6.7



Moray-Buchan Boxcox data Correlation Linkage Diagram

Figure 6.8



Moray-Buchan Boxcox data Robust Correlation Linkage Diagram

subgroups, one consisting of Ca, Sr and Mg probably represents metamorphosed limestones, the second comprising Rb, K, Ba/Be, U, La, Y and Zr is probably related to felsic igneous rocks and metasediments.

Robust estimation causes Al, B and Ba to be included in the correlation linkage diagrams and Be to be excluded. Robust estimation causes correlations between Al, B and Ba and other variables to increase because the minor anomalies (those object vectors which are anomalous in only one of the two variables) which are the cause of the low correlations are downweighted. Conversely, correlations between Be and other variables have decreased because the major anomalies (those object vectors which are anomalous in both variables) which are the cause of the high correlations are downweighted.

Differences between the correlation linkage diagrams for logten and boxcox transformed data are small in comparison with the raw data diagram.

A full analysis and interpretation of the MB correlation matrices along the same lines as the SIND data is impossible. The conclusions reached from the analysis of the smaller data set may instead be applied to the larger with the proof of the validity of those conclusions being in the results of their application to the multivariate techniques described in subsequent chapters.

6.8 Conclusions

The sample correlation matrix should reflect the structure of the background sample population, and must not be distorted by the influence of anomalous object vectors. With this aim in mind we can now categorically state that the raw correlation matrix (assuming anomalies exist) is totally unsatisfactory. There are then two paths to follow in achieving the above aim. The first is simply to use robust estimation of the correlation matrix based on the raw data.

The second path involves data transformation followed by robust estimation. Boxcox transformation is superior to logten in achieving both univariate and multivariate distributional normality and consequently has a correlation matrix which is likely to characterize background population structure the better. Robust estimation is in some ways the "insurance" or the icing-on-the-cake in the sense that any anomalous object vectors that still remain after transformation get downweighted.

Intuitively the second path probably results in a correlation matrix that best represents background population structure. If the aim of the data analysis is anomaly detection the best combination for use in multivariate techniques is the boxcox robust correlation

matrix with the raw data - this will give the best possible chance of a high anomaly contrast. If the aim of the data analysis is determination of background relationships and associations then the boxcox robust correlation matrix should be used in conjunction with the boxcox transformed data - this will ensure that anomalous object vectors have the least possible effect on the analysis.

CHAPTER SEVEN

GRAPHICAL REPRESENTATION OF MULTIVARIATE DATA

7.1 Introduction

The graphical examination of data has long been regarded by statisticians as an extremely useful technique and most would advocate it as an essential precursor to more formal analysis.

Although attempts at graphic depiction of empirical data go back to the 10th or 11th Centuries A.D. (Beniger and Robyn, 1978), it was only after the work of Crome and Playfair in the late 18th and early 19th Centuries that the use of graphs and charts for data display became accepted practice (see Fienberg, 1979).

Playfair's work provides excellent examples of good graphics. He gave us the bar chart in his "Commercial and Political Atlas", (1786) and the pie chart and circle graph in "The Statistical Breviary", (1801). The latter publication and "An Inquiry into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations", (1805) show some of the earliest attempts at the graphical

depiction of multivariate data. Subsequent developments include those of Fourier (cumulative frequency curve) and Quetelet (graphs of frequency curves, plots of histograms with limiting normal curves). In 1849 Fletcher published the first statistical map in a statistical journal, although such maps had appeared elsewhere as early as 1819. In 1857 Florence Nightingale (1857, 1858) introduced the coxcomb chart to describe by month the causes of mortality in the British Army during the Crimean War. The coxcomb is the forerunner of the modern day rose chart and other graphs used to present cyclic phenomena. The "Statistical Atlas of the United States Based on the Results of the Ninth Census" (Walker, 1874) contained the first examples of population pyramids and bilateral frequency polygons.

Despite its use throughout the 19th Century by influential men and women, graphical data display came to geology relatively late. Harker's (1909) interest in the use of graphical techniques in geology set the scene and by the time Bowen's classic work "The Evolution of the Igneous Rocks" was published in 1928, graphical methods for studying the variation of rock composition (frequency distributions, bivariate scatter plots and triangular diagrams) were firmly established. These graphical techniques still remain the most widely used methods in geochemistry today (Howarth and Turner, 1986 in press).

The rapid spread of the use of computers for statistical analysis in the early 1960's led to an upsurge in work

involving multivariate analysis. This, in turn, led to various proposals for representing multivariate data in only two dimensions. The formal numerical techniques of multivariate data analysis are too often designed to yield specific answers to rigidly defined questions, often requiring some specific assumptions about the data. In many situations, however, the techniques which are needed for an investigation of multivariate data are informal exploratory ones rather than formal confirmatory ones, such as tests of specific hypotheses. Graphical techniques are well suited to this purpose since they tend to be less formal and confining, and enable the analyst to gain insight into the structure of data without imposing restrictive constraints. Once the analyst has gained some insight into the structure and behaviour of the data a better, more rational choice may be made of the formal numerical techniques required to elucidate certain aspects of the data.

The qualities and values of charts and graphs as compared with textual and tabular forms of presentation have been summarized by Schmid (1954) as follows:

1) In comparison with other types of presentation, well designed charts are more effective in creating interest and in appealing to the attention of the reader.

2) Visual relationships, as portrayed by charts and graphs, are more clearly grasped and more easily

remembered.

3) The use of charts and graphs saves time, since the essential meaning of large masses of statistical data can be visualized at a glance.

4) Charts and graphs can provide a comprehensive picture of a problem that makes possible a more complete and better balanced understanding than could be derived from tabular or textual forms of presentation.

5) Charts and graphs can bring out hidden facts and relationships and can stimulate, as well as aid, analytical thinking and investigation.

7.2 Displays for Multivariate Data

This chapter is concerned with techniques for the graphical representation of multivariate data. By comparison with data for one or two variables, multivariate data both benefit greatly from graphical displays and also present significant difficulties and pitfalls. The benefits accrue because there is often a great deal of information in multivariate data making inspection of a simple table of values a challenging task. The essential problem, however,

is that our most common plotting techniques, such as scatter plots and histograms are directly useful only for one or two variables. We must bear in mind that we are trying to use a fundamentally two dimensional plot to represent data which have intrinsically more than two dimensions. No single multivariate plot is likely to convey all the relevant information about a given data set; however, some will perform better than others. For effective data analysis it is necessary to integrate graphical displays with model fitting, data transformation and other analytical techniques. In addition, the suitability of specific methods for a given set of data depends on the number of observations, the number of variables and other properties of the data.

Chambers and Kleiner (1982) group multivariate plotting methods into two classes:

1) Multiple scatter plots; and

2) Symbolic plots.

In the first class the actual plots are two-variable scatter plots. A set of these is generated to represent higher dimensional data either directly or through derived variables. In the second class the data values are not used as coordinates in scatter plots, but as parameters to select or draw graphical symbols. The first class is not

discussed here, partly due to the drawbacks mentioned in Chambers and Kleiner (1982) and partly because the two techniques on which I want to concentrate are members of the second class.

It should be noted that all the techniques discussed below have been developed by statisticians and primarily used by data analysts in the fields of medicine, psychology, the biological and the social sciences. To my knowledge, apart from two papers, by Garrett (1983) and Howarth (1984), the techniques of graphically representing multivariate data have not been used in the geological sciences.

7.3 Symbolic Plots

The methods introduced below associate a p-parameter symbol with each p-dimensional observation (see Kleiner and Hartigan, 1981 for illustrations of some of these methods).

Let $X = (x_1, x_2, \ldots, x_p)$ be a p-dimensional vector to be described.

1) Profiles represent X by a bar chart with the i-th bar at a height determined by x_i. A common variation consists of using a polygonal line connecting the points whose coordinates are (i, x_i). This is undoubtedly the

most common method of representing multivariate data; its principal disadvantage is its dependence on the order of the variables. The lack of structure also makes it difficult to make comparisons between symbols for many variables.

2) Stars, also called Polygons, represent each variable as a value along p equally spaced rays issuing from the centre of a circle. This technique is similar to profiles except that values are arranged in polar coordinates (ϱ_i , θ_i) where

$$\varrho_i = x_i \text{ and}$$
$$\theta_i = 2\pi\,(i-1)/p, \text{ for } i = 1 \text{ to } p$$

rather than rectangular ones. The points on the rays are usually connected in a polygon. Unless the order of variables is carefully chosen, jagged and confusing pictures can occur for large numbers of variables.

3) Glyphs represent each variable as a ray emanating from a circle of fixed size, the length of the ray corresponding to the value of the variable. It is similar to a profile plot with the bars displaced at various angles. It is easy to compare the same variable in different glyphs using angle as a visual cue, but it is hard to compare different variables within the same

glyph. The arrangement of rays on the circle is arbitrary. The visual impression depends on the order of the variables and becomes increasingly confusing as the number of variables increases.

4) Function Plots (Andrews, 1972) represent each observation X by the finite Fourier series:

$$f(X) = x_1 / \sqrt{2} + x_2 \, Sin(t) + x_3 \, Cos(t) + x_4 \, Sin(2t) + x_5 \, Cos(2t) \ldots .$$

plotted over the range $-\pi \leq t \leq \pi$. For n observations there are n curves and they are best represented superimposed on one set of axes. Function plots are effective when p is large but become cluttered when there are many observations. They are very sensitive to the ordering chosen for the variables.

5) Faces (Chernoff, 1973) represent each observation in p-dimensional space, $p \leq 20$, by a cartoon of a face whose features, such as length of nose and curvature of mouth, correspond to each of the p variables. Thus every multivariate observation is visualized as a computer drawn face. Faces will be discussed in greater detail in Section 7.6.

6) Trees (Kleiner and Hartigan, 1981) represent each variable as the length of a branch of a stylized tree whose structure is determined by applying a hierarchical clustering algorithm to the variables. Trees are discussed in greater detail in Section 7.4.

With the exception of trees, in all the symbolic plots described above, the order of assigning variables to features in a symbol is arbitrary, yet the shape of the symbol and, to a large extent, the effectiveness of the whole display, can depend critically on the assignment. Replotting the display with a different assignment might emphasize different features and relationships in the data. Faces compound this disadvantage by treating the different variables in a highly asymmetric way.

7.4 Trees

To overcome the disadvantage of the arbitrary assignment of variables to symbol features, Kleiner and Hartigan (1981) suggested that a hierarchical clustering of the variables be used as an aid to assignment. This approach could be used in conjunction with many of the graphical techniques used to represent multidimensional data (Jacob,

1981). Kleiner and Hartigan developed trees as graphical displays specifically designed to incorporate the results of the clustering.

Assume that the input data consist of n observations with p variables or dimensions. All variables are assumed to be measured on a comparable scale - preliminary standardization may, therefore, be necessary. A hierarchical clustering of the variables is performed on a p x p similarity matrix (see Section 7.5). The result determines the basic structure of the trees for the individual observations, in that each tree has the same topology as the dendrogram (or cluster tree) obtained by the hierarchical clustering of the variables. The trees are drawn, on this basis, as follows.

1) Branch thickness is proportional to the number of variables "above" it. A variable is defined to be "above" a branch if the path from the leaf corresponding to that variable to the bottom of the tree passes through that branch. Therefore each leaf has unit width.

2) The angle between adjacent branches at a division is a linear function of the cluster strength between the variables above the two branches.

3) Branch direction - at each division it is necessary to decide which branch goes to the left and which to the

right. When there are many variables, simply copying the order directly from the dendrogram was found to lead to overlaps. The problem may be alleviated by using the following rules.

Define the stem of a tree as the path beginning at the bottom of the tree and following the thicker branch at each division until both branches are of the same thickness. The stem alternates direction at each division as it moves up the tree. At divisions not involving the stem, the thicker branches will move away from the stem. If two branches are of equal thickness, an arbitrary choice is made as to which branch goes to the left and which goes to the right. The resulting tree is, therefore, not completely independent of the original ordering of the variables.

4) The angle of a branch to the vertical is proportional to the thickness of the branch, except on the stem where this angle is inversely proportional to the thickness of the branch. In each case the sum of the two angles to the vertical is given by rule 2 (above). The reason for differentiating the stem from the other branches is to force the stem to remain nearly vertical, while it is the offshoots that are forced nearly vertical for other branches, as in real trees.

Rules 1 through 4 remain the same for all symbols

constructed from a given data set.

5) Each observation (or sample) is subsequently represented, using the template tree generated from rules 1 through 4, in such a way that the length of a branch is proportional to the average value of all variables "above" that branch. Thus limbs of the tree grow and shrink as corresponding clusters of variables grow and shrink in absolute value.

7.5 Clustering

The basis for the construction of the trees, and the reason why the technique is so innovative, lies in the use of a hierarchical clustering algorithm to aid assignment of variables to tree parameters.

Beginning with the p x p similarity matrix between the variables, hierarchical clustering methods merge variables or groups of variables which are most similar and thereby proceed from the initial stage of p variables to the final stage in which all variables are in a single group. The measure of similarity or closeness between one variable j and another, k, is commonly taken to be either the correlation coefficient r_{jk} (in which case a large correlation indicates similarity) or the Euclidean distance

d_{jk}

$$d_{jk} = \sqrt{\Sigma (X_{ij} - X_{ik})^2}$$

(in which case a small distance indicates similarity). The complete procedure is described by the dendrogram and hence the trees.

Kleiner's FORTRAN program TREES is used to generate the trees. It gives the user a choice of seven agglomerative hierarchical clustering methods. Differences between the methods arise because of the different ways of defining similarity between a variable and a group containing several variables, or between two groups of variables. The intergroup similarity measures used by these methods satisfy a recurrence formula for the similarity measure $d_{k(ij)}$ between a group k and a group (ij) formed by the merging of groups i and j. The formula is:

$$d_{k(ij)} = \alpha_i \, d_{ki} + \alpha_j \, d_{kj} + \beta \, d_{ij} + \gamma \, | \, d_{ki} - d_{kj} \, |$$

where d_{ij} is the distance between groups i and j and α, β and γ are parameters whose values for the seven methods are

Table 7.1

Recurrence Formula Parameters for various Agglomerative Hierarchical Clustering Methods

1) Nearest Neighbour– Single Linkage (minimum distance)
$\alpha_i = \alpha_j = 0.5 \ , \ \beta = 0 \ , \ \gamma = -0.5$

2) Furthest Neighbour– Complete Linkage (maximum distance)
$\alpha_i = \alpha_j = 0.5 \ , \ \beta = 0 \ , \ \gamma = 0.5$

3) Average Linkage (average distance)
$\alpha_i = \alpha_j = 0.5 \ , \ \beta = \gamma = 0$

4) Median Clustering
$\alpha_i = \alpha_j = 0.5 \ , \ \beta = -0.25 \ , \ \gamma = 0$

5) Weighted Average Linkage
$\alpha_i = n_i \ / \ (n_i + n_j) \ , \ \alpha_j = n_j \ / \ (n_i + n_j)$
$\beta = \gamma = 0$

6) Centroid Clustering
$\alpha_i = n_i \ / \ (n_i + n_j) \ , \ \alpha_j = n_j \ / \ (n_i + n_j)$
$\beta = \alpha_i \ \alpha_j \ , \ \gamma = 0$

7) Wards Method
$$\alpha_i = \frac{n_k + n_i}{n_k + n_i + n_j} \ , \ \alpha_j = \frac{n_k + n_j}{n_k + n_i + n_j}$$
$$\beta = \frac{-n_k}{n_k + n_i + n_j} \ , \ \gamma = 0$$

given in Table 7.1. For more details on the above discussion see Davis (1973) and Everitt (1978).

The dendrogram constitutes a summary of the information contained in the similarity matrix, since the process of amalgamating variables or groups of variables and treating them as a single new group is bound to introduce distortion. The distortion becomes increasingly apparent as sucessive levels of groups are merged. It therefore becomes necessary to find a measure of the distortion (or conversely the goodness-of-fit) of the dendrogram to the similarity matrix. It is possible to measure this distortion by comparing the original similarity matrix with the matrix of cophenetic values (either distances or correlations) generated from the dendrogram. The cophenetic distance (or correlation) between two variables in a dendrogram may be taken as the fusion level at which they both appear, for the first time, in the same cluster. A perfect dendrogram would therefore be one in which each value in the original similarity matrix is the same as the corresponding value in the cophenetic matrix. Thus a measure of goodness-of-fit (or conversely of distortion) of the dendrogram or tree would be the cophenetic correlation coefficient between elements of the two matrices. A good fit is revealed by a high cophenetic coefficient (say, $\geq$ 0.8), in which case one may conclude that imposing a system of nested clusters on the data has not caused undue distortion. If the correlation is much lower, say about 0.6

or 0.7, then the assumption that the data consist of a system of nested clusters is questionable and there is the possibility that the dendrogram, and therefore the trees, are misleading in their topology. It is, however, true to say that distortion of the dendrogram will tend to increase proportionally with the number of variables to be clustered. As a result the cophenetic correlation coefficients will tend to be higher for lower dimensional data and lower for higher dimensional data, and so the above correlation values should only be regarded as rough guidelines. As an example, the cophenetic correlation coefficients for the 6 variable SIND data range from 0.93 to 0.97 while the 24 variable MB data set has coefficients ranging from 0.55 to 0.68, depending on the similarity matrix and the cluster algorithm used.

7.6 Faces

Chernoff (1973) describes a method for graphically representing multivariate data in which each object vector is represented by a cartoon of a face, the features of which are governed by the values taken by particular variables. A full data matrix is thus represented by a collection of such faces which may then be examined in a bid to identify clusters, outliers, etc.

The rationale behind the development of the faces technique lies, according to Chernoff, in the ease with which the human mind is able to grasp the essential regularities and irregularities present in the data as represented. Chernoff postulates that the representation of data in terms of faces may be more useful than others, since people are used to studying and reacting to faces and, hopefully, they are able to filter out insignificant visual phenomena and focus on the potentially important.

An inherent characteristic of the faces method is that some features of the face have a greater distinguishing effect to the analyst than do other features (McDonald and Ayers, 1978, and Huff and Black, 1978). This characteristic may be advantageously exploited if it is desired to highlight or emphasise certain aspects of the data. However, this characteristic may also serve to weight disproportionately a subset of the variables in a visual clustering of the objects. For example, the facial outline may play a more crucial role in clustering than do the individual eye features. Perhaps a skilled and experienced analyst may be able partially to overcome these difficulties. Bruckner (1978) considers that it is advantageous to link facial characteristics with the physical meaning of the variables. For instance, the smile can be used to represent a "success/failure" variable, the eyes can represent "slyness" or a political stance, the forehead may represent intelligence. This approach makes a

virtue out of the unsymmetric relationship between the data and the facial features. However if, as in geochemistry, the variables are measured in the same units, this unsymmetrical relationship makes it impossible to draw quantitative or qualitative comparisons between variables within a face (something it is possible to do with trees). Indeed it is only if one knows the physical characteristics of each facial feature which correspond to the maximum and minimum data values for each variable that the analyst may compare one variable in different faces. This implies that the only possible use of faces is in cluster recognition and as we shall see even in this capacity they may prove misleading.

It is generally considered (Chernoff and Rizvi, 1975; Everitt,1978 and Wang, 1978) that the appearance of a face can vary markedly depending on the assignment of variables to facial features. Chernoff and Rizvi (1975) posed the question as to how much an analysis using faces would be affected by a permutation of the features representing each variable. They designed an experiment to evaluate the conjecture that the observer's reaction to the faces will quickly determine the relevant information with negligible dependence on the choice of features. Their main conclusion was that random permutations of the features to be used tend to affect the error rate in a classification task by a factor of about 25%. The efficiency of the faces technique is therefore strongly dependent on the permutation of

variables to features. A variety of permutations may need to be tried before one can arrive at the best for a given data set. This implies that the analyst has some prior knowledge of the structure or relationships within the data set, an implication that is clearly illogical considering that the aims of the technique are to elucidate those very features.

7.7 Trees, Faces and the SIND Data

7.7.1 SIND Trees

The preliminary cluster analysis may be carried out on a correlation matrix or a distance matrix. If A and B are two standardized variable vectors (i.e. their lengths are both equal to (p-1) where p is the number of variables in the total data set) in n dimensional object space (where n is the number of observations) then the correlation coefficient between A and B is $r_{AB} = \cos \theta_{AB}$ where θ_{AB} is the angle between the two vectors. The Euclidean distance $d_{AB} = 2 (p-1) \sin (\theta_{AB} /2)$ which geometrically is the length of the line connecting the ends of the two vectors.

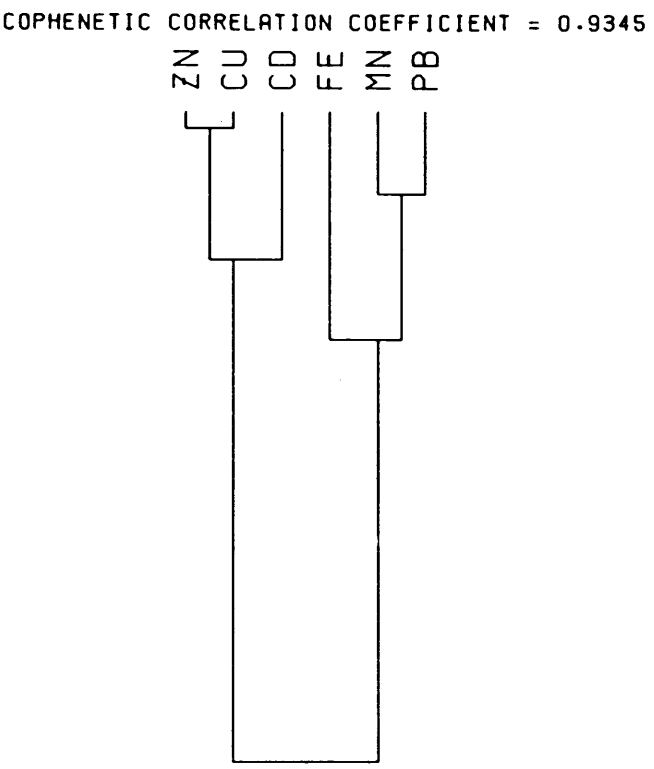
The choice of similarity matrix has no effect on

dendrogram topology but does affect the efficiency of the clustering algorithm as measured by the cophenetic correlation coefficient. The weighted average algorithm works best for both the correlation and the distance matrix. The median clustering, Ward's method and centroid clustering algorithms are the worst. The choice of clustering algorithm has no effect on dendrogram topology. Since the choice of similarity matrix used has little effect on the appearance of the trees and since the correlation coefficient is a more conventional and easily understood measure of similarity than distance, the correlation matrix was used to obtain all TREES program results subsequently presented. Remember that these conclusions are based on a data set of only six variables. How they change with an increase in the number of variables will be illustrated in Section 7.8.

To analyse the changes due to data transformation and to robust estimation of the correlation matrix, six separate runs of the program TREES were carried out. The three data sets (raw, logten and boxcox) were used with both non-robust and robust correlation matrices (see Figures 7.1 through 7.30).

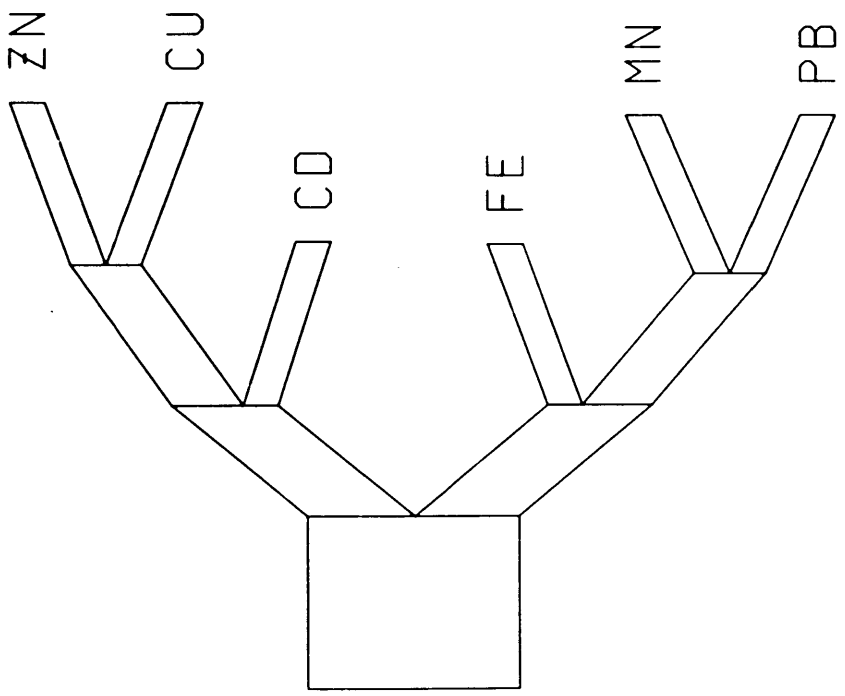
Transformation of the data causes a reduction in the cophenetic correlation coefficient indicating that there is more distortion inherent in the generation of the dendrogram. Boxcox transformation results in a lower cophenetic coefficient than logten transformation. Robust

Figure 7.1



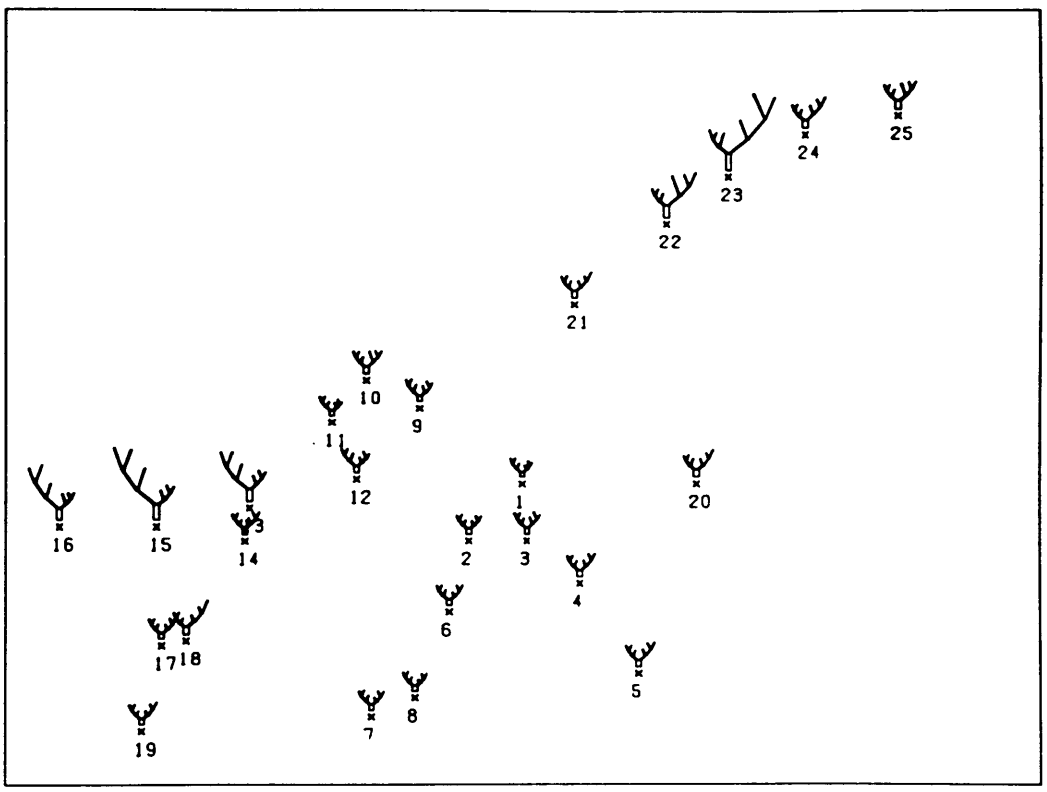
SIND RAW DATA (RAW)

Figure 7.2



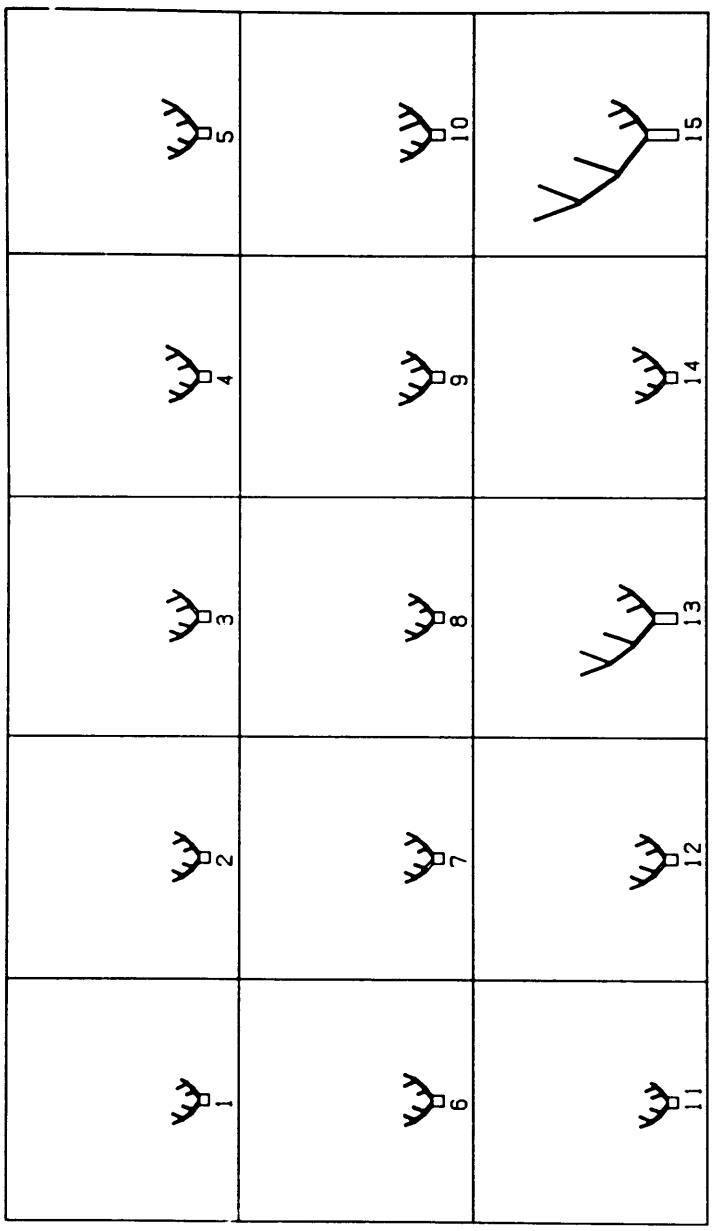
SIND RAW DATA (RAW)

Figure 7.3



SIND RAW DATA (RAW)

Figure 7.4



SIND RAW DATA (RAW)

Figure 7.5

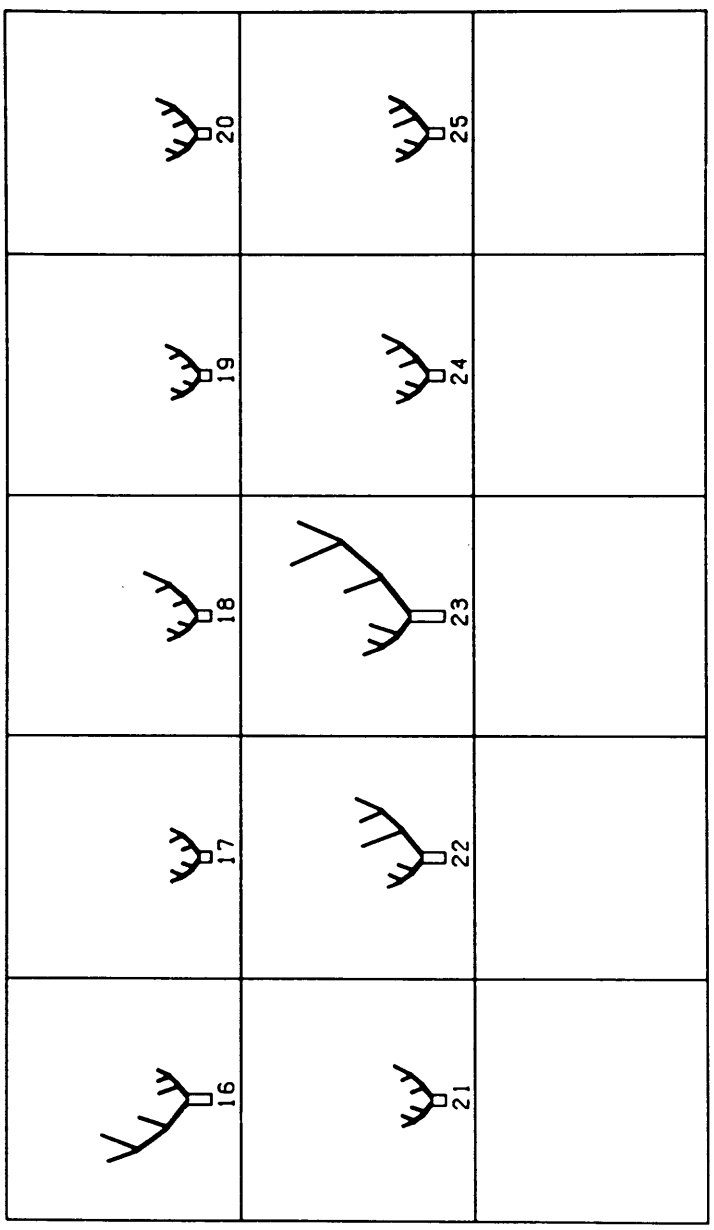
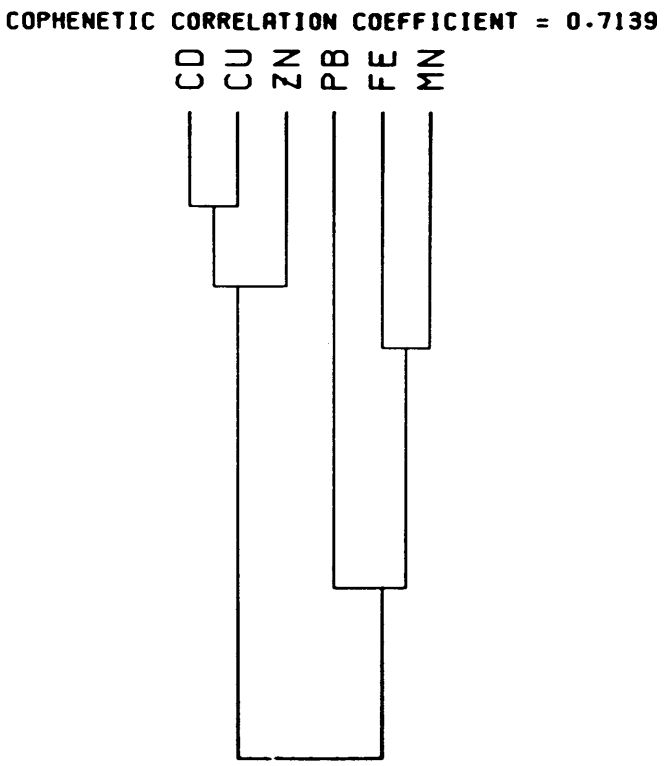
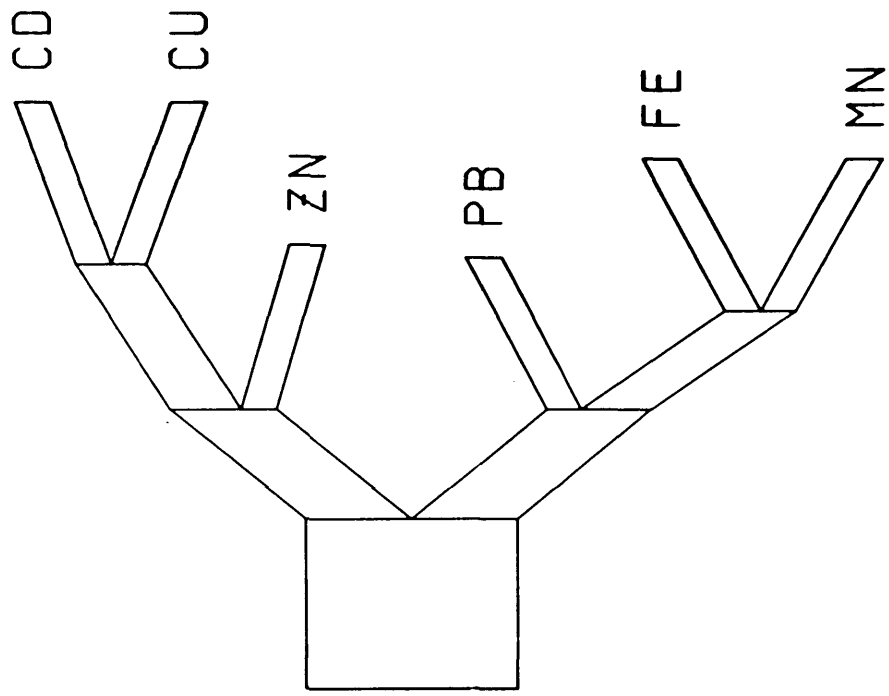


Figure 7.6



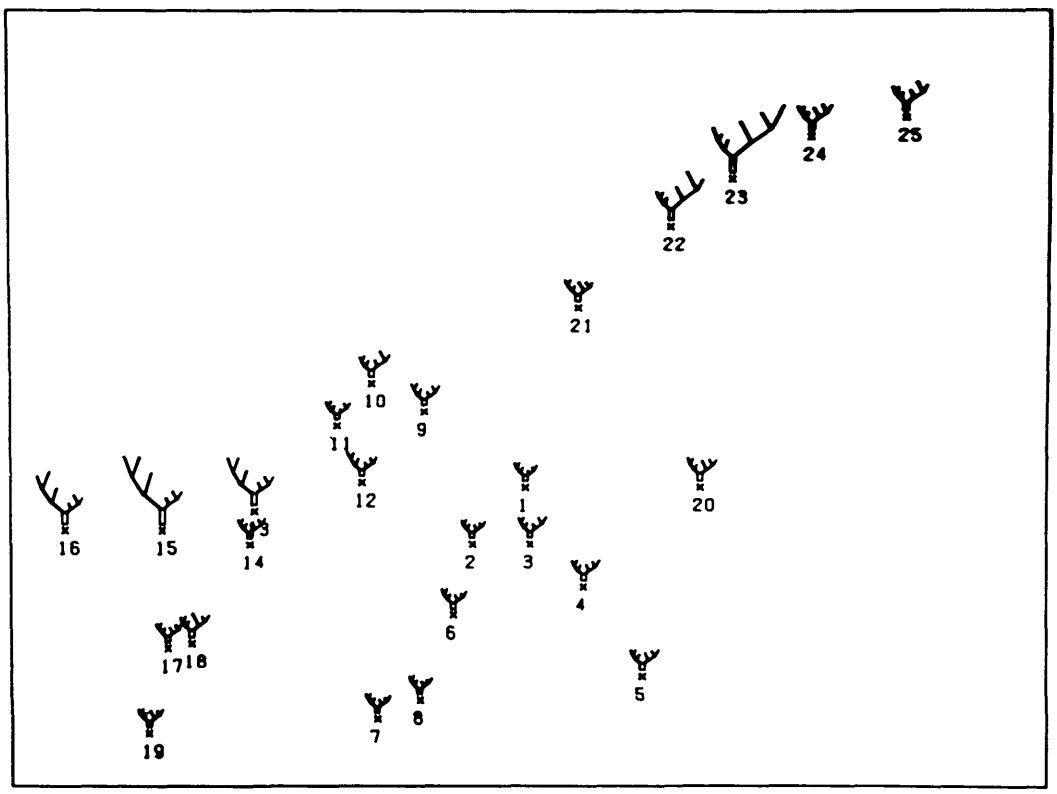
SIND RAW DATA (ROBUST)

Figure 7.7



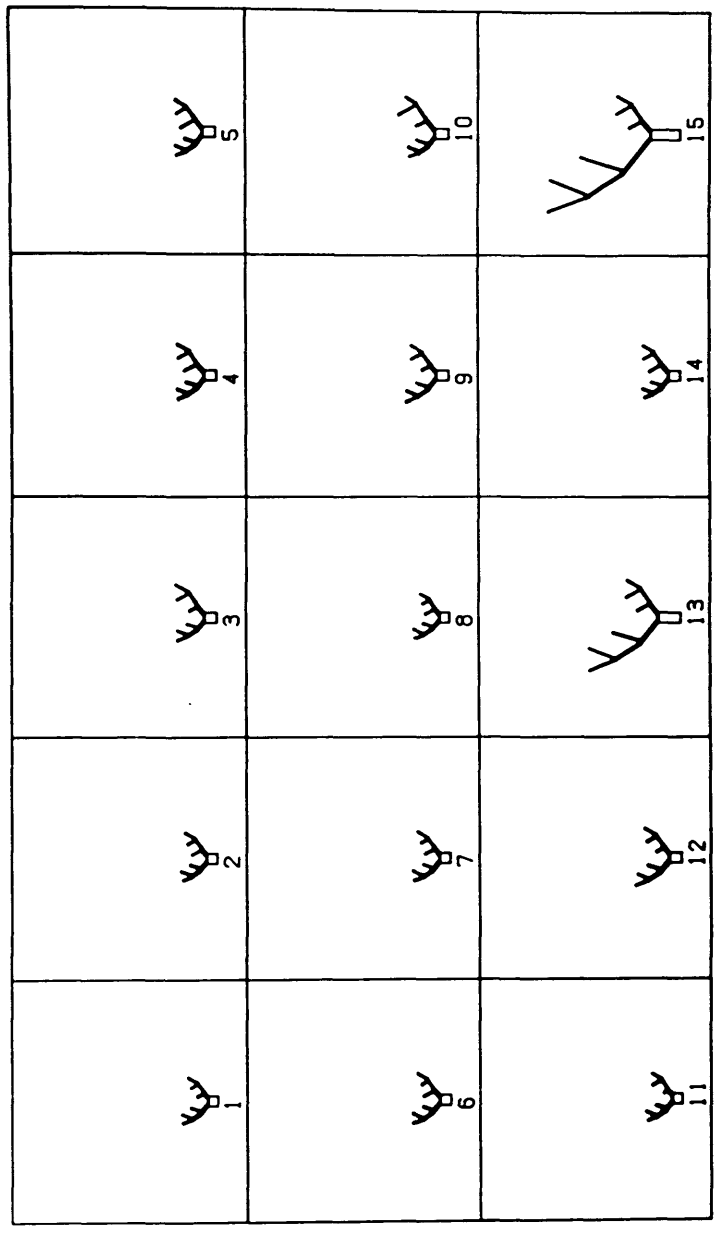
SIND RAW DATA (ROBUST)

Figure 7.8



SIND RAW DATA (ROBUST)

Figure 7.9



SIND RAW DATA (ROBUST)

Figure 7.10

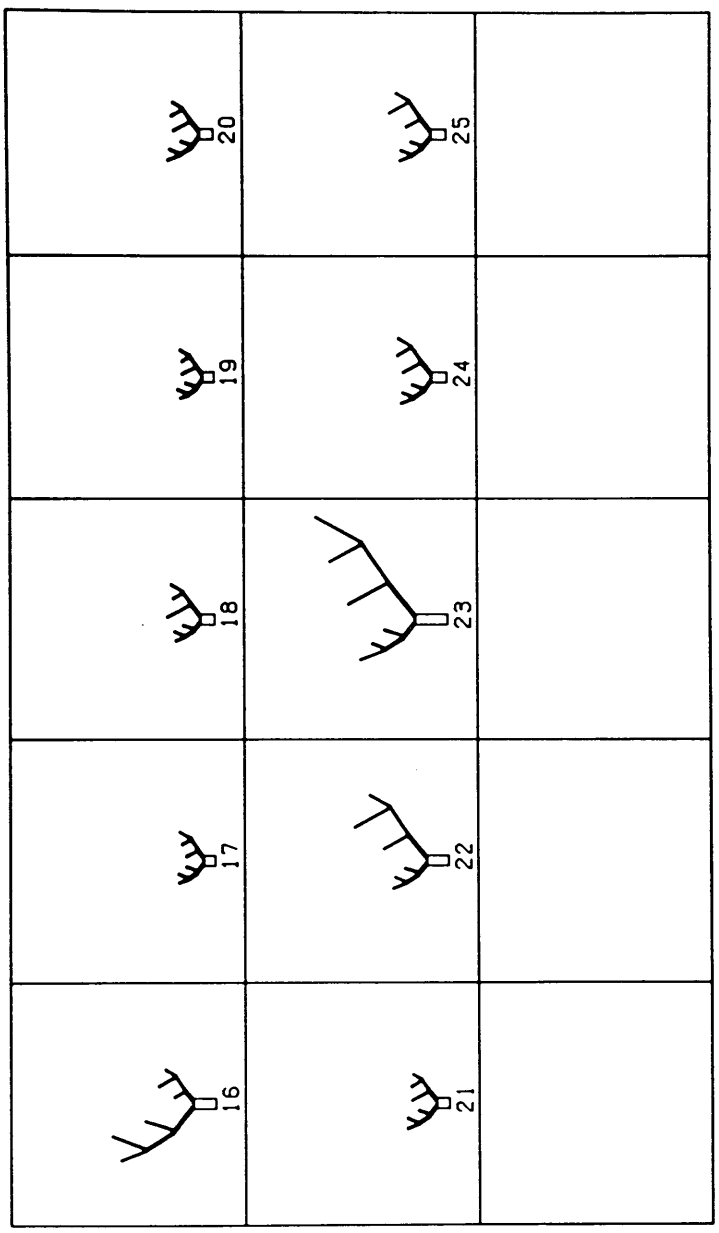
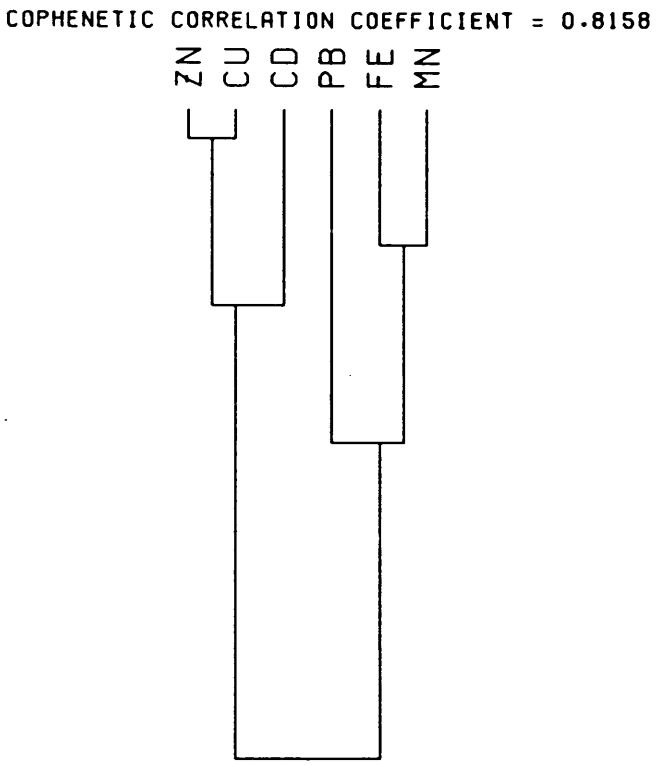
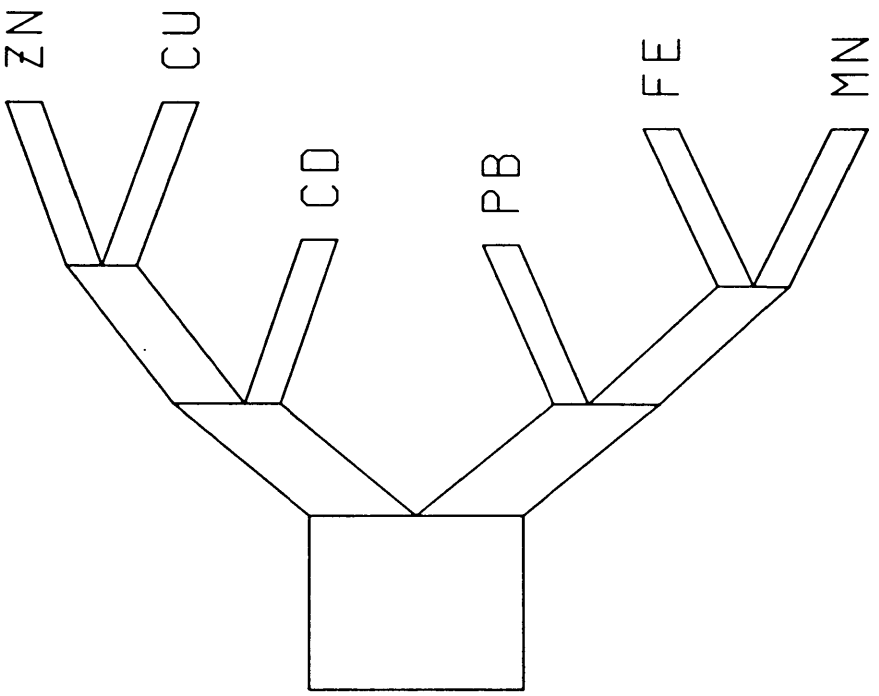


Figure 7.11



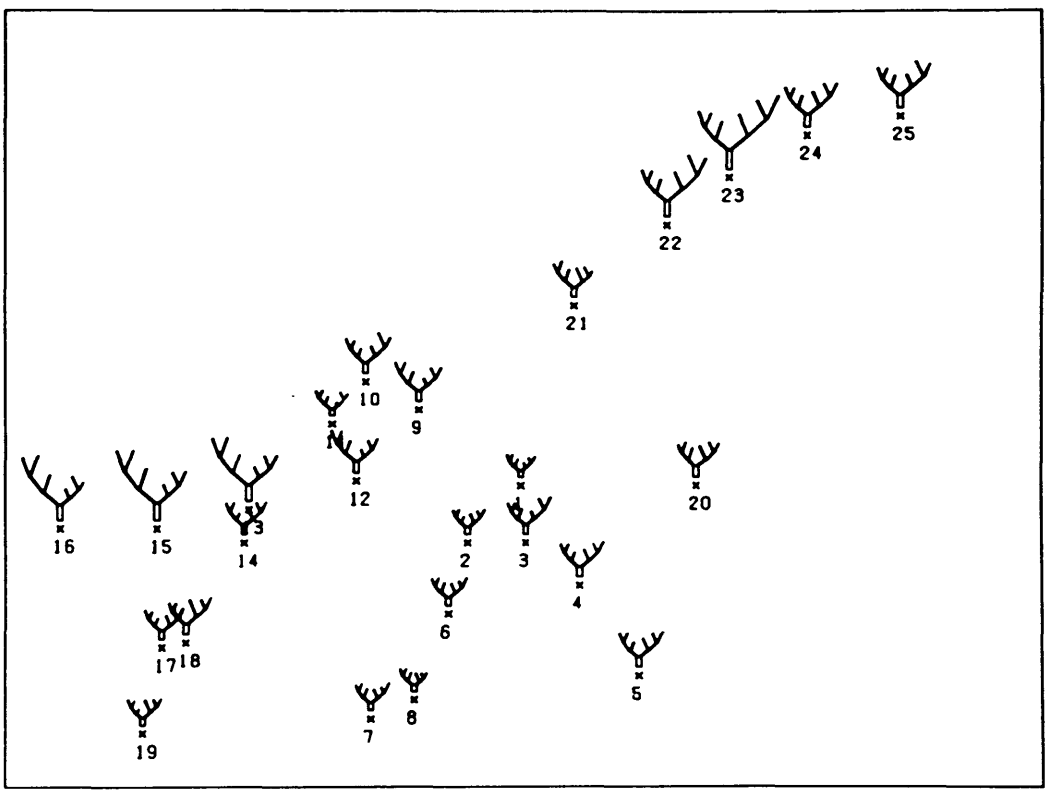
SIND LOGTEN DATA (LOGTEN)

Figure 7.12



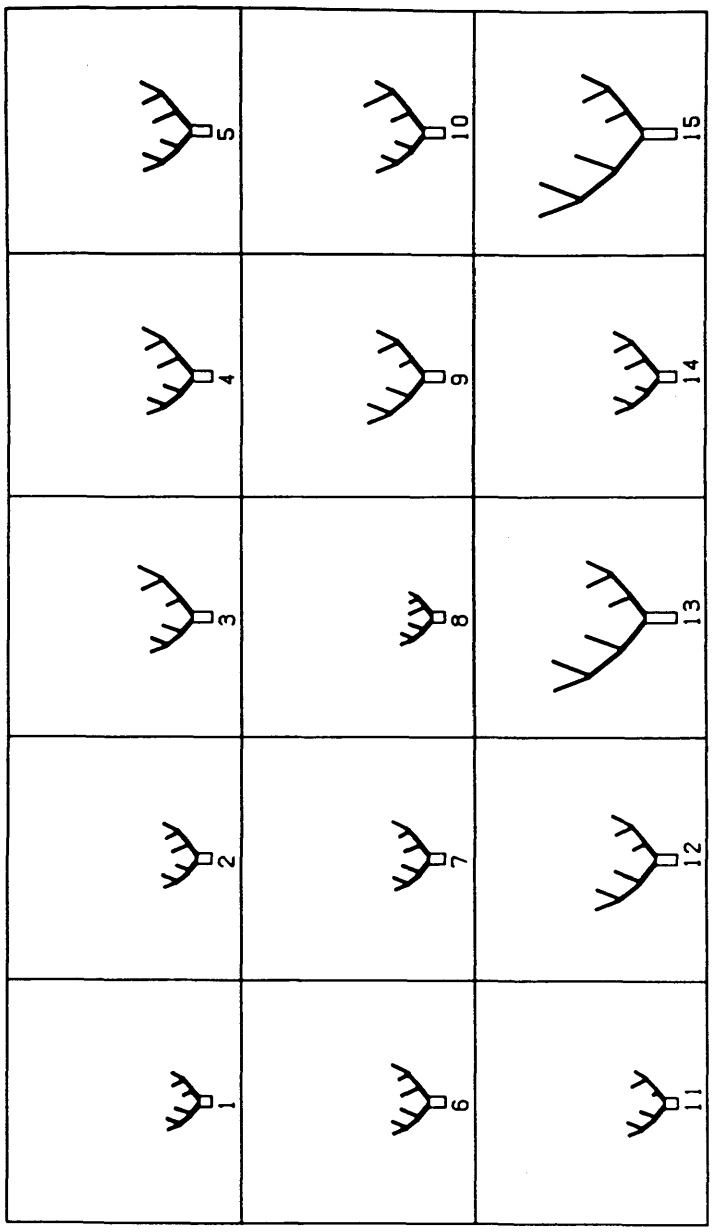
SIND LOGTEN DATA (LOGTEN)

Figure 7.13



SIND LOGTEN DATA (LOGTEN)

Figure 7.14



SIND LOGTEN DATA (LOGTEN)

Figure 7.15

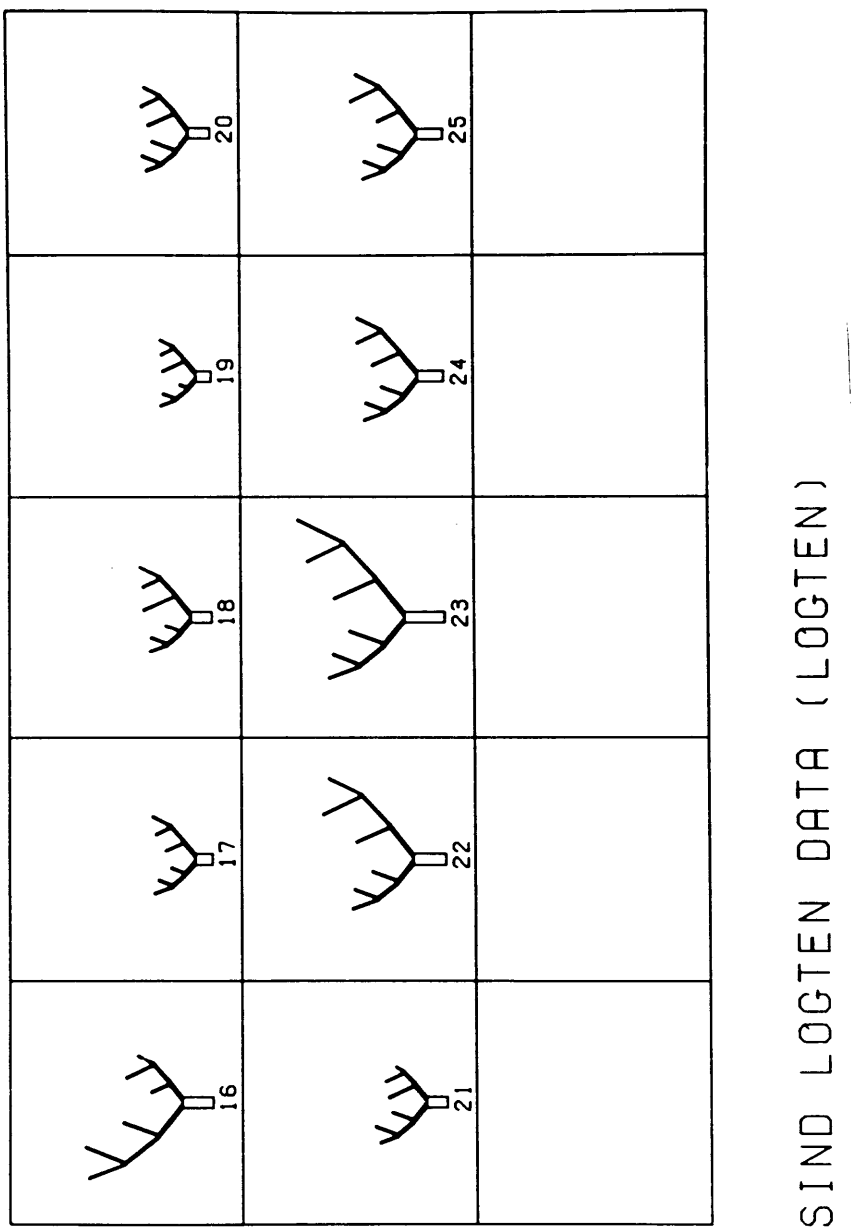
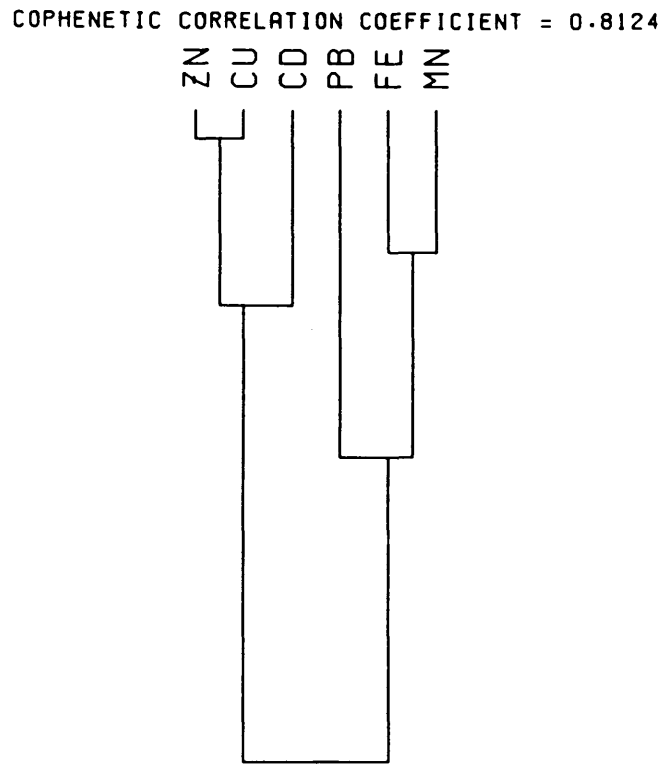
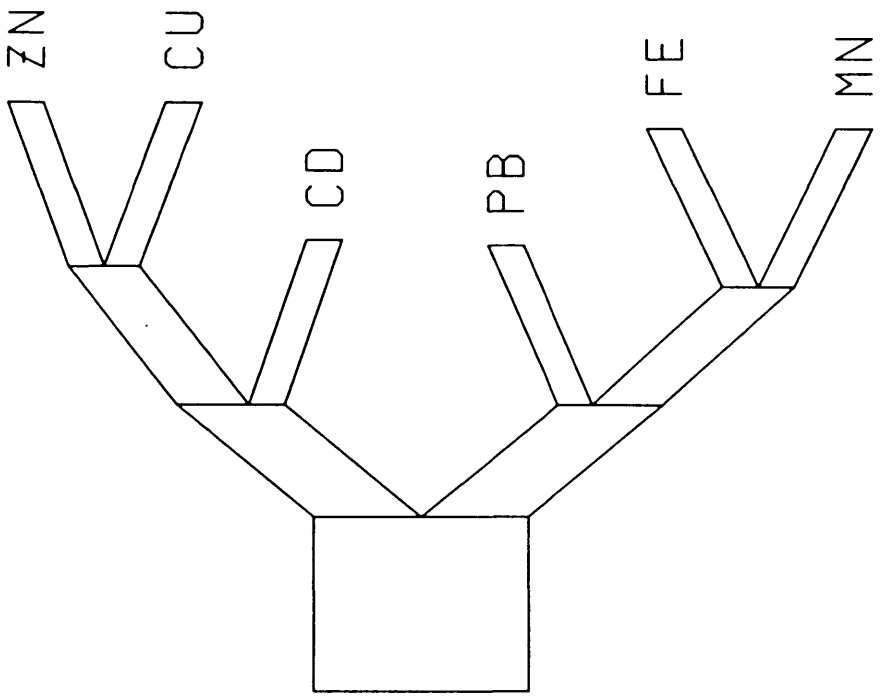


Figure 7.16



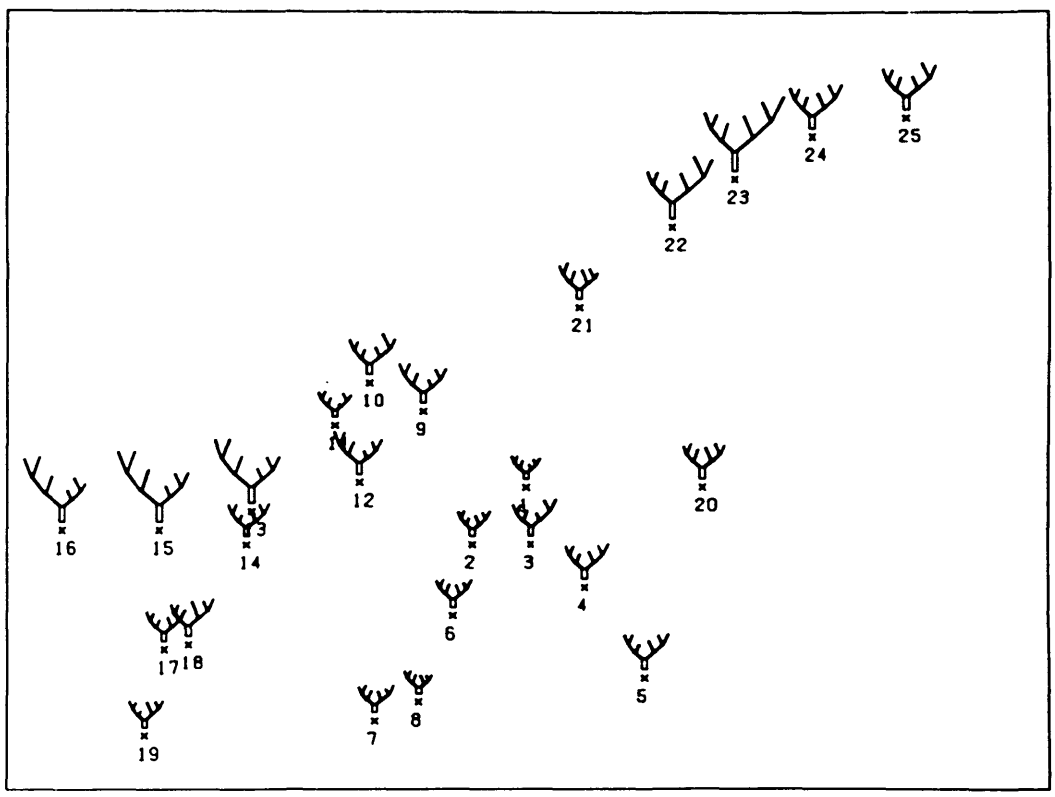
SIND LOGTEN DATA (ROBUST)

Figure 7.17



SIND LOGTEN DATA (ROBUST)

Figure 7.18



SIND LOGTEN DATA (ROBUST)

Figure 7.19

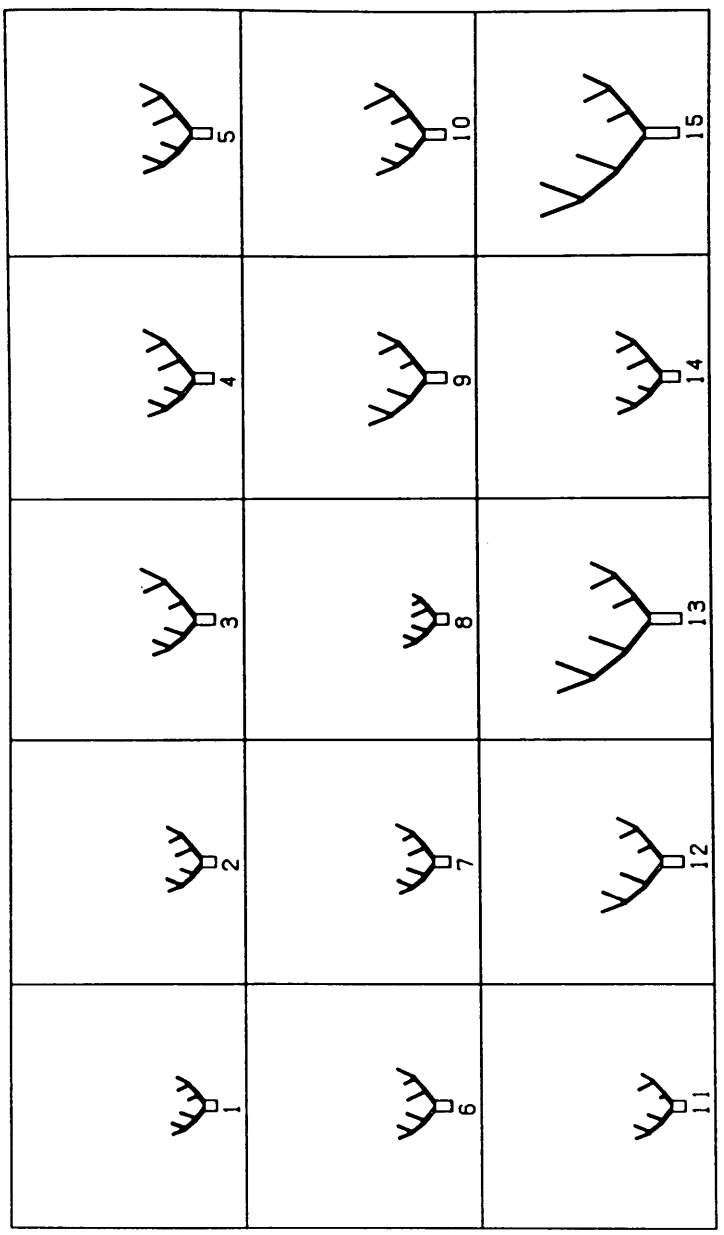


Figure 7.20

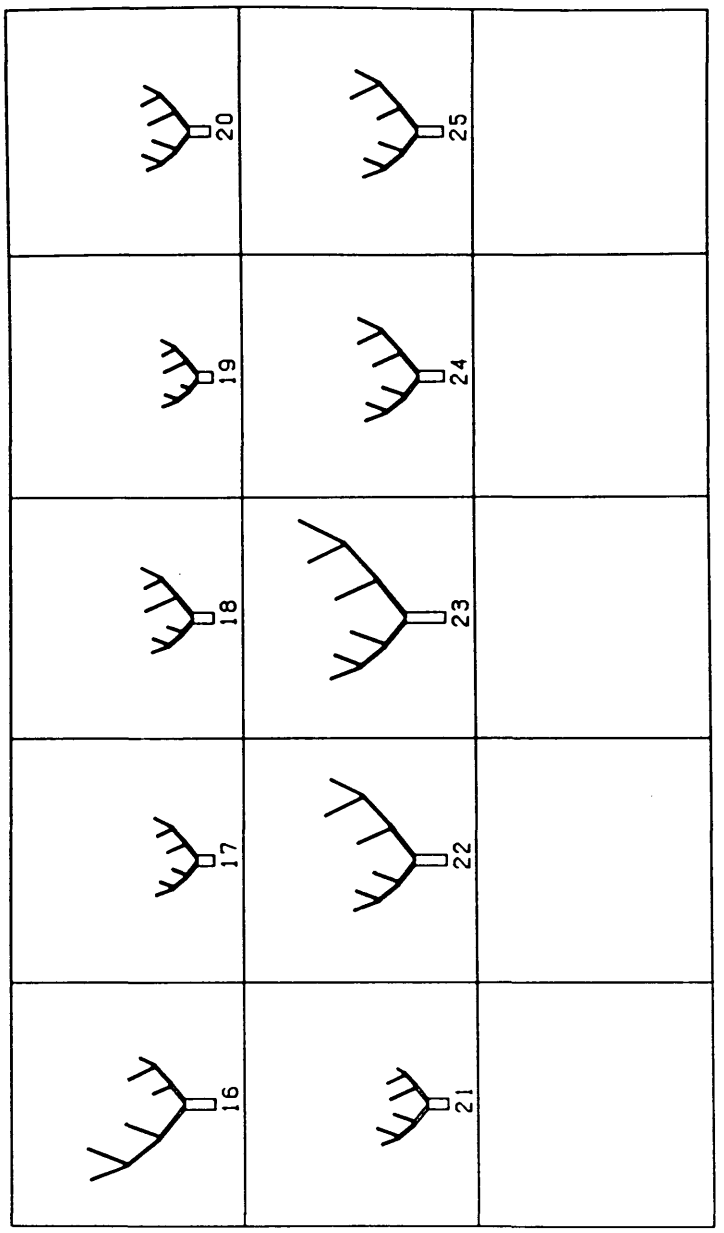
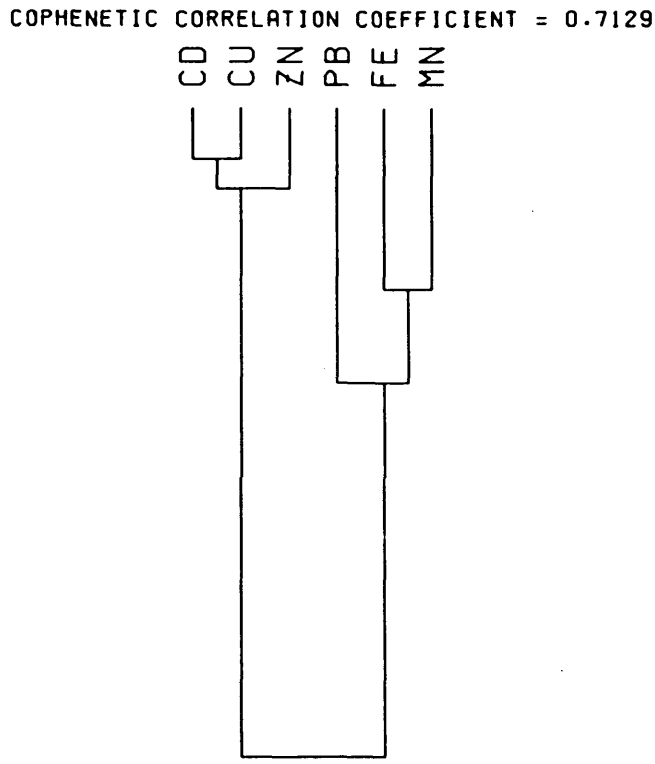
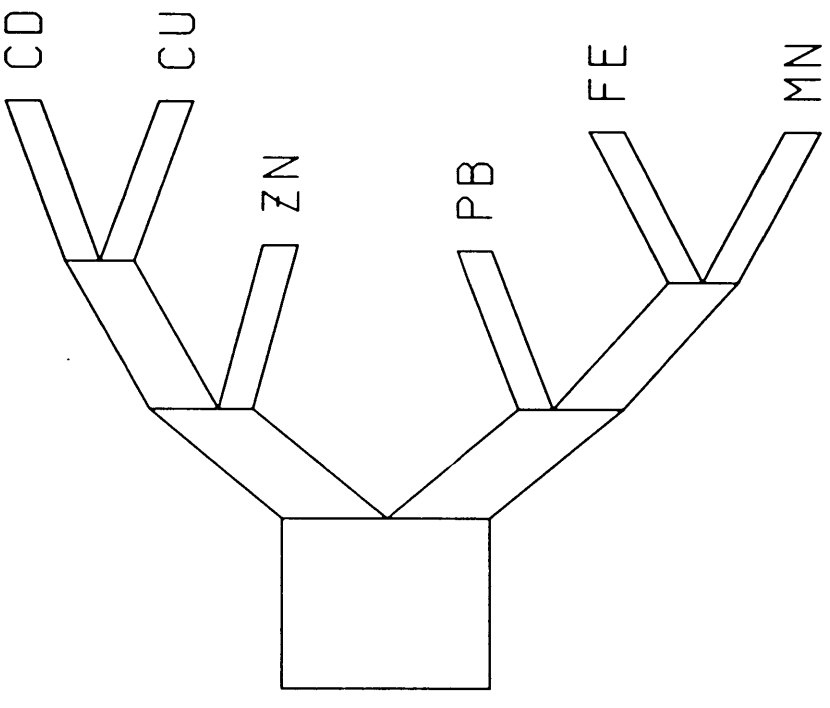


Figure 7.21



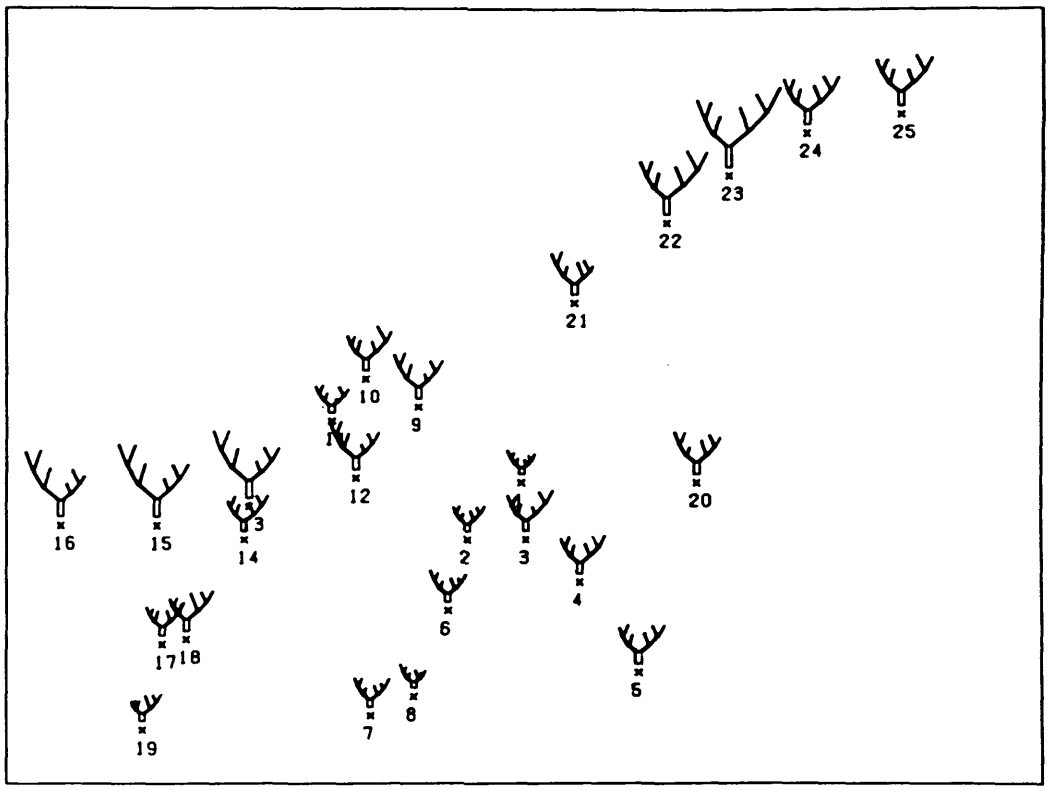
SIND BOXCOX DATA (BOXCOX)

Figure 7.22



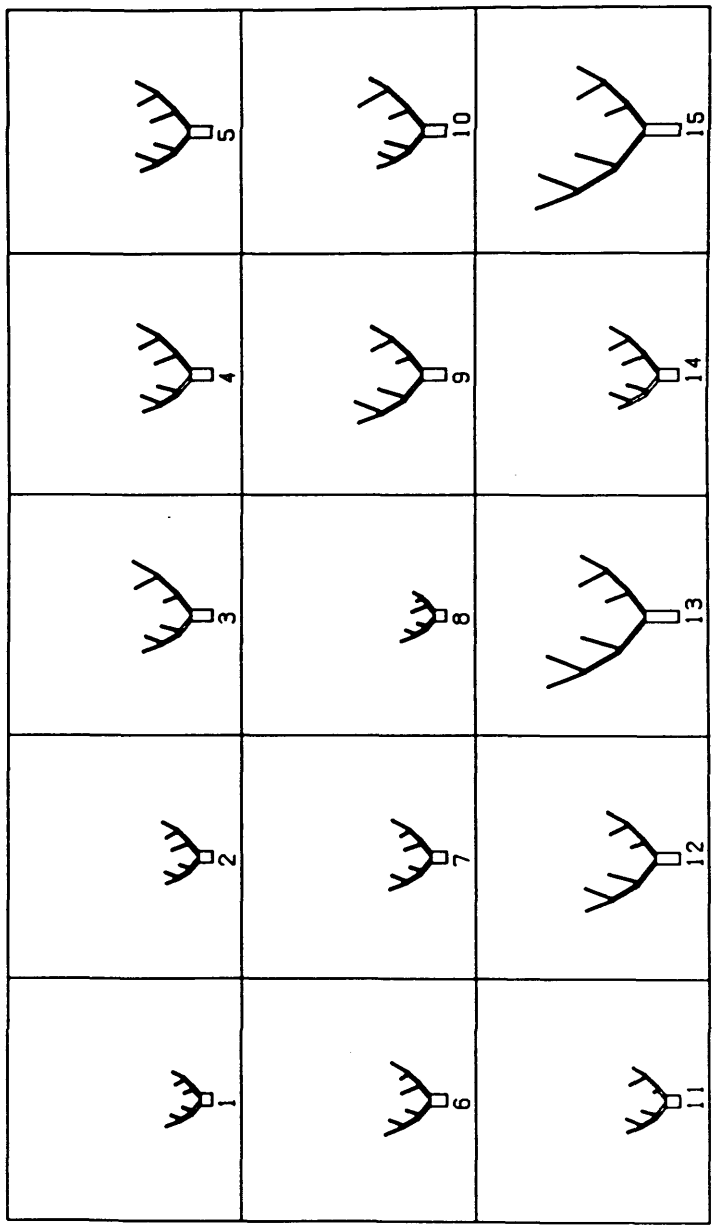
SIND BOXCOX DATA (BOXCOX)

Figure 7.23



SIND BOXCOX DATA (BOXCOX)

Figure 7.24



SIND BOXCOX DATA (BOXCOX)

Figure 7.25

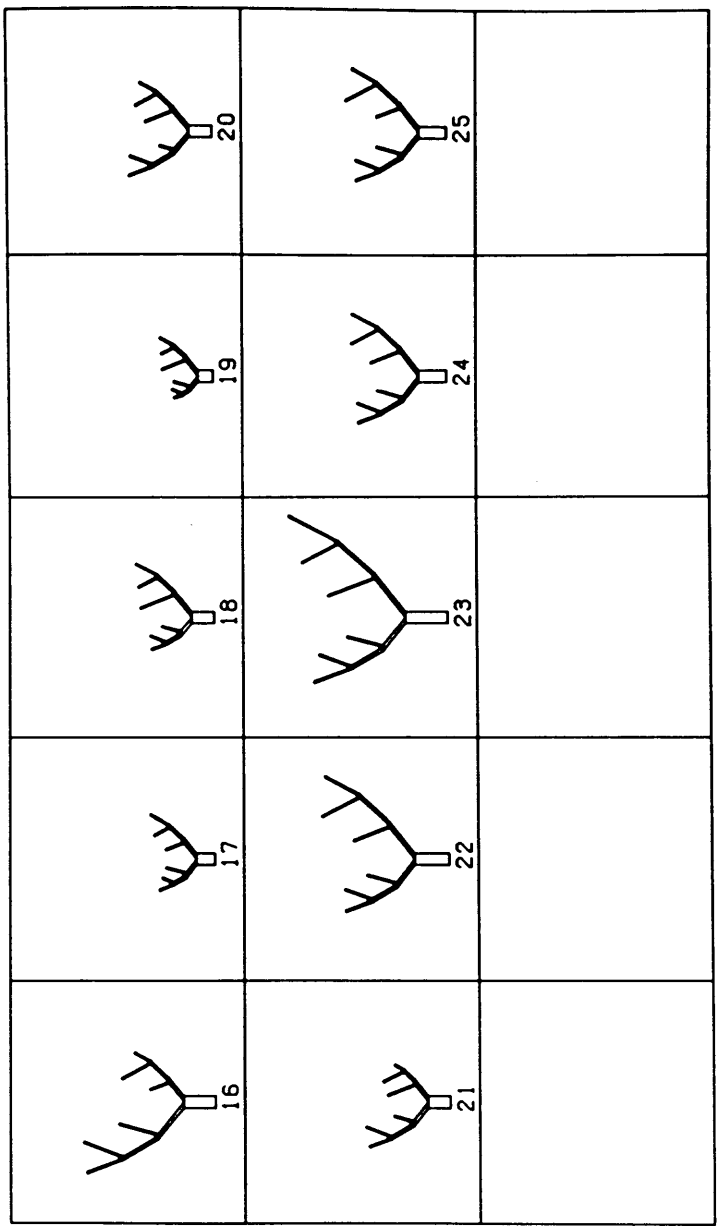
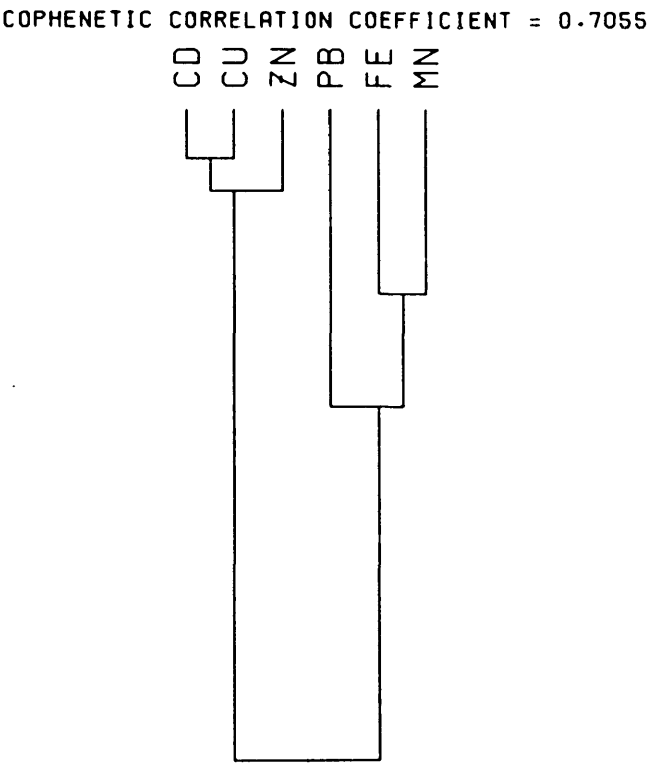
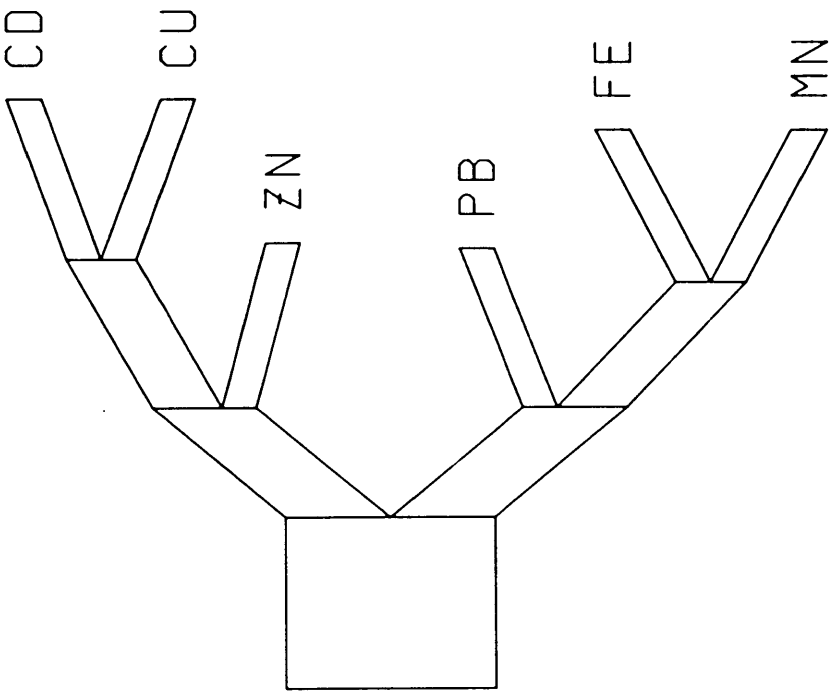


Figure 7.26



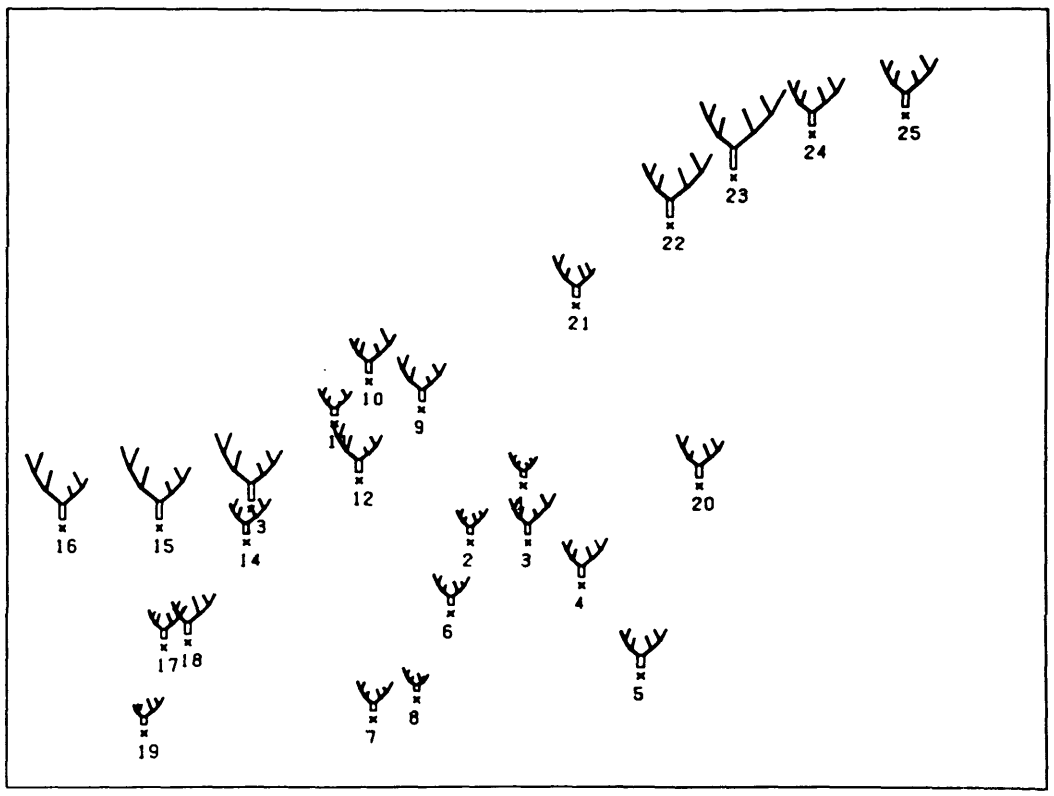
SIND BOXCOX DATA (ROBUST)

Figure 7.27



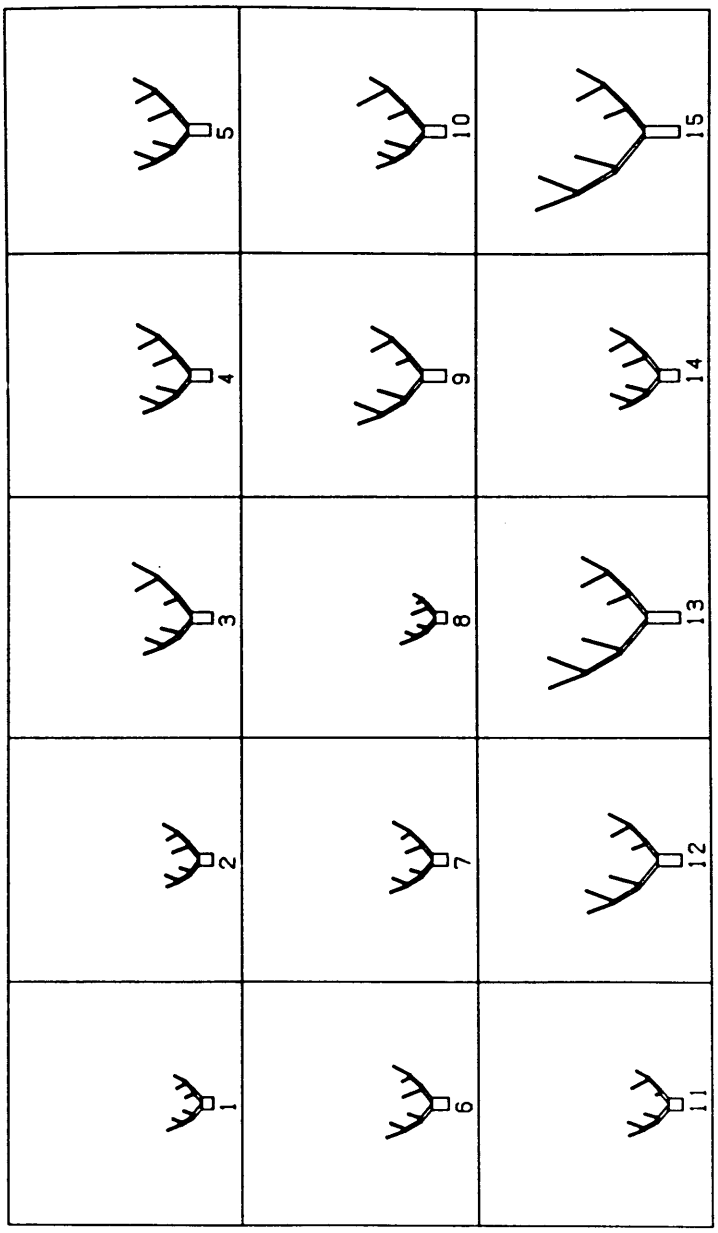
SIND BOXCOX DATA (ROBUST)

Figure 7.28



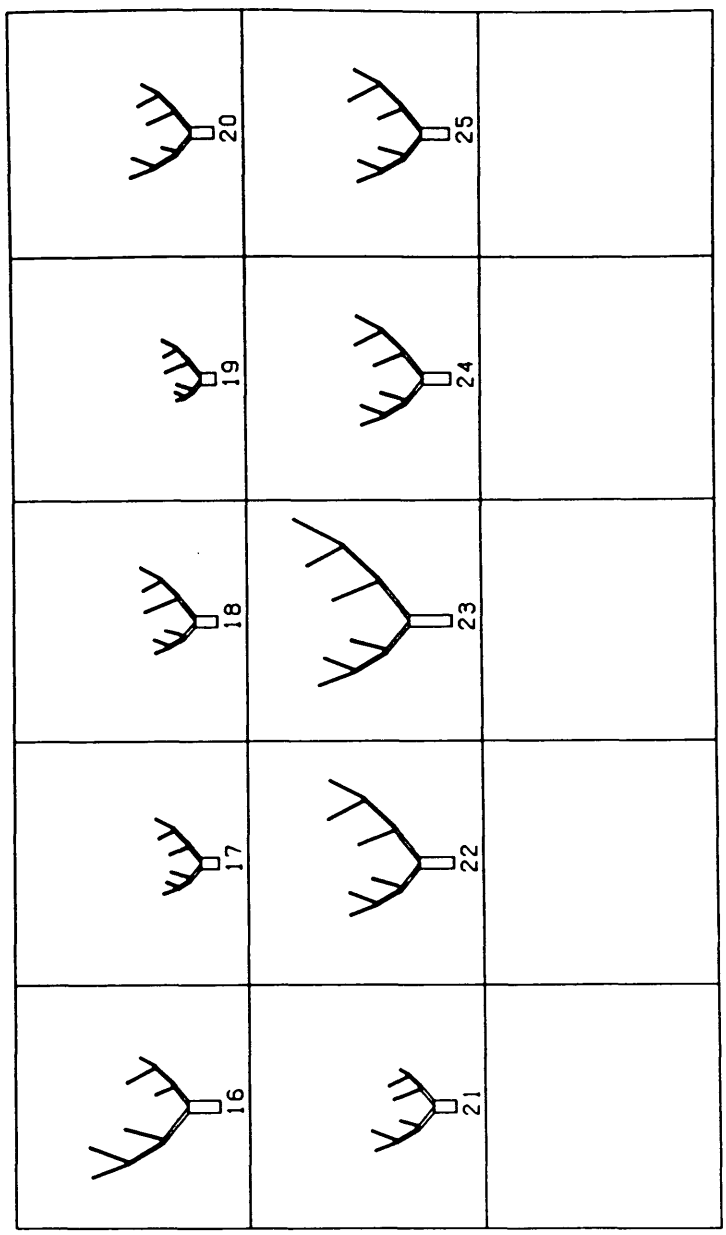
SIND BOXCOX DATA (ROBUST)

Figure 7.29



SIND BOXCOX DATA (ROBUST)

Figure 7.30



estimation of the correlation matrices tends to cause a large decrease in the cophenetic coefficient for raw data, but only slight decreases in those for logten and boxcox data. This may reflect the large changes seen between raw non-robust and robust correlation matrices and the small changes seen between logten and boxcox non-robust and robust matrices (see Chapter 6). Transformation also causes a change in the topology of the dendrogram. The raw non-robust correlation matrix gives a dendrogram with Zn and Cu-plus-Cd on one side and Mn and Pb, together with Fe, on the other. This reflects the two geochemical processes operative in the SIND area. The raw robust correlation matrix gives the same topology as both the boxcox non-robust and robust correlation matrices, i.e. Cd and Cu-plus-Zn on one side with Fe and Mn-plus-Pb on the other. Both the logten non-robust and robust correlation matrices show the same topology, i.e. Zn and Cu-plus-Cd on one side with Fe and Mn-plus-Pb on the other.

The choice of non-robust or robust correlation matrix and the cluster algorithm used affects only the topology of the dendrogram and hence the tree morphology. For the SIND data, only slight changes in the order of the variables on the tree have been caused by either transformation of the data, or the use of robust estimation of the correlation matrix. Compare the TREES results obtained with the raw data with a non-robust correlation matrix and that of raw data with a robust matrix. The slight changes to the

morphology of the trees caused by the use of robust estimation has little visual impact. All the major features of the raw data are revealed. Sample sites 13, 15 and 16 have trees in which the left hand branches are enhanced - this is due to high concentrations of Zn, Cu and Cd. Samples 22 and 23 have trees in which the right hand branches are enhanced due to the high concentrations of Fe, Mn and Pb. All other samples have trees of very similar size representing a relatively uniform background population.

Looking at the equivalent figures for logten and boxcox data it is clear that the choice of transform has little visual effect on the tree morphology while the use of robust estimation has even less effect. Again the major visual impressions are the result of changes in the concentrations of the variables from sample to sample. Sample sites 13, 15, 16, 22 and 23 clearly have the largest trees, although by comparison with raw data they show far less contrast with background sample trees. The five anomalous trees show the same branch length enhancements as do the raw data trees but with far less contrast in comparison to the other branches. The remaining trees have a considerable size spread with sample sites 24, 25, 12, 9 and 10 not that far off sites 22 and 23 in size. The reasons for these observations are clear. Data transformations (see Chapter 5) cause enhancement of background spread and suppression of anomalous object

vectors by their movement in multivariate space towards the background population.

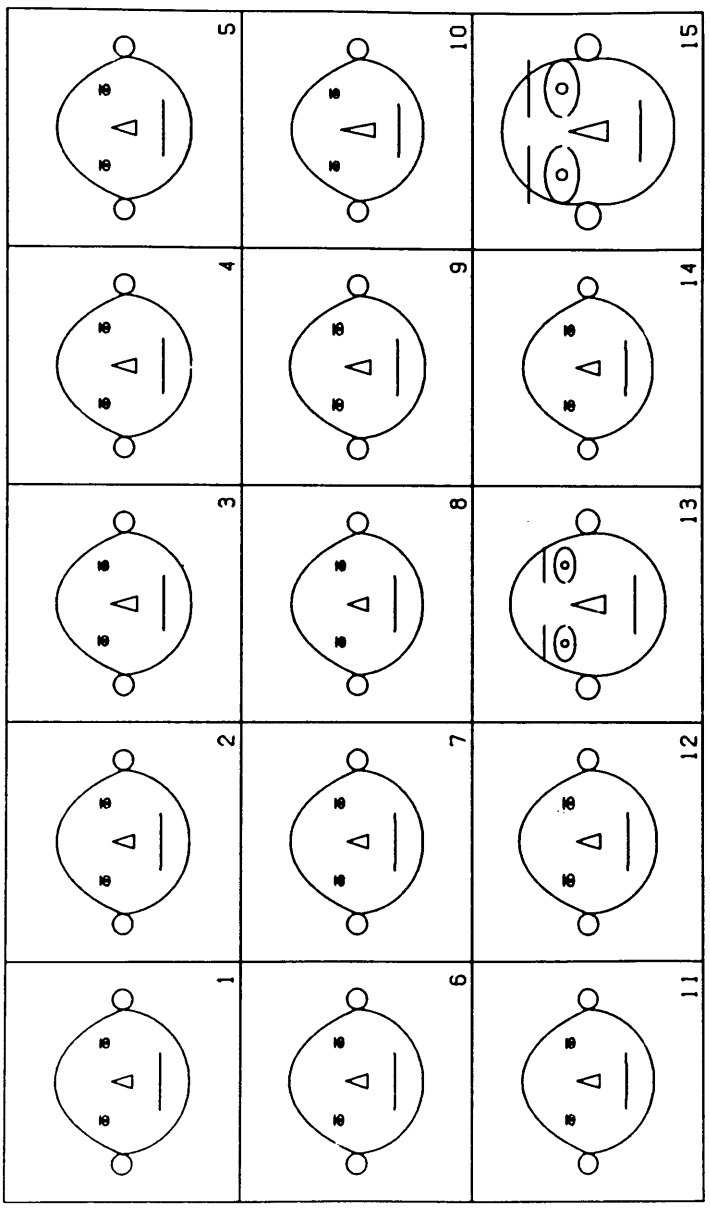
7.7.2 Chernoff Faces and the SIND Data

After much time consuming experimentation, a combination of facial features to variables was chosen which visually emphasised the true anomalies (sites 13, 15 and 16) and suppressed the false anomalies (sites 22 and 23) (see Figures 7.31 and 7.32). A combination was also found which placed the emphasis the other way (see Figures 7.33 and 7.34).

To emphasise the sample sites influenced by mineralization the variables Zn, Cd and Cu were assigned to facial features which have large visual impact e.g the size of the upper face ellipse, the half face height, the size of the eye. The variables Fe, Mn and Pb were assigned to facial features which have small visual impact e.g nose length, nose width, brow height. To place the emphasis on Fe, Mn and Pb and away from Zn, Cd and Cu the facial feature to variable assignments were reversed.

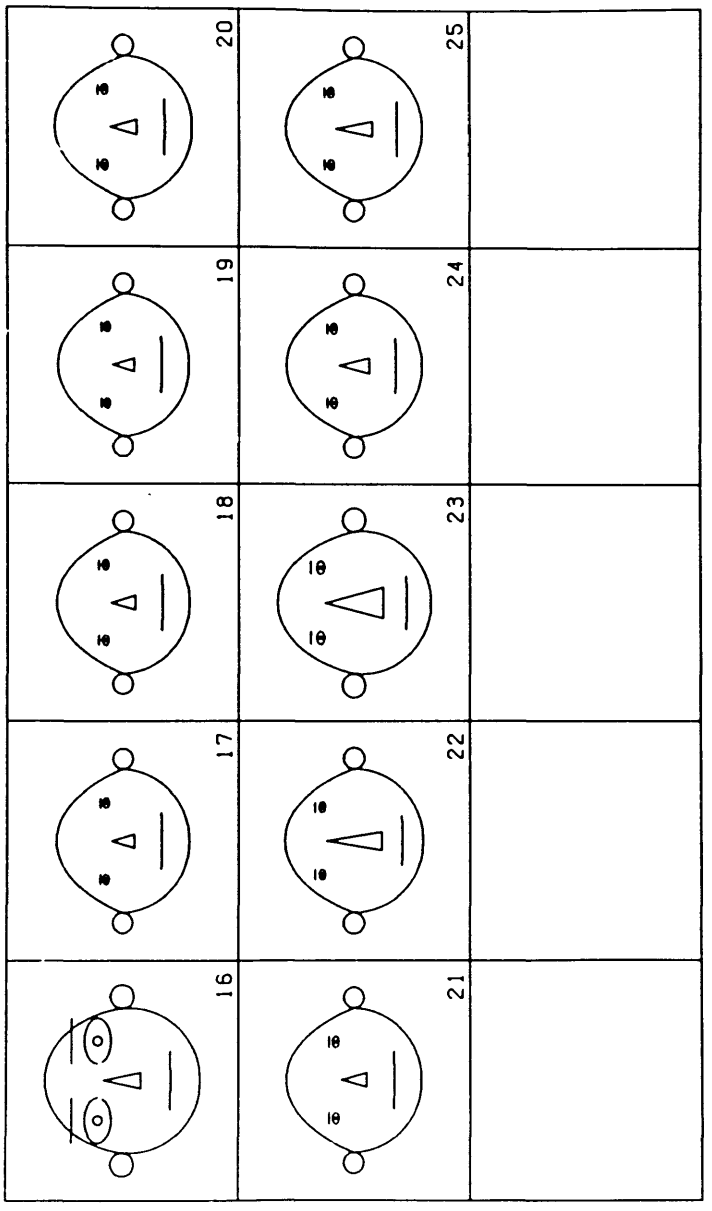
This process implies that the analyst can emphasise or suppress any variable, which in turn implies that deliberate distortion could occur. In order to place the emphasis on the true anomalies the analyst must have prior

Figure 7.31



CHERNOFF FACES SIND RAW DATA

Figure 7.32



CHERNOFF FACES SIND RAW DATA

Figure 7.33

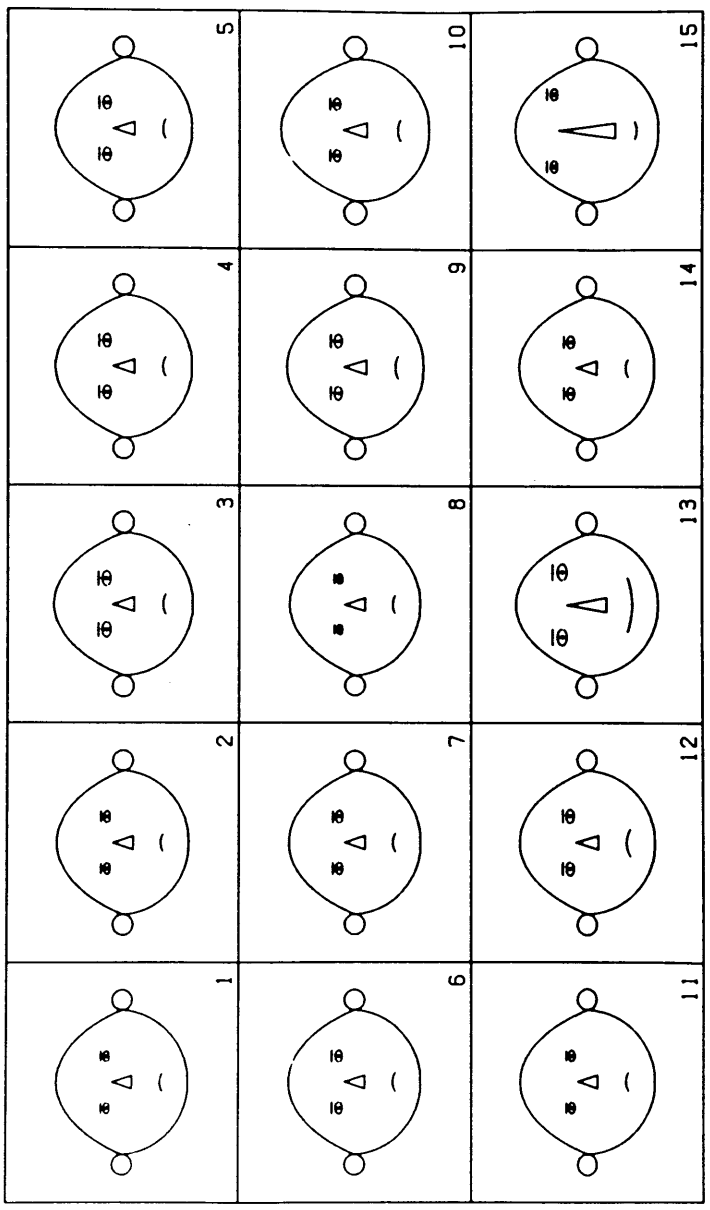


Figure 7.34

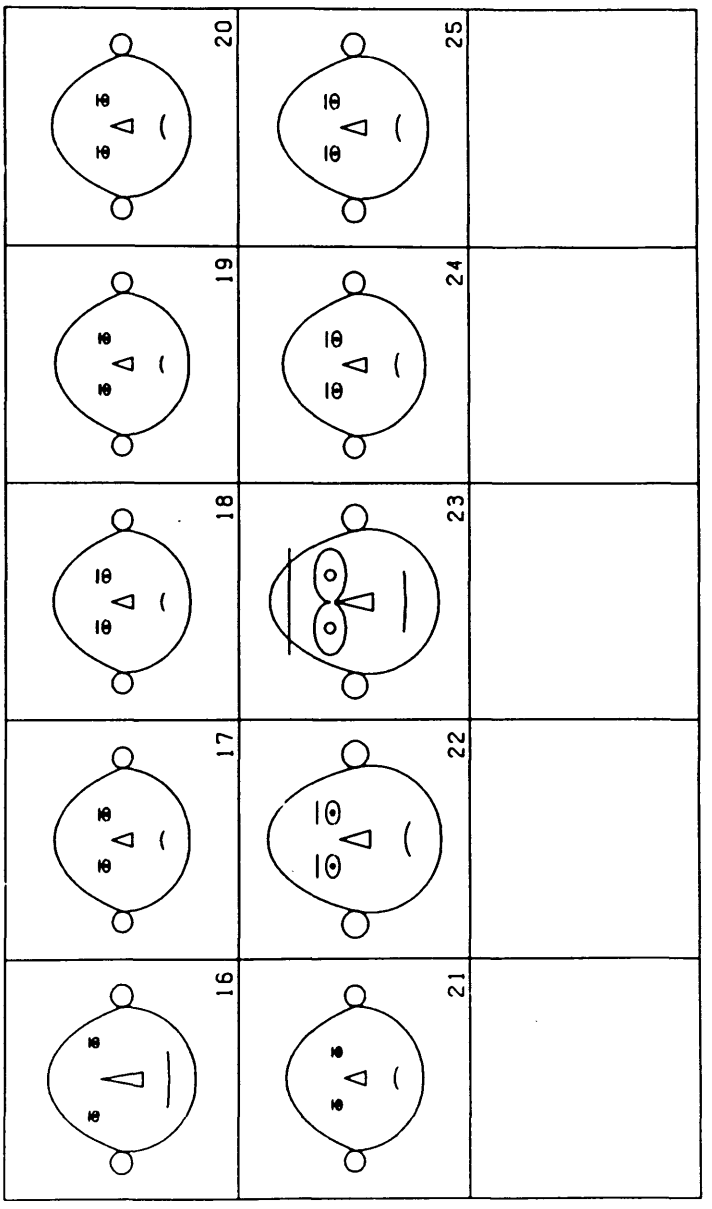
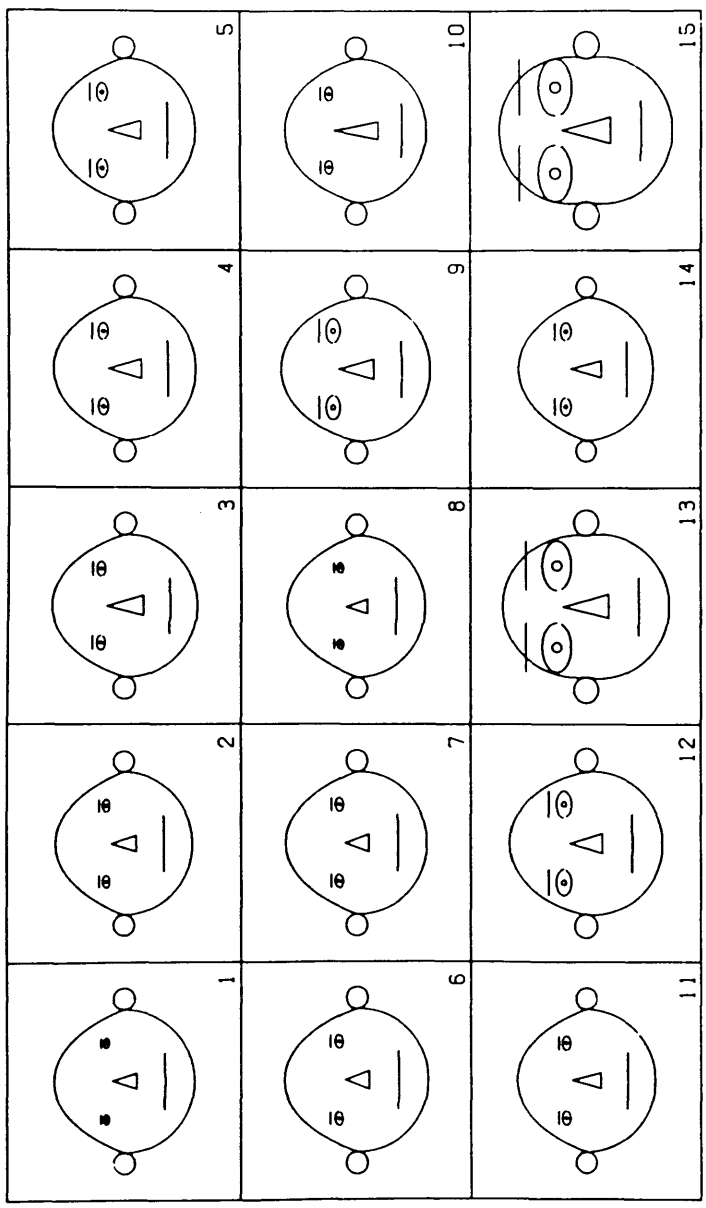
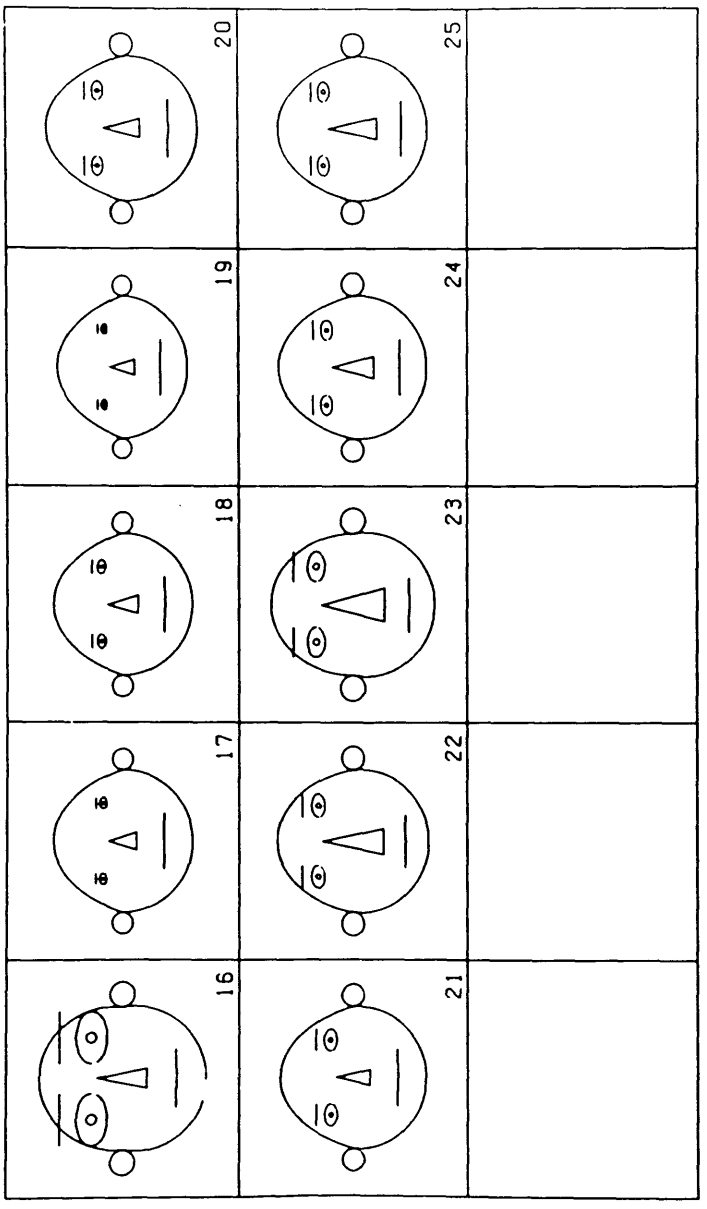


Figure 7.35



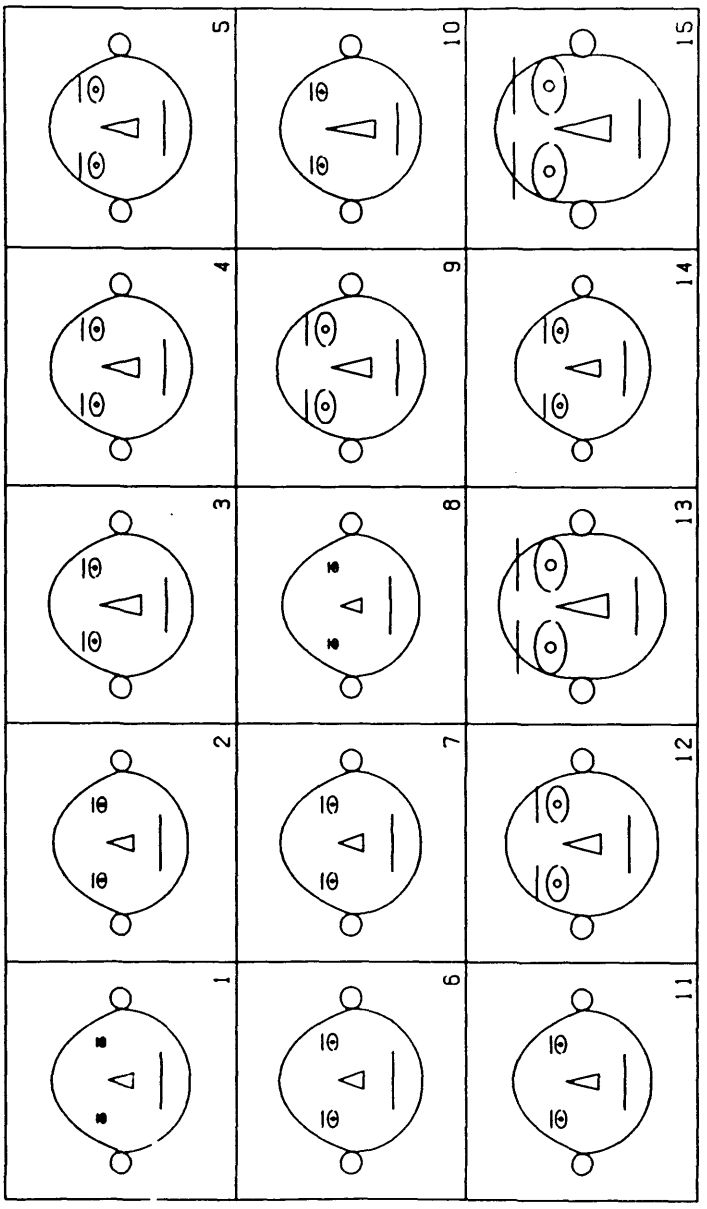
CHERNOFF FACES SIND LOGTEN DATA

Figure 7.36



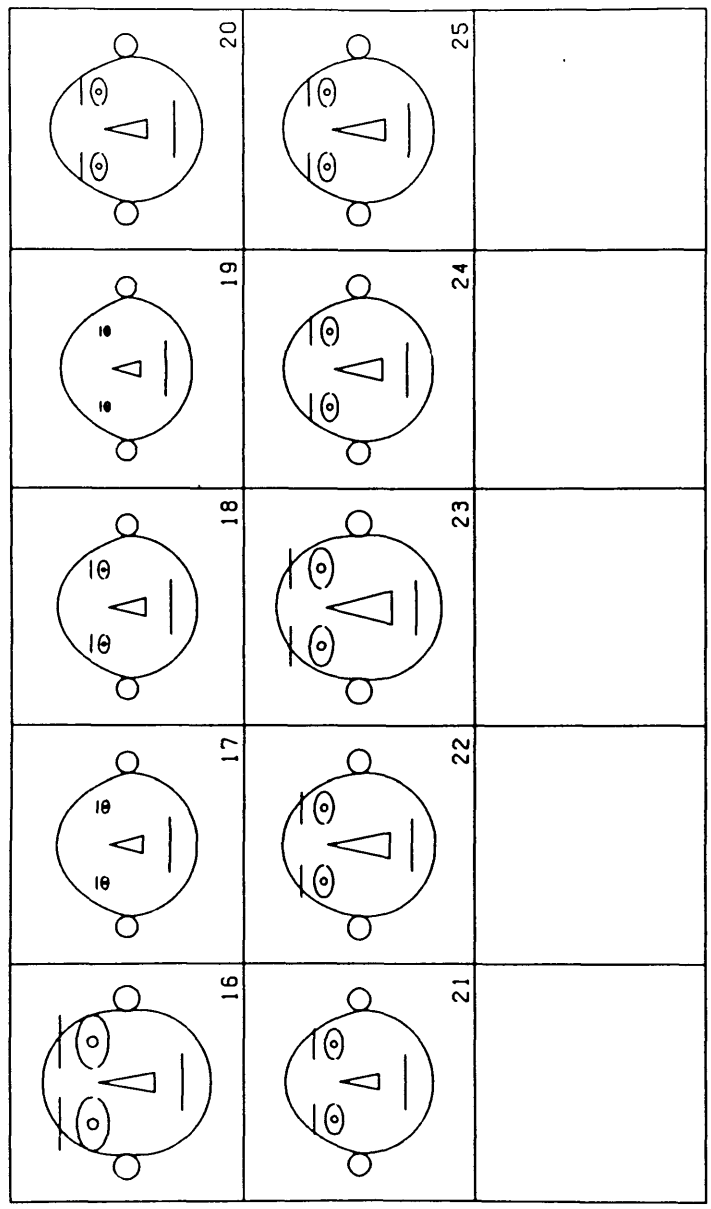
CHERNOFF FACES SIND LOGTEN DATA

Figure 7.37



CHERNOFF FACES SIND BOXCOX DATA

Figure 7.38



knowledge of those variables that characterize the anomalies. This is knowledge which is often impossible to obtain and if obtainable would defeat the purpose of the analysis.

To look at how data tranformations affect Chernoff faces Figures 7.35 through 7.38 have been generated based on the assignment of variables to facial features which emphasise Zn, Cd and Cu. There is very little difference between the faces for logten data and those for boxcox data. Both sets show a facial style for samples 13, 15 and 16 which is distinct from that for samples 22 and 23. However the contrast between these five anomalies and the background samples is very low, especially in the case of sample 22 which could be interpreted as high background. In comparison, the raw data faces not only show the true anomalies by their distinct facial style, but have such a marked anomaly contrast that none of the five outliers could possibly be taken for anything else.

7.8 Trees and the Moray-Buchan Data

As the number of variables in the analysed data increases, the distortion inherent in the dendrogram will increase (see Section 7.5). In addition, as the number of variables increases the choice of cluster algorithm starts

to have a bigger influence on the cophenetic coefficient and the variable groupings. So, for the MB data (24 variables) the cophenetic coefficients are lower than for the SIND data (6 variables) and have a larger spread of values depending on the cluster algorithm chosen. The variable groupings, and particularly the manner in which they are arranged in the dendrogram, change considerably with the algorithm chosen. The algorithm which gives the clearest, most geologically logical and, incidently, the most aesthetically pleasing dendrogram and template tree is the Ward's method. This algorithm had the highest cophenetic coefficient for the boxcox robust correlation matrix but only the third highest for the logten matrix.

As far as the choice of the cluster algorithm is concerned there seem to be no hard and fast rules and thus no substitute for generating the dendrogram and template tree for each algorithm and making the choice on the basis of geological sense, clarity and aesthetics.

Two different TREES runs were completed on the MB data. The first uses logten transformed data with the cluster analysis performed on the logten non-robust correlation matrix (Figures 7.39 through 7.74). The second uses raw data with the cluster analysis performed on the boxcox robust correlation matrix (Figures 7.75 through 7.110). The first is the traditional or classical approach which we compare with an approach which makes use of the lessons learnt in the previous chapters.

Figure 7.39

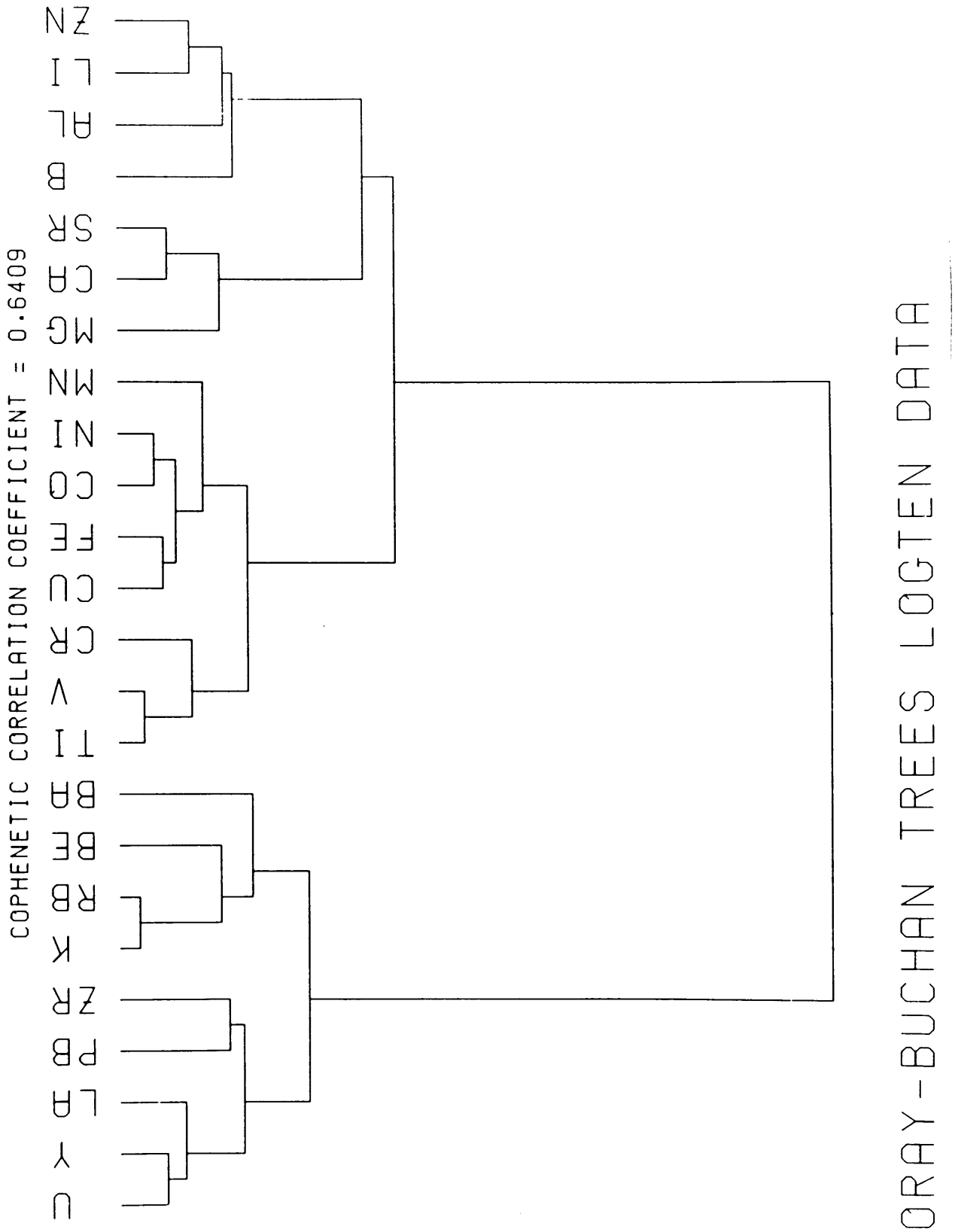
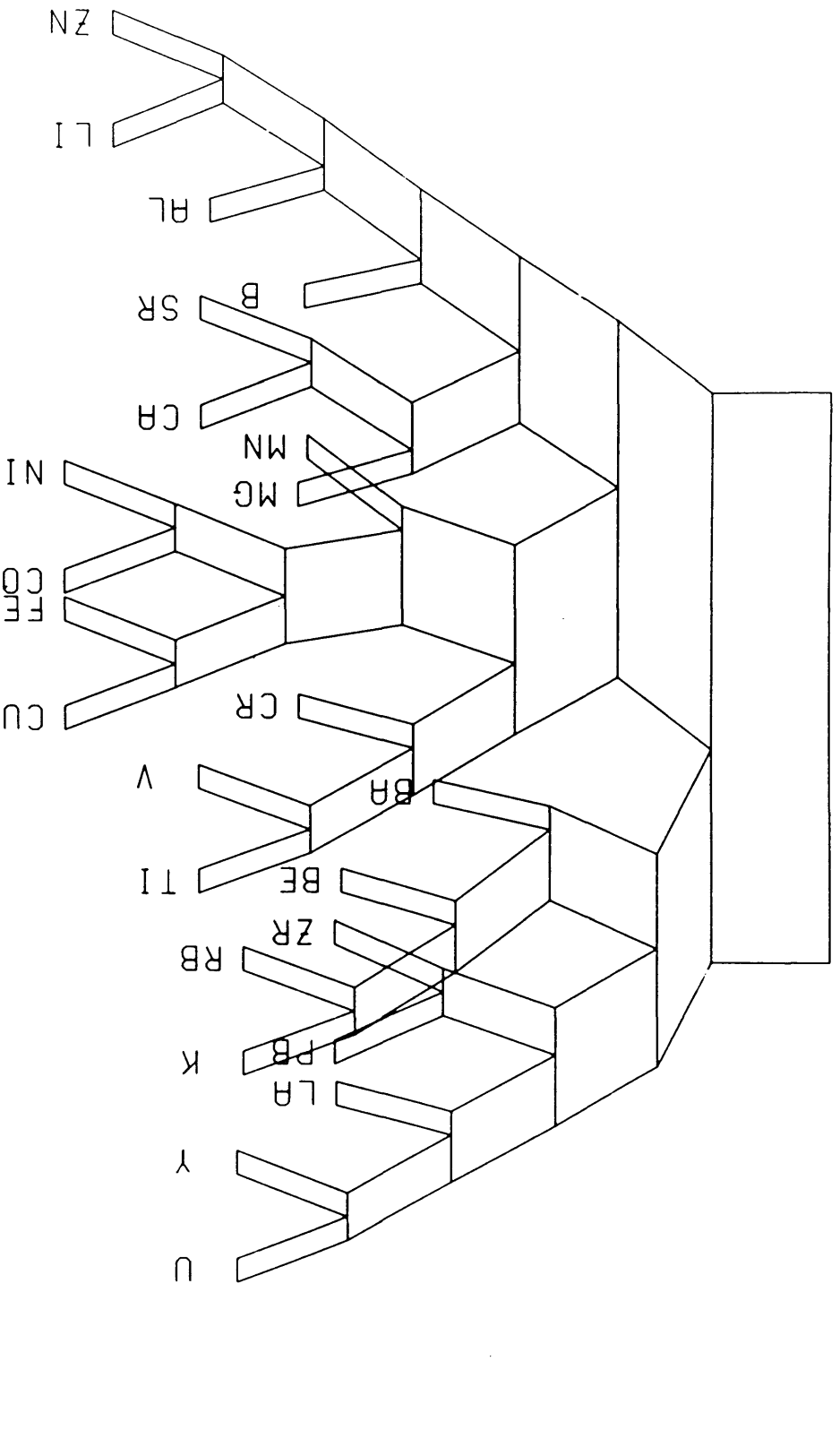
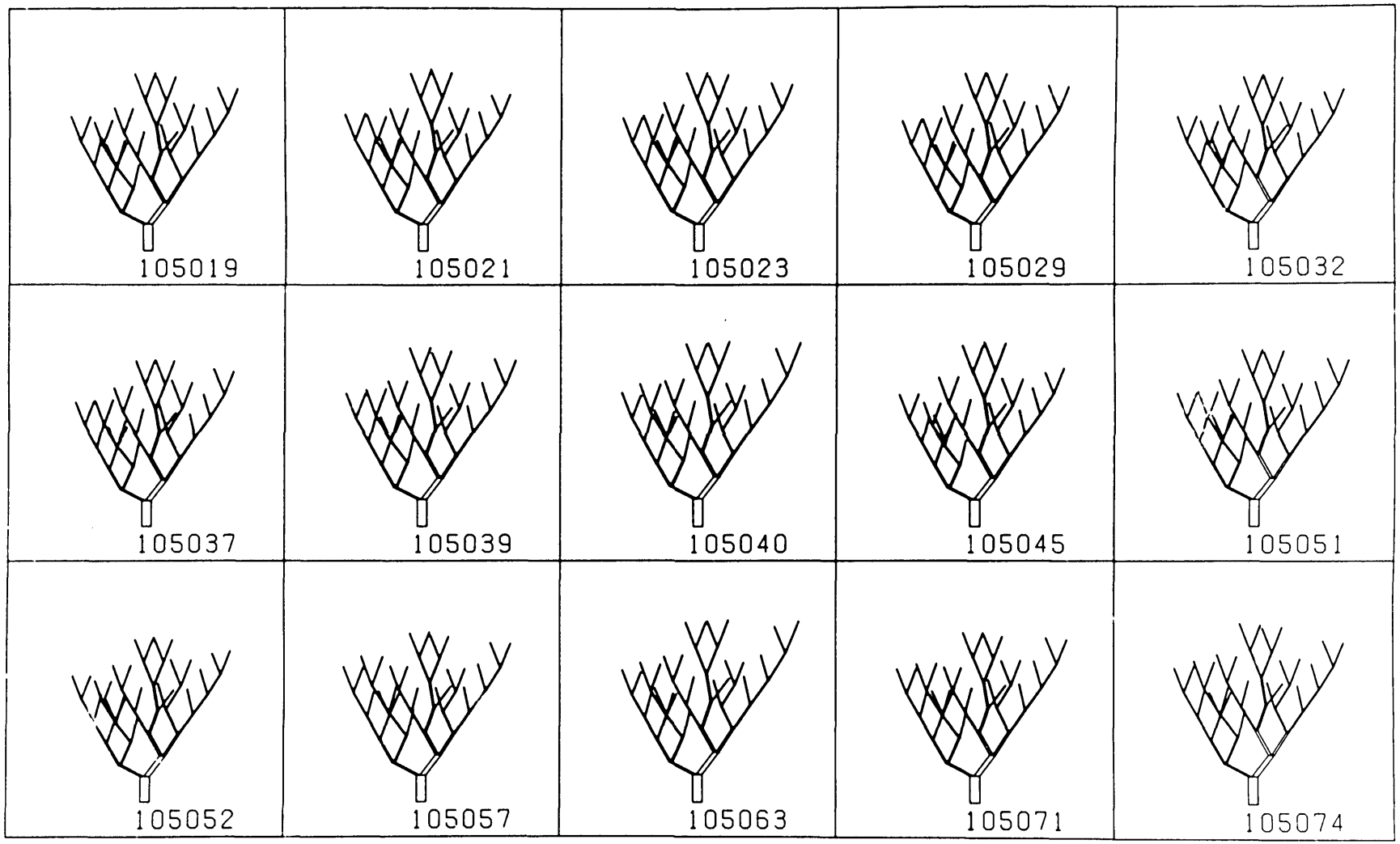

MORAY-BUCHAN TREES LOGTEN DATA
COPHENETIC CORRELATION COEFFICIENT = 0.6409
ZN
LI
AL
B
SR
CA
MG
MN
NI
CO
TE
CU
CR
V
TI
BA
BE
RB
K
ZR
PB
LA
Y
U

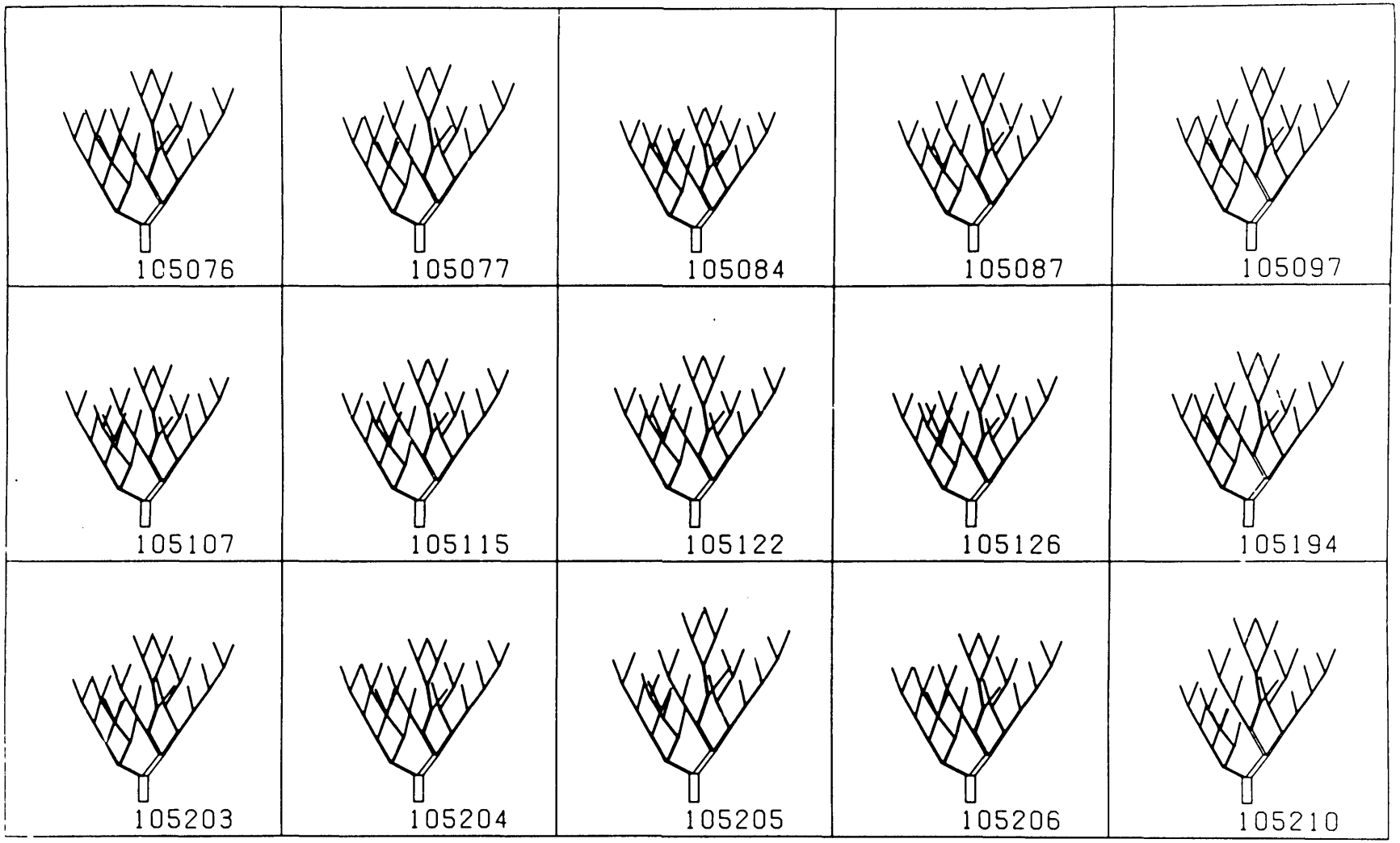
Figure 7.40


ZN
LI
AL
B
SR
CA
MN
MG
NI
CO
FE
CU
CR
BR
V
TI
BE
ZR
RB
PB
K
LA
Y
U
MORAY-BUCHAN TREES LOGTEN DATA



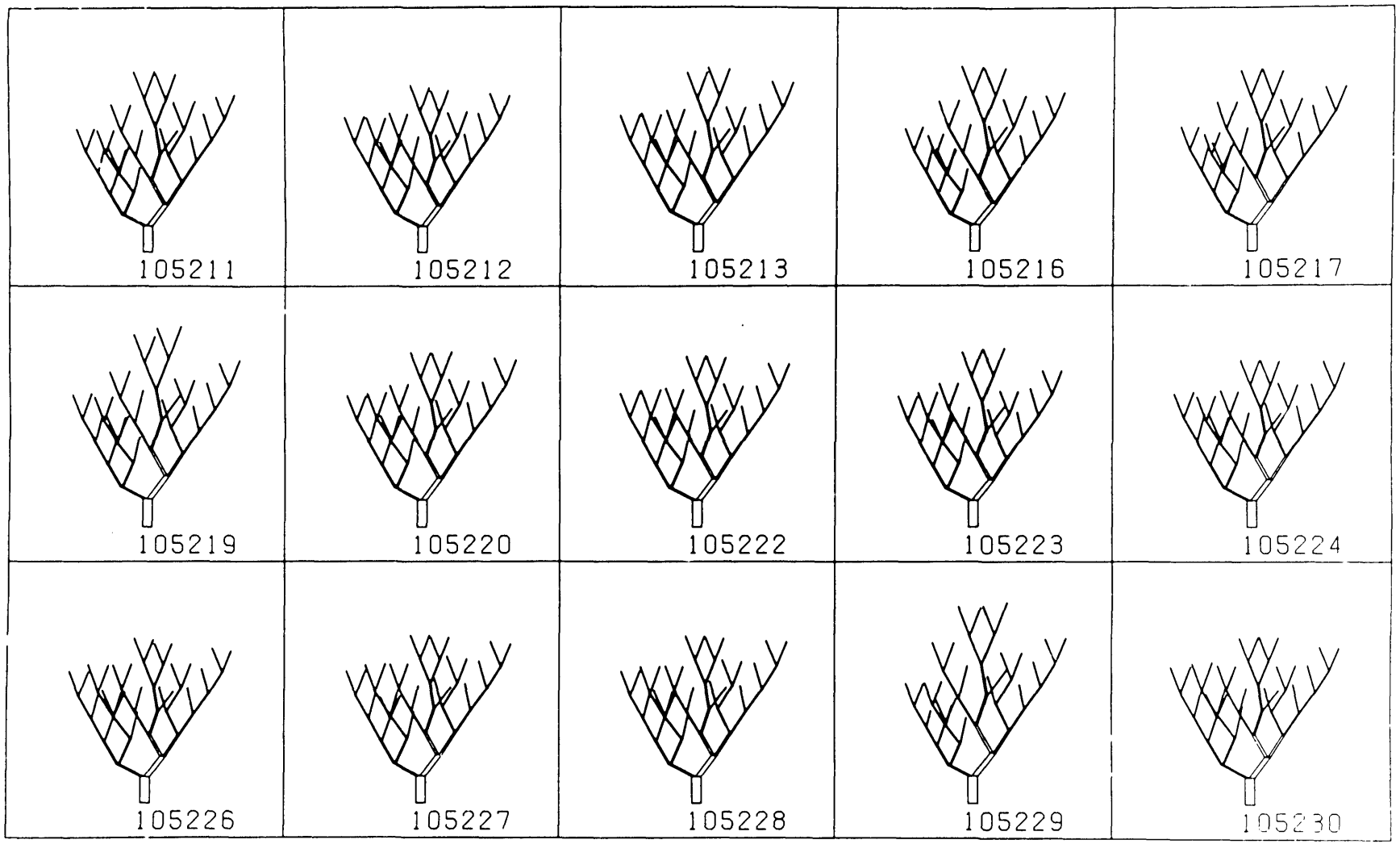
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.41



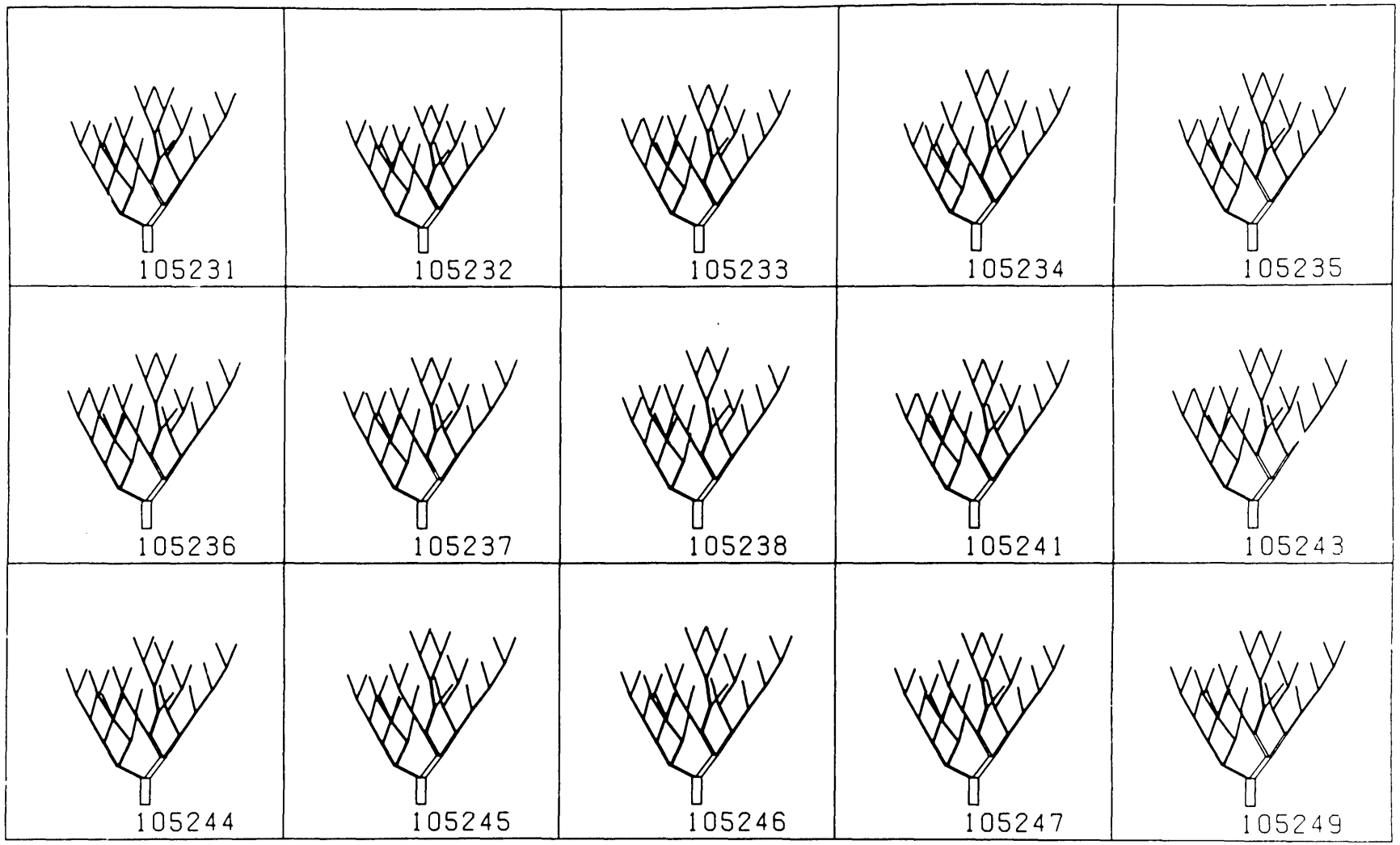
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.42



MORAY-BUCHAN TREES LOGTEN DATA

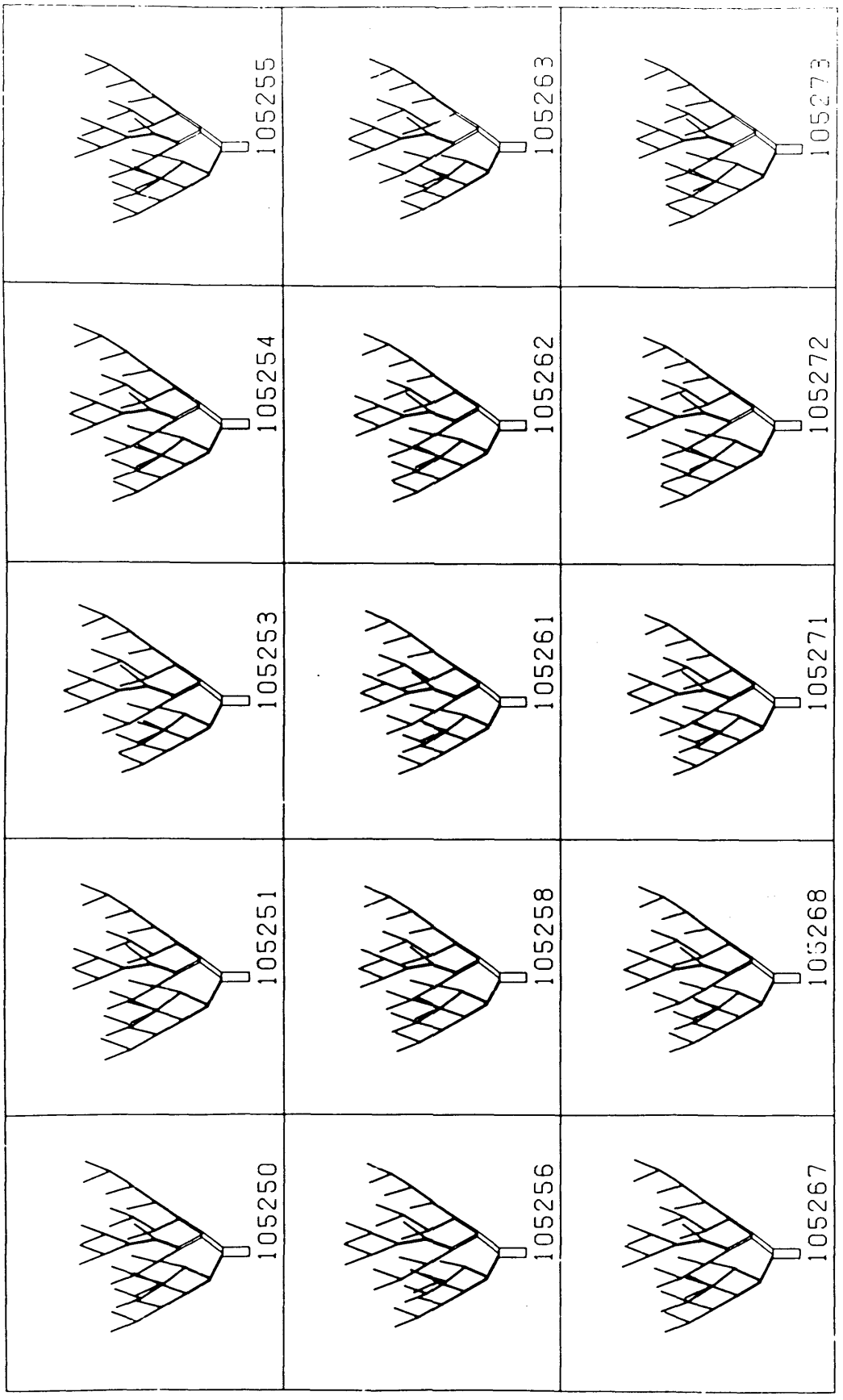
Figure 7.43



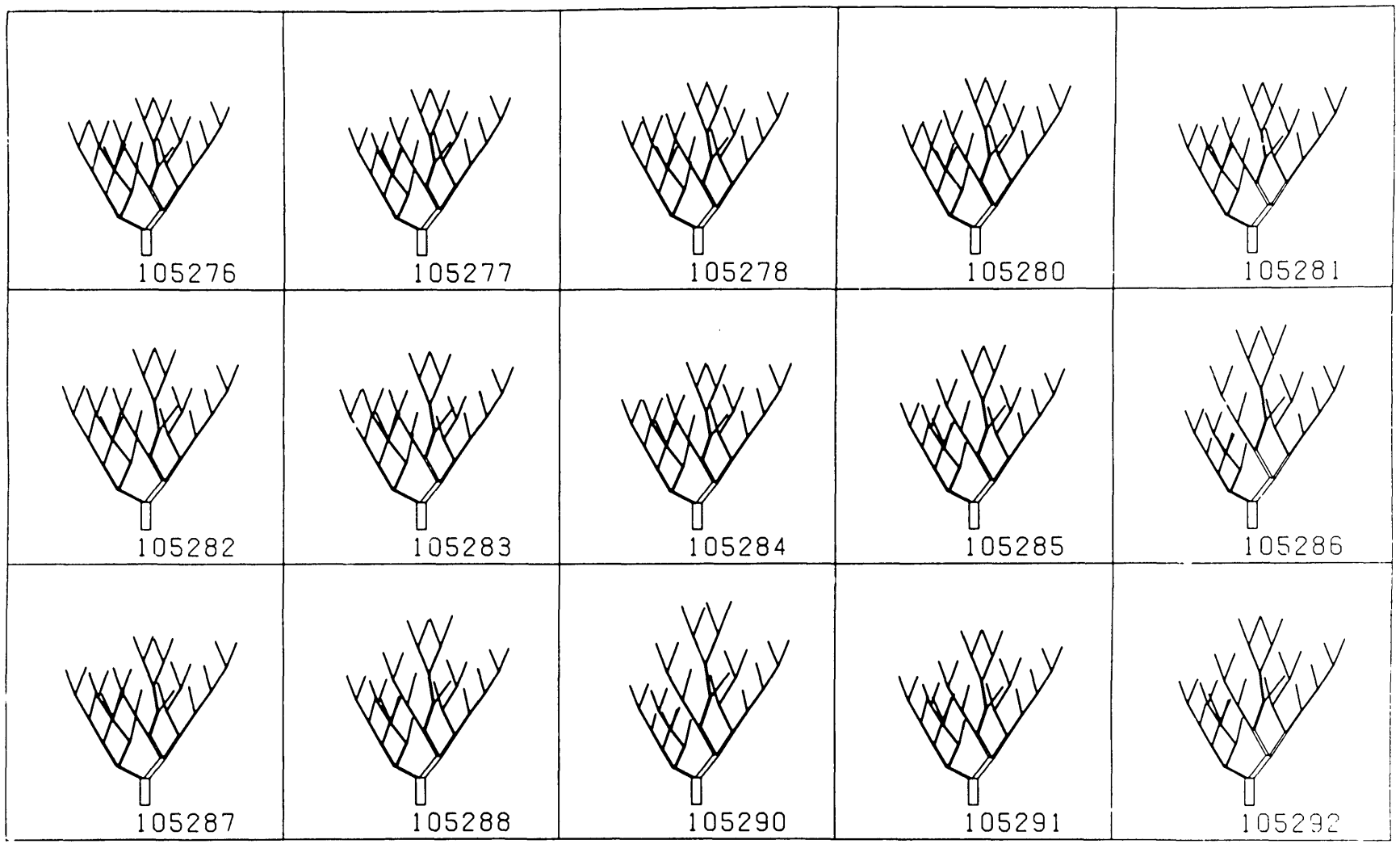
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.44

Figure 7.45



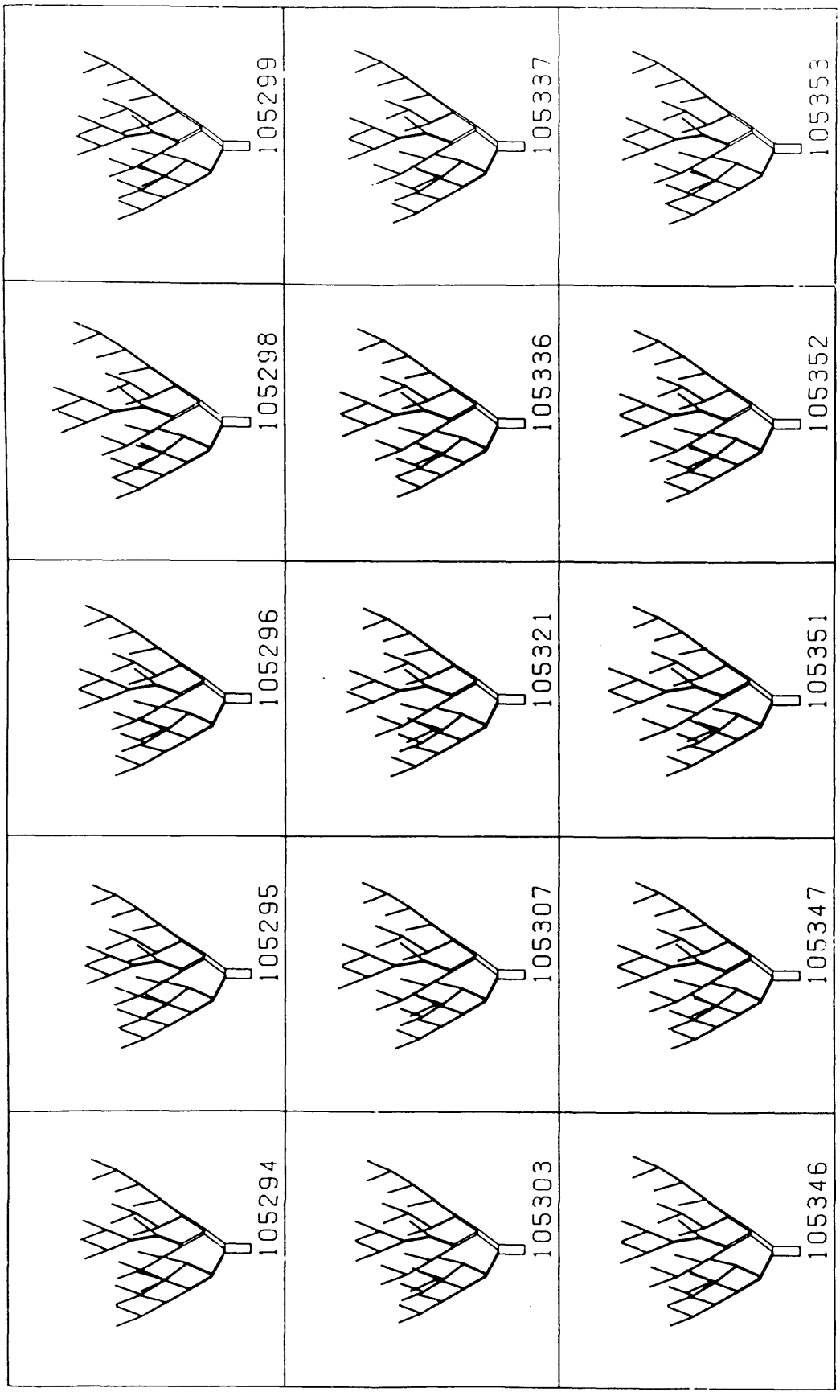
MORAY-BUCHAN TREES LOGTEN DATA



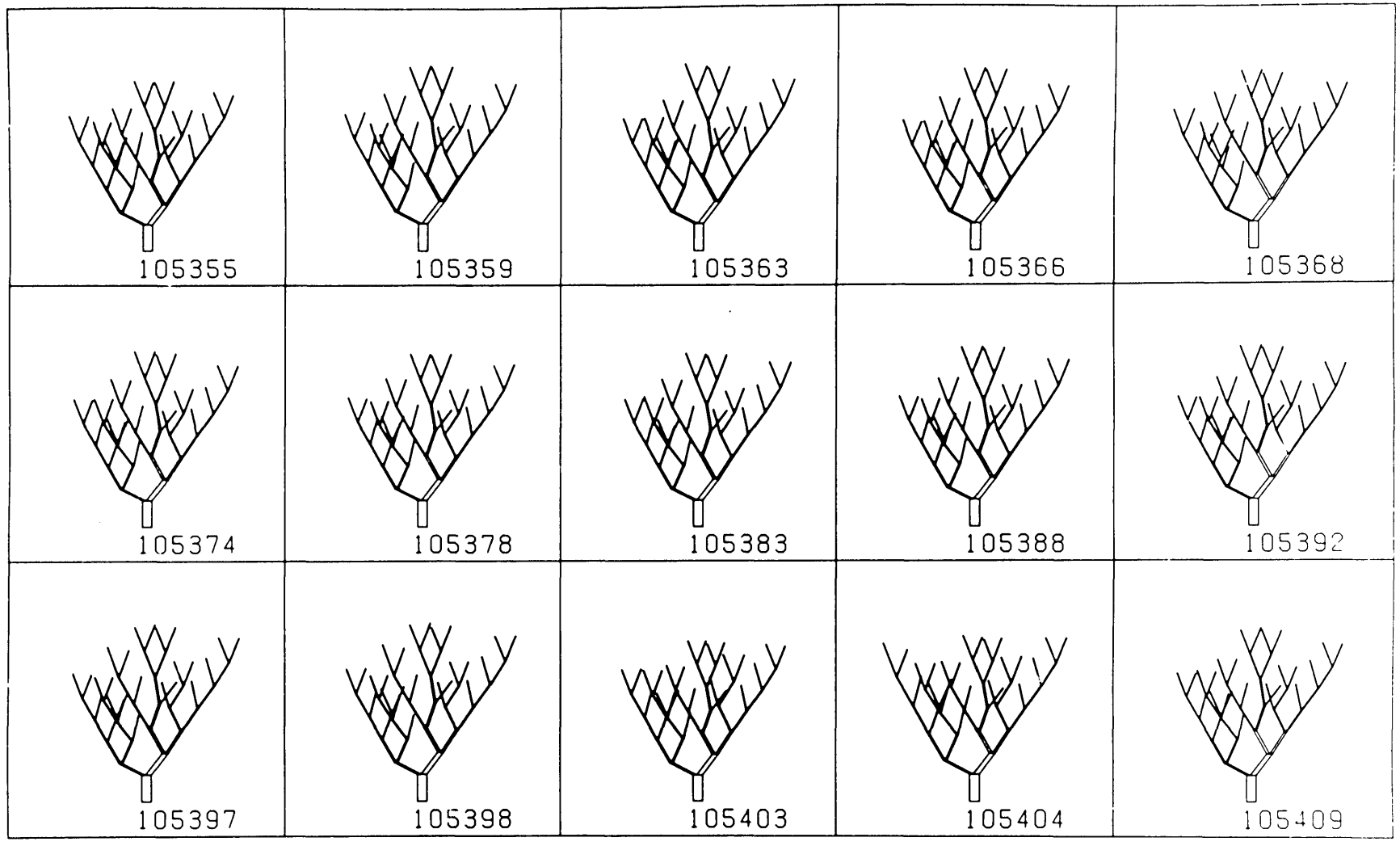
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.46

Figure 7.47



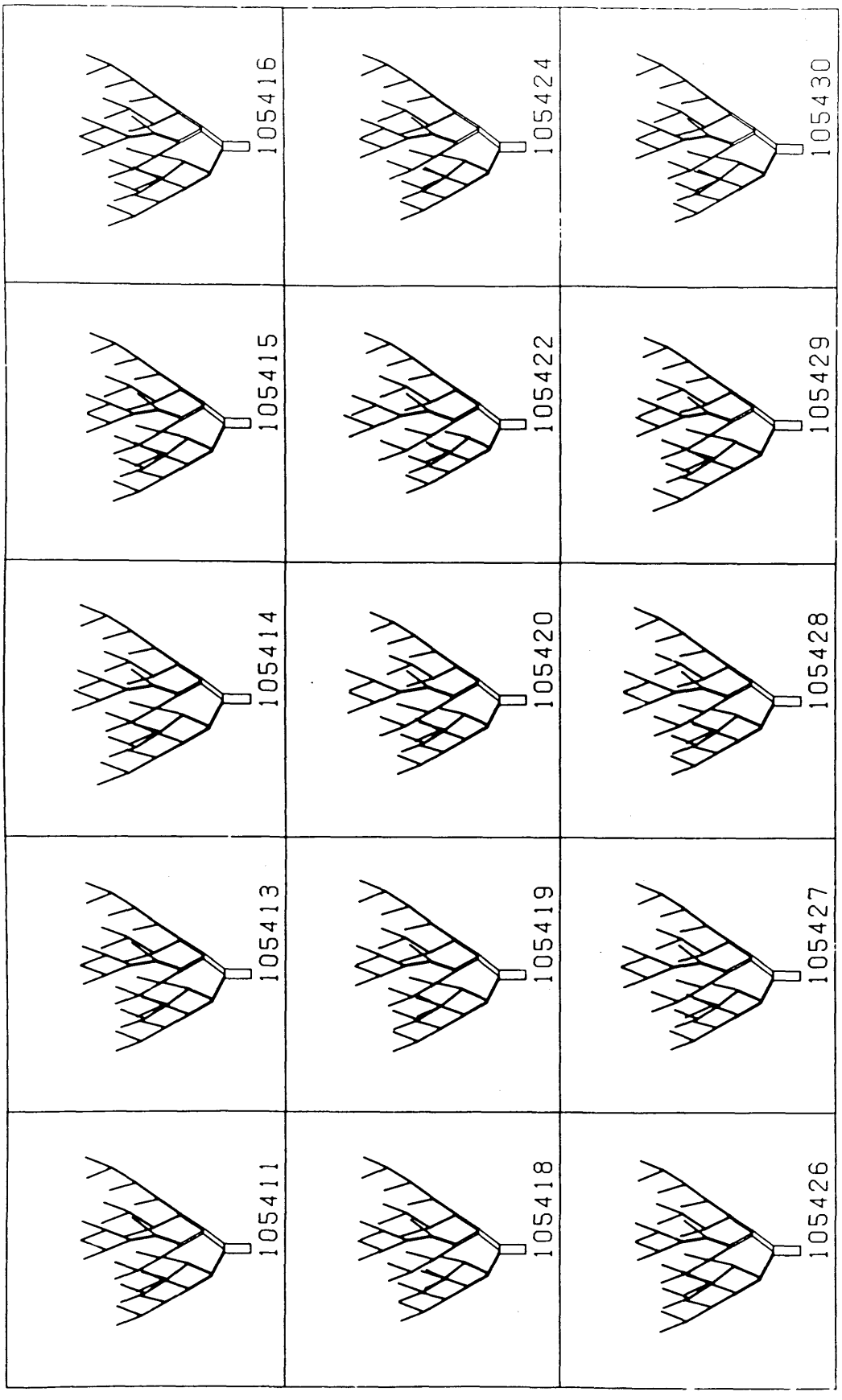
MORAY-BUCHAN TREES LOGTEN DATA



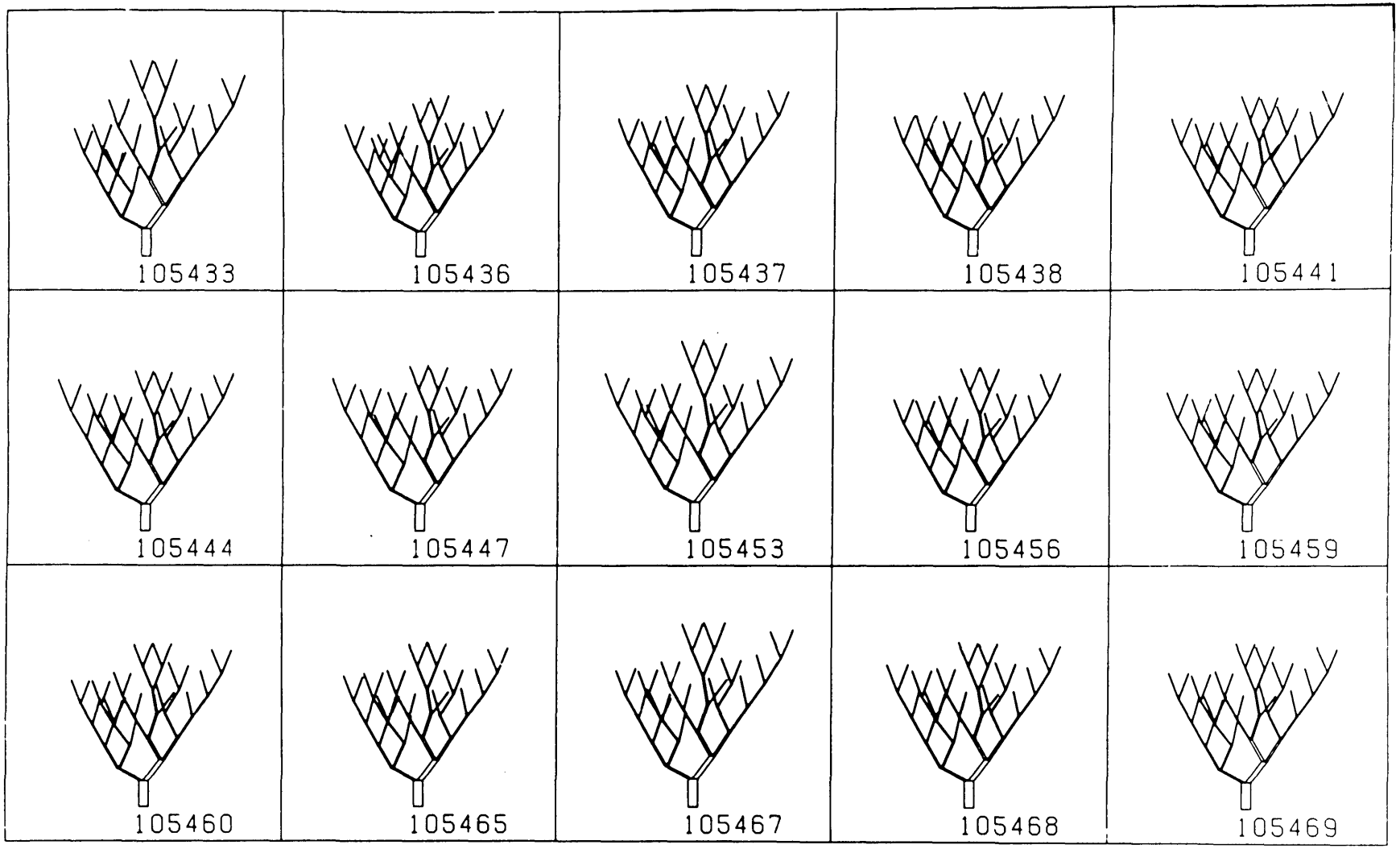
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.48

Figure 7.49



MORAY-BUCHAN TREES LOGTEN DATA



MORAY-BUCHAN TREES LOGTEN DATA

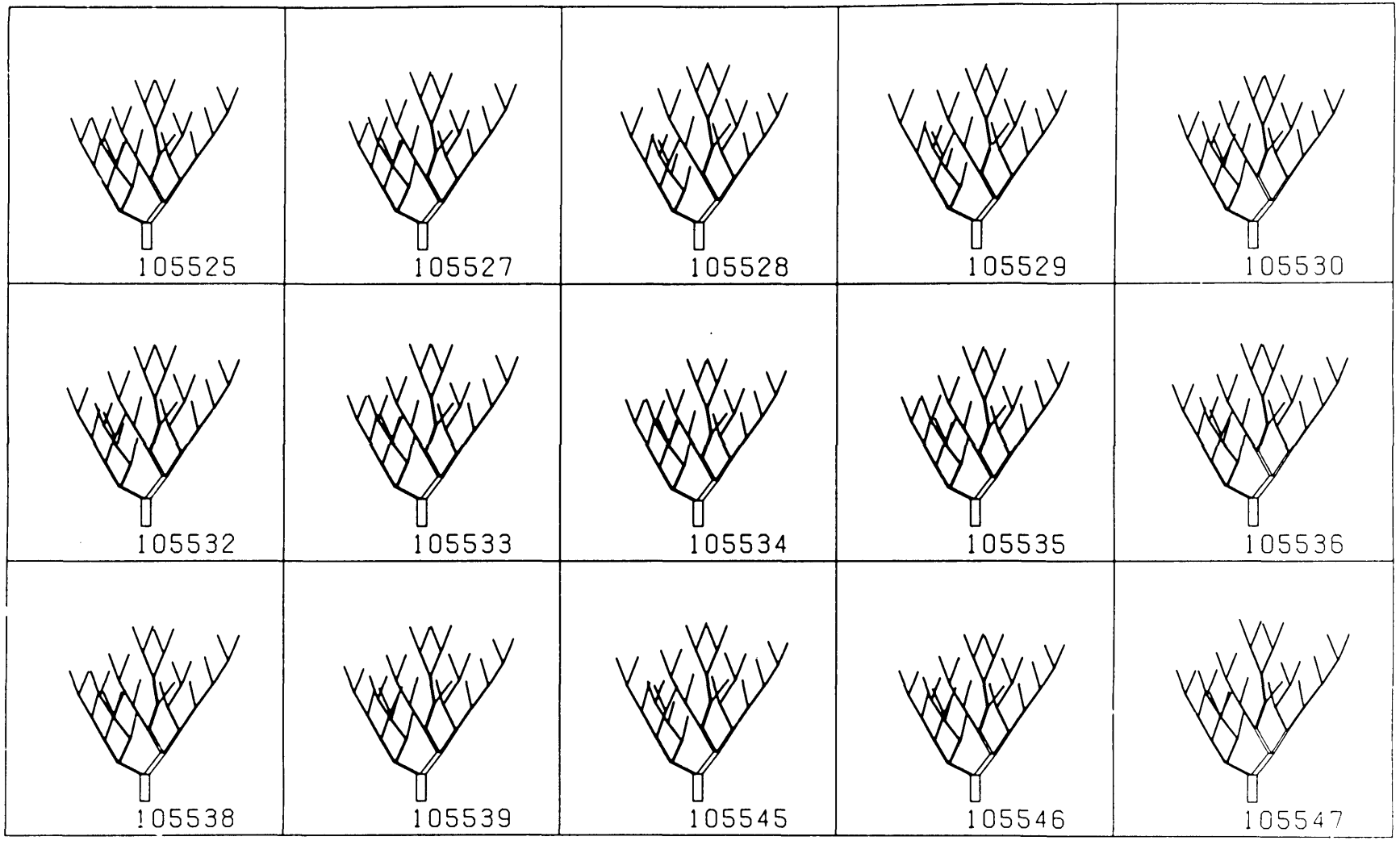
Figure 7.50

MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.51

MORAY-BUCHAN TREES LOGTEN DATA

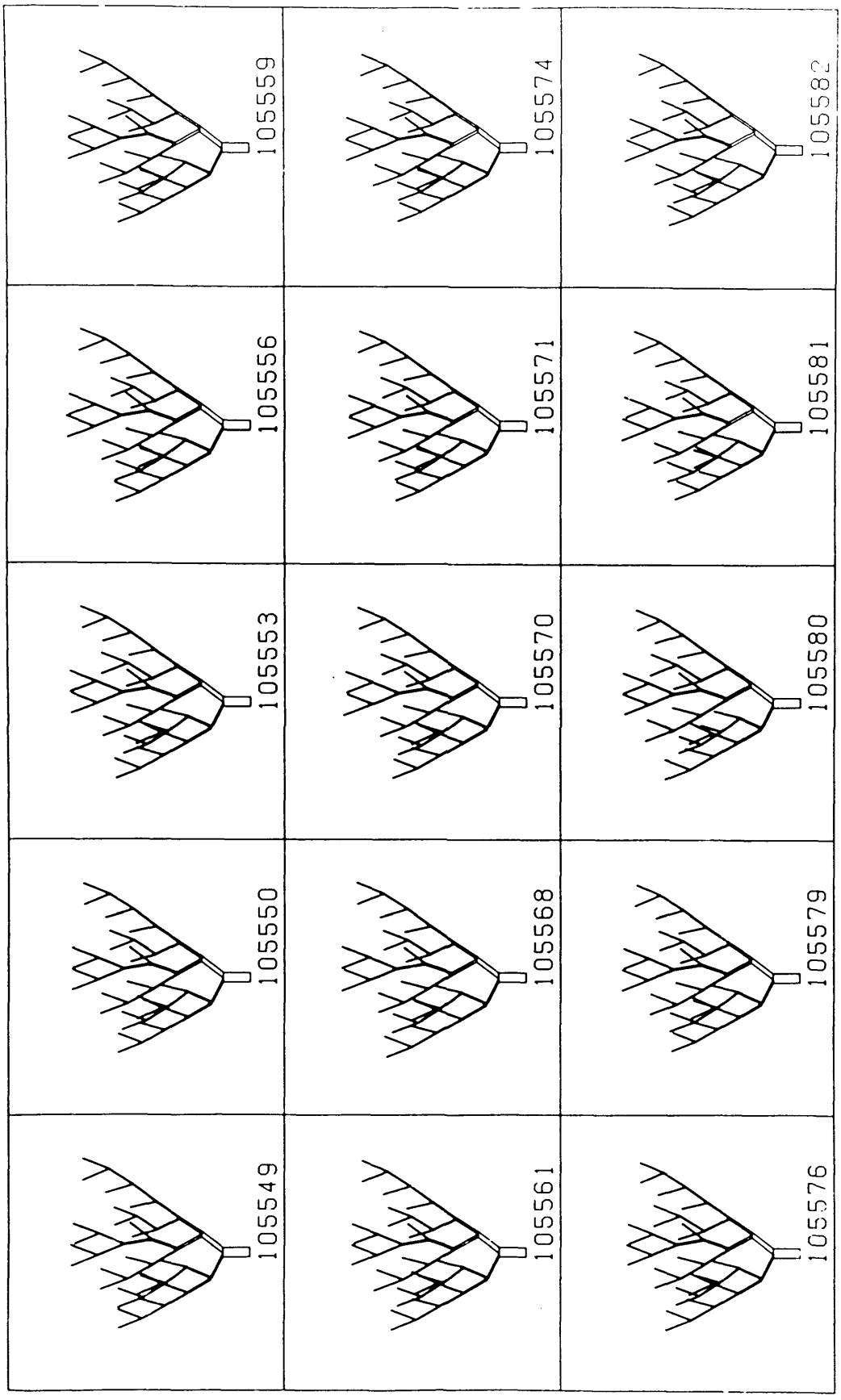
Figure 7.52



MORAY-BUCHAN TREES LOGTEN DATA

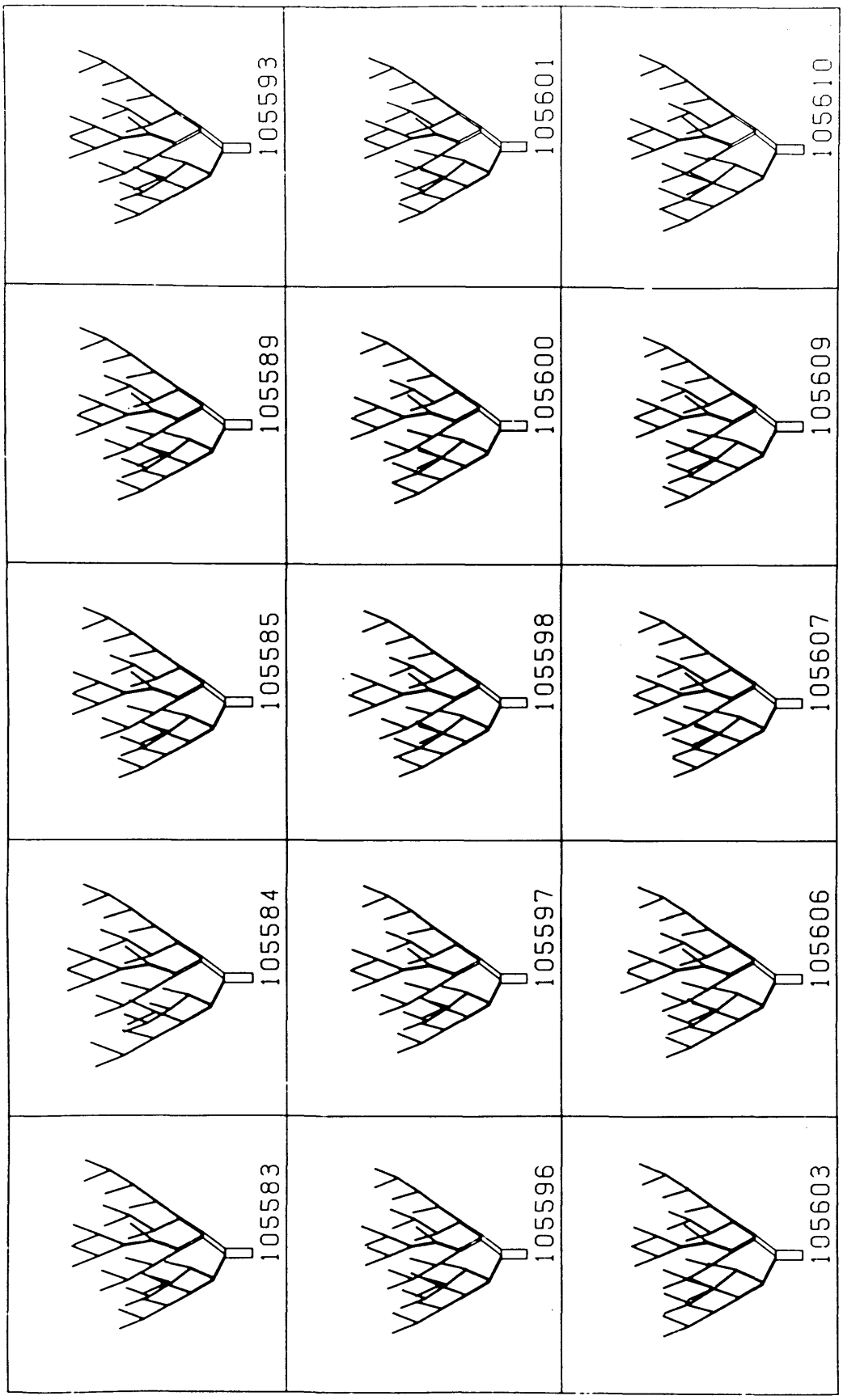
Figure 7.53

Figure 7.54

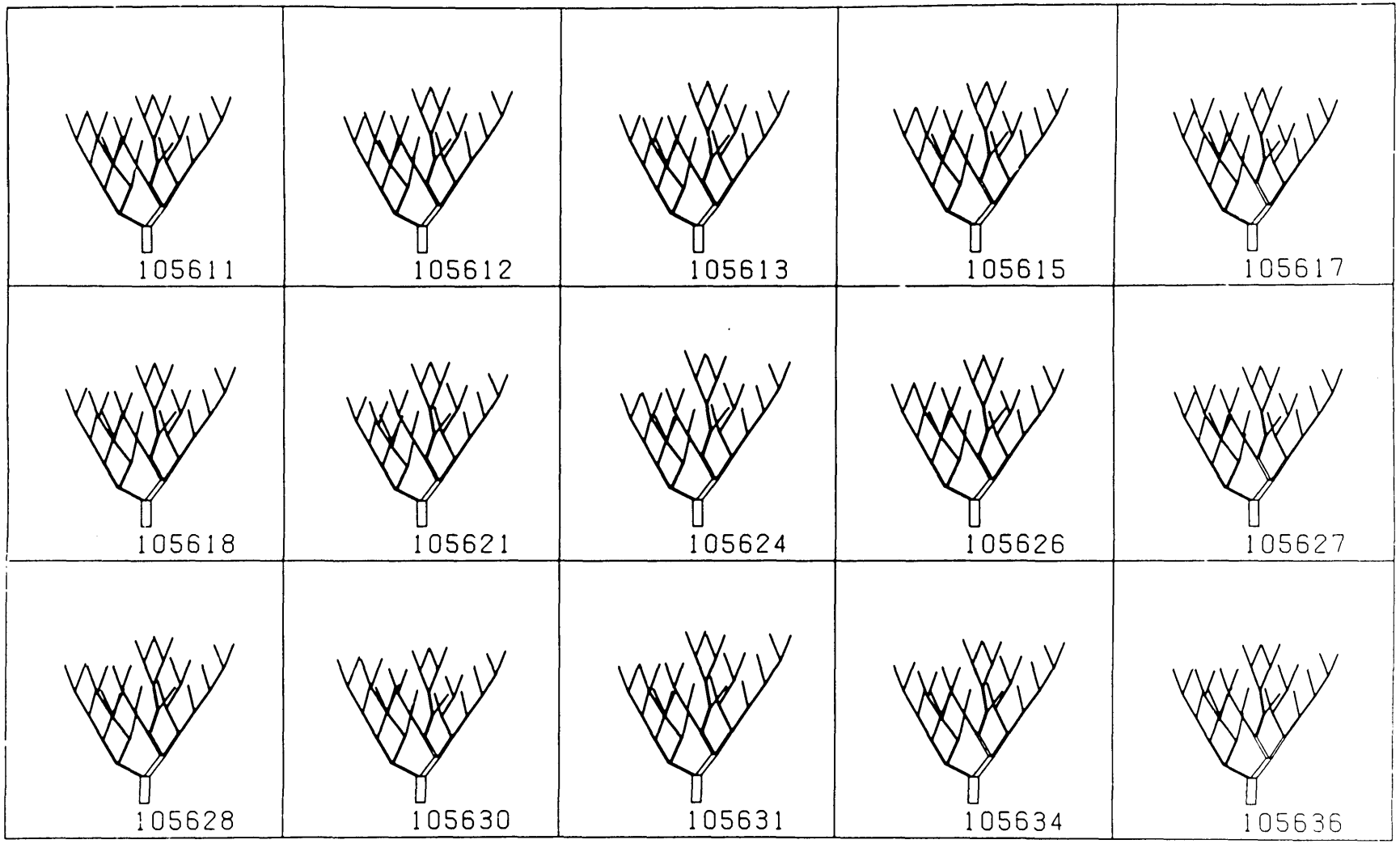


MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.55

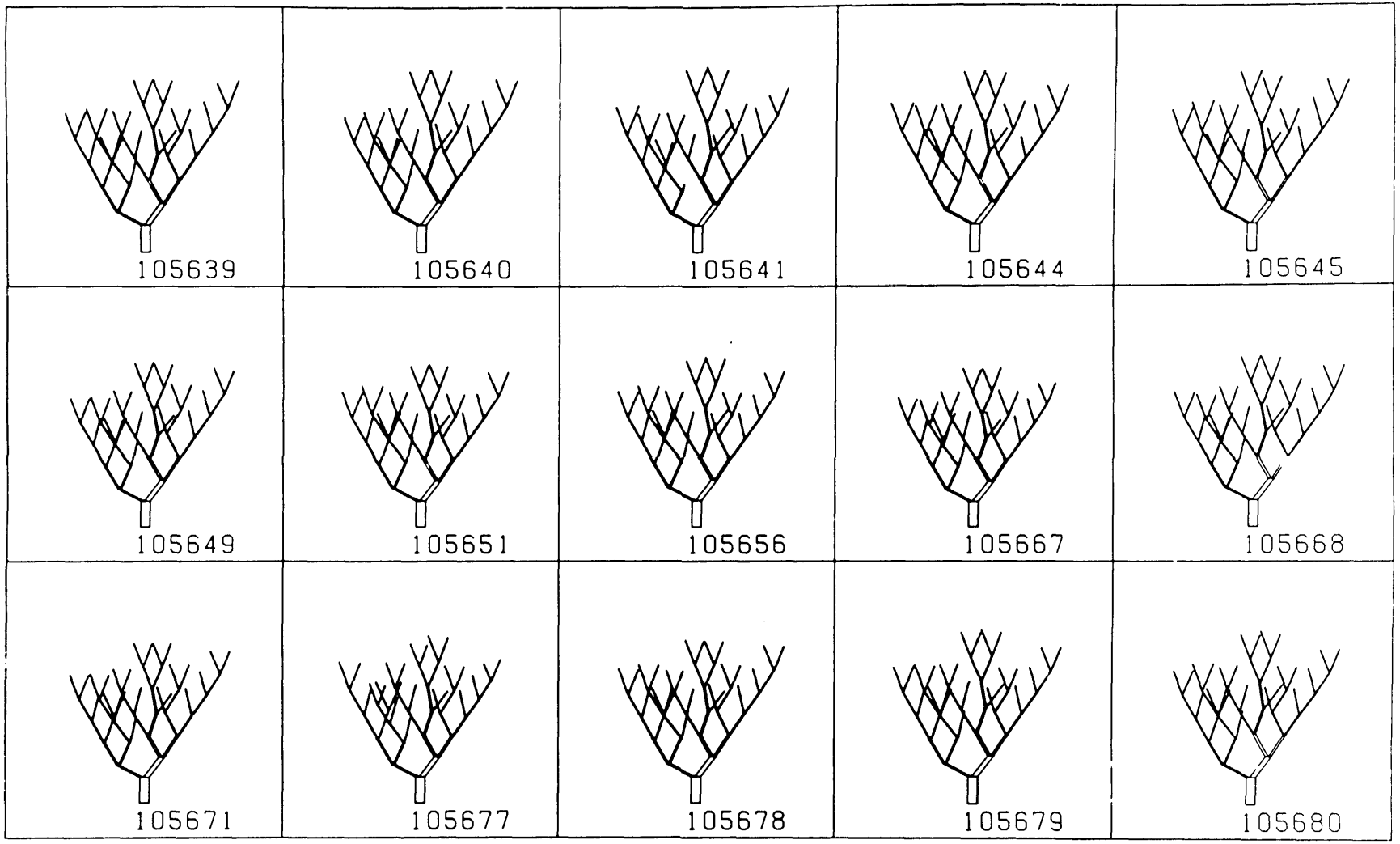


MORAY-BUCHAN TREES LOGTEN DATA



MORAY-BUCHAN TREES LOGTEN DATA

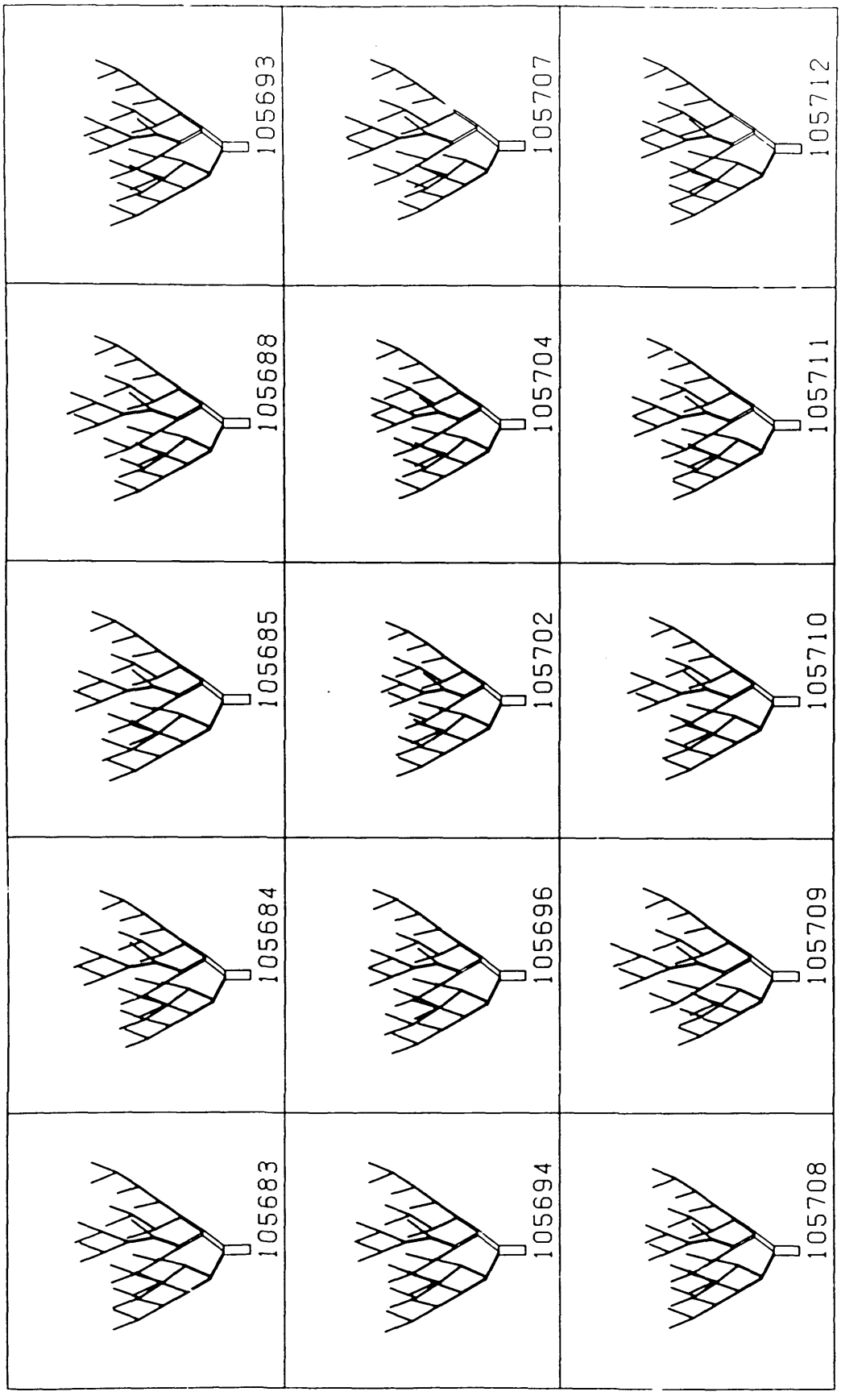
Figure 7.56



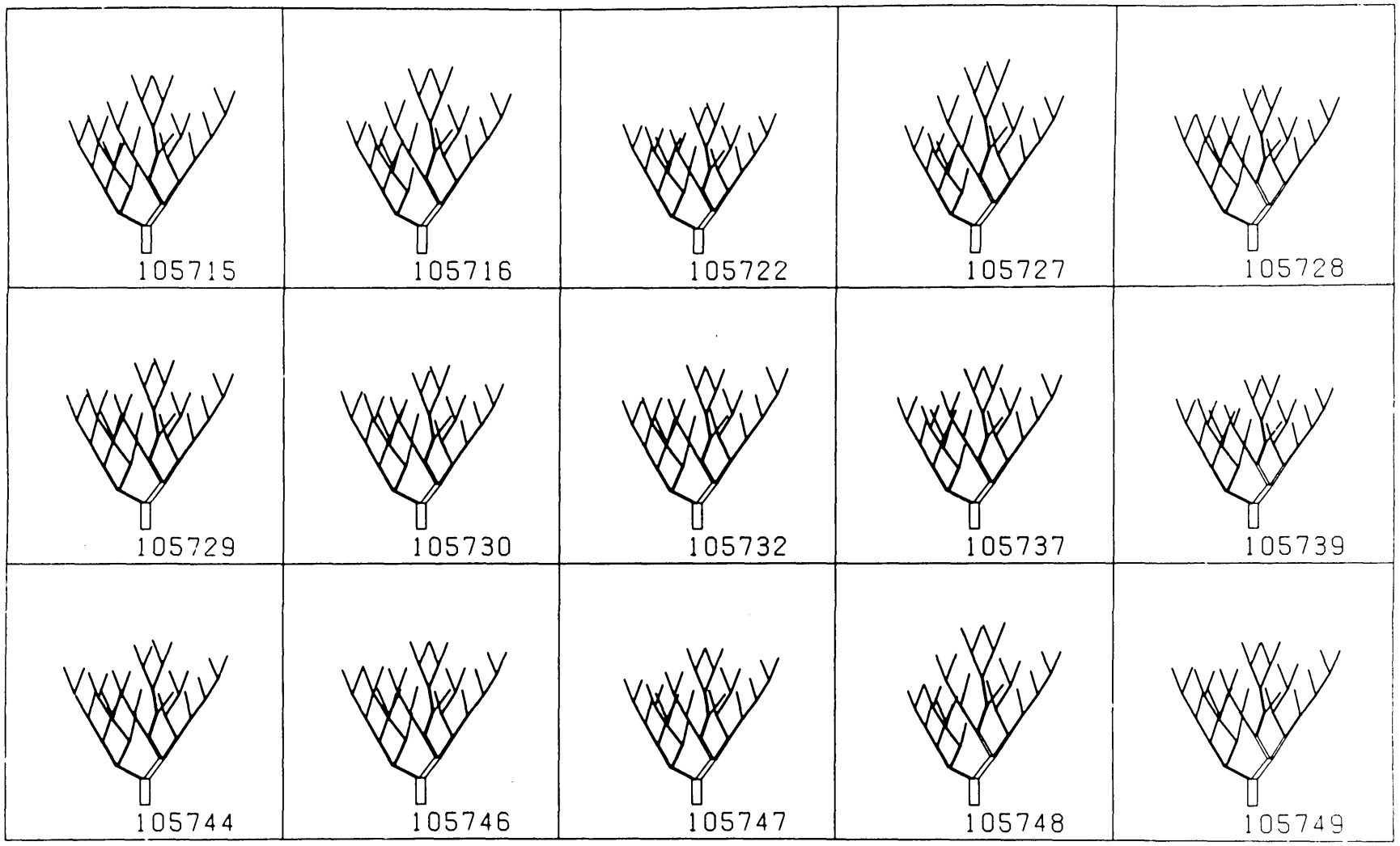
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.57

Figure 7.58

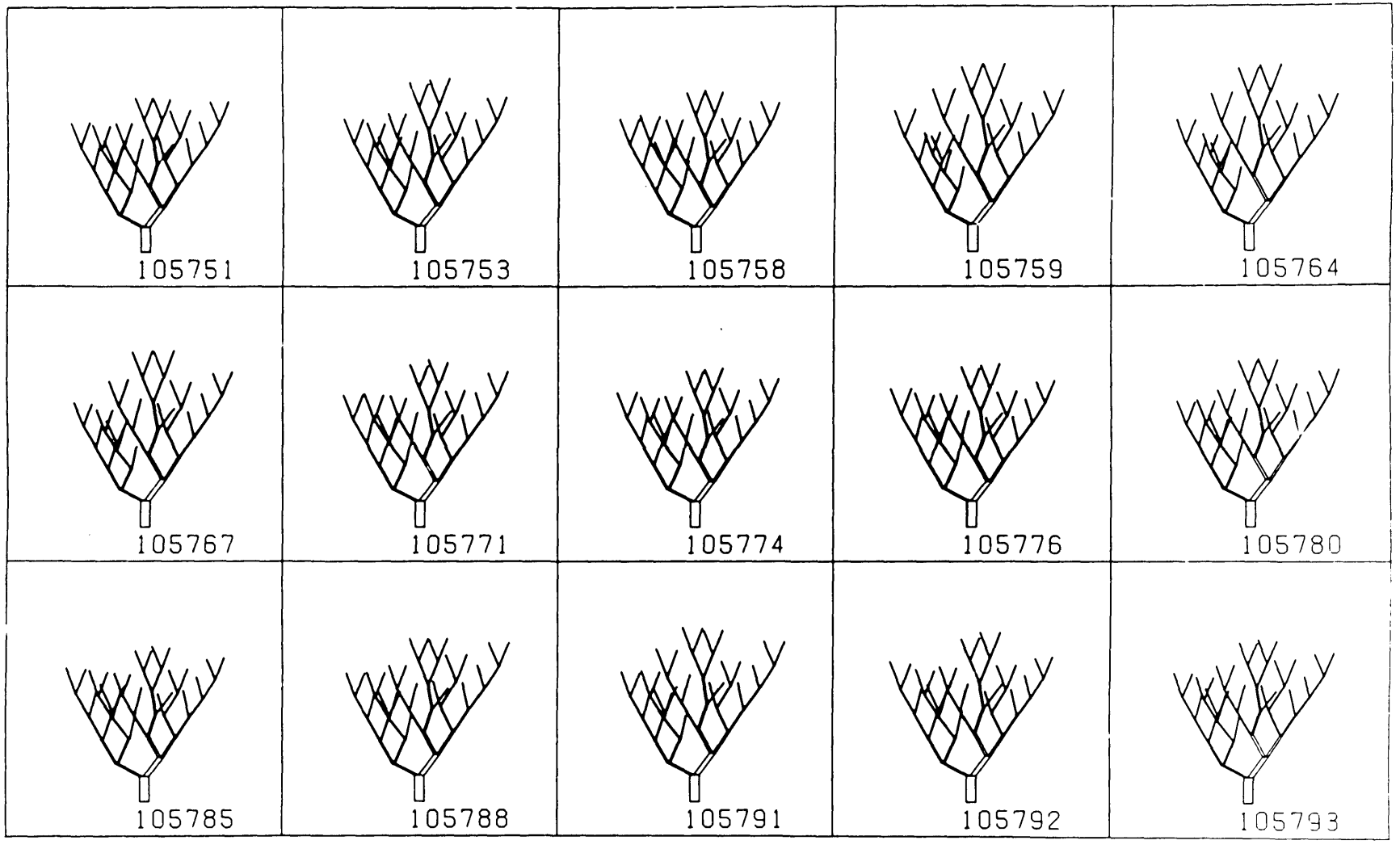


MORAY-BUCHAN TREES LOGTEN DATA



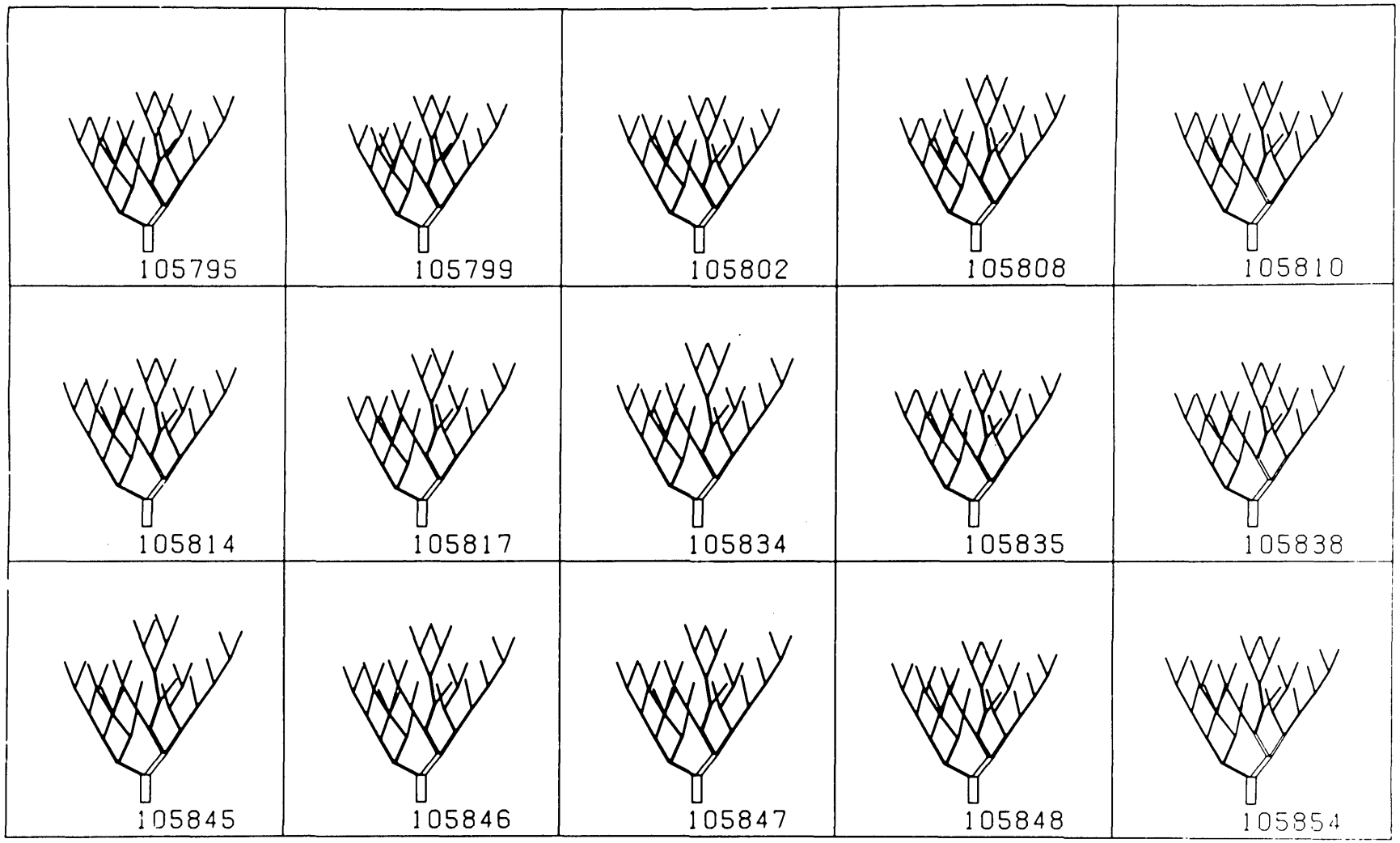
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.59



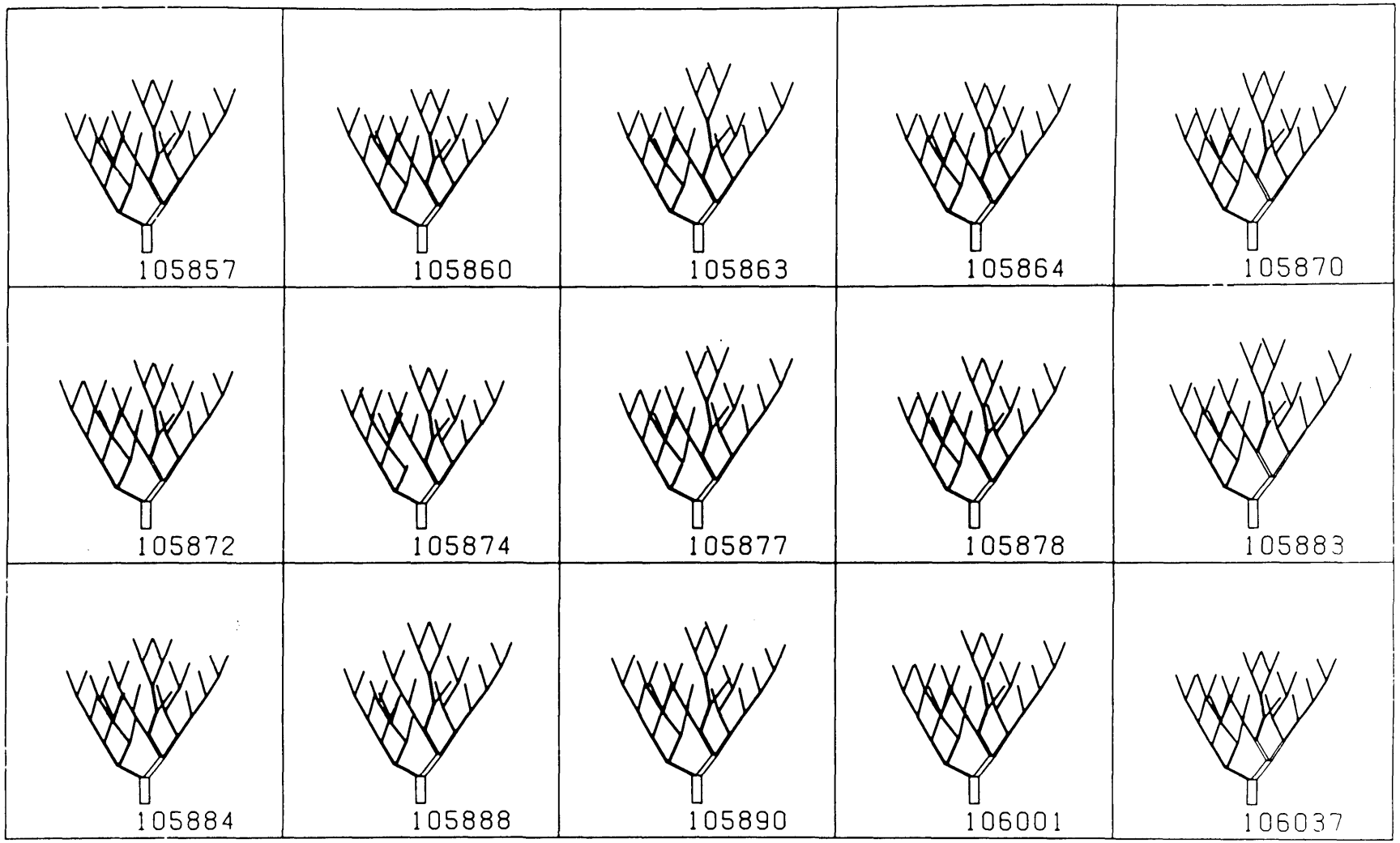
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.60



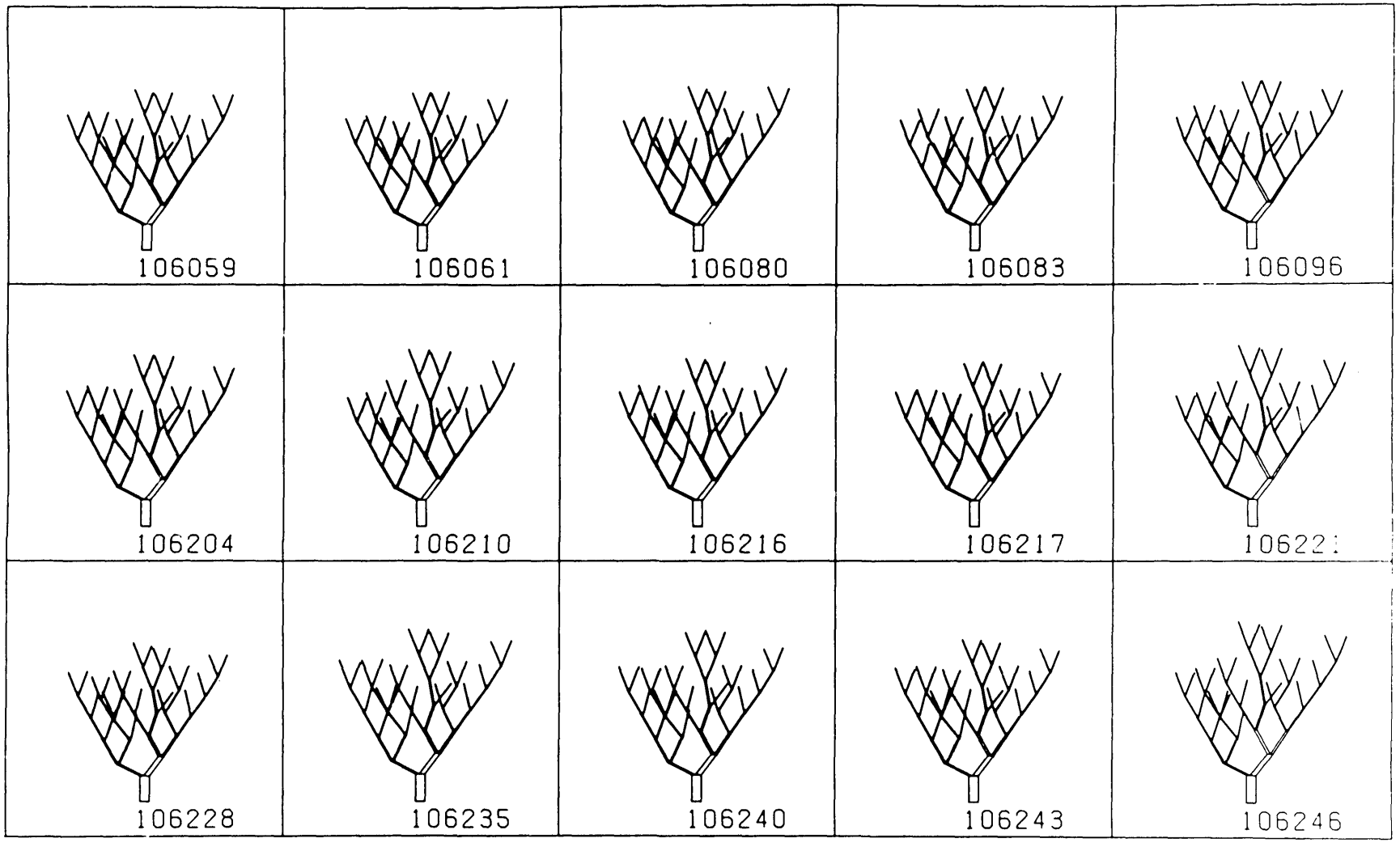
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.61



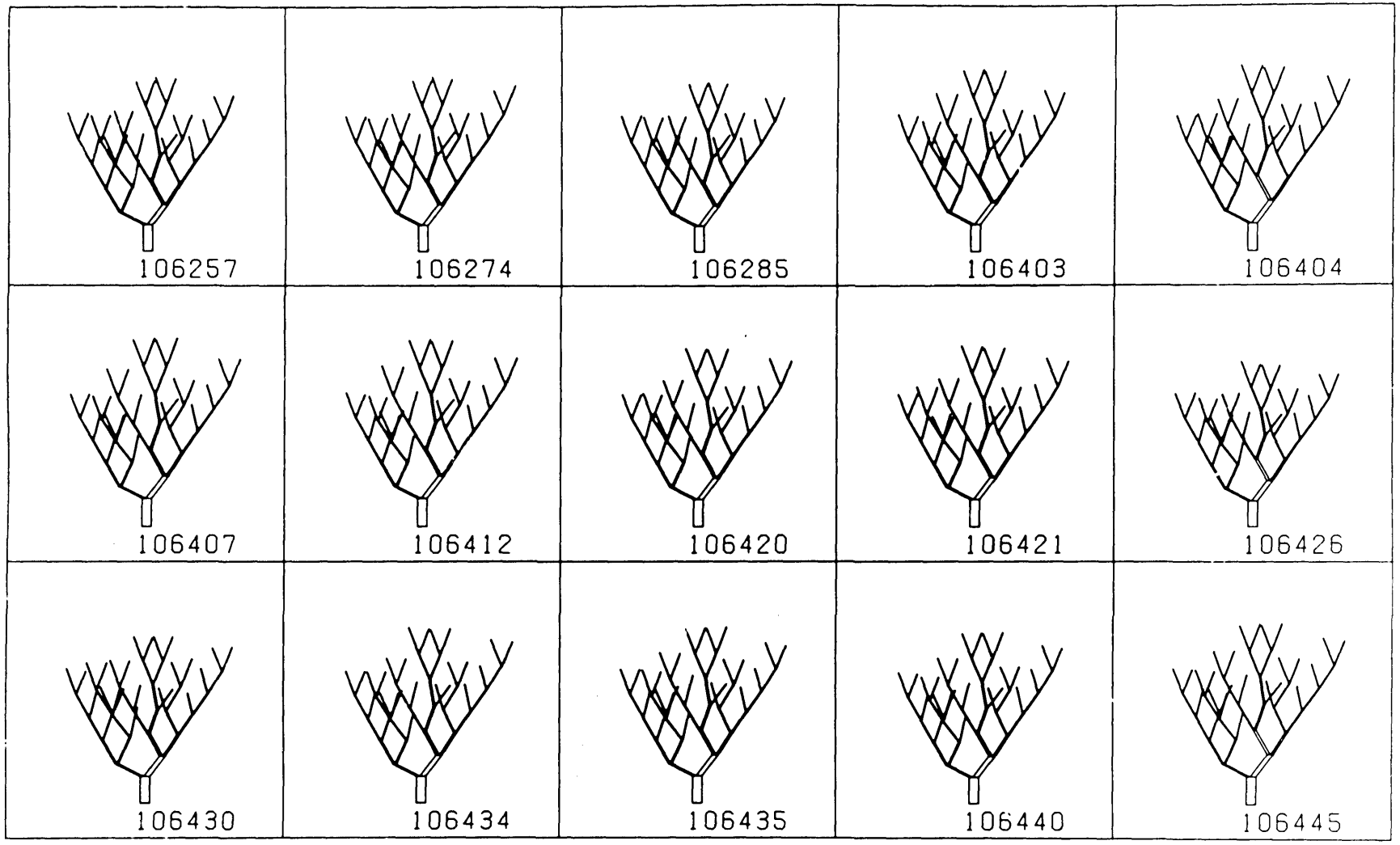
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.62



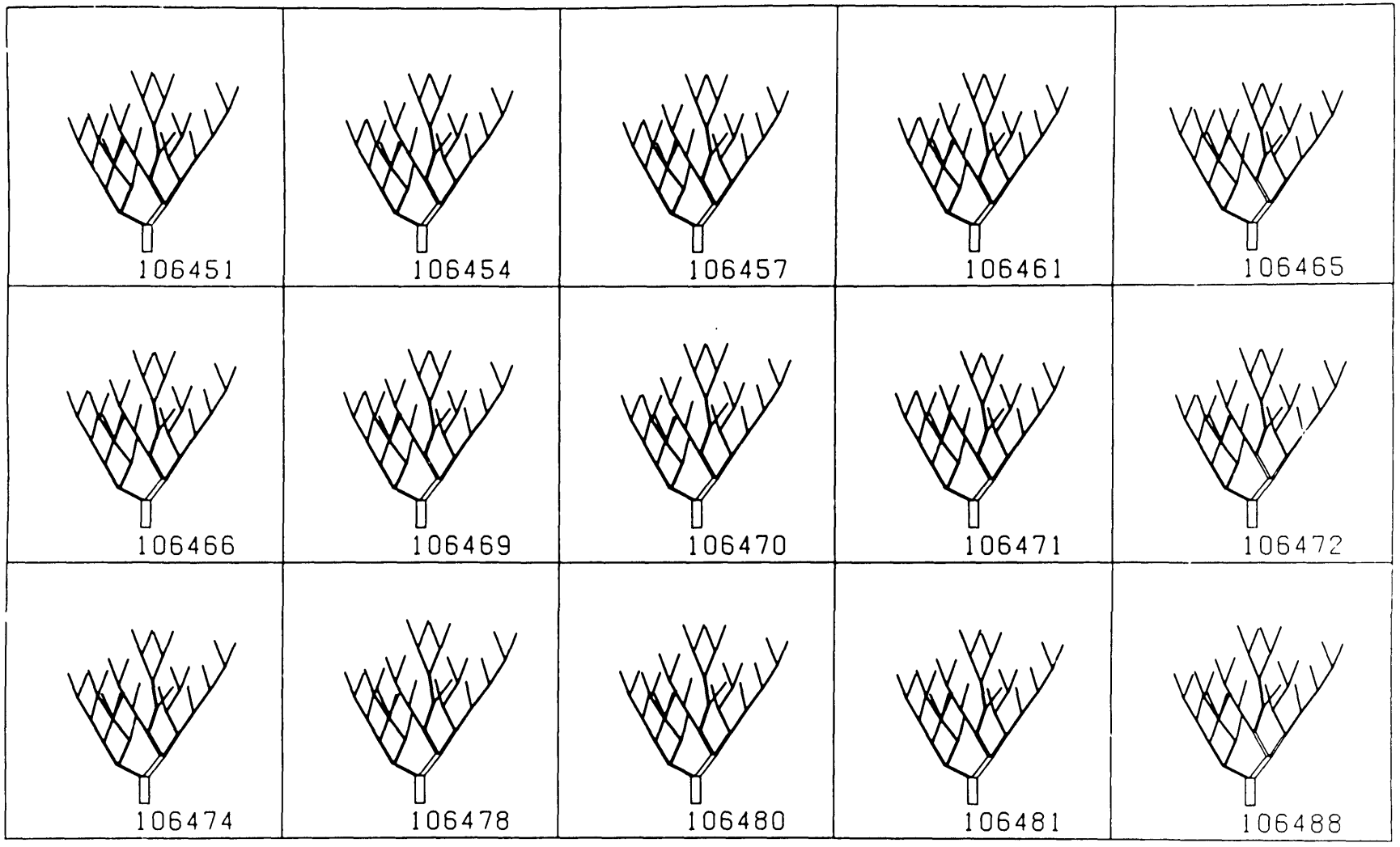
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.63



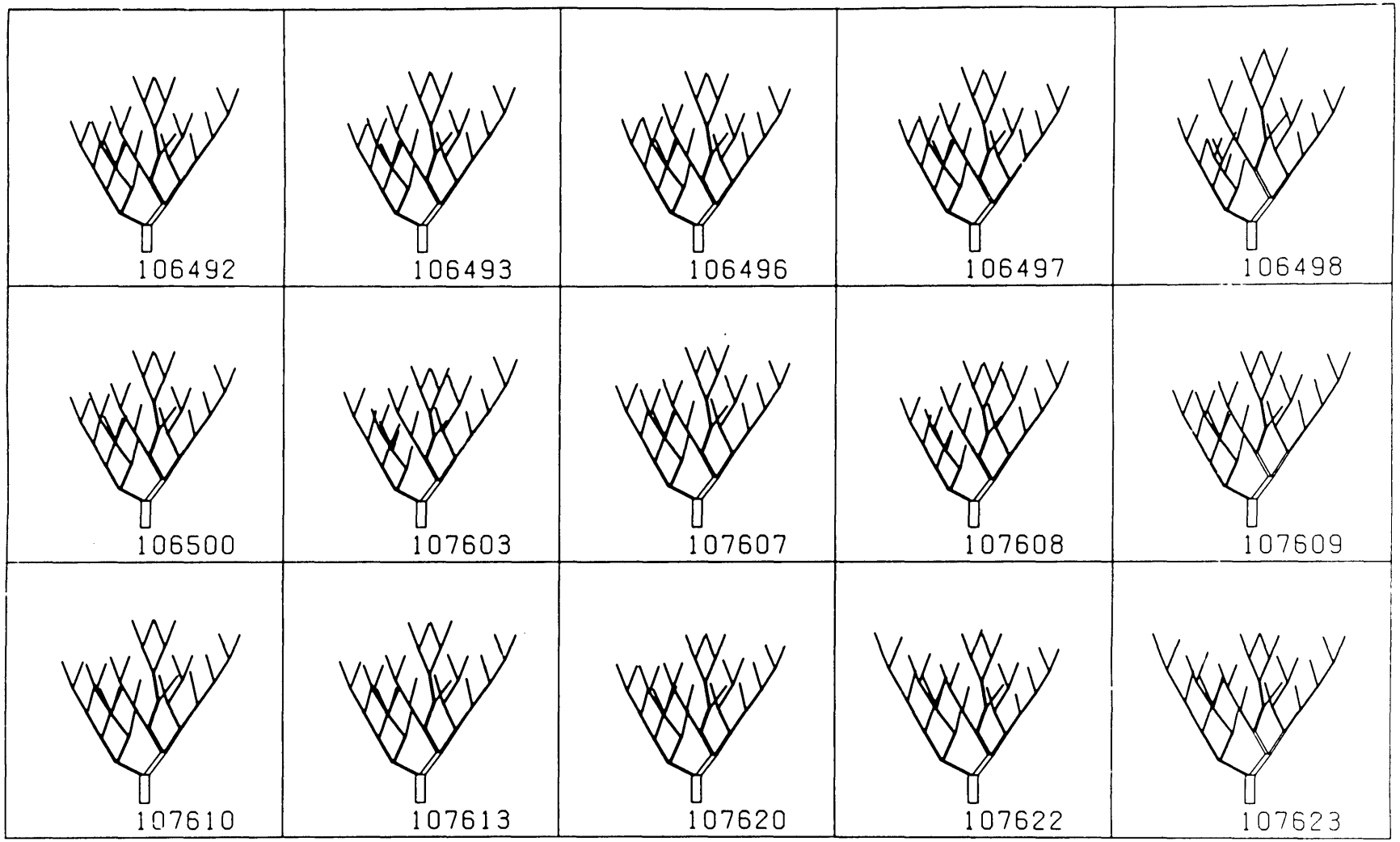
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.64



MORAY-BUCHAN TREES LOGTEN DATA

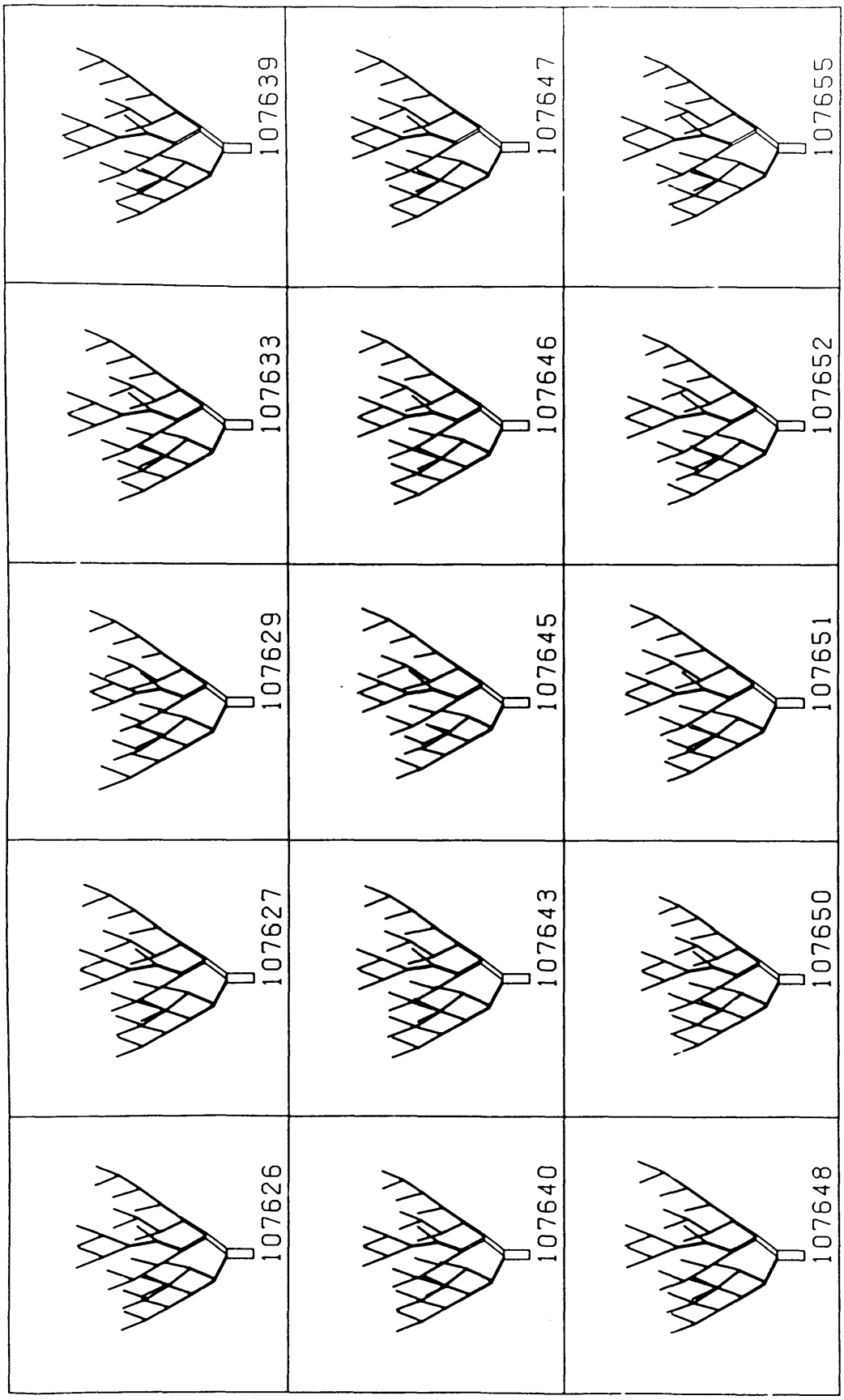
Figure 7.65



MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.66

Figure 7.67

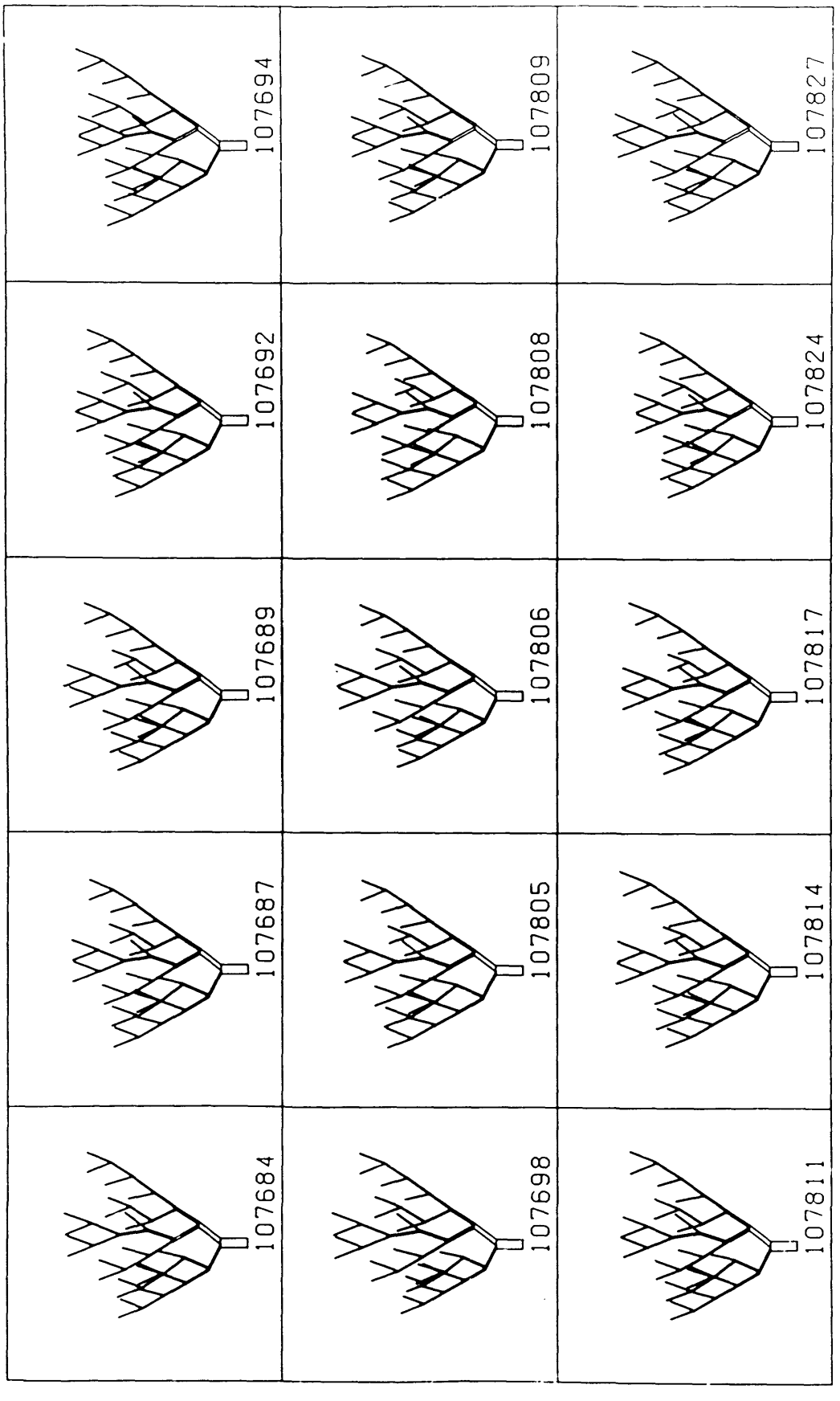


MORAY-BUCHAN TREES LOGTEN DATA

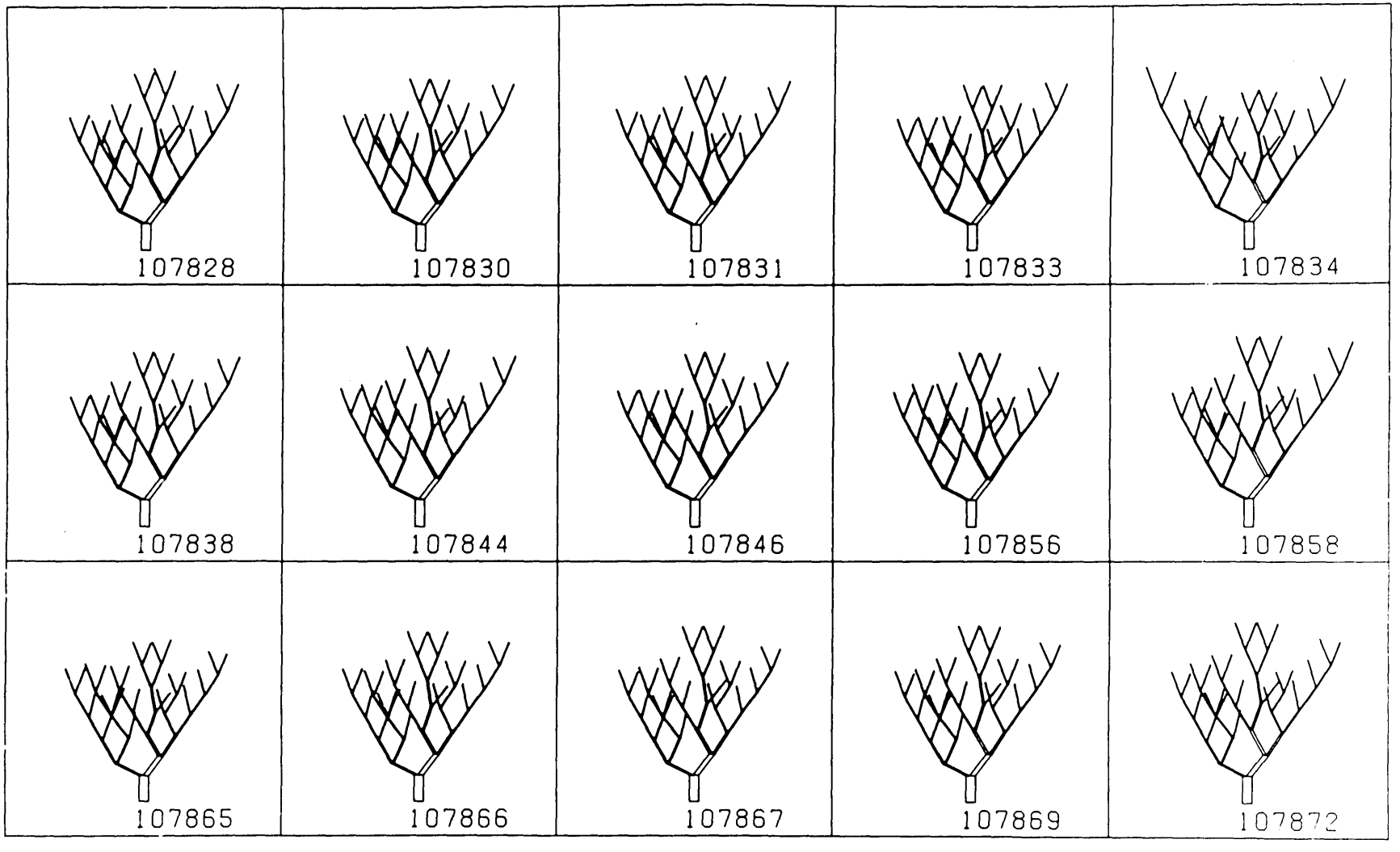
MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.68

Figure 7.69



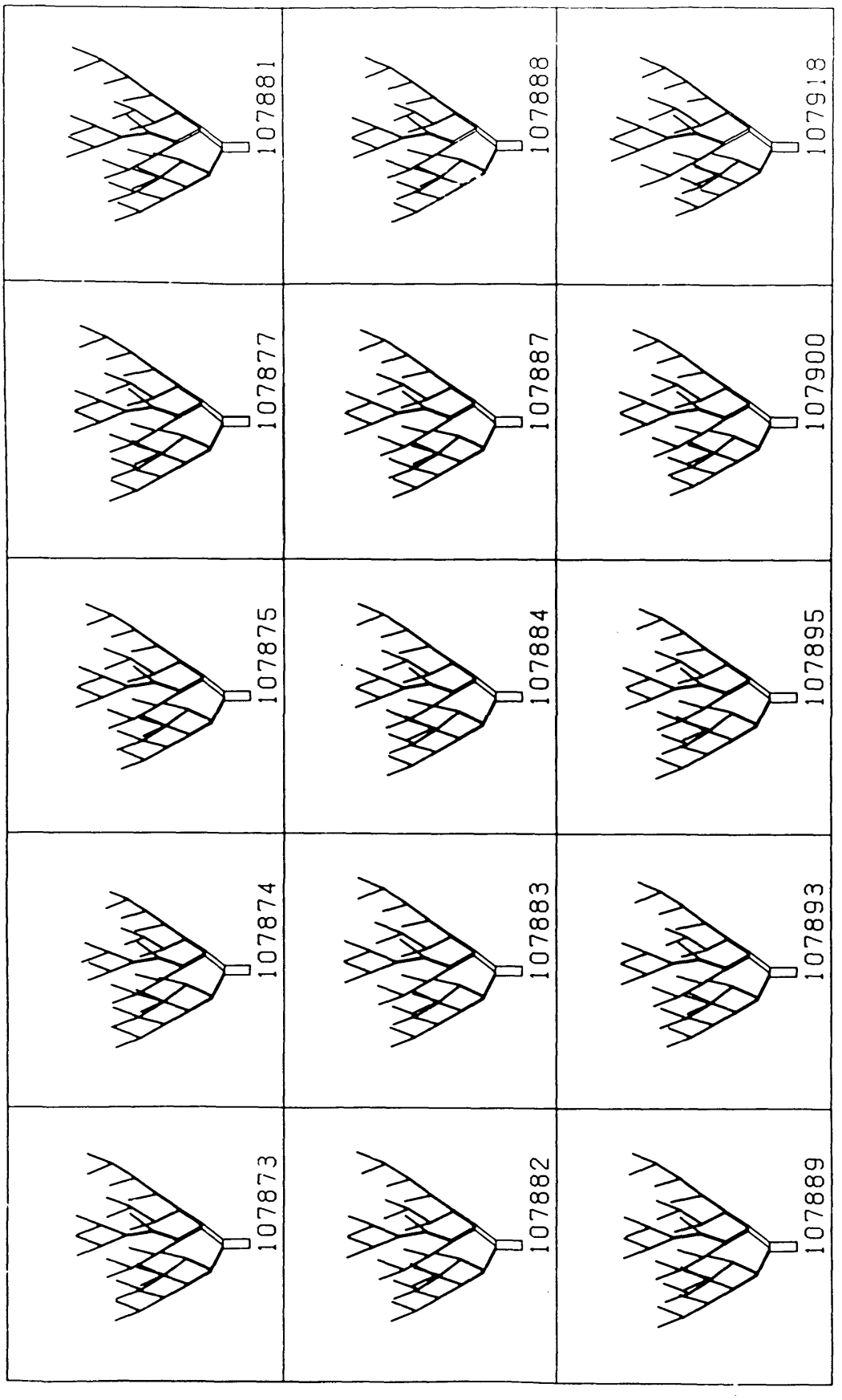
MORAY-BUCHAN TREES LOGTEN DATA



MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.70

Figure 7.71



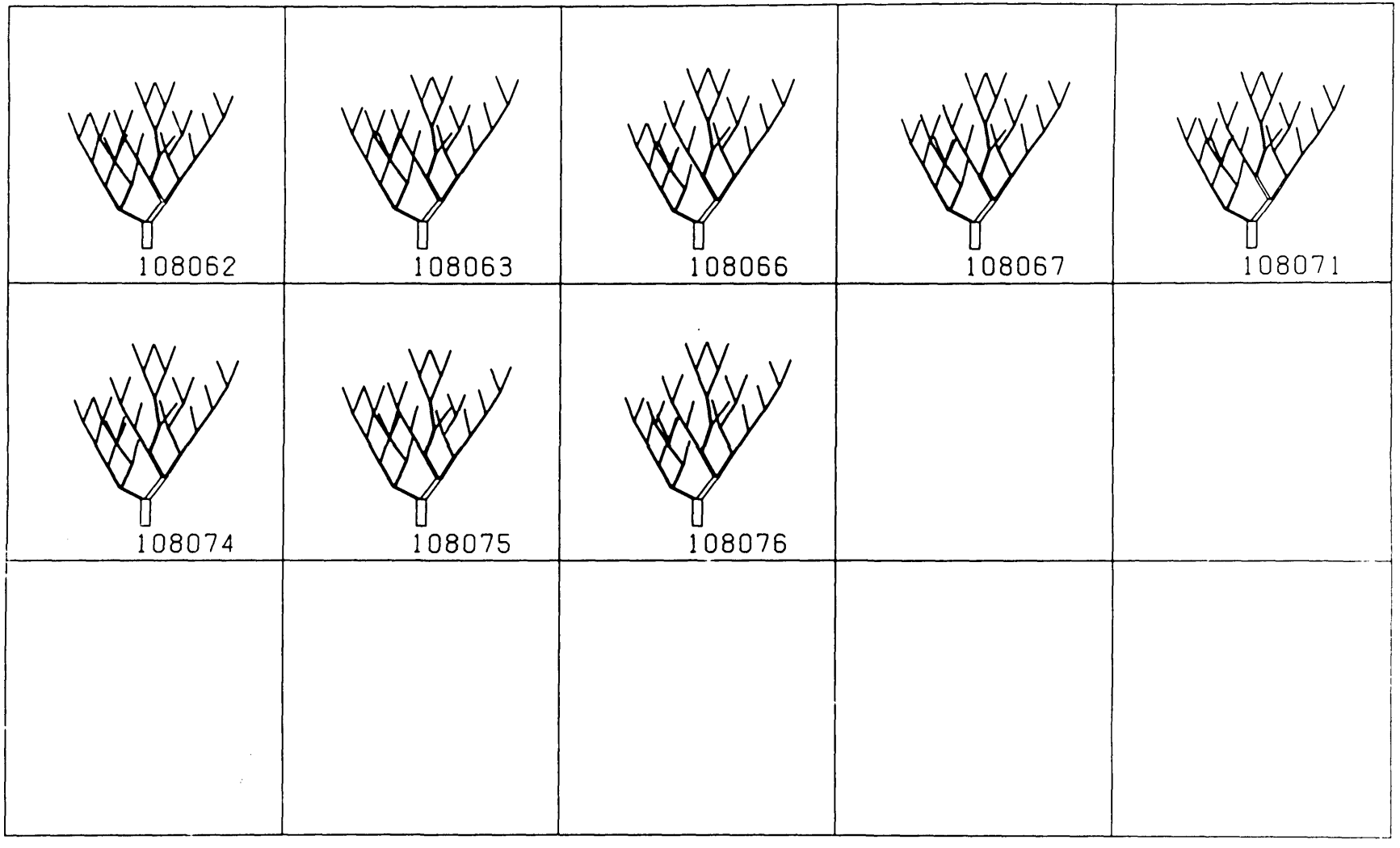
MORAY-BUCHAN TREES LOGTEN DATA

MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.72

MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.73



MORAY-BUCHAN TREES LOGTEN DATA

Figure 7.74

Figure 7.75

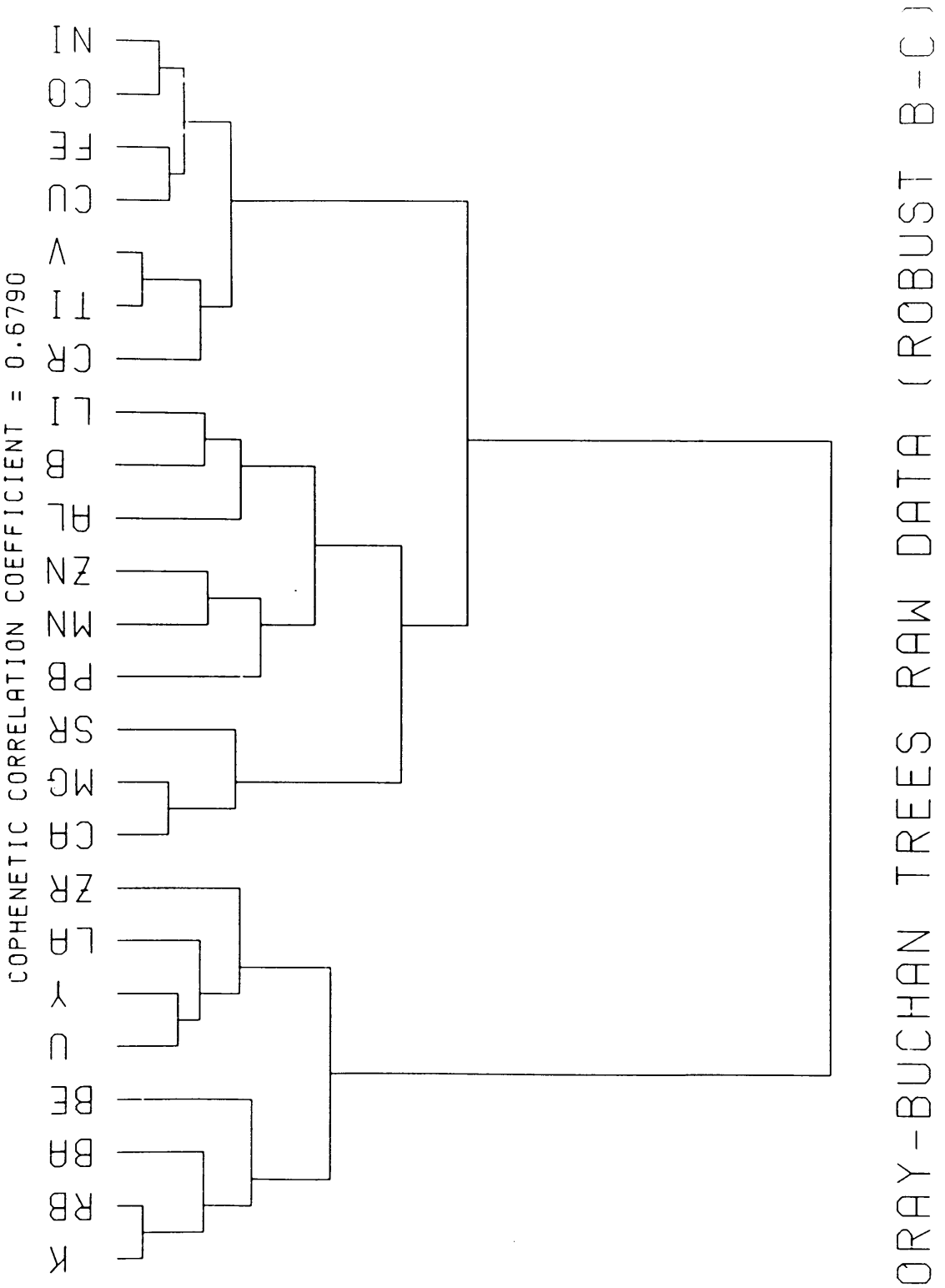
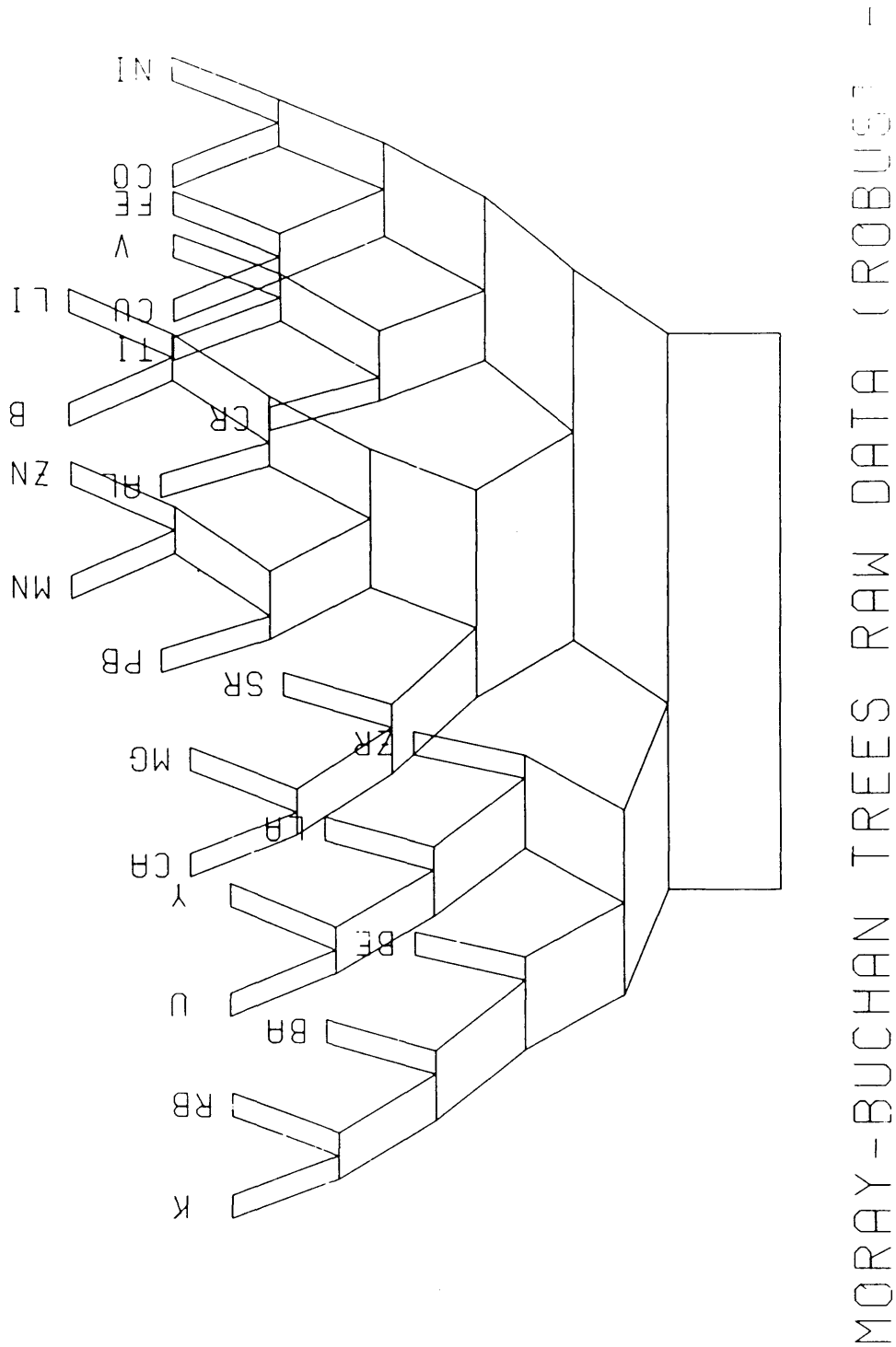

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)
COPHENETIC CORRELATION COEFFICIENT = 0.6790
K
RB
BR
BE
U
Y
LR
ZR
CR
MG
SR
PB
MN
ZN
AL
B
LI
CR
TI
V
CU
FE
CO
NI

Figure 7.76


MORAY-BUCHAN TREES RAW DATA (ROBUST
K
RB
BR
U
BE
Y
CA
LR
MG
ZR
SR
PB
MN
AL
ZN
CR
B
TI
CV
LI
V
EE
CO
N1

MORAY-BUCHAN TREES RAW DATA (ROBUST F

Figure 7.77

MORAY-BUCHAN TREES RAW DATA (ROBUST B

Figure 7.78

MORAY-BUCHAN TREES RAW DATA (ROBUST B-

Figure 7.79

MORAY-BUCHAN TREES RAW DATA (ROBUST E

Figure 7.80

MORAY-BUCHAN TREES RAW DATA (ROBUST B-

Figure 7.81

MORAY-BUCHAN TREES RAW DATA (ROBUST R

Figure 7.82

MORAY-BUCHAN TREES RAW DATA (ROBUST B-

Figure 7.83

MORAY-BUCHAN TREES RAW DATA (ROBUST B-

Figure 7.84

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C

Figure 7.85

MORAY-BUCHAN TREES RAW DATA (ROBUST B

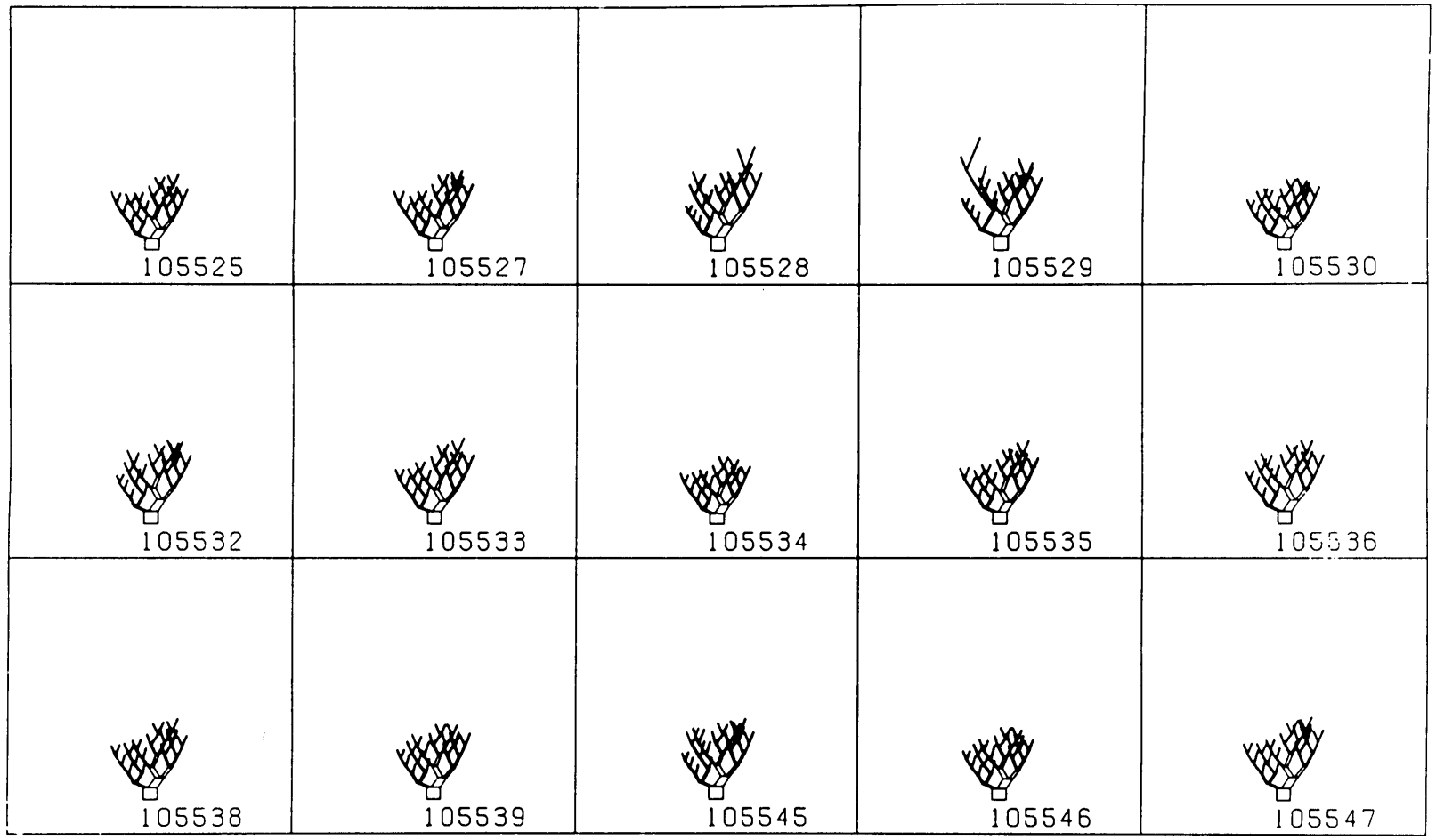
Figure 7.86

MORAY-BUCHAN TREES RAW DATA (ROBUST B-

Figure 7.87

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.88



MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.89

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C

Figure 7.90

MORAY-BUCHAN TREES RAW DATA (ROBUST B-

Figure 7.91

MORAY-BUCHAN TREES RAW DATA (ROBUST B

Figure 7.92

MORAY-BUCHAN TREES RAW DATA (ROBUST B

Figure 7.93

MORAY--BUCHAN TREES RAW DATA (ROBUST E

Figure 7.94

MORAY-BUCHAN TREES RAW DATA (ROBUST B-(

Figure 7.95

MORAY-BUCHAN TREES RAW DATA (ROBUST B-

Figure 7.96

MORAY-BUCHAN TREES RAW DATA (ROBUST B

Figure 7.97

MORAY-BUCHAN TREES RAW DATA (ROBUST F

Figure 7.98

MORAY-BUCHAN TREES RAW DATA (ROBUST B-

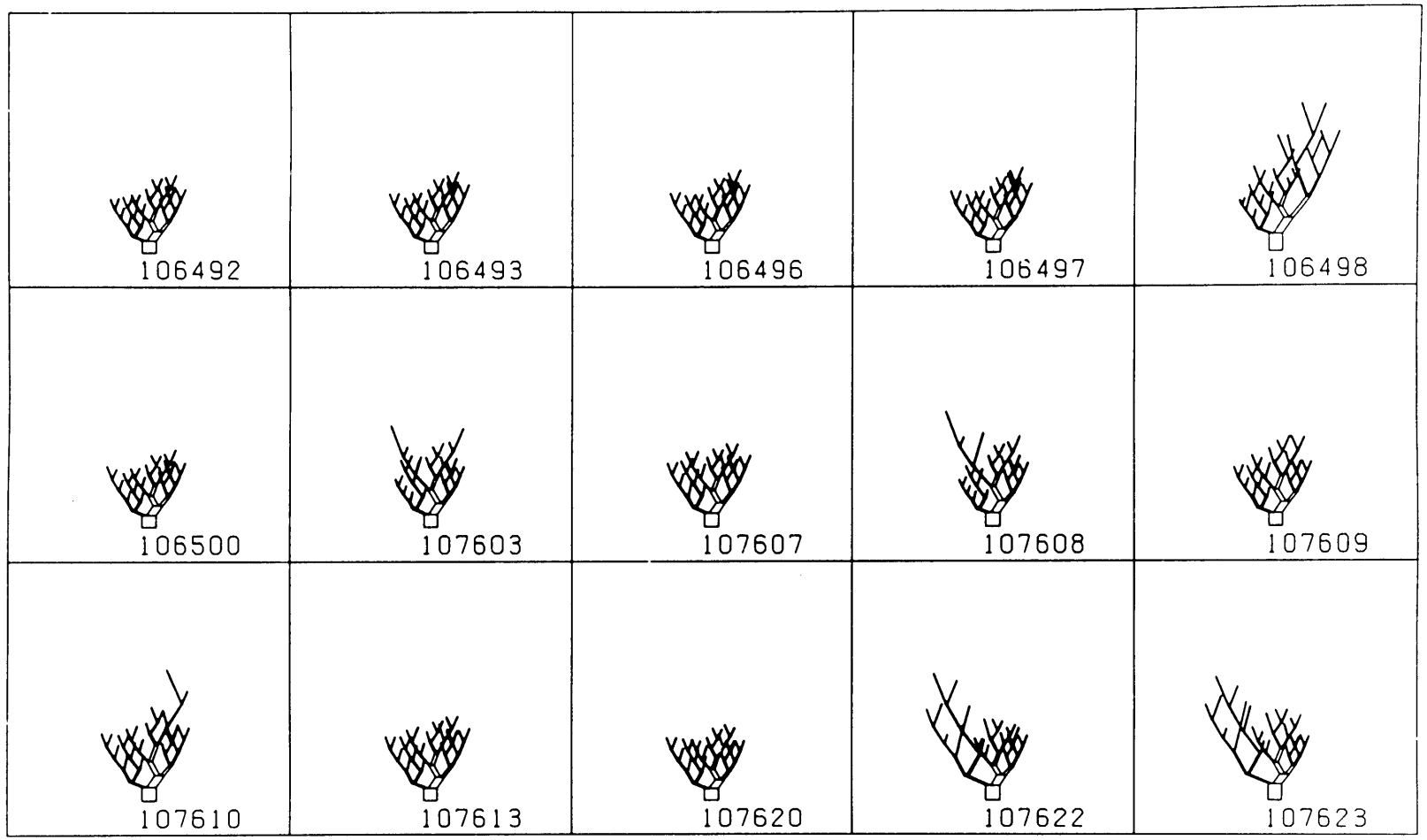
Figure 7.99

MORAY-BUCHAN TREES RAW DATA (ROBUST E

Figure 7.100

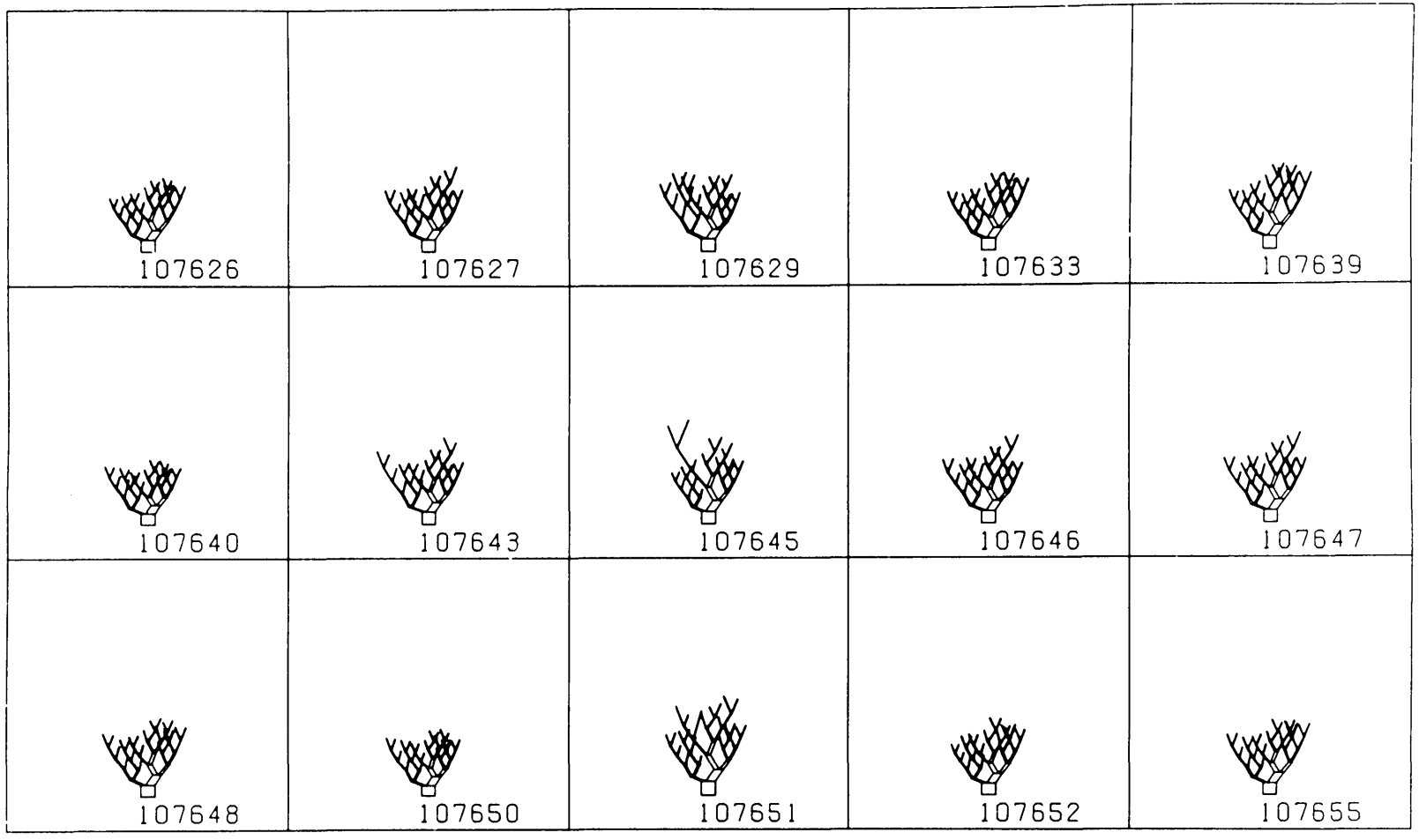
MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.101



MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.102



MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

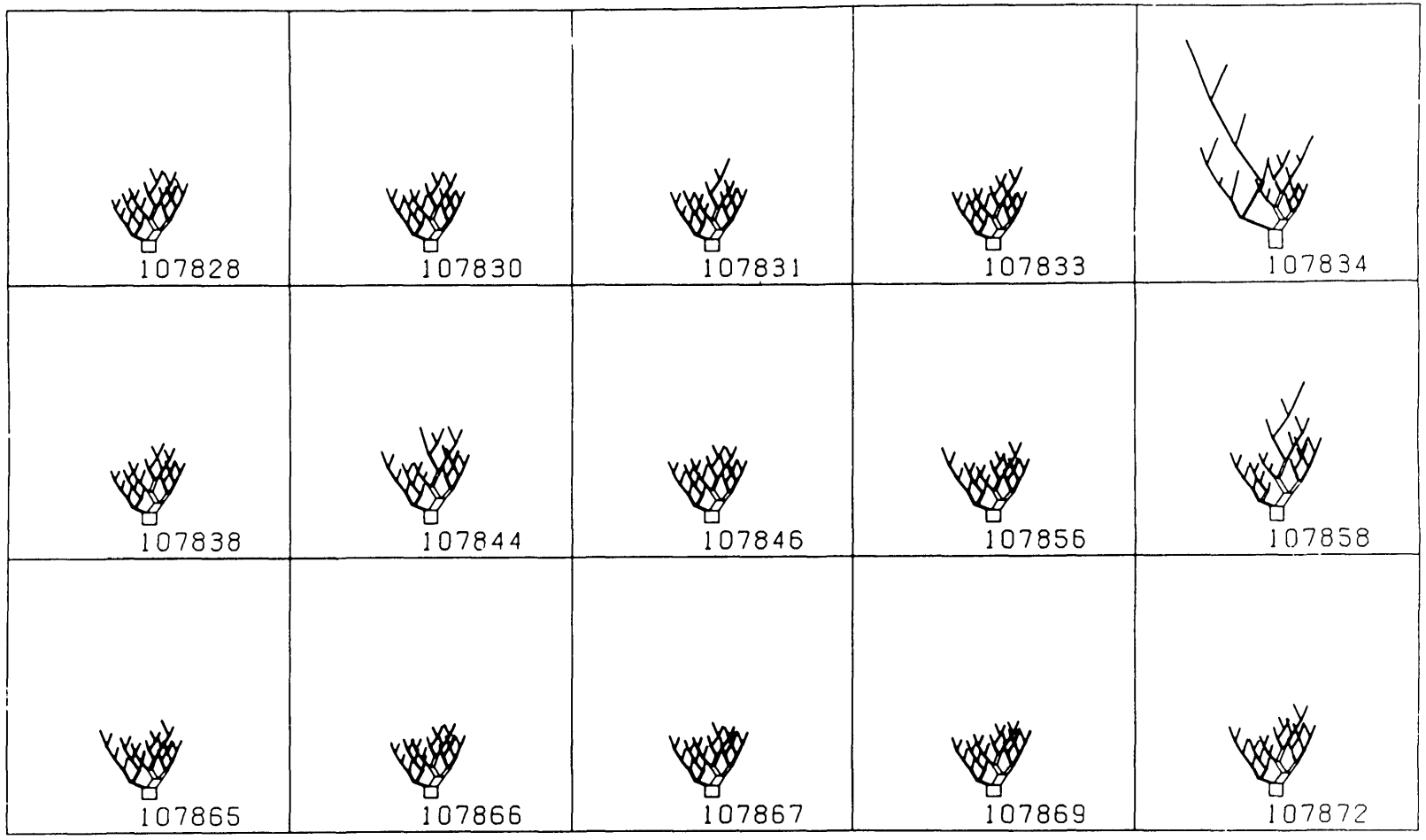
Figure 7.103

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.104

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.105



MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.106

MORAY--BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.107

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.108

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C)

Figure 7.109

MORAY-BUCHAN TREES RAW DATA (ROBUST B-C

Figure 7.110

For both runs Ward's method was used to generate the dendrogram. Using the logten correlation matrix the dendrogram showed the following variable groups:

1) U, Y, La plus Pb, Zr;

2) K, Rb, Be, Ba;

3) Ti, V, Cr plus Cu, Fe, Co, Ni, Mn;

4) Mg, Ca, Sr; and

5) B, Al, Li, Zn.

Using the boxcox robust correlation matrix the dendrogram showed the following variable groups.

1) U, Y, La, Zr;

2) K, Rb, Be, Ba;

3) Ti, V, Cr plus Cu, Fe, Co, Ni;

4) Mg, Ca, Sr; and

5) Pb, Mn, Zn plus Al, B, Li.

Apart from the different classification of Pb and Mn, the two dendrograms are the same in terms of variable groupings. Group one is probably related to felsic igneous sources, although U can also be high in black shales especially where they are mineralized (e.g. Ben Eagach Schist, see Coats et al 1981). Group two is probably related to metsediments, with the presence of Ba indicating a possible association with barian silicates (and hence Aberfeldy style mineralization). Group three is undoubtedly related to basic/ultrabasic igneous sources. Group four is probably related to metamorphosed limestone sources with a possible additional association with basic/ultrabasic sources. Group five is possibly related to the Dalradian metallogenic belt.

Plant et al (1984) propose that B and Li and elements of basic association indicate a sedimentary regime consistent with that of an extensional marginal basin - an environment suitable for Aberfeldy type stratabound mineralization. Plant et al observed that all recorded occurrences of stratabound mineralization in the Dalradian of Scotland occur in regions where B and Li levels are generally high and where Pb, Ba and Zn concentrations are elevated.

The conclusions from the two TREES runs on the MB data are as follows:

1) One of the main conclusions from the analysis of SIND

trees is reinforced, namely that data transformation dramatically reduces anomaly contrast.

2) Any graphical representation of data should enable the main features of that data to be quickly identified. It is an exploratory technique, not one that demands close analysis. With this in mind the logten trees are unsatisfactory because the visual impact of anomalous samples is reduced due to the low anomaly contrast. It is particularly difficult to visually detect anomalies in the variables whose branches are low in the tree structure i.e. Pb-Zr, K-Rb, Be, Ba (tend to be lower than and masked by groups on either side), Mn and Mg.

3) The raw data trees show how useful this method of fast visual analysis can be. The large number of branches presents much less of a problem than with the logten trees. Variables whose branches are low down in the tree can still be relatively easily visualized. Anomaly contrast is maximized through the use of raw data and the visual impact of anomalous samples is immediate. (See Appendix A1 for a list of anomalous samples obtained through analysis of these raw trees.)

7.9 Conclusions

The analysis of Kleiner-Hartigan Trees has made clear that data transformation adversely affects anomaly contrast. If anomaly detection is the aim of data analysis then the generation of trees from raw data with a robust correlation matrix provides a useful preliminary visual guide to those samples which are likely to be of most interest and also, in the cluster dendrogram, gives the major background variable relationships and associations. If determination of background trends is the main aim of the data analysis then boxcox transformed data must be used with its robust correlation matrix.

Changes in tree morphology will tend to reflect not anomalous sites but rather sites which show changes in the relative importance of the variable groupings as shown by the cluster dendrogram. Sample sites which have similar background variable associations will tend to have similar tree morphologies. Let us envisage a situation in which a data set is drawn from a population which is known to show three end-member variable associations. It should be simple to select three samples which conform closely to each of the three end members. It will then be possible to classify all the samples to one of the three end members by looking for visual similarities between each sample tree

and the three trees which represent the end members. In this way an R-mode cluster analysis is used to perform a variation of Q-mode cluster analysis. It has also been found that the variable associations seen in the background cluster dendrogram closely resemble those associations generated from analysis of the component/factor loadings matrix from PCA/FA. It is therefore possible to assign a particular tree morphology to a particular component or factor, and check quickly, by looking for visual similarities, those samples which are likely to have high scores in that component or factor.

CHAPTER EIGHT

PRINCIPAL COMPONENTS ANALYSIS

8.1 Introduction

A distinction is made in this chapter between principal components analysis (PCA) and factor analysis (FA) and no attempt is made, or indeed should be made, to place these specific techniques under the generic term factor analysis (as in Joreskog et al, 1976).

The purpose of both PCA and FA is to interpret the structure within the similarity matrix of a multivariate data set. The technique used involves extraction of the eigenvalues and eigenvectors from the similarity matrix (either the correlation matrix or the covariance matrix).

The history of FA has been described by a number of authors: Nichol et al (1969), Davis (1973), Joreskog et al (1976). Following its introduction in 1904 by C. Spearman, FA, as distinct from PCA, was developed by experimental psychologists to extract fundamental measures of intellect from scores on intelligence tests and other tests of mental ability. It is generally held that a given test does not really measure intelligence, but an individual's test score

usually is related to their mental ability. A person judged to be mentally superior will do better on all (or most) tests than a person considered to be mentally lacking. Psychologists believed FA to be capable of extracting an intelligence factor from these test scores, although none of the individual tests was capable of measuring intelligence directly.

Factor analysis was one of the first multivariate techniques to become widely used among geologists. The first published use of FA in geology was by W.C. Krumbein in 1957. Factor analysis is often regarded as a panacea for the geologist faced with complex data and little insight into it's structure, or any idea of how to obtain this insight. This, together with the complexity of the underlying hypothesis, the assumptions implicit in the factor model and the complexity of the jargon inherited from the early experimental psychologists has led to the technique of FA being frequently misunderstood and misapplied.

The purpose of this chapter is to explain the distinction between FA and PCA, examine some of the techniques for interpreting the results of PCA/FA and to evaluate the use of robust PCA/FA.

Both PCA and FA can be performed in variate space (R-mode) using a between sample similarity matrix e.g. the Pearson correlation matrix, or in sample space (Q-mode) using an angular measure of similarity e.g. $\cos^{-1}\theta$, where

θ is the angle between the variable vectors.

R-mode analysis investigates interrelationships between variables and is used primarily to identify variable groupings within a data set and as a means of reducing the dimensionality of a data set.

Q-mode analysis investigates the interrelationships between samples. The objectives of Q-mode analysis are much the same as those of cluster analysis i.e. to arrange a suite of samples into a meaningful order so the relationship between one sample and another may be deduced. Q-mode analysis is usually applied to fairly small data sets where it is used to redescribe variation in the data in terms of a few unrelated (and uncorrelated) hypothetical "end-member" samples or factors.

In this chapter discussion is restricted to R-mode analysis, since in exploration geochemistry element grouping and dimensionality reduction are of most interest.

The following works have provided the basis for much of the theoretical descriptions found in this chapter. Davis (1973) presents a very lucid description of FA and PCA. He also gives both a mathematical and geometric explanation of eigenanalysis which is the key to both FA and PCA. Temple (1978) presents a very convincing case against the use of FA and criticizes some of the long held conventions of both FA and PCA. Mancey (1980) gives a very interesting and instructive critical review of FA and PCA in geology. Howarth and Sinding-Larsen (1983) present a different

approach to the description of FA and PCA with many numerical examples of relevance to exploration geochemists. Joreskog et al (1978) devote a whole book to a description of the theory and practice of FA in geology.

8.2 FA and PCA

Let X define an n x p data matrix, where n is the number of objects and p is the number of variables or dimensions.

Let Y define an n x p normalized data matrix where y_{ij} is derived from x_{ij} by:

$$Y_{ij} = (x_{ij} - \bar{x}_j) / \sqrt{(n - 1)} \qquad \text{Equation 8.1}$$

where $\bar{x}_j$ is the mean of the jth variable.

The basic mathematical model for PCA is then:

$$Y = F\ B' \qquad \text{Equation 8.2}$$

where F is the n x p matrix of component scores and B is the p x p normalized and scaled loadings matrix.

In R-mode analysis the covariance matrix S of the original data X is defined by:

$$S = Y' Y \qquad \text{Equation 8.3}$$

and combining Equations 8.2 and 8.3 gives:

$$S = B F' F B' \qquad \text{Equation 8.4}$$

It can be shown (Joreskog et al, 1978) that the covariance matrix of the component scores equals the identity matrix:

$$F' F = I \qquad \text{Equation 8.5}$$

The component scores are, therefore, uncorrelated - this is one of the basic assumptions of PCA. From Equations 8.4 and 8.5 we have:

$$S = B B' \qquad \text{Equation 8.6}$$

The mathematical model of R-mode PCA can be expressed geometrically. In the same way that sample observations (objects) are commonly plotted in variate space with the variables represented by axes at 90° (orthogonal) to one another, so can the variables be represented by points or vectors in object space in which the n object axes are mutually orthogonal. Suppose the p variables of the n x p normalized data matrix Y are represented by p vectors (lines of known length and direction radiating from the origin) V_1, V_2,..., V_p in n-dimensional object space (see Figure 8.1). The angle (θ) between any two vectors i and j is related to the correlation between the two associated variables, such that θ_{ij} is the angle whose cosine is equal to the correlation coefficient r_{ij}. If r_{ij} is zero then θ will be 90° i.e. vectors are orthogonal; if r_{ij} is one then θ will be 0° and the vectors will be coincident. The squared lengths of the vectors are equal to the variances of the variables of the original data matrix X.

The configuration of the vectors is a graphical representation of the similarities between variables. During PCA (or FA) the configuration of vectors remains unchanged. One can think of the PCA process as inserting a new group of p orthogonal component vectors into the existing group of variable vectors and producing transform equations to define the relationships between the two. The first new vector is inserted in a position such that the

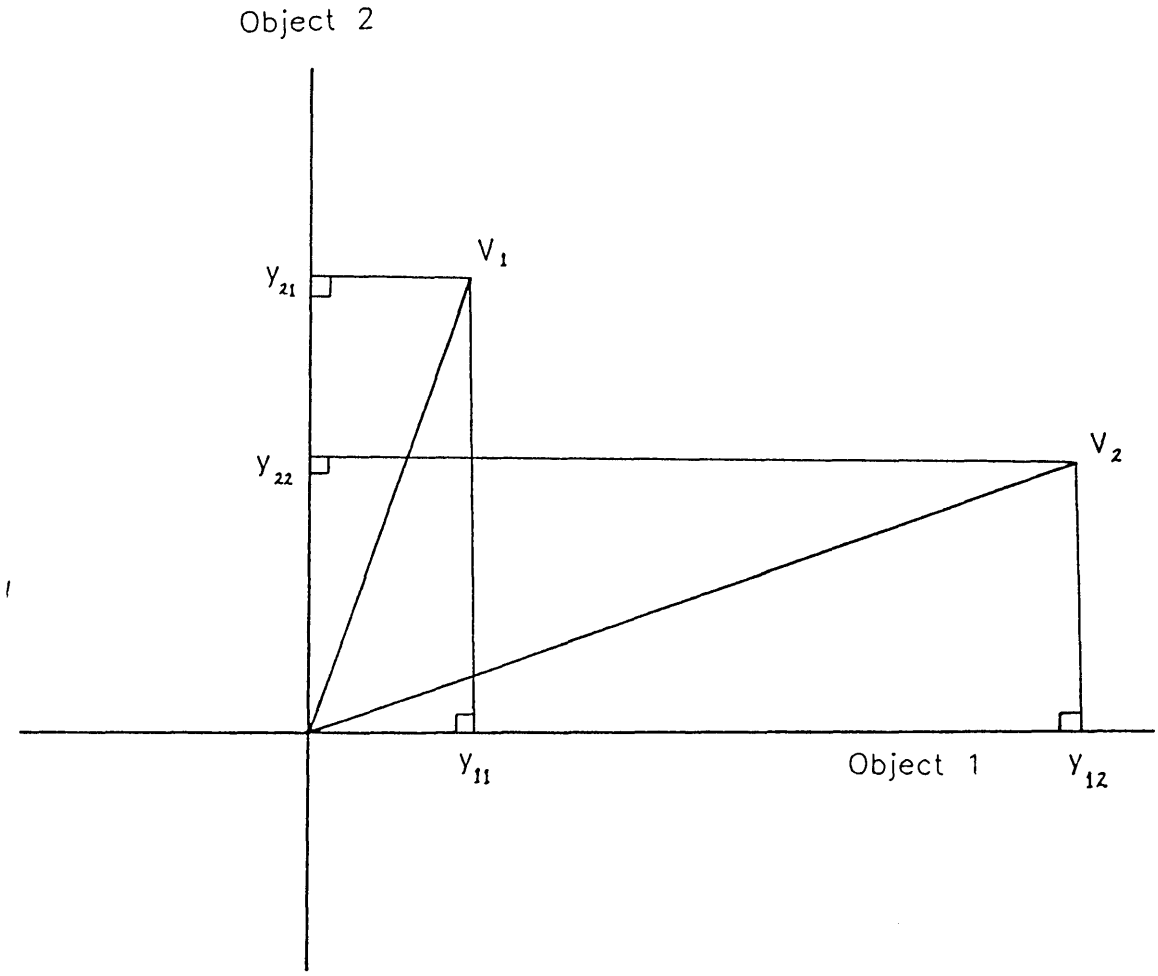


Figure 8.1

$$Y = \begin{matrix} Y_1 & Y_2 \\ Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \\ \\ Y_{n1} & Y_{n2} \end{matrix}$$

Length of vector V_1 in object space $= \sqrt{\sum_{i=1}^{n} Y_{i1}}$

Length of vector V_2 in object space $= \sqrt{\sum_{i=1}^{n} Y_{i2}}$

Where n = no. of objects.

angles it makes with the p existing variable vectors (V_1 to V_p) are all minimized. The second new vector is inserted in a position orthogonal to the first, and such that it also minimizes the angles it makes with the p original variable vectors. Subsequent vectors are inserted using the same criteria.

The generation of these new component vectors is achieved by calculation of the eigenvectors and eigenvalues of the covariance matrix of the original data matrix X (see definitions in the Glossary). Any component vector, say the i-th, can be defined by a direction term - the normalized eigenvector P_i whose length is unity - and by a second term specifying the length of the component vector. The latter term is the square root of the corresponding eigenvalue $K^{1/2}$. As we have seen, each eigenvector is successively positioned so as to minimize the angles it makes with the original variable vectors, given the constraint of orthogonality with one another. This also means that the eigenvectors are positioned so as to successively coincide with the direction of maximum variance given the constraint of orthogonality. The eigenvector associated with the largest eigenvalue determines the direction of maximum variance of the variable vectors in object space (Figure 8.2) or of the object vectors (data points) in variate space. This variance, which is equivalent to the sum of the squared projections of all the variable vectors on the eigenvector, is largest along the first eigenvector (Figure

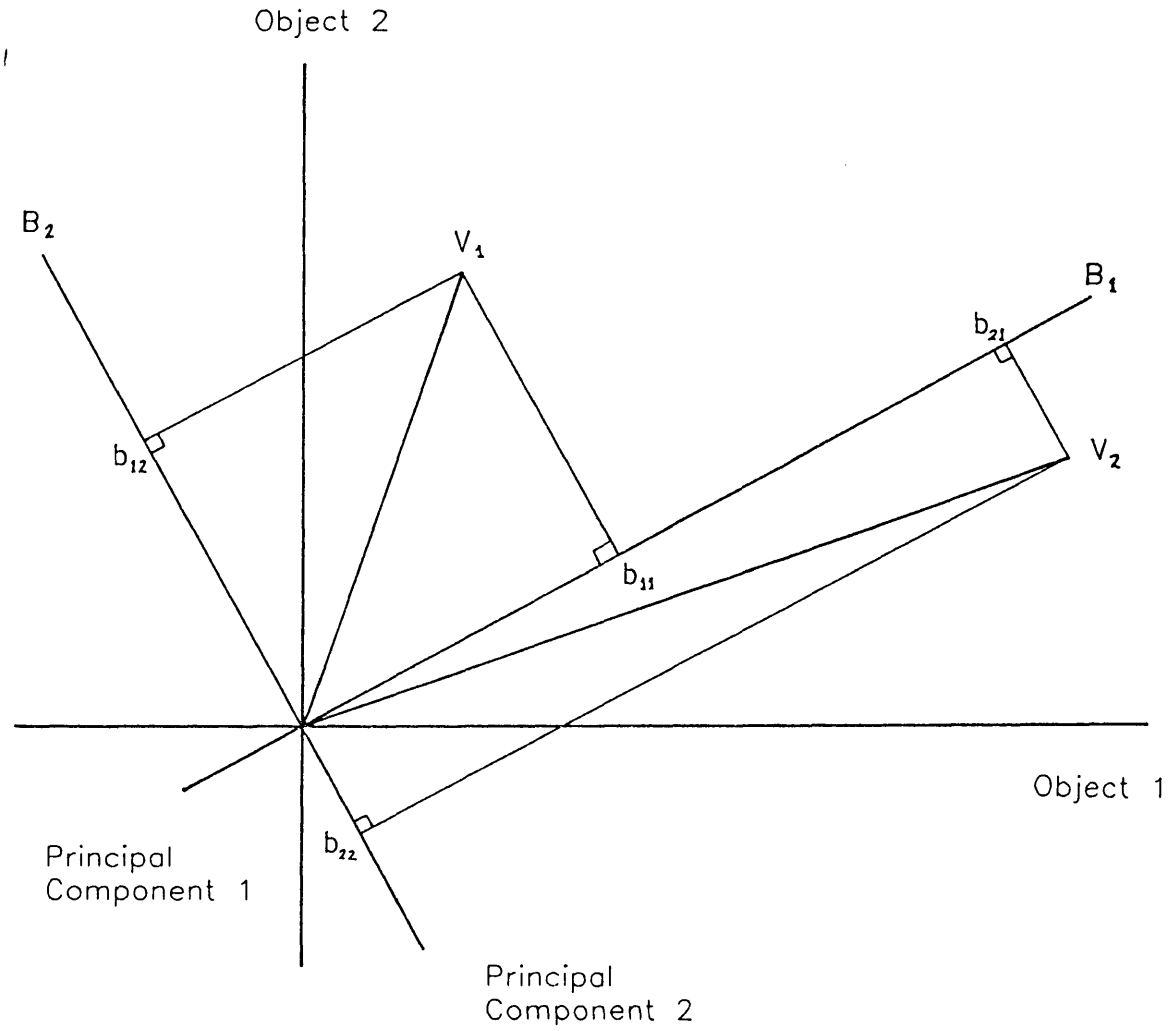


Figure 8.2

$$B = \begin{matrix} B_1 & B_2 \\ b_{11} & b_{12} \\ b_{21} & b_{22} \end{matrix}$$

Length of vector V_1 in component space = $\sqrt{\sum_{i=1}^{k} b_{1i}}$

Length of vector V_2 in component space = $\sqrt{\sum_{i=1}^{k} b_{2i}}$

Where k = no. of principal components.

8.2). This variance is equal to the associated eigenvalue. The eigenvector associated with the second largest eigenvalue locates the direction of maximum variance orthogonal to the first and so on.

The best way to explain the geometrical relationship between the eigenvectors and the original variable vectors is to visualize the eigenvectors in variate space (Figure 8.4). The coordinates of each of the n objects of the transformed data matrix Y can also be visualized as vectors or points in variate space. Assuming that the variables are normally distributed, the data points can be enclosed by equal density contours. For bivariate, normally distributed, variables the contours will form ellipses about the origin. If the variables are uncorrelated, the ellipses will describe a circle; if perfectly correlated they will degenerate into a line. For more than two variables, the data points can be viewed as forming a p-dimensional hyperellipsoid, where p equals the number of variables. The major and minor axes of the hyperellipsoid can be found by calculating the eigenvectors and eigenvalues of the covariance matrix of the original data, X. The eigenvector associated with the largest eigenvalue will define the major axis of the hyperellipsoid, which is also the direction of maximum variance of the data points. The eigenvector associated with the second largest eigenvalue defines the minor axis of the hyperellipsoid orthogonal to the first which accounts for the maximum of

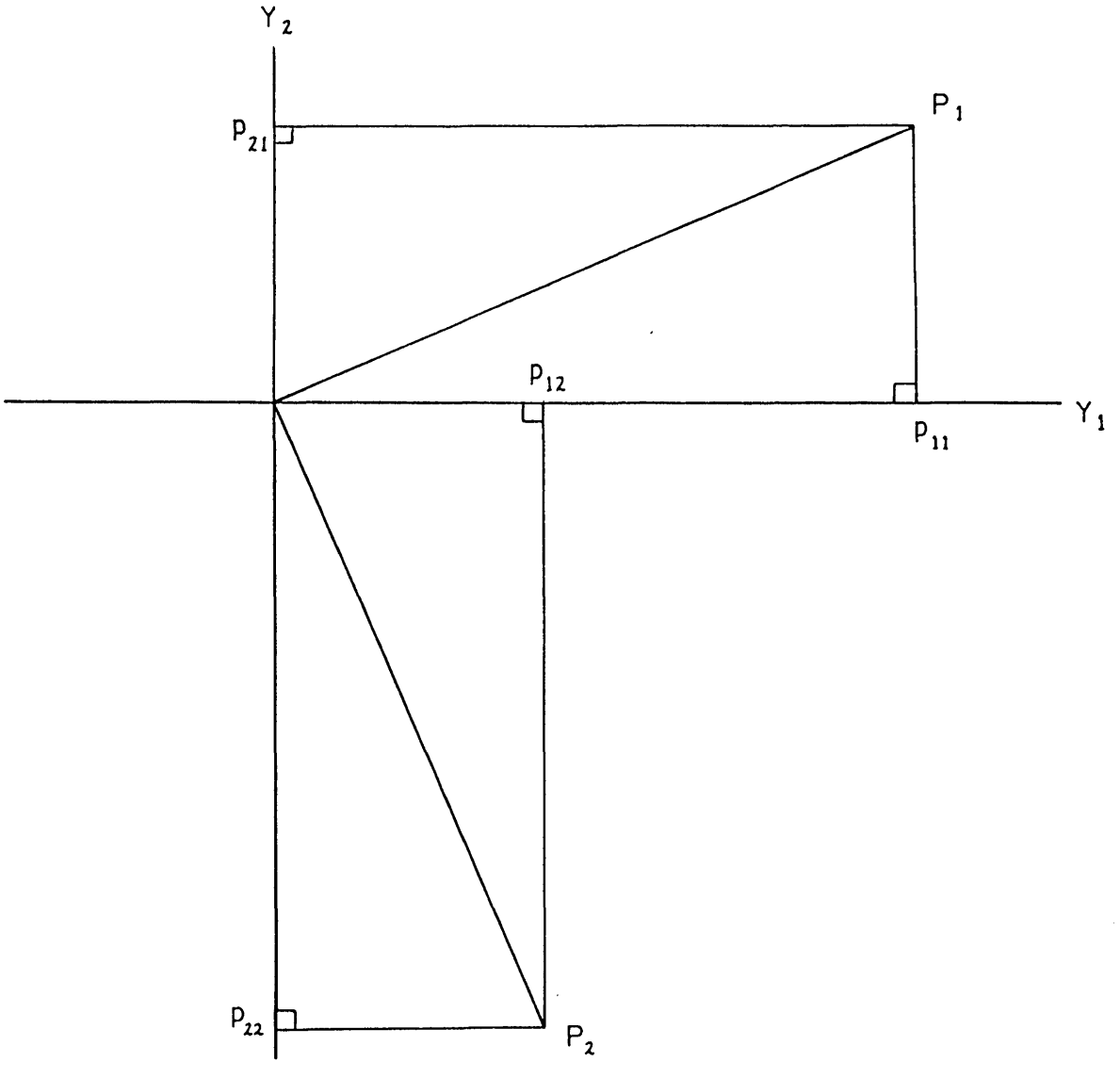


Figure 8.3

$$P = \begin{matrix} & P_1 & P_2 \\ & p_{11} & p_{12} \\ & p_{21} & p_{22} \end{matrix}$$

Length of vector P_1 in variate space $= \sqrt{\sum_{i=1}^{p} p_{i1}} = 1.0$

Length of vector P_2 in variate space $= \sqrt{\sum_{i=1}^{p} p_{i2}} = 1.0$

Where p = no. of variables

Length of vector Y_1 in component space $= \sqrt{\sum_{i=1}^{k} p_{1i}}$

Length of vector Y_2 in component space $= \sqrt{\sum_{i=1}^{k} p_{2i}}$

Where k = no. of components

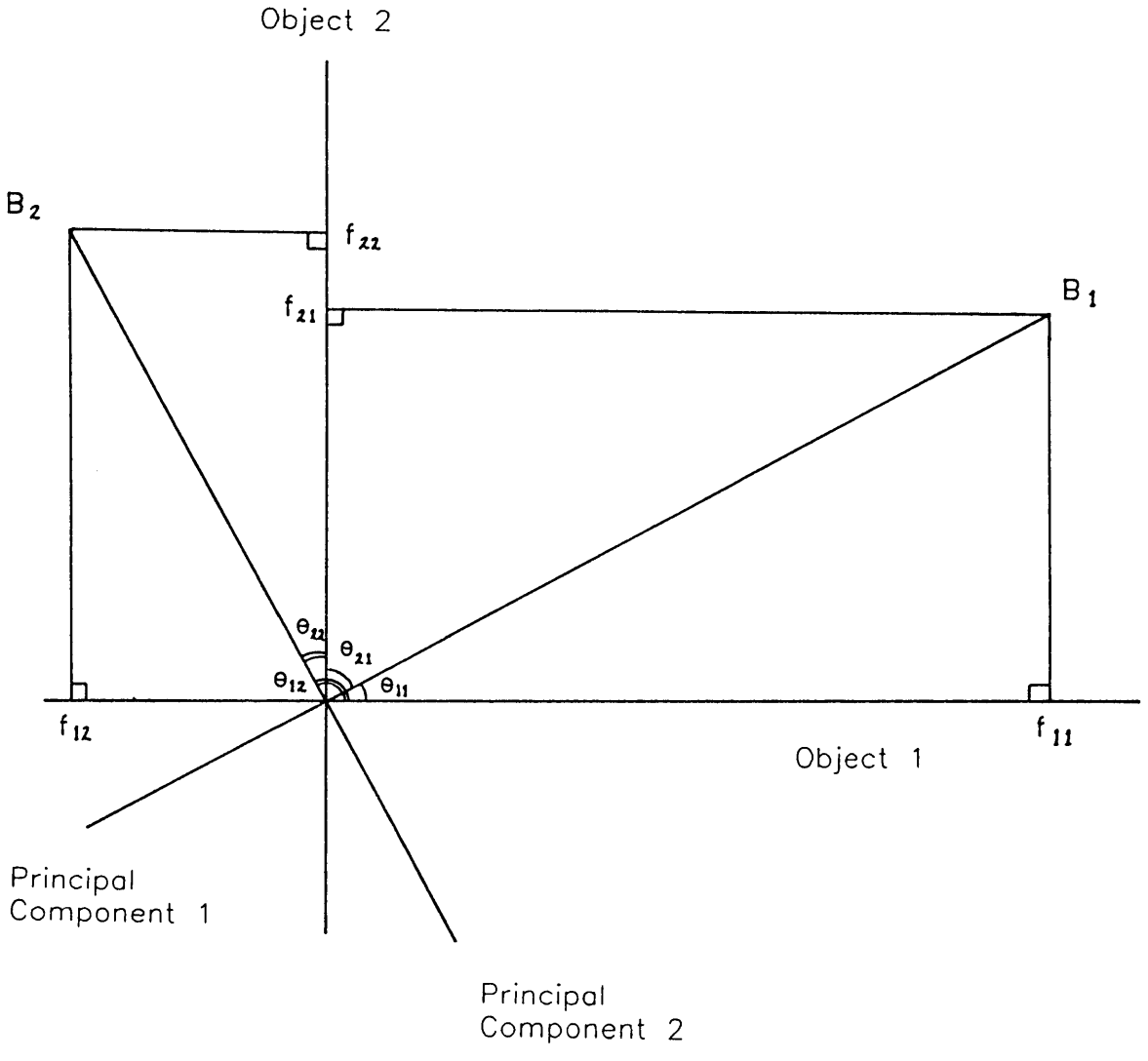


Figure 8.4

$$F = \begin{matrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ & \\ f_{n1} & f_{n2} \end{matrix}$$

Cos θ_{11} = f_{11} assuming length of B_1 is unity.

the remnant variance in the data points, and so on.

The location of eigenvectors may be viewed as a problem of rotation or transformation. The original axes, representing the variables, are rotated to new positions according to the criteria reviewed above. The normalized eigenvector matrix P is a p x p matrix consisting of p columns each representing the coordinates in p variate space of each of the p normalized eigenvectors (see Figure 8.4). A normalized eigenvector is one whose length is unity. Recall that the eigenvector gives only the direction of the component vector, while the square root of the corresponding eigenvalue gives its length.

The elements of the normalized eigenvector matrix P can be considered as the cosines of the angles of rotation from the old to a new system of coordinate axes (see Figure 8.4). For instance:

$$\cos \theta_{11} = \frac{p_{11}}{\text{Length of } P_1} = p_{11} = r_{11}$$

since the length of the first eigenvector P_1 is unity. Also, since the cosine of the angle between two vectors is equal to their correlation coefficient, the value p_{11} can be equated to the correlation coefficient r_{11} between the first variable and the first eigenvector.The other elements of the eigenvector matrix may be analysed in a similar

manner. We can therefore regard the p x p matrix of normalized eigenvectors as a matrix of correlations between the p variables (Y_1, Y_2,..., Y_p) and the p new eigenvectors (P_1, P_2,..., P_p) and as such will contain only values betwen -1 and +1. We can think of the p eigenvectors in the same way as the p variable vectors - as p new variables. The elements of the eigenvector matrix are usually termed loadings of the original variables on the new variables - the whole matrix being termed the loadings matrix. The loadings matrix is conventionally arranged so that the original variables are represented by the rows and the components by the columns of the matrix.

The relation between the raw data matrix X and their eigenvectors and eigenvalues is usually expressed by:

$$S\ P = P\ K \qquad \text{Equation 8.7}$$

where S is the covariance matrix of the raw data X, P is the matrix of normalized eigenvectors and K is the diagonal matrix of eigenvalues. The matrix P has the following properties:

$$P\ P' = P'\ P = I \qquad \text{Equation 8.8}$$

$$P' = P = P^{-1}$$

It is, however, usual in both PCA and FA to scale the normalized eigenvectors so that their length, instead of being unity, is equal to the magnitude of their eigenvalue. This is achieved by multiplying the loadings matrix P by the diagonal matrix of the square roots of the eigenvalues:

$$B = P\ K^{1/2} \qquad \text{Equation 8.9}$$

Previously when we wanted to completely describe the set of component vectors we required two matrices: the normalized eigenvector matrix P to describe direction, and the diagonal eigenvalue matrix K to describe length. Now we can completely describe the direction and length of the set of component vectors by one matrix - the normalized and scaled loadings matrix, B.

The sum of squared loadings across each row of the loadings matrix B is equal to the variance accounted for by the original variable corresponding to that row. The sums of squares are termed communalities (see Figure 8.2). This may be shown by the equation:

$$B\ B' = S \qquad \text{Equation 8.10}$$

which is equivalent to Equation 8.6 (above). The communalities of the matrix B are given by the diagonal elements of the matrix S and are, therefore, the variances of the original variables (X_1, X_2,..., X_p). Consequently, the sum of the communalities of the matrix B is equal to the trace of the matrix S (i.e. the sum of the diagonal elements of S) which is equal to the total variance of the data, X. The square roots of the communalities give the lengths of the variable vectors (V_1, V_2,..., V_p) in component space (see Figure 8.2).

The sum of the squared loadings down each column of the matrix B is equal to the variance along the eigenvector corresponding to that column (see Figure 8.2). This variance is equal to the associated eigenvalue. This can be shown by the equation:

$$B' B = K \qquad \text{Equation 8.11}$$

The square root of the eigenvalue can be used as the standard deviation of the "new" variable represented by the eigenvector. The square root of each eigenvalue is also equal to the length of the associated eigenvector.

Note that the sum of the communalities of any loadings

matrix is equal to the sum of the associated eigenvalues, both of which are equal to the total variance of the original data from which the matrix was generated.

It is necessary at this point to introduce the concept of component or factor scores, since they will be mentioned in subsequent sections. Scores will be fully explained in Section 8.8. In the same way that the data matrix Y represents the coordinates of the p variable vectors V_1, V_2,..., V_p in n-dimensional object space (see Figure 8.1), so can we generate a new data matrix F comprising the coordinates of the p new orthogonal component vectors B_1, B_2,..., B_p in n-dimensional object space (see Figure 8.3). This new data matrix F is termed the matrix of component scores. Alternatively the component scores may be envisaged as the coordinates of each of the n data points in the p-dimensional space of the orthogonal component vectors.

In summary, the eigenvectors of a covariance matrix represent a set of orthogonal axes to which the variable or object vectors may be referred. They are a set of linearly independent basis vectors whose positions are located so as to account for the maximum variance (given by the associated eigenvalues) of the original data matrix.

8.3 Standardization and the Similarity Matrix

PCA and FA are concerned with interpreting the structure of the variance-covariance matrix obtained from a collection of multivariate observations. It is frequently the case in geology that the variables are expressed in a number of different units of measurement and are thus not all directly comparable. It is possible to convert a data set into standardized or unitless form by subtracting from each observation the mean of the data set and dividing by the standard deviation. This procedure ensures that all variables are directly comparable.

If it is decided to perform PCA or FA on a standardized data matrix the first step, as we have seen, is to calculate the variance-covariance matrix. It is, however, not necessary to standardize each individual item of data before computing this matrix, because the covariance matrix of standardized data is identical to the correlation matrix of the original (unstandardized) data. In practice the correlation matrix is almost always used and its use is assumed in later discussions.

Both the covariance matrix and the correlation matrix are parametric summaries of the data and so the data should be normally distributed for the most reliable results. A data transformation procedure may, therefore, be necessary

(see Chapter 5). However, data transformation has adverse effects on the component scores - an important analytical tool in exploration geochemistry (see section 8.9). A preferred alternative is to use a robust estimate of the correlation matrix, thus maintaining the integrity of the original data values.

The fact that the covariance matrix of standardized data can be calculated without actually standardizing the data can, and has in some cases, led to miscalculation of component scores. Trochimczyk and Chayes (1978) point out that data transformed by the loadings matrix into component space (i.e. component scores) should be uncorrelated. Spurious correlations between components can result if loadings are calculated from the correlation matrix (R) and scores from unstandardized data (X), or loadings from the covariance matrix (S) and scores from standardized data (Z). Correlated scores can also result from the use of a robust correlation matrix, where the correlations are due to the presence of outliers in the data which are not accounted for in the robust matrix. Trochimczyk and Chayes propose adherence to the following rules:

1) If the eigenanalysis is performed on the correlation matrix then the data should be standardized prior to the scores being calculated.

2) If the covariance matrix is used the unstandardized

data should be used to compute the scores.

Let us now see what effect the use of a standardized data matrix has on the PCA/FA model as described in section 8.2.

Let Z define an n x p standardized data matrix where z_{ij} is derived from x_{ij} by:

$$z_{ij} = (x_{ij} - \bar{x}_j) / s_j$$

where $\bar{x}_j$ and s_j are the mean and standard deviation respectively of the jth variable.

The covariance matrix of standardized data is usually denoted by R, the symbol for the correlation matrix of the original data X. The matrix R is defined by:

$$R = Z' Z / \sqrt{(n - 1)} \qquad \text{Equation 8.12}$$

The basic mathematical model for PCA/FA now becomes:

$$Z = F A'$$

where A is the p x p normalized and scaled loadings matrix generated from the correlation matrix R.

The relation between standardized data and their eigenvectors and eigenvalues is expressed similarly to Equation 8.7:

$$R\ Q = Q\ \Lambda \qquad \text{Equation 8.13}$$

where Q is the normalized eigenvector matrix and Λ is the diagonal matrix of eigenvalues. Since the normalized eigenvectors are orthogonal, the matrix Q, like its equivalent the matrix P, has the property:

$$Q\ Q' = Q'\ Q = I$$

$$Q = Q' = Q^{-1}$$

However, $Q \neq P$ and $\Lambda \neq K$.

It is usual, as explained above, to scale the normalized eigenvectors:

$$A = Q \Lambda^{1/2} \qquad \text{Equation 8.14}$$

where A is the scaled and normalized loadings matrix. The communalities of the matrix A are equal to the variances of the original standardized variables. Therefore, the communalities of the matrix A are, by definition (since standardized data have mean equal to zero and standard deviation equal to one) always equal to unity. This is the case in PCA where the number of principal components always equals the number of variables. As we shall see in section 8.4, in FA, where the number of principal components is less than the number of variables, the factors (as the components are termed in FA) cannot account for all the original variance of the variables and so the communalities will be less than one.

The "standardized" equivalents of Equations 8.10 and 8.11 are respectively:

$$A' A = R \qquad \text{Equation 8.15}$$

$$A A' = \Lambda \qquad \text{Equation 8.16}$$

8.4 Factor Analysis

PCA consists of a linear transformation of p original variables to p new variables (components), where each new variable is a linear combination of the old. The process is performed in a fashion that requires that each new variable accounts for, successively, as much of the total variance as possible. When p new variables have been computed, all of the original variance will be accounted for. Nothing is said about probability, hypotheses or testing because PCA is not, strictly, a statistical procedure. Rather, it is a mathematical manipulation. However, it assumes some of the properties of statistical procedures when decisions are made to discard some components as being of little significance.

FA is somewhat different in that it is commonly regarded as a statistical technique. It relies on a set of assumptions about the nature of the parent population from which the samples were drawn. In FA the relationship within a set of p original variables is regarded as reflecting the correlations of each of the p variables with k mutually uncorrelated underlying factors. The usual assumption is that $k < p$. Total variance in the p original variables is therefore derived from variance in the k factors, but in addition a contribution is made by unique sources which

independently affect the p original variables. The k underlying factors are referred to as common factors, while the independent contribution is summarized as a unique factor. The factor model may be expressed as:

$$X = F B' + E \qquad \text{Equation 8.17}$$

where X is an n x p raw data matrix, F is an n x k matrix of factor scores, B is a p x k matrix of normalized and scaled factor loadings and E is an n x p matrix of unique error terms (residuals).

The hypothesis underlying FA may be expressed in the following way. The observed p x p covariance matrix S of the original raw data is the product of a p x k factor loadings matrix B, multiplied by its transpose B', plus a p x p matrix of unique variances E:

$$S = B B' + E \qquad \text{Equation 8.18}$$

Multiplying a p x k matrix by its transpose will create a p x p matrix which, however, will have only k positive eigenvalues and associated eigenvectors. If k = p, the matrix E will vanish and the problem is equivalent to PCA.

In FA, however, $k < p$ and estimates must be obtained of the factor loadings matrix B and the matrix of unique variances E. The factor model requires that k, the number of factors, be known prior to analysis. This implies that the analyst has some insight into the probable nature of the factors and can predict a suitable number of factors to be extracted.

The eigenanalysis in FA is the same as for PCA; however the operation is performed always on the standardized covariance (or correlation) matrix R.

The loadings matrix is arranged so that there are k columns (relating to the number of factors) and p rows (relating to the number of original variables). If we square the elements in the loadings matrix and sum within each row, we obtain the communalities of that matrix. If we were to extract p factors from a p x p covariance matrix, the communalities would be equal to the original variances. Because we are working with standardized variables (and therefore the correlation matrix) these communalities would all be unity. However, if (as in FA) we extract only k factors, the resulting communalities will be less than the original variances and can provide an index to the efficiency of the reduced set of factors. The residual variance for each variable is equal to one minus the communality for that variable and is equivalent to the unique factor associated with that variable.

As might be expected, the magnitude of the communalities

is dependent on the number of factors that are retained. This poses one of the major problems associated with FA: how many factors should be retained? Unfortunately there is no universally agreed method for determining the number of retained factors. The best empirical guide is from the eigenvalues, as these indicate the amount of variance accounted for by successive factors. Mancey (1980) found the most useful criterion to be Cattell's Scree-test. This involves plotting the ordered eigenvalues for each factor (expressed as a percentage of the sum of the eigenvalues) against the corresponding factor number. A break in slope is associated with the correct number of factors which should be used to represent the data. Another commonly used method is to retain all factors with eigenvalues greater than one. In other words, retain all factors which contain greater variance than the original standardized variables.

8.5 Uses of PCA and FA

PCA and FA are used in two principal ways:

1) as a means to identify variable groupings within the data and

2) as a means of reducing the dimensionality of the

data.

Let us treat these in order.

Variable Grouping

The normalized and scaled loadings matrix A contains the correlation coefficients (i.e. the cosines of the angles) between each variable vector and each component vector. A group of variables whose vectors all make small angles with, say, the first eigenvector, will have high loadings in the first column of the loadings matrix A and may be interpreted as being associated for some reason. In this sense PCA/FA performs a similar function to cluster analysis. However, in most geological applications the extracted components are assigned to certain geological processes, influences or environments. For example, suppose the first component extracted from a data set of marine sediments had high loadings in Mn, Co, Ni, Cu and Zn; then this component might be interpreted as an authigenic Mn oxide component representing the adsorption by Mn as it comes out of solution of Co, Ni, Cu and Zn. Sample sites with a high score on this particular component can be pinpointed easily on a map of the scores and may be interpreted as locations where authigenic Mn oxide

precipitation is most active.

In most applications the components are interpreted in terms of lithological or secondary environment associations. However, it is often the case when analysing geochemical data from a mineral exploration program that an attempt is made to identify a "mineralization" component, with the hope that sample sites with high scores on this component will reveal locations of potential mineralization. In most circumstances the extraction of a mineralization component from the sample correlation matrix is very unlikely, because components represent overall trends, not minor associations. Indeed, in practice, it is advisable to produce a correlation matrix prior to PCA/FA which best characterizes the overall trends and background relationships in the data and supresses any anomalous trends. This may be achieved by Boxcox transforming the raw data and then, as an extra precaution, calculating a robust correlation matrix from the transformed data (see Chapter 5 for more details).

Reduction of Dimensionality

Reduction of dimensionality is assumed as part of the FA model. The analyst must specify k, the number of factors to retain, where k is less than p, the dimensionality of the

original data.

In PCA reduction of dimensionality may be achieved simply by using only those eigenvectors (components) corresponding to the largest eigenvalues.

A new data matrix of reduced dimensionality can be obtained by computing the scores on the retained factors or components. Maps of these scores may be used to show the spatial distribution of the variable groupings associated with each component or factor. While it is usually the scores on the first few components which are studied in this way, Gnanadesikan and Kettenring (1972) and Gnanadesikan (1977) have suggested that the scores on the principal components corresponding to the smallest eigenvalues would be of interest in detecting multivariate outliers. This is because the first eigenvector lies in such a position as to explain the most variation, i.e. it lies along the major axis of the data cloud in p-dimensional variate space. Any points which depart significantly from this axis could be considered as outliers and will show up best on the minor axes orthogonal to this axis, i.e. on components with small eigenvalues. So in general the extreme values (outliers) should show as high scores on the last few components (see Section 8.10 for more details).

8.6 Summary of the Differences Between PCA and FA

Although the techniques of PCA and FA have much in common and are often referred to in the literature as one and the same process, there are important mathematical and conceptual differences between them. According to Joreskog et al (1976) the contrast is that in PCA components are determined so as to account for the maximum variance of all the observed variables, whereas in FA a given number of factors (less than the number of variables) are defined so as to account maximally for the common intercorrelations between the observed variables. Thus, PCA can be said to be variance oriented, whereas FA is correlation oriented.

In PCA the total variance of the original variables is completely accounted for by all the p components. In FA, however, total variance is partitioned between that due to intervariable correlations (the common factors measured as communalities) and that due to variable variances (the unique factor). Temple (1978) comments that the partitioning of observed variances into communalities and uniqueness is such an obscure concept that its interpretation does not appear to have been attempted in geological applications of FA. The distinction is often, in practice, abandoned: the unique factor is set to zero and the PCA solution $Q \Lambda^{1/2}$ is taken as the estimate of the

loadings matrix A.

FA assumes the existence of a small number of common factors. In principle this number is specified before the analysis and is less than the number of original variables. This assumption may be justified in some branches of science, but most geological problems are so complex that the number of underlying factors cannot be known in advance, nor can it even be known whether this number is smaller or larger than the number of observed variables. PCA, however, makes no assumptions about the number of components to be extracted. When PCA is being used as a dimensionality reduction method, only those eigenvectors corresponding to the largest eigenvalues may be used in the interpretation, but these eigenvectors are not the only significant ones.

8.7 Rotation

The estimation of the loadings matrix in FA does not yield a unique solution. This indeterminacy means that the k orthogonal factors can be rigidly rotated without affecting the estimates of communalities or uniqueness. Advantage is taken of this indeterminacy for the factors to be rotated to more "meaningful" or "interpretable" positions relative to the original variables. Various

methods of rotation are available, the most commonly used being the orthogonal Varimax method of Kaiser which maximizes the variances of the squared factor loadings on the original variables, that is, tends to produce a mixture of ones and zeros (or their approximations) among the loadings, so that the factors can be easily interpreted in terms of the original variables. Alternatively, the orthogonality of the factors can be abandoned (so that the factors become intercorrelated), in which case the factors can be obliquely rotated to whatever position the analyst thinks fit.

Temple (1978) presents a convincing argument against factor rotation. He dismisses the technique of oblique rotation as departing too far from the original data. In contrast, Khaleelee (1969) investigated FA and methods of factor rotation for the interpretation of stream sediment data and came to the conclusion that oblique rotation gave the most interpretable results and was preferrable because the underlying geological factors were likely to be correlated. Despite this, I take a similar view to Temple that oblique rotation is too subjective a technique to be of consistent use in PCA or FA.

Temple is also opposed to orthogonal rotation and makes the point that the eigenvectors are meaningful only in their original relation to the variate axes and to the corresponding eigenvalues: if you rotate the first eigenvector then it is no longer in the direction of

maximum variance. Temple also makes the point that the purpose of rotation is to move the factor axes to positions that correspond to the centre of clusters of variable vectors. However, rotation of the factor axes does not alter the relative positions of the vectors in space: clusters remain clusters whether or not a factor has been rotated into them. Rotated factors are, therefore, no better than unrotated ones for depicting the spatial relationships of variables in factor space.

Despite Temple's arguments, factor rotation (usually the orthogonal Kaiser's Varimax method) is a common practice in the geosciences and is generally considered to provide a more interpretable solution.

8.8 Component and Factor Scores

Principal component scores are the projection of the original objects onto the new component axes. Thus, just as each sample has a value for each variable, so it has a score for each component. The methods for calculating component scores for PCA and FA are mathematically distinct from each other.

In PCA the n x p matrix of component scores F is calculated from the n x p matrix of original data X and the p x p normalized loadings matrix P:

$$F = X\ (P')^{-1} \qquad \text{Equation 8.19}$$

However, since $P' = P = P^{-1}$ then Equation 8.19 may be written:

$$F = X\ P \qquad \text{Equation 8.20}$$

If the loadings matrix is generated from the correlation matrix, i.e. if we are using standardized data, then Equation 8.20 becomes:

$$F = Z\ Q \qquad \text{Equation 8.21}$$

If the loadings matrix is scaled as well as normalized the relationship in Equation 8.8 does not operate. In other words $B' \neq B \neq B^{-1}$ and $A' \neq A \neq A^{-1}$. Therefore the jump from Equation 8.19 to Equation 8.20 is no longer operable. We have instead:

$$F = X\ (B')^{-1}$$ Equation 8.22

or, if standardized data is used then we have:

$$F = Z\ (A')^{-1}$$ Equation 8.23

In PCA where all the components are retained the Equations 8.22 and 8.23 are sufficient. However, when fewer factors than variables are used, the solution is more complex. Assuming k factors and p variables the FA model (assuming E is negligible) is:

$$Z = F\ A'$$

Postmultiplication by A gives:

$$Z\ A = F\ A'\ A$$

Finally, after postmultiplication by $(A'A)^{-1}$ we get:

$$F = Z\ A\ (A'\ A)^{-1} \qquad \text{Equation 8.24}$$

This expression is fairly general and can be used for PCA or FA.

From Equation 8.5 we know that factor scores should be uncorrelated. However Trochimczyk and Chayes (1978) note that if loadings are calculated from the correlation matrix and scores from unstandardized data, or loadings from the covariance matrix and scores from standardized data, the result will be to introduce correlations between the transformed variables (scores). Correlations between scores may also be introduced if loadings are calculated from a robust correlation matrix.

8.9 Robust PCA

PCA is a mathematical technique which, unlike FA, does not have to rely on subjective assumptions and hypotheses. PCA is therefore theoretically preferrable to FA.

As we have seen, it is the primary purpose of PCA to determine the background relationships and associations within a particular data matrix. These relationships and associations are characterized by the loadings matrix. It

is, however, the correlation matrix - the summary of the structure of the original data - that governs the loadings matrix. Thus, in order for the loadings matrix to characterize the background relationships and associations it is necessary for the correlation matrix to do so too. As we have seen in Chapter 6, the correlation coefficient is sensitive to the presence of outliers and as a consequence so is the PC solution. Figure 8.5 shows how a single outlier may distort the eigenanalysis. As a result of the presence of outliers the correlation matrix seldom characterizes the main (background) structure of the data, unless either the data are transformed to normality in some manner (see Chapter 5) or a robust estimate of the matrix is generated (see Chapter 6). Once we are sure that the correlation matrix, and hence the loadings matrix, does characterize the background structure of the data, the original data may be used to calculate the component scores. It is appropriate to regard the scores as residuals, i.e. that which remains after a summary or fitted model (characterized by the loadings matrix) has been "subtracted" from the data according to the schematic equation:

$$\text{Residual} = \text{Data} - \text{Fit} \qquad \text{Equation 8.25}$$

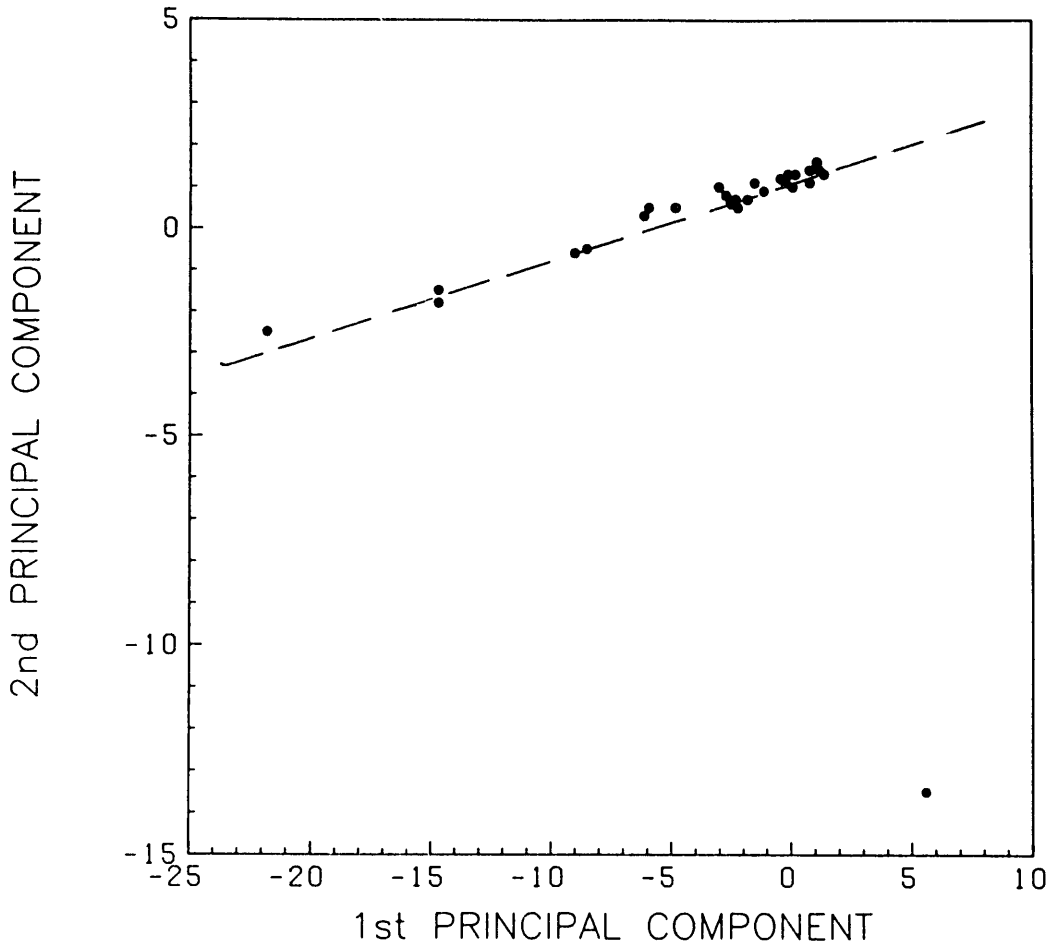


Figure 8.5

Example showing the influence of outliers on principal components analysis. Data are annual observations of 14 economic variables for 29 chemical companies. If the outlier were not present in the lower right corner, the first principal component would follow the dashed line and the second principal component would be negligible because the correlation coefficient of the remaining 28 points is $r = 0.99$ (from Devlin et al 1981)

It seems reasonable that if we can define the background trends of a data set correctly, via the loadings matrix, then any anomalies present in the data will be revealed in the residuals or scores. It also seems reasonable to suppose that the anomaly contrast will be maximized in such a situation. However, if the loadings matrix is distorted by the presence of outliers then the scores are likely to be similarly distorted and the anomaly contrast reduced.

There are two ways to ensure that the correlation matrix represents background trends in a data set: by data transformation to normality, or by a robust estimate of the correlation matrix. If the data are transformed, outliers will tend to be suppressed and background trends accentuated. If PCA is then performed on this transformed data the correlation matrix and hence the loadings matrix will tend to characterize the background structure of the data. Variable associations revealed in the loadings matrix can be regarded as real and not induced by distortions due to anomalies. However, when the transformed data matrix is used to calculate component scores it is extremely unlikely that these scores will reveal any of the outliers that may be present in the untransformed data. This is due simply to the fact that any outliers have been transformed out of the data. It seems logical therefore to calculate component scores from the original untranformed data, so that once a background fit has been "subtracted" from this data any

anomalies present should be revealed in the scores. What is required then is a method of generating a correlation matrix which represents background trends, while the data matrix remains intact. This requirement is met by performing standard eigenanalysis on a robust estimate of the correlation matrix.

8.10 PCA/FA and the SIND Data

The purpose of this comparative analysis is to investigate the effect on the loadings matrix and component scores of:

1) data transformation;

2) robust estimation of the correlation matrix; and

3) rotation of the components.

In addition a solution to the question: "is PCA or FA the better technique?" is attempted.

8.10.1 Data Transformation

To investigate the effect of data transformation on PCA, the loadings matrices and component scores generated from raw, logten and boxcox data are compared. The loadings matrices for the three data sets (Table 8.1) are analysed as follows.

In all three data sets Zn, Cd and Cu have high loadings on PC1 (the change in sign has no significance here). The highest loadings for these three variables are with raw data. Logten and boxcox transformation cause a significant decrease in these loadings with boxcox being slightly lower on average than logten. This decrease is matched by an increase in the loadings for Fe, Mn and Pb. These movements can be explained by looking at the correlation matrices (Table 6.2) from which the loadings matrices were generated. Transformation causes an increase in the correlations between the Zn-Cd-Cu group and the Fe-Mn-Pb group. This has the effect of bringing the variable vectors of the two groups closer together, which means that PC1 (which predominantly follows the Zn, Cd and Cu vectors) will have a greater correlation with the vectors of Fe, Mn and Pb. Greater correlation between component and variable equals greater loading.

For raw data, Pb and Mn have high loadings on PC2 while

Table 8.1

Component Loadings Matrices Raw, Logten, Boxcox SIND Data (Rot')

	PC1	PC2	PC3	PC4
Zn	0.9903	0.0643	−0.0880	−0.0277
	−0.9175	−0.1102	0.2406	0.2359
	−0.8223	0.1335	−0.2644	0.3591
Fe	0.1897	0.5747	−0.7948	−0.0443
	−0.3866	−0.3579	0.3870	0.7565
	−0.4537	0.3228	−0.3322	0.7604
Mn	0.1198	0.8698	−0.2541	−0.4054
	−0.2587	−0.3775	0.8422	0.2845
	−0.2882	0.3881	−0.8352	0.2599
Cd	0.9465	0.2452	−0.1293	−0.1052
	−0.8960	−0.1385	0.2528	0.1905
	−0.9304	0.1295	−0.2410	0.1453
Cu	0.9856	−0.0685	−0.0617	0.0527
	−0.9790	−0.0783	0.0771	0.1442
	−0.9380	0.1308	−0.1274	0.2475
Pb	0.0197	0.9626	−0.2175	0.1598
	−0.0947	−0.9387	0.2665	0.1967
	−0.1222	0.9387	−0.2649	0.1828

for logten and boxcox data only Pb is highly loaded. This again can be explained by examination of the correlation matrices. Pb and Mn have a high correlation in raw data which is significantly reduced on tranformation. The close association of the Pb and Mn vectors in raw data causes the correlations or loadings between these vectors and PC2 (which preferentially follows the Pb vector) to be high. On transformation the association between Pb and Mn is significantly reduced and PC2 follows the Pb vector to the exclusion of all others.

Component scores can best be analysed using probability plots (Figures 8.6 through 8.11). The high loadings for Zn, Cd and Cu on PC1 should mean that those samples with high concentrations in those elements will be revealed by high scores on that component. This is true for all three data sets although the contrast between the anomalous samples (numbers 13, 15 and 16) and the background is much reduced (in comparison with the raw data Q-Q plot) for logten and boxcox transformed data. This is due partly to the reduced contrast between anomalous samples and background in the transformed data, and partly to the increased Fe, Mn and Pb loadings and decreased Zn, Cd and Cu loadings on PC1.

The split between PC2, PC3 AND PC4 of the high loadings for Fe, Mn and Pb means that no single Q-Q plot can reveal both of the relevant anomalous samples (22 and 23). For the raw data the Q-Q plot for PC2 scores (highly loaded in

Figure 8.6

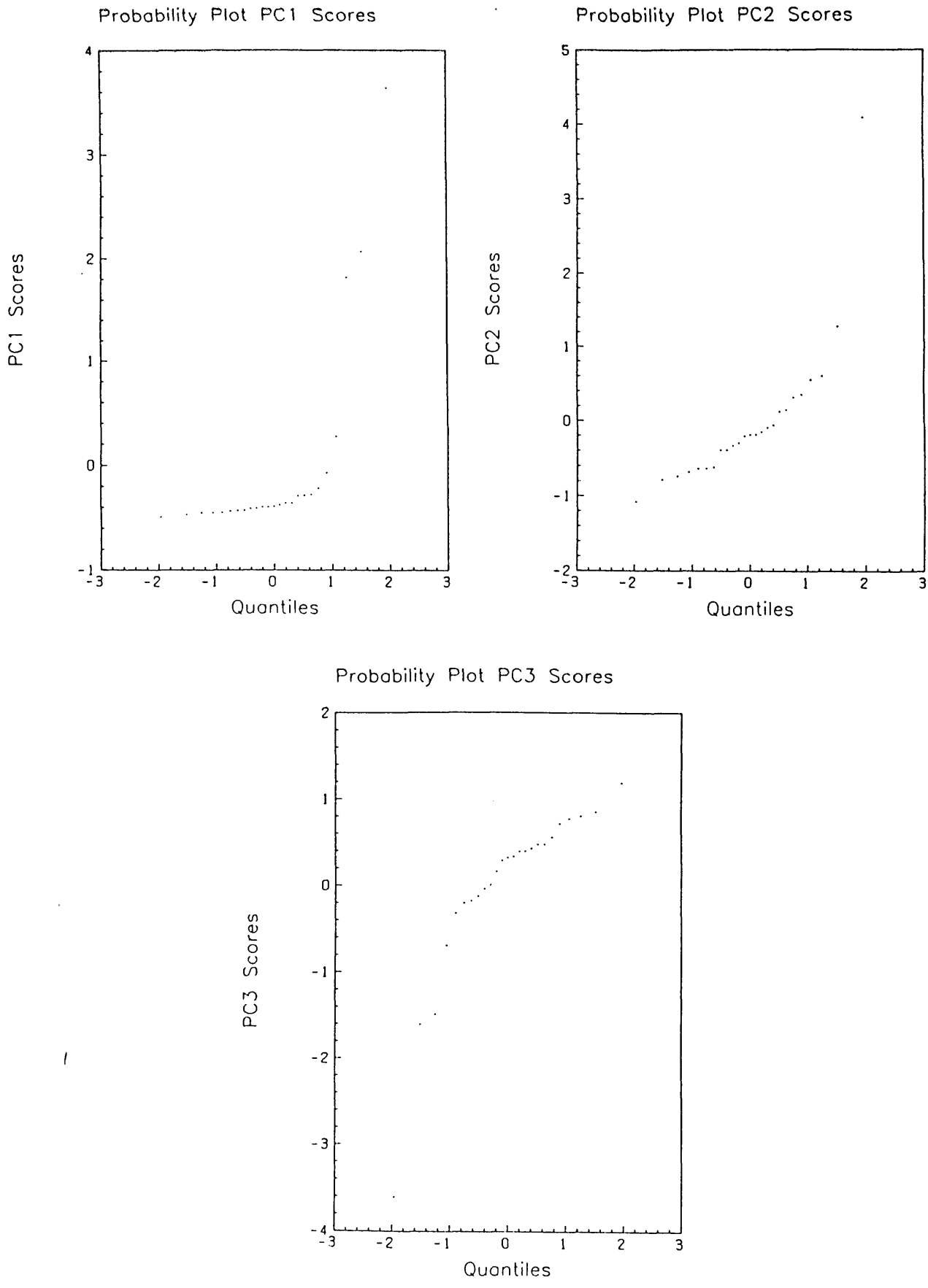

Probability Plot PC1 Scores
PC1 Scores
Quantiles
Probability Plot PC2 Scores
PC2 Scores
Quantiles
Probability Plot PC3 Scores
PC3 Scores
Quantiles

Figure 8.7

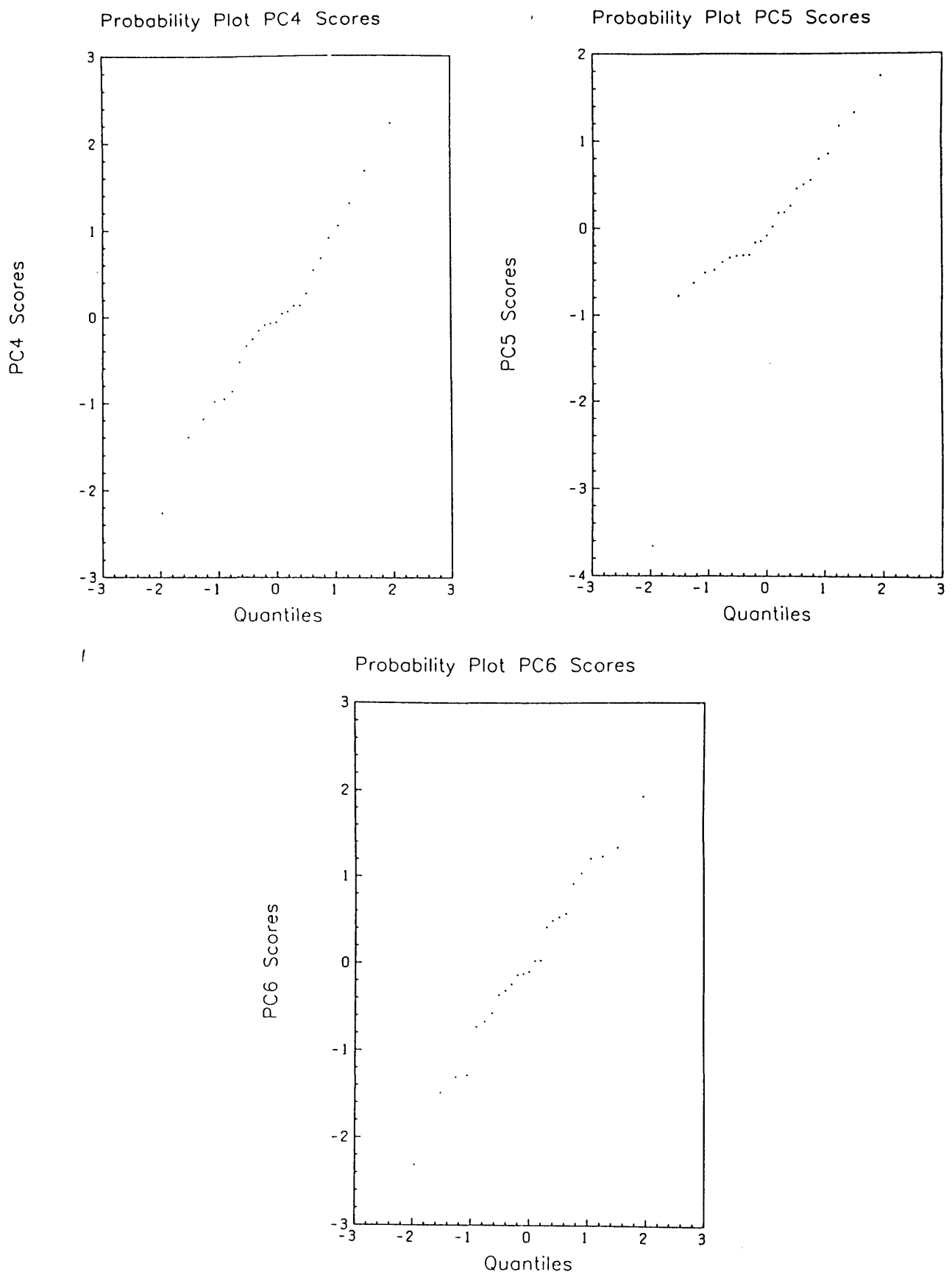

Sind Raw Data Rot'
Probability Plot PC4 Scores
PC4 Scores
Quantiles
Probability Plot PC5 Scores
PC5 Scores
Quantiles
Probability Plot PC6 Scores
PC6 Scores
Quantiles

Figure 8.8

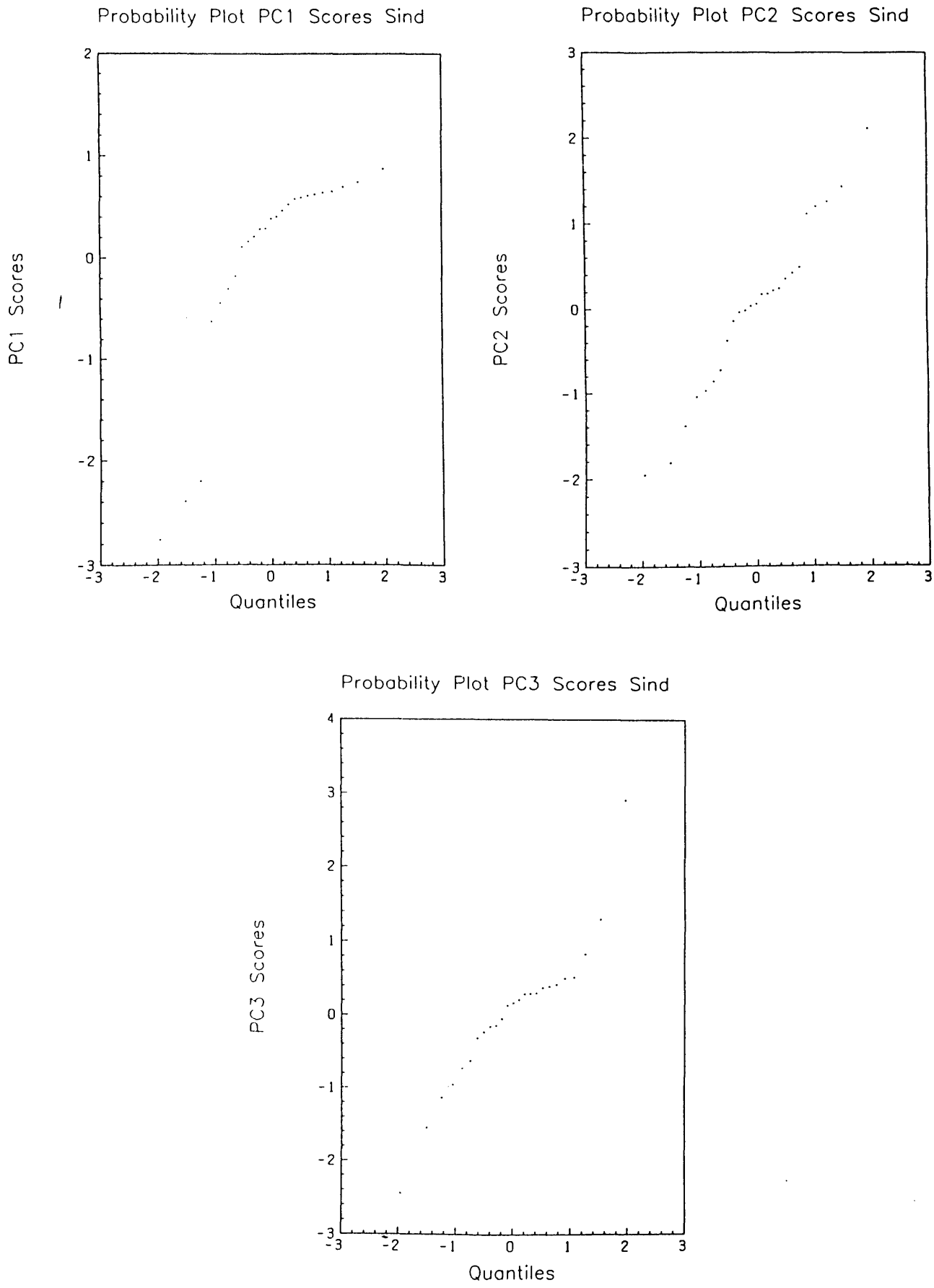

Probability Plot PC1 Scores Sind
PC1 Scores
Quantiles
Probability Plot PC2 Scores Sind
PC2 Scores
Quantiles
Probability Plot PC3 Scores Sind
PC3 Scores
Quantiles

Figure 8.9

Logten Data Rot'

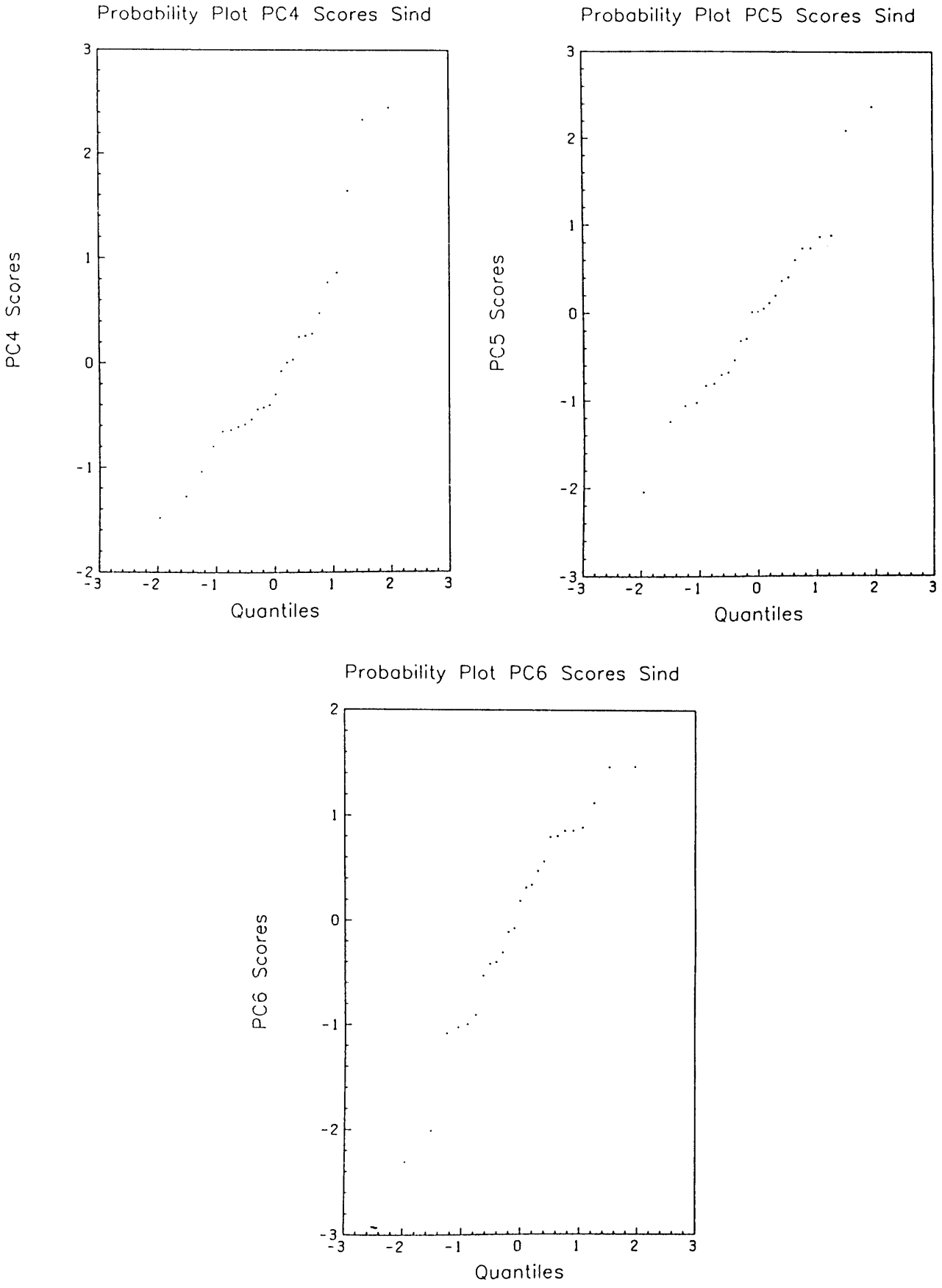


Figure 8.10

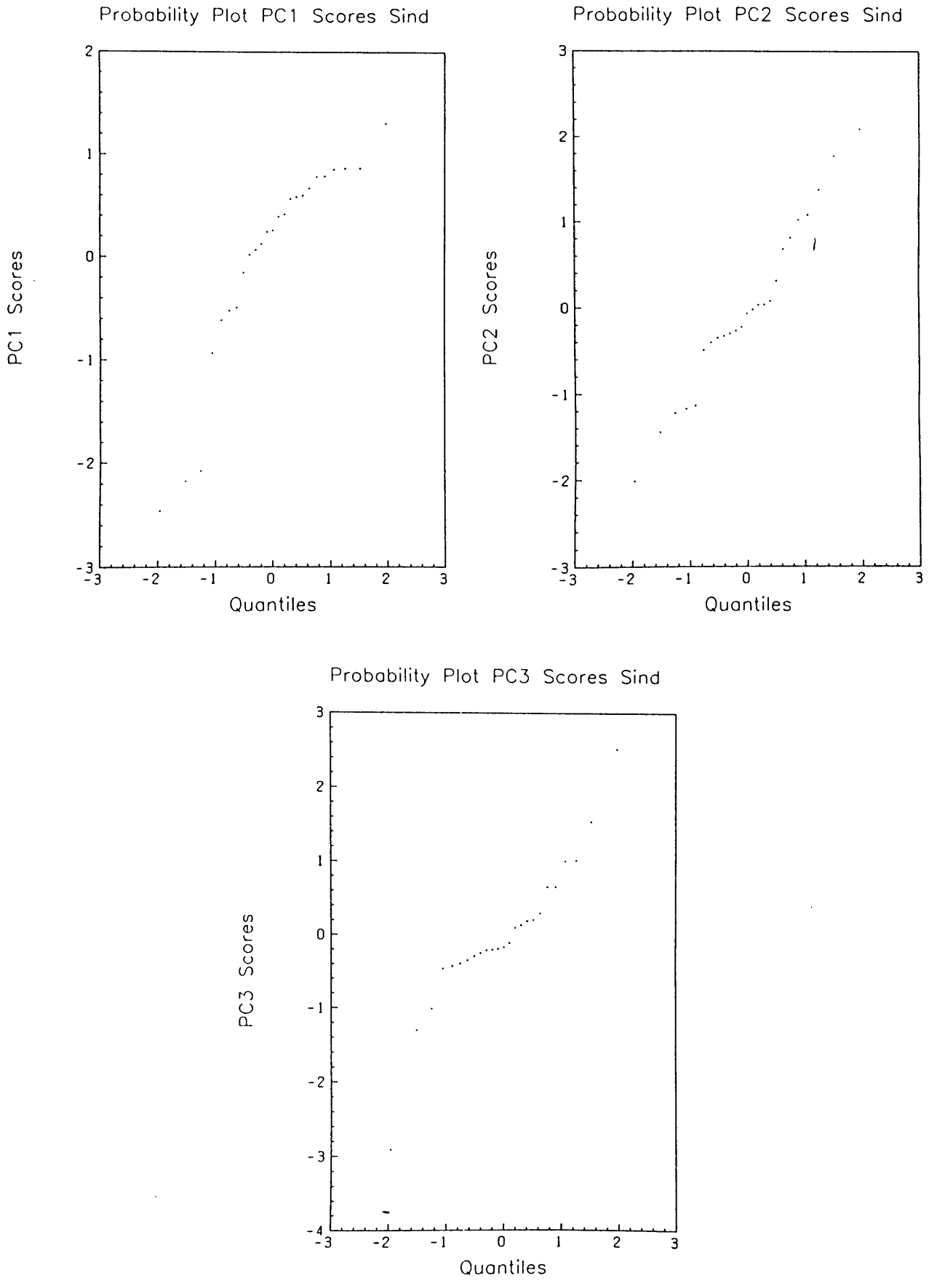
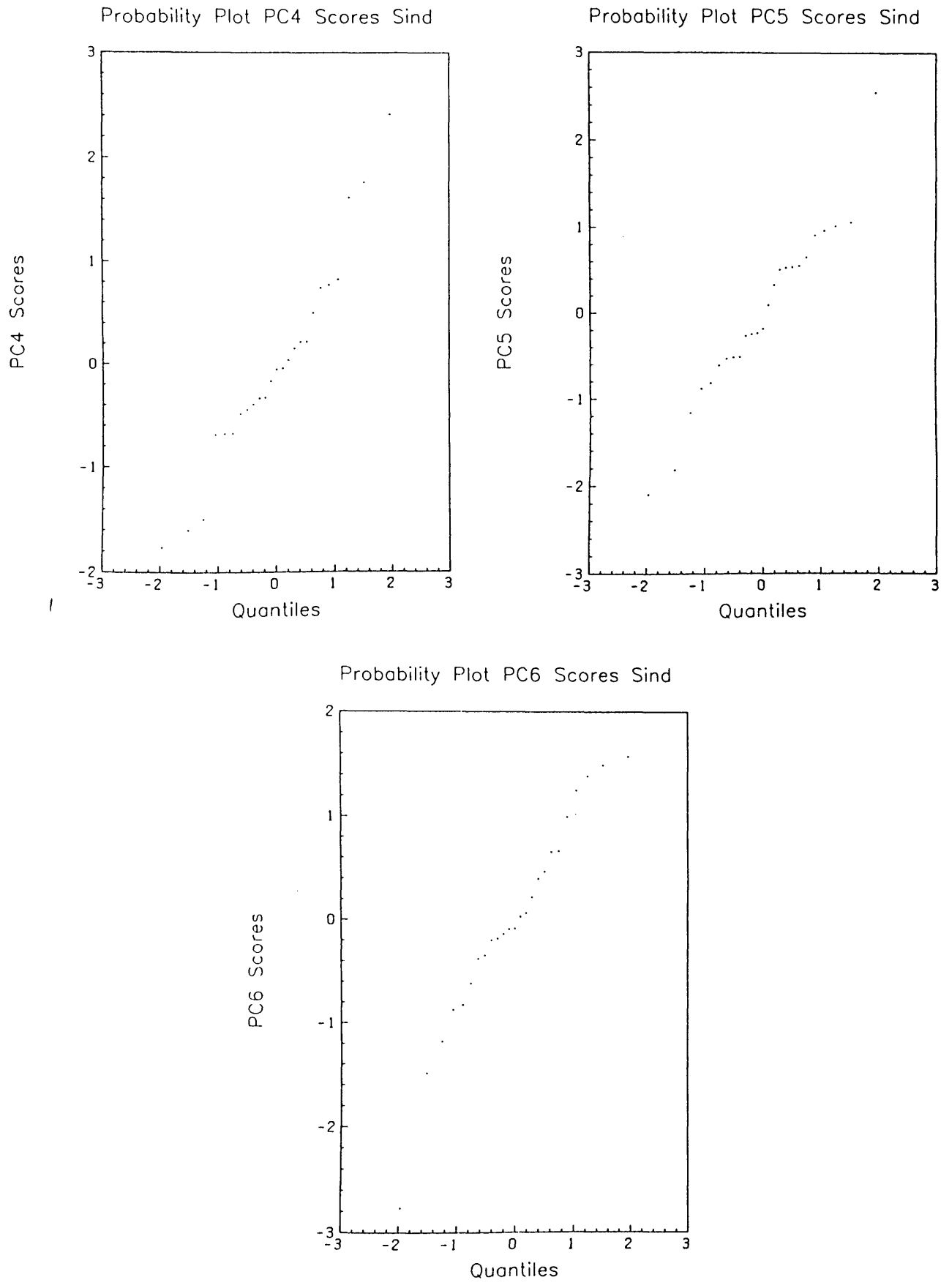

Probability Plot PC1 Scores Sind
PC1 Scores
Quantiles
Probability Plot PC2 Scores Sind
PC2 Scores
Quantiles
Probability Plot PC3 Scores Sind
PC3 Scores
Quantiles

Figure 8.11


Probability Plot PC4 Scores Sind
PC4 Scores
Quantiles
Probability Plot PC5 Scores Sind
PC5 Scores
Quantiles
Probability Plot PC6 Scores Sind
PC6 Scores
Quantiles

Mn and Pb) reveals sample 23, while PC3 scores (highly loaded in Fe) reveal sample 22, both with reasonable contrast. For logten and boxcox transformed data PC3 scores (highly loaded in Mn) reveal sample 23 amongst other "anomalies" and in a low contrast situation. The Q-Q plots for PC2 AND PC4 show close approximation to distributional normality.

8.10.2 Robust Estimation of the Correlation Matrix

To investigate the effect that robust estimation of the correlation matrix has on PCA, loadings matrices and scores generated using a robust correlation matrix are compared with those generated using a non-robust matrix (see Tables 8.2 through 8.4 for raw, logten and boxcox non-robust and robust loadings matrices respectively).

Robust estimation causes significant changes to the loadings matrix for raw data (see Table 8.2). These changes can be explained by reference to the relevant correlation matrix (Table 6.3). On PC1 the robust loadings matrix shows a dramatic decrease in the loading for Zn, while Cd and Cu remain highly loaded. Robust estimation of the correlation matrix results in the high correlations between Zn-Cd and Zn-Cu being significantly reduced. However the correlation between Cd and Cu remains high. On

Table 8.2

Non-Robust and Robust Component Loadings Matrices SIND Raw Data (Rot")

	PC1	PC2	PC3	PC4	PC5
Zn	0.9903	0.0643	−0.0880	−0.0277	0.0340
	0.5333	−0.0079	0.3188	0.3115	0.7189
Fe	0.1897	0.5747	−0.7948	−0.0443	0.0037
	0.2256	0.1553	0.2913	0.8998	0.1744
Mn	0.1198	0.8698	−0.2541	−0.4054	0.0125
	0.1758	0.1431	0.9200	0.2723	0.1674
Cd	0.9465	0.2542	−0.1293	−0.1052	0.1131
	0.9364	0.0565	0.1805	0.2232	0.0847
Cu	0.9856	−0.0685	−0.0617	0.0527	−0.1282
	0.8704	0.0721	0.0897	0.1011	0.3811
Pb	0.0197	0.9626	−0.2175	0.1598	0.0128
	0.0572	0.9848	0.1139	0.1177	0.0051

Table 8.3

Non-Robust and Robust Component Loadings Matrices SIND Logten Data (Rot^n)

	PC1	PC2	PC3	PC4	PC5
Zn	−0.9175	−0.1102	0.2406	0.2359	0.1238
	−0.9187	−0.0989	0.2406	0.2342	0.1292
Fe	−0.3866	−0.3579	0.3870	0.7565	−0.0167
	−0.3838	−0.3482	0.3892	0.7638	−0.0154
Mn	−0.2587	−0.3775	0.8422	0.2845	−0.0186
	−0.2549	−0.3511	0.8561	0.2803	−0.0168
Cd	−0.8960	−0.1385	0.2528	0.1905	−0.2787
	−0.9014	−0.1219	0.2442	0.1889	−0.2780
Cu	−0.9790	−0.0783	0.0771	0.1442	0.0423
	−0.9785	−0.0744	0.0811	0.1463	0.0449
Pb	−0.0947	−0.9387	0.2665	0.1967	−0.0110
	−0.0849	−0.9429	0.2585	0.1917	−0.0092

Table 8.4

Non-Robust and Robust Component Loadings Matrices SIND Boxcox Data (Rot")

	PC1	PC2	PC3	PC4	PC5
Zn	−0.8223	0.1335	−0.2644	0.3591	−0.3269
	−0.8205	0.1192	−0.2701	0.3582	0.3331
Fe	−0.4537	0.3228	−0.3322	0.7604	−0.0364
	−0.4470	0.3072	−0.3401	0.7673	0.0372
Mn	−0.2882	0.3881	−0.8352	0.2599	−0.0339
	−0.2812	0.3509	−0.8552	0.2556	0.0338
Cd	−0.9304	0.1295	−0.2410	0.1453	0.0757
	−0.9371	0.1058	−0.2300	0.1441	−0.0771
Cu	−0.9380	0.1308	−0.1274	0.2475	−0.0048
	−0.9369	0.1223	−0.1337	0.2484	0.0097
Pb	−0.1222	0.9387	−0.2649	0.1828	−0.0170
	−0.1077	0.9450	−0.2538	0.1751	0.0159

robust estimation, therefore, PC1 retains its association with the Cd and Cu vectors but loses its association with Zn which is accounted for by PC5.

The high non-robust correlation between Mn and Pb (Table 6.3) is reduced on robust estimation to a low coefficient. This results in principal PC2 no longer having a high loading for Mn. Instead, Mn is highly loaded on PC3.

Tables 8.3 and 8.4 show non-robust and robust loadings matrices for logten and boxcox data respectively. Robust estimation of the correlation matrix clearly has very little effect on the loadings. Again, the correlation matrices give the explanation. We have seen from Chapter 6 that there is very little change in the logten and boxcox correlation matrices on robust estimation - this explains a similar lack of change in the corresponding loadings matrices. The structure of the correlation matrices can explain the individual loadings. The correlations between the variables Zn, Cd and Cu are high - this leads to their having high loadings on PC1. Fe, Mn and Pb have no very high correlations with any other variable and so are loaded on three separate principal components.

A comparison is made of the Q-Q plots for component scores generated from non-robust loadings matrices (Figures 8.6 through 8.11) with those generated from robust matrices (Figures 8.12 through 8.17).

We would expect large changes in the scores for raw data since the two loadings matrices are so different. However,

Sind Robust Raw Data Rot" **Figure 8.12**

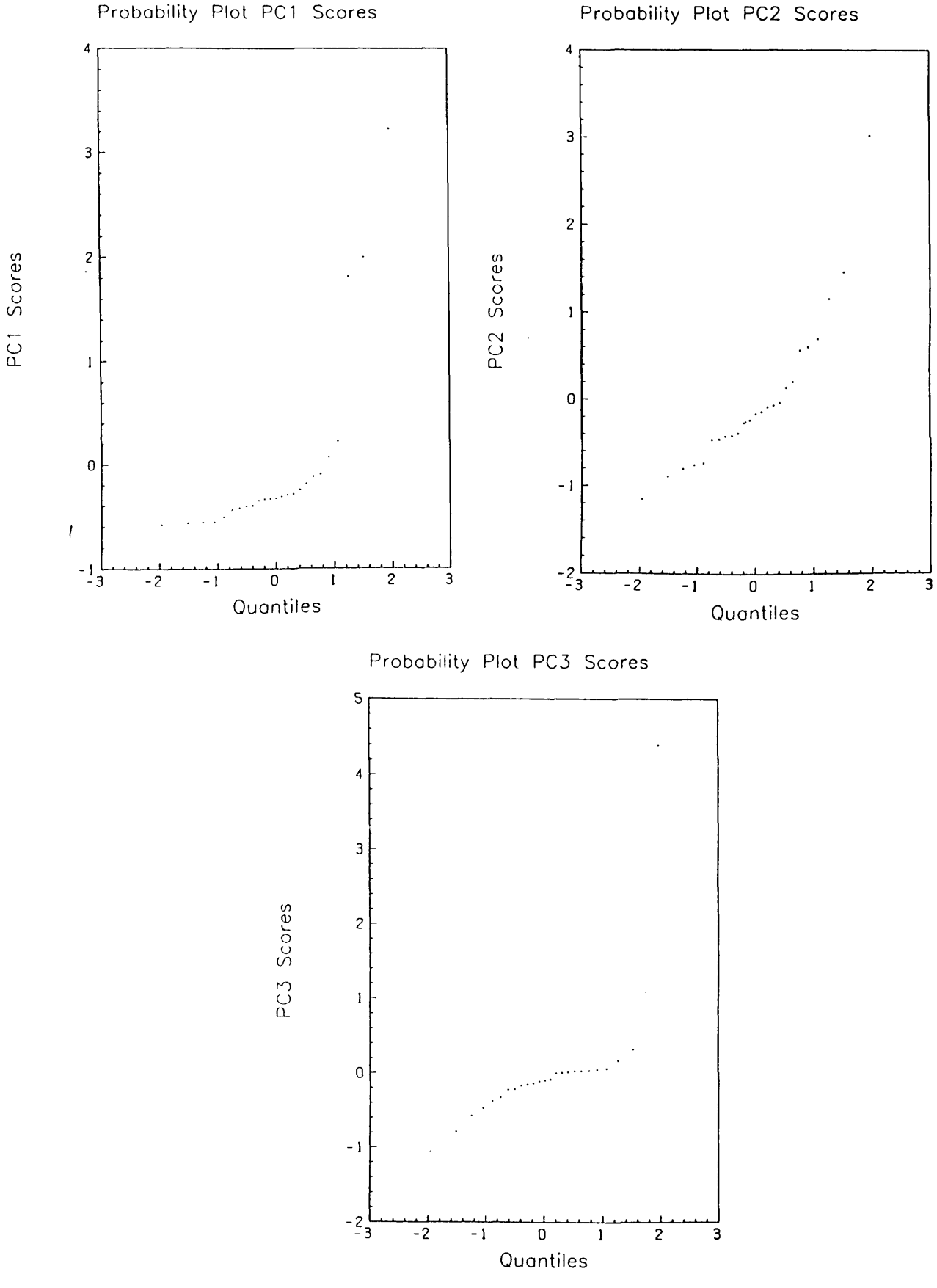
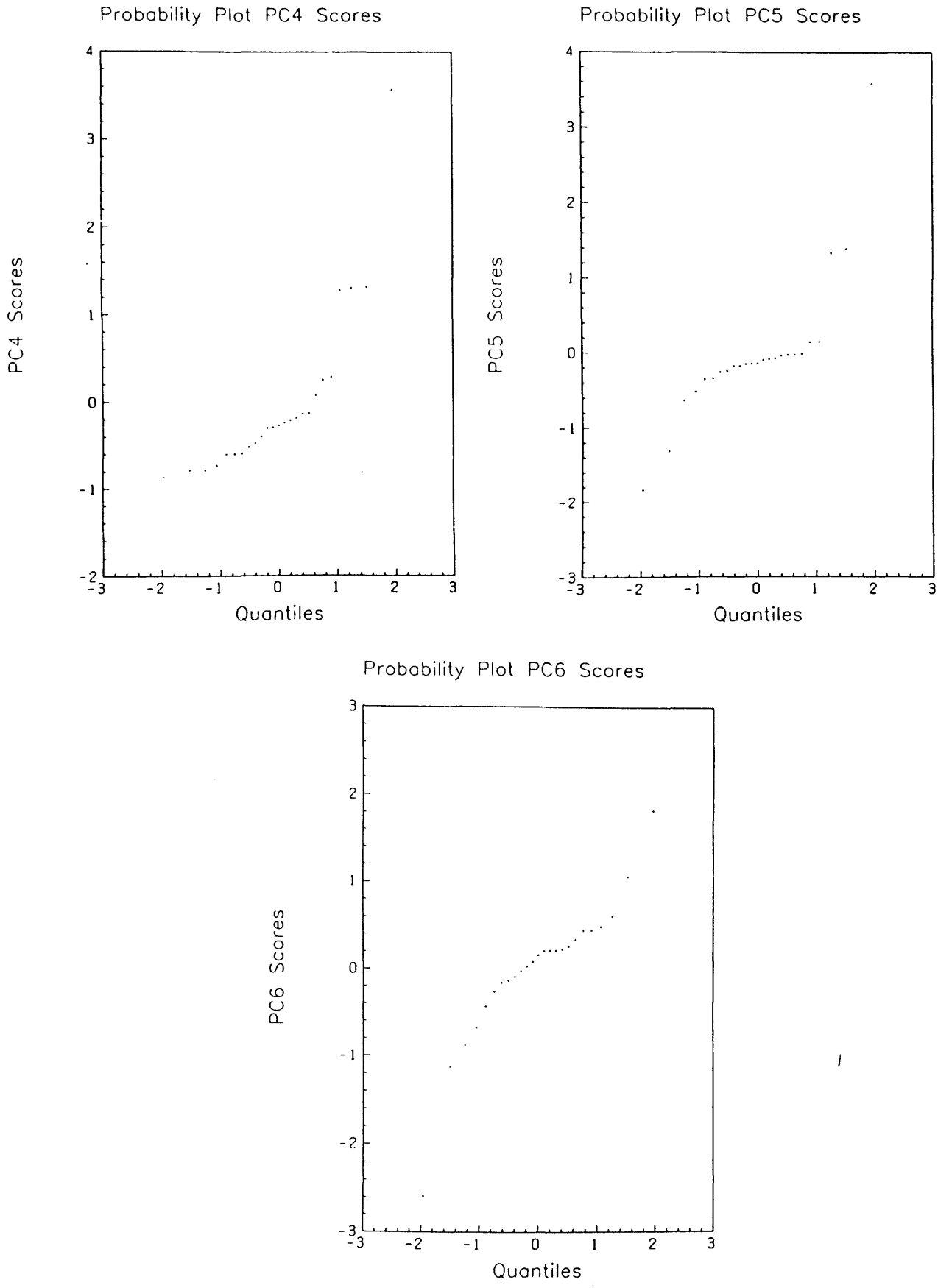


Figure 8.13


Probability Plot PC4 Scores
PC4 Scores
Quantiles
Probability Plot PC5 Scores
PC5 Scores
Quantiles
Probability Plot PC6 Scores
PC6 Scores
Quantiles

Robust Logten Data Rot^n **Figure 8.14**

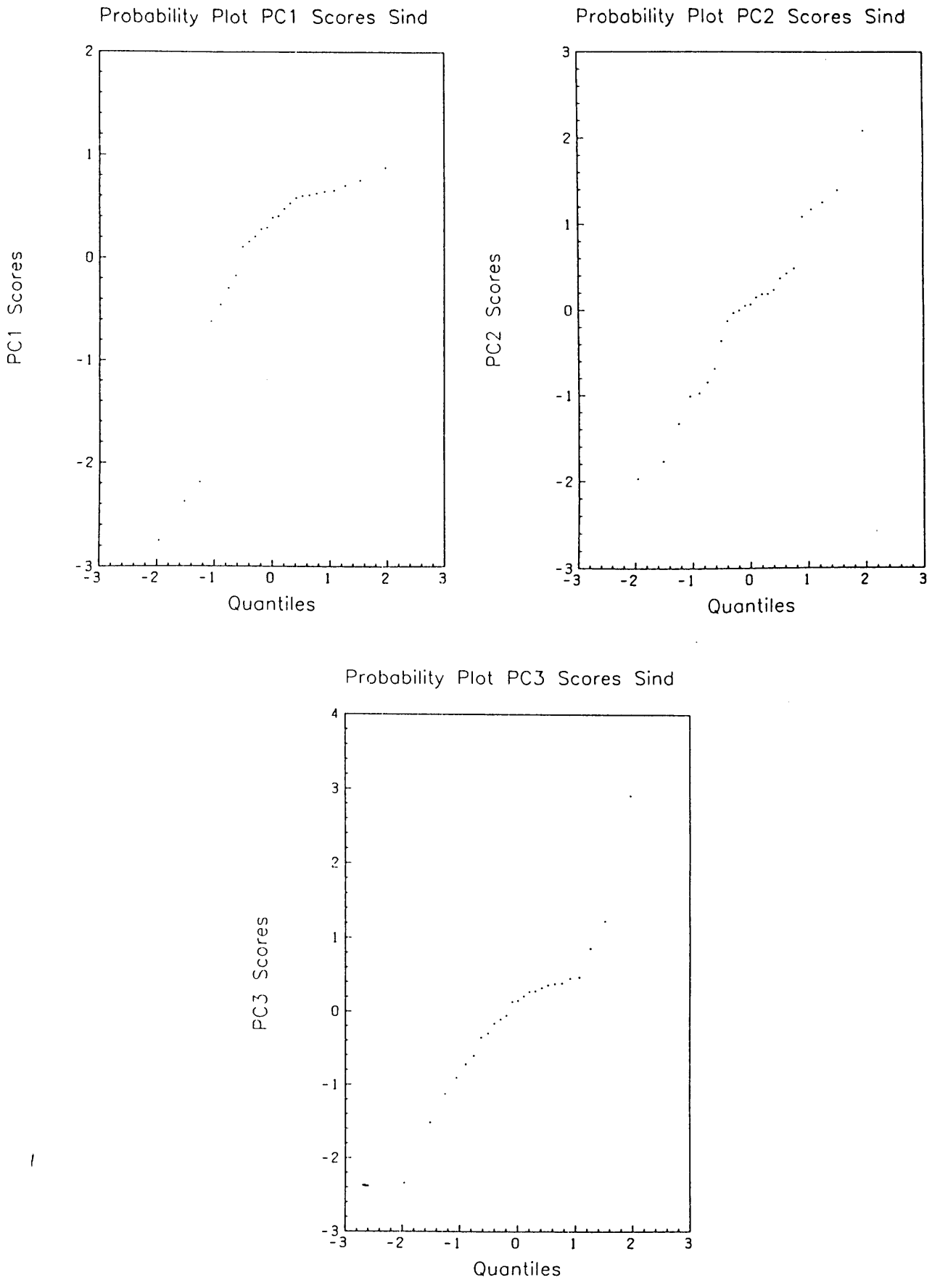
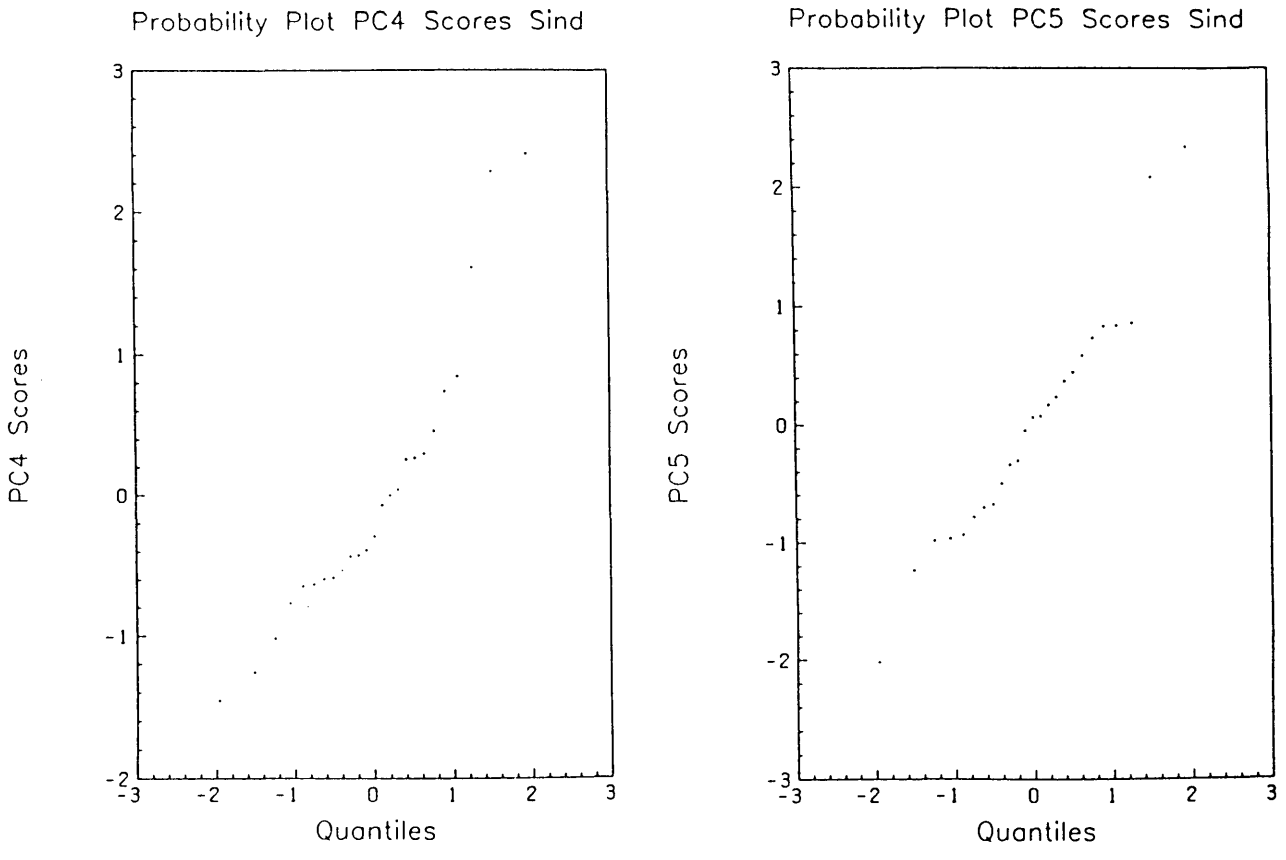
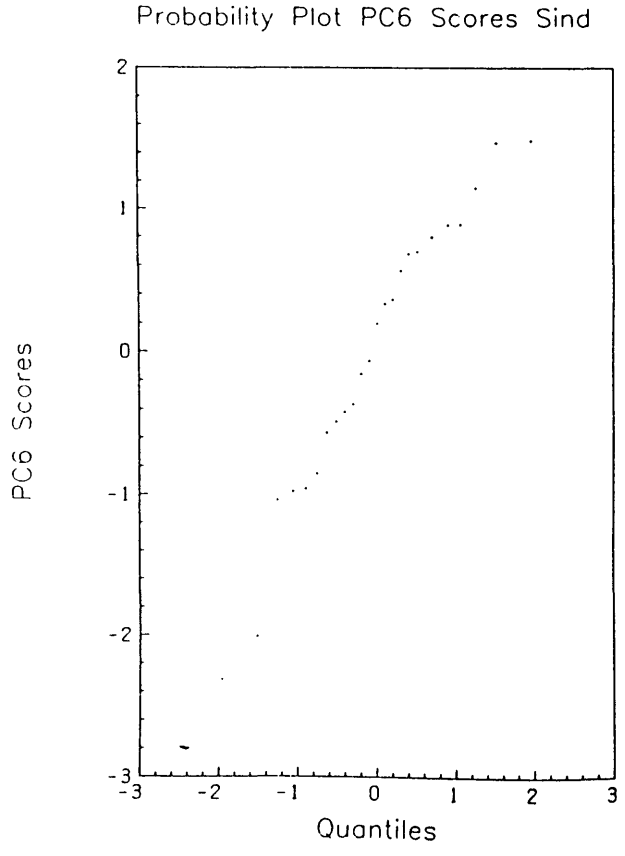


Figure 8.15


Probability Plot PC4 Scores Sind
PC4 Scores
Quantiles
Probability Plot PC5 Scores Sind
PC5 Scores
Quantiles


Probability Plot PC6 Scores Sind
PC6 Scores
Quantiles

Robust Boxcox Data Rot"

Figure 8.16

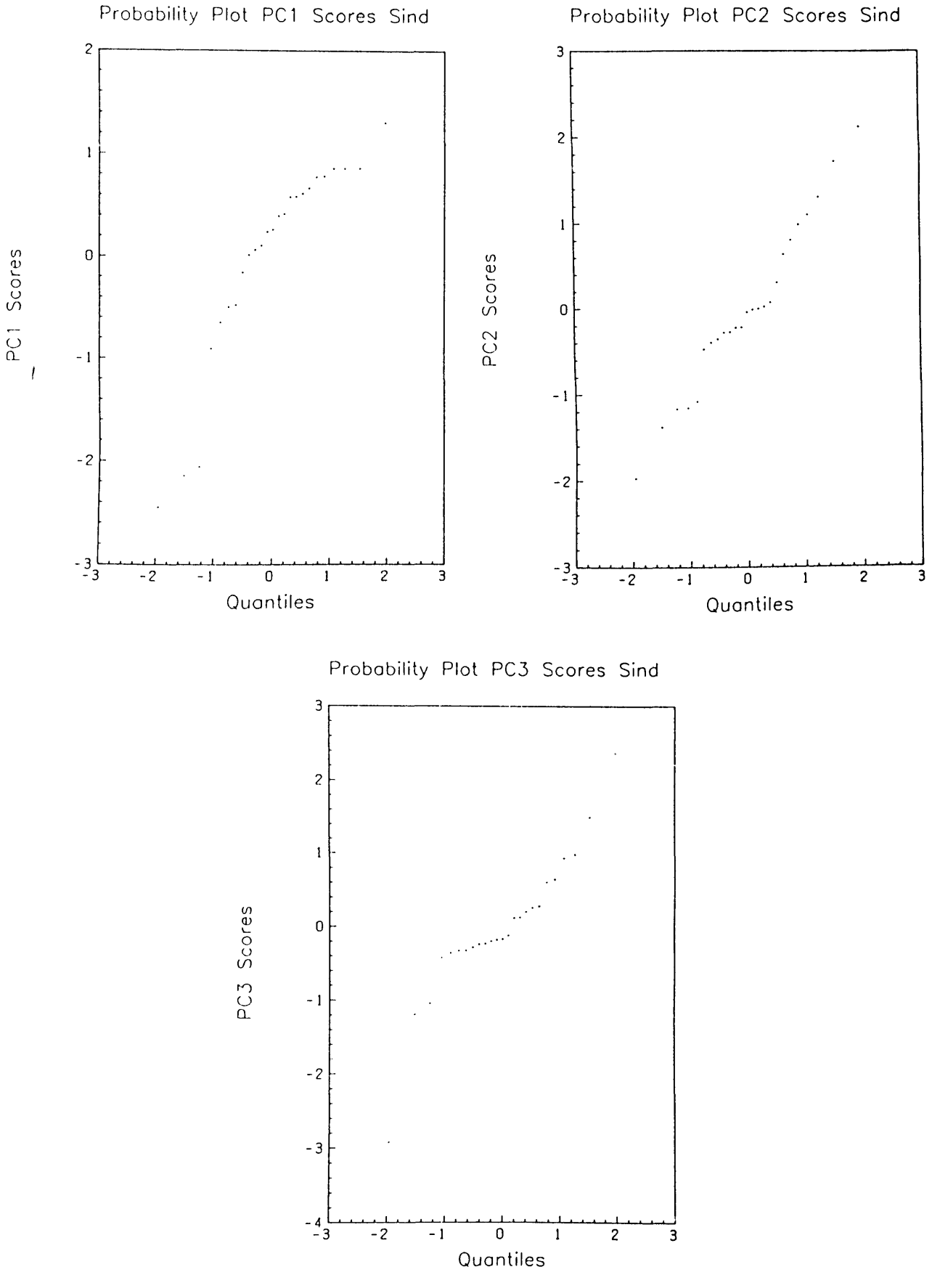
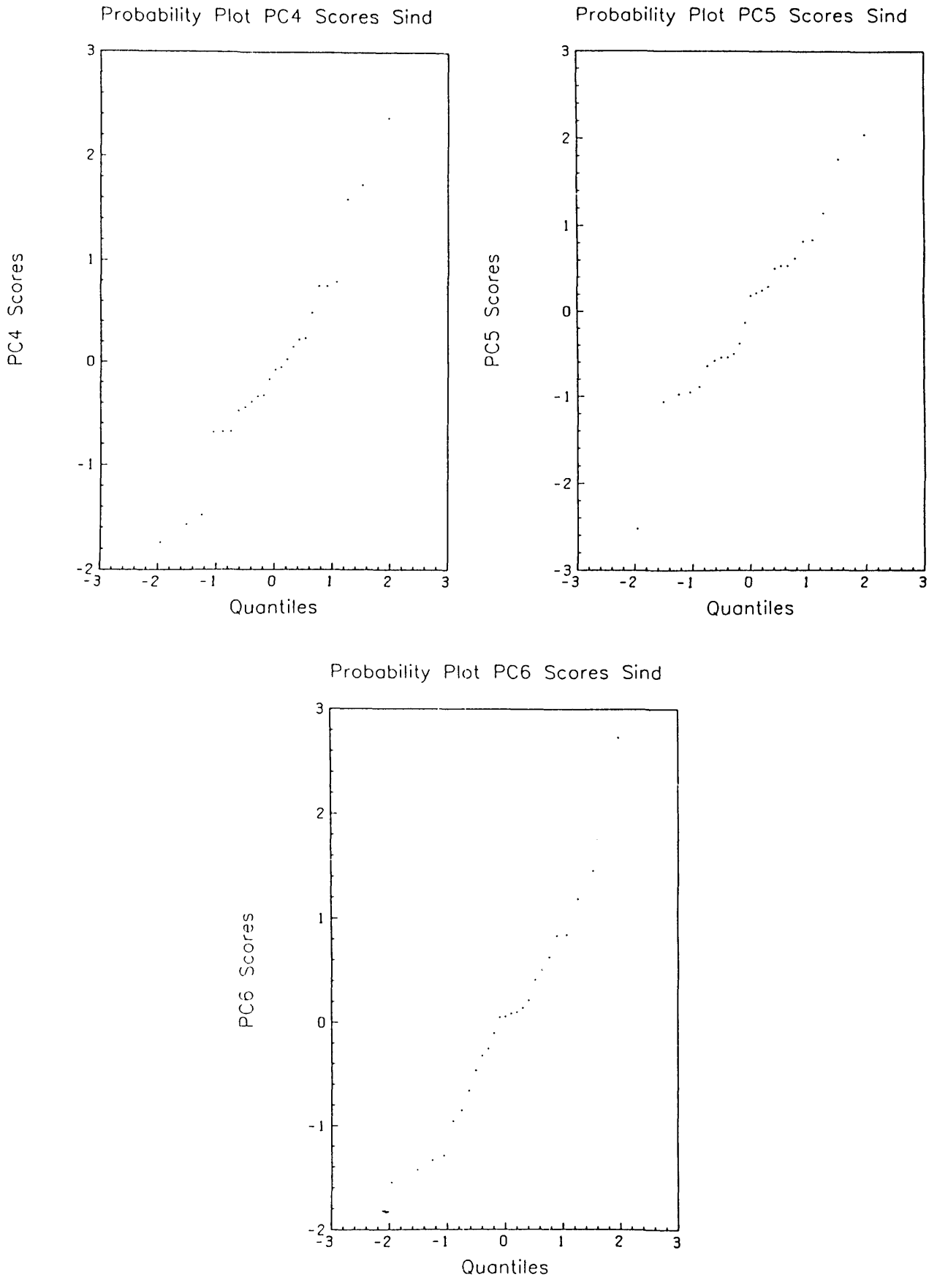


Figure 8.17


Probability Plot PC4 Scores Sind
PC4 Scores
Quantiles
Probability Plot PC5 Scores Sind
PC5 Scores
Quantiles
Probability Plot PC6 Scores Sind
PC6 Scores
Quantiles

the robust PC1 scores reveal anomalous samples 13, 15 and 16 with only slightly less contrast than the non-robust scores. Robust PC2 scores (highly loaded in Pb) reveal sample 23 with less contrast than non-robust scores (highly loaded in Pb and Mn). Robust PC3 scores reveal sample 23 with much greater contrast, due to the higher loading on Mn, than do either robust or non-robust PC2 scores. Non-robust PC3 scores (highly loaded in Fe) reveal sample 22 with about the same contrast as robust PC4 scores (highly loaded in Fe). Robust PC5 scores (highly loaded in Zn) reveal samples 13, 15 and 16 with high positive values, and samples 22 and 23 with high negative values.

As would be expected from the analysis of the loadings matrices the logten and boxcox robust component score Q-Q plots show very little difference to the equivalent non-robust plots.

In conclusion, the use of robust estimation in PCA makes a large difference to the component loadings for raw data but the changes make little difference to the interpretation of their component scores. This may be because the SIND data are very simple in structure. The effect of robust estimation of the correlation matrix on PCA of logten and boxcox data is negligible.

8.10.3 Rotation

A comparison of each rotated loadings matrix (Table 8.1) with its equivalent non-rotated matrix (Table 8.5) shows that the latter bears little resemblance to the former and the differences between the two appear unsystematic.

Only the first two components of each non-rotated loadings matrix are of any interest. PC1 is highly loaded in all six variables with Zn and Cu being the two highest in both non-robust and robust raw and logten matrices, with either Cd or Fe being the third largest. With boxcox data Zn, Cu and Fe have the highest loadings on PC1 in both the non-robust and robust loadings matrices. The loadings on PC1 are all positive for the raw matrix so we would expect anomalous samples to be revealed by high positive scores (see Figure 8.18). The loadings are all negative for logten and boxcox matrices and so we would expect any anomalies to be revealed by high negative scores.

The second principal component is more important than the first since it differentiates between the "true" and the "false" anomalies. For non-robust and robust raw and boxcox loadings matrices, Zn, Cd and Cu are negatively loaded and Fe, Mn and Pb positively loaded on PC2. For both logten matrices the signs are reversed. This causes anomalous object vectors 13, 15 and 16 to have high

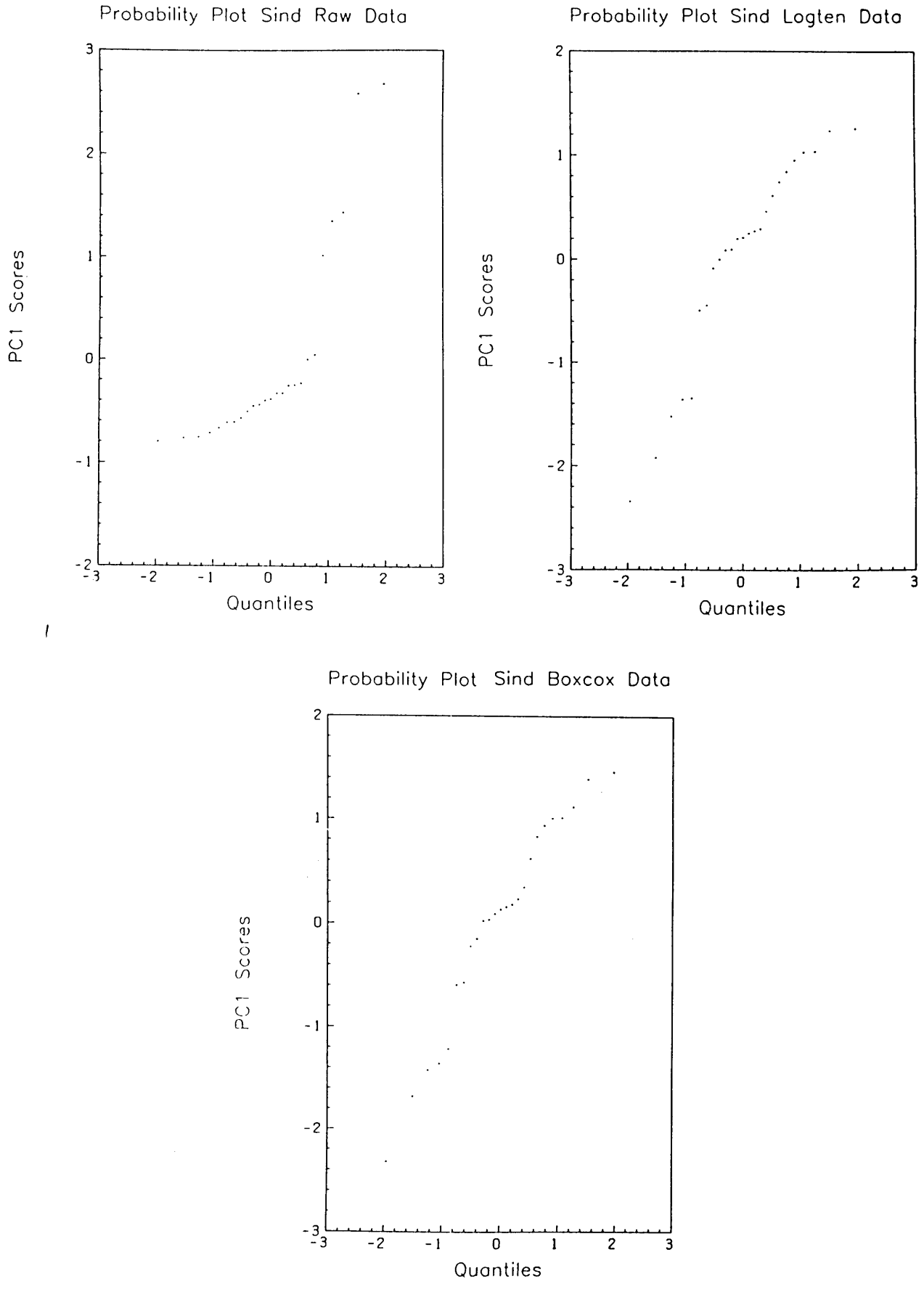
Table 8.5

Non--Rotated Component Loadings Matrices Raw, Logten, Boxcox SIND Data

	PC1	PC2	PC3	PC4
Zn	0.8449	−0.5276	0.0278	0.0065
	−0.9148	0.3515	−0.0091	−0.0022
	−0.9289	−0.2695	−0.0485	0.0646
Fe	0.7063	0.5462	−0.4500	−0.0162
	−0.8559	−0.3115	0.2013	0.3587
	−0.8891	0.1617	−0.2066	0.3554
Mn	0.6606	0.6685	0.2071	−0.2674
	−0.7783	−0.4743	0.3053	−0.2727
	−0.7927	0.4385	−0.3037	−0.2917
Cd	0.9257	−0.3489	0.0803	−0.0513
	−0.9089	0.3276	−0.0388	−0.0919
	−0.8839	−0.3571	0.1286	−0.1972
Cu	0.7551	−0.6388	−0.0255	0.0783
	−0.8497	0.4958	−0.1390	0.0016
	−0.8906	−0.3957	0.1485	0.0118
Pb	0.5699	0.7435	0.1799	0.3001
	−0.5862	−0.6903	−0.4235	−0.0181
	−0.6079	0.7024	0.3676	0.0315

PC1 Scores No-Rot"

Figure 8.18



negative scores and vectors 22 and 23 to have high positive scores on PC2 for raw and boxcox data (see Figure 8.19). Again, the situation is reversed for logten data.

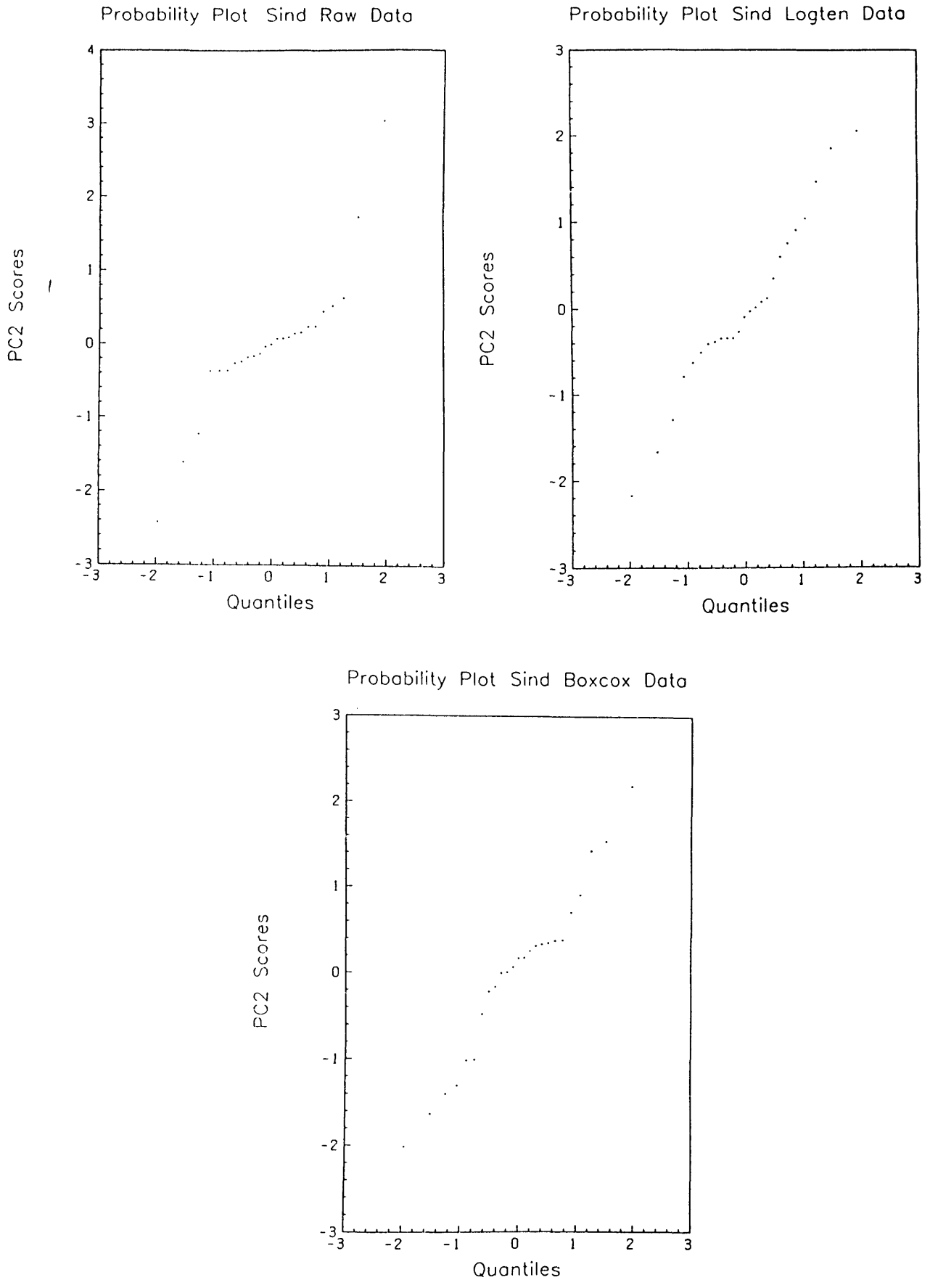
A comparison of non-rotated raw, logten and boxcox Q-Q plots of PC2 scores (Figure 8.19) shows that logten and boxcox scores closely approximate distributional normality so that despite the five anomalous object vectors being numerically distinct they are visually inseparable from background - the anomaly contrast is low. The equivalent Q-Q plot for raw data clearly shows a normally distributed 20 sample background population with object vectors 13, 15 and 16 having high negative and 22 and 23 having high positive component scores - the anomaly contrast is high. The Q-Q plots of PC1 scores (Figure 8.18) also show the difference in anomaly contrast between raw and transformed data.

Robust estimation seems to have little effect on non-rotated loadings matrices and therefore on component scores.

8.10.4 PCA or FA - Practice

So far in this chapter we have analysed the SIND data using PCA - all the components have been retained in the final loadings matrix. There are strong theoretical

Figure 8.19


Probability Plot Sind Raw Data
PC2 Scores
Quantiles
Probability Plot Sind Logten Data
PC2 Scores
Quantiles
Probability Plot Sind Boxcox Data
PC2 Scores
Quantiles

arguments (see Section 8.6) against the use of FA but as we have seen from the above section, PCA does leave something to be desired as far as interpretability goes. As a result, FA was performed on the raw, logten and boxcox SIND data sets with both non-robust and robust estimation of the correlation matrix. In all cases rotated factors were used.

The number of factors retained is equal to the number of eigenvectors whose eigenvalues are greater than one (see Section 8.4). In the case of the SIND data this is always equal to two, which is also the number of major variable groups seen in the cluster dendrogram (see Chapter 7).

With only two factors retained there is a dramatic increase in the interpretability of the loadings matrices:

1) Zn, Cu and Cd are highly loaded on F1 (see Table 8.6). F1 scores should therefore reveal anomalous object vectors 13, 15 and 16.

2) Fe, Mn and Pb are highly loaded on F2 (see Table 8.6). F2 scores should reveal anomalous object vectors 22 and 23.

3) On data transformation, significant changes are observed in the factor loadings. These changes are easily explained when Table 8.6 is analysed with the aid of the corresponding correlation matrices (Table 6.2).

Table 8.6

Factor Loadings Matrices SIND Raw, Logten, Boxcox Data (Rot")

	FACT1	FACT2
Zn	0.9904	0.1063
	−0.9399	−0.2840
	−0.9130	0.3190
Fe	0.2186	0.8657
	−0.4845	−0.7713
	−0.6306	0.6473
Mn	0.1073	0.9337
	−0.3233	−0.8521
	−0.3916	0.8168
Cd	0.9437	0.2967
	−0.9186	−0.2993
	−0.9272	0.2216
Cu	0.9884	−0.0367
	−0.9751	−0.1302
	−0.9551	0.1938
Pb	−0.0104	0.9367
	−0.0391	−0.9048
	−0.0881	0.9247

4) Robust estimation of the correlation matrices has little effect on the loadings matrices of logten and boxcox data but a large effect on that of raw data (see Tables 8.7, 8.8 and 8.9). These changes observed in the raw loadings matrix on robust estimation are easily explained by reference to the non-robust and robust raw correlation matrices (Table 6.3).

5) The Q-Q plots for raw data non-robust and robust factor scores (Figure 8.20) show the five anomalous object vectors very clearly - the anomaly contrast is very high - with the robust scores having a slight edge over the non-robust.

For logten and boxcox data the non-robust and robust factor score plots (Figures 8.21 and 8.22) show the five anomalies, but with considerably less contrast compared with the raw data plots. Note that the difference between non-robust and robust score plots is negligible, as we suspected from the loadings matrices.

Table 8.7

Non–Robust and Robust Factor Loadings Matrices SIND Raw Data (Rot°)

	FACT1	FACT2
Zn	0.9904	0.1063
	0.8805	0.2709
Fe	0.2186	0.8657
	0.4987	0.6576
Mn	0.1073	0.9337
	0.4343	0.6790
Cd	0.9437	0.2967
	0.8846	0.1521
Cu	0.9884	−0.0367
	0.9234	0.0653
Pb	−0.0104	0.9367
	−0.0901	0.8196

Table 8.8

Non-Robust and Robust Factor Loadings Matrices SIND Logten Data (Rot")

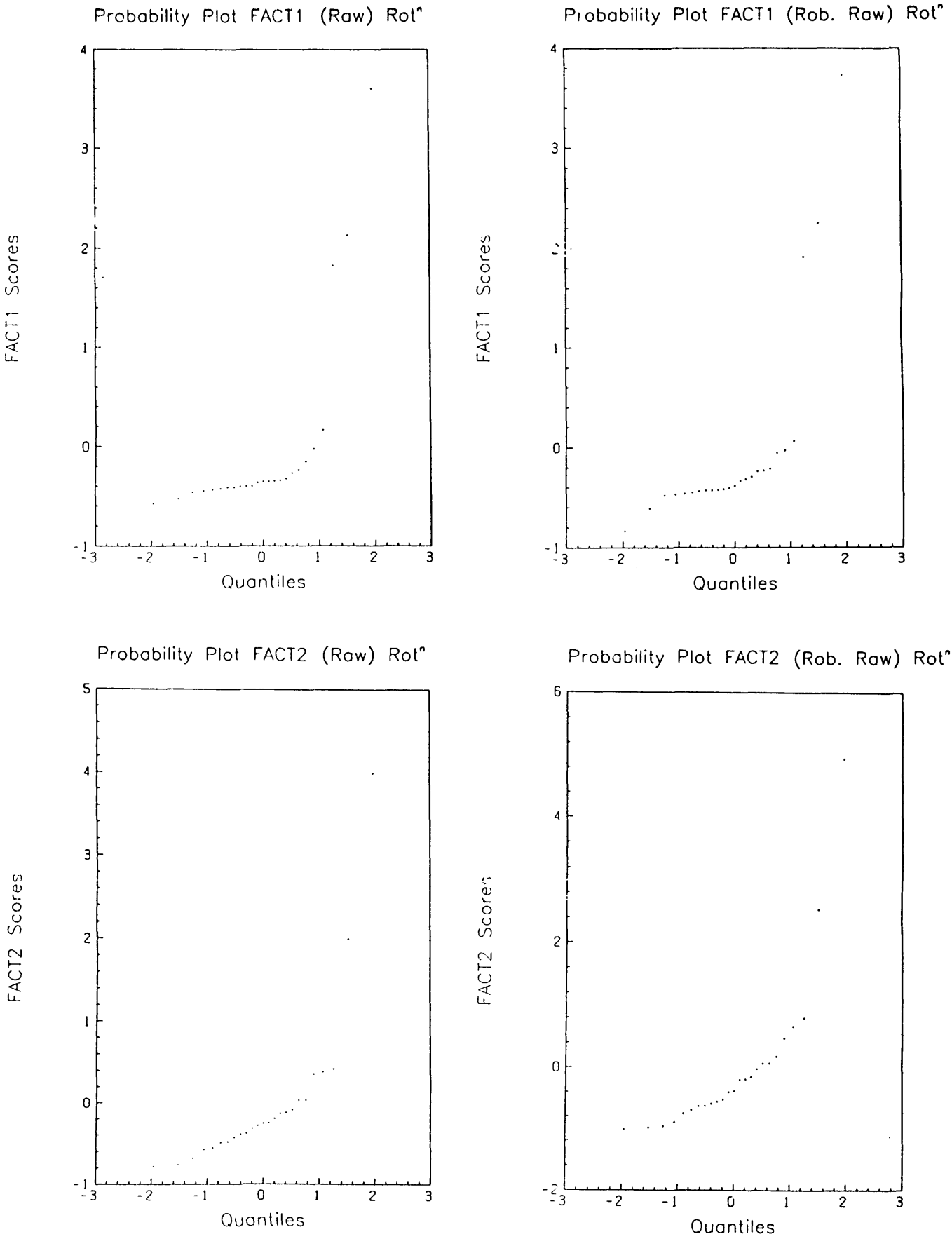
	FACT1	FACT2
Zn	−0.9379	−0.2840
	−0.9399	−0.2756
Fe	−0.4845	−0.7713
	−0.4869	−0.7660
Mn	−0.3233	−0.8521
	−0.3260	−0.8418
Cd	−0.9186	−0.2993
	−0.9239	−0.2809
Cu	−0.9751	−0.1302
	−0.9743	−0.1299
Pb	−0.0391	−0.9048
	−0.0229	−0.8986

Table 8.9

Non-Robust and Robust Factor Loadings Matrices SIND Boxcox Data (Rot")

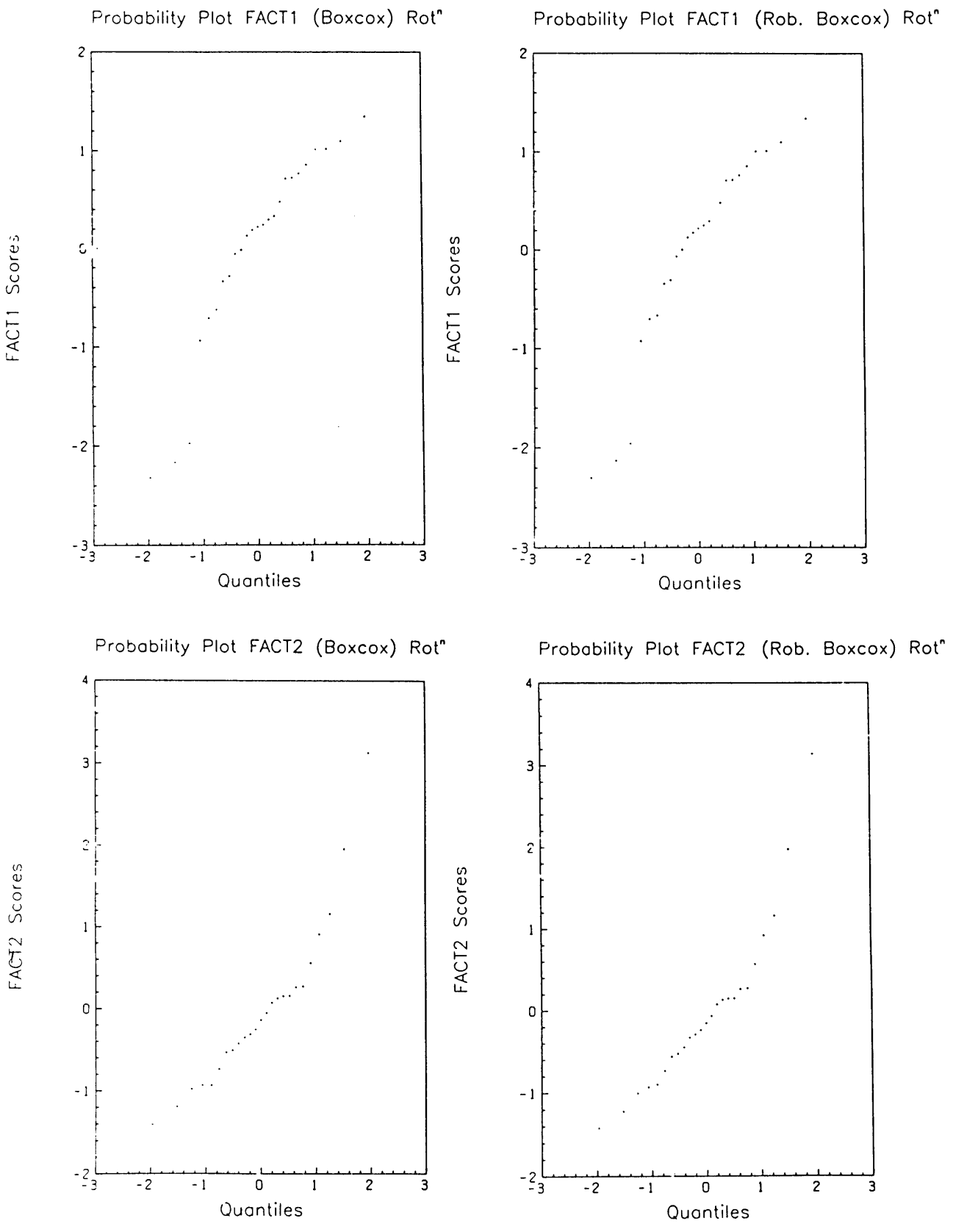
	FACT1	FACT2
Zn	−0.9130	0.3190
	−0.9148	0.3113
Fe	−0.6306	0.6473
	−0.6307	0.6439
Mn	−0.3916	0.8168
	−0.3973	0.7991
Cd	−0.9272	0.2216
	−0.9328	0.1945
Cu	−0.9551	0.1938
	−0.9542	0.1903
Pb	−0.0881	0.9247
	−0.0640	0.9196

Figure 8.20

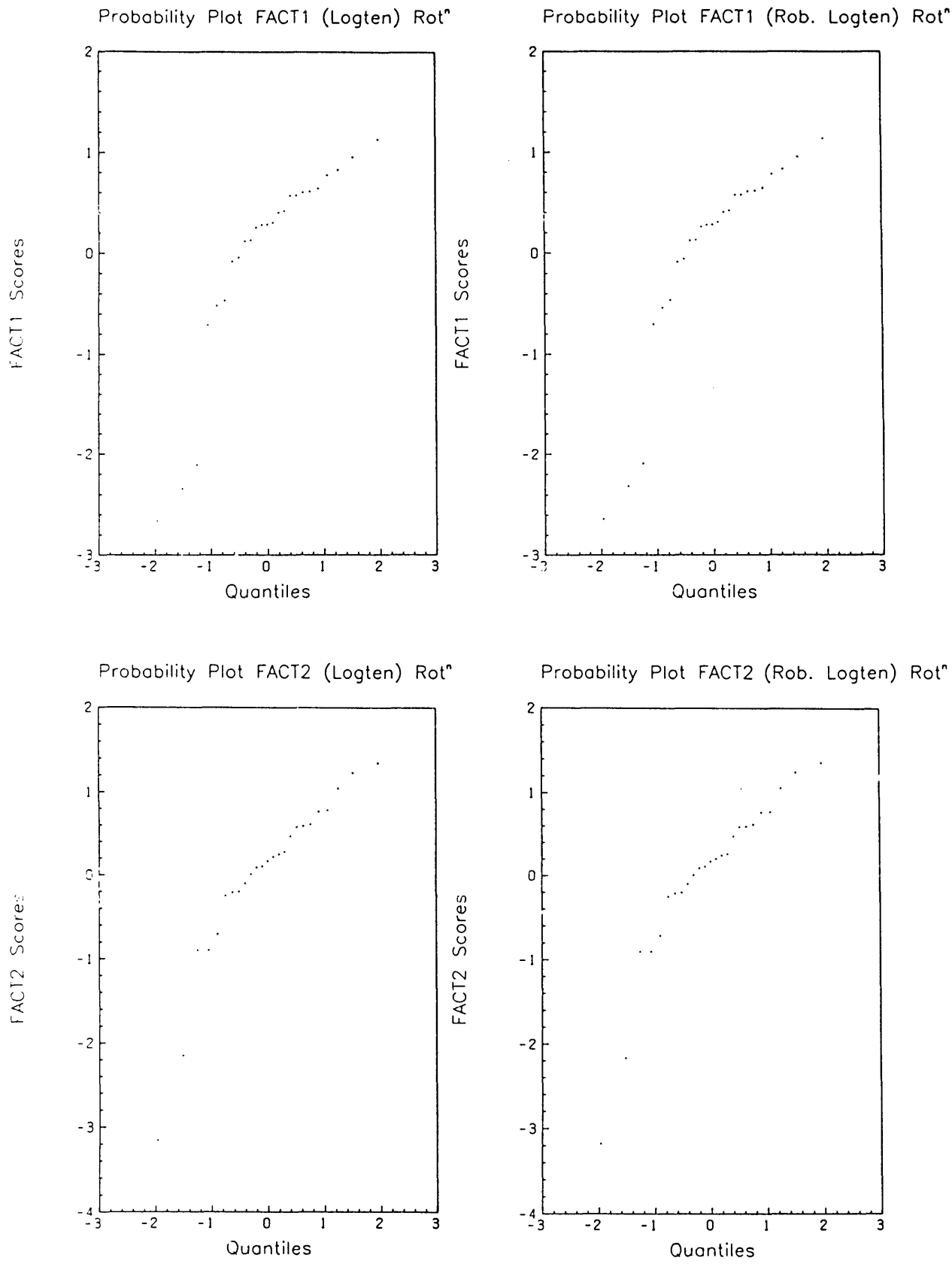

Probability Plot FACT1 (Raw) Rot"
FACT1 Scores
Quantiles
Probability Plot FACT1 (Rob. Raw) Rot"
FACT1 Scores
Quantiles
Probability Plot FACT2 (Raw) Rot"
FACT2 Scores
Quantiles
Probability Plot FACT2 (Rob. Raw) Rot"
FACT2 Scores
Quantiles

SIND Boxcox Data

Figure 8.21



SIND Logten Data Figure 8.22



8.10.5 SIND Data: Conclusions

1) Despite the theoretical arguments against the use of FA, the analysis of the SIND data has shown that, compared with PCA, FA shows an increase in interpretability of the loadings matrix plus simpler and clearer score Q-Q plots. To see whether this conclusion holds up when a larger, more complex data set is used, see Section 8.11.

The key to the successful use of FA over PCA is in the selection of the number of factors to retain. This is especially important when the initial number of variables (p) is large. The criterion used above (the number of eigenvectors whose eigenvalues are greater than one) works well and may be supported by the number of major groups of variables seen in the cluster analysis dendrogram (see Chapter 7).

2) Rotation should be used.

3) Data transformation has a dramatic effect on the factor scores. This is due mainly to the effect of transformation on the data (i.e. background enhancement and anomaly suppression) rather than to changes in the

loadings matrix. If anomaly detection is the aim of the data analysis then data transformation should be avoided. The best anomaly contrast results from the use of raw data with either a non-robust or robust correlation matrix (the latter has a slight edge). If determination of background relationships and associations is the aim of the analysis then the loadings matrix should be generated using the boxcox robust correlation matrix to give the best characterization of background population structure. If scores are required then the boxcox data should be used. If raw data were used the visual emphasis would be on the anomalies and not on background trends.

8.11 PCA/FA and the Moray-Buchan Data

The aims of this section are:

1) To compare a typical classical approach to PCA and FA using logten transformed data, with an approach which draws on the lessons learnt in previous chapters.

2) To determine the relative merits of PCA and FA on a large data set.

8.11.1 PCA - Logten Data

The rotated component loadings matrix generated from the logten non-robust correlation matrix shows the following features:

1) The elements V, Ti, Fe, Ni, Cu, Cr and Co have high negative loadings on PC1. This component obviously has basic/ultrabasic associations.

2) K and Rb have high positive loadings on PC2, which may be regarded as having sedimentary associations.

3) With the exception of PC5 (high positive loadings in Mn and Co) each subsequent component has a high loading (either positive or negative) on only one element. Thus Y is highly loaded on PC3, Zr on PC4, B on PC6, etc. This means that, with the exceptions of PC1, PC2 and PC5, the component scores will only mimic the original (logten) data and will therefore be of little use.

The probability plots of PC1 and PC2 scores (Figure 8.23) show that both have distributions which very closely approximate to normality.

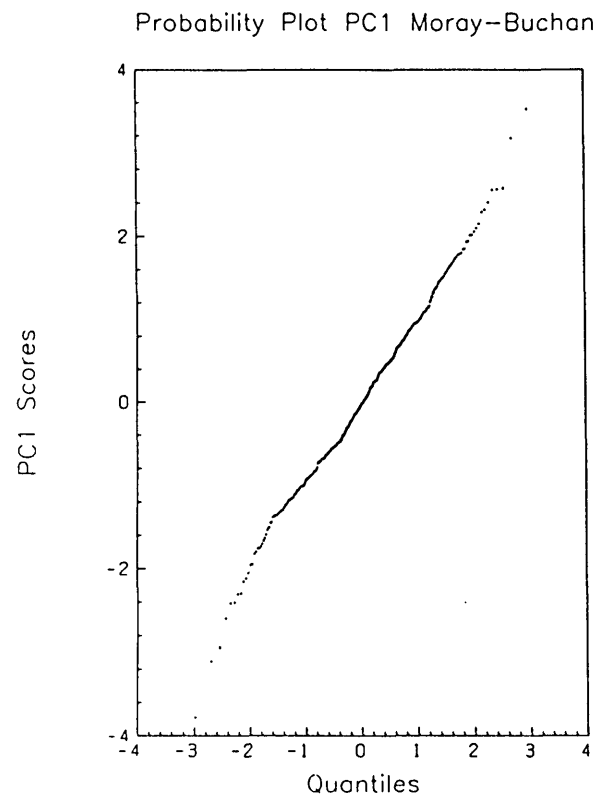
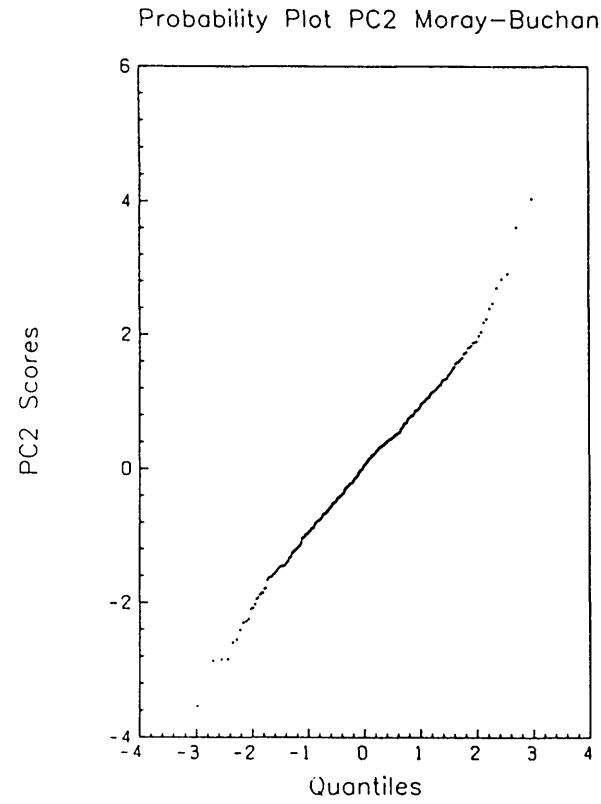

Probability Plot PC1 Moray-Buchan Logten Data (Rot°)
PC1 Scores
Quantiles
4
2
0
-2
-4
-4 -3 -2 -1 0 1 2 3 4

Figure 8.23


Probability Plot PC2 Moray-Buchan Logten Data (Rot°)
PC2 Scores
Quantiles
6
4
2
0
-2
-4
-4 -3 -2 -1 0 1 2 3 4

The point symbol map of PC1 scores was generated with class intervals selected to emphasise the high negative scores - since V, Ti, Fe, Ni, Cu, Cr and Co have high negative loadings on this component. The conclusions from study of the two PC score maps (Figures 8.24 and 8.25) are as follows:

1) The PC1 scores symbol map shows enhancement of elements of basic association over Tayvallich/Crinan subgroup mixed pelites of Upper Middle Dalradian age and over Upper Dalradian greywackes and pelites. Lower Dalradian rocks show a marked depletion of those elements.

2) The PC2 scores symbol map shows enhancement of K and Rb in two localities. One is in sites draining Lochaber subgroup Quartzites of Lower Dalradian age and supplying the River Spey near Rothes. The second is in sites draining Ballachulish subgroup mixed pelites of Lower Dalradian age about 6km north of Keith.

Figure 8.24

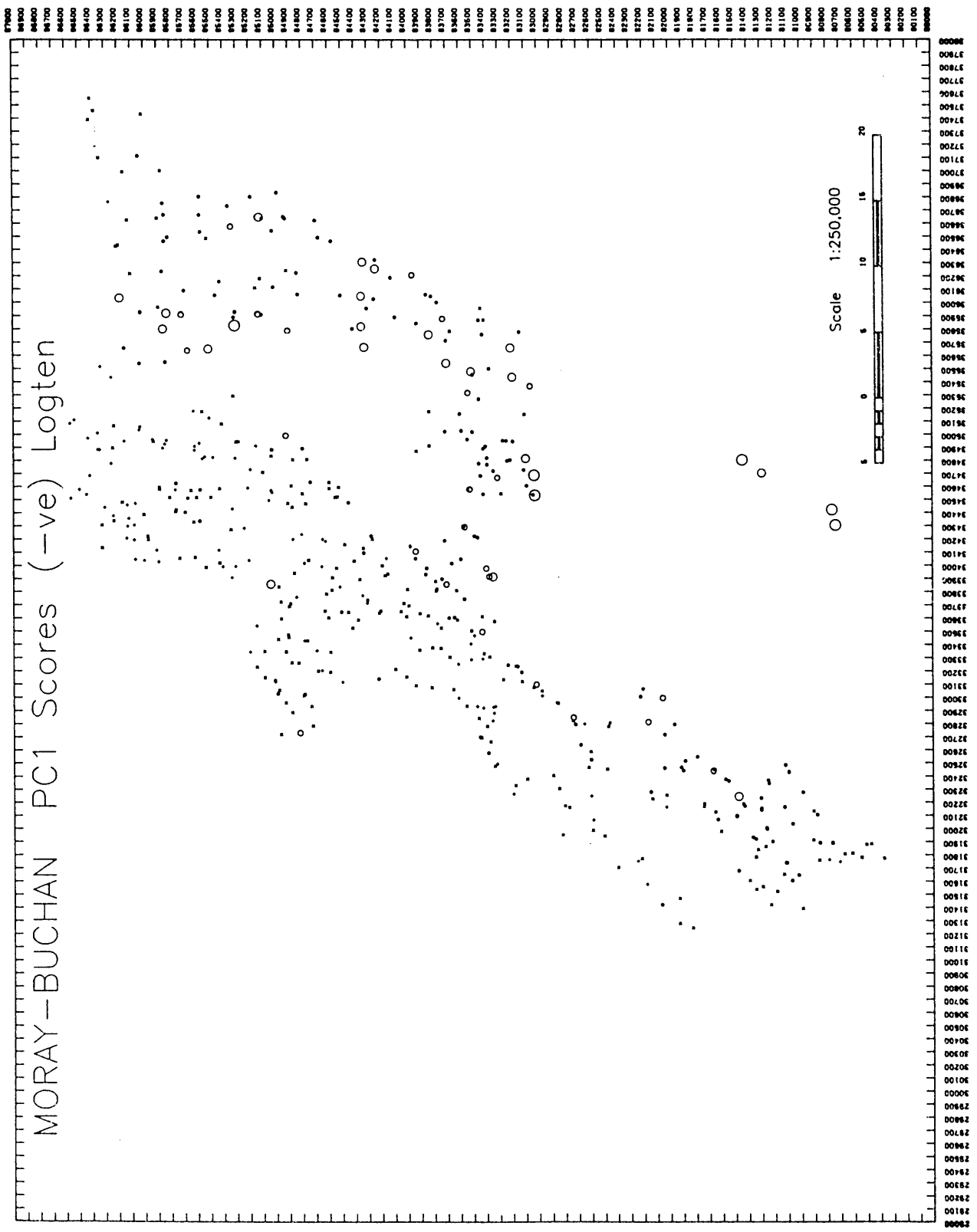
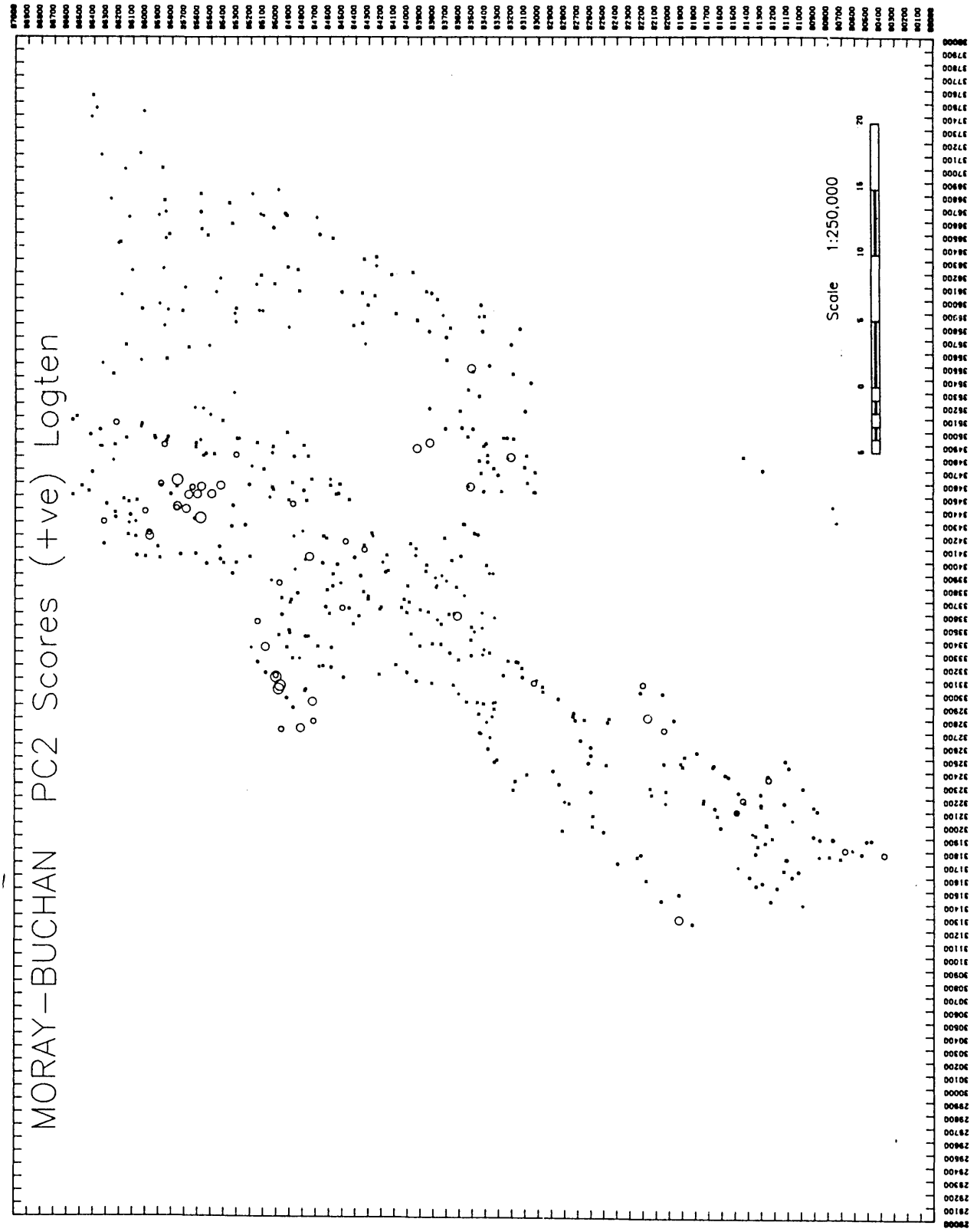


Figure 8.25



8.11.2 PCA - Raw Data

The rotated component loadings matrix generated from the raw robust correlation matrix shows the following features.

1) The elements Ti, V, Fe, Ni, Cu and Cr have high negative loadings on PC1. This component behaves in a very similar fashion to PC1 from the logten PCA run (above).

2) Potassium and Rb have high negative loadings on PC2. Apart from the change of sign this component is very similar to PC2 from the logten PCA run (above).

3) Calcium and Mg have high negative loadings on PC3. From here the similarities between raw robust and logten non-robust loadings matrices end. The former loadings matrix has high single element loadings on PC4 and subsequent components, with Mn on PC4, Y on PC5, Be on PC6, Pb on PC7, Sr on PC8, etc.

The probability plots of PC1 and PC2 scores for raw data with raw robust loadings matrix show distinct negatively skewed distributions (Figure 8.26) indicating that there is an anomalous population of high negative scores for both

Probability Plot PC1 Scores Moray-Buchan Raw Data (Rob. Raw)

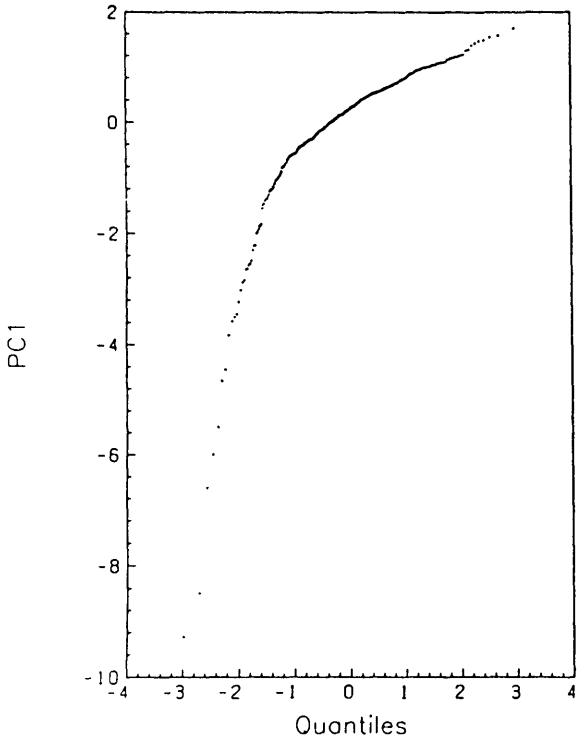
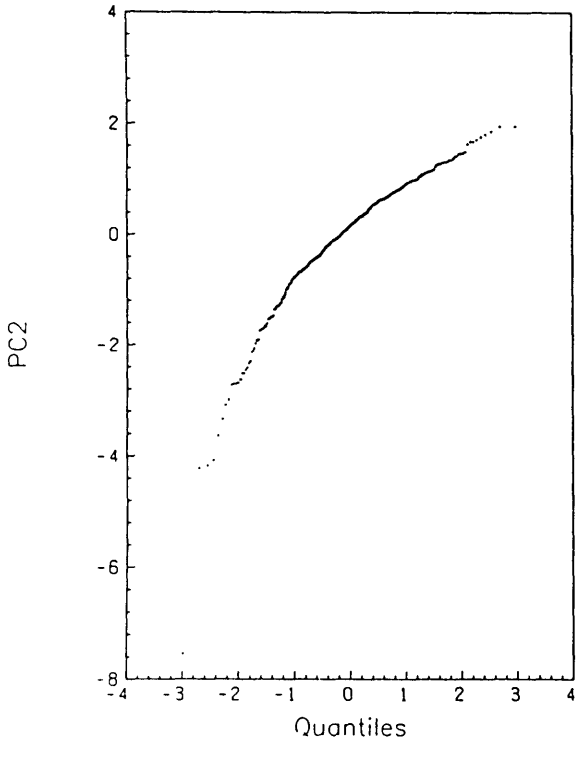


Figure 8.26

Probability Plot PC2 Scores Moray-Buchan Raw Data (Rob. Raw)



principal components. The point symbol maps for PC1 and PC2 (Figures 8.27 and 8.28) show the same trends as those for logten data (see above). Since the choice of class intervals for the point symbol maps is based on percentiles, the appearance of a map is independant of transformation. In other words a point symbol map of a raw variable will be identical to that of the same variable logten transformed. So, apart from the relatively minor changes in the PC1 and PC2 scores caused by the slight differences between the raw robust and logten loadings on PC1 and PC2, the two maps of raw scores will appear very similar to those of logten scores.

The major difference between PCA on logten data using a logten non-robust correlation matrix and PCA on raw data using a raw robust correlation matrix is that in the latter the frequency distributions of the individual component scores tend to be distinctly skewed whereas in the former they tend to closely approximate normality.

8.11.3 FA - Logten Data

Let us now see to what extent the interpretation changes when FA is performed on the same data combinations. Since only the first six of the eigenvectors generated from the logten non-robust correlation matrix have eigenvalues

Figure 8.27

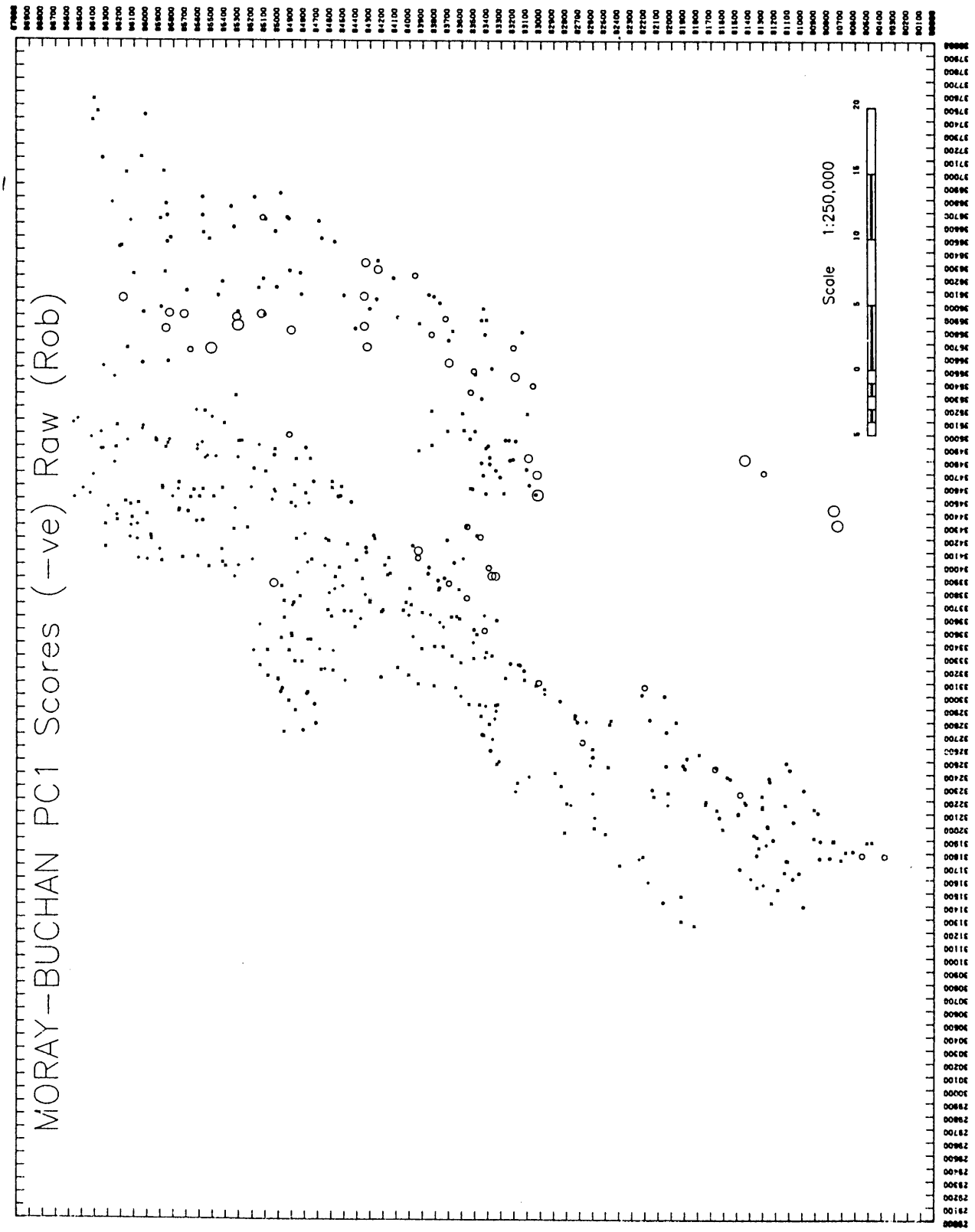
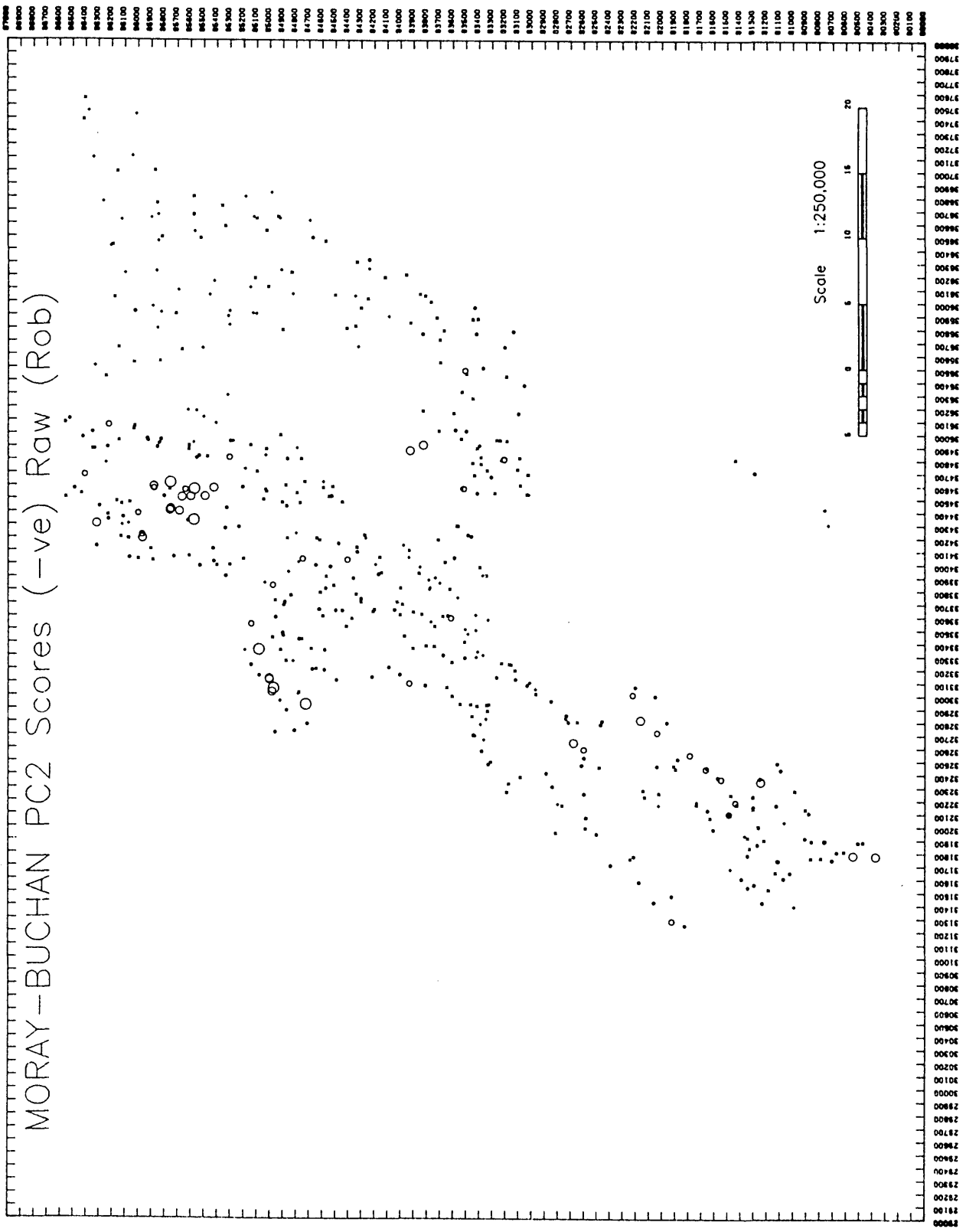


Figure 8.28



greater than one, only six factors are retained in FA. The rotated factor loadings matrix yields the following conclusions:

1) Ti, V, Ni, Cr, Fe, Cu and Co have high negative loadings on F1 which evidently has basic/ultrabasic associations.

2) Ca, Sr and Mg have high negative loadings on F2 while K and Rb have high positive loadings. This demonstrates the antipathetic relationship between these two variable groups - something that has been clearly seen in the correlation linkage diagrams (see Figure 6.5). High negative scores will therefore reveal samples high in Ca, Sr and Mg, while those samples high in K and Rb will be revealed by high positive scores.

3) Y, U, La and Zr have high positive loadings on F3. This factor is probably related to felsic igneous rocks.

4) Li, Al, Cu and Mg have high negative loadings on F4.

5) Mn, Co and Zn have high positive loadings on F5.

6) B and Pb have high positive loadings on F6.

The probability plots of factor scores for logten data

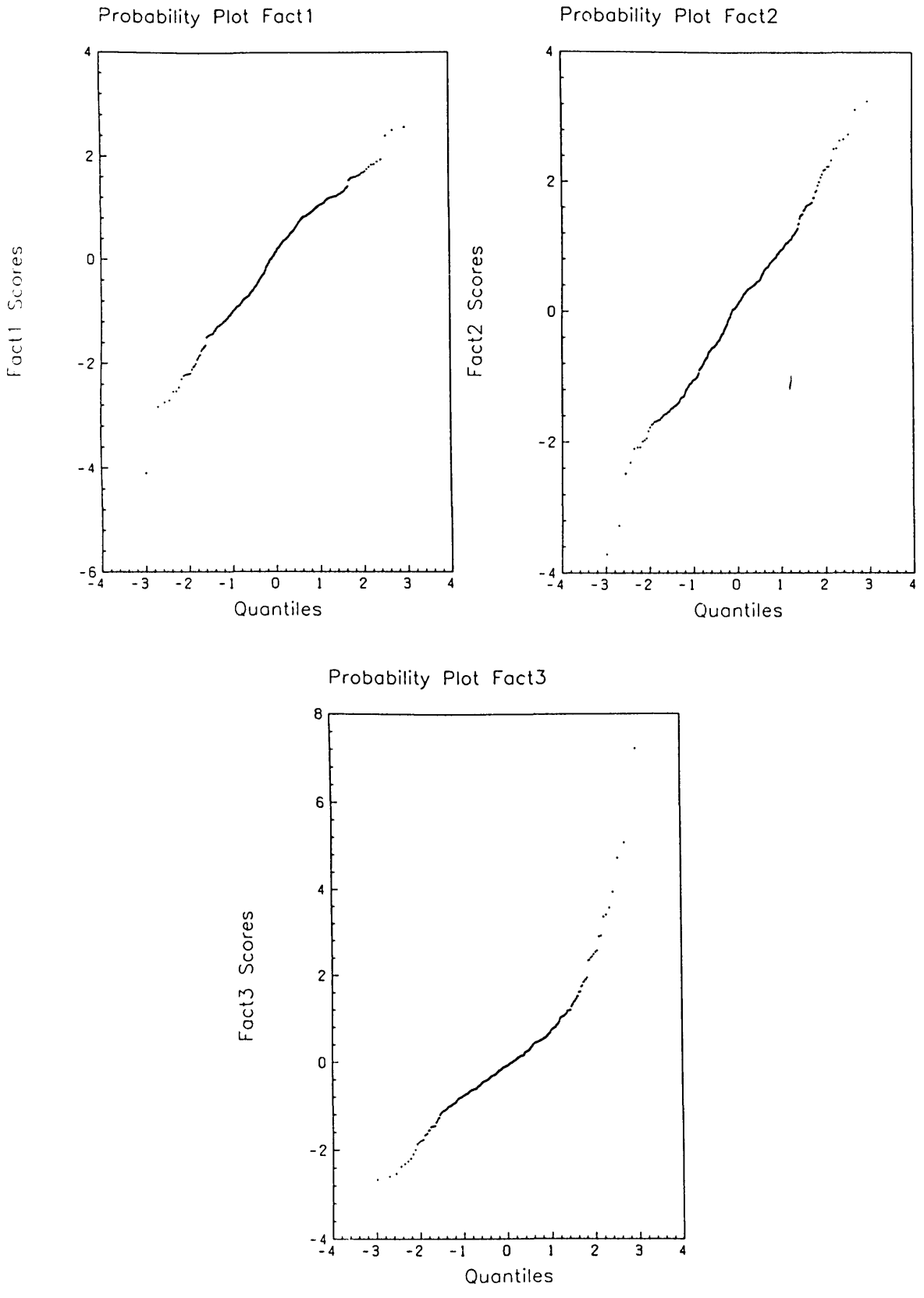
(see Figures 8.29 and 8.30) show only very minor positive or negative skew. The point symbol maps show the following relationships (see Figures 8.31 through 8.37):

1) F1 scores show that the Lower Dalradian and the Lower Middle Dalradian is depleted in elements of basic association. These elements are enhanced generally in the Upper Dalradian, in Tayvallich/Crinan subgroup mixed pelites of Upper Middle Dalradian age, in Easdale subgroup pelites south of Portsoy and in pelites and blackschists of undifferentiated Lower/Middle Dalradian age approximately 10km SE of Dufftown.

2) F2 scores are shown as two maps - one to emphasise the negative scores, the other to emphasise the positive scores. The former shows that Ca, Sr and Mg are depleted in the Lower Dalradian but enhanced in the extreme south of the study area, close to metamorphosed limestone outcrop of undifferentiated Lower/Middle Dalradian age, also about 5km S and SW of Dufftown in undifferentiated Lower/Middle Dalradian age mixed pelites and in Tayvallich/Crinan subgroup mixed pelites of upper Middle Dalradian age.

The second map shows that K and Rb are enhanced in Lochaber subgroup quartzites near Rothes, and in Ballachulish subgroup mixed pelites about 6km north of Keith. A third collection of dominant scores occurs

Moray–Buchan Logten Data (Rot°) **Figure 8.29**



Moray–Buchan Logten Data (Rot°) Figure 8.30

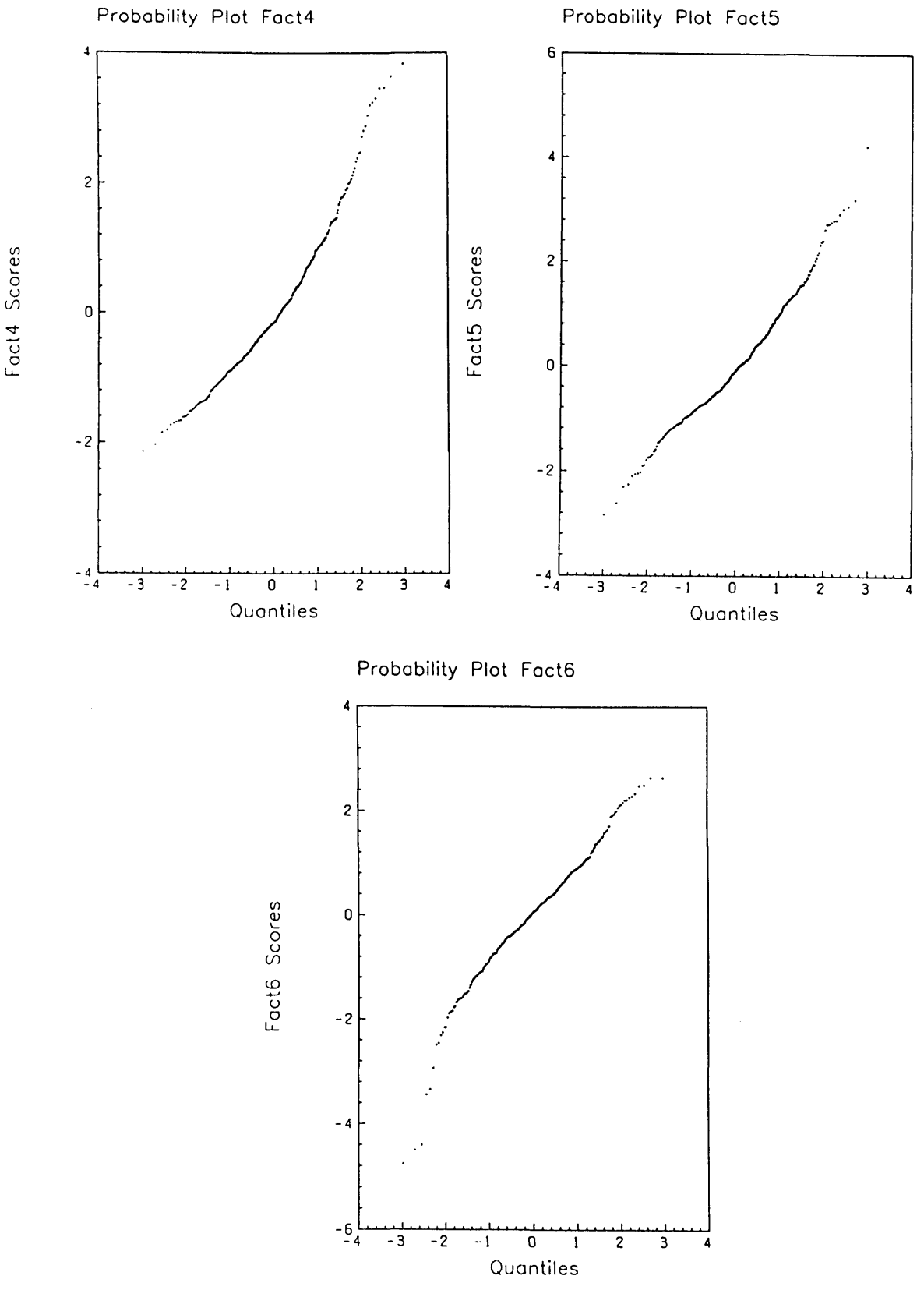


Figure 8.31

Figure 8.32

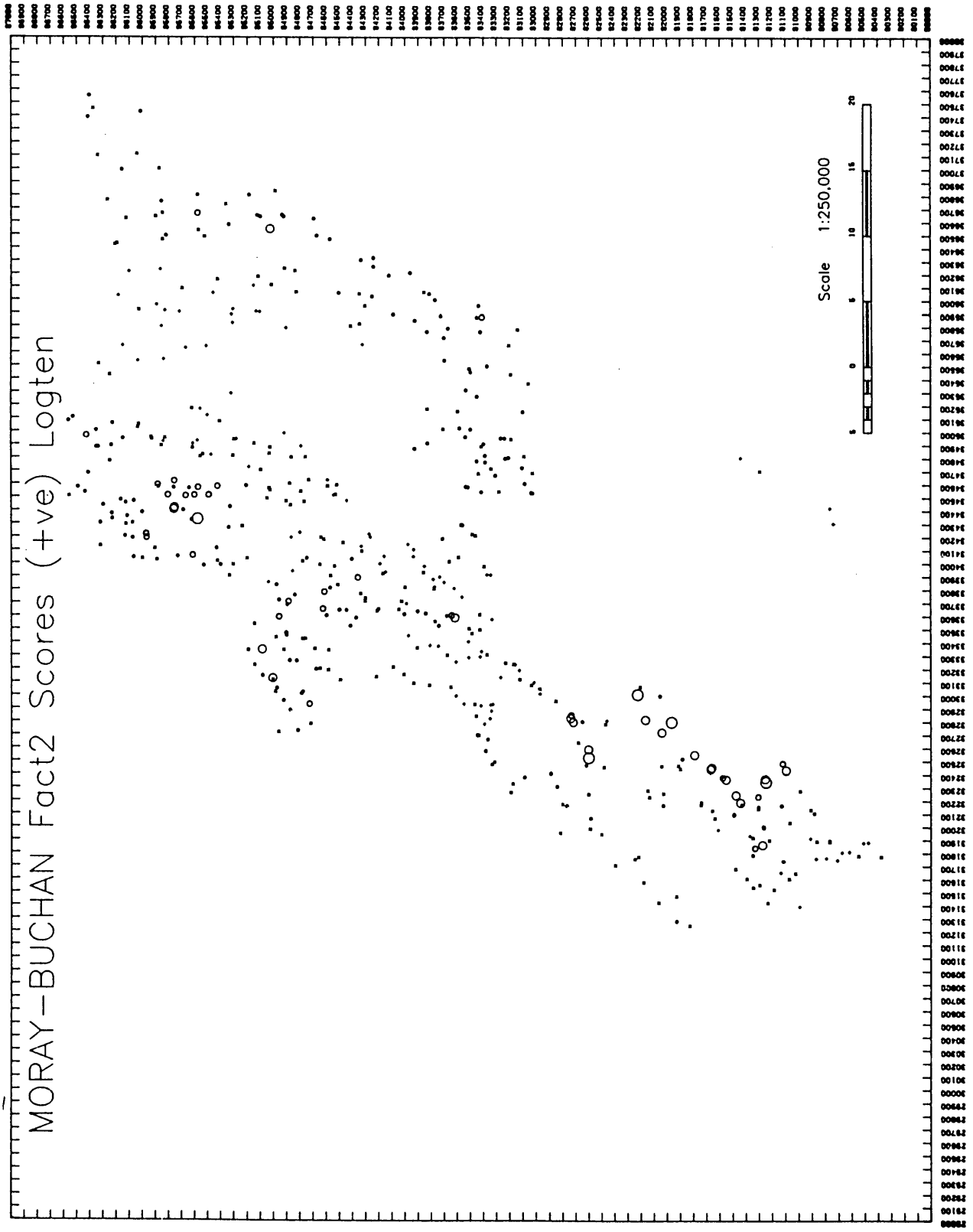


Figure 8.33

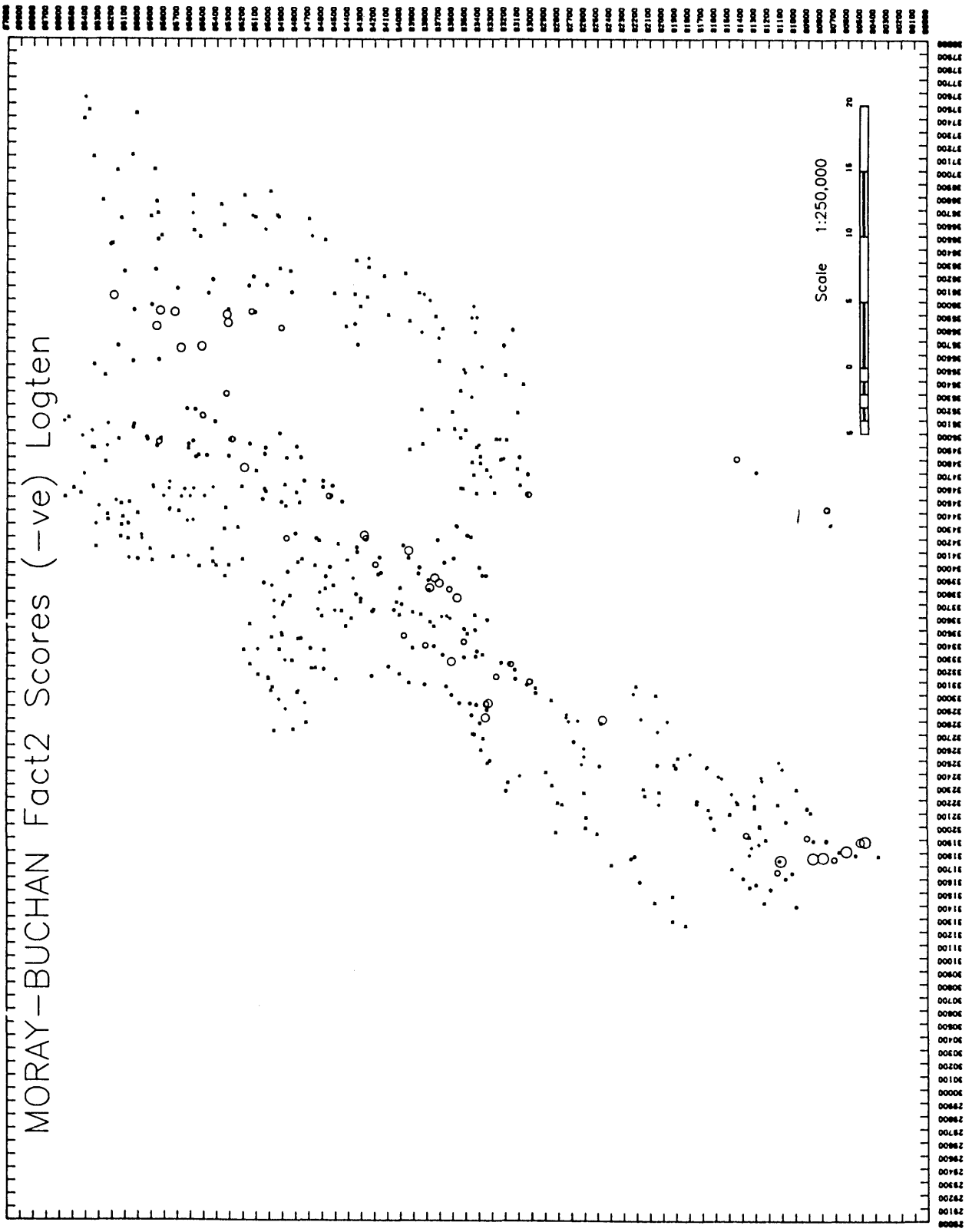


Figure 8.34

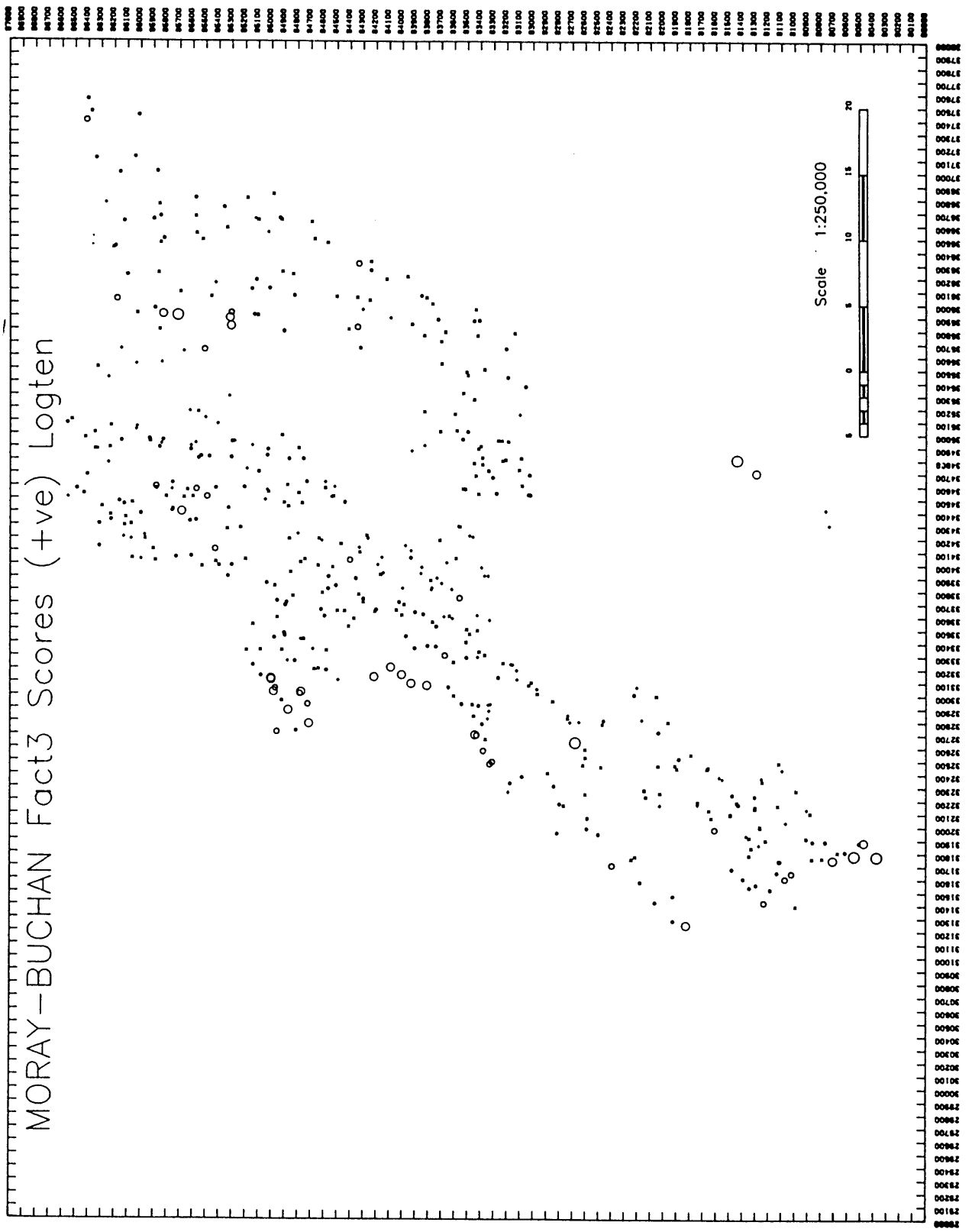


Figure 8.35

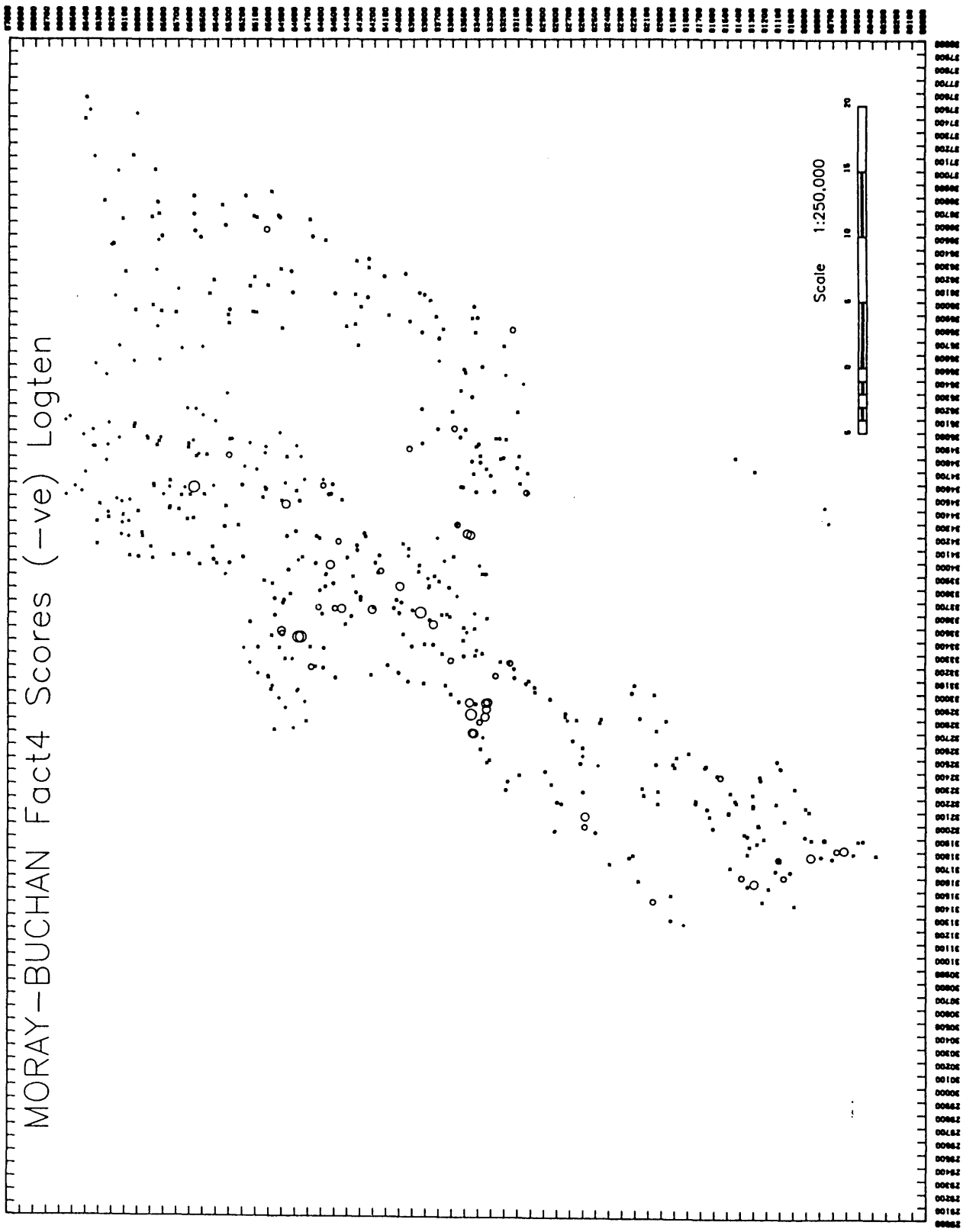


Figure 8.36

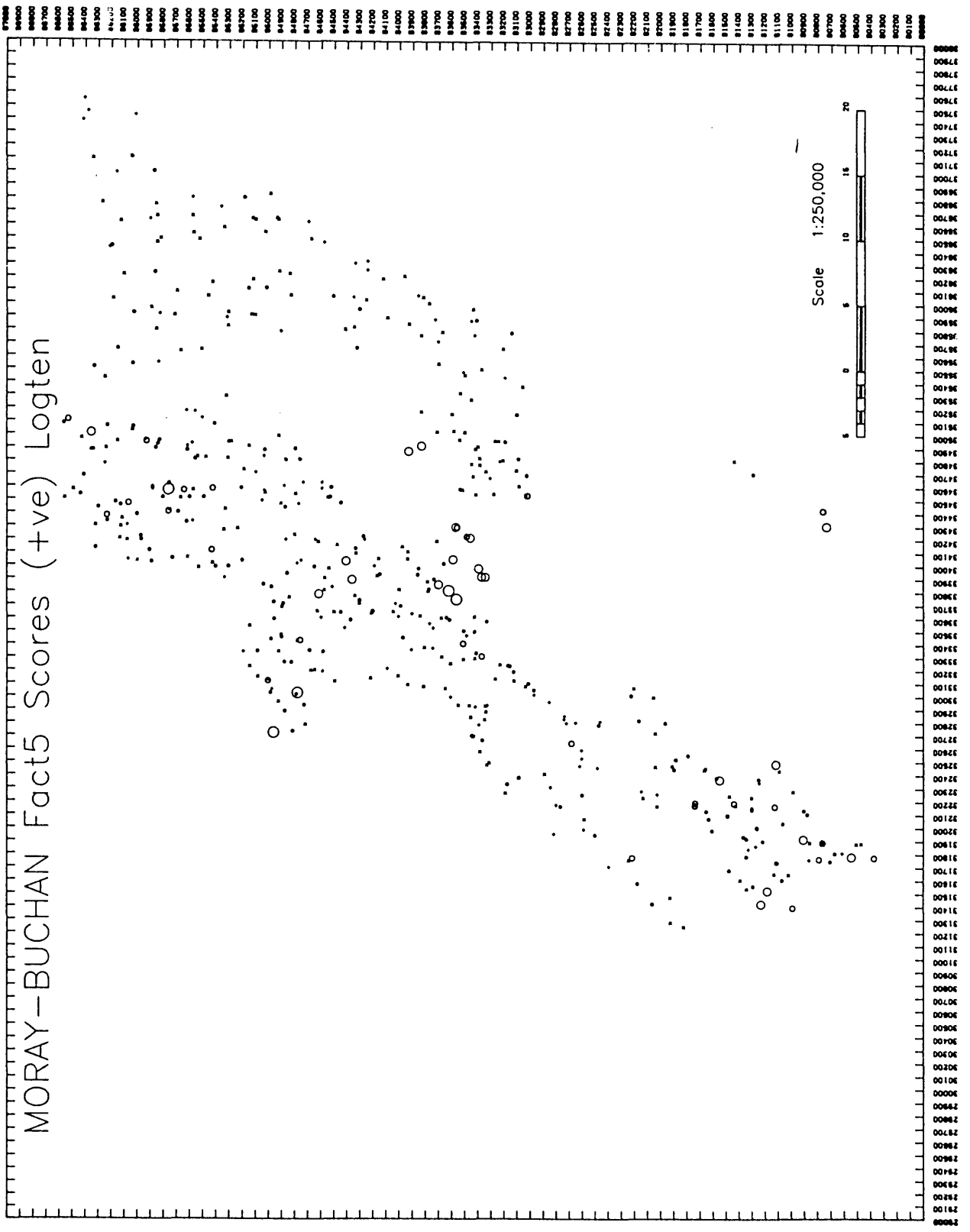
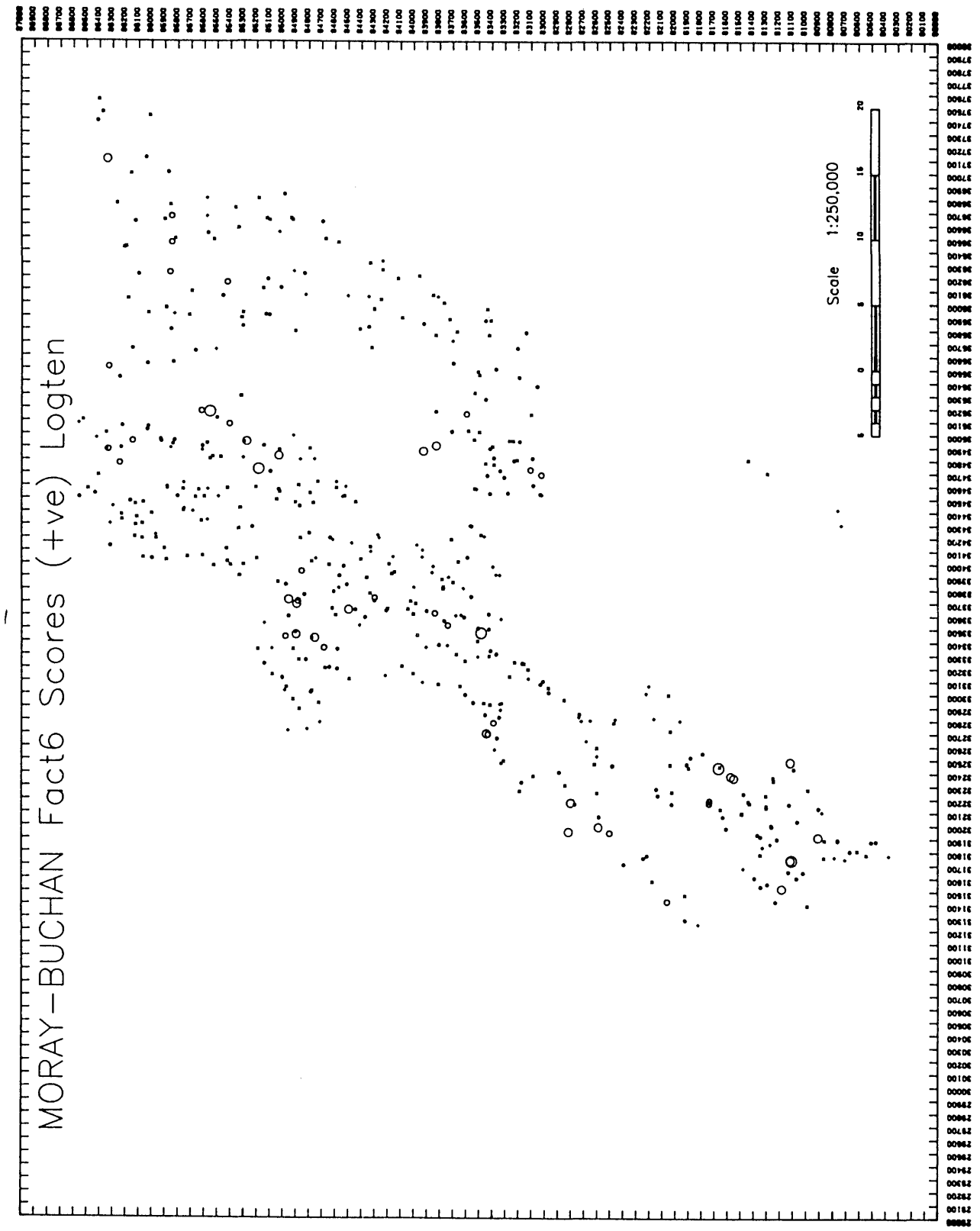


Figure 8.37



about 10km E to SE of Tomintoul over undifferentiated Lower/Middle Dalradian pelites, quartzites and blackschists.

3) F3 scores show that Y, U, La and Zr have a close association with granitic outcrop, being enhanced near both the Cairngorm and the Ben Rinnes granites. In addition, these elements are enhanced over Lochaber subgroup quartzites near Rothes, and over Tayvallich/Crinan subgroup mixed pelites. Both the latter occurences seem to be unrelated to obvious felsic igneous sources and could indicate that these elements have another association, possibly with blackschists.

4) The three subsequent symbol maps do not show any dominant trends.

8.11.4 FA - Raw Data

Let us now compare the above results with those from FA on a raw robust correlation matrix with scores generated using raw data. The conclusions from the analysis of the rotated factor loadings matrix with seven retained factors are as follows:

1) Ti, V, Cu, Fe, Ni, Al and Cr have high negative loadings on F1. Despite the removal of Co and the inclusion of Al this factor still has basic/ultrabasic associations.

2) K, Rb and Ba have high negative loadings on F2, and no longer share this factor with Ca and Mg. Note that Ba is included in this factor suggesting that the scores on this factor may reveal more information about barian-silicates than would the equivalent factor in the logten loadings matrix.

3) Ca and Mg have high negative loadings on F3. Sr is not significantly loaded on this factor.

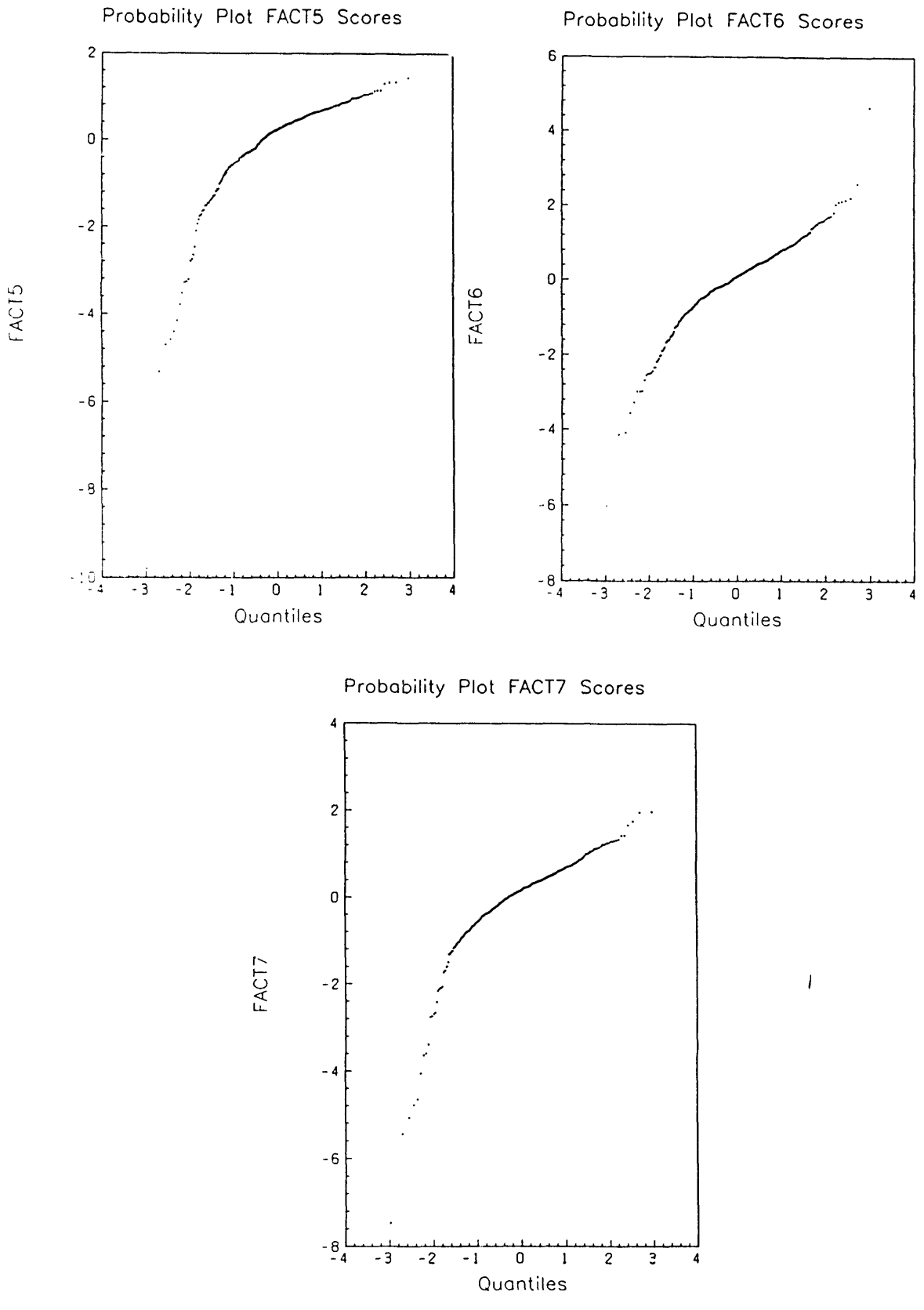
4) Y, U, La and Zr have high positive loadings on F4. This factor is similar to F3 in the logten loadings matrix (above).

5) Mn and Co have high negative loadings on F5.

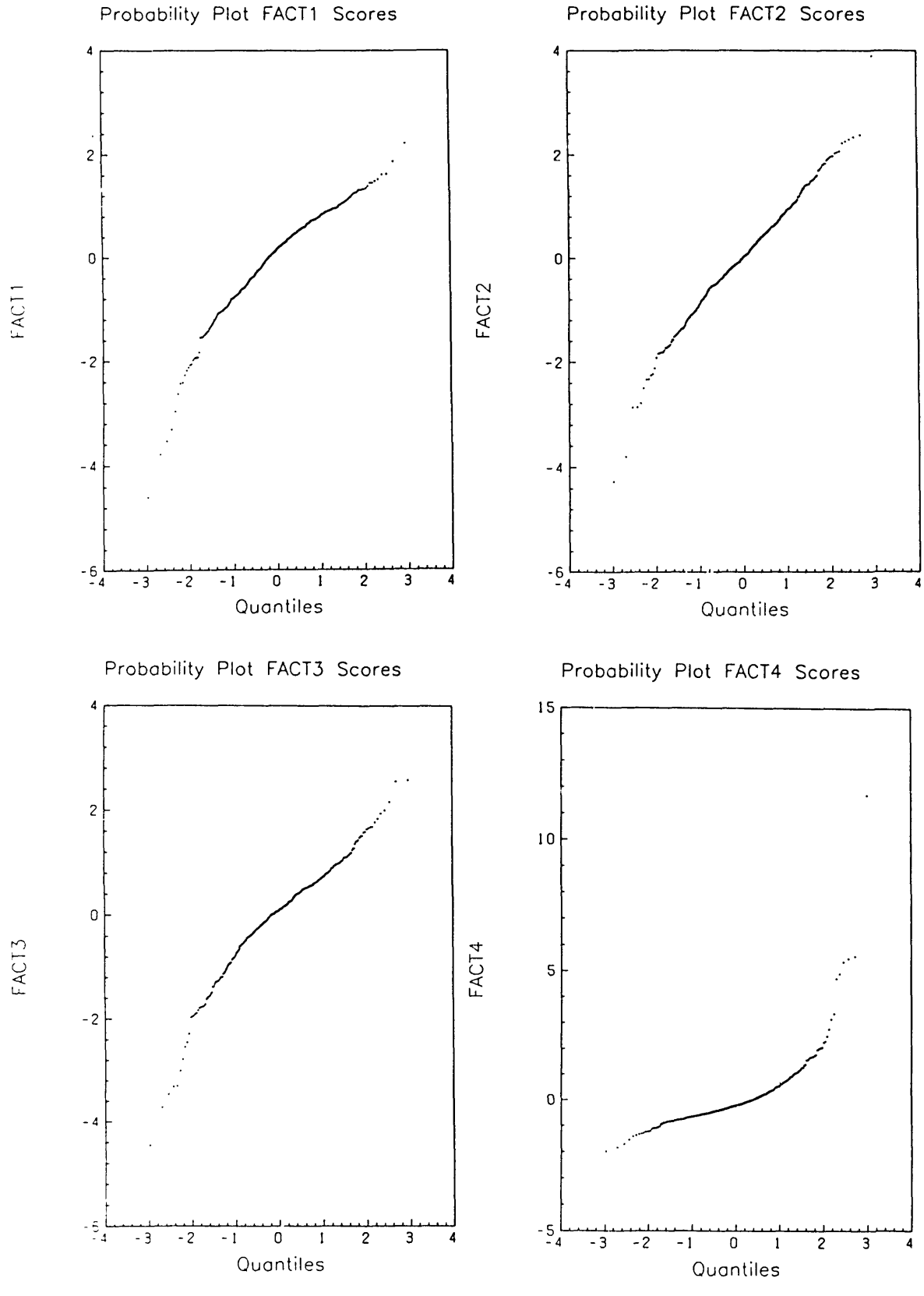
6) B, Be and Li have high negative loadings on F6.

7) Pb and Zn have high negative loadings on F7.

The probability plots of the first three factor scores (see Figures 8.38 and 8.39) show only slight negative


Probability Plot FACT5 Scores
FACT5
Quantiles
Probability Plot FACT6 Scores
FACT6
Quantiles
Probability Plot FACT7 Scores
FACT7
Quantiles

Moray–Buchan Raw Data (Rob. Raw) **Figure 8.39**



skews, F4 scores show a strong positive skew and scores on F5, F6 and F7 show strong negative skews. The sign of the skew is dependent on the sign of the dominant loadings in the factor concerned.

The point symbol maps for raw data scores (Figures 8.40 through 8.46) show close similarities with those for logten data:

1) Raw data score maps for factors one to four correspond to logten data score maps for factors one to three (bearing in mind that logten factor two has both high positive and high negative loadings) with only minor differences.

2) F5 scores show that Mn and Co have a fair spread of highs over the study area. There is one dominant concentration of large values about 10km SE of Dufftown over undifferentiated Lower/Middle Dalradian blackschists and pelites.

3) F6 scores show that B, Be and Li are generally enhanced over blackschists of the Blair Atholl subgroup of the Lower Dalradian and blackschists of undifferentiated Lower/Middle Dalradian age.

4) F7 scores show that Pb and Zn have a scatter of large values over the study area, with the main collection to

Figure 8.40

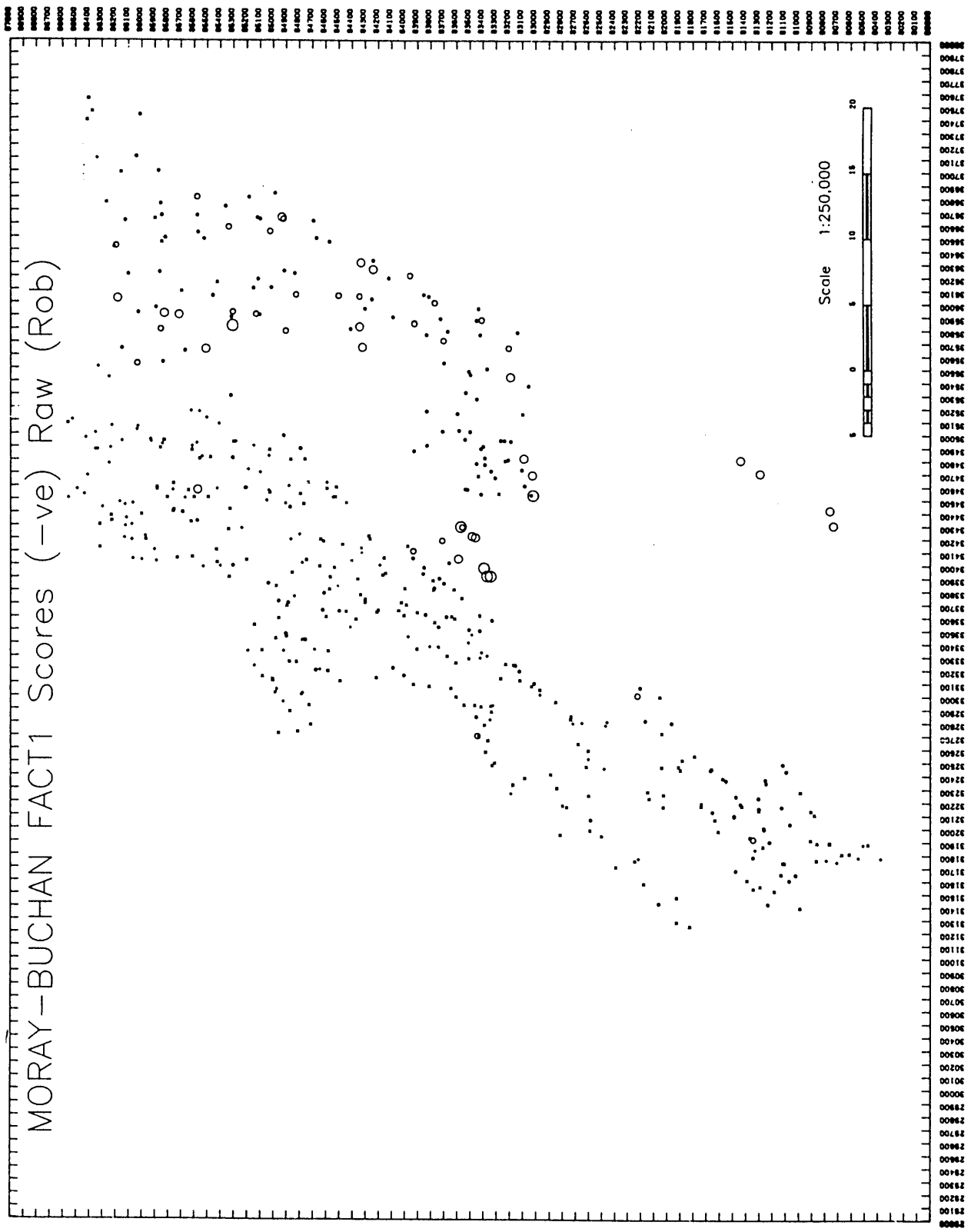


Figure 8.41

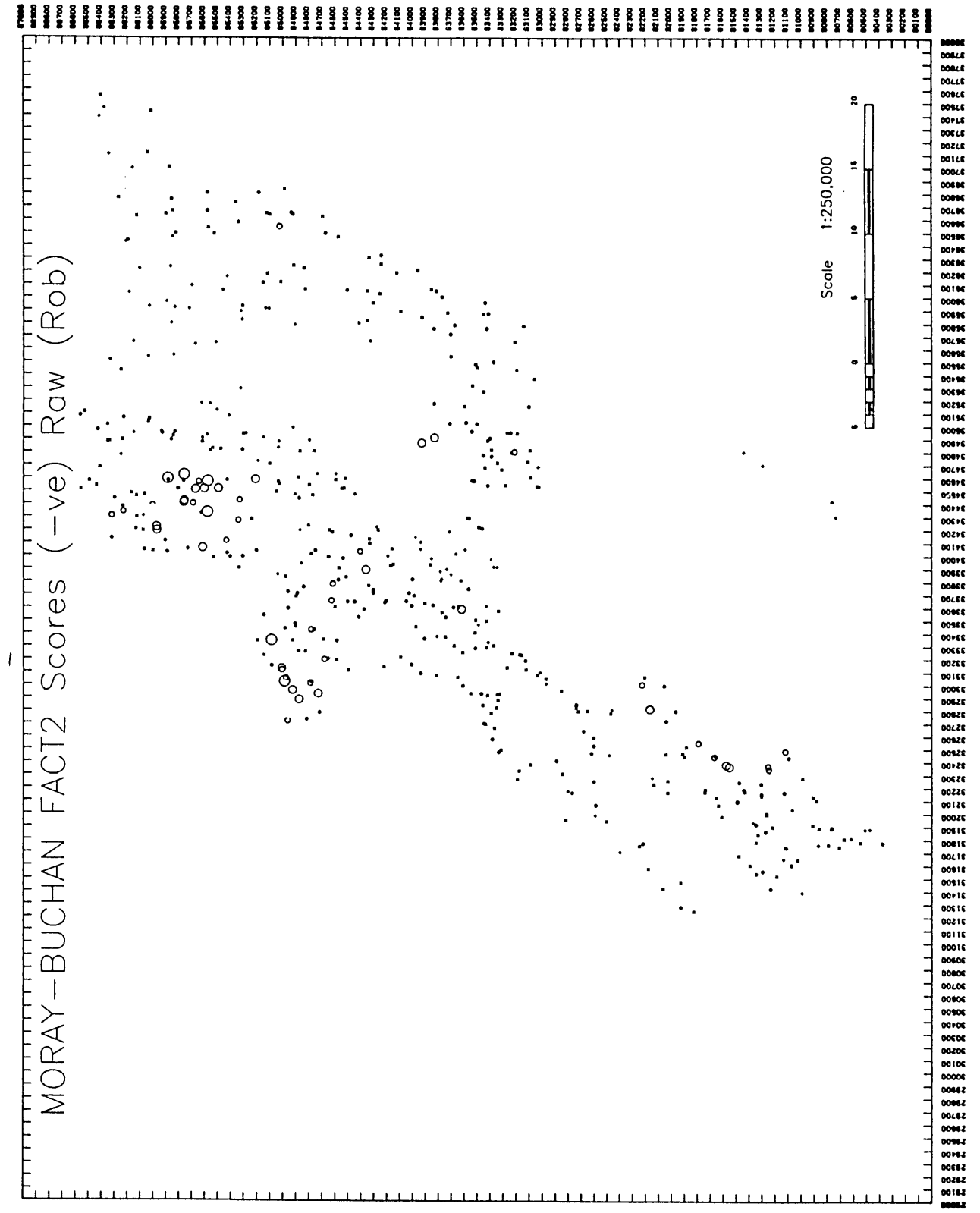


Figure 8.42

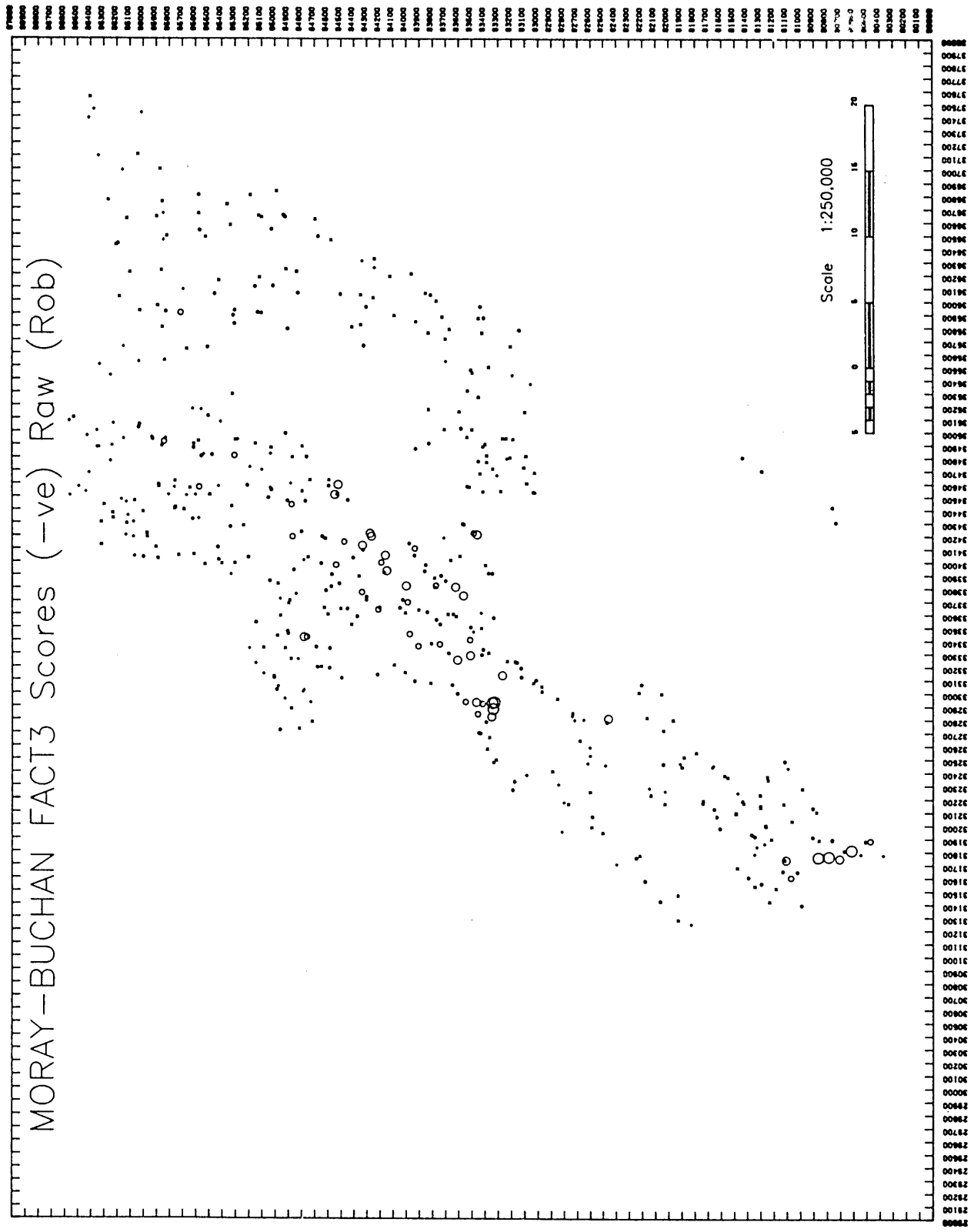


Figure 8.43

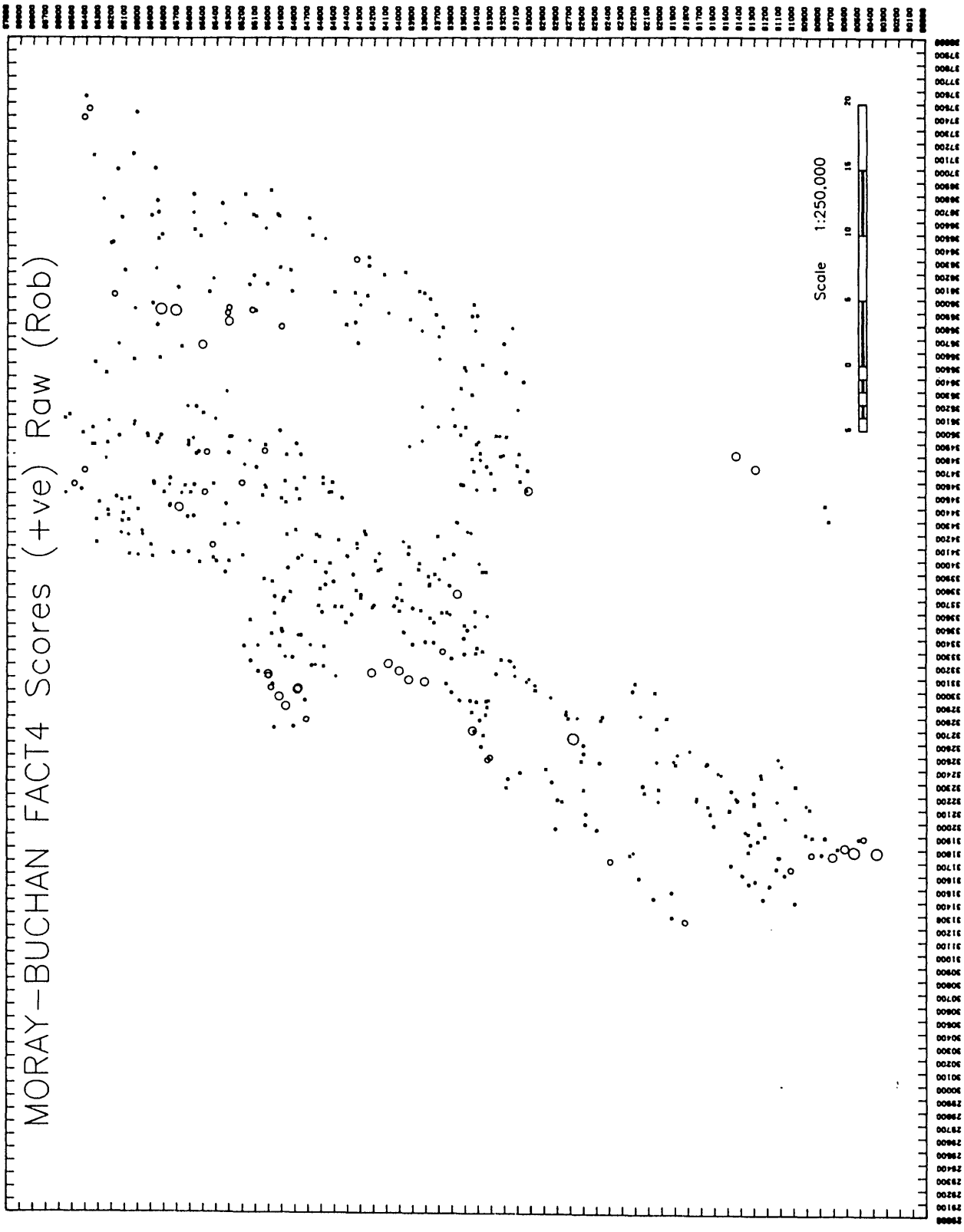


Figure 8.44

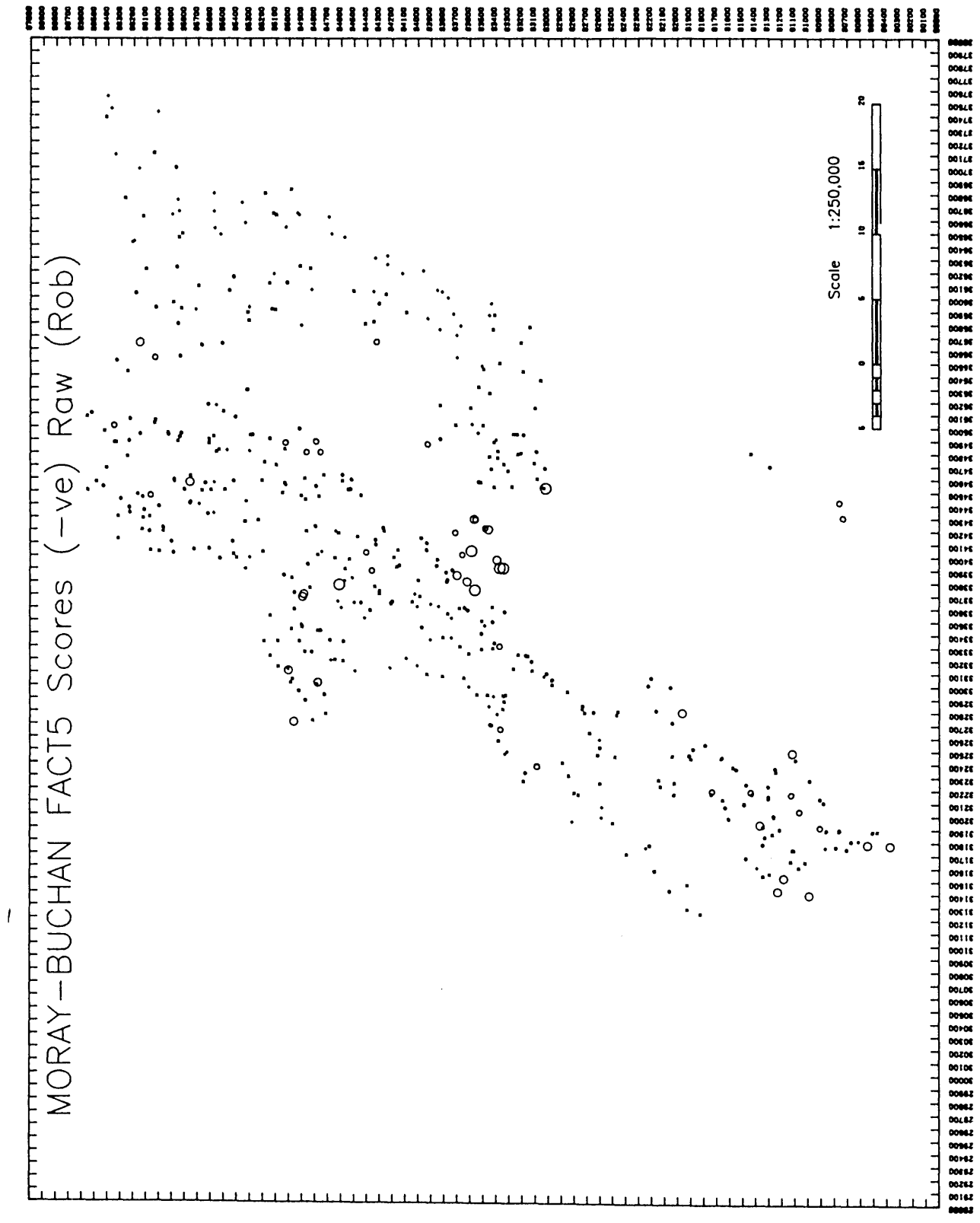


Figure 8.45

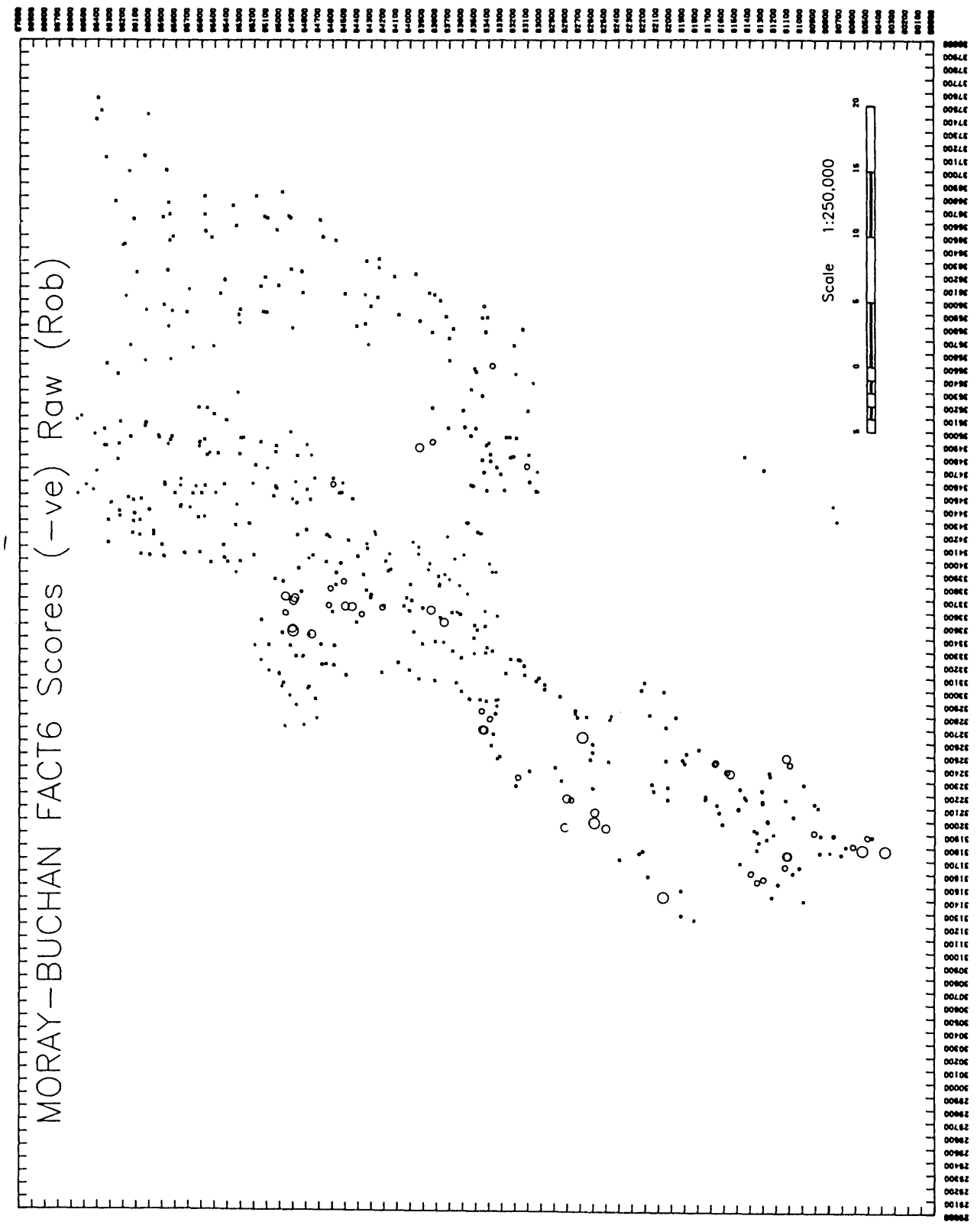
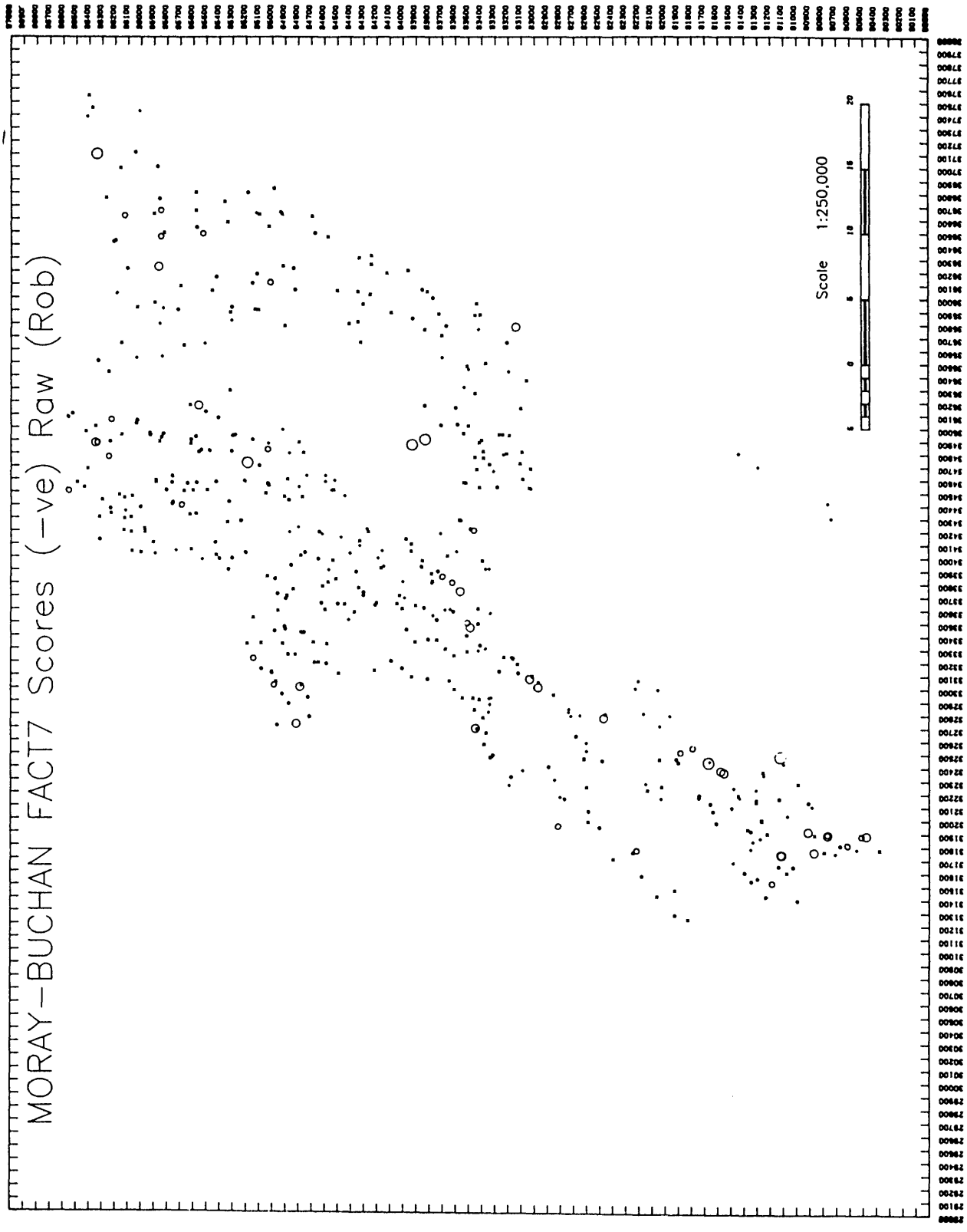


Figure 8.46



the S and E of Tomintoul on pelites and blackschists of undifferentiated Lower/Middle Dalradian age. In addition there are two large values about 5km W of Huntly over Tayvallich/Crinan subgroup mixed pelites.

8.11.5 Moray-Buchan Data: Conclusions

1) The two aims of PCA/FA are determination of variable associations and dimension reduction. Bearing these in mind, FA is preferable to PCA because the use of a small number of factors forces a grouping of the variables and by definition reduces the number of dimensions. In PCA the tendency is for only one variable to be highly loaded on a component - producing a set of scores that is of little more use than the original data. In FA the reduction of dimensions forces the generation of variable groups upon the data. The variable groups or associations tend to reflect the structure of the correlation matrix from which the factor loadings matrix was generated. So the variable associations generated from the raw non-robust correlation matrix will tend to be distorted by the presence of outliers in the raw data and will not reflect true background associations. A robust or transformed (or both) correlation matrix in which outliers have been suppressed will tend to produce

variable associations which reflect the background structure of the data.

2) The choice of which correlation matrix to use to generate a factor loadings matrix is a subjective one. It is possible to get a measure of the efficiency of the reduced set of factors from the communalities of the factor loadings matrix. The sum of the communalities divided by p (the number of variables) is equal to the proportion of the total variance accounted for by the retained factors. This proportion will tend to increase as the number of retained factors increases. In the case of the MB data it is the factors generated from the boxcox robust correlation matrix which account for the highest proportion of total variance assuming the number of retained factors is constant. The following table shows the percentage of the total variance accounted for by 7 factors using six different correlation matrices:

Correlation Matrix	% of Total Variance Accounted for by 7 Factors
Raw	69.75
Robust Raw	72.47
Logten	74.44
Robust Logten	75.67

Boxcox	74.93
Robust Boxcox	76.05

These figures show that the use of a robust matrix is more efficient than a non-robust matrix. They also show the benefit of transformation. However, the variable groupings revealed by the six factor loadings matrices differ considerably. The factor loadings matrix which gives the most geologically meaningful variable groupings may not be the most efficient. The ultimate choice of correlation matrix must therefore depend largely on a subjective judgement. In the case of the MB data the variable groupings for the raw robust loadings matrix were preferable to those for the logten or boxcox robust matrices.

3) Once the appropriate correlation matrix has been decided, and the factor loadings matrix calculated, the factor scores are generated. The raw, logten or boxcox data set may be used for the calculation irrespective of the origin of the correlation matrix used to generate the loadings matrix. If either of the transformed data sets are used the scores tend to exhibit some of the characteristics of that data. The MB logten scores tend to have univariate and multivariate ($D_m^{2/3}$) frequency distributions which are close to normal and which

exhibit low anomaly contrast. On the other hand the raw data scores tend to show skewed distributions and exhibit high anomaly contrast. As explained above, the point symbol maps presented do not make clear this contrast. However, it is possible to generate symbol maps based on statistical parameters which emphasise distributional anomalies, in which case the raw score maps would be of considerably more interest and use than the logten score maps.

4) To summarize, the preferred course of action involves generating a rotated factor loadings matrix which is not only efficient in terms of accounting for the total variance, but reveals geologically meaningful variable groupings. To gain efficiency and to reduce distortion due to outliers, a robust loadings matrix is recommended. To obtain the most geologically meaningful variable groupings a comparison of the loadings matrices generated from the raw robust and the boxcox robust correlation matrices is recommended.

If anomaly detection is the aim of the analysis then the factor scores should be calculated using the raw data to give maximum anomaly contrast. The scores may be treated as a transformed and reduced data set, and univariate and multivariate statistics may be performed upon them.

If determination of background associations and

relationships is the aim of the analysis then the factor scores should be calculated using the boxcox data.

CHAPTER NINE

MULTIPLE REGRESSION ANALYSIS

9.1 Introduction

Scientists of all fields are often interested in the explicit relationships which exist between two or more sample sets of observations, especially if preliminary tests have shown that their means are significantly different. As a result, models are built to explain such differences. A useful technique in this respect is that of linear regression.

The term regression was introduced to statistics by Sir Francis Galton in a series of papers published late last century. He used the word more in the sense of its dictionary definition: "backward movement; retreat; return of curve; reversion". His observations concerned the tendency for the heights of offspring of parents of a given height to "regress" or move towards the average population height. Current statistical practice is to use the term to refer to the fitting of curves to sample observations.

Regression analysis is now widely used as an aid to the interpretation of geochemical data, particularly with a

view to detecting those samples in which element concentration is attributable to processes of mineralization, in contrast to more widespread geochemical associations.

9.2 The Regression Model

Multiple linear regression is use to study the linear dependency of one variable (called the response variable) on a set of other variables (called the explanatory variables). The mathematical model for multiple regression is (in vector form):

$$Y = XB + E \qquad \text{Equation 9.1}$$

where Y is an n x 1 vector of observable response observations, X is an n x p matrix of observable explanatory variables, B is a p x 1 vector of unobservable regression coefficients and E is an n x 1 vector of unobservable residuals. The following assumptions are made about E :

1) E_i, i = 1 to n are random variables with mean zero and a common but unknown variance σ^2 ;

2) E_i and E_j, $i \neq j$, are independent; and

3) E_i, i = 1 to n are normally distributed.

The unobservable parameters **B** and **E** are estimated by **b** and **e** respectively, where **b** is a p x 1 vector of estimated regression coefficients b_0 - b_q (where q is the number of explanatory variables = p-1) and **e** is an n x 1 vector of observed residuals. Thus the mathematical model becomes:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$$

and

$$\mathbf{e} = \mathbf{Y} - \mathbf{Xb}$$

The <u>least squares estimator</u> (LSE) and maximum likelihood estimator of **B** is **b** where:

$$\mathbf{b} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Y} \qquad \text{Equation 9.2}$$

This **b** minimizes the <u>residual sum of squares</u> (RSS)

$$e^2 = (Y - Xb)' (Y - Xb)$$

and is the minimum variance unbiased estimator of **B** .

9.3 Standardization

Marquardt and Snee (1975) consider standardization (see also Section 8.3) of the variables essential. They say that ill-conditioning of the (X' X) matrix resulting from failure to standardize is all the more insidious because it is not due to any real defect in the data, but only to the arbitrary origins of the scales on which the explanatory variables are expressed. The relative importance of regression coefficient estimators generated from standardized data (termed standardized estimators) can be easily established by direct comparison - something which is impossible when the estimators are unstandardized.

The (X' X) and (X' Y) matrices generated from raw data are matrices of uncorrected sums of squares and sums of products (SSSP). If the raw data are standardized then Y^* defines an n x 1 vector of standardized response observations where y_i^* is derived from y_i by

$$y_i^* = (y_i - \bar{y}) / s_y$$

where $\bar{y}$ and s_y are the mean and standard deviation respectively of the response variable. In the same way X^* defines an n x q vector of standardized explanatory variables where x_{ij}^* is derived from x_{ij} by

$$x_{ij}^* = (x_{ij} - \bar{x}_j) / s_j$$

$$i = 1,2,\ldots,n$$

$$j = 1,2,\ldots,q$$

where $\bar{x}_j$ and s_j are the mean and standard deviation respectively of the jth explanatory variable. The matrix $(X^* \cdot X^*)$ is then the uncorrected SSSP matrix for the standardized explanatory variables. However, since the data are standardized, this matrix is also the corrected SSSP matrix which, if divided by (n-1), is the variance-covariance matrix of the standardized data. It is not, however, necessary to standardize each variable before computing this variance-covariance matrix, because it is equivalent to the correlation matrix of unstandardized data. Equation 9.2 then becomes:

$$a = (X^{*\prime} X^{*})^{-1} X^{*\prime} Y^{*}$$

where both $(X^{*\prime} X^{*})$ and $(X^{*\prime} Y^{*})$ are subsets of the correlation matrix of the n x p total unstandardized data.

The standardized regression coefficient estimators a can be related to the original estimators b by use of the ratio of unstandardized standard deviations

$$a_j = b_j s_j / s_y, \qquad j = 1,2,\ldots,q$$

The a's may be transformed back to the original units:

$$b_j = a_j s_y / s_j$$

$$b_0 = \bar{Y} - \Sigma b_j \bar{X}_j , \qquad j = 1,2,\ldots,q$$

In subsequent discussions in which (X' X) and (X' Y) are used it will be assumed that they are in correlation form unless otherwise specified.

The least squares estimators have some good properties aswell as some less desirable ones. The desirable properties include the facts that:

1) **b** is an unbiased estimator of **B** : E(**b**) = **B** ;

2) the RSS is minimized; and

3) the variance of **b** is minimum among all linear unbiased estimators.

However, it can be shown that in situations in which the (X' X) matrix is ill-conditioned (i.e. exhibits multicollinearity) due to high correlations between the explanatory variables - a common occurrence in geochemistry - the least squares solution yields estimators whose absolute values may be too large and whose signs may actually reverse with negligible changes in the data. In such a situation of unstable estimators little confidence can be placed in the resultant residuals. Since we know that the variance of the LSE is minimized given the condition of unbiasedness, it seems logical, when looking for an improvement over least squares, to relax this condition and focus instead on minimizing $E(L^2)$, the mean squared distance between the estimates **b** and the true unknown regression coefficients **B** , where

$$L^2 = (\mathbf{b} - \mathbf{B})' (\mathbf{b} - \mathbf{B})$$

and $$E(L^2) = \sigma^2 \text{ TRACE } (X' X)^{-1}$$

$E(L^2)$ is also termed the mean squared error of $\mathbf{b}$, denoted MSE($\mathbf{b}$). If the ordered eigenvalues of $(X' X)$ are

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q > 0$$

then $$E(L^2) = \sigma^2 \Sigma(1 / \lambda_i)$$

9.4 Ridge Regression

Hoerl and Kennard (1970a) describe a procedure for removing the effects of multicollinearity by finding another vector of estimators $\mathbf{b}^*$, such that the squared distance between $\mathbf{b}^*$ and $\mathbf{B}$ is less than that between $\mathbf{b}$ and $\mathbf{B}$. The new vector $\mathbf{b}^*$ is found by moving in such a direction in parameter space that L^2 decreases. The movement will cause an increase in the RSS. Hoerl and

Kennard suggest that this procedure may be implemented by altering the matrix (X' X) by adding a small positive constant k to each of the diagonal elements:

$$b^* = (X'X + kI)^{-1} (X' Y) \qquad k \geq 0$$

The new matrix (X'X + kI) is better conditioned than (X' X) and should give more stable estimates of B .

The optimal value for k, which will minimize MSE(b^*) cannot be determined with certainty bcause it depends on the unknown coefficient vector B and unknown error variance σ^2. In practice k must be determined subjectively or estimated from the data. Hoerl and Kennard (1970b) proposed the use of a ridge trace using certain guidelines. However, using the ridge trace is still an inherently subjective procedure. As a result several authors have provided and evaluated algorithms for the automatic selection of k (see Hoerl and Kennard, 1980). Several of these algorithms, k_b (Lawless and Wang, 1976), k_{at} (Hoerl and Kennard, 1976) and k_a (Hoerl et al, 1975) were evaluated in the course of this work using a variety of both small and large geochemical data sets and were found to produce, without exception, erratic and inflated values for k. This may be a function of the data rather than the algorithms, and suggests that the algorithms are not tested

rigorously enough.

The ridge trace is a graphical representation of the change in the values of the estimators with the change in k. The trace serves to portray the complex interrelations that exist between non-orthogonal explanatory vectors and the effect of these interrelations on the estimation of **B** . It must be stressed that, when undertaking analysis of the ridge trace, the estimators should be in standardized form (**a**). This is because in this form they are dimensionless, and may, therefore, be directly compared with each other.

From a practical point of view, the value of ridge regression lies in the increased confidence which one can place on estimators generated from highly correlated data, and hence on the residuals.

9.5 Robust Regression

The linear regression model is one of the most widely used tools in statistical analysis and the least squares method is a very popular estimation technique for this model. But in spite of its obvious advantages, the LSE suffers a dramatic lack of robustness. Outliers, which result from skewed or heavy tailed distributions, or which are simply bad data points due to errors, have an unusually

large influence on the LSE. That is, the outliers pull the least squares "fit" towards them. Examination of the resulting residuals is misleading because such outliers will tend to have residuals which appear normal. Accordingly, robust methods have been created to modify least squares schemes so that the outliers have much less influence on the final estimates.

There are a number of ways of approaching this problem.

1) Outliers may be detected in some manner and trimmed from the original data set to yield a pure background sample. Least squares regression is then performed on the trimmed data (see Garrett et al, in press). The disadvantages of outlier rejection procedures compared to robust procedures are discussed in Sections 3.3.6 and 6.3.

An outlier rejection facility is incorporated into the ridge regression program BRIDGE (and used to perform the subsequent analyses) in which the residual probability plot is used to detect anomalous objects which may then be eliminated from the data. When the analyst is satisfied that the least squares estimators are no longer influenced by anomalies (i.e. when the residuals closely approximate to a normal distribution), the total data may be resubstituted into the "background" regression equation to yield a complete set of residuals which are uninfluenced by outliers.

2) Iteratively reweighted least squares (IRLS) regression involves the iterative downweighting of anomalous residuals until convergence of estimated regression coefficients occurs. The weighted linear regression model takes the form:

$$W^{1/2} = W^{1/2} XB + W^{1/2} E \qquad \text{Equation 9.3}$$

where $W^{1/2}$ is a known n x n diagonal matrix of weights. If we redefine the matrices: $Y^* = W^{1/2} Y$; $X^* = W^{1/2} X$ and $E^* = W^{1/2} E$ then Equation 9.3 becomes:

$$Y^* = X^* B + E^*$$

which transforms the IRLS model to an unweighted LS problem. Computationally, then, B can be estimated by multiplying Y_i and each element of X_i including the constant by $W_{ii}^{1/2}$, and solving the resulting LS problem:

$$b = (X^{*\prime} X^{*})^{-1} X^{*\prime} Y^{*}$$

and the residuals are:

$$E^{*} = Y^{*} - X^{*}b = W^{1/2} (Y - Xb)$$

The weights are generated from weight functions very similar in form to those functions studied in Chapter 3 for calculating robust location estimates. The measure of deviation U_i is taken to be equal to the scaled weighted residual at each iteration.

$$U_i^{(k+1)} = E_i^{*(k)} / s$$

Thus the weights $w(U_i)$ for each iteration are calculated from the scaled residuals from the previous iteration. The scaling factor s is defined once at the first iteration (the unweighted LS iteration):

$$s = \text{med} [|E_i - \text{med} (E_i)|] / 0.6745$$

The 0.6745 divisor makes this an approximately unbiased estimate of scale when the error model is Gaussian.

In practice IRLS regression is unsatisfactory in certain circumstances. If ordinary LS is used initially, prior to the first iteration, any outliers present can, in certain cases, draw the regression hyperplane towards them, with the result that they tend to have low residuals. Subsequent iterations will tend to reinforce this trend. This process will tend to occur more in cases where outliers are major ones (i.e. anomalous in a high proportion of variables rather than just one or two). A robust start, where the initial estimates of the regression coefficients are generated from a robust correlation matrix, would help solve this problem. It does not, however, alter the basic critisisms of the technique. It can also be argued that if a robust start is used, why not finish in the same way - a procedure which is advocated below.

3) Instead of downweighting objects on the basis of their deviation from the regression hyperplane, downweighting should be based on the Mahalanobis distance attributed to each object vector. The latter is the basis for the generation of the robust correlation matrix (see Chapter 6). If all outliers, both major and minor, are downweighted (not just those which happen to fall distant

from the LS regression hyperplane) then a true background population will be defined, and LS regression will describe a background fit. Residuals from this background fit will be more geologically accurate and should have maximized anomaly contrast.

9.6 Regression and the SIND Data

The aims of this section are:

1) to compare non-robust with robust ridge regression

2) to evaluate the effects on regression analysis of data transformation.

For this comparative analysis a ridge solution is always sought since the correlation matrices consistently display multicollinearity. For the reasons given in the previous section, IRLS regression is not performed. Instead, a robust calculation of the correlation matrix is undertaken using the process specified in Chapter 6.

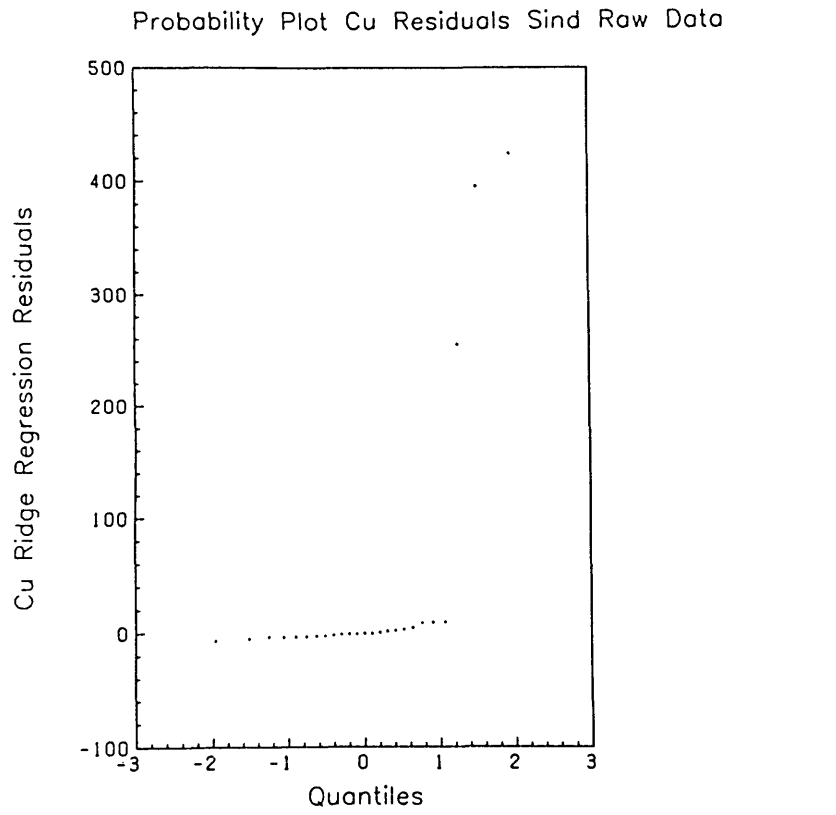
9.6.1 Robust versus Non-robust Regression

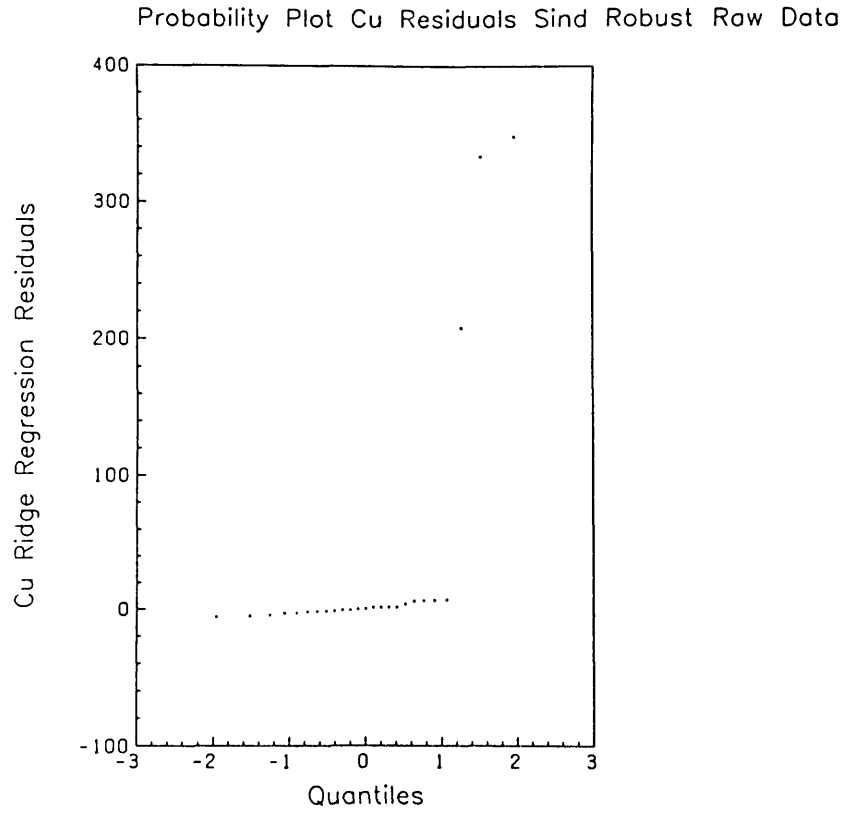
Non-robust ridge regression analysis was first performed on SIND raw data with Cu as the response variable and Zn, Fe, Mn, Cd and Pb as the explanatory variables. As would be expected from the correlation matrix, the variables which best explain Cu are Zn ($r = 0.969$) and Cd ($r = 0.906$). The first pass of the program BRIDGE reveals sample site 16 as an exceptionally high anomaly. When this observation is eliminated from the data and the analysis repeated we find that the R^2 value has increased significantly and the variance of error (σ^2) has decreased dramatically. The latter parameter is equal to the sum of the squared residuals divided by the degrees of freedom ($n - q - 1 = 18$) and since the observation with the largest residual has been eliminated a decrease in σ^2 is logical. The second pass then reveals sample sites 13 and 15 as having slightly anomalous positive residuals, with site 12 having a high negative residual. Observations 13 and 15 are eliminated from the data and a third pass analysis is performed. The variance of error is further reduced but R^2 decreases, probably due to the elimination of the anomalous observations 13 and 15. The residual probability plot at the end of the third pass shows an approximately normal frequency distribution, indicating that no further

anomalies exist, and that the regression equation characterizes the background relationship between Cu and the other variables. The total data is then resubstituted into this "background" equation and the residuals recalculated, with the result that those for the three anomalous observations 13, 15 and 16 are greatly increased in comparison to background observations (see Figure 9.1).

We can now compare the approach outlined above with one in which each pass of the program BRIDGE is based on a robust correlation matrix. Zn and Cd are again revealed to be the best explanatory variables for Cu. The variance of error at the first pass of the robust regression is very similar to that for the non-robust solution at the third and final pass (see above), in which all outliers have been eliminated. So we see immediately that robust regression is doing automatically what needed to be done manually in three steps using a non-robust approach. The robust residuals also show much the same degree of anomaly contrast as the resubstituted non-robust residuals (compare the Q-Q plots in Figure 9.1). The three anomalous observations 13, 15 and 16 are eliminated from the second pass, but, since the first pass was based on a robust correlation matrix which effectively discounts the influence of those observations, the changes are minimal. The correlation matrix shows very little change and both the ridge and the non-ridge solutions are very similar. The residual probability plot shows a close resemblance to

Figure 9.1


Probability Plot Cu Residuals Sind Raw Data
Cu Ridge Regression Residuals
500
400
300
200
100
0
-100
-3
-2
-1
0
1
2
3
Quantiles


Probability Plot Cu Residuals Sind Robust Raw Data
Cu Ridge Regression Residuals
400
300
200
100
0
-100
-3
-2
-1
0
1
2
3
Quantiles

distributional normality. Resubstitution of the total data into the regression equation generated in the second pass gives almost identical residuals to the first pass (see Figure 9.1).

The conclusions from this comparative analysis are as follows:

1) A non-robust approach to regression analysis only produces a satisfactory answer if some sort of iterative trimming is undertaken. In the case of the SIND data three iterations were required to produce a regression equation which properly characterized the background relationships between Cu and the other variables. It is only at this third pass that the observations 13, 15 and 16 are revealed as coincident anomalies.

2) Using robust regression analysis a background equation is ensured at the first pass and the three anomalous observations are revealed in the residual probability plot with a dramatic anomaly contrast.

3) A one stage, automatic procedure is obviously better in practice than a multi-stage manual process, but the latter does have theoretical drawbacks aswell. The observations eliminated at each pass are selected on the basis of having high absolute residuals. This process

does not select as anomalous those observations which lie on or near the regression hyperplane and which have a large influence on the analysis. This objection is similar to that against the IRLS regression procedure (see Section 9.5). A robust estimate of the correlation matrix, on the other hand, will reduce the influence of these troublesome objects even before the analysis has begun.

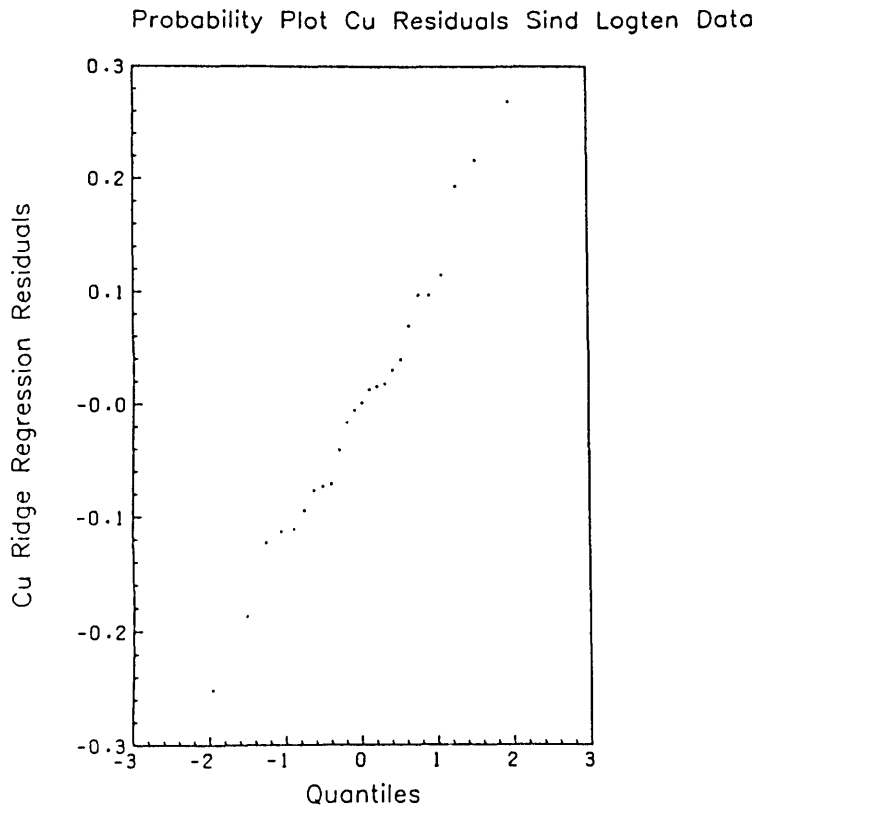
9.6.2 Transformation and Regression

To study the effect of transformation on non-robust and robust ridge regression analysis, two further sets of analyses were performed, one on SIND logten transformed data, the other on SIND boxcox transformed data.

The non-robust and robust ridge regression analyses for both sets of data produced normally distributed residual probability plots at the first pass (Figures 9.2 and 9.3). The difference between the non-robust and robust ridge regression analyses was negligible due to the close similarity between the non-robust and robust correlation matrices for the transformed data sets (see Chapter 6).

The use of transformation is, on the basis of this study, of no use in anomaly detection. It is not as if the anomalous observations 13, 15 and 16 are detected but not

Figure 9.2


Probability Plot Cu Residuals Sind Logten Data
Cu Ridge Regression Residuals
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-3
-2
-1
0
1
2
3
Quantiles

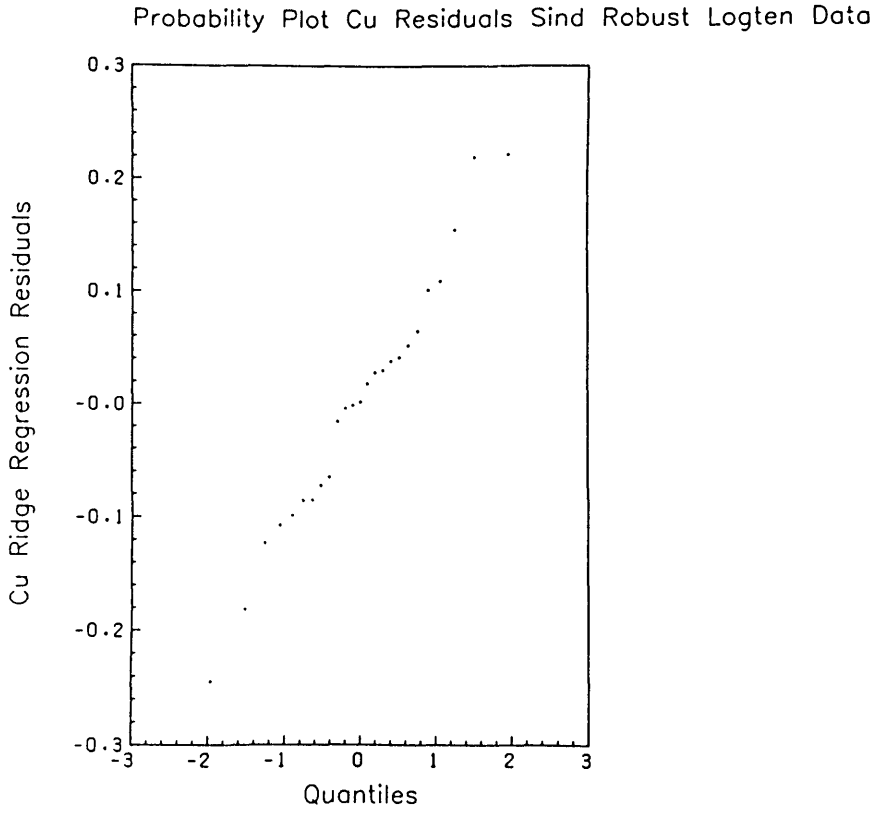
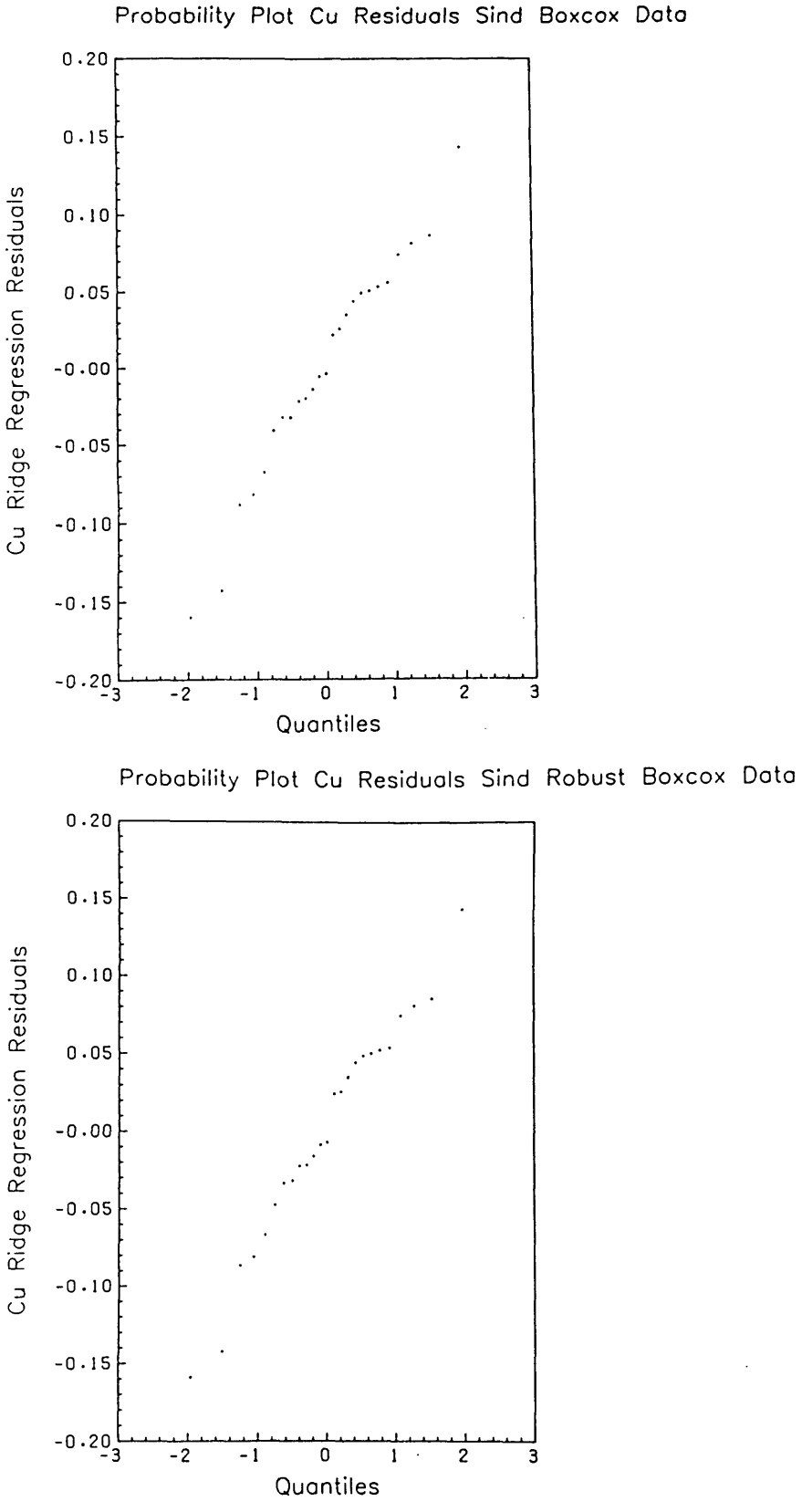

Probability Plot Cu Residuals Sind Robust Logten Data
Cu Ridge Regression Residuals
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-3
-2
-1
0
1
2
3
Quantiles

Figure 9.3


Probability Plot Cu Residuals Sind Boxcox Data
Cu Ridge Regression Residuals
0.20
0.15
0.10
0.05
-0.00
-0.05
-0.10
-0.15
-0.20
-3
-2
-1
0
1
2
3
Quantiles
Probability Plot Cu Residuals Sind Robust Boxcox Data
Cu Ridge Regression Residuals
0.20
0.15
0.10
0.05
-0.00
-0.05
-0.10
-0.15
-0.20
-3
-2
-1
0
1
2
3
Quantiles

revealed in the residual probability plot because of their suppression and amalgamation with background - they just are not detected at all.

9.7 Robust Ridge Regression and the Moray-Buchan Data

The aim of this section is to see whether the conclusions reached in the comparative study of the small SIND data set can be applied to the larger and more complex MB data set. To this end, non-robust and robust ridge regression analyses were carried out on raw and logten transformed data each with three separate response variables - Ba, Zn and Pb - to give twelve sets of results. To save computer processing time, especially in the case of robust regression analysis, preliminary runs were carried out to ascertain which variables were the best at explaining the relevant response variable. The combinations are as follows:

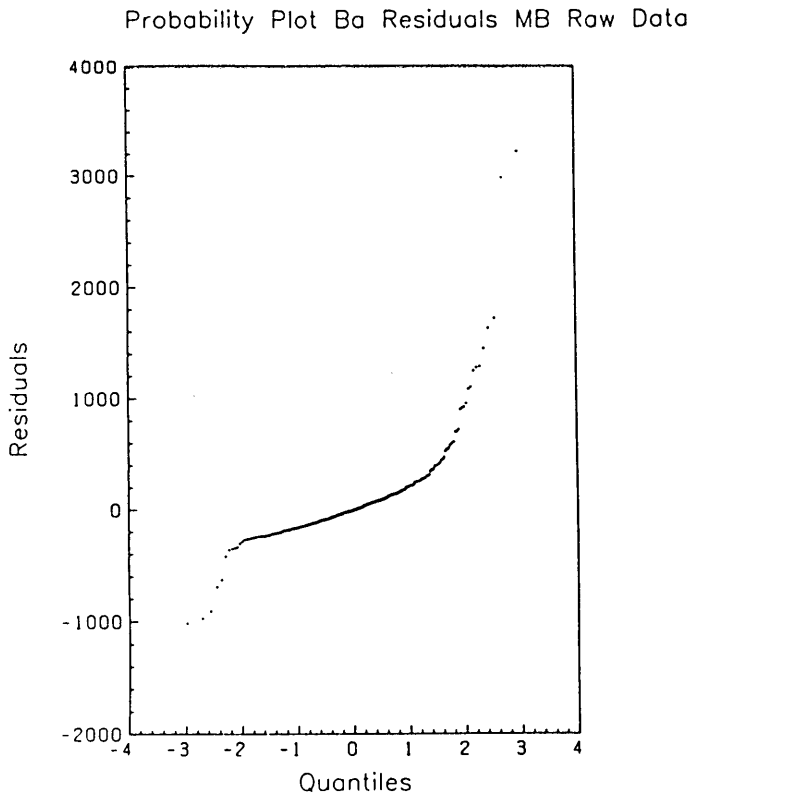
Raw Data	Ridge Regression	Ba	K Rb
		Zn	K Li Mn Pb Zr
		Pb	Li U Zn
	Robust Ridge Regression	Ba	K Rb Y
		Zn	Co Li Mn Pb

		Pb	Ba Zn
Logten Data	Ridge Regression	Ba	Al Fe Mn Rb
		Zn	Li Mn Pb
		Pb	Mg U V Zn
	Robust Ridge Regression	Ba	K Rb U Y
		Zn	Co Cu Li Mn Pb
		Pb	Ba Mg U Zn

As expected, the situation is nowhere near as clear as with the SIND data. Comparison of the probability plots of non-robust with the robust residuals from the raw data (see Figures 9.4 through 9.6) show that, as with the SIND data, the two techniques come to very much the same answer. However, in each robust analysis, factors such as R^2, σ^2, the correlation matrix and the ordering of the anomalous residuals change very little in comparison with the equivalent non-robust analysis. This suggests that the robust analysis obtains the correct estimation of the underlying relationship the first time whereas with the non-robust analysis it is only by manual trimming of anomalous residuals that the correct solution is attained.

The use of logten transformed data seems to suppress the anomaly contrast of positive residuals (see Figures 9.7 through 9.9), which are the ones we are interested in. Logten transformation also tends to enhance any negative residuals. Because of these tendencies, data

Figure 9.4


Probability Plot Ba Residuals MB Raw Data
Residuals
4000
3000
2000
1000
0
-1000
-2000
-4
-3
-2
-1
0
1
2
3
4
Quantiles

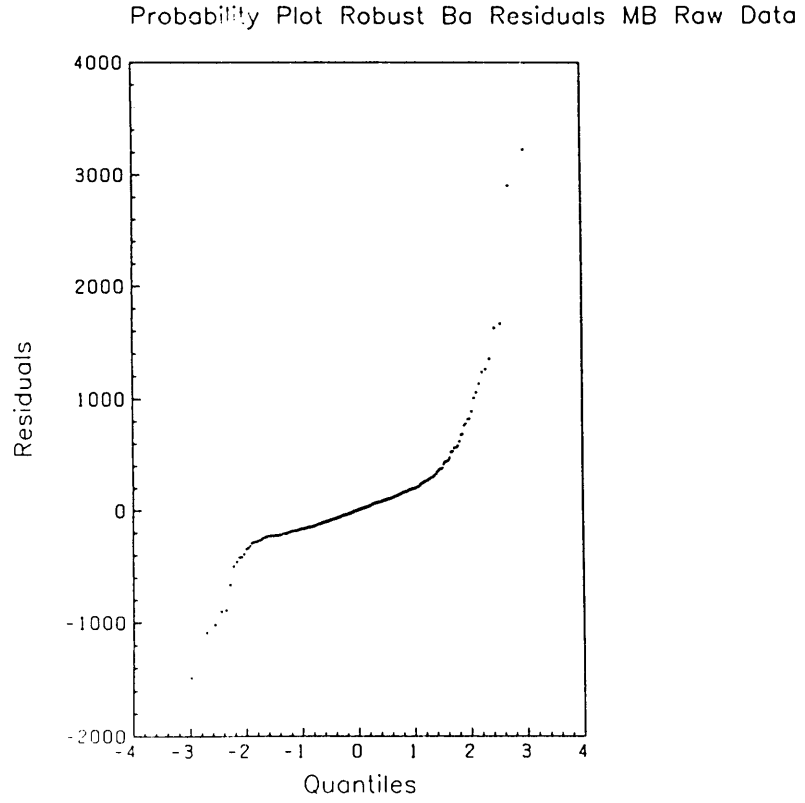
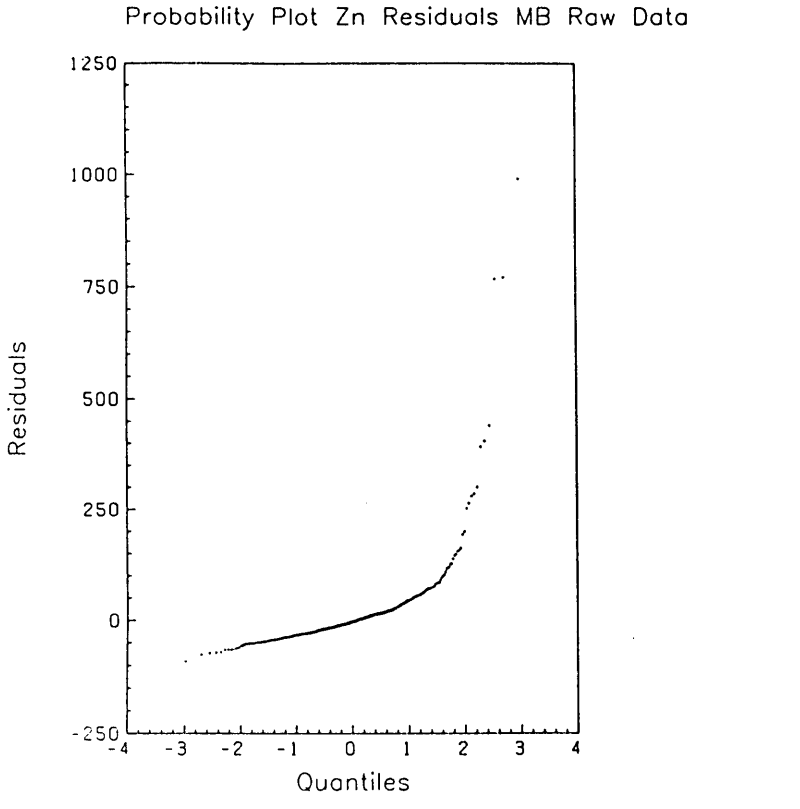

Probability Plot Robust Ba Residuals MB Raw Data
Residuals
4000
3000
2000
1000
0
-1000
-2000
-4
-3
-2
-1
0
1
2
3
4
Quantiles

Figure 9.5


Probability Plot Zn Residuals MB Raw Data
Residuals
1250
1000
750
500
250
0
-250
-4
-3
-2
-1
0
1
2
3
4
Quantiles

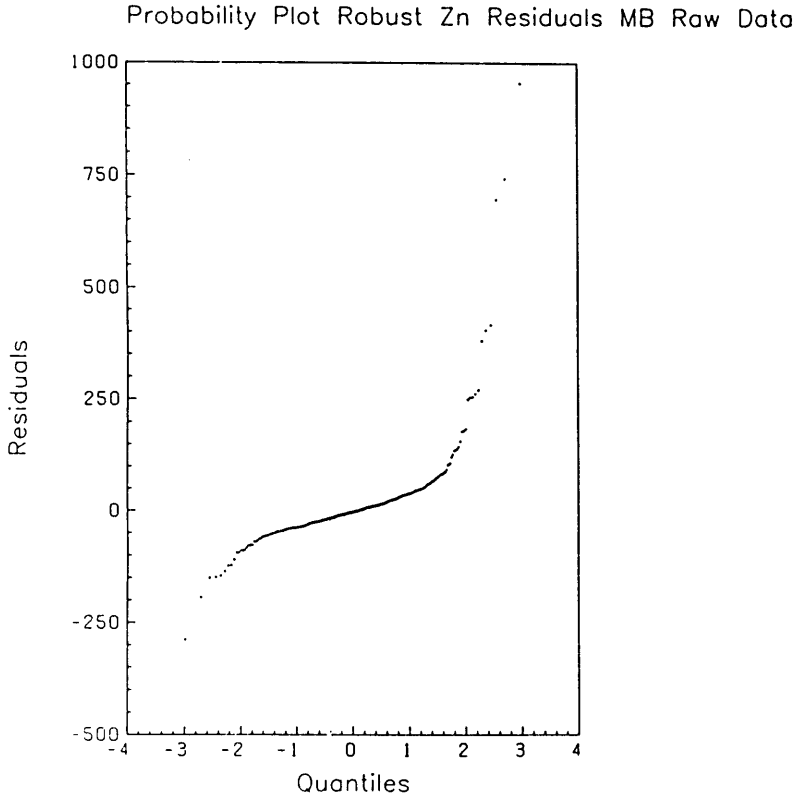
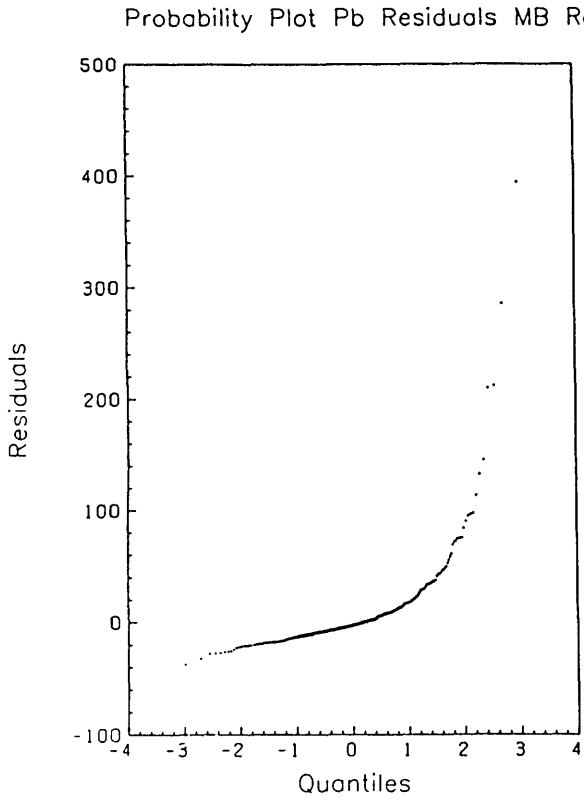

Probability Plot Robust Zn Residuals MB Raw Data
Residuals
1000
750
500
250
0
-250
-500
-4
-3
-2
-1
0
1
2
3
4
Quantiles

Figure 9.6


Probability Plot Pb Residuals MB Raw Data
Residuals
500
400
300
200
100
0
-100
-4
-3
-2
-1
0
1
2
3
4
Quantiles

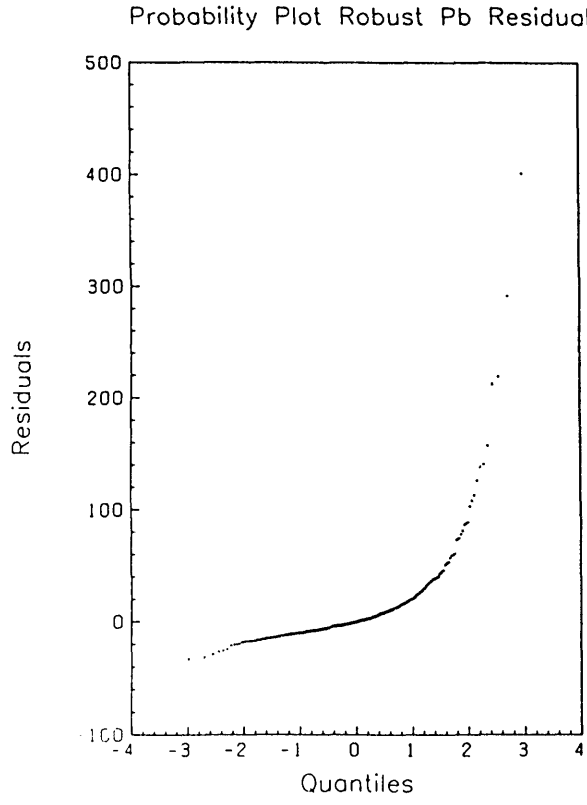
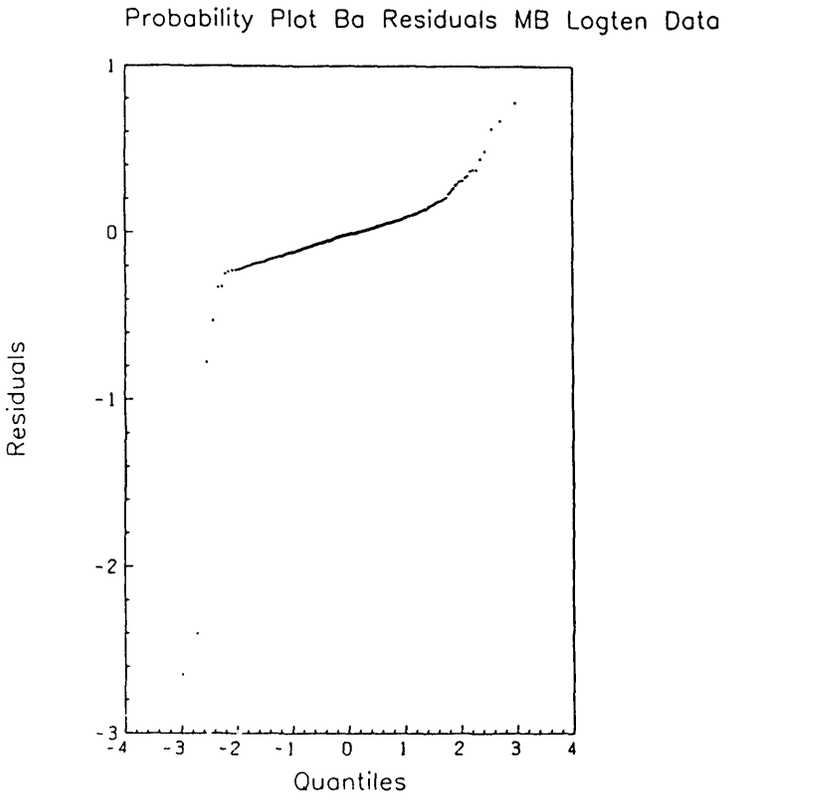

Probability Plot Robust Pb Residuals MB Raw Data
Residuals
500
400
300
200
100
0
-100
-4
-3
-2
-1
0
1
2
3
4
Quantiles

Figure 9.7


Probability Plot Ba Residuals MB Logten Data
Residuals
1
0
-1
-2
-3
-4 -3 -2 -1 0 1 2 3 4
Quantiles

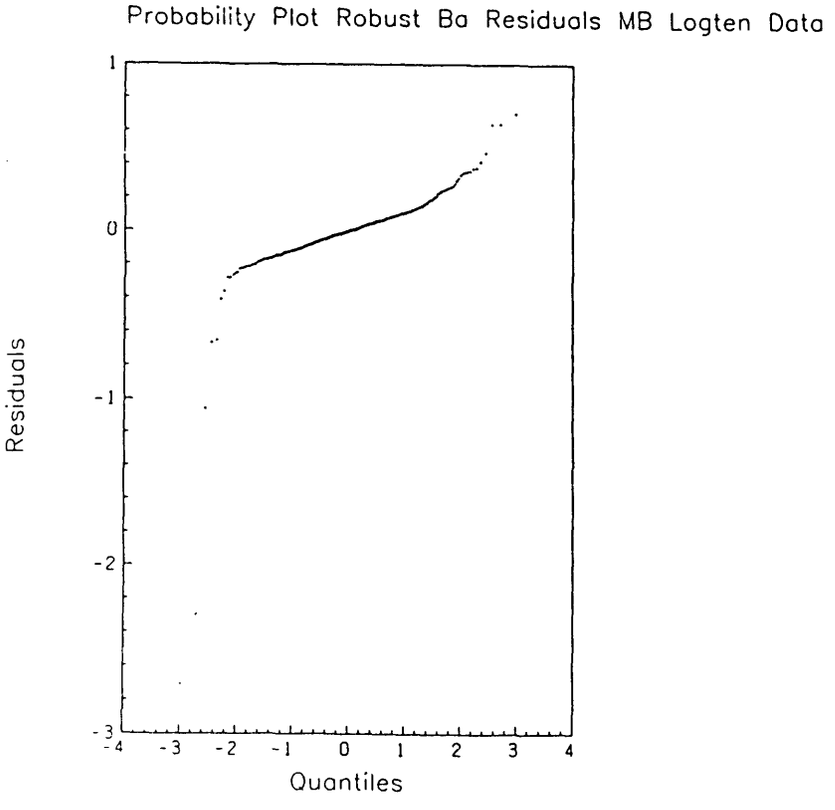
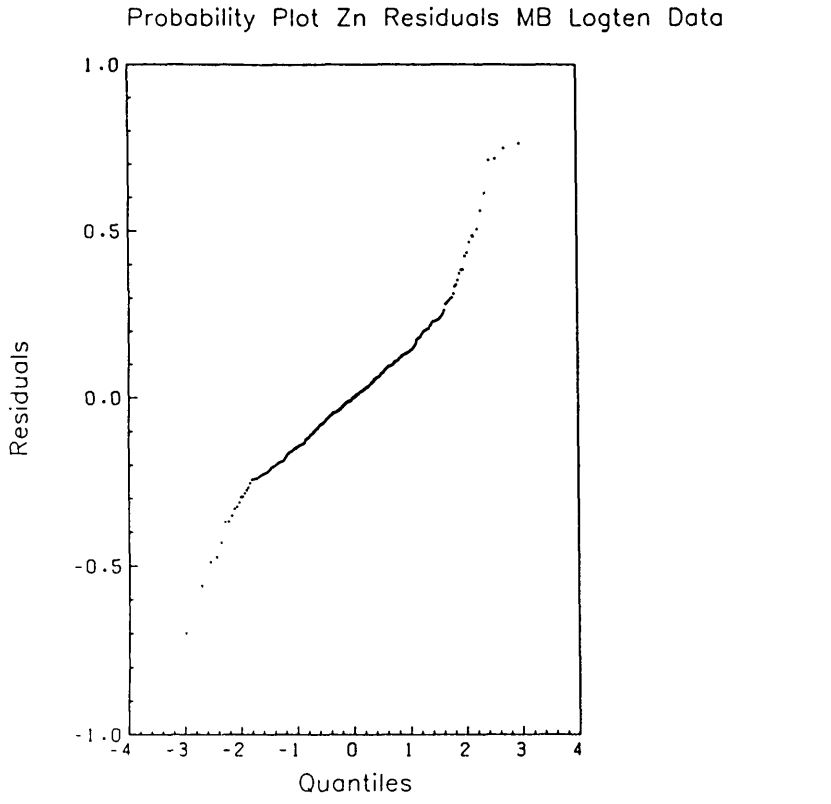

Probability Plot Robust Ba Residuals MB Logten Data
Residuals
1
0
-1
-2
-3
-4 -3 -2 -1 0 1 2 3 4
Quantiles

Figure 9.8


Probability Plot Zn Residuals MB Logten Data
Residuals
1.0
0.5
0.0
-0.5
-1.0
-4
-3
-2
-1
0
1
2
3
4
Quantiles

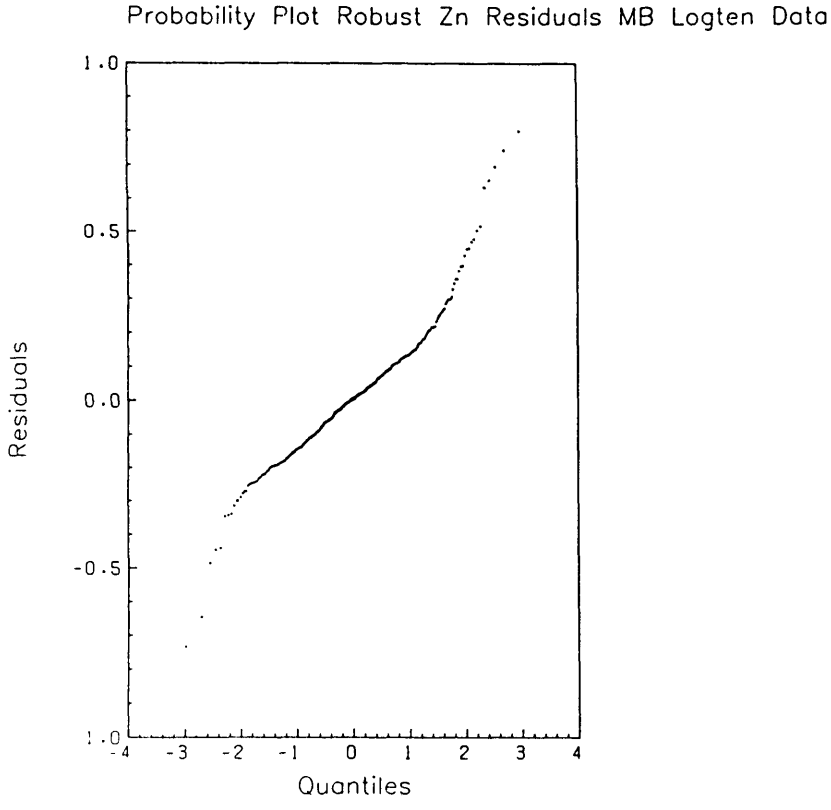
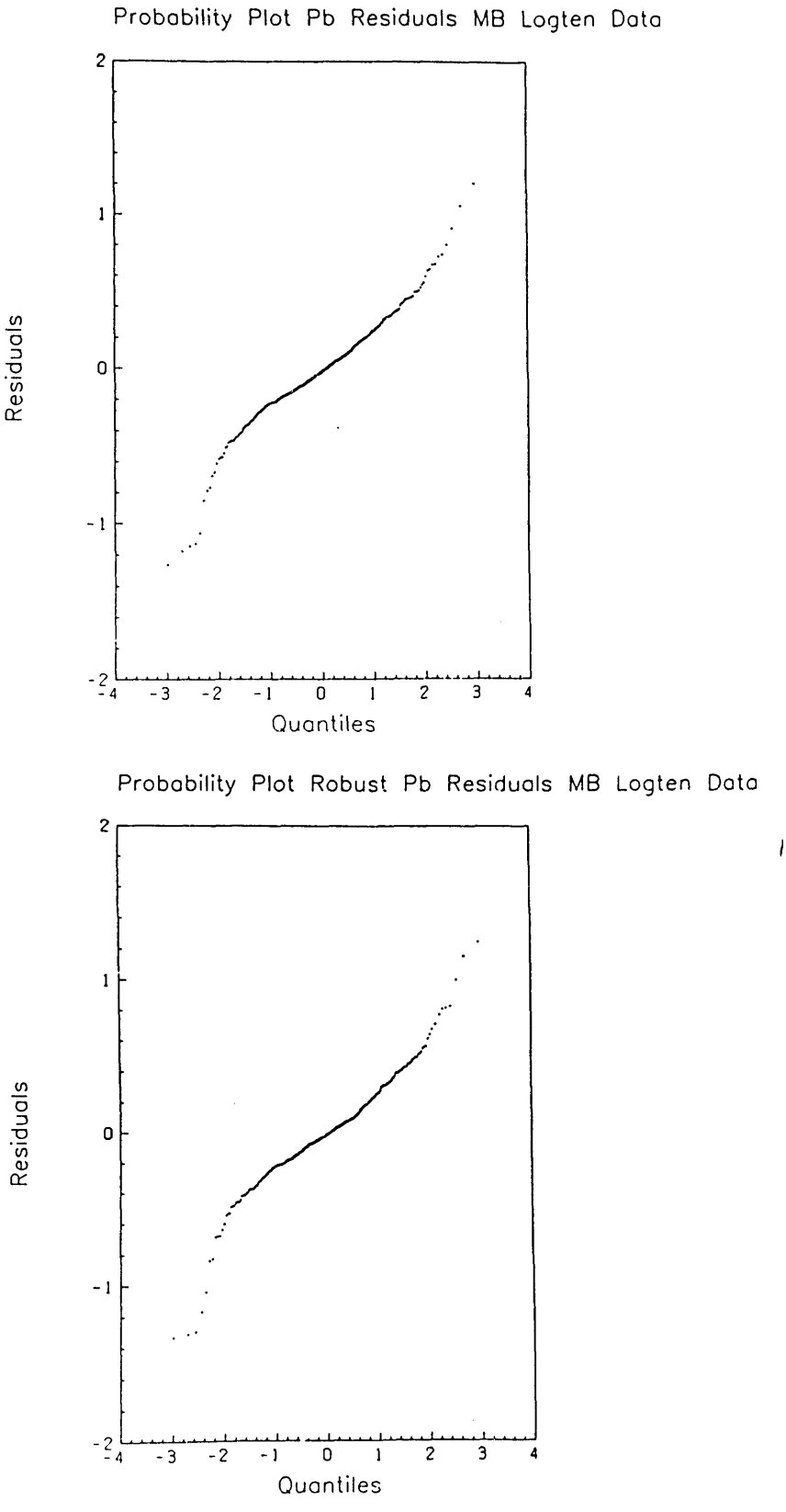

Probability Plot Robust Zn Residuals MB Logten Data
Residuals
1.0
0.5
0.0
-0.5
-1.0
-4
-3
-2
-1
0
1
2
3
4
Quantiles

Figure 9.9


Probability Plot Pb Residuals MB Logten Data
Residuals
Quantiles
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
Probability Plot Robust Pb Residuals MB Logten Data
Residuals
Quantiles
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4

transformations are not recommended, although it must be stated that the point symbol maps of logten residuals reveal very much the same information as do those of raw residuals.

The point symbol maps of Ba, Zn and Pb residuals for non-robust/robust raw/logten data (Figures 9.10 through 9.21) reveal very similar information. This is summarized as follows:

1) There is a dramatic concentration of large Ba residuals over the Lower Dalradian Lochaber subgroup quartzites near Rothes. There are also a few coincident large values associated with south draining streams off Lower Dalradian Ballachulish subgroup mixed pelites 3km to 5km north of Keith. Ba, K and Rb show large elemental concentrations over these two areas with the emphasis for K and Rb being on the Ballachulish subgroup, and the emphasis for Ba being on the Lochaber quartzites. Thus K and Rb explain most of the large Ba concentrations over the Ballachulish area, while over the Lochaber quartzites there is a distinct dichotomy between observed and predicted Ba.

2) The main concentration of large Zn residuals is associated with streams draining undifferentiated Lower/Middle Dalradian pelites and quartzites approximately 10km E and SE of Tomintoul. Additional

Figure 9.10

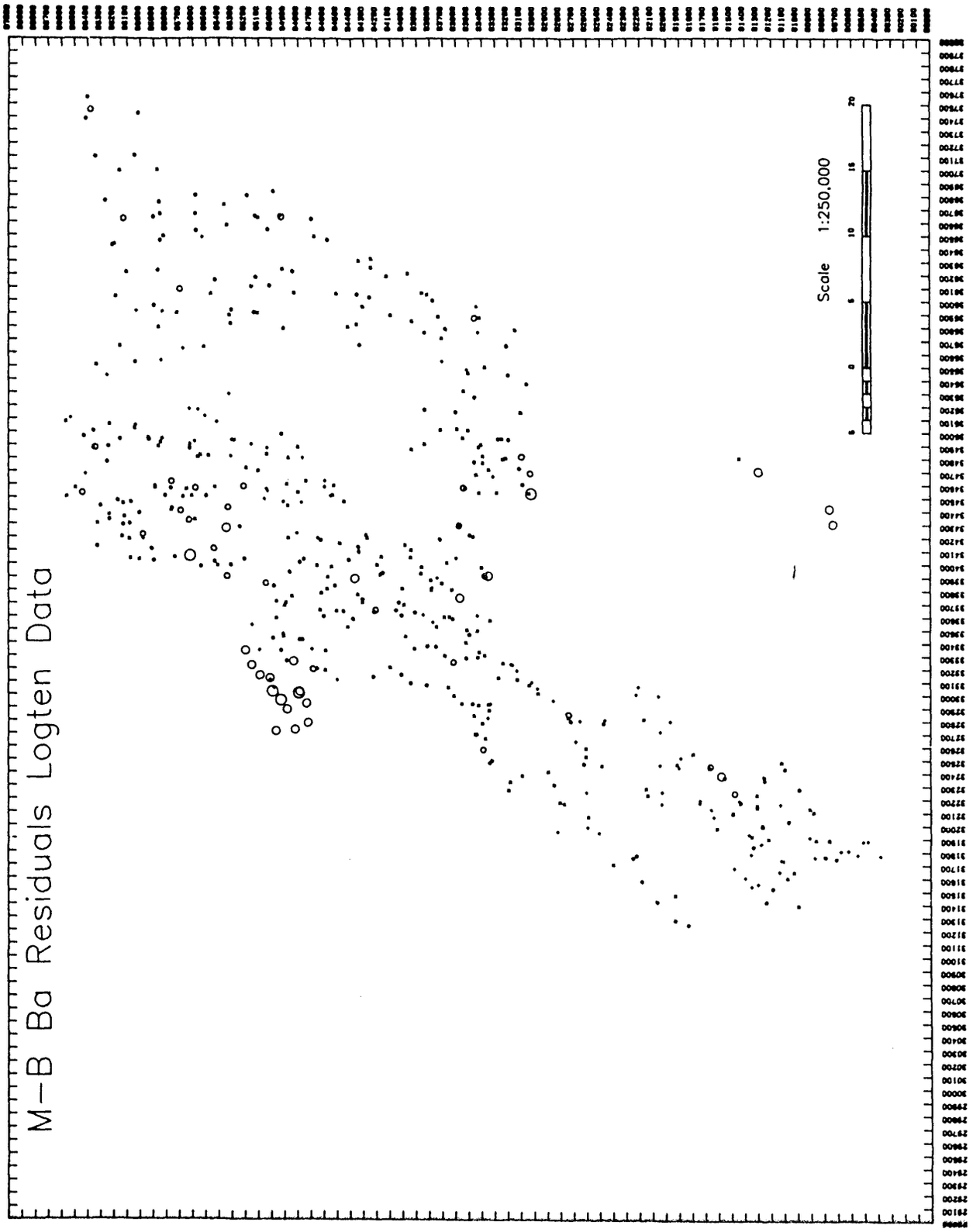


Figure 9.11

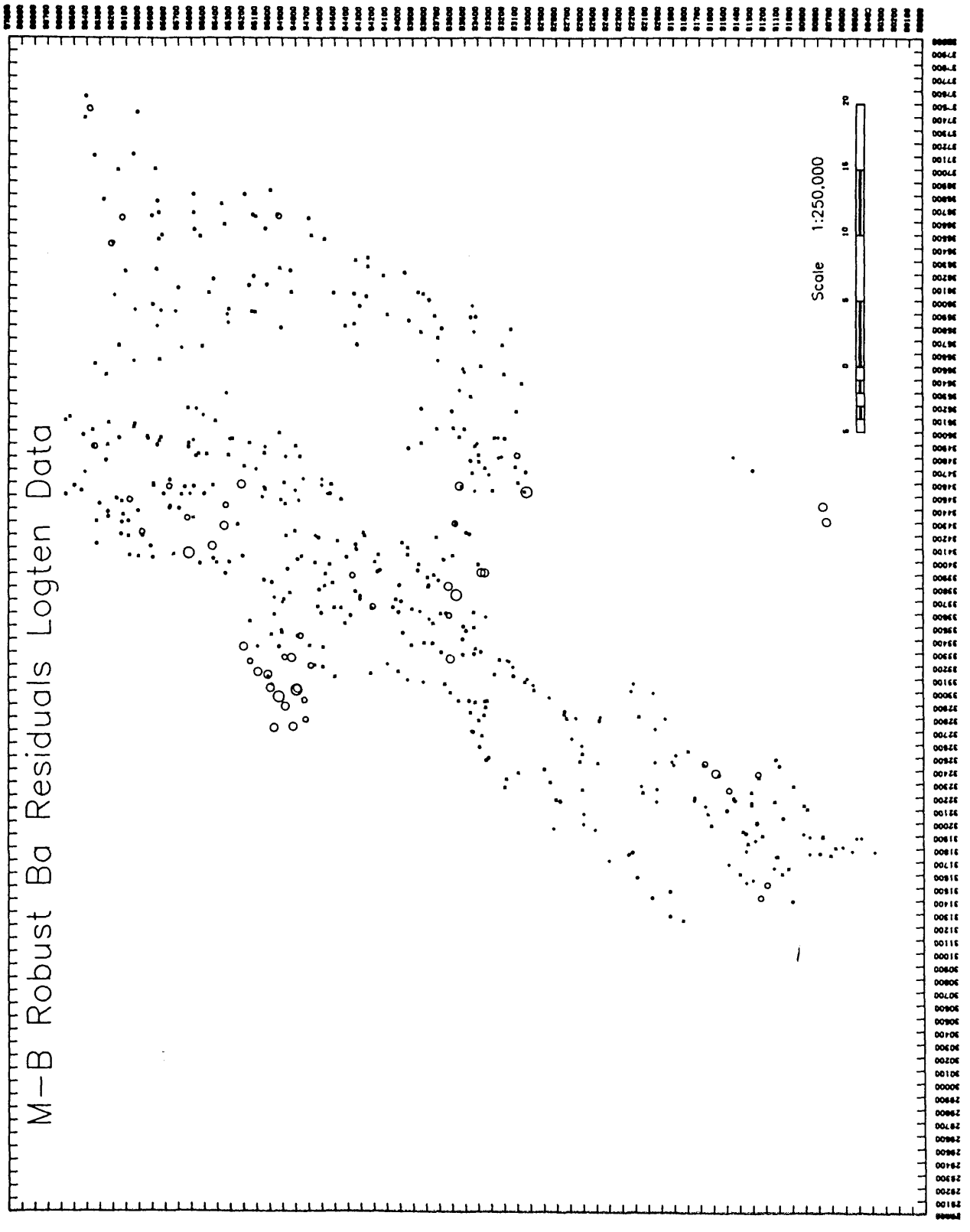


Figure 9.12

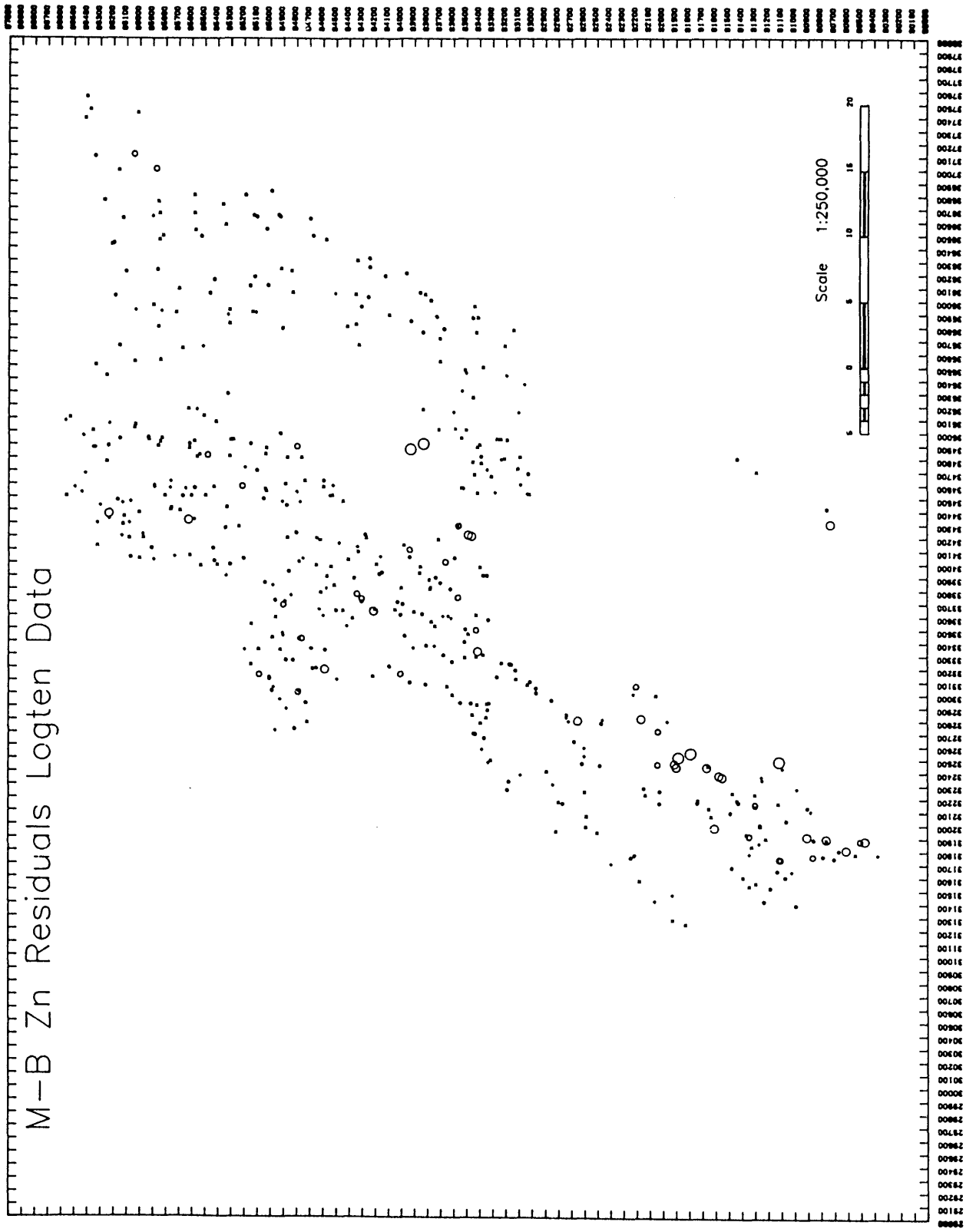


Figure 9.13

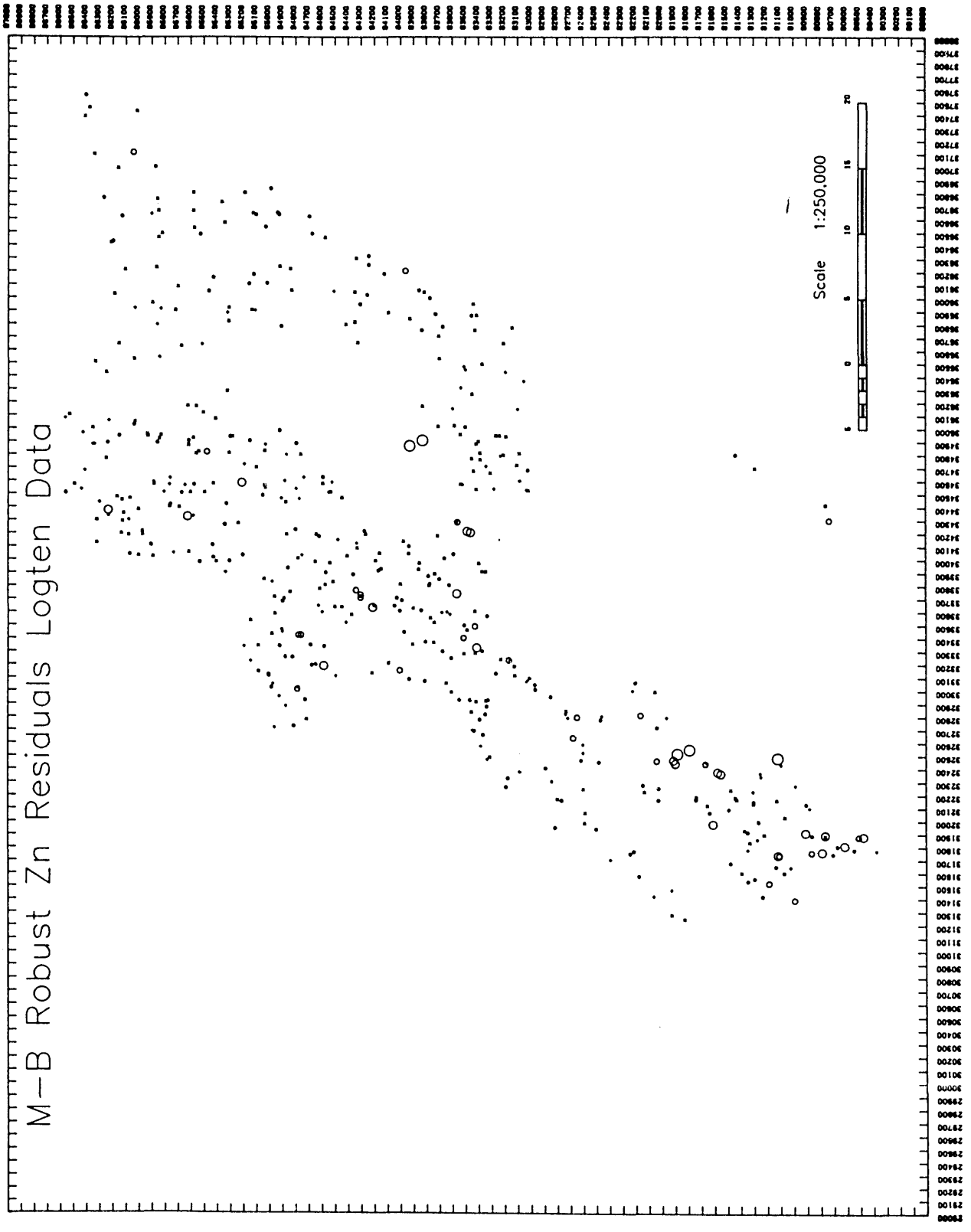


Figure 9.14

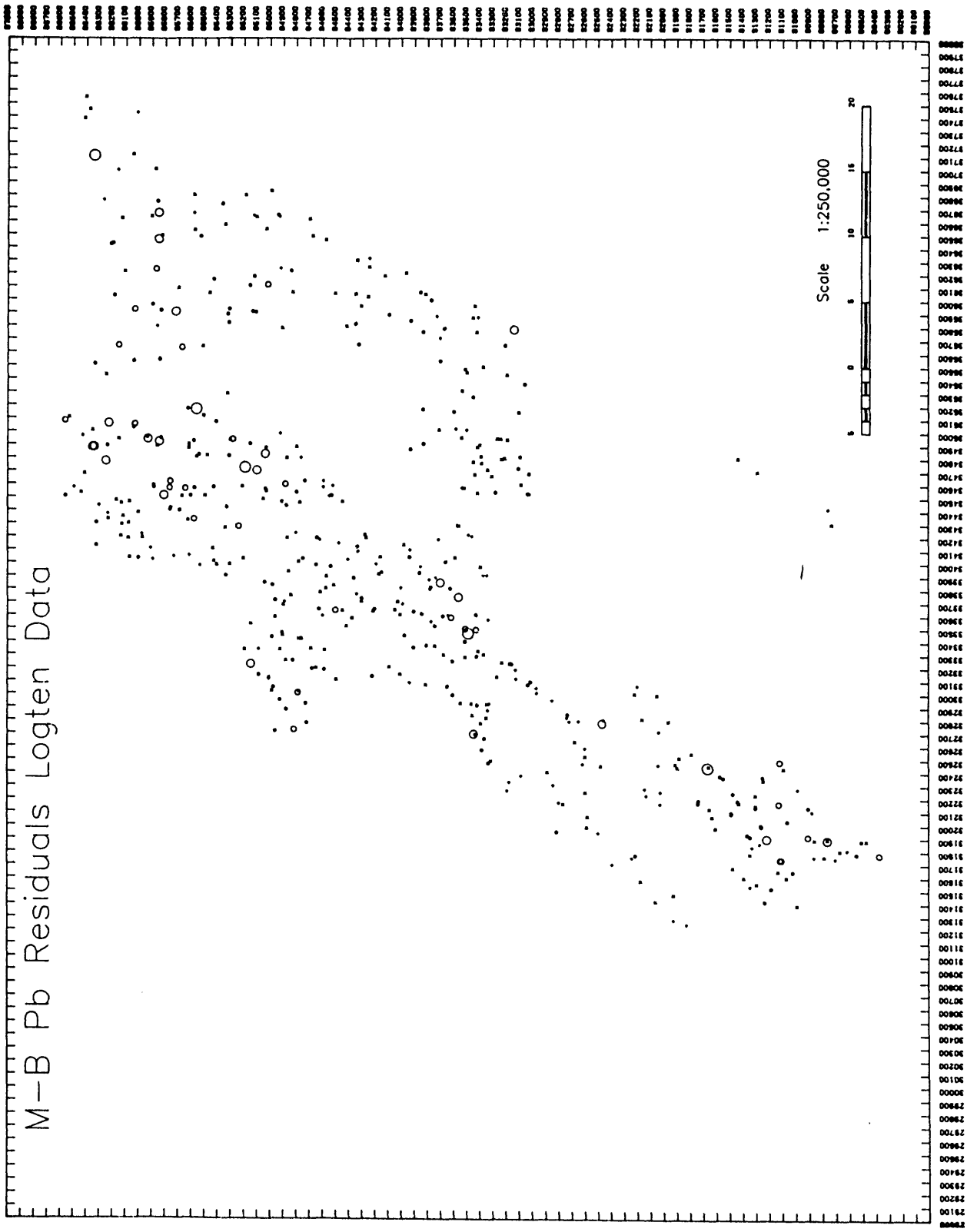


Figure 9.15

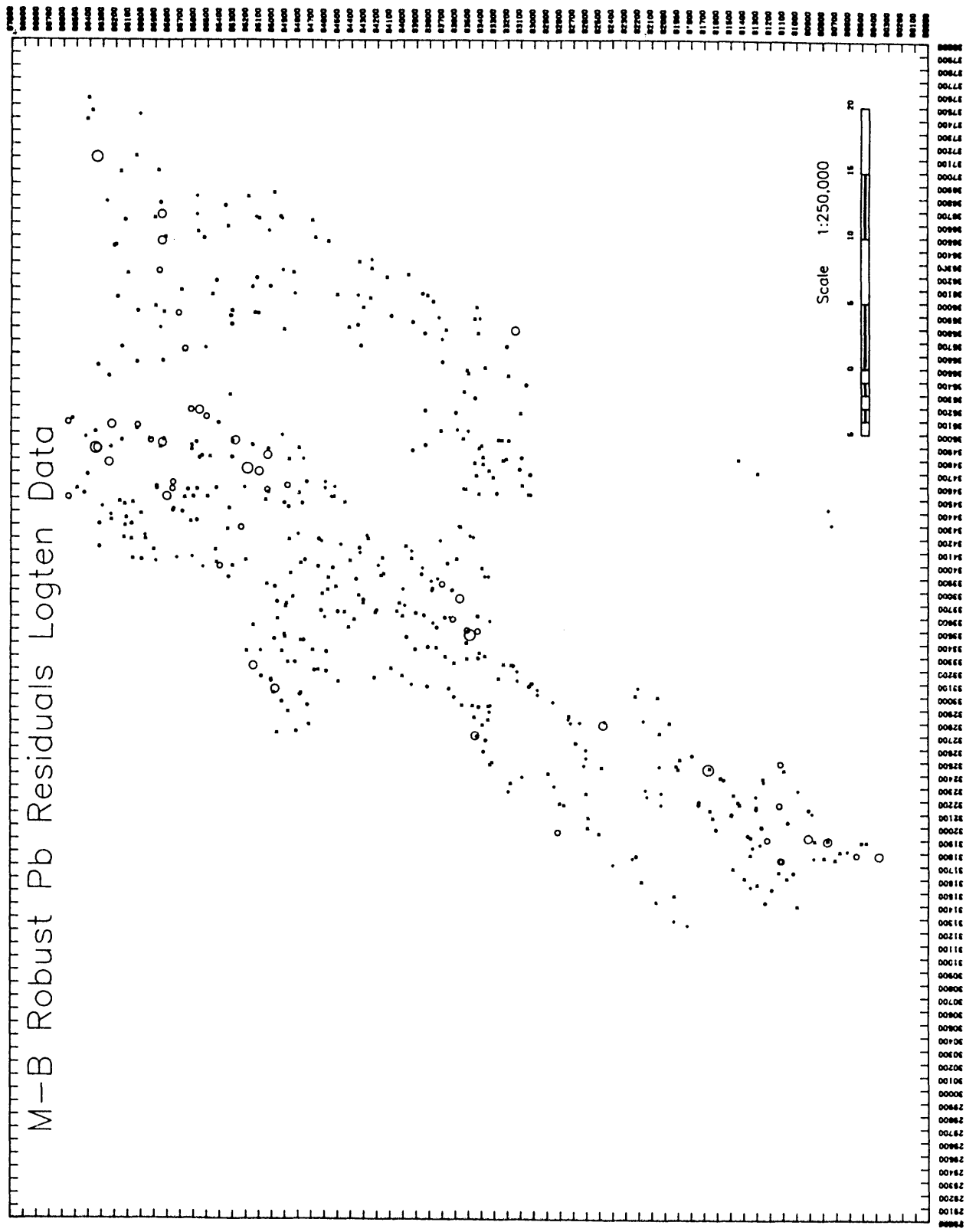


Figure 9.16

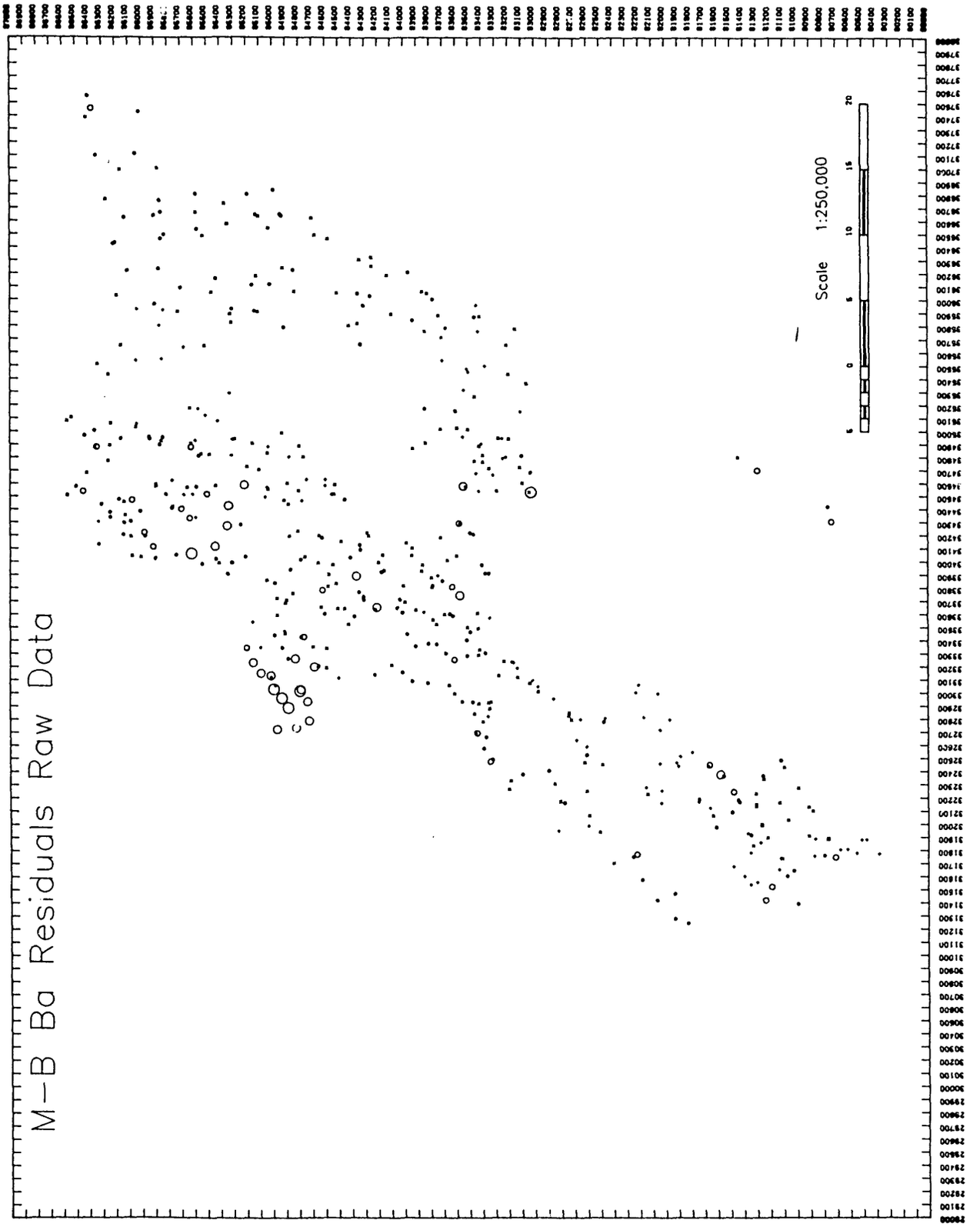


Figure 9.17

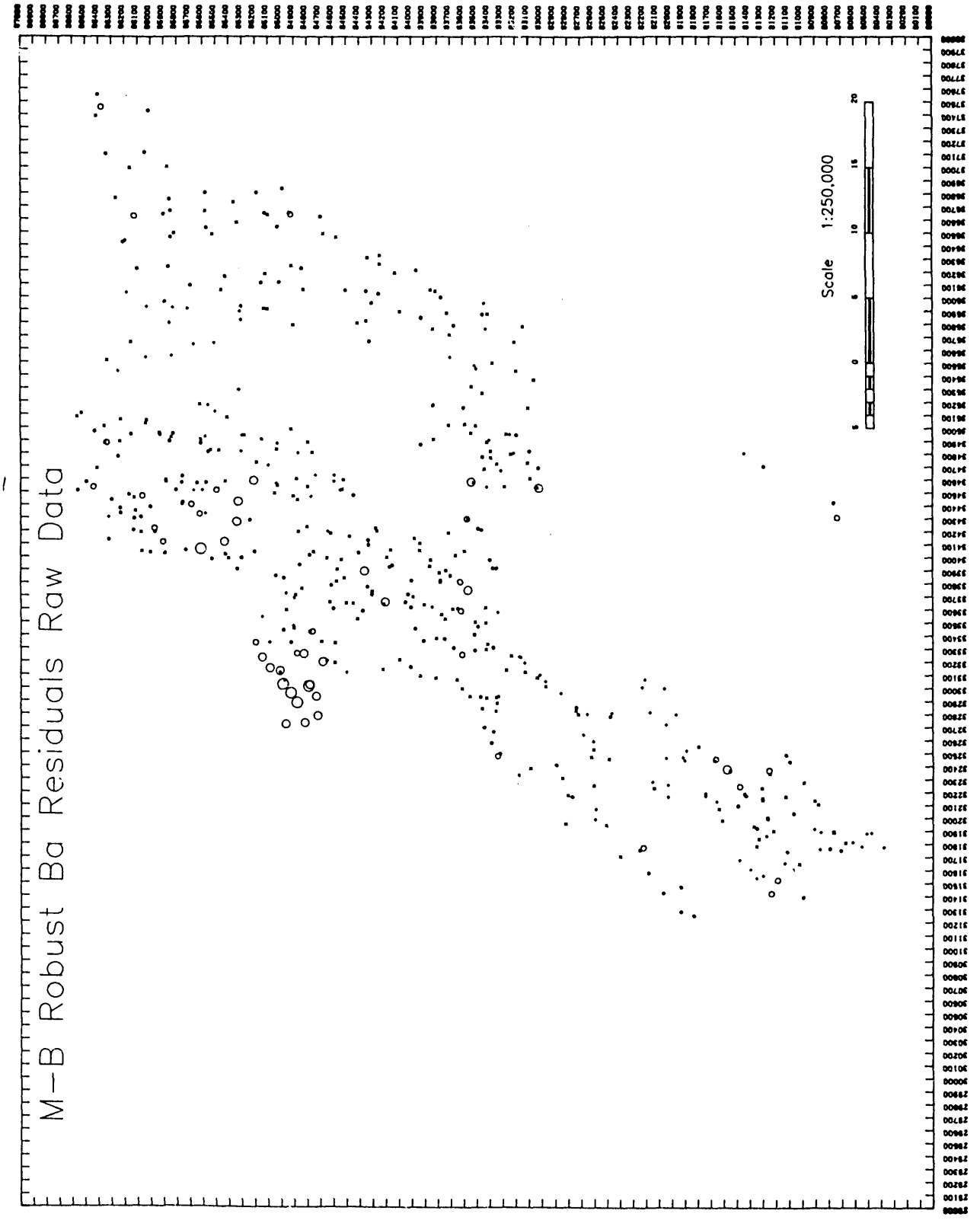


Figure 9.18

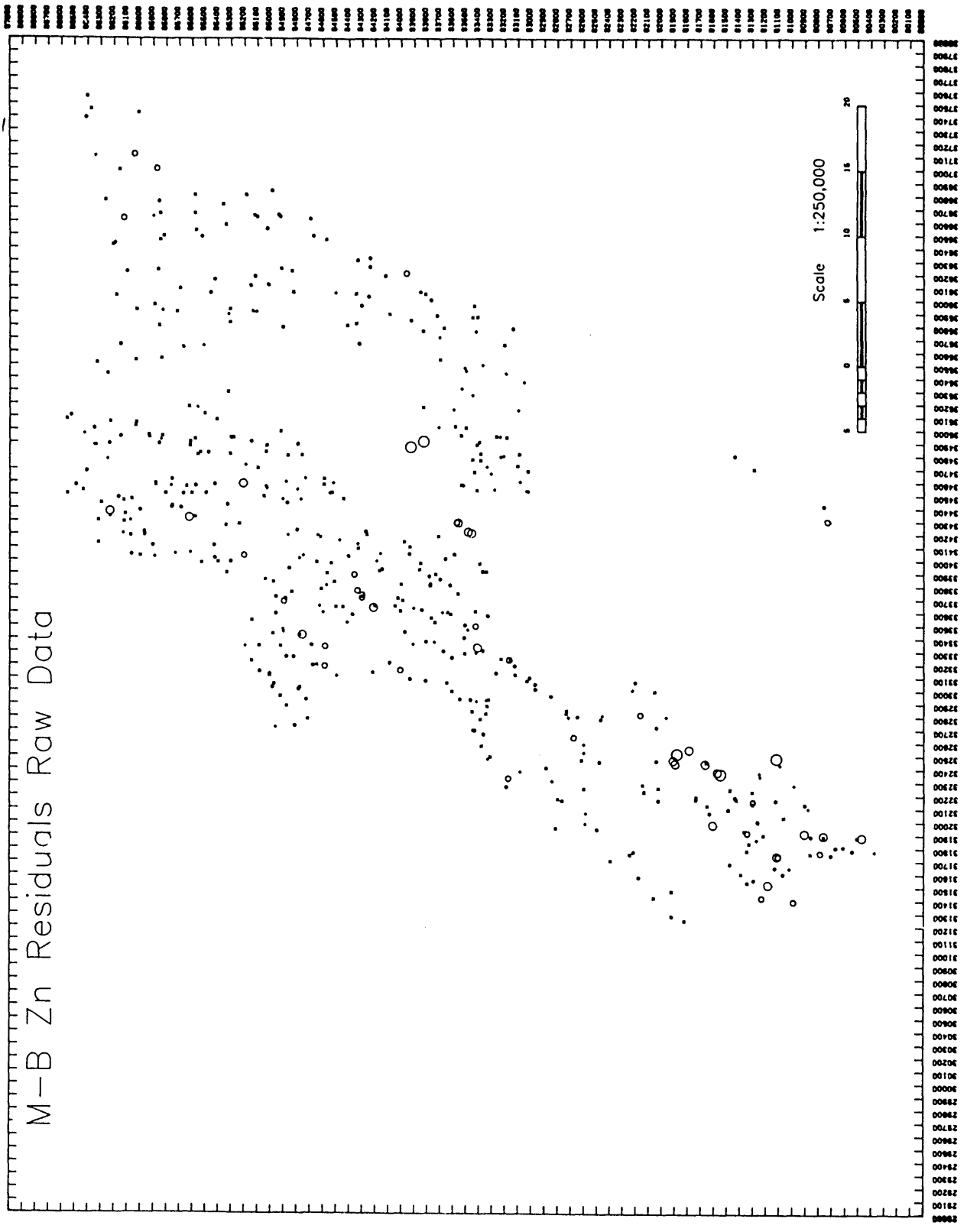


Figure 9.19

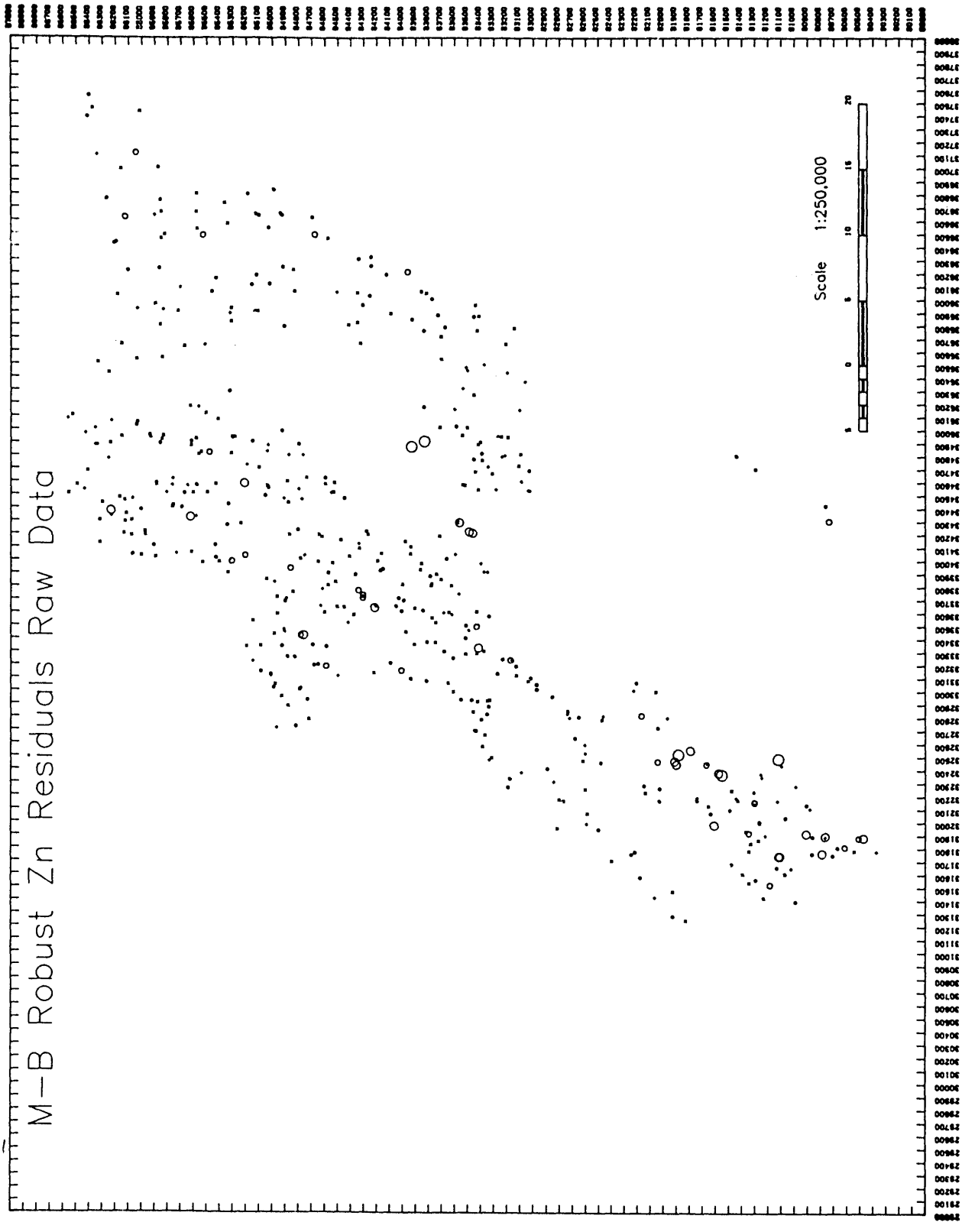


Figure 9.20

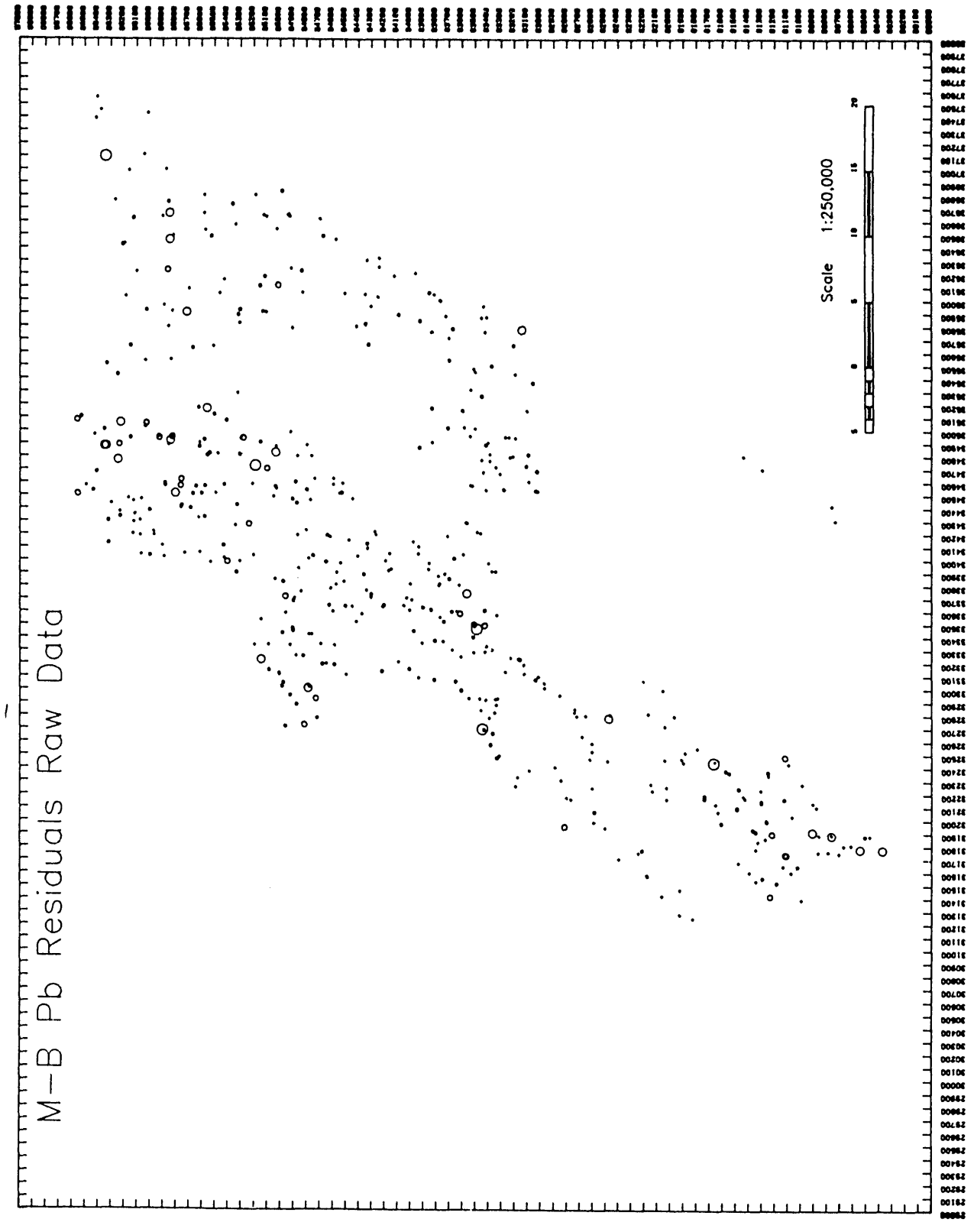
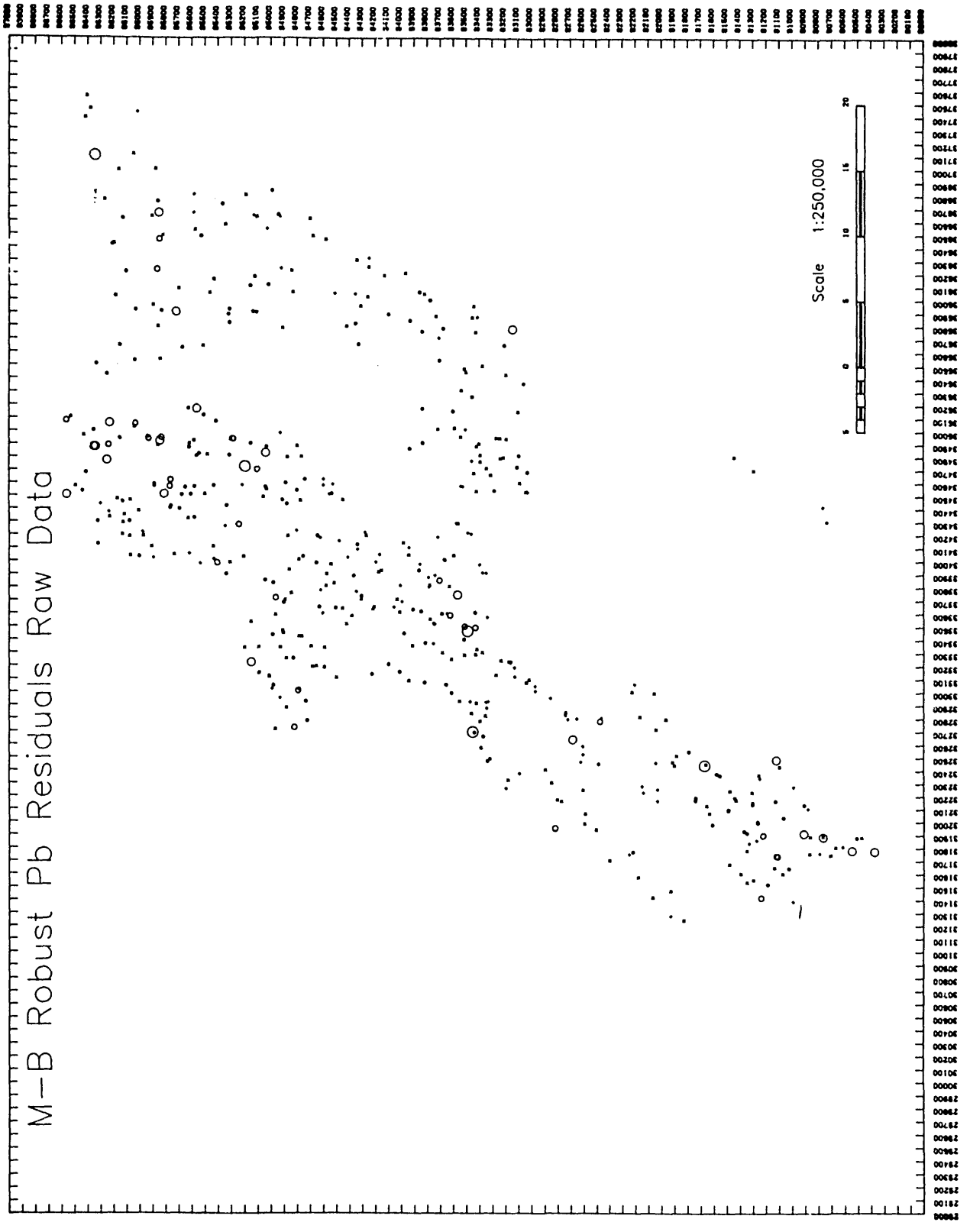


Figure 9.21



large values occur about 3km W of Huntly over Middle Dalradian Tayvallich/Crinan mixed pelites and about 10km ESE of Dufftown over undifferentiated Lower/Middle Dalradian blackschists.

3) Pb residuals show a wide scatter of large values. These are mainly associated with: streams draining the eastern outcrop of Lower Dalradian Ballachulish subgroup mixed pelites; Middle Dalradian Islay subgroup quartzites and Easdale subgroup pelites in the north of the study area; Tayvallich/Crinan mixed pelites NW of Aberchirder; and in Upper Dalradian greywackes NE of Aberchirder and S of Banff.

9.8 Conclusions

The ridge regression techniques that we have seen produce such clear cut results with the small SIND data set, have given, with the much larger MB data set, results that are complimentary but considerably less clear. Robust regression has not only major theoretical advantages but in practice has been seen to be more efficient, as accurate and more consistent than non-robust regression with manual trimming.

Data transformations are not recommended because of the effect they have on regression residuals. It is not the use of the transformed correlation matrix to generate the regression fit that is not recommended, it is the use of the transformed data to generate the residuals that can result in distortions. In fact, as we have concluded for the SIND data, a robust correlation matrix - or a transformed, (preferably boxcox transformed), matrix - is ideal for determining the regression fit. The residuals, once the fit has been "subtracted" from the data, are governed by the form the data takes. If the data are in raw, untransformed, state then the residuals will display maximum anomaly contrast. If the data are transformed, then anomalous residuals tend to be suppressed and anomaly contrast is substantially reduced.

CHAPTER TEN

SUMMARY, FURTHER WORK and CONCLUSIONS

10.1 Summary

A comparative study of several univariate and multivariate statistical techniques, new to the analysis of geochemical data, was carried out using the SIND test data set (25 observations, 6 elements). The techniques were then applied to a large data set (503 observations, 24 elements) drawn from the BGS geochemical atlas database for the Moray-Buchan area of NE Scotland, using an approach based on the conclusions from the SIND data analysis. The results were compared with a traditional approach.

Robust statistics form the main theme which runs through the whole thesis. Robust statistics are equivalent to parametric statistics, but are designed to perform in a manner which is resistant to the distorting presence of outliers. Maximum likelihood or M-estimates of standard univariate and multivariate parametric statistics perform well on geochemical data which display moderate to extreme positive skew.

The traditional logarithmic transformation procedure, used to deskew or normalize frequency distributions, is inferior to the more flexible power transformation of Box and Cox (1964). Parametric statistics generated using transformed data are resistant to outliers, and are comparable with robust statistics generated using untransformed data. Due to the normalizing process of the power transformation, robust statistics applied to transformed data prove to give little extra resistance. The process of deskewing or normalizing the frequency distribution of a set of data involves the suppression of the outliers and the spread (or enhancement) of background observations. The contrast between outliers and background is thus reduced.

The graphical technique of probability plots is briefly described as a useful aid for rapidly viewing the distributional characteristics of a set of univariate or multivariate data.

Most multivariate statistical techniques calculate, from the input data set, a summary of the structure of those data which is then used as the basis for the technique. This summary usually takes the form of the correlation matrix. Like any parametric statistic, the correlation matrix is subject to distortions in the presence of outliers. Robust estimation or, preferably, data transformation followed by robust estimation of the correlation matrix, solves this problem and provides a way

to "robustify" standard multivariate statistical techniques.

Residuals may be defined as those which remain after a summary or fitted model has been subtracted out of the data according to the schematic equation:

$$\text{residual} = \text{data} - \text{fit}$$

Study of residuals is a major facet in exploratory data analysis (Tukey, 1977). This analysis should take advantage of the tendency of robust techniques to provide a clear separation between dominant background behaviour and anomalous behaviour in the data. A fitted model is determined by the dominant structure of the data. If this structure is distorted by the presence of outliers, then the fitted model, and consequently the residuals, will also be distorted. Robust estimation results in a fitted model that defines the true background structure of the data. Residuals from this robust fit will, in turn, be robust. However, the residual is determined not only by the accurate estimation of the fit, but by the correct use of the data. The use of transformed data yields residuals in which the contrast between true anomalies and background (termed anomaly contrast) is low. Anomaly contrast is maximized with the use of raw data.

Trees are a recently developed graphical technique for representing multivariate data based on a cluster analysis (Kleiner and Hartigan, 1981). The technique may be performed using a transformed robust correlation matrix to generate the cluster dendrogram, with the individual trees developed using the raw data to give maximized anomaly contrast. Trees provide a useful exploratory means of viewing the data, and of making a preliminary determination of those samples which are likely to be of further interest. The cluster dendrogram, on which the topology of the trees is based, provides a classification of the variables into groups which are close together in multidimensional space. The variable groups obtained using cluster analysis compare closely with those from PCA.

Factor Analysis proved superior to PCA and was used to determine and quantify variable associations present in the data. FA performs best using a transformed robust correlation matrix, to generate the factor loadings, and the raw data to generate the factor scores. Since a transformed robust correlation matrix is used, the variable associations revealed in the factor loadings matrix will reflect major background relationships. One should not expect that minor associations will be revealed, especially if they are present in only a minority of the data (as is usually the case with mineralization associations). If we regard the factor loadings matrix, generated in the fashion described above, as a robust fit, then we can regard the

factor scores, calculated using the raw data, as robust residuals which exhibit maximized anomaly contrast. FA is a powerful technique which is too often thought of as a universal panacea in geochemical data processing, and consequently is often applied inappropriately and out of context. If the technique is fully understood and properly applied, the results can be exceedingly useful as part of a unified data analysis methodology.

Robust regression analysis necessitates the robust estimation of the correlation matrix of the input data internally to the program, as opposed to input of a robust matrix and data separately. This means that observations are weighted, according to their Mahalanobis distance, prior to regression analysis, as opposed to the more common practice of weighting according to the residual from the regression hyperplane. The latter procedure was found to be theoretically and practically unsound. The use of raw data, instead of transformed data, yields robust residuals which display maximized anomaly contrast. Robust regression proved to be more efficient, more consistent and as accurate as non-robust regression with manual trimming of high residuals.

Residuals from both FA and regression analysis can be mapped to show their spatial distribution.

10.2 Further Work

Continuation of the work described in this thesis could be along two avenues: statistical programming; and data analysis methodology.

On the programming side, further work could be towards the formation of a single piece of software, which would be interactive and flexible, incorporating all the techniques described in the thesis.

On the methodology side, further work could involve the incorporation of geomorphometric parameters into the regression model to improve the proportion of the total sum of squares of the response variable explained by the regression (i.e. R^2). Such parameters could be: area of drainage basin upstream from sample site; the proportion of that area cut by each of the rock types present in the study area; drainage basin relief; drainage density.

10.3 Conclusions

1) Robust statistics must be accepted into routine geochemical data analysis. They have been proved to perform better than classical statistics under conditions which are prevalent in geochemical data.

2) If anomaly detection is the aim of the analysis then the raw data must be used to generate residuals from the fitted model.

3) If determination of background relationships and associations is the aim of the analysis, perhaps with a view to geochemical mapping, then boxcox transformed data should be used to generate residuals.

4) The techniques described in this thesis should not be regarded as separate methods treated in isolation, but should be used together to gain a clearer and more accurate view of the behaviour of the database under review. For example, FA and cluster analysis are complimentary techniques, and the fact that they both show similar patterns is strong evidence that those patterns are real. Achieving such corroboration is a good reason for performing both methods of analysis.

5) Raw data distributions tend to be positively skewed to varying degrees.

6) Logarithmic transformation is too rigid a normalization method.

7) Boxcox power transformation optimizes the normality of the data distribution and is a superior technique to log transformation.

APPENDICES

A.1 Back-transformation of statistical parameters obtained from logarithmic or power transformed data.

Suppose our observations are $x_1, \ldots, x_n$ and that taking natural logarithms gives

$$z_1 = \ln x_1, \ldots, z_n = \ln x_n$$

where the z's appear to be a representative sample from a normal distribution. This normal distribution has unknown mean μ and variance σ^2 estimated respectively by

$$\bar{z} = 1/n \; \Sigma \, z_i$$

$$s^2 = 1/(n-1) \; [(\Sigma \, z_i^2 - n\bar{z}^2)]$$

The distribution from which the x's come - called the lognormal distribution - has mean and variance estimated by

$$\exp\left(\bar{z} + 0.5s^2\right)$$

and $$(e^{s^2} - 1)e^{s^2} + 2\bar{z}$$

respectively.

If, instead of logarithms, we use a power tranformation $z = (x^{\lambda} - 1)/\lambda$ such that the z's are normally distributed for some value of λ , then the mean and variance of the distribution from which the x's come are estimated respectively by:

$$(1 + \lambda z)^{1/\lambda} + s^2 (1 - \lambda) (1 + \lambda z)^{(1/\lambda)-2}$$

and $$s^2 (1+ \lambda z)^{(2/\lambda) - 2}$$

A.2 Moray-Buchan anomalous samples detected using TREES analysis

Sample No.	Anomalous Features
105040	Zn B Li
105057	K Rb
105063	Zn B Li
105126	Pb
105205	Mn minor Pb Zn + mafics
105210	Ca Mg minor mafics
105219	Co Ni + subsidiary mafics
105229	as 105219
105230	Ba + subsidiary Rb
105236	B Li Al
105243	B Li Al
105253	Sr + subsidiary Ca Mg
105267	B Li
105286	Mafics + Ca Mg Sr
105290	as 105286
105292	Ca Mg Sr + Mn Zn Pb
105298	Mn + Al B Li + mafics (but low Cr Ti V)
105404	Ca Mg
105418	Ca Mg

105424	Ca Mg
105430	Ca Mg
105436	Pb
105467	B Li Al
105453	Pb + Cu Fe + Al B Li
105470	B Li Al
105472	B Li Al
105474	Mafics + Ca Mg
105493	Ca Mg
105508	B Li (very low K Rb)
105528	as 105508
105529	Y La
105584	Y La
105617	K Rb
105641	Mn
105645	K Rb
105685	K Rb
105709	Mafics + Ca Mg
105727	Cr + other mafics + minor Ca Mg
105759	Ti V + minor Ca Mg
105764	as 105759
105799	Pb
105802	Ba
105814	K Rb
105847	B Li
105872	K Rb Ba
105860	Ba

105883	B Li
106498	Ti V Cr + other mafics + Mn
107603	Ca Sr + La + Li
107608	Ca Sr
107610	B
107622	U Y La + Be K Rb
107623	U Y La + very high Be Rb K
107645	Ca Mg
107814	Zn + B Li
107834	massive U Y La Zr + Be K Rb + Li
107858	Zn Mn Pb
107881	Zn + B Li
107889	B Li
107918	Mafics + Ca Mg
107970	Y La
108000	Ti V
108039	Ca Sr

BIBLIOGRAPHY

Ahrens, L.H., 1954. The lognormal distribution of the elements. Geochim. et Cosmochim. Acta., 5: 49-73.

Ahrens, L.H., 1954. The lognormal distribution of the elements. Geochim. et Cosmochim. Acta., 6: 121-131.

Aitchison, J., 1981. A new approach to null correlations of proportions. Math. Geol., 13:

Aitchison, J., 1984. The statistical analysis of geochemical compositions. Math. Geol., 16: 531-564.

Andrews, D.F., 1972. Plots of high dimensional data. Biometrics, 28: 125-136.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W., 1972. Robust estimates of location - survey and advances. Princeton University Press, Princeton, N.J., 373pp.

Atherton, M.P., 1977. The metamorphism of the Dalradian rocks of Scotland. Scott. J. Geol., 13: 331-370.

Beniger, J.R. and Robyn, D.L., 1978. Quantitative graphics in statistics: a brief history. Amer. Statist., 32: 1-11.

Bowen, 1928. The evolution of the igneous rocks.___

Box, G.E.P. and Cox, D.R., 1964. An analysis of transformations. J. Royal Stat. Soc., Series B, 26: 211-243.

Bruckner, L.A., 1978. On Chernoff faces. In: P.C.C.Wang (Editor), Graphical representation of multivariate data. Academic Press, N.Y., pp.93-121.

Campbell, N.A., 1979. Canonical variate analysis: some practical aspects. Unpubl. Ph.D. Thesis. University of London.

Chambers, J.M. and Kleiner, B., 1982. Graphical techniques for multivariate data and for clustering. In: P.R.Krishnaiah and L.N.Kanal (Editors), Handbook of statistics, Vol. 2. North-Holland Publishing Company, pp.209-244.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A., 1983. Graphical methods for data analysis. Duxbury, Boston, 395pp.

Chapman, R.P., 1976a. Some consequences of applying lognormal theory to pseudolognormal distributions. Math. Geol., 8: 209-214.

Chapman, R.P., 1976b. Limitations of correlation and regression analysis in geochemical exploration. Trans. I.M.M., 85: B279-283.

Chayes, F., 1954. The lognormal distribution of the elements: a discussion. Geochim. et Cosmochim. Acta., 6: 119-120.

Chernoff, H., 1973. The use of faces to represent points in K-dimensional space graphically. J.A.S.A., 68: 361-368.

Chernoff, H. and Rizvi, M.H., 1975. Effect on classification error of random permutations of features in representing multivariate data by faces. J.A.S.A., 70: 548-554.

Coats, J.S., Smith, C.G., Fortey, N.J., Gallagher, M.J., May, F. and McCourt, W.J., 1980. Stratabound barium-zinc mineralization in Dalradian schist near Aberfeldy, Scotland. Trans. I.M.M., 89: B110-122.

Coats, J.S., Smith, C.G., Gallagher, M.J., May, F. and McCourt, W.J., 1981. Stratabound barium-zinc mineralization in Dalradian schist near Aberfeldy, Scotland: final report. I.G.S. Mineral Reconnaissance Programme Rept. No. 40, 116pp.

Cressie, N. and Hawkins, D.M., 1980. Robust estimation of the variogram: I. Math. Geol., 12: 115-125.

Davis, J.C., 1973. Statistics and data analysis in geology. Wiley, N.Y., 550pp.

Devlin, S.J., Gnanadesikan, R. and Kettenring, J.R., 1981. Robust estimation of dispersion matrices and principal components. J.A.S.A., 76: 354-362.

Everitt, B.S., 1978. Graphical techniques for multivariate data. Heinemann, London, 117pp.

Fettes, D.J., 1970. The structural and metamorphic state of the Dalradian rocks and their bearing on the age of emplacement of the basic sheet. Scott. J. Geol., 6: 108-118.

Fienberg, S.E., 1979. Graphical methods in statistics. Amer. Statist., 33: 165-178.

Gallagher, M.J., 1984. Barite deposits and potential of Scotland. Trans. I.M.M., 93: A130-132.

Garrett, R.G., 1983. Opportunities for the 80's. Math. Geol., 15: 385-398.

Garrett, R.G., Goss, T.I. and Poirier, P.R., in press. Multivariate outlier detection - an application to robust regression in the earth sciences. Technometrics.

Gnanadesikan, R., 1977. methods for statistical data analysis of multivariate observations. Wiley, N.Y., 331pp.

Gnanadesikan, R. and Kettenring, J.R., 1972. Robust estimates, residuals and outlier detection with multiresponse data. Biometrics, 28: 81-124.

Govett, G.J.S., Goodfellow, W.D., Chapman, R.P. and Chork, C.Y., 1975. Exploration geochemistry - distribution of elements and recognition of anomalies. Math. Geol., 7: 415-446.

Halls, C., Reinsbakken, A., Ferriday, I., Haugen, A. and Rankin, A., 1977. Geological setting of the Skorovas orebody within the allochthonous volcanic stratigraphy in the Gjersvik Nappe, Central Norway. In: Volcanic processes in ore genesis. Special paper No. 7, I.M.M./Geol. Soc. (Lond), pp.128-151.

Harker, A., 1909. The natural history of igneous rocks. Methuen, London, 400pp.

Harris, A.L. and Pitcher, W.S., 1975. The Dalradian supergroup. In: A.L.Harris (Editor), Precambrian: A correlation of precambrian rocks in the British Isles. Special Rept. No. 6, Geol. Soc. (Lond), pp.52-75.

Healy, M.J.R., 1968. Multivariate normal plotting. Appl. Statist., 17: 157-161.

Hoaglin,D.C., Mosteller, F. and Tukey, J.W., 1983. Understanding robust and exploratory data analysis. Wiley, N.Y., 447pp.

Hoerl, A.E. and Kennard, R.W., 1970a. Ridge regression: biased estimation for non-orthogonal problems. Technometrics. 12: 55-67.

Hoerl, A.E. and Kennard, R.W., 1970b. Ridge regression: applications to non-orthogonal problems. Technometrics. 12: 69-82.

Hoerl, A.E. and Kennard, R.W., 1976. Ridge regression: iterative estimation of the biasing parameter. Communications in Statistics - Theory and Methods, A5: 77-88.

Hoerl, A.E. and Kennard, R.W., 1980. Ridge regression - 1980. Advances, algorithms and applications. Unpubl. annotated bibliography.

Hoerl, A.E., Kennard, R.W. and Baldwin, K.F., 1975. Ridge regression: some simulations. Communications in Statistics, 4: 105-123.

Hogg, R.V., 1974. Adaptive robust procedures: a partial review and some suggestions for future applications and theory. J.A.S.A., 69: 909-927.

Howarth, R.J. (Editor), 1983. Statistics and data analysis in geochemical prospecting. Handbook of exploration geochemistry. Vol. 2. Elsevier, Amsterdam, 437pp.

Howarth, R.J., 1984. Statistical applications in geochemical prospecting: a survey of recent developments. J. Geochem. Expl., 21: 41-61.

Howarth, R.J. and Earle, S.A.M., 1979. Application of a generalized power transformation to geochemical data. Math. Geol., 11: 45-62.

Howarth, R.J. and Sinding-Larsen, R., 1983. Multivariate statistics. In: R.J.Howarth (Editor), Statistics and data analysis in geochemical prospecting. Handbook of exploration geochemistry, Vol. 2. Elsevier, Amsterdam, pp.207-289.

Howarth, R.J. and Turner, M.St.J., in press. Statistical graphics in geochemical journals. Math. Geol., 19: 1-24.

Hoyle, M.H., 1973. Transformations - an introduction and a bibliography. Int. Stat. Rev., 41: 203-223.

Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Statist., 35: 73-101.

Huber, P.J., 1981. Robust statistics. Wiley, N.Y., 308pp.

Huff, D.L. and Black, W., 1978. A multivariate graphic display for regional analysis. In: P.C.C.Wang (Editor), Graphical representation of multivariate data. Academic Press, N.Y., pp.199-218.

Jacob, R.J.K., 1981. Comment. On: B.Kleiner and J.A.Hartigan, Representing points in many dimensions by trees and castles. J.A.S.A., 76: pp.270-272.

Joreskog, K.G., Klovan, J.E. and Reyment, R.A., 1976. Geological factor analysis. Elsevier, Amsterdam, 178pp.

Kendall, M.G. and Buckland, W.R., 1982. A dictionary of statistical terms. Longman, N.Y., 213pp.

Khaleelee, J., 1969. The application of some data processing techniques to the interpretation of geochemical data. Unpubl. Ph.D. Thesis, University of London.

Klau, W. and Large, D.E., 1980. Submarine exhalative Cu-Pb-Zn deposits - a discussion of their classification and metallogenesis. Geol. Jahrb., 40: 13-58.

Kleiner, B. and Hartigan, J.A., 1981. Representing points in many dimensions by trees and castles. J.A.S.A., 76: 260-276.

Kollung, S., 1979. Stratigraphy and major structures of the Grong district, Nord-Trondelag. Nor. Geol. Unders. Bull., No. 354: 1-51.

Krige and Magri, 1982. Studies of the effects of outliers and data transformation on variogram estimates for a base metal and gold ore body. Math. Geol., 14:557-564.

Large, D.E., 1980. Geological parameters associated with sediment hosted submarine exhalative Pb-Zn deposits. An empirical model for mineral exploration. Geol. Jahrb. 40: 59-129.

Lawless, J.F. and Wang, P., 1976. A simulation study of ridge and other regression estimators. Communications in Statistics - Theory and Methods. A5: 307-323.

Mancey, S.J., 1980. Computer based interpretation of large geochemical data sets. Unpubl. Ph.D. Thesis, University of London.

Mancey, S.J. and Howarth, R.J., 1980. Power transform removal of skewness from large data sets. Trans I.M.M., 89: B92-97.

Marquardt, D.W. and Snee, R.D., 1975. Ridge regression in practice. Amer. Statist. Assoc., 29: 3-19.

McDonald, G.C. and Ayers, J.A., 1978. Some applications of the Chernoff faces: a technique for graphically representing multivariate data. In: P.C.C.Wang (Editor), Graphical representation of multivariate data. Academic Press, N.Y., pp.183-197.

Nichol, I., Garrett, R.G. and Webb, J.S., 1969. The role of some statistical and mathematical methods in the interpretation of regional geochemical data. Econ. Geol., 64: 204-220.

Nightingale, F., 1857. Mortality of the British army. London.

Nightingale, F., 1858. Notes on matters affecting the health, efficiency and hospital administration of the British army. London.

Plant, J.A. and Moore, P.J., 1979. Regional geochemical mapping and interpretation in Britain. Phil. Trans. R. Soc. Lond., B288: 95-112.

Plant, J.A., Smith, R.T., Stevenson, A.G., Forrest, M.D. and Hodgson, J.F., 1984. Regional geochemical mapping for mineral exploration in northern Scotland. In: Prospecting in areas of glaciated terrain. I.M.M., pp103-120.

Playfair, J., 1786. Commercial and political atlas. London.

Playfair, J., 1801. The statistical breviary. London.

Playfair, J., 1805. An inquiry into the permanent causes of the decline and fall of powerful and wealthy nations. London.

Russell, M.J., Hall, A.J., Willan, R.C.R., Allison, I., Anderton, R. and Bowes, G., 1984. On the origin of the Aberfeldy celsian + baryte + base-metal deposits, Scotland. In: Prospecting in areas of glaciated terrain. I.M.M., pp.159-170.

Schmid, C., 1954. Handbook of graphical presentation. Ronald Press, N.Y.

Shaw, D.M., 1961. Element distribution laws in geochemistry. Geochim. et Cosmochim. Acta., 23: 116-134.

Smirnov, 1963. Statistical estimation of the geochemical background in geochemical prospecting. Geochemistry, 3: 353-364.

Smith, C.G., Gallagher, M.J., Coats, J.S. and Parker, M.E., 1984. Detection and general characteristics of stratabound mineralization in the Dalradian of Scotland. Trans. I.M.M., 93: B125-133.

Smith, C.G., Pitfield, P.E.J., Burley, A.J., Howard, S.H.D., Parker, M.E., Michie, U.McL. and Fortey, N.J., 1977. Investigation of stratiform sulphide mineralization in parts of central Perthshire. I.G.S., Mineral Reconnaissance Programme Rept. No. 8, 83pp.

Smith, C.G., 1977. Investigation of stratiform sulphide mineralization in parts of the Dalradian of central Perthshire, Scotland. Trans. I.M.M., 86: B50-51.

Temple, J.J., 1978. The use of factor analysis in geology. Math. Geol., 10: 379-387.

Trochimczyk, J. and Chayes, F., 1978. Some properties of principal component scores. Math. Geol., 10: 43-52.

Vistelius, A.B., 1960. The skew frequency distribution and the fundamental law of the geochemical processes. J. Geol., 68: 1-22.

Walker, F.A., 1874. Statistical atlas of the United States based on the results of the nineth census. U.S. Bureau of the Census, Washington D.C.

Wang, P.C.C. (Editor), 1978. Graphical representation of multivariate data. Academic Press, N.Y., 278pp.

Willan, R.C.R., 1980. Stratabound sulphide mineralization in the Dalradian supergroup of the Grampian Highlands, Scotland. Nor. Geol. Unders. Bull., No. 300: 241-258.

Wilson, M.J., Berrow, M.L. and McHardy, W.J., 1970. Lithiophorite from the Lecht mines, Tomintoul, Banffshire. Min. Mag. 37: 618-623.


Spey
1
2
3
4
5
6
7
8
9
A
B
C
D
E
F
G
H
I
J
R
U

Figure 2.2 Moray-Buchan study area and summary geology

A	Old Red Sandstone	
B	Southern Highland Group	
C	Tayvallich/Crinan subgroup	
D	Easdale subgroup	Argyll group
E	Islay subgroup	
F	Blair Atholl subgroup	
G	Ballachulish subgroup	Appin group
H	Lochaber subgroup	
I	Undifferentiated Lower/Middle Dalradian	
J	Moine	
U	Basic/ultrabasic	
R	Granitic	

TOWNS:

1 Aberchirder	2 Banff
3 Cullen	4 Dufftown
5 Huntly	6 Keith
7 Portsoy	8 Rothes
9 Tomintoul	

Key for point symbol maps:

+	< 20%
×	20% to 60%
o	60% to 90%
o	90% to 95%
O	95% to 99%
O	> 99%

APPENDIX 3

SUMMARY OF FORTRAN PROGRAMS USED

Program: WTHGRP

Purpose: to generate robust estimates of the means, standard deviations and correlation matrix of the input data.

The input to this program is an n x p data matrix X. The basic equations for defining robust estimators of means and covariances are:

$$X^{*} = \frac{\Sigma\ (W_i\ X_i)}{\Sigma\ W_i} \qquad \text{Equation A.1}$$

and

$$S^{*} = \frac{\Sigma\ W_i^{2}\ (X_i - X^{*})\ (X_i - X^{*})'}{(\ \Sigma\ W_i^{2} - 1)} \qquad \text{Equation A.2}$$

where

$$W_i = w\ (\ D_m^{*}\ ) = \psi\ (\ D_m^{*}\ )\ /\ D_m^{*} \qquad \text{Equation A.3}$$

and

$$D_m^* = [(X_i - \bar{X}^*)' S^{*-1} (X_i - \bar{X}^*)]^{1/2}$$

Equation A.4

when $\Sigma W_i = n$, i.e. all weights for n objects are unity, then the above equations define the classical estimates.

The ψ function is defined in Equation 6.9, Section 6.4.2.2. The univariate equivalents of equations 1, 2 and 4 are:

$$\bar{x}^* = \frac{\Sigma (w_i x_i)}{\Sigma w_i}$$

Equation A.5

$$s^* = \sqrt{\frac{\Sigma w_i^2 (x_i - \bar{x}^*)^2}{(\Sigma w_i^* - 1)}}$$

Equation A.6

$$D_m^* = [(x_i - \bar{x}^*)^2 / s^{*2}]^{1/2}$$

Equation A.7

1) Order the data values and generate quartiles for each variable;

2) calculate median and 0.74 * interquartile range for each variable.

3) calculate univariate D_m^* vector (Equation A.7) based on

the median and 0.74 * interquartile range as initial robust estimates of $\bar{x}^*$ and s^* respectively;

4) calculate new univariate robust means (Equation A.5) using weights generated from the current D_m^* vector;

5) calculate new univariate robust standard deviations (Equation A.6) using weights generated from the current D_m^* vector;

6) test for convergence: convergence is taken to occur when $| (s^{*(k-1)} - s^{*(k)}) |$ is less than or equal to 0.0001 * $s^{*(k-1)}$; where $s^{*(k)}$ is the current robust estimate of standard deviation. If convergence occurs or the number of iterations is equal to three go to STEP 8.

7) Calculate new D_m^* vector based on current estimates of $\bar{x}^*$ and s^*, and go to STEP 4.

8) Print out both classical and robust univariate statistics.

9) One of two paths may be taken to generate an initial estimate for the robust covariance matrix.

a) Robust covariances may be calculated from the classical correlation coefficients and robust standard

deviations using the following equation:

$$COV_{ij} = r_{ij} * s_i^* * s_j^*$$

b) Robust covariances may be recalculated using as weights the product of the weights given by the individual univariate estimators (STEPS 4 to 7). This is done using Algorithm AS41 from J. R. Statist. Soc. C. (1971) Vol. 20, No. 2.

10) Calculate multivariate D_m^* vector (Equation A.4) based on the vector of robust univariate means $\bar{X}^*$ and the robust covariance matrix S^* generated in STEP 9.

11) Calculate the new vector of robust means (Equation A.1) with weights generated from the current estimate of D_m^*;

12) calculate the new robust covariance matrix (Equation A.2) with weights generated from the current estimate of D_m^*;

13) test for convergence: convergence is taken to occur when all the absolute differences between a current covariance estimate and its previous estimate are less than or equal to 0.01 times the previous estimate. If convergence occurs or the number of iterations is equal to 10 go to STEP 15.

14) Calculate new D_m^* vector based on current estimates of the robust mean vector $\bar{X}^*$ and the robust covariance matrix S^* and go to STEP 11.

15) Multivariate statistics are printed out plus a list of the non-unit weights and their sample identities. The full robust D_m^2 vector is listed and a Q-Q plot of robust $D_m^{2/3}$ values is plotted (see Section 4.5).

16) The robust covariance and correlation matrices are written to separate files in packed lower triangle form to be used subsequently by other statistical procedures.

Program: LAMT

Purpose: to perform an optimal power transformation of a data matrix.

The basis of this program is the FORTRAN 77 code presented by Howarth and Earle (1979). The three optimization criteria that may be used (see section 5.3) are all based on parametric statistics which are subject to distortions in the presence of outliers in the data. It is necessary to trim or downweight the outliers prior to calculation of lambda.

1) calculate the multivariate D_m^* vector by iteration using a robust start (see program WTHGRP).

2) Plot the Q-Q plot of robust $D_m^{2/3}$ values;

3) ask the user to trim any multivariate outliers from the data. If the Q-Q plot shows a close approximation to a multivariate normal population for $D_m^{2/3}$ then go to STEP 5.

4) Calculate new D_m^* vector based on trimmed data and go to STEP 2.

5) Calculate optimal value for lambda for each variable based on the trimmed data;

6) transform the total data using the value for lambda generated from the trimmed data in STEP 5.

7) Output the transformed data to file and print diagnostic information containing the lambda value, skewness and kurtosis of the transformed data for each variable.

Program: TREES

Purpose: to generate a cluster dendrogram and Kleiner-Hartigan Trees of each object based on the input data.

Program TREES initially performs an R-mode agglomerative hierarchical cluster analysis. Agglomerative hierarchical techniques involve a series of successive fusions of variables into groups or clusters, such that the initial p variables are ultimately reduced to a single cluster containing all variables.

The basic procedure with all agglomerative methods is similar. They begin with the computation of a similarity or distance matrix between the variables. The end product of the methods is a dendrogram showing the successive fusions of variables, which culminates at the stage where all the variables are in one group. At any particular stage the methods fuse variables or groups of variables which are closest (or most similar). Differences between the methods arise because of the different ways of defining distance (or similarity) between a variable and a group containing several variables or between two groups of variables (see Section 7.5).

The dendrogram forms the template for the construction

of the Kleiner-Hartigan Trees (Section 7.4).

1) Generate a similarity matrix. This may be either a correlation or a distance matrix. The matrix may either be calculated from the input data or read directly from file.

2) Perform one of seven agglomerative clustering procedures (see Section 7.5 and Table 7.1). The algorithms used are from Spath (1980).

3) Generate the cophenetic correlation coefficient between corresponding values in the original similarity matrix and the cophenetic matrix.

4) Plot the cluster dendrogram with its cophenetic coefficient. The user is given the option to choose another clustering method. Then a template tree is plotted based on a dummy object vector of ones. Each tree is adapted from the dendrogram and includes the information in each object vector, based upon the rules outlined in Section 7.4. The subroutines for construction of the trees are from Kleiner and Hartigan's original code (Kleiner and Hartigan, 1981).

5) Once the user is satisfied with the cluster method used, the trees for each object are plotted in a gridded format. If the input data set contains sample site coordinates, the

trees may be plotted as a map.

Program: PCA

Purpose: to perform Principal Components Analysis or Factor Analysis.

The input to PCA may be just a correlation matrix, a correlation matrix and a data matrix, or a data matrix only. In all cases the correlation matrix (whether input or generated from the data matrix) is used to calculate the component or factor loadings matrix. If a data matrix is input, there is the option to calculate component or factor scores and to write them to a separate file. Rotation of the eigenvectors is optional (see Chapter 8).

The basic equation for both PCA and FA using a standardized data matrix Z (see Section 8.3) is:

$$Z = F A'$$

where F is the n x p matrix of component or factor scores and A is the p x p normalized and scaled loadings matrix generated from the correlation matrix.

1) If necessary generate the correlation matrix from the input data.

2) Compute the normalized loadings matrix Q and the vector of eigenvalues Λ . This is achieved using a subroutine given by Davis (1973).

3) Scale the normalized loadings matrix to produce A, the scaled and normalized loadings matrix.

$$A = Q \Lambda^{1/2}$$

4) Calculate the communalities of the loadings matrix A. These are the squared loadings summed across each row of A.

5) Print out the matrix A and the communalities.

6) If required, rotate the loadings matrix A. This uses the varimax rotation subroutine given by Davis (1973).

7) Print out the rotated loadings matrix A.

8) If a data matrix is input, calculate the component or factor scores according to the equation:

$$F = Z A (A' A)^{-1}$$

Note that since the correlation matrix is used the standardized data matrix Z must be used to calculate F. Subroutines to transpose and invert a matrix are from Davis

(1973).

9) The component or factor scores are written to a separate file for subsequent manipulation.

Program: RIDGE

Purpose: to perform a ridge regression analysis of an input set of data.

The first task of any regression program is to calculate the vector of unknown regression coefficients **b** from the equation:

$$\mathbf{b} = (X' X)^{-1} (X' Y) \qquad \text{Equation A.8}$$

where X and Y are unstandardized data matrices. It is more convenient computationally to generate the vector of unknown standardized regression coefficients **a** from the equation:

$$\mathbf{a} = (X' X)^{-1} (X' Y) \qquad \text{Equation A.9}$$

where X and Y are in standardized form, and (X' X) is the correlation matrix of the explanatory variables, and (X' Y) is the vector of correlations between the response variable Y and the explanatory variables. In ridge regression a slight adaptation is made to Equation A.9 by the addition of a constant k to the diagonal elements of

the correlation matrix:

$$\mathbf{a} = (X' X + kI)^{-1} (X' Y) \qquad \text{Equation A.10}$$

The standardized regression coefficients may be transformed to original units using the equations given in Section 9.3. The vector **b** generated in this way may then be used to calculate the predicted response vector and hence the residuals.

1) Generate the correlation matrix of the input data.

2) Calculate the standardized ridge regression coefficients for each of 14 values of k in Equation A.10. k = 0.0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0.

3) The coefficients are printed in tabular form and plotted in graphical form against their appropriate value for k. In this way the user may select the optimum value for k to use with the data.

4) The standardized regression coefficients for the ordinary least squares (OLS) solution when k = 0.0 are now transformed to original units. The predicted response vector is then generated using the data matrix, and the residual vector is calculated.

5) Analysis of variance (ANOVA) is then performed on this OLS solution. The ANOVA table is printed out.

6) An optimum value for k is then specified by the user and STEP 4 and STEP 5 repeated for this new set of standardized regression coefficients.

7) Regression coefficients and their standard errors are calculated and printed for both the OLS and the ridge solution (see Turner, 1980).

8) A quantile-quantile plot of the ridge regression residuals is plotted and the user is given the option to trim any outliers from the data and to perform a repeat analysis on the basis of this reduced data matrix.

9) If trimming is done, the correlation matrix of the reduced data set is calculated and the analysis proceeds from STEP 2.

10) If trimming is not required then the program is terminated with the observed, predicted and residual response vectors being written to file.

REFERENCES

Davis, J.C., 1973. Statistics and data analysis in geology. Wiley, N.Y., 550pp.

Spath, H., 1980. Cluster analysis algorithms. Wiley, N.Y., 226pp.

Turner, M.S., 1980. A comparative study of multiple regression techniques in geochemistry. Unpubl. M.Sc. Thesis. University of London.