Evaluation of tip resistance and side friction based on spectral clustering and hidden Markov chain model

Jiaqi Chang, Lei Fu, Wen Cheng

Tongji University

Question re-description

The question is to calculate the ultimate axial pile capacity (Q_u) of a pile using cone penetration tests, i.e., tip resistance (q_c) and side friction (f_s), of other 6 boreholes around the location of the pile. The location of the pile and other 6 boreholes are shown in Fig. 1. An empirical approach to calculate the Q_u is suggested using q_c and f_s value which can be find in competition question files. The q_c and f_s values of the 6 boreholes are available. Thus, the point is to evaluate the q_c and f_s values of the pile location using the q_c and f_s values of the 6 boreholes.

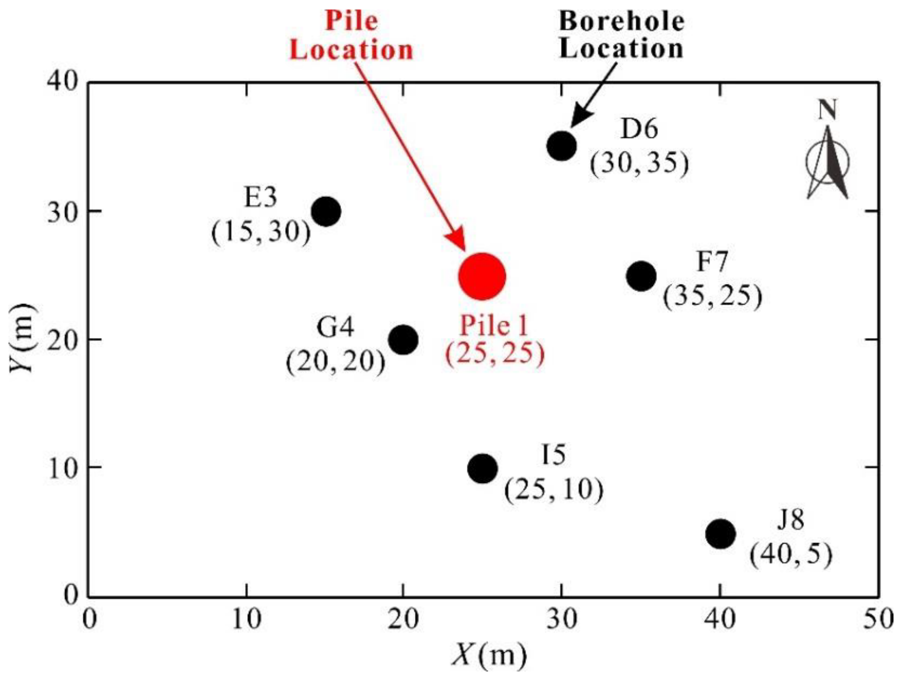


Fig. 1 location of the pile and the 6 boreholes

Question analysis

Soil is layered in the ground. The mechanical properties of the same type of soil are close. If the soil type distribution along the depth of the pile location are clear, the q_c and f_s values are clear and the Q_u can be calculated. The q_c and f_s values are widely used in the construction of engineering to gain the type of soil along the depth (Robertson, 1990). Typical methods to determine the type of the soils are based on the experience of engineers according to the regulations of the values of the q_c and f_s, which is subjective (Shusheng Lv et al., 2021). An objective and data-based method (Ogen et al., 2019), spectral clustering (von Luxburg, 2007), is applied to gain the soil type using q_c and f_s values to gain the soil type distribution of the 6 boreholes. Then a hidden Markov chain model is applied to evaluate the soil type distribution at the location of the pile according to the distribution at the 6 boreholes.

Spectral clustering

Spectral clustering is an unsupervised machine learning method. The method is applied here to classify the type of soil using q_c and f_s values. The classifications may be different from the existing soil classifications since the existing soil classifications considers many other soil properties such as cohesion and interior friction angle. Because the calculation of Q_u only needs q_c and f_s values, the classifications from q_c and f_s values can be more suitable.

The principle and steps of spectral clustering are as follows. Let G=(V, E) be an undirected graph with vertex set V={1, 2, ……, n} and edge set $E \subset V \times V$. The graph is assumed weighted, that is, a non-negative weight $w_{ij}=w_{ji} \geq 0$ is associated with each edge $(i, j) \in E$ between two vertices i and j. For $(i, j) \notin E$, w_{ij} is set to 0. The weighted adjacency matrix of the graph is the matrix:

$$W = \left(w_{ij}\right) \in \mathbb{R}^{n \times n}$$

The degrees $d_i = \sum_{j=1}^{n} w_{ij}$ are the elements of the diagonal matrix

$$D = diag(d_i),\ \ d_i = (W1)_i,\ \ where 1 := (1, \ldots, 1)^T \in \mathbb{R}^n.$$

The (unnormalized) Laplacian of the graph is given by matrix L=L(W),

$$L = D - W,\ \ i.e.,\ \ L(W) = diag(W1) - W.$$

The Laplacian matrix L is symmetric and positive semi-definite; since $L\mathbf{1}=0$, $\lambda_1=0$ is the smallest eigenvalue of L. Note that the matrix L does not depend on diagonal elements of the matrix W, which means that self-edges do not change the graph Laplacian. The graph Laplacian and its eigenvectors provide important instruments to spectral clustering, as stated by the following.

Let $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ be the eigenvalues of L. If λ_2 is a simple eigenvalue, then the corresponding eigenvector is known as the Fiedler vector. In spectral graph theory (Fiedler M., 1973), it is common to compute the second smallest eigenvalue of L and label the positive components of the Fiedler vector as belonging to one subset and the negative ones to another subset, and in this way obtaining a natural partition of the graph. However, this becomes unreliable when a small perturbation of the weights yields a coalescence of the eigenvalues λ2 and λ3. More generally (von Luxburg, 2007), for a subset of vertices $C \subset V$, we here denote the indicator vector $\mathbf{1}_C$ as the vector whose lth entry is equal to 1 if $v_l \in C$ and is equal to zero otherwise.

The multiplicity k of the eigenvalue 0 equals the number of connected components C1, ..., Ck in the graph. The eigenspace of the eigenvalue 0 is spanned by the indicator vectors $\mathbf{1}_{C1}$, ..., $\mathbf{1}_{Ck}$.

Nonempty sets C1, ..., Ck form a clustering of the graph if

$$C_i \cap C_j = \varnothing \quad for\ i, j = 1, \ldots, k,\ i \neq j \quad and \quad \bigcup_{i=1}^{k} C_i = V.$$

Similar to the case of two clusters, this result motivates an algorithm for clustering a graph into k components. (Liu J. et al., 2018) is an example for the classical k-means algorithm. Analogous results and algorithms are extended to the normalized Laplacian.

$$L_{Sym} := D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$

Algorithm 1 Unnormalized spectral clustering algorithm.

Input: weight matrix W, number k of clusters

Output: Clusters $C_1, \ldots, C_k$

Begin

Compute the Laplacian matrix $L = D - W$

Compute the eigenvectors $x_1, \ldots, x_k$

4: Set $X = [x_1 | x_2 | \dots | x_k]$

5: For $i = 1, \dots, n$, define $r_i \in \mathbb{R}^k$ the vector given by the i-th row of X

6: Cluster the points $(r_i)_{i=1,\dots,n}$ by the k-means algorithm into k clusters $C_1, \dots, C_k$

7: Return $C_1, \dots, C_k$

Hidden Markov chain model

Hidden Markov chain model are often used to describe the relationship between hidden state chain and visible state chain. In this question, the hidden state is the type of the soil which cannot be seen, and the visible states are the q_c and f_s values. Each hidden state has the probability of several visible states which called the transition probability. The visible states are continuous values so the mean and variance of q_c and f_s are the parameters describing the transition probability. The principle and steps of hidden Markov chain model are as follows.

As one of the important kinds of rock and soil variability, the phenomenon of irregular distribution of soil mass is widely existed in practice. In the existing researches on soil variability, the method of inferred stratigraphic strata adopted in a considerable part of the researches cannot reflect the randomness of stratigraphic distribution and cannot deal with the more complex stratigraphic distribution forms (Blanchin and Chilès, 1993; Calcagno et al., 2008). In recent years, Markov chain model (MC) are used to simulate the stratigraphic distribution along the depth.

To explain the definition of MC model, Suppose K is a MC value on a pile, then two properties of K need to be satisfied:

$$P(k) > 0, \forall x \in \Omega$$

$$P(k_i) = P(k_i | k_j,\ k_j \in N_i)$$

Where Ni denotes neighborhood sets of the given depth i. That is, any variable in the MC is positive and only affected by its neighboring variables. In detail, the probability distribution of any variable in MC under all given variables is equal to the probability distribution under given its neighboring variables. For example, the random variable at a specific depth is correlated with other corresponding variables of certain spatially adjacent depths, which are referred as neighbors of depth (Elfeki A et al., 2001; Elfeki A M M et al., 2005).

In the so-call hidden Markov chain (HMC) model, an configuration of all states of soils x = (x1, x2, x3,..., xN) can be assumed, which is unobservable by the site experiment. Then the emitted field P(y|x) is generated by the following equation:

$$P(y|x) = \prod_{j=1}^{s} P(y_j|x_j)$$

Where y is the observation value. In this equation, each pair of (xj, yj) indicates that the corresponding features yj is available from the assumptive label xj, by given the local neighborhood configuration x∂j. In order to determine a HMC model, the following three sets of parameters are needed:

The state transition probability matrix A which denotes the probability that the model transitions between states. Each value in A can be defined as:

$$a_{ij} = P(x_j|x_i)$$

The output observation probability matrix B which is used for the calculation of the probability that the model obtains each observed value based on the current state. As similar, the value in the matrix B can be detailed as following:

$$b_{ij} = P(y_j|x_i)$$

In the so-call model of Gaussian hidden Markov chain (GHMC), the formula can be rewritten as:

$$b_{ij} = P\{y_j|(\mu_i, \sigma_i)\}$$

where the (μl, σi) indicates the parameter set of the Gaussian component density function.

The initial state probability for representing the probability of initial state at the initial moment, which can be written as:

$$\pi = P(y_1, y_2, \dots, y_k)$$

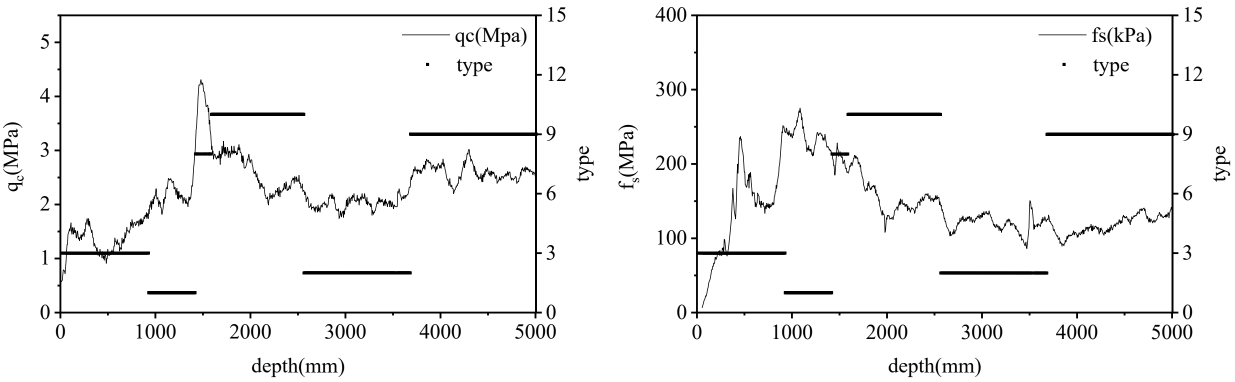
where k is the total number of the observed variables.

To sum up, the GHMC models are usually represented by λ=[A, B, π]. Thus, by implementing the GHMC theory, the correlation of the observable and unobservable variables can be accessed. And the parameter prediction of the target site will be introduced in the following section.

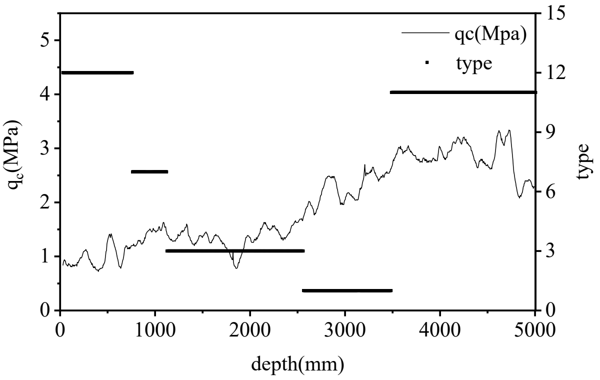
Modeling process

Spectral clustering of soils based on q_c and f_s

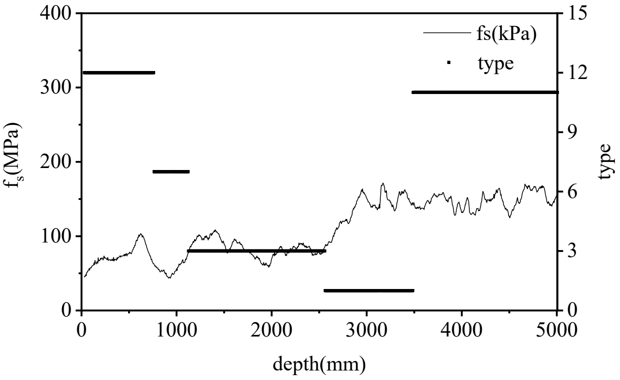
The values of q_c and f_s are given with matching depth. Besides the q_c and f_s, depth is also an important parameter that should be considered. Thus, the samples of cluster consist of 3 features: q_c, f_s and depth. The number of samples is 1000 from every 5mm along the depth of 5005mm. The results of clustering are show as Fig. 2. Several clustering methods, such as k-means clustering, density-based clustering, spectral clustering, are used and spectral clustering performs best. The results show that the stratums with similar q_c, f_s and depth cluster into the same type. Totally 14 soil types are clustered and the average of q_c, f_s and depth of the 14 soil types are shown in Table 1. According to the results, all 6 borehole's soil type along depth form 0mm to 5005mm can be got.



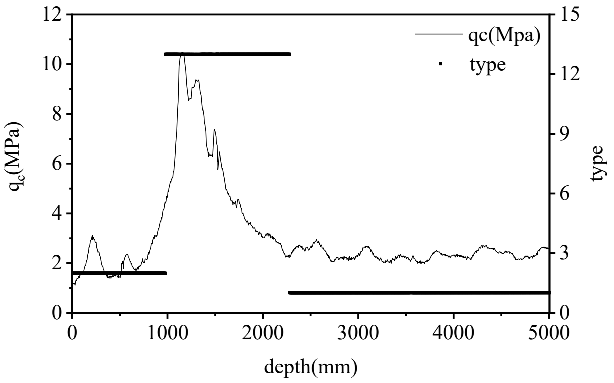
(a) cluster results and q_c data of D6 (b) cluster results and f_s data of D6



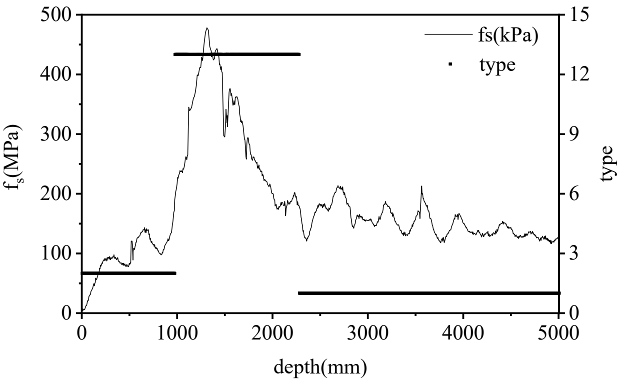
(c) cluster results and q_c data of E3



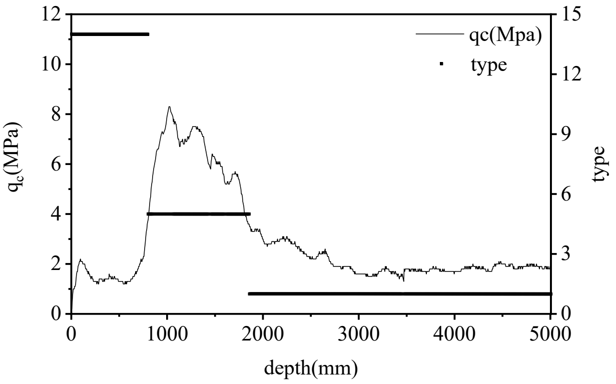
(d) cluster results and f_s data of E3



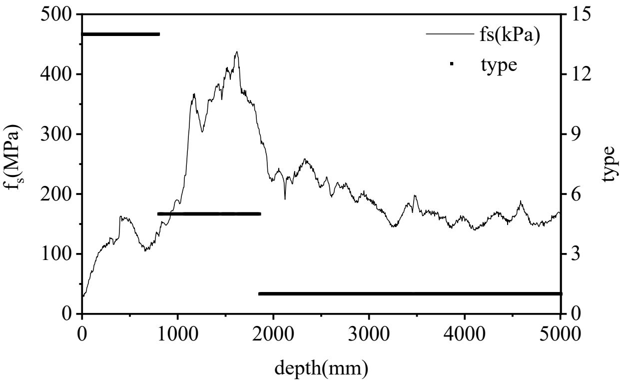
(e) cluster results and q_c data of G4



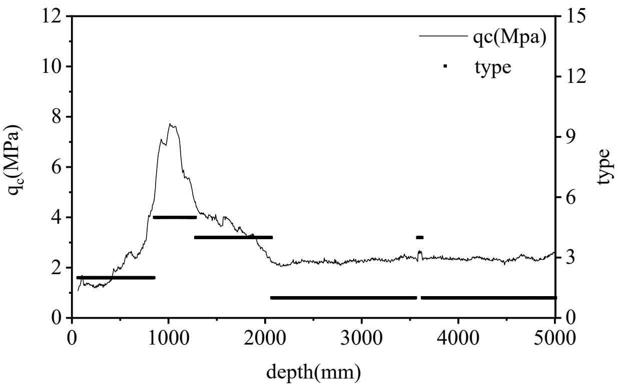
(f) cluster results and f_s data of G4



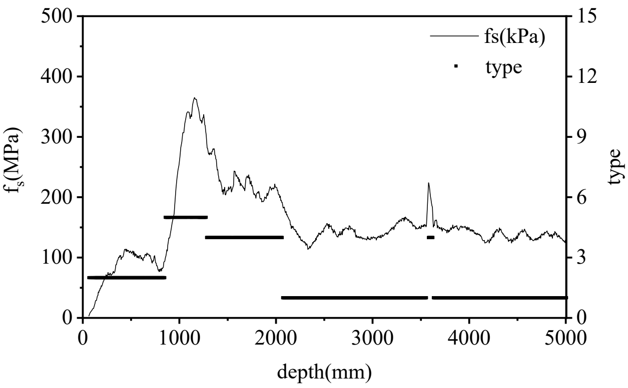
(g) cluster results and q_c data of I5



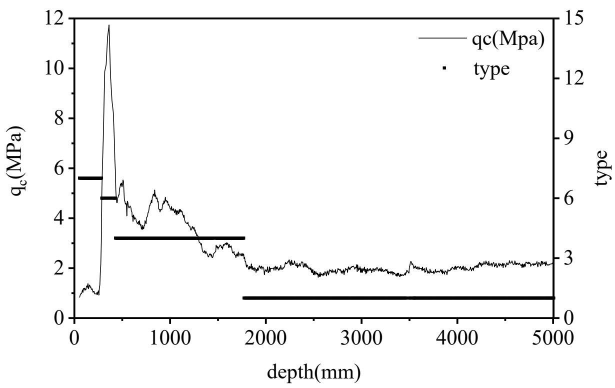
(h) cluster results and f_s data of I5

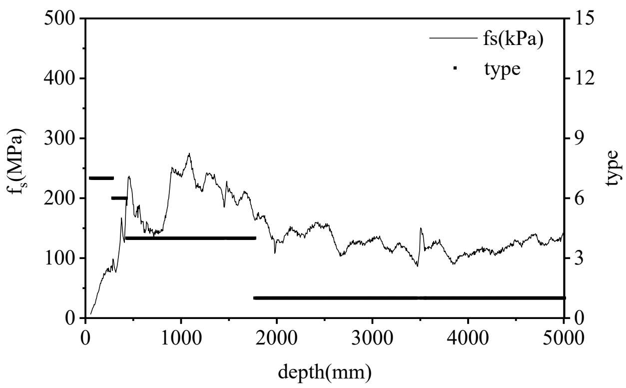


(i) cluster results and q_c data of J8



(j) cluster results and f_s data of J8





(k) cluster results and q_c data of F7 (l) cluster results and f_s data of F7

Fig. 2 Clustering results of the 6 boreholes

Table 1 Average of qc, fs and depth for all soil types

soil type	average q_c (MPa)	average f_s (kPa)	average depth (mm)
1	2.17	156.5	3380.9
2	2.15	89.6	481.0
3	1.37	80.4	1319.5
4	3.67	208.0	1246.9
5	6.22	309.9	1299.2
6	8.25	127.0	360.2
7	1.32	57.5	631.4
8	3.77	144.8	1506.7
9	2.59	152.1	4356.9
10	2.59	129.2	2076.3
11	2.81	148.1	4270.5
12	0.98	73.4	394.6
13	4.93	280.1	1677.7
14	1.59	114.2	403.5

Hidden Markov chain model of soil types

Hidden Markov chain model is used to describe the relationship between soil type and the q_c, f_s values. The 14 soil types are hidden states, the q_c, f_s values are visible states. The probability of stratums at the same depth change from one type to another is the transition probability. The average and variance of soil types describe the emission probability.

The transition of soil type considered here is in horizontal directions between 2 boreholes. Two transition probability matrixes describing two direction, i.e. south-north and east-west are calculated. For south-north transition probability matrix, the transition probability between G4 and D6, I5 and D6, J8 and F7, G4 and E3, I5 and G4, F7 and D6 are weighted based on distance from pile location and summed up. The farther from pile location, the smaller the weight of boreholes transition probability is. The east-west transition borehole groups are E3 and F7, G4 and F7, E3 and D6, I5 and J8. The transition borehole groups are chosen considering the 2 boreholes without other boreholes between them.

After getting the transition probability and emission probability, the Hidden Markov chain model is established. The model can calculate the probability of input data. The input data should be 2*2 matrix. The first row is the q_c and f_s of the borehole, the second row is the q_c and f_s of the stratum at the pile location.

Evaluation of q_c and f_s at the pile location

The process of evaluation of q_c and f_s at the pile location is a trial and error process. At each 5mm along 5005mm, data of q_c and f_s are generated in turn from the minimum to the maximum of all 14 soil types. Then the data of pile location are combined with the data of q_c and f_s of 6 borehole at the same depth forming the input 2*2 matrix. 6 boreholes form 6 input matrixes whose first rows are different, i.e., q_c and f_s of the borehole and the second rows are same, i.e., the q_c and f_s of the pile location. The hidden Markov chain model calculate the probability of the 6 input matrixes and summed up with weights in inverse proportion to the distance from borehole to pile location. The data of q_c and f_s at pile location with maximal probability are chosen as the evaluated data. The detailed process is as follows:

---

*(1) depthi = 5*i, i=1,2,…,1001*

(2) soil type = j at depthi at pile location, j=1,2,…,14

(3) q_c = *minimal* q_c *of soil type j* + $k * step_{q_c}$,

f_s = *minimal* f_s *of soil type j* + $k * step_{f_s}$,

k=1,2,…,n, $step_{q_c} = \frac{maximal\ q_c\ of\ soil\ type\ j - minimal\ q_c\ of\ soil\ type\ j}{n}$

$$step_{f_s} = \frac{maximal\ f_s\ of\ soil\ type\ j - minimal\ f_s\ of\ soil\ type\ j}{n}$$

(4) borehole = m, data$_m$ = $\begin{bmatrix} q_{cm} & f_{sm} \\ q_c & f_s \end{bmatrix}$, *m=1,2,…,6*

(5) probability(i,j,k) = $\sum_{m=1}^{6} Markov(data_m) * weight_m$

(6) get the q_c, f_s with maximal probability

The evaluated q_c, f_s and the CPT data of pile location are shown in Fig. 3. In order to compared with the traditional interpolation method, the interpolated q_c, f_s and the CPT data are shown in Fig. 4. From the results, the evaluated data are closer to the CPT data than the interpolated data. The evaluated q_c matches the CPT data well while the evaluated f_s does not match the CPT data very well. The depth range where evaluated f_s does not match the CPT are (1230,1515), (2505,3750) and (3750,5005). The average of q_c and f_s are (2.47,53.5), (2.90,83.5) and (2.44, 39.7). These 3 types of soil do not belong to the former 14 types because their f_s values are too small and the 6 boreholes cannot reflect the existence of the low-f_s soil type.

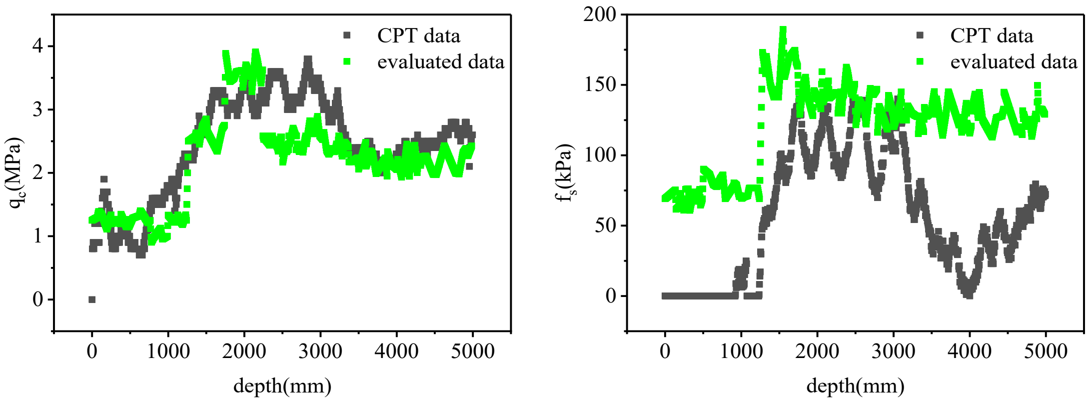


Fig. 3 Evaluated data and CPT data

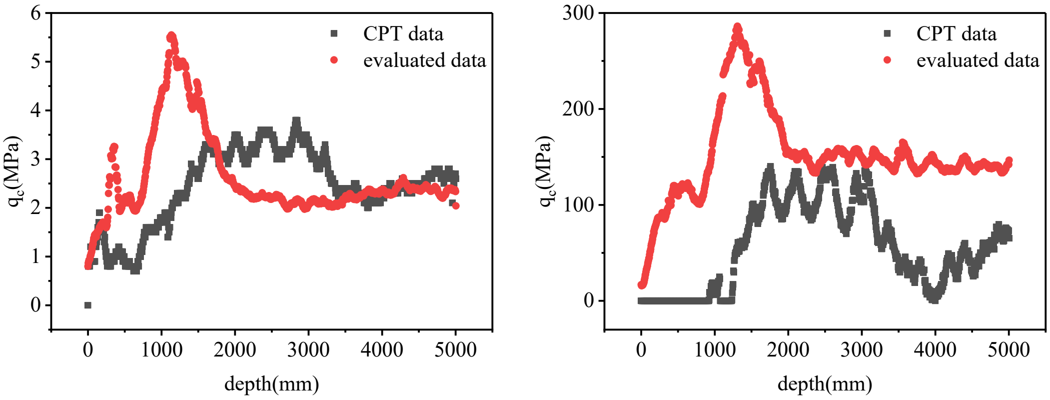


Fig. 4 Interpolated data and CPT data

Evaluate the ultimate bearing capacity

The suggested empirical approach is used to evaluate the ultimate bearing capacity.

$$Q_u = q_p A_p + \sum_{i=1}^{n} f_{pi} A_{si} \quad (1)$$

$$q_p = (q_{c1} + q_{c2})/2 \ \leq 15MPa \quad (2)$$

$$f_p = k_c f_s \ \leq 120kPa, k_c = 0.2 - 1.25 \quad (3)$$

where Q_u is the ultimate axial pile capacity, q_{c1} is minimum of the average of q_c values of zones ranging from 0.7 to 4D below pile tip, D is pile diameter, and q_{c2} is average minimum q_c values 8D above the pile tip.

In this question, D=400mm and length of pile is 4500mm so the q_{c1} is minimum of the average of q_c values of zones ranging from 4780mm to 6100 below pile tip. After calculation, q_{c1} =2.35MPa, q_{c2} =2.39MPa, q_p =2.46MPa. A_p=0.1257m^2. k_c is valued according to the literature. Then Q_u is calculated out with value of 645.46kPa.

Reference

Ogen, Y., Zaluda, J., Francos, N., Goldshleger, N., Ben-Dor, E. Cluster-based spectral models for a robust assessment of soil properties[J]. Geoderma, 2019, 340, 175-184.

Shusheng Lv, Peishuai Chen, Min Qiu, Zhao Li. Soil Classification Method and Experimental Research from Cone Penetration Test Based on Hierarchical Clustering Algorithm[J]. Science Technology and Engineering, 2021, 21(07): 2609-2615.

Robertson, P.K. SOIL CLASSIFICATION USING THE CONE PENETRATION TEST[J]. Canadian Geotechnical Journal, 1990, 27, 151-158.

von Luxburg, U. A tutorial on spectral clustering. Statistics and Computing, 2007, 17, 395-416.

Fiedler M. Algebraic connectivity of graphs[J]. Czechoslovak mathematical journal, 1973, 23(2): 298-305.

Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395-416.

Liu J, Han J. Spectral clustering[M]. Data Clustering. Chapman and Hall/CRC, 2018: 177-200.

Blanchin R, Chilès J P. The Channel Tunnel: Geostatistical prediction of the geological conditions and its validation by the reality[J]. Mathematical geology, 1993, 25(7): 963-974.

Calcagno, P., Chilès, J.-P., Courrioux, G., Guillen, A., 2008. Geological modelling from field data and geological knowledge: Part I. Modelling method coupling 3D potential-field interpolation and geological rules[J]. PHYSICS OF THE EARTH AND PLANETARY INTERIORS, 2008, 171: 147-157.

Elfeki A, Dekking M. A Markov chain model for subsurface characterization: theory and applications[J]. Mathematical geology, 2001, 33(5): 569-589.

Elfeki A M M, Dekking F M. Modelling subsurface heterogeneity by coupled Markov chains: Directional dependency, Walther's law and entropy[J]. Geotechnical & Geological Engineering, 2005, 23(6): 721-756.