

## STATISTICAL TREATMENT OF CONE PENETROMETER TEST DATA

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

## ABSTRACT

Several statistical techniques have been applied to cone penetrometer test (CPT) data. The analysis studies the variability of cone profiles, identifies different types of layering and performs trend analysis. Autocorrelation and spectral analysis have been performed on the data. Methods of ascertaining the stationarity or non-stationarity of data are discussed and ways of stationarizing data are also presented. The distance of perfect correlation for the CPT data is obtained using the variance function while highlighting its importance.

## INTRODUCTION

The natural heterogeneity of the soil, the limitation of data availability, soil disturbance and measurement errors are the basic kinds of uncertainty in a soil profile modelled from a site investigation. It should be emphasized that there is nothing random about soil properties if every point in the ground could be tested. However, this is not practical and economical and, therefore, only selected areas are tested and hence arises the problem of limitation of data availability in soil investigations. These uncertainties provide ample evidence as to why the modelling of the stochastic character of soil properties is important in geotechnical practice. Presently, continuous cone penetrometer test (CPT) soundings are essentially used as a tool to identify different soil layers and as an approximate predictor of soil strength properties, such as cohesion, friction angle, relative density, etc., all on a deterministic basis. The almost continuous data that is available from the CPT could be most efficiently used, yielding important information if the random nature of the data variation is recognized from a probabilistic point of view. A comparison of the continuous cone profiles obtained from adjacent boreholes provides sufficient evidence as to the spatial randomness of soil properties and it is amply evident that any attempt at a deterministic description is far from accurate. Statistical methods in the analysis of soil data is therefore of great value to geotechnical engineers as it enables many of the uncertainties to be quantified in terms of probabilities.

The analysis of CPT data presented herein, discusses methods of removing anomalies from cone profiles and also cautions against problems resulting from such filtering. The variability of the soil is assessed using the coefficient of variance and the different types of layering in the sand is identified. Trend lines are determined from a simple regression analysis and are used to identify different layers. Various methods of removing trends from non-stationary data are also presented while also describing a statistical test for verifying stationarity or homogeneity of soil records. Autocorrelation and spectral analysis are performed on the data after trend removal while highlighting some of the uses of such an analysis. The autocorrelation function obtained for the trend removed data is also fitted by an exponential trigonometric function which could be used for analytical purposes if necessary. The distance of perfect correlation or the scale of fluctuation has been determined for the different layers using the variance function to quantify soil variability.

The main aim of this paper is to demonstrate and discuss the various statistical methods that can be applied to data obtained from the cone penetrometer in order to gain a better understanding of the variation of soil properties with depth.

## STATISTICAL ANALYSIS OF McDONALD FARM DATA

All data analyzed herein have been obtained from the McDonald Farm in-situ research site of the University of British Columbia, near the Vancouver International Airport on Sea Island. CPT data from several probings, extending to more than 30 meters, are available. These soundings include data such as end bearing, sleeve friction, excess pore pressure on the face and behind the face, temperature and inclination. The soil stratification is basically sands of different density to 15 meters with clay and silt extending below it. The data has been acquired at 2.5 cm intervals using a cone with an area of 10 sq.cm. and a sleeve area of 150 sq.cm., penetrating at 2 cm. per second. A general discussion of CPT testing at the McDonald Farm research site is given by Campanella et al., 1983. The analysis in this paper will only deal with the cone bearing data in the sand layer to a depth of 15 meters (Fig. 1).

## FILTERING

It is possible that the soil being tested includes anomalies, like very thin lenses of clay or sand, or pockets of gravel and it may be necessary to remove these from the profile. A high CPT bearing value representing a thin sand lens cannot be made use of in a foundation design but knowledge of its existence will be very important in determining the driveability of a pile or caisson. Therefore, it is apparent that filtering of these anomalies will have to be handled with care, and the degree of filtering is highly situation dependent. Filtering can be performed based either on the mean or median but the median method is highly recommended (Vivatrat, 1978) since the cutoff point which is based on the median is not dependent on the very high peaks and low troughs of the data, unlike the method based on the mean. In the above method the soil is divided into sublayers of 25 cm. and the median  $M_i$  and standard deviation  $S_i$  of each layer are calculated and for each sublayer  $i$ , a standard deviation  $S_d$  is defined as,

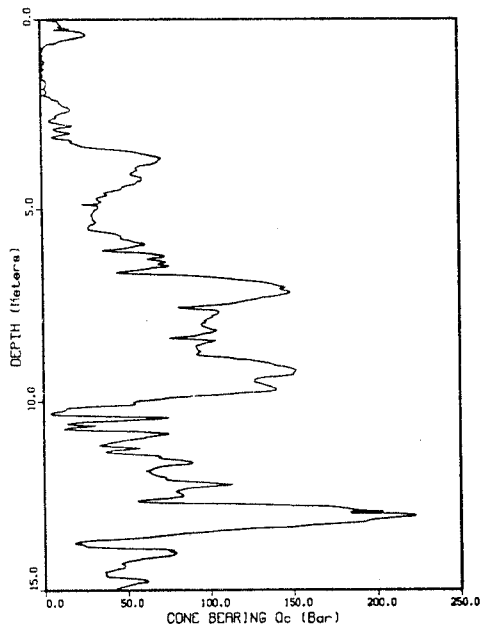


Fig. 1. Cone Bearing Profile

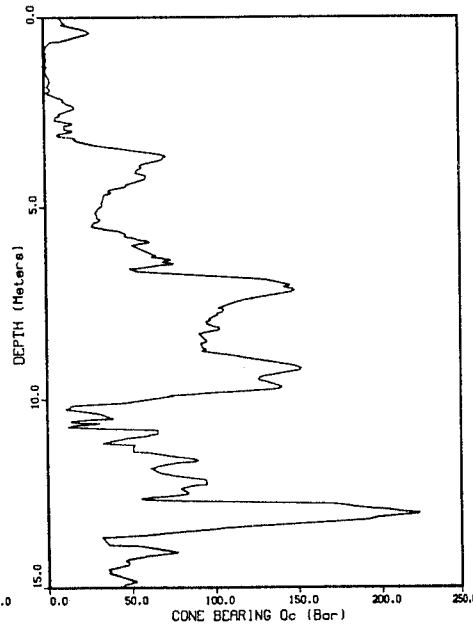


Fig. 2. Filtered Cone Bearing Profile

$$S_d = \text{Minimum of } \left\{ \frac{1}{2}(S_i + S_{i+1}), \frac{1}{2}(S_{i-1} + S_{i+1}), \frac{1}{2}(S_i + S_{i-1}) \right\} \quad (1)$$

Filtering of data in a particular sublayer was performed by removing data outside the region  $M_d \pm S_d$ . The filtered CPT profile in Fig. 2 can be compared to the unfiltered data in Fig. 1 and it is evident that the filtering has resulted in the removal of several extreme values. It is important that the sublayers considered above are not very thick, so that the assumption of stationarity in a thin region is reasonable and justifiable even in a soil exhibiting a trend. Ideally, filtering should only be used on stationary or trend removed data. The effect of filtering is usually not very significant in sand but is often especially effective in clay. Moving average methods may also be used for filtering purposes, but should be used with caution and only as a rough guide to remove unwanted peaks or obvious noise in the data.

## VARIATION OF SOIL PROPERTIES

To assess the variability of the soil the coefficient of variation was calculated, dividing the soil layer into sublayers of 25 cm. The coefficient of variation is defined as the ratio between the standard deviation and the mean of a set of data. Once again, it is recommended that these sublayers be as thin as possible so that the range of soil properties within a layer is not very high and any trends are a minimum. It is also important that each of these sublayers comprise at least ten data points so that the coefficient of variation provides a reasonable picture of the actual variation. As can be seen from Fig. 3 the soil between 5 to 10 meters is relatively uniform with a low coefficient of variation, while the soil between ground level and 5 meters and also between 10 and 15 meters depicts erratic variation.

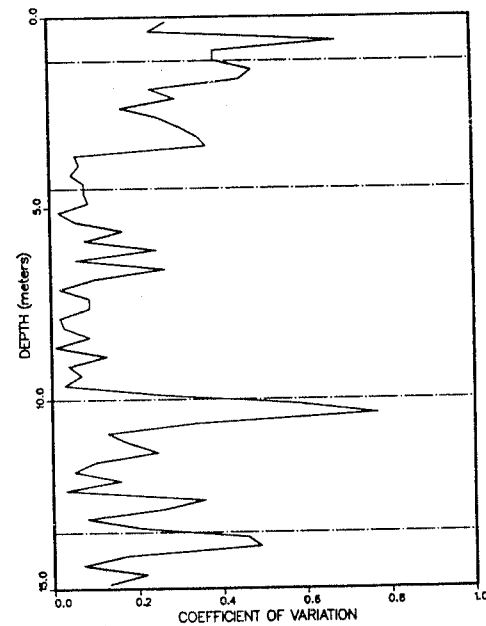


Fig. 3. Distribution of Coefficient of Variation with Depth.

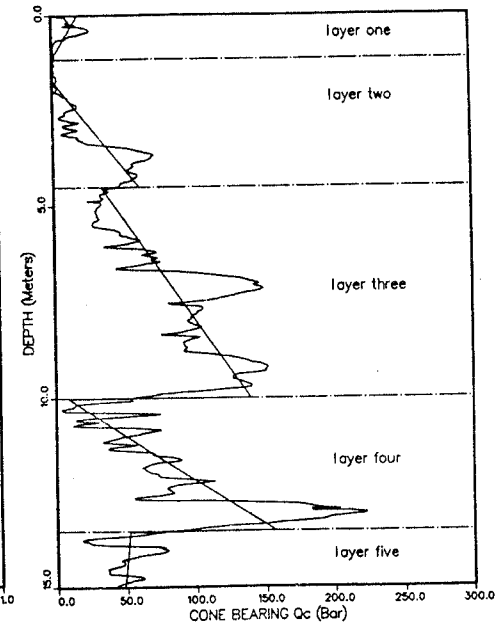


Fig. 4. Trend Lines of the Different Layers.

## TREND ANALYSIS

A visual inspection of the cone profile (Fig. 1) and the variation of the coefficient of variation with depth (Fig. 3) gives an indication of the different types of soil layers existing in the 15 meters. In the McDonald Farm site data five different types of layers were visually identified in the top 15 meters, as follows; Layer 1 from 0 to 1.15 meters, layer 2 from 1.15 to 4.50 meters, layer 3 from 4.50 to 10.0 meters, layer 4 from 10.0 to 13.5 meters and layer 5 from 13.5 to 15.0 meters. The very high coefficient of variation from 9.5 to 11 meter depth could be the basis for choosing another, rather thin, layer. Trend lines are obtained for each of these layers using regression analysis. Layer 1 consists of fill material and exhibits a negative gradient, or decreasing cone bearing with depth. As illustrated in the figure there is a sudden drop in the cone bearing value at 10 meters, and thereafter increases with depth suggesting that layer 4 belongs to a different depositional period. Layer 5 is evidently the transition layer preceding the clay at 15 meters. Once the layering is identified each of these layers can be analyzed individually. Most of the analysis to follow deals with the layer 3 data and was felt most appropriate since it is the thickest layer, containing the highest number of data.

It is evident that the data shows a linear trend and therefore the data are non-stationary. In the simplest terms if the data has a non-constant mean it is termed non-stationary. Soil properties are highly dependent on depth and therefore although data is stationarized, it is also very important to study the variation of soil properties with depth. In this respect the most appropriate tool in overcoming non-stationarity is trend removal using regression analysis. The resulting residuals from such an analysis are the stationary variables while the deterministic trend is the regression line obtained. To eliminate the linear trend of our data the trend component was estimated by the classical linear model of least squares. The trend lines obtained after regression are illustrated in Fig. 4, for the different layers identified. The estimate of the trend will only be approximate, if the residuals are autocorrelated. However, the improvement to the linear regression model by considering autocorrelated variables is minimal if the number of data points exceed about one hundred (Davis, 1978).

## TEST FOR STATIONARITY AND TREND REMOVAL

A statistical test known as the "RUN" test could be used to appraise the homogeneity of data (Alonso and Kriezek, 1975). RUN is defined as a sequence of events of the same type. The criteria chosen here is comparing the mean of the selected thickness against the global mean of the entire layer. There are two types of events possible; the local mean of a selected thickness being below or above 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". In applying this test to layer three data, the global mean was calculated to be 87 bar or 8.7 MPa. The soil layer was divided into sub-layers of thickness 25 cm and their local means with respect to the global mean and the number of runs are indicated in Fig. 5. The number of runs was obtained as three and it is apparent from tables relating the number of runs required for homogeneity for different levels of significance (Swed and Eisenhart, 1943), that the soil cannot be considered homogeneous even at very low significance levels. This result was as expected for a soil layer exhibiting a trend.

The general model of a cone profile can be expressed as,

$$\text{DATA} = \text{TREND} + \text{RESIDUAL} \quad (2)$$

To perform any residual analysis of the data the trend of the data will have to be removed, transforming the non-stationary data to stationary residuals. The most efficient method of removing the trend from data is by performing a regression analysis and this has already been discussed. The other methods which are used to stationarize data are standardizing and differencing.

Standardizing is performed by subtracting the mean of the data and dividing by the standard deviation, which results in a variable with zero mean and unit standard deviation. The standardized variable  $(X_s)_i$  is given by

$$(X_s)_i = (X_i - X_{av})/\sigma \quad (3)$$

where  $\sigma$  and  $X_{av}$  are the standard deviation and mean respectively, of the entire layer.  $X_i$ 's are the data of the layer. The technique of standardizing to stationarize data, was used in the evaluation of the variance function, explained in a later section.

Differencing is another technique by which a trend could be removed. While a first degree differencing would eliminate a linear trend, higher degree differencing is required to eliminate trends of higher order. If  $X_i$  and  $X_{i-1}$  are adjacent data recordings exhibiting a linear trend, the stationarized and differenced data  $(X_d)$  are given by,

$$(X_d)_i = X_i - X_{i-1} \quad (4)$$

As can be seen from the above equation, differencing transforms the data, completely and loses continuity with the original data which is not desirable.

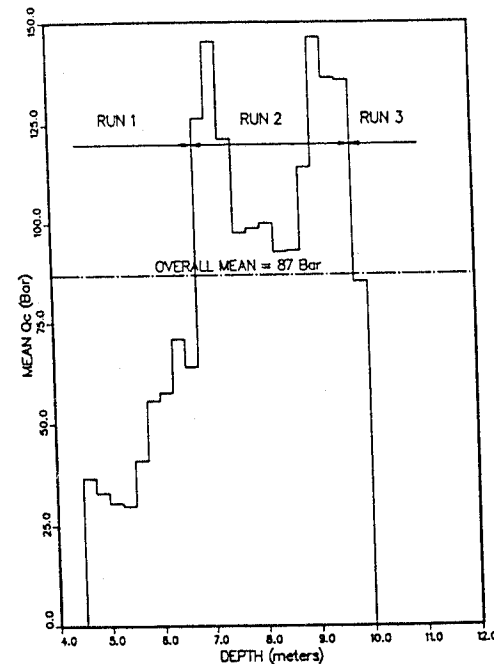


Fig. 5. Distribution of RUNS for the Data in Layer 3

Therefore, the more important purpose of differencing is to determine the level of differencing required to stationarize a set of data. The level of differencing required could be determined by evaluating R from Eq. (5) as follows;

$$R = \frac{N/6}{\sum_{k=0} \rho_k} \quad (5)$$

where,  $\rho_k$  is the autocorrelation function which will be described in the next section and N the number of data. R should be calculated for different levels of differencing 'd'. The value of 'd' at which R begins to increase, is the level of differencing required to stationarize the data.

#### AUTOCORRELATION FUNCTION AND SPECTRAL DENSITY FUNCTION

An autocorrelation function which is not solely dependent on the lag or the spacing of data points considered, but also on the position of the data also leads to non-stationarity. In the event that the presence of a trend is not apparent from the data the autocorrelation function and the spectral density function can be used to verify a trend. A linear autocorrelation function or a spectral density function with a non-zero intercept confirms the presence of a trend. The autocorrelation at a lag k is defined as,

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

where the autocorrelation is obtained for different lags k. Lag is the spacing between data points considered. For example when k=1 the interval between data points considered is equal to 2.5 cm. For k=2 the interval is equal to 5 cm, since the interval of CPT data recorded is 2.5 cm. The resulting variation of the autocorrelation versus lag k is termed the autocorrelation function ( $\rho$ ). N is the number of data and  $X_{av}$  is the mean of the data  $X_i$ .

The spectral density function (S) is the Fourier transform of the autocorrelation function and is defined as,

$$S = 2 \int_{-\infty}^{\infty} \rho_k \exp(-i\omega k) dk \quad (7)$$

where,  $\omega$  is the angular frequency.

The spectral density function and the autocorrelation function of the residuals of layer 3 are illustrated in Figs. 6 and 7 respectively. It can be observed from Fig. 6 that the trend removal has resulted in a spectral density function which has a zero intercept. The power spectral estimate is an indication of the frequency content of the data or the variability of the data, a higher frequency meaning a more variable soil. For the detrended data the dominating frequencies are between 0 to 1.35 cycles/meter with the energy content being more or less uniformly distributed as could be expected for a random process. The frequencies in the above range are quite uniformly distributed and are attributed to the fact that the soil in layer 3 is of a low variability as was found from the coefficient of variation diagram (Fig. 3).

Several forms of autocorrelation functions have been suggested by Vanmarcke (1978) to express autocorrelations in soil properties. The decaying fashion of the stationarized data of layer three as shown in Fig. 7, suggest an autocorrelation function of the exponential trigonometric form,

$$\rho_z = \exp(-a|z|) \cos(2\pi bz) \quad (8)$$

where a and b are constants and z the distance of correlation.

The approximate curve with parameters  $a = 1.18 \text{ meter}^{-1}$  and  $b = 0.419 \text{ meter}^{-1}$  is also illustrated in Fig. 7. Although layer three has 220 data points (4.5 meters) the autocorrelation function has been calculated only up to 90 lags (2.25 meters) since the accuracy of it decreases with increasing number of lags in excess of one fourth of the total number of data (Davis, 1973). That is, it is exact only up to an interval of 1.37 meters. However, the approximate curve appears to be valid even in the interval range of 2.25 meters and there seems to be no loss of generality. The approximate curve enables the analytical treatment of the autocorrelation function and if greater accuracy is required, a more accurate form could be developed.

#### SCALE OF FLUCTUATION

The scale of fluctuation  $\delta$ , or the distance of perfect correlation could be defined in terms of the autocorrelation function;

$$\delta = \int_{-\infty}^{\infty} \rho_z dz \quad (9)$$

Vanmarcke has also suggested a method of deriving the scale of fluctuation from the variance function which expresses the decay of the standard deviation with increasing number of data. The variance function F, is defined as

$$F = (\sigma_z/\sigma)^2 \quad (10)$$

where  $\sigma_z$  is the standard deviation of a sublayer of thickness z and  $\sigma$  is the standard deviation of the layer.

The procedure in obtaining the variance function is as follows. The data is first considered in pairs (n=2) and a moving average series for the data is obtained which is for n equal to two or z equal to 5 cm. The standard deviation of this series is then calculated. Thereafter, a moving average series is obtained with data considered in clusters of three (n=3), or z equal to 7.5 cm,

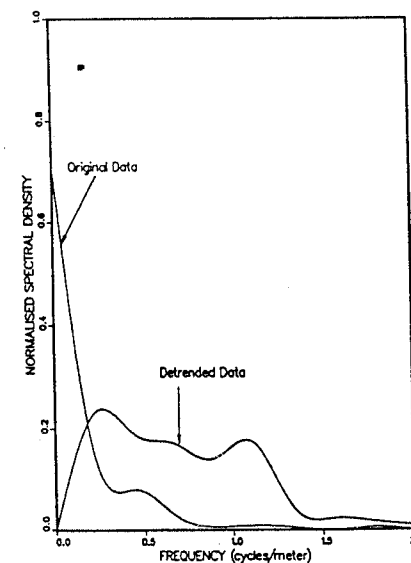


Fig. 6. Spectral Density Function.

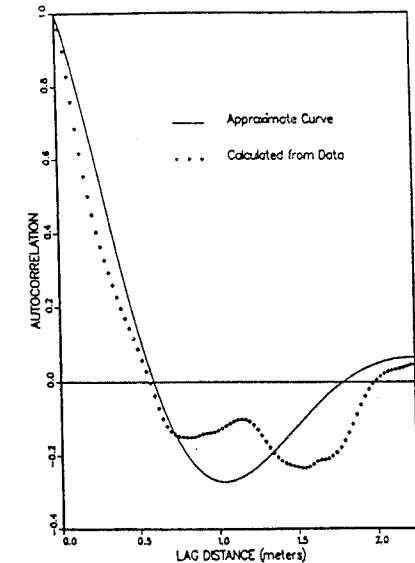


Fig. 7. Autocorrelation Function.

and the standard deviation of this series is calculated. The above procedure is continued for increasing  $n$ , and for each  $z$  the variance function can be obtained from the definition already given. It is obvious that with increasing  $z$ , the moving average series will be less dispersed resulting in a decreasing standard deviation and thus a decreasing variance function. The variance function  $F$  has a maximum value of unity. When  $z$  is very large the value of  $F \cdot z$  will approach the value of the scale of fluctuation. For the different layers of the CPT profile the variation of  $F \cdot z$  is illustrated in Fig. 8,  $F$  increases to a maximum and then drops off when the decrease in  $F$  is significant in comparison to the increase in  $z$ . The scale of fluctuation is equal to the maximum value of  $F \cdot z$ , and its value for the five different layers are obtained as layer 1 - 29 cm, layer 2 - 62 cm, layer 3 - 71 cm, layer 4 - 40 cm, and layer 5 - 13 cm (Fig. 8). Layer 3 has the lowest variability as observed from the low coefficient of variation (Fig. 3) and the high scale of fluctuation of 71 cm. confirms this. Thus, sampling in this layer could be as large as 70 cm. Layer 5 has the largest variability and requires a sampling interval of at least 13 cm.

The importance of the scale of fluctuation is apparent when two different test methods are being compared. It is also recommended that the sampling distance should be less than the scale of fluctuation (Vanmarcke, 1978) 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 the scale of fluctuation is advisable.

**ARIMA MODEL FITTING.** Methods of autoregressive integrated moving average (ARIMA) model fitting were used on the McDonald Farm CPT data according to Wu (1985). The ARIMA was fitted to the data in order to obtain the autocovariance function and the random testing error. The random testing error for the layer 3

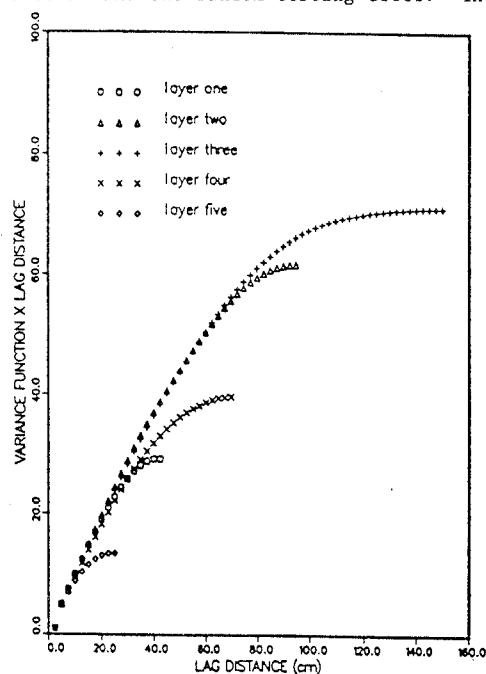


Fig. 8. Scale of Fluctuation of the Different Layers.

data was as low as one percent and this reflects the consistency of CPT results. Subsequent to the publication dealing with ARIMA model fitting by Wu (1985), he has questioned the validity of the assumption that the autocorrelation is independent of the random error, and this shortcoming must be studied further.

#### CONCLUSIONS

1. If filtering is performed on CPT data to remove anomalies, it has to be done in a way that important information is not lost and in this sense the method of filtering is highly situation dependent. The median method of filtering is recommended since the cut off point which is based on the median is not affected by very high or very low values, unlike the method based on the mean.

2. The "RUN" test has been used to determine the non-homogeneity or the non-stationarity of the data. Even at very low significance levels, the number of runs required for homogeneity of layer 3 data is more than the number of runs obtained in the analysis, which was as expected since the data exhibits a trend.

3. Non-stationary data can be transformed to stationary data, i.e., trend removal, using standardizing, differencing and regression analyses. The method of trend removal by regression analysis seems the most suitable since the resulting trend also gives an idea of the variation of soil property with depth. The resulting residuals can be further studied using autocorrelation and spectral analysis, yielding important information which can be related to the variation of soil properties.

4. The autocorrelation function of the data was fitted using a theoretical autocorrelation function of the exponential trigonometric form,  $\rho_z = \exp(-a|z|) \cos(2\pi bz)$  with  $a = 1.18 \text{ m}^{-1}$  and  $b = 0.419 \text{ m}^{-1}$ . This theoretical form could be used for analytical purposes like obtaining the scale of fluctuation.

5. The scale of fluctuation or the distance of perfect correlation obtained for the different layers were, layer 1 - 29 cm, layer 2 - 62 cm, layer 3 - 71 cm, layer 4 - 40 cm, and layer 5 - 13 cm. Layer 3 has the highest scale of fluctuation and appropriately the lowest variation. When two different test methods are compared it is important that the sampling distance be less than the scale of fluctuation. For optimum sampling benefit using the same test method, spacing should be at least equal to the scale of fluctuation.

#### ACKNOWLEDGEMENTS

The financial support of the Natural Sciences and Engineering Research Council, Canada and the International Center for Ocean Development is gratefully appreciated.

#### REFERENCES

- Alonso, E.E. and R.J. Krizek. (1975). "Stochastic Formulation of Soil Properties", Proc. Second ICASP, Aachen, Germany, Vol. 2, pp. 10-32.
- Campanella, R.G., P.K. Robertson and D.J. Gillespie. (1983). "Cone Penetration Testing in Deltaic Soils", Canadian Geotechnical Journal, Vol. 20, No. 1, pp. 23-35.
- Davis, J.C. (1978). Statistics and Data Analysis in Geology, Wiley
- Swed, F.S. and C. Eisenhart. (1943). "Tables for Testing Randomness of Grouping in a Sequence of Alternatives", Annals of Mathematical Statistics, Vol. 14, pp. 66-87.
- Vanmarcke, E.H. (1977). "Probabilistic Modelling of Soil Profiles", ASCE, Vol. 103, No. GT11, pp. 1227-1246.
- Vanmarcke, E.H. (1978). "Probabilistic Characterization of Soil Profiles", Proc. ASCE - Specialty Workshop on Site Characterization and Exploration, Northwestern University, Illinois, USA, pp. 199-219.
- Vivatrat, V. (1978). "Cone Penetration in Clays", Doctoral Dissertation, M.I.T., USA.
- Wu, T.H. (1985). "Use of Time Series in Geotechnical Data Analysis", Geotechnical Testing Journal, ASTM, Vol. 8, No. 4, pp. 151-158.