최상부분집합이 고려된 능형회귀를 적용한 현장관입지수에 대한 통계적 예측기법 개발 및 적용

이항로[1] · 송기일[2*] · 김경열[3]

[1]비회원, 인하대학교 토목공학과 통합과정
[2]정회원, 인하대학교 토목공학과 부교수
[3]정회원, 한국전력 전력연구원 차세대송변전연구소 책임연구원

Development and implementation of statistical prediction procedure for field penetration index using ridge regression with best subset selection

Hang-Lo Lee[1] · Ki-Il Song[2*] · Kyoung Yul Kim[3]

[1]Graduate Student, Dept. of Civil Engineering, Inha University
[2]Associate Professor, Dept. of Civil Engineering, Inha University
[3]Principal Researcher, Power Transmission Laboratory, KEPCO Research Institute

*Corresponding Author : Ki-Il Song, ksong@inha.ac.kr

Abstract

The use of shield TBM is gradually increasing due to the urbanization of social infrastructures. Reliable estimation of advance rate is very important for accurate construction period and cost. For this purpose, it is required to develop the prediction model of advance rate that can consider the ground properties reasonably. Based on the database collected from field, statistical prediction procedure for field penetration index (FPI) was modularized in this study to calculate penetration rate of shield TBM. As output parameter, FPI was selected and various systems were included in this module such as, procedure of eliminating abnormal dataset, preprocessing of dataset and ridge regression with best subset selection. And it was finally validated by using field dataset.

Keywords: Field penetration index, Statistical prediction procedure, Prediction model, Best subset selection, Ridge regression

초 록

사회기반시설의 지중화로 인하여 쉴드 TBM 적용이 점차 확대되고 있는 추세다. 합리적인 공기기간 및 공사비 산정을 위해 쉴드 TBM의 실굴진율을 정확하게 예측하는 것은

OPEN ACCESS

Journal of Korean Tunnelling and Underground Space Association
19(6)857-870(2017)
https://doi.org/10.9711/KTAJ.2017.19.6.857

eISSN: 2287-4747
pISSN: 2233-8292

Received September 11, 2017
Revised September 25, 2017
Accepted October 10, 2017

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright ⓒ 2017, Korean Tunnelling and Underground Space Association

매우 중요한 사안이라 할 수 있다. 이러한 이유로 국내에서는 지반의 물성을 합리적으로 반영한 쉴드 TBM의 실굴진율 예측모델이 필요한 상황이다. 본 연구는 쉴드 TBM의 순굴진율 산정을 위해 현장 데이터베이스를 기반으로 현장관입지수의 통계적 예측절차를 모듈화 하였다. 출력인자로 현장관입지수를 선정하였고, 비정상치 제거 및 전처리 그리고 최상부분집합선택이 고려된 능형회귀를 적용한 예측시스템을 모듈에 포함하였다. 또한 현장 굴진 데이터를 활용하여 예측모델의 적용성을 확인하였다.

주요어: 현장관입지수, 통계적 예측절차, 예측 모델, 최상부분집합선택, 능형회귀

1. 서 론

가스, 통신 및 전력 등 사회기반시설의 지중화로 인하여 쉴드 TBM (Tunnel boring machine)적용이 점차 확대되고 있다. 이러한 흐름에 대응하여 현장의 합리적인 공기 및 공사비를 산정하기 위해서는 쉴드 TBM의 실굴진율(Advance rate)을 정확하게 예측하는 것이 매우 중요하지만 국내에서는 지반의 물성을 합리적으로 반영한 쉴드 TBM의 실굴진율 예측모델이 필요한 상황이다.

쉴드 TBM의 굴진성능 예측을 위해 크게 기계인자, 지반인자 그리고 다운타임으로 분류할 수 있으며 이와 관련된 세부인자들이 존재한다. 그러나 이와 같이 다양한 인자들을 모두 고려하는 것은 현실적으로 어렵기 때문에 Table 1과 같이 다양한 경험 및 이론식을 적용하여 굴진성능을 예측하고 있다. Farmer and Glossop (1980)과 Hughes (1986)은 커터 당 추력과 암반의 강도특성을 이용하여 굴진성능을 예측하였으나 이들은 지반의 절리특성을 고려하지 못한다는 한계가 있으며 Bruland (1998)가 제시한 NTNU모델은 다양한 인자를 고려하였다는 점에서 합리적인 예측이 가능하지만 특수 실내실험으로 인한 입력인자를 얻기가 쉽지 않고, Open TBM에 특화되어 있기 때문에 국내의 쉴드 TBM 현장에 적용하기에는 다소 어려움이 있다. 최근에는 Hassanpour et al. (2009)와 Delisio et al. (2013) 등과 같이 비교적 획득이 용이한 지반 및 기계인자를 사용하여 예측모델을 개발하였지만 역시 국내의 지반특성과는 다소 거리가 있으며 터널굴진면의 매핑 등 국내 현장에서 현실적으로 획득하기 어려운 입력인자가 포함될 수 있다. 국내에서는 KICT 모델이 개발되어 적용되고 있으며 NTNU 모델과 같이 입력인자를 획득하기 위해서 특수 실내실험이 요구된다(Chang et al., 2007). 위의 배경으로 비추어 볼 때 국내 현장에서 획득이 용이한 인자와 국내에 맞는 데이터베이스를 고려한 실굴진율 예측모델이 필요하다고 할 수 있다.

예측모델을 구축하기 위한 방법 중 최소제곱법을 적용한 선형회귀분석이 있다. 그러나 이 기법은 입력인자 간 다중공선성(Multicollinearity)으로 인해 예측력이 떨어질 수 있으며, 단계적선택 및 최상부분집합선택 등의 변수선택기법을 통하여 입력인자 간 연관이 있는 변수를 제거하여 예측력을 보완할 수 있다. 그러나 선형모델에 대한 최상의 부분집합을 구축했을지라도 입력인자 간에 완전히 독립적이지 않을 수 있기 때문에 이에 대한 추가적인 보완이 필요하다고 할 수 있다. 또한, 현장 데이터베이스의 비정상치를 제거하는 작업에서부터 전처리, 데이터 분할 및 예측모델 구축 및 검증까지 단계적으로 나누어서 분석하는 것은 번거로울 수 있으며 데이터의 성격이 달라지면 다시 분석해야하는 어려움이 있을 수 있다. 본 연구에서는 위의 필요성에 근거하여 최근 시공이 완료된 현장

의 데이터베이스를 기반으로 현장관입지수의 통계적 예측절차를 모듈화 하고자 하였다. 특히, 변수선택기법의 예측력을 보다 향상시키기 위해서 최상부분집합을 고려한 능형회귀(Ridge regression)를 제안하였다. 또한 구축된 예측모델을 현장 굴진 데이터를 활용하여 적용성을 확인하였다.

Table 1. Summary of TBM performance prediction models

Model	Predicted value	Machine parameters	Rock mass parameters
Farmer and Glossop (1980)	Penetration rate (m/h)	Cutter force	Tensile strength
Hughes (1986)	Penetration rate (m/h)	Cutter force	Uniaxial compressive strength
NTNU (Bruland, 1998)	Penetration rate (m/h), Advance rate (m/h)	Cutter force, RPM, Cutter spacing, Cutter size and shape, Cutterhead power	Uniaxial compressive strength, DRI, Number of joint sets, Joint frequency, Joint orientation, Porosity
Chae et al. (2005)	Penetration rate (m/h)	Number of cutting per revolution, Penetration per revolution, RPM	-
Gong and Zhao (2009)	Bore-ability index (kN/cutter/mm/rev)	Cutter force, RPM	Uniaxial compressive strength, Volumetric joint count, Brittleness index, Joint orientation
Hassanpour et al. (2010)	Field penetration index (kN/cutter/mm/rev)	Cutter force, RPM	Uniaxial compressive strength, Joint spacing
Satar Mahdevari et al. (2014)	Penetration rate (m/h)	Specific energy, Thrust force, Cutterhead power and torque	Uniaxial compressive strength, Brazilian tensile strength, Brittleness index, Joint spacing and orientation

2. 출력인자 선정

암반의 Bore-ability는 TBM과 암반과의 상호작용의 결과를 종합적으로 나타내는 지표라고 할 수 있다. 기존의 연구자들은 Bore-ability를 정량적으로 표현하기 위해서 각기 다양한 지표를 제시해왔으며 그 중 Hamilton and Dollinger (1979)는 Bore-ability를 정량적으로 표현하기 위해 다음과 같이 현장관입지수를 제안하였다.

$$FPI = \frac{F_n}{P_e} \tag{1}$$

여기서, $F_n(kN/cutter)$은 커터 작용력, $P_e(mm/rev)$는 커터헤드가 한번 회전하는데 관입된 깊이, 즉 관입깊이이며, $FPI(kN/cutter/mm/rev)$는 현장관입지수 즉, 굴진의 난해성을 의미하는 지표이다. 예를 들어, 현장관입지수가 증가하면 단위깊이 당 상대적으로 높은 커터 작용력이 요구되기 때문에 장비의 굴진이 어렵게 되며

반대로 현장관입지수가 감소하게 되면 상대적으로 낮은 커터 작용력으로 동일한 단위깊이를 관입할 수 있기 때문에 굴진이 용이해진다고 표현할 수 있다.

최근의 연구된 Gong et al. (2007)에 의하면, 현장관입지수는 커터 작용력이 증가함에 따라 치핑(Chipping)효율이 높아져 관입깊이가 비선형적으로 증가하는 경향을 실험을 통해 규명하였으며 이는 앞서 제시하였던 식 (1)과 달리 현장관입지수가 커터 작용력에 따라 값이 다를 수 있음을 의미한다. 그러나 Hamidi et al. (2010)은 관입깊이가 1 mm/rev 이상인 경우에는 커터 작용력과 관입깊이의 관계는 선형으로 근사할 수 있다고 하였으며, 이는 기울기를 의미하는 현장관입지수가 일정한 값으로 수렴한다고 볼 수 있다. 기계인자인 커터의 너비, 직경 및 간격 등이 고려되지 못한 경우라도 커터헤드의 직경이 비슷한 쉴드 TBM이 적용된다면 지반인자를 통해 현장관입지수의 예측이 가능하다고 알려져 있다(Hamidi et al., 2010). 그러나 커터헤드의 토크 또한 커터 작용력 및 RPM 등과 같이 쉴드 TBM의 굴진율에 주요한 영향을 미치는 요소이기 때문에 토크의 영향을 고려한 연구가 추가적으로 필요하다고 할 수 있다.

현장관입지수는 쉴드 TBM의 커터 작용력과 RPM이 설계단계에 결정이 되면 예측된 현장관입지수를 사용하여 다음의 식 (2)와 같이 굴진율 산정도 가능할 수 있다.

$$PR = \frac{F_n \bullet RPM}{FPI} \tag{2}$$

Table 2. Results of statistical analysis and R squared (R^2) suggested by Hassanpour et al. (2009)

Characteristics of rockmass	Parameters	R squared (R^2)			Regression type
		PR (m/h)	P_e (mm/rev)	FPI $(kN/cutter/mm/rev)$	
Uniaxial compressive strength		0.634	0.445	0.697	Log
Joint frequency	$Spacing$ (m)	0.533	0.342	0.639	Linear
	RQD $(\%)$	0.431	0.277	0.636	Linear
	K_s tot	0.381	0.280	0.344	Log
Rock mass classification systems	$Basic\ RMR$	0.377	0.172	0.571	Linear
	RMR 89	0.550	0.311	0.688	Linear
	GSI	0.506	0.300	0.669	Linear
	Q	0.394	0.263	0.624	Linear
Rock mass strength parameters	Q_c	0.563	0.417	0.718	Linear
	σ_{cm} (MPa)	0.550	0.339	0.718	Linear
	UCS_{rm} (MPa)	0.559	0.423	0.709	Log
	$RMCI$	0.621	0.447	0.748	Linear
	Mean	0.508	0.335	0.647	

여기서, PR은 굴진율을 의미하며 시간 당 거리로 표현할 수 있다.

Hassanpour et al. (2009)은 통계분석을 통해 굴진성능을 나타내는 출력인자와 다양한 지반인자와의 관계를 분석하였다(Table 2). 세 가지 출력인자들 중에서 현장관입지수의 결정계수(R^2)는 0.647로 가장 높은 값을 나타내었으며 이는 현장관입지수가 다양한 지반인자를 잘 표현하는 변수임을 의미한다고 판단할 수 있다. 본 연구에서는 통계학적 결과와 활용성을 고려하여 현장관입지수를 예측모델의 출력인자로 선정하기로 결정하였다.

3. 예측기법 및 모델선정 기준

선형회귀는 기본적으로 최소제곱법을 근거하나 더 나은 예측력과 모델의 해석력을 위해 추가적인 분석기법이 존재한다. 입력인자의 부분집합(Subset)을 선택하는 방법, 입력인자들의 추정계수를 0으로 수축시키는 방법 그리고 기존 입력인자들의 조합으로 주성분 인자를 새로 구성하여 분석하는 방법 등이 있다. 본 연구에서는 최상부분집합선택을 고려한 능형회귀(Ridge regression)방법을 적용하고자 하였으며, 먼저 최적의 예측모델을 선정하기 위한 기준이 필요하므로 이와 관련된 지표를 조사하여 정리하였다.

3.1 예측모델의 선정 지표

선형회귀모델에 사용되는 입력인자를 선정하는 과정은 모델의 예측력을 높이는데 중요한 단계라 할 수 있다. 입력 또는 출력인자가 서로 관련이 거의 없는 경우 불필요하게 복잡한 모델이 되어 관여한 훈련데이터에 대해서는 잘 예측할 수 있지만 테스트데이터에 대한 오차는 반대로 커질 수 있기 때문이며 일반적으로 이를 과대적합이라 불린다. 또한, 입력인자간의 상관관계가 높은 경우 다중공선성으로 인하여 계수추정치의 분산을 증가시키기 때문에 오히려 예측력이 떨어지는 문제가 발생할 수 있다. 그렇기 때문에 지반인자의 최상부분집합을 선정하여 예측오차를 최대한 줄일 수 있는 방안이 필요하다고 할 수 있다.

예측모델을 평가하는 기준은 Table 3과 같이 다양한 방법이 존재한다(James et al., 2014). 훈련 잔차제곱합(Train RSS)과 결정계수(R^2)는 입력인자의 수가 동일한 모델 간의 평가가 가능한 특징이 있으나, 서로 다른 경우에는 입력인자의 수가 많은 모델을 선정하기 때문에 부분집합을 선정하는 기준으로는 부적합하다고 할 수 있다. 입력인자의 수가 증가함에 따라 페널티를 부여해 모델을 간접적으로 평가하는 방식으로 Akaike (1974)가 제안한 Akaike information criterion (AIC), 수정된 결정계수(R^2_{adj}) 그리고 Mallow (1973)이 제안한 멜로우즈 Cp 등이 있으며, 이들은 서로 다른 개수의 입력인자를 가진 모델 간의 비교가 가능하다는 특징이 있다(James et al., 2014). 이와 반대로, 데이터 마이닝 분야에서 주로 쓰이며 모델의 예측오차를 직접적으로 평가하는 지표인 교차검증을 적용한 방법이 있다. 여기서 교차검증이란 데이터셋을 동일한 크기의 k개의 그룹으로 나눈 후 각 하나의 그룹을 나머지 그룹들로 적합한 모델을 검증하는 방식을 말한다, 이와 같이 각 단계마다 얻어진 오차의 제곱을 평균한 값을 교차검증 평균제곱오차(CV MSE)라 하며, 이를 척도로 예측력을 직접적으로 평가할 수 있다. 본 연구

에서는 p개의 지반인자의 가능한 모든 부분집합, $\sum_{i=1}^{p}{}_{p}C_{i} = 2^{p} - 1$ 개 중에서 가장 작은 CV MSE를 갖는 부분집합을 최상부분집합으로 선택하기로 결정하였다.

Table 3. Summary of statistical indicators

Applicable condition	Evaluation method	Indicators
The same number of input parameters	Indirect	R^2
	Direct	$Train\ RSS$
The different number of input parameters	Indirect	C_p, AIC, R^2_{adj}
	Direct	$CV\ MSE$

3.2 계수추정치 수축 기법

입력인자의 부분집합을 선택하는 대신에 p개의 입력인자를 모두 사용하여 예측력을 개선하는 방법으로 능형회귀(Hoerl and Kennard, 1970)가 있다. 이 기법은 일종의 규칙화(Regularization)라는 방법을 사용하여 예측모델의 계수추정치를 0으로 수렴하게하며 이는 모든 입력인자를 사용하면서 다중공선성 및 모델의 과대적합으로 인한 계수추정치의 분산이 커지는 문제를 상당히 줄일 수 있는 것으로 알려졌다(James et al., 2014). 즉, 데이터베이스가 조금만 변경되더라도 계수추정치가 민감하게 바뀌는 현상을 방지하여 높은 예측력을 유지하는 강건(Robust)한 모델을 구축할 수 있다.

능형회귀는 잔차제곱합(RSS)을 최소화하는 최소제곱법에 Lambda(λ)라는 조절파라미터를 추가하여 다음의 식 (3)을 만족하는 조건으로 능형회귀의 계수추정치 $\hat{\beta}^{ridge}$ 을 산정할 수 있다.

$$\hat{\beta}^{ridge} = Argmin_{(\beta)}\left\{\sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} = Argmin_{(\beta)}\left(RSS + \lambda\sum_{j=1}^{p}\beta_j^2\right) \tag{3}$$

여기서, y_i, $\hat{y_i}$는 각각 출력인자의 관측 값, 예측 값을 뜻하며, n과 p는 각각 데이터셋의 수, 입력인자의 수를 의미한다. 최적의 λ을 추정하는 방법은 지반인자의 특정 부분집합에 대하여 CV MSE가 최소가 되는 값으로 찾을 수 있으며 이는 4.3절에서 자세히 정리하였다.

4. 비정상치 제거 및 예측모델 구축

4.1 비정상치 제거

현장에서 획득한 데이터셋의 일부분은 비정상치가 포함될 수 있다. 이는 단순히 입력치를 잘못 입력한 경우 또

는 컴퓨팅에러를 포함하는 외부적인 요인과 회귀모형에 적합하지 않는 데이터를 모두 의미할 수 있다. 단순히 외부적인 요인으로 발생한 데이터는 제거되는 것이 타당하나 회귀모형에 적합하지 않는 데이터의 경우는 다르다고 할 수 있다. Myers (1990)에 따르면, 연구자가 특정 데이터셋이 특이점(Outlier)임을 확신하더라도 이를 제거할 필요는 없다고 하였으며 이 말인즉슨 모델에 적합하지 않은 데이터셋을 무조건 제거할 필요가 없음을 의미할 수 있다. 회귀모델에 적합하지 않더라고 계수추정치의 변동에 큰 영향을 주지 않을 수 있기 때문이다.

인자가 한 개인 단변량 데이터나 두 개의 인자를 가지는 이변량 데이터는 비교적 비정상치를 진단하는 것은 어렵지 않으나, 인자가 세 개 이상인 복잡한 다변량 데이터의 경우에는 정형화된 방법은 없으며, 데이터 특성에 따라 진단하는 것이 현명하다고 할 수 있다(Rahmatullah, 2005; Cousineau and Chartier, 2010). 본 연구에서는 현장에서 획득한 데이터셋의 불필요한 손실을 줄이고자 Cousineau and Chartier (2010)를 참고하여 특이점 중에서 계수추정치에 큰 영향을 주는 영향점(Influential point)을 비정상치로 간주하기로 결정하였다. 여기서 특이점이란 주어진 2차원 회귀모델의 경우, y축 방향을 향하여 비정상적으로 떨어져 있는 점을 의미하며, 계수추정치를 크게 변하게 만드는 데이터셋을 영향점이라 정의한다. 본 연구에서 선택한 방법 이외에도 마할로노비스 거리를 개선한 로버스트 거리, K-평균 클러스터링 등을 활용하여 비정상치를 진단할 수 있는 것으로 알려져 있다(James et al., 2014).

특이점 및 영향점을 진단하는 지표는 여러 종류가 있으나, 본 연구에서는 Table 4에 정리된 외표준화잔차(Externally studentized residuals)와 Welsch (1980)가 제시한 DFFITS를 선정하였다. i 번째의 데이터를 제외한 분산추정치를 사용한 표준화잔차를 외표준화잔차라 하며, DFFITS는 외표준화잔차와 큰 지렛점을 종합적으로 고려한 영향점을 진단하는 지표이다. 큰 지렛점(High leverage point)은 특이점과 반대로 회귀모형의 x축 방향을 향하여 비정상적으로 떨어진 점을 의미한다. Table 4에 정리한 각 지표의 관련 식은 Myers (1990)를 통해 확인할 수 있다.

Table 4. Summary of three indices for diagnosing abnormal data in this study

	Applied indices	Symbol	Thresholds
Outlier	Externally studentized residual	r_i^*	$\left\|r_i^*\right\| > 2.0$
High leverage point	Hat matrix value	h_{ii}	$h_{ii} > \frac{2p}{n}$
Influential point	DFFITS	$DFFITS_i$	$\left\|DFFIT_i\right\| > 2\sqrt{\frac{p}{n}}$

4.2 데이터 전처리

능형회귀에서 최적의 λ을 얻기 위해서 먼저 입력인자의 정규화가 선행되어야 한다. 입력인자로 사용되는 지반인자의 단위는 각기 다르기 때문에 출력인자와의 영향이 실제로 미미한 입력인자일 지라도 상대적으로 스케일

이 크면 모형을 적합할 때 큰 영향을 끼칠 수 있다. 본 연구에서는 다음의 식을 이용하여 입력인자를 정규화 하였다.

$$\bar{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

식 (4)에서 x은 정규화 전의 입력인자이며 x가 $x_{\min}$에 가까울수록 $\bar{x}$은 0으로 수렴하게 되고 반대로 $x_{\max}$에 다가갈수록 $\bar{x}$은 1에 수렴하는 양상을 갖는다. 즉, 모든 입력인자를 0과 1사이의 값으로 정규화 하여 스케일에 따른 영향을 제거하였다.

4.3 현장관입지수에 대한 통계적 예측기법 모듈화

예측모델을 적합하고 예측력을 확인하기까지의 모든 과정을 구현하려면 먼저 데이터셋을 훈련데이터, 검증(Validation)데이터 그리고 테스트데이터로 나누어야 한다. 훈련데이터는 모델을 적합하는데 사용되고 λ와 같이 조절파라미터를 최적화하기 위해서는 검증데이터를 이용하게 된다. 마지막으로 적합된 예측모델에 관여하지 않은 테스트데이터를 사용하여 그 성능을 검증하게 된다. 여기서 중요한 점은 최적의 λ을 결정하는데 이미 사용된 훈련데이터를 사용하면 안 된다는 것이다. 만약에 검증데이터가 아닌 훈련데이터가 λ을 선택하는 기준으로 사용되면 예측모델은 훈련데이터에서만 과대적합이 되고 새로운 데이터에 대해서는 오히려 예측력이 떨어지는 문제가 발생할 수 있기 때문이다. 그러나 데이터셋의 양이 많지 않은 경우라면 데이터셋을 세 그룹으로 나누는 것은 부담이 될 수 있다. 최적의 λ을 선정하는데 $1/3$의 데이터셋을 사용하기 때문에 모델 개발에 사용되는 훈련데이터와 테스트에 사용되는 테스트데이터의 손실이 크기 때문이다. 본 연구에서는 데이터셋의 손실을 줄이고자 CV MSE을 적용하였다. 교차검증을 적용하면 검증데이터를 따로 생성할 필요가 없기 때문에 훈련데이터만으로 모형 적합과 최적의 λ까지 선택할 수 있게 된다.

교차검증에서 나뉘는 그룹의 개수 즉, k에 따라서 다양한 교차검증 방식이 존재하며 이를 k겹 교차검증이라고 한다. 본 연구에서는 현장에서 획득한 데이터셋이 많지 않은 관계로 k를 데이터셋의 개수로 설정한 Leave one out cross validation (Loocv)을 적용하였다.

Fig. 1은 파이썬 프로그래밍 언어(Sanner, 1999)를 사용한 현장관입지수의 통계적 예측절차를 알고리즘으로 나타낸 것이다. 먼저 현장에서 획득한 데이터베이스를 입력하면 통계적 절차에 따라 비정상치 여부를 진단하게 된다. 그 다음으로 전처리(Preprocessing)를 통한 정규화 데이터를 이용하여 임의로 훈련(80%) 및 테스트(20%) 데이터로 나눈다. 훈련 데이터는 최상부분집합선택을 고려한 능형회귀분석에 사용되며 데이터 성격에 따라 사용자가 지정한 다수의 $\lambda \geq 0$이 투입된다. 여러 λ에 대하여 CV MSE가 최소가 되는 모델을 일차적으로 선정하게 되며, 이를 입력인자에 대한 각 부분집합에 대하여 반복 실시한다. 각 부분집합에 대한 CV MSE가 최소가 되는 모델 중에서 가장 작은 CV MSE이 되는 모델이 최종 선정된다. 마지막으로 테스트 데이터를 적용하여 예측성능을 평가한 후 종료된다.

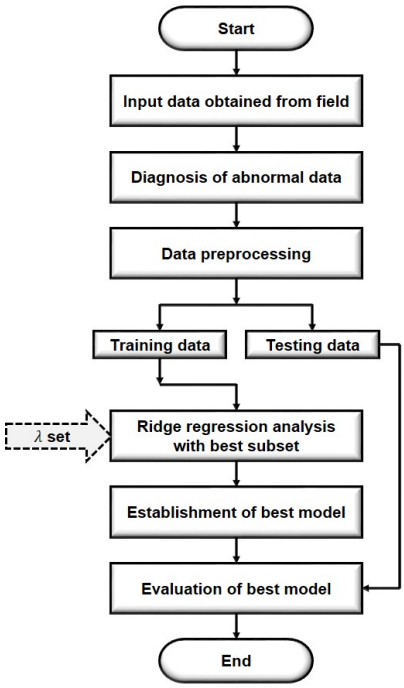

Fig. 1. Statistical procedure for the prediction of field penetration index

5. 현장 굴진데이터 적용

모듈화 한 통계적 예측기법을 국내 쉴드 TBM 현장에서 획득한 굴진 데이터에 적용하였고, 구축된 모델의 예측성능을 검증하고자 하였다. 본 현장은 변전소 간의 전기공급을 위한 전력구 공사현장으로 총 연장은 5,094 m에 달하며 그 중 쉴드 TBM이 적용된 구간은 2,258 m이다. 설계단계에서 시추공 샘플을 이용하여 실내실험을 실시한 결과로 27개의 데이터셋(비정상치 포함)을 확보하였으며 일축압축강도, 코어 회수율, RQD, RMR, 루전값, 흡수율, 탄성파 속도, 변형계수가 포함되었다. 출력인자는 문헌조사를 근거하여 현장관입지수를 선택하였으며 주어진 식 (2)에 대입하여 현장관입지수를 산정하였다.

현장의 굴진데이터의 기술통계는 Table 5에 정리하였고, 적용된 TBM의 사양은 Table 6에 요약하였다. 암반강도의 분포는 대략 20~95 MPa 정도이며, RMR의 경우에는 54~84% 범위로 암반등급이 양호하고 우수한 축에 속한다고 볼 수 있다(Bieniawski, 1973). 현장에 적용된 장비는 EPB타입 쉴드 TBM이며 추력과 RPM은 최대 9,600 kN, 9 rev/min까지 운전이 가능한 것으로 조사되었고 실제 굴착은 평균 2,718 kN의 추력과 7.17 rev/min의 RPM을 적용한 것으로 나타났다.

현장에서 획득한 데이터에는 기계적 오류 및 모델에 적합하지 않는 비정상치가 존재할 수 있으므로 비정상치 진단을 수행하였으며, Fig. 2와 같이 결과를 그래프로 나타내었다. 그래프의 가로 및 세로축은 Table 4의 지렛점과 외표준화잔차를 나타내며, 원의 크기는 DFFITS의 절댓값의 크기를 의미한다. 분석 결과, 2번 및 26번 데이터

가 특이점과 영향점을 동시에 만족하였으며 이는 의도적인 굴진속도 감소에 기인하는 것으로 판단되었다. 10번 데이터는 큰 지렛점과 영향점에는 속하지만 특이점에는 해당되지 않기 때문에 분석에 그대로 적용되었다.

Table 5. Summary of dataset obtained from field

Parameters	N	Minimum	Maximum	Mean	Std.deviation
UCS (MPa)	25	20.15	95.26	51.81	21.06
TCR (%)	25	94	100	99.56	1.53
RQD (%)	25	47	100	77.16	18.28
RMR (%)	25	54	84	68.6	8.78
Lv	25	0.01	15.89	1.95	3.37
Ar (%)	25	0.12	2.91	0.54	0.66
Ev (m/sec)	25	3,748	5,252	4,629	456
Dm (MPa)	25	14,821	77,036	41,445	17,852
Thrust (kN)	25	1,532	3,454	2,718	528
RPM (rev/min)	25	7.07	7.26	7.17	0.04
FPI (kN/cutter/mm/rev)	25	12.47	68.74	33.38	15.49

Lv = Lugeon value; Ar = Absorption ratio; Ev = Elastic wave velocity; Dm = Deformation modulus

Table 6. Specification of shield TBM applied to field

Specification of TBM	
Type	EPB shield TBM (upgrade)
Machine diameter (mm)	3,330
Cutter diameter (mm)	350 mm
Number of disc cutter	26
Total thrust (kN)	9,600
RPM (rev/min)	9

한편, 최상부분집합을 고려한 능형회귀분석을 실시하기 위해서 λ을 약 0~3 범위에서 30개를 로그스케일로 선정하였으며, 비정상치가 제거된 25개의 데이터를 토대로 통계절차에 따른 분석을 실시하였다. 그 결과 RMR, 탄성파 속도, 코어회수율, 루전값이 지반인자의 최상의 부분집합으로 분석되었으며 이들의 조합으로 λ에 따른 CV MSE의 변화추이를 Fig. 3에 나타내었다. λ가 0에서 서서히 증가할수록 CV MSE가 점점 감소하다가 0.204일 때 최소치를 나타내었고, 이를 기준으로 다시 증가하는 경향을 보였다. CV MSE가 감소하는 구간에 대해서는 다양한 원인이 있을 수 있으나, 최상의 부분집합이더라도 서로 간의 완전한 독립적인 인자가 아닐 수 있으므로, λ을 서서히 가함으로써 입력변수간의 관계를 감소시키는 효과가 있는 것으로 분석되었다.

Fig. 4는 λ에 따른 최상부분집합의 계수추정치 변화를 나타낸 그래프이며, λ가 0에서 증가하다가 CV MSE가 최소가 되는 지점($\lambda = 0.2$)에서 예측성능이 가장 높은 예측모델을 도출하였다.

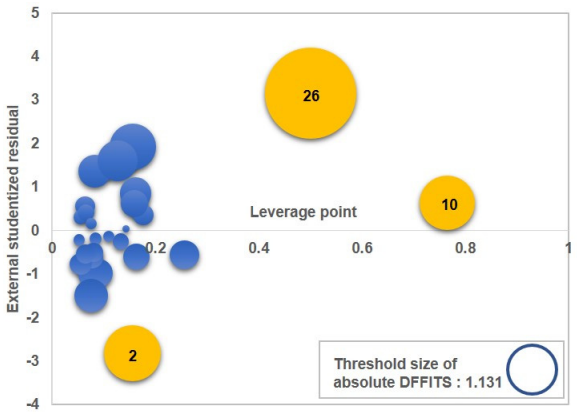

Fig. 2. Diagnosis for abnormal dataset

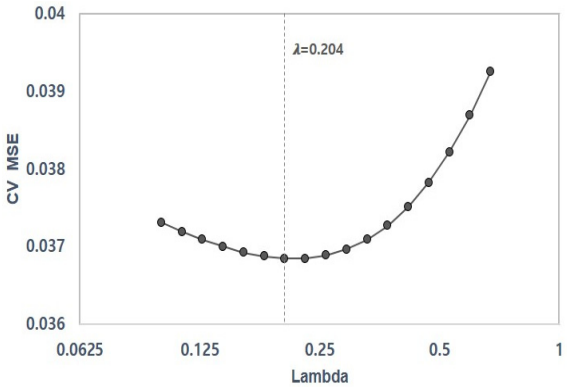

Fig. 3. The relationship between λ and CV MSE for best subset (Case 1)

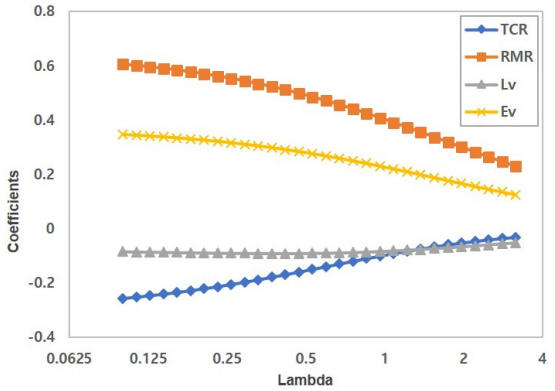

Fig. 4. The tendency of coefficient estimate depending on λ for best subset (Case 1)

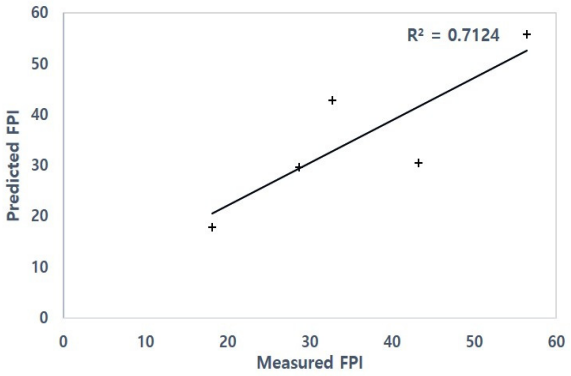

Fig. 5. Comparison between measured FPI and predicted FPI using ridge regression with best subset for Case 1

이와 같이 선정된 예측모델을 테스트 데이터에 적용하고자 하였다. 본 연구에서는 예측오차를 나타낼 수 있는 지표로 절대평균백분율오차를 선정하였으며 식 (5)와 같이 표현할 수 있다.

$$MAPE = \frac{|y_i - \hat{y}_i|}{y_i} \cdot 100 \tag{5}$$

여기서, MAPE는 절대평균백분율오차를 의미하며, y_i, $\hat{y}_i$는 각각 출력인자의 실제 값과 예측 값을 나타낸다. 위의 지표로 예측모델을 테스트한 결과, 약 13.309%의 MAPE를 나타내었으며 Fig. 5와 같이 예측 FPI와 실제 FPI 사이의 결정계수는 0.712로 분석되었다.

그러나 데이터를 훈련 및 테스트 데이터로 분할을 할 때마다 데이터의 표본이 매번 바뀌기 때문에 예측성능이 다르게 나타날 수 있다. 그렇기 때문에 네 번의 Case (2~5)를 추가적으로 반복하여 분석하였으며 이에 대한 결과를 Table 7에 정리하였다. 총 다섯 번의 통계치를 평균한 결과, 평균 MAPE는 17.311%, 표준편차는 7.94임을 확인할 수 있었으며, Case 2의 최상부분집합은 나머지 Case와 상이한 결과를 나타냄을 확인할 수 있었다. 이는 다양한 원인이 있을 수 있으나 임의적으로 데이터를 분할할 때마다 지반인자의 평균 및 표준편차 등 표본의 통계치가 동일하지 않아 서로 다른 부분집합을 보인 것으로 판단되었다.

Table 7. Results of model testing with arbitrary partitioning of dataset for Case (1~5)

Case	MAPE (%)			Best subset	CV MSE	λ
	Least squares	Best subset selection	Best subset + Ridge			
1	24.7	18.95	13.309	TCR, RMR, Lv, Ev	0.037	0.204
2	35.645	23.943	23.737	TCR, RMR, Ev	0.029	0.06
3	35.069	27.826	27.232	TCR, RMR, Lv, Ev	0.036	0.168
4	14.279	14.659	14.32	TCR, RMR, Lv, Ev	0.040	0.206
5	9.263	7.73	7.957	TCR, RMR, Lv, Ev	0.040	0.168
Mean	23.791	18.622	17.311			
Std.dev	11.938	7.866	7.94			

또한, 본 연구에서는 최상부분집합을 고려한 능형회귀모델을 비교하기 위해 최소제곱 그리고 최상부분집합선택만을 적용한 결과를 함께 나타내었다. 모든 입력인자로 최소제곱법을 적용했을 때의 평균 MAPE는 23.971%의 오차를 나타내었고 최상부분집합선택 그리고 최상부분집합을 고려한 능형회귀모델의순서로 평균 MAPE가 작아지는 것을 확인할 수 있었다. 즉, 최상부분집합을 고려한 능형회귀가 최소제곱법보다 6.48% 더 작았고, 최상부분집합선택보다는 1.311% 더 낮은 평균 MAPE를 보임을 확인할 수 있었다.

7. 결 론

본 연구는 쉴드 TBM의 순굴진율 산정을 위해 현장 데이터베이스를 기반으로 현장관입지수의 통계적 예측절차를 단독적으로 실행할 수 있도록 모듈화 하였다. 문헌조사를 통하여 현장관입지수를 출력인자로 선정하였고, 비정상치 제거 및 전처리 그리고 최상부분집합이 고려된 능형회귀 등을 모듈에 포함하였다. 또한, 테스트 데이터를 사용하여 구축된 모델에 대한 적용 및 검증을 수행하였다. 분석결과를 다음과 같이 요약할 수 있다.

1. 국내 전력구 데이터의 비정상치 진단여부를 검사한 결과, 총 27개 중 두 개의 데이터(2번 및 26번)에서 특이점과 영향점을 동시에 만족하였으며 이는 장비의 의도적인 굴진속도 감소에 기인하는 것으로 판단되었다.

2. 국내 전력구 데이터를 적용하여 최적의 모델을 구축한 결과 RMR, 탄성파 속도, 코어회수율, 루전값이 지반인자의 최상의 부분집합으로 분석되었으며, CV MSE가 최소가 되는 지점($\lambda = 0.204$)에서 예측성능이 가장 높은 최적의 예측모델을 도출하였다. 그러나 현장 데이터베이스의 양이 다소 작기 때문에 지반인자에 대한 최상의 부분집합이 굴진성능의 주요인자로 의의를 두기는 어려우며, 지속적인 데이터베이스 구축을 통하여 현장관입지수와 지반인자간의 상관성을 일반화 할 필요가 있다.

본 연구에서 제시한 모델을 테스트 데이터에 적용한 결과 최소제곱법보다 평균 MAPE가 6.48% 더 작았고, 최상부분집합선택보다는 1.311% 더 낮은 평균 MAPE를 나타내었다. 이를 통해 본 연구에서 개발한 현장관입지수에 대한 통계적 예측기법이 잘 작동함을 확인할 수 있었다.

본 모델은 사용된 데이터베이스에 성격에 따라 적용 범위가 다르기 때문에 다양한 지반특성 및 장비특성에 따른 데이터베이스를 구축할 필요가 있다.

또한 본 연구에 사용된 데이터베이스는 다운타임이 반영되지 않은 현장관입지수를 사용하였기 때문에, 추 후 다운타임이 반영된 지반조건 별 평균 현장관입지수를 도입하여 쉴드 TBM의 실굴진율을 예측할 수 있는 일반화 모델이 필요할 것으로 판단된다.

감사의 글

본 연구는 국토교통부(국토교통과학기술진흥원) 건설기술연구사업인 “도심지 소단면(Φ 3.5 m급) 터널식 공동구 설계 및 시공 핵심기술 개발(15SCIP-B105148-01)” 연구단을 통해 수행되었습니다. 연구지원에 감사드립니다.

Reference

1. Akaike, H. (1974), “A new look at the statistical model identification”, IEEE Transactions on Automatic Control, Vol. 19, No. 6, pp. 716-723.
2. Bieniawski, Z.T. (1973), “Engineering classification of jointed rock masses”, Civil Engineer in South Africa, Vol. 15, No. 12, pp. 335-343.
3. Bruland, A. (1998), “Hard rock tunnel boring: Advance Rate and Cutter Wear”, Doctoral Thesis, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, pp. 6-26.
4. Cousineau, D., Chartier, S. (2010), “Outliers detection and treatment: a review”, International Journal of Psychological Research, Vol. 3, No. 1, pp. 58-67.
5. Chae, J.S., Lee, D.H., Lee, S. (2005), “A calculation of shield TBM advance rate in the shield tunnel”, Journal of The Korea Institute for Structural Maintenance and Inspection, Vol. 9, No. 1, pp. 35-41.
6. Chang, S.H., Choi, S.W., Bae, G.J., Jeon, S.W. (2007), “A parametric study of rock properties and mechanical cutting conditions for deriving an optimum design model of a TBM cutterhead equipped with disc cutters”, Journal of The Korean Society of Civil Engineers, Vol. 27, No. 1C, pp. 87-98.

7. Delisio, A., Zhao, J., Einstein, H.H. (2013), "Analysis and prediction of TBM performance in blocky rock conditions at the Lötschberg Base Tunnel", Tunnelling and Underground Space Technology, Vol. 33, pp. 131-142.
8. Farmer, I.W., Glossop, N.H. (1980), "Mechanics of disc cutter penetration", Tunnels and Tunnelling, Vol. 12, No. 6, pp. 22-25.
9. Gong, Q.M., Zhao, J., Jiang, Y.S. (2007), "In situ TBM penetration tests and rock mass boreability analysis in hard rock tunnels", Tunnelling and Underground Space Technology, Vol. 22, No. 3, pp. 303-316.
10. Gong, Q.M., Zhao J. (2009), "Development of a rock mass characteristics model for TBM penetration rate prediction", International Journal of Rock Mechanics and Mining Sciences, Vol. 46, No. 1, pp. 8-18.
11. Hamidi, J.K., Shahriar, K., Rezai, B., Rostami, J. (2010), "Performance prediction of hard rok TBM using Rock Mass Rating system", Tunnelling and Underground Space Technology, Vol. 25, No. 4, pp. 333-345.
12. Hamilton, W.H., Dollinger, G.L. (1979), "Optimizing tunnel boring machine and cutter design for greater boreability", Rapid Excavation and Tunneling Conference, Proceedings, Vol. 1, pp. 280-296.
13. Hassanpour, J., Rostami, J., Khamehchiyan, M., Bruland, A. (2009), "Developing new equations for TBM performance prediction in carbonateargillaceous rocks: a case history of Nowsood Water conveyance tunnel", Geomechanics and Geoengineering, Vol. 4, No. 4, pp. 287-297.
14. Hoerl, A.E., Kennard, R.W. (1970), "Ridge regression: Biased estimation for nonorthogonal problems", Technometrics, Vol. 12, No. 1, pp. 55-67.
15. Hughes, H.M. (1986), "The relative cuttability of coal-measures stone", Mining Science and Technology, Vol. 3, No. 2, pp. 95-109.
16. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), An Introduction to Statistical Learning: With Applications in R, Springer, New York, pp. 205-227.
17. Mahdevari, S., Shahriar, K., Yagiz, S., Shirazi, M.A. (2014). "A support vector regression model for predicting tunnel boring machine penetration rates", International Journal of Rock Mechanics and Mining Sciences, Vol. 72, pp. 214-229.
18. Mallows, C.L. (1973), "Some comments on Cp", Technometrics, Vol. 15, No. 4, pp. 661-675.
19. Myers, R.H. (1990). Classical and Modern Regression with Applications, PWS-KENT Publishing Company, Boston, pp. 222-227, 251-258.
20. Rahmatullah Imon, A.H.M. (2005), "Identifying multiple influential observations in linear regression", Journal of Applied statistics, Vol. 32, No. 9, pp. 929-946.
21. Sanner, M.F. (1999), "Python: a programming language for software integration and development", J Mol Graph Model, Vol. 17, No. 1, pp. 57-61.
22. Welsch, R.E. (1980), "Regression sensitivity analysis and bounded-influence estimation", Evaluation of Econometric Models, Academic Press, pp. 153-167.