

Prepared for ISOPT1 - Specialty Session on "The Use of Statistics in Evaluating Penetration Test Results".

#62

APPLICATIONS OF STATISTICAL METHODS IN SITE CHARACTERIZATION

R . G. Campanella and Damika S. Wickremesinghe, Department of Civil Engineering, University of British Columbia, Vancouver, B. C., Canada V6T 1W5.

1.0 Introduction

The natural variability of the soil, the limitation of available data, soil disturbance while testing or sampling and measurement errors, all contribute to the uncertainty of soil property evaluation. If every point in the ground could be tested, soil properties could be known at all intended locations. However, in practice this is not feasible and the need arises to treat this variation as random. In this regard statistics and probabilistic methods become a very appropriate tool in characterizing such variations.

This presentation deals with the Cone Penetrometer Test (CPT), the soundings of which are primarily used to identify soil stratigraphy. All data analysed herein have been either obtained from the McDonald Farm insitu research site of the University of British Columbia or from its newer site adjacent to the Arthur Laing Bridge, both situated at the Vancouver International Airport on Sea Island. The data have been acquired at 2.5 cm intervals using a cone with a base area of 10 sq.cm., penetrating at 2 cm. per second. Methods of filtering CPT data using statistical methods, trend analysis to characterize different types of layering, the concept of the scale of fluctuation and methods of eliminating random error will be discussed in the initial sections of the paper together with a method of determining the optimum number of samples required to identify a layer of soil, at a certain confidence level and precision. The last part of the paper will discuss the interpolation of soil property values in two dimensions considering the correlation between points. The importance of the autocorrelation function and the variogram function will be highlighted with an application to a two dimensional interpolation problem.

2.0 Filtering of CPT Profiles

Filtering is done to eliminate extremes in order to identify trends and is a process which requires engineering judgement to ensure that finite layers are not removed from the profile. It should be emphasized that only distinct anomalies in data are removed. To make sure that the latter condition mentioned above is satisfied, the soil profile is divided into thin layers of 25 cm., each sublayer consisting of ten data points. There are several options available in the computer routine developed for filtering CPT data. These being, sublayer thickness, median or mean filtering, width of filtering window, method of replacement of filtered data etc.. Figures 1a and 1b and 2a and 2b illustrate the effect of filtering on two different CPT profiles which have

been filtered using the median method. A distinct advantage of the median method, is that it is not affected by the values of data extremities unlike the mean method, which is a function of all the data including extreme values on both the low and high sides. Methods of replacement of filtered data include substitution by the mean or median, the average of the immediately adjacent unfiltered data points or even data removal. The degree of filtering (DF) which is dependent on the width of filtering window, depends on the intensity of filtering required and also on the variability of the profile itself. DF is the number of standard deviations above and below the median (or mean as the case may be) which forms the filtering window and the data falling outside this range are removed. The effect of filtering is expressed as a filtering ratio, defined by the ratio of the number of filtered points to the total number of data. The results also give the number of data points removed from each sublayer, so that it may be evident if a significant thin layer has been removed. The profiles in Figures 1a and 2a have used the median method with a DF of 2 and filtered data have been replaced by the average of adjacent points. The filtering ratios of profiles 1 and 2 were found to be, .09 and .08 respectively. If it is felt that too many data have been filtered out from a particular sub layer, the filtering ratio can be reduced by increasing DF, that is by increasing the window within which data are not removed. It should be reiterated once again, that application of filtering is highly situation dependent and requires good engineering judgement, keeping in mind its prime purpose.

3.0 Trend Analysis

As mentioned before, the main purpose of the CPT is to identify different types of soil layers in a stratum. Soil properties are highly depth dependent, and in most profiles a significant depth dependency is observed as can be seen from Fig.3b. The breaks in the trend will also indicate the different kinds of layering and a closer examination of the cone bearing log will indicate the approximate layer start and layer end depths of the sublayers within the entire profile. At present the friction ratio is essentially used to differentiate different layers (Fig.3a) and as can be observed it does not exhibit a trend within a soil type. It has been found by the authors that the simple statistical parameter, the coefficient of variation is a very good indicator for the above purpose of identifying different types of layering. The coefficient of variation is the ratio between the standard deviation and the mean, and its variation with depth is illustrated in Fig.4a. Comparison of Figs.4a and 4b illustrate that the different layers identified possess different degrees of variation, the average values of which are tabulated in Table 1.

The importance of trend analysis in two and three dimensional analysis will be discussed later. Most dimensional statistical methods, deal with stationary data (data with no trend) which is referred to as homogeneous data in two and three dimensional analysis. Non-stationary data with which geotechnical engineers deal with regularly, can be transformed to stationary data by removing the trend as follows;

$$\text{RESIDUAL} = \text{DATA} - \text{TREND} \quad (1)$$

The deterministic component, the trend, may be obtained by some form of a least squares regression technique. The appropriate trend for any profile provides a fit that gives the best correlation coefficient and the least variance. The residual is the stationary component which is used in correlation analysis, for various purposes such as interpolation. The final estimated value in such a procedure is the sum of the regressed trend term and the correlated residual term. Details of this procedure is described in a later section.

4.0 Scale of Fluctuation

In order to describe a soil profile completely, the scale of fluctuation, δ , is required in addition to the mean and standard deviation. (Vanmarke, 1977). The scale of fluctuation, (δ), is an indication of the degree of variability of a profile. A highly variable profile will have a low δ while a slowly varying profile will result in a high δ . The scale of fluctuation of any stratum is inversely proportional to the coefficient of variation, as indicated in Table 1. The scale of fluctuation is also referred to as the distance of perfect correlation since it is the distance within which the soil property shows relatively strong correlation from point to point. When two different test methods are being compared, it is recommended that the sampling distance be less than δ , so that comparison is being done in a region of perfect correlation. The opposite is true when sampling is performed using the same equipment, where for optimum sampling benefit, a spacing greater than δ is advisable. The maximum values of the curves (Fig.5) for the different layers of Fig.3, are the respective scales of fluctuation for the different layers, and is also tabulated in Table 1. A detailed explanation of obtaining δ is given in Campanella and Wikremesinghe (1987). It can be observed from Table 1 that the soil that is most variable is layer 1, as given by the average coefficient of variation of 0.21. In keeping with the above argument it is the layer with the lowest δ of 20 cm. Similarly, layer 3 has the smallest coefficient of variation (0.09) and the highest δ of 71cm. Therefore for an efficient testing program using the same test equipment, sampling can be performed at spacings as much as 72 cm. in layer 3, while in layer 1 it drops down to 20 cm., solely due to the fact of its high variability. All values of δ so far discussed and appearing in Table 1 has been derived for the cone bearing value. Table 2 has these values compared to the δ values for sleeve friction. Since the sleeve friction at a particular depth is an average value over a finite length one might expect a much higher δ (due to a lower variability caused by the averaging effect) for sleeve friction than for cone bearing. However, this will be true only if the cone bearing measures the bearing at a point. Results in Table 2 indicate that the δ values for the two soil parameters are approximately equal and thus suggesting that the cone bearing too is indicative of an averaged value over a region which is likely almost the same as that of friction, and does not indicate the bearing at a particular point.

5.0 Optimum Sample Spacing

The CPT at UBC performs data logging at a depth interval of 2.5 cm. For logging purposes the spacing is ideal, because it provides almost a continuous profile. However, if a soil parameter is to be estimated from a profile the spacing may need to be adjusted, depending on the soil variability and the required confidence level of the estimate. It is assumed that any layer is fully characterized when the average value obtained from the data Q_{av} for that layer is within 10 percent (Δ) of the actual average, X_{av} , which is unknown. The tolerance allowed is therefore 10%. In other words the precision is 90%.

The sample size (n) needed to estimate the mean to the above precision, with a confidence level of $(1 - \gamma)$ is,

$$n = (Vt_{n-1}^{\gamma})/\Delta \quad (2)$$

where V is the coefficient of variation and t_{n-1}^{γ} is the students-t variate with $(n - 1)$ degrees of freedom. It is seen from the above expression that the number of samples is dependent on the variability of the soil and the level of confidence and precision required of the estimate. Table 3 has tabulated some values to show the above relationship. For the same tolerance of ± 10 and a confidence level of 95%, the sample spacing required in the soil of high variability is 2.94cm while for a soil of low variability it is 8.3 cm. However, if a higher confidence level of 99% is required in the more variable soil, spacing will have to be increased to 4.2 cm. Similarly, if the engineer requires to reduce the tolerance by half, in order to increase the precision of the estimate, the sample spacing will have to be reduced to 0.5 cm., for the same confidence level of 99%. However, the sample spacing in the less variable soil, the sample spacing required for the same confidence level and a precision of ± 0.05 is 1.5 cm. This clearly illustrates the three factors which contribute to the selection of sample spacing in a soil stratum; soil variability, precision of the estimate and confidence level of the estimate.

6.0 Errors in Testing and Data Scatter

The scatter in geotechnical data is obtained from three sources; actual variability of soil properties, random measurement error and bias. The bias is a systematic error introduced by systematic influences in testing. Measurement errors with non zero mean are considered as biases. While there are suggested ways of removing the random error, the bias can be determined only in relative terms, that is in comparison with a result obtained more accurately using a better test method. For example, if the undrained strength of a clay has been determined both from the triaxial test in the laboratory and in the field from the vane test, the bias in the vane test can be determined relative to the laboratory values. In the absence of such results the only error that can be removed from data scatter is the random error. Random

measurement error or noise is assumed to be independent from point to point, whereas actual soil properties are not. The covariance $C(h)$ which is used for this purpose is defined below.

The covariance at lag h ($C(h)$), is given by,

$$C(h) = (N - h)^{-1} \sum_{i=1}^{N-h} (X_i - X_{av})(X_{i+h} - X_{av}) \quad (3)$$

where X_i is the measured soil property value at a point 'i', X_{av} is the mean of the data and N is the total number of data. If the associated error at point 'i' is E_i , the actual value at that point (X_c)_i is given by,

$$(X_c)_i = X_i + E_i \quad (4)$$

Considering covariances (C) of the above equation, it can be shown that,

$$C(X_c) = C(X) + C(E) \quad (5)$$

Since the random error E is independent from point to point, $C(E)$ will have a value not equal to zero only at zero distance, while the soil property variation will have a maximum value of $C(X)$ at zero distance and a slowly decaying function, with increasing distance. This is because soil properties at points closer to each other show a stronger correlation than points further apart. The random measurement error would therefore be the difference between the variance of the data (covariance at zero separation distance) and the value of the decaying covariance function, at the point where it meets the ordinate. This method was first introduced by Beacher (1978) and has been performed on two cone penetrometer tests and one standard penetration test (SPT). As expected the random measurement error for the CPT was very low compared to that of the SPT which is high at approximately 25%. The random measurement error of the CPT has also been obtained using Box - Jenkins (1976) methods of Time Series analysis (Wu et al. 1986) using the statistical package SAS. Although the results from this analysis were approximate, due to the assumptions employed, it compared well with the low values obtained using the above method. These low values were in the range of 2 - 4% of the measured soil property values.

7.0 Interpolation of Soil Property Values considering Correlation

As already mentioned in an earlier section, the limitation of data availability at a site results in interpolating soil property values at untested locations. Traditionally, geotechnical engineers are very conservative and would typically use very low bounds of the soil parameters in design and analysis. Independence of soil property values between points is assumed and interpolation methods such as least squares regression and distance and simple weighting functions are used. The mean of a data set is obtained from a simple weighting function where the weights are all equal to the reciprocal of the number of data. An obvious shortcoming of all the above methods, is that redundant information is not discriminated against. That is a cluster of n data points located very close to each other, will get the same weight for each data point as for a single data point located at the same distance from the point to be estimated, but at a different location. Therefore when n is very large, the estimation will almost totally depend on the cluster of data points, completely neglecting the effect of the isolated data point. This is actually a hypothetical extreme case, but it clearly exemplifies the shortcoming of such methods. In contrast, methods which account for correlation overcomes this drawback.

Soil property values, situated closer to each other are expected to be related more to each other, compared to points which are separated, wider apart. This relationship between data points is expressed by a correlation function, either in the form of the autocorrelation function ρ or the variogram function γ which are defined below.

The autocorrelation at a lag h , $\rho(h)$ is defined as,

$$\rho(h) = \frac{1}{N} \sum_{i=1}^{N-h} (X_i - X_{av})(X_{i+h} - X_{av}) / (N-h) \sum_{i=1}^{N-h} (X_i - X_{av})^2 \quad (6)$$

The variogram function at lag h , $\gamma(h)$ is defined as

$$\gamma(h) = \frac{1}{2(N-h)} \sum_{i=1}^{N-h} (X_i - X_{i+h})^2 \quad (7)$$

where, X_i are the data points, X_{av} the mean of the data and N is the total number of data.

To develop an autocorrelation function which closely resembles the actual process in the field, in an analytical form, a sizeable data set will be required. However, in geotechnical engineering, this requirement will rarely be satisfied in a typical project, except for large projects like a site investigation for an offshore oil platform, an earth dam or a very high risk project. Therefore, it is very important that the test locations be chosen in such away as to optimize the information that could be derived from the investigation.

The approach of separating the trend (if a trend exists) from the observed data, and performing interpolations on the correlated residuals, has been carried out on some data obtained from the McDonald Farm site. Seven CPT's were performed at 5 meter spacings (Fig.6). The profiles were divided into layers and where applicable the trend was removed. Several models were attempted to model the field autocorrelation function and the one with the closest fit was an exponential sinusoidal function in the form,

$$\rho(x,y) = \exp[-(x/q + y/r)].\text{COS}(x/q + y/r) \quad (8)$$

where, q and r are constants and x and y are the horizontal and vertical distances respectively. Fig.6 also shows the variability across the site at McDonald Farm, in the form of the seven cone bearing profiles obtained at locations A,B,C,D,E,F and G which were spaced at 5 meter intervals. Figs. 7 and 8 illustrate the interpolated values together with the two immediately adjacent cone profiles. Point M is located exactly between D and E while point N is located 2 meters from E towards F. The interpolated values clearly indicate the effect of correlation because the value at M is clearly not the mean of the data at D and E. It should be emphasized that the autocorrelation or variogram function must be determined for the trend removed data, or otherwise the interpolation would result in significant error.

8.0 Micro Computer Programs

Micro Computer Programs have been developed to perform all the above statistical procedures, and are adaptable for various data formats. These programs which have been written in Fortran77 and compiled using Microsoft Fortran, are all interactive and are very flexible with several options available to the user.

9.0 References

- Box G. E. and Jenkins G. M. (1978) - "Time Series Analysis", Forecasting and Control, Holden-Day, San Francisco CA.
- Baecher G.B. (1978) - "Analysing Exploration Strategies", Site Characterization, ed. C. H. Dowding, NSF-ASCE.
- Campanella R. G. and Wickremesinghe D. S. (1978) - "Statistical Treatment of Cone Penetrometer Test Data", ICASP5, Vancouver, B.C., Canada.
- Vanmarke E. H. (1977) - "Probabilistic Modeling of Soil Profiles", ASCE, Vol.103, No.GT11.
- Wu T.H. and El-Jandali A. (1985) - "Use of Time Series in Geotechnical Data Analysis", Geotechnical Testing Journal ASTM, Vol.8, No.4.

Parameter	Layer 1	Layer 2	Layer 3	Layer 4
Scale of Fluc. δ (cm)	20	60	72	40
Aver. Coeff. of Variation	.21	.17	.09	.19

Table 1. Variation of Scale of Fluctuation and Average Coefficient of Variation for Profile in Figure 3b.

Scale of Fluctuation, δ	Layer 1	Layer 2	Layer 3	Layer 4
δ for Bearing (cm)	20	60	72	40
δ for Friction (cm)	20	53	64	33

Table 2. Comparison of the Scale of Fluctuation obtained for Cone Bearing and Friction for Profile in Fig.3b.

Layer (m)	Tolerance	Conf. Level %	n	Spacing(cm)
4.50 - 5.0 Coefficient of Variation = 0.11	$\pm .10$	90	5	12.5
		95	7	8.3
		99	11	5.0
	$\pm .05$	90	15	3.5
		95	20	2.5
		99	34	1.5
10.0 - 10.5 Coefficient of Variation = 0.20	$\pm .10$	90	13	4.2
		95	18	2.94
		99	30	1.72
	$\pm .05$	90	43	1.2
		95	64	0.8
		99	110	0.5

Table 3. Spacing of Samples for a Given Precision and Confidence Level for data in

Fig.3b.

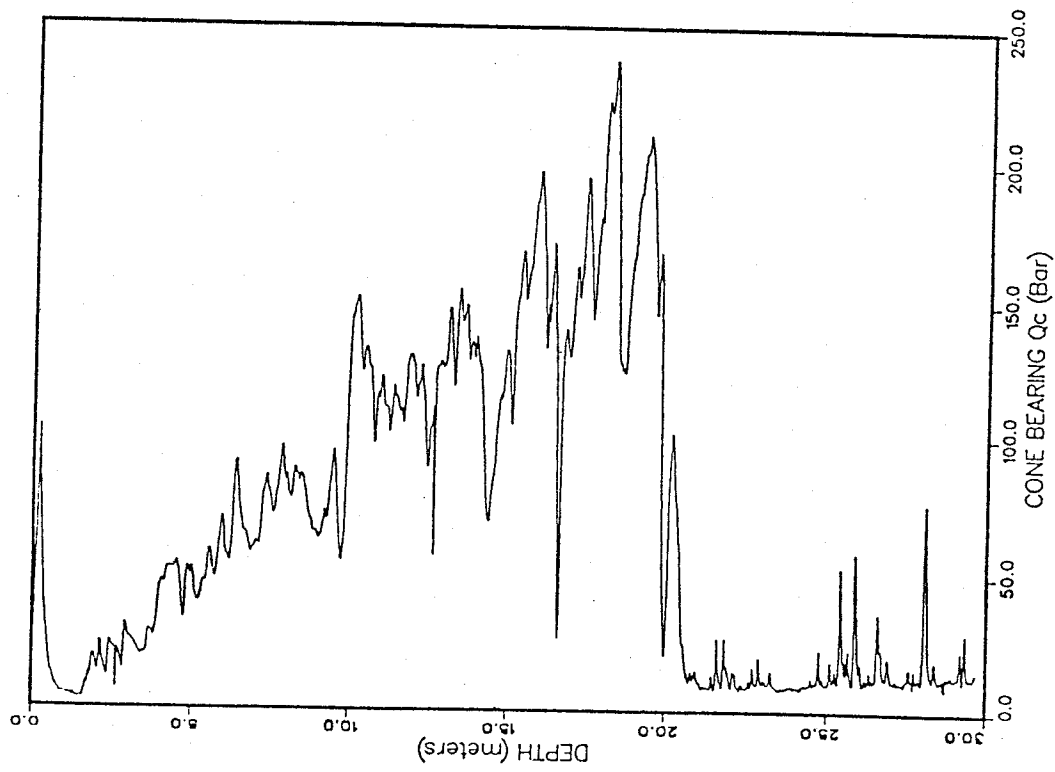


Fig.1a - Cone Bearing Profile (1) at Laing Bridge Site

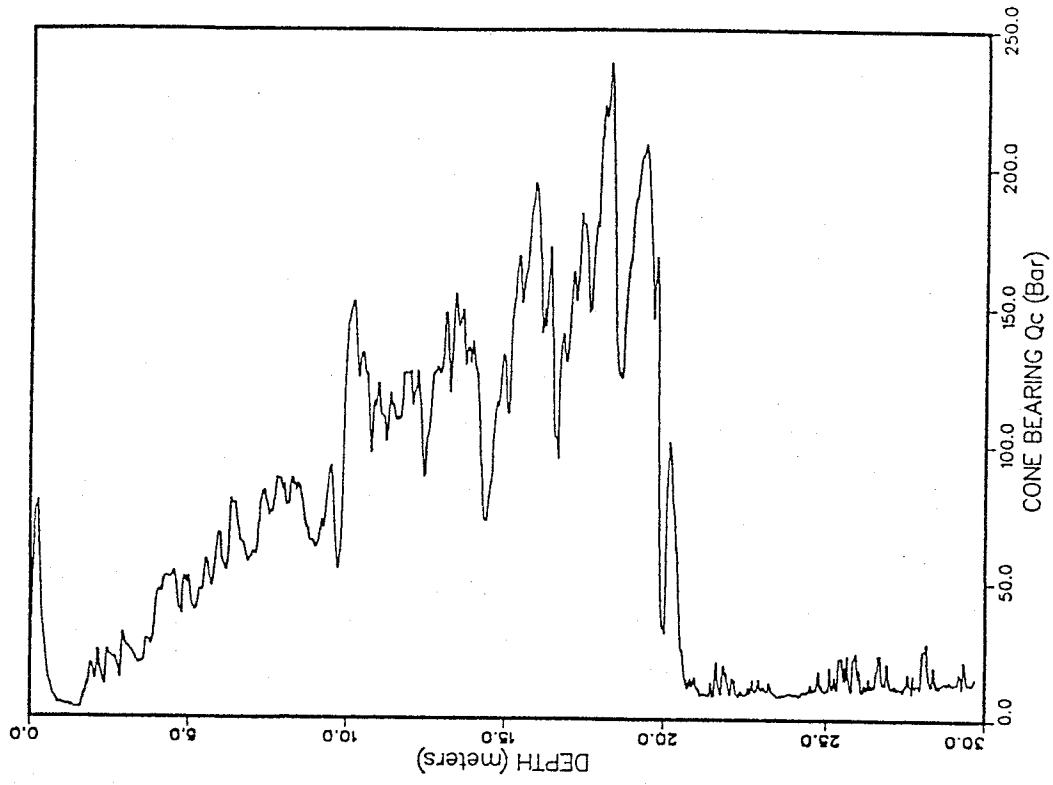


Fig.1b - Filtered Cone Bearing Profile of Fig.1a

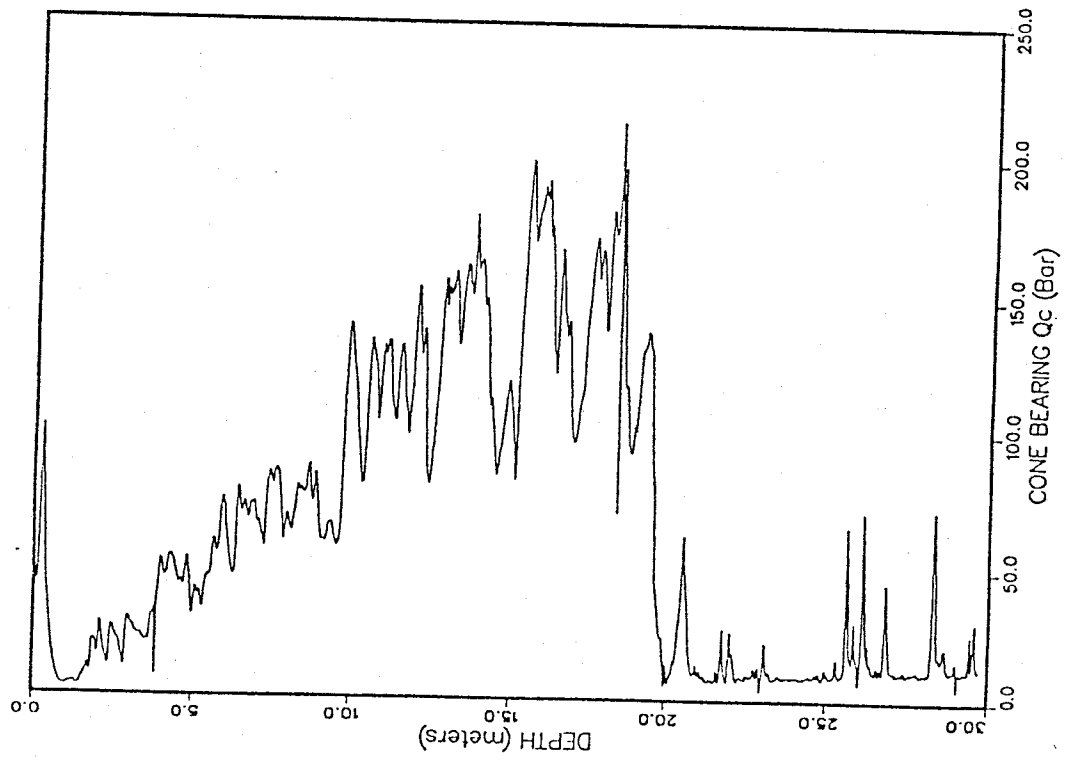


Fig.2a -- Cone Bearing Profile (2) at Laing Bridge Site

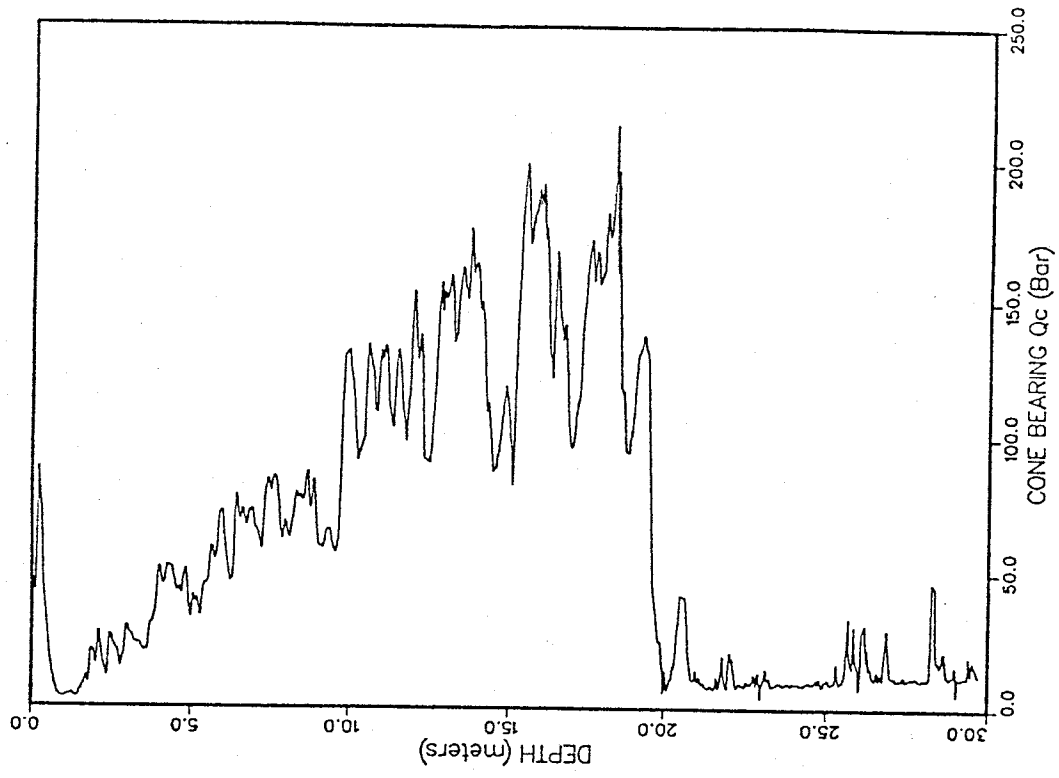


Fig.2b -- Filtered Cone Bearing Profile of Fig.2a

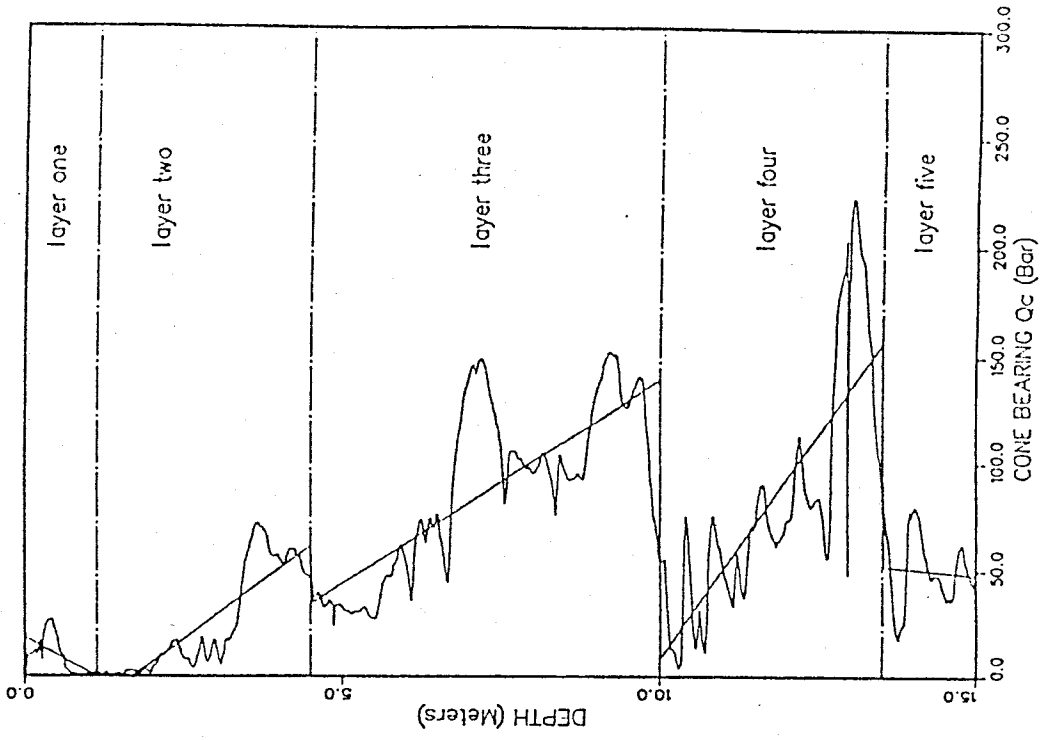


Fig.3a - Friction Ratio Profile at McDonald Farm

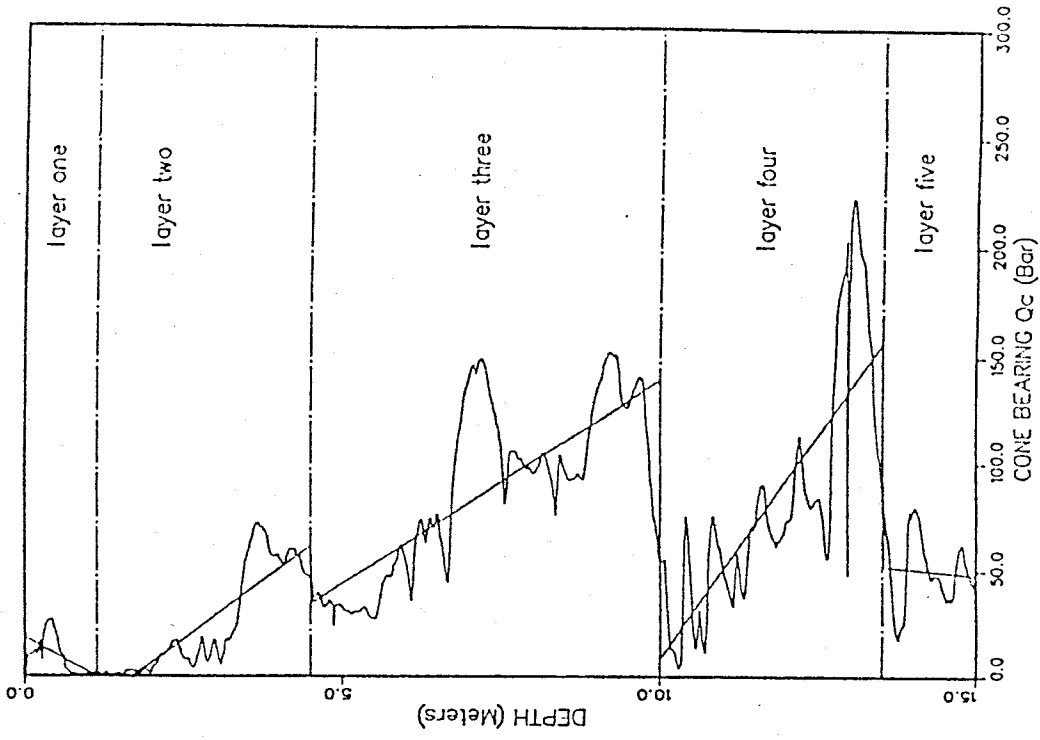


Fig.3b - Cone Bearing Profile at McDonald Farm

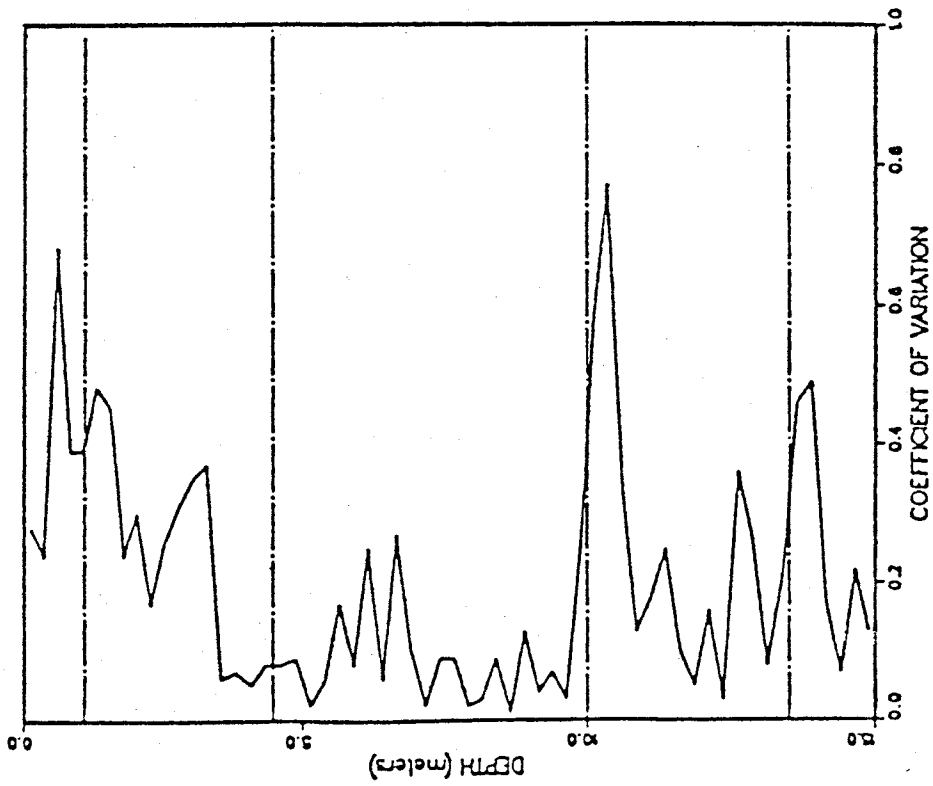
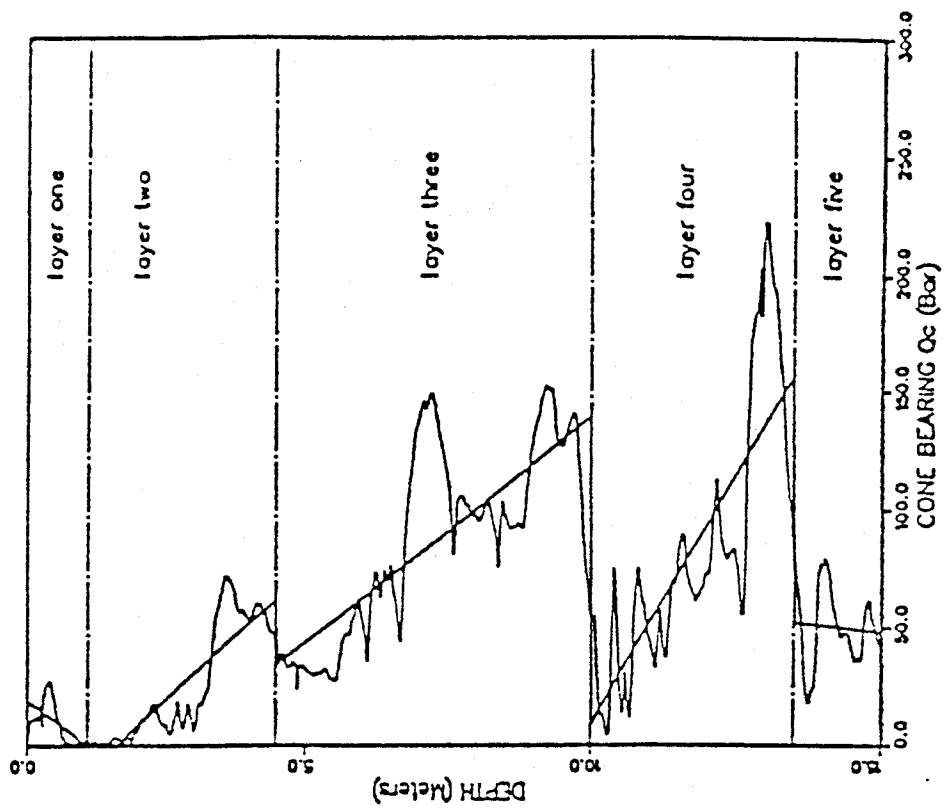


Fig.4 - Comparison of Coefficient of Variation with Trend

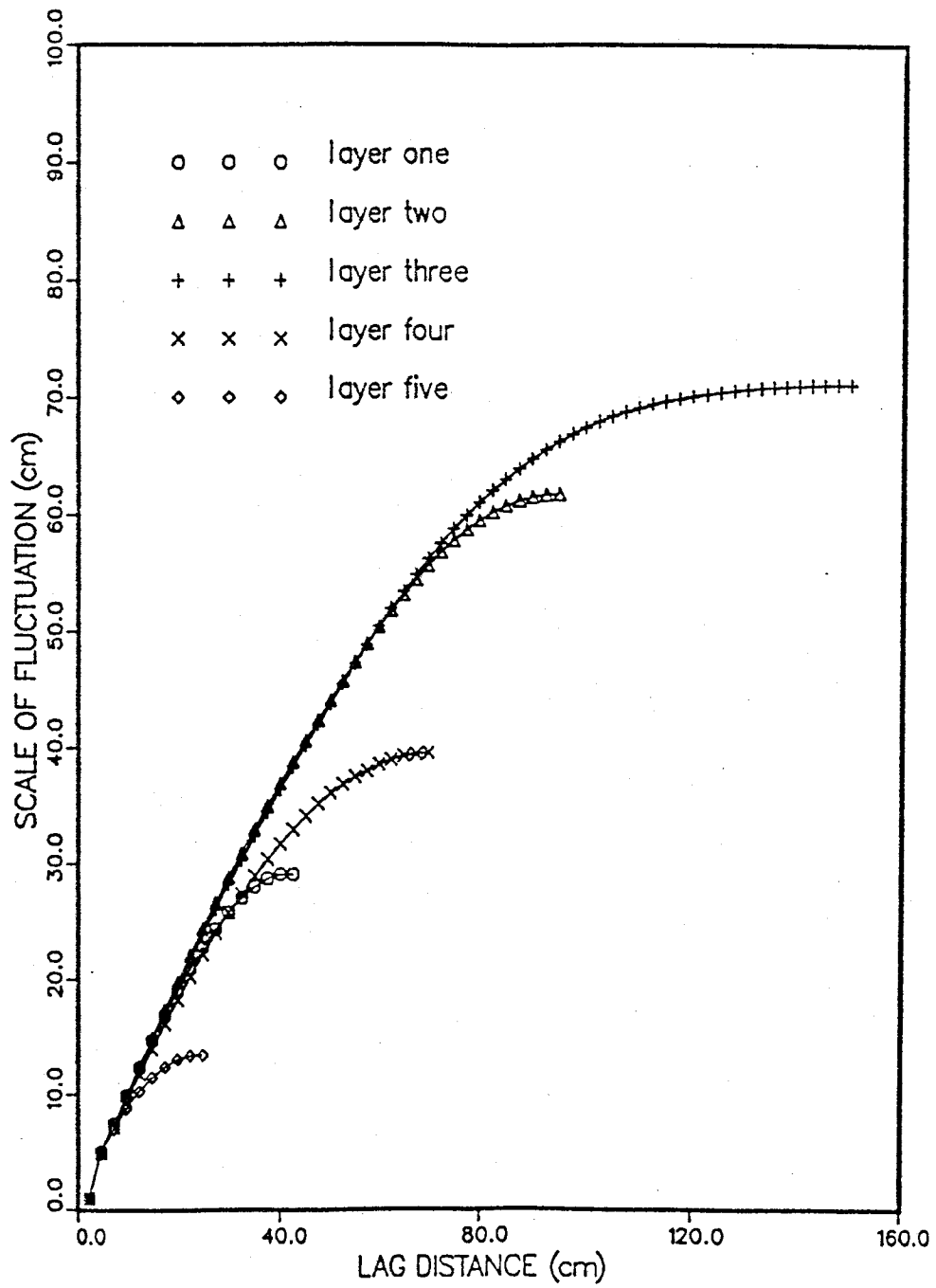


Fig.5 - Scale of Fluctuation for the different layers at McDonald Farm

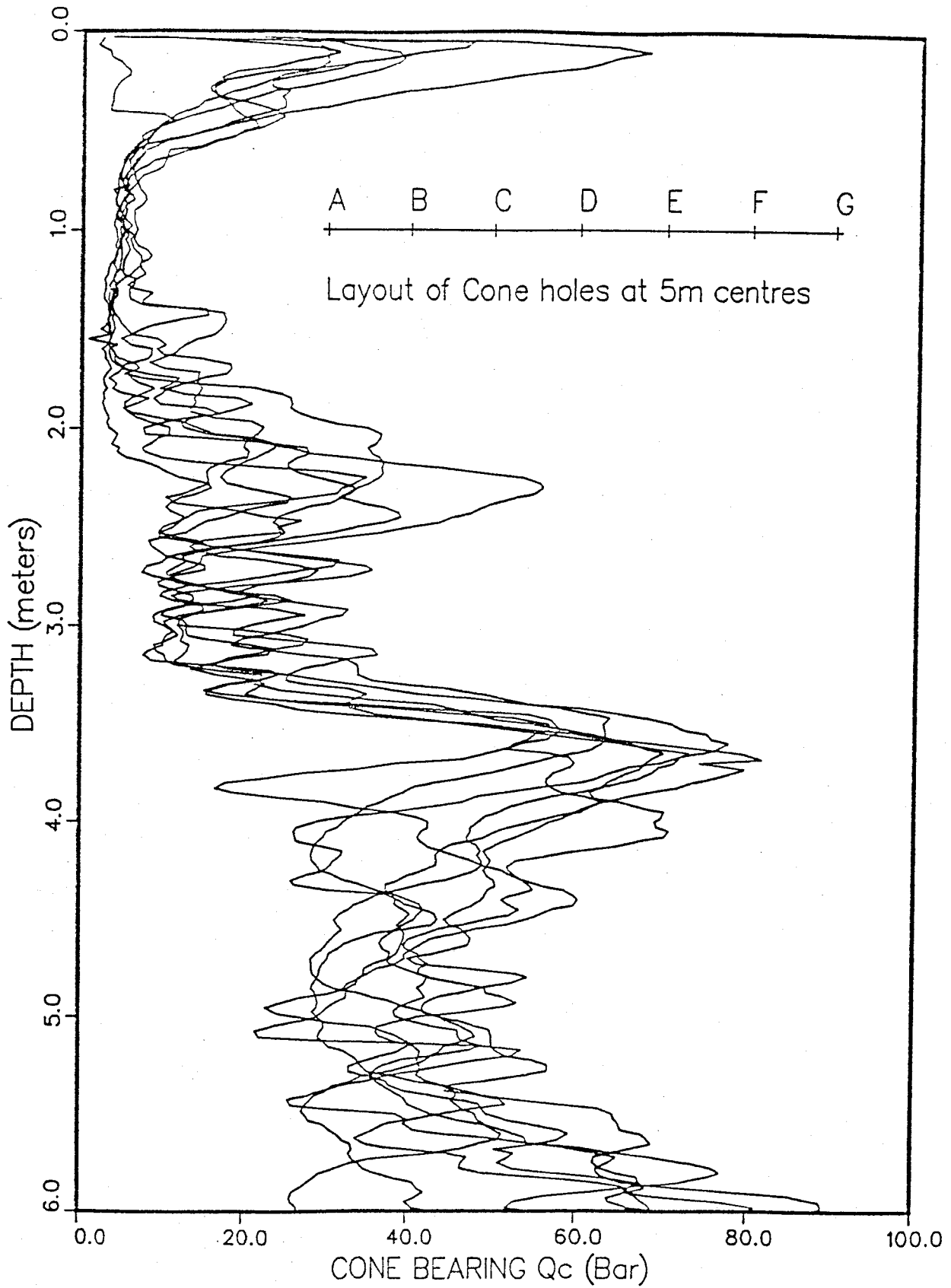


Fig.6 - Variation of Cone Bearing at locations A through G as illustrated above.

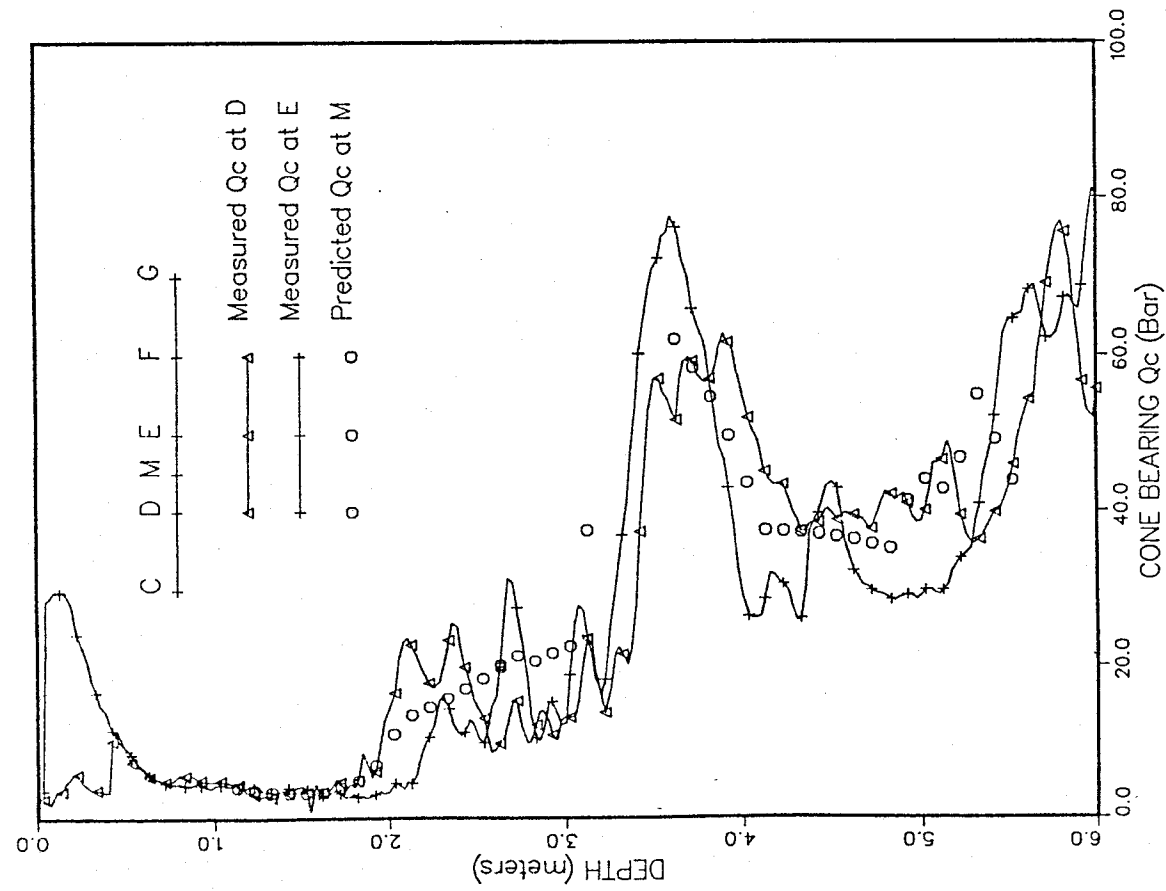


Fig.7 - Interpolated Cone Bearing profile at M and measured values at D and E.

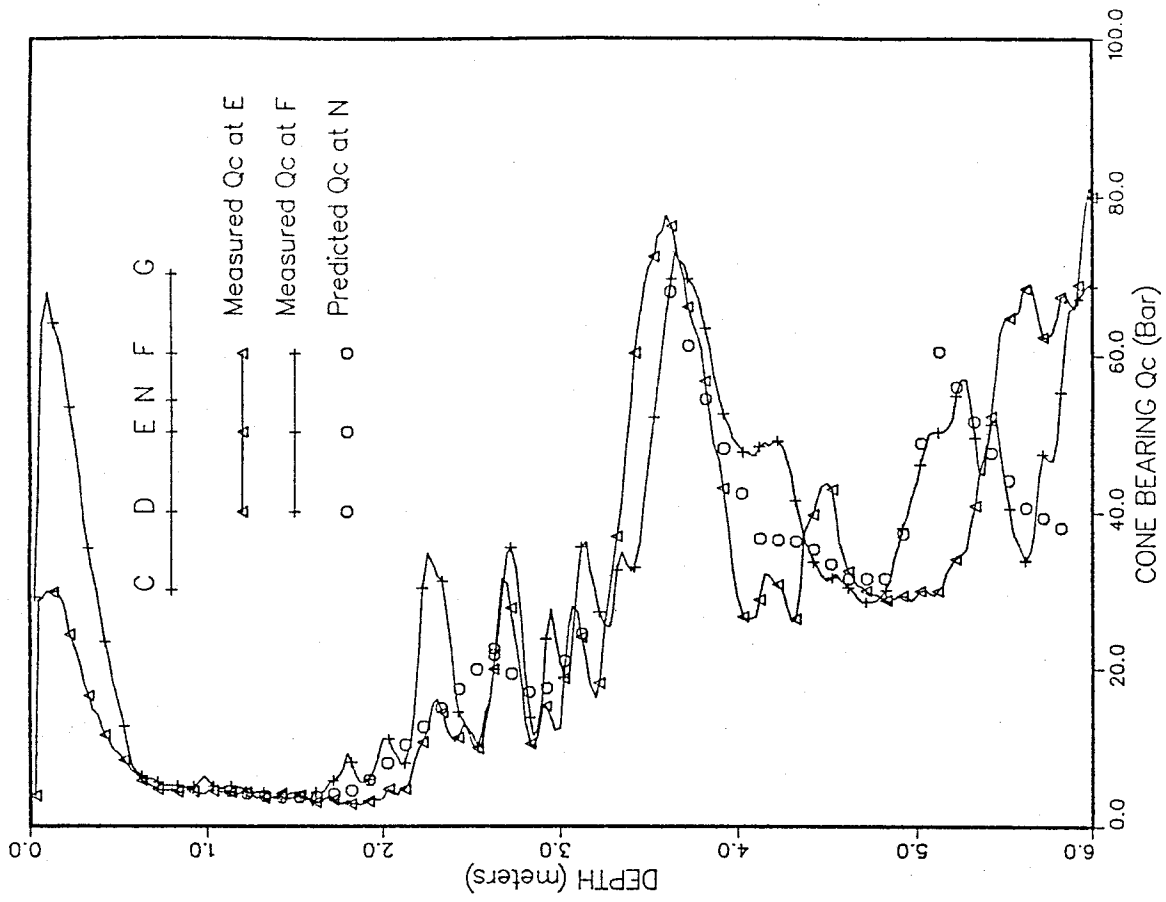


Fig.8 - Interpolated Cone Bearing profile at N and measured values at E and F.