Research Paper

Applicability of Statistics-based forward Osmosis Module Models

Jongmin Jeon¹ · Noori Kim¹ · Joon Young Choi² · Suhan Kim^{1,*}

¹Department of Civil Engineering, Pukyong National University

²Institute of Technology, Hyorim industries, Inc.

(Received October 14, 2019; Revised October 31, 2019; Accepted November 1, 2019)

Objectives : Since full-scale forward osmosis (FO) process designs are based on the arrangement of membrane FO modules, the FO module model to predict the performance of the modules is necessary. Through a previous study, a FO module model based on the fundamentals of FO process accompanied by a fitting process to reflect the module geometry was developed, but it has a drawback of long calculation time. Thus, a statistics-based model was developed to shorten the calculation time.

Methods : The drawback of statistical-based models is a ‘overfitting problem’, which means the accuracy of the model becomes poor if it is applied to the data not used to train the model. In this study, four regression models without considering the fundamentals of FO process were developed and trained with the datasets of 10, 30, 50, and 70 randomly selected from the total 116 datasets which consisted of the flow, pressure, and concentration of influent and effluent from the FO module. The model prediction errors with the datasets used and not used to develop the model (train data and test data, respectively) were compared to estimate overfitting tendency.

Results and Discussion : When 70 data were selected as test data, the model prediction error with test data decreases from 10.16 to 8.43% as the model accuracy increases. These errors are similar to or less than the error (10.13%) of the theoretical model. When the most accurate model is tested, the prediction error with test data decreases from 12.99 to 8.43% as the number of train data increases from 10 to 70. These errors become close to the error range with train data (6.38 to 7.72%).

Conclusions : The regression analyses revealed that the overfitting problem decreased as the number of train data and the accuracy of the model was increased. Especially, the accuracy of the statistics-based regression model is higher than that of the theoretical model and the overfitting problem was negligible. Thus, the statistics-based FO module model can be applied to the design of full-scale FO processes.

Key Words : FO Module Model, Statistics, Regression Model, Overfitting

* Corresponding author E-mail: suhankim@pknu.ac.kr Tel: 051-629-6065 Fax: 051-629-6063

연구논문

통계기반 정삼투 모듈 모델의 적용성 검토

전종민¹ · 김누리¹ · 최준영² · 김수한^{1,*}

¹부경대학교 토목공학과

²(주)효림산업

목적: 실규모 정삼투(FO) 공정 설계는 막모듈의 배열을 통해 이루어지므로, 막모듈의 성능을 예측할 수 있는 FO 모듈 모델의 역할은 필수적이다. 기존 연구를 통해 정확도가 90%가 넘는 FO 모듈 모델을 이론적 배경 및 모듈 형상을 반영하는 fitting process를 거쳐 만들었지만 계산 속도가 최대 수백초까지 걸리는 단점이 있어서, 이를 개선하기 위한 통계 기반 FO 모듈 모델을 개발하고 적용성을 평가하였다.

방법: FO 공정 관련 이론적 배경을 전혀 고려하지 않고 모델 개발에 사용된 데이터(train data)의 예측력을 최대한 하도록 하는 통계 기반 모델의 단점은 외부 데이터(test data)에 대한 예측력이 떨어지는 과적합이므로, 본 연구에서는 이론적 배경을 무시하고 만들어진 4개의 회귀 모델에 대해 예측정확도를 분석한 후, FO 공정의 유입수와 유출수의 유량, 압력, 농도값으로 각각 구성된 116개의 데이터셋에서 train data의 수를 10개, 30개, 50개, 70개로 하였을 때 train data와 test data의 오차를 비교하는 방법으로 과적합도를 분석하였다.

결과 및 토의: 회귀 모델의 예측력이 증가함에 따라 test data의 오차는 10.16%에서 8.43%로 감소했는데(train data의 수가 70인 경우), 이론적 모델의 경우(10.13%)와 비슷하거나 낮은 값을 보였다. 예측력이 가장 좋은 모델을 사용한 경우, train data 수가 10에서 70까지 증가하는 동안 test data의 오차가 12.99%에서 8.43%로 줄어들었고, train data의 오차(6.38-7.72%)와 비슷해지는 경향을 보였다.

결론: 통계 기반 FO 모듈 모델의 적용성을 평가했을 때, train data의 수가 많고 모델의 예측력이 높을수록 과적합도가 낮다는 결론을 얻었다. 특히, 기존 연구에서 도출된 이론 기반 모델과 비교했을 때도 test data의 오차율이 낮았기 때문에, 과적합도로 인한 적용성 문제는 전혀 없다고 볼 수 있다.

주제어: 정삼투 모듈 모델, 통계, 회귀모델, 과적합

1. 서론

정삼투(forward osmosis, FO) 공정은 분리막을 사이에 두고 고농도 용액과 저농도 용액이 흐를 때, 삼투압을 구동력으로 하여 저농도 용액의 물이 고농도 용액으로 투과되는 공정으로 역삼투(reverse osmosis, RO) 공정에 비해 상대적으로 작은 압력이 필요하다. 그러므로, FO 공정은 저에너지로 운전이 가능하다는 장점이 있다.¹⁻⁴⁾ 이러한 장점 덕분에 FO 공정은 하수처리와 담수화 같은 수처리 분야나 식품 공정 등과 같은 다양한 분야에 적용하기 위한 연구가 진행되고 있다.⁵⁻⁸⁾

RO 공정과 마찬가지로 실규모 적용을 위한 FO 공정의 최소 단위는 모듈이고, 이들 모듈을 직렬, 병렬로 배열하여 시스템을 설계하게 된다. RO 공정의 경우 막 제조사에서 제공된 설계프로그램을 통해 모듈을 배열했을 때의 운전

결과 값을 예측하여 설계에 활용한다.⁹⁾ FO 공정의 경우 모듈 실험과 모델링 연구가 활발히 진행되고 있지만,¹⁰⁻¹³⁾ RO 공정과는 달리 막모듈 제조사가 많지 않고 공개된 설계 프로그램도 없는 실정이다. FO 공정을 설계하기 위해선 RO 공정과 마찬가지로 모듈을 직렬, 병렬로 배열했을 때의 유량, 압력 변동을 예측해야 하고, 이를 위해 FO 모듈의 성능을 예측할 수 있는 모델 개발이 필수적이다.

FO 모듈 성능 예측 모델(이하 FO 모듈 모델) 개발을 위해서는 FO 막에서 일어나는 내부농도분극(internal concentration polarization, ICP)과 외부농도분극(external concentration polarization, ECP) 현상, 삼투압과 외부압력차로 인해 발생하는 물 투과, 이온 농도차로 인해 발생하는 이온 투과 등의 물질 전달 현상을 모듈 크기로 확대하는 과정이 필요하고, 이는 사전 연구 결과를 담은 문헌에 잘 기술되어 있다.¹⁴⁾ 사전 연구에서는 모델의 정확성을 높이기 위해 모듈 실험 결과 값과 모델

값의 오차를 줄이기 위한 fitting parameter를 도입하였고, 30개 이상의 실험 데이터를 확보하면 정확도 90% 이상의 FO 모듈 모델을 개발할 수 있다. 이 모델을 기반으로 모듈을 직렬, 병렬 배열했을 때의 물질평형을 고려하면 RO 공정 설계 프로그램과 같은 기능을 가지는 FO 공정 설계 프로그램을 만들어 설계할 수 있다.

기 개발된 FO 모듈 모델은 FO 막 내외부에서 일어나는 현상, 각종 물질전달 및 물질평형식 등 20여개의 비선형 연립방정식을 풀이하는 과정에서 Microsoft Excel의 Solver 기능 중 Evolutionary 알고리즘을 적용했는데, 이 알고리즘은 복잡한 연립방정식의 해를 정확하게 구할 수 있다는 장점이 있지만 계산 시간이 상대적으로 오래 걸린다는 단점이 있다. 따라서, 이 FO 모듈 모델과 연계된 FO 공정 설계 프로그램의 계산 시간도 오래 걸려(컴퓨터 속도에 따라 수십에서 수백 초), 계산 시간 1초 내외의 RO 공정 설계 프로그램과 비교하면 사용성이 떨어진다.

FO 공정 설계 프로그램의 계산 시간 단축을 위한 방법들 중에서는 FO 공정의 이론, 물질전달, 물질평형 등 이론적 수식들을 계산하는 알고리즘을 최적화하여 속도를 향상시키는 방법과, 이론적 수식들의 도움 없이 실험 데이터를 기반으로 모델을 만드는 방법, 두 가지가 있다. 실험 데이터를 기반으로 모델을 만드는 방법의 경우 재해예방, 수처리 관련 분석 등 다양한 분야에서 활용되어지고 있다.¹⁵⁻¹⁹⁾ 본 연구에서는 두 번째 방법, 즉, 계산 시간을 차지하는 가장 큰 이유가 되는 이론적 수식들을 배제하고 실험값을 활용해서 통계적인 방법으로 FO 모듈 모델을 개발하여 이론적 수식 기반으로 개발된 기존 모델과의 예측 성능을 비교하고자 한다. 이론에 기반하지 않고 통계적인 방법으로 실험값을 잘 예측하도록 만들어진 모델의 가장 큰 단점은 모델 개발을 위한 학습(train)에 사용된 데이터는 잘 예측하나, 그 외의 데이터에 대한 예측 정확도는 높지 않은 과적합(overfitting) 문제이다. 본 연구에서는 과적합 관점에서 통계기반 모델에 대한 분석을 실시하여 FO 모듈 모델 개발 시 통계를 기반으로 하는 모델 적용이 가능한지의 여부를 고찰하고자 한다.

2. 연구방법

2.1. 모델 구축

모델 구축을 위해 사용된 데이터는 기존 문헌¹⁴⁾에 정리된 FO 모듈 실험 데이터 셋 116개이다. 각 데이터 셋은 Fig.1에서와 같이 모듈에 유입되는 유도용액(draw solution)과 원수(feed solution), 모듈에서 유출되는 희석 유도용액(diluted draw solution)과 농축수(concentration)의 유량(Q), 압력(P), 이온농도(C) 데이터, 즉 총 12개(4개의 흐름별로 3개 데이터씩)의 데이터로 구성되어 있다. Fig.1에서 ds, fs, dd, c는 각각

Fig. 1. Input and output flows of FO module.

유도용액, 원수, 희석 유도용액, 농축수를 나타낸다. 모듈에 입력되는 6개의 입력값(원수와 유도용액 데이터)을 이용해 6개의 결과값(농축수와 희석 유도용액 데이터)을 출력하도록 모델을 구성하게 되는데, 본 연구에서는 6개의 입력값을 이용하여 정삼투 수투와 플럭스(J_w)를 구하는 모델에 대해서 다루고자 한다.

막모듈의 면적을 A라고 하면, 수투와 플럭스는 식 1과 같이 표현된다.

$$J_w A = Q_{dd} - Q_d = Q_f - Q_c \quad (1)$$

식 1과 116개의 데이터 셋을 이용하여 수투와 플럭스의 실측값을 계산한 후, 이를 예측하는 회귀모델식을 다양하게 구축하여 예측 정확도를 분석하였다. 회귀모델식은 6개의 입력값을 독립변수로 하는 1차 다항식 형태로부터, 각 독립변수의 제곱 및 상호곱을 추가 입력변수로 하는 2차 다항식 형태까지 설정하여 예측 정확도를 향상시켰고, 각 입력변수의 p-value를 분석하여 0.05보다 높은 경우에는 모델에서 배제시켜가면서 모델의 과적합 정도를 최소화시켰다. 모델 구축 시 116개의 데이터 셋을 모두 사용하였고, 예측 정확도는 식 2와 같이 normalized root mean square of errors (NRMSE)로 표현하였다.

$$NRMSE(\%) = \frac{RMSE}{mean} \times 100 \quad (2)$$

식 2에서 RMSE는 오차제곱합의 제곱근을, mean은 각 독립변수의 평균값을 의미한다.

2.2. 모델의 과적합도 분석

서론에서 언급했듯이 통계기반 회귀모델은 모델 개발에 사용된 데이터(train data)는 잘 예측할 수 있지만 그 외의 데이터(test data)는 잘 예측하지 못하는 과적합 문제를 갖고 있다. 따라서, FO 모듈 모델을 회귀분석을 이용해서 만들었을 때 과적합 문제가 어떻게 발생하는지 분석하는 것이 중요하다. 본 연구에서는 2.1절에서 언급된 116개의 데이터 셋으로부터 train data를 각각 10, 30, 50, 70개 셋을 무작위로 추출하여 만들어진 회귀 모델의 train data와 test data의 오차(NRMSE)를 구하여 비교하고자 한다. 이때, train data가 어떤 조합으로 추출되는 지에 따라 오차가 달

Table 1. Summary of the regression analysis using various equation types.

	Model 1		Model 2		Model 3		Model 4	
	a_n	P-value	b_n	P-value	c_n	P-value	d_n	P-value
Q_f	0.05916	2.82×10^{-2}	0.07084	1.23×10^{-3}	0.0632	2.98×10^{-4}	0.07794	1.14×10^{-5}
P_f	6.48301	2.13×10^{-19}	6.49165	3.12×10^{-32}	6.4535	8.93×10^{-41}	5.19188	4.43×10^{-18}
C_f	-0.07381	1.10×10^{-15}	-0.07095	2.20×10^{-17}	-0.0698	5.71×10^{-21}	-0.06809	3.81×10^{-21}
Q_d	0.05197	7.62×10^{-1}	-	-	-	-	-	-
P_d	-0.09091	9.79×10^{-1}	-	-	-	-	-	-
C_d	0.01533	5.27×10^{-47}	0.01540	3.67×10^{-49}	0.0265	4.43×10^{-30}	0.02674	1.68×10^{-31}
C_d^2	-	-	-	-	-9.9×10^{-6}	4.62×10^{-9}	-1.2×10^{-5}	4.60×10^{-11}
$P_f C_d$	-	-	-	-	-	-	0.00275	2.25×10^{-3}
intercept	2.13688	6.14×10^{-3}	2.14520	3.82×10^{-3}	0	-	0	-
NRMSE	9.15%		9.19%		8.17%		7.69%	

라질 수 있기 때문에 이 과정을 100번씩 반복하여 그 결과를 분석하였다. 반복 과정은 Microsoft Excel의 Visual Basic for Application (VBA)를 사용해서 코딩하여 시행하였다.

3. 결과 및 고찰

3.1. 회귀모델 구축

이론적 배경을 배제하고 순수 통계 기반으로 FO 모듈 모델을 만들기 위해, 6개의 입력값(원수와 유도용액의 압력, 유량, 이온농도 값)을 모두 입력변수로 하는 1차 다항식 형태의 수식(식 3)부터 회귀 분석을 실시하였다. 일반적으로 통계 기반 모델을 만들기 위해서는 1차 다항식 형태의 수식부터 시작하여 입력 값들을 제거하거나 서로 곱해주는 등의 과정을 거쳐 오차를 줄여나간다. 아래 언급된 Model 1-4는 위와 같은 과정을 진행하며 만든 모델이다.

$$J_{w, \text{mod1}} = a_0 + a_1 Q_f + a_2 P_f + a_3 C_f + a_4 Q_d + a_5 P_d + a_6 C_d \quad (3)$$

식 (3)에서 $a_n(n = 0-6)$ 은 상수 계수로 모델값과 실측값의 오차를 최소화시키기 위해 최소자승법을 통해 구하게 된다. 모델의 정확도를 더 높이기 위해서는 각 입력변수의 최대 차수를 증가시키거나, 로그나 지수함수 등 비선형 항을 추가시킬 수 있는데, 수식의 복잡성이 올라가서 과적합도를 증가시킬 위험이 있기 때문에 최대 차수를 2차까지로 제한하는 것을 원칙으로 삼았고, 이론에 근거한 기존 FO module model(이하 이론적 모델)의 NRMSE 범위인 10.01%¹⁴⁾보다 회귀모델의 오차가 큰 경우 회귀식의 최대 차수를 3차 이상으로 높이기로 하였다.

식 3의 회귀 모델을 Model 1이라고 했을 때의 회귀 분석 결과는 Table 1에 정리되어 있다. 모델 예측결과의 정확도

를 나타내는 NRMSE 값은 9.15%로 이미 이론적 모델의 정확성을 약간 상회하는 결과를 보이고 있다. 다만, Q_d 와 P_d 의 계수에 대한 p-value(계수가 0이 될 확률을 의미하며, 통계학적으로 0.05가 넘지 않아야 회귀모델의 계수로서 의미가 있다고 보는 경우가 일반적임)가 0.05를 훨씬 넘어가기 때문에, 이들 변수를 제외하고 Model 2를 만들면 식 (4)와 같다($b_n(n = 0-4)$ 은 상수 계수).

$$J_{w, \text{mod2}} = b_0 + b_1 Q_f + b_2 P_f + b_3 C_f + b_4 C_d \quad (4)$$

변수의 개수가 줄어 자유도가 줄어들었기 때문에 NRMSE 값이 0.04% 늘었지만, 모델의 정확도 차이는 거의 없다고 볼 수 있다. 왜냐하면, Model 1에서 회귀모델에 큰 영향을 미치지 않는 변수(p-value가 높은 변수)를 제거하여 Model 2를 만들었기 때문이다. 본 절의 서두에서 언급되었듯이, 회귀모델의 최대 차수를 2차까지 증가시켜보기로 했기 때문에 Model 2에서 남은 변수의 제공으로 만들어진 새로운 변수항들(즉, Q_f^2 , P_f^2 , C_f^2 , C_d^2)을 추가하여 모델을 만든 후, p-value가 0.05가 넘는 값을 제외하고 만든 Model 3은 식 (5)와 같다($c_n(n = 0-5)$ 은 상수 계수).

$$J_{w, \text{mod3}} = c_0 + c_1 Q_f + c_2 P_f + c_3 C_f + c_4 C_d + c_5 C_d^2 \quad (5)$$

Model 3의 NRMSE 값은 8.17%로 제공항 도입에 의해 약 1%의 정확도 개선이 이루어졌다. 각 1차 항의 곱도 2차 항에 해당되기 때문에, 이들을 변수로 추가하여 p-value가 0.05보다 높은 값들을 제외하면 Model 4와 같은 식 형태가 된다(식 (6)에서 $d_n(n = 0-6)$ 은 상수 계수).

$$J_{w, \text{mod4}} = d_0 + d_1 Q_f + d_2 P_f + d_3 C_f + d_4 C_d + d_5 C_d^2 + d_6 P_f C_d \quad (6)$$

Fig. 2. Experimental flux and model flux according to the regression model type.

Table 2. NRMSE values of regression models and theoretical model.

No. of train data	NRMSE of Model 1		NRMSE of Model 2		NRMSE of Model 3		NRMSE of Model 4		NRMSE of Theoretical Model	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
10	6.50	30.30	7.00	17.19	7.40	16.58	6.38	12.99	9.28	12.30
30	8.06	11.49	8.56	10.54	7.60	9.36	7.71	9.44	9.86	10.32
50	8.69	10.40	8.85	10.13	7.90	8.93	7.47	8.79	10.05	10.12
70	8.91	10.16	8.98	9.91	8.07	8.73	7.72	8.43	9.98	10.13

Model 4의 NRMSE 값은 7.69%로 Model 3과 비교했을 때 0.5%의 유의미한 정확도 개선이 이루어졌다. Fig. 2는 모델 플럭스와 실험 플럭스 비교를 통해 모델의 개선 정도를 나타낸 것이다. 각 모델의 정확도를 나타내는 NRMSE 값만 보았을 때는, 위에 기술된 단순한 절차를 통해 만들어진 통계 기반 회귀모델이 상대적으로 FO 공정의 작동 원리에 대한 이해 및 수치해석을 통해 만들어진 이론적 모델보다 정확도가 높다는 결론을 얻게 된다.

3.2. 회귀모델의 과적합 문제 분석

FO 공정의 작동원리 등 이론에 대한 지식을 배제하고 단순 절차로 만든 다변수 2차 회귀 모델이 이론적 모델의 정확도를 상회한다는 결론을 3.1절에서 얻었다. 그러나 서론에서 밝힌 바와 같이, 통계 기반 모델의 단점은 모델의 파라미터(즉, 독립변수 항의 계수값)를 결정하는 데 사용된

데이터(train data)가 아닌 외부 데이터(test data)를 적용할 경우 정확도가 떨어지는 과적합 문제이기 때문에, 본 절에서는 각 회귀모델의 과적합 문제를 분석하였다.

Table 2는 3.1절에서 도출된 4개의 회귀모델 및 기존 문헌¹⁴⁾의 이론적 모델에서 train data와 test data의 NRMSE 값을 비교한 것이다. 2.2절에 언급되었듯이, 총 116개의 데이터 셋으로부터 10개, 30개, 50개, 70개의 train data를 무작위로 추출하여 Model 1-4의 파라미터 값을 최소자승값을 통해 얻은 후, train data와 test data의 NRMSE 값을 찾는 과정을 100회 반복한 후 그 평균값을 Table 2의 표에 나타낸 것이다.

Test data의 개수가 50개 이하인 경우 회귀 분석으로 만들어진 모델들의 train data와 test data의 차이는 1.03-23.80% 범위였고, 이론적 모델의 경우 그 차이는 0.07-3.02%였다. 이론에 근거한 모델의 경우, train data의 범위를 벗어나더라도 동일한

Fig. 3. NRMSE values of test and train data according to the regression model type.

이론의 적용을 받기 때문에 test data의 오차가 train data의 경우보다 크지 않게 되는 것이다. Test data의 개수가 10개인 경우는 데이터 수가 너무 적기 때문에 train data를 통해 만들어진 회귀모델들의 외부 데이터 적용성이 현저히 떨어지는, 과적합 사례를 명백하게 보여주고 있다. Table 2는 train data의 개수가 증가할수록, 회귀 모델의 정확도가 증가할수록 과적합도가 줄어드는 결과를 잘 보여준다.

Table 2에 제시된 NRMSE 값들은 100개의 회귀 분석을 통해 얻은 평균값이기 때문에 train data와 test data의 차이가 과소평가될 수 있다. 따라서, Fig. 3과 4의 전체 데이터를 보면 회귀모델의 정확도와 test data의 개수가 과적합도에 미치는 영향을 좀 더 구체적으로 볼 수 있다.

Fig. 3은 train data의 개수가 30개인 경우 각 회귀모델에 따른 train data와 test data의 NRMSE 값을 나타낸 것이다. 각 그래프에서 원점과 우측 상단 끝점을 연결한 기울기가 +1인 대각선의 의미는 train data와 test data의 NRMSE 값이 완전히 일치한다는 의미이다. 즉, 데이터 포인트가 대각선에 가까이 위치할수록 과적합도가 줄어든다. 데이터 포인트가 대각선의 상부에 위치하면 test data의 오차가 train data의 오차보다 크다는 의미이기 때문에 과적합도가 높은 것이고, 대각선의 하부에 위치하면 test data의 오차가 모델 파라미터 도출에 사용된 train data보다 낮은, 확률 상으로 드물게 일어나는 현상이 나타난 것이다. Fig. 3의 (a)-(d) 모

두 test data의 오차가 train data의 오차보다 작은 경우가 적은 빈도로 발생하고 있음을 보여 준다. 한편, 회귀 모델의 정확도가 높아질수록 데이터 포인트가 대각선에 가깝게 모여드는 현상이 Fig. 3을 통해 명백하게 나타난다. 다만 Model 1과 2를 비교했을 때, Model 1의 정확도가 다소 높음에도 불구하고 오히려 데이터 포인트가 대각선에 모이는 경향이 Model 2의 경우보다 약한 것을 볼 수 있다. 이는 3.1절에서 토의했듯이, Model 1에 사용된 독립변수들 중 일부(Q_d 와 P_d)가 모델의 종속변수 값을 예측하는 데 영향력이 작기 때문에, 이들로 인해 모델 최적화에 사용되지 않은 외부 데이터를 예측하는 데 있어 오차를 증가시킬 수 있는 요소가 되기 때문이다.

Fig. 4는 예측 정확도가 가장 높은 Model 4에 대해서 train data 수가 10개에서 70개로 증가하는 동안 과적합도가 감소하는 경향을 잘 보여주고 있다. 50개 이상의 데이터(train data)를 회귀 모델 파라미터를 결정하는 데 사용하면, 외부 데이터(test data)를 예측하는 데 심각한 오차가 발생하지 않음을 알 수 있다. Fig. 4의 (c)와 (d)에 의하면 train data의 최대 오차는 10% 미만이고 test data의 최대 오차도 약 12%, 10%로 FO 공정 설계를 위한 모듈 모델로 활용하는 데 큰 무리가 없다. 즉, 이론적 배경 없이 회귀 모델을 만들어도 모델 파라미터 결정을 하는 데 필요한 실험 데이터의 양과 질에 따라 양호한 성능을 도출할 수 있

Fig. 4. NRMSE values of test and train data according to the number of train data.

고, 무엇보다 통계적 모델 개발 시 가장 우려되는 과적합 문제도 크게 심각하지 않음을 알 수 있다.

4. 결론

본 연구에서는 FO 공정을 설계하기 위한 프로그램을 개발하기 위한 핵심 요소인 FO 모듈의 성능을 예측하는 모델을 개발하기 위한 두 가지 방법, 이론에 근거한 수치 해석(이론적 모델), 통계에 기반한 회귀 분석(회귀 모델)에 대해서 비교하였다. 실용적인 관점에서 보면, 이론적 모델의 계산 시간이 최대 수백 초 정도로 오래 걸리기 때문에 이를 개선하기 위한 모델을 찾는 과정에서 회귀 모델의 적용성을 검토한 것이다.

회귀 모델의 경우 실측값을 예측하기 위한 오차를 최소화하기 위해 파라미터 값들을 최적화하기 때문에, 모델을 구축하기 위한 데이터에 대해서는 높은 예측력을 보이지만 외부 데이터에 대해서는 예측력이 떨어질 수 있는, 즉 과적합 위험도가 있다. 본 연구에서는 회귀 모델 개발 시 과적합이 어느 정도로 나타나고, 이를 줄이기 위해서 어떤 점을 주의해야 하는 지에 대해서 분석하였다. 과적합은 모델 개발에 사용된 데이터의 수가 늘어날수록, 모델에 사용된 독립변수의 영향력(즉, 낮은 p-value)이 클수록, 모델의 정확도가 높을수록 낮게 나타났다. 본 연구에 사용된 116개 데

이터 셋을 사용해 분석한 결과, 모델 개발에 사용된 데이터 수를 50개 이상으로 하고, 독립변수 항의 차수를 2차 항까지 두면, 이론적 모델과 비교했을 때 정확도가 더 높으면서 과적합도도 낮은 모델을 개발할 수 있었다.

예측 모델을 개발하는 모든 상황에서 이런 결과를 얻기는 어려운 것이라 보이지만, 본 연구의 결과는 이론적 배경 없이도 제대로 작동하는 모델을 개발하는 것이 적어도 FO 공정에서는 가능하다는 결론을 얻었다. 본 연구에서 설명한 회귀 모델을 만들어가는 절차는 입력 값으로부터 결과 값을 얻는 상관관계(수식)를 만들고 그 예측결과와 실측값과의 오차를 최소화시키는 최적화 과정을 거치는 것이다. 이론에 근거하여 수치 모델을 유도하고 이를 풀어내는 전통적인 모델이 실용성(즉, 높은 예측 능력 및 낮은 과적합도) 측면에서만 보았을 때는 통계기법으로 개발된 모델에 의해 대체될 수도 있다는 결론이 된다. 향후 통계기법으로 개발된 모델을 FO 공정 설계프로그램에 탑재한 후, 이론적 모델과 비교분석을 통해 활용성을 평가할 예정이다.

Acknowledgement

본 결과물은 환경부의 재원으로 한국환경산업기술원의 플랜트 연구 사업의 지원을 받아 연구되었습니다(1485016285).

References

1. B. C. Kim, C. H. Boo, S. Y. Lee, S. K. Hong, Evaluation of forward osmosis (FO) membrane performances in a non-pressurized membrane system, *Journal of Korean Society on Water Environment*, 28(2), 292-299(2012).
2. J. Jeon, B. Park, Y. Yoon, S. Kim, An optimal design approach of forward osmosis and reverse osmosis hybrid process for seawater desalination, *Desalin. Water Treat.*, 57(55), 26612-26620(2016).
3. M. Elimelech, W. A. Phillip, The future of seawater desalination: energy, technology, and the environment, *Science*, 333(6043), 712 - 717(2011).
4. L. Chekli, S. Phuntsho, J. E. Kim, J. Kim, J. Y. Choi, J. S. Choi, S. Kim, J. H. Kim, S. Hong, J. Sohn, H. K. Shon, A comprehensive review of hybrid forward osmosis systems: Performance, applications and future prospects, *J. Membrane Sci.*, 497, 430 - 449(2016).
5. T. Y. Cath, A. E. Childress, M. Elimelech, Forward osmosis: principles, applications, and recent developments, *J. Membrane Sci.*, 281(1-2), 70 - 87(2006).
6. S. Kim, S. Paudel, G. T. Seo, Forward osmosis membrane filtration for microalgae harvesting cultivated in sewage effluent, *Environ. Eng. Res.*, 20(1), 99-104(2015).
7. A. Achilli, T. Y. Cath, E. A. Marchand, A. E. Childress, The forward osmosis membrane bioreactor: a low fouling alternative to MBR processes, *Desalination*, 239(1-3), 10-21(2009).
8. J. Jeon, J. Y. Choi, J. Sohn, S. Kim, Performance analysis of a spiral wound forward osmosis membrane module, *J. Korea Soc. Environ. Eng.*, 40(12), 481-486(2018).
9. J. Jung, N. Kim, J. Jeon, S. Kim, Reliability Assessment of Reverse Osmosis System Projection Programs, *J. Korea Soc. Environ. Eng.*, 41(1), 42-47(2019).
10. J. E. Kim, G. Blandin, S. Phuntsho, A. Verliefde, P. Le-Clech, H. K. Shon, Practical considerations for operability of an 8" spiral wound forward osmosis module: Hydrodynamics, fouling behaviour and cleaning strategy, *Desalination*, 404, 249-258(2017).
11. J. R. McCutcheon, M. Elimelech, Modeling water flux in forward osmosis: implications for improved membrane design, *AIChE J.*, 53(7), 1736-1744(2007).
12. J. E. Kim, S. Phuntsho, S. M. Ali, J. Y. Choi, H. K. Shon, Forward osmosis membrane modular configurations for osmotic dilution of seawater by forward osmosis and reverse osmosis hybrid system, *Water Res.*, 128, 183 - 192(2018).
13. S. Benavides, W. A. Phillip, Water recovery and solute rejection in forward osmosis modules - Modeling and bench-scale experiments, *J. Membrane Sci.*, 505, 26 - 35(2016).
14. J. Jeon, J. Jung, S. Lee, J. Y. Choi, S. Kim, A simple modeling approach for a forward osmosis system with a spiral wound module, *Desalination*, 433, 120-131(2018).
15. S. Yoon, S.-R. Lee, S.-H. Kang, D.-W. Park, A statistical mobilization criterion for debris-flow, *Journal of the Korean Geotechnical Society*, 31(6), 59-69(2015).
16. T. Jeong, J. Ryoo, Chemical characteristics of precipitation by using the statistical analysis in Jeonbuk area, Korea, *J. Korea Soc. Environ. Eng.*, 24(4), 623-632(2002).
17. W.-B. Lee, J. Kim, Prediction of BaP and total PAH in soil from Pyr concentration using regression analysis, *J. Korea Soc. Environ. Eng.*, 39(3), 118-123(2017).
18. S. Subramani, R. C. Panda, Statistical regression and modeling analysis for reverse osmosis desalination process, *Desalination*, 351, 120-127(2014).
19. M. Li, Predictive modeling of a commercial spiral wound seawater reverse osmosis module, *Chem. Eng. Res. Des.*, 148, 440-450(2019).