Geomechanics and Engineering, *Vol. 29, No. 5 (2022) 523-533*
https://doi.org/10.12989/gae.2022.29.5.523

Sequential prediction of TBM penetration rate using a gradient boosted regression tree during tunneling

Hang-Lo Lee[1a], Ki-Il Song*[2], Chongchong Qi[3b] and Kyoung-Yul Kim[4c]

[1]*Disposal Performance Demonstration Research Division, Korea Atomic Energy Research Institute, Daejeon, 34057, Republic of Korea*
[2]*Department of Civil Engineering, Inha University, Incheon 22212, Republic of Korea*
[3]*School of Resources and Safety Engineering, Central South University, Changsha, 410083, China*
[4]*Next Generation Transmission & Substation Laboratory, KEPCO Research Institute, Daejeon, 34057, Republic of Korea*

(Received August 27, 2021, Revised January 12, 2022, Accepted January 13, 2022)

Abstract. Several prediction model of penetration rate (PR) of tunnel boring machines (TBMs) have been focused on applying to design stage. In construction stage, however, the expected PR and its trends are changed during tunneling owing to TBM excavation skills and the gap between the investigated and actual geological conditions. Monitoring the PR during tunneling is crucial to rescheduling the excavation plan in real-time. This study proposes a sequential prediction method applicable in the construction stage. Geological and TBM operating data are collected from Gunpo cable tunnel in Korea, and preprocessed through normalization and augmentation. The results show that the sequential prediction for 1 ring unit prediction distance (UPD) is $R^2 \geq 0.79$; whereas, a one-step prediction is $R^2 \leq 0.30$. In modeling algorithm, a gradient boosted regression tree (GBRT) outperformed a least square-based linear regression in sequential prediction method. For practical use, a simple equation between the R^2 and UPD is proposed. When UPD increases R^2 decreases exponentially; In particular, UPD at $R^2=0.60$ is calculated as 28 rings using the equation. Such a time interval will provide enough time for decision-making. Evidently, the UPD can be adjusted depending on other project and the R^2 value targeted by an operator. Therefore, a calculation process for the equation between the R^2 and UPD is addressed.

Keywords: construction stage; gradient boosted regression tree; penetration rate; sequential prediction; tunnel boring machine

1. Introduction

With the successful introduction of mechanized tunneling methods in geotechnical projects, estimation of a tunnel boring machine (TBM) performance has become an essential step for establishment of the construction period and time schedule of a project. TBM performance has been evaluated as several factors such as penetration rate(PR), penetration per revolution, boreability index, and thrust per disc cutter (Benato and Oreste 2015, Chang *et al.* 2006, Gong and Zhao 2009). For evaluation of the TBM performance, experimental studies using a linear cutting machine have been comprehensively performed (Ozdemir and Wang 1979, Sanio 1985, Sato 1991, Snowdon *et al.* 1982). The experimental studies primarily focused on TBM cutterhead design such as determination of disc cutter spacing (Cho *et al.* 2010), the number of cutter (Rostami 1997), and cutterhead layout (Huo *et al.* 2011). However, because the experimental results are based on the intact rock or simple jointed rock, it has limitation to a prediction of a field TBM performance at in-situ rock. As an alternative approach, empirical method based on field data has also been carried out because field data are involved with several factors of geological and geotechnical properties and TBM operating condition along the tunnel alignment (Bruland 1999, Hassanpour *et al.* 2011, Yagiz 2008).

To utilize an empirical-based model, a site investigation is generally preceded. However, as the geological survey such as in-situ coring are partially conducted, it is difficult to represent the detailed geological characteristics along the tunnel, especially in the unknown region between boring holes (Mazzoccola *et al.* 1997). Furthermore, TBM performance is dependent on a TBM operator's skill (Hammerer 2015). Therefore, empirical models are mainly limited to the rough prediction of construction period in the design stage (Hassanpour *et al.* 2016).

In construction stage, the expected PR and their trends may be changed during tunneling due to operator's skills and the gap between the expected and actual geological conditions. Monitoring these changes during tunneling is very important in terms of adjusting the excavation plan in real-time. Although similar study pertaining to statistical based prediction of TBM thrust has been conducted (Lee *et al.* 2021), the PR prediction in construction stage has not been extensively performed.

*Corresponding author, Professor
E-mail: ksong@inha.ac.kr
[a]Ph.D.
E-mail: hanglolee@kaeri.re.kr
[b]Professor
E-mail: chongchong.qi@esu.edu.cn
[c]Ph.D.
E-mail: solasido@kepco.co.kr

ISSN: 2005-307X (Print), 2092-6219 (Online)

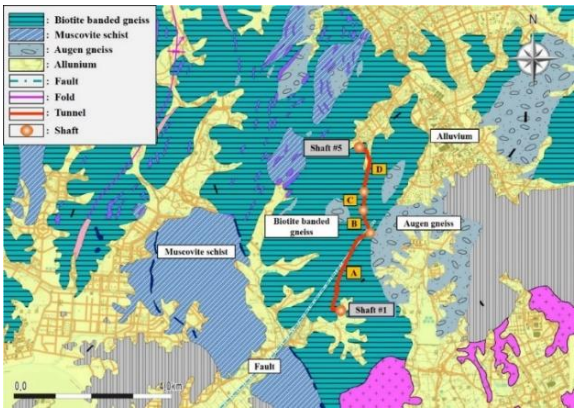


Fig. 1 Geological map of GCT project

In recent years, machine learning (ML) techniques have been used to examine the complex relationship between inputs and targets in the geomechanics engineering field, such as clogging evaluation of tunneling (Bai *et al.* 2021), ground settlement induced by tunneling (Zhang *et al.* 2020). Although ML based model for PR prediction in design stage have been proposed (Gao *et al.* 2021, Gao *et al.* 2019, Yagiz and Karahan 2015), a prediction of PR in construction stage using ML techniques have not been comprehensively performed. Among the various ML algorithm, ensemble-based gradient boosted regression tree (GBRT) provides exclusive advantages: (1) it is a series of ensemble model that combines several single decision trees to improve the prediction performance and resist outliers, (2) it also has been validated to have a better accuracy compared with a different single model (Pedregosa *et al.* 2011).
Hence, a sequential prediction method applicable in the construction stage is proposed. The method is applied to the Gunpo cable tunnel in Korea and validated by comparing it with the one-step prediction method. In addition, the comparative analysis is performed between a GBRT and least square-based linear regression algorithm. For practical purpose, PR predictions based on unit prediction distance is examined and discussed.

2. Description of the site and data collection

2.1 Gunpo cable tunnel

The Gunpo Cable Tunnel (GCT) was designed as a 5.09 km long power supply tunnel connecting the West Seoul substation and the Sanbon substation in South Korea. The geological formations of the project generally comprise pre-cambrian biotite banded gneiss and some intrusive rocks, particularly, a small amount of quartzite and limestone, which are interspersed throughout the metamorphic rock (Fig. 1).

The GCT is divided into four sections according to construction method: Section-A (2.21 km) as shield TBM, Section-B and D (0.75 km, 1.63 km) as open-cut, and Section-C (0.46 km) as semi-shield TBM. Section-A considered in this study is located at a depth of 45–51.5 m below the ground surface, which is lower depth than the groundwater level. Since the coefficient of permeability is the range from 3.95×10^{-5} to 1.26×10^{-4} cm/s, the influence of groundwater infiltration seems to be negligible. The rock mass rating of surrounding rock belongs to 41-60 RMR, which is classified as '*Fair*' condition (Bieniawski 1989). An earth pressure balance (EPB) shield type was used in the Section-A, and the excavation diameter is 3.4 m. The primary specification for the shield is summarized in Table 1.

Table 1 Main specifications of EPB shield TBM

Parameter	Value
Excavation diameter	3.41 m
Cutter diameter	330 mm
Number of disc cutter	26
Maximum cutterhead thrust	9,600 kN
Maximum cutterhead torque	1,250 kN-m
Maximum cutterhead revolution per minute	9.0 rev/min

2.2 Rock properties and TBM operating conditions

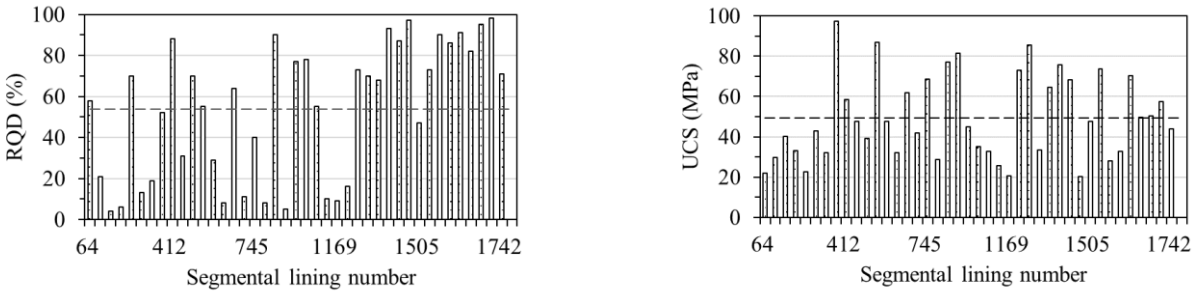
To determine the quantitative characteristics of the rock mass for Section-A, 49 NX-size core samples were obtained. The discontinuity of rock mass has a huge influence on crack initiation and propagation, thus affecting the penetration rate of TBMs (Gong *et al.* 2005, Gong *et al.* 2006). This can be mostly quantified by joint spacing, joint orientation, and rock quality designation (RQD) (Deere 1968). The RQD has been identified as the major factor that influences the PR of TBMs and it has the advantage of being easy to calculate in-situ without additional equipment (Hassanpour *et al.* 2010). RQD provide a quantitative estimate of rock mass quality from drill cores. It is defined as the length of intact core pieces longer than 10 cm over the total length of core (Eq. (1)). The core rock should be NX size and drilled with a double-tube core barrel.

$$RQD = \frac{\sum Length\ \ of\ \ core\ \ piecies > 10\ \ cm\ \ length}{Total\ \ length\ \ of\ \ core\ \ run} \quad (1)$$

The rock strength is also known to be a major factor in the PR of TBMs (Rostami 1997). When the disc cutter indents the rock mass, the stress must be higher than the rock's strength. The rock strength thus is directly proportional to the PR of TBMs. Uniaxial compressive strength (UCS) is a representative factor that measures the strength of rocks. UCS is the ultimate strength when a failure occurs on the specimen and is calculated using the failure load (P) and cross area of the specimen (A) (Eq. (2)).

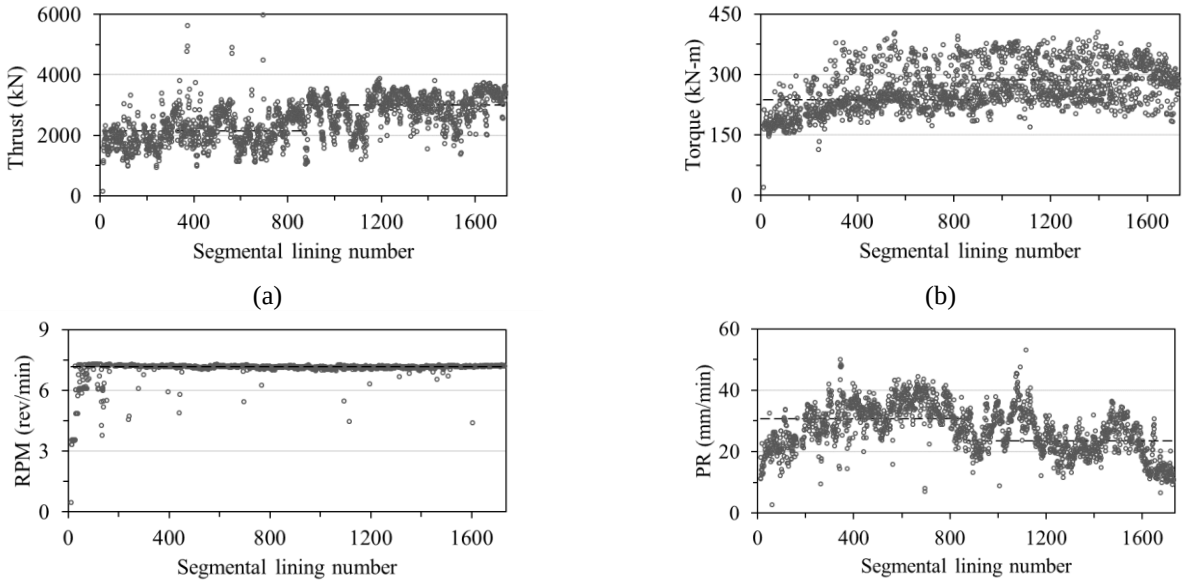
$$UCS = \frac{P}{A} \quad (2)$$

This study followed the criteria of the American Society for Testing and Material (ASTM) D-2938 to obtain the UCS. Among the 49 samples, 41 were tested except for weathered or fractured samples. Fig. 2 shows the distribution of RQD and UCS along the tunnel alignment.



(a) (b)

Fig. 2 Distribution of (a) RQD and (b) UCS based on the segmental lining number



(a) (b)

(c) (d)

Fig. 3 Distribution of (a) thrust, (b) torque, (c) RPM and (d) PR based on the segmental lining number

The TBM used is equipped with a system that measures TBM operational values at intervals of 20 mm. To facilitate data processing, the operational values corresponding to each segmental lining were averaged and recorded in a comma-separated values format. Operational parameters include thrust, torque, and RPM. The thrust and torque exerted to tunnel face gradually increased as TBM progressed (Fig. 3). On the other hand, PR increased in the first half and decreased in the second half. It is believed that it occurred because rock quality in the second half section is higher than condition in the first half tunnel.

3. Methodology

3.1 Tree-based GBM

In machine learning technique, ensemble based models has been evaluated to have a better accuracy compared with a different single model such as artificial neural networks and support vector machine (Pedregosa *et al.* 2011). In practice, the ensemble approach depends on combining several relatively weak and simple models to ensure a stronger ensemble performance. The most prominent example of this machine learning ensemble technique is the gradient boosting machine (Friedman 2001).

The gradient boosting machine (GBM) is based on constructive strategies for ensemble formation. The main principle behind boosting is to combine a new model to the ensemble sequentially. In each particular iteration, a weak model called the base-learner is trained with respect to the residual of the entire ensemble that has been learned so far.

The GBM, a type of boosting, is an algorithm that allows the new base-learner to be maximally associated with the negative slope of a loss function called the pseudo residuals. The loss function, which is related to the entire ensemble, is subject to the user's choice and has the potential to exploit various loss functions depending on the nature of the problem. To use a particular loss function, a loss function and its corresponding negative slope function must be specified and therefore we used the GBM algorithm in this study. The commonly known loss functions include squared-error L2 loss, absolute L1 loss, Huber loss, and quantile loss functions (Friedman 2001,

Specify training data $(x_i, y_i)_{i=1}^n$, differentiable loss function L, iteration M

Choose the initial model $\hat{F}_0$

$$\hat{F}_0(x) = \underset{\rho}{argmin} \sum_{i=1}^{n} L(y_i, \rho)$$

For $m = 1$ to M

Calculate the negative gradient, called pseudo residuals

$$-\mathrm{g}_m(x_i) = -\left[\frac{\partial\ L\left(y_i, F(x_i)\right)}{\partial\ F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

Fit the base learner h_m to pseudo residuals

Compute the step size ρ_m by solving the following one-dimensional optimization problem

$$\rho_m = \underset{\rho}{argmin}\left[\sum_{i=1}^{n} L\left(y_i, \hat{F}_{m-1}(x) - \rho \mathrm{g}_m(x_i)\right)\right]$$

Renew the model

$$\hat{F}_m(x) = \hat{F}_{m-1}(x) - \rho_m h_m(x)$$

Output the $\hat{F}_m(x)$

Fig. 4 GBM algorithm suggested by Freidman (2001)

Koenker and Hallock 2001). Specific GBMs can also be designed through other base-learner models. The commonly used base-learners are classified into three distinct categories: linear model, smooth model, and decision trees.

This high flexibility allows the GBM to be easily customizable for specific data-driven tasks. This gives a considerable amount of freedom in the model design; thus, one can choose the appropriate loss function for the problem of trial and error. However, the boosting algorithm is relatively simple to implement and allows for experimenting with various model designs. The GBM has also been very successful in the geotechnical engineering field as well as in solving various machine learning and data mining issues (Qi *et al.* 2018).

Gradient boosting is an approach for searching an approximation $\hat{F}$ in the form of a weighted sum of function h, called the basic learner as shown in Eq. (3)

$$\hat{F}_m(x) = \sum_{m=1}^{M} \rho_m h_m(x) + const. \tag{3}$$

It operates by starting with a constant function $\hat{F}_0(x)$, and then expands it incrementally in a greedy stagewise as shown in Eqs. (4) and (5)

$$\hat{F}_0(x) = \underset{\rho}{argmin} \sum_{i=1}^{n} L(y_i, \rho) \tag{4}$$

$$\hat{F}_m(x) = \hat{F}_{m-1}(x) + \underset{\rho_m, h_m}{argmin}\left[\sum_{i=1}^{n} L(\ y_i, F_{m-1}(x_i) + \rho_m h_m(x_i))\ \right] \tag{5}$$

where h_m is a base-learner function, introduced to discriminate it from the entire ensemble function estimate $\hat{F}$. Unfortunately, seeking the function h at each iteration for a loss function L is a computationally infeasible optimization problem in general. To address this problem, choosing a new function h_{m} which is the most parallel to the following negative slope $(-\mathrm{g}_m)$, called a pseudo residual, was proposed as following Eq. (6)

$$-\mathrm{g}_m(x_i) = -\left[\frac{\partial\ L(y_i, F(x_i))}{\partial\ F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{6}$$

Substituting the above equation in Eq. (3), we get Eqs. (7) and (8)

$$\hat{F}_m(x) = \hat{F}_{m-1}(x) - \rho_m \mathrm{g}_m(x_i) \tag{7}$$

$$\rho_m = \underset{\rho}{argmin}\left[\sum_{i=1}^{n} L\left(y_i, \hat{F}_{m-1}(x) - \rho \mathrm{g}_m(x_i)\right)\right] \tag{8}$$

where ρ_m is the step size. For finite data, if we can choose the function h closest to the gradient of the loss function L, ρ may be calculated from the line search using Eq. (8). The algorithm is summarized in Fig. 4.

The GBM typically uses Classification and Regression Tree (CART) decision trees as base-learners. It is called the gradient boosted regression tree (GBRT). Let J_m be the number of terminal leaves on the mth decision tree. The mth decision tree as shown in Eq. (9), splits the input space into disjoint regions, R_{1m}, R_{2m}, . . . , $R_{J_m m}$ and predicts a constant value b_{jm} in each subspace

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm}\, \mathbf{1}\left(x \in R_{jm}\right) \tag{9}$$

where $\{R_{jm}\}_1^{J_m}$ is the region defined by the terminal node of the mth decision tree. The indicator function $\mathbf{1}(\cdot)$ has a value of 1 if the argument is true, otherwise it is zero. Then, as shown in step 3 in Fig. 5, the tree-based $h_m(x)$ is multiplied by ρ_m and finally expressed as Eqs. (10) and (11)

$$\hat{F}_m(x) = \hat{F}_{m-1}(x) - \sum_{j=1}^{J_m} \gamma_{jm}\, \mathbf{1}\left(x \in R_{jm}\right) \tag{10}$$

$$\gamma_{jm} = \underset{\gamma}{argmin}\left[\sum_{x_i \in R_{jm}} L\left(y_i, \hat{F}_{m-1}(x) - \gamma\right)\right], \quad for \quad \gamma_{jm} = \rho_m b_{jm} \tag{11}$$

Non-parametric models such as GBM have non-adjustable parameters from a training data, which are called hyper-parameters. This must be defined before building the model. Iteration M in the GBM, one of the representative hyper-parameter, means the number of single decision tree. It controls the complexity of ensemble trees. The maximum number of leaf node and learning rate are also another essential hyper-parameter which are respectively regularizes the complexity of trees and controls the degree of contribution to single decision tree.

3.2 Data preparation

Huge amount of data is one of the major requirements in machine learning. A lack of data can lead to over-fitted models. However, only 41 core rock samples were obtained from the site investigation along the tunnel alignment, which is much less than the 1,641 of TBM operational data. Loss of operational data between boreholes is inevitable owing to the mismatch between geological information and operational data. Therefore, both operational and geological data should be properly modified to utilize all the available data.

In this study, a weighting variable expressed by a distance was introduced. The weighting variable of the *n*th segmental lining is defined as the distances between two adjacent boreholes from the position of the *n*th segmental lining (see Fig. 5). The weighting variable makes it possible to retrieve the geological information of the two adjacent core rock for the *n*th segmental lining. Introducing the weighting variable, available variables for arbitrary *n*th segmental lining have three groups (total 9 variables): TBM operational variables (thrust, torque, and RPM), geological variables (uniaxial compressive strength, and RQD of rear and front core samples), and weighting variables (negative and positive distances). Through this procedure, a number of data is augmented from 41 to 1641.

The performance of tree-based GBM (or GBRT) are sensitive to the scales of input variables. If a scale for input variable is too large, the GBRT tends to overestimate the importance of corresponding input variable. To prevent the problem, all the inputs variables were normalized using the Eq. (12) in the range from 0 to 1 for eliminating a scale effect

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{12}$$

3.3 Process of sequential prediction

PR prediction method in design stage predicts the PR for an entire target section in one-step (one-step prediction method) using a constructed model (see Fig. 6(a)). As mentioned in Introduction, it is restricted to the rough prediction of construction period because of limited geological information and operator's skill; hence it is not proper method for application in construction stage. Additional consideration of a TBM monitoring data such as thrust, torque, and RPM during tunneling is important for accurate prediction during tunneling. Sequential prediction makes the monitoring data reflected in a prediction model in real-time.

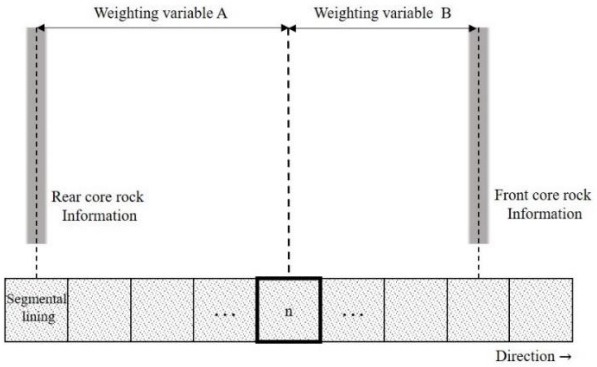


Fig. 5 Weighting variables for n-th segmental lining

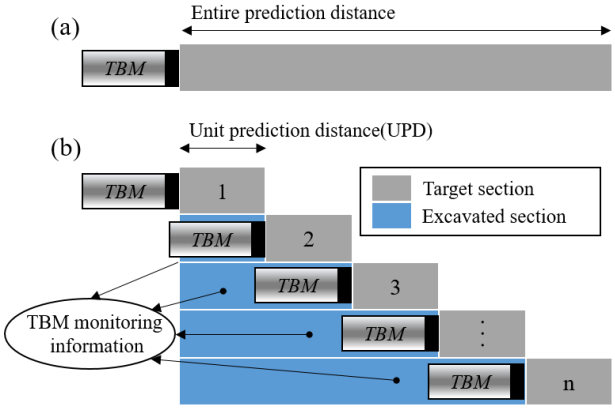


Fig. 6 (a) One-step prediction and (b) sequential prediction method

In the process of the sequential prediction, a prediction model is renewed by using additional monitoring data in the excavated section (Fig. 6(b)); subsequently, predicts PRs using subsequent input data obtained from the section within unit prediction distance (UPD). This cycle is repeated until TBM excavation is completed. Here, it was assumed that the expected operating condition within UPD is similar with recent TBM operating condition. When a tunnel excavation is finished, all of the predicted values are evaluated all at once. The strength of sequential prediction is to stabilize the variance of the model's coefficients for a PR prediction, which makes the accuracy of prediction performance improves.

To progress the sequential prediction method, models should be optimized automatically. The most important task in model fitting is to optimize hyper-parameter set because the model performance varies significantly depending on the hyper-parameter set. Although there are several criteria for selecting a hyper-parameter set, k-fold cross validation (k-fold CV) is commonly used in the data mining field. The principle of k-fold CV is to divide the training data into k

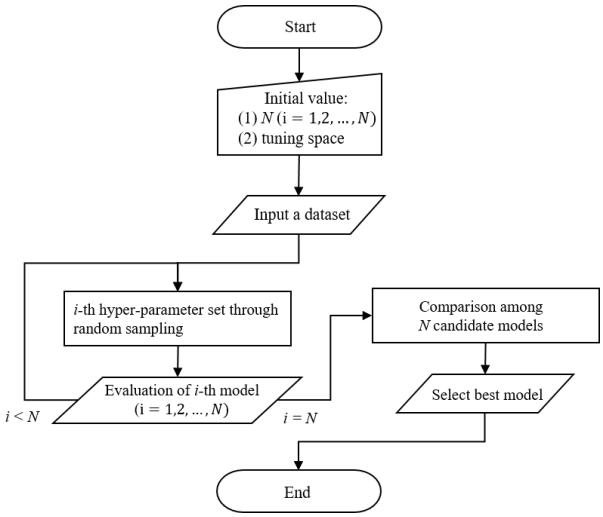


Fig. 7 Automated algorithm for model construction

folds, and then one fold is used to validate a model which was fitted by the remaining folds. In general, a value of 5 or 10 is used as k values in the engineering field. Herein, we selected the value 5 for the analysis.

To achieve the optimum hyper-parameter, the GBRT-base algorithm was developed based on the above criterion (Fig. 7). Once the data is entered in the algorithm, a model having a particular hyper-parameter set is constructed. Here, the hyper-parameter set is obtained through random sampling within a tuning space (see Table 2). This candidate model is validated by a 5-fold CV. This process is repeated N times ;and then the best model, which is the highest performance for CV, is selected among the N candidate models,

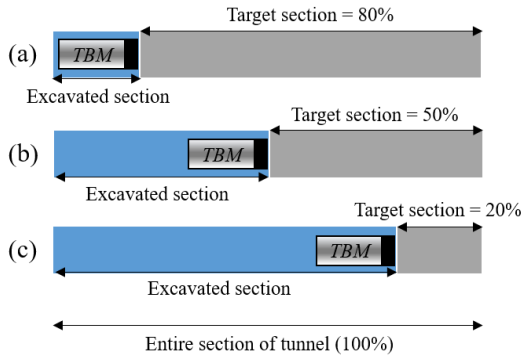


Fig. 8 Three simulation cases based on the portion of future target section in entire section

3.4 Model assessment

A common definition of the coefficient of determination (R^2) is the proportion of variation of the response variable that can be explained by the model. It primarily agrees with this statement only if the predicted value by the training data comes from an ordinary regression, which is made by the training data. However, this is not the case for applying models for testing data. Special care should be placed on the performance of the model on its accuracy and precision, and not on how well it explains the variation in specific data (Alexander *et al.* 2015). Kvålseth (1985) defines R^2, with the sum of squared residual (SSR) and the sum of squared deviation (SSD) as shown in Eq. (13)

$$R^2 = 1 - \frac{\text{Sum of squared residual (SSR)}}{\text{Sum of squared deviation (SSD)}} \quad (13)$$

Where y_i is the measured target variable, $\bar{y}$ is the mean of measured values, and $\hat{y}_i$ is the corresponding predicted value. Eq. (13) measures the magnitude of residuals from the model over the magnitude of the residuals for a null model, where all predictions are the constant value averaged ($\bar{y}$). The best possible score is 1.0 if the value predicted from the model is equal to the measured value. In the case that a null model always predicts the average of measured values, ignoring the features of the input variables, would get an R^2 value of zero. Sometimes, R^2 may show negative values for testing data when the model has worse than the null model (Alexander *et al.* 2015).

4. Results and discussions

4.1 Results of one-step and sequential prediction

Depending on the portion of future target section in entire section, three cases of 80, 50, and 20% are chosen to implement the sequential prediction (Fig. 8). Herein, the target section is considered as a new project with similar geological conditions and TBM specifications of excavated section.

The modeling package used in this study is Scikit-learn ver. 0.24.1 in Python ver. 3.8.8 (Pedregosa *et al.* 2011). Initial models are constructed using the data from the excavated section. As initial values for model construction in Fig. 7, the number of hyper-parameter set (N) is set to 200, The tuning spaces are specified through trial-and-error and by referring to the literature (Lu *et al.* 2019). The range of maximum number of leaf node is 2-5, the number of base

Table 2 Tuned GBRT models through 5-fold cross validation (CV)

Training data	Hyper-parameter			R^2	
	Max. num. of leaf node	Num. of single decision tree	Learning rate	Training set in CV	Validation set in CV
20%	4	243	0.095	0.94	0.72
50%	4	243	0.095	0.86	0.71
80%	4	243	0.095	0.86	0.76

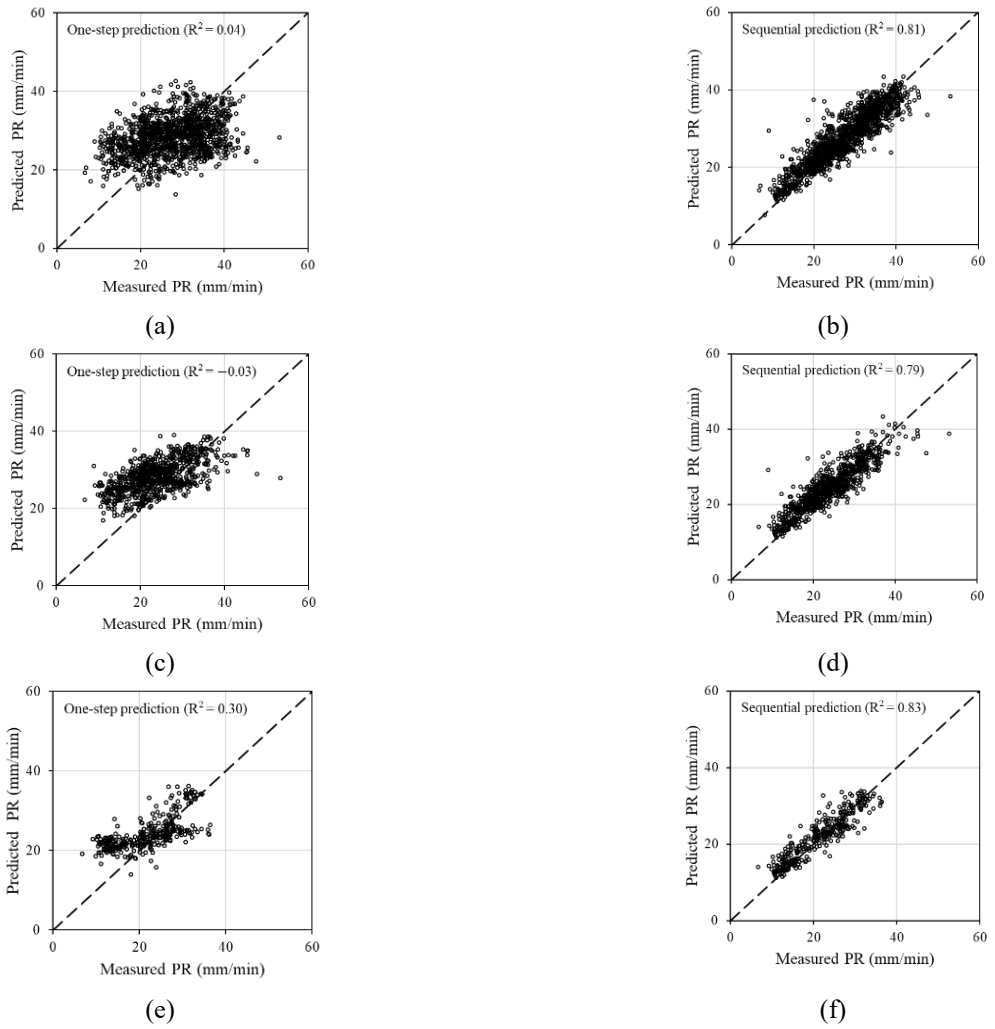


Fig. 9 Scatter plots of measured and predicted PRs according to the initial target section of (a)-(b) 80%, (c)-(d) 50%, and (e)-(f) 20%

learner is 1-250, and the learning rate is 0.01-0.1. The results of tuned GBRT models using 5-fold cross validation is summarized in Table 2.

The three initial models for each cases are applied to the future target section. As mentioned in Fig. 6, the simulation was tested in two prediction approach; (1) one-step prediction, (2) sequential prediction. Here, the unit prediction distance (UPD) is designated as 1 ring of segmental lining.

The scatter plots of measured and predicted values based on the cases of target section 20, 50, and 80% are shown in Fig. 9. Here, the dashed line shows the relationship $y = \hat{y}$. Data points for a good prediction would lie close to dashed line. The sequential prediction performed better than the one-step prediction regardless of the portion of target section. The R^2 from one-step prediction is less than 0.3 even after the model validation. Furthermore, it gives negative values in the case of target section 50%, which means that the predictive performance is worse than the null model that always predicts the mean of measured PR values. However, the R^2 from the sequential prediction is higher than 0.79. It is believed that it occurred because the PR prediction is a kind of extrapolation problem where the geological and TBM operating conditions may be out of range. To describe the difficulty of this problem, the extrapolation prediction is

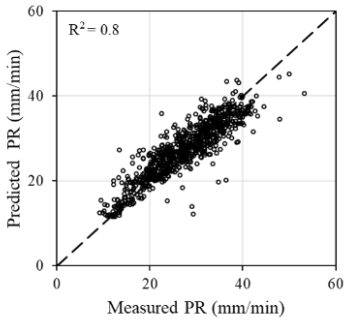


Fig. 10 Scatter plots of measured and predicted PRs for one-step prediction in interpolation approach

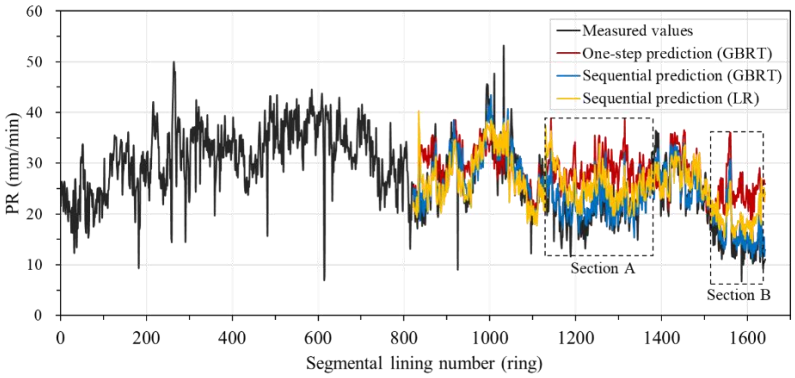


Fig. 11 PR curves based on prediction method and modeling algorithm in the case of target section 50%

compared with interpolation prediction. To simulate the interpolation approach, 50% of entire data were randomly extracted and applied to the one-step prediction method. Although the R^2 for extrapolation approach showed -0.03 as mentioned in Fig. 9(c), R^2 for interpolation approach is 0.79 (Fig. 10) which is considerably higher than that from extrapolation approach. This indicated that one-step prediction is not always guaranteed to ensure the accuracy of PR prediction.

The portion of the target section had almost no influence on the R^2 value (see Fig. 8). When the portion of target section decreases, the sum of squared residual (SSR) decreases, but the sum of squared deviation (SSD) also decreases proportionally (see Eq. (13)). Here, SSD is kind of a prediction error using a null model that always predicts the average of measured values. It can be believed that if the lower the SSD, the lower the difficulty of the prediction problem. Therefore, the results mean that the sequential prediction guarantees the quality of predictive performance regardless of the portion of target section in spite of decreasing the SSR.

4.2 Comparison between two different modelling algorithm

For comparison between modeling algorithms in sequential prediction condition, the GBRT is compared with least square-based linear regression(LR). As shown in Fig. 11, the average value for the GBRT algorithm is 24 mm/min which is significantly similar to the average value for measured PR (24 mm/min) in target section 50%. In addition, the PR curve for GBRT closely resembles the trend of measured PR. However, the PR curve for LR algorithm relatively shows a stiffed pattern, and the average 28 mm/min. In particular, over-predictions for LR are observed in the Section A and B (see Fig. 11). It is believed to be due to the simplicity of the LR model, which does not take into account complex relationship between inputs and output. Evidently, the performance for this sequential prediction using LR shows higher than the one-step prediction using GBRT. In particular, the prediction error for one-step prediction was relatively high at 1200-1600 rings. As shown in Figs. 2 and 3, although higher thrust and torque exerted to the section of 1200-1600 rings, PR was conversely lower than the other section due to the relatively high rock quality designation(RQD). It is believed that it occurred because such a one-step prediction method is not adaptive to upcoming geological conditions during tunneling. This finding suggests that the prediction method have more influence on the prediction performance than the modeling algorithm.

4.3 Prediction depending on unit prediction distance

Till now, the performance for sequential prediction was examined when the unit prediction distance(UPD) is 1 segmental lining (units: ring). Practically, the optimal UPD should be determined. In this study, R^2 is evaluated based on the UPD in target section 50%. The UPD ranges from 1 to maximum UPD. Herein, the R^2 at maximum UPD

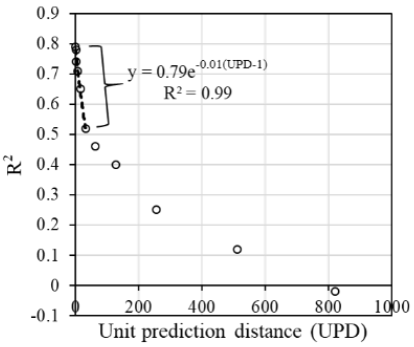


Fig. 12 Changes in R2 based on UPD in target section 50%

exactly same with R^2 for one-step prediction method because the distance to the target section is equal to maximum UPD. As shown in Fig. 12, when UPD increases, the R^2 decreases exponentially. In general, the minimum acceptable level of R^2 in the data mining field is 0.6(Alexander *et al.* 2015). According to the condition, the UPD for this project can be 28 rings using the following equation

$$R^2 = 0.79e^{-0.01(UPD-1)} \quad (14)$$

It should be noted that Eq. (14) is only valid under the condition of $R^2 \geq 0.6$. Such a time interval for 28 rings (28×1.2=34 m) can provide the manager and operator with enough time for decision-making. Here, the UPD may be changed depending on the R^2 value targeted by the TBM operator.

Evidently, a threshold UPD changes according to a tunnel project due to the different geological and operational conditions. Given that the relationship between UPD and R^2 is an exponential, the equation for arbitrary project can be established

$$R^2 = \alpha e^{-\beta(UPD-1)} \quad (15)$$

Where, α is the R^2 when UPD is 1 ring, β is the reduction coefficient, defined in this study. The estimation of α and β is approximated by the following process:

(1) Collect the data from the section where the tunneling is completed immediately involving geological and TBM operational information.
(2) Obtain the preprocessed data through a normalization and augmentation process (see Section 3.2).
(3) Divide the data into training and testing set without random sampling.
(4) Construct a GBRT model using the training set (see Fig. 7).
(5) Test the GBRT model with the testing set using sequential prediction method; subsequently, evaluate the R^2 according to UPD. It should be noted that the investigation range of UPD is only valid under the condition of $R^2 \geq 0.6$.
(6) α and β are approximated by a regression of Eq. (15).

Using the Eq. (15), TBM operators can find the UPD according to R^2 being targeted.

4.4 Significance and limitation of the study

Predicting PR in construction stage is a difficult problem owing to the gap between the expected and actual geological conditions, and operator's skills (Abate *et al.* 2019) Hence, the one-step prediction applied to design stage is not proper to the application during tunneling. Although monitoring the changes in PR is a crucial task for rescheduling the excavation plan during tunneling, the PR prediction in construction stage has not been extensively performed. Therefore, the sequential prediction method based on a GBRT algorithm was proposed for application to construction stage. For practical use, a simple equation between R^2 and UPD was introduced. Evidently, the UPD can be adjusted depending on other projects and the R^2 value targeted by an operator. Therefore, a calculation process for the equation between the R^2 and UPD has been addressed. The sequential prediction method can be used for predicting PR corresponding a unit prediction distance in real-time. Therefore, this study will be useful for rescheduling the excavation plan in real-time by adjusting expected PRs. It should be noted that this study is valid when the TBM excavation is stable because the authors did not consider downtime factors such as rock reinforcement, TBM jamming, and disc cutter replacement.

5. Conclusions

A sequential prediction method has been proposed to apply in the tunneling stage. This method was simulated to the Gunpo cable tunnel in Korea and evaluated by the comparison with the one-step prediction method. In modeling algorithms, gradient boosted regression trees had better performance than the least-square-based linear regression. This result suggests that the prediction method has more influence on the prediction performance than the modeling algorithm. For practical purposes, the PR prediction based on the unit prediction distance (UPD) was examined. Using the relation, a simple equation between R^2 and UPD was proposed involving the estimating process. The UPD can be adjusted depending on the R^2 value targeted by an operator. The sequential prediction method can be applied to other TBM projects in the tunneling stage if the geological and TBM operating conditions are obtained in real-time. Therefore, the method will be useful for rescheduling the excavation plan in real-time by adjusting expected PRs.

Acknowledgments

This research was supported by a grant (20SCIP-B105148-06) from the Construction Technology Research Program, funded by the Ministry of Land, Infrastructure, and Transport of the Korean government. This research was supported by a grant (21SCIP-B146946-04) from Smart Civil Infrastructure Research Program funded by Ministry of Land, Infrastructure and Transport of Korean Government.

References

Abate, G., Corsico, S., Grasso, S., Massimino, M.R. and Pulejo, A. (2020), *Analysis of The Vibrations Induced by a TBM to Refine Soil Profile during Tunneling: The Catania Case History*, *Tunnels and Underground Cities: Engineering and Innovation Meet Archaeology, Architecture and Art (1st Ed.)*, CRC Press, London, UK.

Alexander, D.L., Tropsha, A. and Winkler, D.A. (2015), "Beware of R^2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models", *J. Chem. Inform. Model.*, **55**(7), 1316-1322. https://doi.org/10.1021/acs.jcim.5b00206.

Bai, X.D., Cheng, W.C., Ong, D.E. and Li, G. (2021), "Evaluation of geological conditions and clogging of tunneling using machine learning", *Geomech. Eng.*, **25**, 59-73. https://doi.org/10.12989/gae.2021.25.1.059.

Benato, A. and Oreste, P. (2015), "Prediction of penetration per revolution in TBM tunneling as a function of intact rock and rock mass characteristics", *Int. J. Rock Mech. Min. Sci.*, **74**, 119-127. https://doi.org/10.1016/j.ijrmms.2014.12.007.

Bieniawski, Z.T. (1989), *Engineering rock mass classifications: a complete manual for engineers and geologists in mining, civil, and petroleum engineering*, John Wiley & Sons, Canada.

Bruland, A. (1999), "Hard Rock Tunnel Boring Advance Rate and Cutter Wear", Norwegian Institute of Technology (NTNU), Trondheim, Norway.

Chang, S.H., Choi, S.W., Bae, G.J. and Jeon, S. (2006), "Performance prediction of TBM disc cutting on granitic rock by the linear cutting test", *Tunn. Undergr. Sp. Tech.*, **21**, 271, https://doi.org/10.1016/j.tust.2005.12.131.

Cho, J.W., Jeon, S., Yu, S.H. and Chang, S.H. (2010), "Optimum spacing of TBM disc cutters: A numerical simulation using the three-dimensional dynamic fracturing method", *Tunn. Undergr. Sp. Tech.*, **25**, 230-244, https://doi.org/10.1016/j.tust.2009.11.007.

Deere, D.U. (1968). *Geological Consideration. Rock Mechanics in Engineering Practice*, New York, USA.

Friedman, J.H. (2001), "Greedy function approximation: a gradient boosting machine", *Annal. Stat.*, 1189-1232.

Gao, B., Wang, R., Lin, C., Guo, X., Liu, B. and Zhang, W. (2021), "TBM penetration rate prediction based on the long short-term memory neural network", *Undergr. Sp.*, **6**, 718-731. https://doi.org/10.1016/j.undsp.2020.01.003.

Gao, X., Shi, M., Song, X., Zhang, C. and Zhang, H. (2019), "Recurrent neural networks for real-time prediction of TBM operating parameters", *Automat. Constr.*, **98**, 225-235. https://doi.org/10.1016/j.autcon.2018.11.013.

Gong, Q.M., Zhao, J. and Jiao, Y.Y. (2005), "Numerical modeling of the effects of joint orientation on rock fragmentation by TBM cutters", *Tunn. Undergr. Sp. Tech.*, **20**(2), 183-191. https://doi.org/10.1016/j.tust.2004.08.006.

Gong, Q.M., Jiao, Y.Y. and Zhao, J. (2006), "Numerical modelling of the effects of joint spacing on rock fragmentation by TBM cutters", *Tunn. Undergr. Sp. Tech.*, **21**(1), 46-55. https://doi.org/10.1016/j.tust.2005.06.004.

Gong, Q. and Zhao, J. (2009), "Development of a rock mass characteristics model for TBM penetration rate prediction", *Int. J. Rock Mech. Min. Sci.*, **46**(1), 8-18. https://doi.org/10.1016/j.ijrmms.2008.03.003.

Hammerer, N. (2015), *Influence of steering actions by the machine operator on the interpretation of TBM performance data*. University of Innsbruck: Innsbruck, Austria.

Hassanpour, J., Rostami, J., Khamehchiyan, M., Bruland, A. and Tavakoli, H.R. (2010), "TBM performance analysis in pyroclastic rocks: a case history of Karaj water conveyance tunnel", *Rock Mech. Rock Eng.*, **43**(4), 427-445. https://doi.org/10.1007/s00603-009-0060-2.

Hassanpour, J., Rostami, J. and Zhao, J. (2011), "A new hard rock TBM performance prediction model for project planning", *Tunn. Undergr. Sp. Tech.*, **26**(5), 595-603. https://doi.org/10.1016/j.tust.2011.04.004.

Hassanpour, J., Vanani, A.G., Rostami, J. and Cheshomi, A. (2016), "Evaluation of common TBM performance prediction models based on field data from the second lot of Zagros water conveyance tunnel (ZWCT2)", *Tunn. Undergr. Sp. Tech.*, **52**, 147-156. https://doi.org/10.1016/j.tust.2015.12.006.

Huo, J., Sun, W., Chen, J. and Zhang, X. (2011), "Disc cutters plane layout design of the full-face rock tunnel boring machine (TBM) based on different layout patterns", *Comput. Ind. Eng.*, **61**, 1209-1225. https://doi.org/10.1016/j.cie.2011.07.011.

Koenker, R. and Hallock, K.F. (2001), "Quantile regression", *J. Economic Perspectives*, **15**, 143-156.

Kvålseth, T.O. (1985), "Cautionary note about R^2", *The American Statistician*, **39**(4), 279-285.

Lee, H.L., Song, K.I., Qi, C., Kim, J.S. and Kim, K.S. (2021), "Real-time prediction of operating parameter of TBM during tunneling", *Appl. Sci.*, **11**, 2967. https://doi.org/10.3390/app11072967.

Lu, X., Zhou, W., Ding, X., Shi, X., Luan, B. and Li, M. (2019), "Ensemble learning regression for estimating unconfined compressive strength of cemented paste backfill", *IEEE Access*, **7**, 72125-72133. https://doi.org/10.1109/ACCESS.2019.2918177.

Mazzoccola, D., Millar, D. and Hudson, J. (1997), "Information, uncertainty and decision making in site investigation for rock engineering", *Geotech. Geol. Eng.*, **15**, 145-180. https://doi.org/10.1023/A:1018499222495.

Ozdemir, L. and Wang, F.D. (1979), Mechanical tunnel boring prediction and machine design. Nasa Sti/Recon Technical Report N, **80**, 16239.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011), "Scikit-learn: Machine learning in Python", *J. Machine Learn. Res.*, **12**, 2825-2830.

Qi, C., Fourie, A., Chen, Q. and Zhang, Q. (2018a), "A strength prediction model using artificial intelligence for recycling waste tailings as cemented paste backfill", *J. Cleaner Production*, **183**, 566-578. https://doi.org/10.1016/j.jclepro.2018.02.154.

Rostami, J. (1997), "Development of a force estimation model for rock fragmentation with disc cutters through theoretical modeling and physical measurement of crushed zone pressure", Ph.D. Dissertation, Colorado School of Mines. Colorado

Sanio, H. (1985), "Prediction of the performance of disc cutters in anisotropic rock", *Int. J. Rock Mech. Min. Sci. Geomech.*, 153-161.

Sato, K. (1991), "Prediction of disc cutter performance using a circular rock cutting rig", *Proceedings of the first international symposium on mine mechanization*, Golden, Colorado.

Snowdon, R., Ryley, M. and Temporal, J. (1982), "A study of disc cutting in selected British rocks", *Int. J. Rock Mech. Min. Sci. Geomech.*, 107-121.

Yagiz, S. (2008), "Utilizing rock mass properties for predicting TBM performance in hard rock condition", *Tunn. Undergr. Sp. Tech.*, **23**, 326-339. https://doi.org/ 10.1016/j.tust.2007.04.011.

Yagiz, S. and Karahan, H. (2015), "Application of various optimization techniques and comparison of their performances for predicting TBM penetration rate in rock mass", *Int. J. Rock Mech. Min. Sci.*, **80**, 308-315. https://doi.org/10.1016/j.ijrmms.2015.09.019.

Zhang, P., Wu, H.N., Chen, R.P. and Chan, T.H. (2020), "Hybrid meta-heuristic and machine learning algorithms for tunneling-induced settlement prediction: A comparative study", *Tunn.*

Undergr. Sp. Tech., **99**, 103383, https://doi.org/10.1016/j.tust.2020.103383

IC