STATISTICAL CHARACTERIZATION
OF SOIL PROFILES USING IN SITU TESTS

by

DAMIKA SAMPATH WICKREMESINGHE

B. Sc. (Civil Engineering) University of Moratuwa, Sri Lanka

M. Eng. (Geotechnical Engineering) Asian Institute of Technology, Thailand

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF CIVIL ENGINEERING

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 1989

67

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Civil Engineering

The University of British Columbia

2324 Main Mall

Vancouver, Canada

V6T 1W5

Date: _17 July 1989._

Abstract

Several statistical procedures that would enhance the site characterization capabilities of insitu test data with special emphasis on the cone penetrometer test have been proposed and presented.

Two methods to identify different soil layers from a profile have been described. One of these procedures is based on the effects of the individual parameters, namely, cone bearing, sleeve friction and pore pressure, while the other method employs a multivariate scheme of analysis, which has the capability of handling all three or any two parameters, simultaneously. The advantages of these statistical methods over the conventional methods of soil layer identification, have also been highlighted. Critical levels of the values of the Intraclass Correlation coefficient and the D^2 statistic have been proposed for the identification of layer boundaries as primary or secondary for both sand and clay type soils.

Methods of trend analysis have been proposed while the complications arising from the presence of correlations have been discussed. The role played by methods of statistical filtering and smoothing, in the identification of trends, have also been illustrated. Statistical procedures have been proposed, for the purpose of verification of non-stationarity or stationarity, in the event it cannot be determined from a visual inspection.

The need for the consideration of geotechnical data as random has been emphasized, together with applications of random field theory in the determination of exceedance probabilities of given threshold values over spatial averages of a soil layer. A computationally more convenient method for the determination of the scale of fluctuation has been proposed while emphasizing its importance in several areas of applications, with respect to the cone penetration test.

Time Series methods have been employed in order to model the stationary component of soil profiles and also have been extended to obtain the measurement noise of different test methods. A comparison of the measurement noise of different insitu testing devices, obtained by the time series method has been compared to a procedure based solely on the autocorrelation function of the data, resulting in a good agreement. The relatively low value of measurement noise obtained for the cone penetration test confirms its superiority over other insitu testing methods like the field vane test which gave fairly high estimates of the measurement noise.

A two dimensional interpolation procedure considering the correlation between data points has been recommended. This procedure which uses the autocorrelation function, has been applied to a set of cone penetrometer test data and the results of which have been compared with the actual profile at that location. The reasonable comparison of the predicted with the actual, clearly indicate the need for the consideration of correlations if they do exist, in interpolating geotechnical data in two or three dimensions.

IBM - PC compatible interactive micro computer programs have been developed in order to perform most of the techniques proposed in the thesis. These programs cater to any type of data format and have several inbuilt options available to the user. Detailed user manuals for these programs are also available.

Table of Contents

List of Tables

ix

List of Figures

Acknowledgement

DEDICATION

This dissertation is lovingly dedicated to my mother, Sylvia, who was back home in Sri Lanka, battling bravely against cancer for almost three years while I was pursuing my studies at UBC. She died on the 22 nd of December, 1987.

Chapter 1

Introduction

1.1 The Need for a Statistical Approach

Soil properties are highly variable and exhibit considerable variation from point to point. Most of these variations cannot be quantified and therefore, it becomes very important that the maximum amount of information is derived from an available set of data to reach conclusions on the characteristics of a soil profile. In geotechnical engineering, it is common to assume that the risk of failure is a function of the factor of safety, but the fact that is often neglected is that the risk of failure also depends on the accuracy with which the factor of safety is determined.

The variability or uncertainty in soil profile modeling has been explained by Vanmarke (1977) to comprise the following. The main source of variability is the natural inherent heterogeneity caused by the differences in particle size, mineral composition and stress history which are all mainly due to various geological influences. These also give rise to trends in both the horizontal and vertical directions, with the effect in the vertical dimension generally being more significant. Limited availability of data is the second source for the uncertainty since soil properties have to be deduced from field or laboratory tests on a limited number of samples. This problem can be averted or reduced by increased sampling, but since economics play a vital role this option is not always feasible except for major projects. Statistical and probabilistic approaches can maximize information that could be derived from a given set of data, and are therefore equivalent in some ways to performing more tests, if results are

to be analyzed solely on a deterministic basis. Measurement errors caused by man and machine are the third source of uncertainty in geotechnical test data. These are caused by factors such as sample disturbance, inaccuracies in testing procedures and human errors. All of the above uncertainties contribute to the belief that a stochastic approach employing statistical methods is the most efficient way of dealing with soil test data. There is nothing random if all the points in the ground could be tested accurately. However, this is not a feasible idea both practically and economically, giving rise to the need for the consideration of stochastic approaches in analyzing geotechnical test data.

This thesis examines a series of statistical procedures that may be useful to enhance the identification of soil profiles and thereby increase the site characterization capabilities of the cone penetrometer test (CPT). The high repeatability of the CPT and its capability of sampling at close intervals have caused it to emerge as one of the most widely used in situ testing methods. The large data base that results from the CPT provides an ideal tool for statistical applications. Traditionally, geotechnical engineers have been conservative, with most of the designs and analyses being based on fairly scanty data, acquired by methods about which much was not known in terms of theoretical basis. Over the years, more sophisticated testing methods both in the laboratory and the field have evolved, and the increased knowledge of soil behavior has led to the development of more elaborate theories and models, capable of predicting stress - strain characteristics of soils more accurately. In spite of all these technological and intellectual advancements, the uncertainty and the highly variable nature of soil behavior are still present. The overall accuracy of design analyses have not improved significantly, caused largely by the reluctance of the geotechnical engineer to replace the traditional deterministic methods by probabilistic techniques.

In the last decade, geotechnical engineers have been confronted by new challenges,

due to increased demands from field situations. Complicated high risk structures such as foundations for nuclear power plants and deep foundations for offshore oil platforms were required to be constructed. No more could a structure be built on the best site, but instead, the need of the day is more for the construction of the structure at the available site which may comprise a soil stratum with adverse conditions. These immense challenges have rendered the statistical and probabilistic approaches as ideal tools to supplement the traditional deterministic methods of analysis and design. Additional conservatism results in additional costs, and in very large projects as those mentioned earlier, this would be an unaffordable luxury in terms of economics. Thus the use of probabilistic methods which use statistical techniques to quantify uncertainties and risks is attractive.

Statistics enable the acquisition of a better understanding of limited data, permitting a better description of site characteristics, which in turn results in analysis and design requiring less conservatism. In the light of the above considerations the cone penetrometer test and statistical methods seem to be ideal partners which would enable the enterprising geotechnical engineer to meet the present day challenges and achieve the goals of dynamic design incurring the least cost, and most importantly at a reduced risk of failure.

1.2 Scope of the Thesis

This thesis examines several techniques which may be used for the statistical characterization of soil profiles with special emphasis on applications to data obtained from the CPT. These techniques are new to geotechnical engineering and would enhance the soil profile characterization capabilities of in situ testing methods. These techniques can also be applied to data obtained from other devices such as geophysical logging equipment which samples at reasonably close intervals.

Most statistical methods depend on the stationarity of data and likewise the techniques used in this thesis too will often refer to stationarity of profiles. In view of this frequent reference, it is appropriate to make a precise definition and a clear distinction between one dimensional and two dimensional stationarity, right at the outset.

Most soil properties increase in value with depth giving rise to trends. The nonstationary nature of soil property values are caused by these trends. Therefore, soil properties in the depth dimension can be expressed to comprise of two components as follows;

SOIL DATA = TREND + RESIDUAL

The resulting residual after trend removal fluctuates around the trend and is stationary. This explanation can also be extended to the horizontal dimension where applicable.

The main emphasis in this dissertation is on the investigation of single profiles in the depth dimension where the concern of stationarity will be in the depth dimension. In terms of first moments, stationarity implies a constant mean although in a wider sense, a stationary data set is defined as one which also has a constant variance and an autocorrelation function which is dependent only on the separation distance (lag distance) between data points. Two dimensional data analysis is performed in Chapter 6 where correlations in both the vertical and horizontal dimensions are determined. In such situations the stationarity of concern will be in the vertical and horizontal dimensions, although the basic definition remains the same. Figure 1.1 illustrates these two types of stationarity more clearly. In the one dimensional (depth dimension, z direction in Fig. 1.1) situation, the stationarity of concern will be for individual profiles. For example, the stationarity of profile V1 is independent of profile V2, V3 and V4 which are handled individually in one dimensional analysis. In the two dimensional analysis, not only will the concern be on the stationarity in
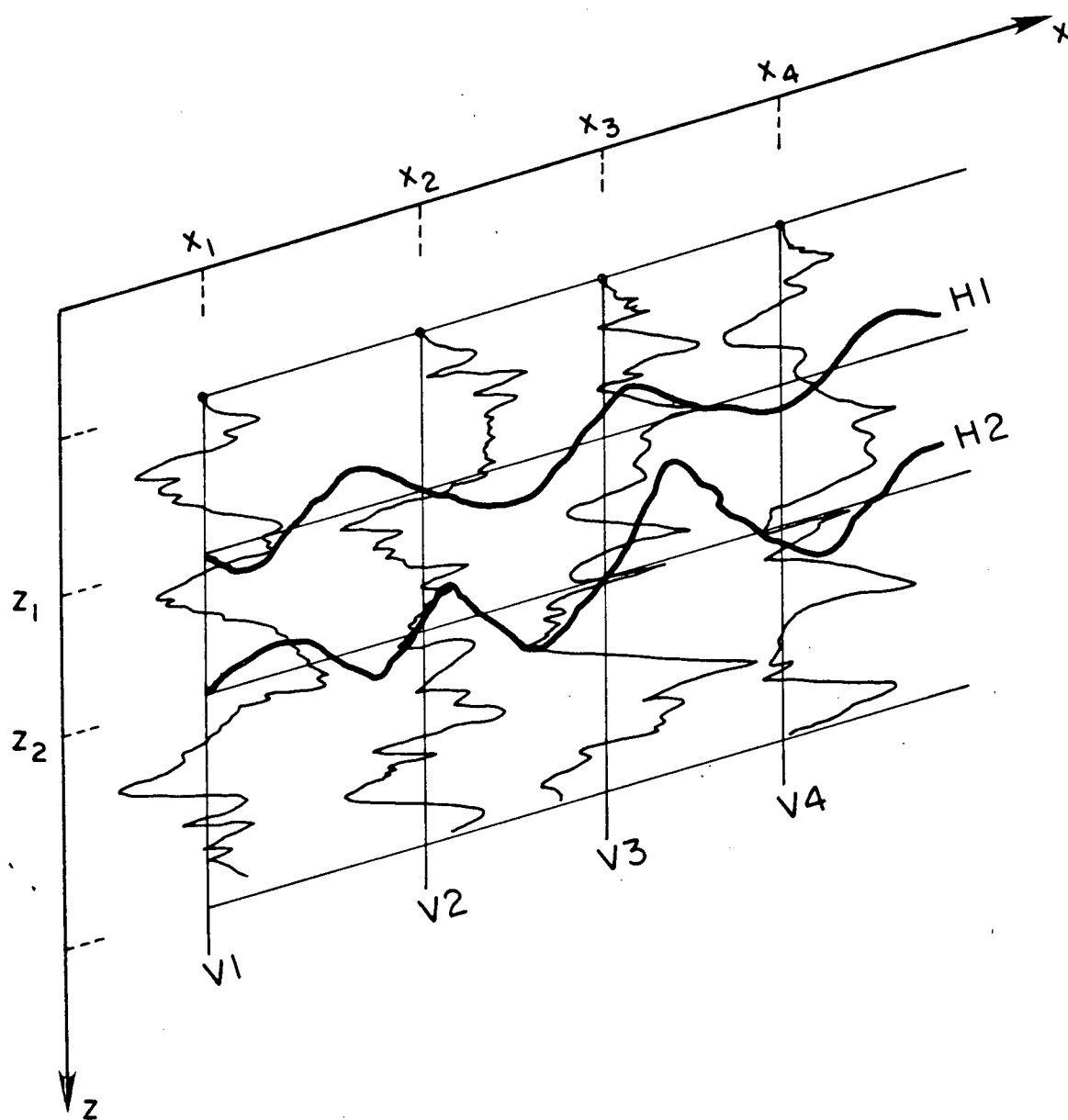
Figure 1.1: Illustration of Stationarity in One and Two Dimensions.

the vertical dimension but also on the stationarity of the generated profiles H1, H2 etc., at different depths z_1, z_2 and so on, respectively. The idea of stationarity and methods of determination of stationarity for both cases discussed above will be dealt with more thoroughly in the relevant sections of the thesis.

Soil profiles are highly heterogeneous and may consist of several substrata which exhibit different characteristics from layer to layer. It is of prime importance that each of these sublayers is identified prior to any design or analysis, since a mere visual inspection of the profile may not lead to the proper delineation of layers.

The cone penetration test is an ideal tool for the purpose of discriminating between different layer types, due to its capability of sampling at close intervals and also because of its high repeatability. Prior to any statistical analysis, it is necessary to divide the entire profile into statistically homogeneous sublayers, based on the mean, the variance and the trend. Well established empirical charts based on the friction ratio and the pore pressure ratio exist to classify different soil types present in a soil profile but at times, these methods are unable to determine layering accurately. The statistical methods to be described can be used to supplement the information obtained from the classical methods of layer identification. The three statistical techniques that will be tested for layer identification are the T Ratio, Intraclass Correlation Coefficient and the D^2 statistic.

The T Ratio or the Intraclass Correlation Coefficient can be used to investigate any one of the three main parameters obtained from the CPT, namely the cone tip bearing, sleeve friction and pore pressure, in determining the layer boundaries. A multivariate analysis which uses the D^2 statistic will be employed to investigate the combined effects of bearing, friction and pore pressure, together or for any combination of two of the above parameters. This type of analysis which considers two parameters is in a way equivalent to the conventional friction ratio method, if the two variables

considered are cone bearing and friction. On the other hand, it is also equivalent to the soil classification based on the pore pressure ratio, if the two variables considered are cone bearing and pore pressure. In all of the above three methods, namely, T Ratio, Intraclass Correlation Coefficient and the multivariate analysis using the D^2 statistic, a window of a pre-determined width will be passed along the data profile and the statistics on either side of the window center will be investigated. If the engineer requires a detailed analysis, which requires a more severe discrimination between layer types, a narrower window width can be chosen. All of the above three methods will be applied on three sets of CPT data, in an attempt to illustrate the advantages of the statistical methods over the conventional method, in identifying different layers in a soil stratum.

Once the layers are identified, the different trends and other properties of these sublayers will have to be characterized. Soil properties are highly depth dependent and therefore significant trends in the vertical dimension can be expected. Methods of trend analysis which essentially use regression techniques to describe different layers will be described. A measured soil property is made up of three parts: namely, the deterministic trend, the residual and the error term. The trend obtained from a regression analysis will be accurate only if it has succeeded in absorbing all the correlations present and if it has not, the residuals will also have to be considered in obtaining accurate estimates. The difficulties arising when dealing with geotechnical test data will be looked into in detail and methods of overcoming these problems will be highlighted. If layering in soil profiles is correctly identified, it is reasonable to assume that trends in geotechnical data will be linear or curvilinear. The curvilinear trend could be fitted with a polynomial of the second degree while the linear trend can be modeled by a straight line. Two applications on real data illustrate how the best model could be chosen. In some instances, a mere visual inspection may not

reveal the presence of a trend in a soil layer. In such cases, it may be a good idea to perform the 'RUN' test to determine the stationarity of a soil layer. Two applications of the 'RUN' test are used to illustrate the use of this method.

It is common that a CPT profile may consist of anomalies or extremities. In such cases, statistical filtering based either on the median or the mean can be employed in order to remove these anomalies. While the degree of filtering required is highly situation dependent, it has to be exercised with utmost caution, so that genuine data giving rise to actual thin layers are not removed. The main purpose of filtering should be to act as an aid in identifying trends. Methods of smoothing such as moving average smoothing and Fourier smoothing will be examined, together with applications, with a view to illustrate the effects of smoothing on the identification of trends.

Natural heterogeneity of soils, limitation of data availability, soil disturbance during testing, etc., all contribute to the uncertainty of soil data, which lead to the belief that the most appropriate method of analyzing soil data is by considering it as random. Different applications of the theory of random fields to cone penetration test data will be investigated in order to obtain a better understanding of the soil profile characteristics.

The scale of fluctuation is a parameter of great potential in the statistical characterization of soil profiles. The concept of this parameter and its multiple applications to CPT data will be investigated in detail in this dissertation. The idea of the scale of fluctuation was introduced to geotechnical engineering by Vanmarke (1977), but since then, other researchers have not made use of it. The present work has recognized its potential and will apply the concept to CPT data, from the point of soil variability and also as a tool to study the averaging effects of cone bearing, sleeve friction and pore pressure. The method of derivation used is a practical variant of the method

first used by Vanmarke (1977), and is very advantageous for computerization of this procedure. The advantages of the proposed method will be highlighted together with a comparison with the original method. The variance function from which the scale of fluctuation is derived and the scale of fluctuation itself are also used to investigate the correlation effects between spatial averages. For example, in the computation of foundation settlements, the effect of the settlement of an adjacent footing is very rarely considered in classical foundation engineering. However, correlations do exist and have to be considered if accurate estimates are required. Exceedance probabilities of spatial averages over threshold values with respect to layer thickness, threshold value considered and variability of the soil layer will be described with applications on CPT profiles. The concept of exceedance probabilities are of great concern to geotechnical engineers who are concerned about the magnitudes of the disturbing force and the available soil strength, especially in slope stability analysis.

The geotechnical engineer is very concerned that redundant data are not gathered in any site investigation program. Too much data acquisition will result in over expenditure while the effects of lesser exploration and testing than required could be drastic and may even result in catastrophic consequences. Therefore, it is very important to strike a balance between the two and with this aim, the optimum sampling distance has been derived as a function of three main factors; soil variability, the desired accuracy of the estimate and the confidence based on the estimate.

Time series methods will be used to demonstrate the beneficial use of autoregressive and moving average models to represent the stationary component of a soil profile after trend removal. Time series methods have also been used to evaluate the measurement noise in order to draw conclusions on the efficiency of different test methods. These results have also been compared to a different technique recommended by Baecher (1985). The other major component of errors in soils data is the

bias error which can only be determined with respect to a different test method and the evaluation of even an approximate estimate of the random error would be useful in determining the quality of a given set of data.

Soil properties are highly correlated especially in the vertical dimension. Any interpolation method similar to regression will be accurate only in the absence of correlation. This is a very common assumption which is often violated in geotechnical engineering when estimating soil properties at untested locations. Therefore, any two-dimensional interpolation procedure which has the depth dimension as one of its axes, will necessarily have to consider correlations, if accurate estimates are required. A new method of formulating the two dimensional autocorrelation functions applicable to geotechnical test data analysis will be proposed in order to perform interpolation considering the correlation between points. Different types of autocorrelation and semi-variogram functions which can be used to model soil property correlations in one, two or three dimensions will also be presented. Autocorrelation functions for a given set of two dimensional CPT data have been developed and interpolation performed. The results obtained will be compared with actual CPT data, in making evaluations of the recommended interpolation procedure.

1.3 Organization of the Thesis

Chapter 2 of the thesis will present univariate and multivariate statistical methods which can be used for the identification of layering in a soil profile. These methods include the T Ratio, Intraclass Correlation Coefficient and the D^2 statistic.

Methods of Trend Analysis will be described in **Chapter 3**, with the objective of obtaining a better understanding of the characteristics of the soil profile. The effect of smoothing and filtering on trends and a method for determining the stationarity of a soil profile based on the 'RUN' test will also be explained in **Chapter 3**.

Several applications of random field theory on CPT data will be detailed in **Chapter 4**. The concept of the scale of fluctuation in evaluating the variability of a soil profile, applications of exceedance probabilities and the effect of variability on the optimum sample spacing for a given layer will also be presented in **Chapter 4**.

The role of time series methods in the interpolation of one dimensional geotechnical test data will be discussed in **Chapter 5** together with approximate methods which can be used for the estimation of the random noise component of different in situ testing devices.

Chapters 1 to 5 deal with data profiles in the depth dimension and therefore are all one dimensional types of analyses.

Chapter 6 presents a procedure for the interpolation of soil property values in two dimensions with one of these dimensions being the depth. Correlation of soil property values between data points is considered in the analysis.

Chapter 7 describes a simple case history involving soil densification where statistical methods such as layer identification, trend analysis and the concept of the scale of fluctuation have been used to verify the effects of soil improvement.

Chapter 8 presents the final conclusions of this dissertation.

1.4 Interactive Micro Computer Programs

Several interactive IBM - PC compatible micro computer programs have been developed to accommodate different data formats. These programs which have been written in Microsoft Fortran are very flexible with several options available to the user.

The programs include procedures for layer identification, statistical filtering, smoothing and trend analysis. The determination of the scale of fluctuation and the evaluation of stationarity using the 'RUN' test to check for stationarity can also be performed

using these programs.

The detailed manuals of these programs are available in the Department of Civil Engineering of the University of British Columbia.

1.5 In Situ Testing Devices

Traditionally, in situ testing methods have been used by geotechnical engineers to gain a better understanding of the qualitative characteristics of the subsoil. Modern techniques have resulted in improved methods of testing and sophisticated data acquisition systems which have stimulated rapid development in in situ testing methods over the years, especially so in the last decade. It is possible that in the not so distant future, in situ testing methods will play a more dominant role in geotechnical engineering. Mitchell et al.(1978) have listed four main reasons, supporting this prediction. They are,

(i) The ability to determine properties of soils, such as sands and offshore deposits, that cannot be sampled in the undisturbed state.

(ii) The ability to test a larger volume of soil than can be conveniently tested in the laboratory.

(iii) The ability to avoid some of the difficulties of laboratory testing, such as sample disturbance and the proper simulation of in situ stresses.

(iv) The increased cost effectiveness of an exploration and testing program using in situ testing methods.

Difficulties such as the inability to independently vary stress direction and stress paths, the unknown effects of principal stress rotation during testing, the inability to control drainage independently and the semi empirical nature of interpretation methods are some of the shortcomings of most in situ testing techniques. These shortcomings inhibit the development of a theoretical background which would be able

to explain fully the behavior of a soil element adjacent to an in situ testing device. If this adversity is overcome, it may be possible to replace all the empirical correlations available at present, with theoretical expressions having a sound fundamental basis.

Most of the data to be dealt with in this thesis have been obtained from the cone penetrometer test. The CPT is becoming increasingly popular as an in situ test for site investigation and geotechnical design, because of its high repeatability and relatively low measurement noise (Wu, 1986). As a logging tool for geotechnical engineering purposes, it is efficient with respect to the delineation of stratigraphy and in its capability of performing simultaneous measurements of data on several channels.

Results from the CPT have been used to develop empirical correlations with soil parameters such as friction angle, relative density and shear strength. The seismic cone which is an improvement of the basic CPT, measures the shear wave velocities from which the maximum shear moduli can be estimated. In recent times, researchers have developed more rational correlations between the standard penetration test (SPT) 'N' value with the cone bearing obtained from the CPT. As a result of these correlations, Seed's original liquefaction curves based on the SPT, have been extended for use with the CPT, enhancing its capabilities. In spite of all these advances, the CPT is primarily an efficient logging tool and interpretation charts have been developed to identify soil strata based on the friction ratio and cone bearing. Detailed descriptions of these procedures are available in Campanella et al. (1983, 1988) and Robertson (1983). Descriptive accounts of the different types of equipment and testing procedures, and methods of data acquisition are available in Robertson and Campanella (1986).

The cone used for this research at the University of British Columbia has a cone tip of 10 cm^2 base area with an apex angle of 60°. It is illustrated in Fig. 1.2. The

swage fitting to lock
14 conductor cable

wires spliced to cable
inside tube

seismometer

slope sensor

Quad ring

strain gages for
friction load cell

equal end area
friction sleeve
(150cm² area)

temperature sensor

pressure transducer

strain gages for cone
bearing load cell

O-rings

Quad ring

porous plastic

60°cone
35.68mm O.D.

small cavity
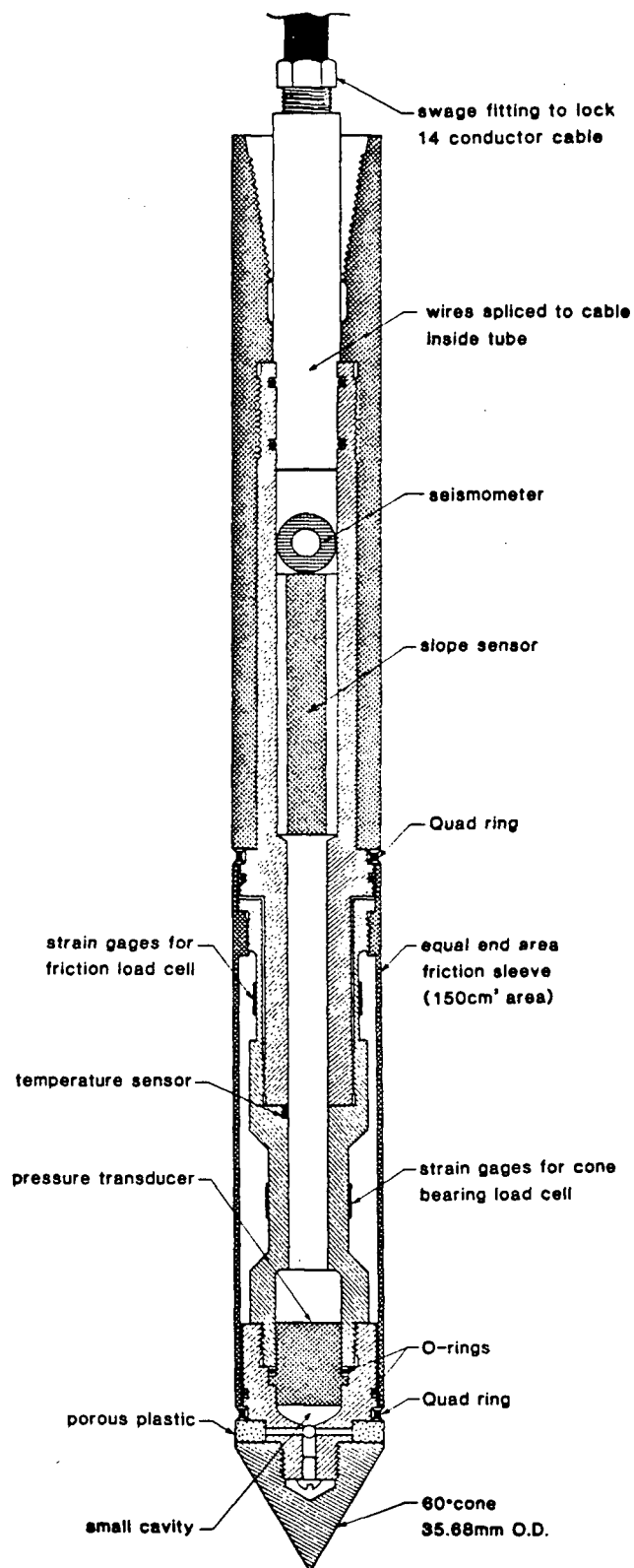
Figure 1.2: Cone Penetrometer used at UBC.

friction sleeve located immediately behind the cone tip has a standard area of 150 cm^2. The cone is made to penetrate at 2 cm/sec and has the facility to sample on six different channels at 2.5 cm intervals, measuring the cone bearing, sleeve friction, pore pressure at the tip and behind the sleeve, temperature and inclination.

The most widely used in situ testing device in North America is the standard penetration test (SPT). Many soil properties have been correlated to the 'N' value obtained for the SPT. More details of the SPT are available in any text on foundation engineering. The flat plate dilatometer (DMT) is another test which is known for its repeatability. Like the CPT, several empirical correlations have also been established for the DMT. The most important parameters that could be derived from the DMT are the friction angle in sands, the lateral stress coefficient and the over consolidation ratio which are all dependent on three index parameters. These parameters are also used for soil classifications. More details of the DMT is available in Jamiolkowski et al.(1985).

There are three types of pressuremeters being used in practice at present. They are the Menard pre-bored pressuremeter, the self-bored pressuremeter and the full-displacement pressuremeter. The more important parameters which can be obtained from the pressuremeter are the shear modulus and the lift off pressure which could be related to the lateral stress in the ground.

The field vane test is used to determine the undrained shear strength of cohesive soils in the field. This test measures the shear strength of soils, both in the undisturbed and remolded states, and hence could be used for determining the sensitivity of the soil. However, the performance of the vane test is questionable due to the high disturbance of the soil around the apparatus, and the ensuing measurement noise which is found to be very significant.

The dynamic cone penetrometer (DCPT) is somewhat similar to the SPT since it

measures the number of blowcounts required for each foot of penetration. However, unlike the SPT, it uses a dynamic load to drive the rods, thus causing an enormous disturbance in the surrounding soil. Inconsistency and the very low repeatability of the DCPT are the causes for its limited usage in geotechnical site investigations.

The screw plate test is a modified form of the plate load test with the ability of performing unloading and reloading cycles of load on the soil at depth. The load - settlement curve which is the main product from such a test, is used to obtain the vertical modulus of deformation of the soil, which in turn is used to determine settlement characteristics of soils under load.

A fairly recent innovation in in situ testing devices, is the Iowa Stepped Blade (Handy et al., 1982) which is used to estimate the in situ horizontal stress. This instrument is made up of sections of varying blade thickness along its length and the horizontal stress of the soil can be measured at the center of each of these sections of different thicknesses. The resulting extrapolated value of stress at zero blade thickness gives a reasonable estimate of the in situ horizontal stress.

More comprehensive details of all of the above in situ testing methods are found in Jamiolkowski et al. (1985). The insitu testing devices briefly described above are only those directly related to geotechnical investigations. The statistical methods presented in this thesis are also applicable to geophysical logging techniques such as gamma ray, sonic, nuclear and electrical logging etc.. A comprehensive description of most of the geophysical logging devices are given in Telford et al.(1976).

1.6 General Geology and Site Descriptions

The statistical techniques presented in this thesis will be applied to data obtained from several research sites in the lower mainland in British Columbia. They are,

- McDonald Farm

- Haney Site

- Tilbury Island

- Langley (Upper 232nd)

- Strong Pit

- B. C. Hydro Railway Site

- Annacis Island Site

The general location of the above sites are shown in Fig. 1.3.

1.6.1 McDonald Farm

McDonald Farm is located at the northern edge of Sea Island in the municipality of Richmond. The island is one of several that make up the Fraser River delta. The general geology consists of deltaic distributory channel fill and overbank deposits which overlie post glacial estuarine and marine sediments (Armstrong, 1978). The general stratigraphy of the sites consists of a soft organic clay in the top 2 m underlain by loose to dense coarse sand upto about 15 m. The soft normally consolidated clayey silt which lies below this sand extends to a depth of 300 m (Greig, 1985).

1.6.2 Haney Site

The Haney site is located in the Haney Slide site and is situated approximately 30 km east of Vancouver almost directly below the town centre of Haney. The site is a remnant of the Haney slide of January, 1880. The general geology consists of interbedded marine, glaciomarine and glacial sediments of the Fort Langley Formation. The soil profile consists of a fill in the top 2 m underlain by a meter of sand which overlies a sandy silt to silty clay extending to 30 m (Greig, 1985).
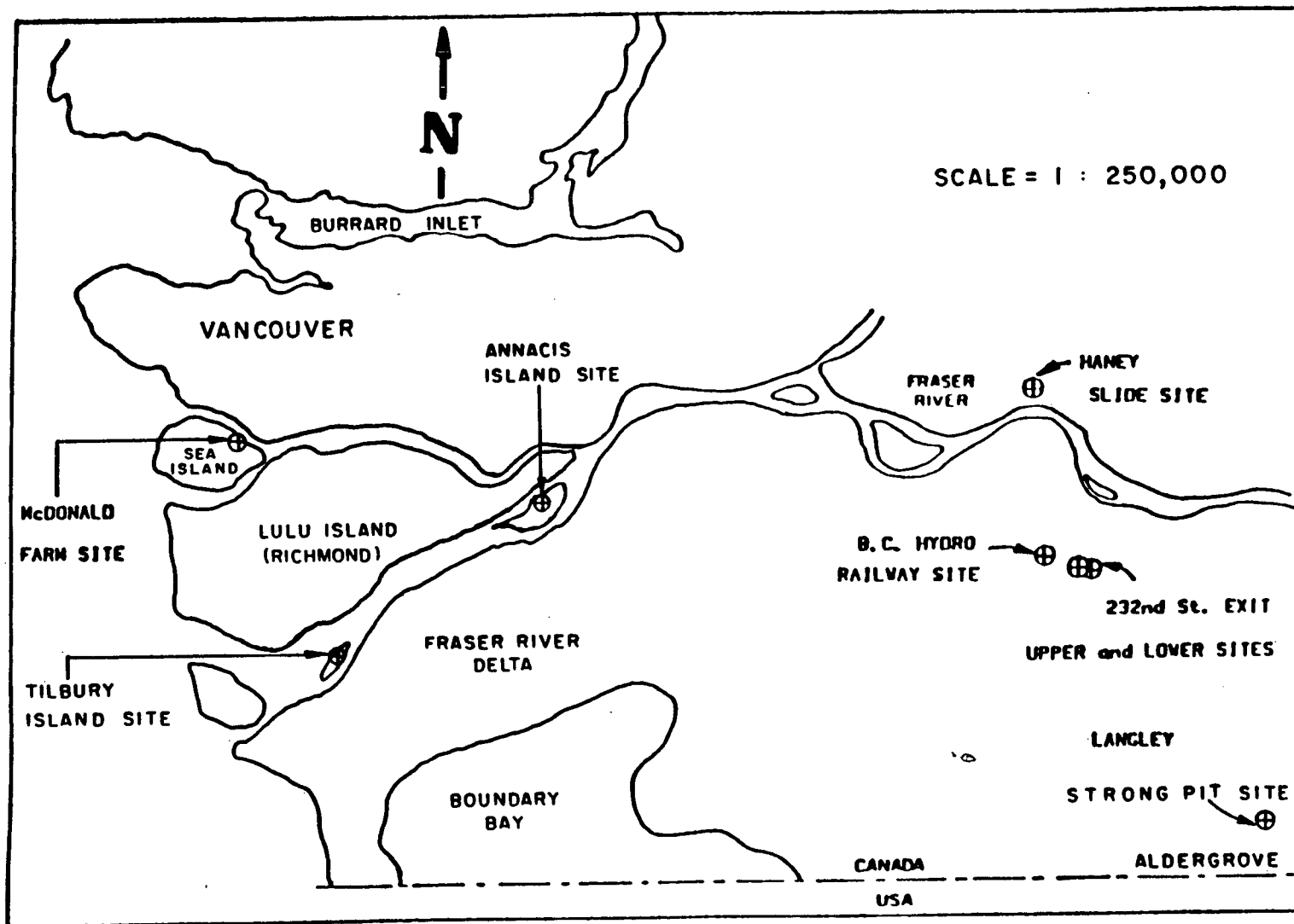
Figure 1.3: Location Map of Research Sites.

1.6.3 Tilbury Island

The site from which the data was obtained is situated immediately adjacent to the B.C. Hydro LPG plant located towards the north-eastern side of Tilbury Island. It consists of overbank sandy silt to silty loam (about 2 m thick) overlying 15 m or more of deltaic and distributory channel fill including tidal flat deposits. These are mainly interbedded fine to medium sand with intrusions of slight silt lenses.

1.6.4 Langley Site

The site is located at the 232nd St. exit of the Trans Canada Highway in Langley. It is about 1 km east of the B. C. Hydro railway site. This site is on a compacted clay fill that forms the approach for the 232nd St. overpass and lies at the western extent of the Fort Langley Formation (Greig, 1985). This formation has recorded at least three advances and retreats of a valley glacier and comprises of interbedded marine, glaciomarine and glacial sediments (Armstrong, 1978). The stratigraphy consists of a 2.5m compacted organic clay fill followed by a overconsolidated silty clay beween approximately 2.5 and 7.5 m. This is underlain by a normally consolidated silty clay with occasional sand lenses, extending below 20 m.

1.6.5 Strong Pit

This site is located at the Strong Gravel Pit near Aldergrove (in British Columbia) which is in the central Fraser Valley in the Fort Langley glaciomarine deposits. The stratigraphy at Strong Pit consists of a outwash sandy gravel (Sumas Formation) in the top 1.5m. Below the sandy gravel is an overconsolidated clay extending past 10 m with a thin layer of sand at about 9 m.

1.6.6   B. C. Hydro Railway Site

This site is located at the base of a 5m cut adjacent to the Trans Canada Highway in Langley. It is situated approximately 100 m west of the B. C. Hydro railway overpass (Greig, 1985). The site is located at the eastern extent of the Capilano sediments which consist of raised deltas, intertidal and beach deposits and glaciomarine sediments (Armstrong, 1978). The top 2.5 m of this site consists of mixed gravel and sand fill underlain by a lightly overconsolidated silty clay with occasional silty sand layers between 2.5 and 10 m. The layer below this is a normally consolidated silty clay extending beyond 30m.

1.6.7   Annacis Island Site

The site in which the soil compaction was performed is situated at the north side of Annacis Island along the north channel crossing and immediately east of the Alex Fraser Highway (Gray Beverage canning plant site). It constitutes an artificial promonotory built by infilling with dredged sand behind a rockfill dike. Investigations indicated that the site was covered by a 1.8 to 2.4 m thick sand fill on top of a 2.4 to 3.9 m thick clayey silt underlain by an alluvial sand extending below 10 m.

1.7   Review of Literature on Statistical Methods

The pioneering work of statistical applications to soil test data was performed by Lumb (1966) who investigated the variability of natural soils. Subsequent work by Lumb (1967,'70,'74,'75) all refer to basic statistical applications which covered topics such as the sampling patterns, identification of trends, distribution function of soil properties and the precision and accuracy of soil tests, etc..

Kay and Krizek (1971) considered the effects of correlations of estimates and also the coefficient of variation in deriving probability distributions of soil properties.

Holtz and Krizek (1971) made use of statistical parameters in large projects such as dam and bridge construction. Krizek's real contribution to the field of statistics and probability in geotechnical engineering is found in Alonso and Krizek (1975) who considered soil properties as random variables. The autocorrelation function was used to express soil property correlations while also introducing the spectral density function as an adequate descriptor of soil properties. Rizkallah et al.(1975) have made use of regression techniques in obtaining soil parameters, neglecting effects of correlations. Rizkallah et al.(1979) have also used the concept of energy to perform comparisons between the static cone penetrometer and the dynamic cone penetration test, employing methods of multiple regression.

The knowledge in the field of statistics in soil engineering has been enhanced significantly, by the contribution of Baecher (1982) who was especially interested in probabilistic site exploration problems, and emphasized the need for the identification of statistically homogeneous layers prior to any analysis. Baecher and Ingra (1979) and Baecher (1981), considered the autocorrelation function in expressing soil property correlations. The sources of data scatter and the different types of errors encountered in geotechnical data have been described by Baecher (1984a, 1984b) who described a procedure by which the measurement noise of data could be determined. In addition to detailed descriptions of error analysis and uncertainty in geotechnical engineering, the importance of the autocorrelation function and the correlation coefficient of soil property values in the determination of accurate estimates for problems of bearing capacity and settlement have been reiterated in Baecher (1985).

Several applications of simple statistical procedures, similar to the evaluation of probability distribution functions, correlation of soil properties such as shear strength through methods of regression, etc., are found in Cheong at al.(1980), Haldar (1981),

Asaoka et al.(1982), Krahn et al.(1983) and Anderson et al.(1984) and Johannesson (1985). All of these publications, have concentrated on basic statistical aspects, disregarding the more sophisticated problems in the characterization of soil profiles.

The first major attempt in introducing Bayesian concepts to geotechnical engineering was made by Tang (1971). This is an excellent introduction to Bayesian evaluation and information for foundation engineering. Since then Bayesian methods in the estimation of soil properties have also been used by Veneziano et al.(1975), where modeling has been performed, accounting for soil property uncertainty, with the exponential function being used to represent soil property correlation. In more recent times Vita (1984a, 1984b) has used Bayesian methods, incorporating both the soil property variability and the uncertainty factor caused by the limitation of data availability.

The real major contribution to the area of 'modern' statistical soil profile modeling was made by Vanmarke (1977a) who used the theory of random fields, in the derivation of various parameters which definitely improved the knowledge in problems of site characterization. In addition to the introduction of the scale of fluctuation to geotechnical engineering, theories on the probabilities of exceedance and the correlation of spatial averages (Vanmarke - 1978a, 1978b) have also been introduced. These concepts have been applied in the estimation of the reliability of earth slopes (Vanmarke - 1977b). The text *Random Fields; Analysis and Synthesis* (Vanmarke - 1983), is a significant contribution to the understanding and application of random field theory, which is relatively new in geotechnical engineering.

Tang (1984a, 1984b, 1987) has also applied the theory of random fields to geotechnical test data. In addition, Tang (1979) has performed a probabilistic evaluation of penetration resistances, considering uncertainties of the inherent spatial variability of soil and location of sampling point on the determination of unbalanced moments on

gravity platforms. Tang et al.(1984) also give a conclusive description of a procedure for the probabilistic evaluation of gravity platforms.

Wu et al.(1985) used methods of Time Series Analysis on geotechnical data to represent soil property variation with autoregressive and moving average models. The importance of probabilistic and statistical approaches has been highlighted in Wu (1973) and also in the extensive probabilistic site exploration studies (Wu - 1981), including methods suited for offshore conditions (Wu - 1986).

In the area of two and three dimensional site characterization, the only recognizable attempt since Baecher (1982) has been made by Tabba et al.(1981a, 1981b) who used polynomial equations to represent the autocorrelation functions. A similar approach has also been used by Kulatilake et al.(1987,'88). Prior to this work in the field of interpolation of soil test data, Tabba and Yong (1979) described a procedure, where the maximum likelihood function was used to estimate trend coefficients. Yong (1984) also analyzed the probabilistic nature of soil properties such as shear strength, consolidation characteristics, Atterberg limits and chemical composition, before concluding most appropriately, that the degree to which samples are representative of the soil stratum under investigation is the most important factor that controls the determination of the real soil property.

With the advent of geostatistics in the field of geology and mining (Agterberg, 1970,'74), geotechnical engineers too have begun in recent years, to appreciate its practical advantages (Soulie, 1984). In addition to the procedure to identify soil layers using the variogram, its versatility in the area of soil property interpolation, seems promising (Christakos, 1985,'87). Although variogram modelling and techniques of 'Kriging' have not been used in geotechnical engineering they have found wide applications in soil science (Webster, 1980,'85), mining and geology (Journel et al., 1978, Davis et al., 1978, Delfiner, 1973,'76).

Chapter 2

Identification of Soil Layers

2.1 Introduction

2.1.1 General

The proper identification of soil layers is important for design in order to obtain reasonable engineering parameters for the different layers. Accurate identification of sublayers is also important from a statistical view point, since all statistical analyses have to be performed on essentially statistically homogeneous layers.

The cone penetration test (CPT) performs data logging at close intervals and does simultaneous measurements on several channels yielding, for example, values of cone bearing, sleeve friction and pore pressure. The electrical cone penetrometer is essentially a logging tool and proper methods of layer identification should be a high priority. At present, this important task is performed by visual inspection of the various soil parameter profiles and by studying the variation of either the friction ratio, R_f (the ratio between the sleeve friction and the cone bearing) or the pore pressure ratio, B_q (the ratio between the excess pore pressure and corrected cone bearing) with the cone bearing.

A low friction ratio with high bearing is evidence of soil which is predominantly granular and a high ratio with low bearing implies a soil which is mainly cohesive with composites and silty soils lying somewhere between. The ratios obtained are then used together with the cone bearing to predict the particular type of soil encountered, from well established soil classification charts such as the one shown in Fig. 2.1.

Figure 2.1: Simplified Soil Classification Chart for the CPT (after Robertson and Campanella, 1986)

2.1.2 Soil Classification Chart for the CPT

The classification chart given in Fig. 2.1 identifies soil types as indicated, but fails to identify soil sublayers which may be present in a layer of sand, silt or clay. Once the main soil types are determined, correlation graphs are used to obtain appropriate values of friction angle and relative density in conjunction with cone bearing. The resulting values of different friction angles and relative densities would give an indication as to the different sublayers present in the stratum. The above procedure is subjective and could result in erroneous demarcation of sublayer boundaries and could be improved if a method can be deployed to identify different sublayers. Having a knowledge of the existing sublayers, the classification chart based on the friction ratio and the correlation graphs of relative densities and friction angles could then be used to determine the different parameters. The classification chart covers a wide range of values, especially in the case of sands. For example, a soil having a cone bearing value between 70 and 180 bar and a R_f value less than 1.4% is classified as sand while a soil which has a cone bearing value between 180 bar and 500 bar with a R_F value of less than 2.0% is classified as a gravelly sand. It is obvious that a soil encompassing such a wide range of bearing values will have several sublayers of different relative densities and friction angles which could be obtained from correlation graphs. As mentioned previously, this procedure would be simplified and made less ambiguous, if the sublayers could be accurately identified.

The friction ratio is a function of the sleeve friction and cone bearing. The cone precedes the sleeve which is also much longer, resulting in the sleeve friction being indicative of an averaged value. This effect certainly imposes a mechanical limitation on discernible layer thickness.

A survey involving fifteen people who have all had prior experience with the CPT chart, was conducted to ascertain the subjectivity involved in its use. The results

indicated a fair amount of discrepancies in the layers identified by the fifteen subjects.

2.1.3 Improved Methods of Layer Identification

With the intention of identifying layering more accurately, several statistical procedures will be described and followed up with several applications to illustrate the advantages and superiority of such methods.

The statistical methods proposed will also enable the engineer to decide on the different number of layers he could select, by inspecting the statistics of the sub-regions within the main layer. If the design requires more detail and sophistication, a number of layers based on less critical limits can be chosen, while for a general design for a low risk structure, the layering can be based only on the more critical or the highest peaks of the statistic profile. Details of these procedures will be described later in the chapter.

2.1.4 Characteristics of a CPT Profile

The data from a typical CPT profile comprise of the following characteristics.

(a) The data may have a lot of short range variations and a search for a longer pattern is difficult. Such variations are often erratic and may be regarded as noise.

(b) The data are highly irregular and often consist of sharp changes. There is no way of representing these variations as functional forms between soil parameters and depth, unless different layers are clearly identified depending on the acceptance level of the engineer.

(c) The data are multivariate, that is bearing, friction and pore pressure have an influence on the soil type encountered. It may be that the dependency on bearing is

much higher than the other parameters, but there is no a priori reason for accepting it. Furthermore, since the CPT gives all these data, it is always more meaningful to base predictions on the maximum amount of information that can be derived.

From the above explanation, it is evident that a statistical method of identifying different types of layers is justifiable. Most of the statistical and probabilistic methods rely on the homogeneity or stationarity criteria of soil properties within a sublayer and the proper identification of layers becomes vitally important. In addition to the univariate and multivariate statistical methods, a detailed procedure of nonlinear optimization techniques was also applied in an attempt to estimate soil layer boundaries, but this method was not successful as described below.

2.2 Nonlinear Optimization Techniques

Traditional trend analysis techniques use the concept of minimizing the squared differences, which is known as least squares. This is also a process of linear optimization in the case of linear trend analysis since the layer start and end depths are known. The optimization equation or the equation to be minimized can be simply expressed for each layer which has been decided a priori. The details of these procedures are described in the chapter on Trend Analysis. As explained in that chapter, the squared differences are minimized with respect to the unknown regression coefficients. However, the simplicity of this procedure is lost, when the different layer depths are unknown and the layer depths, too, become variables, resulting in the linear problem being transformed into a more complicated nonlinear optimization problem. The nonlinear optimization function F_L can be expressed as,

$$F_L = \sum_{l=1}^{l_N} \sum_{i=1}^{N} \left(Q_i - \hat{Q} \right)^2 \qquad (2.1)$$

where, $\hat{Q}$ is the regressed estimate, l_N is the number of layers and N is the number of data in a particular layer. Three different types of trends were tried for the estimation of $\hat{Q}$;

(a) Linear trend with continuity at the border (Fig. 2.2a).

(b) Linear trend with discontinuity at the border (Fig. 2.2b).

(c) Linear constant trend (Fig. 2.2c).

If l_N is the number of layers (starting and ending co-ordinates unknown) the number of unknowns (coefficients of the trend line) would be, $2l_N$ for case (a), $3l_N - 1$ for case (b) and $2l_N - 1$ for case (c). Due to the nature of the optimization equation above, it is impossible to determine the partial derivatives of F_L with respect to the variables in closed form. Therefore, two nonlinear optimization routines available in the UBC MTS system were used. Out of these two, the routine NLPQ0 (Vaessen, 1984) calculates the derivatives within it while the routine POWEL (Vaessen, 1984) does not require the partial derivatives of the optimization function. For the iterative procedures adopted in these routines to be efficient and yield satisfactory results, both these programs need good starting points for the variables. Due to the irregular shape of the function F_L, convexity is not assured, and invariably the local minimum is not the global minimum. The original intention was to perform the analysis for different layer numbers (l_N) and to investigate the minimized value of the function $F_L(F_{L_{min}})$ in order to select the layer number which results in the lowest $F_{L_{min}}$. A detailed analysis was done for different types of profiles and in about ninety percent of the cases it was found that the optimum layer depths obtained for both NLPQ0 and POWEL were highly sensitive to the prescribed starting values. As a result of this adversity the idea of using nonlinear optimization techniques for estimating layer boundaries was abandoned in favor of the preferred simple statistical methods to be
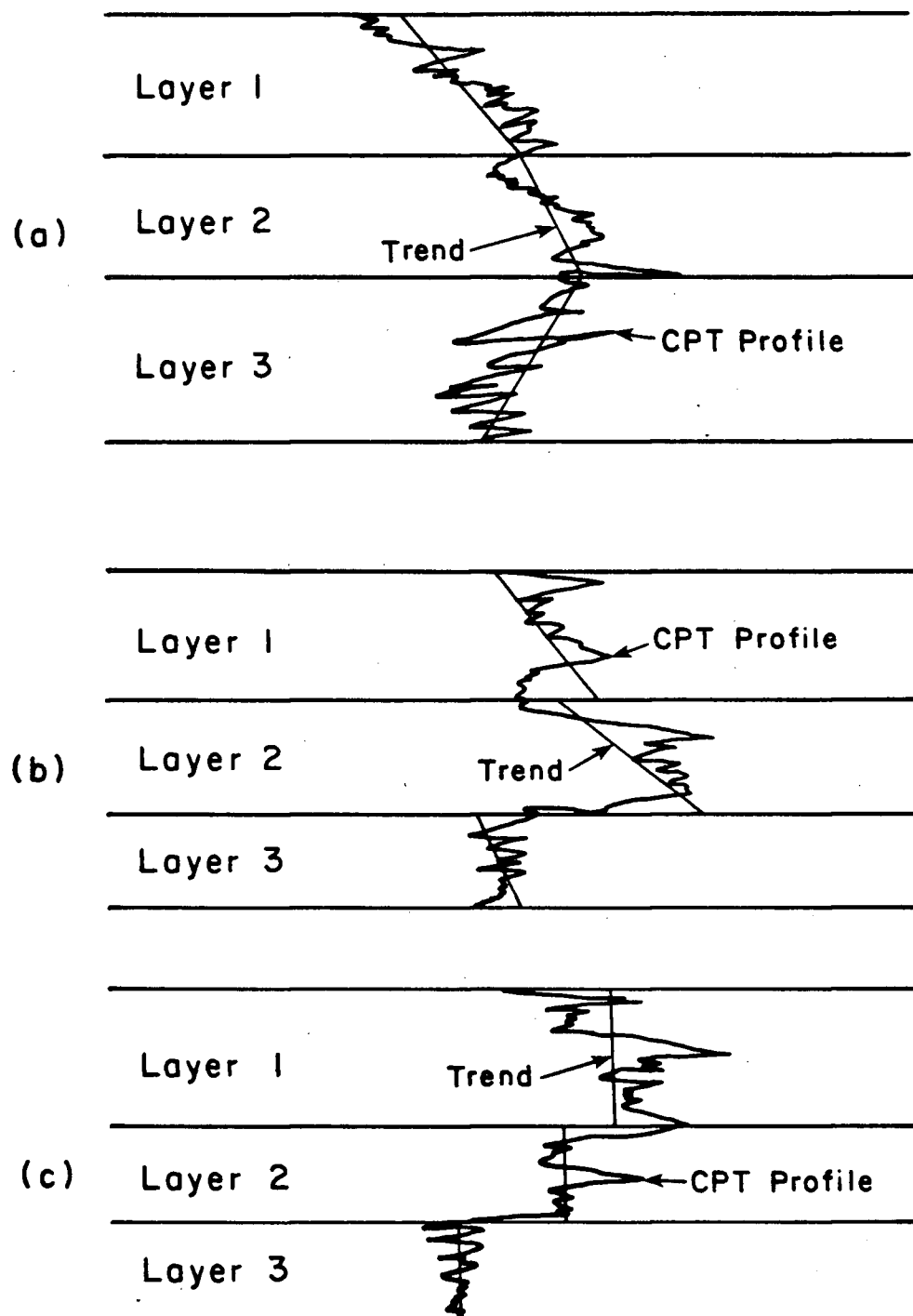
Figure 2.2: Different Types of Trend Patterns Used for Non Linear Optimization: (a) Continuous Trend Lines with Non Constant Mean (b) Discontinuous Trend Lines with Non Constant Mean (c) Discontinuous Trend Lines with Constant Mean.

described in section 2.3.

2.3 Layer Boundary Location Using Statistical Methods

Consider a section of a transect along which a property such as bearing has been recorded at a number of sampling points and within which the presence of a soil boundary can be expected. The effect of the boundary is to divide the sampling points into two groups which can be verified for distinctness. This effect of variability can be assessed by comparing the difference between the two classes. The larger the difference between the two classes and lesser the variation within them, the better is the classification. This effect can be measured using either the T Ratio or the Intraclass Correlation Coefficient which will be explained in sections 2.3.2.1 and 2.3.2.2 respectively. For multivariate records, the bearing, friction and pore pressure are used together to determine the D^2 statistic (section 2.3.3.1) which is used to obtain optimal boundary demarcations. These methods have never been used in geotechnical engineering or in a wider sense in Civil Engineering where statistical methods have not made significant inroads.

2.3.1 Moving Window

In analyzing long profiles where the presence of several boundaries are suspected it is not practical to consider the entire profile to investigate for individual boundaries. Similarly, it is also not advisable to bracket segments of data arbitrarily (Webster, 1973). In order to avoid the above impracticalities, a 'window' of fixed width (W_D) is made use of and the exposed portion of the data within the window is examined, with the center point of the window d_o, being a potential boundary. This 'window' is moved along the profile in steps equal to the sampling spacing and at each point d_o (the center of the window), the two sets of data one above and one below d_o

are examined for distinctness, using any one of the following statistics; the T Ratio or the Intraclass Correlation Coefficient for univariate data and the D^2 statistic for multivariate data. The variations of the above statistics are plotted against d_o, with the maxima or peaks of these giving the optimal layer boundaries. If only the layers with highly dissimilar characteristics are required, only those d_o values which have the highest values of the statistic need be chosen. However, if a more elaborate layer identification is necessary, even d_o values giving moderately high values should be selected.

The width of the window is another matter of concern and it should ideally lie between two limits. It should not be too wide, so that it includes more than one boundary and on the other hand, should not be too narrow because if it is so the values of the statistic will be strongly influenced by noise, rendering any interpretations of the calculated statistic almost impossible. Soulie (1984) states that an approximate estimation of the average layer thicknesses of a stratum can be obtained from the autocorrelation function which is defined and described in section 4.3.4. A conservative value of about fifty percent of the above estimate is recommended for the window width in order to alleviate the possibility of missing layers. However, if the spacing between boundaries differs significantly, it is advisable to use a fairly low window width, to avoid missing any layer boundaries. If relatively sharp changes are present between soil types and the distance between layers do not change too much, the autocorrelation function will decrease steadily with increasing lag distance, from a maximum of unity to a minimum value and fluctuate around this minimum. In practice the autocorrelation function first decreases gradually, and then fluctuates around some minimum, giving several local maxima and minima. The distance at which the first minimum is reached can be taken as the expected average distance between layers. As mentioned previously, half of this distance is a safe estimate for the

window width. If the window widths are narrow, the number of data on either side of d_o will be also low, resulting in the additional restraint that the data be multivariate normal (Johnson and Wichern - 1982). For a one meter window width, each sample will have twenty data points (0.5 m divided by the 0.025 m CPT data interval) and for all practical purposes, could be assumed as normally distributed, without serious adverse consequences. As a result of this restriction, it is recommended that window widths of less than 1.0 m not be selected. In the event it is absolutely necessary to select narrow window widths, the data should be verified for normality.

2.3.2 Univariate Records

The two statistics used to identify soil layers from single records are;

(a) T Ratio

(b) Intraclass Correlation Coefficient ρ_I

2.3.2.1 T Ratio

On either side of the window center d_o, there will be two samples, Ω_1 and Ω_2. Let $\bar{Q}_1$ and $\bar{Q}_2$ be the means of the samples and $\sigma_1{}^2$ and $\sigma_2{}^2$ be the variances with n_1 and n_2 the sample sizes, respectively;

where,

$$\sigma_1{}^2 = \frac{1}{(n_1 - 1)} \sum_{i=1}^{n_1} (Q_i - \bar{Q}_1)^2 \tag{2.2}$$

$$\sigma_2{}^2 = \frac{1}{(n_2 - 1)} \sum_{i=1}^{n_2} (Q_i - \bar{Q}_2)^2 \tag{2.3}$$

For two samples with equal number of data points, n, on either side of the window center, d_o, a pooled combined variance, $\Upsilon_w{}^2$, can be defined as,

$$\Upsilon_w{}^2 = \frac{n}{2n-1}\sigma_1{}^2 + \frac{n}{2n-1}\sigma_2{}^2 \qquad (2.4)$$

In the above equation $\sigma_1{}^2$ and $\sigma_2{}^2$ can be expected to be reasonably homogeneous, if the window widths are not too wide. The T Ratio can now be defined as,

$$T = \sqrt{\frac{n}{2}}\left(\frac{\bar{Q}_1 - \bar{Q}_2}{\Upsilon_w}\right) \qquad (2.5)$$

Equation 2.5 is a modified form of the one given by Webster(1968) which is a general expression for samples with unequal number of data. One requirement for the best possible differentiation of any two adjacent layers, is that the difference between the means $(\bar{Q}_1 - \bar{Q}_2)$, be maximum. If the two samples Ω_1 and Ω_2 are clearly distinct, another requirement is that the individual variances of the two segments, $\sigma_1{}^2$ and $\sigma_2{}^2$, be relatively low, implying that the weighted pooled variance given by Eq. 2.4, also be appreciably low. Considering the aforementioned requirements, the T ratio given by Eq. 2.5 will necessarily have to peak at potential layer boundaries. The T ratio, thus obtained for different values of d_o, gives an indication of the layer boundaries of the profile.

2.3.2.2   Intraclass Correlation Coefficient (ρ_I)

As for the previous case, let σ_1^2 and σ_2^2 be the variances of samples Ω_1 and Ω_2 and the pooled combined variance $\Upsilon_w{}^2$, given by Eq. 2.4. The between class variance $\Upsilon_b{}^2$ is the variance of the combined sample given by,

$$\Upsilon_b{}^2 = \frac{1}{n_1 + n_2 - 1} \sum_{i=1}^{n_1 + n_2} (Q_i - \bar{Q})^2 \qquad (2.6)$$

where, $\bar{Q}$ is the mean of all the data Q_i with $i = 1, 2 \ldots (n_1 + n_2)$.

For equal number of data in each sample,

$$\Upsilon_b{}^2 = \frac{1}{2n - 1} \sum_{i=1}^{2n} (Q_i - \bar{Q})^2 \qquad (2.7)$$

The Intraclass Correlation Coefficient ρ_I is defined by (Webster, 1968),

$$\rho_I = \frac{\Upsilon_b{}^2}{\Upsilon_b{}^2 + \Upsilon_w{}^2} \qquad (2.8)$$

It is evident that if each sample Ω_1 and Ω_2 has minimum variability, $\sigma_1{}^2$ and $\sigma_2{}^2$ in Eq. 2.4 will both approach zero and so will $\Upsilon_w{}^2$. In addition, if the difference between the samples is not significant, $\Upsilon_b{}^2$ in Eq. 2.7 will also approach zero. Since $\Upsilon_w{}^2$ and $\Upsilon_b{}^2$ are both positive quantities, $\Upsilon_b{}^2$ in Eq. 2.8 will approach zero faster than the quantity $(\Upsilon_b{}^2 + \Upsilon_w{}^2)$. Therefore, for two such samples, Ω_1 and Ω_2 on either side of d_o, ρ_I will approach zero. This is for the extreme case and in general, if the differences between the samples are not significant and they possess some variability, then $\Upsilon_b{}^2$ and ρ_I are not significantly greater than zero. The other scenario is when the two

samples Ω_1 and Ω_2 have minimum variability but are significantly different in respect to their mean values. In this circumstance, $\Gamma_w{}^2$ will approach zero as before, while $\Gamma_b{}^2$ will have some value, resulting in ρ_I being equal to unity. ρ_I will therefore always lie between these two hypothetical extremes of zero and unity. In reality, a relatively high value of ρ_I at a particular depth d_o will indicate the presence of a layer boundary at that point. As with the T Ratio, the value of ρ_I can be plotted against depth, in order to determine the best layer boundaries along the profile.

Several applications of using both the T Ratio and the Intraclass Correlation Coefficient will be illustrated later.

2.3.3 Multivariate Records

The CPT performed at UBC performs data logging on several channels, the cone bearing, sleeve friction and the pore pressure being the most important of these from an engineering point of view. All these parameters exhibit a different kind of behavior in different types of soils, and therefore, any method which considers the combined effects of cone bearing, sleeve friction and pore pressure together in one analysis, will definitely be the more efficient and accurate method due to the additional information contained in such an analysis. While the T Ratio and the Intraclass Correlation Coefficient contain the variance (second moment) and the mean (first moment) of the data the D^2 statistic also includes the covariances of the different variables in addition to the mean and the variance. In contrast to these statistical methods the conventional method of layer identification using the CPT chart entails only the mean values of the parameters. In this regard it is obvious that the statistical methods should be better with the D^2 statistic being the most superior.

2.3.3.1 D^2 Statistic

The D^2 statistic gives most weight to those variates that discriminate best between segments. Problems arise if there are several variates, in comparison to the number of data (Rao - 1952). However, in the problems dealt with here, this is not of major concern even for the case of narrow window widths, since the number of variates do not exceed three.

The use of the discriminant function may be considered in terms of a sample Ω_1 consisting of m variates, which form a cluster of points in m - dimensional space. Another sample Ω_2 may be described similarly by the same m variables in m - dimensional space. The determination of a (m - 1) - dimensional plane that separates the two clusters of points is the discriminant function (Harbaugh and Merriam, 1968). The D^2 is the distance between the multivariate means of the two dimensional sample spaces Ω_1 and Ω_2, implying that greater the value of D^2, the more distinct the two samples would be (Rao - 1965). This is illustrated in Fig. 2.3 for the case of two variables, α_1 and α_2 (m = 2).

The D^2 statistic is given by,

$$D^2 = \{\bar{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_2\}^{\mathbf{T}}[\mathbf{W}]^{-1}\{\bar{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_2\} \qquad (2.9)$$

where, $\{\bar{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_2\}$ is the column matrix of the mean differences of the variates in the two samples. For the case with m variables, $\{\bar{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_2\}$ is a m x 1 matrix. $[\mathbf{W}]$ is the pooled variance - covariance matrix of the samples Ω_1 and Ω_2. For layer identification purposes using the cone, the maximum number of variates (m) will be equal to three. Let the set of n_1 data points from Ω_1 and n_2 data points from Ω_2 be described by the following variables;
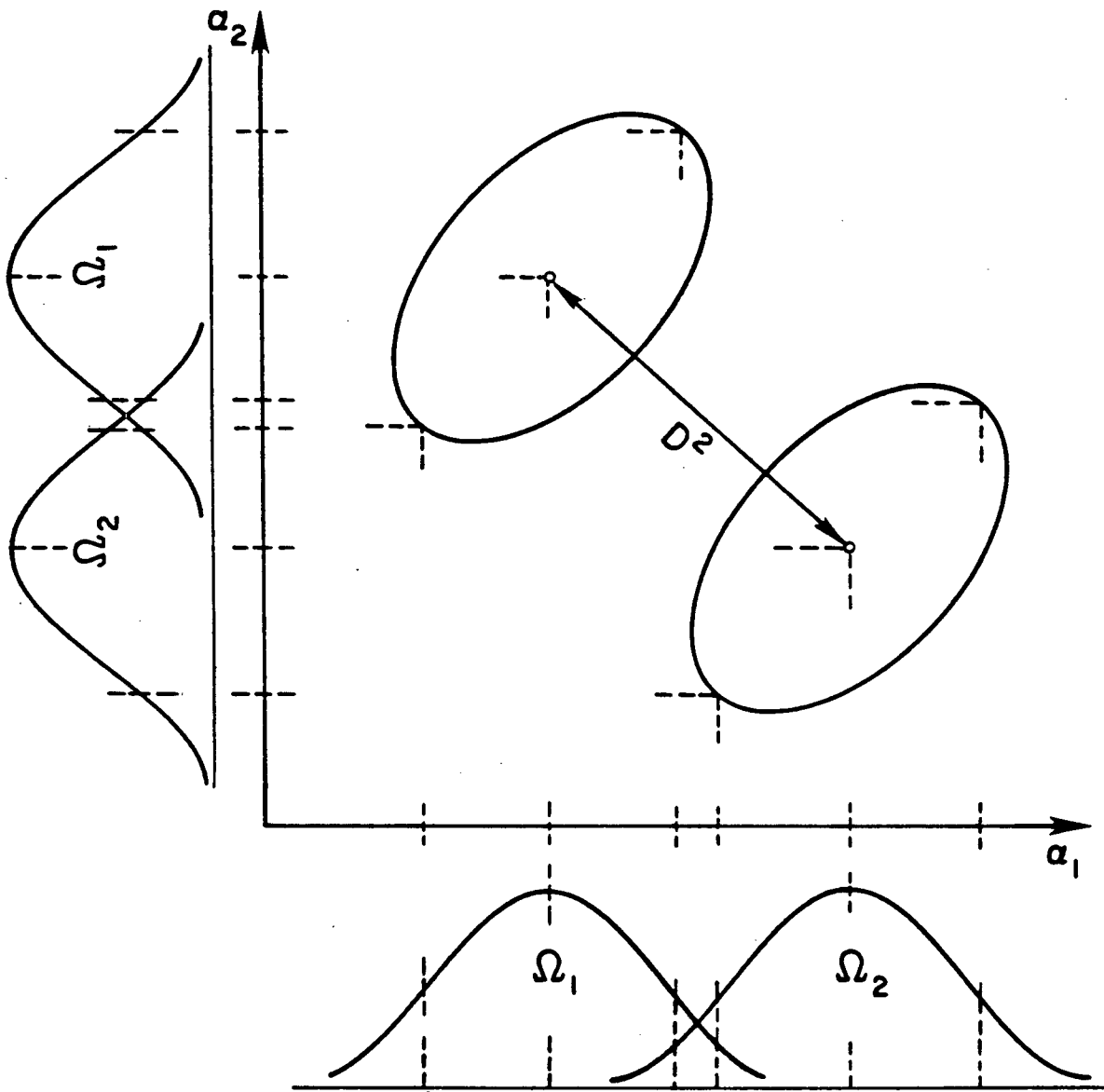
Figure 2.3: D^2 Statistic for Two Samples Ω_1 and Ω_2 with Two Variates α_1 and α_2 (after Harbaugh and Merriam, 1968).

q_1, f_1, u_1 being the bearing, friction and pore pressure in Ω_1

q_2, f_2, u_2 being the bearing, friction and pore pressure in Ω_2

The means of the respective parameters in sample Ω_1 are given by, $\bar{q}_1$, $\bar{f}_1$ and $\bar{u}_1$ and for sample Ω_2, by $\bar{q}_2$, $\bar{f}_2$ and $\bar{u}_2$, and their variances by, $\sigma_{q_1}{}^2$, $\sigma_{f_1}{}^2$, $\sigma_{u_1}{}^2$, $\sigma_{q_2}{}^2$, $\sigma_{f_2}{}^2$ and $\sigma_{u_2}{}^2$.

The mean differences of the variates of the two samples are given by,

$$\Delta q = \bar{q}_1 - \bar{q}_2 \qquad \Delta f = \bar{f}_1 - \bar{f}_2 \qquad \Delta u = \bar{u}_1 - \bar{u}_2 \qquad (2.10)$$

The covariances are given by,

$$\sigma_{q_1 f_1}{}^2 = \frac{\sum_{i=1}^{n_1} q_{1i} f_{1i}}{n_1} - \frac{\sum_{i=1}^{n_1} q_{1i} \sum_{i=1}^{n_1} f_{1i}}{n_1{}^2} \qquad (2.11)$$

$$\sigma_{q_1 u_1}{}^2 = \frac{\sum_{i=1}^{n_1} q_{1i} u_{1i}}{n_1} - \frac{\sum_{i=1}^{n_1} q_{1i} \sum_{i=1}^{n_1} u_{1i}}{n_1{}^2} \qquad (2.12)$$

$$\sigma_{f_1 u_1}{}^2 = \frac{\sum_{i=1}^{n_1} f_{1i} u_{1i}}{n_1} - \frac{\sum_{i=1}^{n_1} f_{1i} \sum_{i=1}^{n_1} u_{1i}}{n_1{}^2} \qquad (2.13)$$

$$\sigma_{q_2 f_2}{}^2 = \frac{\sum_{i=1}^{n_2} q_{2i} f_{2i}}{n_2} - \frac{\sum_{i=1}^{n_2} q_{2i} \sum_{i=1}^{n_2} f_{2i}}{n_2{}^2} \qquad (2.14)$$

$$\sigma_{q_2 u_2}{}^2 = \frac{\sum_{i=1}^{n_2} q_{2i} u_{2i}}{n_2} - \frac{\sum_{i=1}^{n_2} q_{2i} \sum_{i=1}^{n_2} u_{2i}}{n_2{}^2} \qquad (2.15)$$

$$\sigma_{f_2 u_2}{}^2 = \frac{\sum_{i=1}^{n_2} f_{2i} u_{2i}}{n_2} - \frac{\sum_{i=1}^{n_2} f_{2i} \sum_{i=1}^{n_2} u_{2i}}{n_2{}^2} \qquad (2.16)$$

The pooled weighted variances are given by,

$$\Gamma_q{}^2 = \frac{n_1}{n_1 + n_2 - 1} \sigma_{q_1}{}^2 + \frac{n_2}{n_1 + n_2 - 1} \sigma_{q_2}{}^2 \qquad (2.17)$$

$$\Gamma_f{}^2 = \frac{n_1}{n_1 + n_2 - 1}\sigma_{f_1}{}^2 + \frac{n_2}{n_1 + n_2 - 1}\sigma_{f_2}{}^2 \tag{2.18}$$

$$\Gamma_u{}^2 = \frac{n_1}{n_1 + n_2 - 1}\sigma_{u_1}{}^2 + \frac{n_2}{n_1 + n_2 - 1}\sigma_{u_2}{}^2 \tag{2.19}$$

Similarly, the pooled weighted covariances are given by,

$$\Gamma_{qf}{}^2 = \frac{n_1}{n_1 + n_2 - 1}\sigma_{q_1 f_1}{}^2 + \frac{n_2}{n_1 + n_2 - 1}\sigma_{q_2 f_2}{}^2 \tag{2.20}$$

$$\Gamma_{qu}{}^2 = \frac{n_1}{n_1 + n_2 - 1}\sigma_{q_1 u_1}{}^2 + \frac{n_2}{n_1 + n_2 - 1}\sigma_{q_2 u_2}{}^2 \tag{2.21}$$

$$\Gamma_{fu}{}^2 = \frac{n_1}{n_1 + n_2 - 1}\sigma_{f_1 u_1}{}^2 + \frac{n_2}{n_1 + n_2 - 1}\sigma_{f_2 u_2}{}^2 \tag{2.22}$$

If equal number of data points are considered in samples Ω_1 and Ω_2 (as usually the case is), $n_1/(n_1 + n_2 - 1)$ and $n_2/(n_1 + n_2 - 1)$ in Eqs. 2.17 to 2.22 can be approximated by,

$$\frac{n_1}{n_1 + n_2 - 1} = \frac{n_2}{n_1 + n_2 - 1} = 0.5 \tag{2.23}$$

The variance covariance matrix $[\mathbf{W}]$ can be now formulated and is comprised of the elements derived above,

$$[\mathbf{W}] = \begin{bmatrix} \Gamma_q{}^2 & \Gamma_{qf}{}^2 & \Gamma_{qu}{}^2 \\ \Gamma_{qf}{}^2 & \Gamma_f{}^2 & \Gamma_{fu}{}^2 \\ \Gamma_{qu}{}^2 & \Gamma_{fu}{}^2 & \Gamma_u{}^2 \end{bmatrix} \tag{2.24}$$

$\{\bar{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_2\}$ in Eq. 2.9 is given by,

$$\{\bar{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_2\} = \left\{ \begin{array}{c} \Delta q \\ \Delta f \\ \Delta u \end{array} \right\} \quad (2.25)$$

In using the D^2 statistic, to identify soil layer boundaries, the window is moved along the data profile, with d_o the mid-point of the window separating the two samples and for each d_o, the value of D^2 is calculated and plotted against depth. The peaks of the ensuing plot would illustrate the best positions of the layer boundaries. If only a few boundaries are needed, the points at which the highest D^2 values occur can be selected. If in the engineer's mind more layers are needed, the less critical D^2 values too can be used in order to obtain more layer demarcations.

2.4 Application to CPT Profiles

The above concepts of statistically identifying layers have been applied to three sets of data in order to illustrate the advantages of the methods explained above. The locations from which the data have been obtained are given below and their geology and site descriptions are given in section 1.7.

(a) McDonald Farm Site

(b) Haney Site

(c) Tilbury Island Site

All the data have been obtained using a cone of sectional area $10cm^2$ with penetration at 2 *cm/sec* and data logging being performed at 2.5 cm intervals. The McDonald Farm site is predominantly sand, the Haney site predominantly clay and the Tilbury Island site is mainly silt and sand. These particular sites were selected as

they cover a wide area of soil types which could be encountered in a site investigation and in the event the statistical methods prove successful in all three sites, it becomes possible to infer that the proposed methods are applicable to any type of soil profile.

2.4.1 McDonald Farm Site

The McDonald Farm, typically consists of sand and sandy silts in the top 15 m with clayey soils extending below the sand. The cone bearing, sleeve friction, pore pressure and friction ratio profiles are illustrated in Fig. 2.4.

At the outset, an autocorrelation analysis was performed for the three variables and the variation of the function with lag distance (separation distance between points) is illustrated in Fig. 2.5. The purpose of this was to determine an optimal window width, W_D, which ideally should lie between two limits: not too wide, in order to avoid the possibility of missing thinner layers and not too narrow, in order to minimize noise in the calculated statistics. The plot of the autocorrelation function in Fig. 2.5 results in three different initial minimum points for the three variables, indicated by the arrows which read as 6.82 m for cone bearing, 2.74 m for friction and 5.22 for pore pressure. The multivariate analysis requires a single value for W_D since all three variables are handled simultaneously, while for the univariate analysis, three different widths can be used for the three variables. However, it is suggested to decide on a single W_D even for the univariate case to facilitate comparisons between variates. The more serious consequence of choosing an incorrect W_D is that if it is too wide, potential layers will be missed. The consequence of missing layers is highly undesirable and has to be avoided. As explained earlier, the upper limit of the W_D will have to be below the minimum value of 2.74 m, and preferably about half of 2.74 in order to avoid any possibility of missing any dominant layers.

An initial value of 0.5 m was selected for W_D to illustrate the effects of noise,
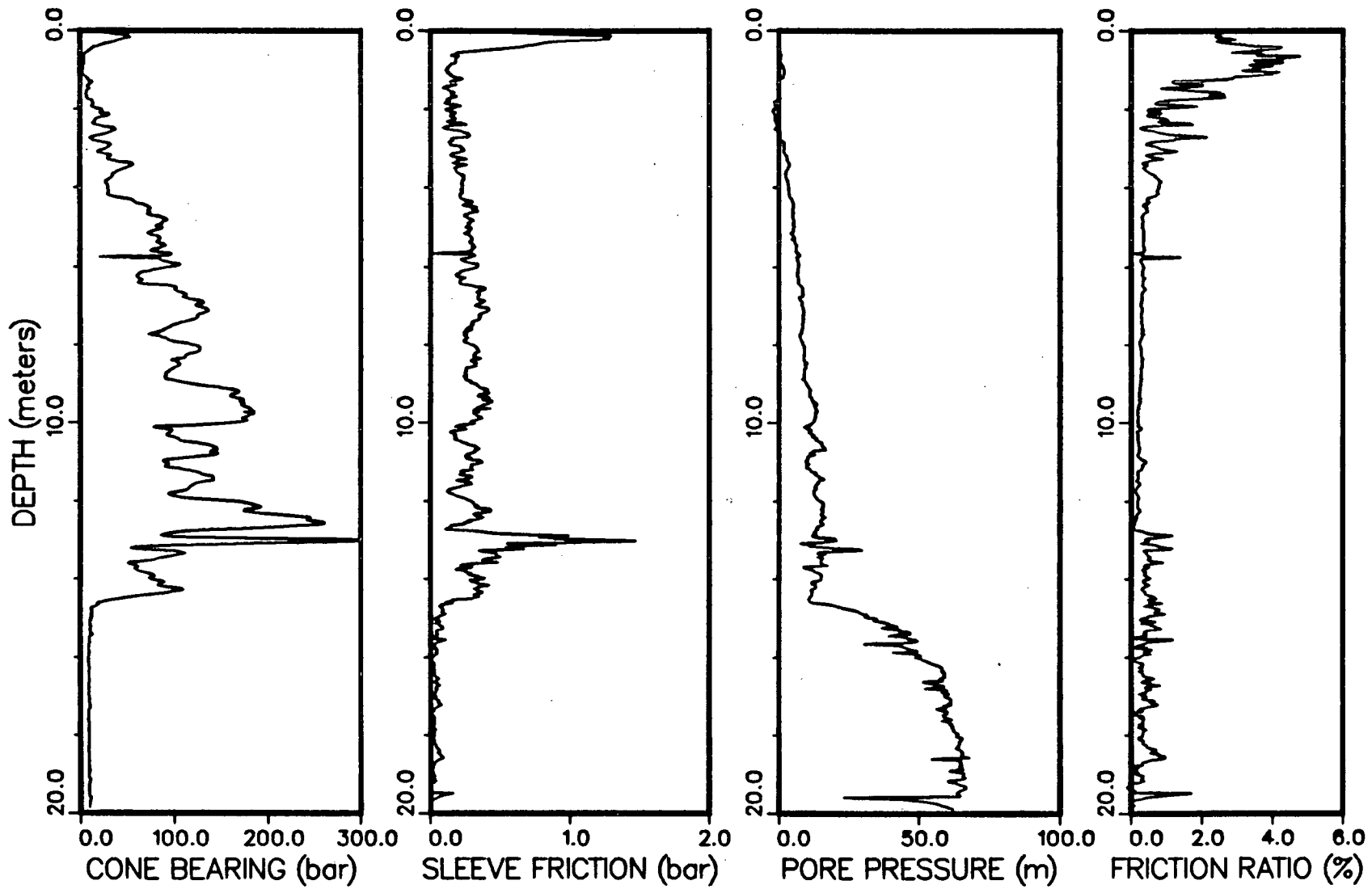
Figure 2.4: Cone Bearing, Sleeve Friction, Pore Pressure and Friction Ratio Profiles at McDonald Farm.
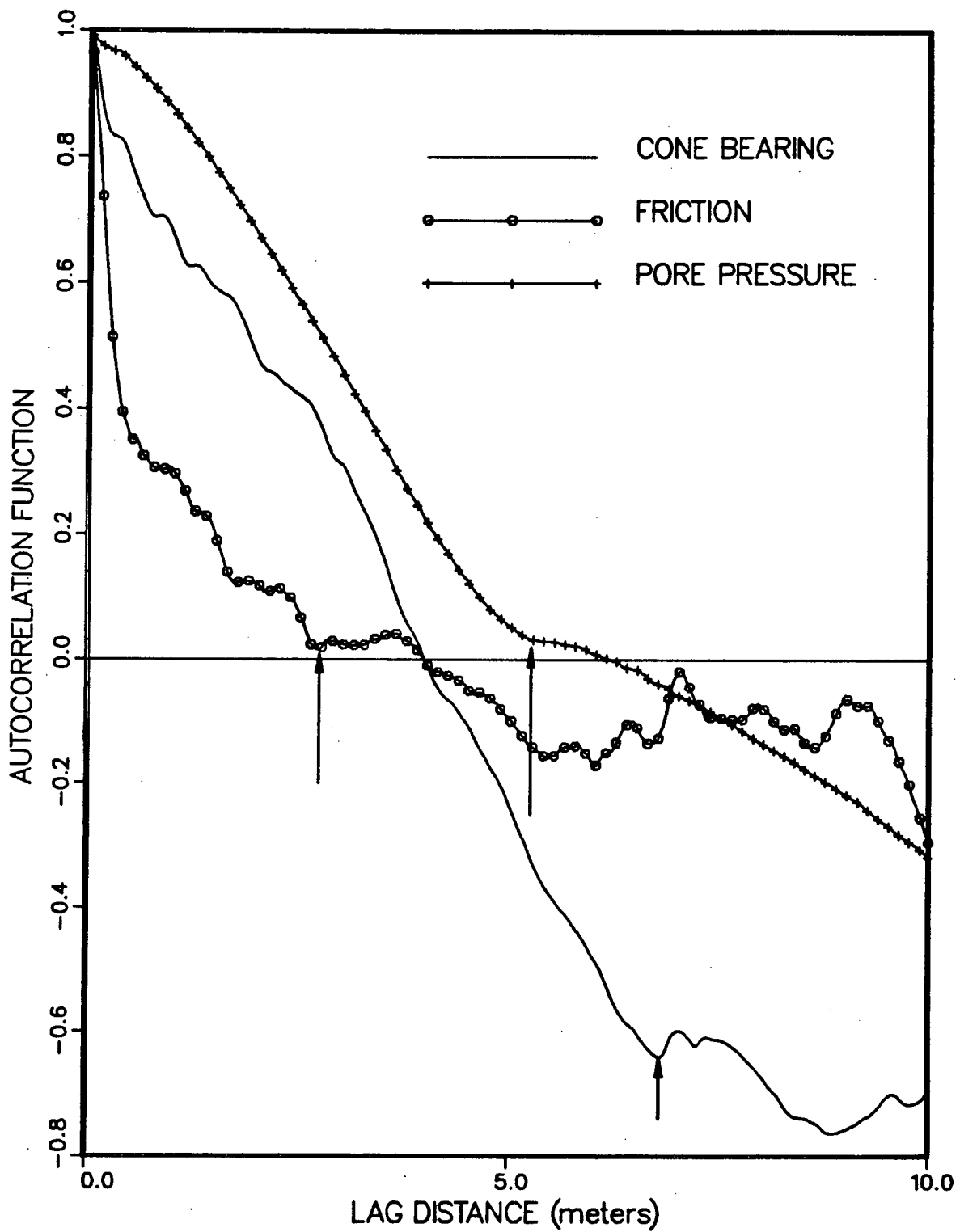
Figure 2.5: Autocorrelation Functions of Cone Bearing, Sleeve Friction and Pore Pressure Profiles at McDonald Farm.

when the thickness is low. A W_D of 3.0 m was also used to illustrate the more crucial effect of missing layers with wide window widths. These effects will be discussed later with the appropriate figures. As a consequence of further investigations, the most adequate W_D was selected as 1.5 m and all the detailed analyses and comparisons to follow will deal with this window width.

The variations of the Intraclass Correlation Coefficient (ρ_I) with depth for bearing, friction and pore pressure are illustrated in Figs. 2.6, 2.7 and 2.8. The variation of the T ratio for the three properties are given in Figs. 2.9 to 2.11. The results for the case with W_D of 3.0 m, for both T Ratio and ρ_I are illustrated in Figs. 2.12 and 2.13 respectively. By comparing Figs. 2.6 to 2.11 for $W_D = 1.5$ m to the corresponding Figs. 2.12 and 2.13 for $W_D = 3.0$ m, it is evident that a W_D of 1.5 m is superior due to the apparent convenience in which layer boundaries can be picked up without the risk of missing out thinner layers. Figures 2.14 and 2.15 illustrate the variation of bearing, friction and pore pressure overlaid on one another for the T Ratio and ρ_I respectively, for a W_D of 1.5 m.

The following depths have been obtained as the most critical layer depths, considering the T Ratio for cone bearing, sleeve friction and pore pressure in Fig. 2.14.

Cone Bearing : 0.65, 0.93, [4.35], 6.60, [9.05], 10.03, 11.93, 12.80, [14.50] m.

Friction : 0.65, [4.33], 6.53, 7.35, [9.05], 9.90, [14.53], 17.70 m.

Pore Pressure : 3.13, [4.30], 5.43, 6.48, 8.03, [9.10], 11.98, [14.78], 17.80 m.

The bracketed values indicate the depths at which the statistic attains a high magnitude for at least two variates.

For the Intraclass Correlation Coefficient (Fig. 2.15), the results are as follows;

Cone Bearing : 1.25, [4.35], 6.60, [9.05], 10.03, 11.98, 12.80, [14.50], 15.48 m.
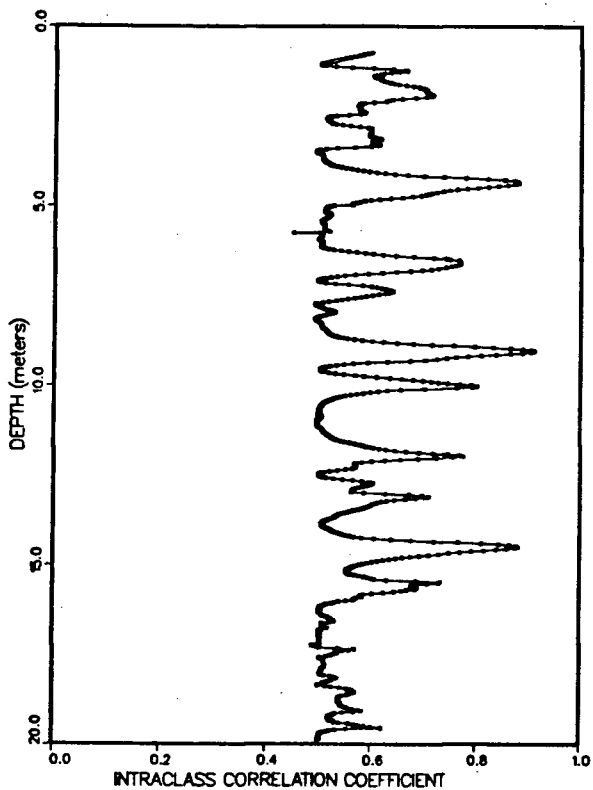
Figure 2.6: Intraclass Correlation Coefficient for Cone Bearing at Mc-Donald Farm for a Window Width of 1.5 meters.
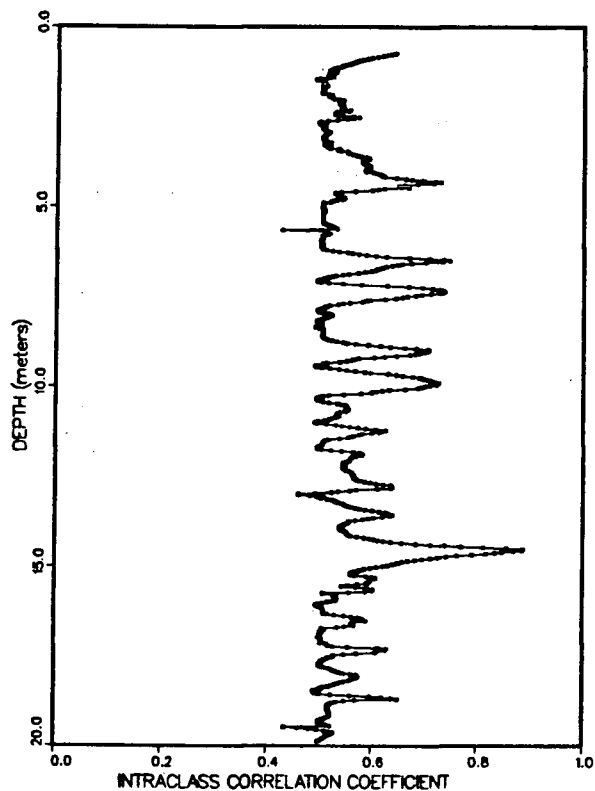
Figure 2.7: Intraclass Correlation Coefficient for Sleeve Friction at McDonald Farm for a Window Width of 1.5 meters.
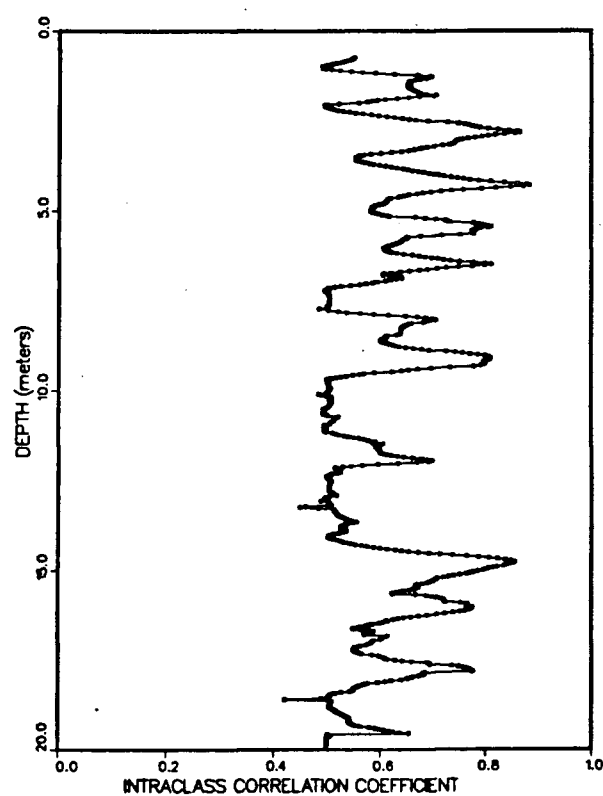
Figure 2.8: Intraclass Correlation Coefficient for Pore Pressure at Mc-Donald Farm for a Window Width of 1.5 meters.
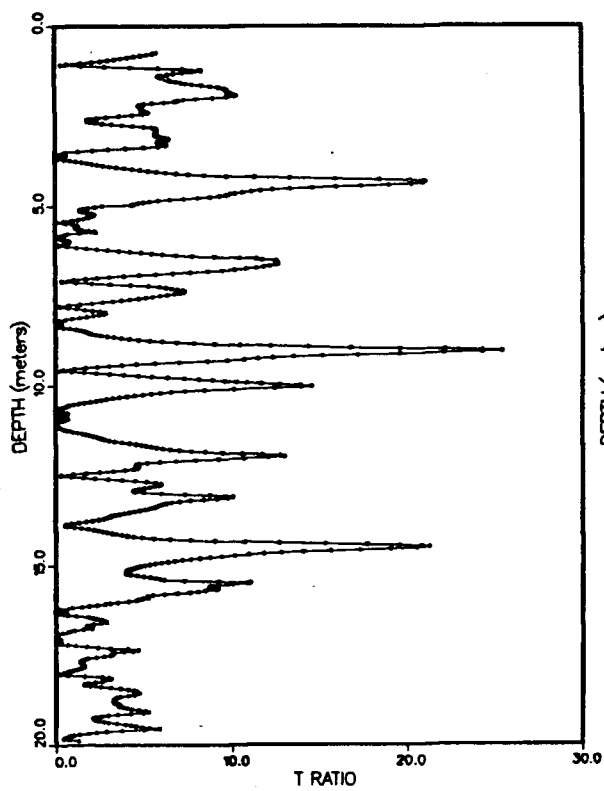
Figure 2.9: T Ratio for Cone Bearing at McDonald Farm for a Window Width of 1.5 meters.
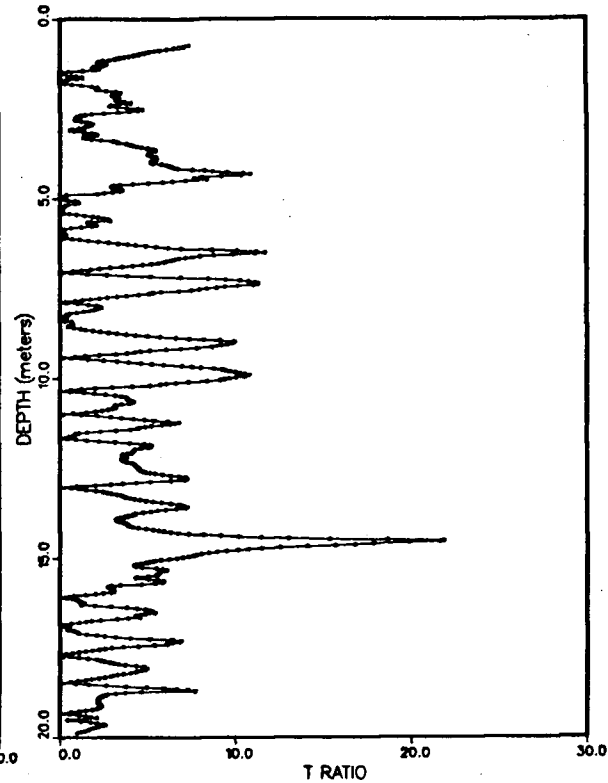
Figure 2.10: T Ratio for Sleeve Friction at McDonald Farm for a Window Width of 1.5 meters.
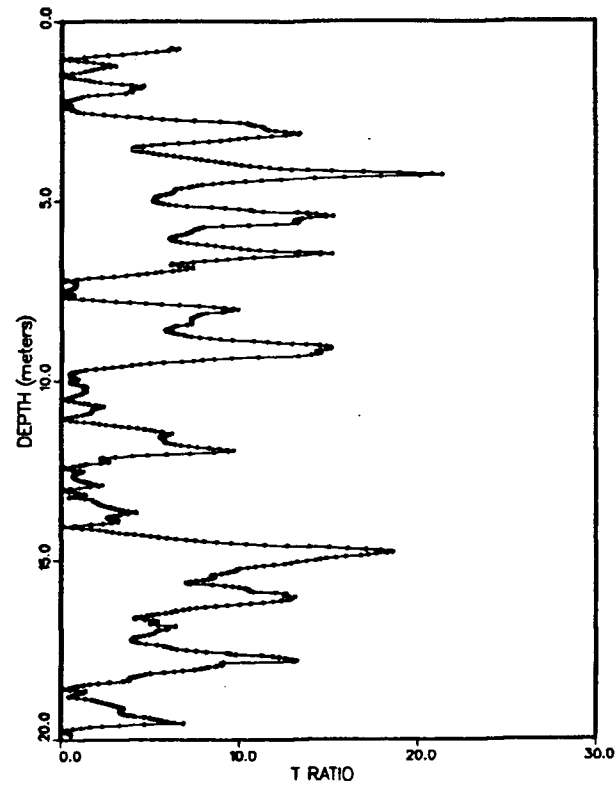
Figure 2.11: T Ratio for Pore Pressure at McDonald Farm for a Window Width of 1.5 meters.
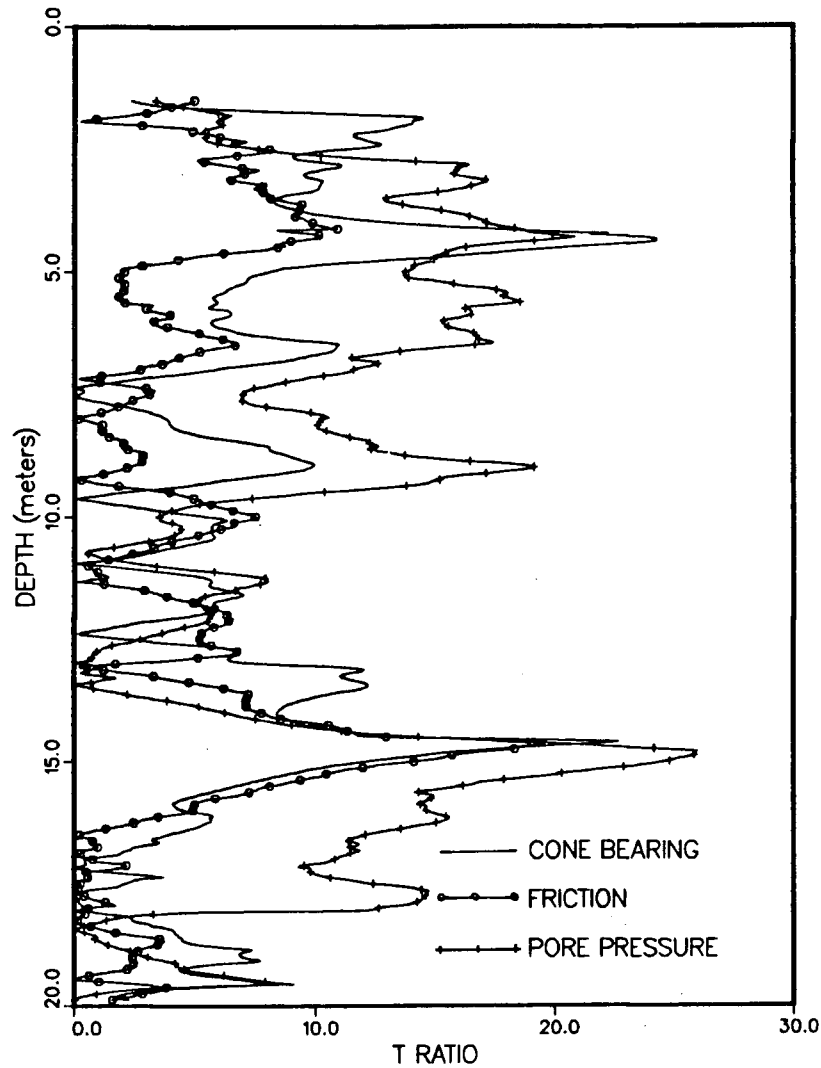
Figure 2.12: T Ratio for Cone Bearing, Friction and Pore Pressure at McDonald Farm for a Window Width of 3 meters.
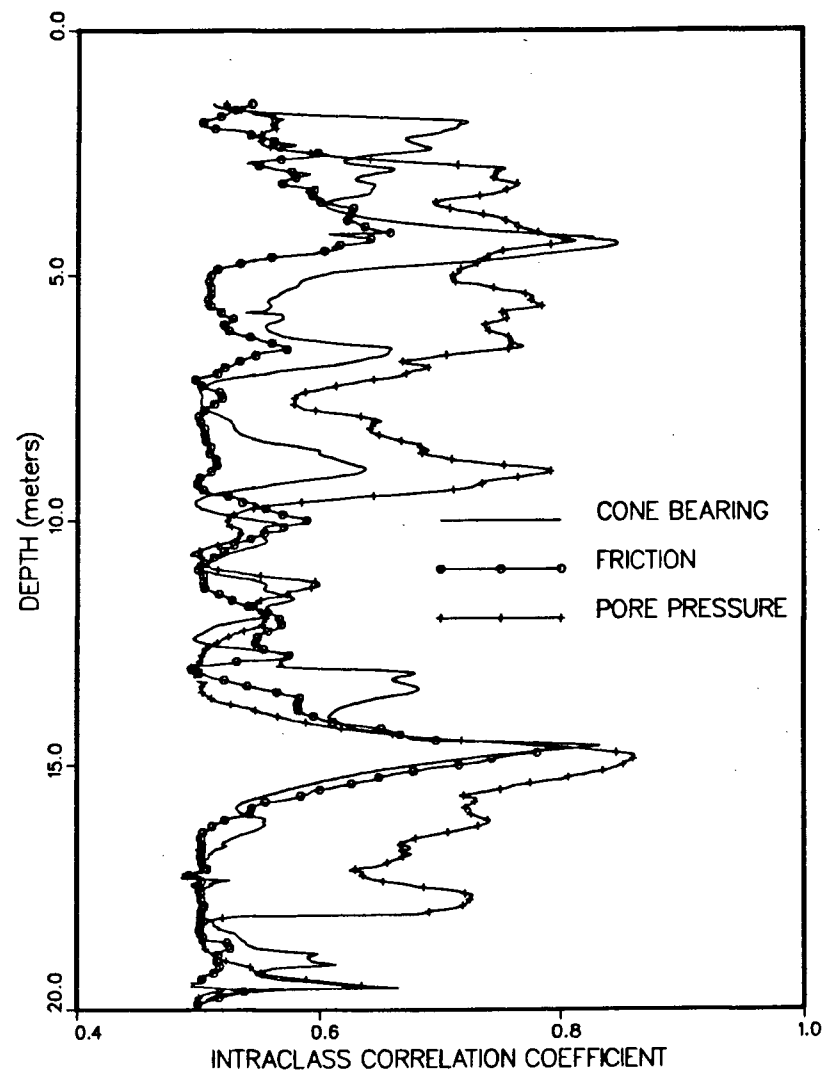
Figure 2.13: Intraclass Correlation Coefficient for Cone Bearing, Friction and Pore Pressure at McDonald Farm for a Wndow Width of 3 meters.
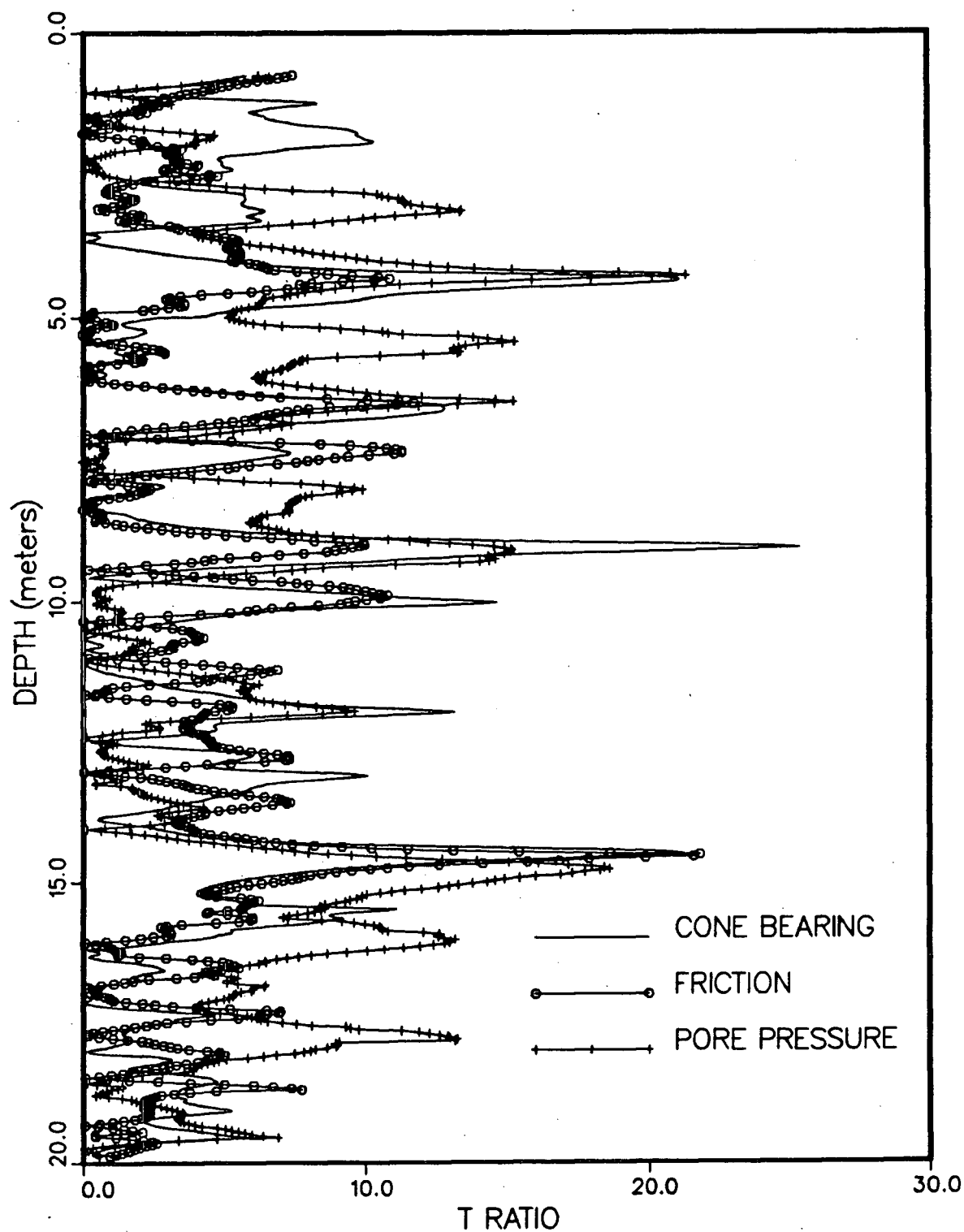
Figure 2.14: T Ratio for Cone Bearing, Friction and Pore Pressure at McDonald Farm for a Window Width of 1.5 meters.
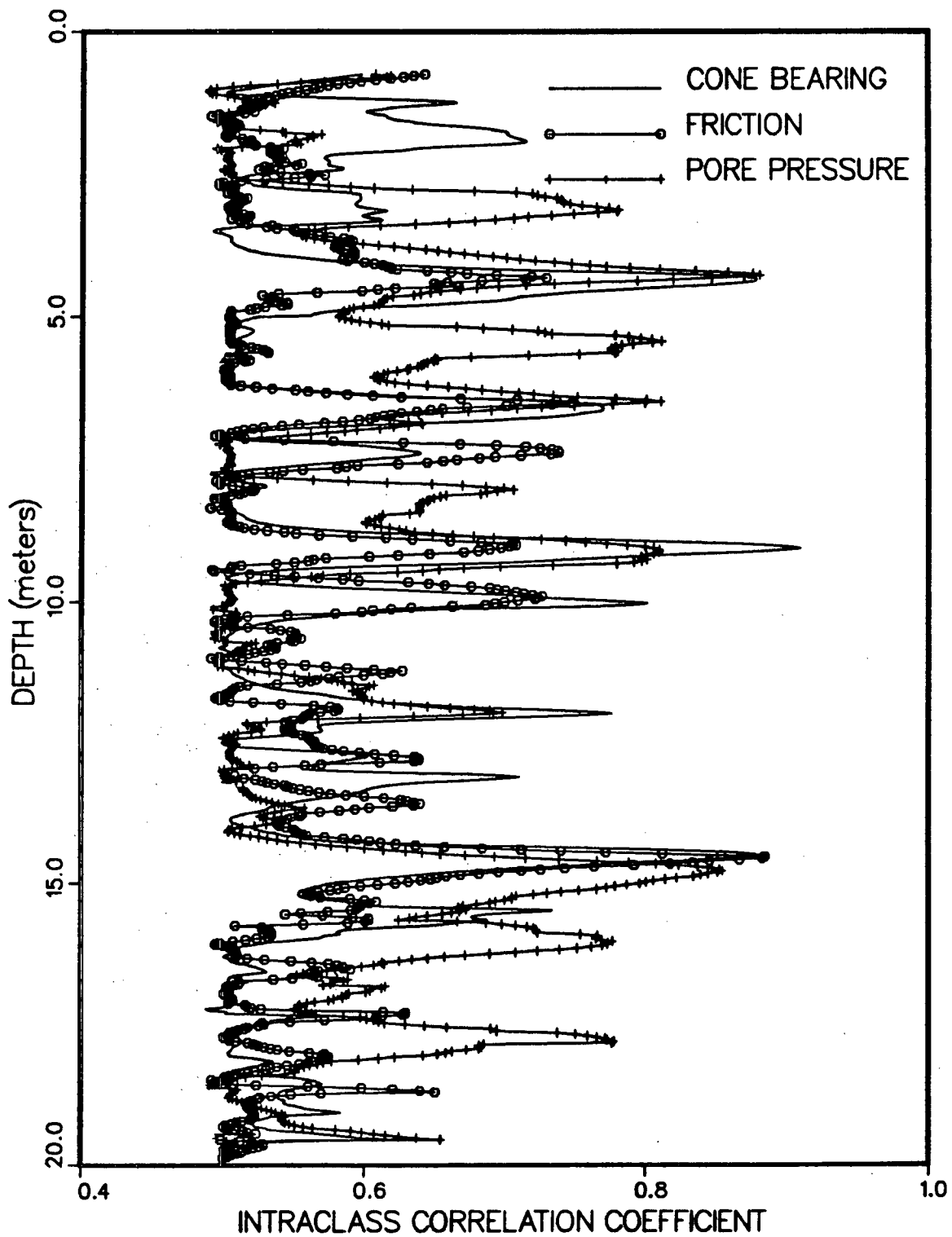
Figure 2.15: Intraclass Correlation Coefficient for Cone Bearing, Friction and Pore Pressure at McDonald Farm for a Window Width of 1.5 meters.

Friction : 0.63, [4.33], 6.50, 7.38, [8.98], 9.90, [14.53], 17.70 m.

Pore Pressure: 0.63, [4.28], 5.43, 6.48, [9.10], 11.98, [14.68], 16.03, 17.80 m.

Here too, the bracketed values indicate the depths at which the highest values of ρ_I occur for at least two variates. These results agree appreciably well with the depths obtained using the T Ratio. The layer boundaries that could be selected from the results of the Intraclass Correlation Coefficient are 4.30, 9.05 and 14.60 m and agree well with those obtained for the T ratio. In situations where the T Ratio and ρ_I of the variables considered are not in total agreement, the decision of layer boundaries will have to be based on the results of the multivariate analysis. From the above results, it can be concluded that the two statistics, the Intraclass Correlation Coefficient and the T Ratio are appropriate statistics for discriminating between layers.

A multivariate analysis was performed for the three variables and the D^2 was calculated. The variation of D^2 is illustrated in Fig. 2.16 and the most prominent peaks for layer differentiation was obtained as 4.30, 9.10, 14.60 m. Another possible but less dominant layer boundary can be found at 10.0 m. These results agree with the ones obtained for the univariate analysis except for the 12.85 m depth where only the cone bearing suggested a layer boundary for the univariate analysis. The D^2 statistic for W_d value of 0.5, illustrated in Fig. 2.17, shows the effect of noise.

It is widely accepted that the friction ratio in conjunction with the cone bearing, is a reliable method for identifying different layers in a soil stratum. If the CPT classification chart was used for layer boundary location, the layer boundaries would have been identified at 1.0, 4.25, 13.0 and 14.75 m. Furthermore, these estimates of layering are based on a fair amount of judgement, leaving room for subjectivity and inconsistency. Table 2.1 gives a comparison of the layer boundaries identified by the statistical methods with that based on the CPT classification chart. The statistical
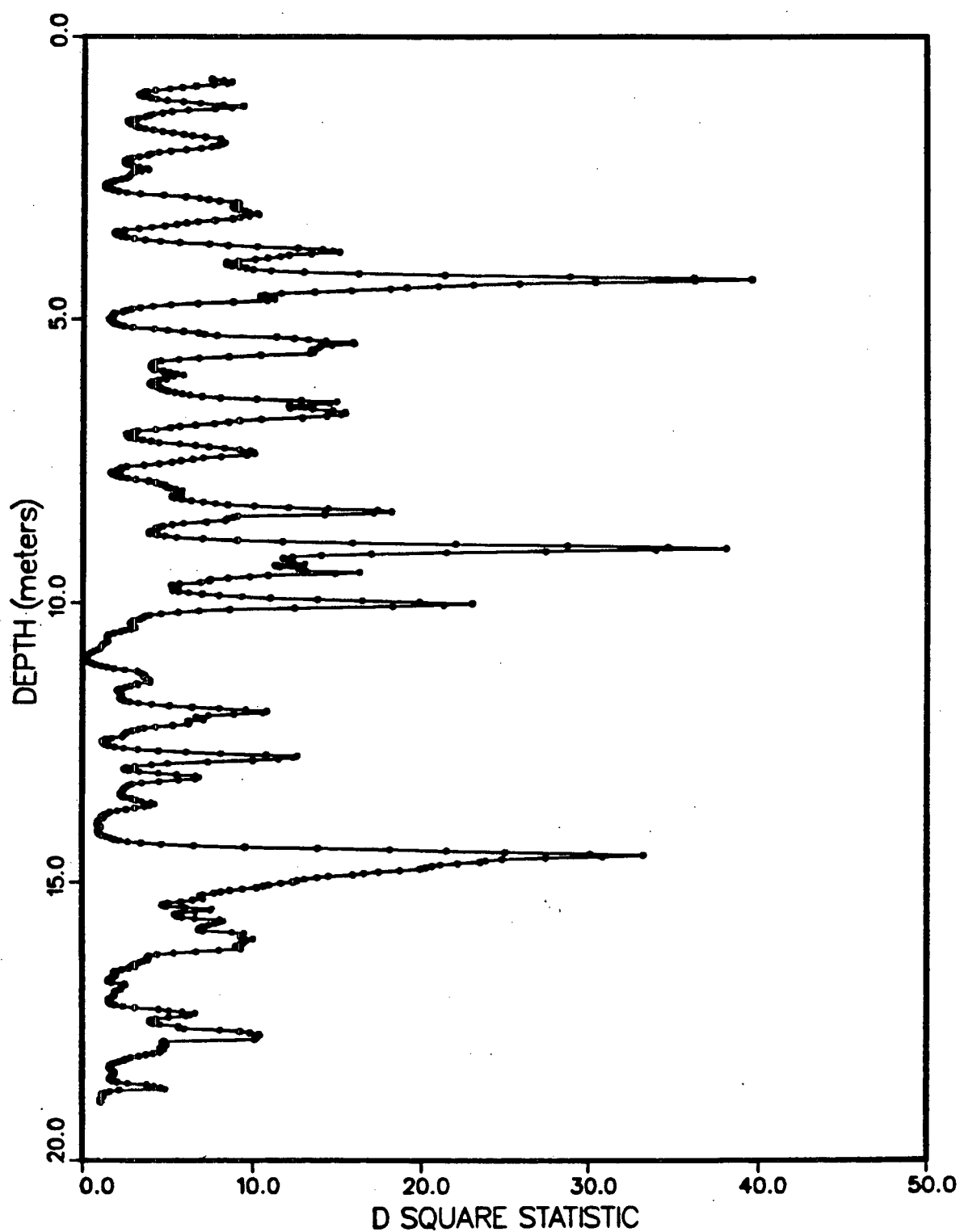
Figure 2.16: D^2 Statistic from Multivariate Analysis at McDonald Farm for a Window Width of 1.5 meters.
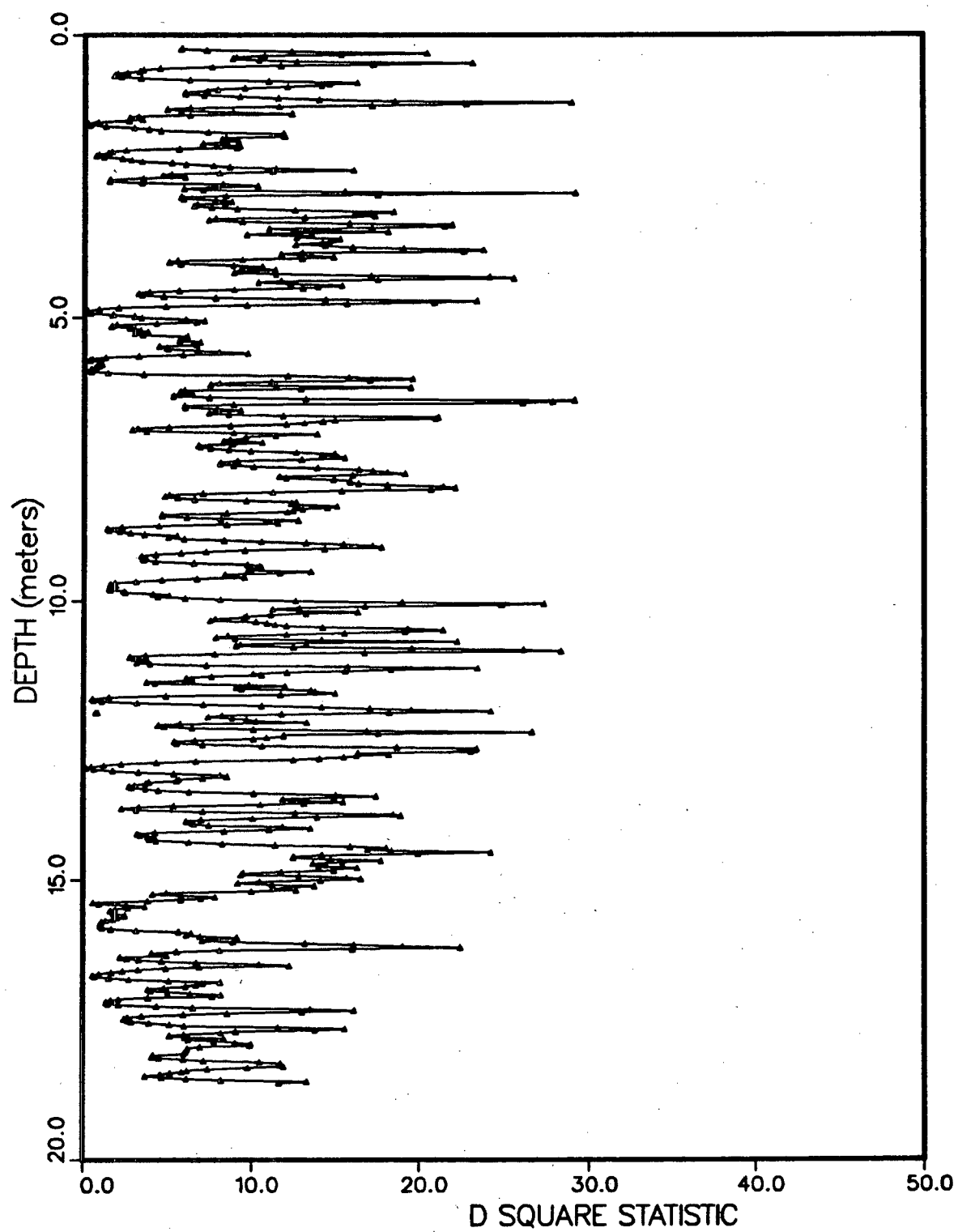
Figure 2.17: D^2 Statistic from Multivariate Analysis at McDonald Farm for a Window Width of 0.5 meters.

techniques proposed can be used to identify two types of layers as follows. The primary layer boundaries define depths with relatively large values of the T Ratio, ρ_I and D^2 while the secondary layer boundaries define depths which have lesser magnitudes of the above statistics. The significance of these magnitudes will be discussed in a subsequent section. As can be seen from Table 2.1, the conventional method has failed to detect the layer boundaries at 6.50, 9.05 and 17.80 m. In addition, the proposed statistical methods were also able to supplement information on layering by assigning exact depths which are generally estimated by eye if the method based on the CPT chart is used. The statistical methods of layering are based on some specific numerical values, alleviating the possibility of erroneous classification due to misjudgement. Once the pattern of detailed layering is recognized, correlation graphs can be used with less uncertainty to obtain values of friction angle, relative density, etc., for the different sublayers. The detailed layering pattern will also be useful for engineering design, since the engineer is informed of the different layers in existence, so that averaging of properties can be done for the statistically homogeneous layers determined from above.

2.4.2 Haney Site

The measured cone bearing, sleeve friction, pore pressure and friction ratio profiles of the Haney site are illustrated in Fig. 2.18. The soil conditions at Haney is predominantly clay alternating between clayey silt and silty clay. As for the previous example an autocorrelation analysis was performed on the three variables and the expected layer thicknesses were not similar. The lowest value was obtained for friction with a value of approximately 3.8 m. Therefore, a window width W_D, of 2.0 m was selected to avoid the possibility of missing layer boundaries.
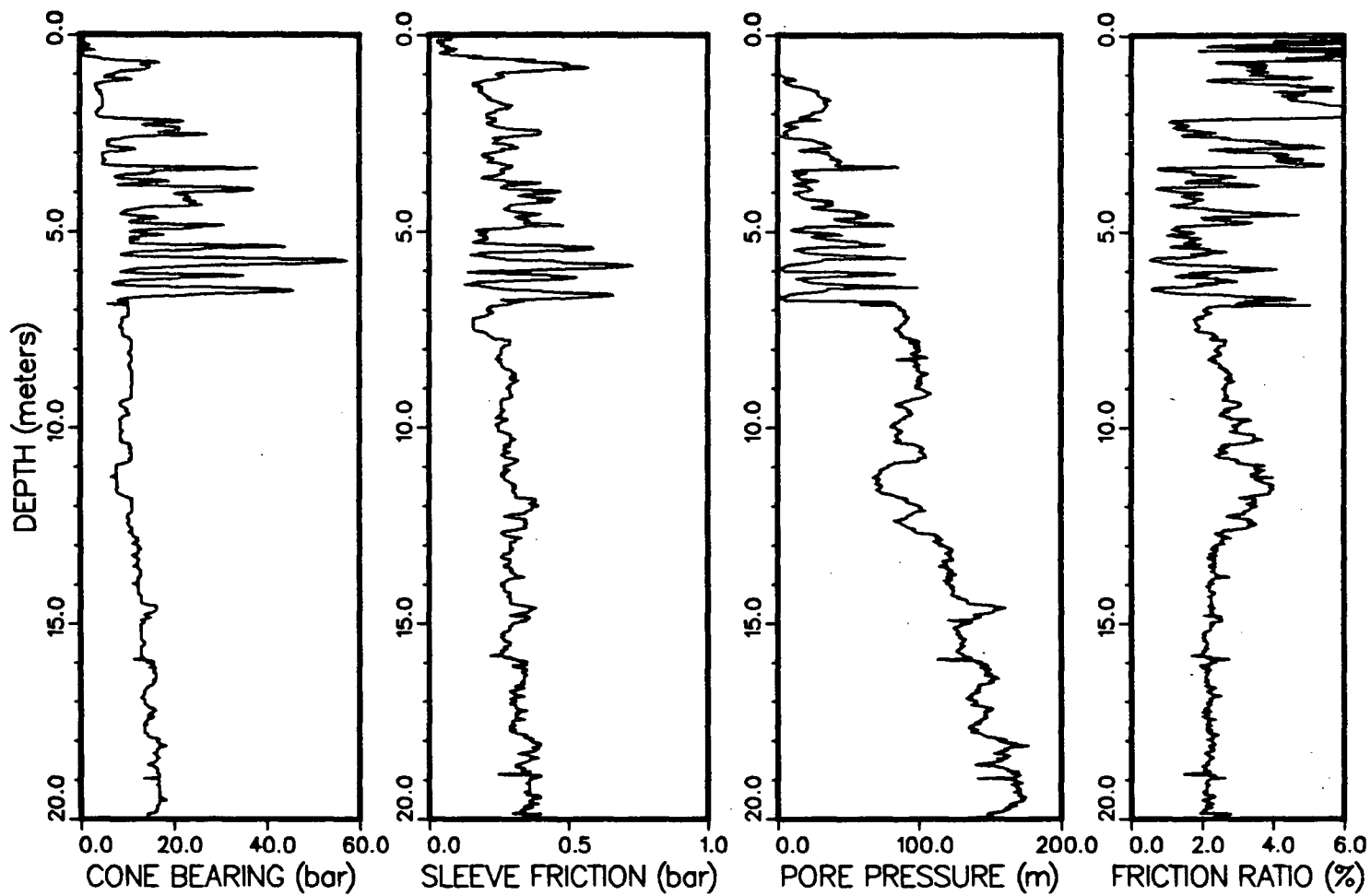
Figure 2.18: Cone Bearing, Sleeve Friction, Pore Pressure and Friction Ratio Profiles at Haney Site.
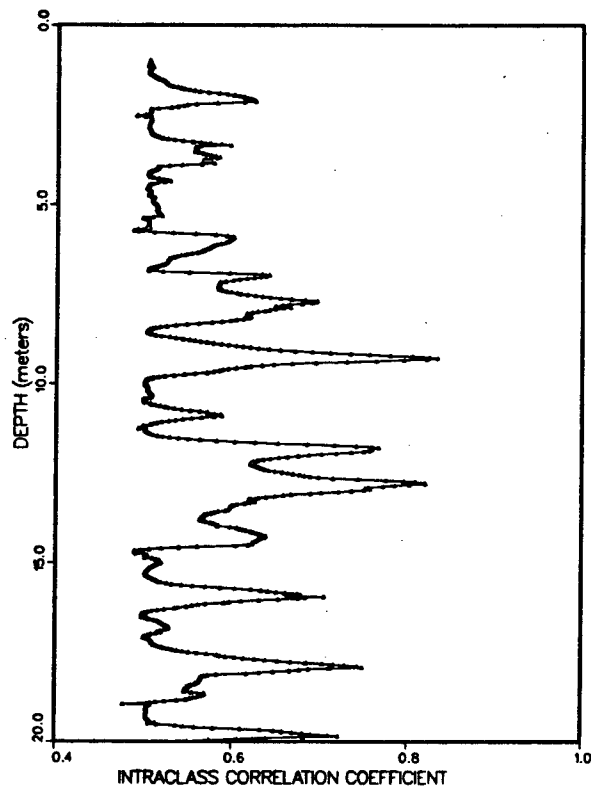
Figure 2.19: Intraclass Correlation Coefficient for Cone Bearing at Haney Site for a Window Width of 2.0 meters.
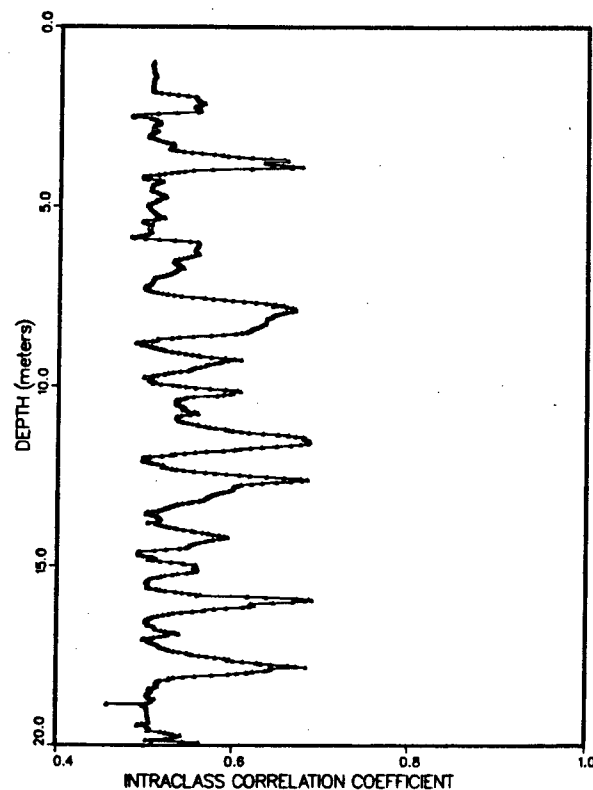
Figure 2.20: Intraclass Correlation Coefficient for Sleeve Friction at Haney Site for a Window Width of 2.0 meters.
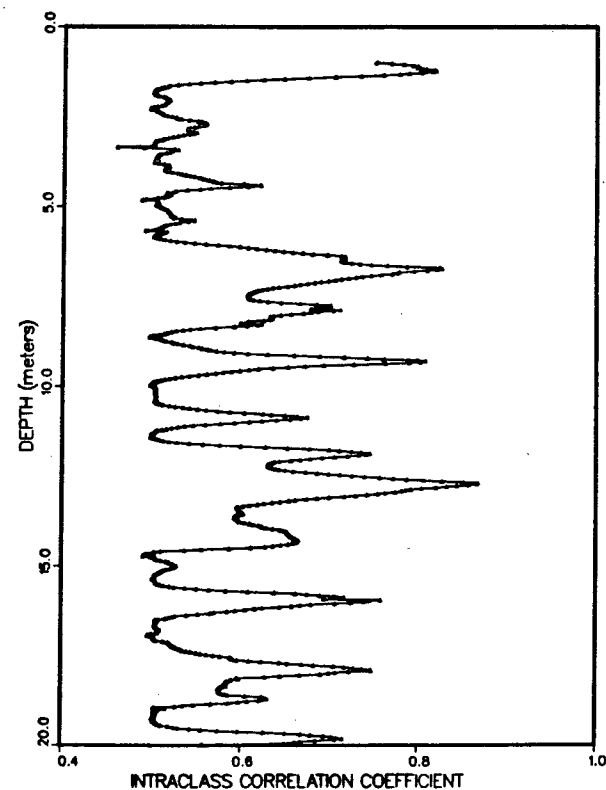
Figure 2.21: Intraclass Correlation Coefficient for Pore Pressure at Haney Site for a Window Width of 2.0 meters.

The Intraclass Correlation Coefficient, ρ_I, was calculated and the best layer boundaries for the three variables (Figs. 2.19 to 2.21) were obtained as follows;

Cone Bearing : 7.70, 9.30, 11.80, 15.95, 17.93 m.

Friction : 3.70, [6.70], 7.68, 9.28, 11.55, [12.65], 15.95, 17.83 m.

Pore Pressure : [1.10], [6.73], 7.78, 9.30, 11.88, [12.70], 15.88, 17.93 m.

Table 2.2 includes the best layer selections from the three statistical methods as well from the classification chart. based on the above depths which have been obtained for individual parameters. The combined plot of ρ_I for the three variables is illustrated in Fig. 2.22.

The T Ratio profiles (Fig. 2.23) suggest similar layer boundaries as those obtained from the Intraclass Correlation Coefficient.

The D^2 determination, considering all three variables, resulted in the following layer boundaries (Fig. 2.24);

[1.25], [6.73], 7.78, [9.30], [12.68], 15.95 and 17.93 m.

The bracketed values above indicate the depths where D^2 attained relatively high magnitudes. These depths are also categorized into two types of layering with the bracketed depths indicating the main layer boundaries and the other depths representing less prominent boundaries.

Table 2.2 shows that the statistical methods are in agreement with the identification of layer boundaries using the CPT chart on qualitative basis, although the less dominant boundaries at 9.30, 15.95 and 17.93 m are not picked up by the latter method. In contrast to the subjectivity of picking layer boundaries using the R_f profile in conjunction with the bearing profile, the other statistical profiles, namely the Intraclass Correlation Coefficient, T ratio and especially the D^2, recognize the layer boundaries distinctly and conveniently.
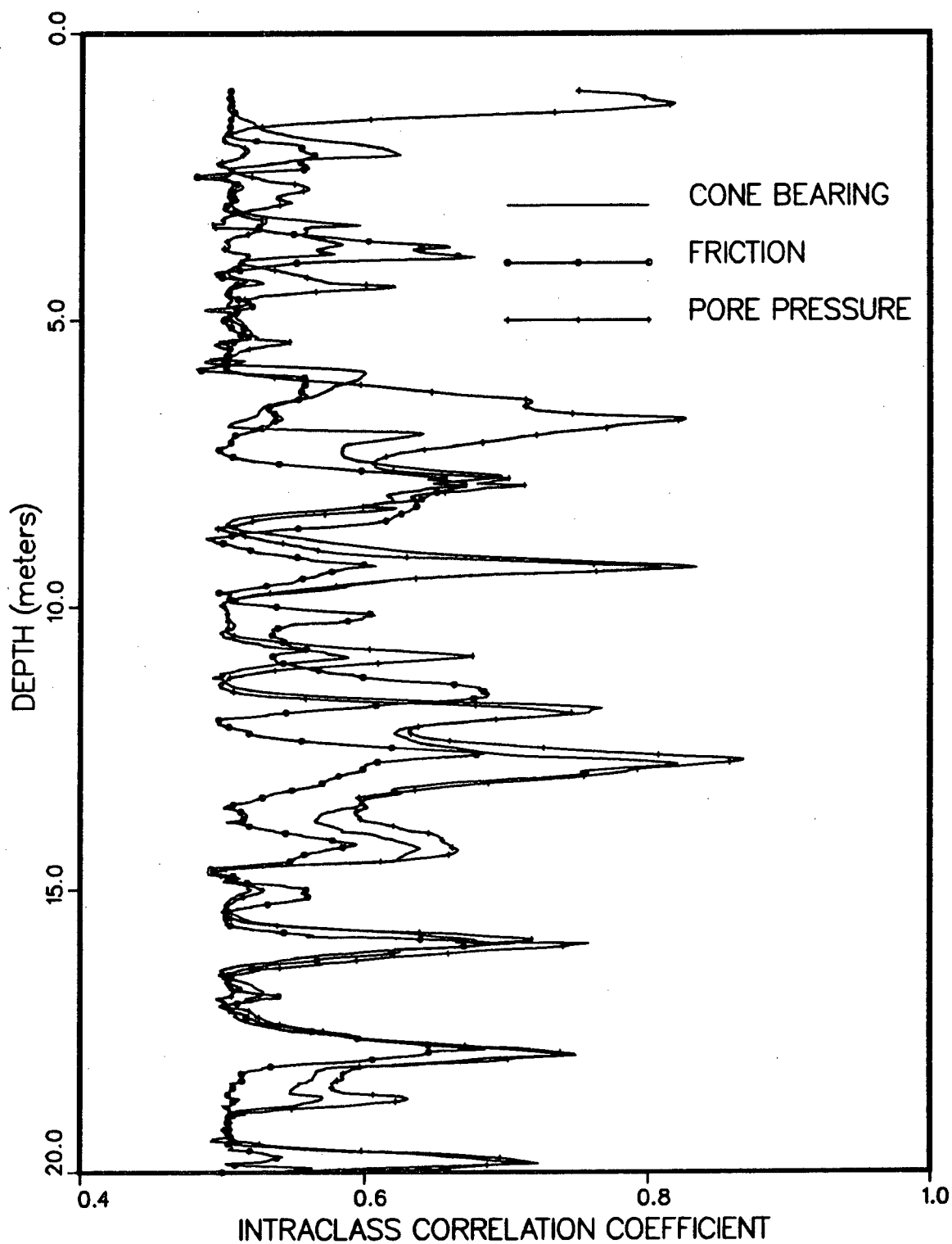
Figure 2.22: Intraclass Correlation Coefficient for Cone Bearing, Friction and Pore Pressure at Haney Site for a Window Width of 2.0 meters.
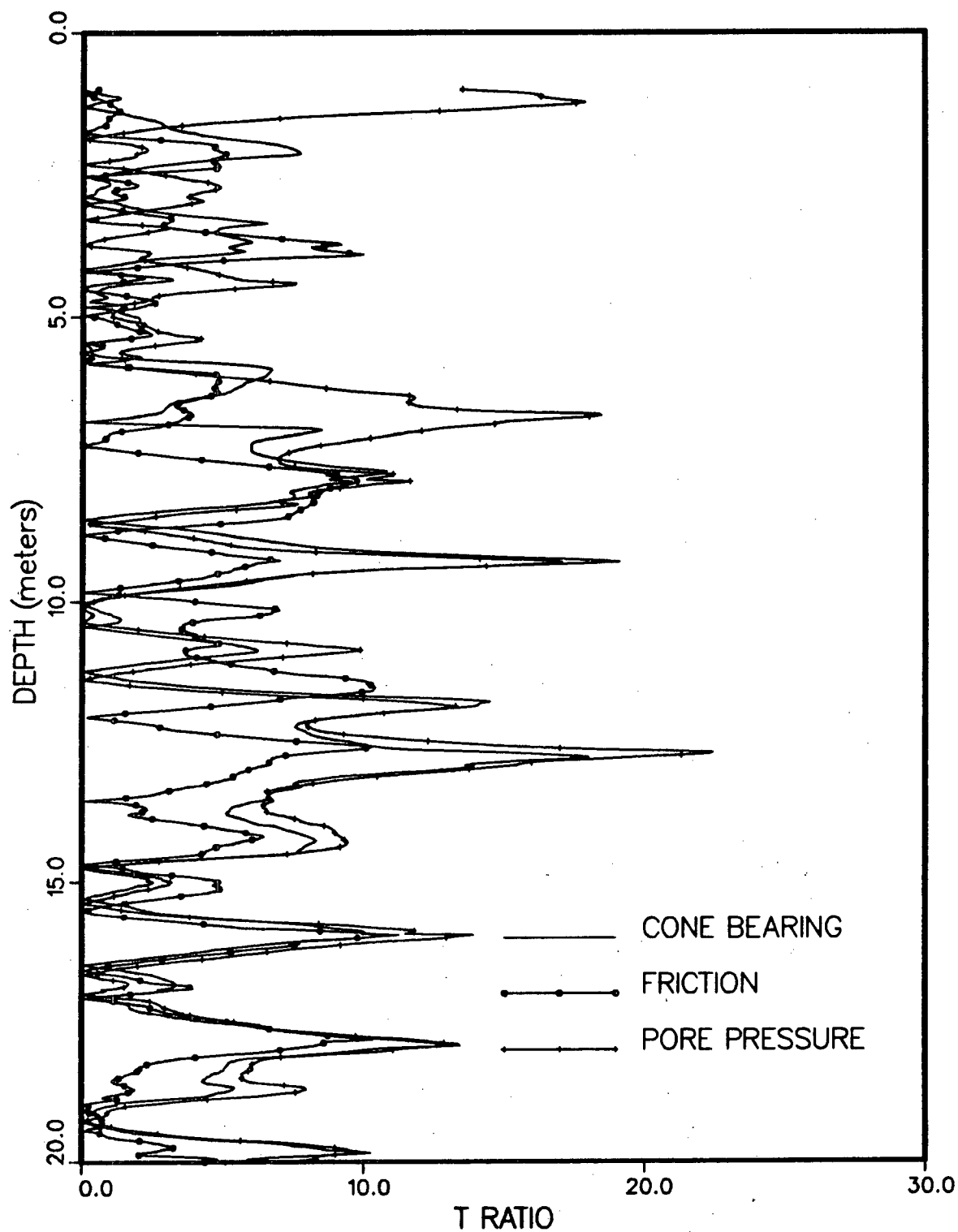
Figure 2.23: T Ratio for Cone Bearing, Friction and Pore Pressure at Haney Site for a Window Width of 2.0 meters.
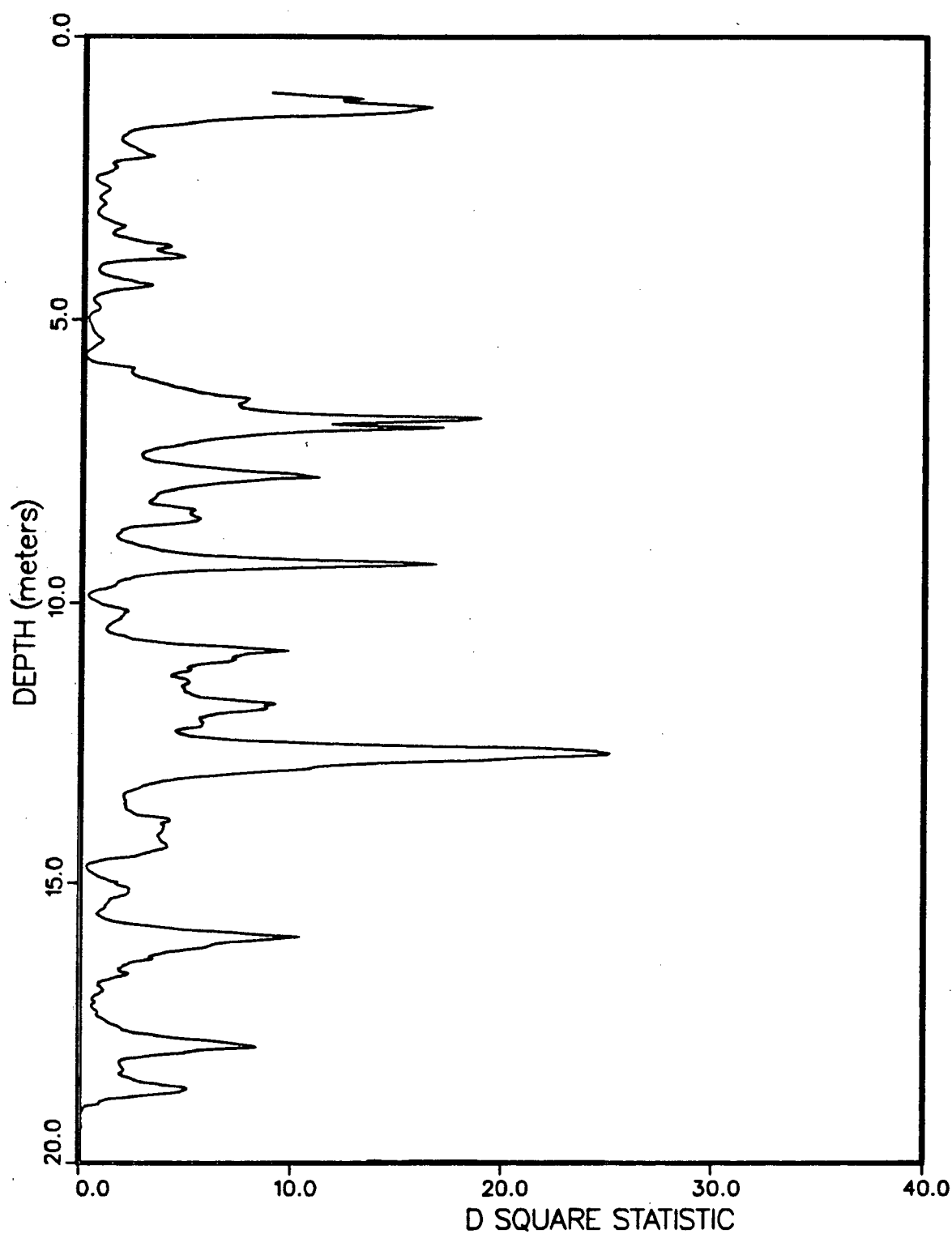
Figure 2.24: D^2 Statistic from Multivariate Analysis at Haney Site for a Window Width of 2.0 meters.

2.4.3 Tilbury Island Site

The Tilbury Island profile (Fig. 2.25) is predominantly sand with some surface silt. Pore pressure data were not available and only the cone bearing and friction will be considered. Based on an autocorrelation analysis of the variables concerned, a W_D of 2.0 m was selected for the analysis. The variation of the Intraclass Correlation Coefficient (Fig. 2.26), picked the following boundaries for bearing and friction.

Cone Bearing : [1.26], [2.13], [7.58], 9.70, [11.70], 12.68, 17.18 m.

Friction : [1.23], [2.13], [7.55], [11.60], 12.67, 17.17 m.

The D^2 profile (Fig. 2.27), gave the following boundaries;

[1.28], [2.18], [7.78], [11.75], 12.68 and 17.20 m.

Table 2.3 shows the comparison of the two methods and the results are agreeable qualitatively although the statistical methods have the added advantage of recognizing additional sublayering. Even in situations such as this, the advantages of the statistical methods are self explanatory from Table 2.3, with its ability of picking specific layers. It should be reiterated that although the traditional method picks the layer boundaries by judgement supported by the classification chart, the statistical methods of identification entail no errors caused by incorrect judgement. As in the previous examples, once the layer boundaries are specifically determined, the well established methods in in situ geotechnical engineering can be used to characterize these layers.

2.4.4 Primary and Secondary Layer Boundaries

The statistics already described will have varying magnitudes depending on the power of discrimination between layers. The higher the value of the statistic at peaks of the statistic profile, the greater is the chance for a layer boundary to occur at that point.
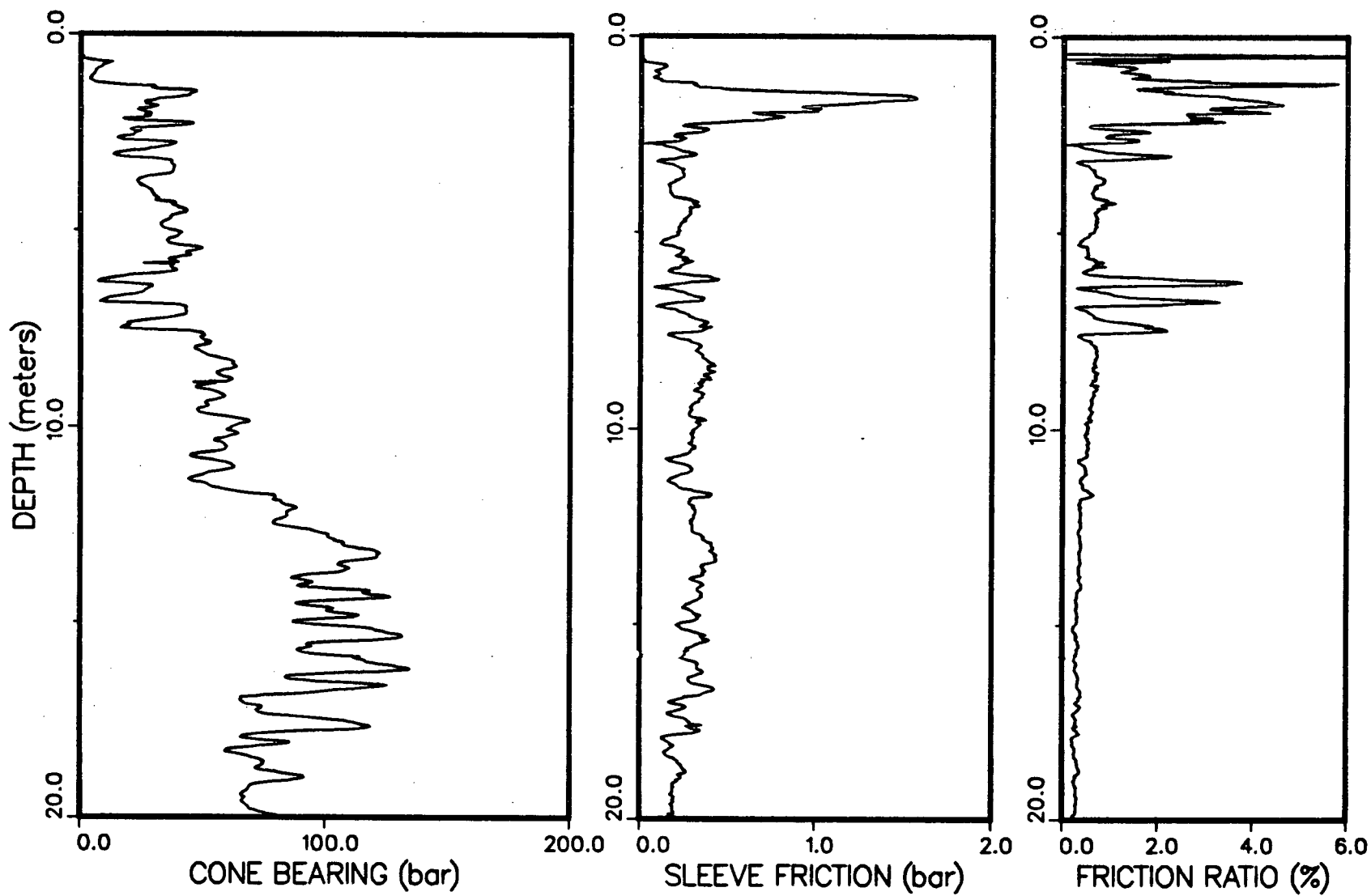
Figure 2.25: Cone Bearing, Sleeve Friction and Friction Ratio Profiles at Tilbury Island.
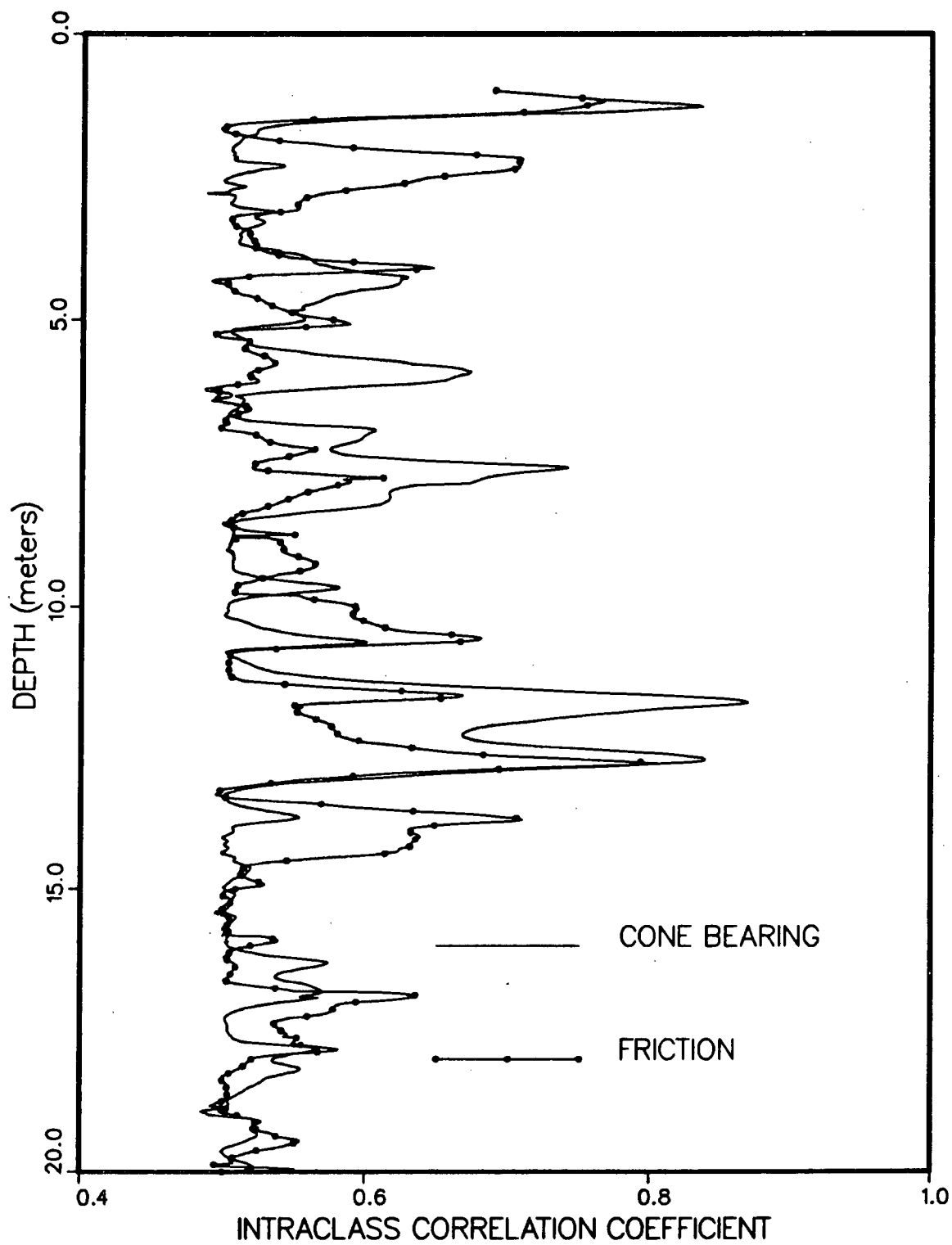
Figure 2.26: Intraclass Correlation Coefficient for Cone Bearing and Friction at Tilbury Island for a Window Width of 2.0 meters.
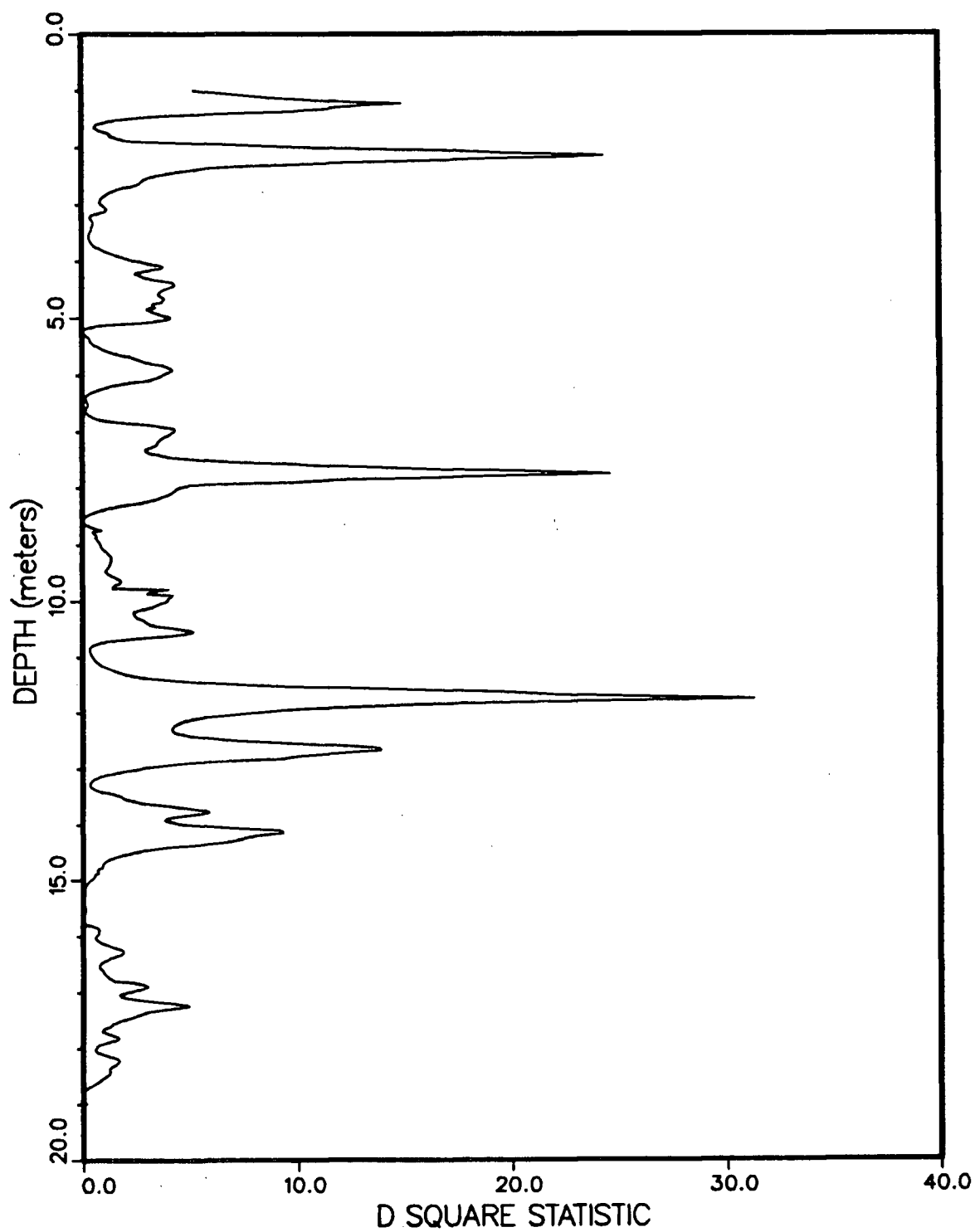
Figure 2.27: D^2 Statistic from Multivariate Analysis at Tilbury Island for a Window Width of 2.0 meters.

While there will be very high peaks of these statistics, it is also possible to have peaks possessing significantly lower magnitudes. Different ranges of these peaks may be established, with each range depicting sublayer boundaries. However, the peaks of the profiles that are of concern in this study will be the group of the maximum peaks and the peaks that fall into a lower range of magnitudes of the statistics. On the above lines, two types of layering can be defined as follows;

(a) A primary layer boundary is found at a depth at which the statistic under consideration attains a very high value (points giving the highest peaks).

(b) A secondary layer boundary is found at a depth at which the statistic attains a peak, but not to the extent as given by (a) above. These peaks will be in a range lower than that of the maximum peaks.

This concept of primary and secondary layer boundaries is analogous to the layering and sublayering in geotechnical engineering. In the overall sense, when the variation of both the univariate and multivariate statistics are considered in layer boundary identification, there are further requirements which have to be met in deciding which is a primary and which is a secondary layer boundary. However, the above definition would suffice at this stage and the detailed requirements as mentioned above will be described in a subsequent section. The values of the depths obtained from the CPT chart in Tables 2.1, 2.2 and 2.3 have been based on the results of the survey conducted to evaluate the subjectivity involved in the use of the chart. The values decided by the subjects in the survey consisted of a fair amount of dispersion and therefore the depths selected represent the approximate means of the different layer boundary depths obtained from the survey.

Table 2.1: Comparison of the Layer Boundaries Identified Using the CPT Chart with the Proposed Statistical Methods, for McDonald Farm Data.

Depth (m)	Soil Characteristics		Layering and Soil Type from CPT Chart	Proposed Statistical Methods					
	Cone Bearing (bar)	Friction Ratio %		Primary Layer Boundaries			Secondary Layer Boundaries		
				T	ρ_I	D^2	T	ρ_I	D^2
	< 30	4.0	Organic Silty Clay						
	30 – 40	0.75-2.0	Silty Sand and Sandy Silt						
5	80 – 140	0.4	Medium to Dense Sand						
10									
	80	0.6-1.0	Silty Sand						
15									
	10	0.8	Silty Clay						
20									

Table 2.2: Comparison of the Layer Boundaries Identified Using the CPT Chart with the Proposed Statistical Methods, for the Haney Data.

Depth (m)	Soil Characteristics		Layering and Soil Type from CPT Chart	Proposed Statistical Methods					
	Cone Bearing (bar)	Friction Ratio %		Primary Layer Boundaries			Secondary Layer Boundaries		
				T	ρ_I	D^2	T	ρ_I	D^2
	0-5	2-6	Organic Silty Clay						
5	20-50	2-6	Clayey Silt						
10	10	2.5-3.5	Silty Clay						
15	12-18	2	Sandy Silt						
20									

Table 2.3: Comparison of the Layer Boundaries Identified Using the CPT Chart with the Proposed Statistical Methods, for the Tilbury Island Data.

Depth (m)	Soil Characteristics		Layering and Soil Type from CPT Chart	Proposed Statistical Methods			
	Cone Bearing (bar)	Friction Ratio %		Primary Layer Boundaries		Secondary Layer Boundaries	
				P_I	D^2	P_I	D^2
	30	2 - 6	Gravelly Sand to Sand				
5	20 - 35	0.6-1.0	Silty Sand to Sandy Silt				
10	60	0.5	Sand to Silty Sand				
15	70 - 125	0.4	Sand				
20							

2.5   Types of Profiles where T Ratio, ρ_I and D^2 are Unable to Detect Layer Boundaries

The applications of the layer identification statistics, namely, the T Ratio, ρ_I and the D^2 statistic have clearly indicated their superiority over the method popularly used at present, for discriminating layers. All these methods are based on the first and second moments of the samples on either side of the window center, d_o.

In the event of the presence of two adjacent layers, one with an increasing trend and the other with a decreasing trend (Type I, Fig. 2.28) the means of the two layers could possibly be approximately equal. The discriminating statistics depend on the difference of the means of the two adjacent layers and at the expected boundary at A in Fig. 2.28, the T Ratio, ρ_I and D^2 would all reach a minimum, instead of a maximum that is normally expected at a layer boundary. This type of profile is rare and none of the data available at UBC exhibited such a behavior. Therefore, a profile was simulated to reflect the above type of behavior. The variations of the T Ratio, ρ_I and D^2 statistics for the simulated profile is illustrated in Figs. 2.29 to 2.31. While ρ_I attains a minimum at approximately 10.0 m (around point A), the T Ratio and D^2 statistics reach zero. The latter two statistics become zero because the cone bearing immediately following the 10 m depth was simulated to have a trend exactly opposite to that of the cone bearing just prior to that depth. This is a hypothetical case, and in reality what could be expected is a minimum value as for ρ_I. The effect of Fourier smoothing on the statistics are also illustrated in Figs. 2.29 to 2.31. It is evident that smoothing does not significantly improve the identification efficiency of the three statistics discussed.

The other type of layer boundary which could be expected not to be detected using the above statistics is in a profile where the gradients on either side of a potential boundary, d_o, barely change with the mean on either side being approximately equal.
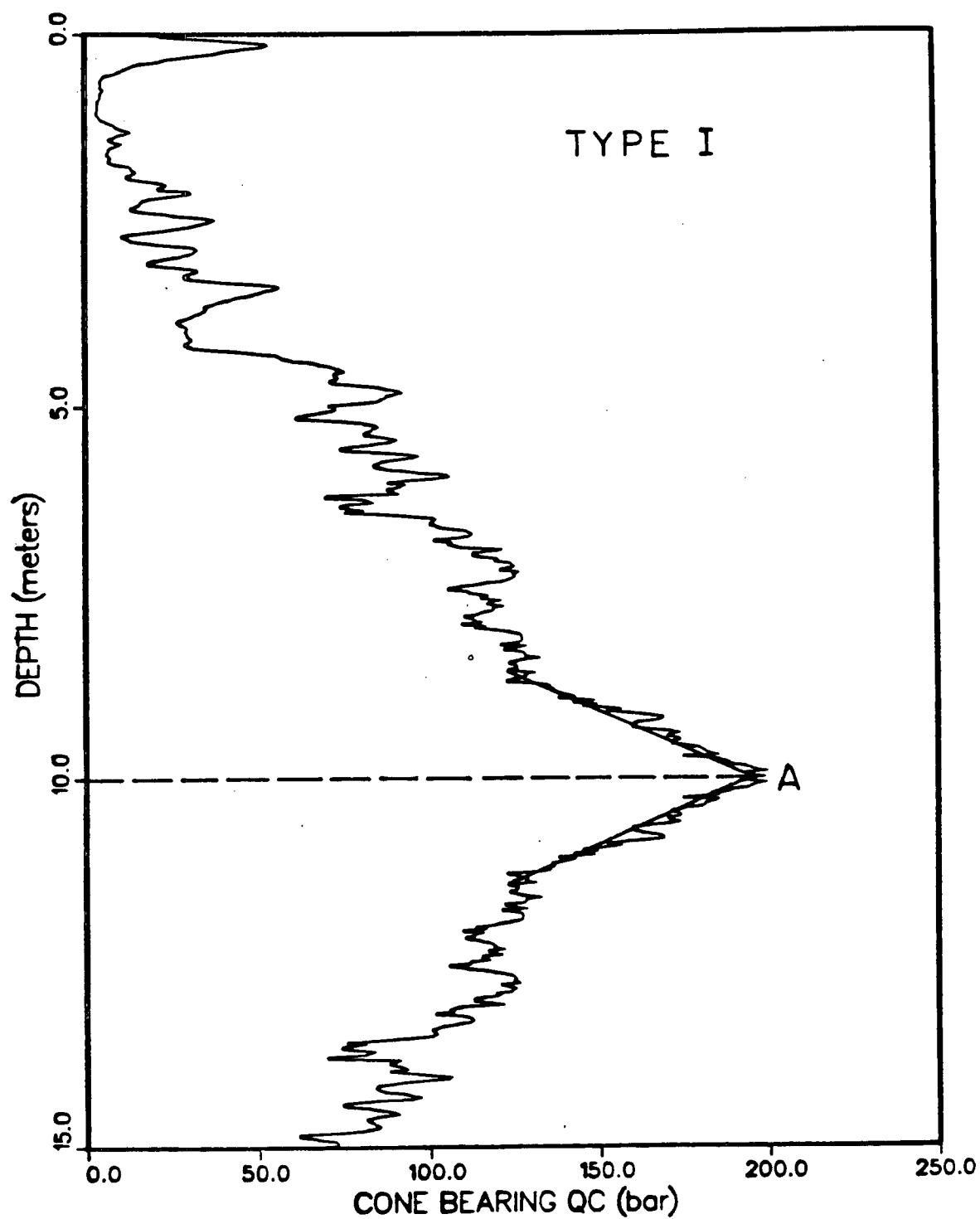
Figure 2.28: Type I Profile
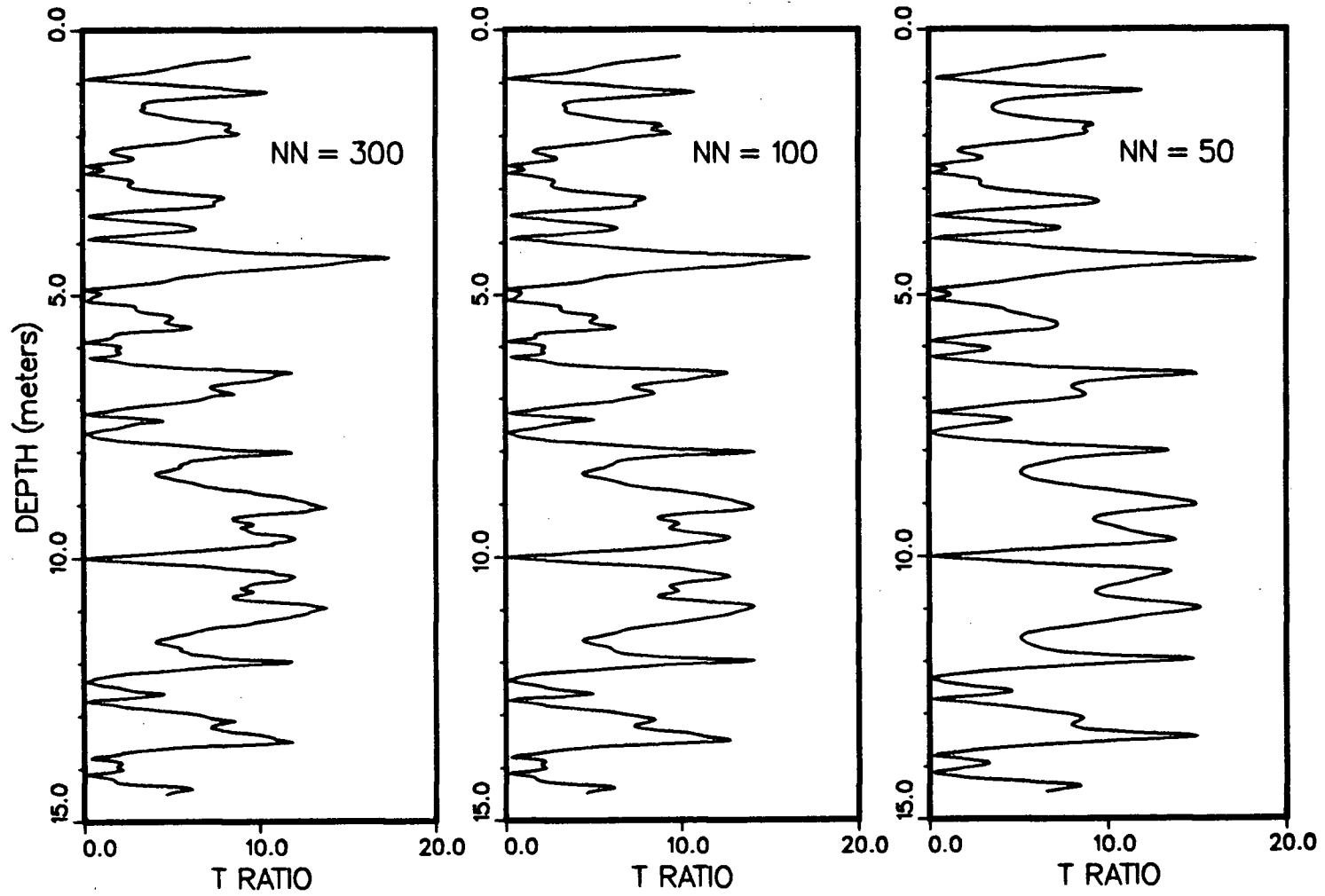
Figure 2.29: T Ratio for Type I Profile for Different Degrees of Fourier Smoothing. NN = 300 Represents the Unsmoothed Profile, NN = 100 Represents the Profile with 100 of the 300 Possible Harmonics Considered and NN = 50 with only 50 Harmonics (highest degree of smoothing).
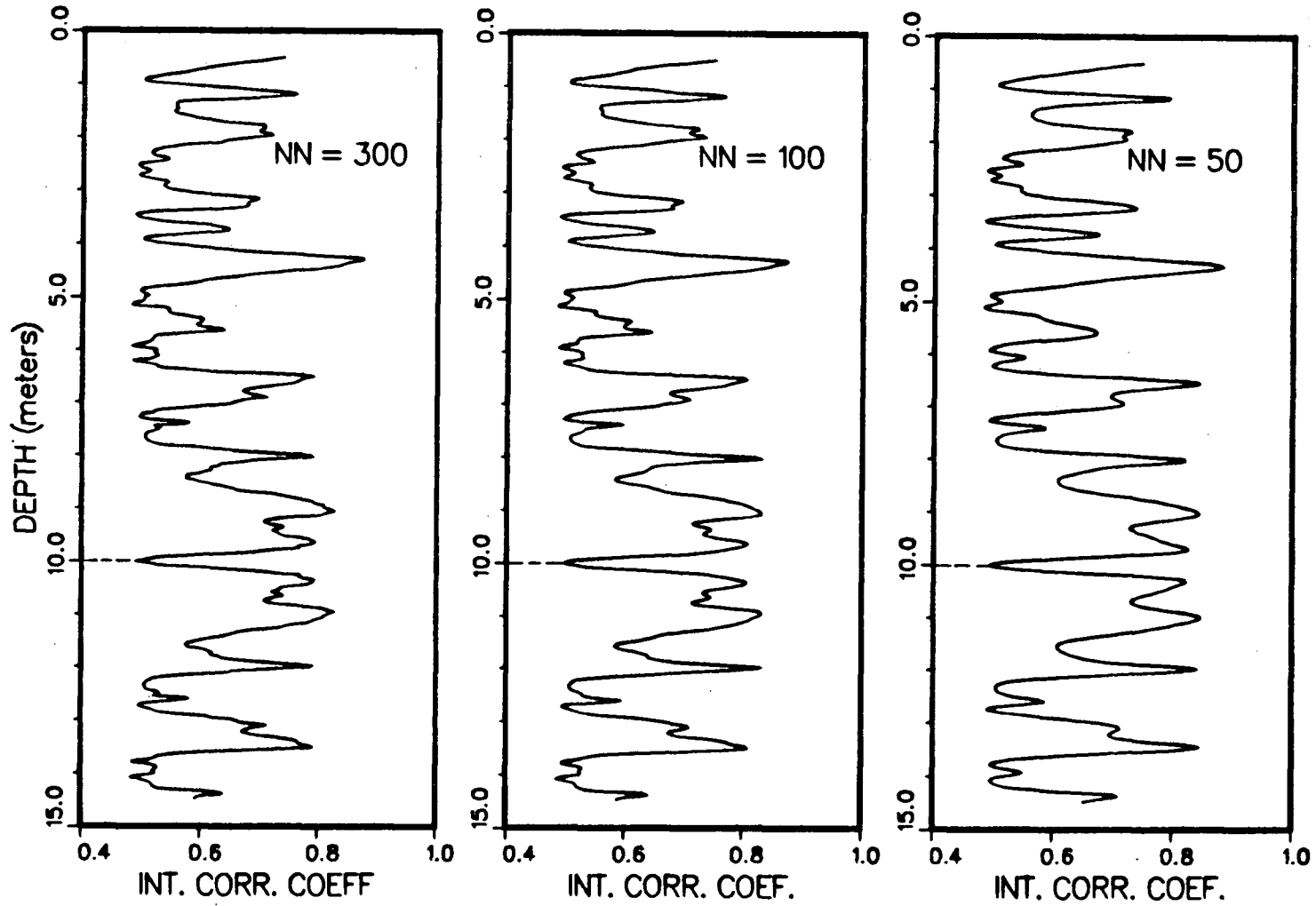
Figure 2.30: Intraclass Correlation Coefficient for Type I Profile for Different Degrees of Fourier Smoothing. NN = 300 Represents the Unsmoothed Profile, NN = 100 Represents the Profile with 100 of the 300 Possible Harmonics Considered and NN = 50 with only 50 Harmonics (highest degree of smoothing).
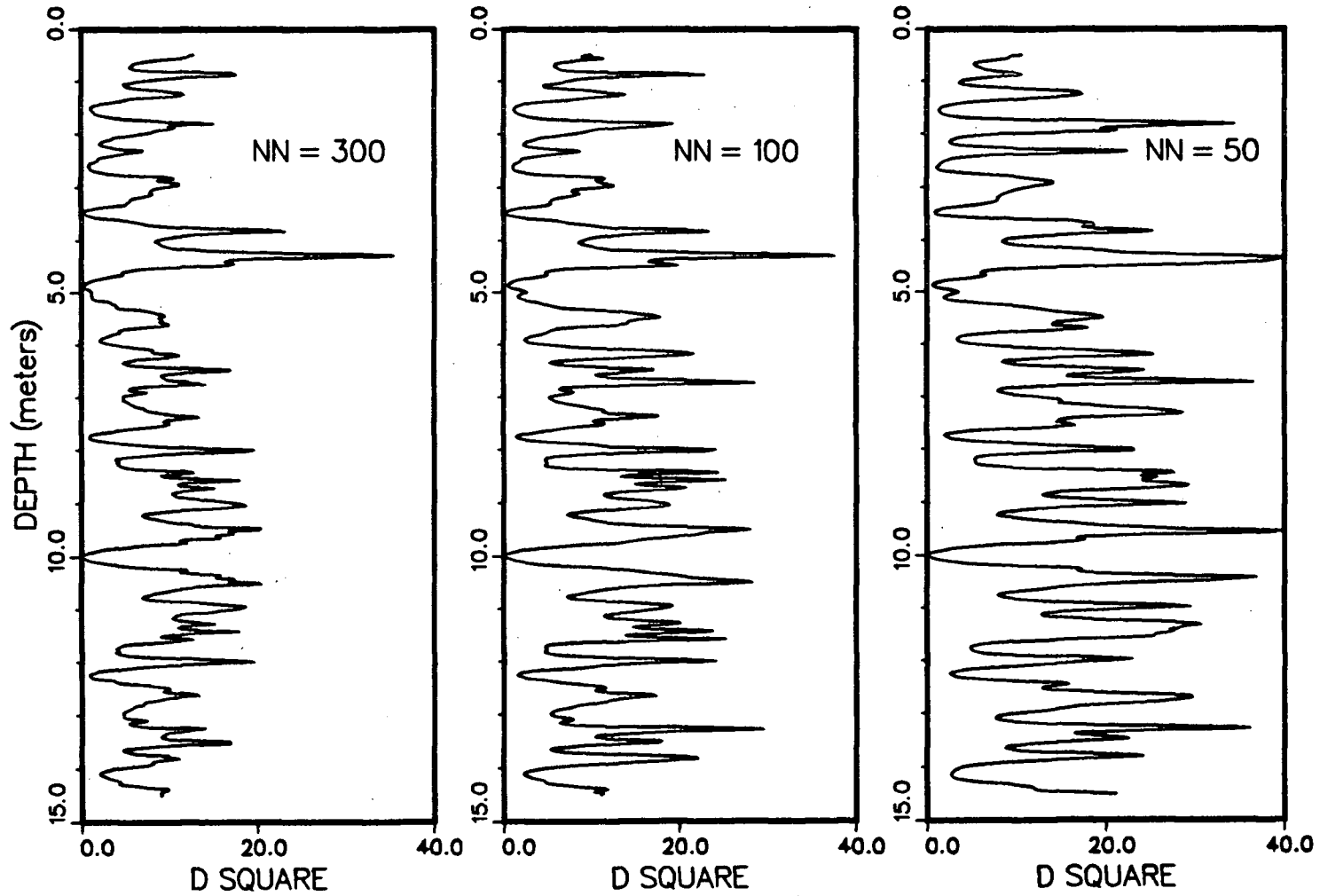
Figure 2.31: D^2 Statistic for Type I Profile for Different Degrees of Fourier Smoothing. NN = 300 Represents the Unsmoothed Profile, NN = 100 Represents the Profile with 100 of the 300 Possible Harmonics Considered and NN = 50 with only 50 Harmonics (highest degree of smoothing).

This type of boundary will fall into the category of secondary layer boundaries. A profile to represent the above behavior (Type II) was also simulated and is illustrated in Fig. 2.32 and the point of concern is around B, where a layer boundary can be expected due to the change in the gradient. The T Ratio, ρ_I and D^2 statistics for the profile in Fig. 2.32 is illustrated in Figs. 2.33 - 2.35 which do not possess peaks at the expected layer boundary near B. A suggested method to identify layer boundaries in profiles similar to Types I and II is discussed below.

It has to be reiterated that the Type I (Fig. 2.28) and Type II (Fig. 2.36) profiles are synthesized profiles to illustrate the effectiveness of the Gradient method more clearly.

2.5.1 Change of Gradient Between Layers

Soil properties are highly depth dependent and more commonly linearly depth dependent. A linear description requires two statistics to fully describe its behavior, namely, the intercept and the gradient. Therefore, the intercept and the gradient are two possible statistics which could be used to identify different layers, when statistics such as the T Ratio, ρ_I and D^2 fail in performing this task, as for the cases illustrated in the previous section.

An appreciable change in the intercept would be reflected in the change of the means and in turn would be reflected in the increased values of the T Ratio, ρ_I and D^2. However, as was illustrated before, the change in gradient alone may not be reflected in the above discriminating statistics. In an attempt to arrest the above problem of obtaining layer boundaries in situations such as Type I and Type II profiles, the window was moved along the data profile and the gradients on either side of its center, d_o, were investigated. This study indicated that the absolute value of the difference of the gradients representing the linear influence before and after
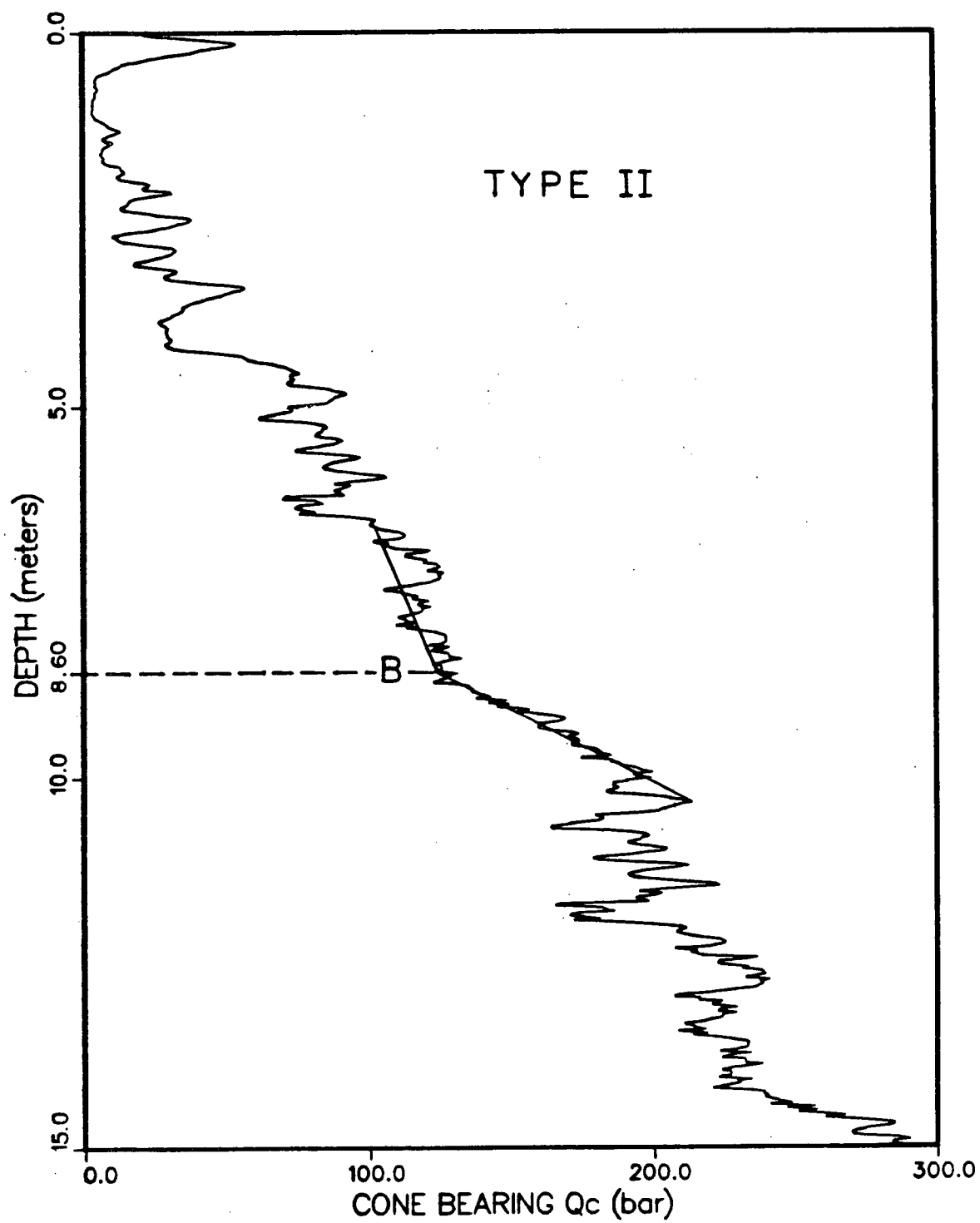
Figure 2.32: Type II Profile.
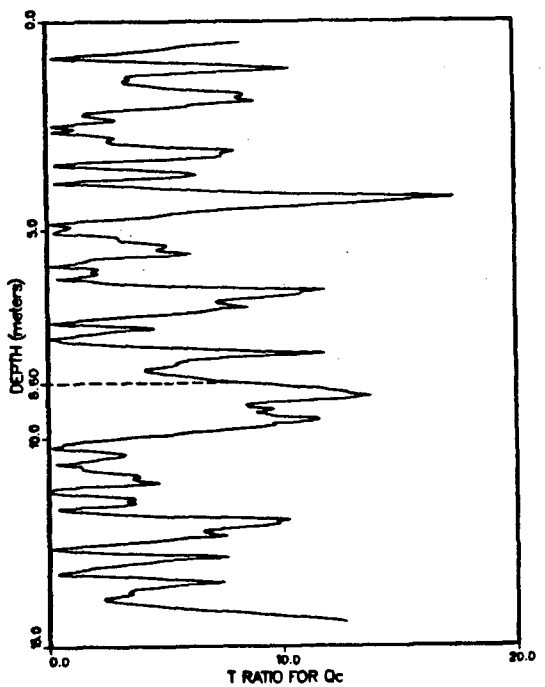
Figure 2.33: T Ratio for Type II Profile.
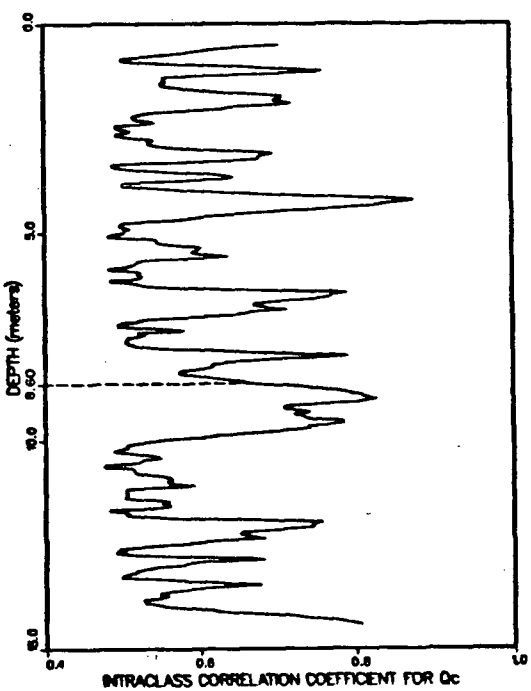
Figure 2.34: Intraclass Correlation Coefficient for Type II Profile.
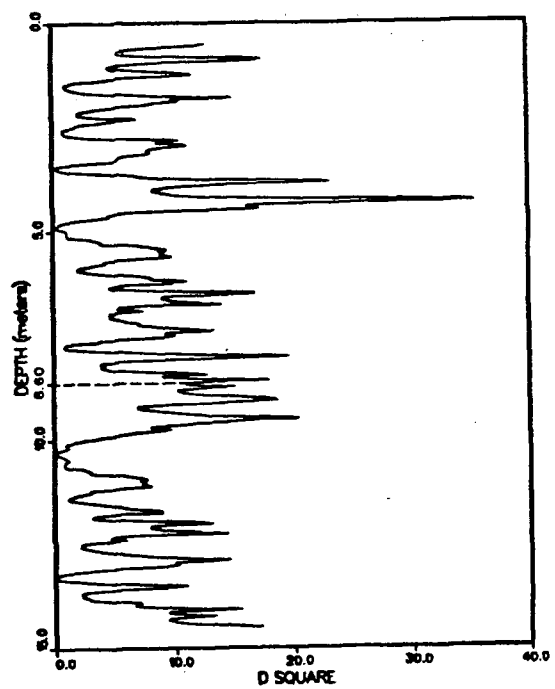
Figure 2.35: D^2 Statistic for Type II Profile.
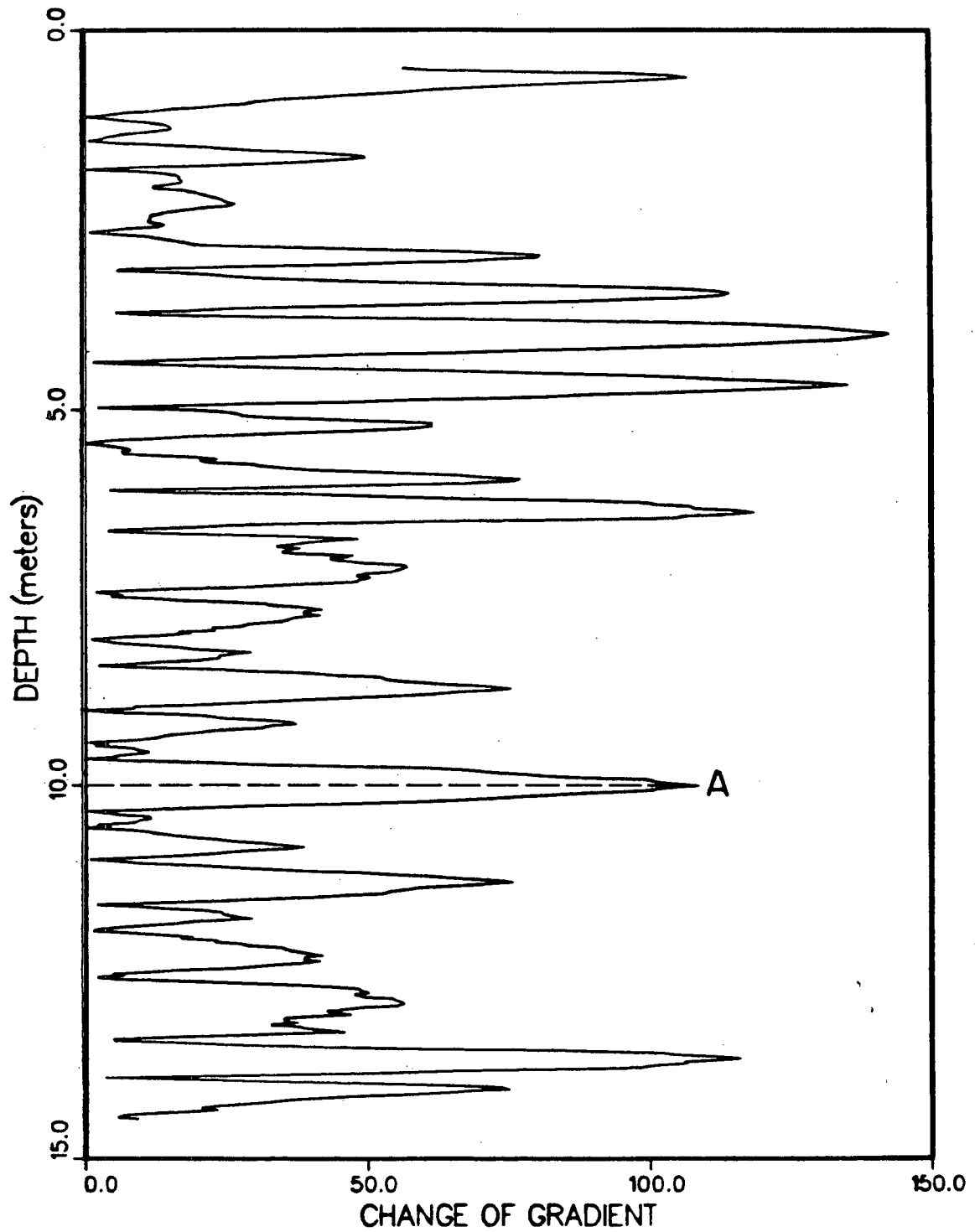
Figure 2.36: Change of Gradient for Type I Profile.

d_o, would be a reliable statistic to define a layer boundary of the types discussed. The change of gradient profile for the Type I profile is illustrated in Fig. 2.36 and for the Type II profile, it is given in Fig. 2.37. It is evident from the above figures that the change of gradient reaches a local maximum at 10.0 m (point A) for Type I and at 8.60 m (point B) for the Type II profile, providing sufficient evidence as to its capability of picking up layers of the above types.

Figure 2.38 illustrates that the effect of smoothing on the change of gradient is negligible. The T Ratio, ρ_I and D^2 statistics (Fig. 2.29 - 2.31) of the Type I profile also exhibited a lack of sensitivity to Fourier smoothing, providing evidence to the understanding that CPT data is relatively noise free. More elaborate descriptions of filtering and smoothing methods are described in Chapter 3.

2.6 Sensitivity of Window Width

As discussed in section 2.3.1, the effect of a narrow window width (W_D) is the introduction of noise into the statistic under consideration. However, this would not affect the selection of layer boundary depths. In contrast, the choice of a wider window width could lead to the possibility of missing out possible layer boundaries. The latter option could result in serious consequences and has to be avoided by choosing a low value for W_D based on the autocorrelation function of the variables concerned.

The effect of W_D on the layer boundary depth was investigated in detail for selected depths for a soil stratum comprised mainly of sand (Mc Donald Farm) and of clay (Haney), and the results for the Intraclass Correlation Coefficient are tabulated in Table 2.4 and Table 2.5 for the two sites respectively.

The results indicate that the primary layer boundary depth is not highly sensitive to the window width chosen. However, this does not preclude the possibility of missing out layers if too wide a W_D is selected, with the secondary layer boundaries being
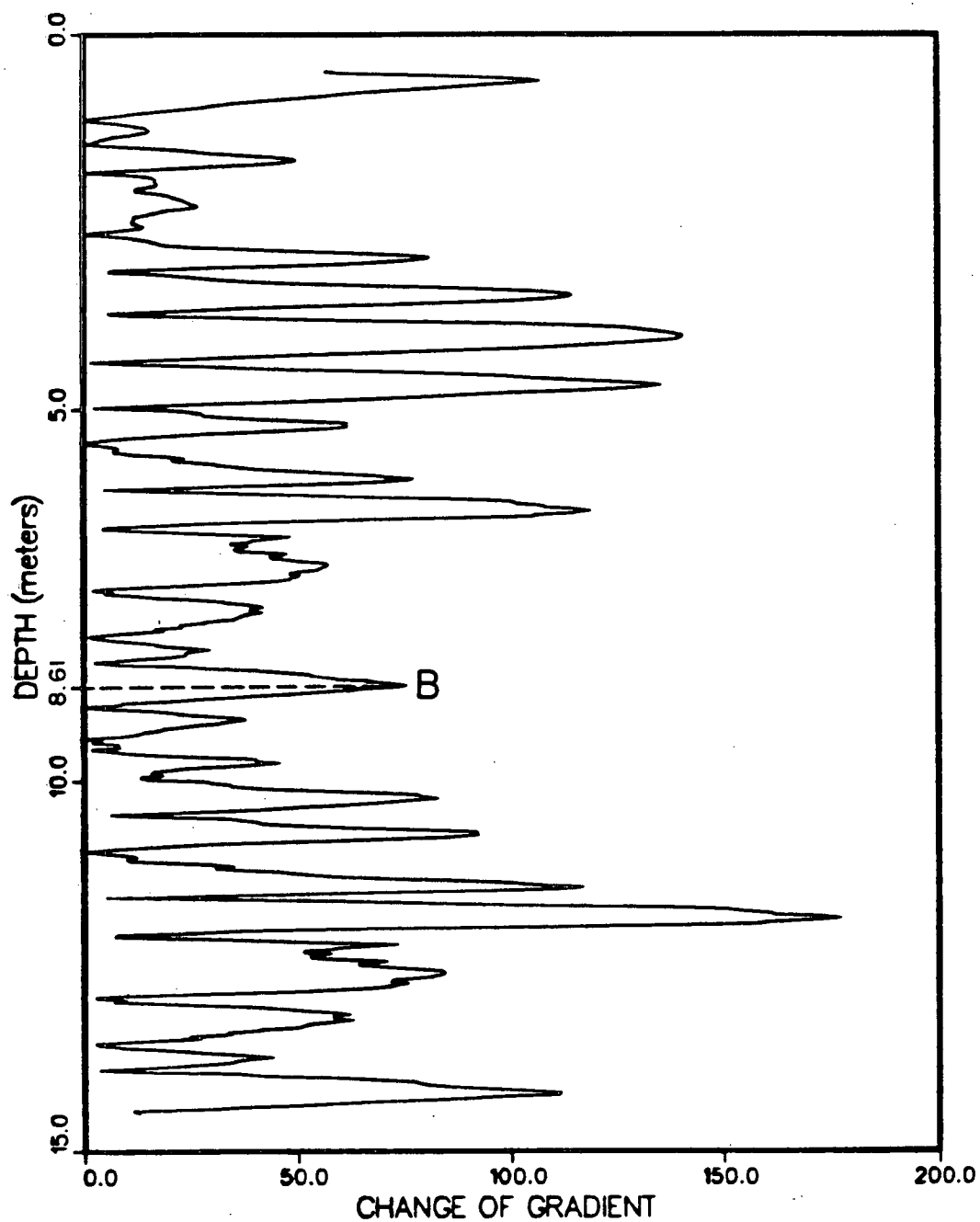
Figure 2.37: Change of Gradient for Type II Profile.
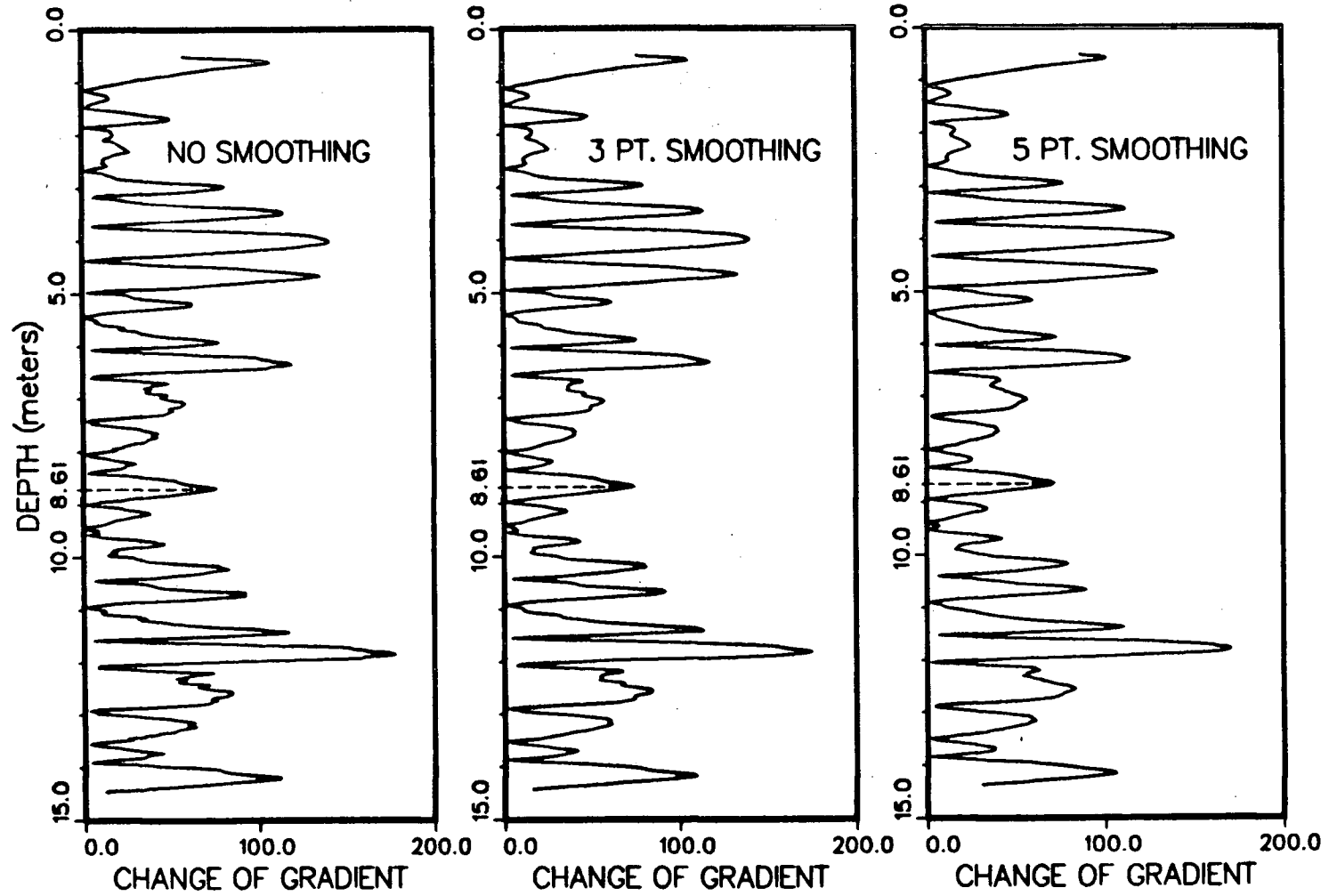
Figure 2.38: Change of Gradient for Type II Profile for Different Degrees of Moving Average Smoothing.

Table 2.4: Effect of Window Width on Primary Layer Boundary Depth for the Intraclass Correlation Coefficient for McDonald Farm.

Window Width (m)	1.0	1.5	2.0	2.5
Layer Boundary Depth (m)	4.30	4.35	4.35	4.35
Intraclass Correlation Coefficient	0.8743	0.8783	0.8586	0.8725
Layer Boundary Depth (m)	9.05	9.05	9.03	9.00
Intraclass Correlation Coefficient	0.8975	0.9097	0.8771	0.8460
Layer Boundary Depth (m)	14.48	14.50	14.50	14.48
Intraclass Correlation Coefficient	0.8776	0.8809	0.8696	0.8723

Table 2.5: Effect of Window Width on Primary Layer Boundary Depth for the Intraclass Correlation Coefficient for Haney Site.

Window Width (m)	1.0	1.5	2.0	2.5
Layer Boundary Depth (m)	6.70	6.78	6.73	6.75
Intraclass Correlation Coefficient	0.7752	0.7200	0.7051	0.7122
Layer Boundary Depth (m)	9.30	9.30	9.30	9.25
Intraclass Correlation Coefficient	0.7742	0.8197	0.8317	0.7626
Layer Boundary Depth (m)	12.73	12.75	12.70	12.75
Intraclass Correlation Coefficient	0.7927	0.7963	0.8016	0.7506

Table 2.6: Effect of Window Width on Primary Layer Boundary Depth for the D^2 Statistic for McDonald Farm.

Window Width (m)	1.0	1.5	2.0	2.5
Layer Boundary Depth (m)	4.30	4.30	4.33	4.33
D^2 Statistic	24.99	30.15	26.53	27.67
Layer Boundary Depth (m)	9.05	9.10	9.03	9.03
D^2 Statistic	31.47	35.99	24.26	16.06
Layer Boundary Depth (m)	14.50	14.60	14.53	14.50
D^2 Statistic	34.46	33.21	30.10	27.17

more susceptible. The non sensitivity of the primary layer boundary depth to small changes in window width clearly reveals the robustness of the Intraclass Correlation Coefficient as an adequate parameter. A similar type of sensitivity analysis was also performed for the D^2 statistic, the results of which are tabulated in Tables 2.6 and 2.7 for the two sites.

These results in Tables 2.6 and 2.7 also indicate that the depths of layer boundaries are not highly sensitive to the window width chosen. The absolute values of the D^2 however, seem to be dependent on the window width, with no specific pattern of variation with increasing or decreasing W_D. This phenomenon is not of real concern, as long as the depths of local maxima (peaks) of the D^2 profile are not significantly dependent on the width of window selected.

The secondary layer boundary depths are somewhat sensitive to W_d for both the Intraclass Correlation Coefficient and D^2 and therefore the use of a W_D based on the autocorrelation function will avoid the possibility of missing such boundaries.

Table 2.7: Effect of Window Width on Primary Layer Boundary Depth for the D^2 Statistic for Haney Site.

Window Width (m)	1.0	1.5	2.0	2.5
Layer Boundary Depth (m)	6.75	6.73	6.73	6.73
D^2 Statistic	28.63	25.19	18.89	15.52
Layer Boundary Depth (m)	9.25	9.28	9.30	9.25
D^2 Statistic	19.34	20.77	16.82	18.90
Layer Boundary Depth (m)	12.70	12.68	12.68	12.65
D^2 Statistic	24.89	21.53	25.16	27.65

2.7 Establishment of Critical Statistics

The establishment of critical acceptance levels for the layer identification statistics is desirable to alleviate the need for picking out layers based on judgement of the variation of these statistics. As a result of the investigation of several data sets representing different soil types, certain criteria have been developed to achieve the above purpose. It is suggested that a combination of the univariate and multivariate methods would be the optimum way by which certain guidelines could be established.

2.7.1 Univariate Analysis

The results obtained for the three sites already discussed reflect the close association of the T Ratio and ρ_I. The Intraclass Correlation Coefficient is a normalized form of a variability parameter with a typical range between zero and unity, rendering it especially convenient for comparison purposes. The T Ratio, however, suffers from the fact that it is somewhat dependent on the magnitude of the units of the data considered. For the aforementioned reasons, it is proposed that the Intraclass Correlation

Coefficient is more preferable to be used as a layer boundary identifying statistic, for the univariate case.

A closer inspection of the ρ_I profiles for the three data sets at McDonald Farm, Haney and Tilbury Island reveals the presence of certain bounds which can be identified for the purpose of defining primary and secondary layer boundaries. The critical ranges of these limits are given in Table 2.8. These limits need not be applied stringently but an allowance of approximately $\pm$ 10% would be permissible. It should be reiterated that the fulfillment of these requirements alone is not sufficient, and for a layer to be identified as primary or secondary, the limits on the requirements for D^2 described in section 2.7.2 should also be satisfied.

Table 2.8: Critical Levels of the Intraclass Correlation Coefficient for the Definition of Primary and Secondary Layer Boundaries.

Boundary Type	Range of ρ_I
Primary	$\rho_I \geq 0.80$
Secondary	$0.80 > \rho_I \geq 0.65$

2.7.2 D^2 Statistic for Multivariate Analysis

In contrast to the Intraclass Correlation Coefficient the D^2 statistic is a function of the cross correlation structure of the variables. The correlation structure of sand and clay are very different and, therefore, the critical limits will be different for different soil types. In addition to the sites already described, data from several other sites were also analyzed to obtain appropriate limits for the definition of primary and secondary layer boundaries. Two levels of maxima for these data are tabulated in Table 2.9

which also includes the soil type that predominates the stratum under consideration.

The information given in Table 2.9 indicates relatively high values of D^2 for sand compared to clay and will have to be given due consideration in the formulation of significance levels. The discrepancy in their relative magnitudes is due to the variances and covariances of the variables in sand being relatively higher than their counterparts in clay. The ranges given in Table 2.9 lead to Table 2.10 which tabulates approximate critical levels for the definition of primary and secondary layer boundaries for both sand and clay type soils.

Similar to limits on ρ_I, these levels are not to be considered very stringently. These should only serve as a rough guide and aide for the purpose of layer boundary delineation and a tolerance of $\pm 10\%$ is similarly applicable. The criteria for the D^2 statistic has to be used in conjunction with the Intraclass Correlation Coefficient if any confidence is to be placed on the recommended procedure of identifying layers and classifying them as primary and secondary.

2.7.3   Combined Critical Limits Based on ρ_I and D^2

The multivariate analysis contains more information than the univariate analysis because the D^2 statistic considers the combined effects of the variables including their correlations. The results from the multivariate analysis therefore, deserves a higher recognition when discrepancies between the two methods occur. The results of the applications of these methods to the three sites indicated discrepancies at times, with the D^2 statistic suggesting the presence of a primary layer boundary while ρ_I revealed a secondary layer boundary or vice versa, although both of these types of events were the exception, rather than the rule. In such situations the result obtained from the multivariate analysis should be given a higher weight due to the higher information content it carries. In all types of incidents similar to the above, it

Table 2.9: Levels of the D^2 Statistic for the Definition of Primary and Secondary Layer Boundaries.

Data File	Soil Type	Maximum Level 1	Maximum Level 2
Mc Donald Farm	Sand	20 - 25	15 - 20
Tilbury Island	Sand	20 - 30	8 - 15
Laing Bridge	Sand	20 - 30	10 - 20
Mc Donald Farm	Clay	none	8 - 12
Haney	Clay	15 - 20	8 - 12
Strong Pit	Clay	15 - 20	6 - 10
Langley	Clay	15 - 20	8 - 15
Langley	Sand	20 - 40	none

Table 2.10: Critical Levels of the D^2 Statistic for the Definition of Primary and Secondary Layer Boundaries.

Soil Type	Range for Primary Boundary	Range for Secondary Boundary
Sand	$D^2 \geq 20$	$20 > D^2 \geq 10$
Clay	$D^2 \geq 12$	$12 > D^2 \geq 7$

was found that when the multivariate statistic attained a peak the univariate also did so, although one may have a magnitude representing a primary layer boundary while the other may predict a secondary layer boundary. In such extreme situations the following guideline can be used, so that the process of layer boundary discrimination is consistent.

(a) If the D^2 statistic indicates the presence of a primary layer boundary, verify whether ρ_I indicates at least a secondary layer boundary. If so, the above depth is a primary layer boundary.

(b) If the D^2 statistic indicates the presence of a secondary layer boundary, this depth will indicate a secondary layer boundary, irrespective of the boundary suggested by ρ_I.

2.8 Conclusions

The three examples described above, provide sufficient evidence that statistical methods should be employed in identifying layer boundaries. These statistical methods using univariate (Intraclass Correlation Coefficient and T Ratio) and multivariate (D^2) methods have a sound fundamental basis for discriminating between different

soil layers. The main conclusions are as follows;

(i) The primary layer boundaries from the statistical methods agree with the main layering obtained from the CPT classification chart. The main advantage of the statistical methods being the capability of picking depths devoid of the subjectivity involved in the latter. In addition, the proposed statistical techniques have the capability of picking secondary layers which are commonly known as sublayers in geotechnical engineering.

(ii) The Intraclass Correlation Coefficient (Univariate Analysis) and D^2 (Multivariate Analysis) are robust statistics in that the primary layer boundaries are not highly sensitive to small changes in the window width.

(iii) The proposed critical limits for ρ_I and D^2 are recommended to be used to differentiate between primary and secondary layer (main layers and sublayers) boundaries. For the D^2 statistic, different sets of limits have been recommended for sand and clay type soils due to the dissimilarity of their correlation structure. In the event of contradicting results for the ρ_I and D^2 statistics, the latter should be given priority due to its relatively higher information content. It has to be emphasized that the particular soil has to be identified as predominantly clay or sand, prior to arriving at conclusions based on the value of the statistic concerned.

(iv) In rare situations (Type I and Type II profiles) where ρ_I and D^2 fail in discriminating between soil layer boundaries, the gradient method is recommended. For the gradient method to be valid, the profile should necessarily exhibit a trend which can be verified by regression analysis.

In conclusion, the proposed statistical methods are recommended to be used in conjunction with the results of the CPT chart in order to obtain layer boundary depths with minimum amount of subjectivity.

Chapter 3

Trend Analysis

3.1 Introduction

Trend Analysis is used to describe large scale variations of a variable or group of variables in space and can be solved using the method of regression. Simple regression is used to describe the variation of a single variable while multiple regression is used when a group of variables is considered. Soil properties are highly depth dependent and the procedures of trend analysis will enable to evaluate the pattern of this dependency.

It is often advantageous to separate the spatial variation of a geologic variable into two or more components. If systematic changes in the average or mathematical expectation exist, the main component will be the trend. Deterministic functions such as polynomials can be employed to represent a trend. The second component is the residual and is generally treated as random. In general, the data can be expressed as,

$$DATA = TREND + RESIDUAL \qquad (3.1)$$

In any type of soil profile, a concern of great interest is the evaluation of stationarity, since once it is established, many statistical analyses can progress from there onwards. The presence of a trend or non-stationarity often is not apparent from a visual inspection, and a statistical test such as the RUN test may be used to verify this condition.

Trend removal using regression methods is the most widely used technique to

obtain stationary residuals. Data can also be made stationary by differencing which is widely used in Time Series methods and hence will be described in Chapter 5 of the thesis.

Smoothing and filtering techniques would also enhance the identification of trends. Smoothing essentially removes high frequencies from a data set and results in a more uniform profile. Statistical filtering would remove extremities or anomalies in data, enabling easier visualization of trends in profiles.

3.2   Smoothing and Filtering of Cone Profiles

Filtering is performed to eliminate extremities of data in order to identify trends more accurately. It is important that the process of filtering removes only distinct anomalies and does not remove thin layers. Filtering also requires engineering judgement and the particular method of filtering adopted is highly situation dependent. A thin anomaly of high bearing or strength in a soft clay could be removed without jeopardizing the design, since the strength availability for the foundation will not depend on this thin layer. However, the awareness of a thin layer of high strength, if present, might be important in the determination of the driveabilty of a pile. This emphasizes how situation dependent filtering is and the kind of engineering judgement the process of filtering requires.

3.2.1   Smoothing

The present literature is not clear as to the difference between filtering and smoothing. It is the author's opinion that in geotechnical engineering, processes such as three point, five point and seven point moving averages are methods of smoothing similar to Fourier smoothing. The procedures of autoregressive (AR), moving average (MA) and autoregressive moving average modeling (ARIMA) in time series analysis, also

fall into the category of smoothing. Techniques of smoothing, as the name indicates, smooths out a profile by removing the high frequency content and alter the entire set of data. Smoothing methods will even cause the 'good' and acceptable data to be modified, which of course is not desirable. In contrast to these methods, 'filtering' would filter out only the anomalies which fall outside a selected window width while the data inside this window remain unchanged.

3.2.1.1   Moving Average Smoothing

Methods of three point, five point and seven point smoothing would consider data in groups of three, five and seven, respectively, and the output of the complete data set would be modified; the smoothness of the profile increasing with the number of data considered in a group. In other words, the seven point smoothing would result in the smoothest curve of the above three methods while the three point smoothing would result in the least smooth curve.

A typical smoothing equation of degree m, for any point 'i' is expressed as follows;

$$Q_{fi} = a_{i-m}Q_{i-m} \ldots \ldots + a_{i-2}Q_{i-2} + a_{i-1}Q_{i-1} + a_iQ_i$$
$$+a_{i+1}Q_{i+1} + a_{i+2}Q_{i+2} \ldots \ldots + a_{i+m}Q_{i+m} \tag{3.2}$$

where a_i's are the coefficients with $\sum_{i=1}^{m} a_i = 1$.

For example, for three point smoothing, $a_i = 0.5$ and $a_{i-1} = a_{i+1} = 0.25$. For five point smoothing $a_i = 0.4$, $a_{i-1} = a_{i+1} = 0.2$ and $a_{i-2} = a_{i+2} = 0.1$ with the weighting of the coefficients being inversely proportional to the distance from the centered point 'i'. There is also the more simplistic version of moving average smoothing where the smoothed value is equal to the simple average of the values around 'i' which are grouped together. It is evident from Eq. 3.2 that Q_{fi} would replace Q_i in the smoothed profile irrespective of whether it is an anomaly or not. It is this adversity of losing

genuine and reliable data to modified values that renders the technique of moving average smoothing unsatisfactory. If the intention of the smoothing is merely to have a clearer picture of the profile, it may be acceptable, but is not suitable if further analysis is to be made using the smoothed profile. The moving average smoothing procedure is illustrated in Fig. 3.1 for different degrees of smoothing MA, with MA = 1 referring to the raw cone data. These clearly indicate how the smoothness of the profile increases with increased degree of smoothing, MA.

3.2.1.2 Fourier Smoothing

Fourier smoothing, or more appropriately Fourier Transform of a data set transforms the entire data set. Fourier analysis is a technique whereby the profile or curve is expressed as a sum of sinusoids (sine and cosine curves) of varying number of harmonics. If a data set consists of 'n' number of points, the transform with exactly n/2 harmonics if n is even, and (n+1)/2 if n is odd would produce the original profile almost exactly. Decreasing the number of harmonics would result in a smoother profile. This is a procedure where high frequencies are removed from the profile, retaining the lower harmonics. Figure 3.2 illustrates the effect of Fourier transforming the data, where M is number of harmonics used with M = 800 referring to the original data profile with 1600 data points. Profiles with M equal to 200, 100 and 50 are the transformed profiles which have used 200, 100 and 50 harmonics respectively. The increased smoothness with decreasing harmonic number is apparent from the figures. The effect of the removal of the high frequency content is illustrated in Fig. 3.3 which shows the spectral density function (Bendat and Piersol, 1971) at different frequencies. The entire region, comprising of zones A, B and C, depicts the spectra of the unsmoothed original data, with regions A and B representing the spectra with the high frequencies removed (MA $\leq$ 200). Zone A represents the spectra of
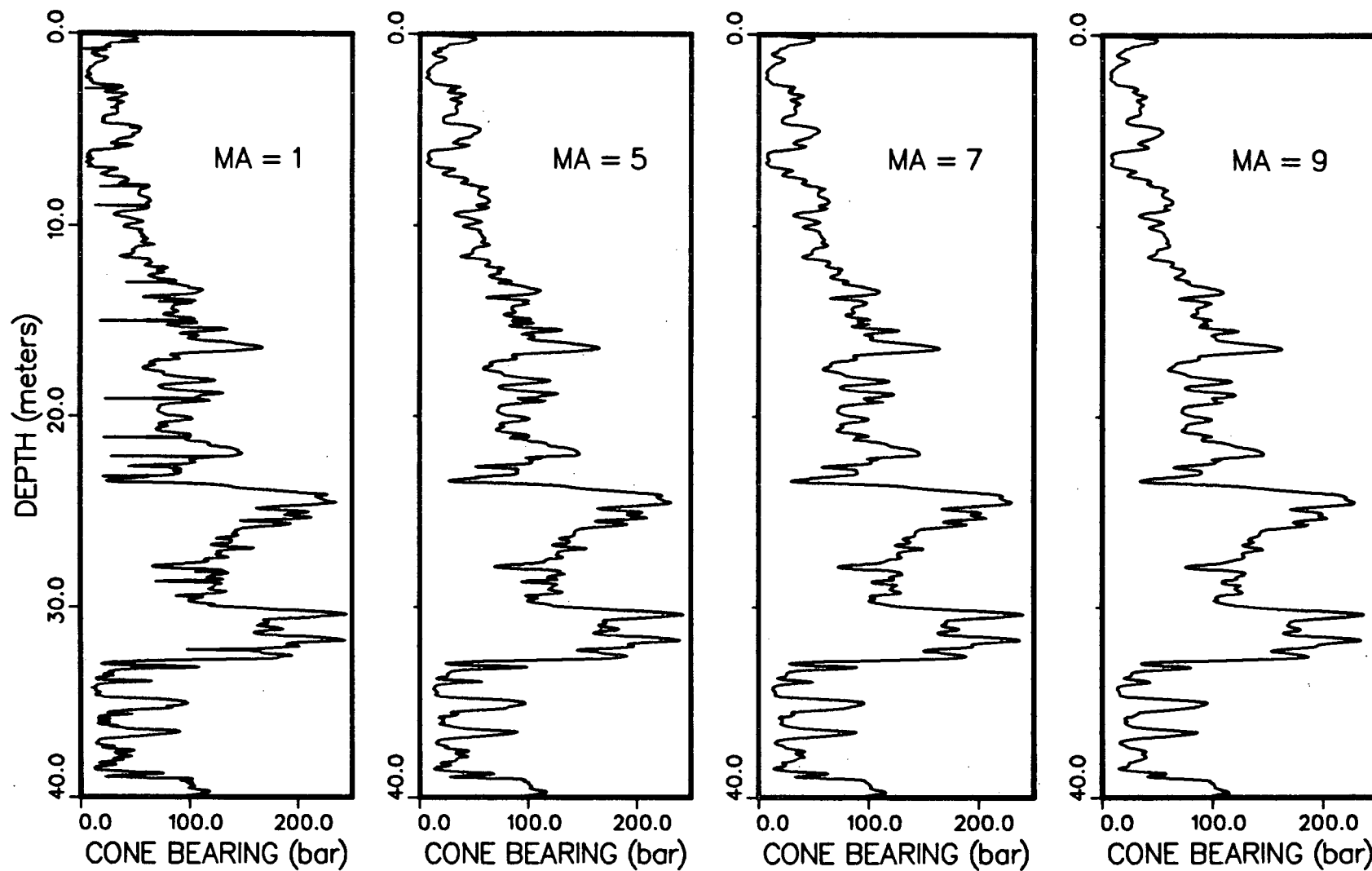
Figure 3.1: Cone Bearing Profiles at Tilbury Island after Moving Average Smoothing with MA = 1, 5, 7 and 9.
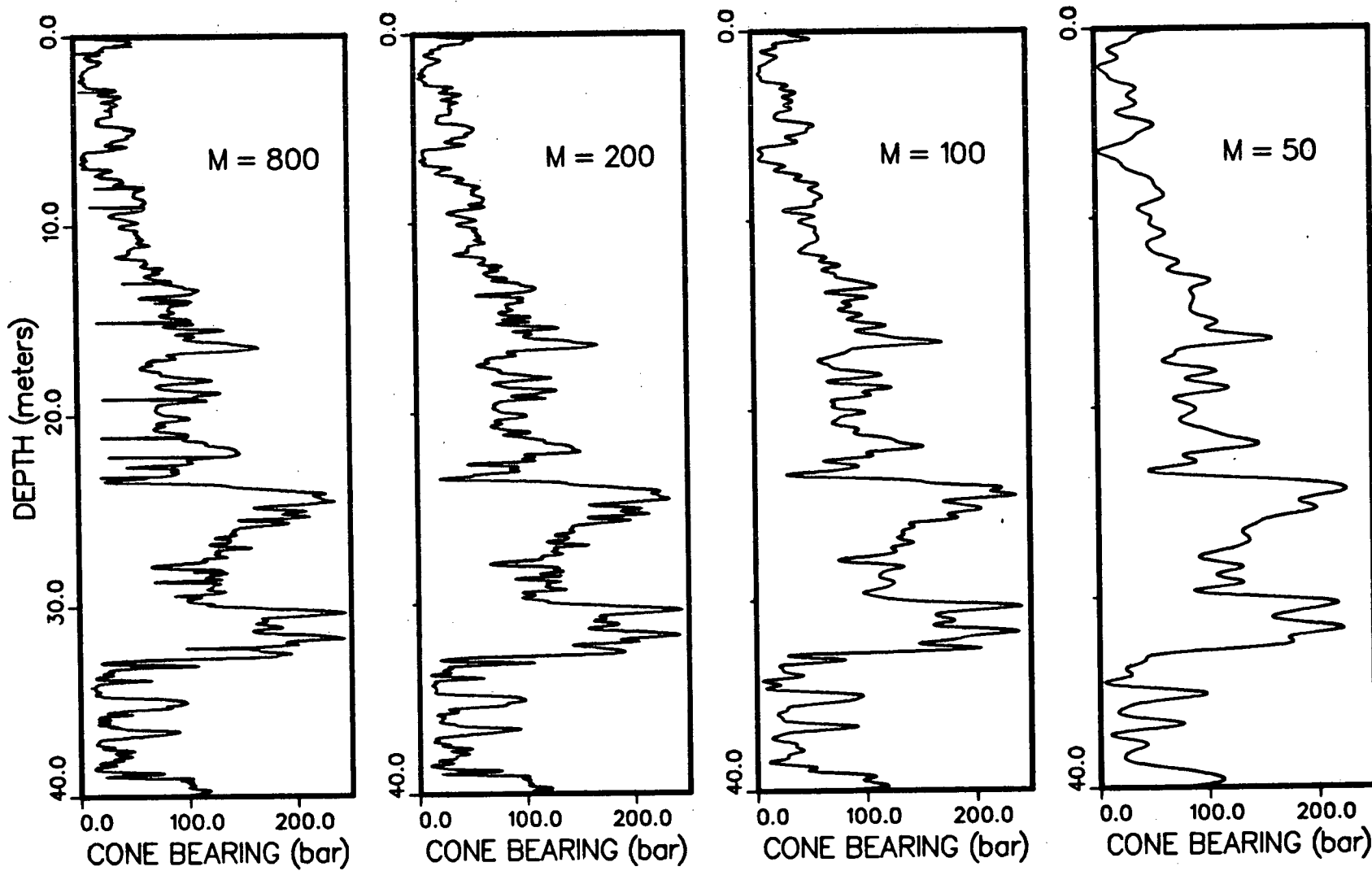
Figure 3.2: Cone Bearing Profiles at Tilbury Island after Fourier Smoothing with M = 800, 200, 100 and 50.
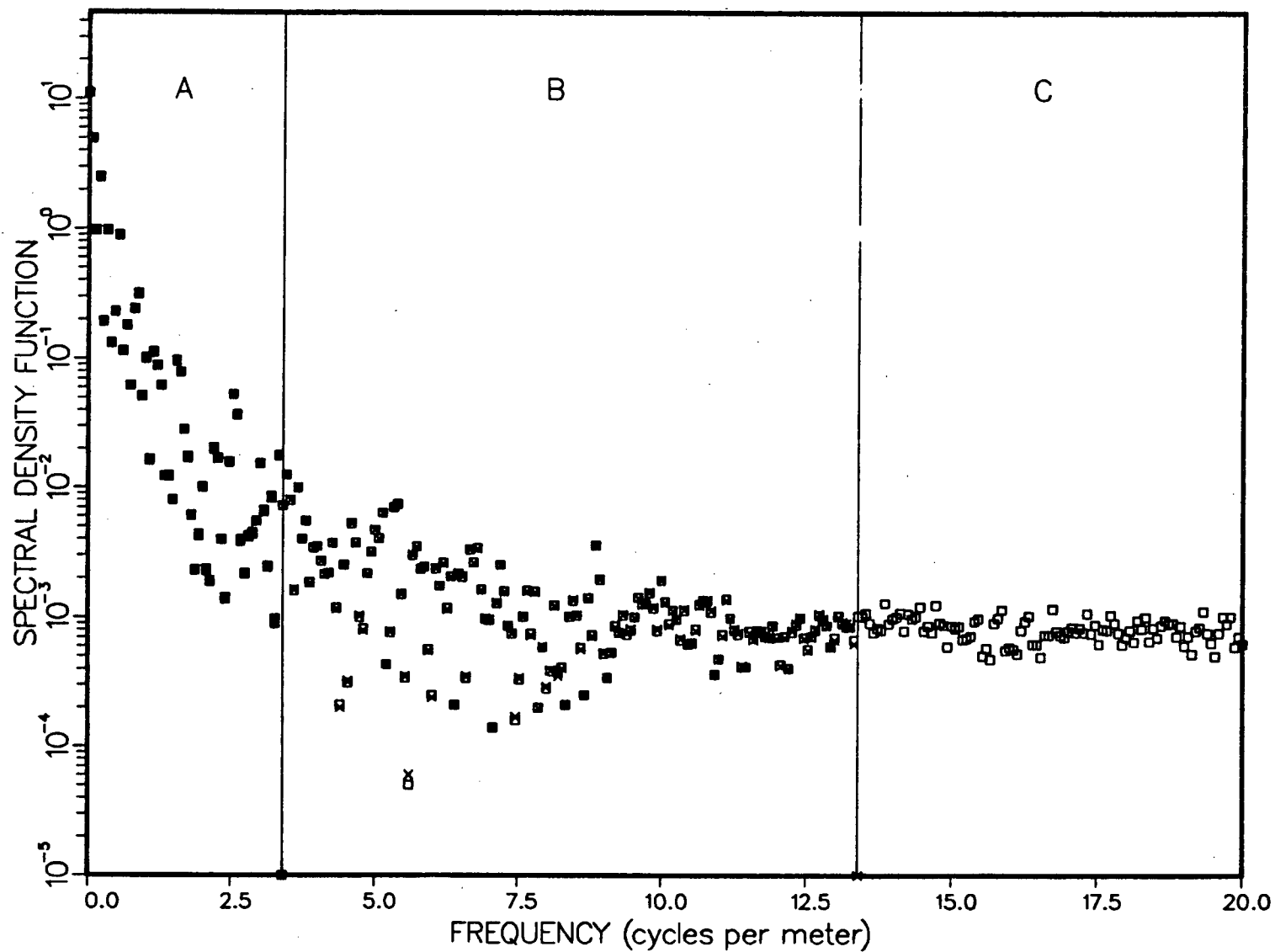
Figure 3.3: Spectral Density Function of Fourier Smoothed Data.

profile with $M \leq 50$ with only the lowest frequencies. This clearly illustrates how the higher frequencies are removed with decreasing number of harmonics employed in the smoothing process.

3.2.1.3 Autoregressive Integrated Moving Average Smoothing

The Autoregressive Integrated Moving Average (ARIMA) process can also be used for smoothing. This is essentially a process where a value at a point is expressed as a function of the properties at adjacent points (similar to other methods of smoothing) and suffers from the same shortcomings. This procedure is described in detail in Chapter 5 and hence will not be elaborated upon here.

3.2.2 Statistical Filtering

Different methods of statistical procedures can be used to filter out extremities from a CPT profile. In the simplest case, the soil profile can be divided into sub layers of some thickness and the statistics of each of these layers calculated. The statistics, namely the mean, median and the standard deviation, are used to develop an acceptance band for a given depth. The recommended procedure has several options for both the filtering procedure and the method of replacement of data points outside the band. It is well accepted that the value of bearing obtained from the CPT depends both on the immediately past and future values. Therefore, when considering a sublayer in which the cone tip is advancing, it is reasonable and logical to also consider the immediately adjacent sublayers.

The entire soil profile or layer is divided into several sub layers, the width of which is a variable. However, the width should not be too high to avoid the possibility of missing a definite layer by mis-recognizing it as an anomaly, and secondly, because

soil properties often exhibit a trend with depth. If the layers are too thick, that is each sublayer comprising of too many points, the data in that region may not be stationary and the ensuing statistics will be inaccurate. A sublayer of lesser thickness will alleviate this problem. It is also important that the sublayers not be too narrow to ensure that the bias of the statistics to be calculated are within acceptable limits of reliability. From the above explanation, it is evident that a compromise has to be reached as to the optimum thickness of a sublayer. An increased thickness of the sublayer can introduce the possibility of filtering out actual thin layers which is highly undesirable. It has to be reiterated that the selection of the sublayer thickness is situation dependent. If the intention of filtering is solely for the purpose of inspecting trends, a large value can be selected. Considering all the above requirements, ten data points, that is a thickness of sublayer of twenty five centimeters is recommended for purposes of removing extremeties in data, while a thickness in the region of twenty five (10 data points) to seventy five centimeters (30 data points) is a good choice for purposes of trend evaluation. The technique proposed in the thesis is a modification of the method given by Vivatrat (1978) and Campanella and Wickremesinghe (1987). The detailed procedure is described below;

(a) Select a band width (ten data points is a good choice)

(b) For a sublayer 'i', calculate the standard deviations of layers 'i', 'i-1' and 'i+1', given by σ_i, σ_{i-1} and σ_{i+1}, respectively. The first layer will not have a (i-1)th layer and the last will not have a (i+1)th layer.

(c) Obtain a representative standard deviation σ_{si} defined by σ_{ai} or σ_{bi}, whichever is lesser, where

$$\sigma_{ai} = \left(\sigma_{i-1}{}^2 + \sigma_i{}^2\right)^{1/2} \tag{3.3}$$

$$\sigma_{bi} = \left(\sigma_{i+1}{}^2 + \sigma_i{}^2\right)^{1/2} \tag{3.4}$$

(d) Calculate for layer 'i', the mean (m_{ai}) and median (m_{di}) considering all data in all three layers 'i-1', 'i' and 'i+1' as a group. For the first layer, only 'i' and 'i+1' will be considered while for the last layer, only 'i' and 'i-1' will be considered. The median of any data set is not dependent on the extremities of the data. That is, it is not affected by low or high values, since it is the middle term of an ordered data set. The mean, however, is a function of all the data and as a result, it can be severely affected by data which are high or low.

(e) Compute for layer 'i', the band width W_{bi} such that the data outside of it will be replaced or removed.

If the mean method is adopted,

$$W_{bi} = m_{ai} \pm (BS)\sigma_{ai} \tag{3.5}$$

If the median method is adopted,

$$W_{bi} = m_{di} \pm (BS)\sigma_{ai} \tag{3.6}$$

where BS is a constant to be decided depending on the degree of filtering required. $0.5 < BS < 1.5$ has been found to be a reasonable range with BS = 0.5 resulting in a high degree of filtering and a value of BS = 1.5 resulting in a low degree of filtering. As Eqs. 3.5 and 3.6 indicate, the band width W_{bi} will depend both on the method of filtering (mean or median) and the chosen value of BS. Since the mean

is a function of all the data in the layer, an extremely high or extremely value will affect its value and thereby also affect the band width, and this is not desirable. This effect will be prominent in a variable soil or even in a less variable soil if wider layer thicknesses have been chosen for the sublayers. However, in the case of a less variable soil or where sublayers are relatively low in thickness (around ten data points or 25 cm), this adversity is not of great concern as some of the results to be presented, will indicate. The median is not affected by extreme data points in a sample and hence the above problem does not apply if the median method of filtering is adopted. In general, for purposes of trend evaluation, a relatively low value of BS can be selected since losing thin layers is not of great concern. However, for purposes of filtering out extremeties, a high value of BS should be selected, alleviating the possibility of filtering out actual thin layers present in the profile.

(f) Replace or remove data outside the acceptance band W_{bi}.

Removal of data is an option but generally not recommended because it reduces the original data file creating practical difficulties. Replacement of data can be performed in several ways. Replacement by the mean of the immediately preceding and imme-diately following unfiltered data points is a good choice, since the substitution is totally dependent on the closest neighboring points which are within the acceptable limits. Other possibilities of substitution are by the mean or median of the sublayer in which the filtered data points occur. The latter option is not suitable if the local region under consideration has significant trends. It should be noted that removal or replacement of data will not change W_{bi} in Eqs. 3.5 and 3.6.

If thin layers are thought to be present, it is advisable to use the mean method, since the possibility of such thin layers being removed is avoided. Another safeguard against such an adversity is to reduce sublayer thickness, but under no circumstances

should it be less than eight data points for reasons already explained. The extent of filtering can be expressed by PF (as a percentage) which is given by,

$$PF = \frac{N1}{NUM} \times 100 \tag{3.7}$$

where N1 is the number of data outside the acceptance band and NUM is the total number of data in the profile. A high value of PF indicates that a significant number of data have been filtered and a low value reflects the opposite.

Statistical filtering has been performed on data from Tilbury Island for both the mean and median methods. The method of replacement used is by substituting the mean of the adjacent unfiltered data points. Figures 3.5 - 3.8 illustrate the filtered data profiles of Fig. 3.4. As can be observed, for an intense filtering of BS = 1.0, the PF value obtained for the median method is 20.12 (Fig. 3.5) and 25.81 for the mean method (Fig. 3.6). The higher value of PF for the mean method is also evident when BS = 1.5, by comparing Figs. 3.7 and 3.8. For the higher value of BS (=1.5), PF drops to 9.50 for the median and to 10.37 for the mean method, indicating a reduced degree of filtering.

3.2.3 Evaluation of Stationarity of a Soil Layer

For various reasons statistical applications like trend analysis, determination of parameters like the scale of fluctuation (Chapter 4), time series analysis (Chapter 5) and for the interpolation problem considering correlations (Chapter 6), it is important to evaluate whether a particular soil layer is stationary in the mean. Statistical methods such as autocorrelation and variogram analysis are also performed on stationary data and there is a significant difference between the above functions for stationary and non stationary data. Figures 3.9 and 3.10 illustrate the McDonald Farm site cone bearing profiles which will be tested for stationarity using the RUN test. Figures 3.11 and 3.12 illustrate the different autocorrelation and variogram functions obtained for
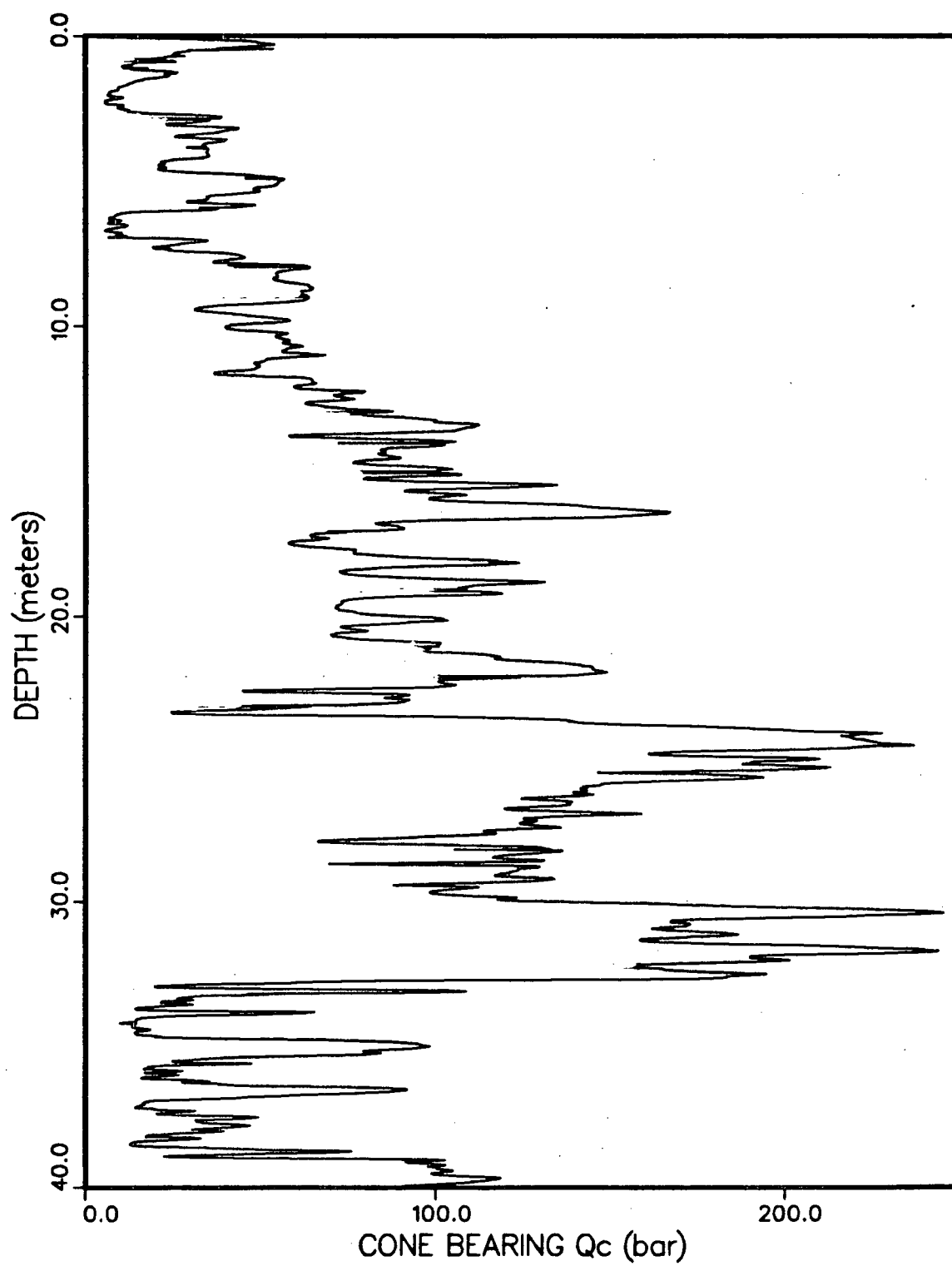
Figure 3.4: Cone Bearing Profile at Tilbury Island.
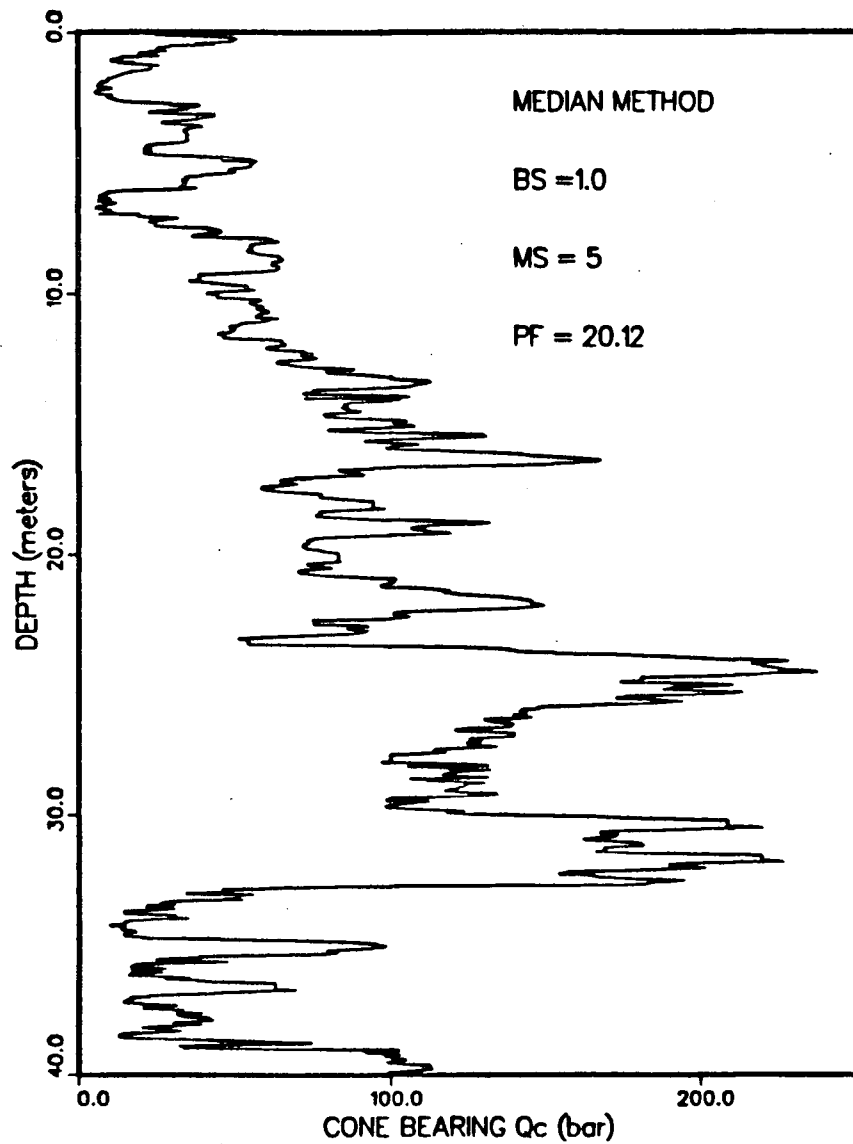
Figure 3.5: Statistically Filtered Cone Profile using the Median Method with BS = 1.0 and MS = 5.
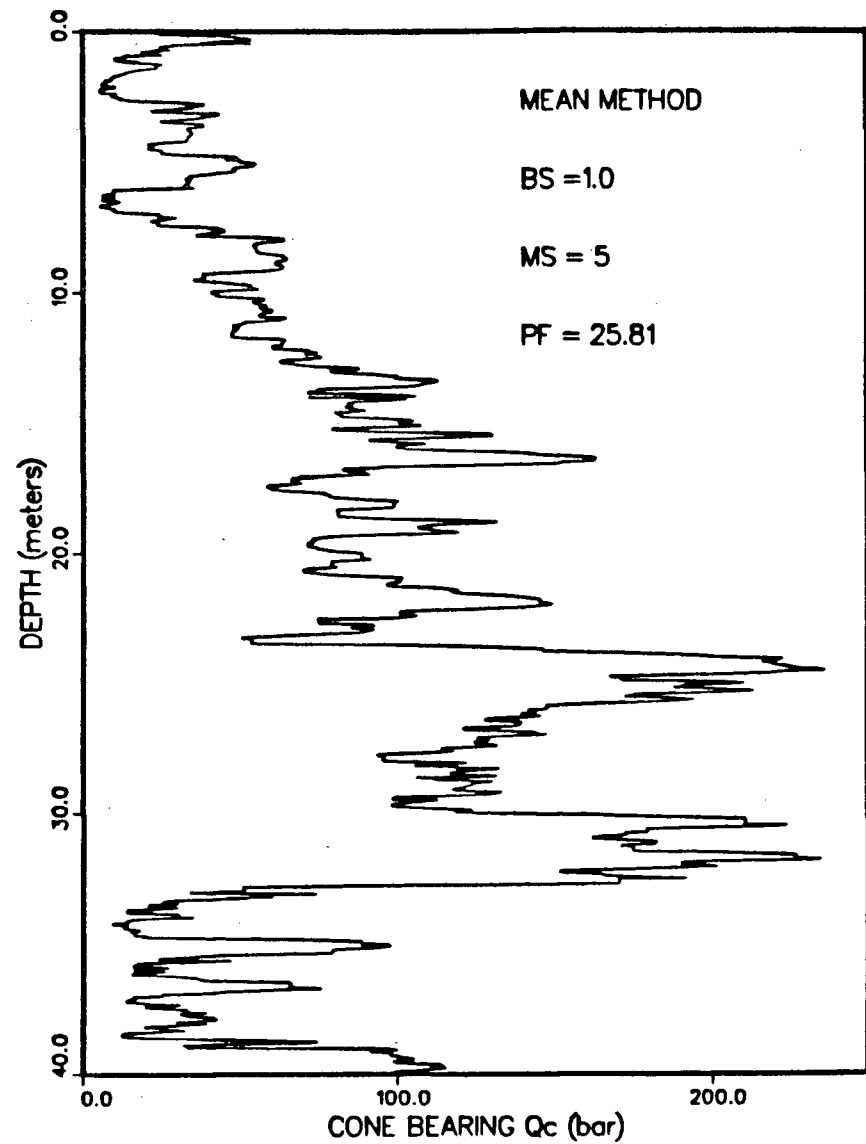
Figure 3.6: Statistically Filtered Cone Profile using the Mean Method with BS = 1.0 and MS = 5.
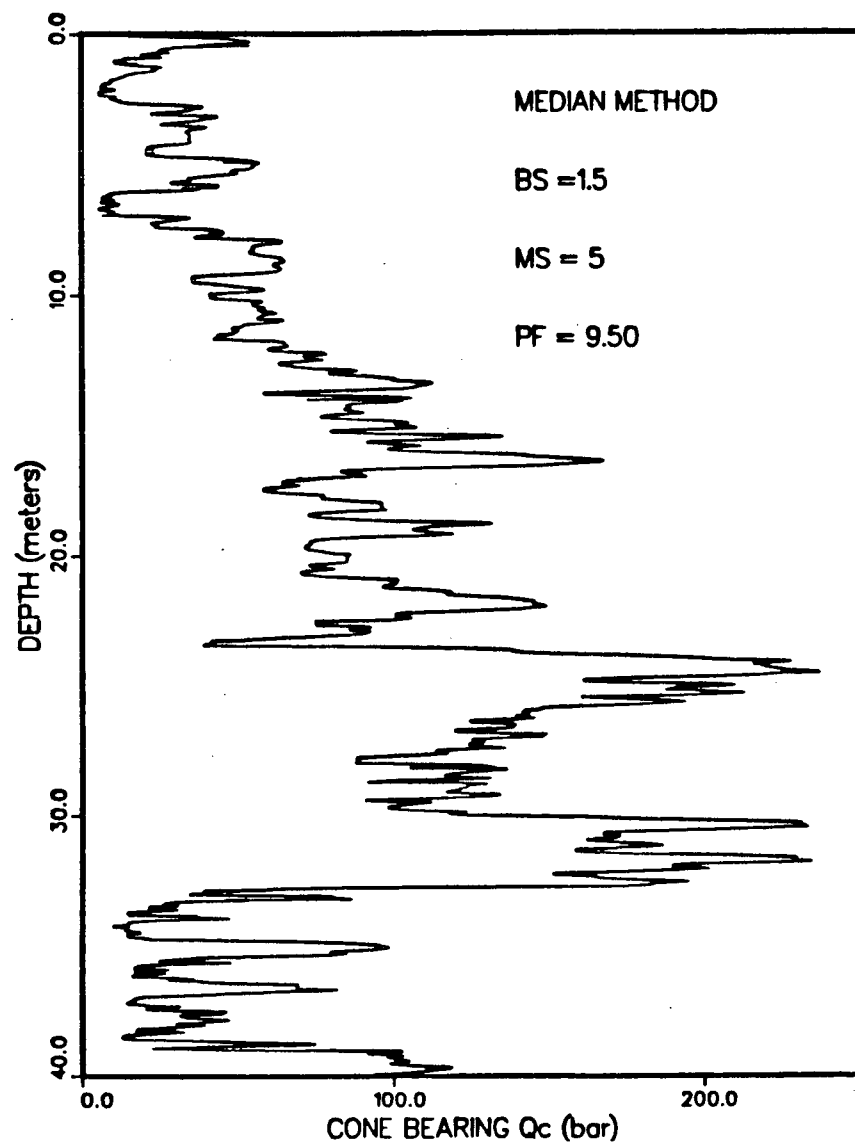
Figure 3.7: Statistically Filtered Cone Profile using the Median Method with BS = 1.5 and MS = 5.
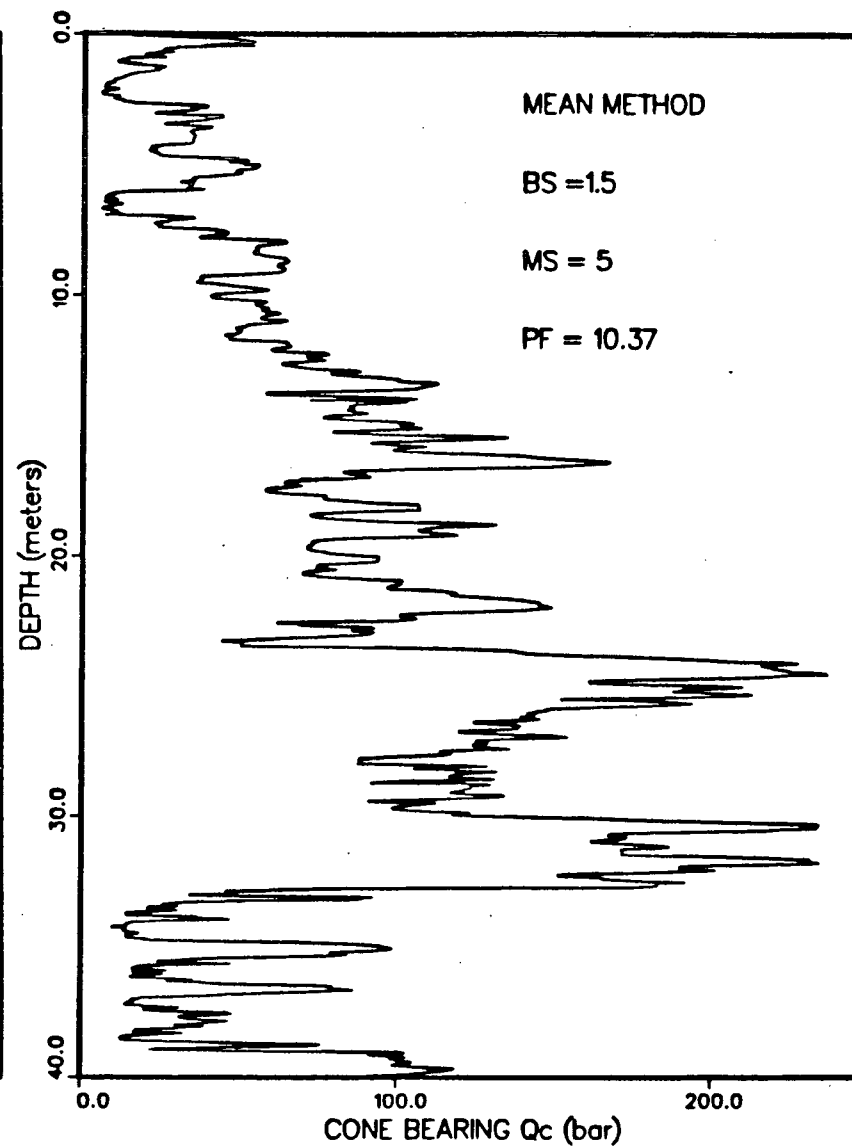
Figure 3.8: Statistically Filtered Cone Profile using the Mean Method with BS = 1.5 and MS = 5.
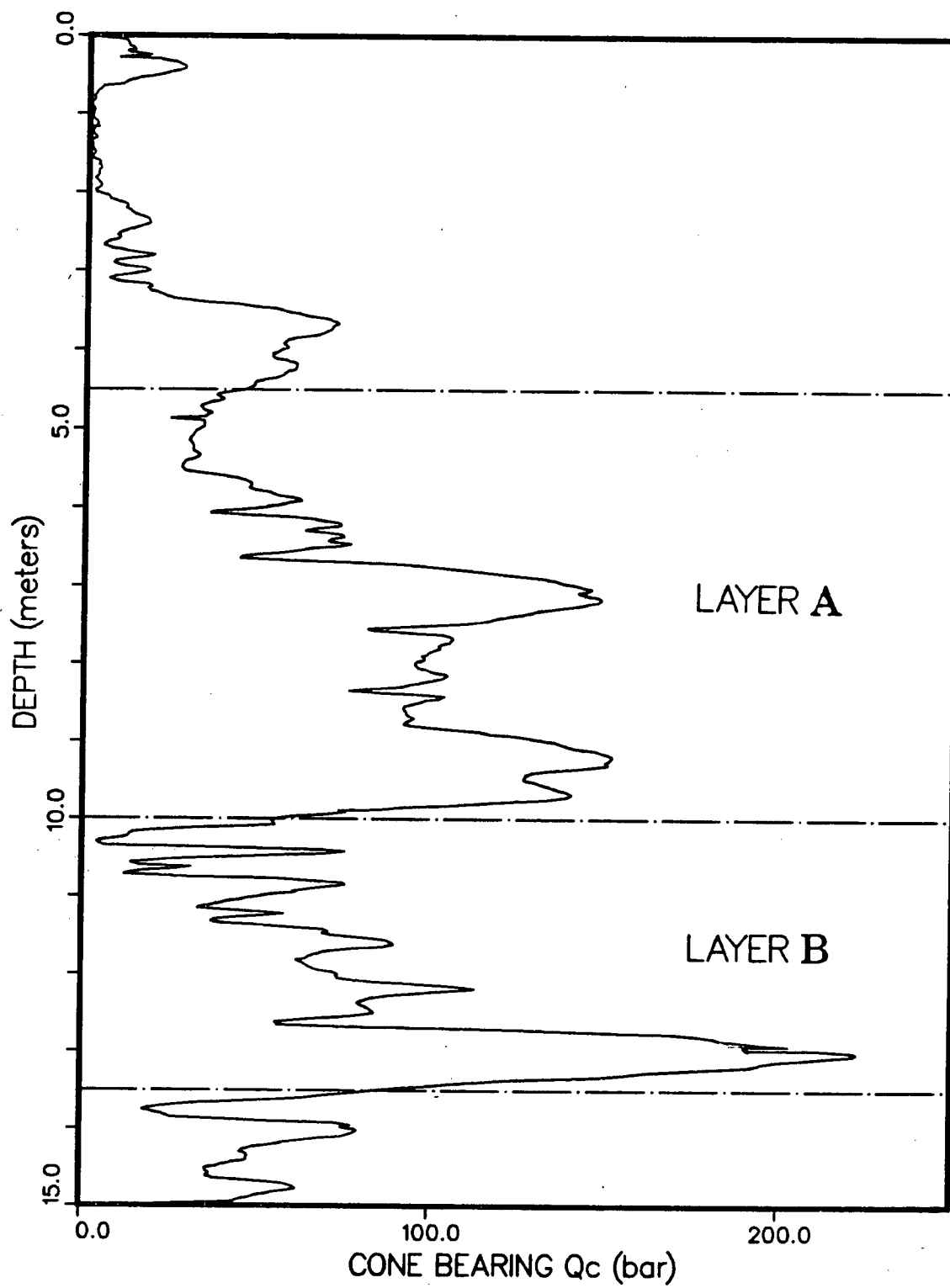
Figure 3.9: Different Layers in 15 meter Cone Bearing Profile.
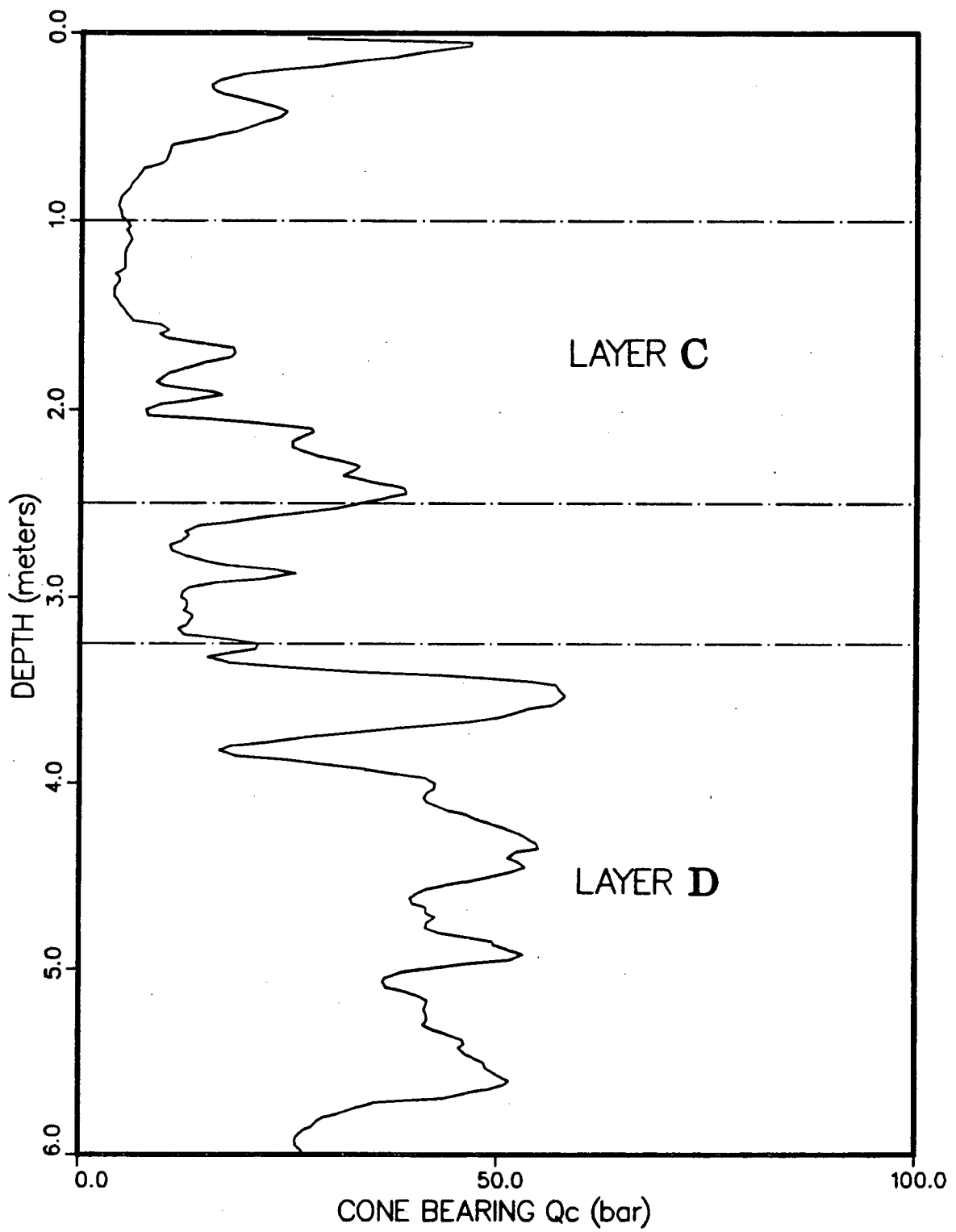
Figure 3.10: Different Layers in 6 meter Cone Bearing Profile.
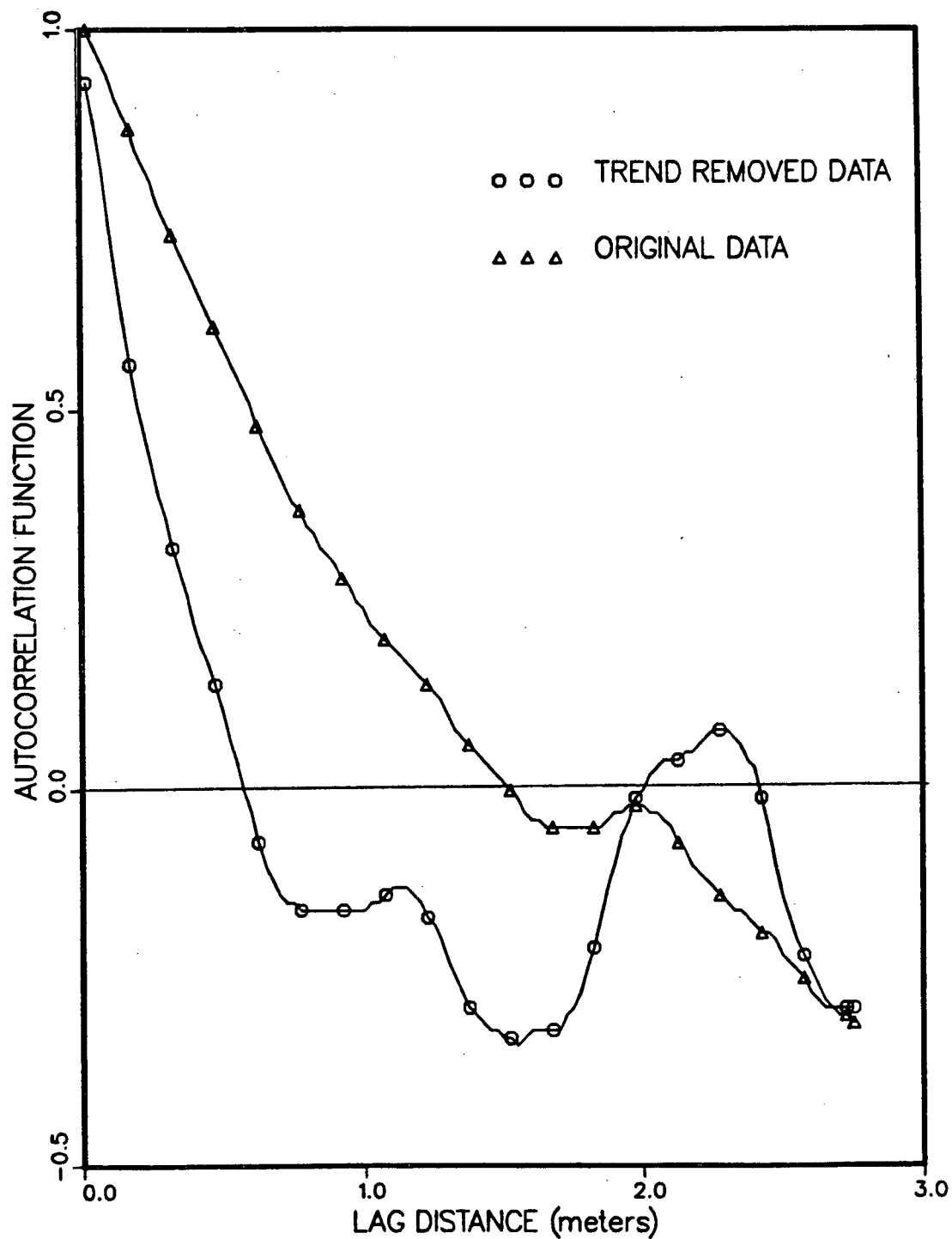
Figure 3.11: Autocorrelation Function of Original and Trend Removed Data of Layer A.
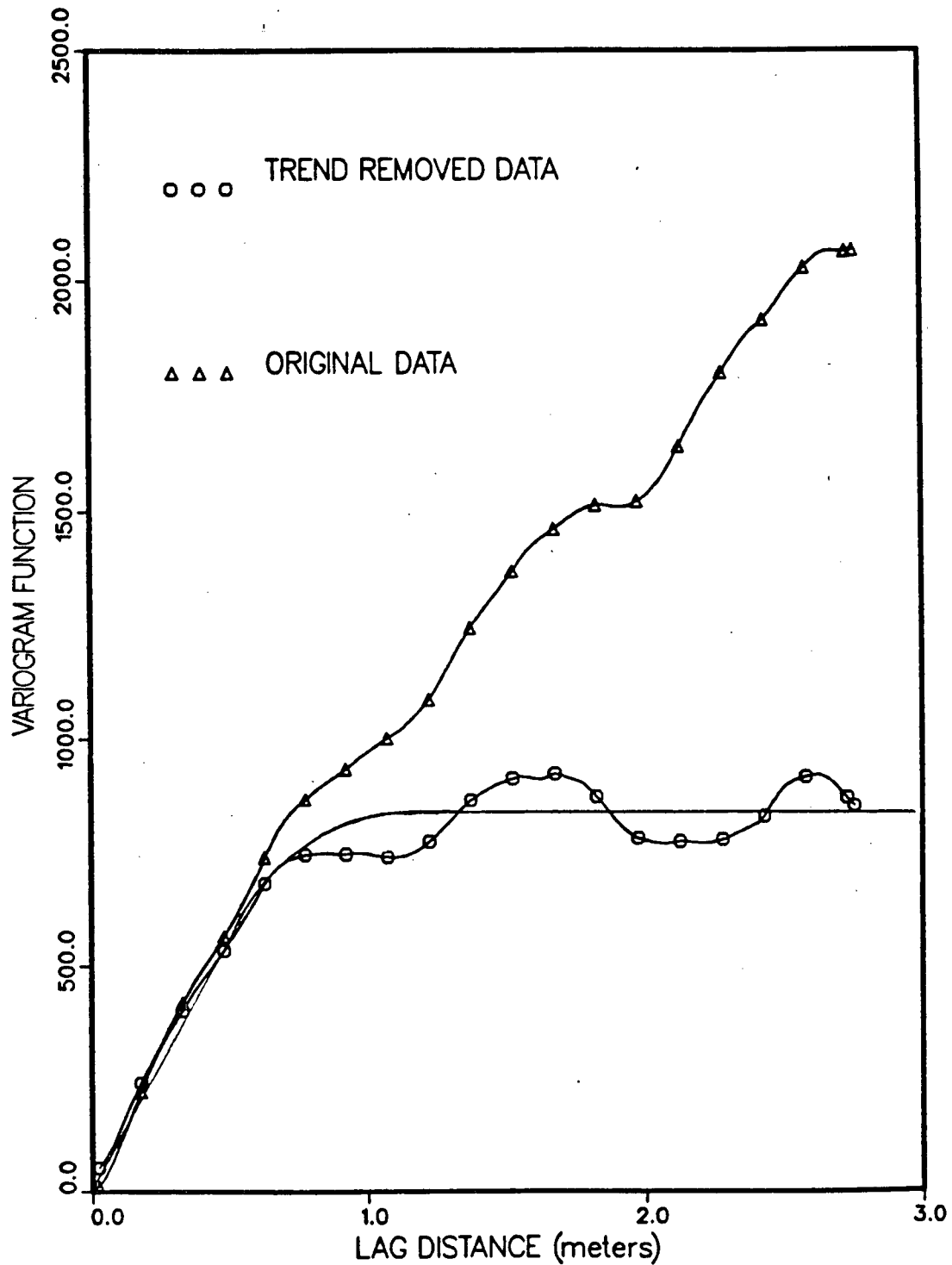
Figure 3.12: Variogram Function of Original and Trend Removed Data of Layer A.

the Layer A data in Fig. 3.9 for both stationary (trend removed) and non-stationary data. These two figures illustrate the importance of removing the trend, if there is one present, prior to any analysis. Methods such as trend removal, described in this chapter and techniques of differencing explained in Chapter 5 can be used to stationarize data. Although a visual inspection of a profile may at times give an indication of the stationarity or non-stationarity, it is not always the case, and in instances such as above, the RUN test becomes a convenient and simple tool in assessing it.

RUN is defined as a sequence of events of the same type (Bury, 1975). The criterion chosen here is comparing the mean of a selected sublayer against the global mean or the overall mean of the entire layer. There are two type of events possible: the local mean of the selected thickness being above the global mean, and the local mean being below the global mean. A sequence of events where the local mean is evaluated to be above or below the global mean is termed a RUN. A similar test can also be done for the standard deviation, but is not recommended due to the reason that the standard deviation is a second moment statistic. The reliability of second moment statistics on small samples is low due to the increased variability as compared to a large sample. In the situation under study, the entire soil layer is the large sample whose variability is not adequately represented by the smaller sublayers consisting of fewer data points.

In the applications to follow, the sublayers were selected to comprise ten data points, that is, a sublayer thickness of 25 cm. It is important that these sublayers are not too thick, so that each sublayer could be assumed to be approximately stationary and not too narrow, since if it is, there will be too little data rendering it impossible to get good estimates of statistics such as the mean. Considering the above factors, ten data points seemed to be a reasonable choice. The number of RUNS with respect to the mean and standard deviation can be determined but as mentioned before, the RUN test based on the standard deviation is not recommended due to the reason

that the standard deviation of the sublayers will not be representative of the standard deviation of the entire layer.

Once the number of RUNS with respect to the mean is determined, it is compared to the values in the Tables of the RUN test (Swed and Eisenhart, 1943) where the number of RUNS required for stationarity or homogeneity for different significance levels are tabulated. More details of the RUN test are available in Bury (1975), Alonso and Krizek (1975) and Campanella and Wickremesinghe (1987).

The RUN test has been performed on two sets of data, the profiles of which are given in Figs. 3.9 and 3.10. All layers depicted in the figures have been identified using methods described in Chapter 2. The detailed results of Layer A of Fig. 3.9 are tabulated in Table 3.1 and the summarized results of Layers A and B are tabulated in Table 3.2. The m and n values in the tables refer to the number of values above and below the global mean and they are interchangeable.

As can be seen from Table 3.2, Layer A fails the test for stationarity at all levels of significance of 1%, 5% and 10%. Layer B fails the test for stationarity at significance levels of 5% and 10% and is at the border of acceptance even at a low level of significance of 1%. These results are as expected for a profile exhibiting a trend. Figure 3.13 illustrates the distribution of RUNS of Layer A with respect to the global mean of that layer, 87 bar.

The detailed results of Layer D of Fig. 3.10 are tabulated in Table 3.3 while summarized results of both layers C and D, in Table 3.4. Layer C data is non-stationary due to the presence of a trend while the Layer D data is stationary at all levels of significance since it does not possess a trend and is clearly homogeneous. Fig. 3.14 illustrates the distribution of RUNS for the Layer D data of Fig. 3.10. The Layer D data have also been used for the interpolation problem described in Chapter 6 of the thesis, and unlike Layer C, there was no need to consider the cone bearing
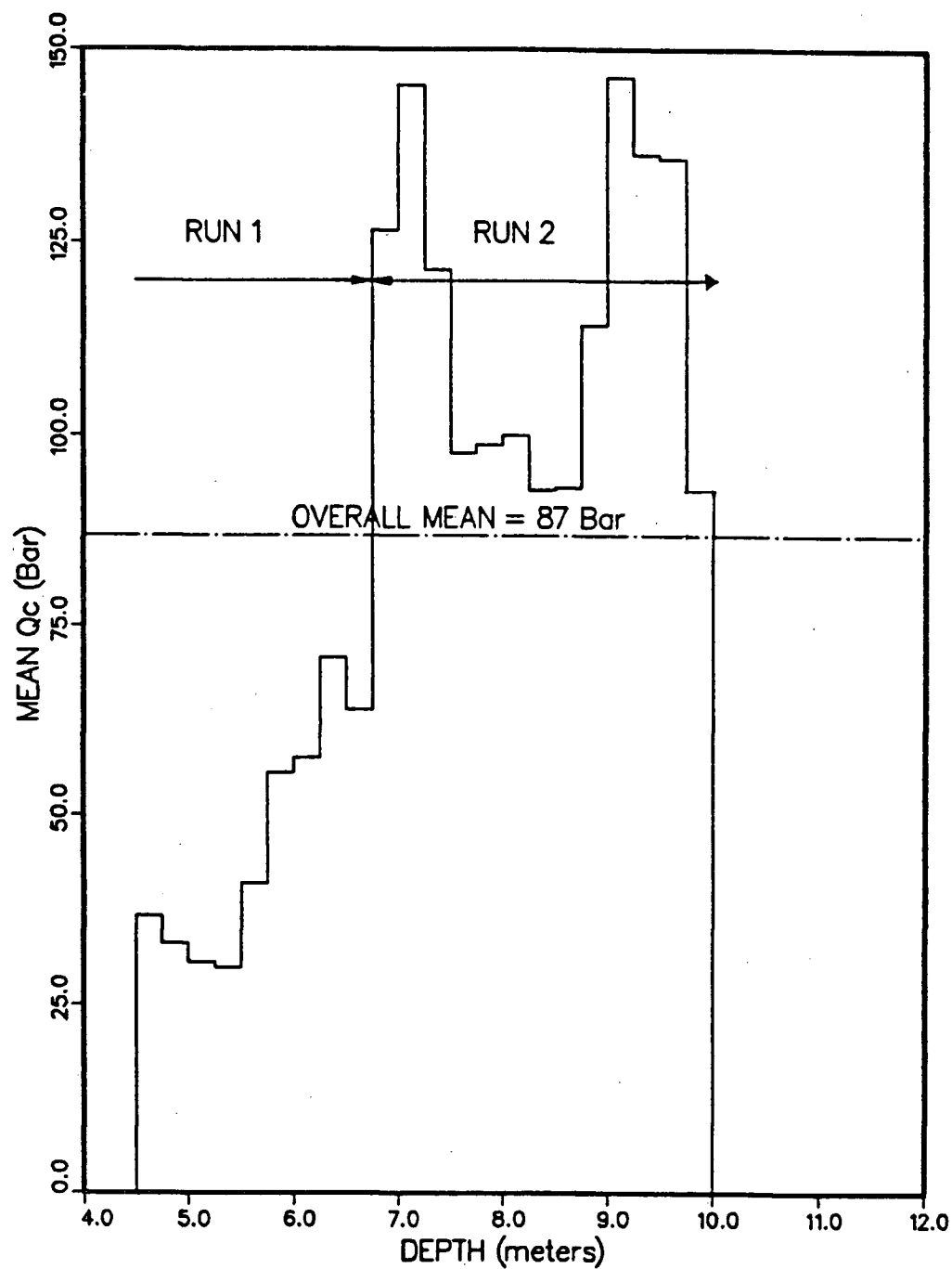
Figure 3.13: Distribution of RUNS for Data of Layer A.
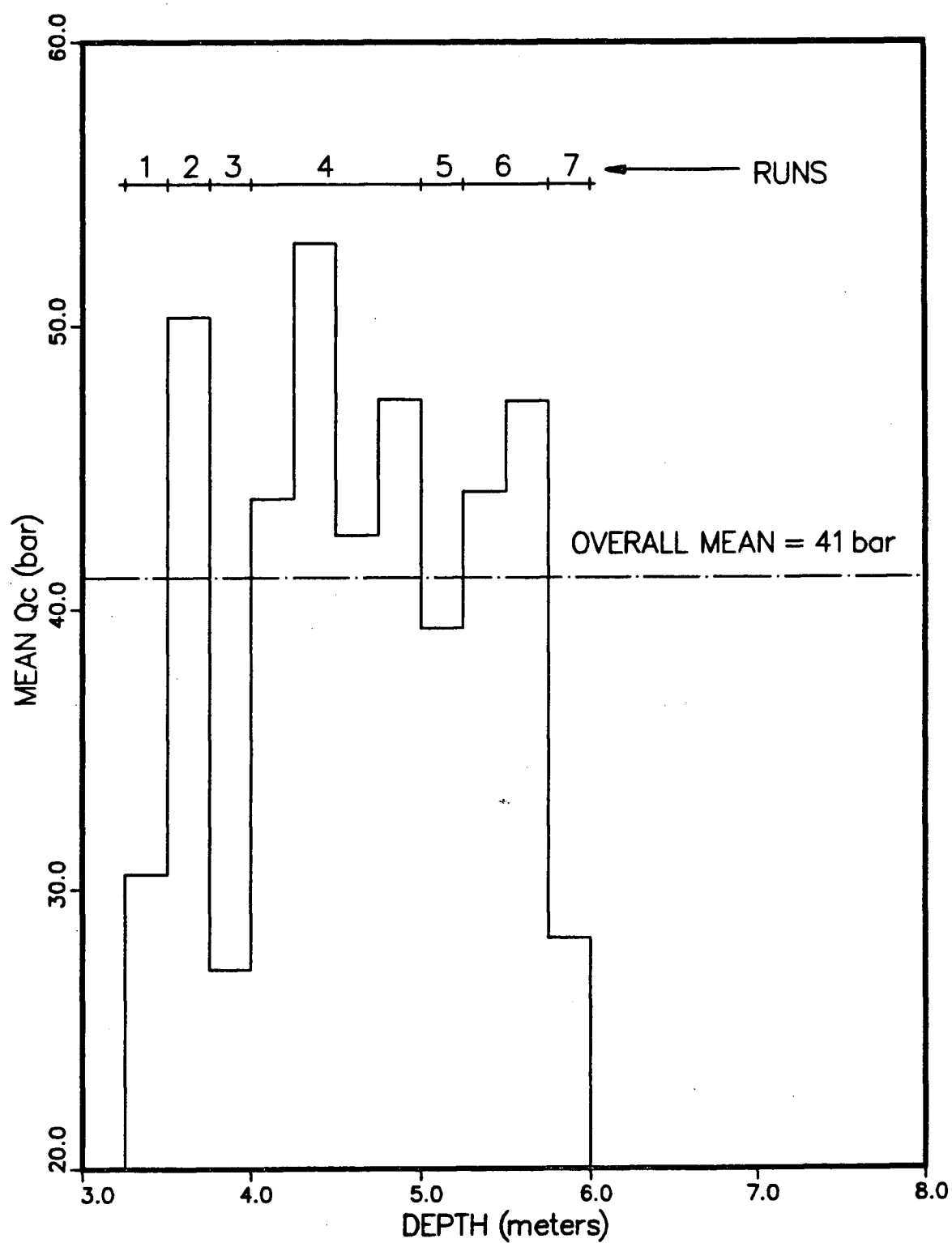
Figure 3.14: Distribution of RUNS for Data of Layer D.

value in two parts; namely the stationary residual and the non-stationary trend. This is an ideal example where the RUN test was used to verify the stationarity of a soil layer.

As Fig. 3.12 illustrates, stationarity or non-stationarity can also be verified using the variogram which is defined in section 6.2. The variogram (or semi-variogram) function attains a constant value (sill) for a stationary set of data. For non-stationary data the variogram function is of a continuously increasing nature (Fig. 3.12). However, this method suffers from the deficiency that particular levels of significance of acceptance or rejection of stationarity can not be established. In contrast, the RUN test affords this capability and the acceptance levels of stationarity can be based on the problem at hand and the required confidence.

The above examples illustrate the applicability of the RUN test and is a convenient method to determine the homogeneity of a soil layer. Once this condition is verified, and if found to be non-stationary, the data can be made stationary by the popular method of trend analysis or by differencing. Different applications use different methods of stationarizing data. For example, while methods of time series analysis use differencing it is more convenient to use trend analysis for soil property interpolation considering correlations. These will be described in Chapters 5 and 6 respectively.

Table 3.1: Results of Run Test for Layer A Data in Fig. 3.9.

Sub - Region	Mean	Standard Deviation
1	37.89	3.57
2	33.15	3.01
3	30.76	0.84
4	30.22	1.71
5	39.04	7.53
6	55.15	4.81
7	55.59	13.51
8	70.57	3.82
9	61.45	13.44
10	121.99	14.07
11	145.48	2.15
12	125.46	9.40
13	97.44	9.09
14	99.50	2.24
15	99.89	3.41
16	92.76	8.52
17	93.46	1.25
18	110.01	14.33
19	144.75	6.21
20	138.72	10.13
21	135.21	4.61
22	92.90	23.46
Entire Layer	87.0	39.5

Table 3.2: Comparison of the Actual Number of Runs (Mean) with the Number of RUNS Required for Stationarity for Different Levels of Significance for Layer A Data in Fig. 3.9.

Layer	RUNS Obtained for Data	m	n	RUNS Required for Stationarity		
				90%	95%	99%
A	2	13	9	7 - 15	6 - 16	5 - 17
B	2	5	9	4 - 10	3 - 11	2 - 11

Table 3.3: Results of Run Test for Layer D Data in Fig. 3.10.

Sub - Region	Mean	Standard Deviation
1	30.53	14.98
2	50.31	7.99
3	27.10	8.21
4	43.94	2.75
5	52.91	1.22
6	42.63	2.94
7	47.43	4.12
8	39.32	2.14
9	44.18	2.11
10	47.37	4.59
11	28.27	2.29
Entire Layer	41.06	10.60

Table 3.4: Comparison of the Actual Number of RUNS (Mean) with the Number of RUNS Required for Stationarity for Different Levels of Significance for Layer D Data in Fig. 3.10.

Layer	RUNS Obtained for Data	m	n	RUNS Required for Stationarity		
				90%	95%	99%
C	2	2	4	5	5	5
D	7	7	4	3 - 8	2 - 9	2 - 9

3.3 Least Squares and Regression in Trend Analysis

The least squares approach takes into account the location of sample points with respect to the estimated point and also the inter-relationships between sample points themselves. The major shortcoming of such an approach is that it neglects the structure of the variable under study, assuming that the variables considered in the analysis are uncorrelated and have a common variance. However, the treatment of the variables as non correlated results in the estimator being non-optimal. If correlation between the residuals exists, the traditional least squares methods do not yield satisfactory results, and hence the generalized least squares techniques will have to be used (Draper and Smith, 1966). The method of generalized least squares incorporates the variance-covariance matrix into the calculation process of the regression coefficients, the details of which will be described in this section.

A statistical model which is used to represent the trend of any soil parameter is termed linear if the dependent variable (soil parameter) can be expressed as a linear combination of the unknown coefficients and the independent variables (data co-ordinates). In some cases of trend analysis, a non-linear trend in the form of a second or third degree polynomial may be more appropriate to model the trend. In classical least squares estimation, the main objective is to minimize the function S where,

$$
\begin{aligned}
S &= E\left[Q_i - \hat{Q}_i\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)^2
\end{aligned}
\tag{3.8}
$$

where n is the total number of data, $\hat{Q}_i$ is the regressed value at location i and Q_i is the actual observed value of the soil parameter. The regressed value $\hat{Q}_i(= f\{X; \beta_j\})$, where, X is the spatial co-ordinate system and β_j's being the regression coefficients.

In classical least squares estimation procedures, the following assumptions are

made on the residuals (Myers, 1986).

(a) Constant Variance

(b) Independence

(c) Zero Mean

(d) Normally Distributed

The regression coefficients β_j are maximum likelihood estimates, only if (a), (b) and (c) above are satisfied. The assumption of normality is required if certain tests like the Student's t test or the F test are to be used to verify the effectiveness of the regression. The condition of zero mean will be satisfied only if there is a constant term (β_o) in the regression equation. For example, if the intercept term in the expression for linear regression is made to be zero, condition (c) above will not be satisfied. When Eq. 3.8 is minimized, subject to,

$$\frac{\partial S}{\partial \beta_j} = 0 \tag{3.9}$$

the values of $\beta_j's$ from above gives the regression coefficients.
In matrix form, it can be expressed as,

$$\{\hat{\beta}\} = \left[X^T R^{-1} X\right]^{-1} \left[X^T\right] [R]^{-1} \{Q\} \tag{3.10}$$

where, $\{X\}$ is the co-ordinate matrix given below and $\{Q\}$ is the soil parameter matrix and $\{R\}$ is the correlation matrix of the residuals. For a typical two dimensional linear problem,

$$[X] = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{bmatrix} \tag{3.11}$$

and,

$$\{Q\} = \begin{Bmatrix} Q_1 \\ Q_2 \\ \vdots \\ \vdots \\ Q_n \end{Bmatrix} \tag{3.12}$$

where, n is the number of data points where soil properties are known. The regressed estimate (Q_o), at any point (x_o, y_o), is given by,

$$Q_o = \begin{bmatrix} 1 & x_o & y_o \end{bmatrix} \begin{Bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \end{Bmatrix} \tag{3.13}$$

As mentioned before, if there is found to be significant correlation among the residuals, $\beta_j's$ will be accurate only if it is considered and Eq. 3.10 expresses the form of generalized least squares. If correlation among residuals is absent or negligible, and the residuals are of constant variance (σ^2), {R} can be expressed as a n × n diagonal matrix in the form,

$$[R] = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ldots & \ldots & 0 \\ 0 & 0 & 1 & \ldots & \ldots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad (3.14)$$

whereby, the expression for $\{\hat{\beta}\}$ in Eq. 3.10 will reduce to,

$$\{\hat{\beta}\} = \left[X^T X\right]^{-1} \left[X^T\right] \{Q\} \qquad (3.15)$$

which gives the classical least squares estimates. The form of Eq. 3.10 allows different functions to be used for the correlation matrix $[R]$. This function can be the autocorrelation function, the covariance function or the variogram function. The definitions of the autocorrelation and covariance functions are given in Chapter 4 and the variogram function is described in detail in Chapter 6.

3.3.1 Verification of Residuals for Non Constant Variance

In order for the traditional least squares method to be valid, two criteria have to be satisfied. That is, the residuals should have a constant variance and possess negligible correlation. The residual e_i, which is the difference between the actual value Q_i and the regressed value $\hat{Q}_i$, can be expressed as,

$$e_i = \hat{Q}_i - Q_i \qquad (3.16)$$

It is generally assumed that e_i is normally distributed and have constant variance, σ^2, given by,

$$\sigma^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - p} \qquad (3.17)$$

where, p is the number of unknown coefficients β_j to be estimated from the regression. For the verification of the constant variance of the residuals, residual plotting is adopted. The ordinate of such a plot will be a function of the residual in the form $e_i / V[e_i]$, where the variance of the residuals, $V[e_i]$ can be obtained from Eq. 3.18 below. It should be noted that $V[e_i]$ will be equal to σ^2 if the variance is constant.

$$V[e_i] = (1 - m_{ii}) \sigma^2 \tag{3.18}$$

where, m_{ii} is the diagonal element of the matrix $[M]$, given by,

$$[M] = [X] \left[X^T X\right]^{-1} \left[X^T\right] \tag{3.19}$$

As can be seen from Eqs. 3.18 and 3.19, $V[e_i]$ is dependent on the form of $[X]$. If large variations of $V[e_i]$ are not expected, e_i / σ^2 may be used instead of $e_i / V[e_i]$, as the ordinate of the residual plot which would have $\hat{Q}_i$ or the independent variables X_i as the abscissa. The reason for plotting the residual function against $\hat{Q}_i$ and not against Q_i, is because the e_i 's are usually correlated to Q_i , while e_i's and $\hat{Q}'_i s$ are independent. The ensuing plots, which are possible outcomes of such an analysis, are illustrated in Figs. 3.15a to 3.15d with the range of the residuals falling in the shaded bands. Figure 3.15a depicts a case where the variance is constant, while Fig. 3.15b illustrates the case where the variance is non constant. Figure 3.15c is a typical illustration when the linear effect of X_j has not been removed and Fig. 3.15d illustrates the need for extra terms in the regression equation or expresses the need for transformation of variables (Myers, 1986). While outcomes in Figs. 3.15c and 3.15d can be remedied by adding extra terms to the regression equation, the situation in Fig. 3.15b can be handled by deviating from the simple methods of least squares and performing weighted least squares. (Myers, 1986).
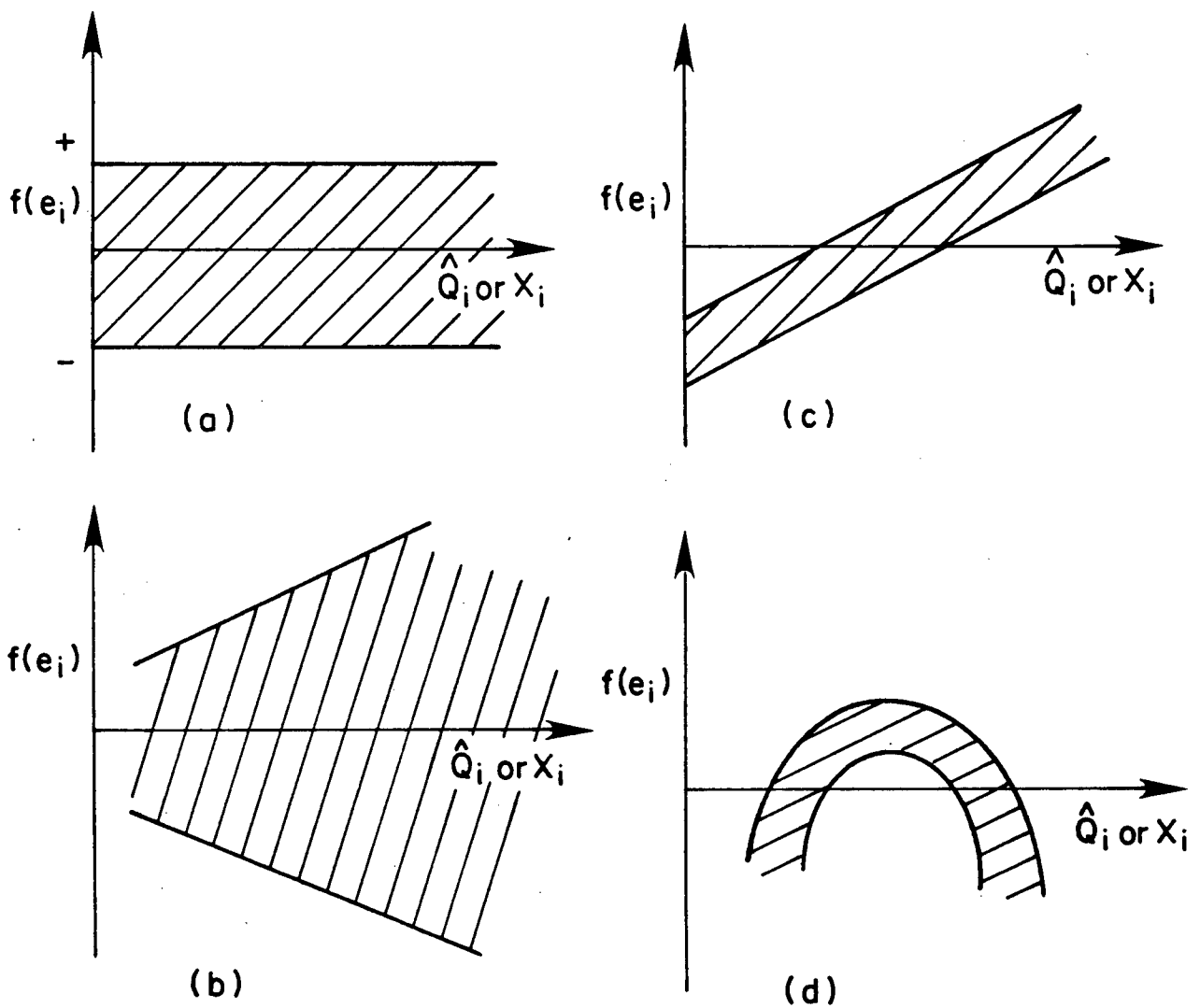
Figure 3.15: Different Relationships of the Residual Function with the Dependent Variable or the Independent Variable.

3.3.2 Verification of Residuals for Correlation

The presence or the absence of correlation (autocorrelation) is verified by using the Durbin - Watson Statistic (Durbin and Watson, 1951), d_o, which is defined as,

$$d_o = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (3.20)$$

This value of d_o should be compared with the values in the tables (" Testing for Serial Correlation in Least Squares Regression II ", - Durbin and Watson). The tables provide upper limits d_u and lower limits d_l for different number of predictor variables used in the model, for three different significant levels, 5%, 2.5% and 1%. If $d_o > d_u$, the autocorrelation of residuals is negligible and if $d_o < d_l$, the correlation is significant. If $d_u > d_o > d_l$, the test is said to be inconclusive and it is conservative to consider correlation (autocorrelation).

In the event significant correlation of the residuals exists, the most convenient method of interpolation is for it to be done in two parts. The non-stationary component, which is the trend, can be determined by the classical method of least squares and the stationary component can be subjected to methods where correlation is considered. Procedures of correlation analysis is described in Chapter 6 of the thesis.

3.3.3 Statistical Tests to Measure Efficiency of Fit

The quality of a fit of a regression analysis is measured by the multiple correlation coefficient, R_θ^2 (Graybill, 1961 and Brooke and Arnold, 1985), which is given by,

$$R_\theta^2 = \frac{\sum_{i=1}^{n} \left(\hat{Q}_i - \bar{Q}\right)^2}{\sum_{i=1}^{n} \left(Q_i - \bar{Q}\right)^2} \qquad (3.21)$$

where, $\bar{Q}$, is the mean of the data given by,

$$\bar{Q} = \frac{1}{n} \sum_{i=1}^{n} Q_i \qquad (3.22)$$

and $\hat{Q}$ is the regressed estimate. Brooke and Arnold (1985) suggest that $R_\theta{}^2$ should be at least 0.5 for any confidence to be placed on the regression. $R_\theta{}^2$ satisfies a Beta distribution with parameters $0.5\,\nu_1$ and $0.5\nu_2$ where ν_1 is equal to (k-1) and ν_2 is equal to (n-k), k being the number of regression coefficients in the equation.

The F test (Brooke and Arnold, 1985) can be used to verify whether the model is parsimonious (optimum number of variables in the regression equation). The most efficient model is not necessarily the model that gives the highest $R_\theta{}^2$ value and the minimum variance of the residuals. It is also important that the optimum number of variables is used. The above test evaluates in a quantitative way the improvement gained by the addition of extra variables and its significance on the estimation. A different type of F test can also be used to verify the efficiency of the coefficient, β_1 in simple regression.

The 't' test (Brooke and Arnold, 1985) is performed on the coefficients of the regression to evaluate its efficiency. If the fitted regression is acceptable at a particular significance level, the null hypothesis $H_1 : \beta_j = 0$ for all $j = 1, ...k$ should be rejected. In the case of simple regression, the F test on coefficients is equivalent to the 't' test ($F = t^2$). More details of both the F and 't' test are available in texts by Myers(1986), Brooke and Arnold (1985) and Draper and Smith (1981).

3.3.4 Application of Trend Analysis to a CPT profile

Methods of trend analysis described in the preceding sections have been applied to the CPT profile in Fig. 3.9. Using methods of layer identification already described in Chapter 2, two prominent layers have been identified between 4.50 m and 10.0 m (Layer A) and 10.0 m and 13.5 m (Layer B) considering the cone bearing profile.

The statistics obtained for the Layer A data for a simple linear trend (model 1) and for a curvilinear trend (model 2) are given below.

Layer A - Linear Trend (Model 1)

Model 1 is given by,

$$Q_i = \beta_o + \beta_1 y_i \tag{3.23}$$

$n = 221$ $\qquad\qquad R_\theta{}^2 = 59.14$ $\qquad\qquad d_o = 0.03$

Sum of squares of residuals, $\sigma_{r1}{}^2 = 139776$

Variance of Data, $\sigma^2 = 34410$

The statistics of the regression coefficients are given in Table 3.5.

Table 3.5: Statistics for Layer A with Linear Trend.

Coefficient	Mean	Standard Deviation	t Ratio
β_o	-51.46	7.91	- 6.51
β_1	19.06	1.07	17.89

Layer A - Curvilinear Trend (Model 2)

Model 2 is given by,

$$Q_i = \beta_o + \beta_1 y_i + \beta_2 y_1{}^2 \tag{3.24}$$

$n = 221$ $\qquad\qquad R_\theta{}^2 = 66.70$ $\qquad\qquad d_o = 0.04$

Sum of squares of residuals, $\sigma_{r2}{}^2 = 114587$

Variance of Data, $\sigma^2 = 34410$

The statistics of the regression coefficients are given in Table 3.6.

Higher order polynomials resulted in higher R_θ values but were not selected due to the significant correlation between the independent variables, giving rise to multi-collinearity (Myers, 1986). The improvement in the model caused by a change from a linear (model 1) to a simple curvilinear was verified using the F test given by,

Table 3.6: Statistics for Layer A for Curvilinear Trend.

Coefficient	Mean	Standard Deviation	t Ratio
β_o	-286.08	34.64	- 8.26
β_1	87.09	9.87	8.82
β_2	-4.69	0.68	-6.92

$$F_{\nu_1,\nu_2} = \frac{(\sigma_{r1}^2 - \sigma_{r2}^2)/(k_2 - k_1)}{\sigma_{r1}^2/(n - k_1)} \tag{3.25}$$

where, for the case considered, k_1 and k_2 (the number of coefficients in models 1 and 2 respectively) take the values two and three, respectively; ν_1 is equal to $(k_2 - k_1)$ and ν_2 takes the value $(n - k_1)$.

The value obtained from Eq. 3.25 (F = 39.46) is greater than the value of F_{crit} obtained (3.82) at 95% significance from the tables, suggesting that model 2 is more appropriate. The 't' test on the coefficients also indicates the adequacy of the model. The t Ratios for the coefficients are all greater than $t_{crit}(= 1.645$, from t tables), at a significant level of 95% for a degree of freedom $(n - k)$, k being the number of coefficients; two for model 1 and three for model 2. However, the Durbin - Watson statistic, d_o, is low $(d_o < d_l)$, indicating the presence of autocorrelation among the residuals. Due to the presence of correlation among the residuals, the exact value of the coefficients will be given by Eq. 3.10 and not Eq. 3.15. In geotechnical applications, however, there will be only one realization at a given point, rendering it impossible to evaluate the covariance matrix {R}. Therefore, the only option will be to perform interpolation on the residuals, separately. In the present example, only trend analysis will be performed and applications of residual correlation analysis will be explained in Chapter 6. With increasing degree of the polynomial used for the fit, the correlation among the residuals decreases. This behavior is reflected in the

value of d_o which increases from 0.03 for model 1 to 0.04 in model 2. However, as explained before, the degree of the polynomial used for a trend is not only limited by considerations of multi - collinearity in higher order models, but is also restricted from a practical point of view. Therefore, an intelligent compromise would be to tolerate correlations among the residuals instead of deciding on a model which exhibits multi-collinearity.

The analysis was also carried out for Layer B (10.0 m - 13.5 m) in a similar manner. As for the Layer A data, the only drawback was the presence of correlation among the residuals. Once again higher order polynomials were not considered due to the high correlation among the independent variables. All other statistical tests were satisfied with the curvilinear model being more superior. As for the Layer A analysis, the value of d_o increases from 0.07 for model 1 to 0.08 in model 2. The F value obtained by using Eq. 3.25 was 9.65, reflecting the superiority of model 2 over model 1.

The statistical details pertaining to Layer B are listed below.

Layer B - Linear Trend (Model 1)

As before, model 1 is given by,

$$Q_i = \beta_o + \beta_1 y_i \tag{3.26}$$

$n = 141$ $\qquad\qquad R_\theta{}^2 = 63.30$ $\qquad\qquad d_o = 0.07$

Sum of squares of residuals, $\sigma_{r2}{}^2 = 158868$

Variance of Data, $\sigma^2 = 432761$

The statistics of the coefficients are given in Table 3.7.

Layer B - Curvilinear Trend (Model 2)

Model 2 is given by,

$$Q_i = \beta_o + \beta_1 y_i + \beta_2 y_1{}^2 \tag{3.27}$$

Table 3.7: Statistics for Layer B with Linear Trend.

Coefficient	Mean	Standard Deviation	t Ratio
β_o	-430.76	33.50	- 12.86
β_1	43.78	2.84	15.42

$n = 141$ $\qquad R_\theta{}^2 = 65.80$ $\qquad d_o = 0.08$

Sum of squares of residuals, $\sigma_{r2}{}^2 = 147992$

Variance of Data, $\sigma^2 = 432761$

The statistics of the coefficients are given in Table 3.8.

Table 3.8: Statistics for Layer B for Curvilinear Trend.

Coefficient	Mean	Standard Deviation	t Ratio
β_o	894.70	419.00	2.14
β_1	-183.31	71.62	-2.56
β_2	9.66	3.04	3.257

The linear trends of Layers A and B are illustrated in Fig. 3.16 and the curvilinear trends in Fig. 3.17. The distributions of the residuals for the linear and curvilinear models considered fall into the category given in Fig. 3.15a. Therefore, there was no need for weighted regression to be performed and the simple regression procedure was adequate to model the profiles in Layer A and Layer B.

3.3.4.1 Lower Confidence Limit of Cone Bearing

The foregoing applications illustrated how trend lines in profiles can be obtained. In geotechnical engineering, the matter of greater concern is generally the establishment of lower bounds at a particular significance level. For example, a 90% lower confidence limit for a particular layer of soil will indicate the boundary above which the soil parameter under consideration will lie ninety times out of hundred.
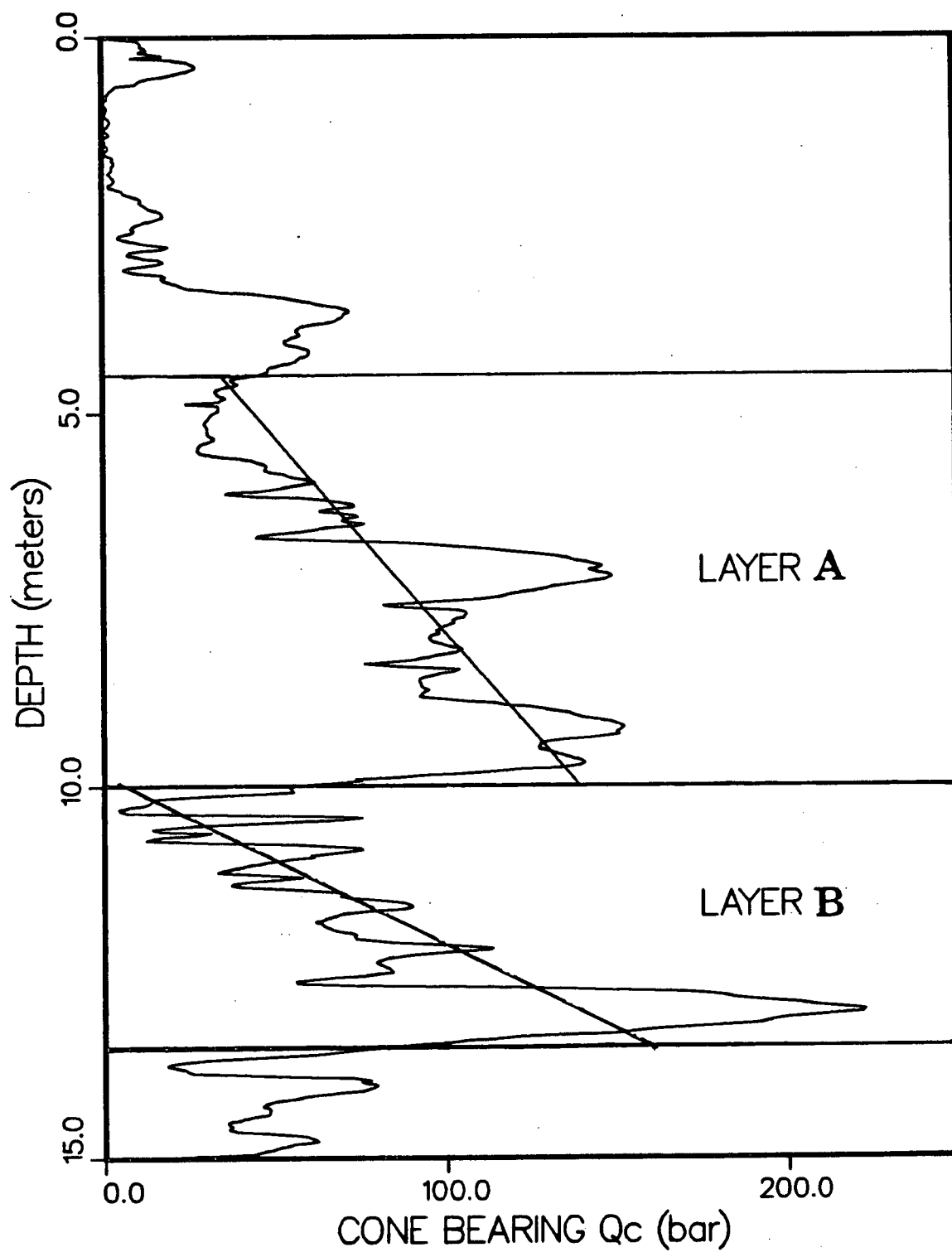
Figure 3.16: Linear Trends of Layer A and Layer B.
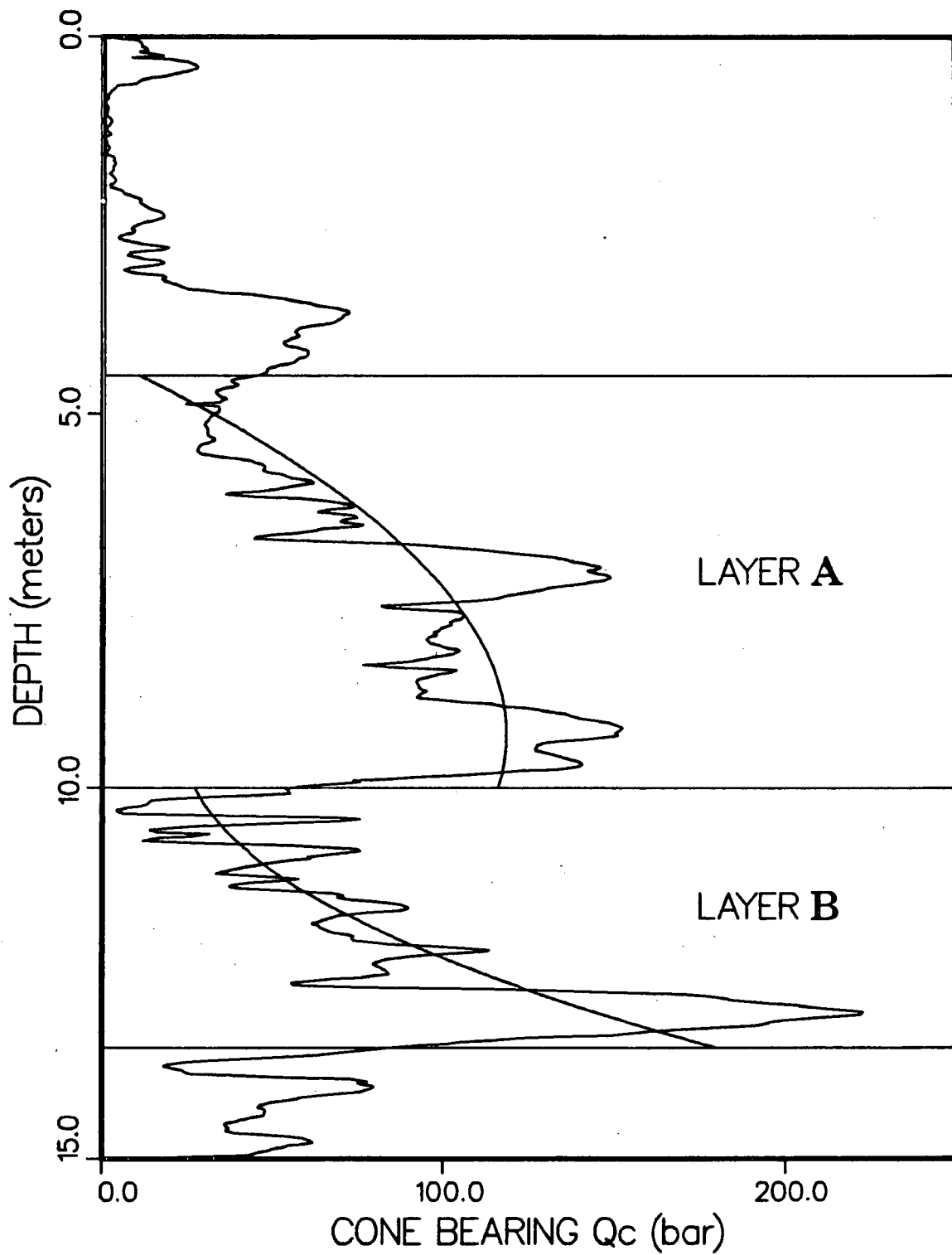
Figure 3.17: Curvilinear Trends of Layer A and Layer B.

Figures 3.18 and 3.19 illustrate the lower 95% confidence limits of Layer A and Layer B, for data assuming linear and curvilinear trends, respectively. The lower 95% is a highly conservative limit and if the engineer is willing to increase the element of risk in the design, it can be lowered to 90% or even 80%, shifting the lower limit desired towards the trend line. Once the lower confidence limit is decided upon, the graphs which correlate relative density and friction angle with cone bearing can be overlaid on the lower confidence bands, enabling the engineer to obtain reliability estimates on these soil parameters which will be used in design.

For data assuming a linear trend, the lower confidence estimate of the trend $(\tilde{Q}_i)$ at a significant level of $(1 - \alpha)$ is given by (Brooke and Arnold, 1985),

$$\tilde{Q}_i = \hat{Q}_i - t_{\alpha/2.(n-2)}\sqrt{\left(1 + \frac{1}{n} + \frac{(Q_i - \bar{Q})^2}{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}\right)\sigma^2} \qquad (3.28)$$

where σ^2 is the variance of the residuals given by Eq. 3.17 and $t_{\alpha/2(n-2)}$ is obtained from Student's 't' tables.

Similarly, the lower confidence estimate for data assuming a curvilinear trend $(\tilde{Q}_i')$ represented by a second degree polynomial is similarly given by,

$$\tilde{Q}_i' = \hat{Q}_i - t_{\alpha/2.(n-3)}\left(1 + \frac{1}{n} + \left([1 \quad z_i \quad z_i^2]\begin{bmatrix} n & \sum z_i & \sum z_i^2 \\ \sum z_i & \sum z_i^2 & \sum z_i^3 \\ \sum z_i^2 & \sum z_i^3 & \sum z_i^4 \end{bmatrix}^{-1}\begin{Bmatrix} 1 \\ z_i \\ z_i^2 \end{Bmatrix}\right)\right)\sigma$$
$$(3.29)$$

where, $z_i's$ are the depth co-ordinates.
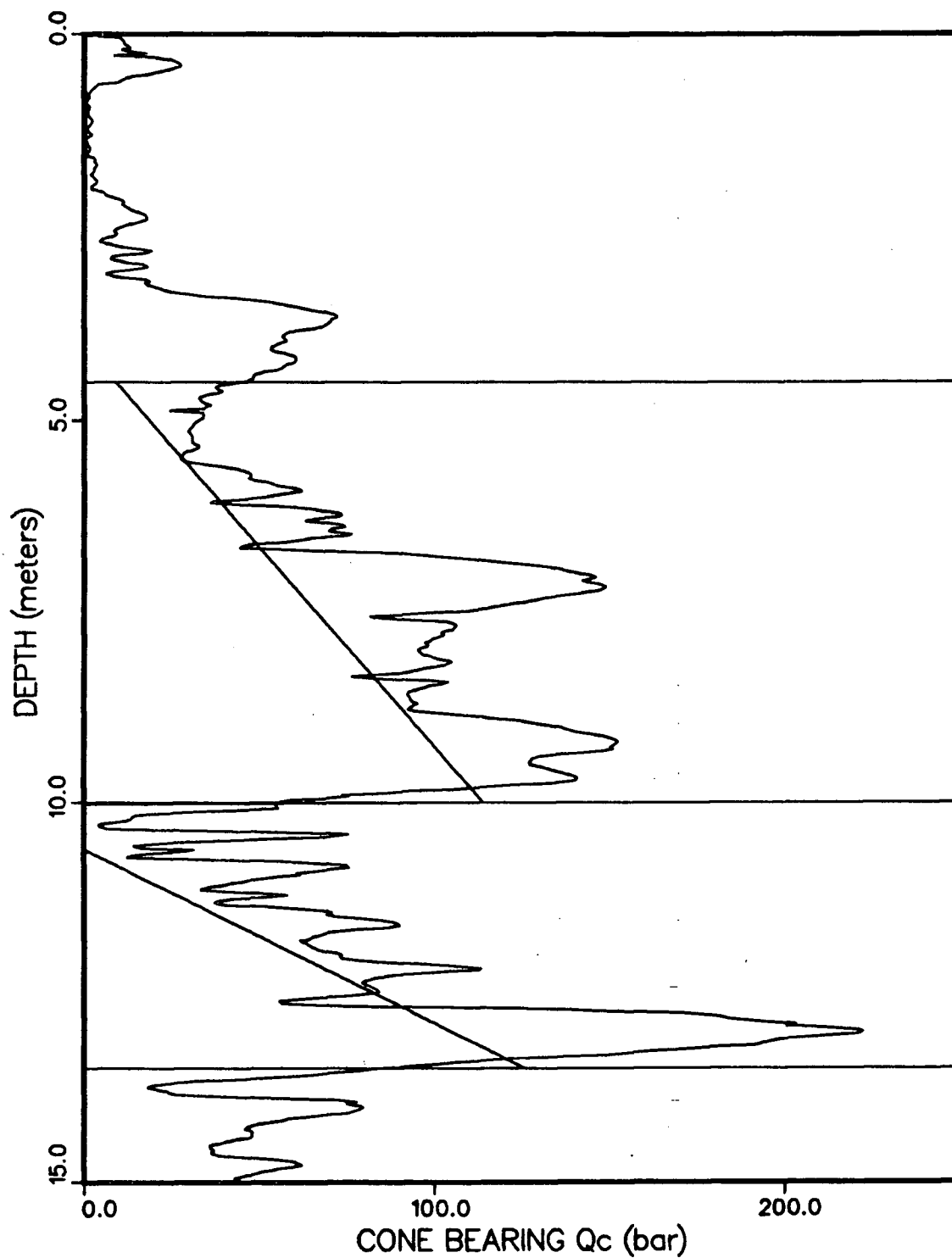
Figure 3.18: Lower 95% Confidence Estimate of Bearing for Linear Trend.
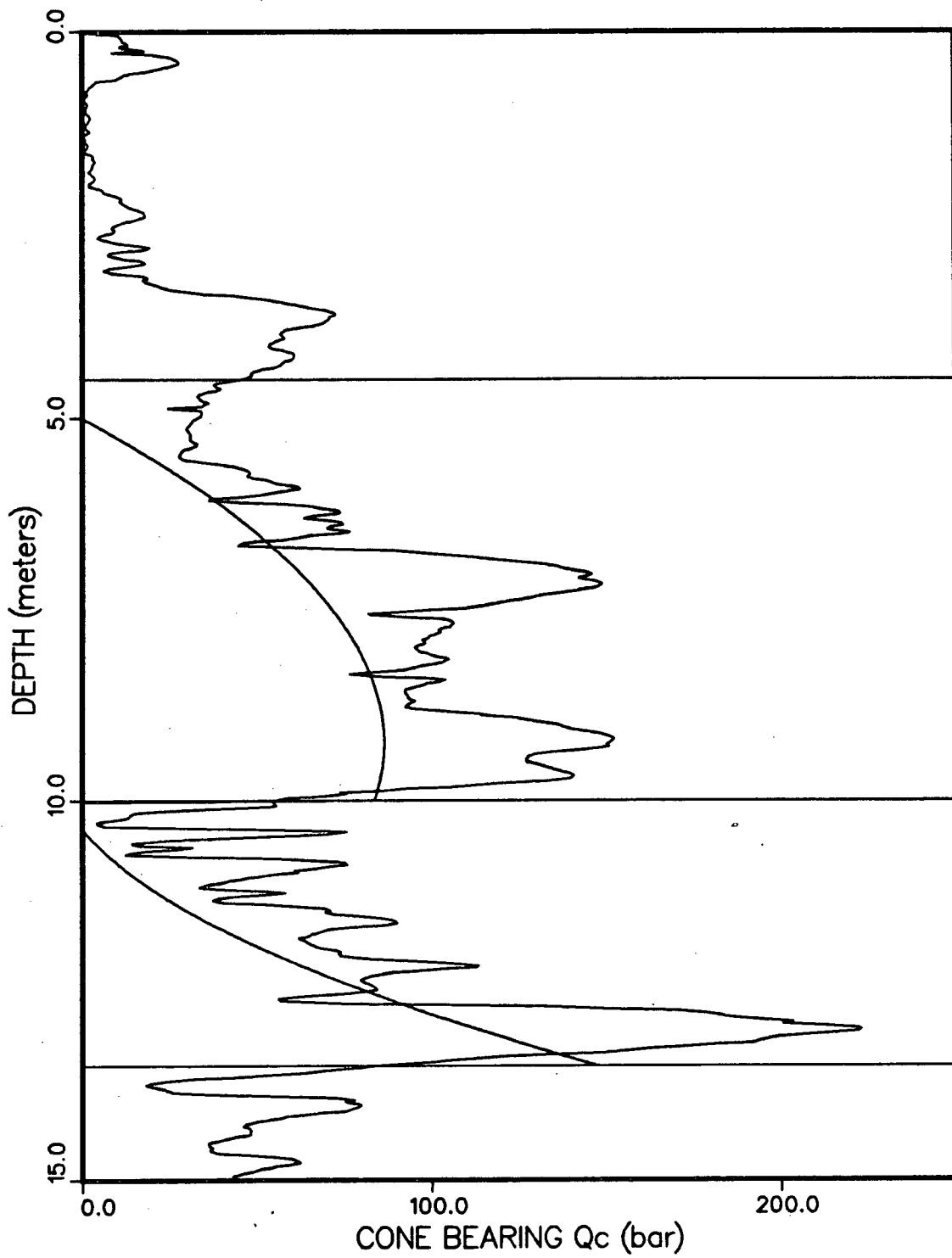
Figure 3.19: Lower 95% Confidence Estimate of Bearing for Curvilinear Trend.

3.4 Conclusions

The main conclusions drawn from this chapter are;

(i) Methods of moving average smoothing and Fourier smoothing are only suitable for the evaluation of trends in a qualitative manner. These methods are not recommended for other areas of applications where analysis of data is performed, because they modify even the acceptable ('good') data points.

(ii) Statistical methods of filtering are recommended over smoothing methods because the removal or substitution of extreme data points is done with some statistical basis. The median method of filtering is preferable to the mean method, since the mean is dependent on extreme data points while the median is not. Substitution of filtered data points is best performed by the mean of the adjacent two unfiltered data points.

(iii) The RUN test has proven to provide an efficient way of determining the stationarity of soil profiles.

(iv) From a statistical standpoint the method of trend analysis using regression techniques is a convenient method of expressing soil property dependence with depth. A linear trend or a polynomial of degree two will be sufficient to model the variation of soil properties with depth.

(v) Geotechnical engineers can use the lower confidence limits of their preference for design purposes, such that a high percentage of the soil property value of concern will be above an acceptable lower limit.

Chapter 4

Random Field Theory in Geotechnical Data Analysis

4.1 Introduction

Modeling the stochastic character of soil properties is very important in geotechnical engineering. The natural heterogeneity of the soil, soil disturbance during testing or extraction of samples, measurement errors caused both by man and machine and most importantly the limitation of data availability, all give rise to uncertainties in soil parameter estimation. There would be nothing random in the distribution of soil properties if all the points in the ground could be tested accurately. However, this is not feasible both practically and economically and hence the need for the treatment of in situ soil data considering as if it were random has arisen.

This chapter will investigate different types of applications from the point of random fields. Parameters such as the variance function and the scale of fluctuation will be investigated from a geotechnical engineering point of view. A different method of obtaining the scale of fluctuation will be proposed and application areas of this parameter such as averaging effects of the cone bearing, sleeve friction and pore pressure from the CPT will be explored. The influence of trend on the scale of fluctuation will also be discussed with specific examples to illustrate its significance. This chapter will then look at the application areas of correlations between spatial averages and exceedance probabilities of CPT profiles. Most of the theories used are extensions of those derived by Vanmarke (1983). The effects of soil variability, accuracy and

confidence levels of estimates on the optimum sample spacing for a given site will be described to show how the collection of unnecessary data can be avoided.

4.2 Parameters Required to Fully Identify a Soil Stratum

In geotechnical engineering, it is common to divide the heterogeneous soil stratum into statistically homogeneous layers. The means or some lower bound values of these statistically homogeneous layers are then considered for design and analysis, neglecting the effect of variation or fluctuation about these values. A constant mean, constant standard deviation and an autocorrelation function which is independent of the location and is a function only of the separation distance (lag distance) in the depth dimension, are necessary requirements for homogeneity or stationarity. If a soil stratum exhibits varying types of trends at different depths, it can be divided into distinct layers, each identified by a particular trend: linear, curvilinear, etc.. A trend in actual effect is a non constant mean and, therefore, in keeping with the above definition of stationarity, it will have to be removed for the soil layer to be classified as stationary. These layers are then treated individually, in order to derive their respective statistics.

In addition to the mean ($\bar{Q}$), two other parameters are required to describe the spatial variability of a soil property characteristic which is to be treated as random (Vanmarke, 1977). One of these parameters is the standard deviation (σ_Q), which measures the degree to which the actual values differ from the mean. The coefficient of variation (η) is a standardized form of a variability factor, which gives the relationship between the mean and the standard deviation.

The coefficient of variation, η, is defined as,

$$\eta = \frac{\sigma_Q}{\bar{Q}}$$

(4.1)

The scale of fluctuation (δ) is the other important parameter. This measures the distance within which the soil properties show strong correlation in the vertical or horizontal direction. The emphasis in this chapter will be on the scale of fluctuation in the vertical dimension. If two points in a soil layer lie closer than its scale of fluctuation, the soil property values at both these points will be on the same side of the mean (either both above or both below). It is in this sense that δ is also known as the distance of perfect correlation. A low value of the scale of fluctuation means rapid fluctuations of the property value about the mean (high variability) and a high value of δ reflects the slowly varying nature of the property value about the mean (low variability). The above explains why it is important to consider the scale of fluctuation in addition to the mean and the standard deviation, when a soil profile needs to be fully characterized. The name of this important parameter (δ) is somewhat misleading since a higher value of the scale of fluctuation reflects a lower variability, and vice versa. In this regard, although it seems more appropriate for δ to be referred to as the ' scale of uniformity ', for the sake of consistency with the present literature, this thesis will continue to refer to it as the scale of fluctuation.

4.3 Scale of Fluctuation

4.3.1 Spatial Averaging

The scale of fluctuation is a parameter which describes spatial variability. Therefore, it is important to acquire a complete understanding of the effect of spatial averaging prior to discussing the derivation and merits of the scale of fluctuation.

Within a small volume or element , soil property values are approximately uniform and less variable. However, among the group of these small elements, some may have lower average values while some may possess higher average values. As a result, the within element variability will definitely be lower than the between element variability.

This phenomenon has been effectively used to identify different types of layering in Chapter 2. If the elements considered are large, the above concept will not be true since in the larger elements, the internal variations will balance out such that the average property values from one large element to another will not differ too much (Baecher, 1985). The averages of large volumes will be approximately equal to those of the smaller volumes but the standard deviation which reflects the variability will be significantly different. The variation of the standard deviation from one small element to the next will be greater than if the elements were larger.

The extent of the variability of the standard deviation is dependent on the structure of the spatial variability of the soil property value under consideration and is expressed by the variance function, $\Gamma^2(.)$, which will be defined and explained in section 4.3.2.

4.3.2 Variance Function (Γ^2)

The recommended procedure (Vanmarke, 1977) of determining the scale of fluctuation is in terms of the variance function which adequately explains the effects of spatial averaging. The less frequently used method uses the autocorrelation function to derive the scale of fluctuation. The details of the latter method and its drawbacks will be described in section 4.3.4.

The procedure of obtaining the scale of fluctuation (δ) in terms of the variance function is as follows;

The data are first considered in pairs ($n = 2$) and a new series of data comprising of the respective averages of the adjacent data points are derived. The length of averaging will be equal to the spacing of data points, (Z_2). The standard deviation (σ_2) of this derived series is then calculated. The standard deviation of this series

(n = 2) will be less than the standard deviation of the original data set, σ_1, because of the cancelling out of fluctuations due to spatial averaging. The above procedure is extended to the case n = 3, where three adjacent data points will be averaged to obtain the derived series for the case n = 3. The corresponding standard deviation of this series, σ_3, is calculated with the spacing, Z_3, being equal to twice the spacing between data points. For a typical CPT sounding which samples at 2.5 cm, Z_2 will be equal to 2.5 cm and Z_3 will be 5.0 cm. This procedure is repeated for $n = 4, 5, 6, \ldots\ldots$ until n approaches the total number of data points, N. The effect of spatial averaging will be more significant with increasing n with,

$$\sigma_1 > \sigma_2 > \sigma_3 > \ldots\ldots\ldots > \sigma_N$$

For each n, the variance function, $\Gamma^2(Z_n)$, can be calculated as,

$$\Gamma^2(Z_n) = \frac{\sigma_n{}^2}{\sigma_1{}^2} \tag{4.2}$$

where, $\sigma_n{}^2$ is the variance (squared of the standard deviation) of the derived moving average series of degree n and $\sigma_1{}^2$ is the variance of the original data. If the spacing of the data is d, Z_n in Eq. 4.2 will be equal to (n - 1)d. The variance function given above can be determined for different lag distances (separation distances), Z. Figure 4.1 illustrates a typical variation of $\Gamma^2(.)$ which has a maximum value of unity, decaying towards zero for increasing lag distance values Z.

From Vanmarke (1977), for large values of Z (very large n) the variance function will become inversely proportional to Z and can be expressed as,

$$\Gamma^2(Z) = \frac{\delta}{Z} \tag{4.3}$$

The the relationship in Eq. 4.3 can also be expressed as (Vanmarke, 1977),

$$\Gamma(n) = \left(\frac{\delta}{nd}\right)^{\frac{1}{2}} \qquad (4.4)$$

where, n is the number of data points which are averaged and d is the sampling interval. The next step is to fix n (n^*) and observe $\Gamma(n^*)$ and the scale of fluctuation, δ, will be given by,

$$\delta = \Gamma^2(n^*)n^*d \qquad (4.5)$$

The point of maximum curvature is a suitable point to obtain n^* (Vanmarke, 1988). For the variance function profile given in Fig. 4.1, $n^* = 48$ (n^*d $= 0.025 \times 48 = 1.2$ m) and $\Gamma^2(n^*) = 0.27$. Therefore, the scale of fluctuation (δ) is given by (Eq. 4.5),

$$\delta = 1.2 \times 0.27 = 0.324m = 32.4cm \qquad (4.6)$$

In the original method described above, the value of Γ^2 is selected from the curve at a reasonably high value of Z, where there is a distinct change in the curve (point of maximum curvature). This has been verified by Vanmarke (1988).

A practical variant of the above method of determining δ is proposed and used in this thesis. It makes use of Eq. 4.3 directly and is very convenient for computer applications. At large values of Z, the function $\Gamma^2(Z).Z$ reaches a peak and this maximum value gives a good approximation for δ. Figure 4.1 gives a typical variation of a variance function, $\Gamma^2(Z)$, with a maximum value of unity and gradually decreasing with increasing distance, Z. Figure 4.2 shows the variation of $\Gamma^2(Z).Z$ with Z and as can be observed, $\Gamma^2(Z).Z$ reaches a maximum value of 31.25 cm, which compares well with the value (32.4 cm) obtained from the method recommended by Vanmarke (1977). The proposed method allows a consistent and objective determination of δ.
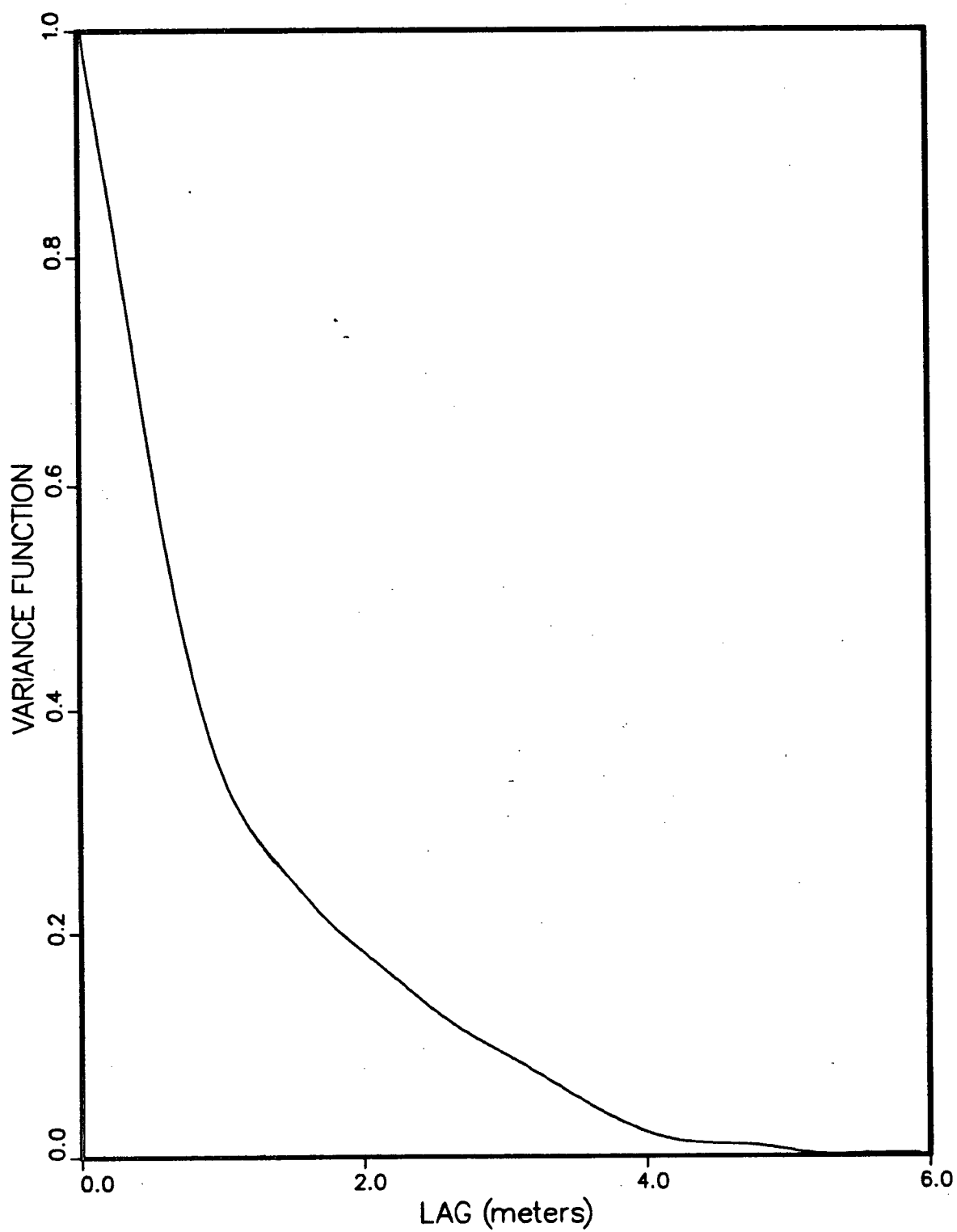
Figure 4.1: Variance Function of Haney Data for Layer Between 9.3 and 15.51 meters.
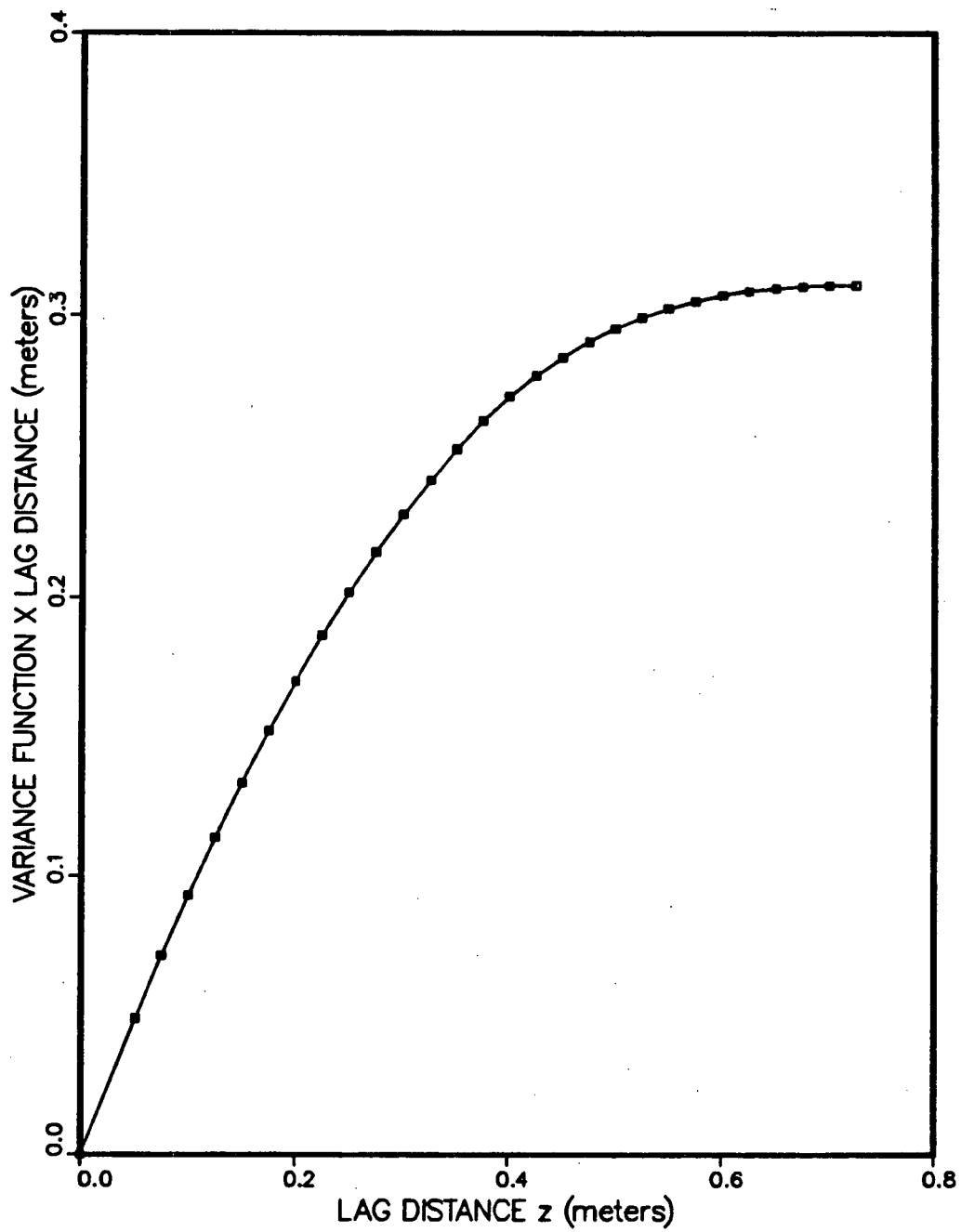
Figure 4.2: Variation of the Variance Function × Lag Distance ($\Gamma^2.Z$) for Haney Data Between 9.3 and 15.51 meters.

Table 4.1: Comparison of the Two Methods for Obtaining the Scale of Fluctuation.

Data	Soil Property	Scale of Fluctuation (cm)	
		Vanmarke's Method	Proposed
Haney 2 (9.3 - 15.51 meters)	Cone Bearing	32.40	31.25
	Sleeve Friction	33.60	35.99
	Pore Pressure	26.68	30.45
Langley 3 (2.60 - 10.80 meters)	Cone Bearing	23.04	24.08
	Sleeve Friction	41.79	39.98
	Pore Pressure	21.98	17.91
Strong Pit 1 (5.25 - 10.12 meters)	Cone Bearing	27.99	26.22
	Sleeve Friction	37.59	36.88
	Pore Pressure	17.75	13.76
Vanmarke (1977)	Cone Bearing	120.0	98.2

The effectiveness and the advantage of of the proposed method of obtaining δ has been acknowledged by Vanmarke (1988). The above value of the scale of fluctuation (31.25 cm) relates to the cone bearing at Haney 2. Figures 4.3, 4.4 and 4.5 illustrate the CPT profiles for the Haney, Langley and Strong Pit sites for which the δ values have been calculated. The values of δ which have been obtained for the different sites listed above are tabulated in Table 4.1 for cone bearing, sleeve friction and pore pressure. It also gives a comparison of the proposed method with Vanmarke's (1977) method. All the results in Table 4.1 indicate the adequacy of the method suggested.

4.3.3 Removal of Trend

Soil properties are highly depth dependent and hence CPT parameters such as cone bearing, sleeve friction and pore pressure exhibit significant trends with depth. If the trends are not significant, data can be considered as stationary. In the presence of significant trends, these will have to be removed prior to determining the scale
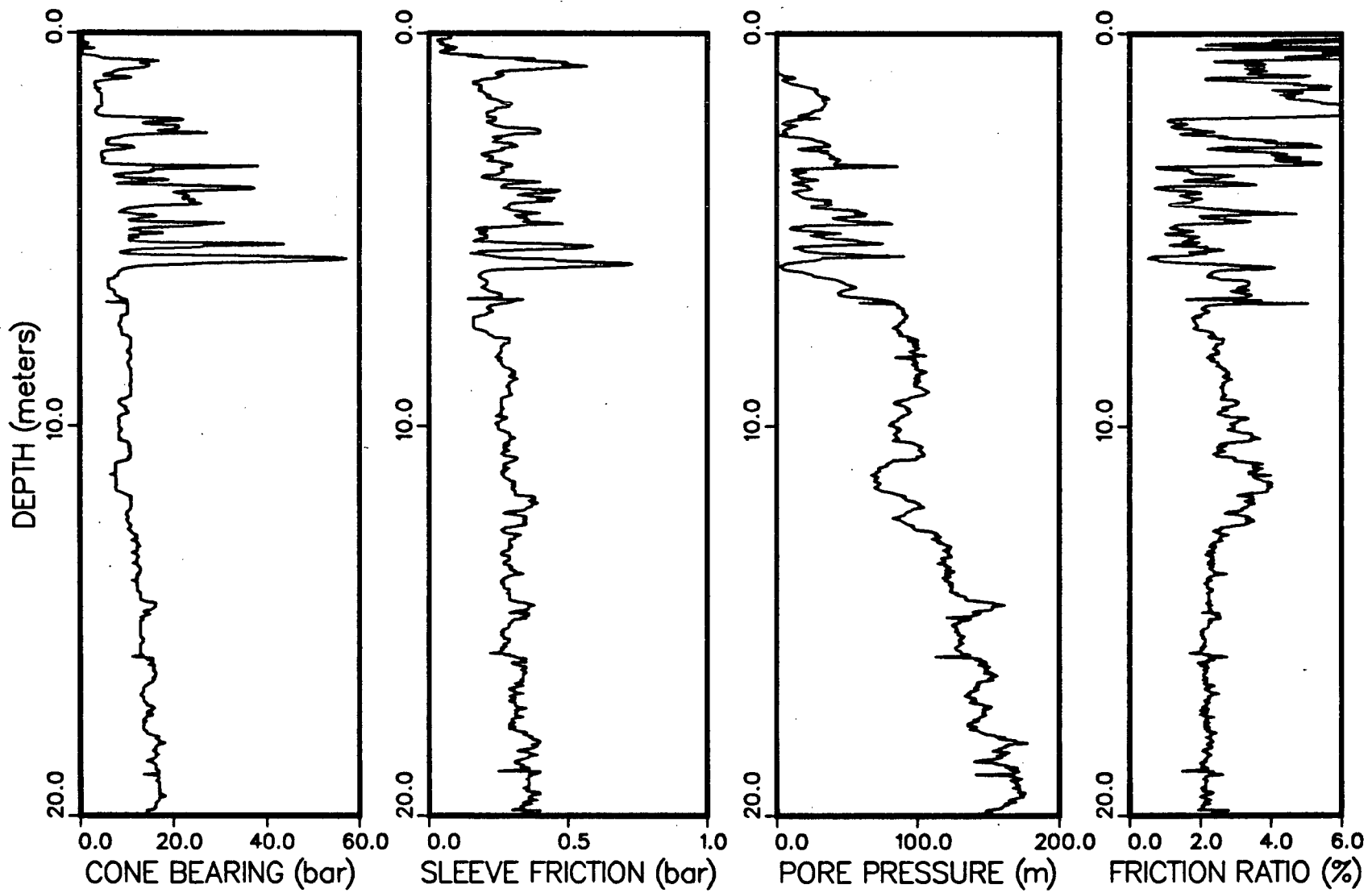
Figure 4.3: Cone Bearing, Sleeve Friction, Pore Pressure and Friction Ratio Profiles at Haney Site.
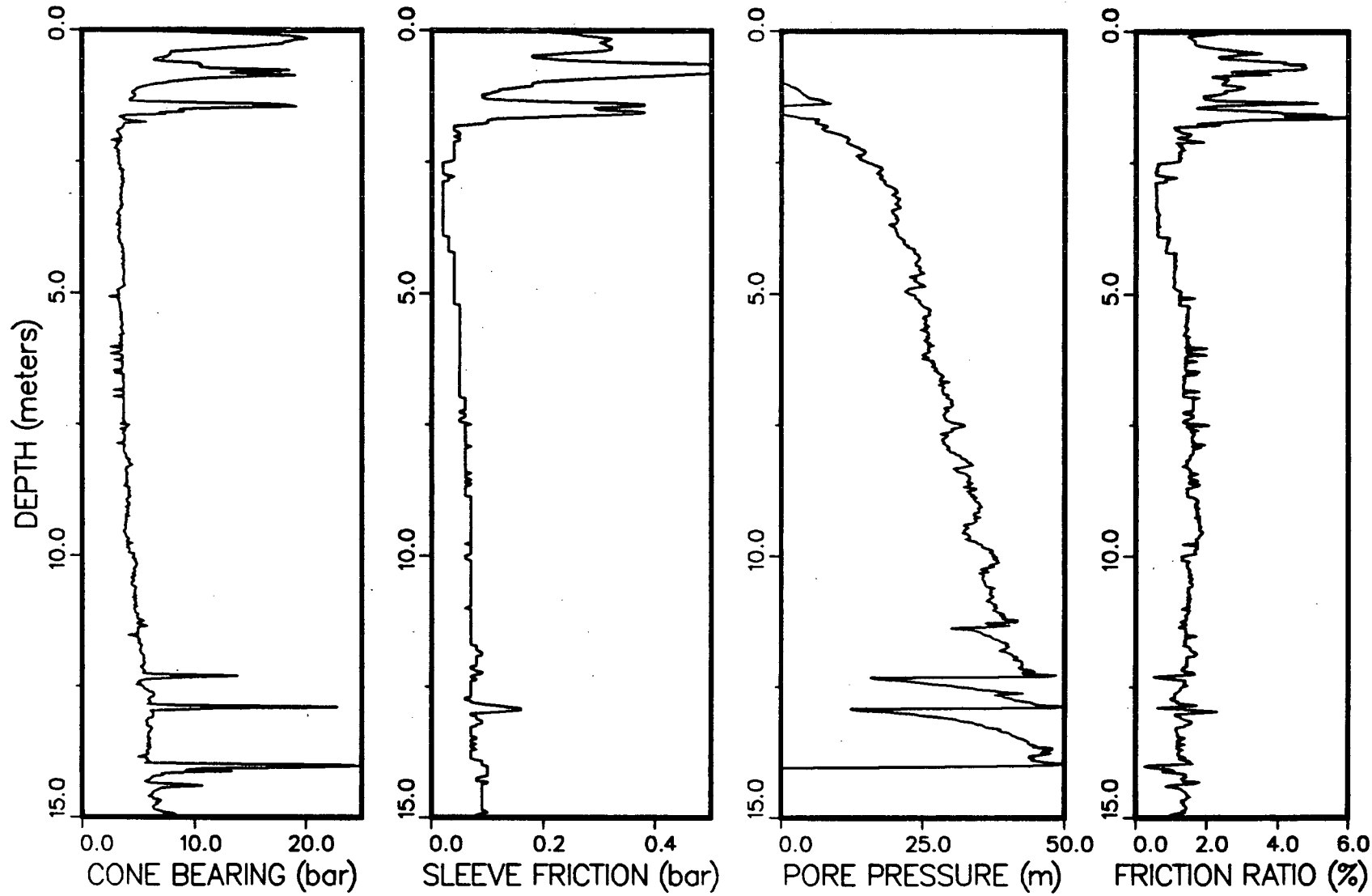
Figure 4.4: Cone Bearing, Sleeve Friction, Pore Pressure and Friction Ratio Profiles at Langley Site.
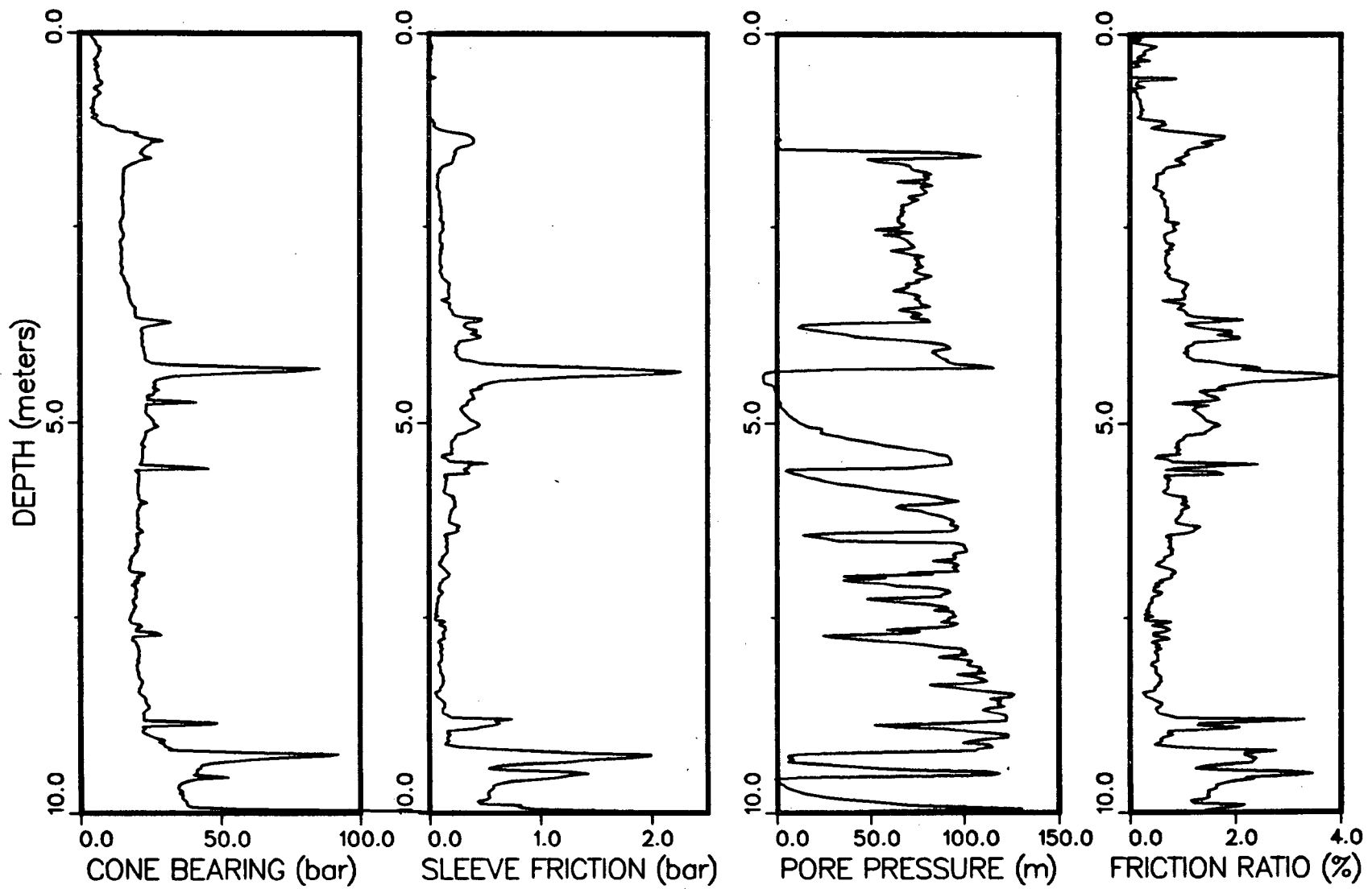
Figure 4.5: Cone Bearing, Sleeve Friction, Pore Pressure and Friction Ratio Profiles at Strong Pit.

of fluctuation. Different methods of trend removal have already been described in Chapter 3. A pre-requisite to this is the identification of statistically distinct layers, using the techniques described in Chapter 2.

Cone bearing and sleeve friction often exhibit linear trends while it is usual for the pore pressure profile to possess a curvilinear trend as illustrated in Figs. 4.3 to 4.5. If a linear trend is used on a pore pressure profile, the resulting residuals will not be stationary, due to the curvilinear effects of the trend not being removed. This will give rise to an incorrect value for δ. The decision as to what type of trend removal is necessary could be taken by inspecting the residuals as explained in Chapter 3. If a linear trend removal results in a stationary set of residuals, curvilinear trend analysis is unnecessary.

In certain profiles where the curvature of the pore pressure profile is prominent, the scales of fluctuation obtained for the linear trend removed data and the curvilinear trend removed data will show an appreciable difference, with the latter method giving lower values. As explained earlier, the curvilinear trend would always be more suitable, since it also includes the linear case as a subset. Table 4.2 provides sufficient evidence as to why a curvilinear trend removal method has to be adopted, if one exists. The δ for the curvilinear trend is less than that for the linear trend, except for the result of Langley 3 data. The reason for the above is that the pore pressure trend is generally much better represented by a curvilinear trend, than by a linear trend. The Langley 2 data seem to have approximately equal δ values for the two methods of trend removal, suggesting that the pore pressure profile in this case may be adequately represented by either a linear or a curvilinear trend. In Tables 4.1 and 4.2, Haney 1, Haney 2 and Haney 3 refer to different profiles obtained at the Haney site while Langley 1, Langley 2 and Langley 3 represent different cone profiles obtained from the Langley site. The cone bearing and sleeve friction results indicated in Table

Table 4.2: Comparison of the Scale of Fluctuation for Pore Pressure Obtained by Linear Trend Removal and Curvilinear Trend Removal.

Data	Layer Depths (m)	Scale of Fluctuation for Pore Pressure (cm)	
		Linear Trend	Curvilinear Trend
Haney 1	7.15 - 13.12	34.85	27.21
Haney 2	9.30 - 15.50	34.21	30.45
Haney 3	13.52 - 22.50	35.34	28.45
Langley 1	2.60 - 10.75	46.08	26.15
Langley 2	2.60 - 10.60	37.37	37.33
Langley 3	2.60 - 10.60	17.91	18.41
Strong Pit	5.25 - 10.12	20.18	13.76

4.1 are for a linear trend while the scale of fluctuation results for pore pressure are for a curvilinear trend. A description of the above sites are given in section 1.7.

4.3.4 Relationship with the Autocorrelation Function

The scale of fluctuation can also be expressed in terms of the autocorrelation function (ρ). The autocorrelation function, as the name suggests, is the function which gives the correlation of a variable with the corresponding variable at different locations. For example, the autocorrelation function at a particular separation distance (lag distance) is the correlation of all data points separated by that distance. Equation 4.7 explains this relationship more clearly. Soil properties generally show stronger correlation for closely located points, with the correlation decaying for increased lag distance. The autocorrelation function, which is the standardized form of the auto-covariance function, has a maximum value of unity with the possibility of minimum values even in the negative region for large lag distances.

The autocorrelation function at a lag distance $l(\rho(l))$ is defined as,

$$\rho(l) = \frac{N}{N - h} \frac{\sum_{i=1}^{N-h}(Q_i - \bar{Q})(Q_{i+h} - \bar{Q})}{\sum_{i=1}^{N}(Q_i - \bar{Q})^2} \tag{4.7}$$

where, N is the total number of data, Q_i is the soil property value at location i and $\bar{Q}$ is the mean of the data. If d is the sample spacing (depth increment) between data points, h in Eq. 4.7 is equal to l/d.

Vanmarke (1978) expresses the scale of fluctuation, δ, in terms of the autocorrelation function, $\rho(l)$, as given below.

$$\delta = \int_{-\infty}^{+\infty} \rho(l)dl \tag{4.8}$$

Since the autocorrelation function given by Eq. 4.7 is an even function, δ can also be expressed as,

$$\delta = 2 \int_{0}^{\infty} \rho(l)dl \tag{4.9}$$

The above relationship of δ with the autocorrelation function is approximate because the value of $\rho(l)$ given by Eq. 4.7 will be accurate only at reasonably low values of the lag distance l. At increased values of l (large h), $N - h$ will be low and therefore the value of $\rho(l)$ will be biased. Agterberg (1970) suggests that for a data set with N data points, the value of $\rho(l)$ will be a reasonable estimate only for $l < (N/4)$. Expressing in terms of distance, the maximum lag distance to which ρ will be accurate, will therefore, be equal to $(N/4).d$. As a result of this, the value of the integral in Eq. 4.9 with an upper limit of infinity cannot be expected to give a good result. However, at large lag distances, the autocorrelation functions of most soil profiles tend to show a cyclic effect with values close to zero, both from the positive and negative sides,

thereby forcing $\rho(l)$ at large values of l not to have any significant influence on the estimation of δ from Eq. 4.9.

Vanmarke (1978) has also established relationships between possible autocorrelation functions and the respective scales of fluctuations which could be derived from them. These are listed in Table 4.3. The constants of the different relationships are denoted by a, b, c, k and m while ΔZ is the lag distance.

4.4   Applications of the Scale of Fluctuation

4.4.1   Comparison of Bearing, Friction and Pore Pressure

The sleeve friction value obtained from the CPT is an average value of the sleeve friction extending along the 13.4 cm length of the friction sleeve. Due to the averaging effect, the scale of fluctuation could be expected to be higher than if it were measuring point values. This is because the averaging process would cancel out fluctuations, resulting in a lower variability. The cone bearing, however, is expected to give a lower scale of fluctuation since it is believed that bearing measures values at the cone tip. Results in Table 4.4 agree with the above explanation, because the δ values for bearing is less than the corresponding values for friction. It is also interesting to note that the δ values for bearing are not as low as those for pore pressure. The pore pressure measures values over the length of the sensing element, which is about 5 mm, and for all practical purposes, can be considered as measuring values at a single point. If the cone bearing too was measuring values at a single point, the δ value relating to it also should be as low as that of pore pressure. It is evident from Table 4.4 that the bearing from the CPT is indicative of a value which is averaged out over some length. However, the fact that the bearing δ value is not as high as that for friction also suggests that the bearing value from the CPT is representative of a value averaged over some length which is less than the averaging length for friction. Table

Table 4.3: Relationship Between Different Autocorrelation Functions and the Scales of Fluctuation (after Vanmarke - 1978)
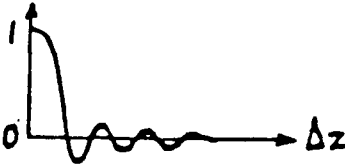
Function $\rho\,(\Delta Z)$	Shape	δ
$\mathrm{EXP}\left[\frac{-\lvert\Delta Z\rvert}{a}\right]$		$2a$
$\mathrm{EXP}\left[-\left(\frac{\Delta Z}{b}\right)^2\right]$		$\sqrt{\pi}\,b$
$\mathrm{EXP}\left[\frac{-\lvert\Delta Z\rvert}{c}\right]\cos(\Delta Z/c)$		c
$\mathrm{EXP}\left[\frac{-\lvert\Delta Z\rvert}{k}\right](1+\lvert\Delta Z\rvert/k)$		$4k$
$[\mathrm{SIN}\,(\Delta Z/m)]\,/\,(\Delta Z/m)$		πm

Table 4.4: Comparison of Averaging Dimensions for Cone Bearing, Sleeve Friction and Pore Pressure for the CPT at UBC.

Parameter	Length (cm)	Area (cm^2)
Cone Bearing	3.1	10.0
Sleeve Friction	13.4	150.0
Pore Pressure	0.5	5.6

Table 4.5: Comparison of the Scale of Fluctuation for Cone Bearing, Sleeve Friction and Pore Pressure.

Data	Layer Depths (m)	Scale of Fluctuation (cm)		
		Cone Bearing	Sleeve Friction	Pore Pressure
Haney 1	7.15 - 13.12	31.60	38.00	27.21
Haney 2	9.30 - 15.50	31.25	35.99	30.45
Haney 3	13.52 - 22.50	29.06	35.34	28.40
Langley3	2.60 - 10.60	24.08	39.98	17.91
Strong Pit	5.25 - 10.12	26.22	36.88	13.76

4.4 indicates the approximate averaging lengths and areas for bearing, friction and pore pressure obtained from the CPT at UBC.

The above argument that the cone bearing value is also indicative of an averaged value over a finite length, instead of the value at a point, is supported by the fact that the cone bearing at the tip is dependent not only on the soil property at the cone tip, but also on values immediately behind, in front of, and around the tip location. The low δ values for pore pressure could also be indicative of the highly variable nature of this measurement since it also reflects diffusion and pore pressure dissipation effects.

Figure 4.6 illustrates the relationship of the variance function × lag distance ($\Gamma^2(Z).Z$) for the Strong Pit data for bearing, friction and pore pressure. The maxima of the respective curves give the scale of fluctuation values in Table 4.5. These values have been obtained using a linear trend for bearing and friction and a curvilinear
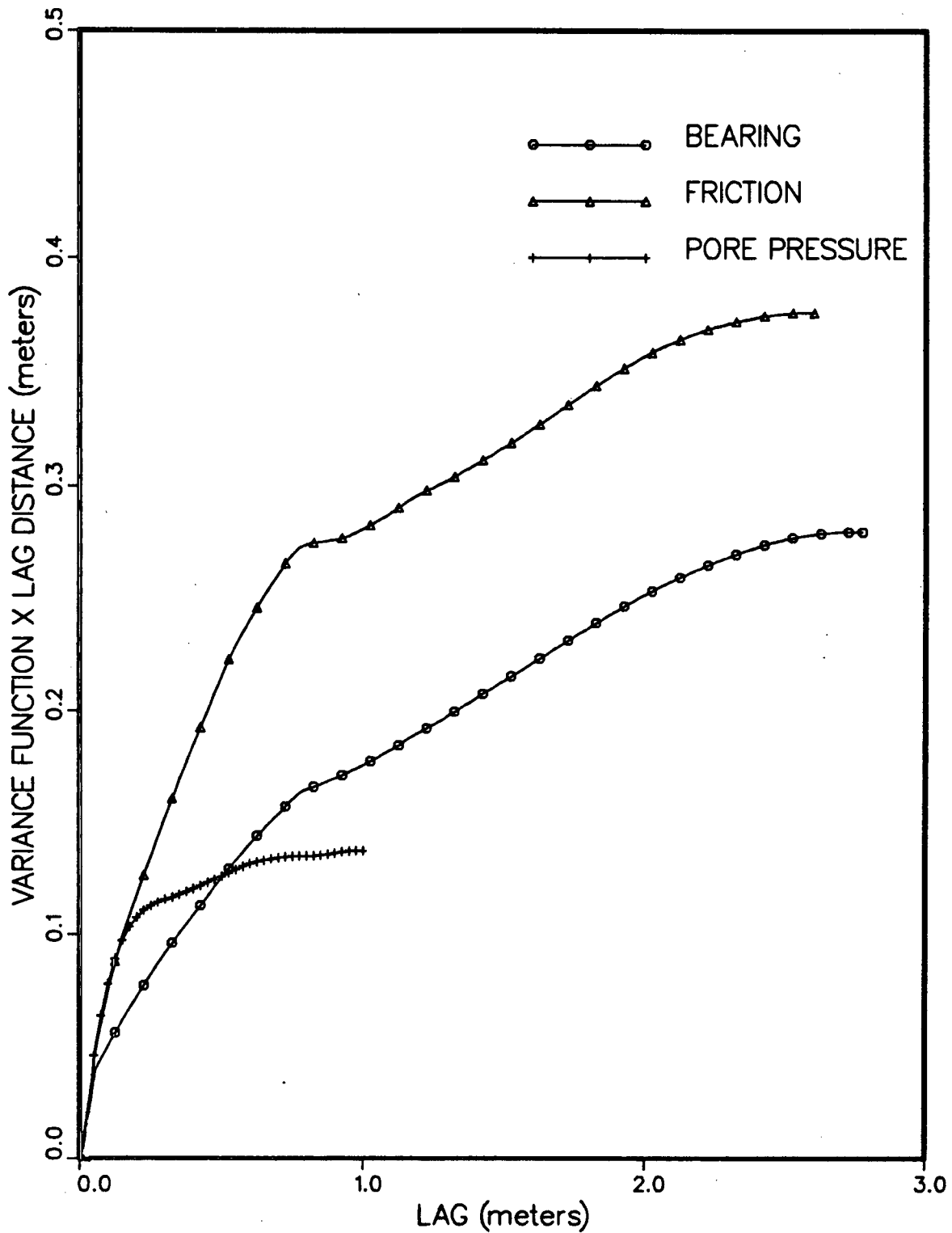
Figure 4.6: Variation of the Variance Function × Lag Distance ($\Gamma^2.Z$) for Strong Pit Data Between 5.0 and 10.0 meters for Bearing, Friction and Pore Pressure.

trend for pore pressure.

4.4.2 Scale of Fluctuation and Variability

As mentioned in section 4.3.3, the scale of fluctuation of a layer represents the variability of a layer. The higher the variability of a layer, the more fluctuations about the mean are expected, resulting in a relatively low value of δ. On the other hand, a slowly fluctuating component about the mean represents low variability, giving rise to a relatively high value of δ. It can therefore, be expected that δ is related to the coefficient of variation.

Figure 4.8 illustrates the coefficient of variation profile for the bearing profile given in Fig. 4.7 which has been divided into three layers. The average values of the coefficient of variation (η) and the δ values of the three layers are tabulated in Table 4.6. The variation of $\Gamma^2(Z).Z$ with Z from which the δ values in Table 4.6 were derived, is illustrated in Fig. 4.9. The results in Table 4.6 clearly indicate an inverse relationship of δ with the coefficient of variation. Layer 2, which has the lowest variability ($\eta = 0.078$), has the highest δ value of 41.75 cm. Layer 3 has the second highest variability ($\eta = 0.180$) and also the second highest value of δ (32.63 cm). Layer 1 is the most variable ($\eta = 0.190$) and appropriately, it has the lowest δ value of 29.37 cm. The variabilities of Layers 1 and 3 are, however, not very different and so are their scales of fluctuation, providing further evidence of the close relationship between the coefficient of variation and the scale of fluctuation.

4.4.3 Effect of Sample Spacing

In addition to assessing soil profile variability and averaging characteristics of properties such as cone bearing, sleeve friction and pore pressure, the concept of the scale of fluctuation can also be used to determine an optimum sample spacing for a given
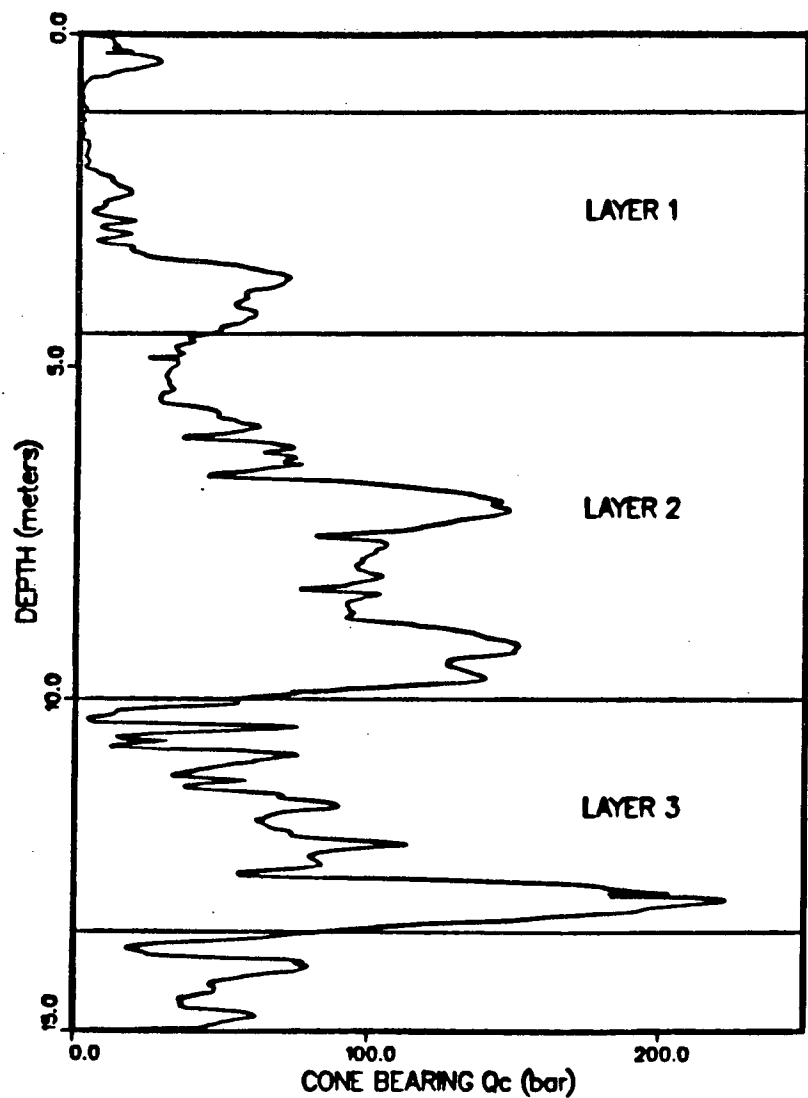
Figure 4.7: Cone Bearing Profile
at McDonald Farm.
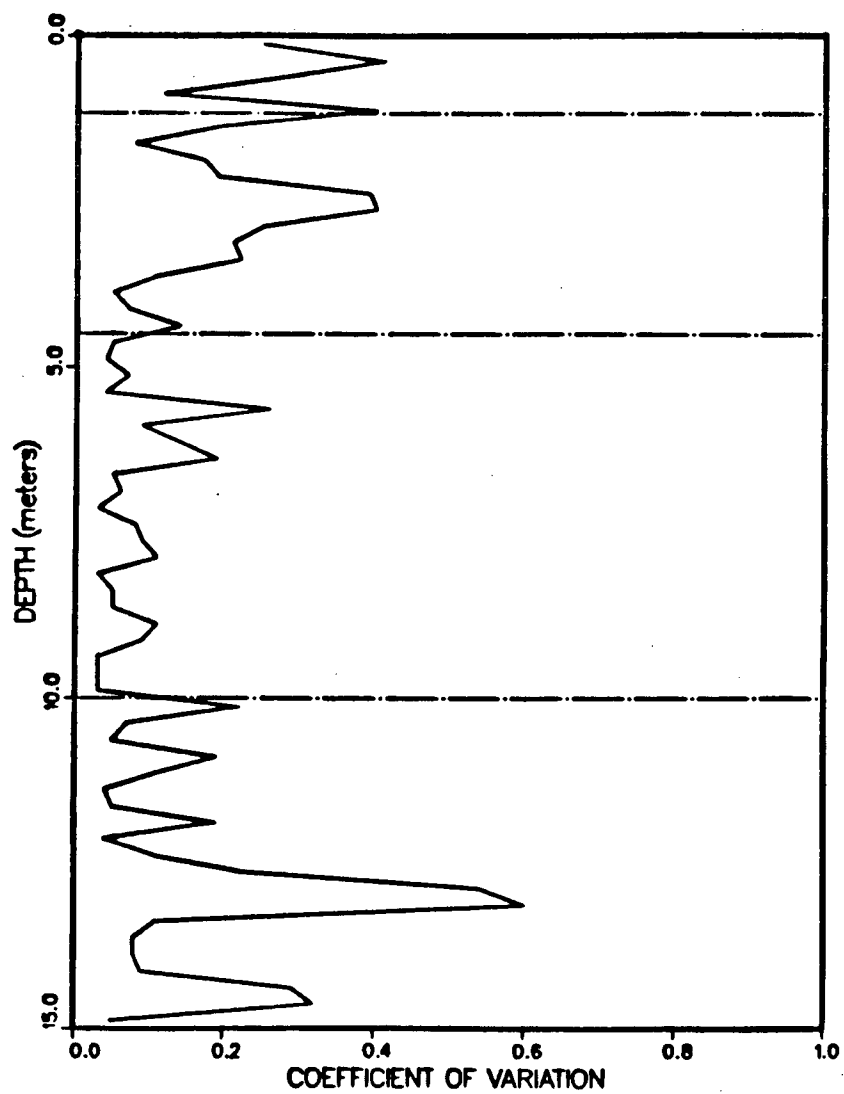
Figure 4.8: Coefficient of Variation Profile
at McDonald Farm.
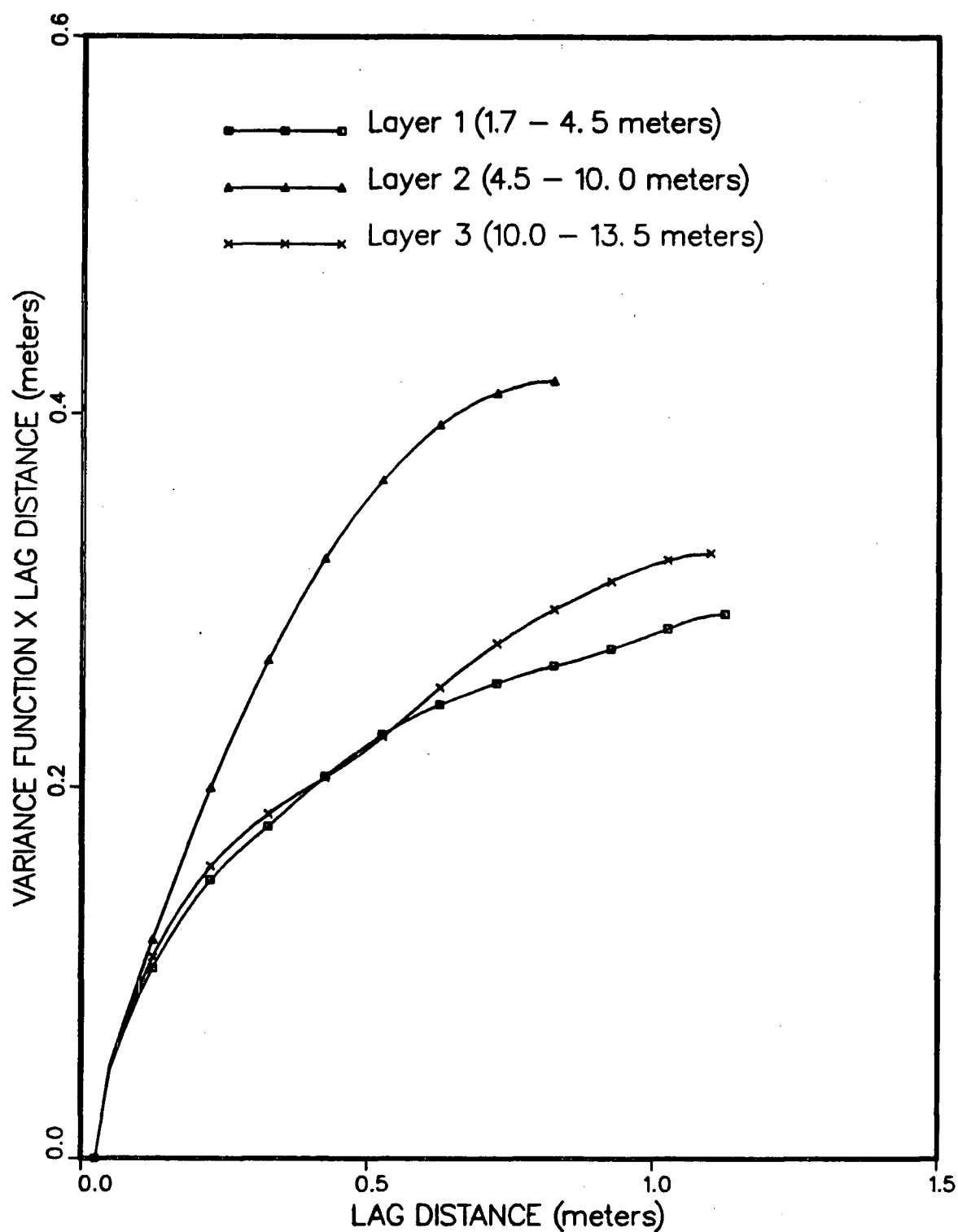
Figure 4.9: Variation of the Variance Function × Lag Distance ($\Gamma^2.Z$) for the Three Layers Identified at McDonald Farm.

Table 4.6: Relationship of the Scale of Fluctuation and the Coefficient of Variation for McDonald Farm.

Layer	Layer Depths (m)	Scale of Fluctuation δ (cm)	Coefficient of Variation η
1	1.17 - 4.50	29.37	0.190
2	4.50 - 10.12	41.75	0.078
3	10.12 - 13.50	32.63	0.180

site. Closer spacing than optimum gives rise to redundant data and unnecessary expenditure of time and effort and therefore, it is advisable to sample at the optimum spacing in order to characterize a soil profile.

Typically, CPT data recording at UBC is performed at 2.5 cm intervals. In order to study the effect of sample spacing on the scale of fluctuation, data sampling was performed at 2 mm for a sounding performed at the B. C. Hydro Railway Site. A description of this site is given in section 1.7. With the data obtained at this close spacing, intermittent data points were systematically removed to form data sets with different sample spacings and the scale of fluctuation was calculated for each spacing. The δ values for cone bearing so calculated for different spacings are given in Table 4.7. The condition of the above site was uniform in the depth interval considered and, therefore, as Table 4.7 indicates, the scale of fluctuation is fairly insensitive to the sample spacing. In a soil profile where the variability is more pronounced, the scale of fluctuation could be expected to be more sensitive to the sample spacing, and is also expected to increase with increased sample spacing. Table 4.8 exhibits this feature for Layer 2 data in Fig. 4.7. For this case, δ is equal to 41.74 cm for a spacing of 2.5 cm. and increases to 48.74 cm for a spacing of 12.5 cm. The reason for this is that the higher sample spacing in a fairly variable soil would be unable to pick up the "real" fluctuations in between these points, thus causing the scale of fluctuation to increase. The Railway site data is of such low variability that even increased sample

Table 4.7: Effect of Sample Spacing on the Scale of Fluctuation for Lower 232 Data.

Spacing of Data Points (cm)	Scale of Fluctuation δ (cm.)
0.2	35.73
0.4	35.21
0.6	35.48
0.8	36.25
1.0	34.55
1.2	35.61
1.6	35.73
2.0	35.03
4.0	33.23
6.0	35.58
8.0	35.88
10.0	36.92
12.0	37.35

spacing does not have any uniform effect on the scale of fluctuation. Instead, the δ value seems to fluctuate within a narrow band.

The importance of the scale of fluctuation is apparent when two different test methods are being compared. In this type of situation it is recommended that the sampling interval should be less than the scale of fluctuation (Vanmarke, 1978) so that comparison is being done in a similar zone. The opposite is true when sampling is performed using the same equipment, where for optimum sampling benefit a spacing

Table 4.8: Effect of Sample Spacing on the Scale of Fluctuation for Layer 2 Data of McDonald Farm Site given in Fig. 4.7.

Spacing of Data Points (cm)	Scale of Fluctuation δ (cm.)
2.5	41.74
5.0	43.30
7.5	43.96
10.0	46.59
12.5	48.74

greater than the scale of fluctuation is advisable (Campanella and Wickremesinghe, 1987).

4.5 Correlation Between Spatial Averages

It is common in geotechnical engineering practice to determine coefficients of variation, probability density functions, etc., of data sets of measured soil properties without much emphasis on the characteristics of spatially averaged soil properties. The average shear strength on a failure surface and the average shear velocity in a soil stratum, which are of great concern to the geotechnical engineer, are some examples of these. Very often, it is assumed that the mean of a spatially averaged soil property does not depend on the averaging dimensions in a statistical homogeneous medium. The aspect of independence between soil property values and averaging dimensions is very desirable, although it is often violated. Lack of correlation will only be exhibited by soil properties for which element averages combine linearly. Vanmarke (1978) also explains that the spatial averages will have narrower probability density functions than the corresponding 'point' values.

The correlation coefficient concept derived by Vanmarke (1984) provides the basis for new methodology to analyze a wide range of stochastic problems in all three spatial dimensions. The almost continuous profile obtained from the cone penetration test (CPT) provides an ideal data base for such considerations in the vertical direction. This procedure introduced by Vanmarke (1984) can be used in the field of numerical methods in geomechanics, by generating the matrix of correlation coefficients between pairs of local averages of some soil property associated with different elements along the vertical axis.

Considering Layers A' and B' in Fig. 4.10, let y_1, y_2, y_3, y_a, y_o and y_b be the distances illustrated in Fig.4.10 and $\Gamma^2(y_1)$, $\Gamma^2(y_2)$, $\Gamma^2(y_3)$, $\Gamma^2(y_a)$, $\Gamma^2(y_o)$ and $\Gamma^2(y_b)$ be their
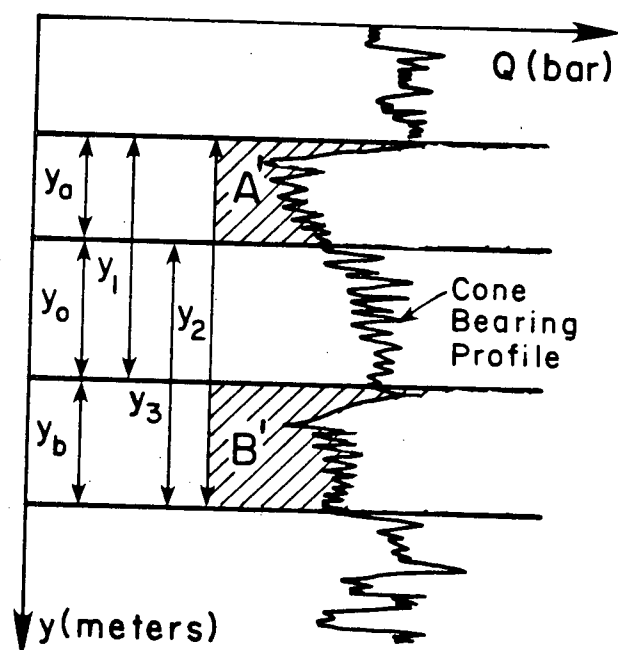
Figure 4.10: Different Layers Used for the Determination of the Coefficient of Correlation Between Spatial Averages.

Table 4.9: Correlation Coefficients for Layer A in Fig. 4.11.

y_a (m)	y_b (m)	y_o (m)	ρ_{ab}
1.0	1.0	0.30	.4320
		0.50	.3500
		0.80	.0215
		1.00	.0125
0.8	1.4	0.50	.2504
		1.00	.1463
		1.50	-.0227

Table 4.10: Correlation Coefficients for Layer B in Fig. 4.11.

y_a (m)	y_b (m)	y_o (m)	ρ_{ab}
1.0	1.0	0.20	.4497
		0.40	.3651
		0.60	.2853
		0.80	.2321
		1.00	.2167

respective variance functions.

The correlation coefficient ρ_{ab} between the spatial average of Layer A' (Q_a) and the spatial average of Layer B' (Q_b) in Fig.4.10 is given by,

$$\rho_{ab} = \frac{y_o{}^2\Gamma^2(y_o) - y_1{}^2\Gamma^2(y_1) + y_2{}^2\Gamma^2(y_2) - y_3{}^2\Gamma^2(y_3)}{2(\sqrt{(y_a{}^2\Gamma^2(y_a) \cdot y_b{}^2\Gamma^2(y_b))})} \qquad (4.10)$$

The derivation of Eq. 4.10 is given in Appendix A.

The above concept of determining the coefficient of correlation has been applied to a cone bearing profile obtained from Tilbury Island (Fig. 4.11). The upper 20 m of this profile was used in Chapter 2 (Fig. 2.25). Two distinct layers have been identified between 24.8 - 30.0 m (Layer A) and between 30.0 - 40.0 m (Layer B),
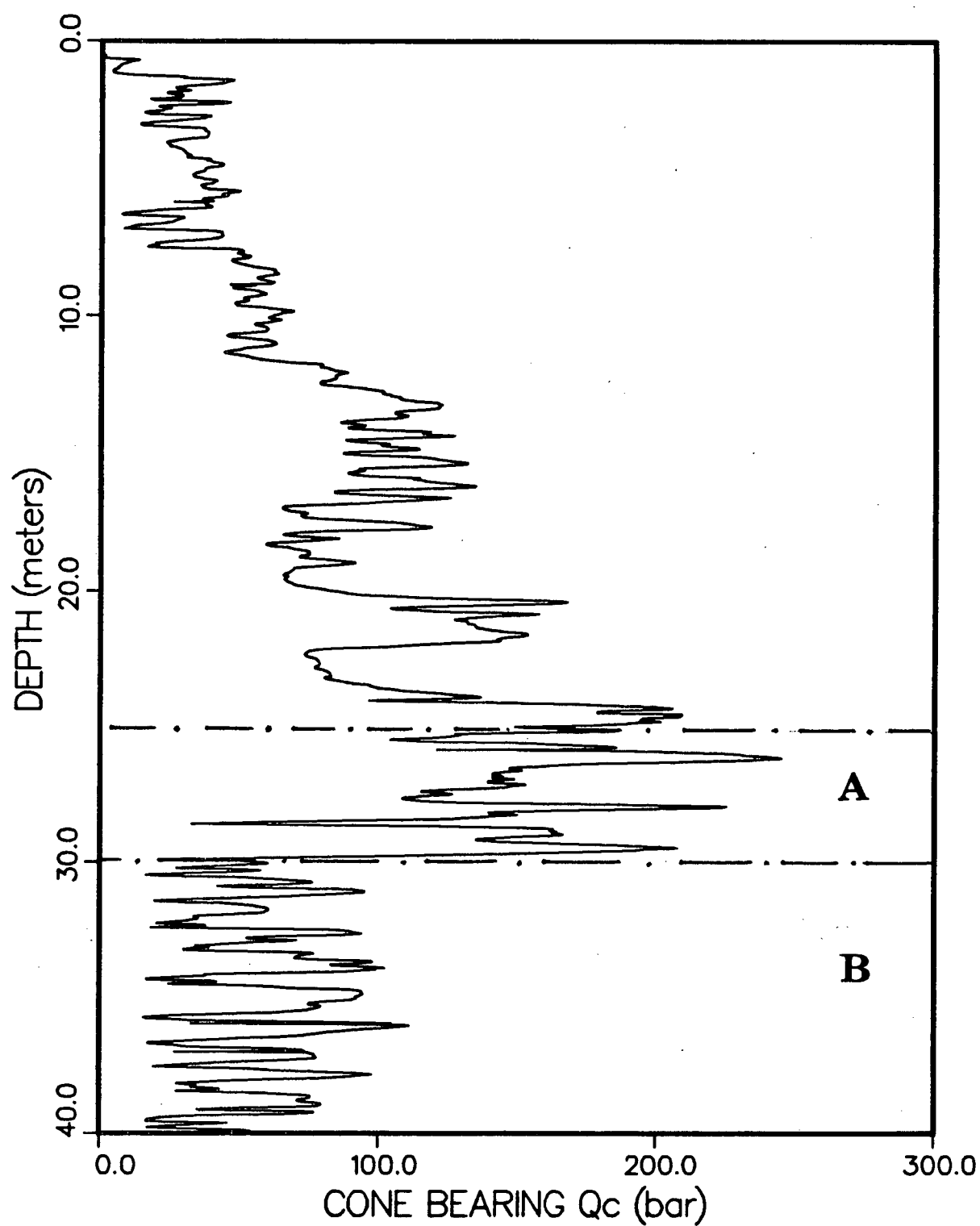
Figure 4.11: Cone Bearing Profile at Tilbury Island with Layers A and B.

using the methods described in Chapter 2. The stationarity of these layers have been confirmed using the RUN test described in Chapter 3.

As can be seen from Eq. 4.10, the only parameters needed to calculate the correlation coefficient are the thicknesses of the layers and their respective variance functions. The variance functions of these two layers are illustrated in Figs. 4.12 and 4.13, together with the autocorrelation functions which will be required for the determination of exceedance probabilities to be described in section 4.6. The correlation coefficients for different sublayer thicknesses (y_a and y_b) and layer separation distances (y_o) are tabulated in Table 4.9 for Layer A and Table 4.10 for Layer B. The discrepancy of the correlation values for similar separation distances in the two layers is due to the difference in the decay pattern of the respective variance functions (Figs. 4.12 and 4.13). Figures 4.12 and 4.13 show the variation patterns of the autocorrelation function and the variance function for lag distances of up to 2.0 and 4.0 meters, respectively. This is due to the reason that the autocorrelation function has been considered to be accurate up to 40% of the length of the data set.

As mentioned previously, it is customary in geotechnical engineering to assume independence of soil properties, that is, considering the correlation coefficient to be zero. This is incorrect as demonstrated above, and could result in significant error of the estimates being calculated. In a geotechnical engineering study the estimate under consideration may be the settlement of a foundation.

4.6   Exceedance Probabilities

The theory of random functions could be applied to find the exceedance probabilities of a CPT profile. Instead of considering the entire soil layer, geotechnical performance may have to be evaluated based on the exceedance of some value q within a region of the soil profile. The slope stability problem in geotechnical engineering, where the
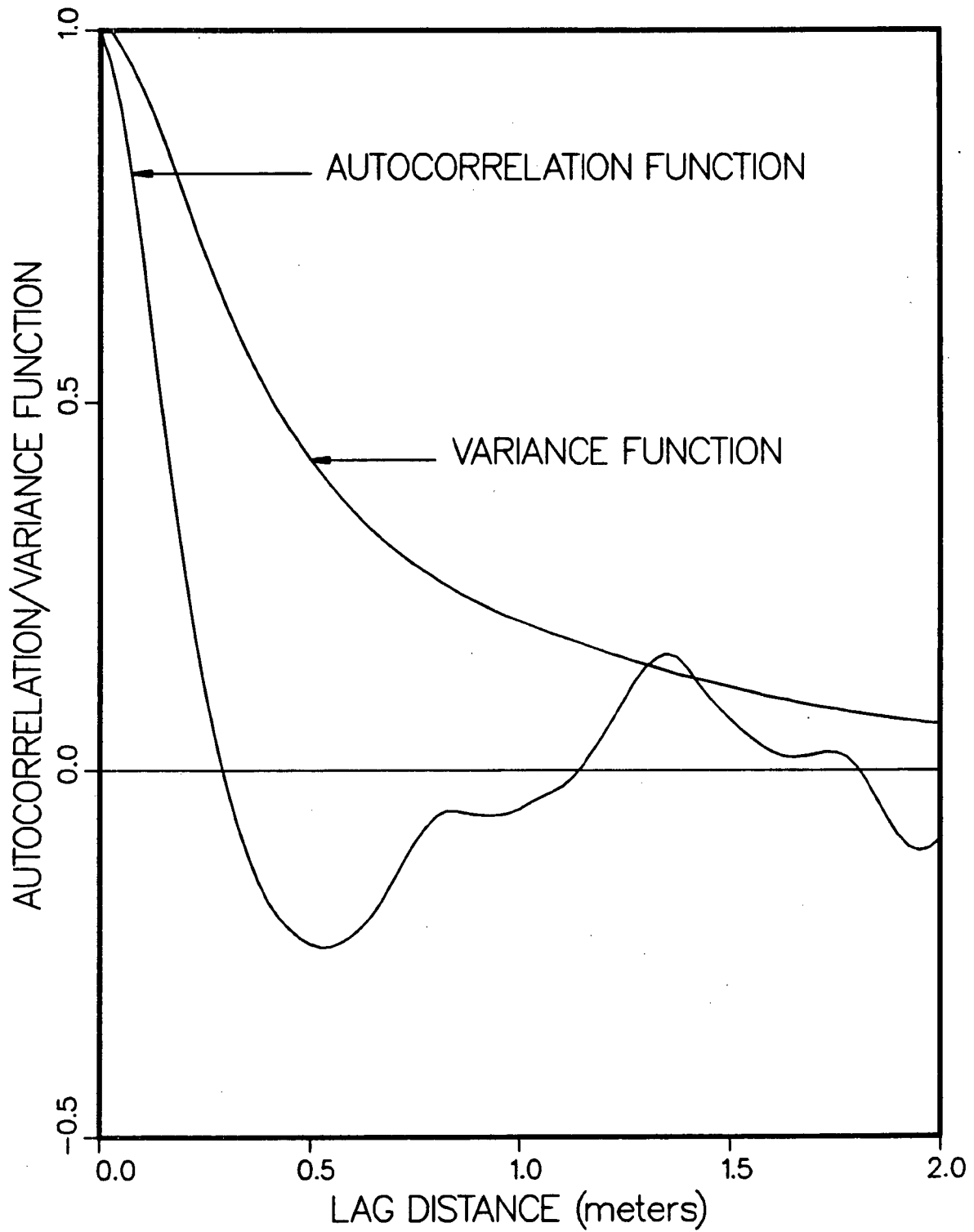
Figure 4.12: Autocorrelation Function and the Variance Function of Layer A (25.0 - 30.0 meters) of Tilbury Island.
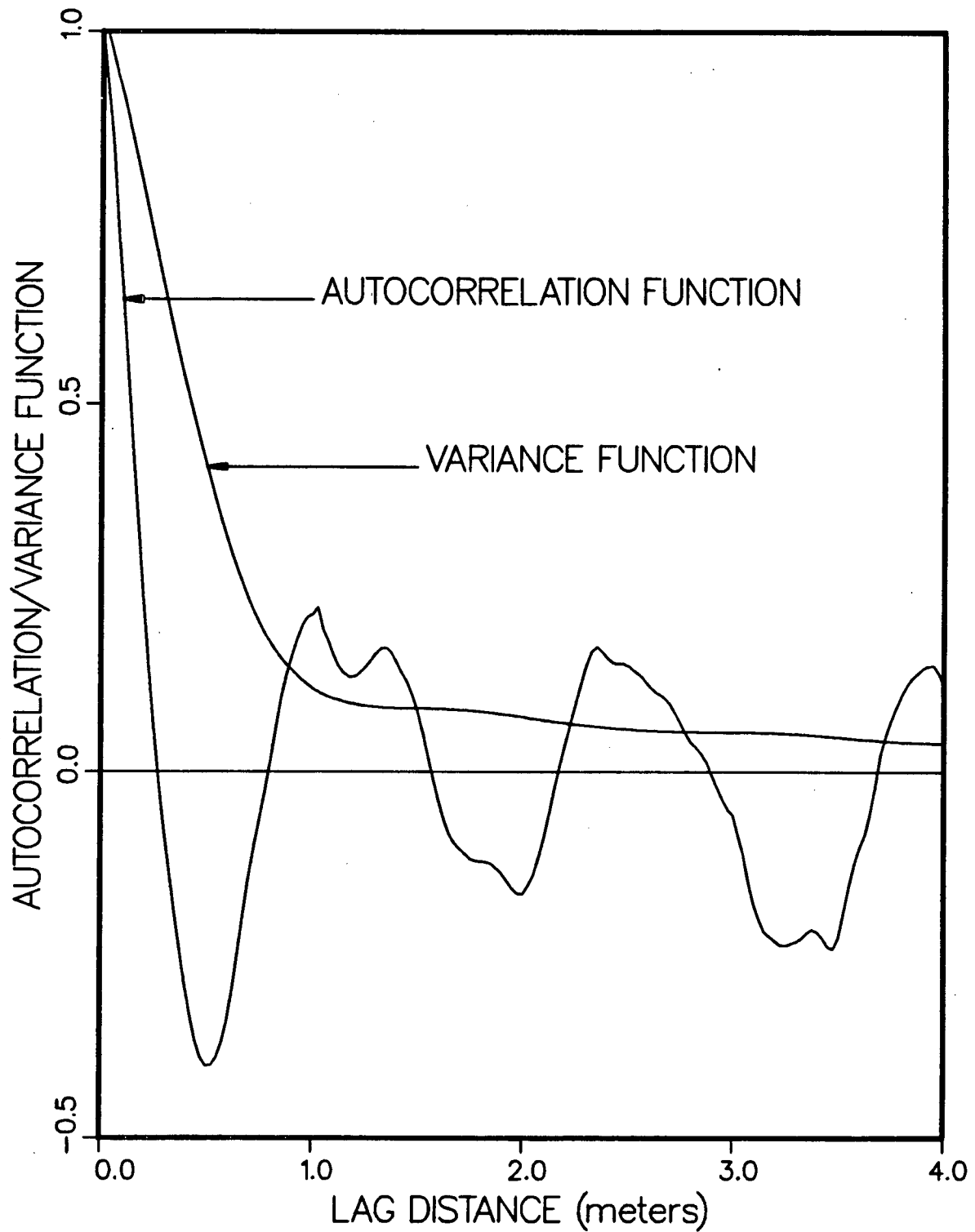
Figure 4.13: Autocorrelation Function and the Variance Function of Layer B (30.0 - 40.0 meters) of Tilbury Island.

main concern is the non exceedance of the available shear strength by the disturbing force, is an example. If this is not satisfied even in a very thin layer of soil, this thin layer is liable to progressive failure. The theory was originally introduced by Vanmarke (1983) and was eventually extended to cover multi-layered systems by Tang et al.(1987).

The mean rate of upcrossings (ν_q^+) above a threshold value q in a local region of length D is given by (Vanmarke, 1987 and Tang et al., 1987),

$$\nu_q^+ = \frac{\sqrt{1 - \rho(D)}}{\sqrt{2\pi} D \Gamma(D)} \exp\left\{ -\frac{(q - \bar{Q})^2}{2\sigma^2} \Gamma^2(D) \right\} \tag{4.11}$$

The derivation of Eq. 4.11 is given in Appendix B.

There will be many segments of length D within the domain length L, and hence the probability of non exceedance (P_L) for all such segments within the entire layer of length L will be approximately given by,

$$P_L = \exp(-\nu_q^+ L) \tag{4.12}$$

Therefore, the probability that the average of a local interval of length D will exceed (probability of exceedance) a threshold value q is given by,

$$P_E = 1 - \exp(-\nu_q^+ L) \tag{4.13}$$

From Eqs. 4.11 and 4.13 it is evident that the probability of exceedance is dependent on the local region of length D, mean ($\bar{Q}$) and standard deviation (σ_Q) of the entire layer, value of the autocorrelation function of the layer at D (ρ_D), square root of the

variance function at D, ($\Gamma(D)$), the threshold value q and L, the thickness of the domain.

The theory of exceedance probabilities has been applied to the cone bearing profile of Layer A and Layer B in Fig. 4.11. The autocorrelation and variance functions of these two layers are illustrated in Figs. 4.12 (Layer A) and 4.13 (Layer B). The mean and standard deviation of Layer A are 149.30 bar and 43.17 bar, respectively, and for Layer B, the mean is 58.26 bar and the standard deviation is 24.65 bar. Layer B has a higher variability with a scale of fluctuation (δ) of 20.0 cm and a coefficient of variation (η) of 0.411, as compared to Layer A which has a scale of fluctuation of 21.34 cm. and a coefficient of variation of 0.289.

Figure 4.14 illustrates the effect of the length of the local interval D and threshold value (q) on the probability of exceedance for Layer A. For any given local interval of length D, the probability of exceedance of a local average increases as the threshold value approaches the mean value of the layer as shown in Fig. 4.14. For any given q, the probability of exceedance decreases with increasing length of the local interval D, due to the effect of averaging within the local region. Figure 4.15 demonstrates a similar behavior for Layer B, which illustrates the increase of the probability of exceedance as the threshold value, q, approaches the mean value (58.26 bar) of that layer. The increase in the probability of exceedance with decreasing local averaging interval, D, is also evident from Fig. 4.15.

When the length of a domain which comprises the smaller segments of length D is decreased, the probability of exceedance also decreases, as exhibited in Fig. 4.16. The probability of exceedance, P_E, for a given domain of length L, is the cumulative effect of all local intervals of length D (Tang, 1988). The number of local intervals which can be included in a domain increases with increasing L and this is the reason for the increase in P_E for increasing L, for a given threshold value q (Fig.4.16).
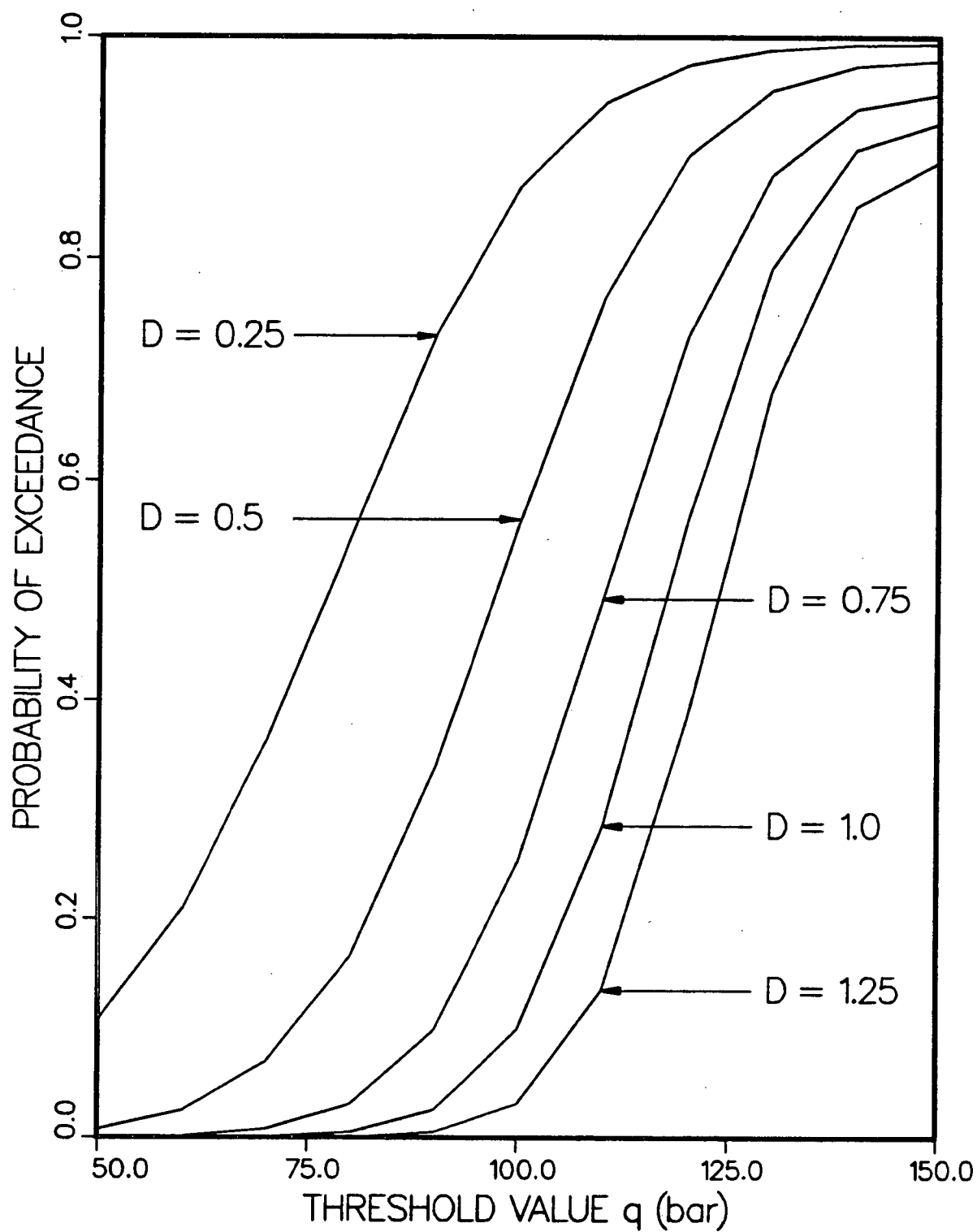
Figure 4.14: Relationship of the Probability of Exceedance with Threshold Value for Different Local Regions of Length D for Layer A at Tilbury Island.
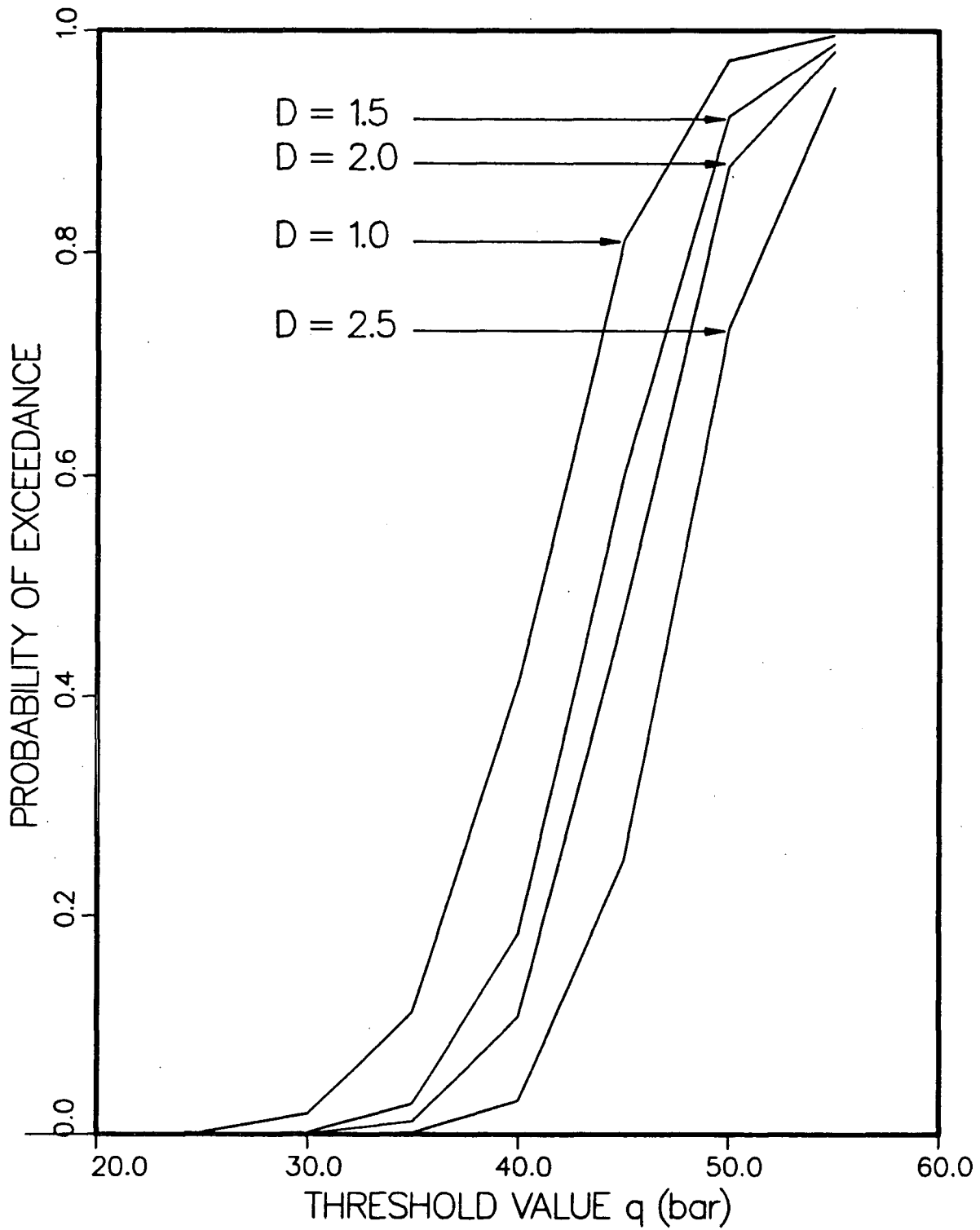
Figure 4.15: Relationship of the Probability of Exceedance with Threshold Value for Different Local Regions of Length D for Layer B at Tilbury Island.
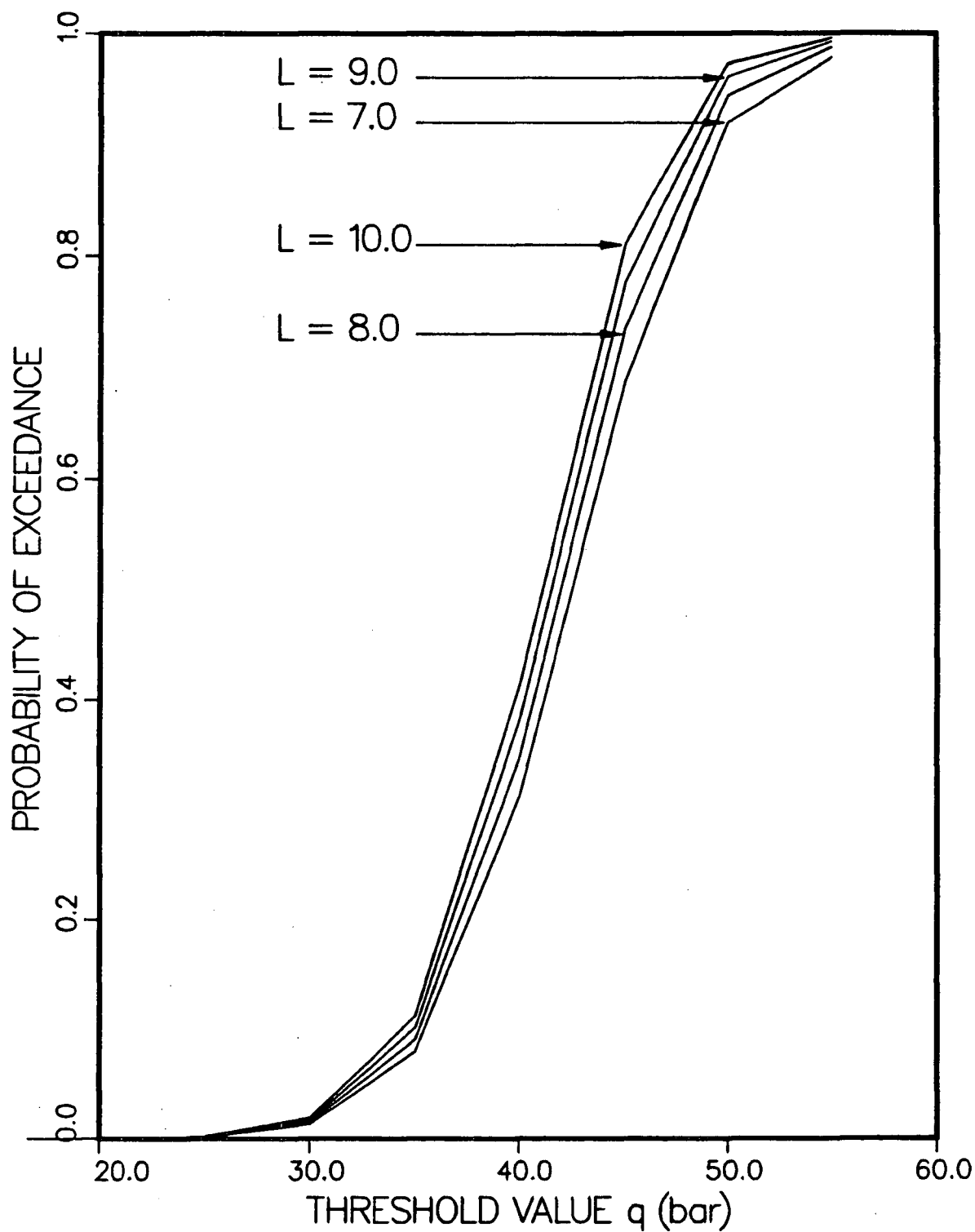
Figure 4.16: Relationship of the Probability of Exceedance with Threshold Value for Different Domain Lengths L for Layer B at Tilbury Island.

Figure 4.17 demonstrates the effect of variability on P_E for local averaging intervals of 0.5 and 1.0 meters for both Layers A and B. For a given value of D, Layer B which has a higher variability ($\eta = .411$ and $\delta = 20.0$ cm) exhibits a higher value of P_E for any given value of $q/\bar{Q}$, as compared to the less variable Layer A ($\eta = .289$ and $\delta = 21.34$ cm.). For comparison purposes, the values of q have been normalized by dividing by the respective means, $\bar{Q}$, to account for the difference of the means of the two layers under consideration. Higher variability reflects more uncertainty, which in turn influences the exceedance probabilities to increase. This phenomenon is amply evident from Fig.4.17. The increase of P_E with decreasing D, is also apparent from Fig.4.17 due to reasons already explained.

4.7   Optimum Sample Spacing

A typical cone penetration test at UBC performs data logging at 2.5 cm. However, if the soil does not exhibit much variability, the sample spacing could be increased without losing much information. In section 4.4.3, a method was described to obtain an optimum sample spacing based on the scale of fluctuation. The technique to be described in this section is more advantageous in that an optimum sample spacing required can be determined based on the confidence level needed for a particular purpose.

The actual mean of the data, Q_θ, is the mean calculated if all the points in a particular sublayer were sampled, and the average calculated. In the actual situation, what is available is an estimate $\bar{Q}$.

Assuming that the data are normally distributed, the limits of $\bar{Q}$ are given by,

$$| Q_\theta - \bar{Q} | = \frac{\sigma}{\sqrt{n}} . t_{n-1}^{\gamma} \qquad (4.14)$$
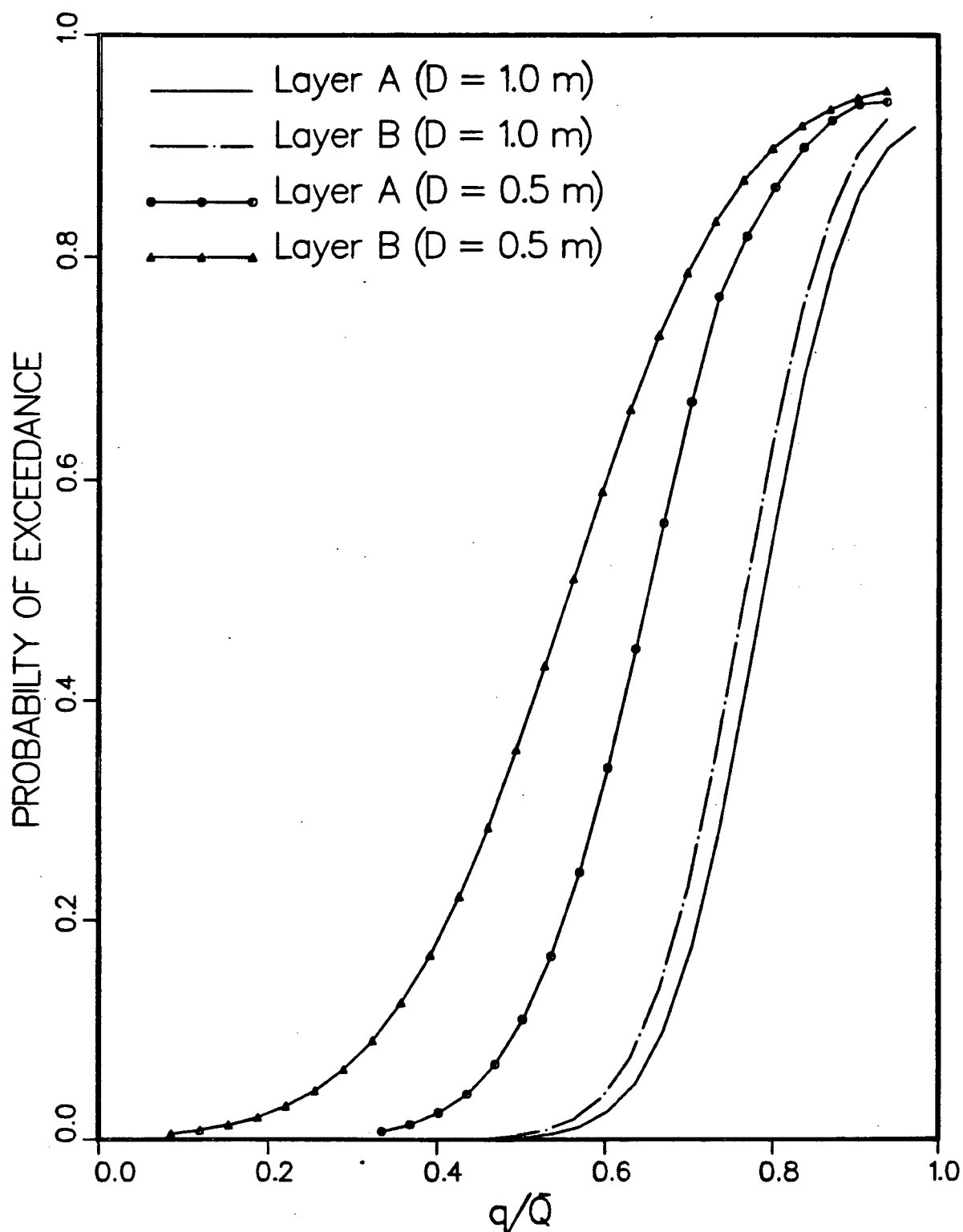
Figure 4.17: Relationship of the Probability of Exceedance with Threshold Value for Layer A (low variabilty) and Layer B (higher variabilty) at Tilbury Island.

or,

$$Q_\theta = \bar{Q} \pm \frac{\sigma}{\sqrt{n}} . t^\gamma_{n-1} \qquad (4.15)$$

where, σ is the standard deviation of the data and t^γ_{n-1} is the Student's 't' variate with $n-1$ degrees of freedom, n being the number of data. The above equations should satisfy,

$$\text{Prob}\{t \geq t^\gamma_{n-1}\} = \gamma \qquad (4.16)$$

where, $(1 - \gamma)$ is the confidence level of the estimation. Let it be assumed that any layer is fully characterized when the mean obtained from the data, $\bar{Q}$, for a particular layer is within $\pm\Delta$ of the actual mean Q_θ. For example, if Δ, hereafter referred to as the degree of tolerance is $\pm.10$, the following condition results: $0.9Q_\theta < \bar{Q} < 1.10Q_\theta$. The tolerance is inversely related to the precision; the higher the tolerance, the lower the precision.

As a result of the above definitions, Δ can be expressed as,

$$\Delta = \frac{Q_\theta - \bar{Q}}{\bar{Q}} \qquad (4.17)$$

The coefficient of variation η is given by,

$$\eta = \frac{\sigma}{\bar{Q}} \qquad (4.18)$$

The sample size (n) required to estimate the mean to the above precision or the sample size required to characterize a soil layer with respect to the mean of a soil

property considered, can be expressed as a function of the degree of tolerance, Δ, with a confidence level of $(1 - \gamma)$.

By combining Eqs. 4.15, 4.17 and 4.18, the sample size (n) is given by,

$$n = \left(\frac{\eta t^{\gamma}_{n-1}}{\Delta}\right)^2 \tag{4.19}$$

According to Eq. 4.19, n depends on three factors;

(a) Variability of the soil layer, expressed by η

(b) Confidence required of the estimate, expressed by t^{γ}_{n-1}

(c) The degree of tolerance allowed, expressed by Δ

The number of samples needed in a given thickness of soil stratum is proportional to the square of the coefficient of variation and the confidence required and inversely proportional to the square of the degree of tolerance. In other words, the sample spacing required, which is the inverse of n, is proportional to the square of the degree of tolerance and inversely proportional to the square of the coefficient of variation and the confidence level.

The above concept has been applied to two sets of data; namely Layer A and the upper five meters of Layer B in Fig. 4.11. These two layers were selected to have the same thickness in order to demonstrate the effect of variability on the optimum sample spacing more explicitly. The results for varying degrees of confidence, 80%, 90% and 95%, for two different degrees of tolerance $\pm$ .05 and $\pm$ .10 are tabulated in Table 4.10 for the two layers considered.

In Table 4.11, for the same tolerance of $\pm$ .10 and a confidence level of 80%, the

sample spacing required in the soil of higher variability ($\eta = .410$) is 17.2 cm while for the soil with a lower variability ($\eta = .289$), it is 35.0 cm. If a higher confidence level of 90% is required, spacing will have to be decreased to 10.4 cm. in the more variable soil and to 20.0 cm. in the less variable soil. Similarly, if the engineer intends to reduce the tolerance by half to $\pm$ .05 in order to increase the precision of the estimate, the sample spacing will have to be reduced to 5.4 cm in Layer A for the same confidence 90%. However, in the more variable soil, the sample spacing required for the same confidence level and tolerance is as low as 2.7 cm. For a higher confidence level of 95%, the spacings required for a tolerance of $\pm$ .05 decrease even further with the low variable soil requiring a spacing of 3.8 cm and the high variable soil a very low 1.9 cm.

The usual sample spacing of 2.5 cm in the high variable soil will result in a confidence level between 90% and 95% for a tolerance of $\pm$ .05 while for a tolerance of $\pm$ .10 the confidence level of the required estimate will be in excess of 99%. In the lower variable soil the confidence level will be significantly higher than 99% for a tolerance level of $\pm$ .10 while for a reduced tolerance of $\pm$ .05 it will be close to 99%. All these significance level are well above what is required for all practical purposes in geotechnical engineering and therefore the sampling interval can be increased at the expense of a decreased confidence. The other option would be to fix a confidence level and study the effect of sampling on the precision (inverse of tolerance) of the estimate.

The above examples clearly illustrate the importance of considering three important criteria when selecting a sample spacing for a soil investigation. They are the variability of the soil, the acceptable precision and the confidence required in the estimate. These criteria have different effects on the sampling rate and perhaps the most important factor is the variability of the soil stratum. In a soil stratum where the

Table 4.11: Effect of Variability on the Optimum Sample Spacing for the Soil Layers between 25.0 - 30.0 meters and 30.0 - 35.0 meters in Fig. 4.11.

Layer (m)	Tolerance Δ	Confidence Level (%)	n	Spacing (cm)
25.0 - 30.0	$\pm.05$	80	58	8.7
		90	93	5.4
Low Variability		95	132	3.8
$\eta = 0.289$	$\pm.10$	80	15	35.0
		90	26	20.0
		95	36	13.8
30.0 - 35.0	$\pm.05$	80	112	4.5
		90	184	2.7
High Variability		95	262	1.9
$\eta = 0.410$	$\pm.10$	80	30	17.2
		90	49	10.4
		95	68	7.4

coefficient of variation, η, is unknown, it will be necessary to perform some tests and obtain an approximate estimate. The spacing required can then be determined using this estimate, and in the event it is greater than the spacing at which the testing has already been performed, there is no need for additional testing. However, if it is not, more testing will have to be done, also enabling a better estimate of the coefficient of variation, which in turn would result in a more accurate estimation of the optimum sample spacing.

It is common to find soil profiles exhibiting a trend, resulting in a fairly high coefficient of variation, thus giving rise to the need of closer sample spacing to ensure a higher accuracy of the estimates. If the above method is to be used to obtain the sample spacing, the uncertainty due to the trend will also have to be taken into account. Another available option would be to consider very thin soil layers, whereby, the effect of violating the assumption of stationarity would not be very drastic.

4.8 Conclusions

The main conclusions of this chapter are,

(i) The proposed method of obtaining the scale of fluctuation in this thesis gives comparable results to those of the original method suggested by Vanmarke (1977). The advantages of the proposed method are its adaptability to computer applications and the consistency of the approach, as compared to the subjectivity involved in the previous method.

(ii) The scale of fluctuation is basically an enhanced estimator of variability. In contrast to the coefficient of variability, it also gives an indication of the spatial variation of soil properties.

(iii) The scale of fluctuation has to be determined on data which have been made stationary by trend removal. A linear trend can be used for the cone bearing and sleeve friction data while a curvilinear trend represents a pore pressure profile more adequately. Effects of non stationarity significantly increases the value of the scale of fluctuation and it is therefore important to select the most appropriate form of trend to remove it.

(iv) Exceedance probabilities of soil properties over threshold values are useful in problems such as slope stability, where the requirement is the non-exceedance of the disturbing force over the available strength. The exceedance probability is strongly dependent on the variability of the soil under consideration. Exceedance probabilities are also heavily dependent on the length of the local region D, which decision is a matter of soil mechanics of sensitivity and progressive failure.

(v) Economics play a vital role in site investigations for large projects, and in this regard, the unnecessary collection of data can be avoided in order to minimize costs. On the other hand, the increased risks involved in having insufficient data or failure to detect anomalous soil zones can be catastrophic. With the above two

considerations, an optimum sampling spacing can be derived. The optimum sampling spacing required to fully characterize a soil profile has been found to be dependent on the accuracy of the estimates required, the confidence level desired, and most importantly on the inherent soil variability.

This chapter has amply demonstrated how applications of random field theory can be effectively extended to analyze cone penetration test data in order to better characterize a soil stratum. In the past, these techniques could not be used due to the lack of sufficient closely spaced data. This is not the case with the emergence and popularity of in situ testing devices such as the cone penetration test (CPT). It is recommended that geotechnical engineers not only use these large data bases for conventional logging purposes, but should also attempt to utilize these data analytically from a statistical aspect, to obtain a better understanding of the characteristics of a soil stratum. Statistical techniques enable the accrual of valuable information at no additional cost and therefore, should be used at every opportunity in supplementing the information gathered from traditional deterministic methods.

Chapter 5

Time Series Methods

5.1 Introduction

A time series relates observations obtained in the past and present with values to be expected in the future. In data analysis dealt with in this thesis, data will be with respect to a spatial coordinate instead of time, although the methods of time series are directly applicable. Therefore " Time Series Methods " in this dissertation actually implies, " Time Series Methods Applied to Spatial Variations ". There are two conditions which have to be satisfied for the application of time series methods: the presence of correlation among data and the requirement that data are at equally spaced intervals. Geotechnical data obtained from in situ test methods satisfy both these conditions and therefore, provide an ideal base for the application of time series methods. There exists a major difference in the application of time series methods in geotechnical engineering and in the classical areas of applications in commerce, economics, etc.. In the latter fields, both interpolation and extrapolation are performed while in geotechnical engineering it only makes sense to carry out interpolation.

In this application, time series methods in geotechnical data analysis can serve two purposes: first to model soil data profiles in order to be able to interpolate between known data points and secondly to estimate the random error component of a data set obtained using a particular test method. Knowledge of the random error not only allows a comparison of the different test methods but also permits the determination of the inherent variability which is important to characterize a soil.

177

Prior to modeling a profile it is necessary that the data be stationary. This can be performed either by using methods of trend analysis described in Chapter 3 or by the technique of differencing to be explained in section 5.2. The types of models which can be used are the Autoregressive (AR) model, Moving Average (MA) model and the Autoregressive Integrated Moving Average (MA) model and these will be described in sections 5.3.1 to 5.3.2. These methods are also referred to as Box - Jenkins methods (Box and Jenkins, 1976) and have been used in the SAS (SASLETS, 1982) package which was employed to perform the modeling to be described in this chapter.

Box - Jenkins methods can also be employed to determine the random testing error of soil test data. These methods have also been made use of by Wu (1985). The random testing estimated from the direct use of time series methods can also be compared to the random error obtained by using the autocorrelation function of the data.

This chapter will also contain a brief review of the types of errors encountered in geotechnical data analysis.

5.2 The Method of Differencing

Although trend removal using linear and non-linear regression techniques are widely used to stationarize data, time series methods use the method of differencing to transform the data to a stationary form.

The method of differencing consists of subtracting values of the observations from one another in some prescribed order. A first order difference transformation is defined as the difference between adjacent observations. Second order differencing consists of taking differences of the single differenced series, and so on. Table 5.1 explains the concept of differencing more clearly.

In Table 5.1, Q_{s-2}, Q_{s-1}, Q_s, Q_{s+1} and Q_{s+2} are sequential data at depths, d_{s-2},

Table 5.1: Effects of the Degree of Differencing on Data.

Depth	Raw Data	Modified Data at Different Degrees of Differencing		
		First Degree	Second Degree	Third Degree
d_{s-2}	Q_{s-2}			
d_{s-1}	Q_{s-1}	$Q_{s-1} - Q_{s-2}$		
d_s	Q_s	$Q_s - Q_{s-1}$	$Q_s - 2Q_{s-1} + Q_{s-2}$	
d_{s+1}	Q_{s+1}	$Q_{s+1} - Q_s$	$Q_{s+1} - 2Q_s + Q_{s-1}$	$Q_{s+1} - 3(Q_s - Q_{s-1}) - Q_{s-2}$
d_{s+2}	Q_{s+2}	$Q_{s+2} - Q_{s+1}$	$Q_{s+2} - 2Q_{s+1} + Q_s$	$Q_{s+2} - 3(Q_{s+1} - Q_s) - Q_{s-1}$

d_{s-1}, d_s, d_{s+1} and d_{s+2}, respectively, and as can be observed, each degree of differencing results in the loss of a single data point.

A first degree differencing removes a linear trend, second degree differencing removes a polynomial trend of order 2, and third degree differencing removes a polynomial of order 3. In most applications of space series (the equivalent of time series where the time domain is replaced by the spatial domain) analysis, stationarity of data is a pre-requisite. In geotechnical engineering, it is customary to divide the entire soil stratum into sub layers exhibiting a similar type of trend which is very often linear, but in rare occasions, curvilinear, necessitating a first or second degree differencing, respectively. A visual inspection of the soil parameter profile will very often give an indication as to the type of differencing required but may not always be the case. The following method can be used to determine the degree of differencing required for the trend removal of a data profile.

For data that have been differenced to different degrees (j = 1, 2, 3, ...), calculate the statistic λ_j defined as (Gottman, 1981),

$$\lambda_j = \sum_{k=0}^{N/6} \rho_{jk} \tag{5.1}$$

where N is the total number of data, k the number of lags and ρ_{jk} is the autocorrelation

coefficient of the data which has been differenced to the jth degree at a lag k. When the data set is over differenced λ_j begins to increase. The degree of differencing required to stationarize the data set is then taken to be as the value of j for which $\lambda_{j+1} < \lambda_j$.

5.3 Types of Models

5.3.1 Autoregressive (AR) Models

A time series or space series can be described as an autoregressive process if the current value of the series Q_s can be expressed as a linear function of the previous values plus a random term a_s.

An AR model of order 'p' [AR(p)] can be expressed as (Box and Jenkins, 1976),

$$Q_s = \phi_1 Q_{s-1} + \phi_2 Q_{s-2} + \phi_3 Q_{s-3} + \cdots\cdots + \phi_p Q_{s-p} + a_s \tag{5.2}$$

where ϕ_i's are the autoregressive coefficients.

An AR(1) model is simply expressed as,

$$Q_s = \phi_1 Q_{s-1} + a_s \tag{5.3}$$

For the special case where ϕ_1 is unity, the random walk model results,

$$Q_s = Q_{s-1} + a_s \tag{5.4}$$

5.3.2   Moving Average (MA) Models

In a Moving Average (MA) model, the current value Q_s can be expressed as a summation of the present and past noise or shock terms (Box and Jenkins, 1976). A MA model of order 'q' [MA(q)] can be expressed as,

$$Q_s = a_s - \theta_1 a_{s-1} - \theta_2 a_{s-2} - \theta_3 a_{s-3} \cdots\cdots - \theta_q a_{s-q} \qquad (5.5)$$

where θ_i's are the moving average coefficients.

A simple MA(1) process can be expressed as,

$$Q_s = a_s - \theta_1 a_{s-1} \qquad (5.6)$$

5.3.3   Combination of AR and MA Models (ARIMA)

In any method of modeling, it is preferable that the least number of parameters be used and in this regard, the ARIMA model which is a combination of both AR and MA models is very useful.

A general form of a (p,q) ARIMA model can be expressed as,

$$Q_s = \phi_1 Q_{s-1} + \phi_2 Q_{s-2} + \cdots + \phi_p Q_{s-p} + a_s - \theta_1 a_{s-1} - \theta_2 a_{s-2} - \cdots - \theta_q a_{s-q} \quad (5.7)$$

5.4   Choice of An Appropriate Model

The choice of the most appropriate model for a data profile is dependent on two functions. Namely,

(a) Autocorrelation Function, ρ_k

(b) Partial Autocorrelation Function, $\rho''(k, k)$

The autocorrelation function has been defined and explained in Chapter 4. The partial autocorrelation function of any two observations, Q_s and Q_{s+k}, is the correlation between these two observations, taking the influence of the intervening observations, Q_{s+1} and Q_{s+k-1} into consideration. If the partial autocorrelation function between observations Q_s and Q_{s+2} ($= \rho''(2,2)$) is needed, not only is the relationship between Q_s and Q_{s+2} required but also the effect of Q_{s+1} on Q_{s+2}. Similarly, if the partial autocorrelation between Q_s and Q_{s+3} is needed, both effects of Q_{s+1} and Q_{s+2} on Q_{s+3} have to be considered. In contrast, the autocorrelation function does not consider the effect of the intervening observations.

The partial autocorrelation function $\rho''(k, k)$, is given by,

$$\rho''(k,k) = \frac{\rho_k - \sum_{i=1}^{k-1} \rho''(k-1,i)\rho_{k-i}}{1 - \sum_{i=1}^{k-1} \rho''(k-1,i)\rho_i} \qquad (5.8)$$

with k = 2, 3, ...
In Eq. 5.8,

$$\rho''(1,1) = \rho_1 \qquad (5.9)$$

$$\rho''(k,i) = \rho''(k-1,i) - \rho''(k,k)\rho''(k-1,k-i) \qquad (5.10)$$

where, $k = 3, 4, \ldots$ and $i = 1, 2, \ldots, k-1$.

The partial autocorrelation function defined above can also be derived directly from the Yule Walker equations (Box and Jenkins, 1976). For example, to obtain $\rho''(2,2)$, the following multiple regression equation need to be solved;

$$Q_{s+2}^* = \phi_1 Q_{s+1}^* + \phi_2 Q_s^* + a_{s+2} \qquad (5.11)$$

where $\rho''(2,2) = \phi_2$.

Similarly, for $\rho''(3,3)$ solve,

$$Q^*_{s+3} = \phi_1 Q^*_{s+2} + \phi_2 Q^*_{s+1} + \phi_3 Q^*_s + a_{s+3} \qquad (5.12)$$

where $\rho''(3,3) = \phi_3$ and $Q^*_{(.)}$ are the respective mean removed values of the $Q_{(.)}$ values.

All higher order partial autocorrelation coefficients can be determined likewise. These procedures are explained in greater detail in Jenkins and Watts (1968) and Box and Jenkins (1976).

The most appropriate model for a given data set can be obtained as follows. For an AR(p) model, the autocorrelation function tails off while the partial autocorrelation function is cut off after lag p. For a MA(q) model, the partial autocorrelation function tails off while the autocorrelation function cuts off after lag q. For an ARIMA(p,q) model, both functions tail off. In most applications of geotechnical engineering, the commonly encountered model is the ARIMA(p,q) model. The cut off levels of the correlation functions are based on the standard errors (σ_s) of the estimates as discussed below.

The standard error of the autocorrelation coefficient, ρ_k is given by (Box and Jenkins, 1976),

$$\sigma_s(\rho_k) = \left(\frac{1}{N} \left[1 + 2 \sum_{i=1}^{k-1} \rho_i^2 \right] \right)^{\frac{1}{2}} \qquad (5.13)$$

Similarly, the standard error of the partial autocorrelation coefficient is given by,

$$\sigma_s \left(\rho''(k, k) \right) = \frac{1}{\sqrt{N}} \qquad (5.14)$$

If the estimated autocorrelation coefficient is less than twice the standard error given by Eq. 5.13, it can be considered as negligible at the 95% significance level. The same criteria applies to the partial autocorrelation coefficient.

The adequacy of ARIMA models are governed by various conditions and the details are available in Box and Jenkins (1976).

5.5 Application of ARIMA Model Fitting

Methods of ARIMA model fitting have been performed on the DMT modulus profile given in Fig. 5.1. All the verifications and procedures already described, have been adopted in developing the most appropriate model to fit the data. The benefits of the use of this technique may not be apparent for CPT data since data logging is performed at very close intervals, but for tests such as the Dilatometer test where data spacing is 0.2 m, or other tests such as the SPT or the Field Vane where spacing is even farther apart, the advantage is that values in between tested points can be interpolated. A requirement of this technique is that the sampling points be equally spaced and most tests performed in situ satisfy this requirement.

5.5.1 Mean Prediction

Regression analysis was used to model the non - stationary part of the data, as it is apparent from Fig. 5.1 that the data exhibit a significant trend. First degree and second degree polynomials were rejected as the multiple correlation coefficient R^2 was very low with a value of 0.48 for the first degree and 0.54 for the second degree. The
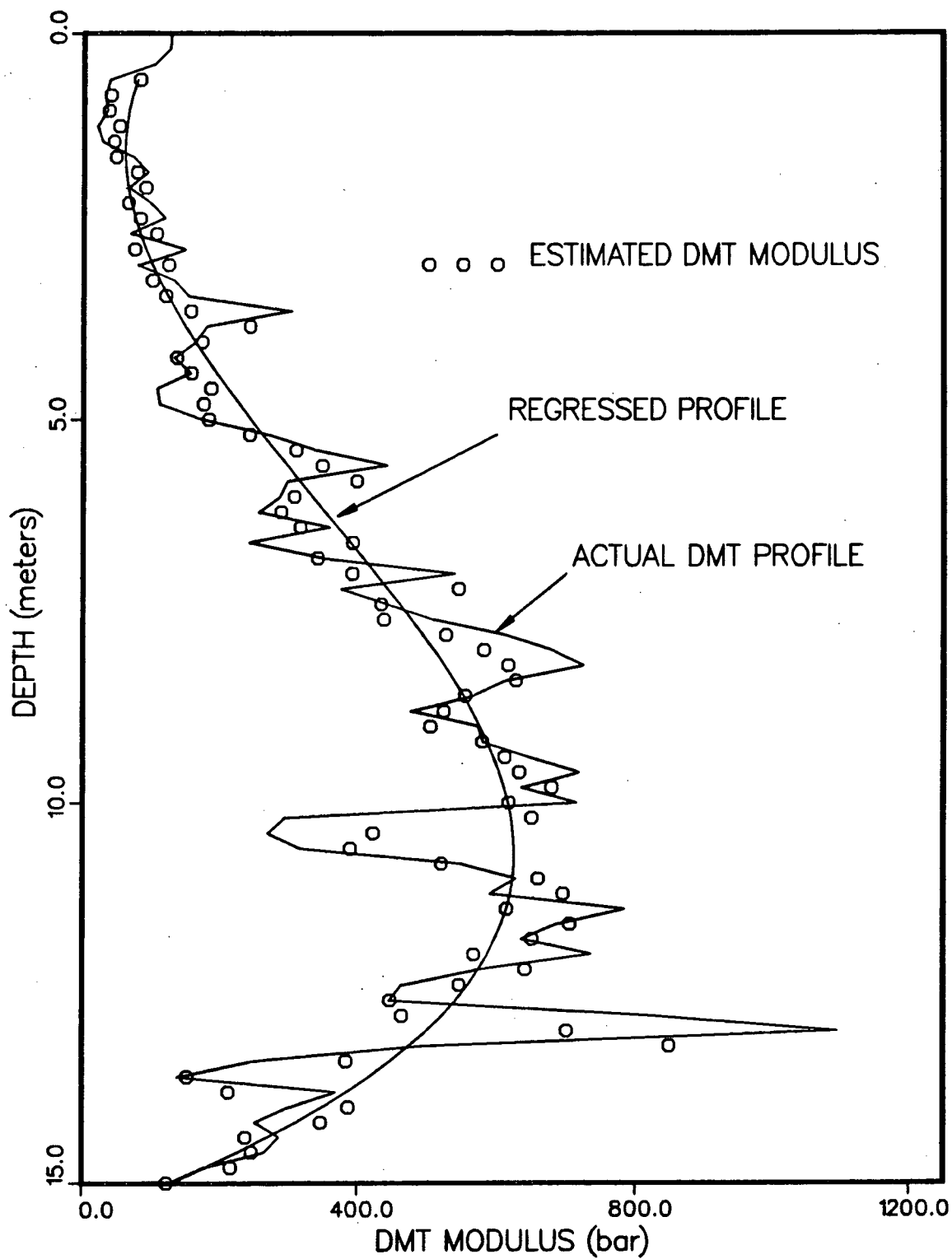
Figure 5.1: Comparison of the Dilatometer Modulus Profile of McDonald Farm with the Regressed Profile and the Estimated Profile.

Table 5.2: Statistics of the Parameters of the Trend.

Coefficient	Mean	Standard Deviation	t Statistic
β_o	110.82	58.15	1.91
β_1	- 71.21	33.80	-2.11
β_2	26.88	5.26	5.11
β_3	- 1.47	0.23	- 6.38

third degree polynomial resulted in a R^2 value of 0.70 and the F test indicated that a fourth degree polynomial would not result in a significant improvement. The model selected was,

$$\hat{Q}_i = \beta_o + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 s_i^3 + \epsilon_i \qquad (5.15)$$

where β_o, β_1, β_2 and β_3 are the regression coefficients, s_i the depth coordinate, ϵ_i the error term and $\hat{Q}_i$, the estimated soil property value at s_i.

The statistics of the parameters of the model are in Table 5.2. The correlation between the regression parameters were negligible and the t statistics of the coefficients were all close to 2.0 or greater, suggesting the adequacy of the model. Once the non-stationary part is determined from the regression equation (Eq. 5.15), the residuals ϵ_i can be obtained. If the residuals are not correlated, the regression estimate obtained from Eq. 5.15 is sufficient for the prediction.

If the residuals are correlated, two options are available to improve the estimates. One is to use generalized least squares to improve the regression coefficients, and the other is to consider the residuals separately and use time series methods to predict the properties at unknown locations. The latter method will be used here.

The correlations of residuals can be checked by using the Durbin - Watson statistic, (d_o), given by,

$$d_o = \frac{\sum_{i=2}^{N} (\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{N} \epsilon_i^2} \qquad (5.16)$$

where ϵ_i's are the residual terms obtained from Eq. 5.15.

In the example considered, $d_o = 1.5$, and from Durbin - Watson tables, the critical value $d_l = 1.51$, even at the 5% significant level, confirming the correlation of the residuals.

The residuals were also checked for the variance pattern and it was revealed that the variance was practically constant with depth, eliminating the need for the use of weighted least - squares approach to obtain regression estimates.

The ARIMA procedure of the SAS package (SASLETS, 1982) was used for the time series (spatial series) analysis. The autocorrelation function was of a gradually decaying type and the partial autocorrelation function cut off after the third lag. It was also found that all the partial autocorrelation coefficients, other than the first and third, were not significant. Considering all of the above, an AR(1,3) model with the following statistics was selected to model the stationary portion.

Table 5.3: Statistics of the Parameters of the Autoregressive Model.

Coefficient	Mean	Standard Deviation	t Statistic
ϕ_1	0.594	0.095	6.25
ϕ_2	-0.240	0.096	-2.51

The correlation (ρ_{13}) between ϕ_1 and ϕ_3 was -0.213, which was well within acceptable limits. The high values of the t ratios of the parameters also indicated the efficiency of the parameters. The constant value estimate was equal to -1.57. The Q statistic (Box and Jenkins, 1976) was calculated for twenty four lags and resulted in a value of 24.95, which was well below the critical value of 33.9 (Chi - Square tables) at the 5% significance level and at 22 (= K - p) degrees of freedom. This verified

that the proposed AR(1,2) model has absorbed all of the correlation remaining in the stationary residuals.

The total prediction from the stationary and non stationary part, the regression estimate of the non stationary part, and the actual DMT profile are illustrated in Fig. 5.1. The prediction is expressed as follows.

Non-stationary component from Polynomial Regression

$$\hat{Q}'_{s_o} = 110.82 - 71.21s_o + 26.88s_o^2 - 1.47s_o^3 \tag{5.17}$$

Stationary component from Time Series Analysis

$$\hat{Q}''_{s_o} = 0.60Q_{s_o-1} - 0.24Q_{s_o-3} - 1.57 \tag{5.18}$$

The total prediction is the sum of Eqs. 5.17 and 5.18, and is given by,

$$\hat{Q}_{s_o} = \hat{Q}'_{s_o} + \hat{Q}''_{s_o} \tag{5.19}$$

5.5.2 Variance Prediction

The variance of the estimates also comprises of two parts; one from the stationary and the other from the non-stationary part of the estimation.

The regression variance $\left(\text{Var}[\hat{Q}'_{s_o}]\right)$ at a point s_o is given by,

$$\text{Var}[\hat{Q}'_{s_o}] = \sigma_r{}^2[\mathbf{S_o}] \left[[\mathbf{C}][\mathbf{C}^{\mathbf{T}}]\right]^{-1} [\mathbf{S_o}]^{\mathbf{T}} \tag{5.20}$$

where $\sigma_r{}^2$ is the variance of the residuals and,

$$[\mathbf{S_o}] = [\,1 \quad s_o \quad s_o^2 \quad s_o^3\,] \tag{5.21}$$

$$[\mathbf{C}] = \begin{bmatrix} 1 & 1 & 1 & \dots & \dots & 1 \\ s_1 & s_2 & s_3 & \dots & \dots & s_N \\ s_1^2 & s_2^2 & s_3^2 & \dots & \dots & s_N^2 \\ s_1^3 & s_2^3 & s_3^3 & \dots & \dots & s_N^3 \end{bmatrix} \tag{5.22}$$

In Eqs. 5.21 and 5.22, s_i can be either the horizontal or the vertical co-ordinate.

The variance of the stationary component can be obtained as follows.

In the example considered, the stationary component is given by Eq. 5.18;

The variance is estimated from the Taylor Series approximation (Bury, 1978) as,

$$\mathrm{Var}[\hat{Q}''_{s_o}] = \sum_{i=1}^{m} \sum_{j=1}^{m} \theta_i \theta_j \mathrm{COV}(\phi_i, \phi_j) \tag{5.23}$$

for the AR(1,3) model under consideration,

$$\mathrm{Var}[\hat{Q}''_{s_o}] = \sum_{i=1,3} \sum_{j=1,3} \theta_i \theta_j \mathrm{COV}(\phi_i, \phi_j) \tag{5.24}$$

where,

$$\theta_1 = \frac{\partial \hat{Q}''_{s_o}}{\partial \phi_1} = Q_{s_o-1} \tag{5.25}$$

$$\theta_3 = \frac{\partial \hat{Q}''_{s_o}}{\partial \phi_3} = -Q_{s_o-3} \tag{5.26}$$

Expanding Eq. 5.24,

$$\mathrm{Var}[\hat{Q}''_{s_o}] = \theta_1^2[\mathrm{Var}(\phi_1)] + \theta_3^2[\mathrm{Var}(\phi_3)] + 2\theta_1\theta_3\rho_{13}[\mathrm{Var}(\phi_1)]^{\frac{1}{2}}[\mathrm{Var}(\phi_3)]^{\frac{1}{2}} \tag{5.27}$$

Substituting for the statistics and parameters of the above equation,

$$\mathrm{Var}[\hat{Q}''_{s_o}] = 0.009Q_{s_o-1}^2 + 0.009Q_{s_o-3}^2 + 0.003Q_{s_o-1}Q_{s_o-3} \tag{5.28}$$

Assuming the variances of the two components are independent, the combined variance of the prediction, $\text{Var}[\hat{Q}_{s_o}]$, is given by,

$$\text{Var}[\hat{Q}_{s_o}] = \text{Var}[\hat{Q'}_{s_o}] + \text{Var}[\hat{Q''}_{s_o}] \qquad (5.29)$$

The 95% confidence band of the estimate based on the above combined variance is illustrated in Fig. 5.2.

5.5.3 Engineering Significance

Once the mean and the variance of the non stationary and stationary components of the estimation have been determined using methods described in sections 5.5.1 and 5.5.2, the engineer is in a position to design at a desired confidence level. For example, the 95% lower bound based on the confidence band established is a value that the traditional geotechnical engineer will be comfortable with. The less conservative engineer who is willing to design with an element of higher risk can design based on 90% or even 80% lower bounds. The level of the lower bound to be decided depends on the type of structure to be designed, the level of uncertainty of the soil parameter under consideration (eg. shear strength, bearing capacity) and on the degree of uncertainty of the load. In offshore structures for example, the highest uncertainty is normally in the design load which is a function of wind speed, wave height etc.. The estimate of the latter quantities can rarely be predicted with a high degree of reliability.

5.6 Errors Encountered in Geotechnical Data

Laboratory and field tests used to measure properties of geotechnical materials are subject to various errors. The scatter in data obtained from various types of in situ
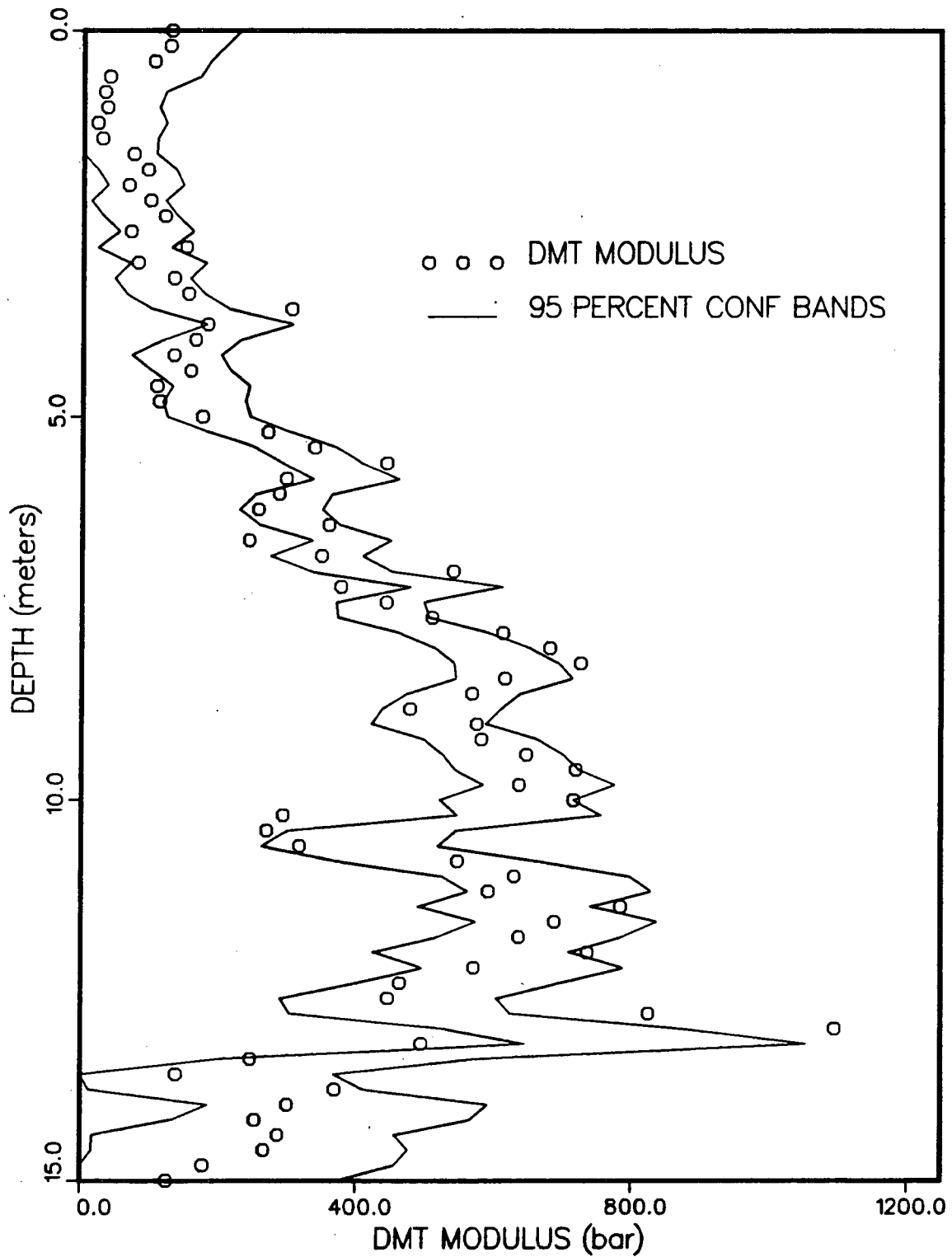
Figure 5.2: 95% Confidence Bands of the Estimated Dilatometer Modulus and the Actual Dilometer Modulus Obtained from Test.

testing methods are due to the inherent variability of the soil, soil disturbance during sampling and errors caused by man and machine. The latter two types of errors can not be identified individually and are therefore lumped together and denoted as the random measurement error or measurement noise (ϵ_r). If the true value of a property at a point 'i', is denoted by $\tilde{Q}_i$, the value measured by a particular test method Q_i, is given by,

$$Q_i = \tilde{Q}_i + \epsilon_r + \epsilon_b \tag{5.30}$$

where,

ϵ_r is the random testing error

ϵ_b is the test method bias which is an unknown constant

The scatter in the test data, Var $[Q_i]$, is given by,

$$\text{Var}[Q_i] = \text{Var}[\tilde{Q}] + \text{Var}[\epsilon_r] \tag{5.31}$$

where,

Var $[\tilde{Q}]$ is the inherent variability of the material

Var $[\epsilon_r]$ is the uncertainty of the testing error

The uncertainty of the random measurement error, ϵ_r, is given by,

$$\text{Var}[\epsilon_r] = \frac{1}{N} \sum_{i=1}^{N} (Q_i - E[Q_i])^2 \tag{5.32}$$

In the above equation, $E[Q_i]$ is the mean of the observed values Q_i, and typically it is non-stationary and is represented by a linear or curvilinear trend. It should be

emphasized that measurements Q_i are not the true values $\tilde{Q}_i$, and the immediate consequence of the above statement is $E[Q_i] \neq E[\tilde{Q}_i]$. Fig. 5.3 illustrates the different components of a data profile.

The bias in the test method, ϵ_b, is expressed as,

$$\epsilon_b = E[Q_i] - \tilde{Q}_i \tag{5.33}$$

The sample variance, $\text{Var}[Q_i]$, in Eq. 5.31 is readily measured from the data and is given by,

$$\text{Var}[Q_i] = \frac{1}{N} \sum_{i=1}^{N} \left(Q_i - \bar{Q} \right)^2 \tag{5.34}$$

The bias may be estimated by comparing the property measured by a given test method with that determined by using a more accurate test method or a reference method. However, any assumed standard does not measure the actual property exactly and, therefore, the bias in a test method can not be evaluated precisely.

The data scatter from in situ tests contains both Q_i and ϵ_r where the effects of sample disturbance and measurement errors are included in ϵ_r. If identical samples are available, ϵ_r may be determined by replicate testing. The inherent variability, which is the error free scatter, introduces uncertainty into the estimate of the average property over a region. Therefore, prior to any type of detailed analysis, it is necessary to isolate the random error from the observed data, if it is found to be significant.

The purpose of obtaining ϵ_r is two-fold:

(i) Permits the comparisons of the efficiency of different test methods: the lower the random error of a test equipment, the higher the efficiency.
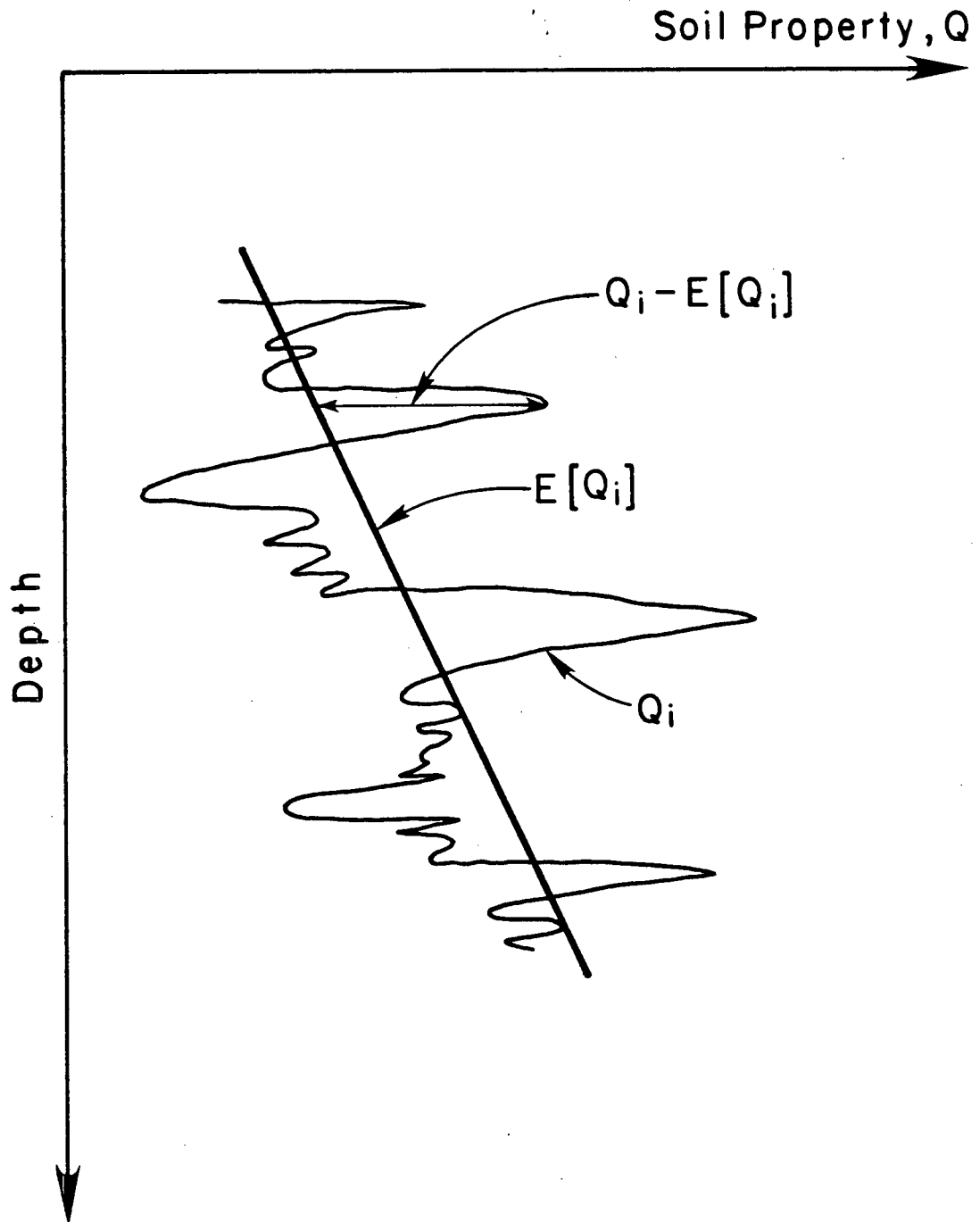
Figure 5.3: Illustration of the Expected Value and Residuals of a Profile Exhibiting a Trend.

(ii) Permits the evaluation of the effect of improvements to test procedures.

In the sections to follow, estimation of ϵ_r will be attempted using two methods: one method employing Box - Jenkins Time Series methods and the other based solely on the autocorrelation function of the data. Comparisons of these two methods will be made on sets of data obtained from different in situ test methods.

5.7 Determination of Random Noise

5.7.1 Random Noise from Time Series Methods

Once a data profile has been modeled using Time Series methods, as described in section 5.4, these procedures can be extended to obtain an estimate of the random error or measurement noise of the data. In this section, these methods has been applied to data from different in situ testing devices.

The bearing profile from the Cone Penetration Test (CPT), the dilatometer modulus values from the Dilatometer Test (DMT), the Dynamic N values from the Dynamic Cone Penetration Test (DCPT) and the undrained shear strength values from the Field Vane Test (FVT) are illustrated in Figs. 5.4, 5.6, 5.8 and 5.10, respectively. These data are all from Mc Donald Farm.

Only the detailed analysis of the CPT data will be described here although the results of all four test data will be discussed in this section. Considering all the factors that contribute to a good model, the ARIMA(1,1,1) was selected to model the linear segment (4.5 - 10.0 meters) of the bearing profile (Fig. 5.4) of the CPT. The autocorrelation function of the above data cut off after one lag, suggesting a MA(1) model, and the partial autocorrelation function also cut off after the first lag suggesting an AR(1) model. A single degree differencing was used for removing the approximately linear trend, and the following parameters resulted:
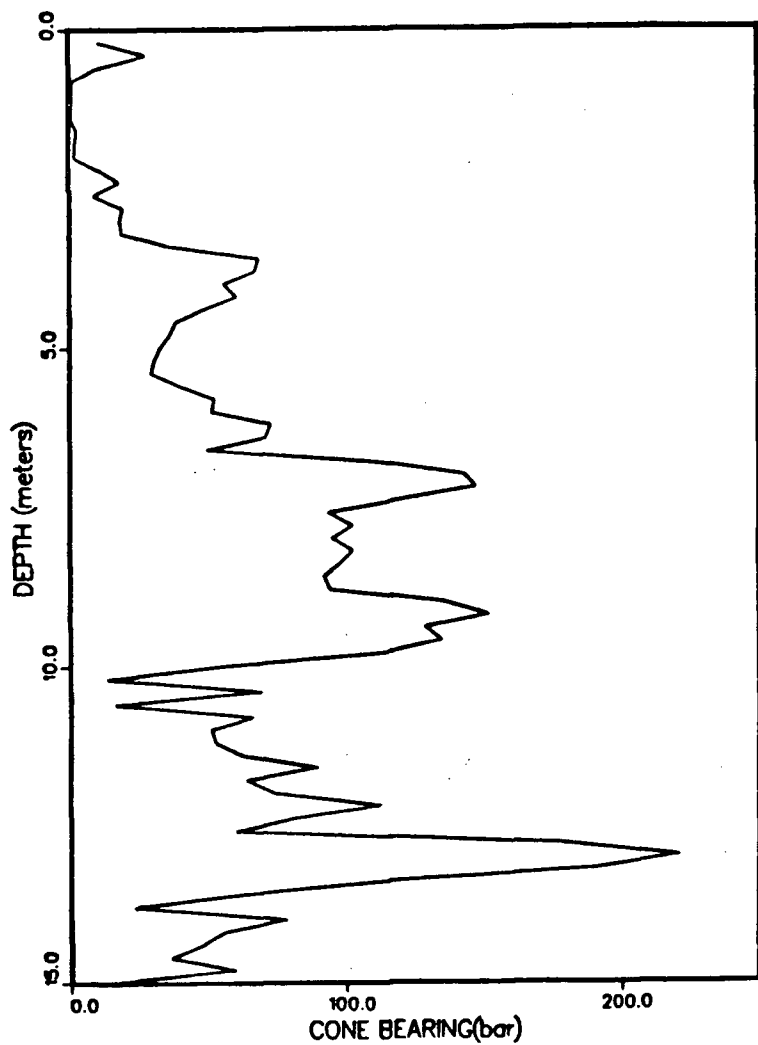
Figure 5.4: Cone Bearing Profile at McDonald Farm.
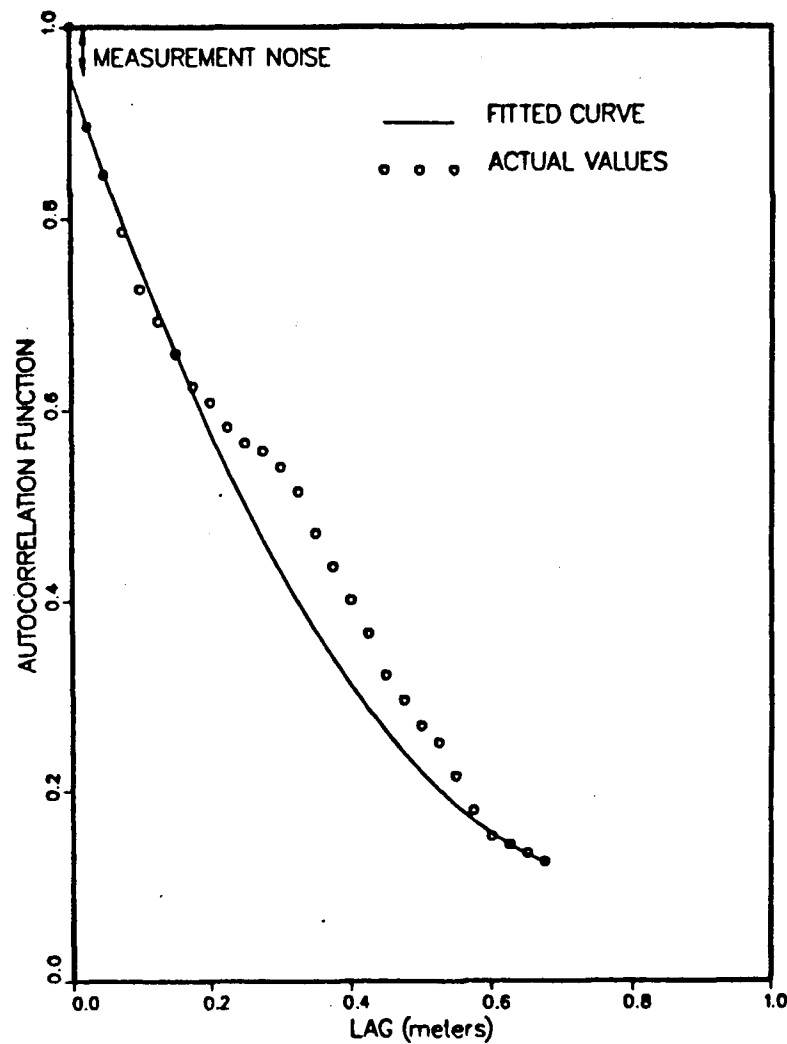
Figure 5.5: Variation of the Autocorrelation Function at McDonald Farm and the Fitted Function for the Determination of Measurement Noise for the Cone Penetrometer Test (CPT).
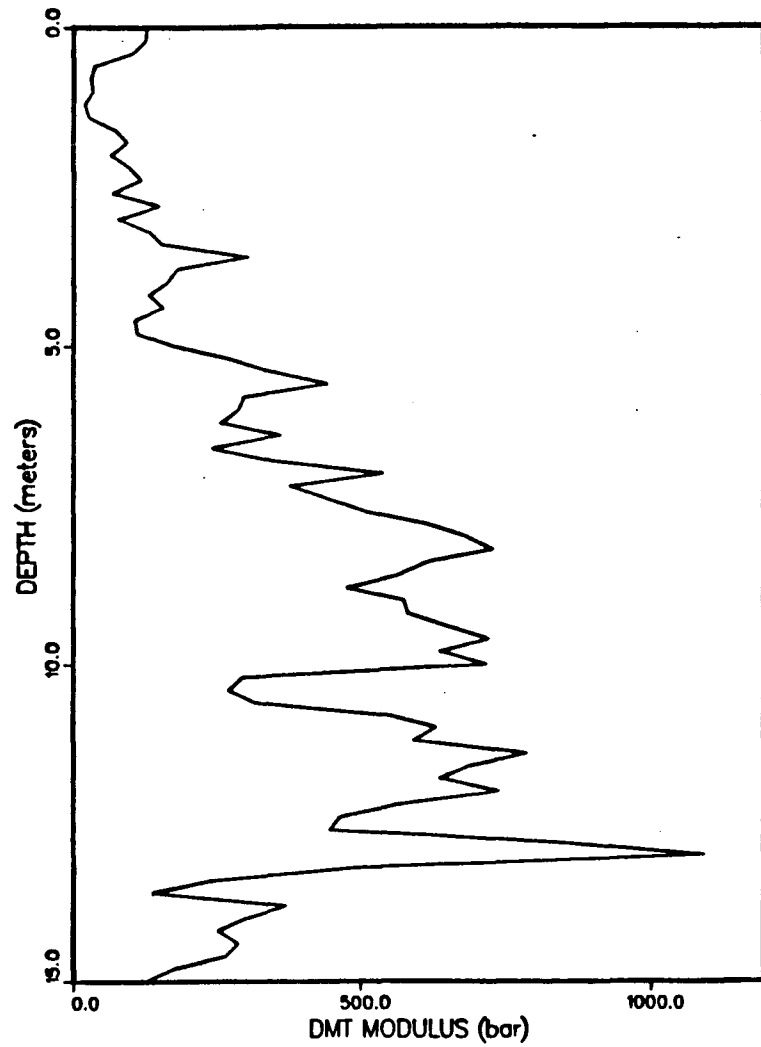
Figure 5.6: Dilatometer Modulus Profile at McDonald Farm.
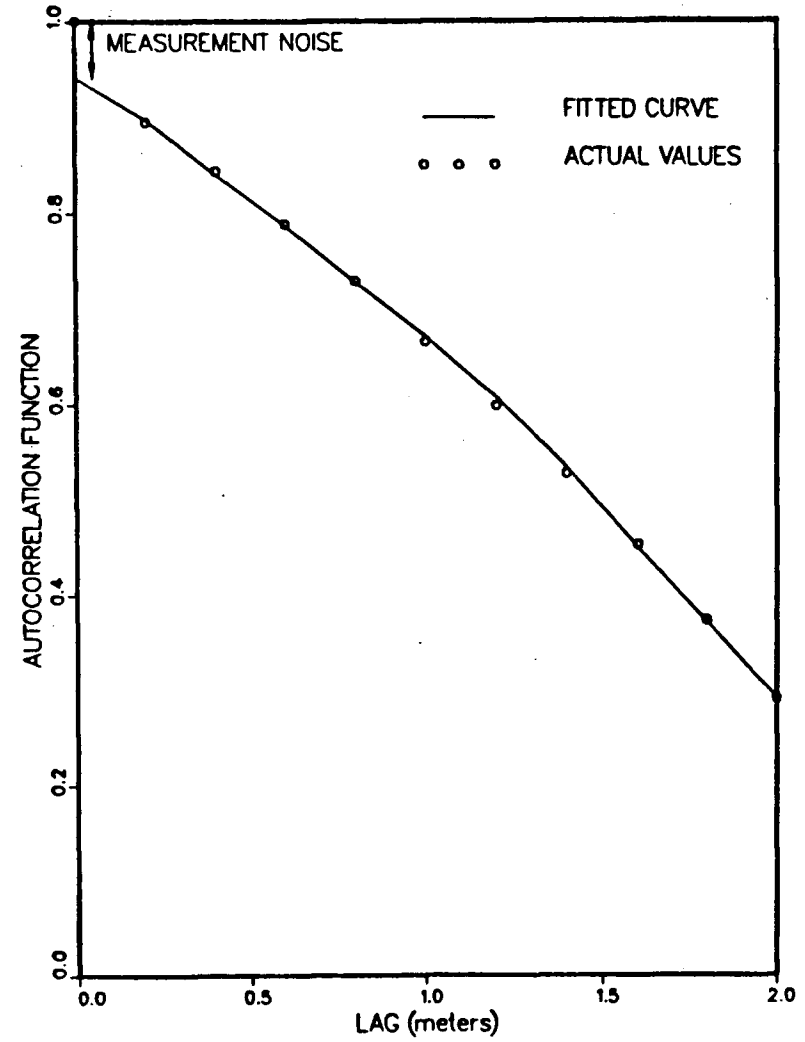
Figure 5.7: Variation of the Autocorrelation Function at McDonald Farm and the Fitted Function for the Determination of Measurement Noise for the Dilatometer Test (DMT).
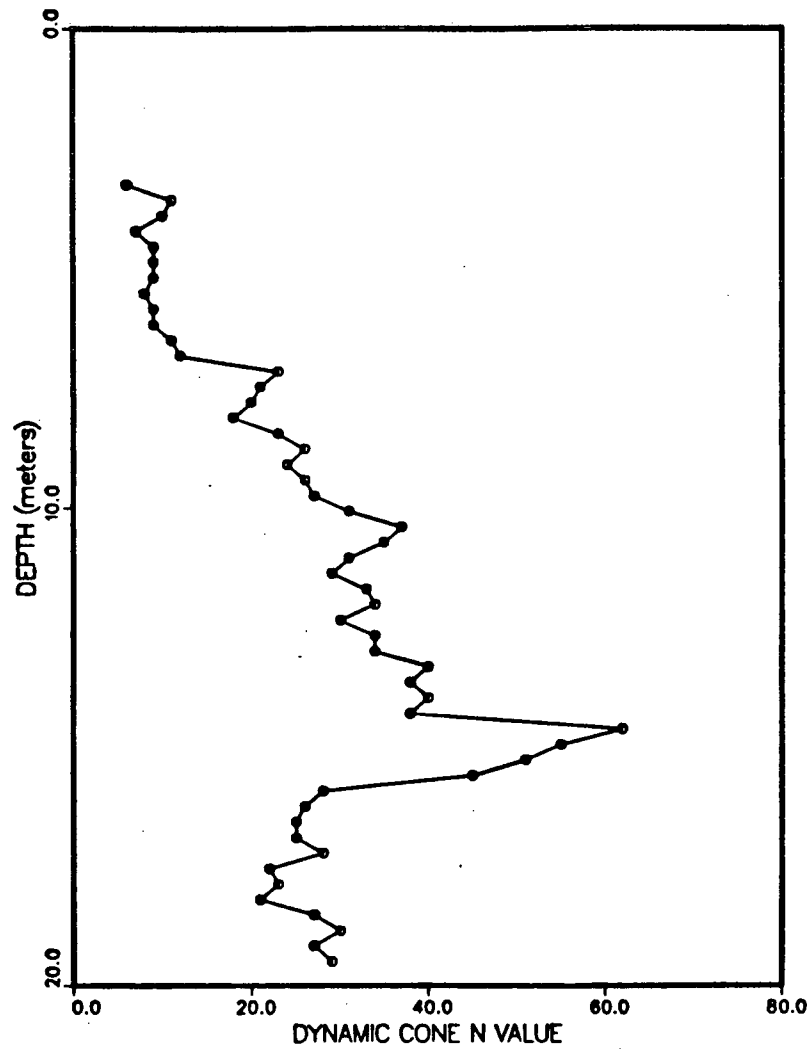
Figure 5.8: Dynamic Cone Penetrometer Test (DCPT) Profile at McDonald Farm.
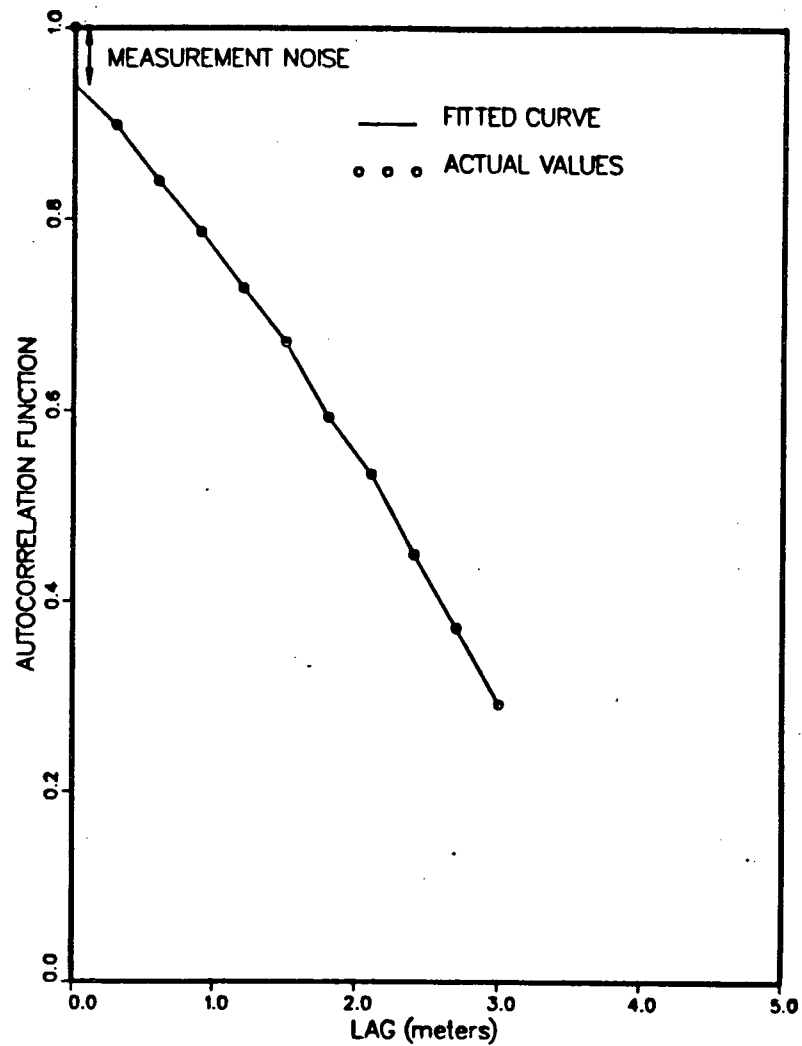
Figure 5.9: Variation of the Autocorrelation Function at McDonald Farm and the Fitted Function for the Determination of Measurement Noise for the Dynamic Cone Penetrometer Test (DCPT).
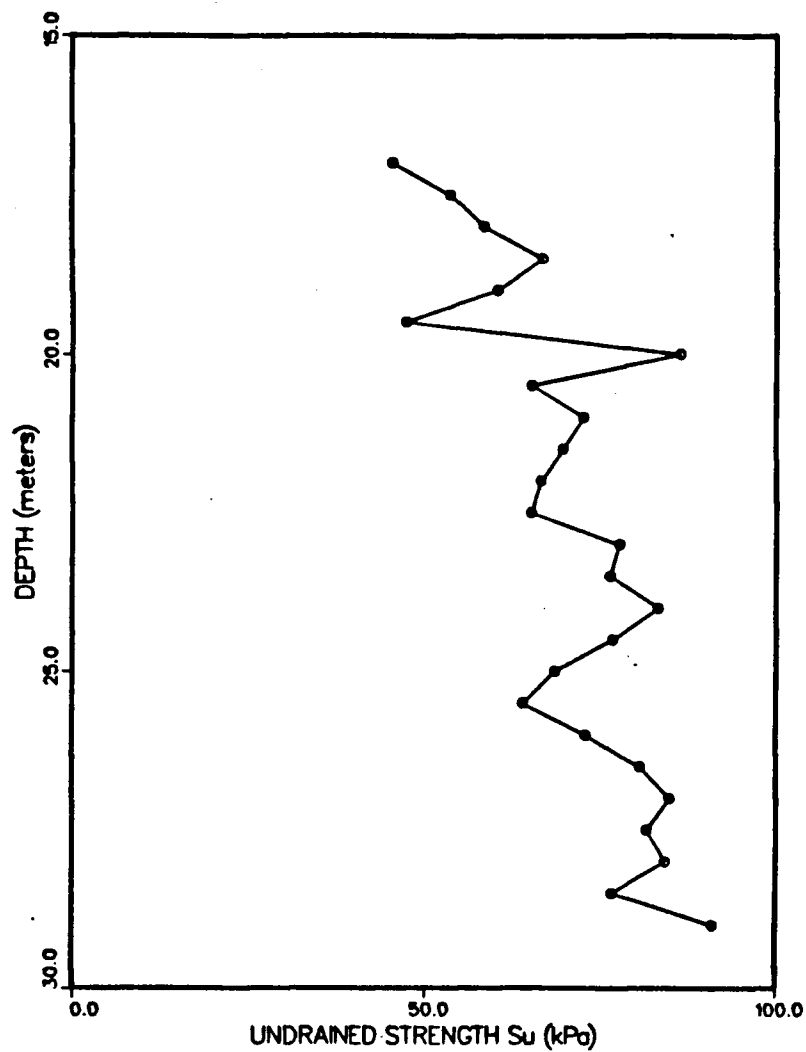
Figure 5.10: Undrained Shear Strength Profile at McDonald Farm Obtained from the Field Vane Test.
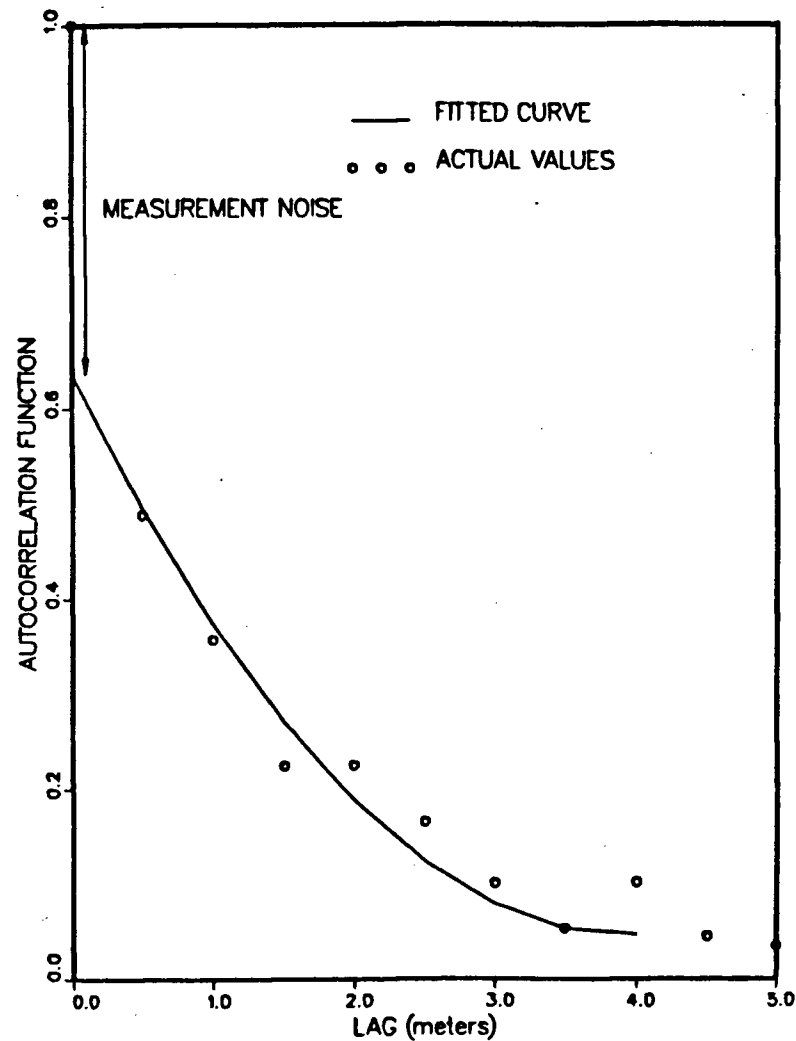
Figure 5.11: Variation of the Autocorrelation Function at McDonald Farm and the Fitted Function for the Determination of Measurement Noise for the Field Vane Test.

MA parameter $(\theta_1) = $ -0.348

AR parameter $(\phi_1) = $ -0.301

According to Box and Jenkins (1976) and Wu(1985), the variance of the data, $\sigma_z{}^2$, is given by,

$$\sigma_z{}^2 = \phi_1 C_1 + \sigma_a{}^2 - \theta_1 C_{za}(-1) + \sigma_r{}^2 \qquad (5.35)$$

where $\sigma_a{}^2$ is the white noise variance, C_1 the value of the autocovariance function at lag 1, $\sigma_r{}^2$ is the estimated variance of the random testing error.

$C_{za}(-1)$ in Eq. 5.35 is given by,

$$C_{za}(-1) = (\phi_1 - \theta_1)\sigma_a{}^2 = 0.047\sigma_a{}^2 \qquad (5.36)$$

Using the ARIMA procedure of the SAS package, $C_1 = 12.51$ and $\sigma_a{}^2 = 874.80$. Substituting the above values in Eq. 5.35,

$$\sigma_r{}^2 = 44.27$$

Therefore, the percentage of random error (ϵ_r^*) is given by (Wu, 1985),

$$\epsilon_r^* = \frac{\sigma_r{}^2}{\sigma_z{}^2} = 5.1\% \qquad (5.37)$$

The same procedure was applied for the other sets of data and the results are tabulated in Table 5.4, which clearly demonstrates the efficiency of CPT with its low percentage of random error. A 5% random error suggests that 95% of the CPT data scatter is due to the inherent variability of the soil tested. The random error obtained for the

DMT and the DCPT were not appreciably higher, while the Field Vane (FVT) result was significantly higher. A high proportion of random error indicates that the data variance is not representative of the inherent variability of the soil. The percentages of data scatter attributable to the inherent variability of the soil were approximately, 95%, 92% and 61%, for the DMT, DCPT and FVT, respectively. It should be noted that the inherent variabilities indicated by the different test methods are not the same because they do not measure the same parameters or give the same properties. However, it is this inherent variability that is important but often overlooked in design considerations.

Table 5.4: Comparisons of the Random Noise Estimates for Different Test Methods

Test Method	CPT	DMT	FVT	DCPT
Model	1,1,1	1,1,1	0,0,2	0,1,2
Parameter 1	- 0.348	0.889	- 0.338	0.302
Parameter 2	- 0.301	0.587	0.355	0.255
Random Error Variance $(\sigma_r{}^2)$	44.62	322.36	50.65	0.6724
Data Variance $(\sigma_z{}^2)$	874.80	5861.14	138.78	8.20
Random Error (ϵ_r^*) %	**5.1**	**5.5**	**36.5**	**8.2**

5.7.2 Random Noise from Autocorrelation Analysis

The random measurement error can also be determined using the autocorrelation function. This method is only dependent on the autocorrelation function and unlike the previous method, it alleviates the need for modeling and performing the other calculations already described.

If the true value of a property is denoted by $\tilde{Q}_i$, the value measured by a particular test method, Q_i, is given by,

$$Q_i = \tilde{Q}_i + \epsilon_r \tag{5.38}$$

where ϵ_r is the error term.

If the autocovariances of Eq. 5.38 are taken,

$$C(Q_i) = C(\tilde{Q}_i) + C(\epsilon) \tag{5.39}$$

The autocovariance of Q_i, $C(Q_i)$, is the variance of the data at zero lag distance (separation distance) and approaches zero as this distance increases. If the lag distance is denoted by s, the autocovariance of the random error term, ϵ_r , which is given by $C(\epsilon_r)$, has a non zero value at s = 0 and zero at all other values of s, since the random error is uncorrelated from point to point. From Eq. 5.39, it can be ascertained that at s = 0, the autocovariance function comprises of two parts; namely, the inherent variability of the soil and the random error term. The autocovariance function is, therefore, a spike at s = 0 and a slowly decaying function for s greater than zero, exhibiting less dependance with increasing separation distance or lag. Considering the above, if the autocovariance function is extrapolated to meet the axis representing the autocovariance function at C_c, the difference between the autocovariance values at s = 0 (C_o) and C_c will be the random noise term (Baecher, 1982). The same procedure is applicable to the autocorrelation function (ρ), which is the standardized form of the autocovariance function (C) and the relationship is given below.

$$\rho_i = C_i / C_o \tag{5.40}$$

For example, the autocorrelation function (Fig. 5.5) of the cone bearing data given in Fig. 5.4, meets the ordinate at 0.95 which is equal to the proportion of the inherent

variability of the of the total data scatter. The random error or measurement noise of the CPT data is therefore, 5%. The above procedure has also been applied to DMT (Fig. 5.6), Field Vane Test (FVT), (Fig. 5.8) and DCPT (Fig. 5.10) data. The actual values of the autocorrelation functions together with the fitted extrapolated curves are illustrated in Fig. 5.5 for the CPT, Fig. 5.7 for the DMT, Fig. 5.9 for FVT and Fig. 5.11 for the DCPT. The results of the above are given in Table 5.5.

5.7.3   Comparison of the Two Methods

The values of the random error obtained using the two methods described in sections 5.6.1 and 5.6.2 are listed in Table 5.5 for the different test types.

Table 5.5: Comparisons of the Random Error Estimates for Different Analysis Methods.

	Random Error %	
Test Method	Autocorrelation Analysis	Time Series Method
CPT	5.0	5.1
DMT	5.8	5.5
Vane	36.0	38.7
DCPT	6.1	8.2

The results in Table 5.5 show a remarkable agreement between the two methods of random error determination, with the CPT giving the lowest component of random error and the FVT resulting in the highest. As mentioned previously, any of these two methods may be used to compare the accuracy of different test types. A measurement noise of 36% is extremely high and is an indication of severe sample disturbance during testing. The above theory has also been applied to a different set of data obtained at a site near the Fraser River Delta. The random error component for the CPT data was an appreciably low 4.8% while the vane test gave a high value of 30%, confirming the earlier findings. The value obtained for the SPT was 14%, which is not as significant

as for the vane, but significantly higher than the measurement noise level of the CPT. The very low random error values obtained for the CPT provide evidence as to the efficiency of this test method.

Due to the possible inaccuracies in modeling, both methods of determining the random error should be considered. Although the autocorrelation method does not incorporate modeling in a strict sense, it makes use of a fitting procedure to obtain the best function for the data. This can also lead to inaccuracies and therefore, it is recommended that the random error from both methods be determined prior to making any decisions.

5.8   Conclusions

The main conclusions which can be drawn from this chapter are;

(i) Time Series methods have been effectively used to model the stationary part of soil data profiles. The benefits of Time Series modeling is more apparent in testing methods where sampling is not performed at close spacings and where interpolation between tested points will be useful. This method can also be used to establish confidence bands which will be useful in engineering practice.

(ii) Time Series methods also provide a convenient way of determining the random error component of an in situ testing technique. For the different tests considered, the results indicated a significantly low random error content (5%) in CPT data, reflecting its superiority over data obtained from other testing methods, such as the Field Vane test and the Dynamic Cone Penetration test which comprise higher proportions of random error.

(iii) The random error derived directly from the autocorrelation function at zero separation distance for the different testing techniques compared appreciably well with that obtained from the Time Series method. It is recommended that both methods

be used if a reliable estimate of the random error is needed. A close agreement of the error values from the two methods will result in an increased confidence of the estimate.

Chapter 6

Interpolation Considering Correlations

6.1 Introduction

In traditional geotechnical engineering, various methods are used to interpolate between known field data values. In most site investigations, economics do not allow the acquisition of a large data base although the engineer would prefer a sizeable base for design. It is this limited data base which causes the variation of soil properties to be considered as random. In reality, there is nothing random about the variation of soil properties, since if every point in the ground can be tested and investigated, it turns out to be a deterministic problem. However, it is not practical to do so, thereby, giving rise to the uncertainty of soil properties at untested locations.

The traditional approach in geotechnical engineering to deal with the limited data base is to interpolate between known points using some simple functions. This approach neglects any correlation between data points. In most cases, the uncertainty is accounted for by adding a factor of safety, sometimes referred to as the factor of ignorance. In a typical site investigation, a borehole or two will be drilled, and in some cases, supplemented by a few cone penetrometer tests. The designer may select a very conservative strength as that representative of the entire site and design for the largest load that is expected to be carried by the foundation, together with an appropriate factor of safety. It is obvious that this is a highly conservative approach but is done because of the highly uncertain nature of soil property variations in three dimensions combined with other adverse factors such as limited data availability and human and

instrument errors encountered during testing. Many sophisticated models to explain the behavior of soils have been developed in recent years but many geotechnical engineers continue to use traditional approaches and are reluctant to sacrifice some of the conservatism and to consider statistical and probabilistic approaches. These latter methods consider the correlations between soil parameters as a vital ingredient.

Soil properties are highly depth dependent. Dependence between points in the horizontal direction too may be present and will have to be considered in any multi-dimensional interpolation procedure. Simple regression techniques assume independence of soil properties and, therefore, the estimates will be biased in the presence of correlation. Regression methods also consider the soil properties at known locations as observed values of a random variable, the distribution of which depends on the co-ordinates of the locations which are not random (Kraus and Mikhail, 1972; Davis, 1978).

As mentioned above, it is common in geotechnical engineering to use simple regression or simple weighting functions in problems of interpolation, disregarding the correlation completely and assuming that soil properties between points are independent. In any three dimensional analysis or a two dimensional analysis which has the depth as one of its co-ordinate axes, it is necessary to consider correlations, if reasonable estimates of soil parameters are needed. Some of the more common methods of interpolations used in geotechnical engineering practice are given in Appendix C.

All the interpolation methods considering correlations do so using the autocorrelation or the semi-variogram function of the data which essentially have to be stationary. Soil data, especially in the depth dimension, are non - stationary and will have to be made stationary using methods of regression already explained in Chapter 3. In the event the ensuing stationary residuals are not correlated, the least squares estimates will be deemed to be satisfactory. If they are correlated, the method to be described

in this chapter will have to be used. The proposed method is a modified form of the procedure referred to as ' Kriging ' in the mineral industry (Matheron, 1963). Although the above technique will be emphasized in this thesis, there also exists a mathematically more rigorous procedure for dealing with non stationary data, known as ' Universal Kriging ' (Matheron - 1967, '70). There also exists a more simplified version of ' Universal Kriging ', the credit for which is due to Gambolati and Volpi (1979). A simpler procedure where the trend and residuals are considered separately will be used in this chapter which will also describe a novel approach to handle two dimensional autocorrelation functions.

6.2 Autocorrelation and Semi-Variogram Functions

The autocorrelation function and the semi-variogram function are two versatile and essential tools which enable the investigation of spatial variation of soil property values. The basic purpose of these functions is to establish the influence of values at any point over values at neighboring points. Soil properties of points at closer distances apart are expected to show a higher correlation than for points which are widely spaced. The autocorrelation function gives this correlation for different values of the distance of separation, or the lag distance.

The one dimensional autocorrelation function ($\rho(l)$) at a lag distance l is defined as,

$$\rho(l) = \frac{N}{N-h} \frac{\sum_{i=1}^{N-h}(Q_i - \bar{Q})(Q_{i+h} - \bar{Q})}{\sum_{i=1}^{N}(Q_i - \bar{Q})^2} \tag{6.1}$$

where, N is the total number of data and l = h.d, d being the sample spacing, and $\bar{Q}$ the mean of the data. The function given in Eq. 6.1 is applicable for anisotropic data in depth or in plan. For isotropic data, l can represent any dimension.

The semi-variogram function at a lag distance l, (γ(h)), is defined as,

$$\gamma(l) = \frac{1}{2(N-h)} \sum_{i=1}^{N-h} (Q_i - Q_{i+h})^2 \tag{6.2}$$

where N, h and $\bar{Q}$ have the same meaning as in Eq. 6.1 and Q_i and Q_{i+h} are the soil property values at locations i and i + h, respectively. The variogram function is equal to twice the value of the semi-variogram function given by Eq. 6.2.

As can be seen from Eq. 6.1, the autocorrelation function is a standardized form of the covariance function, C(h), where,

$$C(l) = \frac{1}{N-h} \sum_{i=1}^{N-h} (Q_i - \bar{Q})(Q_{i+h} - \bar{Q}) \tag{6.3}$$

The denominator of Eq. 6.1 is the variance of the data, σ^2, with,

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left(Q_i - \bar{Q}\right)^2 \tag{6.4}$$

The variance, σ^2, is equal to the covariance at lag zero, C(0). Therefore, ρ(h), can also be expressed as,

$$\rho(l) = \frac{C(l)}{C(0)} \tag{6.5}$$

For the hypothesis of second order stationarity where the mean and the variance of the data are constant, with an autocorrelation function which is only dependent on the lag distance and independent of actual location, it can easily be shown that,

$$\gamma(l) = C(0) - C(l) \tag{6.6}$$

Dividing Eq. 6.6 by C(0) and rearranging terms result in the following relationship between the autocorrelation function, $\rho(l)$ and the semi - variogram function, $\gamma(l)$.

$$\rho(l) = 1 - \frac{\gamma(l)}{C(0)} \tag{6.7}$$

The above relationship shows how these two functions are closely related, enabling a choice for correlation analysis.

6.2.1   Models for the Autocorrelation Function

For analytical purposes, it is necessary for the actual autocorrelation function derived for the data to be fitted with a closed form algebraic, exponential or sinusoidal function. It is common in geotechnical engineering to expect negative values for the autocorrelation function which could therefore be better represented by a sinusoidal function. Some of the more common autocorrelation functions used in the description of geologic data are expressed below. Functions for one dimensional data (Vanmarke, 1978) have already been given in Chapter 4, and the ones to follow are an extension of the 1 - D functions to two dimensions, horizontal and vertical.

$$\rho(\Delta x, \Delta z) = \text{EXP}\left[-(a_1 \mid \Delta x \mid + a_2 \mid \Delta z \mid)\right] \tag{6.8}$$

$$\rho(\Delta x, \Delta z) = \text{EXP}\left[-(b_1 \Delta x^2 + b_2 \Delta z^2)\right] \tag{6.9}$$

$$\rho(\Delta x, \Delta z) = \text{EXP}\left[-(c_1 \mid \Delta x \mid + c_2 \mid \Delta z \mid)\right] \text{COS}\left[\frac{\Delta x}{c_1} + \frac{\Delta z}{c_2}\right] \tag{6.10}$$

$$\rho(\Delta x, \Delta z) = \text{EXP}\left[-\left(\frac{\mid \Delta x \mid}{k_1} + \frac{\mid \Delta z \mid}{k_2}\right)\right]\left(1 + \frac{\mid \Delta x \mid}{k_1}\right) + \left(\frac{\mid \Delta z \mid}{k_2}\right) \tag{6.11}$$

$$\rho(\Delta x, \Delta z) = \text{SIN}\left(\frac{\Delta x}{m_1} + \frac{\Delta z}{m_2}\right)\left(\frac{\Delta x}{m_1} + \frac{\Delta z}{m_2}\right) \tag{6.12}$$

In Eqs. 6.8 to 6.12, Δx and Δz are the distances in the horizontal and vertical directions respectively. The above expressions for possible autocorrelation functions

can also be extended to the three dimensional case by incorporating Δy, to represent the second horizontal axis. For the isotropic case, Δz in the above equations will vanish, and the expressions will only consist of a single dimension term Δx. In geotechnical engineering, however, isotropy in all three dimensions is far from reality, with soil properties being strongly depth dependent, not to mention lateral variations.

In a typical geotechnical exploration program, it is very rare to obtain many data points and as a result, the estimation of a satisfactory autocorrelation function becomes difficult. Furthermore, at larger lag distances, the number of of points available for the calculation of the function is less than for shorter lag distances. Due to this reason, Agterberg (1974) states that the autocorrelation function will be accurate and least biased only up to about one fourth of the maximum separation between data points. Therefore, in order to obtain better estimates for the autocorrelation function, it has been recommended (Baecher, 1980) that a filtering process be performed by assigning a higher weight for points which are spaced closer. This procedure is essentially an application of a modified Bartlet filter (Jenkins and Watts, 1978) to the actual function values, $\hat{\rho}(i)$, obtained and can be expressed as,

$$\hat{\rho(i)} = \frac{\rho_{i-1}N_{i-1} + \rho_i N_i + \rho_{i+1}N_{i+1}}{N_{i-1} + 2N_i + N_{i+1}} \tag{6.13}$$

where, N_i is the number of data points used for making the estimate at i.

6.2.2 Models for the Semi-Variogram Function

The more common models of the semi-variogram can be categorized into two main divisions; namely, models with a sill and models with no sill (Journel and Huijbregts, 1978). The sill is the constant value attained by the variogram at some separation distance or lag. A typical variogram function is illustrated in Fig. 3.12 in Chapter 3 where the trend removed data has a sill while the data with the trend does not

possess a sill. The models to be listed below are all applicable to one dimension with the possibility of extension to two or three dimensions, similar to the autocorrelation function. Let l represent any one of the dimensions, Δx, Δz or Δy and $\gamma^*(l)$ be the normalized form of the semi - variogram, $\gamma(l)$, given by,

$$\gamma^*(l) = \frac{\gamma(l)}{C(0)} \tag{6.14}$$

Models with a Sill

$$\gamma^*(l) = \frac{3l}{2a} - \frac{l^3}{2a^3} \tag{6.15}$$

$$\gamma^*(l) = 1 - \text{EXP}\left(-b\,|\,l\,|\right) \tag{6.16}$$

$$\gamma^*(l) = 1 - \text{EXP}\left(-cl^2\right) \tag{6.17}$$

Models with no Sill

$$\gamma^*(l) = a'l \tag{6.18}$$

$$\gamma^*(l) = b'log(l) \tag{6.19}$$

$$\gamma^*(l) = 1 - c'\frac{\text{SIN}(l)}{l} \tag{6.20}$$

In the above equations, a, b, c, a', b' and c' are constants.

As mentioned before, in most applications of geotechnical engineering, the semi-variogram functions in two perpendicular directions will not be similar, and in such cases, they should be transformed to an equivalent function by methods given in David (1977) and Journel and Huijbregts (1978). It is also important to remove any trend from the data, if it exists, to avoid serious errors in interpolation problems. The effects of trend on the semi-variogram is described in detail by Starks and Fang (1982). In exploration programs where the number of testing locations are limited due economic reasons, the best possible locations in order to obtain the optimal variogram for a site can be determined using methods of linear programming (Warrick and Myers, 1987). Russo (1984) also describes the design of an optimal sampling network for estimating

the variogram. The other advantage of such optimization methods is that they allow the determination of the best possible location for an additional testing location, based on the data already available. However, in all these methods of optimization the number of data points have to be very large, to extents rarely available in normal geotechnical projects. If the number of data points are significant, these optimization methods should give highly beneficial results. Cressie (1985), Christakos (1985) and Sabourin (1976) all give valuable information regarding the estimation of variograms.

6.3 Interpolation Based on the Autocorrelation Function

The procedure of interpolation to be described is valid only for a stationary process. Therefore, in the presence of significant trends (non-stationarity), it will have to be first removed prior to the application on the stationary residuals. During the time of the author's research on this subject at UBC, Kulatilake (1987) has also used a procedure of interpolation considering the trend and residuals separately. However, the autocorrelation function of the residuals are handled in a different way in this thesis.

The basic interpolation relationship is given by (David, 1976),

$$\hat{Q}(s_o) = \lambda_1 Q(s_1) + \lambda_2 Q(s_2) + \lambda_3 Q(s_3) + \ldots\ldots\ldots + \lambda_n Q(s_n) \qquad (6.21)$$

where, $Q(s_1), Q(s_2), \ldots\ldots, Q(s_n)$ are the known soil property values at locations, s_1, $s_2, \ldots\ldots, s_n$. In Eq. 6.21, s_o is the point where the interpolation is required and is a point in space with both a horizontal and a vertical (depth) co-ordinate.

The weights λ_i for $i = 1, 2, 3, \ldots\ldots n$, are obtained from Eq. 6.22 below.

$$\{L\} = [P]^{-1}\{M\} \qquad (6.22)$$

where,

$$[\mathbf{P}] = \begin{bmatrix} 1 & \rho_{Q(s_1)Q(s_2)} & \rho_{Q(s_1)Q(s_3)} & \cdots & \cdots & \rho_{Q(s_1)Q(s_n)} & 1 \\ \rho_{Q(s_2)Q(s_1)} & 1 & \rho_{Q(s_2)Q(s_3)} & \cdots & \cdots & \rho_{Q(s_2)Q(s_n)} & 1 \\ \rho_{Q(s_3)Q(s_1)} & \rho_{Q(s_3)Q(s_2)} & 1 & \cdots & \cdots & \rho_{Q(s_3)Q(s_n)} & 1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \rho_{Q(s_n)Q(s_1)} & \rho_{Q(s_n)Q(s_2)} & \rho_{Q(s_n)Q(s_3)} & \cdots & \cdots & 1 & 1 \\ 1 & 1 & 1 & \cdots & \cdots & 1 & 0 \end{bmatrix} \quad (6.23)$$

$$\{\mathbf{M}\} = \begin{Bmatrix} \rho_{Q(s_o)Q(s_1)} \\ \rho_{Q(s_o)Q(s_2)} \\ \rho_{Q(s_o)Q(s_3)} \\ \vdots \\ \vdots \\ \rho_{Q(s_o)Q(s_n)} \\ 1 \end{Bmatrix} \quad (6.24)$$

$$\{\mathbf{L}\} = \begin{Bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \vdots \\ \lambda_n \\ \mu/\sigma^2 \end{Bmatrix} \quad (6.25)$$

where, s comprises of both a horizontal and vertical co-ordinate. $\{\mathbf{M}\}$ and $[\mathbf{P}]$ in above equations are for the case when the autocorrelation function is used. If the semi-variogram is used, the ρ terms in Eqs. 6.23 and 6.24 will be replaced by γ

terms, thereby causing a change in matrices $[\mathbf{P}]$ and $\{\mathbf{M}\}$. However, the weights (λ_i) obtained by both methods will be identical due to the direct relationship between the autocorrelation and semi-variogram functions, expressed by Eq. 6.7. σ^2 and μ in Eq. 6.25 are the variance of the data and a Lagrange constant, respectively. A detailed derivation of Eq. 6.21 is given in Appendix D. This procedure is an exact interpolation method because if the property value of a known data point which was used for the analysis is determined using the above equations, it will give an identical value. In contrast, regression is not an exact interpolation method.

In any estimation procedure, the variance of the estimator is a very important quantity in order to evaluate the efficiency of the procedure and to establish confidence bands on the estimations.

The estimation variance (σ_e^2) is given by (Appendix D),

$$\sigma_e{}^2 = \sigma^2 \left(1 - \sum_{i=1}^{n} \lambda_i \rho_{Q(s_i)Q(s_o)}\right) - \mu \tag{6.26}$$

If the semi-variogram was used instead of the autocorrelation function, the estimation variance can be expressed as,

$$\sigma_e{}^2 = \sum_{i=1}^{n} \lambda_i \gamma_{Q(s_i)Q(s_o)} - \mu \tag{6.27}$$

6.4 Development of a Two Dimensional Autocorrelation Function

The method of interpolation to be proposed requires the development of a two dimensional autocorrelation function, so that interpolation can be performed in two dimensions (vertical and horizontal). In contrast to one dimensional autocorrelation functions already described, two different types of correlation functions have to be

defined for each lag distance. For example, consider the data array M1 which is comprised of n cone holes at m different depths. At every depth, there exists n data points across the site, while any cone hole has data points at m different depths.

Data Array M1

$$
\begin{array}{ccccccc}
Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \cdots & \cdots & \cdots & Q_{0,n-1} \\
Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & \cdots & \cdots & \cdots & Q_{1,n-1} \\
Q_{2,0} & Q_{2,1} & Q_{2,2} & Q_{2,3} & \cdots & \cdots & \cdots & Q_{2,n-1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
Q_{m-1,0} & Q_{m-1,1} & Q_{m-1,2} & Q_{m-1,3} & \cdots & \cdots & \cdots & Q_{m-1,n-1}
\end{array}
$$

The two types of autocorrelation functions which can be defined for w (horizontal lag) and r (vertical lag) are as follows;

$$\rho(r,w) = \frac{1}{(m-r)(n-w)\sigma^2} \sum_{i=0}^{m-r-1} \sum_{j=0}^{n-w-1} \left(Q_{i,j} - \bar{Q}\right)\left(Q_{i+r,j+w} - \bar{Q}\right) \quad (6.28)$$

and,

$$\rho(-r,w) = \frac{1}{(m-r)(n-w)\sigma^2} \sum_{i=r}^{m-1} \sum_{j=0}^{n-w-1} \left(Q_{i,j} - Q\right)\left(Q_{i-r,j+w} - Q\right) \quad (6.29)$$

where, Q is the soil property, $\bar{Q}$ and σ^2 are the mean and variance of the data respectively.

It should be noted that,

$$\rho(r,w) = \rho(-r,-w) \quad (6.30)$$

$$\rho(-r,w) = \rho(r,-w) \quad (6.31)$$

Now, consider data array M2 which is given below where m and n have the same meaning as before.

Data Array M2

$$
\begin{array}{cccccccc}
Q_{1,1} & Q_{1,2} & Q_{1,3} & Q_{1,4} & \cdots & \cdots & \cdots & Q_{1,n} \\
Q_{2,1} & Q_{2,2} & Q_{2,3} & Q_{2,4} & \cdots & \cdots & \cdots & Q_{2,n} \\
Q_{3,1} & Q_{3,2} & Q_{3,3} & Q_{3,4} & \cdots & \cdots & \cdots & Q_{3,n} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
Q_{m,1} & Q_{m,2} & Q_{m,3} & Q_{m,4} & \cdots & \cdots & \cdots & Q_{m,n}
\end{array}
$$

Let the lag r in the vertical direction be positive from top to bottom, and the lag w in the horizontal direction be positive from left to right. Therefore, for example, the autocorrelation between $Q_{2,1}$ and $Q_{3,3}$ in data matrix M2 will be represented by $\rho(1,2)$ and the autocorrelation between $Q_{1,4}$ and $Q_{3,1}$ will be represented by $\rho(-2,3)$. Here, $\rho(1,2)$ indicates one lag vertically down and two lags horizontally from left to right, while $\rho(-2,3)$ means two lags from bottom to top and three lags from left to right. However, due to the symmetry properties of the autocorrelation function given by Eqs. 6.30 and 6.31, $\rho(1,2)$ and $\rho(-2,3)$ are also equal to $\rho(-1,-2)$ and $\rho(2,-3)$ respectively.

For convenience of representation and manipulation of the procedure, the available data points (Data Array M2) were numbered from top to bottom and proceeding from the left most cone hole, A, to the right most cone hole, G (Fig. 6.1). The numbered data points are shown in data array M3. With this modified notation, $\rho(1, m+3)$ will actually be the autocorrelation function between points 1 and $m+3$ where w, the
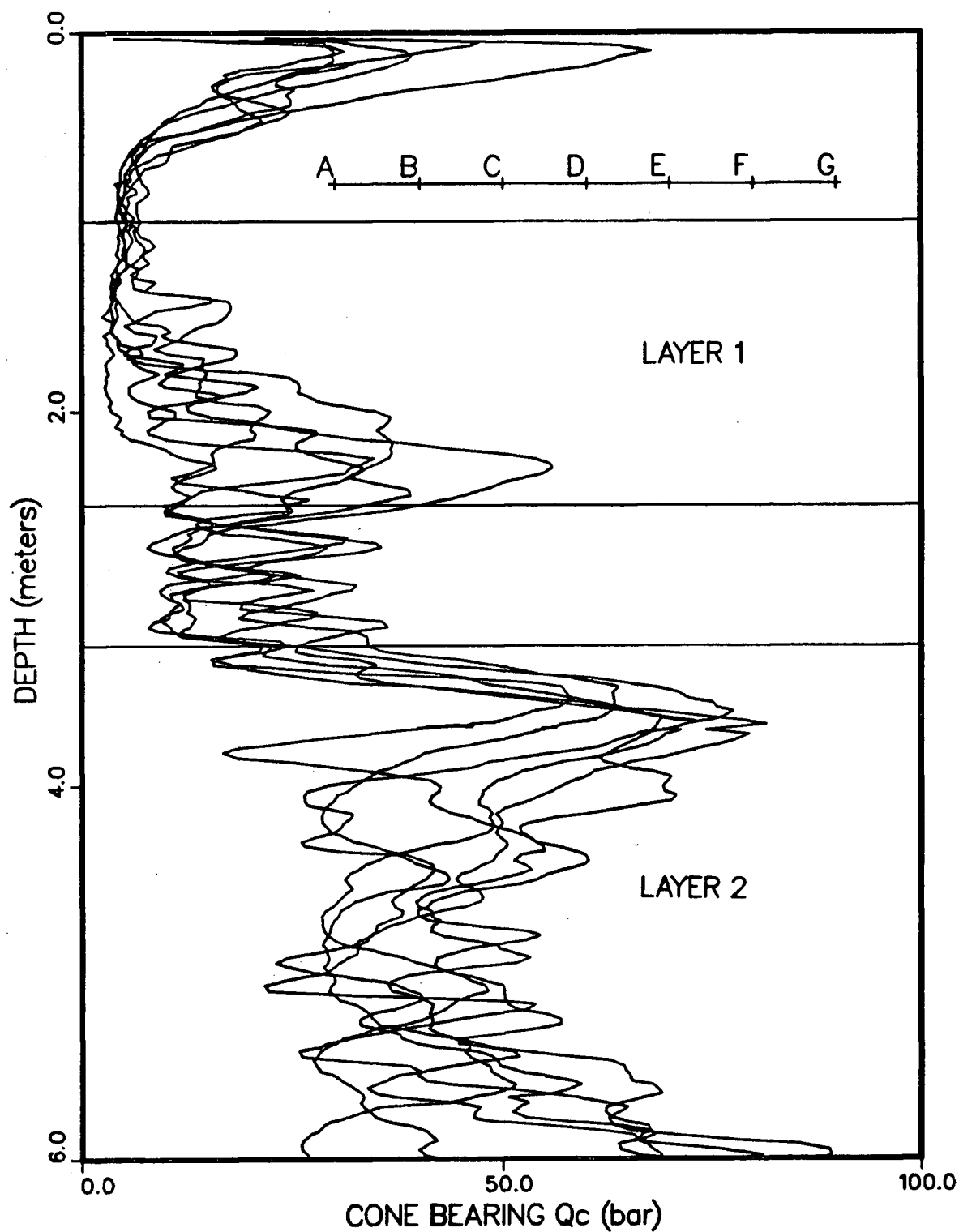
Figure 6.1: Distribution of Cone Bearing Profiles Across the Site Used for the Interpolation at McDonald Farm.

horizontal lag, is equal to unity and the vertical lag (r) is equal to two, both being positive.

Data Array M3

$$Q_{1,1}(1) \quad Q_{1,2}(m+1) \quad Q_{1,3}(2m+1) \quad Q_{1,4}(3m+1) \quad \ldots \quad \ldots \quad \ldots \quad Q_{1,n}$$

$$Q_{2,1}(2) \quad Q_{2,2}(m+2) \quad Q_{2,3}(2m+2) \quad Q_{2,4}(3m+2) \quad \ldots \quad \ldots \quad \ldots \quad Q_{2,n}$$

$$Q_{3,1}(3) \quad Q_{3,2}(m+3) \quad Q_{3,3}(2m+3) \quad Q_{3,4}(3m+3) \quad \ldots \quad \ldots \quad \ldots \quad Q_{3,n}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$Q_{m,1}(m) \quad Q_{m,2}(2m) \quad Q_{m,3}(3m) \quad Q_{m,4}(4m) \quad \ldots \quad \ldots \quad \ldots \quad Q_{m,n}$$

6.5   Application of the Interpolation Procedure

The method of interpolation described was used to interpolate between cone holes obtained across a 30 meter stretch at the McDonald Farm site. Seven CPT's were performed at 5 meter intervals along a straight line. The depth of penetration was 6 meters. The scatter of the seven cone profiles is illustrated in Fig. 6.1 which also gives the layout plan of the cone holes, A through G. In this exercise, the cone profiles at A, B, C, E, F and G will be used to predict the cone profile at D, so that a comparison could be done between the predicted profile and the actual bearing profile obtained. The individual bearing profiles for cone holes A, B and C are illustrated in Fig. 6.2, and for E, F and G in Fig. 6.3. The data in the vertical direction were considered in groups of ten, so that spacing of data points would be 25 cm(2.5×10). As a result, each data point in the vertical dimension represented an averaged cone bearing of a region of 2.5 cm.. This was done to alleviate the possibility of extremities affecting the predicted correlation functions. An increased vertical spacing would also be more preferable, since it is not advisable to have a two dimensional autocorrelation function which has a very high horizontal to vertical lag distance ratio (l_h/l_v). A vertical
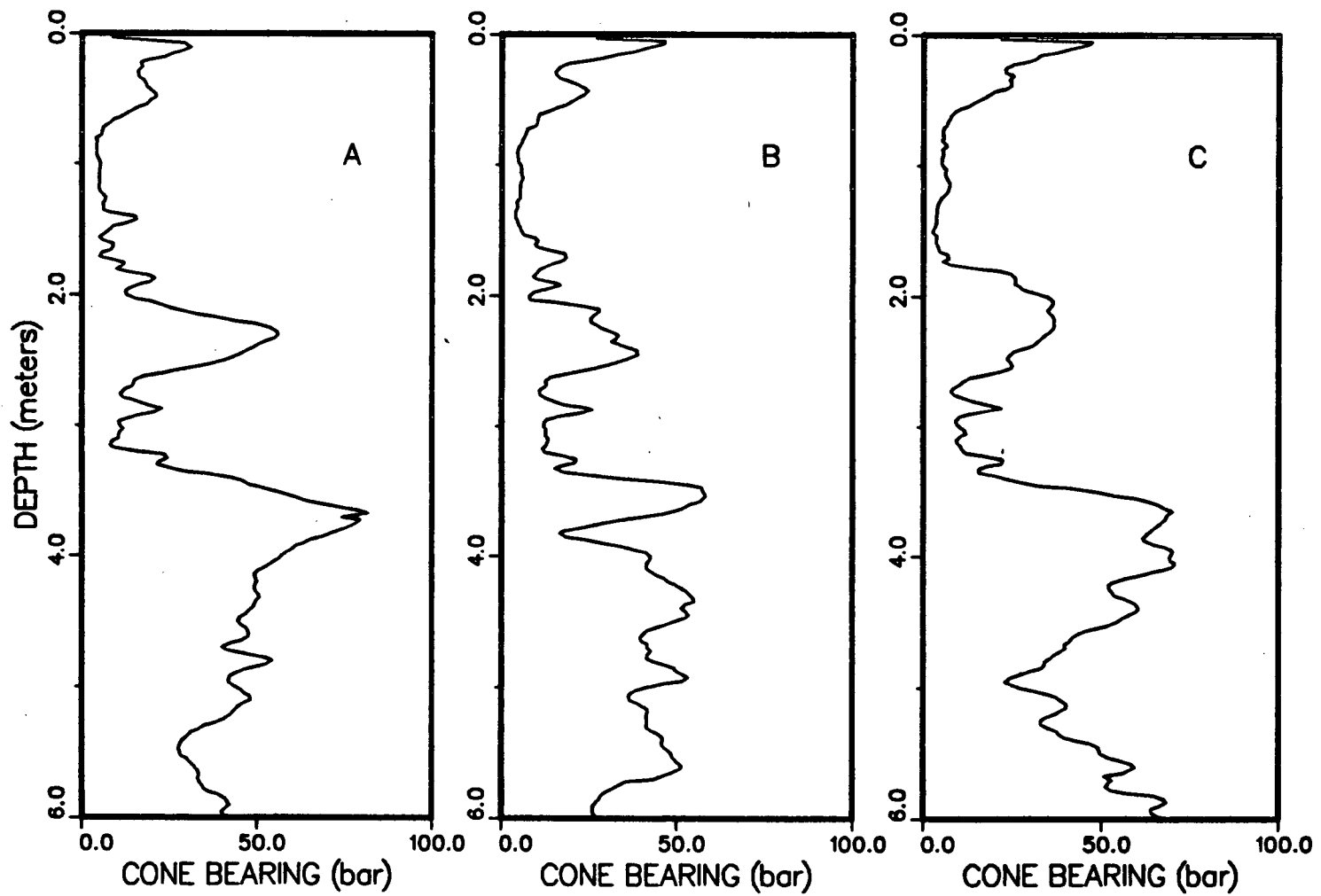
Figure 6.2: Cone Bearing Profiles at Locations A, B and C at McDonald Farm.
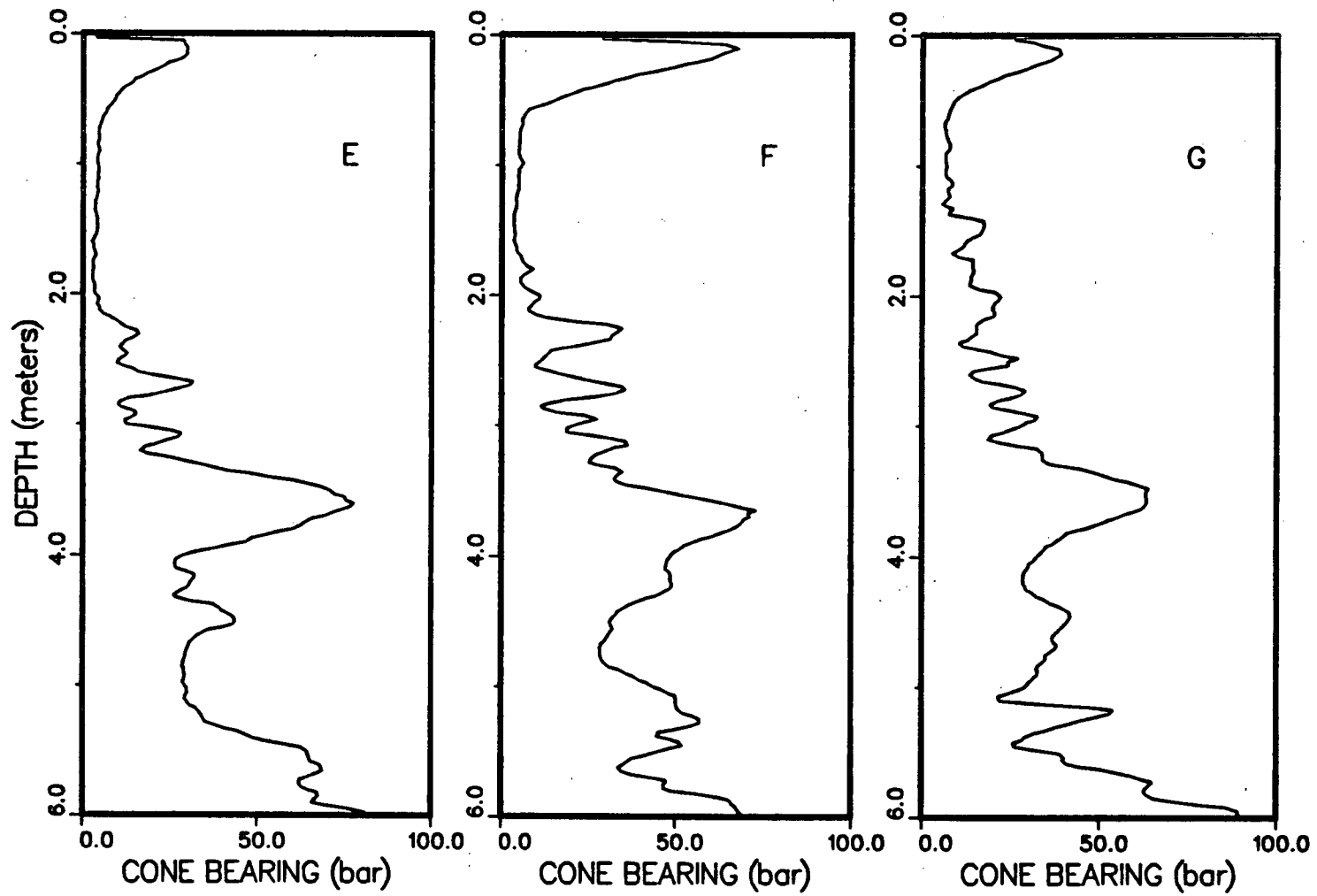
Figure 6.3: Cone Bearing Profiles at Locations E, F and G at McDonald Farm.

spacing of 2.5 cm would result in a l_h/l_v value of 200 while the increased spacing of 25 cm would give a l_h/l_v value of 20, which is more desirable.

Two prominent layers were identified in the 6 m profile (Fig. 6.1) using methods of layer identification discussed in Chapter 2. Layer 1 lies between 1.00 and 2.50 m, and Layer 3 between 3.25 and 6.0 meters. Layers 2 was found to be stationary, while Layer 1 was non-stationary, which was confirmed using the RUN Test at a 95% significant level. A typical data layout is given in Data Array M4. The CPT's at A, B, C, E, F and G will be used to predict the bearing profile at D.

Data Array M4

Depth	A	B	C	**D**	E	F	G
d_1	$Q_{1,1}$	$Q_{1,2}$	$Q_{1,3}$	$\mathbf{Q_{1,0}}$	$Q_{1,5}$	$Q_{1,6}$	$Q_{1,7}$
d_2	$Q_{2,1}$	$Q_{2,2}$	$Q_{2,3}$	$\mathbf{Q_{2,0}}$	$Q_{2,5}$	$Q_{2,6}$	$Q_{2,7}$
d_3	$Q_{3,1}$	$Q_{3,2}$	$Q_{3,3}$	$\mathbf{Q_{3,0}}$	$Q_{3,5}$	$Q_{3,6}$	$Q_{3,7}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
d_m	$Q_{m,1}$	$Q_{m,2}$	$Q_{m,3}$	$\mathbf{Q_{m,0}}$	$Q_{m,5}$	$Q_{m,6}$	$Q_{m,7}$

Let the horizontal dimension have a zero value at A and 30.0 at G. For Layer 1 data, the vertical dimension will be zero at d_1 and 1.50 at d_m. Since layer 1 was non - stationary, several functions were tried out to best represent the trend. Using methods described in Chapter 3, the following model was selected as the most appropriate.

$$\hat{Q} = 7.59 + 21.5y^2 - 0.60xy \qquad (6.32)$$

The ensuing residuals were checked for correlation using the Durbin - Watson statistic (Durbin and Watson, 1951). From tables, d_L was found to be equal to 1.19 and d_u equal to 1.55. The actual value (d_w) obtained was 2.51. Since $4 - d_w = 1.49 < d_u$

Table 6.1: Values of Constants for the Autocorrelation Model for Data in Layers 1 and 2.

	Function	c_1	c_2	c_3	c_4
Layer 1	$\rho(\Delta x, \Delta z)$	18.61	0.49	3.80	0.57
	$\rho(\Delta x, -\Delta z)$	476.48	0.76	5.40	0.72
Layer 2	$\rho(\Delta x, \Delta z)$	7.35	0.46	7.35	0.46
	$\rho(\Delta x, -\Delta z)$	9.11	0.48	9.11	0.48

even for a significance level as high as 97.5%, the autocorrelations of the residuals are significant and have to be considered in any efficient interpolation procedure. The process of verifying the stationarity and correlation of residuals have been already described in Chapter 3.

The two dimensional autocorrelation function of Layer 1 exhibited both negative and positive values, and therefore, it was necessary to model the autocorrelation function, with an exponential sinusoidal type of function which has the capacity to accommodate both positive and negative values. Layer 2 stationary data also exhibited significant correlation. The autocorrelation functions of these data had positive and negative values, emphasizing the need for an exponential sinusoidal function which is given by Eq. 6.33 below. It is somewhat similar to the one given in Eq. 6.10, except that it is more flexible with four constants instead of two.

$$\rho(\Delta x, \Delta z) = \text{EXP}\left[-\left(c_1 \mid \Delta x \mid + c_2 \mid \Delta z \mid\right)\right] \text{COS}\left[\frac{\Delta x}{c_3} + \frac{\Delta z}{c_4}\right] \qquad (6.33)$$

As mentioned before, two types of autocorrelation functions need to be considered and the values of the constants for these two types of functions are tabulated for the two layers in Table 6.1.

Table 6.1 shows that there is a significant difference in the two types of autocorrelation functions for data of both layers, with the difference in Layer 1 data being more significant. All the above functions gave a high multiple correlation coefficient, R^2, in the region of 0.80. The d_w value of the ensuing residuals also suggested that most

of the correlations have been absorbed by the respective autocorrelation functions, reflecting their adequacy. The fit was found to be extremely good especially at the closer lags. In the development of the above correlation functions, all the available data except for the data at D have been considered. However, when calculating the weights from Eq. 6.21 for predicting the bearing profile at D, only the two columns of data immediately adjacent (C and E) to D were considered. The influence of the data points in the other columns were negligible due to the screening effect. For example, column B will be screened by column C and column F will be screened by column E. The data in the two most outer columns will have even a lesser influence, due to the double screening effect. That is, A will be screened by both B and C, while G will be screened by E and F. At this point, it should be emphasized that although only the data in columns C and E have been directly used in the final prediction from Eq. 6.21, all the available data with the exception of the data at D have been used in obtaining the model for the autocorrelation function. For the purpose of inter-polation, $\{P\}$ in Eq. 6.22 will always remain the same if the same data set is used for multiple interpolations. [M] will depend on the point at which interpolation is required, and therefore will change with different points of interpolation, s_o. Once the autocorrelation function for a given set of data is obtained, $\{P\}$ which is dependent on the data points to be used for the interpolation, can be determined. In a similar way, [M] can be determined by substituting the relevant Δx and Δz terms in the derived autocorrelation function (Eq. 6.33). The values of Δx and Δz in the case of [M] will be the respective distances from the point of interpolation $s_o(x_o, z_o)$ to the data points $s_1(x_1, z_1)$, $s_2(x_2, z_2)$, etc.). In $\{P\}$, the values of Δx and Δz will be the respective distances between data points $s_1(x_1, z_1)$, $s_2(x_2, z_2)$ etc..

Table 6.2 has the detailed results for the interpolation at D, together with the variance and 95% confidence bands. The Layer 1 results comprise of two parts: the

estimate from the regression and the estimate from the correlation analysis. Variance of the regressed part (σ_r^2) was obtained from methods given in Chapter 3. The variance of the correlated estimates (σ_e^2) was calculated from Eq. 6.26. The total variance, σ_t^2 was determined from Eq. 6.34 below, assuming that σ_r^2 and σ_e^2 are independent.

$$\sigma_t^2 = \sigma_r^2 + \sigma_e^2 \tag{6.34}$$

Assuming normality, the lower 95% confidence estimate, $\hat{Q}_L(s_o)$, and the upper 95% confidence estimate, $\hat{Q}_U(s_o)$, are given by,

$$\hat{Q}_L(s_o) = \hat{Q}(s_o) - 2\sigma_t \tag{6.35}$$

$$\hat{Q}_U(s_o) = \hat{Q}(s_o) + 2\sigma_t \tag{6.36}$$

Table 6.2: Results for the Interpolation at D

Depth (m)	Regressed Estimate	Correlation Estimate	Total Estimate	σ_e	σ_r	Q_L	Q_U
1.13	7.59	-4.16	**3.43**	1.72	2.09	0.00	11.05
1.38	6.68	-4.78	**1.91**	1.60	1.91	0.00	8.94
1.63	8.47	3.66	**12.13**	6.50	1.66	0.00	28.44
1.88	12.93	-6.99	**5.95**	2.50	1.45	0.00	13.84
2.13	20.09	-4.03	**16.06**	5.60	1.79	1.28	30.82
2.38	29.93	-1.01	**28.83**	5.72	2.89	11.61	46.05
3.38	-	37.12	**37.12**	14.73	-	7.66	63.95
3.63	-	34.49	**34.49**	11.66	-	11.17	57.81
3.88	-	31.80	**31.80**	7.47	-	16.86	46.74
4.13	-	50.31	**50.31**	12.87	-	24.57	76.05
4.38	-	54.96	**54.96**	11.69	-	30.98	78.34
4.63	-	49.97	**49.97**	10.06	-	29.35	69.59
4.88	-	40.64	**40.64**	11.69	-	17.26	64.02
5.13	-	38.47	**38.47**	12.87	-	12.73	64.21
5.38	-	45.74	**45.74**	7.47	-	30.80	60.68
5.63	-	50.49	**50.49**	11.66	-	27.17	73.81
5.88	-	64.81	**64.81**	14.73	-	35.35	94.27

The interpolated profile at D is given in Fig. 6.4 together with the actual measured profile. The two compare well, except between 3.5 and 4.0 meters. The prediction would exhibit even better results if the data in the six cone holes were more correlated. The fairly high variance of the data is caused by the small data base which is generally the case in geotechnical engineering. The estimates, together with the 95% confidence bands are illustrated in Fig. 6.5 which also shows the actual profile at D for comparison purposes. The results are also tabulated in Table 6.2.

As a second exercise, interpolation at point F was performed from the available data at A, B, C, D, E and G. A similar procedure was followed as for the interpolation at D and the results of the prediction are given in Fig. 6.6. The prediction at F models the actual profile satisfactorily, which is a significant improvement over the average profile of the entire site. The average profiles generally used by geotechnical engineers neglect correlation.

Figures 6.7 and 6.8 illustrate predicted profiles at M and N, respectively. As shown in Fig. 6.7, M lies halfway between D and E, and point N in Fig. 6.8 is 2 meters away from E towards F. The immediately adjacent profiles are also illustrated for comparison purposes. The predicted profiles indicate that they are significantly different from the mean values of the two adjacent profiles.
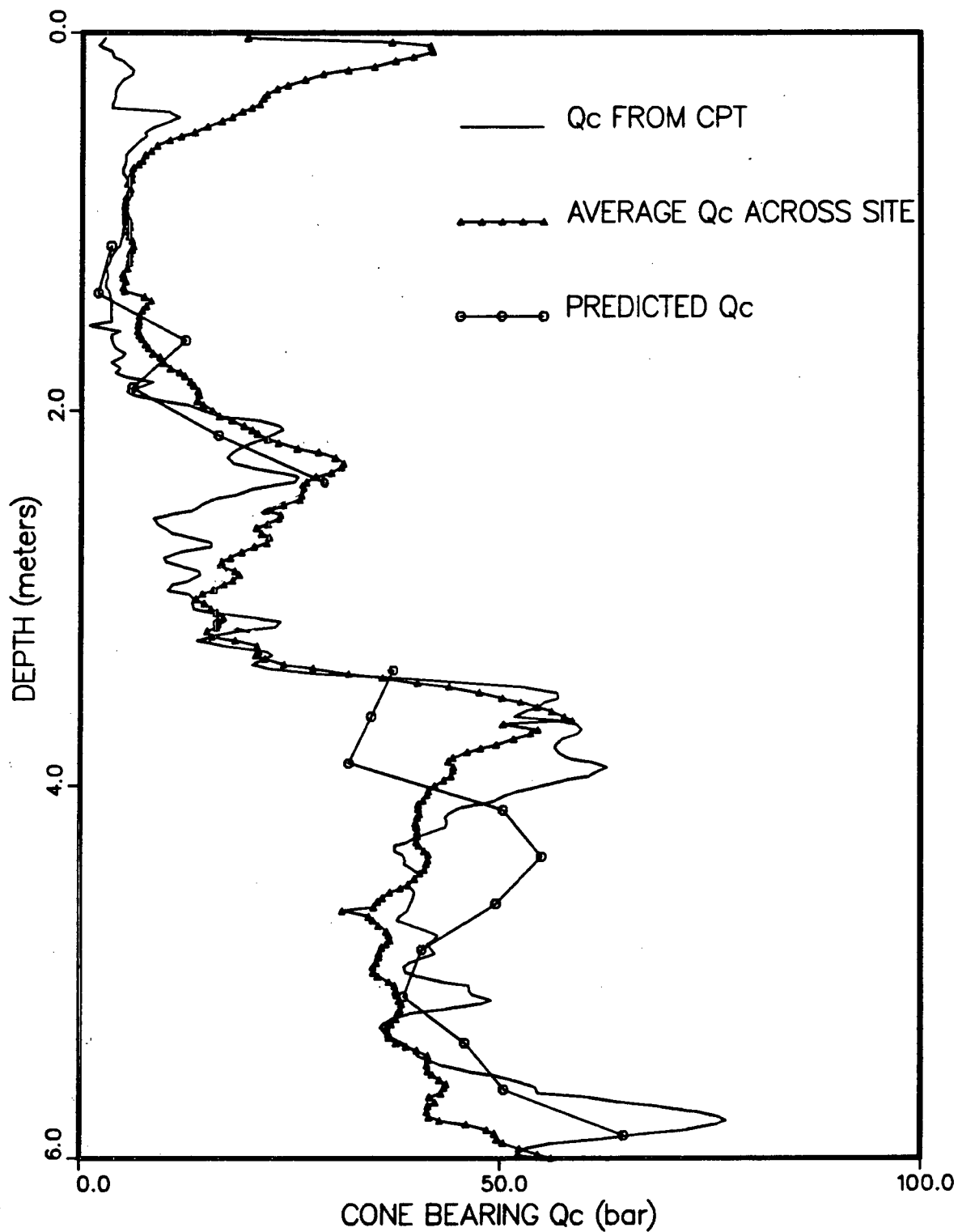
Figure 6.4: Comparison of Predicted Cone Bearing at D with the Actual Cone Bearing Profile and the Average Cone Bearing Profile Across the Site.
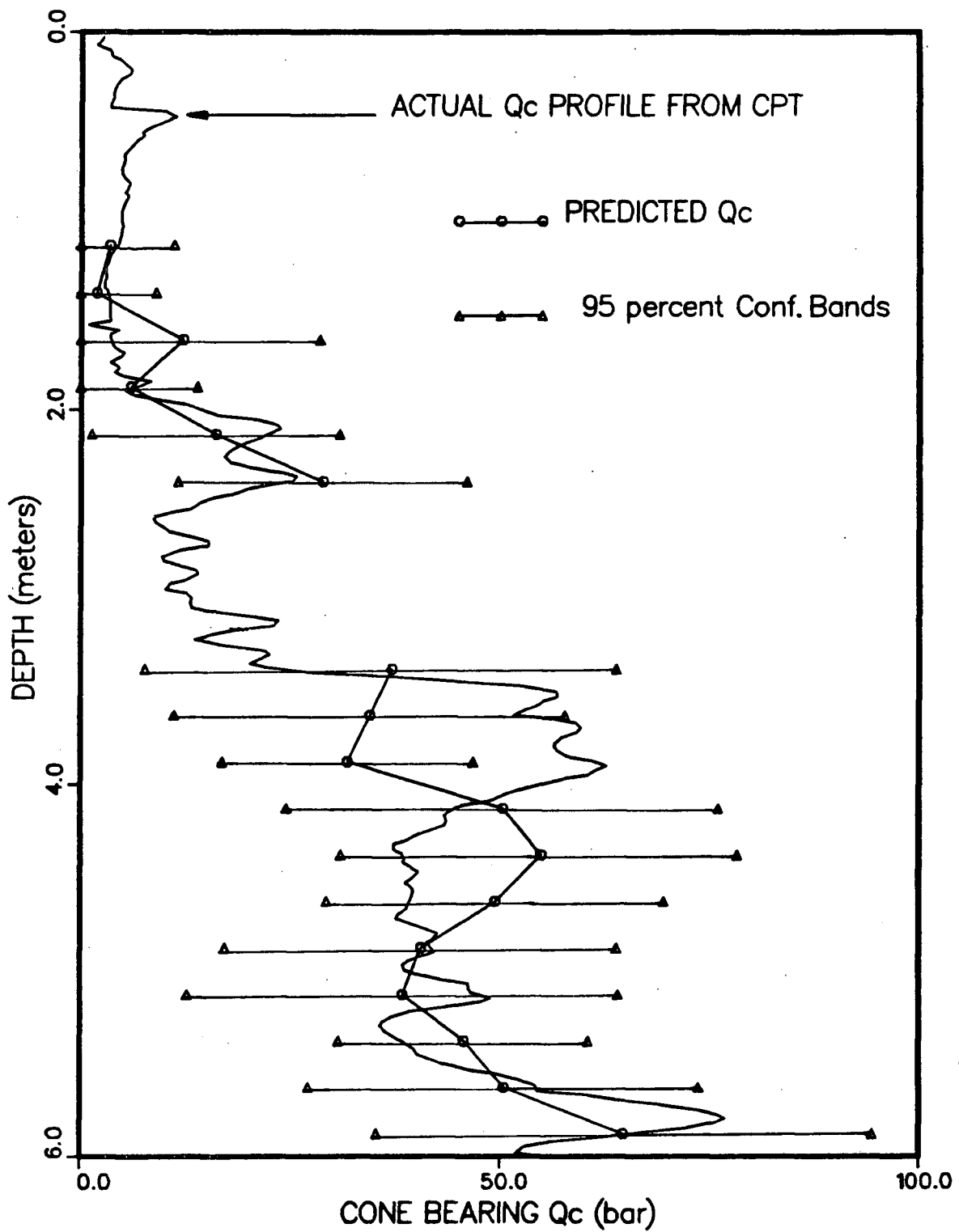
Figure 6.5: Confidence Bands of Predicted Profile at D and the Predicted and Actual Cone Bearing Profiles.
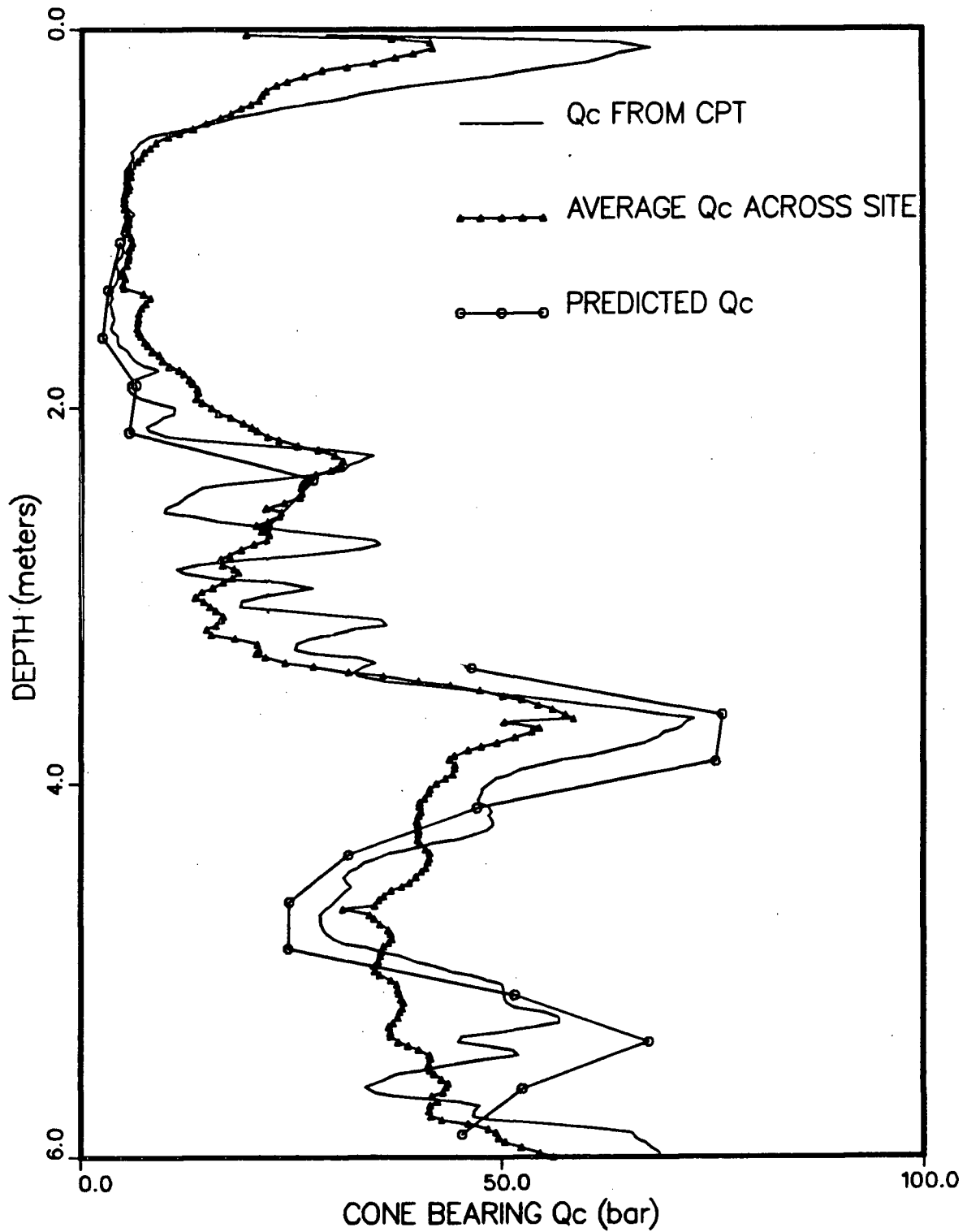
Figure 6.6: Comparison of Predicted Cone Bearing at F with the Actual Cone Bearing Profile and the Average Cone Bearing Profile Across the Site.
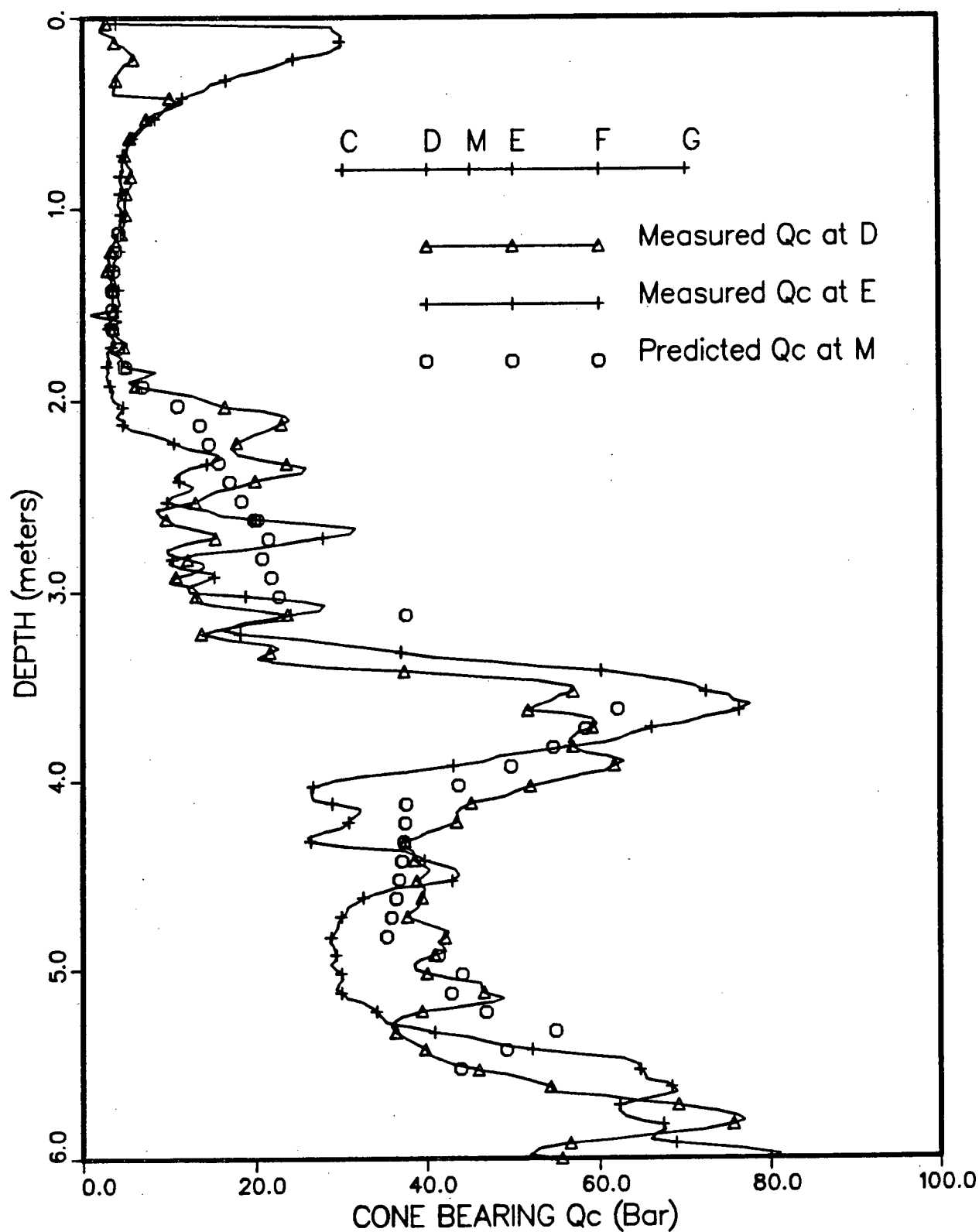
Figure 6.7: Predicted Cone Bearing at M with the Adjacent Cone Bearing Profiles at D and M.
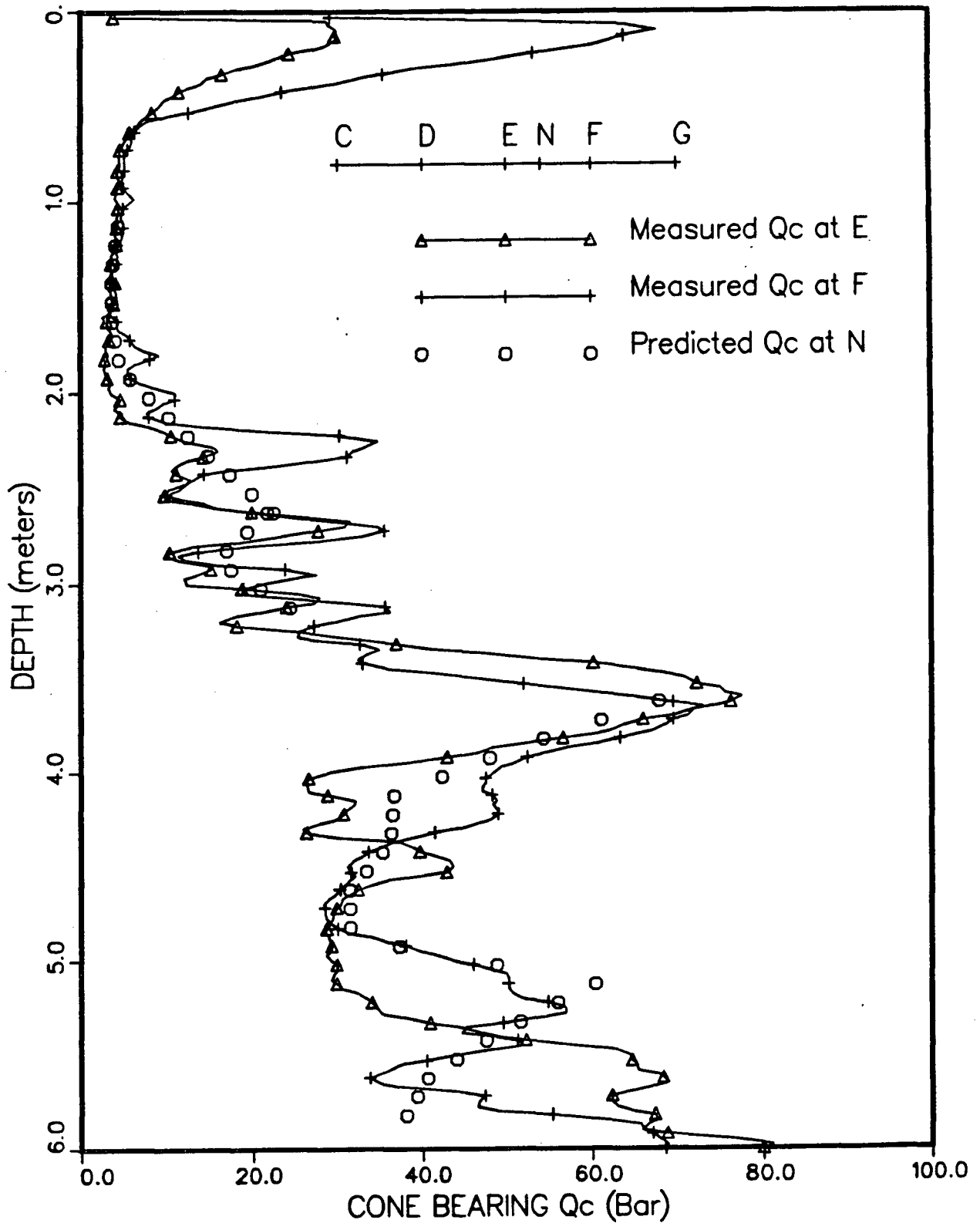
Figure 6.8: Predicted Cone Bearing at N with the Adjacent Cone Bearing Profiles at E and F.

6.6 Conclusions

The main conclusions of this chapter are listed below;

(i) The proposed method of developing the two dimensional autocorrelation function provides a convenient and logical way for dealing with the two types of different correlations present in geotechnical data.

(ii) The autocorrelation of two dimensional soil test data is best represented by exponential sinusoidal functions, due to its capability of having both positive and negative values.

(iii) The fitting of the best possible function for the autocorrelation coefficients of the data was found to be the most tedious part of the analysis process. In situations where the correlation coefficient of the fit is not high, a higher weight can be given to points which are closer to the estimation point, since the points which are farther away from the estimation point have a lesser effect due to screening.

(iv) The applications of the proposed procedure have indicated the need for the consideration of correlations, if they are found to exist. Often, it may be found that the correlation is negligible; in which case, it is sufficient to perform a deterministic trend analysis. However, if the correlation is appreciable, it has to be considered in any interpolation procedure where reasonable estimates are desirable.

(v) The procedure of interpolations considering correlations, allows the designer to interpolate at some location based on limited data and yet with a confidence level in mind. In the traditional method neglecting correlations, the only options available to the engineer are either the use of the mean or more likely the minimum profile and vary the factor of safety accordingly.

(vi) One of the two major shortcomings in this interpolation procedure is that data points have to be regularly spaced. Generally, this requirement is satisfied in the vertical dimension but rarely so in the horizontal dimension. In such situations

data will have to be classified into groups having similar horizontal spacing. The other drawback of this technique is that it is only applicable to large geotechnical projects with a sizeable data base.

What this chapter has demonstrated is a simple and efficient procedure of correlation analysis which can be easily employed for large geotechnical problems where interpolations may be required with a reasonable degree of confidence.

Chapter 7

Statistical Methods to Evaluate Soil Densification: A Case History

7.1 Introduction

7.1.1 General

Some of the statistical techniques described in the thesis have been applied to a ground improvement case history involving the Franki Tri Star probe (Massarsch and Vanneste, 1988). The site in which the soil compaction was performed is situated at the north side of Annacis Island along the north channel crossing and immediately east of the Alex Fraser Highway (Gray Beverage canning plant site). It has to be emphasized that the applications described in this chapter does not encompass all the techniques proposed and presented in this thesis. However, some of the techniques such as layer identification, trend analysis and the concept of the scale of fluctuation have been used to assess the effects of compaction on soil variability. The effects of soil densification were investigated with respect to distance from the Tri Star probe location as well as with respect to elapsed time of densification. The location of the CPT's conducted before and after densification are given in Fig. 7.1.

CPT data given by CT1, CT2 and CT3 were used for investigating the effect of time on soil improvement and CT1, CT3, CD1, CD2, CD3 and CD4 were used to study the effect of distance on densification. The above CPT profiles can be described as follows:
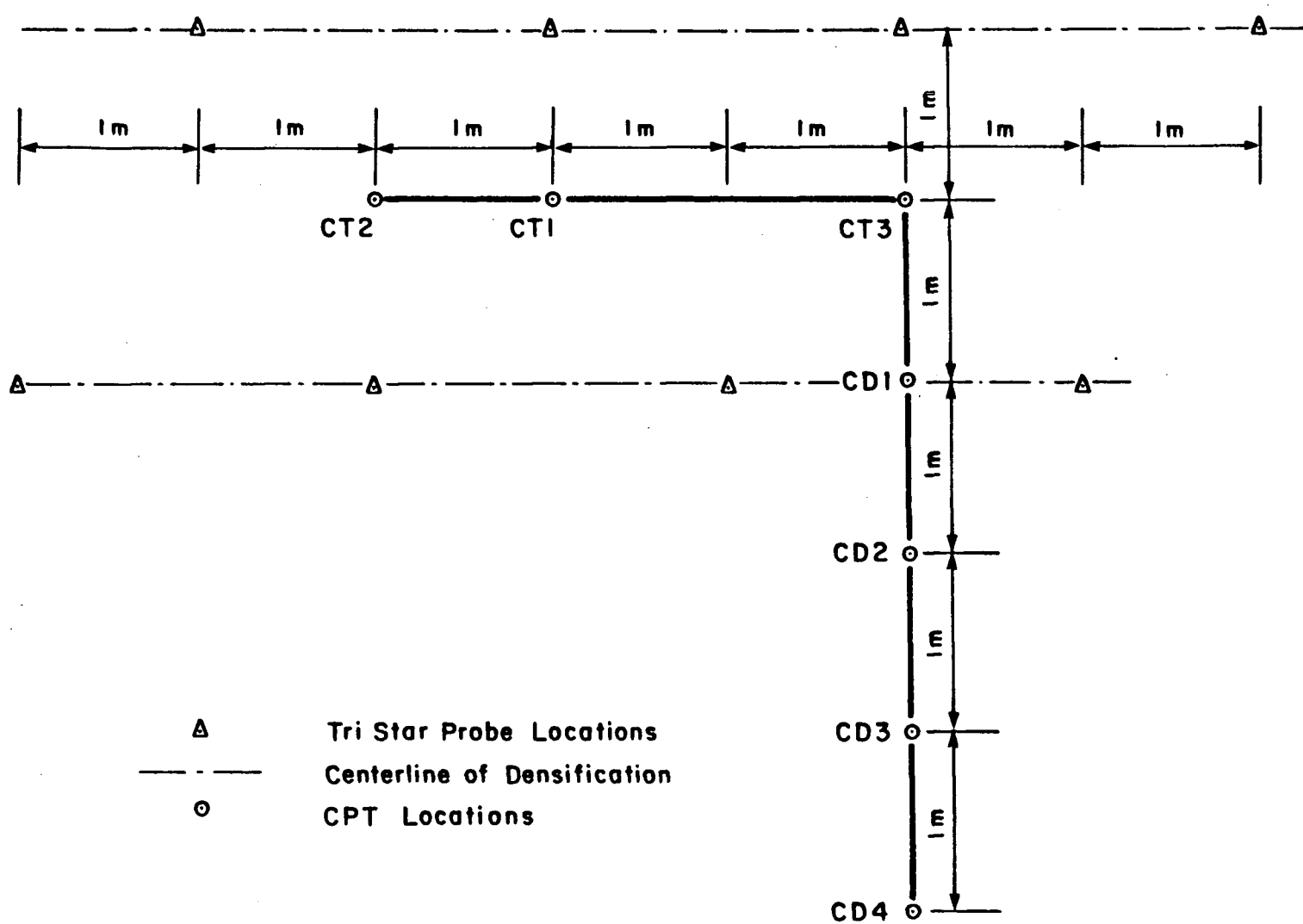
Figure 7.1: Location Plan of CPT Soundings and Tri Star Probe Locations.

CT1 - before densification

CT2 - 67 days after densification

CT3 - 82 days after densification

CD1 - 1 m away from CT1

CD2 - 2 m away from CT1

CD3 - 3 m away from CT1

CD4 - 4 m away from CT1

Note that CD1, CD2, CD3 and CD4 soundings were obtained 82 days after densification. CT1, CT2 and CT3 were equidistant to probe locations.

In the analysis to follow, it was assumed that there was no appreciable inherent variability across the site. This assumption would not have been necessary if all the above points were also tested prior to densification. However, this assumption is not of much concern here since the the main purpose of this exercise was not to evaluate the effectiveness of the testing program or the efficiency of the Tri Star probe but merely to demonstrate the applicability and usefulness of some of the statistical methods described in the thesis.

7.1.2   Site Description

Preliminary investigations indicated that the site was covered by a recently placed 1.8 to 2.4 m thick sand fill on top of a 2.4 to 3.9 m thick clayey silt underlain by an alluvial sand extending below 10 m in depth. The water table was located about 2.0 m below the existing ground surface. It was required to densify the saturated alluvial sand lying approximately between 5.0 and 10.0 m due to its susceptibility to liquefaction in the event of a strong earthquake (Massarsch and Vanneste, 1988).

7.1.3   Tri Star Probe

The Tri Star probe which was selected for the densification, was inserted vertically using a heavy vibrator. It consists of three long steel plates, approximately 20 mm thick and 500 mm wide, welded along a common edge at an angle of 120 degrees. The length of the probe used was 12 m. The compaction process can be divided into three main phases (Massarsch and Vanneste, 1988): probe penetration which takes about two to three minutes to reach the desired depth of 10 m, steady state vibration during which period the tip of the probe was kept at 10 m for a pre-determined duration, and the extraction phase. To minimize the possibility of decomposition of the soil due to probe extraction, it was not withdrawn in one continuous movement but was performed in stages by stopping for a certain time at different depths on its way to the surface.

7.2   Identification of Layers

The CPT profiles at the different locations included cone bearing, sleeve friction and pore pressure (Fig. 7.2). In this study however, only the cone bearing results were analyzed because the main concern of this investigation was to study the effects of densification which would be best represented by the improvement in cone bearing stress. The profiles illustrated in Fig. 7.3 indicate the highly variable nature of the soil stratum in the top 10 m and also exhibit the presence of several layers.

Two of the statistical methods described in Chapter 2 were used for the identification of the layering and included the Intraclass Correlation Coefficient (section 2.3.2.2) and the Gradient method (section 2.5.1). A closer inspection of the bearing profiles indicated that it exhibited similar characteristics to the Type I profile (Fig. 2.28) at certain depths. As already described in section 2.5.1, this type of layer
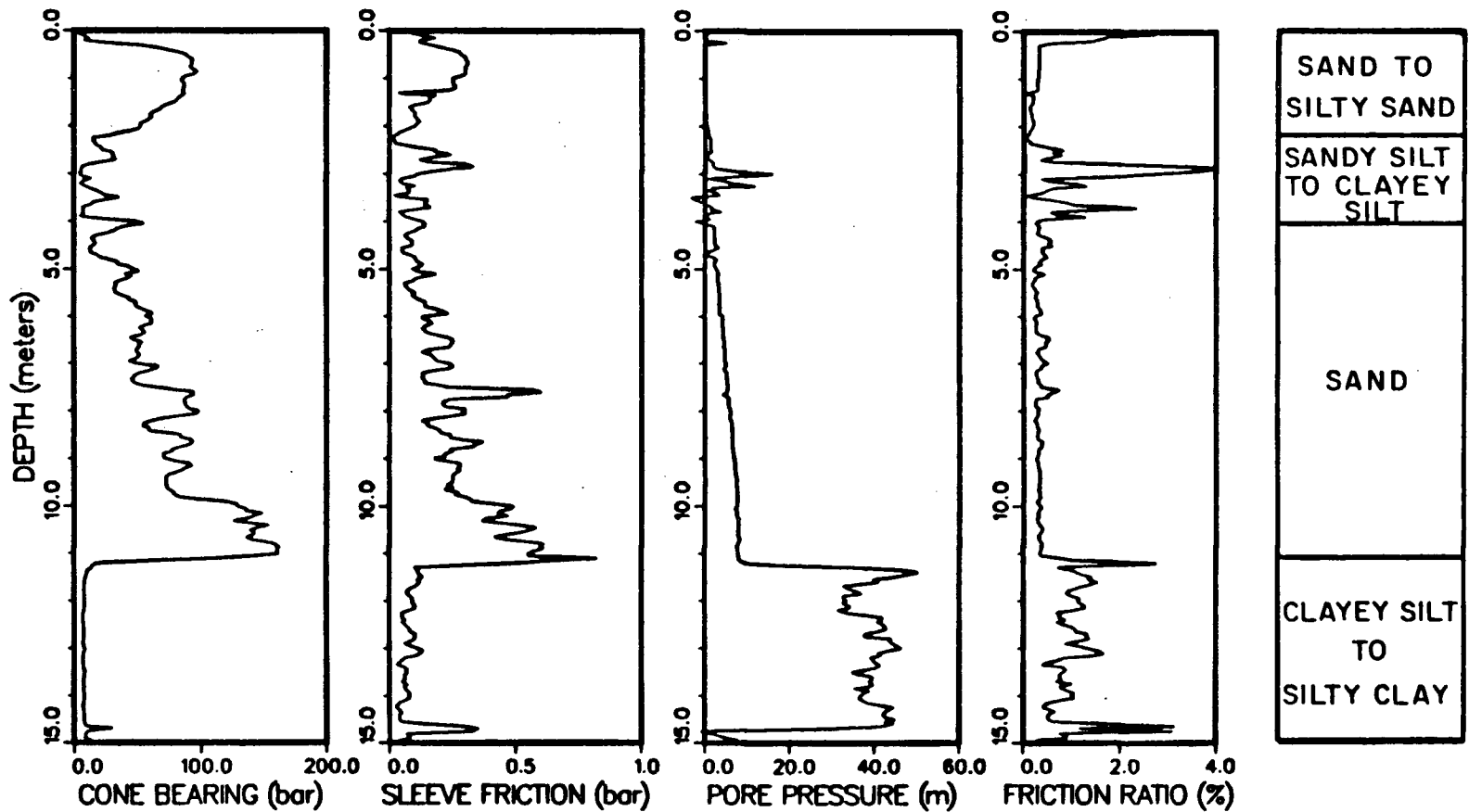
Figure 7.2: Cone Bearing , Sleeve Friction, Pore Pressure and Friction Ratio Profiles of CT1
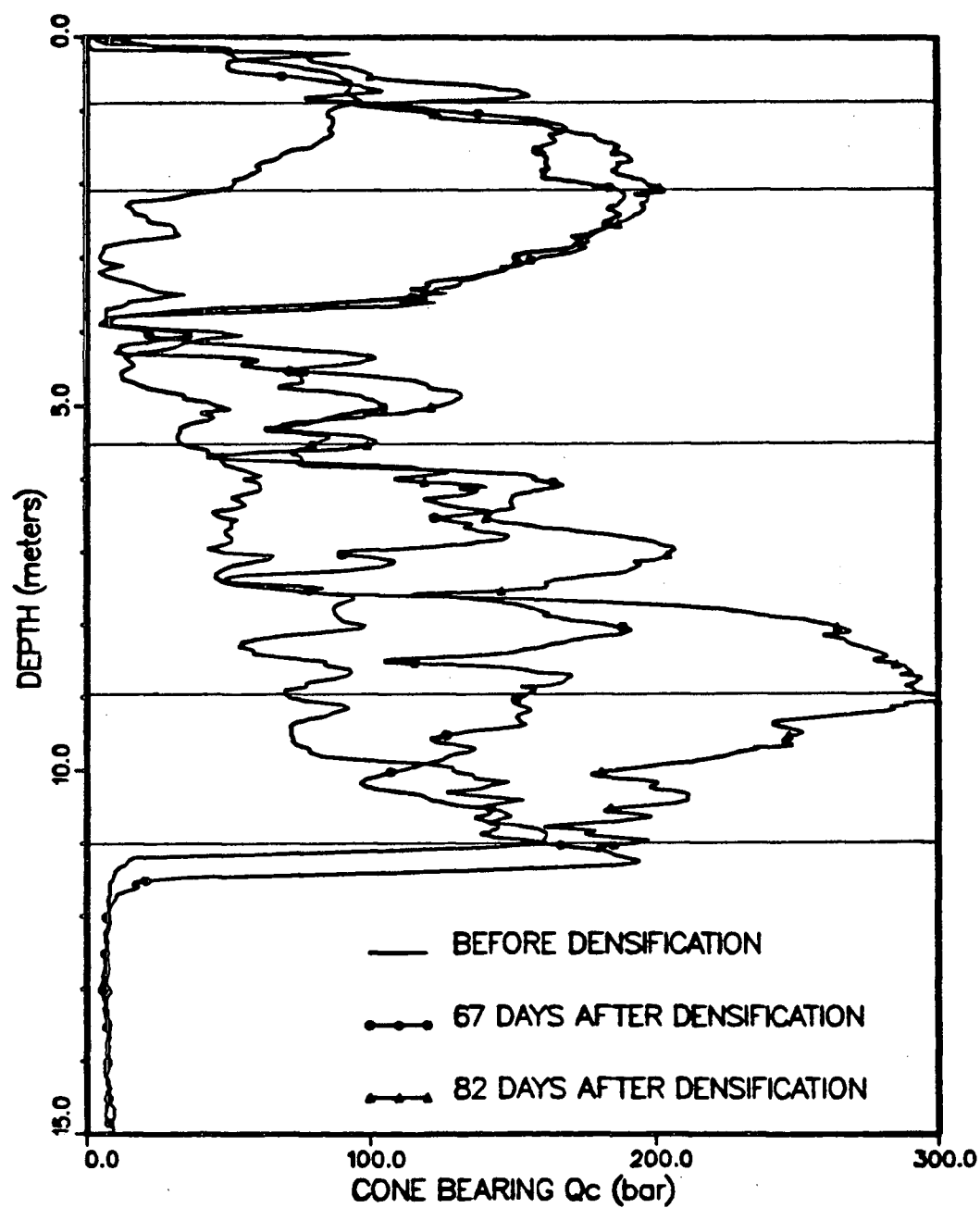
Figure 7.3: Cone Bearing Profiles Before Densification (CT1) and 67 Days (CT2) and 82 Days (CT3) after Densification.

boundary was not picked up by the Intraclass Correlation Coefficient, thus necessitating the need for the use of the Gradient method. Densification causes varying degrees of variability in the profile and in this respect statistical techniques such as the Gradient method would be most suitable in picking out different layers. The conventional method of selecting layers (using Friction Ratio - Bearing Classification chart) could not perform this task efficiently, since it can not differentiate layering by considering the difference of variability between the layers. The Gradient method of layer discrimination was chosen to pick sublayer boundaries due to the highly non-uniform nature of densification. The effect of mixing of soils also contributed to the above. A visual inspection of the profiles (Fig. 7.2) indicated the presence of thin layers and therefore a window thickness of 0.5 m was selected.

For practical convenience of comparison between the different profiles similar layer boundaries were selected for all profiles based on the layer boundaries determined for CT1, CT2 and CT3 (Table 7.1). Based on the results given in Table 7.1, the following depths were decided upon.

Layer 1 :	0.0 - 0.85 m
Layer 2 :	0.85 - 2.10 m
Layer 3 :	2.10 - 5.45 m
Layer 4 :	5.45 - 9.00 m
Layer 5 :	9.00 - 11.00 m

7.3 Trend Analysis

Once the layers were identified it was necessary to investigate the type of depth dependency of each layer. Trend analysis methods described in Chapter 3 were used for this purpose and linear trends were found to be satisfactory in all cases with correlation coefficients in excess of 0.70 (section 3.3.3). It was found in some cases,

Table 7.1: Layer Boundaries Based on Statistical Methods for CPT Profiles CT1, CT2 and CT3.

Depth (m)	BEFORE DENSIFICATION (CT1)	NUMBER OF DAYS AFTER DENSIFICATION	
		67 (CT2)	82 (CT3)
0			
	(0.85)	(0.83)	(0.92)
	(2.20)	(2.15)	(2.10)
			(3.90)
5			(5.35)
	(5.45)	(5.60)	
		(8.70)	
	(8.90)		(9.15)
10			
	(11.0)	(11.0)	(10.95)
15			

that curvilinear trends would only marginally increase the efficiency of the fit and therefore was abandoned in favor of linear trends which were more convenient for comparison purposes. A combination of linear and curvilinear trends for different sublayers of the same stratum is not recommended here since it only complicates the procedure and the comparison of the trends of different sublayers. As a result of the above considerations, linear trends were found to be most suitable. The relatively low thicknesses of the layers selected also helped to ensure the adequacy of linear trends and alleviated the need for the use of curvilinear trends. If the optimal layer boundaries were individually selected for the different profiles, the respective correlation coefficients would be increased to values in the range between 0.74 and 0.80. However, this marginal reduction of the correlation coefficient compensates for the additional practical convenience gained in selecting similar layer boundaries for all profiles, facilitating easier comparison.

7.4   Effect of Densification with Time

Figure 7.3 illustrates the cone bearing profiles before densification (CT1), 67 days after densification (CT2) and 82 days after densification (CT3). Mitchell and Solymar (1984) report that freshly deposited or densified sand may exhibit substantial stiffening and strength increase with time up to several months. This effect is amply evident from the bearing profiles in Fig. 7.3 although it seems to be exaggerated in profile CT3. While a proportion of the increase in CT3 can be attributed to the time effect on densification, the rest of the increase could be due to the reported lowering of the water table and the subsequent gain in strength of the sand. Another possibility for this significant change could be a result of the natural soil variability across the site.

Figure 7.4 shows how statistical filtering and smoothing improve the trend iden-
tifying capability of a variable profile. The data given in Fig. 7.3 were first filtered
using the median method (section 3.2.2) and were subsequently smoothed by the
method of Fourier transforms (section 3.2.1.2) as discussed in Chapter 3. The initial
process of filtering enabled the extremeties of the data to be filtered out. The median
method was used for this purpose with a filtering window, BS = 1.5 (a low degree
of filtering) and replacement of filtered points were performed by the substitution of
the mean of the the adjacent two unfiltered data points (section 3.2.2). As explained
in section 3.2.2 a low degree of filtering (BS = 1.5) was used in order to avoid the
possibility of missing out actual layers. Due to the low degree of filtering used, any
trends present in the profile were not immediately apparent and therefore the profile
was subsequently smoothed using Fourier transforms. It has to be reiterated that
procedures of filtering and smoothing should be used with utmost caution, with its
main purpose being to facilitate the easier identification of trends and is definitely
not used for any analytical purposes.

Figure 7.5 illustrates the coefficient of variation (section 4.2) with depth prior
to filtering and smoothing. It shows clearly that the variability has decreased after
densification, and that the effect of time on variability is minimal. The effect of den-
sification on variability can be more efficiently captured using the scale of fluctuation
to be discussed in section 7.4.2.

7.4.1   Evaluation of Trend and Confidence Estimates

Figure 7.6 illustrates the trend of the cone bearing before and after densification
(CT1 and CT2) for the different layers mentioned in section 7.2. The values of the
correlation coefficients for all the linear trends were higher than 0.70, suggesting the
adequacy of the fit. The trends in Layer 1 are similar but indicated no improvement
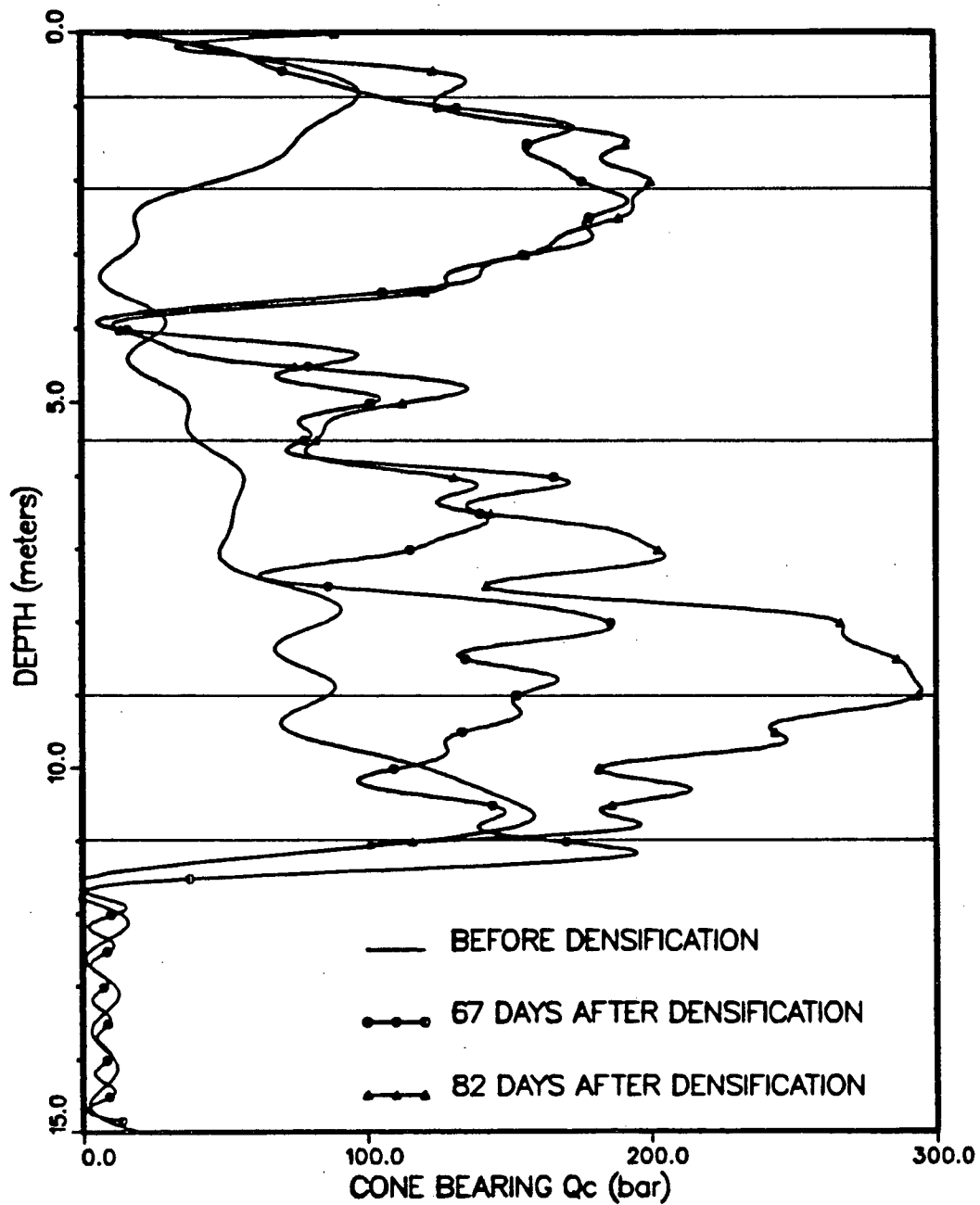
Figure 7.4: Filtered (BS = 1.5) and Fourier Smoothed Profiles of Fig. 7.3.
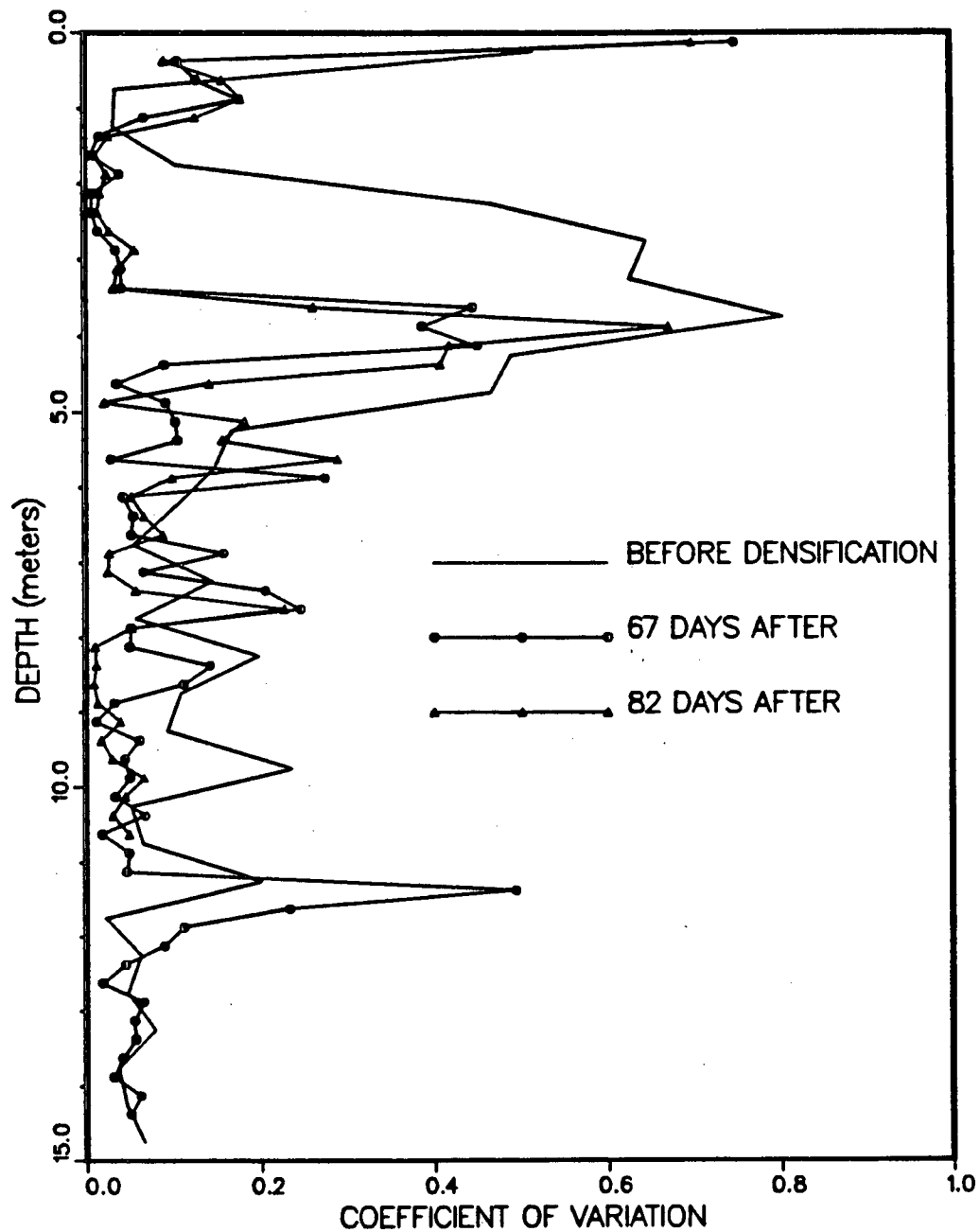
Figure 7.5: Coefficient of Variation Profile of CT1, CT2 and CT3.
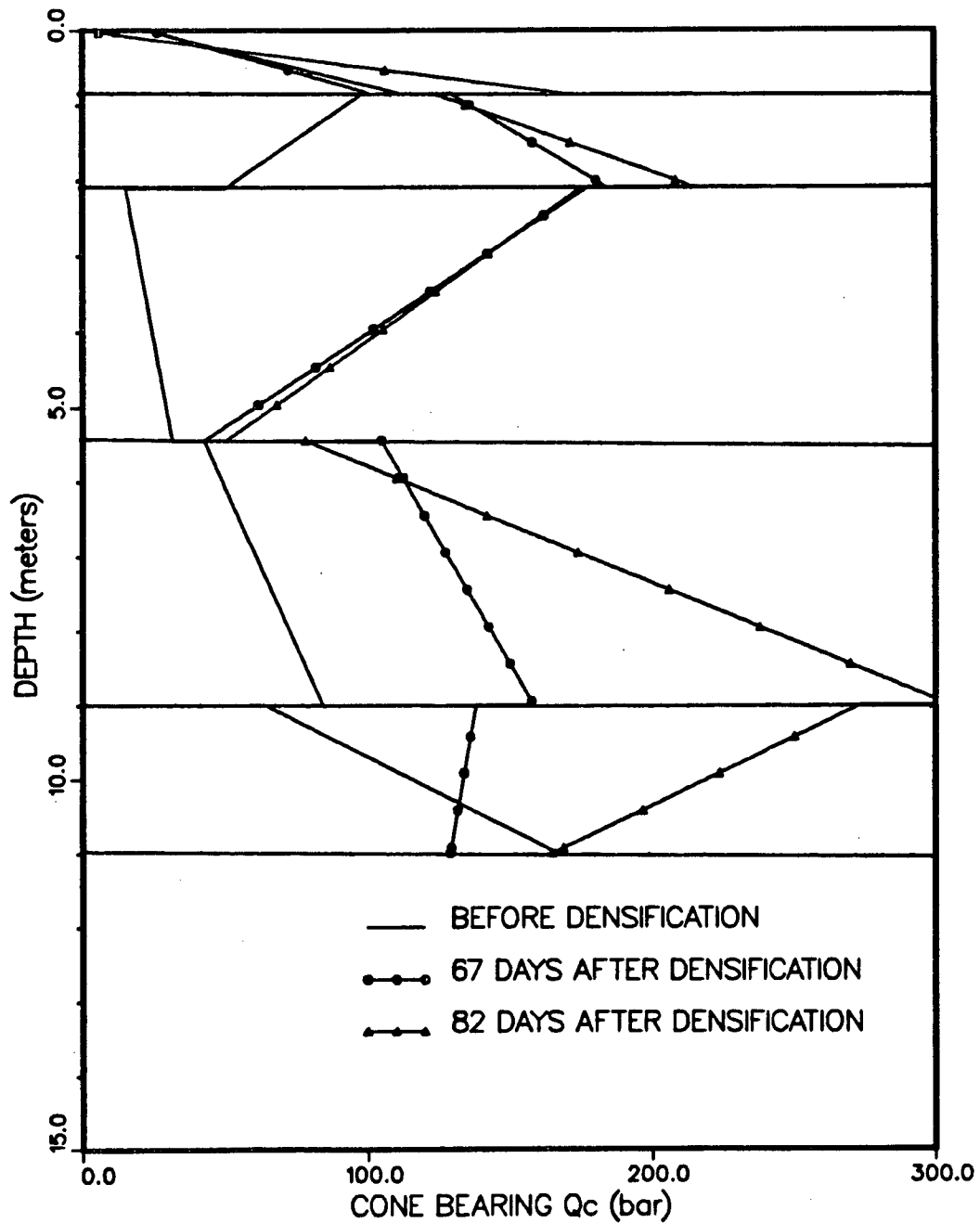
Figure 7.6: Trend Lines of CT1, CT2 and CT3.

with densification. Layer 2 shows a marked improvement with the bearing increasing with depth while Layer 3 reflects a decreasing trend which was significantly higher than the trend prior to densification. In all of the above three layers time effects were not apparent. The main layer of concern (Layer 4) shows a marked increase for both post densification profiles. The 67 day profile (CT2) has improved by approximately 100% throughout the layer while the 82 day profile (CT3) has increased by about 50% at the beginning of the layer (5.45m) and by as much as 300% at the layer end depth (9.0 m). Layer 5 also shows improvement, but with a negative trend. Since the probe tip was at a depth of 10 m, the soil between 9.0 and 11.0 m could be expected to possess effects of soil mixing. A comparison of Fig. 7.6 with the cone bearing profiles in Fig. 7.2 reveals the apparent ease with which the improvement could be judged from the trend lines. The 'RUN' test (section 3.2.3) was performed to determine the non-stationarity of the different layers, and all the layers selected revealed that the cone bearing data was non-stationary with the existence of significant trends.

Methods described in section 3.3.4.1 were used to obtain the confidence estimates of cone bearing (Eq. 3.28) and Fig. 7.7 gives the lower 95% confidence estimate which represents a lower bound where 95% of the cone bearing values lie above. The trend line is a 50% confidence estimate or a mean line where 50% of the data will lie below and 50% will lie above. The 95% confidence estimate is a good value to be used for design purposes. This lower bound of 95% or even a less conservative value of say 90% or 80% can also be used for design. Such lower bound values can also be used in contract specifications for soil densification projects.

The ensuing residuals after trend removal (section 3.3.1) indicated that the variance was approximately constant, justifying the use of simple regression (section 3.3) to identify the trends. The very low correlation of the residuals were verified using the Durbin-Watson statistic described in section 3.3.2. The latter two verifications
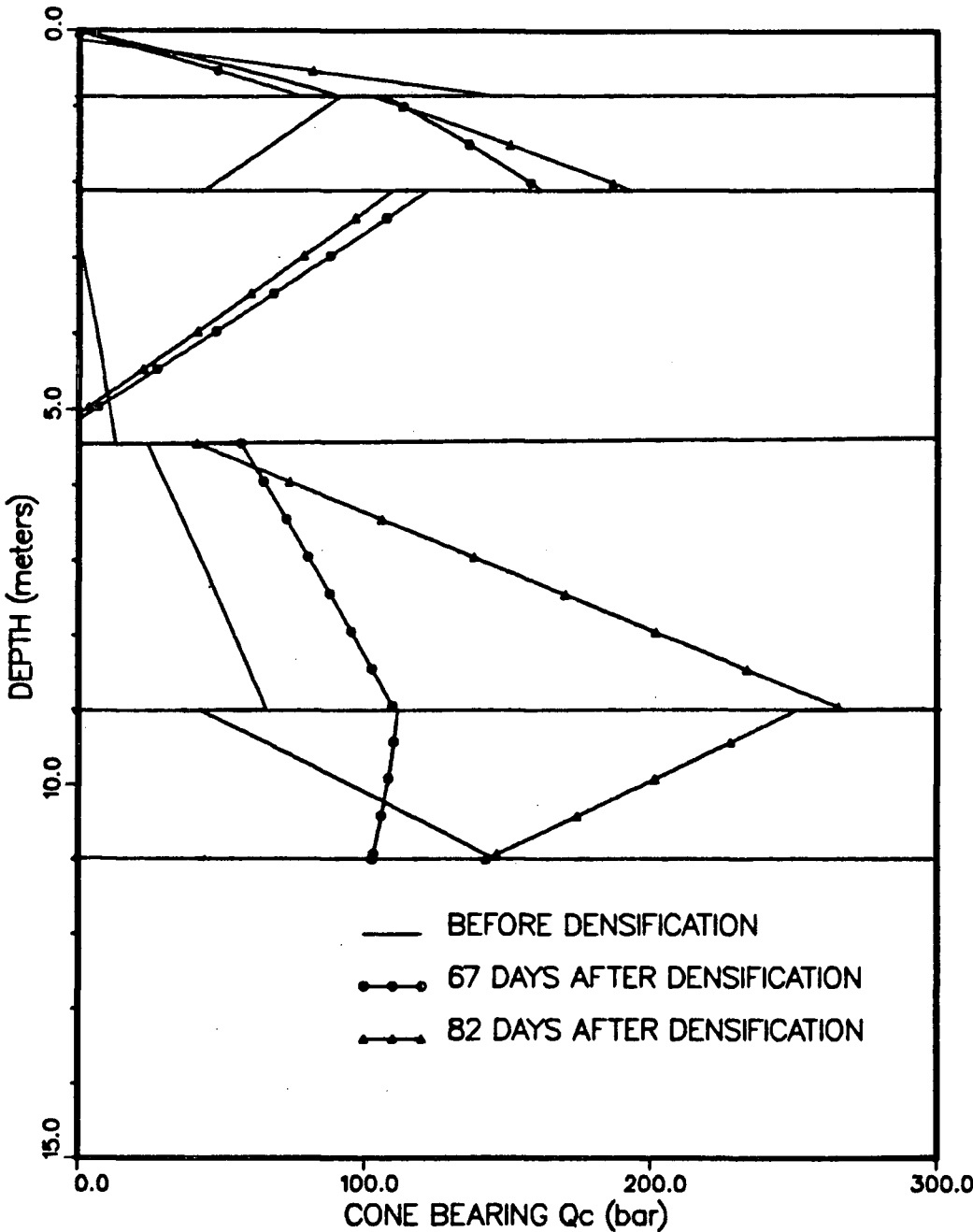
Figure 7.7: 95% Confidence Estimates of Cone Bearing for CT1, CT2 and CT3.

Table 7.2: Scale of Fluctuation for Layer between 5.45 m and 9.00 m for profiles CT1, CT2 and CT3.

CPT Profile	Scale of Fluctuation (cm)
Before Densification (CT1)	21.82

After Densification
 	
    67 Days After Densification (CT2)
    82 Days After Densification (CT3)
	
	33.21
	36.00

suggested the adequacy of simple regression to model the soil property variation.

7.4.2 Scale of Fluctuation

The scale of fluctuation (section 4.3) was used to study the effect of densification on variability as a function of time. Table 7.2 lists the values of the scale of fluctuation for the three different times for the layer between 5.45 m and 9.0 m, since the main purpose of the project was to densify this particular layer. The increase in the scale of fluctuation indicates the reduction in variability. Table 7.2 suggests that densification results in the reduction of variability although time does not seem to have a significant influence as indicated by the marginal increase of the value of the scale of fluctuation from 33.21 cm for the bearing profile 67 days after compaction to 36.00 cm for the profile obtained 82 days after compaction.

The scale of fluctuation is a more reliable estimator of variability, since it takes into account the spatial variability of the soil parameter under consideration.

7.5 Influence of Densification with Distance

Figure 7.8 illustrates the cone bearing profiles before and after densification (82 days after). CT1 and CT3 were along the centerline of densification while the other profiles CD1, CD2, CD3 and CD4 were located 1, 2, 3 and 4 m away from the centerline, respectively. The coefficient of variation profile given in Fig. 7.9 again reflects a decrease in variability after densification and illustrates how this effect was less with increasing distance away from the centerline. This was as expected since the cone profiles beyond 2 m from probe location indicate little, if any, effect of the densification. Once again, it should be pointed out that the evaluation of variability from the coefficient of variation profile was tedious, necessitating the use of the scale of fluctuation as the descriptor of soil variability. Here too the nature of the residuals were such that the use of simple regression to represent the trend was adequate and therefore a similar procedure to that described in section 7.4.1 was used.

7.5.1 Evaluation of Trend and Confidence Estimates

Figure 7.10 shows how the effect of improvement decreases with increasing distances of 1 and 2 m away from the centerline. Although the trends of CD1 (1 m away) and CD2 (2 m away) were higher than that of the profile prior to densification (CT1), they were less than that of profile CT3 which was located along the centerline. The trends of the profiles CD3 (3m away) and CD4 (4m away) shown in Fig. 7.11 were virtually similar to the trend prior to densification, suggesting that the densification pattern and procedure adopted were effective only up to approximately 2 m. The slope reversals of the trend lines of some layers should also be noted. Figure 7.12 illustrates the lower 95% confidence estimates of cone bearing for profiles CT1, CT3, CD1 and CD2 while Fig. 7.13 shows the same estimates of the trends for profiles CT1, CT3, CD3 and CD4. Fig. 7.13 reflects how the confidence estimates of CD3 and CD4
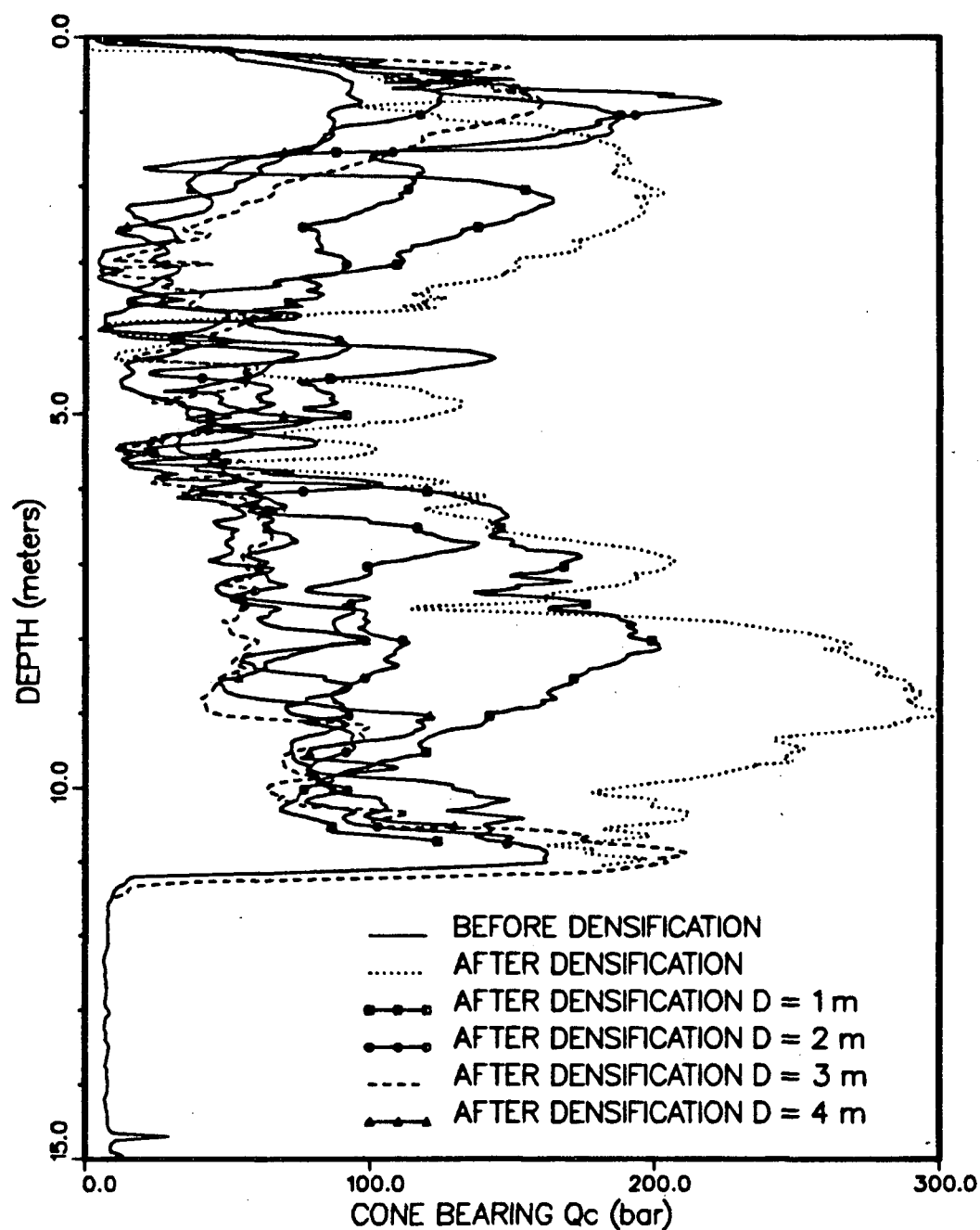
Figure 7.8: Cone Bearing Profiles before Densification (CT1), at Centerline after Densification (CT3) and 1, 2, 3 and 4 m away from Centerline after Densification (CD1, CD2, CD3 and CD4).
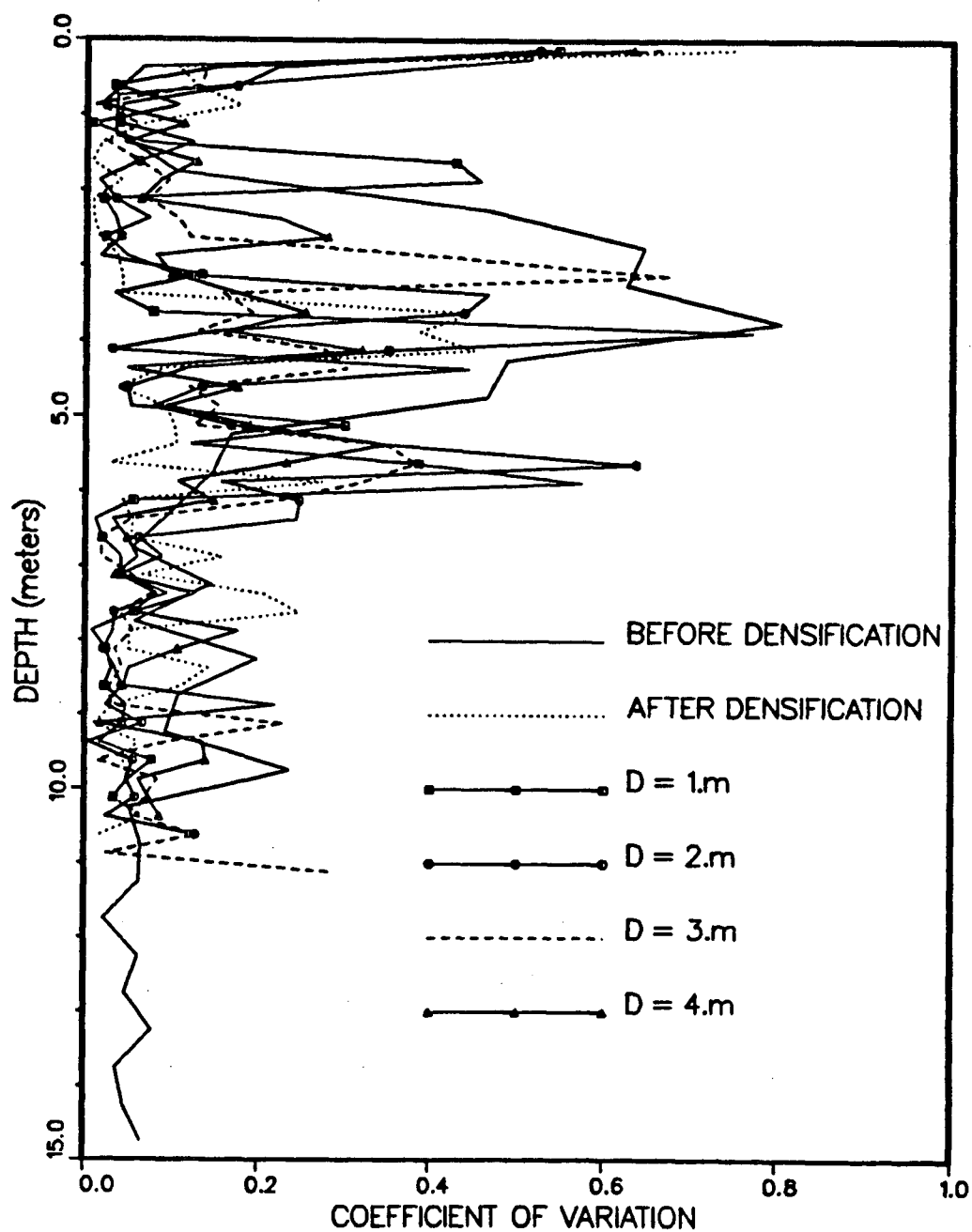
Figure 7.9: Coefficient of Variation Profile of CT1, CT3, CD1, CD2, CD3 and CD4.
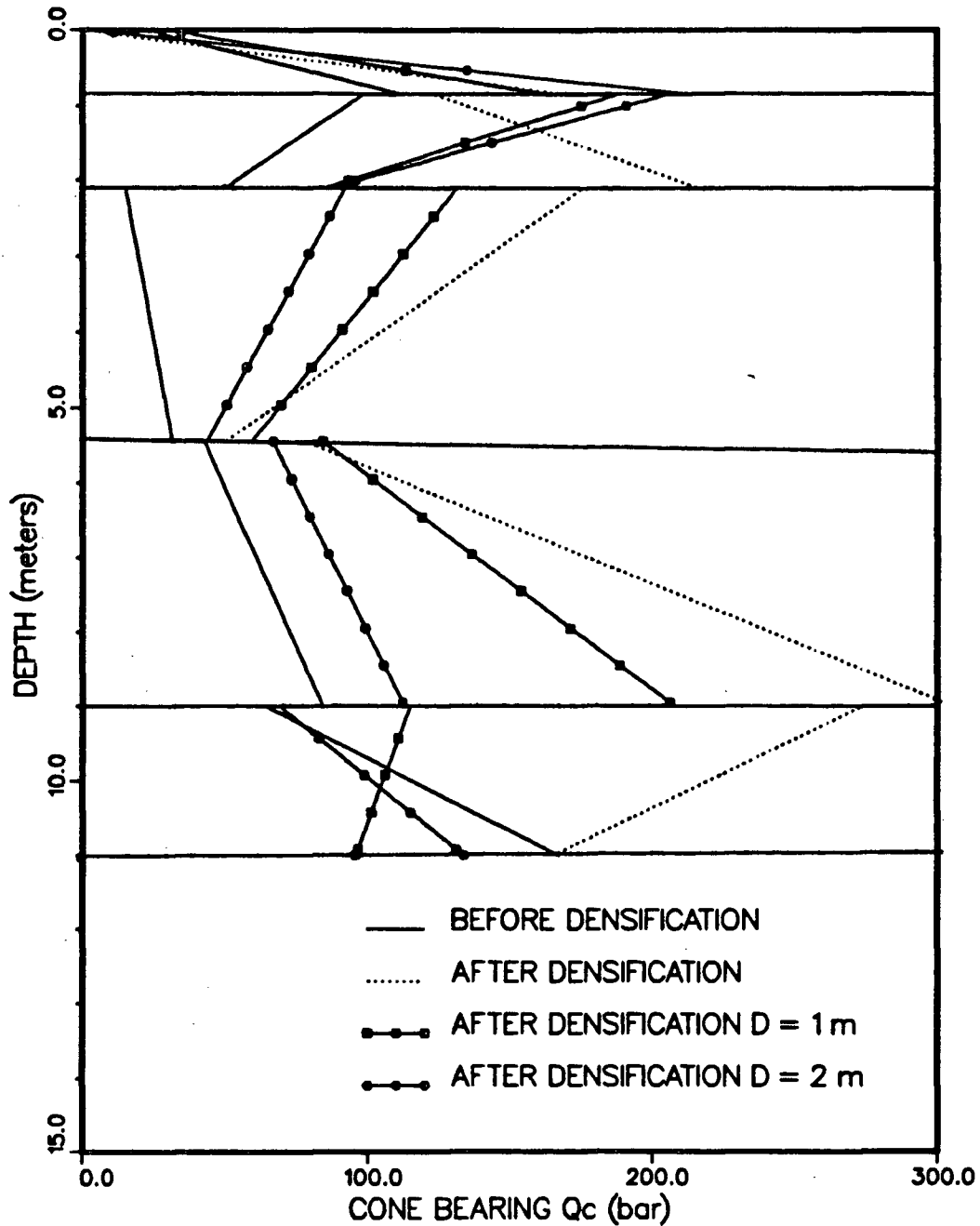
Figure 7.10: Trend Lines Before and After Densification along Centerline (CT1 and CT3) and 1 and 2 m away from Centerline (CD1 and CD2).
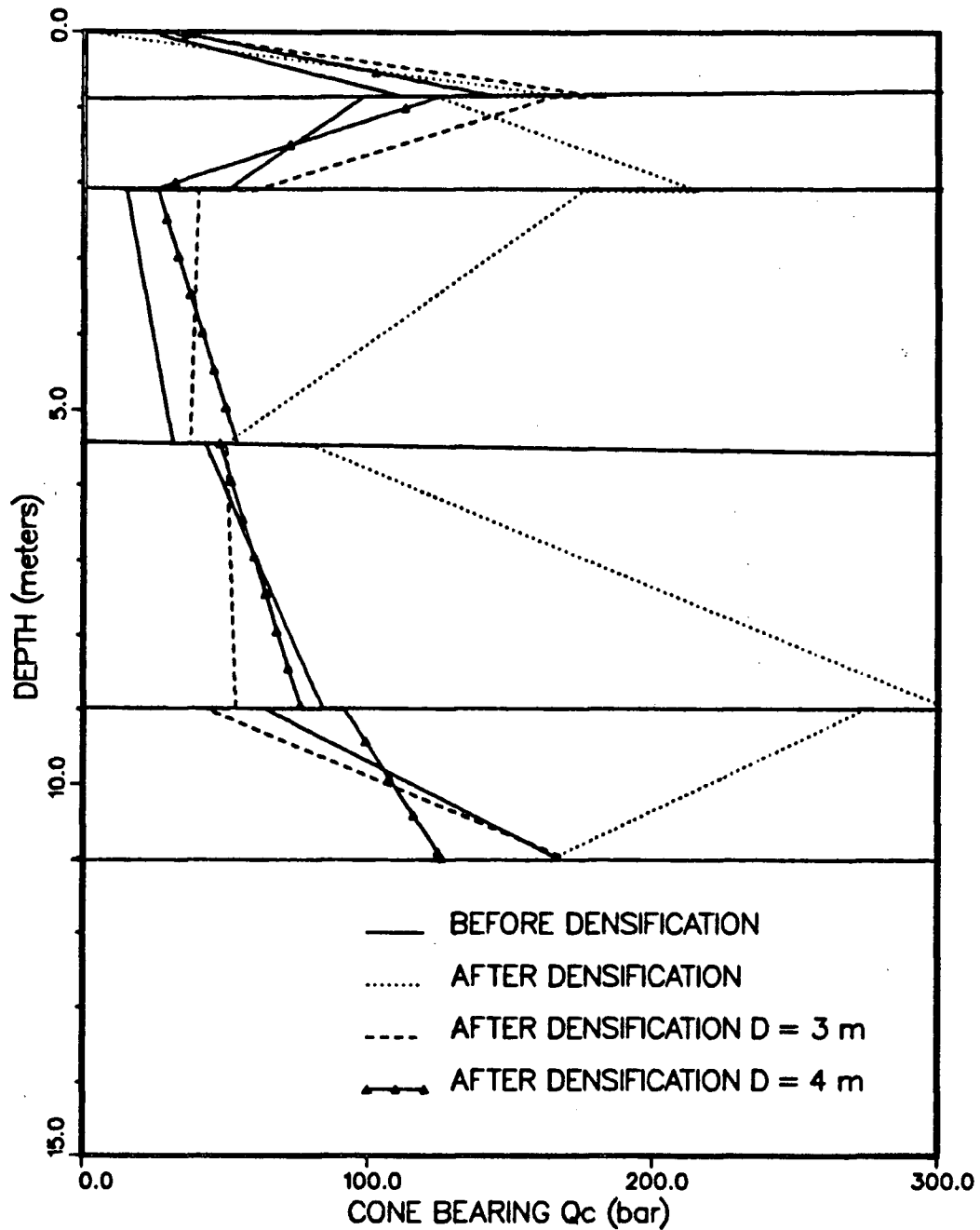
Figure 7.11: Trend Lines Before and After Densification along Centerline (CT1 and CT3) and 3 and 4 m away from Centerline (CD3 and CD4).
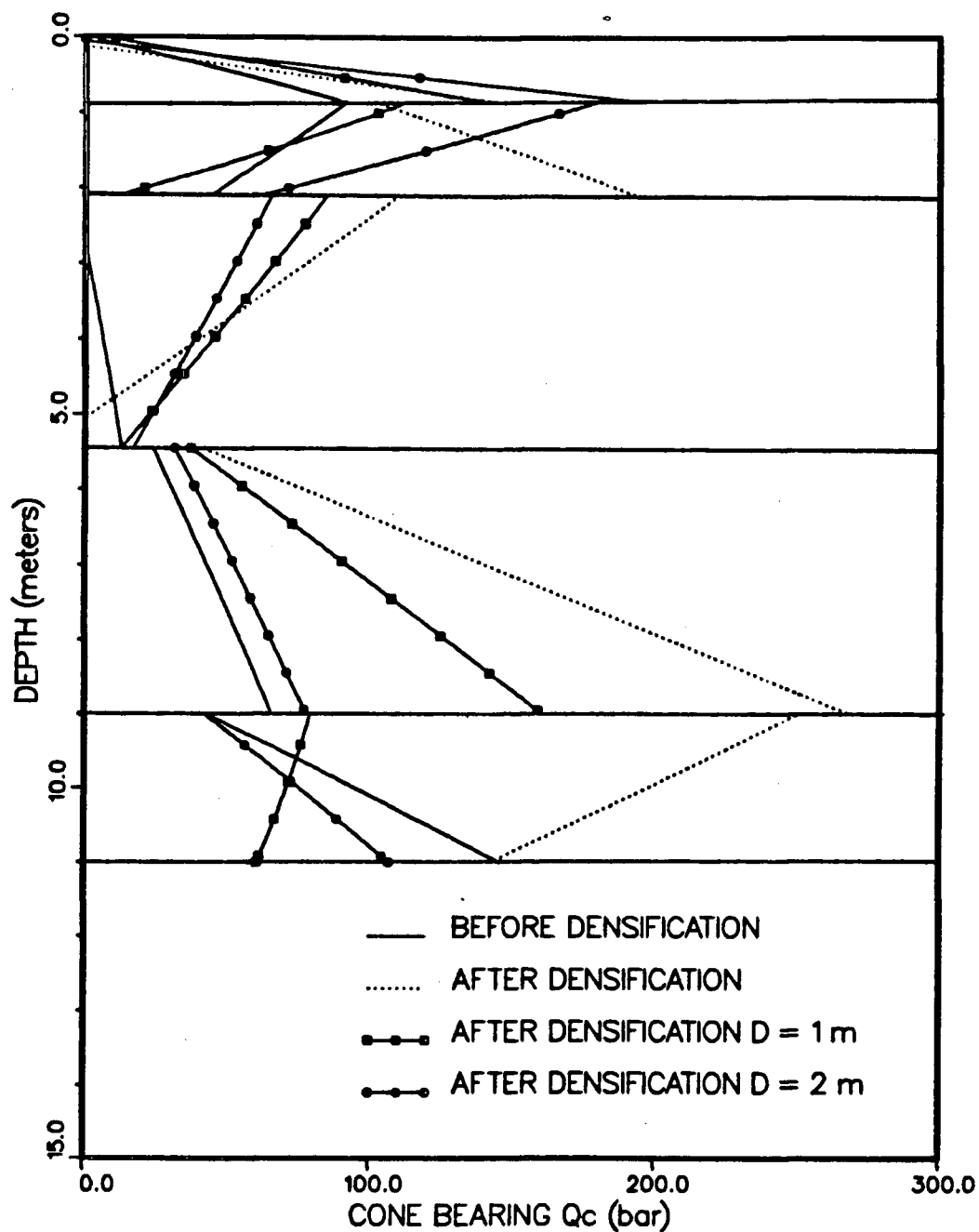
Figure 7.12: 95% Confidence Estimates of Cone Bearing Before and After Densification along Centerline (CT1 and CT3) and 1 and 2 m away from Centerline (CD1 and CD2).
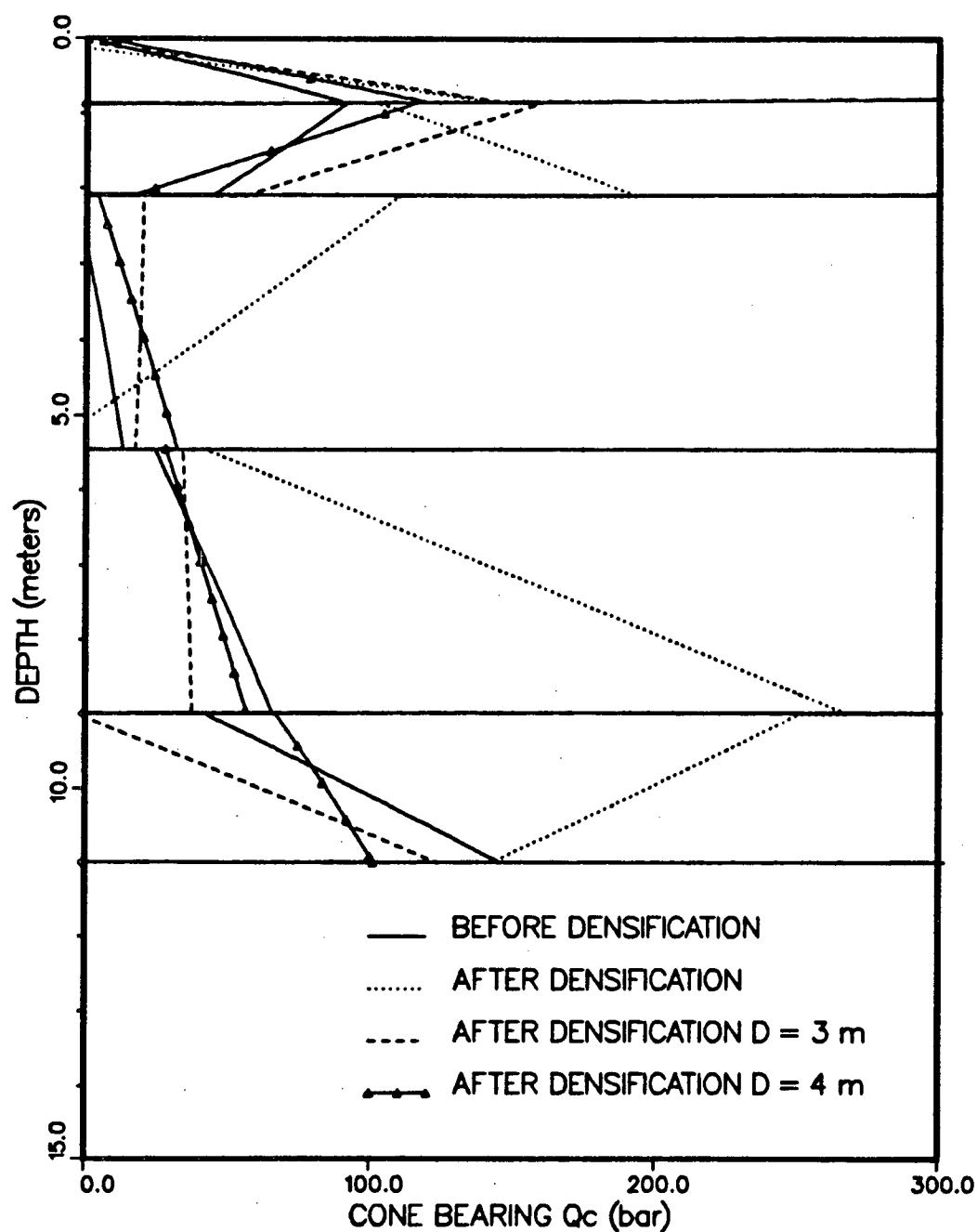
Figure 7.13: 95% Confidence Estimates of Cone Bearing Before and After Densification along Centerline (CT1 and CT3) and 3 and 4 m away from Centerline (CD3 and CD4).

Table 7.3: Scale of Fluctuation for Layer between 5.45 m and 9.00 m for profiles CT1, CT3, CD1, CD2, CD3 and CD4.

CPT Profile	Scale of Fluctuation (cm)
<u>Before Densification</u> (CT1)	21.82
<u>After Densification</u>	
At Centerline of Densification (CT3)	33.21
1 m from Centerline of Densification (CD1)	27.91
2 m from Centerline of Densification (CD2)	25.06
3 m from Centerline of Densification (CD3)	23.29
4 m from Centerline of Densification (CD4)	19.04

approach those of the virgin state (CT1). Here too, this lower bound can be used for design and specification purposes.

The similarity of profiles CT1, CD3 and CD4 also provides justification to the assumption that there was no appreciable inherent soil variability across the relatively close spacings considered in this study.

7.5.2 Scale of Fluctuation

Similar to the investigation on time effects on densification, the scale of fluctuation (section 4.3) was used to study the effect of densification on variability as a function of proximity to the centerline of densification. Values of the scale of fluctuation which were obtained for the layer between 5.45 and 9.0 m are given in Table 7.3.

As observed from Table 7.3, the scale of fluctuation increases (variability decreases) after densification, and decreases towards the value of scale of fluctuation prior to densification. In other words, the variability of the soil at locations 3 and 4

m away from the centerline are similar to that before densification. This is in agreement with the earlier observation that the trends of the bearing profiles 3 and 4 m away from the centerline are approximately equal to those prior to densification.

7.6 Conclusions

The main conclusions which could be drawn from this simple case history are:

(i) Statistical methods of layer identification provide a good tool to identify sublayers present in a soil stratum subjected to densification.

(ii) The improvement in cone bearing was clearly evident from the trend lines of the different layers and was much less tedious than analyzing the effects of densification directly from the cone bearing profiles.

(iii) The lower 95% estimate of bearing provides the geotechnical engineer with a value for design considerations, and this lower bound was conveniently obtained using statistical methods. This lower bound could also be used for compaction control and in contract specifications as opposed to the traditional minimum value.

(iv) The scale of fluctuation has proved to be an efficient indicator of soil uniformity (inverse of variability) in contrast to the coefficient of variation which does not consider the spatial effects of variability.

(v) Statistical methods such as layer identification, trend analysis and the scale of fluctuation have effectively demonstrated their ability to assess variability characteristics of soil profiles. Without the use of these methods, evaluating the amount of improvement caused by densification and ascertaining the effectiveness of the Tri Star probe would be difficult and highly judgemental due to the highly variable and non-uniform nature of the site considered in this case history.

Chapter 8

Summary and Conclusions

The main purpose of this research was to develop and evaluate statistical approaches that could be applied to soil test data with the aim of enhancing the site characterization capabilities of in situ testing devices with special emphasis on the CPT. This thesis has amply demonstrated how statistical methods can be used on large data bases which result from close sample spacing during the cone penetration test. The statistical methods developed can not only provide additional information at a given site, thereby allowing the reduction of uncertainty involved in the estimation of soil properties, but also provide an efficient way to ascertain the variation of soil properties with depth across the site. The ensuing results from such statistical analyses can then be used to supplement the results obtained by other coventional methods commonly used in geotechnical engineering practice.

The cone penetration test has the capability of sampling at very close intervals in the vertical direction and hence provides a detailed description in the depth dimension. Soil strata are highly heterogeneous especially in the vertical direction and the conventional methods of layer identification based on the deterministic CPT classification chart are at times ambiguous due to the subjectivity involved in its use. Statistical methods have therefore been proposed in order to increase the reliability of layer dilineation. Once the different layers in a stratum have been identified methods of trend analysis can be used to obtain a better understanding of soil properties and their variation with depth. Confidence estimates of the soil properties can then be determined for design using reliability approaches. Several applications of

Random Field theory have been extended to provide another means of obtaining an enhanced knowledge of the soil property variation with depth. In this respect the scale of fluctuation provides an extremely efficient basis for ascertaining the variability characteristics of various soil parameters. The amount of variability in a data set is affected by factors other than the inherent variability of the soil which the geotechnical engineer is concerned about. Part of this variability is caused by the measurement error which is sometimes referred to as random noise. This is mainly a result of errors caused by man and machine and could be approximately determined by methods of time series analysis.

In addition to a detailed description of soil properties in the depth dimension, the geotechnical engineer involved in a typical site investigation will naturally be concerned about the variability across the site. All the techniques presented in the thesis can be similarly extended to the horizontal direction, provided a sufficient data base exists. In situations where enough data exist to establish a statistical model horizontally, the proposed two dimensional interpolation procedure considering soil property correlations can be used to obtain estimates and confidence limits of soil property values at untested locations.

Identification of Soil Layers

Different statistical techniques employing univariate and multivariate methods of analysis have been used to identify the soil layers present in a profile. The classical method of identifying soil layers based on the friction ratio is inadequate at times due to the subjectivity involved in the use of the interpretation chart. The statistical methods proposed, have proven to be a good substitute. A multivariate analysis which has the capability of considering the cone bearing, sleeve friction and pore pressure simultaneously, has been shown to be more advantageous than the univariate methods to discriminate between soil layers. Different levels for the values of the Intraclass

Correlation Coefficient and the D^2 statistic enabled boundaries to be classified as primary or secondary for both clay and sand type soils. The location of the primary boundary layers was particularly effective since it was insensitive to the selected window width which was based on the autocorrelation function of the parameters concerned. This provided further evidence to the robustness of the proposed statistical measures of soil layer dilineation. A method based on the gradient of the trend also proved to be successful in identifying layer boundaries in rare situations where the statistical methods were inadequate.

Trend Analysis and Filtering

Trend analysis techniques have been effectively used to describe the characteristics of different layers identified in a stratum. Methods of overcoming difficulties in regression analysis have been explained in detail. Techniques of statistical filtering and smoothing are sometimes required to remove extremeties in data sets in order to establish the trend of the data. Filtering methods must be applied with the utmost caution since the exact statistical parameters selected for the filtering process are highly situation dependent and the possibility of missing out a very thin layer with significantly different characteristics to the layers above and below, should be avoided. The median method of filtering was found to be more advantageous since, unlike the mean, the median is not a function of the extreme data points contained in a selected sublayer. Precaution should also be taken so that the sublayer depths chosen are not so wide as to miss actual layers in the soil stratum. On the other hand, too narrow a sublayer will result in biased statistics rendering the filtering process unreliable. Considering the above two limitations, an optimum width of 25 cm (10 data points) has been recommended for the depth or thickness of the filtering sublayer or window. Very often, the presence of a trend in a profile may not be apparent from a visual inspection. The 'RUN' test has been effectively used in this thesis, in order to verify

the stationarity or non-stationarity of a soil profile. Although the variogram function of a data set can also be used to verify stationarity, the 'RUN' test was more useful in the sense that specific levels of significance can be established for the acceptance or rejection of stationarity of a data set.

Applications of Random Field Theory

The geotechnical engineer is concerned that redundant or excess data is not collected in a site investigation since it costs both time and money. In this regard a simple statistical procedure has been proposed to estimate the optimum sample spacing in a given soil profile. The optimum sample spacing thus obtained was shown to be a function of the variability of the soil layer, the required confidence in the estimate and the degree of tolerance allowed. As the variability of a layer increases a closer spacing of data is required in order to obtain an estimate of a soil parameter at a given confidence and tolerance (or precision).

The natural heterogeneity of the soil, limited data availability and errors caused by man and machine, all contribute to the uncertainty in soil data and has resulted in geotechnical test data being treated as random. The application of random field theory to CPT data has been used to clearly demonstrate how the use of statistics and probability can provide additional information from a given set of data. This results in a less conservative analysis and a greate economy in design. The correlation coefficient between spatial averages and the probability of exceedance have been obtained for different profiles demonstrating its use for site characterization and design purposes. In a classical deterministic analysis these would not be normally considered, with the possible consequences of inaccurate results.

Scale of Fluctuation and Soil Variability

The concept of the scale of fluctuation has been shown to be an excellent indicator

of soil variability in the sense it considers the effects of spatial variability in contrast to the standard coefficient of variability which does not consider spatial variability. The proposed method of calculating the scale of fluctuation compared well with that originally proposed by Vanmarke (1977) and was found to be a more suitable method from the aspect of computational convenience. Several applications described in the thesis reveal the correlation between the variability and the scale of fluctuation and the sensitivity of it to variability. Studies on the scale of fluctuation have effectively demonstrated the averaging characteristics of cone bearing, sleeve friction and pore pressure measured by the CPT. It can be concluded that while the pore pressure is indicative of a measurement made at a point, the cone bearing is indicative of a value averaged over a finite length which is less, but comparable to the averaging distance of the sleeve friction. The scale of fluctuation has also been made use of to obtain an optimum sampling spacing for a given soil layer.

Measurement Noise of Geotechnical Test Data

The application of Time Series analysis to CPT data, illustrated its capability of adequately modeling the stationary component of a soil profile. The random error or the measurement error term obtained for different test methods compared well with that obtained from the autocorrelation function. The five percent random error obtained for the CPT clearly indicated its superiority over other testing methods like the field vane which gave a high value in excess of thirty percent. This reflected the high random error associated with vane testing in contrast to the CPT.

Two Dimensional Correlation Analysis

Soil properties are highly depth dependent. Therefore, any two dimensional interpolation procedure which includes the depth as one of the dimensions should consider

soil property correlation if accurate estimates are required. The need for the consideration of two different types of autocorrelation functions for the representation of two dimensional soil property variation has been highlighted. The two dimensional autocorrelation function which was used for the interpolation yielded satisfactory results, with a good comparison between the predicted and actual profiles. This clearly indicated the need for the consideration of soil property correlations, in the event they do exist, if better estimates are desired.

Case History to Evaluate Densification Effects

A case history concerning site densification was used to show how some of the statistical methods proposed and presented in this thesis can be used to evaluate the effects of soil densification in a more quantitative manner. It has to be emphasized that the techniques used for the case history certainly do not encompass the whole range of applications described in the thesis, but only provide a simple demonstration of how effectively statistical methods can be used to assess soil variability. Methods of layer identification and trend analysis proved to be efficient tools for this purpose and the scale of fluctuation was found to be an ideal tool to assess soil variability. It is clearly evident from this example that in soil profiles of such high non-uniformity, it is impossible to draw reliable conclusions without the use of statistical techniques.

Interactive Micro Computer Programs

Most of the techniques described in the thesis have been performed using IBM-PC compatible interactive micro computer programs which have been developed by the author. These programs are adaptable to different data formats with several options available to the user. Detailed manuals with specific worked examples have also been prepared and is available in the Department of Civil Engineering at UBC.

Scope of Statistical Methods

Finally, this thesis has effectively demonstrated how statistical techniques can be applied to in situ test data to enhance site characterization which, to date, is primarily performed using deterministic approaches. It is the author's belief that, in the light of the sophistication in design and analysis of geotechnical structures, it should not be long before statistical and probabilistic procedures will begin to supplement the deterministic approaches used to-day. Since probabilistic and statistical procedures result in reduced risk and less conservatism due to the additional information gathered from such analyses, the emergence of these methods in geotechnical engineering will be inevitable. In the past the reason for the reluctance in the use of statistical methods in geotechnical data analysis was the lack of an adequate data base. The emergence of in situ testing devices such as the CPT with its capability of sampling at close intervals provides the luxury of larger data bases, paving the way for the use of statistical methods.

Bibliography

[1] **Agterberg, F. P. (1970)** - " Autocorrelation Functions in Geology ", *Geostatistics*, Ed. D. F. Merriam, Plenium Press, NY, USA, pp. 113 - 114.

[2] **Agterberg, F. P. (1974)** - *Developments in Geomathematics* , Elsevier Scientific Publishing Co., NY, USA.

[3] **Alonso E. E. and Krizek, R. J. (1975)** - " Stochastic Formulation of Soil Properties ", *Proceedings of the Second International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP2)*, Aachen, West Germany, Vol. 2, pp. 10 - 32.

[4] **Anderson, L. R., Sharp, K. D., Bowles D. S. and Canfield, R. V. (1984)** - " Application of Methods of Probabilistic Characterization of Soil Profiles ", *Probabilistic Characterization of Soil Properties, Bridge Between Theory and Practice, Proceedings of the ASCE Convention*, Atlanta, Georgia, USA. Eds. D. S Bowles and H. Y. Ko. pp. 90 - 105.

[5] **Armstrong, J. E. (1978)** - " Post Vachon Wisconsin Glaciation, Fraser Lowland, British Columbia ", *Geological Survey of Canada*, Bulletin 332.

[6] **Asaoka, A, and A-Grivas D. (1982)** - " Spatial Variability of the Undrained Strength of Clays ", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 108, No. GT 5, pp. 743 - 757.

[7] **Baecher, G. B. and Ingra, T. (1979)** - " Equivalent Parameters in Soil Profile Modeling ", *Proceedings of the Specialty Conference ASCE ; Probababilistic Mechanics and Structural Reliability*, Tuscon, Arizona, USA, pp. 330 - 337.

[8] **Baecher, G. B., Chan, M., Ingra, T. S., Lee, T. and Nucci, L. R. (1980)** - " Geotechnical Reliability of Offshore Gravity Platforms ", *MITSG 80 - 20 (Research Report)*, MIT Sea Grant College Program, MIT Press, Massachussettes, USA.

[9] **Baecher, G. B. (1981)** - " Optimal Estimators for Soil Properties ", Technical Note, *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 107, No. GT 5, pp. 649 - 653.

[10] **Baecher, G. B. (1982)** - " Statistical Methods in Site Characterization ", *Proceedings of the Engineering Foundation Conference on Updating Subsurface Sampling and Testing for Engineering Purpose*, Santa Barbara, California, USA, pp. 463 - 491.

[11] **Baecher, G. B. (1984)** - " Just A Few More Tests and We'll Be Sure ", *Proceedings of the ASCE Convention ; Probabilistic Characterization of Soil Properties, Bridge Between Theory and Practice*, Atlanta, Georgia, USA. Eds. D. S Bowles and H. Y. Ko. pp. 1 - 18.

[12] **Baecher, G. B. (1984)** - " Simplified Geotechnical Data Analysis ", *Reliability Theory and Its Application in Structural and Soil Mechanics*, Ed. P. Throft Christenson, Martinus Nijhoff Publishers, Hague, Netherlands. pp. 257 - 277.

[13] **Baecher, G. B. (1985)** - " Geotechnical Error Analysis ", *Recent Developments in Measurement and Modeling of Clay Behaviour*, MIT Special Summer Course, MIT, Massachussettes, USA.

[14] **Bendat, J. S. and Piersol A. G. (1971)** - *Random Data Analysis and Measurement Procedures*, Wiley Intersience Publishers, NY, USA.

[15] **Biernatowski, K. (1985)** - " Statistical Characteristics of the Subsoil ", *Proceedings of the Eleventh International Conference on Soil Mechanics and Foundation Engineering (ICSMFE)*, San Fransisco, CA, USA, Vol 2, pp. 799 - 802.

[16] **Box, G. E. and Jenkins, G. M. (1976)** - *Time Series Analysis*, Holden - Day Publishers, San Francisco, CA, USA.

[17] **Brooke, R. J. and Arnold, G. C. (1985)** - *Applied Regression Analysis and Experimental Design* Marcel Decker Inc., NY, USA.

[18] **Bury, K. V. (1975)** - " Statistical Models in Applied Science ", John Wiley and Sons, NY, USA.

[19] **Campanella R. G., Robertson, P. K. and Gillespie, D. J. (1983)** - " Cone Penetration Testing in Deltaic Soils ", *Canadian Geotechnical Journal* , Vol. 20, No. 1, pp. 23 - 35.

[20] **Campanella, R. G. and Wickremesinghe, D. S. (1987)** - " Statistical Treatment of Cone Penetrometer Test Data ", *Proceedings of the Fifth International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP5)*, Vancouver, Canada, Vol. 2, pp. 1011 - 1019.

[21] **Campanella, R. G., Wickremesinghe, D. S. and Echezuria, H. J. (1989)** - " Cone Penetration Test for Site Characterization ", *Twelfth International Conference on Soil Mechanics and Foundation Engineering (ICSMFE)*, Rio De Jeneiro, Brazil (to be published).

[22] **Campanella, R. G. and Robertson, P. K. (1988)** - Special Lecture, " Current Status of the Piezocone Test ", *Proceedings of the First International Symposium on Penetration Testing (ISOPT1)*, Orlando, Florida, USA, Vol. 1, pp. 93 - 116.

[23] **Cheong, H. F. and Subrahmanyam, R. V. (1980)** - " Statistical Analysis of Marine Clay Deposits ", Technical Note, *Journal of the Geotechnical Engineering Division* , ASCE, Vol. 107, No. GT2, pp. 221 - 229.

[24] **Christakos, G. (1985)** - " Recursive Parameter Estimation with Applications in Earth Sciences ", *Journal of Mathematical Geology*, Vol. 17, No. 5, pp. 489 - 515.

[25] **Christakos, G. (1985)** - " Modern Statistical Analysis and Optimal Estimation of Geotechnical Data ", *Journal of Engineering Geology*, Vol 22, pp. 175 - 200.

[26] **Christakos, G. (1987)** - " A Stochastic Approach in Modeling and Estimating Geotechnical Data ", *International Journal for Numerical and Analytical Methods in Geomechanics*, Vol. 11, pp. 79 - 102.

[27] **Cressie, N. (1985)** - " Fitting Variogram Models by Weighted Least Squares ", *Journal of Mathematical Geology*, Vol. 17, No. 5, pp. 563 - 586.

[28] **David, M. (1976)** - " The Practice of Kriging ", *Advanced Geostatistics in the Mining Industry*, Proceedings of the NATO Advanced Study Institute, Rome, Italy. Eds. M. Guarscio et al., D. Reide Publishing Co., Holland. pp. 31 - 48.

[29] **David, M (1977)** - " Geostatistical Ore Reserve Estimation ", *Developments in Geomathematics 2*, Elesevier Scientific Publishing Co., NY, USA.

[30] **Davis, J. C. (1978)** - *Statistics and Data Analysis in Geology*, John Wiley and Sons, NY, USA.

[31] **Davis, M. W. D. and David, M. (1978)** - " Automatic Kriging and Contouring in the Presence of Trends (Universal Kriging Made Simple) ", *Journal of Canadian Petroleum Technology*, Vol. 17, No. 1, pp. 90 - 98.

[32] **Delfiner, P. (1976)** - " Linear Estimation of Non - Stationary Spatial Phenomena ", *Advanced Geostatistics in the Mining Industry*, Proceedings of the NATO Advanced Study Institute, Rome, Italy. Eds. M. Guarscio et al., D. Reide Publishing Co., Holland. pp. 49 - 68.

[33] **Delfiner, P. and Delhomme, J. P. (1973)** - " Optimum Interpolation by Kriging ", *Proceedings of NATO Advanced Study Institute for Display and Analysis of Spatial Data*, Eds. J. C. Davis and M. J. McGullagh, John Wiley and Sons, NY, USA, pp. 96 - 114.

[34] **Draper, N. D. and Smith, H. (1966)** - *Applied Regression Analysis*, John Wiley and Sons, NY, USA.

[35] **Durbin, J. and Watson, G. S. (1951)** - " Testing for Serial Correlation in Least Squares Regression II ", *Biometika*, Vol. 38, pp. 159 - 178.

[36] **Gambolati, G. and Volpi, G. (1972)** - " Groundwater Contour Mapping in Venice by Stochastic Interpolators : Theory ", *Water Resources Research Journal*, Vol. 15, No. 2, pp. 281 - 289.

[37] **Gottman, J. M. (1981)** - *Time Series Analysis* , Wiley and Sons, NY, USA.

[38] **Graybill, F. A. (1961)** - *An Introduction to Linear Statistical Models, Vol. 1* , McGraw Hill Publishers, NY, USA.

[39] **Greig, J. W. (1985)** - *Estimating Undrained Shear Strength of Clay from Cone Penetration Tests*, MASc. Thesis, University of British Columbia, Vancouver, Canada.

[40] **Haldar, A. (1981)** - " Statistical and Probabilistic Methods in Geomechanics ", *Proceedings of the NATO Advanced Study Institute*, University of Minho, Braga, Portugal, pp. 473 - 504.

[41] **Handy, R. L., Remmes B., Moltd S., Luteneggar A. J. and Trott G. (1982)** - " In Situ Stress Determination by Iowa Stepped Blade ", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 108, No. GT 11, pp. 1405 - 1422.

[42] **Harbaugh, J. W., and Merriam, D. F. (1968)** - *Computer Applications in Stratigraphic Analysis*, John Wiley and Sons, NY, USA.

[43] **Hawkins, D. M. and Merriam, D. F. (1973)** - " Optimal Zonation of Digitized Sequential Data ", *Journal of Mathematical Geology*, Vol. 5, No. 4, pp. 389 - 395.

[44] **Hawkins, D. M. and Merriam, D. F. (1974)** - " Zonation of Multivariate Sequences of Digitized Data ", *Journal of Mathematical Geology*, Vol. 6, No. 3, pp. 263 - 269.

[45] **Jenkins, G. M. and Watts, D. G. (1968)** - *Spectral Analysis and Its Applications*, Holden - Day Publishers, San Francisco, CA, USA.

[46] **Johannesson, L. E. (1985)** - " Statistical Analysis of Soundings and Test Results from a Silty Clay ", *Proceedings of the Eleventh International Conference on Soil Mechanics and Foundation Engineering (ICSMFE)*, San Fransisco, CA, USA, Vol 2, pp. 871 - 874.

[47] **Jamalowski, M., Ladd, C. C., Germaine, J. T. and Lancelotta, R. (1985)** - " New Developments in Field and Laboratory Testing of Soils ", State of the Art Paper, *Proceedings of the Eleventh International Conference of Soil Mechanics and Foundation Engineering (ICSMFE)*, San Francisco, CA, USA, Vol. 1, pp. 57 - 153.

[48] **Johnson R. A. and Wichern D. W. (1982),** - *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey, USA.

[49] **Journel, A. G. and Huijbregts, C. J. (1978)** - *Mining Geostatistics*, Academic Press, NY, USA.

[50] **Kay, J. N. and Krizek, R. J. (1971),** - " Estimation of the Mean for Soil Properties ", *Proceedings of the First International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP1)*, Hong Kong, Vol 1, pp. 279 - 286.

[51] **Krahn, J. and Fredlund, D. G. (1983)** - " Variability in the Engineering Properties of Natural Soil Deposits ", *Proceedings of the Fourth International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP4)*, Florence, Italy, Vol 2, pp. 1018 - 1029.

[52] **Kraus, K. and Mikhail, E. M. (1972)** - " Linear Least Squares Interpolation ", *Proceedings of the Twelfth Congress of the International Society of Photogammetry*, Ottawa, Canada, pp. 1016 - 1029.

[53] **Kreyszig, E. (1983)** - *Advanced Engineering Mathematics*, John Wiley and Sons, NY, USA.

[54] **Kulatilake, P. H. S. W. and Ghosh, A. (1988)** - " An Investigation into Accuracy of Spatial Variation Estimation Using Static Cone Penetrometer Data ", *Proceedings of the Conference on Penetration Testing (ISOPT - 1)*, Florida, USA, pp. 815 - 821.

[55] **Kulatilake P. H. S. W. and Miller, K. M. (1987)** - " A Scheme for Estimating the Spatial Variation of Soil Properties in Three Dimensions ", *Proceedings of the Fifth International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP5)*, Vancouver, Canada, Vol. 2, pp. 669 - 677.

[56] **Lumb, P. (1966)** - " The Variability of Natural Soils ", *Canadian Geotechnical Journal*, Vol. 3, No. 2, pp. 74 - 97.

[57] **Lumb, P. (1967)** - " Statistical Methods in Soil Investigations ", *Proceedings of the Fifth Australian - New Zealand Conference in Soil Mechanics and Foundation Engineering*, pp. 26 - 33.

[58] **Lumb, P. (1970)** - " Safety Factors and the Probability Distribution of Soil Strength ", *Canadian Geotechnical Journal*, Vol. 7, No. 3, pp. 225 - 242.

[59] **Lumb, P. (1974)** - " Application of Statistics in Soil Mechanics ", *Soil Mechanics - New Horizons*, Ed. I. K. Lee, Newnes - Butterworth, London, England, pp. 44 - 111.

[60] **Lumb, P. (1975)** - " Statistical Estimation in Soil Engineering ", *Proceedings of the Second International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP2)*, Aachen, Germany, General Report, Session 4, pp. 156 - 181.

[61] **Lumb, P. (1975)** - " Spatial Variability of Soil Properties ", *Proceedings of the Second International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP2)*, Aachen, Germany, Vol. 2, pp. 397 - 421.

[62] **Martin, R. L. (1974)** - " On Spatial Dependence, Bias and the Use of First Spatial Differences in Regression Analysis ", *Area*, Vol. 6, pp. 185 - 195.

[63] **Massarsch R. and Vanneste G. (1988)** - " Tri Star Vibro - Compaction, Annacis Island, B. C., Canada ", *Franki International Technology (FIT) Internal Report* .

[64] **Matheron, G. (1963)** - " Principles of Geostatistics ", *Journal of Economic Geology*, Vol. 58, pp. 1246 - 1266.

[65] **Matheron, G. (1967)** - " Kriging or Polynomial Interpolation ", *Journal of the Canadian Institute of Mining and Metallurgy (CIMM)*, Vol. 70, pp. 240 - 244.

[66] **Matheron, G. (1970)** - " Random Functions and their Applications in Geology ", *Geostatistics*, Ed. D. F. Merriam, Plenum Press, NY, USA, pp. 79 - 87.

[67] **Myers, R. H. (1986)** - *Classical and Modern Regression Analysis with Applications*, Duxberg Press, Massachussettes, USA.

[68] **Mitchell, J. K. and Villet, W. C. B. (1978)** - " The Measurement of Soil Properties Insitu ", *Report Prepared for the US Department of Energy, Contract W-7405-ENG-48*, University of California, Berkely, CA, USA.

[69] **Mitchell, J. K. and Solymar Z. V. (1984)** - " Time Dependent Strength Gain in Freshly Deposited or Densified Sand ", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 110, No. GT 11, pp. 1559 - 1576.

[70] **Rao, C. R. (1952)** - *Advanced Statistical Methods in Boimetric Research*, John Wiley and Sons, NY, USA.

[71] **Rao, C. R. (1965)** - *Linear Statistical Inference and Its Applications*, John Wiley and Sons, NY, USA.

[72] **Rice, S. O. (1944)** - " Mathematical Analysis of Random Noise ", *Bell System Technical Journal*, Vol. 24, pp. 282 - 332.

[73] **Rizkallah, V., Kramer, H. and Maschuwitz, G. (1979)** - " Comparison Between Results of Penetration Tests with the Static Penetrometer and the Heavy Dynamic Penetrometer ", *Proceedings of the Third International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP3)*, Sydney, Australia, Vol. 1, pp. 212 - 220.

[74] **Rizkallah, V. and Nimr, A. E. (1975)** - " Applicability of Regression Analysis in Soil Mechanics with the Help of Data Banks ", *Proceedings of the Second International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP2)*, Aachen, Germany, Vol. 1, pp. 423 - 438.

[75] **Robertson, P. K. (1982)** - *In Situ Testing of Soil with Emphasis on Its Application to Liquefaction Assessment*, Doctoral Dissertation, University of British Columbia, Vancouver, Canada.

[76] **Robertson, P. K. and Campanella, R. G. (1986)** - *Guidelines for Use, Interpretation and Application of the CPT and CPTU*, Soil Mechanics Series, No. 105, Department of Civil Engineering, University of British Columbia, Vancouver, Canada.

[77] **Russo, D. (1984)** - " Design of an Optimal Sampling Network for Estimating the Variogram ", *Journal of the Soil Science Society of America*, Vol. 48, pp. 708 - 716.

[78] **Sabourin, R. (1976)** - " Application of Two Methods for the Interpretation of the Underlying Variogram ", *Advanced Geostatistics in the Mining Industry*, Eds. M. Guarascio et al., pp. 101 - 109.

[79] **SASLETS User's Guide (1982)** - *Econometric and Time Series Library*, Statistical Analysis Systems Institute, Cary, NC, USA.

[80] **Soulie, M. (1984)** - " Geostatistical Applications in Geotechnics ", *Geostatistics for Natural Resources Characterization*, Part 2, Eds. G. Verly et al., pp. 703 - 730.

[81] **Starks, T. H. and Fang, J. H. (1982)** - " The Effect of Drift on the Experimental Semi - Variogram ", *Journal of Mathematical Geology*, Vol. 14, No. 4, pp. 309 - 319.

[82] **Swed, F. S. and Eisenhart, C. (1942)** - " Tables for Testing Randomness of Grouping in A Sequence of Alternatives ", *Annals of Mathematical Statistics*, Vol. 14, pp. 66 - 87.

[83] **Tabba, M. M. and Yong, R. N. (1979)** - " Statistical Analysis of Geotechnical Records ", *Proceedings of the Third Engineering Mechanics Specialty Conference, ASCE*, pp. 331 - 334.

[84] **Tabba, M. M. and Yong, R. N. (1981a)** - " Mapping and Predicting Soil Properties : Theory ", *Journal of the Engineering Mechanics Division, ASCE*, Vol. 107, No. EMD 5, pp. 773 - 793.

[85] **Tabba, M. M. and Yong, R. N. (1981b)** - " Mapping and Predicting Soil Properties : Applications ", *Journal of the Engineering Mechanics Division, ASCE*, Vol. 107, No. EMD 5, pp. 795 - 811.

[86] **Tang, W. H. (1971)** - " A Bayesian Evaluation of Information for Foundation Engineering Design ", *Proceedings of the First International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP1)*, Hong Kong, Vol. 1, pp. 174 - 185.

[87] **Tang, W. H. (1979)** - " Probabilistic Evaluation of Penetration Resistances ", *Journal of the Geotechnical Engineering Division, ASCE*, Vol. 105, No. GTD 10, pp. 1173 - 1191.

[88] **Tang, W. H., Michols, K. A. and Kjekstad, O. (1984)** - " Probabilistic Stability Analysis of Gravity Platforms ", *Norwegian Geotechnical Institute (NGI) Publication*, No. 152.

[89] **Tang, W. H. and Sidi, I. (1984)** - " Random Field Model of A Two State Medium ", *Proceedings of the Fourth ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability*, Berkely, CA, USA, Ed. Y. K. Wen, pp. 210 - 214.

[90] **Tang, W. H. (1984)** - " Principles of Probabilistic Characterization of Soil Properties ", *Probabilistic Characterization of Soil Properties, Bridge Between Theory and Practice, Proceedings of the ASCE Convention*, Atlanta, Georgia, USA. Eds. D. S Bowles and H. Y. Ko. pp. 74 - 89.

[91] **Tang, W. H. and Sidi, I. (1987)** - " Average Property in a Random Two State Medium", *Private Communication.*

[92] **Tang, W. H. (1988)** - *Private Communication.*

[93] **Telford, W. M., Geldart, R. E., Sheriff, R. E. and Keys, D. A. (1976)** - *Applied Geophysics*, Cambridge University Press, Cambridge, England.

[94] **Vanmarke, E. H. (1975)** - " On the Distribution of the First Passage Time for Normal Stationary Random Processes ", *Journal of Applied Mechanics, ASCE Transactions*, Vol. 42, pp. 215 - 220.

[95] **Vanmarke, E. H. (1977)** - " Probabilistic Modeling of Soil Profiles ", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 103, No. GT 11, pp. 1227 - 1246.

[96] **Vanmarke, E. H. (1977)** - " Reliability of Earth Slopes ", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 103, No. GT 11, pp. 1247 - 1265.

[97] **Vanmarke, E. H. (1978)** - " Probabilistic Characterization of Soil Profiles ", *Proceedings of the ASCE Specialty Conference on Site Characterization and Exploration*, Northwestern University, Illinois, USA, pp. 199 - 219.

[98] **Vanmarke, E. H. (1978)** - *Averages and Extremes of Random Fields*, Research Report R79 - 43, Department of Civil Engineering, MIT, Massachussettes, USA.

[99] **Vanmarke, E. H. (1983)** - *Random Fields : Analysis and Synthesis*, MIT Press, Massachussettes, USA.

[100] **Vanmarke, E. H. (1988)** - *Private Communication.*

[101] **Vaessen, W. (1984)** - " UBC NLP - Nonlinear Function Optimization ", Published by UBC Computing Centre.

[102] **Veneziano, D. and Faccioli, E. (1975)** - " Bayesian Design of Optimal Experiments for the Estimation of Soil Properties ", *Proceedings of the Second International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP2)*, Aachen, Germany. Vol. 2, pp. 191 - 214.

[103] **Veneziano, D. (1979)** - " Statistical Estimation and Data Collection : A Review of Procedures for Civil Engineers ", *Proceedings of the Third International Conference on the Application of Statistics and Probability in Soil and Structural Engineering (ICASP3)*, Sydney, Australia, Vol. 1, pp. 247 - 262.

[104] **Vita, C. (1984)** - " Route Geotechnical Characterization and Analysis ", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 110, No. GT 12, pp. 1715 - 1734.

[105] **Vita, C. (1984)** - " Landform Based Characterization of Soil Properties ", *Probabilistic Characterization of Soil Properties, Bridge Between Theory and Practice, Proceedings of the ASCE Convention*, Atlanta, Georgia, USA. Eds. D. S Bowles and H. Y. Ko. pp. 170 - 182.

[106] **Vivatrat, V. (1978)** - *Cone Penetration in Clays*, Doctoral Dissertation, MIT, Massachussettes, USA.

[107] **Warrick, A. W. and Myers, D. E. (1987)** - " Optimization of Sampling Locations for Variogram Calculations ", *Water Resources Research Journal*, Vol. 23, No. 3, pp. 496 - 500.

[108] **Webster, R. and Beckett, P. H. T. (1968)** - " Quality and Usefulness of Soil Maps ", *Nature*, Vol. 219, pp. 680 - 682.

[109] **Webster, R. and Wong, I. F. T. (1968)** - " A Numerical Procedure for Testing Soil Properties Interpreted from Air Photographs ", *Photogrammetria Journal*, Elsevier Publishing Co., Netherlands, pp. 59 - 72.

[110] **Webster, R. (1973)** - " Automatic Soil Boundary Location from Transect Data ", *Journal of Mathematical Geology*, Vol. 5, No. 1, pp. 27 - 37.

[111] **Webster, R. (1978)** - " Optimally Partitioning Soil Transects ", *Journal of Soil Science*, Vol. 29, pp. 388 - 402.

[112] **Webster, R. and Burgess, T. M. (1980)** - " Optimal Interpolation and Isarthmic Mapping of Soil Properties III : Changing Drift and Universal Kriging ", *Journal of Soil Science*, Vol. 31, pp. 505 - 524.

[113] **Webster, R. (1985)** - " Quantitative Spatial Analysis of Soil in the Field ", *Advances in Soil Science*, Vol. 3, Ed. B. A. Stewart, Springer - Verlag Publishers, NY, USA.

[114] **Wu, T. H. (1973)** - " Uncertainty, Safety and Decision in Soil Engineering ", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 100, No. GT 3, pp. 329 - 348.

[115] **Wu, T. H. and Wong, K. (1981)** - " Probabilistic Soil Exploration : Case History ", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 107, No. GT 12, pp. 1693 - 1711.

[116] **Wu, T.H. and El - Jandali, A. (1985)** - " Use of Time Series Methods in Geotechnical Data Analysis ", *Geotechnical Testing Journal*, ASTM, Vol. 8, No. 4, pp. 151 - 158.

[117] **Wu, T. H., Potter, J. C. and Kjekstad, O. (1986)** - " Probabilistic Analysis of Offshore Site Exploration", *Journal of the Geotechnical Engineering Div., ASCE*, Vol. 112, No. GT 11, pp. 981 - 1000.

[118] **Wu, T. H. (1986)** - " Reliability Analysis of Foundation Stability for Gravity Platforms in the North Sea ", *Proceedings of the Third Canadian Conference of Marine Geotechnical Engineering*, pp. 165 - 180.

[119] **Yong, R. N. (1984)** - " Probabilistic Nature of Soil Properties ", *Probabilistic Characterization of Soil Properties, Bridge Between Theory and Practice, Proceedings of the ASCE Convention*, Atlanta, Georgia, USA. Eds. D. S Bowles and H. Y. Ko. pp. 19 - 55.

Appendix A

Correlation Between Spatial Averages

Let y_o, y_1, y_2, y_3, y_a and y_b be the distances illustrated in Fig.4.10. The shaded areas A and B can be expressed by the integrals I_a and I_b, where,

$$I_a = \int_{y-y_a/2}^{y+y_a/2} Q(y)dy \tag{A.1}$$

$$I_b = \int_{y-y_b/2}^{y+y_b/2} Q(y)dy \tag{A.2}$$

Similarly, the areas corresponding to y_o, y_1, y_2 and y_3 which are I_o, I_1, I_2 and I_3 respectively, can be so determined.

The following relationships follow from above and Fig.4.10

$$I_1 = I_a + I_o \tag{A.3}$$

$$I_2 = I_a + I_b + I_o \tag{A.4}$$

$$I_3 = I_b + I_o \tag{A.5}$$

Evaluating $I_1^2 - I_2^2 + I_3^2$, results in the following equation.

$$2I_aI_b = I_o{}^2 - I_1{}^2 + I_2{}^2 - I_3{}^2 \tag{A.6}$$

The local spatial averages of the segments y_a and y_b are Q_a and Q_b respectively,

$$Q_a = I_a/y_a \tag{A.7}$$

$$Q_b = I_b/y_b \qquad \text{(A.8)}$$

The variance function, $\Gamma^2(y_a)$, is defined as,

$$\Gamma^2(y_a) = \frac{\sigma_a{}^2}{\sigma^2} \qquad \text{(A.9)}$$

where, σ_a is the standard deviation of the segment y_a and σ is the standard deviation of the whole stratum comprised of all the segments.

Similarly, $\Gamma^2(y_b)$, $\Gamma^2(y_o)$, $\Gamma^2(y_1)$, $\Gamma^2(y_2)$ and $\Gamma^2(y_3)$ can be defined for the respective segments. The procedure for obtaining the variance function is described in detail elsewhere under the section on the Scale of Fluctuation.

The correlation coefficient ρ_{ab} between the spatial average of Layer A (Q_a) and the spatial average of Layer B (Q_b) in Fig.4.10 is given by,

$$\rho_{Q_a Q_b} = \frac{\text{COV}[Q_a Q_b]}{\sigma_a \sigma_b} \qquad \text{(A.10)}$$

where, $\text{COV}[Q_a Q_b]$ is the covariance between Q_a and Q_b.

Consequent to the manipulation of Eqs. A.6 to A.10 it can be shown that,

$$\rho_{ab} = \frac{y_o{}^2 \Gamma^2(y_o) - y_1{}^2 \Gamma^2(y_1) + y_2{}^2 \Gamma^2(y_2) - y_3{}^2 \Gamma^2(y_3)}{2(\sqrt{(y_a{}^2 \Gamma^2(y_a) . y_b{}^2 \Gamma^2(y_b))})} \qquad \text{(A.11)}$$

Appendix B

Probability of Exceedance

A classical formula for the mean rate of crossings of the level q (ν_q), by a stationary random process $Q(l)$ is given by Rice (1945);

$$\nu_q = \int_{-\infty}^{\infty} |\dot{Q}| \, f_{Q,\dot{Q}}(q,\dot{Q}) d\dot{Q} \tag{B.1}$$

where, $f_{Q,\dot{Q}}(Q,\dot{Q})$ is the joint probability density function of $Q(l)$ and its derivative $\dot{Q}(l)$. Since $Q(l)$ is stationary the random variables $Q(l)$ and $\dot{Q}(l)$ are uncorrelated. If $Q(l)$ is Gaussian, independence too is guaranteed; then,

$$\begin{aligned} \nu_q &= f_Q(q) \int_{-\infty}^{\infty} |\dot{Q}| \, f_{\dot{Q}}(\dot{Q}) d\dot{Q} \\ &= f_Q(q) E\left[|\dot{Q}|\right] \end{aligned} \tag{B.2}$$

where, $E\left[|\dot{Q}|\right]$ is the mean of the absolute value of the slope of $Q(l)$. Every q - upcrossing (crossing of the level q with positive slope) is followed by a q - downcrossing, resulting in the mean rate of upcrossings, ν_q^+, being equal to the mean rate of downcrossings, ν_q^-. Hence, it follows that the mean rates of up and down crossings ν_q^+ and ν_q^- are equal to $\nu_q/2$. Therefore,

$$\nu_q^+ = \frac{1}{2} f_Q(q) E\left[|\dot{Q}|\right] \tag{B.3}$$

Since differentiation is a linear operation, if $Q(l)$ is Gaussian, its derivative $\dot{Q}$ too will be Gaussian. Therefore,

$$E\left[|\dot{Q}|\right] = 2 \int_0^{\infty} \frac{\dot{Q}}{\sqrt{2\pi}\sigma_{\dot{Q}}} \exp\left\{\frac{-\dot{Q}^2}{2\sigma_{\dot{Q}}2}\right\} d\dot{Q}$$

$$= \sqrt{\frac{2}{\pi}} \sigma_{\dot{Q}} \tag{B.4}$$

Substituting for $E\left[|\dot{Q}|\right]$ in Eq. B.3,

$$\nu_q^+ = \frac{1}{2} f_Q(q) \sqrt{\frac{2}{\pi}} \sigma_{\dot{Q}} \tag{B.5}$$

If $Q(l)$ is normally distributed,

$$f_Q(q) = \frac{1}{\sqrt{2\pi}\sigma_Q} \exp\left\{-\frac{1}{2}\left(\frac{q - \bar{Q}}{\sigma_Q}\right)^2\right\} \tag{B.6}$$

where, $\bar{Q}$ and σ_Q are the mean and standard deviation of the soil property, Q, of a local region of length D and q is the property value of which the exceedance or non exceedance is of interest, or in short the threshold value.

Substituting for $f_Q(q)$ in Eq. B.5

$$\begin{aligned}
\nu_q^+ &= \frac{1}{2\sigma_Q\sqrt{2\pi}} \exp\left\{-\frac{(q - \bar{Q})^2}{2\sigma_Q{}^2}\right\} \sqrt{\frac{2}{\pi}} \sigma_{\dot{Q}} \\
&= \frac{1}{2\pi} \frac{\sigma_{\dot{Q}}}{\sigma_Q} \exp\left\{-\frac{(q - \bar{Q})^2}{2\sigma_Q{}^2}\right\}
\end{aligned} \tag{B.7}$$

The mean rate of zero crossings, ν_o is obtained when $q = \bar{Q}$ in Eq. B.7 and can be expressed as,

$$\nu_o = \frac{1}{2\pi} \frac{\sigma_{\dot{Q}}}{\sigma_Q} \tag{B.8}$$

thereby, permitting Eq. B.7 to be written as,

$$\nu_q^+ = \nu_o \exp\left\{-\frac{(q - \bar{Q})^2}{2\sigma_Q{}^2}\right\} \tag{B.9}$$

Considering the autocorrelation function $\rho(D)$ and the variance function $\Gamma^2(D)$ for a local region D, the mean rate of crossings, ν_o, can also be expressed as (Rice - 1948);

$$\nu_o = \left\{\frac{1 - \rho(D)}{2\pi^2 D^2 \Gamma^2(D)}\right\}^{\frac{1}{2}} \tag{B.10}$$

From the definition of the Variance function,

$$\sigma_Q^2 = \sigma^2 \Gamma^2(D) \tag{B.11}$$

Substituting for ν_o and $\sigma_Q{}^2$ from Eqs. B.10 and B.11

$$\nu_q^+ = \frac{\sqrt{1 - \rho(D)}}{\sqrt{2}\pi D \Gamma(D)} \exp\left\{ -\frac{(q - \bar{Q})^2}{2\sigma^2} \Gamma^2(D) \right\} \tag{B.12}$$

There will be many segments of length D within the domain length L and hence the probability of non exceedance (P_E), for all such segments within the entire layer of length L, will be approximately given by, (Vanmarke - 1987),

$$P_L = \exp(-\nu_q^+ L) \tag{B.13}$$

Therefore, the probability that the average of a local interval of length D will exceed a threshold value q, will be given by,

$$P_E = 1 - \exp(-\nu_q^+ L) \tag{B.14}$$

From Eqs. B.12 and B.14 it is evident that the probability of exceedance is dependent on the local region of width D, mean $(\bar{Q})$ and standard deviation (σ_Q) of the entire layer, value of the autocorrelation function of the layer at D (ρ_D), square root of the variance function at D, $(\Gamma(D))$, the threshold value q and L, the thickness of the domain.

Appendix C

Interpolation Methods Neglecting Correlation

C.1 Weighting Functions

Regression techniques require an adequate data set for interpolation to be carried out. In geotechnical engineering, the available data are very often scarce, and the only option may be to adopt methods of weighting functions for interpolation. The major drawback in all methods of weighting functions is that redundant information is not discriminated against. If A, B and C are points with known soil properties and are equidistant to a point P, where estimation is required, then these three points are given weights of 1/3. Now, if another known point D is very close to point A, but again equidistant to P, all four points (A, B, C and D) are assigned the same weight of 1/4. Similarly, if there is a cluster of points around point A, thereby increasing the number of data points around that point, equal weights are given to all points since they are equidistant to P. However, it is obvious that the effects of points B and C on the prediction point P should be negligible in the latter case. Therefore, it is very important that if weighting techniques are used for interpolation, the potential problem discussed should be recognized.

C.2 Distance Weighting Functions

A convenient but very approximate way of an estimation procedure is to designate higher weights to those points which are situated closer to the point where estimation is required. In order to accomplish this, inverse distant weighting functions or

inverse squared distance weighting functions are used. These methods again do not discriminate redundant information. For example, for a cluster of points equidistant to the point of estimation, the weights designated will be approximately equal.

The estimated value, $Q^*(x_o)$, at point x_o can be expressed as,

$$Q^*(x_o) = \sum_{i=1}^{n} \lambda_i Q(x_i) \qquad (C.1)$$

where,

$$\lambda_i = \frac{1/L_i^r}{\sum_{i=1}^{n} 1/L_i^r} \qquad (C.2)$$

In the above equation, L_i is the distance between the data point and the point where the estimate is required. If inverse distance weighting is adopted, r = 1 whereas for inverse squared distance weighting, r = 2.

C.3 Functional Weighting Functions

The concept that the weight decreases with increasing distance is used when a distance correlation function (e.g. exponential decay model) is used for interpolation at unknown points, viz;

$$\lambda_i = \frac{f\left(\mid x_o - x_i \mid\right)}{\sum_{j=1}^{n} f\left(\mid x_o - x_i \mid\right)} \qquad (C.3)$$

where, $\mid x_o - x_j \mid = L_j$ is the distance and α is a appropriate constant in the expression, $f(L) = \text{EXP}(-\alpha L)$. It should be noted that the weight λ approaches zero, as the distance L_j increases.

C.4 Simple Weighting Functions

The classical estimation of the mean or the average is also a weighting function with, λ_i in Eq. C.1, given by,

$$\lambda_i = 1/n \qquad (C.4)$$

for i = 1, 2, ...,n with all the weights being equal.

Appendix D

Interpolating Equations Considering Correlations

As in any optimization procedure, an estimator, $\hat{Q}(s_o)$, of a soil property value, $Q(s_o)$, will be termed a " best estimator ", if it minimizes the mean square error (David - 1976).

That is,

$$E\left[Q(s_o) - \hat{Q}(s_o)\right]^2 = \text{MINIMUM} \qquad (D.1)$$

where, $\hat{Q}(s_o)$, is the estimator of $Q(s_o)$ and E[.] is the expected value.

The estimator, $\hat{Q}(s_o)$, will be the " best unbiased estimator " if the following condition is satisfied.

$$E\left[Q(s_o) - \hat{Q}(s_o)\right] = 0 \qquad (D.2)$$

The estimator, $\hat{Q}(s_o)$, can also be expressed as,

$$\hat{Q}(s_o) = \sum_{i=1}^{n} \lambda_i Q(s_i) \qquad (D.3)$$

where n is the total number of data points and s_i's are the locations of the data points, where soil properties are known, with $i = 1, 2, \ldots\ldots, n$. λ_i's are the weights. The estimation variance, $\sigma_e{}^2$, follows from Eq. D.1. After substituting for $\hat{Q}(s_o)$ from Eq. D.3,

$$\sigma_e{}^2 = E\left[Q(s_o) - \sum_{i=1}^{n} \lambda_i Q(s_i)\right]^2 \qquad (D.4)$$

By expanding and taking expectations, the variance, $\sigma_e{}^2$, can be expressed as,

$$\sigma_e{}^2 = \sigma^2_{Q(s_o)} - 2\sum_{i=1}^{n}\lambda_i\sigma_{Q(s_o)Q(s_i)} + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\sigma_{Q(s_i)Q(s_j)} \qquad \text{(D.5)}$$

The necessary condition for the best estimate is that $\sigma_e{}^2$ given by Eq. D.5 be a minimum. In addition, the estimate has also to be unbiased. That is on average, the estimate $\hat{Q}(s_o)$, should be equal to the actual value $Q(s_o)$.

If the estimated value,

$$E\left[\hat{Q}(s_o)\right] = \xi \qquad \text{(D.6)}$$

From Eqs. D.2 and D.3

$$E\left[\sum_{i=1}^{n}\lambda_i Q(s_i)\right] = \xi \qquad \text{(D.7)}$$

Therefore,

$$\sum_{i=1}^{n}\lambda_i E\left[Q(s_i)\right] = \xi \qquad \text{(D.8)}$$

For the unbiased condition,

$$E\left[Q(s_i)\right] = \xi \qquad \text{(D.9)}$$

From Eqs. D.8 and D.9, it could be easily inferred that,

$$\sum_{i=1}^{n}\lambda_i = 1 \qquad \text{(D.10)}$$

Therefore, in order to satisfy the unbiased condition as well as the minimum variance condition, Eq. D.5 will have to be minimized, subject to the condition stipulated by Eq. D.10. This is a case of finding the minimum value of a function of several variables λ_i, when the relationship between the variables are given (Eq. D.10). The above criteria can be solved by using the Lagrange Principle (David - 1973, Kreyszig - 1983). The equation to be solved (Eq. D.5), subject to the restriction in Eq. D.10,

then transforms to a function ψ, that has to be minimized.

The function ψ is given by,

$$\psi = \sigma_e^2 + 2\mu \left(\sum_{i=1}^{n} \lambda_i - 1 \right) \tag{D.11}$$

where, μ is a the Lagrange constant. Substituting for σ_e^2, from Eq. D.5,

$$\psi = \sigma_{Q(s_o)}^2 - 2\sum_{i=1}^{n} \lambda_i \sigma_{Q(s_o)Q(s_i)} + + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \sigma_{Q(s_i)Q(s_j)} + 2\mu \left(\sum_{i=1}^{n} \lambda_i - 1 \right) \tag{D.12}$$

Taking partial derivatives of Eq. D.12 with respect to λ_i's and μ,

For $i = 1, 2, \ldots\ldots, n$

$$\frac{\partial \psi}{\partial \lambda_i} = -2\sigma_{Q(s_o)Q(s_i)} + 2\sum_{j=1}^{n} \lambda_i \sigma_{Q(s_i)Q(s_o)} + 2\mu \tag{D.13}$$

$$\frac{\partial \psi}{\partial \mu} = \sum_{i=}^{n} \lambda_i - 1 \tag{D.14}$$

Equating Eq. D.13 and D.14 to zero, to minimize F,

$$\sum_{j=1}^{n} \lambda_j \sigma_{Q(s_i)Q(s_j)} + \mu = \sigma_{Q(s_o)Q(s_i)} \tag{D.15}$$

Equation. D.15 represents a linear system of n equations for $i = 1, 2, \ldots, n$. Eq. D.10 together with the linear system of n equations given by Eq. D.15, comprise a system of (n+1) equations, with a similar number of unknowns. The unknowns being the n number of λ_i terms and the Lagrange constant μ.

Equations D.15 and D.10 can be expressed in the following matrix form ;

$$[\mathbf{A}]\{\mathbf{B}\} = \{\mathbf{C}\} \tag{D.16}$$

where, $[\mathbf{A}]$ is the covariance matrix given by,

$$[\mathbf{A}] = \begin{bmatrix} \sigma_{Q(s_1)Q(s_1)} & \sigma_{Q(s_1)Q(s_2)} & \sigma_{Q(s_1)Q(s_3)} & \cdots & \cdots & \sigma_{Q(s_1)Q(s_n)} & 1 \\ \sigma_{Q(s_2)Q(s_1)} & \sigma_{Q(s_2)Q(s_2)} & \sigma_{Q(s_2)Q(s_3)} & \cdots & \cdots & \sigma_{Q(s_2)Q(s_n)} & 1 \\ \sigma_{Q(s_3)Q(s_1)} & \sigma_{Q(s_3)Q(s_2)} & \sigma_{Q(s_3)Q(s_3)} & \cdots & \cdots & \sigma_{Q(s_3)Q(s_n)} & 1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \sigma_{Q(s_n)Q(s_1)} & \sigma_{Q(s_n)Q(s_2)} & \sigma_{Q(s_n)Q(s_3)} & \cdots & \cdots & \sigma_{Q(s_n)Q(s_n)} & 1 \\ 1 & 1 & 1 & \cdots & \cdots & 1 & 0 \end{bmatrix} \quad (D.17)$$

$$\{\mathbf{C}\} = \begin{Bmatrix} \sigma_{Q(s_o)Q(s_1)} \\ \sigma_{Q(s_o)Q(s_2)} \\ \sigma_{Q(s_o)Q(s_3)} \\ \vdots \\ \vdots \\ \sigma_{Q(s_o)Q(s_n)} \\ 1 \end{Bmatrix} \quad (D.18)$$

$$\{\mathbf{B}\} = \begin{Bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \vdots \\ \lambda_n \\ \mu \end{Bmatrix} \quad (D.19)$$

If both sides of Eq. D.15 are divided by the variance of the data (σ^2),

$$[\mathbf{P}]\{\mathbf{L}\} = \{\mathbf{M}\} \quad (D.20)$$

where,

$$[\mathbf{P}] = \begin{bmatrix} 1 & \rho_{Q(s_1)Q(s_2)} & \rho_{Q(s_1)Q(s_3)} & \cdots & \cdots & \rho_{Q(s_1)Q(s_n)} & 1 \\ \rho_{Q(s_2)Q(s_1)} & 1 & \rho_{Q(s_2)Q(s_3)} & \cdots & \cdots & \rho_{Q(s_2)Q(s_n)} & 1 \\ \rho_{Q(s_3)Q(s_1)} & \rho_{Q(s_3)Q(s_2)} & 1 & \cdots & \cdots & \rho_{Q(s_3)Q(s_n)} & 1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \rho_{Q(s_n)Q(s_1)} & \rho_{Q(s_n)Q(s_2)} & \rho_{Q(s_n)Q(s_3)} & \cdots & \cdots & 1 & 1 \\ 1 & 1 & 1 & \cdots & \cdots & 1 & 0 \end{bmatrix} \tag{D.21}$$

$$\{\mathbf{M}\} = \begin{Bmatrix} \rho_{Q(s_o)Q(s_1)} \\ \rho_{Q(s_o)Q(s_2)} \\ \rho_{Q(s_o)Q(s_3)} \\ \vdots \\ \vdots \\ \rho_{Q(s_o)Q(s_n)} \\ 1 \end{Bmatrix} \tag{D.22}$$

$$\{\mathbf{L}\} = \begin{Bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \vdots \\ \vdots \\ \lambda_n \\ \mu/\sigma^2 \end{Bmatrix} \tag{D.23}$$

Therefore, the matrix of unknowns, $\{\mathbf{L}\}$, will be given by,

$$\{\mathbf{L}\} = [\mathbf{P}]^{-1}\{\mathbf{M}\} \tag{D.24}$$

$\{\mathbf{M}\}$ and $[\mathbf{P}]$ in above equations are for the case when the autocorrelation function is used. If the semi-variogram is used the σ terms in Eq. D.17 will be replaced by γ terms, thereby causing a change in matrices $[\mathbf{P}]$ and $\{\mathbf{M}\}$. However, the weights (

λ_i) obtained by both methods will be identical due to the direct relationship between the autocorrelation and semi-variogram functions.

In all of the above expressions, s_o is the point where the interpolation is required, while s_1, s_2, $\ldots\ldots$, s_n are the locations where the property values are known. Once the weights, λ are obtained from Eq. D.24, the estimator, $\hat{Q}$ at s_o, can be determined from Eq. D.3 as follows ;

$$\hat{Q}(s_o) = \lambda_1 Q(s_1) + \lambda_2 Q(s_2) + \lambda_3 Q(s_3) + \ldots\ldots\ldots + \lambda_n Q(s_n) \qquad (D.25)$$

where, $Q_{s_1}, Q_{s_2}, \ldots\ldots, Q_{s_n}$ are the known soil property values at s_1, s_2, $\ldots\ldots, s_n$.

The estimation variance ($\sigma_e{}^2$) can be obtained considering Eq. D.5 together with the restrictions imposed by Eqs. D.15 and D.10, and is given by,

$$\sigma_e{}^2 = \sigma^2 \left(1 - \sum_{i=1}^{n} \lambda_i \rho_{s_i s_o} \right) - \mu \qquad (D.26)$$

If the semi-variogram was used instead of the autocorrelation function, the estimation variance can be expressed as,

$$\sigma_e{}^2 = \sum_{i=1}^{n} \lambda_i \gamma_{s_i s_o} - \mu \qquad (D.27)$$