*CORROSION SCIENCE AND TECHNOLOGY, Vol.18, No.2(2019), pp.61~71*
[Research Paper]
pISSN: 1598-6462 / eISSN: 2288-6524
DOI: https://doi.org/10.14773/cst.2019.18.2.61

# 기계학습을 이용한 유동가속부식 모델링: 랜덤 포레스트와 비선형 회귀분석과의 비교

이 경 근† · 이 은 희 · 김 성 우 · 김 경 모 · 김 동 진
한국원자력연구원 원자력재료연구부, 대전광역시 유성구 대덕대로 989번길 111
(2019년 3월 4일 접수, 2019년 4월 17일 수정, 2019년 4월 17일 채택)

# Modeling of Flow-Accelerated Corrosion using Machine Learning: Comparison between Random Forest and Non-linear Regression

**Gyeong-Geun Lee†, Eun Hee Lee, Sung-Woo Kim, Kyung-Mo Kim, and Dong-Jin Kim**
*Nuclear Materials Research Division, Korea Atomic Energy Research Institute (KAERI)*
*111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, Korea*
(Received March 04, 2019; Revised April 17, 2019; Accepted April 17, 2019)

Flow-Accelerated Corrosion (FAC) is a phenomenon in which a protective coating on a metal surface is dissolved by a flow of fluid in a metal pipe, leading to continuous wall-thinning. Recently, many countries have developed computer codes to manage FAC in power plants, and the FAC prediction model in these computer codes plays an important role in predictive performance. Herein, the FAC prediction model was developed by applying a machine learning method and the conventional nonlinear regression method. The random forest, a widely used machine learning technique in predictive modeling led to easy calculation of FAC tendency for five input variables: flow rate, temperature, pH, Cr content, and dissolved oxygen concentration. However, the model showed significant errors in some input conditions, and it was difficult to obtain proper regression results without using additional data points. In contrast, nonlinear regression analysis predicted robust estimation even with relatively insufficient data by assuming an empirical equation and the model showed better predictive power when the interaction between DO and pH was considered. The comparative analysis of this study is believed to provide important insights for developing a more sophisticated FAC prediction model.

***Keywords:*** *FAC, Statistical modeling, Machine learning, Random forest, Non-linear regression*

## 1. 서론

유동가속부식 (Flow-Accelerated Corrosion, FAC)은 금속 배관에서 빠른 유속을 지닌 유체의 흐름에 의하여 금속 표면에 형성된 보호피막이 용해되고, 이로 인하여 지속적으로 감육 (wall-thinning)이 발생하는 현상을 말한다 [1]. FAC가 발생하기 위해서는 특정 재료 및 수력학적 그리고 수화학적 환경 조건을 만족해야 한다. 원자력발전소의 2차 계통인 주증기 배관 (main steam piping) 과 급수 배관 (feedwater piping)은 고온, 고속의 물과 증기가 흐르는 배관으로, 가동조건 및 재질이 FAC 발생조건에 해당하므로 감육으로 인한 파손이 발생할 수 있다. 실제 FAC에 의한 배관 감육으로 인하여 1986년 미국의 Surry 2호기 원자력발전소 및 2004년 일본 Mihama 3호기에서 배관 파단 사고가 발생하였고, 이후 많은 원전 보유국에서 FAC를 정량적으로 예측하고 방지하기 위한 연구를 수행하였다 [2-4]. 축적된 연구 결과를 바탕으로 각 국가에서는 FAC를 예측할 수 있는 전산코드를 개발하여 원전배관 관리에 사용하고 있다 [5-8]. 이러한 전산코드들은 발전소 배관의 형상 및 배치를 실제와 유사하게 구현하고, 편리한 사용자 인터페이스를 가지고 있어 전체 배관관리에 매우 유용하다. 전산코드의 엔진이라고 할 수 있는 FAC 속도 예측 모델은 환경조건을 입력 변수로 받아 FAC의 속도를 예측한다. 이러한 모델은 모사 실험을 통하여 얻어진 부식 기구에 대한 이해를

† Corresponding author: gglee@kaeri.re.kr

바탕으로 개발하기도 하고, 산업현장에서 축적된 검사자료를 기반으로 경험적인 (empirical) 통계 모델을 구축하기도 한다. 예측 모델은 각 코드의 핵심기술로, 전체 전산코드의 성능에 중요한 역할을 담당하고 있으며, 지속적인 개량을 통하여 예측력을 향상시키고 있다.

최근 알파고의 등장 이후 각광받고 있는 기계학습 (machine learning) 기법을 데이터의 예측 모델링에 도입하려는 시도는 과거부터 많이 있었다 [9]. 그 실체를 확인해 보면, 우선 단순한 숫자 데이터를 활용하여 전통적인 방법으로 모델을 구성하는 통계 모델링 (statistical modeling), 인간의 뇌세포와 유사한 신경망 (neural network)을 이용하여 비선형 데이터를 분석하고 예측할 수 있는 기계학습, 그리고 수치형 데이터가 아닌, 영상 인식 및 음성 인식, 언어 분석 등 보다 복잡한 대상까지 적용이 확대된 딥 러닝 (deep learning) 기법으로 분류될 수 있다. 기본적으로 기계학습 단계에 도달하면, 선형 또는 비선형 회귀와 같이 모델식을 가정하는 것이 아니라, 데이터의 특징을 자동으로 구분할 수 있는 기능을 포함하게 된다. 즉, 데이터 간의 경향을 학습하여, 미래의 데이터를 예측할 수 있게 된다.

FAC의 경우, 많은 입력 변수가 필요하고 복잡한 기구에 의하여 해석된다. 기본적으로 2차 계통수의 온도, 유속, pH, 용존산소 농도 (dissolved oxygen, DO), 그리고 배관 재료의 Cr 함량 등 화학 조성, 배관의 물리적 형상 등 상당히 다양한 조건을 포함하고 있다. 특히, 이들 조건이 FAC에 대하여 선형적인 경향을 나타내지 않기 때문에, 정확한 모델의 구축에 어려움이 있다. 최근에 수집된 데이터를 바탕으로 핵심 입력 변수인 온도, 유속, pH, Cr 함량에 대한 비선형 회귀분석을 실시하였다 [10]. 여기에 용존산소의 효과를 추가한 모델을 구축하려 했으나, 문헌별로 실험 결과에 상당한 분산이 존재하여, 하나의 비선형 회귀모델을 만드는 것이 매우 어려웠다. 이에 본 연구에서는 기계학습 기법 중, 가장 사용이 편리하면서도 우수한 성능을 지니는 랜덤 포레스트 (random forest) 방법을 이용하여, 데이터 기반 FAC 예측 모델을 구현하고, 기존의 비선형 회귀분석과 비교하였다. 이를 통하여 데이터 기반 모델링과 경험식 기반 모델링과의 차이점을 분석하였고, 향후 기술 보완시 고려해야 할 점을 논의하였다. 이러한 분석은 지속적으로 발전하고 있는 기계학습 모델의 적절한 적용에 대한 중요한 통찰을 제공할 수 있으며, 좀더 정교한 FAC 예측모델 개발의 기초를 확립하는데 유용하게 사용될 수 있다.

## 2. 연구방법

### 2.1 분석자료

본 분석에 사용된 데이터는 일본 전력중앙연구원에서 작성한 보고서에서 공개한 실험 자료를 이용하였다 [11,12]. 다른 문헌에 비하여 데이터의 실험 조건이 비교적 정확하게 기술되어 있고 다양한 조건에 대하여 실험을 수행하였기에 FAC 전반에 대한 경향을 해석하기 위한 자료로 적합하였다. 이들 자료 모두 유사한 실험장치를 이용하여 결과를 얻었기에, 그 경향성이 뚜렷하게 나타났다. 실제 데이터는 수치형 데이터로 공개되어 있지 않았지만, 비교적 정확하게 그려진 그래프로부터 각 변수의 값을 확인할 수 있었다. 원본 데이터에 대한 정보를 Table 1에 정리하였다.

Table 1에서 주의할 점은 루프를 1회 가동시킬 때 실험을 1 Run이라고 했을 때, Run 1-8 실험까지는 loop A를 이용

**Table 1 Raw data points from References**

| Reference | Data points | Loop type | Run number | Figure numbers |
|---|---|---|---|---|
| 11 | 79 | A | 1,2,3,4 | 5.1, 5.2, 7.1, 7.2, 9, 11, 17, 19 |
| 12 | 183 | A | 5,6,7,8 | 2.4.8, 2.5.6, 2.6.3, 2.7.6 |
| 12 | 76 | B | 9, 10 | 2.8.7, 2.9.6 |

**Table 2 Ranges of variables in dataset FACDB**

| Abbreviation | Variables | Unit | Minimum | Maximum | Average |
|---|---|---|---|---|---|
| FAC | FAC rate | mm/year | 0.00 | 8.06 | 0.87 |
| FR | Flow rate | m/s | 2.7 | 10.5 | 5.3 |
| TW | Water temperature | ˚C | 67.2 | 180.0 | 144.5 |
| PH | pH at 25˚C | | 7.0 | 10.0 | 8.7 |
| DO | Dissolved Oxygen | ppb | 0.0 | 54.2 | 4.8 |
| CR | Cr content | wt% | 0.001 | 0.660 | 0.042 |

하였고, Run 9-10 실험은 loop B를 이용하였다. Loop A는 재순환식 루프로 main tank의 DO 농도와 시편 근처의 DO 농도와는 상당한 차이가 있었다. 참고문헌에서는 시편 근처에서의 DO를 main tank의 DO의 약 1/9로 계산하여 사용하였다 [12]. Loop B는 재순환 펌프를 제거하고 main tank로부터 직접 DO의 농도를 조절하여 공급하는 형태로 개조되었다. 이와 같은 loop A와 loop B 간의 구조적인 차이로 인하여, main tank에서 측정한 DO의 농도에 따른 FAC 속도 실험 결과가 상당히 다르게 나타났다. 본 논문의 목표는 일관성 있는 데이터에 대하여 기계학습을 이용한 FAC 예측모델의 수립이므로, 상대적으로 데이터 수가 적은 loop B의 실험 결과를 제외하고 loop A의 실험 결과인 262개 만을 이용하였다(이하 FACDB로 표기).

그래프 형태의 실험 결과에서 데이터 값을 추출할 때 주의할 점 중의 하나는 데이터의 중복 입력 가능성이다. 하나의 FAC 결과를 얻기 위해서는 5개 이상의 실험 조건이 명기되어야 하는데, 동시에 모든 실험 조건을 2차원 그래프에서 나타내기 어렵다. 즉 한 장의 그래프에 모든 조건을 표현하기 어렵기 때문에, 이해를 돕기 위하여 하나의 결과를 두 개 또는 세 개의 그래프로 나누어서 그리게 된다. 이때, 그래프 별로 중복되어 표시된 포인트를 잘 골라 입력하는 것이 중요하다. 중첩된 포인트는 전체 모델의 경향성에 심각한 왜곡을 야기할 수 있기 때문이다. 실제, 데이터 기반 모델링에서 가장 많은 시간이 투입되는 분야가 바로 데이터의 입력 및 검증 단계이다. 이 단계가 정확하게 수행되지 않을 경우, 분석된 방법의 유효성과 상관없이 잘못된 분석 결과가 나올 수 있기 때문이다. Table 2에 실제 모델링에서 사용된 데이터들의 특성을 범위 별로 나타내었다. 이후 편의를 위하여 각 변수들은 약어로 표시하였다.

### 2.2 분석방법

축적된 데이터 포인트를 이용하여, 다른 조건에서의 결과값을 예측할 때, 결과값이 연속적인 값으로 나오게 되면 이를 회귀 분석 (regression analysis)이라 한다. 본 연구에서는 랜덤 포레스트를 이용한 회귀 분석과 전통적인 방법인 비선형 회귀를 이용하여 회귀 분석을 수행하였다. 본 연구에서는 오픈소스 통계프로그램인 R을 사용하였고 [13], 랜덤 포레스트 회귀를 위해서는 randomForest 패키지 [14], 그리고 비선형 회귀를 위해서는 minpack.lm 패키지를 이용하였다 [15]. 모델의 과적합(overfitting)을 막기 위한 교차검증(cross-validation) 방법으로는 leave-one-out cross-validation (LOOCV) 방법을 이용하였다 [9]. 여기서 교차검증이란 기계학습에서 모델이 적합에 사용된 데이터에 지나치게 과적합(overfitting)되어, 실제 새로운 데이터의 도입 시 예측력이 떨어지는 것을 고려하여 모델을 평가하는 기법이다. LOOCV는 전체 데이터 중 하나를 빼고 나머지 데이터를 이용하여 모델을 적합한 후, 빠진 데이터에 대하여 모델 예측값과 실측값을 비교하는 방법으로 본 연구와 같이 데이터 포인트가 262포인트 밖에 되지 않는 경우, 적합한 방법이라 할 수 있다. LOOCV에 대한 자세한 설명은 참고문헌을 확인하기 바란다 [9].

## 3. 결과 및 분석

### 3.1 Data 결측값 해결 및 특성 분석

통계 분석용 프로그램에 입력된 데이터의 일부를 Fig. 1에 나타내었다. 일부 데이터 포인트에서 유속 FR이 명시적으로 나타나있지 않는 결측값, 즉 NA (not available)로 표기된 경우가 있다. 이는 실제 실험에서 내경 1-3 mm인 관 형태의 시편을 직렬로 네 개를 연결하고, 루프 가동 중에 비파괴적인 방법으로 관의 직경 감소를 연속적으로 측정하여 두께 변화를 계산하였기 때문이다. 전체 루프의 유속과 시편의 최초 내경이 실제 실험에서의 조건으로 사용되었으며, 실험 중에 시편의 내경은 시간의 증가에 따라 내부에 FAC가 발생하여 깎여 나가기 때문에 증가하게 된다. 증가한 직경만큼 실제 유속은 점차 느려 지므로, 시간의 흐름에 따라 FAC 속도도 달라지게 된다. 따라서, 본 자료에서 FR과 FAC는 서로 독립이 아니라 상관관계를 가지는 것을 알 수 있다. 본 연구에서 사용된 DB중 참고문헌 11에 존재하는

| | ID | FAC | TW | FR | PH | DO | CR | source | fig | batch | series | flow_rate_L_per_min | diameter_mm | loop |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | <int> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <int> | <chr> | <chr> | <int> | <dbl> | <dbl> | <chr> |
| 1 | 1 | 0.767 | 125. | 10.2 | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 1 | 2 | 1.6 | A |
| 2 | 2 | 1.82 | 160. | 10.1 | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 1 | 2 | 1.6 | A |
| 3 | 3 | 0.000300 | 172. | NA | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 1 | 2 | 1.6 | A |
| 4 | 4 | 0.153 | 173. | 10.6 | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 1 | 2 | 1.6 | A |
| 5 | 5 | 0.00029 | 175. | 10.6 | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 1 | 2 | 1.6 | A |
| 6 | 6 | 0.394 | 125. | 5.54 | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 2 | 2 | 2.4 | A |
| 7 | 7 | 0.762 | 161. | 5.54 | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 2 | 2 | 2.4 | A |
| 8 | 8 | 0.766 | 172. | NA | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 2 | 2 | 2.4 | A |
| 9 | 9 | 0.511 | 173. | 5.78 | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 2 | 2 | 2.4 | A |
| 10 | 10 | 0.790 | 176. | 5.78 | 9.2 | 0.1 | 0.001 | 10 | Fig5.1+7.1 | RUN1A | 2 | 2 | 2.4 | A |

**Fig. 1 Screenshot of some DB records stored in the FACDB.**

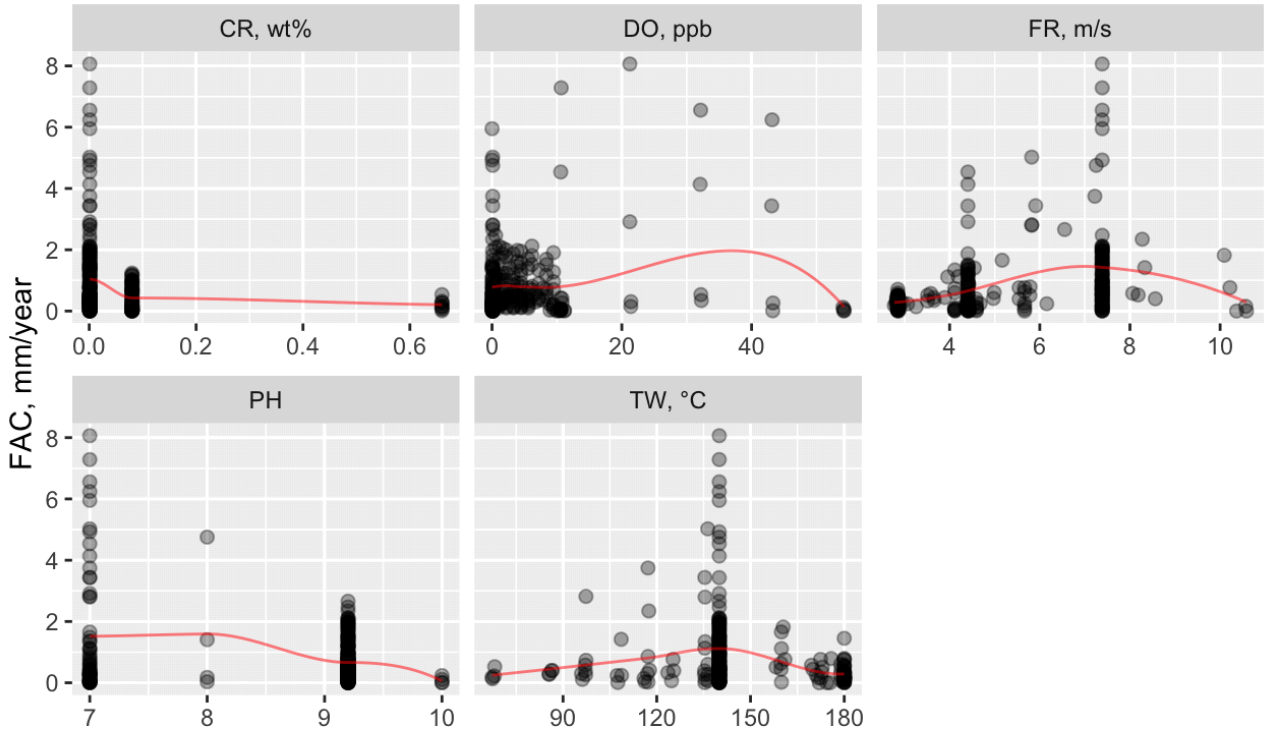


**Fig. 2 Scatter plots of 5 input features on the FAC rate in FACDB.**

총 79개의 데이터 포인트 중 31개가 FR이 결측값으로 존재한다. 이러한 결측값을 처리하기 위해서 해당 포인트 주위의 값에 대하여 내삽을 이용하여 근사하였다. 참고문헌 12의 데이터는 183포인트 전부 다 명시적으로 FR이 제공되지 않으나, 루프의 유속에서 관의 내경에 따라 linear velocity를 계산해주는 그래프를 제공하고 있어, 이를 이용하여 FR을 추정하였다. 실제 실험에서 모든 조건을 명시적으로 측정하는 것이 가장 바람직하지만, 모델 실험의 한계로 인하여 우리가 고려하는 특성에 대하여 결측값이 발생할 수 있다. 이 때 결측값에 대한 적절한 처리가 전체 모델의 예측 성능에 중요한 요소가 되며, 데이터 분석가에 따라 처리 방법의 선택이 다르므로 모델의 결과가 조금씩 달라지는 원인이 되기도 한다.

Fig. 2에 본 모델링에서 고려하고 있는 5가지 변수에 따른 FAC 속도를 262 포인트 전체에 대하여 산포도로 나타내었다. 그래프 중의 실선은 해당 조건에서의 각 포인트들의 평균값을 의미하며, FAC에 미치는 각 변수들의 정성적인 효과를 보여준다. 데이터 상에서 몇 가지 주목할만한 특징은 다음과 같다. CR의 경우, 단지 세가지 조건에서만 실험이 진행된 것을 확인할 수 있다. 따라서 이 조성 범위 안에서 내삽을 진행할 지라도 이에 대한 전체 경향을 정확하게 예측하는 것은 쉽지 않다. DO의 경우, 주로 낮은 DO에서 많은 실험이 수행된 반면, 높은 DO에서는 별로 실험이 수행되지 않아 데이터 포인트의 심각한 불균형 (imbalance)이 존재하고 있다. 이 경우, 데이터가 많은 부분에서의 경향이 데이터가 적은 부분에서의 경향을 압도하게 되어 전체 경향이 제대로 표현되지 않게 된다. 특히 DO가 55 ppm 이상에서는 FAC가 0에 가깝게 떨어지는 것을 알 수 있는데, 이 이상 범위에서는 실제 데이터 포인트가 없으므로, 외삽을 통하여 결과를 예측할 경우 큰 오차가 발생할 것으로 예측된다. 5가지 변수 모두에서 나타나는 중요한 경향은 평균선이 매우 낮게 나타난다는 것이다. 이는 특정 조건을 동시에 만족하는 매우 작은 영역에서만 FAC가 높게 발생하는 것을 의미한다. FAC 모델링이 이 영역을 정확하게 예측하는 것이 목표인데, 실제로 이 영역에 해당하는 데이터 포인트가 그리 많지 않다는 점에서, 데이터 기반 모델링을 구현하는데 어려움으로 작용할 것임을 예상할 수 있다.

### 3.2 Random Forest 회귀 분석

기계학습 방법을 이용한 회귀 분석 중에 최근에 널리 사용되는 방법 중 하나가 바로 랜덤 포레스트 기법이다. 이 랜덤 포레스트 기법을 이해하기 위해서는 먼저 결정 트리 (decision tree)에 대하여 이해하는 것이 중요하다. 결정 트리란, 주어진 입력 값에 대하여 간단한 조건, 보통 2분법으로 연속적으로 나누어 출력 값을 제공하는 방법이다. 이러한 조건은 컴퓨터에 의하여 자동적으로 계산되므로 어떤 문제에 대해서

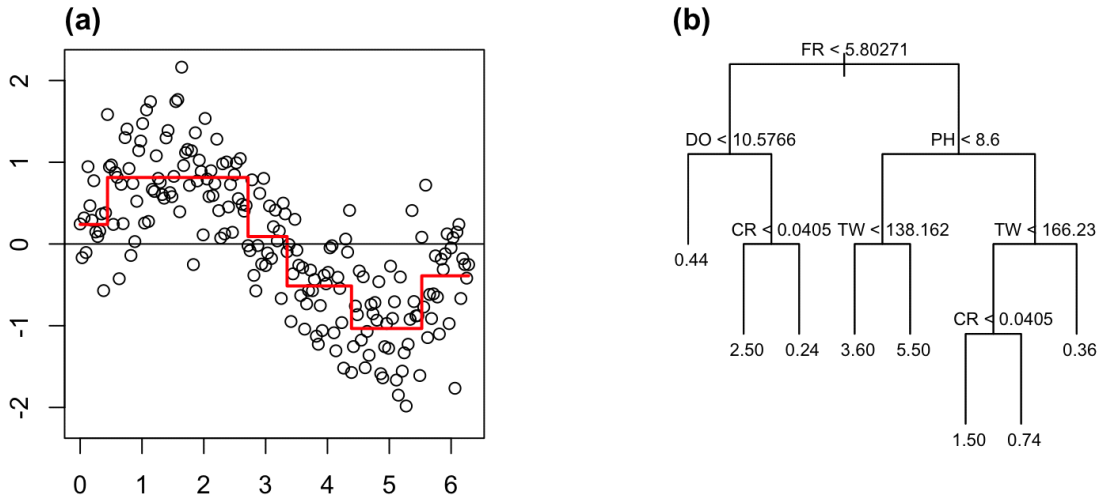


**Fig. 3 Example of decision tree model: (a) sin curve, (b) FACDB.**

도 별 어려움 없이 간단하게 분류 및 회귀가 가능하다. 물론 연속적인 변화량에 대해서는 매끄러운 곡선으로 나타나지 않는 단점이 있지만, 이는 여러 번 분할하여 모델을 구성할 경우, 어느 정도 해결이 가능하다. Fig. 3a에서는 단변수 x의 변화에 따른 y의 변화를 결정 트리를 이용하여 회귀 분석하는 것을 보여주고 있다. 실제로는 sin 곡선과 오차를 결합하여 생성된 모델이지만, 결정 트리를 이용할 경우, 충분한 가지치기를 도입해서 그 값을 근사할 수 있다는 것을 보여준다. Fig. 3b에 결정 트리를 이용하여 FAC를 예측한 간단한 예를 나타내었다. FACDB는 제일 먼저 FR의 특정 값 5.8을 중심으로 나뉠 수 있으며, FR이 5.8 이하인 경우 DO의 농도가 다음으로 중요하고, DO의 농도가 높을 경우에는 그 다음으로 Cr의 농도에 의하여 최종적인 FAC 값이 결정되는 것을 알 수 있다. 반대로, 유속이 5.8 이상인 경우, pH가 다음으로 중요한 변수가 되는 것을 알 수 있다. 실제 복잡한 다중 회귀 분석을 할 때, 각 변수의 특징에 대한 모델을 만들고 그 유효성을 검증하는 데에는 많은 시간과 통찰이 필요하다. 하지만, 결정 트리를 이용하면, 입력 변수, 즉 특성이 너무 많아 모델을 만들 기 어려운 경우에도 쉽고 간편하게 짧은 시간 안에 결과를 얻을 수 있다.

이와 같은 단일 결정 트리의 경우, 약간의 데이터 변화로도 전체 모형의 형태가 달라지게 된다. 즉 비슷한 수준의 정보력을 가지는 두 변수가 있을 때, 첫번째 변수가 어떤 것이 선택되느냐에 따라서 최종적인 트리의 구성이 크게 달라지게 된다. 즉, 학습 데이터에 따라서 민감하게 변화하게 되며, 또한 중간에 오차가 발생하면 다음 단계에서 지속적으로 오차가 전파하는 단점이 있다. 이와 같은 단점을 보완하기 위하여 나온 것이 바로 랜덤 포레스트 방법이다. 랜덤 포레스트 방법은 전체 데이터에 대하여 하나의 트리를 만드는 것이 아니라, 전체 데이터 중 일부의 데이터를 활용한 소규모 트리를 대량으로 만든 후, 새로운 값에 대하여 각 트리의 예측값을 평균 또는 다수결로 투표하여 값을 예측하는 것이다. 즉, 하나의 완벽한 모델을 만드는 것이 아니라, 약간은 부족하지만 다수가 의견을 종합하는 집단 지성을 활용하는 방법으로, 단일 트리에 비하여 월등히 높은 정확성을 보여준다. 특히, 단일 트리에 비교했을 때, 이상치(outlier)에 대하여 상당히 견고한 예측모델을 만들 수 있다. 이 때, 트리의 수는 보통 500 개에서 몇 천 개 정도이며, 데이터는 자체 데이터를 임의로 샘플링하여 재활용하고 입력 특성 중 일부를 생략하여, 일반적인 트리에서 나타날 수 있는 과적합을 억제하고 모델의 예측력을 높이게 된다. 구체적인 수식 및 정보에 대해서는 참고문헌에서 확인할 수 있다 [14].

가장 쉽게 생각해 볼 수 있는 적합 방법은 전체 데이터 FACDB를 이용하는 방법이다. 이 때 총 다섯 가지 입력 변수 (TW, FR, PH, DO, CR)를 이용하여 FAC를 적합하였고, 트리의 수는 1000개, 각 분기점에서 임의로 선택할 수 있는 입력 변수의 수는 2로 설정하였다 (model RF1). 랜덤 포레스트의 장점 중의 하나가 바로 모델의 선택에 필요한 매개변수 (parameter)의 수가 다른 기계학습 방법에 비하여 적다는 것이다. 이는 약간의 매개변수 만으로도 해당 DB에 최적인 랜덤 포레스트 모델을 얻을 수 있다는 것이다. 구축된 모델 RF1에서 적합에 사용된 전체 데이터의 평균제곱근 오차 (root mean square error, RMSE)를 계산하면 0.43이고, 이 모델로 설명 가능한 분산은 69%로 나타난다. Fig. 4에 RF1에서 예측된 FAC 크기에 따른 잔차 (residual)의 분포를 나타내었다. 이상적인 회귀 모델일 경우 등분산 가정에 의하여 전체 분산이 0을 중심으로 고르게 퍼지게 된다.

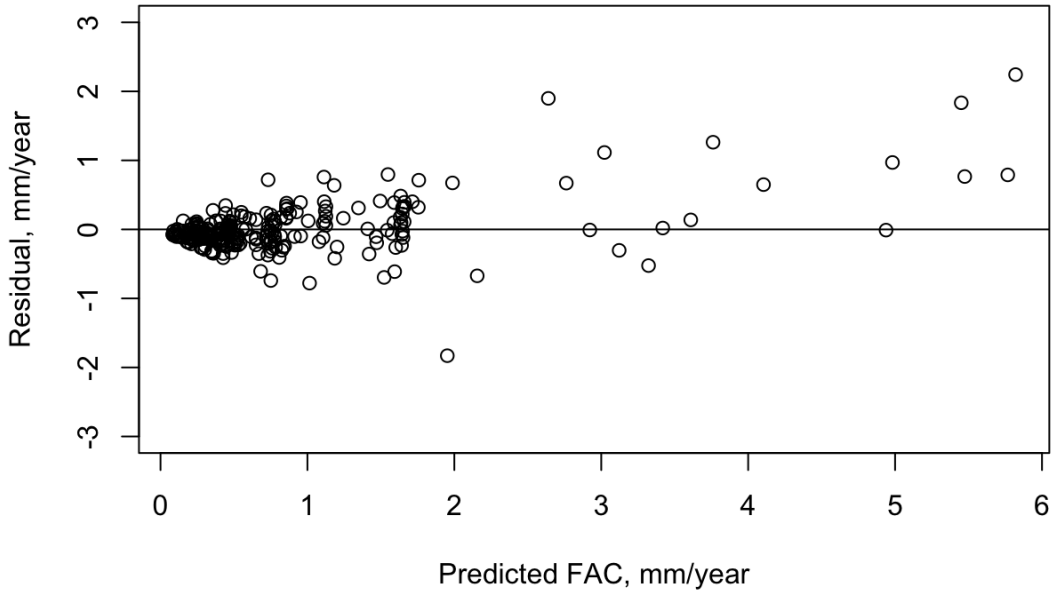


**Fig. 4 Residual plot of model RF1. All data was used to fit the model.**

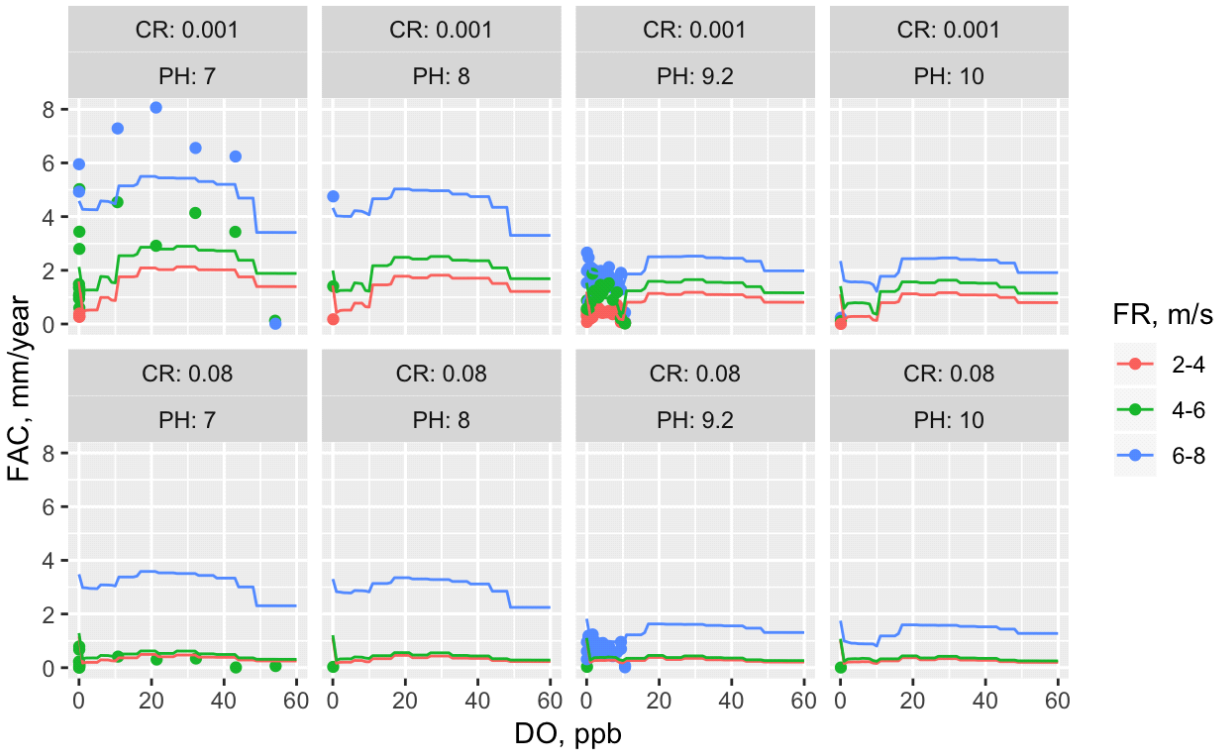


**Fig. 5 DO Effect of RF1 on the FAC at various conditions. The data points are represented as circles, and the predicted values from model RF1 are represented as lines.**

하지만, 본 모델에서는 FAC 예측값이 커질수록 잔차의 폭이 커지며, 특히 양으로 편향이 일어나는 것을 알 수 있다. 또한, 주목할 것은 대부분의 데이터가 FAC가 작은 범위에 주로 존재하며, 큰 FAC 값을 가진 데이터가 많지 않다는 점이다. 따라서 이 모델이 실제 경향을 잘 나타내고 있는지에 대한 검증이 필요하다.

가장 간단하게 생각해 볼 수 있는 검증방법은 실제 데이터와 예측된 값을 직접 도시하여 확인하는 방법이다. 이 방법

의 경우, 입력 변수가 매우 많아지게 되면, 실제 검증하기가 쉽지 않다. 하지만, 본 연구에서는 5가지 변수만을 다루기 때문에, 적절한 범위에서 입력 변수를 선택하여 도시할 경우, 쉽게 경향을 파악할 수 있다. Fig. 5는 랜덤 포레스트로 예측한 DO의 영향을 TW 조건 130 ℃ ~ 150 ℃에서, PH 네 조건과 CR 두 조건 그리고 FR은 세 가지 조건에서 도시한 그림이다. 가장 주목해야 할 점은 데이터 기반 모델에서 외삽의 문제점이다. 실제 DO의 데이터를 보게 되면, 특정 범위 밖에서는 0에 가깝게 DO가 감소하는 것을 알 수 있다. 하지만, 이는 암시적인 (implicit) 결과로, 실제로 실험을 수행하여 데이터를 얻은 것이 아니다. 데이터에 의해서만 결정되는 랜덤 포레스트에서는 명시적인 (explicit) 데이터가 있어야 모델에 반영이 된다. 즉, DO가 높은 구간에서는 데이터가 없어 제대로 된 예측을 하지 못한다. 또한, DO의 효과가 볼록한 곡선을 그리며 휘어지는 것을 알 수 있다. 이는 RF1 모델에서 주어진 데이터에 가장 잘 맞는 과적합 (overfitting)이 발생하였기 때문이다. 이와 같은 과적합은 현재의 데이터는 잘 맞추나, 미래의 데이터에 대해서는 예측력이 떨어지게 된다. 이를 확인하기 위한 간단한 방법이 바로 교차검증이다.

교차검증을 수행하는 다양한 방법 중에 본 연구에서는 LOOCV를 이용하였다. 이는 전체 데이터에서 하나의 데이터를 제외하고 나머지 데이터를 이용하여 모델을 적합한 후, 이 모델을 이용하여 사용하지 않았던 데이터 하나를 예측하는 방법이다. 본 연구에서는 262포인트가 있으므로 총 262번 회귀 분석을 시도해야 한다. 상당한 계산량이 될 수도 있지만, 컴퓨터 계산 속도의 향상으로 그다지 많은 시간이 필요하지는 않았다. LOOCV로 계산한 RF1의 RMSE는 0.67로 계산되었다. 이는 전체 데이터를 사용했을 때 RMSE 0.43에 비하여 많이 증가한 값이다. 이와 같은 증가는, FAC 값이 높은 포인트의 포함 여부에 의하여 전체 모델이 크게 변화하기 때문이다. 일부 포인트의 존재 유무가 전체 모델의 적합성을 결정하게 되므로, 모델의 예측력은 감소하게 된다. 결국 전체 데이터를 이용해서 기계학습 방법을 이용할 때에는, 필연적으로 과적합이 발생하게 되는 것을 확실히 알 수 있었고, 모델의 적합여부를 확인하기 위한 가장 기본적인 단계로 그래프를 이용한 도시와 교차검증을 실시하는 것이 필수임을 확인하였다.

### 3.2 비선형 회귀분석

데이터 포인트가 부족한 실험 자료로부터 일반적인 경향을 표현하기 위해서는 모델식을 가정하는 비선형 회귀 방법이 유리할 수 있다. 특히, FAC는 다양한 입력 변수의 범위 중에서 특정 영역을 동시에 만족할 때만 발생한다. 즉, 개별 입력 변수와 FAC가 서로 선형적인 관계를 가지는 것이 아니라, 입력 변수들간의 특정 교집합에서 주로 발생하게 된

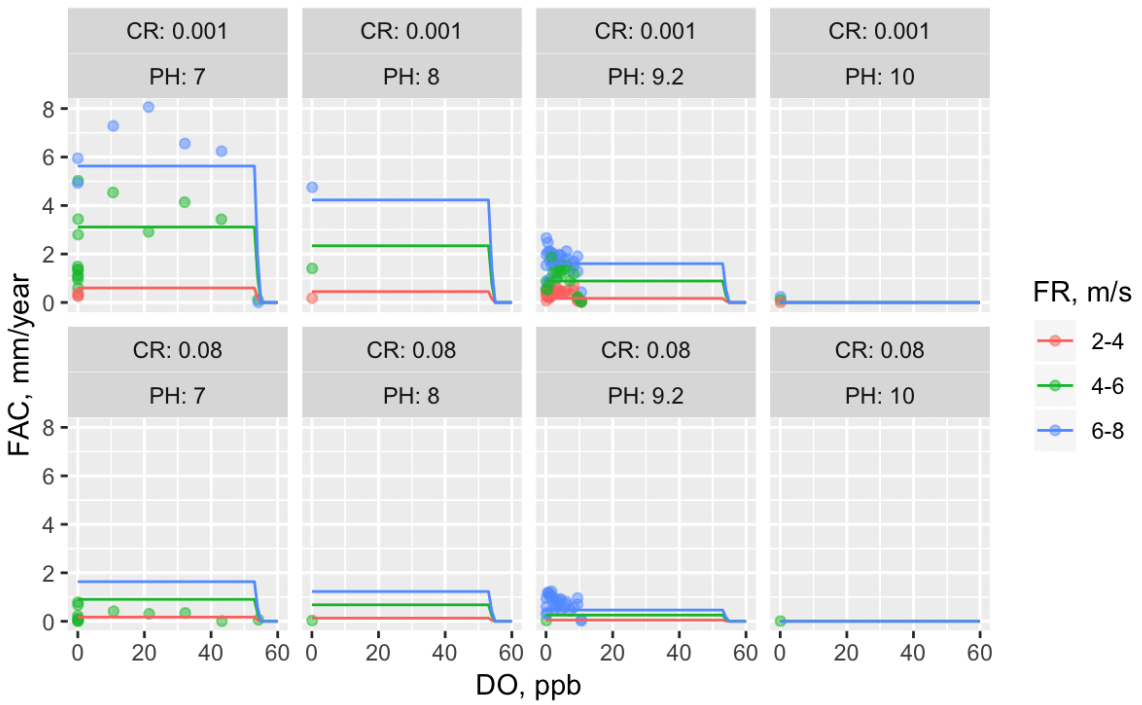


**Fig. 6 DO Effect of NR1 on the FAC at various conditions. The data points are represented as circles, and the predicted values from model NR1 are represented as lines.**

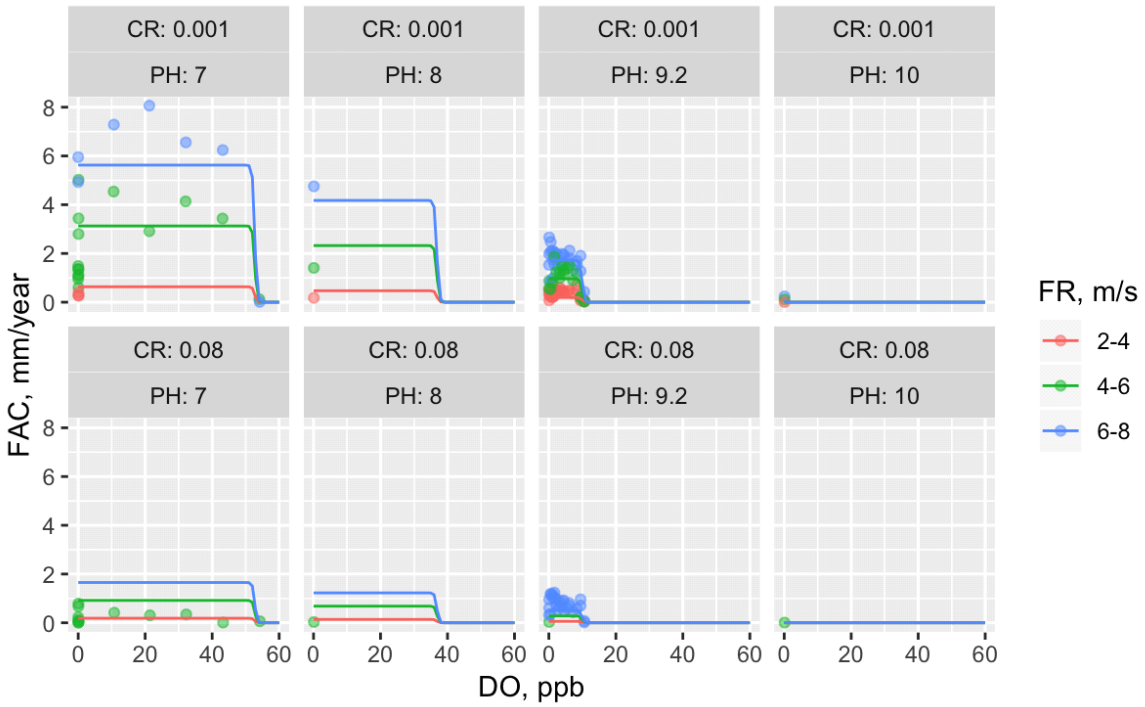


**Fig. 7 DO Effect of NR2 on the FAC at various conditions. The data points are represented as circles, and the predicted values from model NR2 are represented as lines.**

다. 이는 Fig. 2에서 쉽게 확인할 수 있다. 즉, 일부 조건에서만 FAC가 높게 나타나며, 대부분의 조건에서는 평균값이 낮게 나타나는 것이다. 이와 같은 특성을 고려하여, TW, FR, CR, PH 네 변수간의 상호작용을 독립적인 곱셈의 형태로 지정한 비선형 회귀 모델이 많이 쓰이고 있다 [5,10]. 본 논문에서는 DO의 효과까지 포함하여 총 5가지의 입력 변수를 고려한 비선형 회귀모델을 구축하였다 (model NR1). 여기서 DO의 효과는 sigmoid 형태의 함수로 가정하였다. 이에 대해서는 특별한 기구를 고려하지 않고 데이터의 경향을 적절히 고려하여 선정하였다. 다른 입력 변수의 모델식은 참고문헌에서 확인할 수 있다 [10].

D1: f(DO, d1, d2) = 1 / (1 + exp(DO − d1) * d2
CR: Cr content
d1, d2: model parameter for D1
Model NR1: FAC = F1 × T2 × P1 × C1 × D1 (3)

NR1을 이용하여 전체 데이터 FACDB를 사용하여 적합한 경우, RMSE는 0.51로 계산되었다. 그리고 LOOCV로 교차검증한 결과 RMSE는 0.54로, 두가지 방식에서의 RMSE 차이가 그리 크지 않았다. 이는 NR1에서 사용한 총 매개변수의 수가 11개에 불과하기 때문에 데이터 수에 비하여 매우 적어 과적합이 발생하지 않았고, 암시적인 데이터의 경향을 명시적인 모델식으로 지정하여 계산하였기 때문으로 해석된다. Fig. 6에 NR1을 이용하여 계산한 DO의 효과를 나타내었다. DO의 효과를 다른 입력 변수에 독립적으로 설정하였기에 어떤 PH, CR, 및 FR 조건에서도 DO의 효과는 55 ppb 근처에서 0를 나타내게 된다. 하지만, 실제 데이터는 PH가 7에서 9.2로 증가하면서 DO의 한계치가 55 ppb에서 10 ppb로 감소하는 경향을 나타낸다. 이와 같은 DO에 미치는 PH의 상호 작용을 추가로 고려해야 데이터에 더욱 적합한 모델이 될 수 있다.

이 문제를 개량하기 위한 두번째 비선형 회귀모델에서는 DO를 계산하는 모듈 내에 PH의 효과를 추가로 넣어서 모델을 수정하였다 (model NR2). 이 때 계산을 용이하게 하기 위하여 PH의 함수 형태도 약간의 변형을 시도하였다. 새롭게 구성된 모델은 다음과 같다.

P2: f(PH, p1, p2) = -1 / (p2 − p1) * (PH − p1) + 1
(If P2 >= 1 then P2 = 1, and If P2 <= 0 then P2 = 0)
DP1: f(DO, PH, d1, d2, p1, p2, p3) = 1 / (1 + exp((DO - d1 * p) * d2))
(p = p3 * (-1) / (p2 - p1) * (PH - p1) + 1,
If p >= 1 then p = 1 and If p <= 0 then p = 0)
Model NR2: FAC = F1 × T2 × P2 × C1 × DP1 (2)

**Table 3 Comparison of RMSE among various models**

| Model | RMSE with All data | RMSE with LOOCV | Difference |
|---|---|---|---|
| RF1 | 0.43 | 0.67 | 0.24 |
| NR1 | 0.51 | 0.54 | 0.03 |
| NR2 | 0.48 | 0.52 | 0.04 |

수정된 모델 NR2를 이용하여 계산된 DO의 효과를 Fig. 7에 나타내었다. PH의 증가에 따라 DO의 효과가 사라지는 한계치가 변화하는 것을 잘 재현하는 것을 알 수 있다.

최종적으로 결정된 모델 NR2에 대해서, 전체 데이터에 대한 RMSE와 LOOCV로 교차검증했을 때의 RMSE를 Table 3에 정리하여 나타내었다. 과적합이 발생하지 않은 NR1와 NR2는 RMSE 간의 차이가 적은 것을 알 수 있다. NR2는 NR1에 비하여 매개변수가 하나 더 늘어났지만, RMSE가 감소되었고, 실제 데이터를 잘 재현했기 때문에 더욱 적합한 모델임을 알 수 있다. Fig. 8에 NR2의 잔차도를 나타내었다. 잔차도 상으로는 Fig. 4에 표시된 RF1에 비하여 편향성은 없으나, 분산이 더 큰 것으로 보인다. 이는 잔차도가 전체 데이터를 이용하여 작성되었기 때문이다. 즉, NR2의 RMSE 0.48이 RF1의 RMSE 0.43보다 높으므로 잔차도 상에서는 더 큰 분산을 가진다.

최종적으로 선택된 모델 NR2를 이용하여, FACDB 데이터에 대한 적합면의 양상을 Fig. 9에 나타내었다. TW와 FR을 연속변수로 설정하고, PH는 7과 9.2 그리고, CR은 0.001과 0.04 두 조건으로 적합면을 생성하였고, 이 범위와 유사한 데이터 포인트를 함께 표시하였다. 전체 적합면이 데이터 상의 변화 양상을 적절히 재현하는 것을 알 수 있다. 일부 고 CR, 고 PH 조건에서는 NR2가 데이터에 비하여 과소한 값으로 표시되기도 한다. 이는 FR 효과가 선형으로 가정한 모델에서의 문제점으로 보이며, 추가로 개량할 여지가 있다.

종합해서 정리해 보면, 랜덤 포레스트의 경우, 데이터에 의해서만 모델이 구축되므로, 적절한 범위 내에서 데이터가 균형있게 분포할 경우, 매우 우수한 예측능력을 지니지만, 본 연구와 같이 장시간 또는 고난이도의 실험으로 데이터가 부족한 영역이 많이 존재하는 경우, 적절한 모델식을 결합한 비선형 회귀분석이 유리한 것을 알 수 있다. 물론, 데이터가 부족한 부분에 대하여 데이터 강화 (augmentation) 기법과 같이 결측된 데이터를 적절히 추가하는 방식을 이용하여, 기계학습법을 적용할 경우에는 향상된 모델을 얻을 수 있다. 하지만, 해당 내용은 본 논문의 범위를 넘어서기에 추가적인 연구 주제로 고려하고 있다.

만약에 소량 데이터에 대한 모델 구축이 아니라, 실험장비로부터 연속적으로 데이터가 수집되는 경우 또는 현재의 다섯 가지 변수 보다 더 많은 입력 변수를 고려할 경우에는 기계학습을 이용한 방법이 쉽고 간편하면서도 예측력이 높을 수 있다. 현재 KAERI에서는 FAC를 모사할 수 있는 대형설비를 보유하고 있으며, 이 장비로부터 실시간으로 자료를 수집하여 FAC 경향을 파악하고 진단할 수 있는 연구를 추진 중에 있다. 이 때에는 더 발전된 기계학습 방법을 이용한 모델이 더욱 유리할 것으로 사료되며, 이에 대한 추가 연구를 진행할 예정이다.

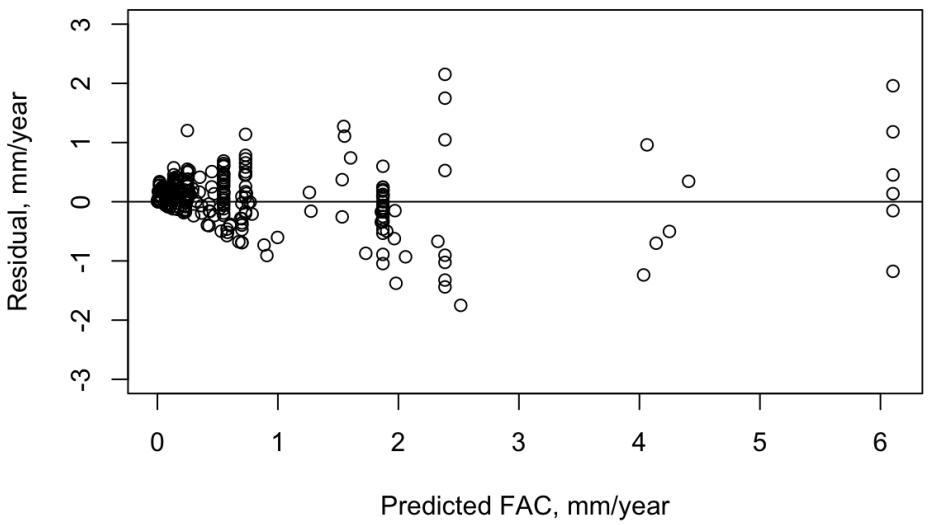


**Fig. 8 Residual plot of model NR2. All data was used to fit the model.**

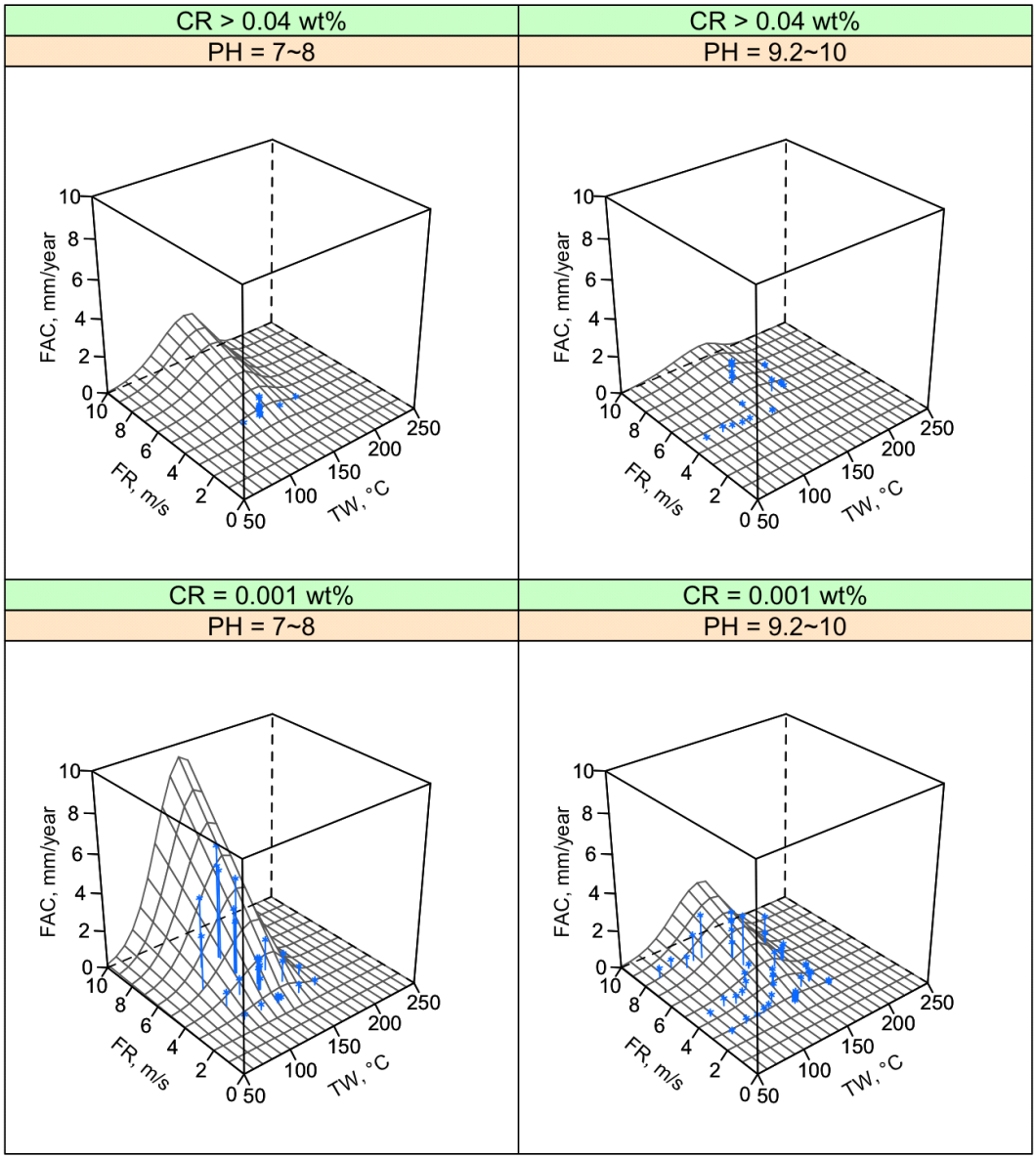


**Fig. 9 The measured data of FACDB and the fitted surface of model NR2. The data were collected under the condition of DO < 1 ppb.**

## 4. 결론

문헌상의 FAC 자료를 조사하고 분석하여 기계학습 방법인 랜덤 포레스트와 전통적인 방법인 비선형 회귀분석을 적용하여 FAC 예측 모델을 개발하였다. 랜덤 포레스트 방법은 다섯 가지 입력 변수에 대한 FAC 경향에 대하여 빠른 시간에 쉽게 계산할 수 있었지만, 실험상의 데이터가 적고 불균형한 분포를 가지고 있어, 교차검증 결과 실제 예측력이 많이 감소하였다. 특히 외삽 조건에서는 상당한 오차를 수반할 수밖에 없어 추가적인 실험 또는 데이터 강화를 이용하지 않고서는 적절한 예측 결과를 얻기 어려웠다. 대조적으로 비선형 회귀분석은 데이터로부터 경험적인 모델식을 가정함으로써 상대적으로 부족한 데이터에서도 유용한 결과를 예측할 수 있었다. 이때 모델에 사용된 각 요인은 각각의 독자적인 함수로 표시되어 곱의 형태로 결합되었으나, DO의 경우 pH와의 상호작용이 필수적으로 고려해야 더욱 적합한 모델이 되었다. 본 연구를 통하여 개발된 모델식은 자료의 지속적인 추가 및 이론적인 기구를 반영한 모델식의 개량을 통하여 정확성과 신뢰성을 향상시킬 예정이다.

## 감사의 글

본 연구는 한국원자력연구원 주요사업과제의 일환으로 수행되었습니다.

## References

1. B. Chexal, J. Horowitz, B. Dooley, P. Millett, C. Wood, R. Jones, M. Bouchacourt, F. Nordmann, P. S. Paul, and W. Kastner, *Flow-Accelerated Corrosion*, EPRI TR-106611-R1 (1998).
2. I. S. Woosey, *Assessment and avoidance of erosion-corrosion damage in PWR feedpipework*, pp. 60 - 65, IAEA Report: IWG-RRPC-88-1, Vienna (1990).
3. G. J. Bignold, K. Garbett, R. Garnsey, and I. S. Woosey, *Erosion-corrosion in nuclear steam generators*, pp. 5 - 18, Water Chemistry of Nuclear Reactor Systems 2, Bournemouth (1981).
4. J. Ducreux, *Theoretical and Experimental Investigation of the Effect of Chemical Composition of Steels on their Erosion-Corrosion Resistance*, Specialists Meeting on Erosion-Corrosion of Steels in High Temperature Water and Wet Steam, Les Renardieres (1982).
5. EPRI TR-103198-P1, *CHECWORKS Computer Program Users Guide*, EPRI (1997).
6. W. Kastner, M. Erve, N. Henzel, and B. Stellwag, *Nucl. Eng. Des.*, **119**, 431 (1990).
7. S. Trevin, M. Persoz, and T. Knook, *Making FAC Calculations With BRT-CICERO™ and Updating to Version 3.0*, pp. 81 - 88, 17th International Conference on Nuclear Engineering Volume 1: Plant Operations, Maintenance, Engineering, Modifications and Life Cycle; Component Reliability and Materials Issues; Next Generation Systems, Brussels, Belgium (2009).
8. K. Hwang, H. Yun, and H. Seo, *Corros. Sci., Tech.*, **17**, 317 (2018).
9. M. Kuhn and K. Johnson, *Applied Predictive modeling*, Springer, New York (2013).
10. G.-G. Lee, E. H. Lee, S.-W. Kim, and D.-J. Kim, *Transactions of the Korean Society of Pressure Vessels and Piping*, **12**, 40 (2016).
11. K. Fujiwara, M. Domae, K. Yoneda, and F. Inada, *Effects of Water Chemistry and Fluid Dynamics on Wall Thinning Behavior <Part 2> - Evaluation of Influential Factors by Flow Accelerated Corrosion Tests*, CRIEPI-report: Q09028, Japan (2009).
12. K. Fujiwara and M. Domae, *Effects of Water Chemistry on Wall Thinning Behavior in Single - Phase and Two - Phase Flow*, CRIEPI-report: Q11025 (2012).
13. R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (2008).
14. A. Liaw, Package 'randomForest', https://cran.r-project.org/web/packages/randomForest/randomForest.pdf (2018).
15. K. M. Mullen, Package 'minpack.lm', https://cran.r-project.org/web/packages/minpack.lm/minpack.lm.pdf (2016).