# A Kernel Regression based Method to Interpret Pile Bearing Capacity

XIAN Jian-tang [1], HU Jin-zheng[1], SUN Yuan[1]
(1, Department of Geotechnical Engineering, Tongji University, Shanghai 210092, P. R. China)

**Abstract:** With the booming of machine learning application, the soil test data is playing an important role in geotechnical engineering. A pile design problem is introduced in the TC304/TC309 Student Contest in NSERIR6. This paper propose a kernel regression based machine learning method to learn the measured cone penetration test (CPT) data and the soil properties at untested points can be interpreted. The proposed method can automatically recognize the complicated non-stationary fluctuation of the soil profile, which are usually neglected in the traditional method. The mean and the standard deviation as a function of the coordinate of the target point can be both calculated. Then the auto-correlation of the soil properties can be represented by virtual CPT data generated from existing measured data. After that, the statistics of the bearing capacity can be calculated. Besides, the samples of the bearing capacity can be generated based on the linear approximation with moment matching technique. The proposed method will bring useful insight for the design of pile foundation with spatially variable soil.
**Keywords:** pile bearing capacity; kernel regression method; spatial variability; non-stationary random field

# 基于核回归的桩基承载力预测方法

冼健棠 [1]，胡金政 [1]，孙源 [1]
（1.同济大学 土木工程学院，上海 210092）

**摘　要：**随着机器学习算法的应用开展，土工试验数据在岩土工程中扮演着愈加重要的角色。本文尝试解答了在第六届全国工程风险与保险研究学术研讨会中举办的 TC304/TC309 学生竞赛提出的有关桩基础设计的赛题。本文提出了一种基于核回归的机器学习方法。通过学习钻孔处的圆锥贯入试验（CPT）数据，本文方法可以用来推测未钻孔取样处的土体性质，从而为桩基础设计提供依据。该方法可以自动感知空间变异性土体随机场的复杂非平稳性，这在传统分析方法中是常常被忽略的。该方法可以计算随空间变化的均值和标准差，同时其复杂的自相关结构也可以通过生成虚拟 CPT 试验数据来模拟。基于模拟 CPT 试验数据即可计算设计桩基础承载力的统计数据。此外，本文还使用了线性近似和矩匹配的方法来生成大量模拟桩基础承载力的样本。本文方法将为空间变异性土中的桩基础设计提供有用的参考。
**关键词：**桩基础承载力；核回归方法；空间变异性；非平稳随机场
**中图分类号：**TU375.4　　**文献标识码：**A

## 1 Introduction

The soil test data is playing an important role in geotechnical engineering. A pile design problem is introduced in the TC304/TC309 Student Contest in the 6th National Symposium on Engineering Risk & Insurance Research (NSERIR6), August 13-15th, 2021, Chongqing University, China. The data are extracted from the A-CPT/232/2500m2 dataset in the 304dB[1]. The task is to interpret the tip resistance ($q_c$) and side friction ($f_s$) at the pile location based upon the available CPT data, and then evaluate the ultimate bearing capacity of this pile foundation. The locations of the piles and the CPT data are shown in Figure 1. One may refer to the TC304 webpage[1] for the detailed introduction of this problem.

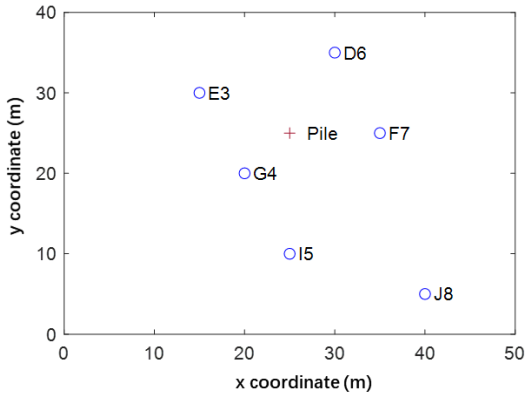


Figure 1. Locations of the piles and CPTs

The soil properties are usually spatially variable (e.g., Phoon and Kulhawy[2]). The most popular model of spatial variability in geotechnical engineering is the random field model (e.g., Vanmarcke[3]). Generally, the random field model is assumed to be stationary, i.e. the mean and the standard deviation are constant and the auto-correlation is only related to the distance between two points (e.g., Cho[4]; Liu et al.[5]). Recently, the studies of non-stationary random field model have increasingly appeared in the literatures (e.g., Wang et al.[6]; Jiang et al.[7]).

The non-parametric regression method is a useful tool to catch the trend line of random process (Härdle[8]). The kernel regression method could consider the weights of the data (Altman[9]), which has yielded encouraging results recently when applying in the geotechnical problems (e.g., Danese et al.[10]; Kordjazi et al.[11]; Kaveh et al.[12]). It may be possible to apply such method to learn the spatially variable soil properties adaptively based on measured data.

From the provided database of six CPTs, it was found that the spatial variability of $f_s$ and $q_c$ is significant and complicated. The CPT data reveals that the random field of the soil parameters is strongly non-stationary. It is tempting to adopt a method that can automatically characterize the sophisticated spatial variability of the soil. The proposed method does not require a specific function form of the non-stationary random field. It is a data-driven method which can automatically capture the complicated non-stationary fluctuation of the provided data of $f_s$ and $q_c$ and then rigorously interpret the complicated stochastic nature of the soil properties of the studied site.

This paper is organized as follows: First, the basic idea of kernel regression method is introduced. Then, the calibration of the kernel regression method is illustrated. Then the mean value and the standard deviation can be calculated based on the kernel regression method. And the method to estimate the soil properties and the bearing capacity based on "Virtual load test" is introduced. Finally, the statistics and the histogram of the bearing capacity are calculated.

## 2 Kernel Regression

In this paper, the kernel regression method is adopted to evaluate both the expectations and variances of the side friction, $f_s$ and the tip resistance, $q_c$ at different given depths. The kernel regression method is a non-parametric regression method to calculate the conditional expectation of a random variable (Kloke et al. [11]). The idea of the kernel regression method is to estimate the trend line based on the data points in the neighborhood of a given depth. One of the most popular expressions of kernel regression is the Nadaraya-Watson kernel regression, which can be seen as a weighted average of the data points (Nadayara [14]). The Nadaraya-Watson estimator for the calculation of the mean value of $f_s$ and $q_c$ at a given location with a coordinate of $\mathbf{x}$, can be expressed as follows:

$$\overline{f}_s(\mathbf{x}) = \frac{\sum_{i=1}^{N} K_{fs}(\mathbf{x}, \mathbf{X}_i) \cdot f_{si}}{\sum_{i=1}^{N} K_{fs}(\mathbf{x}, \mathbf{X}_i)} \quad (1)$$

$$\overline{q}_c(\mathbf{x}) = \frac{\sum_{i=1}^{N} K_{qc}(\mathbf{x}, \mathbf{X}_i) \cdot q_{ci}}{\sum_{i=1}^{N} K_{qc}(\mathbf{x}, \mathbf{X}_i)} \quad (2)$$

where $K_{fs}(\cdot)$ and $K_{qc}(\cdot)$ are the kernel functions for $f_s$ and $q_c$; $N$ is the number of the total data points; $\mathbf{x}$ is the coordinate vector of the target point; $\mathbf{X}_i$ is the coordinate vector of the data point $i$; $f_{si}$ and $q_{ci}$ are the recorded $f_s$ and $q_c$ value of the $i^{\text{th}}$ data point, respectively. The kernel functions $K_{fs}(\cdot)$ and $K_{qc}(\cdot)$ can be seen as the weight of each data points with respect to the target point. In this paper, a three-dimensional problem is

encountered. The following Gaussian kernel function is applied in this paper (Zhong et al. [15]; Ter Haar Romeny [16]):

$$K(\mathbf{x},\mathbf{X}_i)=\exp\left[-\left(\frac{l_h^2}{b_h^2}+\frac{l_v^2}{b_v^2}\right)\right] \quad (3)$$

where $l_h$ is the horizontal distance between $\mathbf{x}$ and $\mathbf{X}_i$; $l_v$ is the vertical distance between $\mathbf{x}$ and $\mathbf{X}_i$; $b_h$ is the horizontal bandwidth; $b_v$ is the vertical bandwidth; the determination of the bandwidth parameters will be described in detail in the following section.

## 3 The Determination of the Bandwidth

To calibrate the model of spatially variable soil, one need to first select an appropriate bandwidth. The bandwidth of the kernel function can cast a significant effect on the final estimation of the conditional expectation as those presented in Eqs. (1) and (2). In terms of the determination of the bandwidth, the cross validation selectors are recognized as one of the most useful methods for a wide range of data sets (Park and Marron [17]). In this paper, the more robust cross validation method with the utilization of the leave-one-out technique is adopted. Hence, the bandwidth can be evaluated by minimizing the value of the equation as shown below:

$$CV(b_h,b_v)=\sum_{j=1}^{6}\sum_{i=1}^{N_j}\left[f_{si}-\bar{f}_{s,-j}(\mathbf{X}_i;b_h,b_v)\right]^2 \quad (4)$$

where $j$ represents the $j^{th}$ borehole and the subscript $-j$ means that the expectation of the variable is estimated leave out the data series of the $j^{th}$ borehole. In this paper, the $b_h$ and $b_v$ for fs are 15.164 m and 0.253 m, respectively and the $b_h$ and $b_v$ for $q_c$ are 14.972 m and 0.293 m, respectively. Substitute these bandwidth values into Eqs. (1) and (2), the expectations of the $f_s$ and $q_c$ can be calculated at the target borehole given the corresponding three-dimensional coordinate. Figure 2 and Figure 3 show the expectations of the $f_s$ and $q_c$ of all the boreholes compared with the real $f_s$ and $q_c$ data. The result presented in Figure 2 indicates that the kernel regression yields a reasonable trend of $f_s$ and $q_c$ which is consistent with that of the real data.

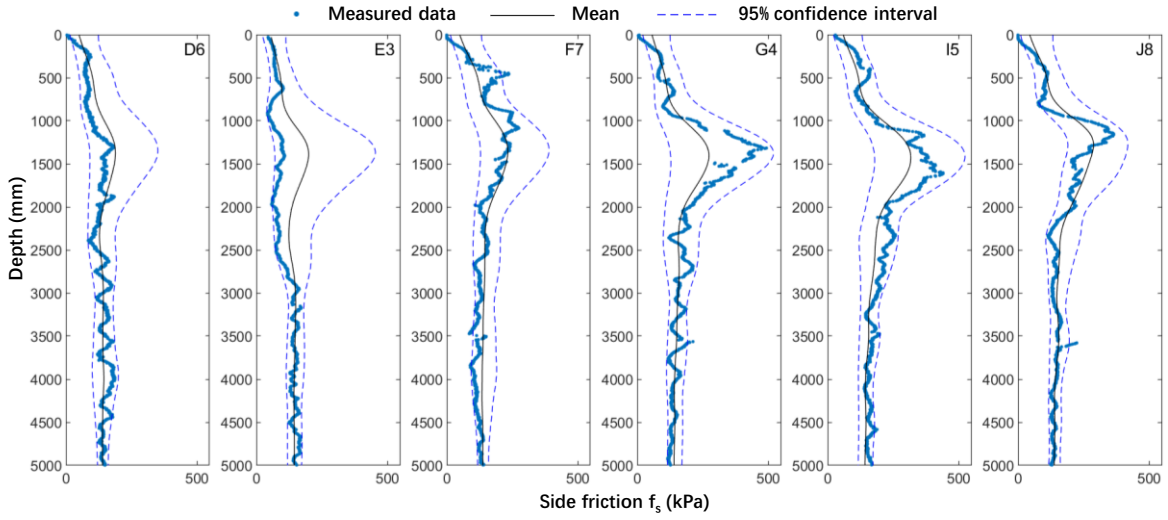


Figure 2. The expectation and 95% confidence level of the $f_s$ of all the bore holes compared with the real $f_s$ data

## 4 Residuals Analysis

The residuals of both $f_s$ and $q_c$ can be calculated as follow:

$$e_{fs,i}=f_s(\mathbf{X}_i)-\bar{f}_s(\mathbf{X}_i) \quad (5)$$

$$e_{qc,i}=q_c(\mathbf{X}_i)-\bar{q}_c(\mathbf{X}_i) \quad (6)$$

Figure 4(a) and Figure 4(b) present the residuals of $q_c$ and $f_s$. From Figure 4(a) and Figure 4(b) we can see that the difference between the yielded expected and recorded values of different boreholes all fluctuate

along the depth. Such variability may be called heteroscedasticity (Buteikis[18]), i.e., the standard deviations of a predicted variable are non-constant. This reflects the non-stationary properties of soil random field. Hence, in order to quantify the depth-variant uncertainty of $f_s$ and $q_c$, the variance or the standard deviation may also be described as a function of the coordinate. Note that the variance can be expressed as the expectation of the square residuals in the regression analysis (Buteikis[18]):

$$\sigma_{fs}^2(\mathbf{x}_i) \approx E\left(e_{fs,i}^2\right) \tag{7}$$

$$\sigma_{qc}^2(\mathbf{x}_i) \approx E\left(e_{qc,i}^2\right) \tag{8}$$

It may be possible to approximate the variance in Eqs. (7) and (8) by regression analysis of the residuals with respect to the coordinate $\mathbf{x}$. Therefore, the series of $e_{fs,i}$ and $e_{qc,i}$ can be used to calculate the variance using the similar procedure of kernel regression method as aforementioned. Then the variance or the standard deviation can be calculated for a given point. Figure 2 and Figure 3 also show the 95% confidence interval of the $f_s$ and $q_c$ of all the boreholes. It can be seen that most measured data points fall into the 95% confidence interval. The uncertainty of $f_s$ is relatively large above the depth of about 1.5m and rapidly reduced as the depth increases. Similar phenomenon is also observed for $q_c$. The reason may be that the properties of the soil encounter a significant change around the depth of 1.5~2.0 m. This is verified in Jaksa[19] that the ground contains several layers. The depth of the surface between Calcareous Mantle and Limy Surficial Layer and that of the surface between Calcareous Mantle and Limy Surficial Layer is around 1~2 m and 2~2.5 m, respectively. This may be the reason of the non-stationary fluctuation of the soil properties. Comparing with traditional methods such as stationary random field, the proposed method can well deal with the non-stationary spatially variable soil.

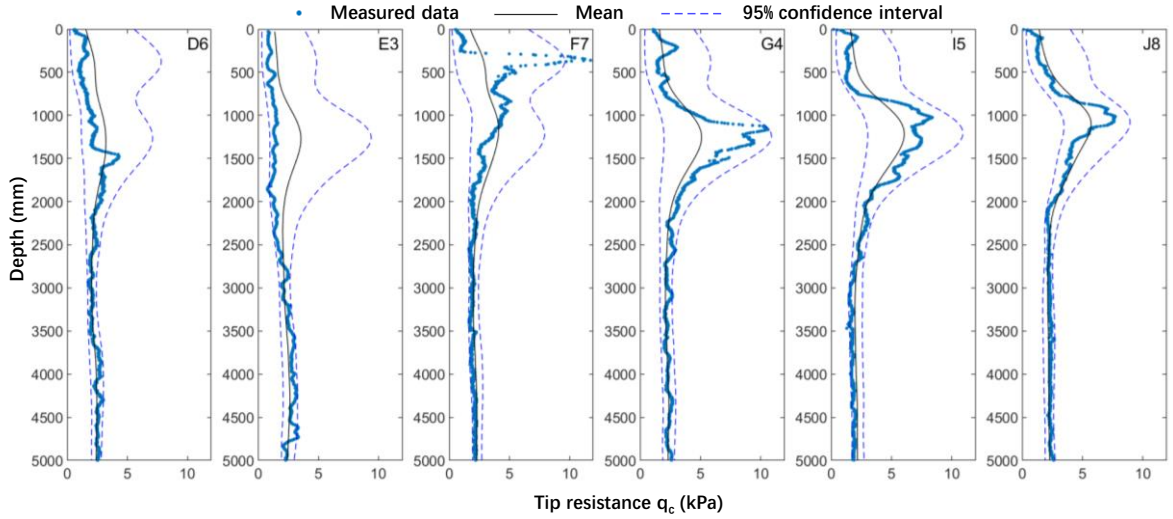


Figure 3. The expectation and 95% confidence level of $q_c$ of all the bore holes compared with the real $q_c$ data

# 5 Estimation of the Bearing Capacity

## 5.1 Generation of the samples of $f_s$ and $q_c$

Assume that $f_s$ and $q_c$ can be modelled as lognormal random variables. The statistics of ln$f_s$ and ln$q_c$ can be calculated based on the mean and the variance of $f_s$ and $q_c$ based on the property of lognormal distribution[20]. In the purpose of predicting the $Q_u$ of the designed location, the samples of predicted $f_s$ and $q_c$ can be generated using the following equations:

$$f_{s,predicted}\left(\mathbf{x}_i\right) = \exp\left[\lambda_{fs}(\mathbf{x}_i) + \varepsilon_{fs,i} \cdot \xi_{fs}(\mathbf{x}_i)\right] \tag{9}$$

$$q_{c,predicted}\left(\mathbf{x}_i\right) = \exp\left[\lambda_{qc}(\mathbf{x}_i) + \varepsilon_{qc,i} \cdot \xi_{qc}(\mathbf{x}_i)\right] \tag{10}$$

where $\lambda_{fs}(\mathbf{x}_i)$ and $\xi_{fs}(\mathbf{x}_i)$ are the mean and the standard

variance of ln$f_s$ at the $i^{\text{th}}$ point, respectively; $\lambda_{qc}(\mathbf{x}_i)$ and $\xi_{qc}(\mathbf{x}_i)$ are the mean and the standard variance of ln$q_c$ at the $i^{\text{th}}$ point, respectively; $\varepsilon_{fs,i}$ and $\varepsilon_{qc,i}$ are the normalized residuals which can be calculated as follows:

$$\varepsilon_{fs,i} = \frac{\ln f_s(\mathbf{X}_i) - \lambda_{fs}(\mathbf{x}_i)}{\xi_{fs}(\mathbf{x}_i)} \tag{11}$$

$$\varepsilon_{qc,i} = \frac{\ln f_s(\mathbf{X}_i) - \lambda_{qc}(\mathbf{x}_i)}{\xi_{qc}(\mathbf{x}_i)} \tag{12}$$

Figure 4(c) and Figure 4(d) show the normalized residuals of $f_s$ and $q_c$ vary along with the depth. It can be seen from Figure 4(c) and Figure 4(d) that the normalized residuals are distributed evenly around 0 and their standard deviations are quite close to 1.

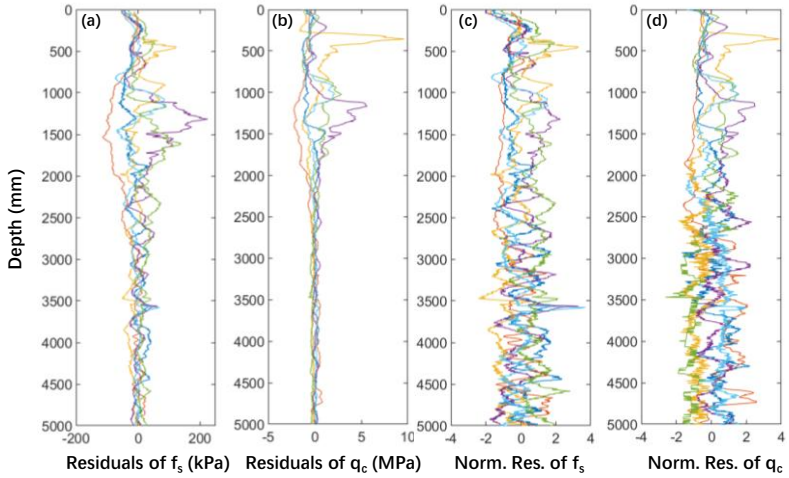


Figure 4. The residuals and normalized residuals of $q_c$ and $f_s$.

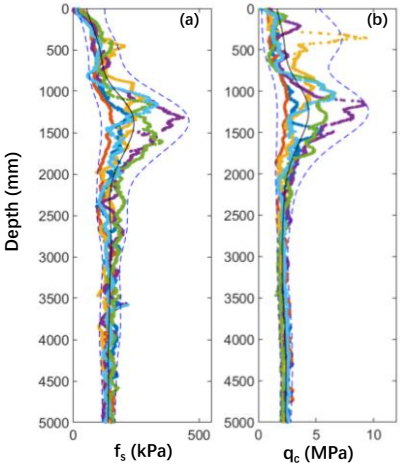


Figure 5. The virtual CPT data

It may be possible to model the normalized residuals as standard normal variables[21]. The cross-correlation coefficient between $f_s$ and $q_c$ can be calculated based on the samples of $\varepsilon_{fs,i}$ and $\varepsilon_{qc,i}$. By generating samples of the normalized residual, the samples of $f_s$ and $q_c$ can be conveniently calculated via Eqs. (9) and (10). However, the autocorrelation of the normalized residuals cannot be easily modeled, which remains a barrier for constructing the Gaussian random field of the normalized residuals. As can be seen in Figure 4(c) and Figure 4(d), the fluctuation of the normalized residuals along the depth is not uniform. This reflects the non-stationary property of auto-correlation of the normalized residuals. Nevertheless, it may be reasonable to consider that the fluctuation of the normalized residuals is similar in this site. The normalized residual at the point of pile can be generated from the normalized residuals of the six boreholes. By utilizing $\varepsilon_{fs,i}$ and $\varepsilon_{qc,i}$ from the six boreholes, the $\overline{f}_s(\mathbf{x}_i)$, $\overline{q}_c(\mathbf{x}_i)$, $\sigma_{fs,i}$ and $\sigma_{fqc,i}$ from the point of designed pile and the six virtual CPTs can be generated via Eqs. (9) and (10). The six generated virtual CPT data are shown in Figure 5. The mean value and the 95% confidence interval are also shown in Figure 5. The similarity between the points of designed pile and various boreholes may be different. Hence the weights of each virtual CPT data are also different. The larger the distance is, the less the chance of observing the corresponding virtual CPT data is. The probability of each simulated virtual CPT data may be modelled using the kernel function of Eq. (3), i.e.:

$$p_k \propto \exp\left[-\left(\frac{l_{ph}^2}{b_h^2} + \frac{l_{pv}^2}{b_v^2}\right)\right] \tag{13}$$

where $p_k$ is the probability of the $k$th borehole; $l_{ph}$ is the horizontal distance between the point of pile and the borehole; $l_v$ is the vertical distance between the point of pile and the borehole, which may be regarded as 0 since the depths of concern are the same in the pile design of this problem. As the horizontal bandwidth of $f_s$ and $q_c$ are very close to each other, here the average value is adopted in Eq. (13).

Based on Eqs. (9) to (12), the ultimate bearing capacity of six virtual CPTs can be calculated, which is called the "virtual load tests":

$$Q_{u,k} = q_p A_p + \sum_{j=1}^{Np} f_{pj} A_{sj} \tag{14}$$

where $Q_{u,k}$ is the ultimate bearing capacity of the $k^{th}$ "virtual load test" and $k$ = 1, …, 6, $q_p$ is the unit end bearing, $A_p$ is the pile end area, As is the surface area along the pile shaft and $f_p$ is the unit shaft friction. The calculation of $q_p$ and $f_p$ is performed using the method suggested in Schmertmann[22] based on the values of $f_s$ and $q_c$ yielded via Eqs. (9) and (10). Let $Q_u$ denote the bearing capacity of the designed pile. Based on Eq. (13), the mean and the standard deviation of $Q_u$ can be estimated as:

$$\mu_{Qu} = \sum_{k=1}^{6} p_k \cdot Q_{u,k} \tag{15}$$

$$\sigma_{Qu} = \sqrt{\sum_{k=1}^{6} p_k \cdot \left(Q_{u,k} - \mu_{Qu}\right)^2} \tag{16}$$

where $\mu_{Qu}$ and $\sigma_{Qu}$ are the mean and the standard deviation of $Q_u$, respectively.

### 5.2 Generation of the samples of $Q_u$

In the above sections, the statistics of $Q_u$ are estimated. However, there is no information concerning the specific distribution of the $Q_u$. Thus, the uncertainty of the predicted $Q_u$ cannot be fully quantified and the value of the $Q_u$ corresponding to a certain confidence level is not available. In order to simulate the shape of the distribution of $Q_u$, it is assumed that $Q_u$ can be estimated via the linear combination of two distributions:

$$Q_u = a \cdot Q_{u,correlated} + b \cdot Q_{u,independent} \tag{17}$$

where $Q_{u,correlated}$ is the ultimate bearing capacity based on the predicted values of $f_s$ and $q_c$ which are fully auto-correlated and $Q_{u,independent}$ is the ultimate bearing capacity based on the predicted values of $f_s$ and $q_c$ which are independent. The samples of $Q_{u,correlated}$ can be obtained via in Eqs. (9) and (10) by simulating the fully correlated normalized residuals. As for $Q_{u,independent}$, the normalized residuals is independently generated for different depths. These two distributions can be regarded as the boundaries of the distribution of $Q_u$ because the actual auto-correlation is between these two cases as shown in Figure 4(c) and (d). The histogram of $Q_{u,correlated}$ and $Q_{u,independent}$ are illustrated in Figure 6(a) and Figure 6(b), respectively.

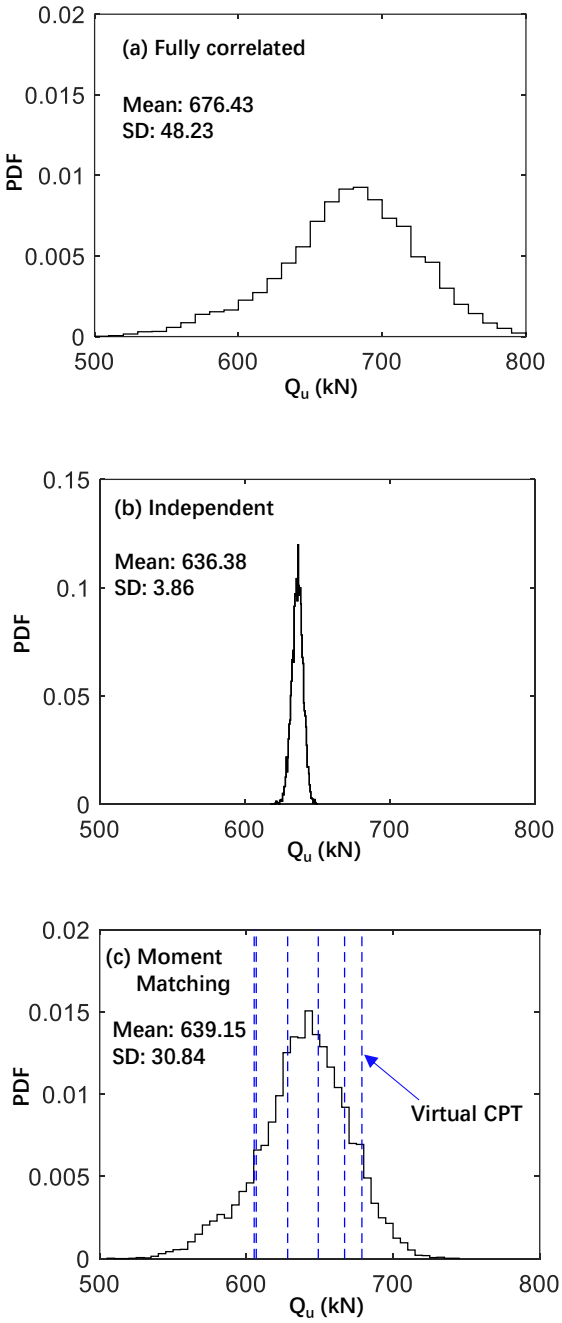


Figure 1. The distribution of $Q_u$ which is based on: (a) fully auto-correlated CPT data; (b) independent CPT; (c) moment matching CPT

Based on the samples of $Q_{u,correlated}$ and $Q_{u,independent}$,

the samples of $Q_u$ , which can give the values of $a$ and $b$ is generated via Eq. (17). To calibrate the values of $a$ and $b$, the moment matching method is adopted[23]. The values of $a$ and $b$ in Eq.(17) are adjusted to ensure that the samples of $Q_u$ could construct a distribution whose mean and standard deviation are as close to those evaluated using Eqs. (15) and (16). The values of $a$ and $b$ can be solved as $a$ = 0.6423 and $b$ = 0.3211. The histogram of $Q_u$ is illustrated in Figure 5(c). The mean and the standard deviation of $Q_u$ are $\mu_{Qu}$ = 639.15kN and $\sigma_{Qu}$ = 30.84kN, respectively.

# 4 Conclusion

1) This paper proposed a kernel regression based method to interpret the soil properties and designed pile bearing capacity. The mean and the variance or the standard deviation of CPT data at the untested points can approximated based on measured data. The complicated non-stationary fluctuation of the soil properties can be automatically sensed by the proposed method.

2) The “virtual load test” generated by virtual CPT data can be simulated based on the measured data from boreholes in the neighborhood. With a method of moment matching, the samples of $Q_u$ can be approximately generated. The proposed method will bring useful insight for design of pile foundation with spatially variable CPT data.

# Acknowledgements:

The participants would like to thank Prof. H.W. Huang and Prof. J. Zhang from Tongji University for their kind support and patient instruction. Meanwhile, the participants are grateful to the considerate work of the organizing committee of NSERIR6. The TC304/TC309 and the contributors of 304dB are also appreciated for their insightful work.

# References:

[1] TC304 Engineering Practice of Risk Assessment & Management[EB/OL].http://140.112.12.21/issmge/tc304.htm, 2021-07–12/2021–05–03.

[2] PHOON K, KULHAWY F H. Characterization of geotechnical variability[J]. Canadian Geotechnical Journal, 1999, 36(4): 612-624.

[3] VANMARCKE E H . Random Fields Analysis and Synthesis. 1983.

[4] CHO S E. Probabilistic assessment of slope stability that considers the spatial variability of soil properties[J]. Journal of Geotechnical and Geoenvironmental Engineering, 2010, 136(7): 975-984.

[5] LIU L L, Cheng Y M, Zhang S H. Conditional random field reliability analysis of a cohesion-frictional slope[J]. Computers and Geotechnics, 2016, 82: 173-186.

[6] WANG Y, ZHAO T, HU Y, PHOON K K. Simulation of random fields with trend from sparse measurements without detrending.[J] Journal of Engineering Mechanics. 2019 Feb 1;145(2):04018130.

[7] JIANG S H, PAPAIOANNOU I, STRAUB D. Optimization of Site-Exploration Programs for Slope-Reliability Assessment[J]. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 2020, 6(1): 04020004.

[8] Härdle W. Applied nonparametric regression[M]. Cambridge university press, 1990.

[9] ALTMAN N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.

[10] DANESE M, LAZZARI M. A kernel density estimation approach for landslide susceptibility assessment[C]//Mountain Risks: Bringing Science to Society. CERG Editions, Strasbourg, Proceedings of International conference of Mountain Risks, Firenze. 2010: 24-26.

[11] KORDJAZI A, NEJAD F P, JAKSA M B. Prediction of ultimate axial load-carrying capacity of piles using a support vector machine based on CPT data[J]. Computers and Geotechnics, 2014, 55: 91-102.

[12] KAVEH A, HAMZE-ZIABARI S M, BAKHSHPOORI T. Patient rule-induction method for liquefaction potential assessment based on CPT data[J]. Bulletin of Engineering Geology and the Environment, 2018, 77(2): 849-865.

[13] KLOKE J, MCKEAN J W, MCKEAN J W. Nonparametric statistical methods using R [M]. CRC Press Boca Raton, 2015.

[14] NADARAYA E A. On estimating regression [J]. Theory of

Probability & Its Applications, 1964, 9(1): 141-2.

[15] ZHONG S, CHEN D, XU Q, et al. Optimizing the Gaussian kernel function with the formulated kernel target alignment criterion for two-class pattern classification [J]. Pattern Recognition, 2013, 46(7): 2045-54.

[16] TER HAAR ROMENY B. M. The Gaussian Kernel [J]. Front-End Vision and Multi-Scale Image Analysis: Multi-Scale Computer Vision Theory and Applications, written in Mathematics, 2003: 37-51.

[17] PARK B U, Marron J S. Comparison of Data-Driven Bandwidth Selectors [J]. Journal of the American Statistical Association, 1990, 85(409): 66-72.

[18] BUTEIKIS A. Practical Econometrics and Data Science [M/OL]. 2020[2021-07–12]. http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/.

[19] JAKSA M B. The influence of spatial variability on the geotechnical design properties of a stiff, overconsolidated clay [D]; University of Adelaide, 1995.

[20] ALFREDO H-S A, WILSON H. Probability concepts in engineering planning and design[M]. New York: John Wily and Sons, 1975.

[21] DRAPER N R., SMITH H. Applied regression analysis [M]. New York: John Wiley & Sons, 1998.

[22] SCHMERTMANN J H. Guidelines for cone penetration test: performance and design[R]. United States. Federal Highway Administration, 1978.

[23] HCINE, M B., BOUALLEGUE, R. Fitting the log skew normal to the sum of independent lognormals distribution[J/OL]. arXiv preprint arXiv:1501.02344, 2015[2021-07–12].