Stratigraphic heterogeneity inference under limited borehole data and massive tunneling data during shield tunneling construction: a semi-supervised learning approach

Xiongyao Xie¹, Changzhi Yang¹, Hongjie Yu², Yu Huang¹, Qiang Wang¹∗

¹1239 Siping Road, Shanghai, China, 200092
²1012 14th Street, Golden, US, CO80401

18310182@tongji.edu.cn

Abstract. It’s essential to identify the soil distribution for shield tunneling construction, especially in mixed ground conditions. Machine learning models have been conducted to infer the stratigraphic heterogeneity, but mostly relying on the geological survey report. The sparse borehole data has limited the identification accuracy of supervised learning models. Inspired by the perception that the shield tunneling data contains the geological information, a semi-supervised learning-based stratigraphic heterogeneity inference approach is presented in this study, which gets the utmost out of the limited borehole data and massive tunneling data. More specifically, a label propagation algorithm (LPA) soil identification model is developed. The rings with borehole data are defined as the labeled rings while other rings are unlabelled. The LPA is then employed to identify the soil distribution on the unlabelled rings with an iterative algorithm as the shield machine driving. A numerical experiment is conducted in Nanning metro line 1. The slurry pressured balanced shield was put forward in the mixed ground containing round gravel and mudstone. The field results show that the LPA approach can identify the mudstone distribution more accurately than traditional supervised learning methods. Input feature importance of tunneling data are also discussed.

1. Introduction
The stratigraphic heterogeneity refers to the simultaneous occurrence of two or more kinds of soil in the tunnel face or alternately along the axis of the tunnel, which may result in severe disturbances in tunneling operations such as tunnel face instability [1], increased cutterhead wear [2], and lower tunneling efficiency[3]. In shield tunneling construction, The soil distribution is determined by the borehole data together with the interpolation method. Usually, the soil distribution along the borehole direction is accurate while the soil distribution along the tunnel axis between boreholes may not be reliable in the geological survey due to the spatial variability of soil.
Random process approaches have been used to describe stratigraphic heterogeneity, such as the Markov chain model, Monte Carlo simulation [4,5]. The random process model is good at the identification of soil distribution with obvious boundary conditions but may lack accuracy in the identification of irregular soil type. On modern shields, hundreds of sensors have recorded massive data regarding tunneling situations and geological information. The soil identification based on tunneling data has been employed recently. Zhou et al [6] applied the artificial neural network (ANN) model to identify the soil distribution in a tunnel section of Guangzhou metro, which employs tunneling data as input and soil distribution obtained from the geological survey as output. Liu et al [7] have conducted a similar work using the improved support vector machine (SVM) model, which
performs well with enough training samples. In these studies, the soil distributions obtained from the geological survey are taken as the true soil distribution and are directly used as the output labels. However, we can only determine the real soil distribution around the borehole location. Zhao et al [8] have realized this problem and just use the soil distribution around the borehole in their ANN model. As there are 4.6 million tunneling data and 88 borehole data in the tunnel section, the feature augmentation method is adopted to gain better identification accuracy. To make full use of the massive tunneling data and limited borehole data, the semi-supervised learning (SSL) approach [9] shows promising potential in the soil identification problem. The SSL approach harvests the idea that the unlabeled data (tunneling data without borehole information), when used in conjunction with a small amount of labeled data (tunneling data with borehole information), can produce a considerable improvement in learning accuracy, which has been widely used in disease classification [10], structural health monitoring [11], and extreme disaster prediction [12].

The remainder of the paper is organized as follows. Section 2 presents the research methodology for the SSL soil identification approach. Section 3 presents the case study in Nanning Metro as well as the comparison with traditional machine learning methods. Section 4 offers a discussion about the early prediction of soil distribution with part of one tunneling data and the input feature importance before the conclusion and future work.

2. Methodology
As one kind of graph-based semi-supervised learning, the label propagation algorithm [13] (LPA) is conducted here as the soil identification approach. The LPA is a fast algorithm for finding communities in a graph that is constructed based on the given data set. Each node in the graph corresponds to each sample in the data set. If there is a strong correlation between two samples, an edge will appear between the corresponding nodes, as illustrated in Figure 1. The strength of the edge is proportional to the similarity between samples. We believe that the response of the shield tunneling boring machine (TBM) is determined by the real soil distribution and TBM operation. The features obtained from the TBM operation as well as the response will be the input features for the LBA-based soil identification model. The borehole data will be regarded as the labeled output, as shown in Figure 1. The purpose of the LBA is to assign the labeled rings to the unlabeled rings.

![Figure 1 Soil distribution identification in shield tunneling construction based on LPA, the nodes are not fully connected for better demonstration](image-url)
We define the data set with borehole information as the labeled data set \( \mathcal{D}_l = \{ (x_i, y_i), (x_j, y_j), \ldots, (x_m, y_m) \} \), and the data set without borehole information as the unlabeled data set \( \mathcal{D}_u = \{ \tilde{x}_i, \tilde{x}_j, \ldots, \tilde{x}_n \} \), where the \( \tilde{x}_i \) is the features of tunneling data and \( y_i \) is the soil type obtained from the borehole results. Usually, the rings with borehole information are much less than those without borehole information (\( l \ll u, l+u = N \), \( N \) is the total rings in the tunnel section). A graph \( G = (V, E) \) is established based on the data set \( \mathcal{D}_l \cup \mathcal{D}_u \), where \( V \) and \( E \) are the node-set and edge set respectively. The node-set can be described as \( V = \{ \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m, \tilde{x}_{i_1}, \tilde{x}_{i_2}, \ldots, \tilde{x}_{i_u} \} \) and the weight of the edge set \( E \) is represented by the affinity matrix \( W = (w_{ij})_{N \times N} \), as defined in Eq. \( \gamma > 0 \) is the hyperparameter that will be determined by numerical experiment. A degree matrix \( D = \text{diag}(d_1, d_2, \ldots, d_N) \), \( d_i = \sum_{j=1}^{N} w_{ij} \) is employed to represent the sum of the weights of all edges connected to the \( i \)th node in the graph. Then a probabilistic transition matrix \( P \) is defined to show the probability to jump from node \( i \) to \( j \), as shown in Eq. \( \text{MERGEFORMAT} \)  

\[
\begin{align*}
    w_{i,j} &= \begin{cases} 
        \exp(-\gamma \| x_i - \tilde{x}_j \|^2), & i \neq j \\
        0, & i = j
    \end{cases} \\
    P_{i,j} &= \frac{w_{i,j}}{d_i} = \frac{\sum_{k=1}^{N} w_{i,k}}{d_i}
\end{align*}
\]

The soil distribution is expressed as \( \tilde{F} = (F_{i_1}, F_{i_2}, \ldots, F_{i_k}) \), where \( F_{i_k} \) means the probability of the input \( \tilde{x}_i \) belongs to soil \( k \). Then the matrix \( F = (\tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_N)^\top \in \mathbb{R}^{N \times K} \) will be the target matrix to be determined by LPA. Each row in the matrix \( F \) represents the probability of each sample that belongs to different soil types. The first \( l \) rows in the matrix \( F \) are the labeled data and the last \( u \) rows are the unlabeled data. The label propagation algorithm is as follows:

**Step 1:** Initialize \( F^{(0)}, t = 0 \).

**Step 2:** \( F^{(t+1)} = PF^{(t)} \).

**Step 3:** \( t = t + 1 \).

**Step 4:** Repeat step 2 until \( F \) converges.

At last, the classes of the unlabeled samples can be predicted using Eq. \( \text{MERGEFORMAT} \)  

\[
\hat{y}_i = \arg \max_{j=\{1,2,\ldots,K\}} F_{ij}^{(t)}, i = l+1, l+2, \ldots, N
\]

### 3. A case study in Nanning metro

#### 3.1. Project description

Nanning Metro Line 1 is the first metro line with twin tunnels in Guangxi province. The section of left line in Nanning Metro Line 1 between Bai Cang Ling Station and Railway Station (BR section) is 1209 m (806 rings in total with segment width of 1.5 m), which is encountered with the mixed ground condition of mudstone and round gravel and is thus selected as a case study. The BR section was excavated with a Herrenknecht SPB shield machine with a diameter of 6.28 m. The tunnels in the BR section are surrounded by complex geological and hydrological conditions. The tunnel passing through geological profiles (illustrated in Figure 2) are mainly round gravel, mudstone, and sand, which are shown in different colors in Figure 2. The mixed ground containing mudstone locates around ring \#120 to ring \#220, and ring \#283 to ring \#470. A total of 36 boreholes (on average 23 rings between boreholes) have been drilled to obtain the geotechnical characteristics of the soils found along the BR section (refer to Table 1). The distance between the ground to the tunnel...
crown ranges from 14 m to 22 m, while the distance between the underground water table to the tunnel
crown is about 1.5m to 9m.

Figure 2 Schematic illustration of the cross-section view in BR section of Nanning metro line 1 # 100
means ring # 100 in the left line

Table 1 Geotechnical characteristics of round gravel and mudstone in the BR section

| Soil          | Soil natural density / $\rho$ g/cm³ | Poisson's ratio / $\nu$ | Cohesion / $c$ kPa | Friction angle / $\phi$ ° | Permeability coefficient / $K$ m/d |
|---------------|-----------------------------------|-------------------------|---------------------|--------------------------|---------------------------------|
| Round gravel  | 2.05                              | 0.27                    | 0.0                 | 35.0                     | 90                              |
| Mudstone      | 2.15                              | 0.20                    | 90.0                | 21.0                     | 0.01                            |

As the round gravel is a kind of water-rich strata with strong permeability, the SPB shield was
conducted in the BR section, which is beneficial for settlement control of the adjacent buildings. There
are two pressurized chambers named the excavation chamber and the working chamber in the SPB
shield machine. The excavation chamber is in the front that is filled with bentonite slurry to support
the tunnel face stability, and the working chamber is behind the excavation chamber with half
pressurized air cushion and half bentonite slurry. These chambers are separated by a submerged wall,
which has a hydraulic connection through an opening at the bottom of the submerged wall via slurry.
There are 16 pairs of double-cylinder are used, which provide the maximum THR of 42.575 MN (350
bar) with a maximum cylinder stroke of 2200 mm. The SPB cutterhead RS is up to 3 rpm while the
maximum AR is 50 mm/min. The rated TOR of the SPB cutterhead is 5.488 MN·m while the breakout
one is 6.619 MN·m.
The SPB worked well in the round gravel area with good settlement control (3 mm on average), stable
excavation efficiency (AR is 30 mm/min on average), and small wear of cutterhead (85% of cutters
can still use after driving 1468 m in the last section). However, in the mudstone rich area, clogging
inevitably occurs due to the clayey content and slurry properties, which resulted in serious
consequences including slurry oozing from the tunnel face to the ground, low excavation
efficiency (AR is smaller than 10 mm/min on average), and high wear of cutterhead (cutterhead jammed in the shield arrival process of the BR section). This paper aims to identify the soil type that
easy to trigger the clogging problem.

3.2. Data used for soil identification
The tunneling parameters were recorded via the programmable logic controller (PLC) every 10
seconds. 395 rings are selected in this study, which locates between ring #40 to ring #450 (some rings
data are not completed). Collected data included the following tunneling parameters: the slurry
pressure in the excavation chamber (SPE), the slurry pressure in the working chamber (SPW), the
cutterhead torque (TOR), the total thrust (THR), and the advance rate (AR). The descriptive statistical
parameters (average, standard deviation, maximum, and range) were determined for each ring to
identify the soil distribution ($\lambda$). The analysis concerned only the driving period (AR>0).
The mixed ratio ($\lambda$) is used to elevate the influence of the mudstone on clogging, as shown in E.q.
\( \lambda = \frac{H_m}{D} \)

\(^{\text{MERGEFORMAT (4)}}\)

where \( H_m \) is the mudstone thickness in the tunnel excavation face and \( D \) is the SPB cutterhead diameter. The \( H_m \) in the ring with borehole information is regarded as the labeled samples. According to the previous research and field observation, we define the clogging soil with \( \lambda > 0.15 \) and normal soil with \( \lambda \leq 0.15 \). To better evaluate the LPA prediction performance, we have made a judgment soil type (clogging or normal) about the rings without borehole information based on the excavated material and field observation. There are 17 rings with borehole information that are taken as the labeled rings \( (l = 17) \) and the other 378 rings without borehole rings are taken as the unlabeled rings \( u = 378 \). The \( l / u = 4.5\% \) that is suitable for the LPA model.

We will conduct two kinds of soil identification models in this study, one will employ the tunneling data of whole rings and the other one will use the first \( M \) minutes tunneling data in each ring for the soil identification. The latter one aims to investigate whether we can realize the early warning of clogging soil type as early as possible.

### 3.3. Results in clogging soil identification

The confusion matrix (Table 2) is used for the assessment of the performance of the LPA method. In this matrix, the Positive (P) value refers to the clogging soil, while the Negative (N) value refers to the normal soil. The model performance is determined using the accuracy, precision, recall, and F1 score (E.q.). The high precision indicates a low false positive (FP) rate; the high recall refers to a low false negative (FN) rate.

| Predicted class | Actual class | \( TP \) | \( TP + TN \) | \( TP + TN + FP + FN \) | \( \times 100\% \) | \( TP \) | \( \times 100\% \) | \( TP + FP \) | \( TP + FN \) | \( \times 100\% \) | \( TP + FN \) | \( \times 100\% \) | \( TN \) | \( \times 100\% \) | \( TP + FN + FP + TN \) |
|----------------|-------------|--------|----------------|---------------------|------------------|--------|----------------|----------------|----------------|------------------|----------------|----------------|--------|----------------|---------------------|
| P: Clogging soil | P: Clogging soil | \( TP \) | \( TP + TN \) | \( TP + TN + FP + FN \) | \( \times 100\% \) | \( TP \) | \( \times 100\% \) | \( TP + FP \) | \( TP + FN \) | \( \times 100\% \) | \( TP + FN \) | \( \times 100\% \) | \( TN \) | \( \times 100\% \) | \( TP + FN + FP + TN \) |
| N: Normal soil | N: Normal soil | \( FN \) | \( FN + TN \) | \( FN + TN + FP + FP \) | \( \times 100\% \) | \( FN \) | \( \times 100\% \) | \( FN + TN \) | \( FN + FP \) | \( \times 100\% \) | \( FN + FP \) | \( \times 100\% \) | \( TN \) | \( \times 100\% \) | \( FN + TN + FP + FP \) |

We first conduct numerical experiments to determine the optimal \( \gamma \) for the LPA model. The candidate values of \( \gamma \) are 1, 5, 10, 20, 40, 60, 80, 100, and we use the whole ring data for the LPA model. Figure 3 shows the F1 score of each \( \gamma \) and we can see that when the \( \gamma = 60 \), the F1 score will be 0.9. Therefore, we selected 60 as the optimal \( \gamma \) value in the following process.

To compare different model performances, we also employ the support vector classification (SVC) model[14] and the random forest (RF) model[15]. These supervised learning models using the labeled rings as the training data set and the unlabeled rings as the test data set. The hyper-parameters are determined via the 5-fold cross-validation with a randomized search method[16]. The hyper-parameters (the number of trees and the maximum depth of the tree) distribution for the RF model is
the random number between 0 to 100. The RBF function is adopted as the kernel function in the SVC model. It has logarithmically spaced $C$ values in the range 1 - 1000 with a vector length of 20, a logarithmically spaced $gamma$ from 0.001 - 1000 with a vector length of 7. Figure 4 summarizes the model performance for clogging soil type identification. It shows that the proposed LPA approach performs the best, followed by the RF model, and finally the SVC model. With the whole ring tunneling data, the LPA model provides good performance with accuracy, precision, recall, and F1 score around 90%. This performance is acceptable in tunneling construction.

3.4. Discussion

3.4.1. Early warning of clogging soil types with the LPA model
As mentioned before, we plan to employ the $M$ minutes tunneling data at the beginning of each ring to identify the clogging soil type. Here we employ the $M = 1, 2, 3, 4, 5, 6, 7, 8, 10$ and check the model performances. Figure 5 shows the model performances with different tunneling data lengths. It shows a great improvement in the model performance at the beginning with the increase in the data length, then the model performance decreases slightly, and finally, the model performance remains stable as the tunneling data length increases. With two minutes of tunneling data, the LPA model can achieve a prediction accuracy of 86% for clogging soil distribution.

3.4.2. Feature importances of the LPA model
To better understand the influence of tunneling data on the identification of clogging soil type, we carry out an analysis of the input features. We present 10 scenarios by dropping or keeping one certain
kind of input feature, and compare the F1 score with the LPA model, as illustrated in Figure 6. For example, we drop the SPE features means we only use the statistical parameters (average, standard deviation, maximum, and range) of SPW, THR, TOR, and AR as model input. It can be found in Figure 6 that by dropping only one kind of feature, there is a little diminution of the F1 score. Among dropping one kind of feature scenario, AR is the most significant one with a decrement of F1 score about 0.07. When we only use one kind of feature to predict soil distribution, the model performs poorly in the cases of dropping SPW and THR. However, the AR can achieve an F1 score larger than 0.9, which means we can just use the AR statistical indexes to predict the clogging soil distribution. The above rules can be explained by field experience during the SPB shield tunneling. The low values of AR are always encountered when clogging occurs and there is a strong negative correlation between the statistical indexes of AR and clogging soil distribution (correlation coefficient around -0.75).

4. Conclusion and future work

This paper presented a methodology for the development and use of a stratigraphic heterogeneity inference method for shield tunneling in the mixed mudstone-gravel ground, which is based on the label propagation algorithm. The comparison of the performances of different machine learning methods on the collected data in Nanjing metro showed that the LPA method high perform the other supervised methods. The LPA method could be used for clogging soil distribution early warning with only two minutes of tunneling data at the beginning of one ring, which realized the accuracy of 0.86. Analysis of the importance of input features on soil distribution prediction revealed the crucial role of the AR in clogging soil prediction.

Although good achievements have been realized in the prediction of clogging soil distribution using the semi-supervised algorithm, there is yet an important need to develop a filed observation method that can identify the soil types as shield machine advance. The LPA-based method seems promising in the early identification of the hard rock TBM project because the rock type can be determined more clearly during the tunneling process.

5. References

[1] X. Xie, Q. Wang, Z. Huang, Y. Qi, Parametric analysis of mixshield tunnelling in mixed ground containing mudstone and protection of adjacent buildings: case study in Nanning metro, Eur. J. Environ. Civ. Eng. 22 (2018) s130–s148.
[2] J. Zhao, Q.M. Gong, Z. Eisensten, Tunnelling through a frequently changing and mixed ground: A case history in Singapore, Tunn. Undergr. Sp. Technol. 22 (2007) 388–400.
[3] N. Tokgöz, I. Serkan Binen, E. Avunduk, An evaluation of fine grained sedimentary materials in terms of geotechnical parameters which define and control excavation performance of EPB TBM’s, Tunn. Undergr. Sp. Technol. 47 (2015) 211–221.
[4] X. Wang, H. Wang, R.Y. Liang, H. Zhu, H. Di, A hidden Markov random field model based approach for probabilistic site characterization using multiple cone penetration test data, Struct. Saf. 70 (2018) 128–138.
[5] H.Z. Cheng, J. Chen, R.P. Chen, G.L. Chen, Reliability study on shield tunnel face using a random limit analysis method in multilayered soils, Tunn. Undergr. Sp. Technol. 84 (2019) 353–363.
[6] X. Zhou, Y. Xia, J. Xue, Neural Network Strata Identification Based on Tunneling Parameters of Shield Machine, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2009: pp. 392–401.
[7] B. Liu, R. Wang, Z. Guan, J. Li, Z. Xu, X. Guo, Y. Wang, Improved support vector regression models for predicting rock mass parameters using tunnel boring machine driving data, Tunn. Undergr. Sp. Technol. 91 (2019) 102958.
[8] J. Zhao, M. Shi, G. Hu, X. Song, C. Zhang, D. Tao, W. Wu, A Data-Driven Framework for Tunnel Geological-Type Prediction Based on TBM Operating Data, IEEE Access. 7 (2019) 66703–66713.
[9] C. Olivier, S. Bernhard, Z. Alexander, Introduction to Semi-Supervised Learning, in: O. Chapelle, B. Scholkopf, A. Zien (Eds.), Semi-Supervised Learn., The MIT Press, 2013: pp. 1–12.

[10] H. Chai, Y. Liang, S. Wang, H. wei Shen, A novel logistic regression model combining semi-supervised learning and active learning for disease classification, Sci. Rep. 8 (2018) 1–10.

[11] D. Sen, A. Aghazadeh, A. Mousavi, S. Nagarajaiah, R. Baraniuk, A. Dabak, Data-driven semi-supervised and supervised learning algorithms for health monitoring of pipes, Mech. Syst. Signal Process. 131 (2019) 524–537.

[12] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, Prabhat, Deep learning and process understanding for data-driven Earth system science, Nature. 566 (2019) 195–204.

[13] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. 76 (2007) 1–12.

[14] Q. Zhang, Z. Liu, J. Tan, Prediction of geological conditions for a tunnel boring machine using big operational data, Autom. Constr. 100 (2019) 73–83.

[15] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[16] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

Acknowledgments
This work was supported by the National Key R&D Program of China under grant 2019YFC0605103, National Nature Science Funds of China under grant 52038008, and Shanghai Science and Technology Development Funds under grant 20DZ1202004. Thanks also go to China Railway 16 Bureau Group Beijing Metro Engineering Construction Co., Ltd. for providing the data about the section from Bai-cang-ling station to Railway station of Nanning metro line 1.