Discussion on Abnormal Values in Static Penetration Data

Liu Xue-xiang*, Wu Dao-xiang
School of Resources and Environmental Engineering, Hefei University of Technology, Hefei 230009, China
*Corresponding author’s e-mail: 407978350@qq.com

ABSTRACT: As the mainstream method of in-situ testing technology, static penetration testing is widely used. At present, the implementation of a large number of projects must be based on its test results. It also makes us have to pay attention to the reliability of its test results. In this paper, based on the previous experience and combined with the static penetration data in Hefei area, the reliability of test data is studied. The abnormal data in the static penetration data is eliminated and the influence of the abnormal data on the calculation results of correlation distance is expounded.

1. Foreword
In daily geotechnical engineering projects, more or less some abnormal data may appear when we measure the basic properties of some basic physical mechanical of the soil layer, which will affect our test results. The abnormal data in the project do not belong to the research object, for they may be the experimental error or caused by some other internal factors. Abnormal data must be identified and deleted before processing the data[4,5]. Therefore, it is necessary to use scientific methods to screen test data for reliability analysis.

Vanmarcke put forward to the Random Field Modeling of Soil Property profile in 1977. It used random function of one or more variables to handle soil parameters,and put forward to the concept of "correlation distance", another index describes the correlation of soil property parameters. Guo Linping and some other people [7] have conducted research on determining the correlation distance around the wave function through the (CPT) test in Lingang, Tianjin. They have analyzed the factors affecting the autocorrelation distance item by item. In addition to the linear regression method (a detrending method) used in the processing of the static penetration data, in some cases, nonlinear regression can be also used to detrend the influence (whose effect is better than linear regression), but this circumstance shall not appear more than twice. Meanwhile, the selection of linear regression and nonlinear regression should be based on the actual situation.

After synthesizing the thoughts of predecessors, correlation distance is an important index to describe the autocorrelation of soil properties. It can be also considered as another important statistical indicator to describe the statistical characteristics of soil parameters except mean and standard deviation[6]. This paper analyzes the selection and removal of abnormal data in the static penetration data in Hefei, and discusses whether the existence of suspicious value affects the calculation of correlation distance.

2. Discrimination of abnormal data

2.1. Grubbs Method
Suppose that the whole subject of study obeys normal distribution, the random subsamples of the sample X are \( x_1, x_2, x_3, \ldots, x_n \), the average value of \( X \) is \( \bar{x} \) and the standard deviation of the sample is \( s_n \). Subsamples are now sorted from large to small \( x(1) \leq x(2) \leq \ldots \leq x(n) \). Use Grubbs method to screen the abnormal data, the screening data is the abnormal value outside the critical value \( G \).

Critical of the cut vale \( G = \left| \frac{x_d - \bar{x}}{s_n} \right| \) (1)

Among which, \( x_d \) is the data to be tested. \( \bar{x} \) is the average of the sample, \( s_n \) is the standard deviation of the sample, \( G \) is the testing index in Grubbs method.

Suppose that \( x(1) \), \( x(2) \) are the rather small value and rather large value in the above mentioned sort. It can be considered that \( x(1) \), \( x(2) \) are the suspicious sample values in the sample.

Suppose that \( x^{(1)} \) is suspicious, there is \( G^{(1)} \geq \frac{|x^{(1)} - \bar{x}|}{s_n} \), and \( x^{(a)} \) is suspicious, there is \( G^{(a)} \geq \frac{|x^{(a)} - \bar{x}|}{s_n} \).

Select confidence coefficient, otherwise select the error rate of Grubbs method, generally, 5% or 1% is selected. The probability meanings is that when the test data is normally distributed, there is \( P(\bar{x} - \frac{\sigma(x)}{\sqrt{N}} \leq \mu \leq \bar{x} + \frac{\sigma(x)}{\sqrt{N}}) = 1 - \alpha = 95\% \) (or 99\%)

According to the critical value \( T_{a,n} \) under a certain probability and a certain sample size, refer to GT data statistics.

There is \( \Delta x = |\bar{x} - x_i| \); There is \( x_{max} = \bar{x} + \Delta x \), \( x_{min} = \bar{x} - \Delta x \); So, the normal values in this sample satisfy the following relationships [4]:

\( T = \frac{\bar{x} - x_i}{s} \) (3) make

\( \Delta x = |\bar{x} - x_i| \) (4)

According to the critical value \( T_{a,n} \) under a certain probability and a certain sample size, refer to GT data statistics.

There is \( \Delta x = T \times s \); There is \( x_{max} = \bar{x} + \Delta x \), \( x_{min} = \bar{x} - \Delta x \); So, the normal values in this sample satisfy the following relationships [4]:
When data of the detected subsamples is in this scope, it can be regarded as normal value; otherwise, it is abnormal value.

3. Detrending analysis

Using random field theory, the premise of the spatial variability of soil is to transform the original data into a stationary or weakly stationary random sequence with 0 mean. The soil parameter function $g(z)$ obtained by analyzing the soil parameter according to the environment $g(z)$ can be divided into two parts:

$$g(z) = r(z) + h(z)$$

In the formula:

$Z$ is depth; $r(z)$ is random function that represent the inherent variability of the soil, $h(z)$ is trend component.

We need stationary random sequence, which is the premise of reliability analysis by random field theory. Detrending is a common method to get stationary random field, which avoids overestimation variability caused by raw data.

3.1. Linear

This paper collects test data of static penetration holes in Hefei and uses it as an example of reliability analysis of geotechnical engineering.

Original static penetration curve is shown in the figure, and a smooth random data with a mean of 0 will be get after detrending. Firstly, we draw a linear trend line in the original static penetration curve, the fitting results are shown in Figure 1, its correlation $R^2$=0.3999.

![Fig1. Linear fitting results](image1)

![Fig2. Curve fitting results](image2)

3.2. Nonlinear

For the above-mentioned data of static hole 1, we draw a quadratic nonlinear trend line according to the requirements of quadratic regression fitting and the fitting results are shown in Figure 2, its correlation $R^2$=0.5891.

The following tale is the static penetration data of the same project, two methods are used on the results of the detrended analysis.

| No. of static penetration hole | Detrending method | Average value | Standard derivation |
|-------------------------------|------------------|---------------|---------------------|
| JT1                           | Linear           | 0             | 0.73                |
|   | Nonlinear | Linear | Nonlinear |
|---|-----------|--------|-----------|
| JT2 | 0 0.63 | 0 0.70  |
|    |           | 0 0.69 |           |
| JT3 | 0 0.29 |        | 0 0.24 |
| JT4 | 0 0.34 |        | 0 0.18 |

Compare the two detrending methods, we can see from Table 1 (the No. in the Table is temporary number) that the standard derivation obtained through nonlinear detrending method (that is quadratic nonlinearity) is obviously smaller than that obtained through linear detrending method. Therefore, the detrending method should be selected according to the actual situation. Not all cases should choose to use the linear detrending method. If higher nonlinear are adopted when using nonlinear detrending method, the standard deviation of random component will be reduced and the correlation coefficient will be increased after detrending, the corresponding mean is no longer 0, and its stationarity will change (which cannot reach the preconditions of the random field theory). And the higher the frequency, the greater the difference between the mean of the random components and 0. It is recommended to use the quadratic nonlinear fitting result when detrending the test data[7].

The abnormal value of the static penetration test result is mainly caused by the following aspects:
1. Thin layer in the soil layer.
2. Transition data between two different layers (leading or lagging data caused by interface effect).
3. The variability of soil is very large, and it is inevitable that local small variation will occur., which is one of the source for the occurring of abnormal data.

The appearing of the above situation in the soil layer is normal, but this is not performed by the basic nature of the soil layer we studied. These data are not the data of the soil to be studied, they shall be processed as abnormal data. According to mathematical statistics, these data are not conform to the requirements of the stationary and individuality of soil random field. Therefore, these data must be removed when studying the reliability of the soil layer.

Figure 3 is the static penetration curve of an old clay soil layer. It can be seen from curve A that the point marked " " at the end of the curve does not belong to the soil layer of the old clay to be studied, remove it immediately and get the curve of Fig. 3(B). Use Grubbs method to calculate and examine the curve, we know that the point marked " " in figure 3(B) is the abnormal data through calculation, it can be deleted to get curve 3(C) without abnormal data. The mean lines of FIG. 3(A),(B) and (C) are respectively:
Fig3. CPT abnormal data deletion comparison chart

\[ y = 0.7929x - 0.2158 \quad R^2 = 0.3283 \quad (A) \]
\[ y = 0.4776x + 1.1505 \quad R^2 = 0.5675 \quad (B) \]
\[ y = 0.3686x + 1.6117 \quad R^2 = 0.739 \quad (C) \]

The standard derivation are respectively 1.458MPa, 0.6496MPa and 0.4147MPa. It can be seen that whether abnormal data is deleted or not has a great impact on the calculation results of mean line and standard deviation.

4. Autocorrelation distance

Autocorrelation distance (\( \delta \)) refers to the minimum distance between any (any is for the soil space) two unrelated points in a soil layer.

The closer the distance between two points is, the greater the correlation is, and the farther the distance is, the smaller the correlation is, which can also be explained as a critical distance. The soil property index within this range is basically related; on the contrary, the soil index outside this range is basically irrelevant. In combination with domestic and foreign literatures, the solutions of autocorrelation distance mainly include average span method, recursive space method, correlation function method and other methods. In this paper, the correlation function method is introduced, as follows:

Calculate the correlation distance \( \delta \) through correlation function \( R(h) \):

1. Select different \( j \) value for \( h = j \Delta z \), calculate the autocorrelation function \( R(h) \) of the sample value:

\[
R(h = j \Delta z) \approx \frac{1}{s^2(n - j - 1)} \sum_{i=1}^{n-j} x(z_i)x(z_{i+j})
\]

Among which:

- \( h \) is the estimated distance; \( \Delta z \) is sampling interval; \( z_i = i \Delta z \), \( z_{i+j} = (i+j) \Delta z \) are the depth coordinates of the \( i \) and \( i+j \) observation point; \( x(z_i) \), \( x(z_{i+j}) \) is the measured value at the depth of \( z_i \), \( z_{i+j} \); \( n \) is sample size; \( s^2 \) is the quadratic difference of the sample, when \( j = 0 \), \( R(h) = 1 \).

2. Use calculation points to draw \( R(h) \sim h \) figure, select suitable mathematical function model to fit it, and the correlation function model and autocorrelation distance \( \delta \) of soil parameters can be obtained. The expressions commonly used to describe the geotechnical parameter correlation function model and the corresponding autocorrelation distance is shown in table 2.
Correlation distance is an index of soil property parameter description like standard deviation and mean value. Abnormal value shall be removed when calculation correlation distance. Here we use the correlation function method to calculate the correlation distance after the removal and before the removal, compare them to find the influence of abnormal data on the calculation of correlation distance. Combined with the static penetration data of Hefei’s geotechnical engineering project (the No. is the temporary summary number), the figure below is calculated and sorted out, and it is a comparison table of the influence of the remove of abnormal data on the calculation of relevant distance.

### Table 2. Common Functional Models and Relevant Distances [8,12]

| Type of the function | Formula | Correlation distance $\delta$ |
|---------------------|---------|-------------------------------|
| SNX                 | $R(h) = e^{-|h|}$ | $\delta = 2/\lambda$ |
| SQX                 | $R(h) = e^{-a|h|^2}$ | $\delta = \sqrt{\pi}a$ |
| CSX                 | $R(h) = e^{-b|h|} \cos(bh)$ | $\delta = 1/b$ |
| SMK                 | $R(h) = (1 + d|h|)e^{-d|h|}$ | $\delta = 1/d$ |
| BIN                 | $R(h) = \begin{cases} 1 - c|h| & (|h| < 1/c) \\ 0 & \text{(other)} \end{cases}$ | $\delta = 1/c$ |

Correlation distance is another indicator to describe the properties of soil layers. The abnormal data in physical concept does not belong to the soil layer; therefore, these abnormal data must be removed. The data of JT13 hole contains it. For the local variation in the soil layer (which is caused by the anisotropy of the soil) is inevitable, so, the test error and local variation that is difficult to distinguish shall be determined by specific circumstances. As can we can seen from Table 3, in many cases, the correlation distance does not change much after removing abnormal value. When there is a small amount of abnormal value in the curve, and the deviation between such abnormal value and the mean

### Table 3. Calculating table of correlation distance

| No.  | Correlation distance | Type of the function |
|------|----------------------|----------------------|
|      | not removed | Remove |                     |
| JT1  | 0.1261    | 0.1559  | (SQX)               |
| JT2  | 0.1769    | 0.2402  | (BIN)               |
| JT3  | 0.1432    | 0.167   | (BIN)               |
| JT4  | 0.322     | 0.4451  | (BIN)               |
| JT5  | 0.2931    | 0.5248  | (BIN)               |
| JT6  | 0.5394    | 0.4728  | (BIN)               |
| JT7  | 0.5789    | 0.5398  | (BIN)               |
| JT8  | 0.2836    | 0.1804  | (BIN)               |
| JT9  | 0.6715    | 0.5961  | (SMK)               |
| JT10 | 0.2303    | 0.2552  | (SMK)               |
| JT11 | 0.4711    | 0.3706  | (SMK)               |
| JT12 | 0.5525    | 0.6842  | (SMK)               |
| JT13 | 0.0942    | 0.472   | (SMK)               |
| JT14 | 0.1386    | 0.1415  | (SMK)               |
value line is small, the removal has little influence on the calculation result of correlation distance.

5. Conclusions
For the identification of abnormal value in static penetration data, physically (and numerical value) remove abnormal value (deviations) firstly. Secondly, remove the trend component of depth direction on the basis of original data, then, use mathematical statistics method for further screening.

Correlation distance is another parameter of soil property. The influence of abnormal data will make the correlation distance lose efficacy. The abnormal value must be removed before calculating, which is the guarantee of the reliability analysis for geotechnical engineering.

Reference
[1] GBT 3359-2009 ,Statistical Processing and Interpretation of Data and Determination of Statistical Tolerance Interval[S].
[2] Huang,S,Q.Detection of outliers in core correction deviation of Rebounded method based on Grubbs[J]. Fujian Construction Science & Technology, 2018(4).
[3] Cheng,Q. Random Field Theory of Soil Parameters and Its Application in Foundation Engineering[D]. 2000.
[4] Wu,N,Q. Reliability Testing of Engineering Survey Test Data[J]. Journal of the Geotechnical Engineering,ASCE.1991, 13(3):93-98.
[5] Gao,D,Z. Reliability Analysis of Static Penetration Test Results[J]. Engineering Investigation, 1981(2):11-16.
[6] Vanmareke Erik H. Probabilistic Modeling of soil Profiles[J].Journal of the Geotechnical Engineering,ASCE, 1977, 103(GTI 1): 1227-1246.
[7] Guo,L,P. et al. A method of calculating scale of fluctuation of soil and its influence factors based on cone penetration test data[J]. Rock and Soil Mechanics, 2017(S1):277-282.
[8] Li,P. Study on Spatial Variability of Unsaturated Clay in Hefei [D].2016.
[9] Tan,X,H, et al. Study on the calculation methods of the correlation distance of soil property parameters[J]. Journal of Hefei University of Technology(Natural Science Edition), 2004, 27(11):1420-1424.
[10] Dou,J,P . Processing Method of Abnormal Data in Static Penetration Curve[J]. Sichuan Architecture, 2002, 15(4):44-45.
[11] Yan,S,W, et al. Rock and Soil Mechanics[J]. Rock and Soil Mechanics, 2007, 28(8):1581-1586.
[12] Hu,L,Q. Study on Autocorrelation Distance Properties of Soil Layer Based on Static Penetration Method[J]. Journal of north China University of Technology, 2013, 25(1):80-84.
[13] Guo, L,Pet al. Application of spatial average method to estimation of scale of fluctuation of soil indexes[J]. Journal of Engineering Geology, 2017(6):1424-1429.
[14] Zhu,Y,Q,Guo,C,X, .Spatial Statistical Characteristic Analysis of Stationary Random Field of Soil Property[J]. Rock and Soil Mechanics, 2003,24(3):455-457.