



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

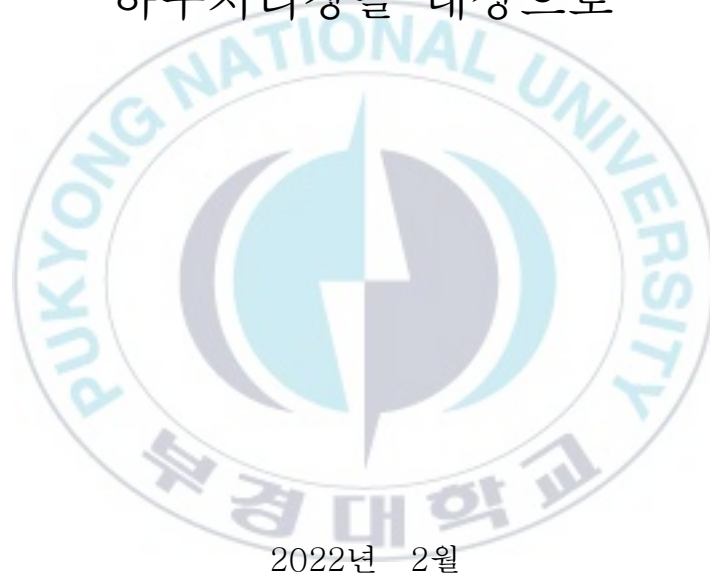
저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공 학 석 사 학 위 논 문

유역관리를 위한 심층신경망 수질 농도
예측 모델 개발에 관한 연구
- 하수처리장을 대상으로 -



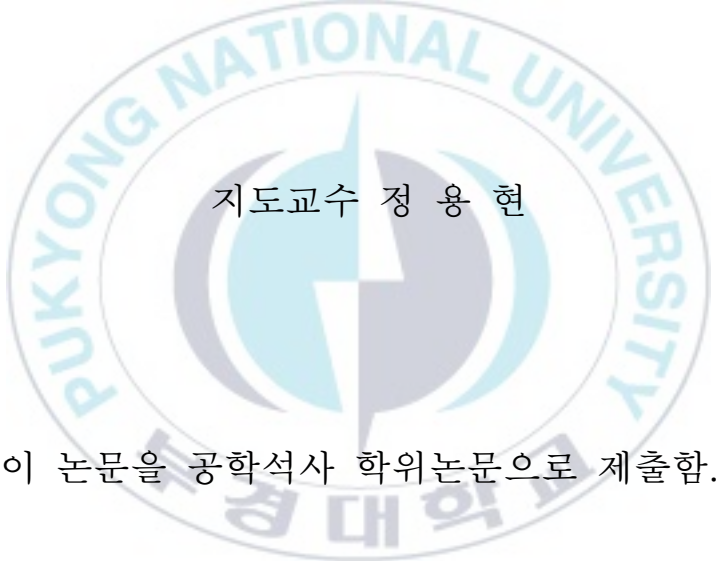
부 경 대 학 교 대 학 원

생 태 공 학 과

이 상 민

공 학 석 사 학 위 논 문

유역관리를 위한 심층신경망 수질 농도
예측 모델 개발에 관한 연구
- 하수처리장을 대상으로 -



지도교수 정 용 현

이 논문을 공학석사 학위논문으로 제출함.

2022년 2월

부 경 대 학 교 대 학 원

생 태 공 학 과

이 상 민

이상민의 공학석사 학위논문을 인준함.

2022년 2월 25일



위 원 장 공학박사 성 기 준 (인)

위 원 공학박사 이 석 모 (인)

위 원 공학박사 정 용 현 (인)

목 차

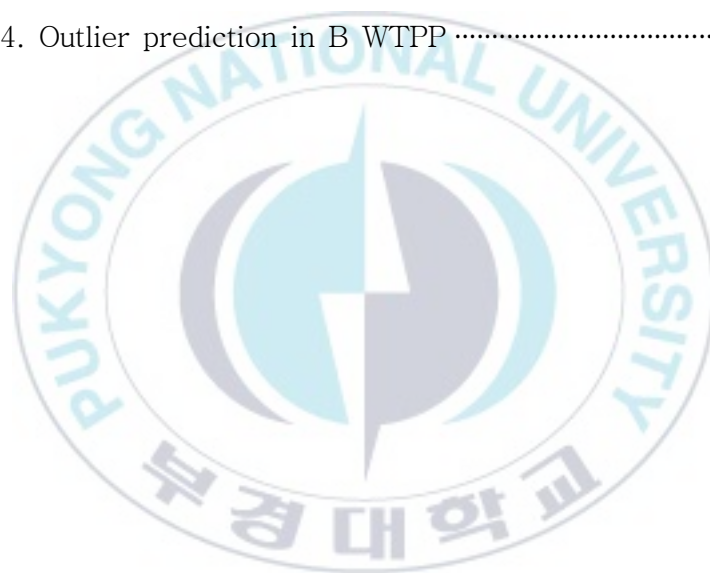
I. 서론	1
II. 이론적 배경	3
1. 하수처리장	3
2. ANN	5
3. DNN	7
4. 신경망을 적용한 선행연구 사례	9
가. 국내 사례	9
나. 해외 사례	12
다. 선행연구와의 차별성	14
III. 연구 방법	15
1. 연구 목적 및 범위	15
2. 자료수집	17
3. 데이터 전처리	17
가. 단계 선택법(Stepwise method)	18
나. 주성분분석(Principal component analysis)	19
다. 파생변수(Derived variable)	20
4. 상관 분석(Correlation analysis)	21
5. 예측 모델 개발 연구	22
가. DNN 모델 개발	22
나. 다중선형회귀 모델 개발	28

6. 모델 검정	29
IV. 결과 및 고찰	30
1. 상관 분석 결과	30
가. A 하수처리장	30
나. B 하수처리장	35
다. C 하수처리장	40
2. 유입수 BOD 예측 모델 개발	45
가. A 하수처리장	47
나. B 하수처리장	51
다. C 하수처리장	56
라. 강우 시 예측력 평가	61
마. 유입수 이상치 예측력 평가	65
V. 결론	69
참고문헌	72

표 목 차

Table 1. Dataset information in A WTPP	31
Table 2. Pearson correlation coefficient in A WTPP variables	33
Table 3. Pearson correlation coefficient adding derived variables of WTPP	34
Table 4. Dataset information in B WTPP	35
Table 5. Pearson correlation coefficient in A WTPP variables	38
Table 6. Pearson correlation coefficient adding derived variables of WTPP	39
Table 7. Dataset information in A WTPP	41
Table 8. Pearson correlation coefficient in C WTPP variables	43
Table 9. Pearson correlation coefficient adding derived variables of WTPP	44
Table 10. Structure and variables for DNN model in A, B, C WTPP to predict daily BOD	46
Table 11. Dataset without derived variable in A WTPP	47
Table 12. Dataset with derived variable in A WTPP	47
Table 13. Comparison models in A WTPP	48
Table 14. Dataset without derived variable in B WTPP	51
Table 15. Dataset with derived variable in B WTPP	51
Table 16. Comparison models in B WTPP	53

Table 17. Dataset without derived variable in C WTPP	56
Table 18. Dataset with derived variable in C WTPP	56
Table 19. Comparison models in C WTPP	58
Table 20. BOD prediction in A WTPP while precipitating	62
Table 21. BOD prediction in B WTPP while precipitating	63
Table 22. BOD prediction in C WTPP while precipitating	64
Table 23. Outlier prediction in A WTPP	67
Table 24. Outlier prediction in B WTPP	68



그 립 목 차

Fig 1. Process of $h()$ activation function	6
Fig 2. Structure of ANN	8
Fig 3. Research procedure	16
Fig 4. PCs orthogonal	20
Fig 5. Structure of DNN in study	23
Fig 6. rectified linear unit(ReLU)	24
Fig 7. Backpropagation algorithm	25
Fig 8. Gradient descent to find optimal weight	26
Fig 9. Training dataset with Batch size	27
Fig 10. Daily BOD in A WTPP	32
Fig 11. BOD Histogram in A WTPP	32
Fig 12. Daily BOD in B WTPP	36
Fig 13. BOD Histogram in A WTPP	36
Fig 14. Daily SS in B WTPP	38
Fig 15. Daily BOD in C WTPP	42
Fig 16. BOD Histogram in C WTPP	42
Fig 17. (a) Before dropout (b) After dropout	46
Fig 18. (a) Original prediction method (b) DNN without derived variable (c) MLR without derived variable(Stepwise method) (d) MLR without derived varibale(PCA) (e) DNN using	

derived variable (f) MLR using derived variable(Stepwise method) (g) MLR using derived variable(PCA)	49
Fig 19. (a) Original prediction method (b) DNN without derived variable (c) MLR without derived variable(Stepwise method) (d) MLR without derived variable(PCA) (e) DNN using derived variable (f) MLR using derived variable(Stepwise method) (g) MLR using derived variable(PCA)	54
Fig 20. (a) Original prediction method (b) DNN without derived variable (c) MLR without derived variable(Stepwise method) (d) MLR without derived variable(PCA) (e) DNN using derived variable (f) MLR using derived variable(Stepwise method) (g) MLR using derived variable(PCA)	59
Fig 21. Bias of predicted values in A WTPP while precipitating ...	62
Fig 22. Bias of predicted values in B WTPP while precipitating ...	63
Fig 23. Bias of predicted values in C WTPP while precipitating ...	64
Fig 24. Normal distribution with standard deviation	65
Fig 25. Bias of predicted outlier values in A WTPP	67
Fig 26. Bias of predicted outlier values in B WTPP	68

**A Study on Forecasting Model Development of Water Quality
Concentration for Watershed Management Using Deep Neural Network
- Focusing on the water treatment power plant -**

Sang Min Lee

*Department of Ecological Engineering, Graduate School,
Pukyong National University*

Abstract

This study developed the 'deep neural network(DNN)' model combined with several non-linear functions part of neural network algorithm to predict BOD in inflow sewage the day that can overcome the limitations of water treatment power plant(WTPP) applied in the past mechanistic and statistical methods and the complex inflow characteristics behavior. To estimate DNN model, this study developed multi linear regression(MLR) model using stepwise method and principal component analysis(PCA) each. And using derived variable as a days ago and two days ago water characteristics. It has related because of continuity of sewage. In result, DNN model used derived variable has the highest predictive power. Also, DNN and MLR model's prediction are increased when used derived variable. Through this study, derived variable as past water characteristics are verified to forecast BOD and make a basis to use DNN model in WTPP.

I. 서론

'유역(Watershed)'이란, 강우시 지면의 유출수가 하천이나 호소 등의 수계로 집수되는 영역으로 집수된 물은 유역 내 산림, 주거지 등을 거치면서 각 지역 오염물질을 포함하게된다(최와박, 2006). 유역 내 인구 증가, 산업 발달 및 도시화의 결과로 유역 내 다량의 오염물질이 수계로 유입되었고 하천 및 호소 수질오염은 국가적 사안이 되었다.

또한 강우시 비점오염원에서 토양 등에 축적되어 있던 오염물질이 일시에 배출되어 유입 부하에 상당한 영향을 미친다. 이러한 비점원오염을 포함한 오염부하량의 양적 증가와 집중화로 인해 환경용량을 초과한 오염물질이 수계에 유입되어 호소나 하천의 부영양화가 심화되고 있다(정 등, 2006).

기존의 유역 내 오염원 관리는 하수처리장의 방류수 수질기준 규제에 의한 점오염원 배출 저감에 주력하였다. 하지만 하수처리장의 분류식 관거에서 나온 우수나 합류식 관거의 월류수가 방류 수역으로 막대한 오염물질을 배출하고 있는 것으로 조사되었다(서 등, 2001).

하수처리장의 효율적이고 안정적인 운전에 대한 요구에 따라 하수 처리공정의 계측, 제어 및 자동화 분야가 90년대부터 활발히 연구되어왔다(최 등, 2007). 그 중 하수의 유입 성상은 하수처리장의 운전 효율을 결정하는 주요한 요인으로 하수처리장의 설계에서부터 운전 시 적절한 전략을 세우는 기초자료로 사용된다. 하수처리장 유입수의 수질 성상 중 생화학적 산소 요구량(biochemical oxygen demand: BOD)은 하수처리장 공정 관리·계획에 가장 중요한 변수로 BOD를 통해 유기물량을 알아내고 유기물 부하율을 조절해 수처리 환경을 조성한다. 하지만 BOD는 미생물이 소비한 산소량으로 당일 측정이 불가능하다. 따라서 하수처리장은 정확한 유기물 값 없이 수처리 후 방류한다. 그러나 현재 유입 하수의 성상을 실시간으로 예측할 수 있는 기술이 미비하여 하수처리장에서는 유입 하수의 평균 농도를 설계 및 운전 전략 결정에 이용하고 있다(정 등, 2006). 또한 하수처리

는 주로 현장 기술자의 경험에 의존하여 공정이 이루어진다(Yoon-Seok et al., 2003). 하지만 하수처리장 가동에 필요한 매개변수들을 전문가의 관습적, 경험적 기술로 처리하게 되면 많은 시간 소요와 함께 객관성의 부족으로 인해 하수처리 효율이 떨어질 수 있다(Maged et al., 2004). 이에 따라 전통 기계론적 방법이 하수처리장 관리에 적용되었으나 모델 비용이 크고 시간이 많이 들었기 때문에 접근성이 쉽지 않았다(Emrah et al., 2008). 또한 유입 성상의 거동은 하수 특성에 따라 변하기 때문에 복잡하며 불투명하여 해석하기 어렵다(김, 2006). 이처럼 유역관리를 위해서는 하수처리장의 처리공정 개선이 필요하고 이를 위해 과거 관측되었던 수질 데이터와 수학적 도구를 기초로한 예측 모델이 개발되어야 한다. 최근 몇 년간 인공지능 기반 관리 방법이 많은 환경 문제에 적용되었고 ‘인공신경망(artificial neural network: ANN)’ 모델이 하수처리장의 새로운 접근 방법으로 제시되었다. 그러므로 본 연구에서는 ANN 모델 중 여러 개의 비선형 함수로 결합된 ‘심층신경망(deep neural network: DNN)’ 모델 알고리즘이 과거 기계론적 방법 및 통계학적 방법으로 적용된 하수처리장의 한계를 극복할 수 있고(Mohamed et al., 1999) 복잡한 유입 성상 거동을 예측할 수 있다고 판단, 유입수의 당일 BOD 예측 DNN 모델을 개발하였다. 선형결합으로 이루어진 회귀 모델로는 예측하기 힘든 유입 성상의 복잡한 물리·화학적 거동을 비선형 함수간 결합으로 구성된 DNN 모델로 예측 가능하다고 판단했기 때문이다. 본 연구를 통해 하수처리장에서 보다 정확한 BOD를 사전에 예측하여 다양한 제어 전략을 제시하고 대응할 수 있으며 유기물 부하율을 적절히 조절하여 공정과정의 효율성을 높일 수 있다. 이는 하수공정과정에 대해 사후대처가 아닌 사전 감지 및 예방을 통한 법적 기준 초과 방류 문제를 해결할 수 있는 도구로 이용가능하고 유역관리에 필요한 점오염원 관리대책에 새로운 방법을 제시할 수 있다. 또한 강우 시, 하수처리장의 유입수 예측을 통해 BOD 변화에 따라 하수처리장으로 유입되지 않고 월류되는 비점오염원의 유기물량을 파악하여 유역관리를 위한 사전 대책의 기초자료로 활용할 수 있다.

II. 이론적 배경

1. 하수처리장

‘하수’란, 사람의 생활이나 경제활동으로 인해 액체성 또는 고체성의 물질이 섞여 오염된 물과 건물, 도로 그 밖의 시설물의 부지로부터 하수도로 유입되는 빗물·지하수를 말한다. ‘개인하수처리시설’은 건물·시설 등에서 발생하는 오수를 침전·분해 등의 방법으로 처리하는 시설을 말한다(하수도법, 2022). 우리나라 하수처리장은 전통적으로 표준 활성 슬러지공법을 통하여 수질오염의 직접적 오염물질인 유기물을 제거했으나 근래에는 질소와 인 제거를 위한 고도처리공법이 늘어나고 있다.

일반적인 하수처리 방법으로는 크게 생물학적, 물리적, 화학적 방법으로 나눌 수 있으며, 이 중 생물학적 처리방법인 생물반응조의 처리효율 및 경제적인 측면에서 가장 널리 이용되고 있다. 미생물의 종류에 따라 혐기성 처리법과 호기성 처리법으로 나눌 수 있는데 도시하수의 경우 호기성 처리법이 주를 이루고 있다(안 등, 2000).

F/M 비(Food-microorganism ratio)는 생물반응조 내 유기물량과 활성오니 미생물량의 비를 나타낸 유기물 부하율로 영양물과 미생물의 양적 균형을 의미하는 중요한 지표이다. 생물반응조의 질산화·탈질 반응의 주요 영향 인자는 BOD로, 만약 유입수의 BOD가 하수처리장 설계기준농도보다 낮으면 빈부하 현상으로 인해 생물 반응조 내 미생물의 번식과 성장이 억제된다. F/M 비는 전통적으로 하수처리장 운전변수로 이용되나 부하의 동특성 때문에 정상적인 공정운전이 어려워 근본적인 대책이 요구되어왔다(허 등, 1999).

높은 F/M 비 환경에선 미생물은 급속히 성장하기 때문에 응집에 필요한 물질을 충분히 분비하지 않고, 분산성 미생물의 증가에 따라 슬러지 용적 지수(Sludge volume index)가 높아져 슬러지 팽화현상(Bulking)이 일어난다(Ning Guo et al., 2017). 특히 분리막 공법(Membrane bio reactors)의 경우 높은 F/M 비는 멤브레인(Membrane) 오염도가 높아지고 질산화에 악영향을 끼치므로 안정적인 질산화와 질소 제거를 위해 적절한 F/M 비를 유지해야 한다(Yin Liu et al., 2012; Yi-Ju Wu et al., 2013). 반대로 낮은 F/M 비 조건에선 높은 활성슬러지와 빈부하에 의한 핀 플록(Pin floc) 발생으로 슬러지 침강성이 악화되었다(박 등, 2006).

이처럼 적정 F/M 비를 유지해야 안정적인 공정뿐만 아니라 공정 효율을 높일 수 있다(Rania Ahmed Hamza et al., 2018; An-jie Li et al., 2011). 그러기 위해선 최적화된 F/M 비의 관리가 이루어져야 하고(안 등, 2011), 따라서 유입수의 BOD를 정확히 파악해야 한다.

2. ANN

ANN은 1943년 워런 맥컬록(Warren McCulloch)와 월터 피츠(Walter Pitts)에 의해 처음 도입되었다. 신경망의 초기 모형은 선형적 모형에만 학습이 가능했고 ANN의 크기에 따라 요구되는 계산량이 기하급수적으로 증가한다는 문제점으로 각광받지 못하였다(Minsky M. and Papert S., 1969). 하지만 1980년대에 들어서 오차 역전파 알고리즘을 바탕으로 한 DNN을 통해 해결함으로써 다시 한 번 관심을 받기 시작했다(Werbos P., 1974). ANN은 충분한 데이터와 복잡성이 주어진다면 신경망은 일련의 설명 및 종속 변수간 관계를 모델링하도록 훈련된다. 이러한 이유로 신경망은 이해, 정의 및 정량화가 어려운 금융, 의학 공학 등 다양한 문제에 유용하게 적용되었다(Jalili et al., 2008).

ANN은 인간 뇌의 처리 과정을 모사한 모델로 뇌 속 뉴런(Neuron)의 역할을 하는 노드(Node)가 복잡한 연결망을 형성하고 있고 노드 간 연결된 연결선의 가중치(Weight)를 고려하여 더한 후 활성화 함수(activation function)에 전환해 출력하는 형태를 가지며(Fig. 1) 식 2.1로 표현할 수 있다. 여기서, w 는 가중치로 신호의 중요도로 해당 노드의 영향력을 결정 짓는 매개변수(Parameter)이다. b 는 편향(Bias)을 나타내는 매개변수로 노드로 입력된 가중 합에 더해주는 값으로 활성화 함수의 조건을 조절한다. 활성화 함수는 비선형함수를 사용하는데, 선형함수를 사용할 경우 층을 아무리 깊게 쌓아도 여러 층을 구성하는 이점을 살릴 수 없게 된다. 가령 식 2.2와 같이 $h(x) = ax$ 라는 활성화 함수로 네 개의 층을 쌓을 때, $b=a^4$ 로 치환하게 되면 $y(x) = bx$ 라는 은닉층이 없는 함수로 표현된다. 따라서 층을 쌓는 이점을 얻기 위해선 반드시 활성화 함수를 비선형함수로 사용해야 한다(사, 2017).

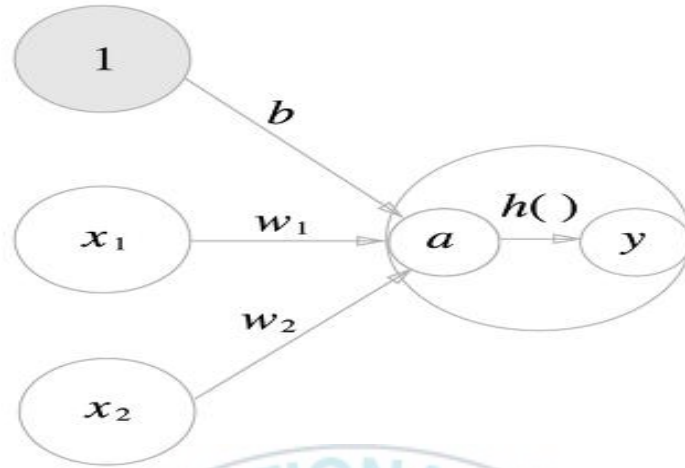


Fig. 1. Process of $h()$ activation function

$$y = h\left(\sum_{i=1}^n w_i x_i + b\right) \quad (2.1)$$

$$h(x) = ax \quad (2.2)$$

$$y(x) = h(h(h(h(x))))$$

$$y(x) = a^*a^*a^*a^*x$$

$$b = a^4$$

$$y(x) = bx$$

3. DNN

DNN은 은닉층을 2개 이상 쌓아 올린 신경망 모델로 구조는 Fig. 2와 같다. 입력층(Input layer)은 처음 데이터를 입력받는 노드들로 이루어진 첫 번째 층으로 각 설명변수에 대응하는 노드들로 구성되며 노드의 수는 설명변수의 개수와 같다. 이들 노드는 각각의 연결 강도를 통해 취합되어 은닉층(Hidden layer)으로 전달한다. 은닉층은 전달받은 가중 합들을 비선형함수로 처리하여 출력층(Output layer)이나 다음 은닉층으로 전달한다. 출력층은 출력변수에 대응되는 노드로서 사용자가 지정한 개수의 출력값을 생성한다. DNN은 여러 층의 비선형성을 통해 차원을 왜곡하고 복잡하지만 정교한 예측이 가능해진다. 하지만 층이 깊어질수록 계산량이 크고 복잡하기에 과대적합(Overfitting)이 발생하기 쉽고 데이터 패턴을 파악하기 위해 빅데이터를 기반으로 한다(Jack, 1996). 과거에는 은닉층이 많이 쌓여있는 구조는 최적화가 어려웠으나, 최근 GPU를 이용하거나 연산속도를 향상시키는 다양한 방법이 개발됨으로써 가능하게 되었다(안, 2016).

그러나 DNN은 예측값을 출력하는 과정에서 수많은 비선형성을 포함하기 때문에 모델 학습 과정 내 판단 근거를 알기 어렵다. 즉, 신경망 모델로 예측력을 높일 수 있지만 해석에 한계가 생긴다. 이러한 ‘블랙박스(Black Box)’ 문제는 신경망의 높은 정확도와 낮은 비용, 그리고 높은 효율성에도 산업에서 신경망 모델만을 이용하여 설비 진단을 하는 것을 지양하게 하는 원인이 된다(이 등, 2020).

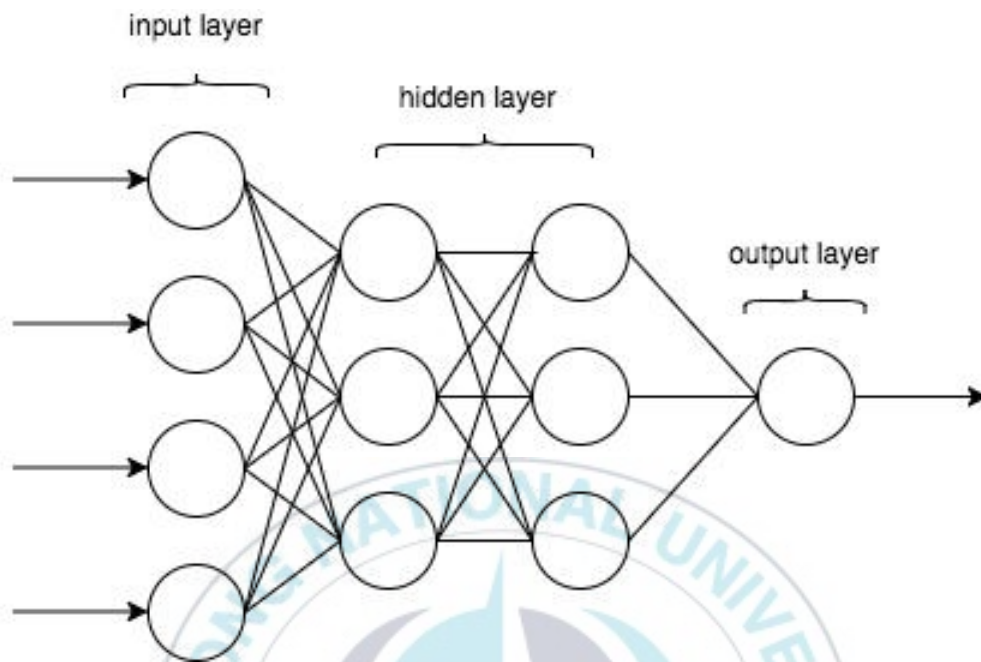


Fig. 2. Structure of ANN

4. 신경망을 적용한 선행연구 사례

가. 국내 사례

국내에서는 환경을 대상으로 신경망을 이용한 예측 모델이 계속해서 연구되어왔다. 하지만 하수처리장을 대상으로 신경망을 이용한 유입수 예측 모델 연구는 많지 않으며 장기간 데이터 기반의 예측 모델이 아닌 2년 이내의 데이터를 이용한 모델이기 때문에 연구 지점의 수질 성상을 파악하고 예측하는데 한계점이 존재했다.

최성원 등(2019)은 경기도 B 시의 월별 자료를 바탕으로 DNN을 활용하여 장래 생활폐기물 발생량을 예측하였다. 설명변수로 월별 인구수, 아동인구, 생산가능인구, 노령인구 데이터를 사용하였다. 60개의 월별 데이터 중 2013년 1월 ~ 2016년 12월까지 48개의 훈련 데이터로 모델을 학습하고 2017년 1월 ~ 2017년 12월까지 12개의 검증 데이터로 모델을 검증하였다. 연구 결과, 예측 기간이 길어질수록 월별 변동 폭이 감소하고 발생량이 특정 수치로 수렴하여 단기예측에 적합하다고 판단했다.

김영선(2019)은 K-NN(K-nearest neighbors), DT(Decision tree), 랜덤 포레스트(Random forest), GB(Gradient boosting), SVM(Support vector machine), 그리고 ANN 모델을 이용해 2016년~2018년 서울시 자치구 폐전자제품 발생량을 예측하는 연구를 수행하였다. 설명변수의 다중공선성을 해결하기 위해 상관관계 분석과 주성분분석을 사용하였고 RMSE와 결정계수를 통해 ANN 모델이 가장 우수한 것을 증명했다.

정형석 등(2006)은 UV 흡광도와 하수의 구성성분 농도와의 상관관계를 조사하였고 구성성분 농도를 효과적으로 추정하기 위해 1차 회귀분석과 ANN을 이용한 통계적 해석을 수행했다. 본 연구는 2005년 10월 1일 ~

11월 21일까지 50일 동안 대도시 A의 하수종말처리장 차집관거 후단의 유입 게이트에서 하수를 채취하여 446개의 유효한 분석 결과를 판단하였다. 이 중 350개를 1차 회귀분석 및 ANN 훈련에 사용하였고, 나머지 96개를 구성성분 농도 예측 가능성 평가에 사용하였다. 연구 결과, 300nm에서의 흡광도와 유량, 그리고 유입 시간을 이용한 ANN 모델이 모든 항목에 대해 흡광도법을 이용한 예측보다 향상된 측정 능력을 보였다.

이경훈 등(2002)은 광주광역시를 대상으로 폐기물 발생량에 영향을 주는 예측인자의 특성을 분석한 후 비선형 예측에 탁월한 예측력을 보이는 신경망 모델을 이용하여 2001년부터 2010년까지 연간 장래 폐기물량을 예측하였다. 예측인자로 인구수, 민수용탄 소비량, 소비자 물가지수, 공장 수, 공장인원수, 지방세 징수금액을 사용하였다. 본 연구는 신경망 내 은닉층을 수를 달리하여 예측 모델을 만들었고 18개 은닉층으로 구성된 모델의 결정계수가 우수하게 나타나 모델 구조에 따라 결과가 달라지는 것을 증명하였다.

김동균 등(2020)은 신경망 회귀 모델을 활용하여 선박의 가치평가 모델을 개발하고자 하였다. 중고 VLCC선을 대상으로 문헌 연구와 다중공선성을 고려하여 VIF가 10 이상인 데이터를 제외한 6가지 변수(VLCC freight, VLCC supply, VLCC scrap value, VLCC new building price, VLCC order book/supply ration, labor rate)를 선정하여 VLCC secondhand price를 평가하였다. 2000년 1월부터 2020년 8월까지 248개의 월별 데이터를 활용하여 VLCC Secondhand Price를 추정하기 위해 선형회귀, 신경망, 랜덤 포레스트 모델을 적용하였다. 전체 데이터 중 198개를 모델 학습에 사용하였으며 50개를 모델 검증에 사용하였다. 데이터의 과적합을 방지하기 위해 정규화를 실행했고 신경망 모델의 경우 은닉층을 각각 1층과 2층으로 구성하고 층 내 노드 수를 각각 6개, 12개로 구성해

총 4개 모델을 만들었다. 연구 결과 6개의 노드로 구성된 2개의 은닉층 모델의 예측력이 가장 뛰어나 신경망 모델의 우수성을 검증했고 신경망 모델 내 은닉층 및 노드 수에 따라 예측력이 달라지는 것을 증명하였다.

김동현 등(2020)은 수위 관측이 어려운 미계측 유역에 대하여 기존 수위 예측 대체 방법으로 DNN 모델을 적용하여 금정산 산지 습지를 대상으로 2017년 4월 20일부터 2018년 7월 31일까지 시간 단위로 측정된 데이터를 활용하였다. 2017년 4월 20일부터 2018년 2월 28일까지 데이터를 학습 데이터로 사용하였고 2018년 3월 1일부터 2018년 7월 31일까지 검증 데이터로 사용했다. 설명변수로 일 평균 강우량, 평균 기온, 최저 기온, 최고 기온, 최대 순간 풍속 기상자료를 활용하였다. 사용된 모델은 다중회귀, 주성분분석, ANN, DNN 총 네 개 모델을 만들어 RMSE와 NRMSE로 모델을 평가했다. 다중회귀모델의 경우 다중공선성 문제를 해결하기 위해 VIF값이 10 이상인 변수들을 제거했다. 연구 결과, DNN 모델이 가장 좋은 예측력을 가졌고 ANN 모델이 두 번째로 높은 예측력을 가지면서 신경망 모델의 예측력을 검증하였다.

이중원 등(2008)은 2005년 1월 2일부터 2006년 12월 31일까지 총 729일간 A 하수처리장 운전 데이터를 사용하여 DNN 모델에 적용하였다. 유입수의 TN, TP, 대장균 농도, 초침 유입수의 BOD, COD, SS 농도, 초침 유출수의 BOD, COD, SS 농도, 생물반응조의 활성슬러지 농도, 하루 전 유출수의 BOD, COD, SS, TN, TP 농도를 입력 변수로 사용하여 유출수의 BOD, COD, SS, TN 농도 값을 출력하였다. 729개의 일별 데이터 중 365개는 훈련 데이터로 나머지 364개는 훈련을 통해 생성된 모델의 테스트를 위해 사용되었다. 연구 결과, 예측값과 실측값의 결정계수가 BOD는 0.9713, COD는 0.8987, SS는 0.6914, 그리고 TN은 0.8435로 상당히 높은 정확성을 보여 비선형성을 지닌 DNN 모델의 효율성을 검증했다.

나. 해외 사례

해외 또한 환경을 대상으로 신경망을 이용한 연구가 계속해서 진행되어왔고 하수처리장을 대상으로 한 신경망 연구도 진행되고 있다. 하지만 해외 연구 역시 적은 수의 데이터를 적용한 모델이라는 한계를 가지고 있었다.

Jalili Ghazi Zaed et al.(2007)는 2004년 ~ 2007년까지 이란 Mashhad 지역의 주별 쓰레기 발생량 데이터를 이용하여 한주 뒤의 쓰레기 발생량을 예측하였다. 과거 12주까지의 주별 쓰레기 발생량 데이터와 예측 한 주 전의 쓰레기를 운반하는데 이용된 트럭 수를 설명변수로 사용하였다. 데이터를 정규화하고 은닉층 내 노드 수를 4 ~ 26개로 변화시켜 가장 성능이 좋은 모델을 선정하였다. 연구 결과 은닉층 내 10개의 노드로 구성된 모델의 결정계수가 0.75로 가장 높은 예측력을 가졌다. 이를 통해 신경망 내 노드 수에 따라 성능이 변화하는 것을 증명했다.

R Noori et al.(2009)는 이란 Tehran을 대상으로 한주 뒤의 쓰레기 발생량을 예측하기 위해 과거 12주 전까지의 주별 쓰레기 발생량 데이터와 예측 한 주 전의 쓰레기를 운반하는데 이용된 트럭 수를 설명변수로 사용하였다. 22개의 노드로 구성된 단일 ANN 모델을 개발하여 3개의 주성분으로 구성된 다중선형회귀 모델과의 예측값을 비교하였다. 연구 결과 ANN 모델의 결정계수가 0.837로 선형회귀 모델보다 우수한 예측력을 가진 것을 증명하였다.

Mohammad Ali Abdoli et al.(2011)는 2001년~2010년까지 120개의 월별 인구수, 최고 기온, 가구 수입 데이터를 설명변수로 사용하여 2011년~2032년까지의 장래 폐기물 발생량을 예측하였다. 전체 데이터 중 84개를 학습 데이터로 사용하였고 18개를 검증 데이터로 사용하여 신경망 모델을 개발하여 선형회귀모델과의 예측력을 비교하였다. 연구 결과, 신경

망 모델이 쓰레기 발생량의 경향(Trend)과 파동을 더 정확하게 예측하였다.

Emrah Dogan et al.(2008)는 2005년의 364개 일별 COD, N, P, SS, 유량(Inflow) 데이터를 활용하여 하수처리장 유입수의 당일 BOD 농도를 예측하였다. 데이터 중 244개는 모델의 학습에 사용하였고 120개는 모델을 검증하는 데 사용하였다. 다중선형회귀 모델과 ANN 모델에 각각 적용하여 선형모델과 비선형모델의 예측 성능을 비교하였다. 연구 결과 은닉층 내 노드가 3개인 ANN 모델의 결정계수가 0.919로 다중회귀분석(0.906)보다 우수한 예측력을 가지는 것을 증명했다.

Mahmoud S. Nasr et al.(2012)는 이집트 EL-AGAMY 하수처리장을 대상으로 COD, BOD, SS 데이터를 설명변수로 사용하여 COD, BOD, SS를 각각 예측하는 ANN 모델을 연구했다. 본 연구는 1년간의 일별 데이터를 활용하여 60%는 모델의 학습 데이터로 사용하였고 20%는 검증 데이터로 사용하였다. 나머지 20%로 검증된 모델을 최종 테스트했을 때 모델의 결정계수가 0.86으로 ANN의 예측력을 검증했다.

Maged M. Hamed et al.(2004)는 10개월간 하수처리장의 일별 BOD, SS 데이터를 설명변수로 사용하여 당일 BOD, SS를 각각 예측하는 두 개의 모델을 개발했다. 모델 학습에 필요한 훈련 데이터를 각각 60%(92개), 70%(108개), 80%(123개), 90%(138개)로 나누었고 나머지를 검증 데이터로 사용했다. BOD를 예측하는 모델은 세 개의 노드가 들어간 ANN 모델을 개발했고, SS를 예측하는 모델은 두 개의 노드가 들어간 ANN 모델을 개발했다. 연구 결과, BOD를 예측한 모델의 결정계수는 0.63 ~ 0.81로 높은 예측력을 나타냈으나 SS를 예측한 모델의 결정계수는 0.45 ~ 0.65로 다소 떨어지는 예측력을 나타냈다. 이를 통해 훈련 데이터와 검증 데이터를 나누는 비율에 따라서 모델 예측력이 달라지는 것을 증명했다.

Mohamed F.Hamoda et al.(1999)는 1995년 9월 13일 ~ 1997년 1월 1일까지 쿠웨이트의 Ardiya 하수처리장의 228개의 일별 기온, 유량, BOD, SS, 첫 번째 공정 이후의 BOD, SS 데이터를 설명변수로 사용하여 하수처리장 내 두 번째 공정 후의 BOD와 SS, 방류수의 BOD, SS 농도를 예측하였다. 본 연구는 노드가 25개로 구성된 은닉층이 1개와 2개인 두 개의 신경망 모델의 예측력을 비교하였는데 은닉층이 1개인 신경망 모델의 결정계수가 더 높았다. 본 연구를 통해 하수처리장을 대상으로 한 신경망 모델의 예측력을 검증하였고 은닉층 수의 증가가 과적합을 일으킬 수 있다는 것을 검증했다.

다. 선행연구와의 차별성

국내 하수처리장 유입수의 BOD 농도 예측방법과 신경망을 통한 예측방법에 관한 비교 연구가 국내에는 많이 부족한 실정이다. 또한, 선행연구 대부분 2년 미만의 데이터를 사용해 예측력을 검증하였는데 이는 유입수의 복잡하고 비선형적 거동을 예측하기에 데이터 신뢰성이 떨어질 수 있는 데이터 크기라 판단된다. 본 연구는 5년 이상의 일별 데이터를 사용하여 모델의 신뢰도를 높였다. 또한 유입수 특성상 처리시설로 연속해서 들어오는 점을 고려하여 과거 수질 데이터와 상관성이 높다고 판단, 하루 전과 이틀 전의 수질 데이터를 파생변수로 추가하여 예측력을 높이하고자 하였다. 마지막으로 강우 시 하수처리장의 유입 BOD 값을 예측하여 BOD 변화에 따른 하수처리장으로 유입되지 않고 월류되는 비점오염원의 유기물량을 예측하여 유역관리를 위한 사전 대처의 기초자료로 활용하고자 하였다.

Ⅲ. 연구 방법

1. 연구 목적 및 범위

본 연구는 당일 BOD 농도를 예측하는 DNN 모델을 개발하여 유입수의 비선형적 특성을 반영하지 못하는 기존 하수처리장의 예측방법의 한계점을 극복하고 신경망 모델의 대체 가능성을 확인하고자 하였다. 이를 통해 하수처리 공정과정에 대해 사후대처가 아닌 사전 감지 및 예방을 통한 안정적인 운전으로 적절한 유기물 부하를 조절하여 공정과정을 효율성을 높이 고자 하였다. 이는 법적 기준 초과 방류 문제를 해결할 수 있는 도구로 이 용할 수 있고 유역관리에 필요한 하수처리장 점오염원 관리대책에 새로운 방법을 제시할 수 있다. 나아가 강우 시 하수처리장의 유입 BOD 값을 예 측하여 BOD 변화에 따른 하수처리장으로 유입되지 않고 월류되는 비점오 염원의 유기물량을 예측하여 유역관리를 위한 사전 대처의 기초자료로 활 용하고자 하였다.

대상이 된 하수처리장은 B 시에서 운영되고 있는 3개 하수처리장(A, B, C)을 대상으로 하였다. 3개 하수처리장을 선택한 이유는 하수처리장마다 위치나 인접한 시설, 주거 환경에 따라 유입 하수의 성상이 달라지기 때문 이다. 따라서 하수처리장마다 분석된 수질 변수(Feature)가 동일해도 변수 간 상관관계 및 데이터 범위가 다르므로 예측 모델에 적용하였을 때 하수 처리장마다 예측 결과가 달라진다. 이렇듯, 다양한 조건의 하수처리장을 대상으로 신경망을 적용하여 다른 모델보다 우수한 예측력이 확인된다면 하수처리장에 대한 신경망 모델 적합성이 더욱 신뢰성 있다고 판단되었다. 따라서 본 연구는 비교 모델로 선형결합 구조인 다선형회귀분석 모델

(Multi linear regression: MLR)을 선정하여 비선형결합 구조인 DNN 모델의 예측력을 비교 및 평가하였다.

본 연구의 절차 개요는 Fig. 3과 같다. 앞서 이론적 배경에서 첫 번째 절차인 하수처리장 및 신경망에 대한 이론적 고찰을 수행하였다.

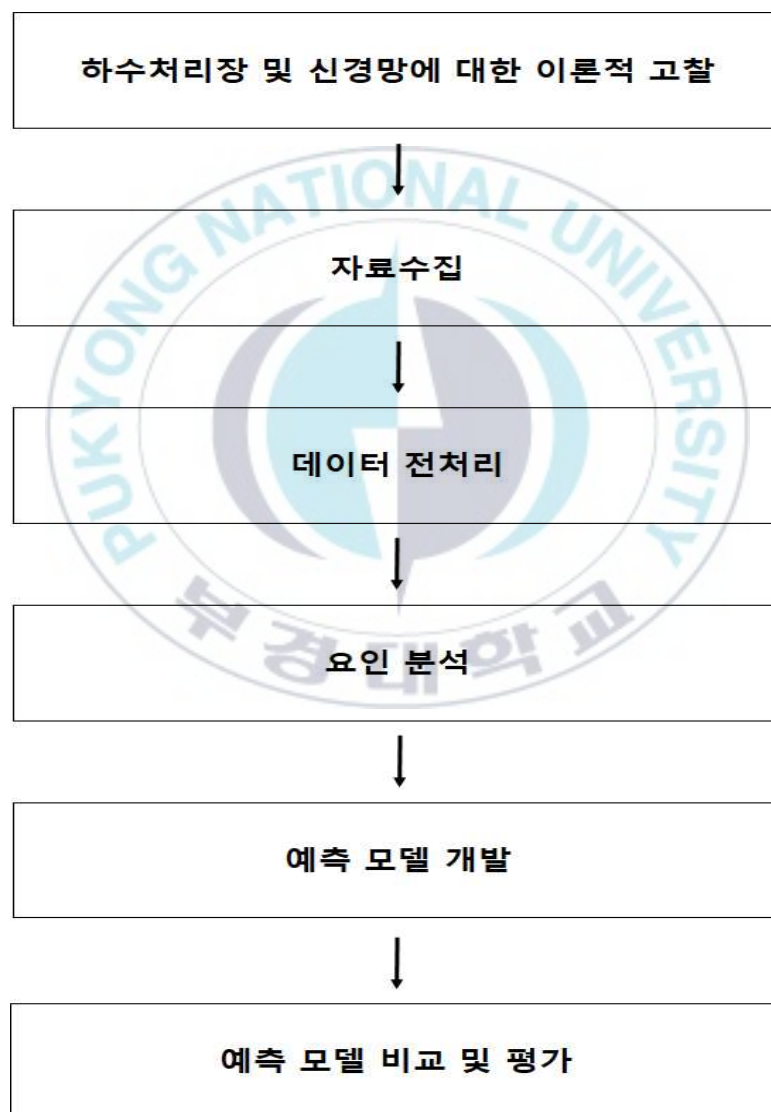


Fig. 3. Research procedure

2. 자료수집

연구에 사용된 데이터는 2015년 1월 1일부터 2020년 9월 30일까지 B시 3개 하수처리장에서 측정된 2,100개의 유입 수질 데이터인 BOD, COD, SS, TN, TP, 유량(Q), 대장균(E.coli)과 기상 데이터인 기온(T), 강수량(P)을 공공 개방 포털에서 확보하였다. 확보된 데이터는 전처리(Preprocessing) 후 예측 모델 개발을 위한 자료로 사용되었다.

3. 데이터 전처리

신경망은 데이터 상태에 따라 예측력이 크게 달라진다. 만약 훈련 데이터셋(Training dataset) 내 데이터 분포의 불균형, 이상치(Outlier) 등이 있다면 과대적합 문제가 발생하게 된다. 또한, 고차원의 데이터를 회귀 모델에 사용 시 불필요한 변수가 존재할 수 있고 다중공선성(Multicollinearity) 문제가 발생할 수 있다. 다중공선성이란, 설명변수 간 연관성이 크고 특성이 유사하여 데이터 분석 시 분산 값과 회귀계수 추정 변동이 커지는 문제다. 회귀분석은 설명변수와 종속변수의 상관관계가 높고 설명변수가 서로 독립일 때 모델의 설명력에 신뢰도가 생기고 높아지기 때문에 다중공선성은 모델의 신뢰도를 저하시킨다(류 등, 2016). DNN 모델의 경우 출력층까지 수많은 비선형 활성화 함수의 복잡한 조합으로 되어 있기 때문에 다중공선성의 문제를 해석 비용으로써 보호할 수 있다(Richard et al., 2000). 하지만 비교 대상인 MLR 모델은 다중공선성 문제에 취약하므로 설명변수 간 연관성이 높은 변수를 소거하거나 새로운 차원에 매핑(Mapping)하여 설명변수 간 독립성을 유지하게 하는 전처리 과정을 수행하는 Feature Engineering을 수행해야 한다. 이는 모델의 예측 수행능력 향상을 위해 주

어진 변수로부터 적합한 변수들로 재구성하는 작업으로, 기존 변수들로부터 새로운 변수들을 만들어내기 위해 산술과 및 집계 연산자와 같은 변환 기능(Transformation functions)을 내포한다. 이러한 변환기능은 변수를 확장 시키거나 변수 간 복잡한 비선형 관계를 모델이 학습하기 쉬운 선형관계로 만드는 데 도움을 준다(Fatemeh et al., 2017). 이를 차원 감소(Dimension reduction)라고 하는 고차원 데이터에 사용되는 전처리 기법으로, 변수를 줄여 연산량을 줄임으로써 계산 효율성을 높이고 다중공선성 문제를 해결하여 모델 성능을 향상시킨다. 본 방법으로 변수 선택(Feature selection)과 변수 추출(Feature extraction)이 있다.

가. 단계 선택법(Stepwise method)

변수 선택이란, 특정 문제를 해결하기 위해 관련된 정보를 포함하는 설명 변수만을 선택하는 방법으로(Nikolay et al., 2006) 유의미한 변수를 선택하고 불필요한 변수를 소거하여 모델의 오차를 최소화할 수 있는 최소한의 설명변수 집합을 선택하는 것을 목적으로 한다(류 등, 2016). 변수 선택 시, 사용자의 주관에 따른 일부 변수만을 선택하거나 통계적 변수 선택 절차를 적용하는 방법이 있다.

단계 선택법은 변수 선택 절차 중 하나로 종속변수에 영향력이 큰 설명변수를 하나씩 선형회귀모형에 추가하였을 때, 특정 설명변수가 종속변수에 유의한 영향을 미치지 못하면 해당 설명변수를 선형회귀모형에서 제외하는 방법을 계속해서 반복하여 최적화한다(송 등, 2005). 하지만 변수 제거 시 해석은 쉬워질 수 있으나 중요 변수가 빠질 수 있고 변수 간 상관성이 강한 경우 어떤 변수를 우선으로 하는지에 대한 명확한 기준이 없기 때문에 변수를 어떻게 선택할 것인지에 대한 논의가 필요하다(김 등, 2018).

나. 주성분분석(Principal component analysis: PCA)

변수 추출이란, 설명변수들의 정보를 최대한 유지한 채 저차원 공간으로 투영하여 새로운 설명변수를 만드는 방법으로, 차원축소 과정을 통해 다중공선성 문제를 해결하고 차원을 줄임으로써 데이터를 좀 더 직관적으로 이해하고 해석할 수 있다. PCA는 대표적인 변수 추출 방법으로 구성성분 간 분산을 최대화하고 불필요한 구성성분을 제거함으로써 다차원 시스템의 복잡성을 감소시킨다(Handan et al., 2005). 기존 데이터셋에서 중요한 정보를 추출하고 이 정보를 주성분(PCs)이라고 하는 직교 변환된 새로운 변수로 만들어 표현한다(Hervé et al., 2010). PCA는 데이터셋을 한 개의 축에 사상시켰을 때, 그 분산이 가장 커지는 축을 제1 주성분으로, n번째로 커지는 축을 제 n 주성분으로 설정한다(Fig. 4). 이러한 과정을 통해 생성된 주성분은 입력 변수의 개수만큼 최대 생성시킬 수 있으므로 주성분의 수는 원래 자료의 변수 개수보다 작거나 같다(홍 등, 2020). 이들 주성분은 서로 선형 연관성이 없는 독립관계이기 때문에 회귀분석 시 발생할 수 있는 다중공선성의 문제를 해결할 수 있다. 그러나 기존 설명변수들의 선형결합을 해석하기 힘들고 각 설명변수가 종속변수에 얼마나 영향을 미치는지에 대한 정보를 잃을 수 있다(Samina et al., 2014). 따라서 변수 추출 과정에서 차원 축소 시 발생하는 정보 손실을 최소화하는 것이 필요하다(R. Noori et al., 2009).

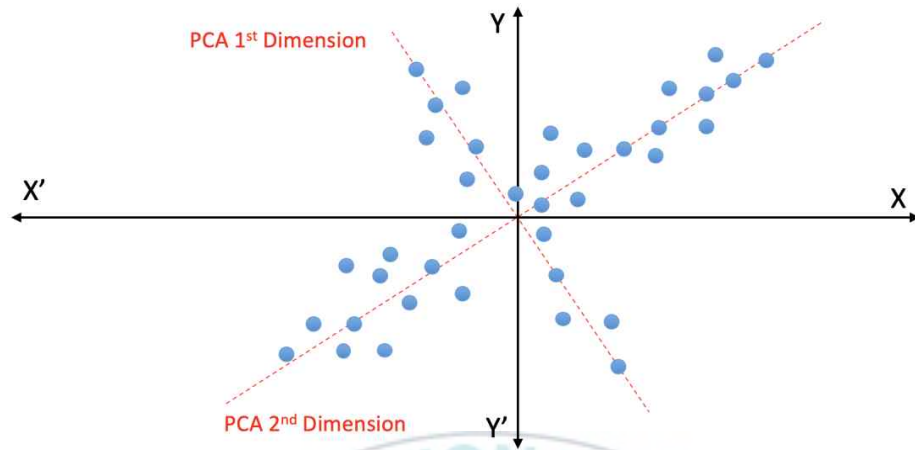


Fig. 4. PCs orthogonal

다. 파생변수(Derived variable)

파생변수란, 데이터 내 새로운 변수로 사용자(분석자)가 직간접적으로 수집한 변수 중에서 상호관련성을 고려하여 만들어낸 제3의 별도 변수이다. DNN 모델의 설명변수들은 활성화 함수를 통해 종속변수와 선형관계가 아닌 특정 비선형 관계로 표현될 수 있으나 파생변수를 사용함으로써 종속변수와 선형관계를 증가시켜 개선된 예측모형을 얻을 수 있다. 본 연구에서는 하수처리장 유입수의 연속성을 고려하여 BOD의 농도가 직전 과거 수질 데이터와 연관이 있다고 판단, 예측일을 기준으로 하루 전(BOD_{t-1} , COD_{t-1} , SS_{t-1} , TN_{t-1} , TP_{t-1})과 이틀 전(BOD_{t-2} , COD_{t-2} , SS_{t-2} , TN_{t-2} , TP_{t-2})의 유입 수질 데이터를 파생변수로서 추가하였다. 파생변수는 예측모형의 신뢰도를 높이는 데 큰 영향을 주기 때문에 유효성을 검증해야하므로 (송과 김, 2013), 본 연구는 파생변수를 추가한 데이터셋과 추가하지 않은 데이터셋을 각각 모델에 적용하여 예측력을 비교 및 평가하였다.

4. 상관 분석(Correlation analysis)

본 연구는 2015년 1월 1일부터 2020년 9월 30일까지의 B 시 3개 하수처리장을 대상으로 총 2,100개의 일별 BOD, COD, SS, TN, TP, 유량, 대장균, 강수량, 기온 데이터를 바탕으로 상관 분석을 시행하였다. 상관 분석으로는 하수처리장의 변수 간 피어슨 상관계수(Pearson correlation coefficient)를 사용하였다. 피어슨 상관계수란, 두 변수의 공분산(Covariance)을 각각의 표준편차의 곱으로 나눈 값으로 +1과 -1 사이의 값을 가진다. +1에 가까울수록 변수 간 강한 양의 상관관계를, -1에 근접할수록 강한 음의 상관관계를 가지고 0에 가까울수록 변수 간 상관관계가 존재하지 않는다고 판단할 수 있다(식 3.1).

$$\rho_{x,y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \quad (3.1)$$

여기서, $cov(X, Y)$: Covariance of X and Y

σ_X : Standard deviation of X

σ_Y : Standard deviation of Y

5. 예측 모델 개발 연구

본 연구는 하수처리장 당일 유입수의 BOD 예측을 목적으로 하수처리장의 일별 데이터에 DNN 모델을 적용하였다. DNN 모델의 예측력을 검증하기 위해 단계선택법과 PCA를 적용한 MLR 모델을 각각 개발하여 DNN 모델과 비교 및 평가하는 데 사용하였다. 또한, 파생변수의 유효성을 검증하기 위해 파생변수를 추가한 데이터셋과 추가하지 않은 데이터셋을 각각 모델에 적용하여 예측력을 비교 및 평가하였다.

가. DNN 모델 개발

본 연구에 사용되는 하수처리장 DNN 모델은 Fig. 5와 같은 구조로 되어 있다. 입력층에 설명변수에 해당하는 수질과 기온 변수를 입력시켜 여러 개의 은닉층 지나며 차원을 왜곡시키고 복잡한 연산과정을 거친 후 최종적으로 출력층에서 예측 BOD 값이 출력된다. 본 연구의 은닉층에 사용된 활성화 함수는 ReLU(Rectified Linear Unit)로 Fig. 6과 같다. 입력값이 0보다 작으면 0으로 출력, 0보다 크면 식 3.1과 같이 입력값으로 출력된다. ReLU 함수는 출력값을 특정 범위 내로 한정하지 않기 때문에 기존 시그모이드(Sigmoid) 함수에서 발생할 수 있는 기울기 소실(Vanishing gradient) 문제를 해결할 수 있으며 함수식이 간단하여 다른 비선형함수보다 연산속도가 빠른 장점이 있다(최와 정, 2020).

모델의 예측력을 높이기 위해선 데이터셋에 맞는 최적의 모델 구조를 설정해야 한다. 사용자는 이를 위해 은닉층 수와 노드 수를 적절하게 조절한다. 만약 노드의 수나 은닉층을 깊게 쌓지 않으면 데이터의 복잡한 변동을 예측하지 못하는 과소 적합(Underfitting) 문제가 발생할 수 있다. 반면 노

드의 수가 지나치게 많고 은닉층을 너무 깊게 쌓으면 훈련에 사용되는 데이터의 변동에만 치우쳐져 훈련 오차는 매우 낮지만 새로운 데이터가 들어왔을 때 오차가 높게 발생하는 과대 적합 문제가 발생할 수 있으므로 사용자는 다양한 구조로 훈련된 모델의 오차를 비교하여 평가해야 한다.

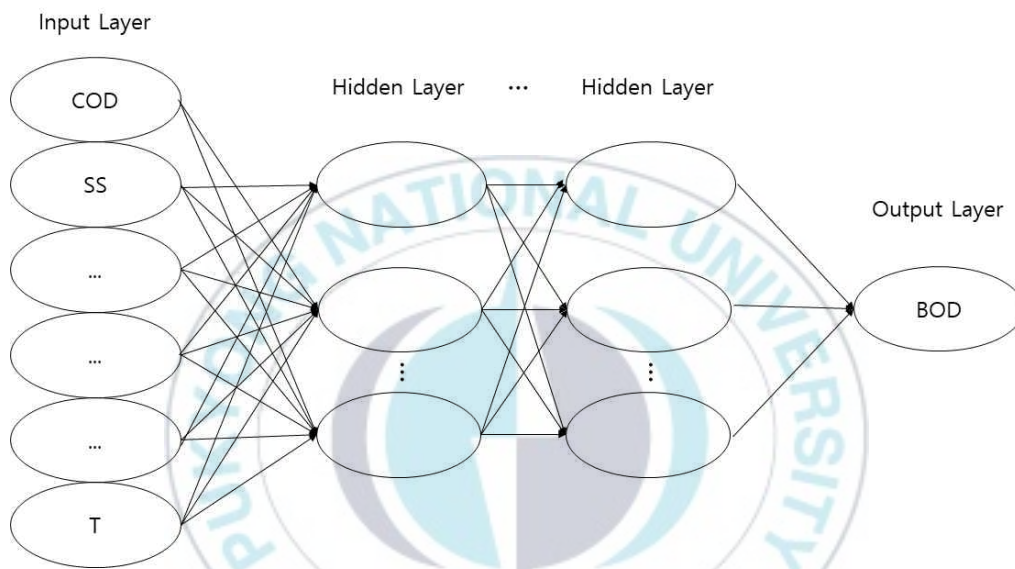


Fig. 5. Structure of DNN in study

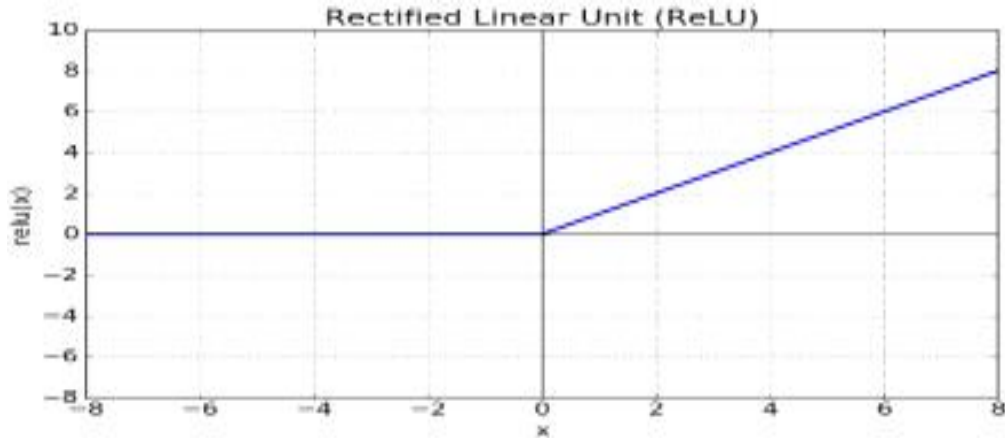


Fig. 6. rectified linear unit(ReLU)

$$ReLU(=y(x)) : \begin{cases} y(x) = x & x \geq 0 \\ y(x) = 0 & x < 0 \end{cases} \quad (3.1)$$

(1) DNN의 학습

신경망 학습은 노드 간 연결된 가중치를 조절하여 출력된 BOD 예측값과 하수처리장의 실측값의 오차를 최소화하는 작업이다. 본 연구의 DNN 모델에 사용된 학습 방법은 '역전파 알고리즘(Backpropagation algorithm)'으로 주어진 실측값에 가장 잘 맞는 노드 간 연결 가중치를 찾아내는 2단계 알고리즘으로 Fig. 7과 같다. 순방향으로 층을 지나면서 오류를 추정한 후, 측정된 오류를 역방향으로 전파하여 연결 가중치를 조정하여 오차를 줄이는 과정을 반복한다. 이 훈련과정은 반복 훈련에 대해 원하는 최적점에 도달할 때까지 계속되며 사용자는 성능 향상을 위해 훈련 반복 수를 늘리거나 더 많은 훈련 데이터를 추가할 수 있다(Mohamed et al., 1999).

경사 하강법(Gradient descent)은 신경망 훈련 시 사용되는 가중치를 편미분하여 오차 함수의 기울기가 낮은 쪽으로 이동시켜 극값을 찾은 뒤 이

것을 최적 가중치로 확정하는 원리로 역전파 알고리즘에서 매우 중요한 개념이다(Fig. 8). 미분 기울기가 양수면 가중치가 커질수록 오차가 커지는 것을 의미하고, 기울기가 음수면 가중치가 커질수록 오차가 작아지는 것을 의미한다. 또한, 기울기 절댓값이 클수록 가파르다는 의미로 최대·최솟값에 해당하는 좌표로부터 멀리 떨어져 있음을 의미한다. 식 3.2와 같이 가중치가 커질수록 기울기가 양의 방향으로 커진다면 음의 방향으로 가중치를 이동시키고 반대로 가중치가 커질수록 기울기가 음의 방향으로 커진다면 양의 방향으로 가중치를 이동시킨다.

학습률(Learning rate)은 경사 하강법에서 시작점이 다음 지점으로 이동할 때의 보폭(Step)으로 이동 속도를 조절해준다. 하지만 설명변수들의 단위나 크기 범위가 다를 경우 변수 간 변동(분산)이 달라진다. 가령 온도의 경우 0.1°C 값이 매우 큰 변화를 의미하지만, 키의 경우 0.1cm 는 큰 변화라 보기 어려우므로 모델 학습 시 데이터의 크기가 다르면 설정된 학습률에 따라 변수의 중요도를 평가하는 데 문제가 생긴다. 따라서 모든 변수를 동일 범위로 스케일링(Scaling)하기 위해 데이터 정규화(Normalization)를 실행했다(식 3.3). 정규화된 데이터는 0~1 사이의 값을 가진다.

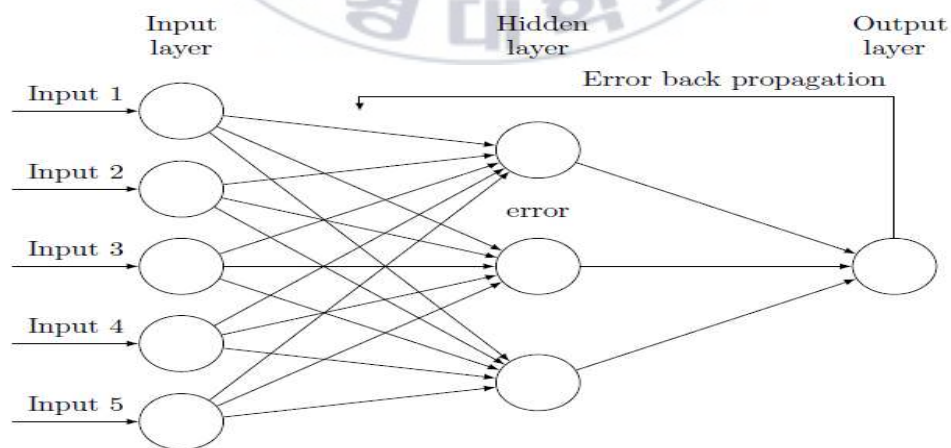


Fig. 7. Backpropagation algorithm

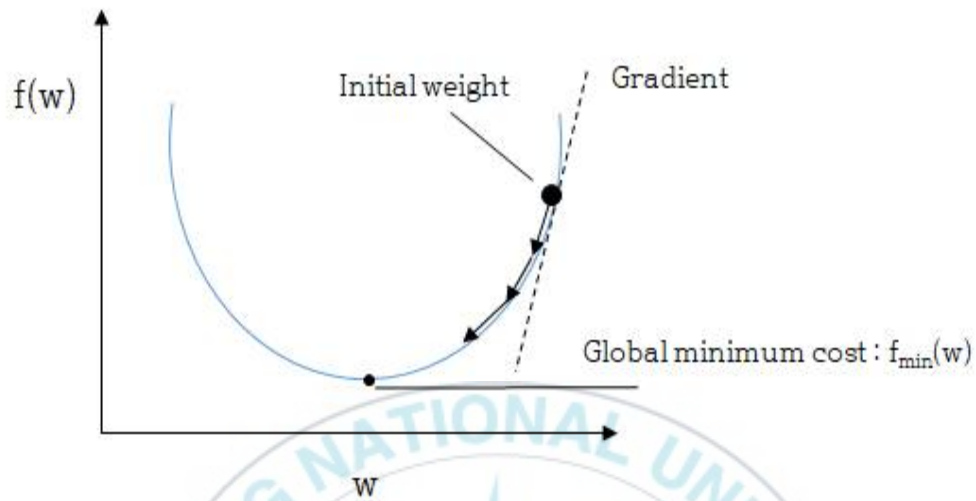


Fig. 8. Gradient descent to find optimal weight

$$w_{i+1} = w_i - \alpha * \frac{df(w_i)}{dw} \quad (3.2)$$

여기서, w_{i+1} : new weight

w_i : before weight

i : ith value, $i = 1, 2, 3, \dots, n$

$f()$: error function

α : learning rate

$$X = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (3.3)$$

여기서, X : normalization value

x : measured value

$x_{\max}$: maximum value in measured values

$x_{\min}$: minimum value in measured values

DNN 모델의 학습은 반복 횟수와 배치 사이즈(Batch size) 조절이 필요하다. 반복 횟수가 적으면 학습이 완전히 이루어지지 않는 과소 적합이 발생한다. 반대로 오차 최적점이 나왔음에도 불구하고 학습이 계속된다면 과대적합 문제가 발생할 수 있다. 배치 사이즈란, DNN 모델의 가중치를 한번 업데이트 시킬 때 사용되는 샘플 묶음을 의미한다. 예를 들어 1,000개의 훈련 데이터가 있는데 배치 사이즈가 20이라면 20개의 데이터마다 모델의 가중치를 수정해 총 50번의 가중치가 수정되고 학습 한 번이 끝나는 것을 의미한다. 단, 반복횟수와 배치 사이즈 크기의 정답은 없으므로 사용자의 경험이나 여러 조건의 결과값을 비교하여 결정한다(Fig. 9).

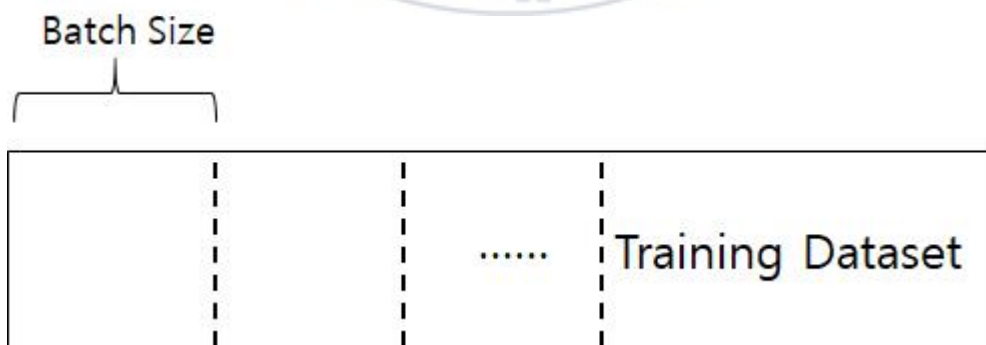


Fig. 9. Training dataset with Batch size

나. 다중선형회귀 모델 개발

DNN의 예측력을 비교 및 평가하기 위해 단계 선택법과 PCA를 이용한 다중선형회귀 모델을 개발하였다. 이들은 모두 선형결합으로 이루어진 회귀 모델이며 다중공선성 문제로부터 자유로워서 DNN 모델의 비선형성과 비교 및 평가하기 적합하다고 판단했다. 두 다중선형회귀 모델 역시 과생 변수를 추가한 데이터셋과 추가하지 않은 데이터셋 각각에 적용하여 과생 변수의 유효성을 검증하였다.

PCA를 적용한 다중선형회귀 모델의 경우 데이터 변동의 누적 설명 비율을 95% 이상이 되는 최소 주성분 수로 설정하였다. 모델에 사용되는 변수들은 표준화(Standardization) 작업을 통해 변수 간 크기 범위를 같이 스케일링하였다(식 3.3).

$$X = \frac{x - m}{\sigma} \quad (3.3)$$

여기서, X : standardization value

x : measured value

m : average of measured value

σ : standard deviation

6. 모델 검증

DNN 모델의 정량적 평가를 위해 예측값과 실측값의 오차 범위를 나타낼 수 있는 평균 제곱근 오차(Root mean square error, RMSE)를 사용하였다(식 3.4). 또한, 모델의 설명력을 검증하기 위해 수정된 결정계수(Coefficient of determination, R^2)를 사용하여 예측력이 가장 우수한 모델을 선정하였다(식 3.5).

$$R'MSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (3.4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.5)$$

여기서, n : the number of values

i : i th value, $i = 1, 2, 3, \dots, n$

y_i : measured value

$\bar{y}_i$: average of measured values

$\hat{y}_i$: output value by predicted model

IV. 결과 및 고찰

1. 상관 분석 결과

본 연구는 B 시 3개 하수처리장을 대상으로 상관 분석을 실시하였다. 하수처리장 세 곳 모두 동일 변수로 구성된 데이터셋을 사용했으나 변수 간 상관관계수가 다르게 나타났다. 이는 유입수 성상을 결정짓는 지역 특성 및 주거 환경에 차이가 있기 때문이라 판단된다. 상관 분석 결과, BOD와 가장 밀접한 연관이 있는 변수는 COD로, 두 변수의 상관성이 높은 이유는 유기물을 제거하는데 필요한 산소 요구량이라는 공통된 특성 때문이라 판단된다. 하지만 하수처리장마다 두 변수 간 상관관계수가 0.8 이상의 강한 상관성을 나타내는 곳이 있었던 반면 0.4 이하의 낮은 값을 가지는 곳이 있었다. 이는 앞서 말한 바와 같이 하수처리장 유입수의 특성을 결정짓는 외부 요인으로 인한 결과로 판단된다.

가. A 하수처리장

2015년 1월 1일부터 2020년 9월 30일까지 A 하수처리장의 일별 데이터는 Table 1과 같다. A 하수처리장의 BOD 변동은 Fig. 10과 같다. 2015년부터 2017년까지 100~150ppm으로 수렴하는 양상을 보이다가 2018년부터 최대 300ppm까지 변동 폭이 커지는 것을 확인할 수 있다. 2019년부터는 변동 폭이 최대 100~200ppm 정도로 일별 최댓값이 감소하였다. 일별 BOD 데이터를 히스토그램으로 나타냈을 때 특정 농도 범위에 데이터가 125~150ppm에 치우쳐 있는 것을 Fig. 11을 통해 알 수 있다. 데이터가 특정 범위에 집

중되어 있으면 학습 시, 해당 범위로 가중치가 조절되어 해당 범위 밖의 데이터가 들어왔을 때 오차가 커지는 과대 적합 문제가 발생할 수 있다.

종속변수 BOD와 설명변수 간 상관관계를 분석한 결과는 Table 2와 같다. BOD는 COD와 약 0.79의 매우 강한 양의 상관관계를 나타내고 있었고 SS, TP, TN 순으로 BOD와 강한 양의 상관관계를 나타내고 있었다. 반면 BOD와 유량의 상관관계는 0.02로 매우 낮았는데, 이는 유량의 변화가 유입수 내 농도 변화에 영향을 주지 않음을 의미한다. 다만 유량은 기온과 0.49의 강한 양의 상관관계를 나타내고 있는데 이는 기온이 높아지면서 지역 물 사용의 증가로 볼 수 있다. 강수량은 BOD와 상관관계수가 -0.06으로 비가 올 때 하수처리장으로 유입되는 BOD 농도를 미미하게 감소시키는 것으로 분석된다. 이는 A 하수처리장은 완전 분류식 하수관거를 사용하기 때문이라고 판단된다.

과생변수를 추가한 후 상관관계를 분석했을 때의 결과는 Table 3과 같다. 하루 전 BOD와 이틀 전 BOD 과생변수 외에는 상관관계수가 0.5 미만으로 종속변수와 큰 유의미한 관계를 가지지 않았다.

Table 1. Dataset information in A WTPP

	BOD	COD	SS	TN	TP	Q (E03)	E.coli (E04)	P	T
mean	142.46	80.52	130.66	34.76	4.00	8.43	1.18	4.64	12.35
std	29.02	14.44	32.56	6.90	1.21	1.39	7.03	16.82	8.68
min	70.10	45.20	66.00	11.24	1.24	5.08	1.30	0.00	-10.20
25%	128.10	72.60	112.00	30.23	3.23	7.52	7.70	0.00	5.20
50%	136.40	78.60	120.00	33.55	3.81	8.28	9.20	0.00	13.40
75%	145.20	85.20	140.00	38.43	4.40	8.97	14.08	0.20	19.40
max	309.60	217.70	400.00	78.18	14.75	14.46	96.00	264.10	28.30

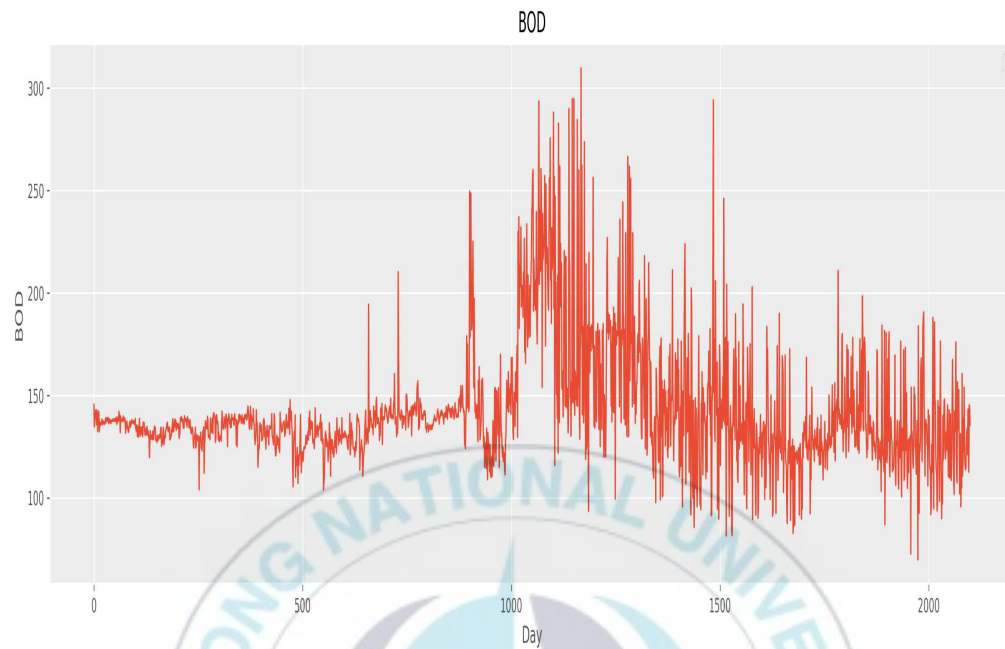


Fig. 10. Daily BOD in A WTPP

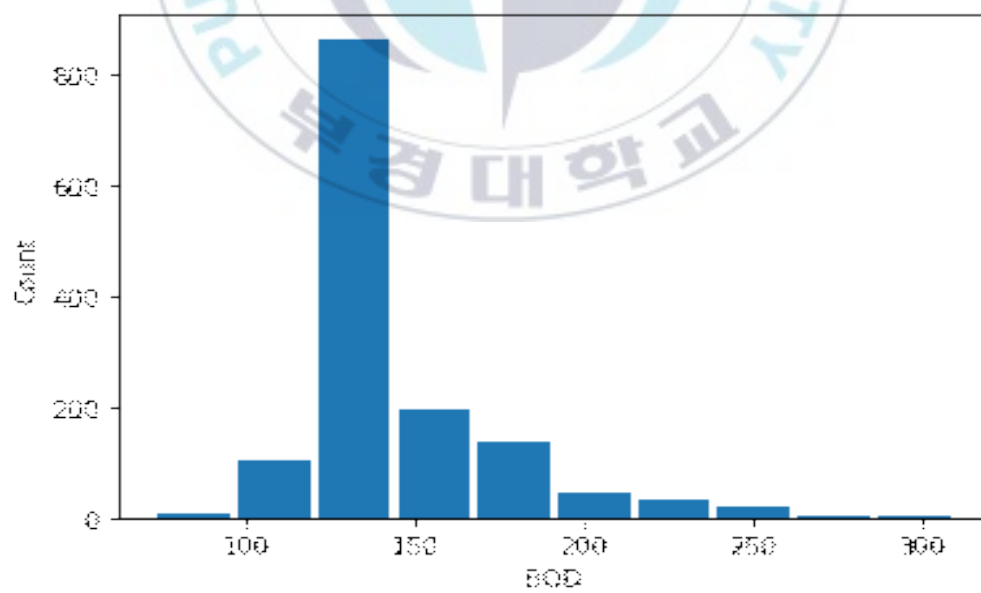


Fig. 11. BOD Histogram in A WTPP

Table 2. Pearson correlation coefficient in A WTPP variables

	BOD	COD	SS	TN	TP	Q	E.coli	P	T
BOD	1.00	0.79	0.64	0.55	0.57	0.02	0.37	-0.06	-0.20
COD	0.79	1.00	0.63	0.60	0.64	-0.02	0.30	-0.07	-0.12
SS	0.64	0.63	1.00	0.54	0.57	0.16	0.51	-0.03	-0.01
TN	0.55	0.60	0.54	1.00	0.67	-0.01	0.27	-0.07	-0.08
TP	0.57	0.64	0.57	0.67	1.00	-0.05	0.30	-0.06	-0.06
Q	0.02	-0.02	0.16	-0.01	-0.05	1.00	0.19	0.29	0.49
E.coli	0.37	0.30	0.51	0.27	0.30	0.19	1.00	-0.01	-0.00
P	-0.06	-0.07	-0.03	-0.07	-0.06	0.29	-0.01	1.00	0.16
T	-0.20	-0.12	-0.01	-0.08	-0.06	0.49	-0.00	0.16	1.00



Table 3. Pearson correlation coefficient adding derived variables of A WTPP

	BOD	BOD _{t-1}	BOD _{t-2}	COD	COD _{t-1}	COD _{t-2}	SS	SS _{t-1}	SS _{t-2}	TN	TN _{t-1}	TN _{t-2}	TP	TP _{t-1}	TP _{t-2}	Q	E.coli	P	T
BOD	1.00	0.57	0.51	0.79	0.40	0.36	0.64	0.42	0.38	0.55	0.36	0.30	0.57	0.38	0.37	0.02	0.37	-0.06	-0.20
BOD _{t-1}	0.57	1.00	0.57	0.37	0.79	0.40	0.41	0.64	0.42	0.36	0.55	0.36	0.37	0.57	0.38	0.02	0.32	-0.06	-0.20
BOD _{t-2}	0.51	0.57	1.00	0.33	0.37	0.79	0.38	0.41	0.64	0.31	0.36	0.55	0.35	0.37	0.57	0.02	0.31	-0.03	-0.21
COD	0.79	0.37	0.33	1.00	0.38	0.31	0.63	0.31	0.24	0.60	0.33	0.26	0.64	0.37	0.32	-0.02	0.30	-0.07	-0.12
COD _{t-1}	0.40	0.79	0.37	0.38	1.00	0.38	0.32	0.63	0.31	0.32	0.60	0.33	0.36	0.64	0.37	0.00	0.22	-0.04	-0.12
COD _{t-2}	0.36	0.40	0.79	0.31	0.38	1.00	0.26	0.32	0.63	0.24	0.32	0.60	0.32	0.36	0.64	-0.01	0.22	-0.02	-0.13
SS	0.64	0.41	0.38	0.63	0.32	0.26	1.00	0.63	0.56	0.54	0.37	0.31	0.57	0.39	0.35	0.16	0.51	-0.03	-0.01
SS _{t-1}	0.42	0.64	0.41	0.31	0.63	0.32	0.63	1.00	0.63	0.35	0.54	0.37	0.37	0.57	0.39	0.18	0.43	-0.02	-0.01
SS _{t-2}	0.38	0.42	0.64	0.24	0.31	0.63	0.56	0.63	1.00	0.29	0.35	0.54	0.34	0.37	0.57	0.17	0.45	-0.03	-0.01
TN	0.55	0.36	0.31	0.60	0.32	0.24	0.54	0.35	0.29	1.00	0.46	0.33	0.67	0.40	0.31	-0.01	0.27	-0.07	-0.08
TN _{t-1}	0.36	0.55	0.36	0.33	0.60	0.31	0.37	0.54	0.33	0.46	1.00	0.46	0.41	0.67	0.40	0.02	0.23	-0.03	-0.07
TN _{t-2}	0.30	0.36	0.55	0.26	0.33	0.60	0.32	0.37	0.54	0.35	0.46	1.00	0.32	0.41	0.67	0.03	0.22	-0.02	-0.08
TP	0.57	0.37	0.35	0.64	0.36	0.32	0.57	0.37	0.34	0.67	0.41	0.32	1.00	0.56	0.47	-0.05	0.30	-0.06	-0.06
TP _{t-1}	0.38	0.57	0.37	0.37	0.64	0.36	0.39	0.57	0.37	0.40	0.67	0.41	0.56	1.00	0.56	-0.02	0.26	-0.02	-0.05
TP _{t-2}	0.37	0.38	0.57	0.32	0.37	0.64	0.35	0.39	0.57	0.31	0.40	0.67	0.47	0.56	1.00	0.01	0.26	0.01	-0.06
Q	0.02	0.02	0.02	-0.02	0.00	-0.01	0.16	0.18	0.17	-0.01	0.02	0.03	-0.05	-0.02	0.01	1.00	0.19	0.29	0.49
E.coli	0.37	0.32	0.31	0.30	0.22	0.22	0.51	0.43	0.45	0.27	0.23	0.22	0.30	0.26	0.26	0.19	1.00	-0.01	0.00
P	-0.06	-0.06	-0.03	-0.07	-0.04	-0.02	-0.03	-0.02	-0.03	-0.07	-0.03	-0.02	-0.06	-0.02	0.01	0.29	-0.01	1.00	0.16
T	-0.20	-0.20	-0.21	-0.12	-0.12	-0.13	-0.01	-0.01	-0.01	-0.08	-0.07	-0.08	-0.06	-0.05	-0.06	0.49	0.00	0.16	1.00

나. B 하수처리장

2015년 1월 1일부터 2020년 9월 30일까지 B 하수처리장의 일별 데이터는 Table 4와 같다. B 하수처리장의 BOD 변동은 Fig. 12와 같다. 2015년부터 2016년까지 150~250ppm으로 수렴하는 양상을 보이다 그 이후로 점점 일별 평균 BOD 농도가 낮아지는 것을 확인할 수 있다. 2019년 이후로는 200ppm 이상 넘지 않으며 최소 50ppm까지 감소하는 추이를 보인다. 일별 BOD 데이터를 히스토그램으로 나타냈을 때 A 하수처리장과 비교해 데이터가 균일하게 분포된 것을 Fig. 13을 통해 확인할 수 있다.

Table 4. Dataset information in B WTPP

	BOD	COD	SS	TN	TP	Q (E04)	E.coli (E04)	P	T
mean	168.62	82.10	98.44	36.05	3.47	8.51	7.63	4.60	12.35
std	39.85	13.96	25.64	5.63	0.69	1.23	2.62	16.77	8.68
min	53.40	41.90	34.00	15.68	1.46	5.21	0.76	0.00	-10.20
25%	141.05	72.90	84.00	32.77	3.06	7.60	6.10	0.00	5.20
50%	162.40	80.60	94.00	36.05	3.46	8.49	7.30	0.00	13.40
75%	195.33	90.10	109.25	38.88	3.84	9.26	8.90	0.20	19.40
max	315.60	182.00	346.70	71.41	9.74	12.89	23.51	264.10	28.30

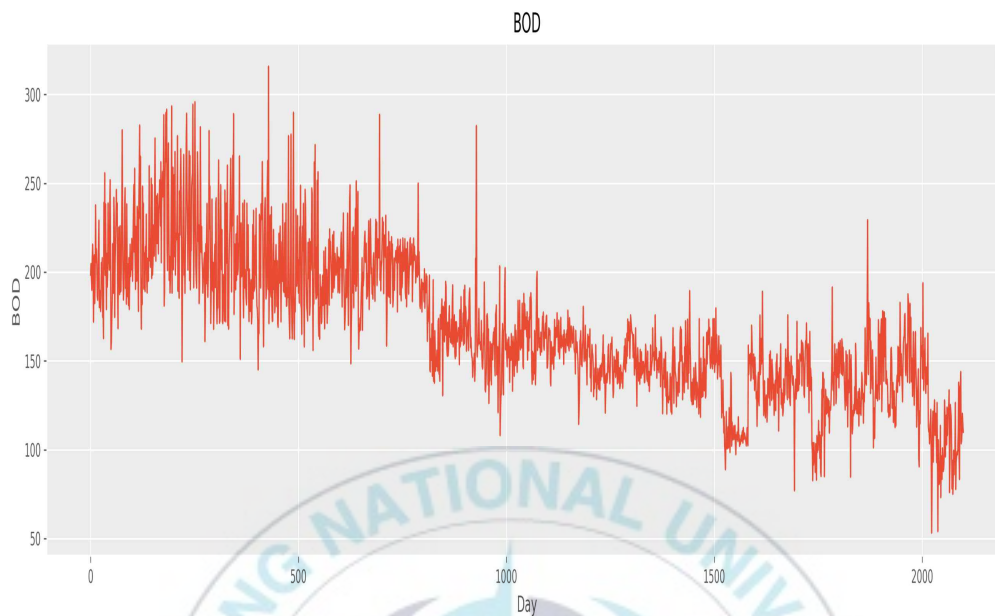


Fig. 12. Daily BOD in B WTPP

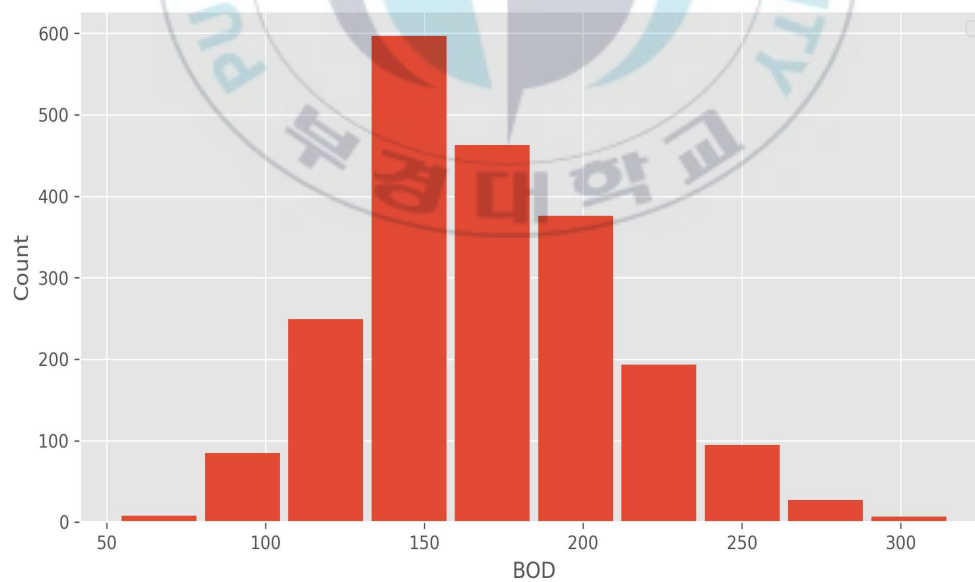


Fig. 13. BOD Histogram in B WTPP

BOD와 설명변수 간 상관관계를 분석한 결과는 Table 5와 같다. BOD는 COD와 0.82의 매우 강한 양의 상관관계를 나타내고 있었고 TP, TN, SS 순으로 BOD와의 양의 상관관계를 가지고 있다. 반면 유량은 -0.61로 BOD와 강한 음의 상관관계를 가지고 있는데 이는 유량이 증가할수록 BOD의 농도가 감소한다는 의미다. 이를 미루어 볼 때 유량이 높은 날은 유기물이 적은 하수가 유입되는 경우가 많다고 생각한다. 유량은 기온과 0.33의 약한 양의 상관관계를 나타내고 있는데 이는 기온 증가로 인한 가정 또는 사업장 내 물 사용 증가로 판단된다. 선행연구 사례에서는 BOD와 SS간 상관관계가 높은 편이지만 본 하수처리장의 경우 BOD와 SS의 상관관계수가 0.21로 매우 낮은 편이었다. 이는 하수처리장에 유입수가 들어오기 전 침사지에서 부유물질들이 상당수 제거되었기 때문이라고 판단된다. 실제 Fig. 14의 SS의 일별 데이터를 확인했을 때 BOD와 달리 값이 튀는 이상치를 제외하면 데이터가 일정 범위 내에서 수렴하는 것을 알 수 있다. 강수량은 BOD와 상관관계수가 -0.05로 비가 올 때 하수처리장으로 유입되는 BOD 농도를 미미하게 감소시키는 것으로 분석된다. 이는 B 하수처리장은 완전 분류식 하수관거를 사용하기 때문이라고 판단된다.

과생변수를 추가한 후 상관관계를 분석했을 때의 결과는 Table 6과 같다. 하루 전 BOD와 COD, 이틀 전 BOD와 COD가 예측 당일 BOD와 0.5 이상의 상관관계수로 상관관계가 유의미하다고 나타났고, 그 외의 과생변수는 0.5 미만의 상관관계수로 유의미한 관계를 얻지 못했다. 이를 통해 B 하수처리장은 가까운 과거의 하수 BOD 농도와 매우 강한 상관관계를 가지고 있고 COD와는 어느 정도 강한 상관성이 있다는 결과를 얻었다.

Table 5. Pearson correlation coefficient in B WTPP variables

	BOD	COD	SS	TN	TP	Q	E.coli	P	T
BOD	1.00	0.82	0.21	0.27	0.50	-0.61	-0.23	-0.05	-0.08
COD	0.82	1.00	0.37	0.41	0.56	-0.47	-0.19	-0.02	-0.09
SS	0.21	0.37	1.00	0.26	0.44	-0.07	0.03	0.10	-0.04
TN	0.27	0.41	0.26	1.00	0.50	-0.17	0.02	-0.04	-0.08
TP	0.50	0.56	0.44	0.53	1.00	-0.44	-0.08	-0.02	-0.12
Q	-0.61	-0.47	-0.07	-0.17	-0.44	1.00	0.21	0.28	0.33
E.coli	-0.23	-0.19	0.03	0.02	-0.08	0.21	1.00	0.02	0.33
P	-0.05	-0.02	0.10	-0.04	-0.02	0.28	0.02	1.00	0.16
T	-0.08	-0.09	-0.04	-0.08	-0.12	0.33	0.33	0.16	1.00

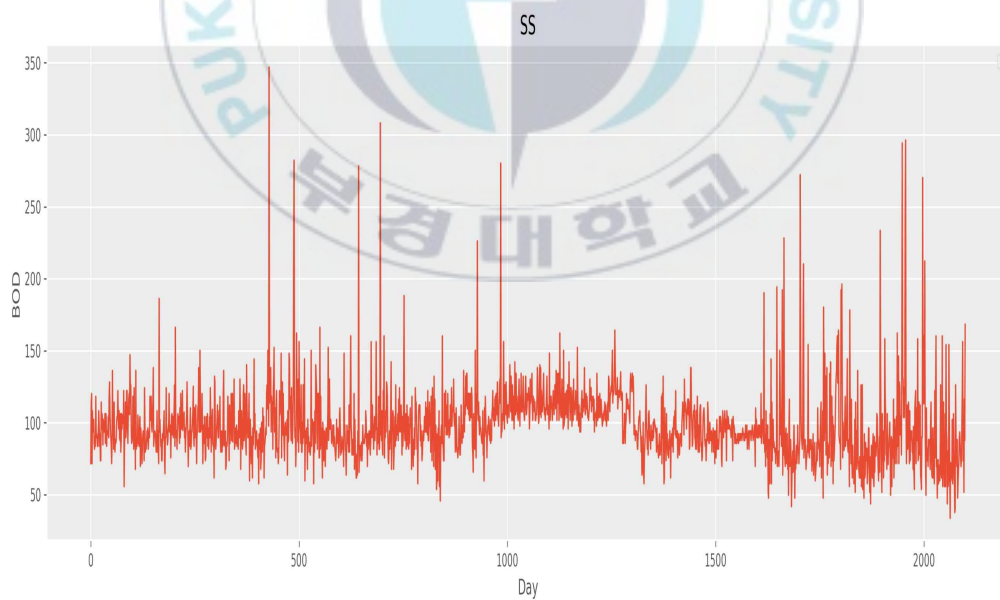


Fig. 14. Daily SS in B WTPP

Table 6. Pearson correlation coefficient adding derived variables of B WTPP

	BOD	BOD _{t-1}	BOD _{t-2}	COD	COD _{t-1}	COD _{t-2}	SS	SS _{t-1}	SS _{t-2}	TN	TN _{t-1}	TN _{t-2}	TP	TP _{t-1}	TP _{t-2}	O	E.coli	P	T
BOD	1.00	0.82	0.77	0.82	0.61	0.54	0.21	0.06	0.05	0.27	0.17	0.14	0.50	0.41	0.39	-0.61	-0.23	-0.05	-0.08
BOD _{t-1}	0.82	1.00	0.82	0.61	0.82	0.61	0.09	0.21	0.06	0.18	0.27	0.17	0.42	0.50	0.41	-0.62	-0.24	-0.04	-0.08
BOD _{t-2}	0.77	0.82	1.00	0.55	0.61	0.82	0.07	0.09	0.21	0.15	0.18	0.27	0.39	0.42	0.50	-0.63	-0.23	-0.04	-0.07
COD	0.82	0.61	0.55	1.00	0.59	0.49	0.37	0.10	0.09	0.41	0.22	0.18	0.56	0.36	0.36	-0.47	-0.19	-0.02	-0.09
COD _{t-1}	0.61	0.82	0.61	0.59	1.00	0.59	0.12	0.37	0.10	0.20	0.41	0.22	0.36	0.56	0.36	-0.48	-0.20	-0.05	-0.08
COD _{t-2}	0.54	0.61	0.82	0.49	0.59	1.00	0.10	0.12	0.37	0.16	0.20	0.41	0.32	0.37	0.56	-0.49	-0.19	-0.04	-0.08
SS	0.21	0.09	0.07	0.37	0.12	0.10	1.00	0.26	0.20	0.26	0.05	0.05	0.44	0.14	0.14	-0.07	0.03	0.10	-0.04
SS _{t-1}	0.06	0.21	0.09	0.10	0.37	0.12	0.26	1.00	0.26	0.04	0.27	0.06	0.11	0.44	0.14	-0.06	0.02	0.03	-0.04
SS _{t-2}	0.05	0.06	0.21	0.09	0.10	0.37	0.20	0.26	1.00	0.02	0.04	0.26	0.09	0.11	0.44	-0.04	0.03	0.03	-0.04
TN	0.27	0.18	0.15	0.41	0.20	0.16	0.26	0.04	0.02	1.00	0.39	0.27	0.53	0.23	0.18	-0.17	0.02	-0.04	-0.08
TN _{t-1}	0.17	0.27	0.18	0.22	0.41	0.20	0.05	0.27	0.04	0.39	1.00	0.39	0.22	0.53	0.23	-0.16	0.01	0.02	-0.07
TN _{t-2}	0.14	0.17	0.27	0.18	0.22	0.41	0.05	0.06	0.26	0.27	0.39	1.00	0.16	0.23	0.53	-0.15	0.00	0.02	-0.08
TP	0.50	0.42	0.39	0.56	0.36	0.32	0.44	0.11	0.09	0.53	0.22	0.16	1.00	0.42	-0.08	-0.44	-0.08	-0.02	-0.12
TP _{t-1}	0.41	0.50	0.42	0.36	0.56	0.37	0.14	0.44	0.11	0.23	0.53	0.23	0.42	1.00	0.42	-0.42	-0.08	-0.04	-0.12
TP _{t-2}	0.39	0.41	0.50	0.36	0.36	0.56	0.14	0.14	0.44	0.11	0.23	0.53	0.23	0.42	1.00	-0.41	-0.09	-0.04	-0.12
Q	-0.61	-0.62	-0.63	-0.47	-0.48	-0.49	-0.07	-0.06	-0.04	-0.17	-0.16	-0.15	-0.44	-0.42	-0.41	1.00	0.21	0.02	0.33
E.coli	-0.23	-0.24	-0.23	-0.19	-0.20	-0.19	0.03	0.02	0.03	0.02	0.01	0.00	-0.08	-0.08	-0.09	0.21	1.00	0.28	0.33
P	-0.05	-0.04	-0.04	-0.02	-0.05	-0.04	0.10	0.03	0.03	-0.04	0.02	0.02	-0.02	-0.04	-0.04	0.02	0.28	1.00	0.16
T	-0.08	-0.08	-0.07	-0.09	-0.08	-0.08	-0.04	-0.04	-0.04	-0.08	-0.07	-0.08	-0.12	-0.12	-0.12	0.33	0.33	0.16	1.00

다. C 하수처리장

2015년 1월 1일부터 2020년 9월 30일까지 C 하수처리장의 일별 데이터는 Table 7로 나타냈다. BOD 변동은 75~170ppm으로 큰 편이나 수렴하는 양상을 Fig 15를 통해 확인할 수 있다. 2018년부터는 BOD 농도가 조금씩 감소하며 75~110ppm으로 수렴하는 추이를 보였다. C 하수처리장의 BOD 데이터를 히스토그램으로 나타냈을 때 양 가장자리 범위를 제외하고 고루 분포된 것을 Fig. 16을 통해 확인할 수 있다.

BOD와 설명변수 간 상관관계를 분석한 결과는 Table 8과 같다. 앞서 A와 B 하수처리장의 경우 BOD와 COD 간 매우 강한 상관관계를 가지고 있었다. 하지만 C 하수처리장의 경우 COD와 0.36으로 다소 낮은 상관관계를 나타내고 있었고 TN, SS, TP, 대장균 순으로 BOD와 양의 상관관계를 나타내고 있었다. 유량과는 -0.45로 다소 높은 음의 상관관계를 가지고 있는데 이는 유량이 증가할수록 BOD의 농도가 감소하는 의미이다. 이는 유량이 증가하는 날은 우수와 같은 유기물이 적은 하수가 유입된다고 판단할 수 있다. 유량은 기온과 0.39로 다소 낮은 양의 상관관계를 나타내고 있는데 이는 기온 증가로 인한 가정 또는 사업장 내 물 사용 증가로 판단된다. 강수량은 BOD와 -0.14의 상관계수로 A, B 하수처리장과 달리 강우 시 하수처리장으로 유입되는 BOD 농도를 보다 감소시키는 것으로 분석된다. 이는 C 하수처리장이 일부 합류식 하수관거를 사용하기 때문이라고 판단된다. 또한, 강수량과 유량의 상관계수가 0.43으로 높은 양의 상관관계를 가지는 것으로 보아 우수가 하수처리장으로 유입되는 것을 확인할 수 있다.

과생변수 추가 후 상관관계를 분석한 결과는 Table 9과 같다. 하루 전 BOD와 이틀 전 BOD의 상관관계가 매우 높게 나타났다. 두 개의 과생변

수를 제외한 다른 변수들은 유량을 제외하고 모두 0.4 이하의 상관계수로 유의미한 관계를 얻지 못했다. 이를 통해 C 하수처리장은 가까운 과거의 하수 BOD 농도와 매우 강한 상관관계를 가지고 있지만 다른 변수와는 상관성이 높지 않다는 결과를 얻었다.

Table 7. Dataset information in C WTPP

	BOD	COD	SS	TN	TP	Q (E04)	E.coli (E04)	P	T
mean	113.34	75.18	147.75	33.31	3.36	7.57	17.45	4.60	12.33
std	28.80	15.06	33.97	5.20	0.64	1.40	2.90	16.77	8.68
min	50.10	34.00	43.00	0.70	0.43	51.01	7.00	0.00	-10.20
25%	90.00	65.20	133.00	30.45	3.00	6.61	16.00	0.00	5.20
50%	113.60	73.60	151.00	32.89	3.25	7.19	17.60	0.00	13.40
75%	136.50	83.00	159.30	35.49	3.68	8.17	19.20	0.20	19.30
max	188.50	196.80	370.00	58.53	7.41	13.63	31.00	264.10	28.30

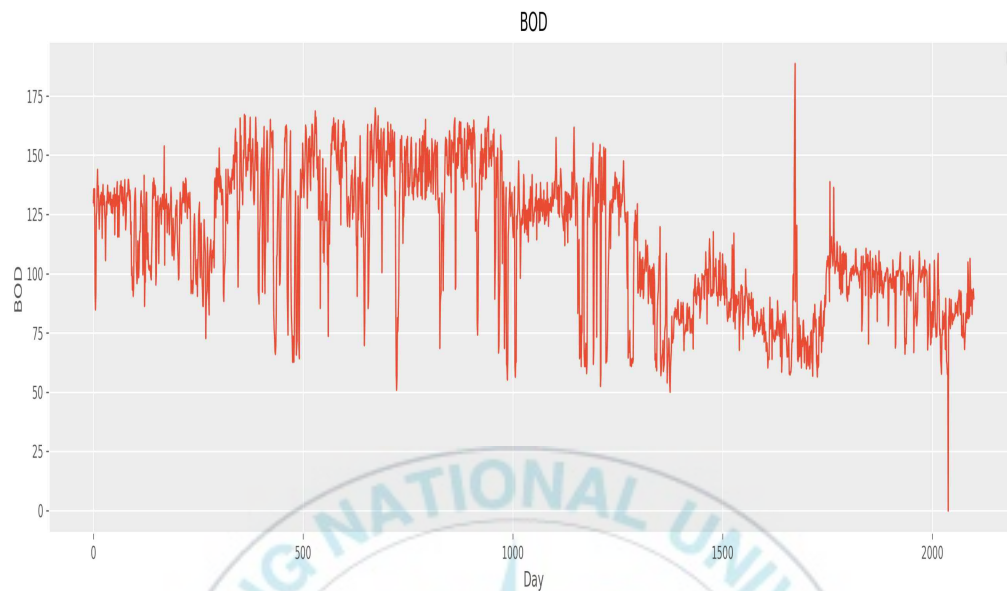


Fig. 15. Daily BOD in C WTPP

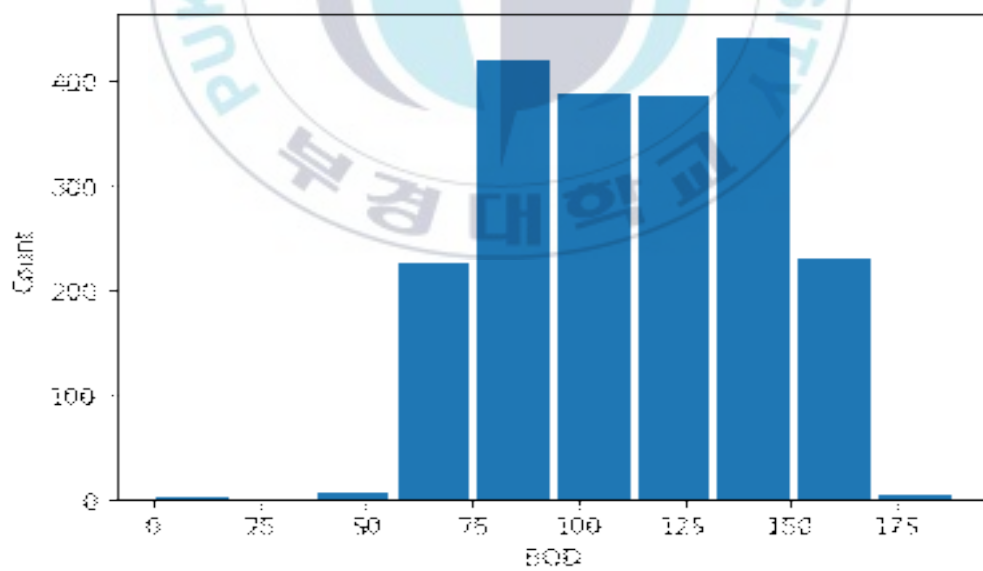


Fig. 16. BOD Histogram in C WTPP

Table 8. Pearson correlation coefficient in C WTPP variables

	BOD	COD	SS	TN	TP	Q	E.coli	P	T
BOD	1.00	0.36	0.30	0.31	0.19	-0.45	0.14	-0.14	-0.14
COD	0.36	1.00	0.41	0.55	0.48	-0.33	0.37	-0.16	-0.14
SS	0.30	0.41	1.00	0.60	0.48	-0.33	0.50	-0.14	0.01
TN	0.31	0.55	0.60	1.00	0.75	-0.37	0.54	-0.18	-0.24
TP	0.19	0.48	0.48	0.75	1.00	-0.24	0.55	-0.13	-0.11
Q	-0.45	-0.33	-0.33	-0.37	-0.24	1.00	-0.34	0.43	0.39
E.coli	0.14	0.37	0.50	0.54	0.55	-0.34	1.00	-0.21	-0.17
P	-0.14	-0.16	-0.14	-0.18	-0.13	0.43	-0.21	1.00	0.16
T	-0.14	-0.14	0.01	-0.24	-0.11	0.39	-0.17	0.16	1.00

Table 9. Pearson correlation coefficient adding derived variables of CWWTP

	BOD	BOD _{t-1}	BOD _{t-2}	COD	COD _{t-1}	COD _{t-2}	SS	SS _{t-1}	SS _{t-2}	TN	TN _{t-1}	TN _{t-2}	TP	TP _{t-1}	TP _{t-2}	Q	E.coli	P	T
BOD	1.00	0.88	0.79	0.36	0.31	0.27	0.30	0.26	0.22	0.31	0.28	0.24	0.19	0.15	0.11	-0.45	0.14	-0.14	-0.14
BOD _{t-1}	0.88	1.00	0.88	0.30	0.36	0.31	0.27	0.30	0.26	0.27	0.31	0.28	0.16	0.19	0.15	-0.37	0.09	-0.07	-0.14
BOD _{t-2}	0.79	0.88	1.00	0.26	0.30	0.36	0.23	0.27	0.30	0.22	0.27	0.31	0.13	0.16	0.19	-0.31	0.05	-0.08	-0.15
COD	0.36	0.30	0.26	1.00	0.80	0.71	0.41	0.35	0.30	0.55	0.49	0.46	0.48	0.42	0.36	-0.33	0.37	-0.16	-0.14
COD _{t-1}	0.31	0.36	0.30	0.80	1.00	0.80	0.36	0.41	0.35	0.50	0.55	0.49	0.45	0.48	0.42	-0.26	0.33	-0.09	-0.14
COD _{t-2}	0.27	0.31	0.36	0.71	0.80	1.00	0.33	0.36	0.41	0.46	0.50	0.55	0.41	0.45	0.48	-0.21	0.29	-0.10	-0.13
SS	0.30	0.27	0.23	0.41	0.36	0.33	1.00	0.81	0.71	0.60	0.56	0.51	0.48	0.44	0.40	-0.33	0.50	-0.14	0.01
SS _{t-1}	0.26	0.30	0.27	0.35	0.41	0.36	0.81	1.00	0.81	0.54	0.60	0.56	0.44	0.48	0.44	-0.24	0.45	-0.07	0.00
SS _{t-2}	0.22	0.26	0.30	0.30	0.35	0.41	0.71	0.81	1.00	0.49	0.54	0.60	-0.40	0.44	0.48	-0.18	0.39	-0.06	0.00
TN	0.31	0.27	0.22	0.55	0.50	0.46	0.60	0.54	0.49	1.00	0.83	0.75	0.75	0.67	0.60	-0.37	0.54	-0.18	-0.24
TN _{t-1}	0.28	0.31	0.27	0.49	0.55	0.50	0.56	0.60	0.54	0.83	1.00	0.83	0.67	0.75	0.67	-0.27	0.49	-0.07	-0.23
TN _{t-2}	0.24	0.28	0.31	0.46	0.49	0.55	0.51	0.56	0.60	0.75	0.83	1.00	0.63	0.67	0.75	-0.19	0.44	-0.06	-0.24
TP	0.19	0.16	0.13	0.48	0.45	0.41	0.48	0.44	0.40	0.75	0.67	0.63	1.00	0.83	0.77	-0.24	0.55	-0.13	-0.11
TP _{t-1}	0.15	0.19	0.16	0.42	0.48	0.45	0.44	0.48	0.44	0.67	0.75	0.67	0.83	1.00	0.83	-0.16	0.49	-0.06	-0.11
TP _{t-2}	0.11	0.15	0.19	0.36	0.42	0.48	0.40	0.44	0.48	0.60	0.67	0.75	0.77	0.83	1.00	-0.09	0.44	-0.03	-0.11
Q	-0.45	-0.37	-0.31	-0.33	-0.26	-0.21	-0.33	-0.24	-0.18	-0.37	-0.27	-0.19	-0.24	-0.16	-0.09	1.00	-0.34	0.43	0.39
E.coli	0.14	0.09	0.05	0.37	0.33	0.29	0.50	0.45	0.39	0.54	0.49	0.44	0.55	0.49	0.44	-0.34	1.00	-0.21	-0.17
P	-0.14	-0.07	-0.08	-0.16	-0.09	-0.10	-0.14	-0.07	-0.06	-0.18	-0.07	-0.06	-0.13	-0.06	-0.03	0.43	-0.21	1.00	0.16
T	-0.14	-0.14	-0.15	-0.14	-0.14	-0.13	0.01	0.00	0.00	-0.24	-0.23	-0.24	-0.11	-0.11	-0.11	0.39	-0.17	0.16	1.00

2. 유입수 BOD 예측 모델 개발

하수처리장의 당일 BOD를 예측하기 위해 DNN 모델과 단계 선택법과 PCA를 각각 적용한 MLR 예측 모델을 개발하였다. 이들 모델은 파생변수의 유효성을 검증하기 위해 파생변수를 추가하지 않은 데이터셋과 추가한 데이터셋을 적용하여 총 6개의 예측 모델을 개발하였다. 파생변수를 추가하지 않은 데이터셋은 8개(COD, SS, TN, TP, 유량, 대장균, 강수량, 기온)의 설명변수를 사용하였다. 파생변수를 추가한 데이터셋은 하루 전과 이틀 전의 수질 성상을 추가로 한 18개(BOD_{t-1} , BOD_{t-2} , COD, COD_{t-1} , COD_{t-2} , SS, SS_{t-1} , SS_{t-2} , TN, TN_{t-1} , TN_{t-2} , TP, TP_{t-1} , TP_{t-2} , 유량, 대장균, 강수량, 기온)의 설명변수를 사용하였다.

가. DNN 모델 개발

DNN 모델의 경우 전체 데이터(2,100개) 중 무작위로 추출한 70%(1,470개)를 훈련 데이터로 사용하여 예측 모델을 개발했다. 개발된 모델의 성능을 검증하기 위해 나머지 30%(630개)를 검증(Validation) 데이터로 사용하였다. 본 연구의 DNN 모델은 반복 횟수를 2,000회, 배치 사이즈를 32로 설정하였고 과도한 학습으로 인한 과대 적합을 방지하기 위해 학습 중 100번 이상 오차가 연속해서 감소하지 않으면 학습이 종료되도록 설정하였다. 정규화를 통해 데이터가 0과 1사이 값을 가지기 때문에 학습률은 0.001로 설정하였다. 신경망은 초기 가중치를 임의의 난수 값을 생성하여 넣게 되는데, 연구의 동일한 결과를 얻기 위해 Random state의 초깃값을 0으로 고정하였다. 은닉층마다 드롭아웃(Dropout)을 0.3으로 설정하였는데, 이는 과대 적합 문제를 해결하는 방법으로 은닉층 노드를 30% 무작위로 제거된 상태

에서 학습을 수행, 실제 검증 시 제거했던 노드를 복원하고 가중치를 연산하여 과대 적합을 방지하는 작업이다(Fig. 17). 유입 BOD 예측을 위한 DNN 모델의 구조를 Table 10으로 나타내었다.

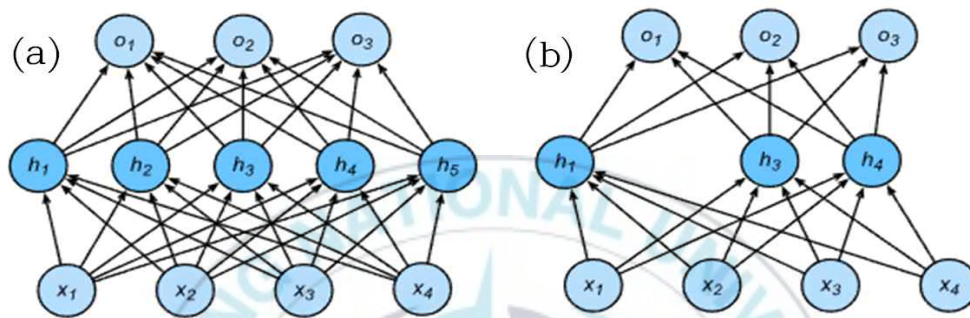


Fig. 17. (a) Before dropout (b) After dropout

Table 10. Structure and variables for DNN model in A, B, C WTPP to predict daily BOD

	target variable	Input variables	Number of nodes		
			Input layer	Hidden layer(I ~ V)	Output layer
without derived variable	BOD	COD, SS, TN, TP, Q, E.coli, P, T	8	256	1
with derived variable	BOD	BOD _{t-1} , BOD _{t-2} , COD, COD _{t-1} , COD _{t-2} , SS, SS _{t-1} , SS _{t-2} , TN, TN _{t-1} , TN _{t-2} , TP, TP _{t-1} , TP _{t-2} , Q, E.coli, P, T	18	256	1

가. A 하수처리장

(1) 다중선형회귀 모델 개발

과생변수를 추가하지 않은 데이터셋을 대상으로 단계 선택법을 적용했을 때 8개 변수 중 TP, 강수량 변수가 다중공선성 문제를 일으켜 소거되었다. PCA를 통한 8개 주성분 중 6개 주성분을 사용하였을 때 전체 데이터 변동의 95.7%를 설명할 수 있었다. 이에 따른 MLR 모델의 설명변수는 Table 11과 같다.

과생변수를 추가한 데이터셋을 대상으로 단계 선택법을 적용했을 때 18개 변수 중 SS_{t-1} , SS_{t-2} , TN_{t-1} , TP, 강수량 변수가 다중공선성 문제를 일으켜 소거되었다. PCA를 통한 18개 주성분 중 12개 주성분을 사용하였을 때 전체 데이터 변동의 95.0%를 설명할 수 있었다. 이에 따른 과생변수를 추가한 MLR 모델의 설명변수는 Table 12와 같다.

Table 11. Dataset without derived variable in A WTPP

MLR model	Input variables
Using Stepwise method	COD, SS, TN, Q, E.coli, T
Using PCA	PCs 1 ~ PCs 6

Table 12. Dataset with derived variable in A WTPP

MLR model	Input variables
Using Stepwise method	BOD_{t-1} , BOD_{t-2} , COD, COD_{t-1} , COD_{t-2} , SS, TN, TN_{t-2} , TP_{t-1} , TP_{t-2} , Q, E.coli, T
Using PCA	PCs 1 ~ PCs 12

(2) 모델 비교 및 평가

하수처리장의 기존 예측 방법을 포함한 7개의 예측 모델의 예측값과 실측값에 따른 R^2 와 RMSE를 통한 정량적 오차 검정을 Table 13과 Fig. 18로 나타내었다. 연구 결과, 파생변수를 추가한 DNN 모델이 가장 뛰어난 예측력을 보였으며 예측 모델 모두 기존 하수처리장의 예측 방법보다 높은 예측력을 보였다. 또한 파생변수를 추가했을 때와 하지 않았을 때 모두 DNN 모델이 MLR 모델보다 높은 예측력을 보였다. 이를 통해 A 하수처리장의 유입 성상을 예측하는데 있어 선형결합 모델보다 비선형적 모델이 더 뛰어난 것을 확인하였다. 다만 파생변수를 추가한 DNN 모델이 파생변수를 추가하지 않은 DNN 모델보다 성능 개선은 약 3.5%로 이는 파생변수 중 과거 BOD 변수 외에는 유의미한 파생변수가 없기 때문이라 판단된다.

Table 13. Comparison models in A WTPP

Forecasting model	R^2	RMSE
Original prediction method	0.632	20.71
DNN without derived variable	0.762	13.63
MLR without derived variable(Stepwise method)	0.712	15.60
MLR without derived variable(PCA)	0.700	15.68
DNN using derived variable	0.797	13.23
MLR using derived variable(Stepwise method)	0.775	13.55
MLR using derived variable(PCA)	0.736	14.74

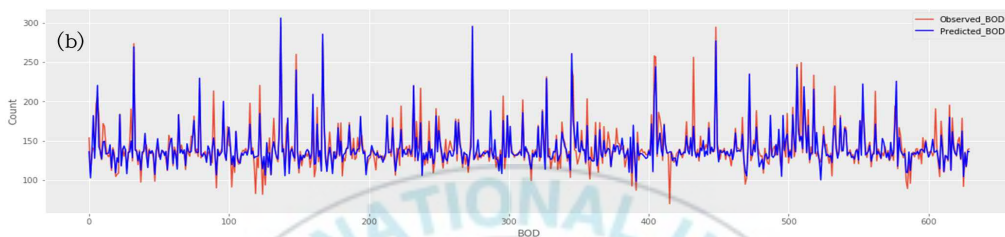
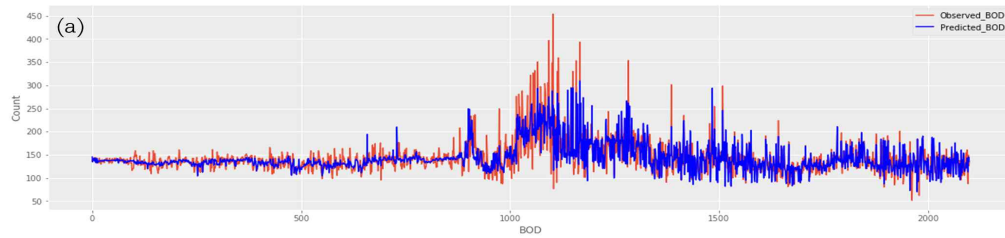
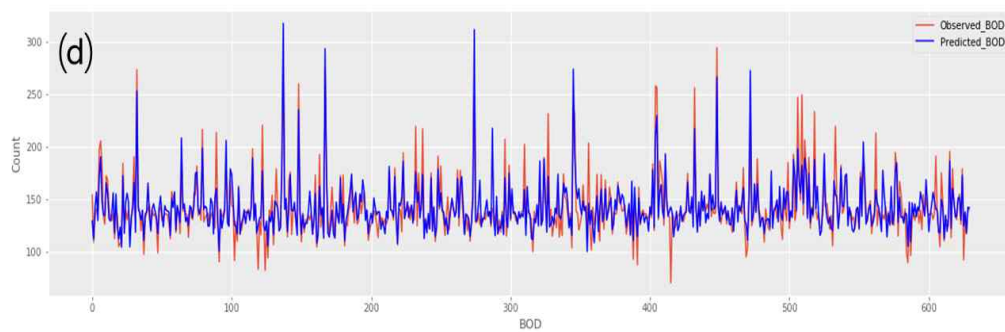
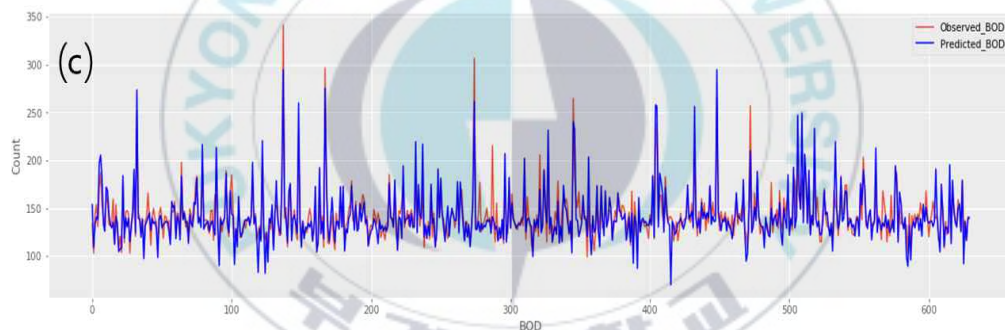
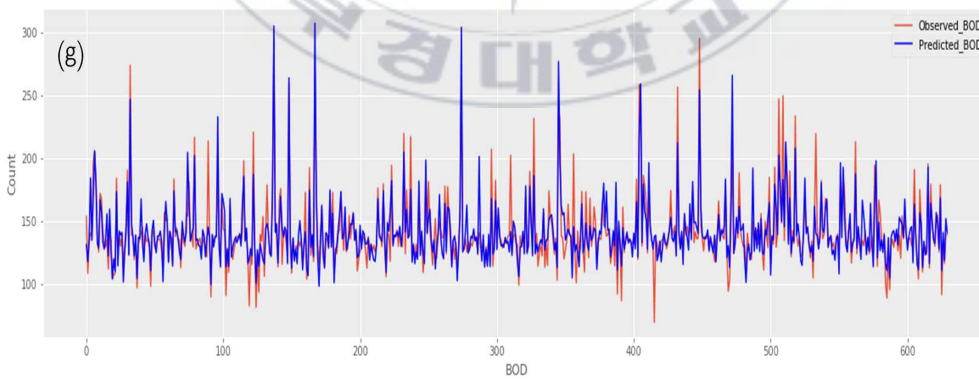
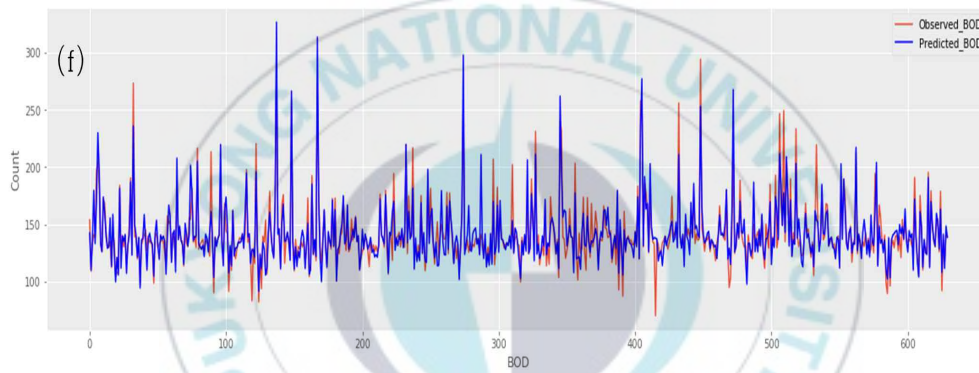
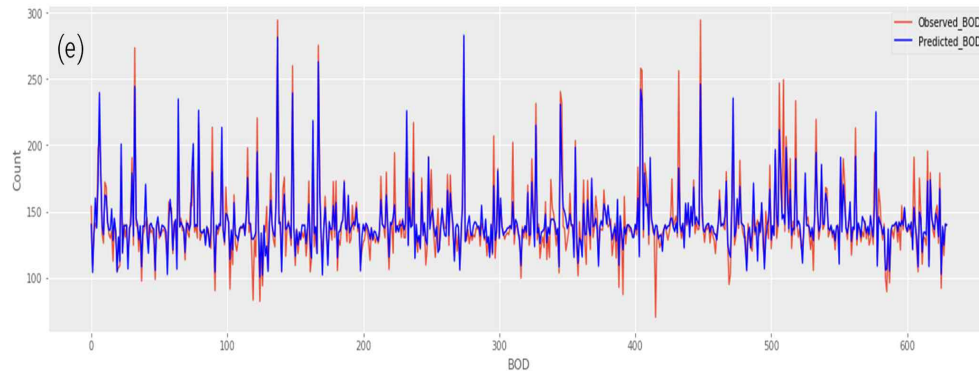


Fig. 18. (a) Original prediction method (b) DNN without derived variable



(c) MLR without derived variable(Stepwise method)

(d) MLR without derived variable(PCA)



(e) DNN using derived variable (f) MLR using derived variable(Stepwise method) (g) MLR using derived variable(PCA)

나. B 하수처리장

(1) 다중선형회귀 모델 개발

파생변수를 추가하지 않은 데이터셋을 대상으로 단계 선택법을 적용했을 때 8개 변수 모두가 유의미한 변수로 나타났다. PCA를 통한 8개 주성분 중 6개 주성분을 사용하였을 때 전체 데이터 변동의 95.6%를 설명할 수 있었다. 이에 따른 MLR 모델의 설명변수는 Table 14와 같다.

파생변수를 추가한 데이터셋을 대상으로 단계 선택법을 적용했을 때 18개 변수 중 SS, SS_{t-2} , TN_{t-1} , TN_{t-2} , TP, TP_{t-2} 변수가 다중공선성 문제를 일으켜 소거되었다. PCA를 통한 18개 주성분 중 13개 주성분을 사용하였을 때 전체 데이터 변동의 95.2%를 설명할 수 있었다. 이에 따른 파생변수를 추가한 MLR 모델의 설명변수는 Table 15와 같다.

Table 14. Dataset without derived variable in B WTPP

MLR model	Input variables
Using Stepwise method	COD, SS, TN, TP, Q, E.coli, P, T
Using PCA	PCs 1 ~ PCs 6

Table 15. Dataset with derived variable in B WTPP

MLR model	Input variables
Using Stepwise method	BBOD _{t-1} , BOD _{t-2} , COD, COD _{t-1} , COD _{t-2} , SS _{t-1} , TN, TP _{t-1} , Q, E.coli, P, T
Using PCA	PCs 1 ~ PCs 13

(2) 모델 비교 및 평가

하수처리장의 기존 예측 방법을 포함한 7개의 예측 모델의 예측값과 실측값에 따른 R^2 와 RMSE를 통한 정량적 오차 검정을 Table 16과 Fig. 19로 나타내었다. 연구 결과, 과생변수를 추가한 DNN 모델이 가장 뛰어난 예측력을 보였으며 과생변수를 추가하지 않은 MLR 모델들을 제외한 네 가지 예측 모델이 기존 하수처리장의 예측 방법보다 높은 예측력을 보였다. 또한, 과생변수를 추가했을 때와 하지 않았을 때 모두 DNN 모델이 MLR 모델보다 높은 결과를 얻었다. 이를 통해 A 하수처리장의 유입 성상을 예측하는 데 있어 선형결합 모델보다 비선형결합 모델이 더 뛰어난 것을 확인하였다. 과생변수를 추가한 DNN 모델의 경우 과생변수를 추가하지 않은 DNN 모델보다 약 8% 개선된 성능을 보였고 과생변수를 추가한 MLR 모델 또한 과생변수를 추가하지 않은 MLR 모델보다 약 15% 개선된 성능을 보였다. 이는 과생변수 중 하루 전과 이틀 전의 BOD와 COD가 종속변수인 당일 BOD와 높은 상관관계를 가졌기 때문이라고 판단된다.

MLR의 경우 과생변수를 추가하지 않은 데이터셋에 PCA를 적용했을 때 단계 선택법보다 예측력이 4% 높게 나왔다. 반면, 과생변수를 추가한 데이터셋의 경우 단계 선택법을 적용했을 때 PCA보다 6.4% 높은 설명력을 가졌다. 이는 설명변수에 따라 생성된 주성분의 설명 가능한 변동이 다르기 때문이라고 판단된다.

Table 16 Comparison models in B WTPP

Forecasting model	R ²	RMSE
Original prediction method	0.813	18.62
DNN without derived variable	0.822	18.29
MLR without derived variable(Stepwise method)	0.744	21.04
MLR without derived variable(PCA)	0.784	19.13
DNN using derived variable	0.901	13.12
MLR using derived variable(Stepwise method)	0.890	13.58
MLR using derived variable(PCA)	0.826	17.11

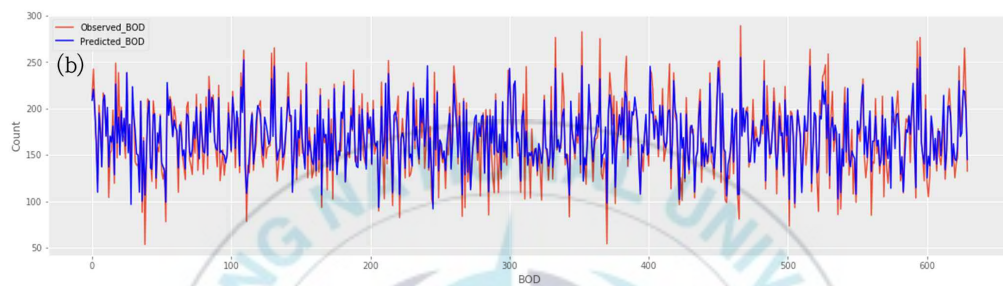
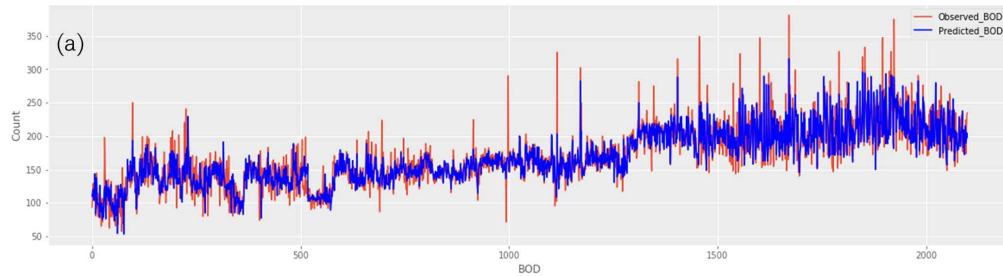
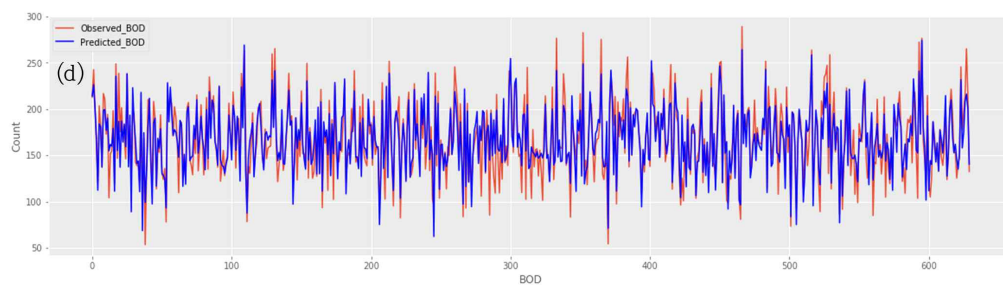
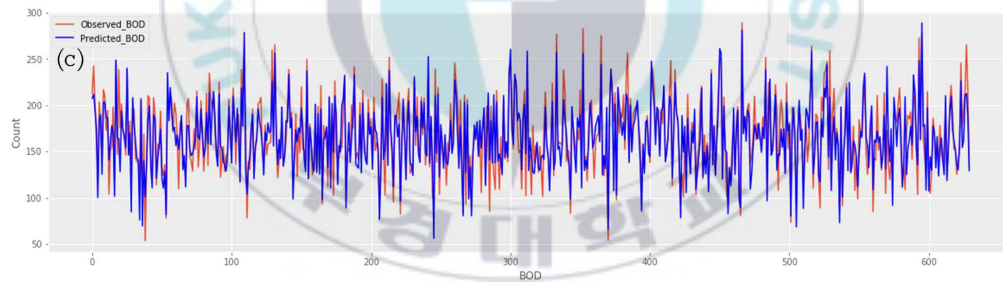
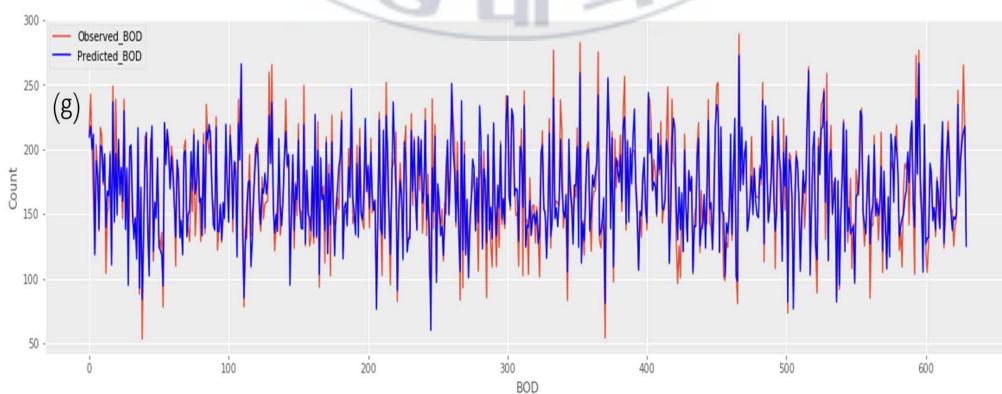
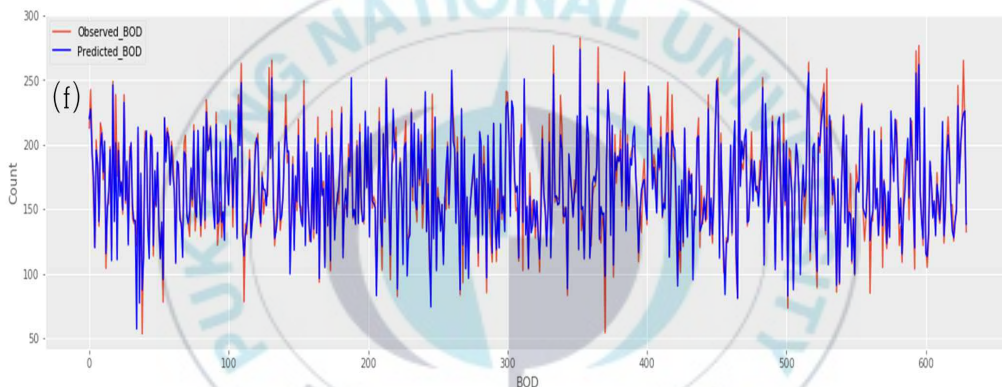
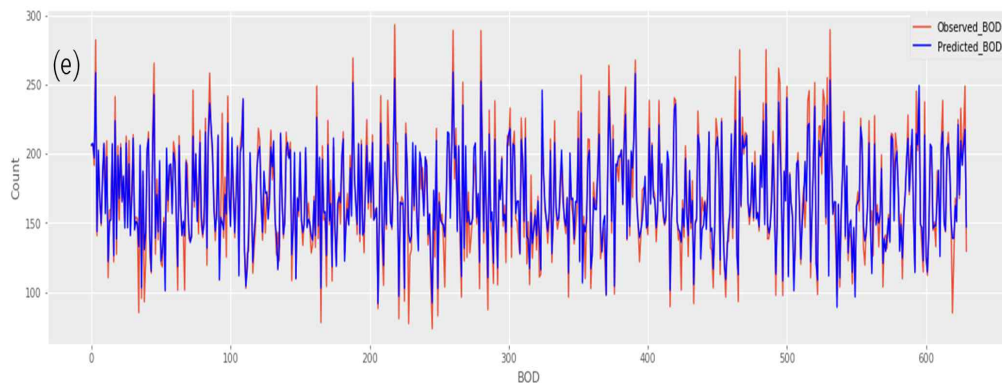


Fig. 19. (a) Original prediction method (b) DNN without derived variable



(c) MLR without derived variable(Stepwise method)

(d) MLR without derived variable(PCA)



(e) DNN using derived variable (f) MLR using derived variable (Stepwise method) (g) MLR using derived variable (PCA)

다. C 하수처리장

(1) 다중선형회귀 모델 개발

과생변수를 추가하지 않은 데이터셋을 대상으로 단계 선택법을 적용했을 때 기온 변수가 다중공선성 문제를 일으켜 소거되었다. PCA를 통한 8개 주성분 중 6개 주성분을 사용하였을 때 전체 데이터 변동의 97.0%를 설명할 수 있었다. 이에 따른 MLR 모델의 설명변수는 Table 17과 같다.

과생변수를 추가한 데이터셋을 대상으로 단계 선택법을 적용했을 때 18개 변수 중 TN_{t-2} , TP, TP_{t-1} , 강수량 변수가 다중공선성 문제를 일으켜 소거되었다. PCA를 통한 18개 주성분 중 11개 주성분을 사용하였을 때 전체 데이터 변동의 95.7%를 설명할 수 있었다. 이에 따른 과생변수를 추가한 MLR 모델의 설명변수는 Table 18과 같다.

Table 17. Dataset without derived variable in C WTPP

MLR model	Input variables
Using Stepwise method	COD, SS, TN, TP, Q, E.coli, P
Using PCA	PCs 1 ~ PCs 6

Table 18. Dataset with derived variable in C WTPP

MLR model	Input variables
Using Stepwise method	BOD_{t-1} , BOD_{t-2} , COD, COD_{t-1} , COD_{t-2} , SS, SS_{t-1} , SS_{t-2} , TN, TN_{t-1} , TP_{t-2} , Q, E.coli, T
Using PCA	PCs 1 ~ PCs 11

(2) 모델 비교 및 평가

하수처리장의 기존 예측 방법을 포함한 7개의 예측 모델의 예측값과 실측값에 따른 R^2 와 RMSE를 통한 정량적 오차 검정을 Table 19와 Fig. 20으로 나타내었다. 연구 결과, 파생변수를 추가한 DNN 모델이 가장 뛰어난 예측력을 보였으나 파생변수를 추가하지 경우 기존 하수처리장에서 사용하는 예측 방법보다 성능이 떨어졌다. 또한, 파생변수를 추가했을 때와 하지 않았을 때 모두 DNN 모델이 MLR 모델보다 높은 결과를 얻었다. 이를 통해 C 하수처리장의 유입 성상을 예측하는 데 있어 선형결합 모델보다 비선형적 모델이 더 뛰어난 것을 확인하였다. 특히, 파생변수를 추가하지 않은 MLR 모델의 경우 성능이 매우 떨어졌으며 파생변수를 추가한 후 성능이 약 67.2% 증가한 모습을 보였으며 DNN 모델 또한 파생변수를 추가했을 때 성능이 약 32.7% 증가하였다. 이는 기존 데이터셋에서 BOD와 높은 상관관계를 가지는 변수가 존재하지 않았으나, 파생변수를 추가했을 때 하루 전과 이틀 전의 BOD가 종속변수인 당일 BOD와 매우 높은 상관관계를 가졌기 때문이라고 판단된다.

MLR의 경우 파생변수를 추가하지 않은 데이터셋에 PCA를 적용했을 때 단계 선택법보다 예측력이 14.1% 높게 나왔지만, 파생변수를 추가한 데이터셋의 경우 단계 선택법을 적용했을 때 PCA보다 1.5% 높은 설명력을 보였다. 이는 설명변수에 따라 생성된 주성분의 설명 가능한 변동이 다르게 나타나기 때문이라고 판단된다.

Table 19. Comparison models in C WTPP

Forecasting model	R^2	RMSE
Original prediction method	0.532	22.35
DNN without derived variable	0.522	20.22
MLR without derived variable(Stepwise method)	0.160	28.05
MLR without derived variable(PCA)	0.301	24.73
DNN using derived variable	0.849	11.53
MLR using derived variable(Stepwise method)	0.832	11.98
MLR using derived variable(PCA)	0.817	12.51

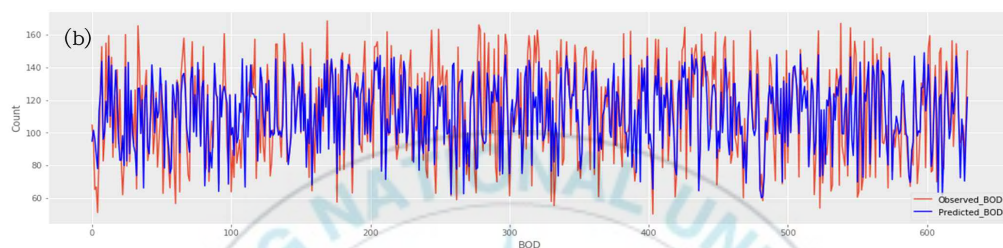
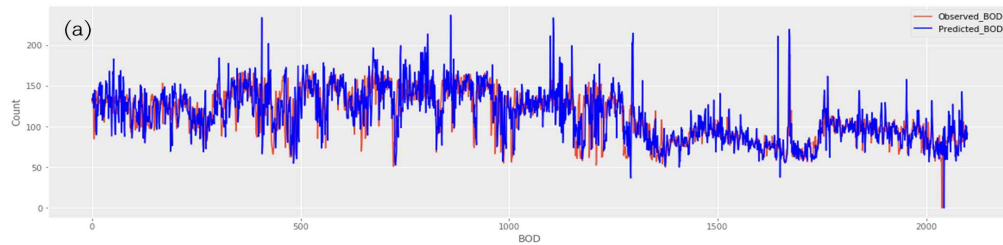
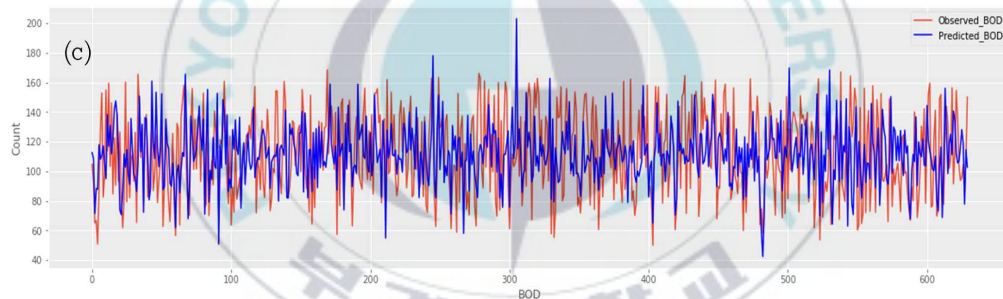
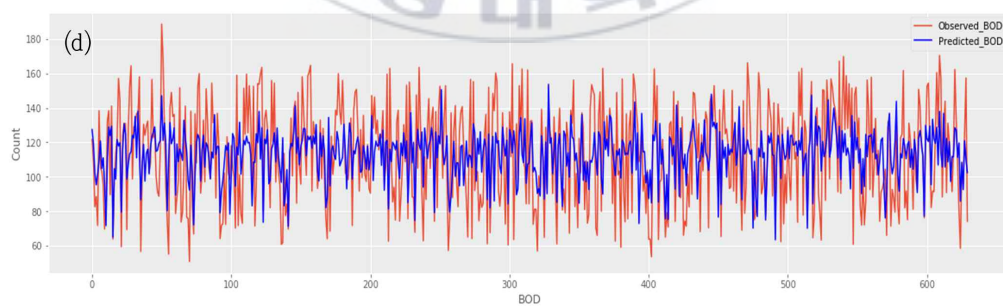


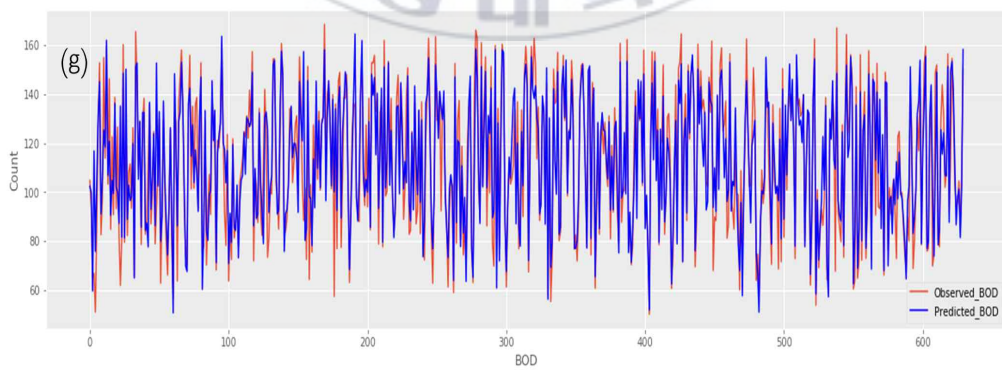
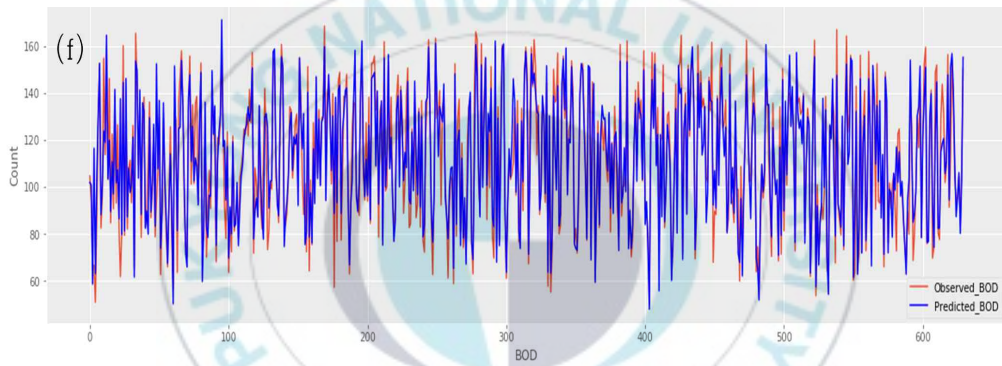
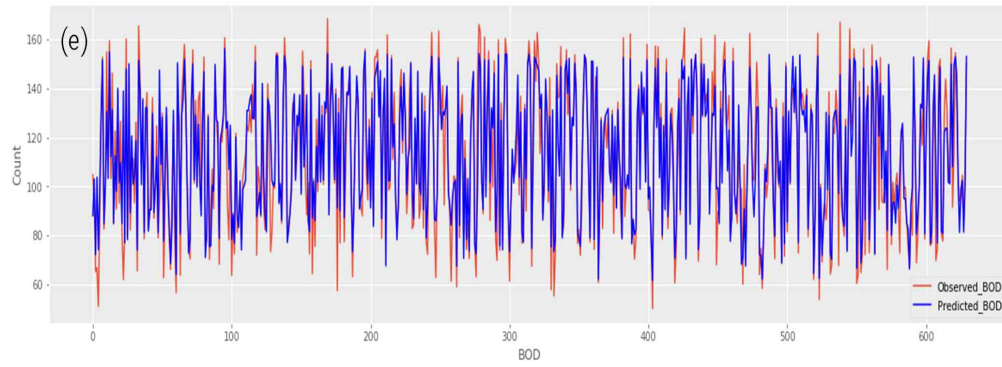
Fig. 20. (a) Original prediction method (b) DNN without derived variable



(c) MLR without derived variable(Stepwise method)



(d) MLR without derived variable(PCA)



(e) DNN using derived variable (f) MLR using derived variable(Stepwise method) (g) MLR using derived variable(PCA)

라. 강우시 예측력 평가

A, B 하수처리장의 경우 완전 분류식 하수관거를 사용하지만 C 하수처리장의 경우 일부 합류식 하수관거를 사용하기 때문에 비가 올 때 유기물이 하수처리장으로 들어가지 못하고 월류되는 현상이 나타날 수 있다. 따라서 예측력이 가장 우수했던 파생변수를 추가한 DNN 모델이 강우 시에도 높은 예측력을 보인다면, 향후 하수처리장으로 유입되지 않고 월류되는 유기물을 예측할 수 있는 기초자료로 활용할 수 있다고 판단하였다. 세 하수처리장의 전체 데이터 중 556개가 강수량이 측정되었고, 그중 검증 데이터셋에 포함된 165개 데이터에 따른 R^2 와 RMSE를 Table 20 ~ 22와 절대 오차 그래프는 Fig. 21 ~ 23으로 나타내었다.

연구 결과, A 하수처리장은 기존 하수처리장의 예측 방법보다 예측력이 약 10.8% 상승, B 하수처리장은 예측력이 약 10.8% 상승, 그리고 C 하수처리장은 예측력이 약 16.7% 상승한 결과를 보였다. 본 연구를 통해 DNN 모델이 강우 시에도 완전 분류식과 부분 합류식 하수처리장 모두 예측력이 우수한 것을 증명했다.

Table 20. BOD prediction in A WTPP while precipitating

Forecasting model	R^2	RMSE
Original prediction method	0.654	21.03
DNN using derived variable	0.762	14.62

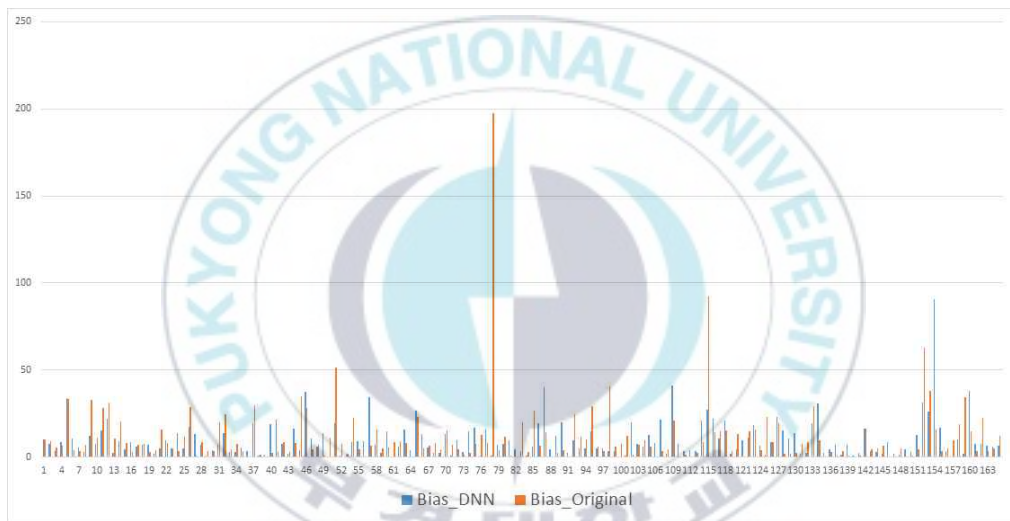


Fig. 21. Bias of predicted values in A WTPP while precipitating

Table 21. BOD prediction in B WTPP while precipitating

Forecasting model	R^2	RMSE
Original prediction method	0.812	19.81
DNN using derived variable	0.920	13.34

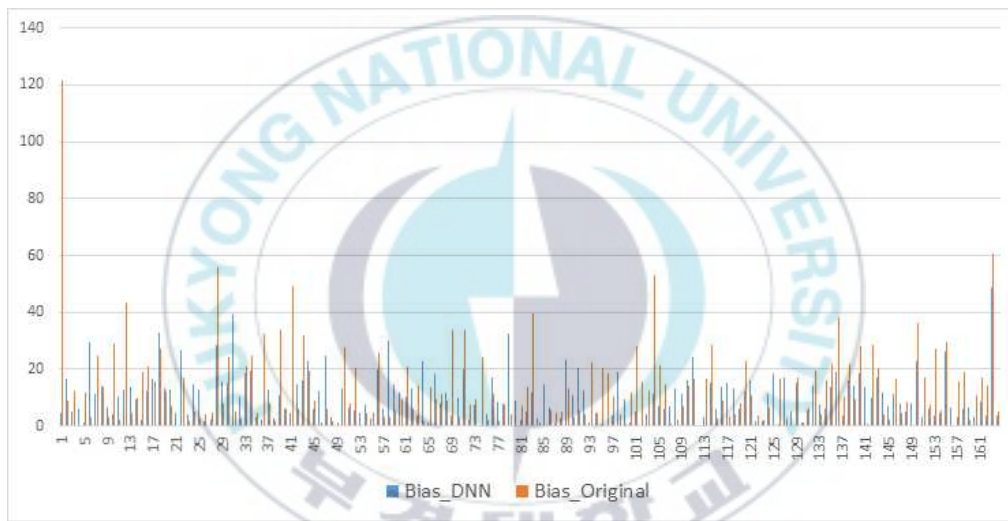


Fig. 22. Bias of predicted values in B WTPP while precipitating

Table 22. BOD prediction in C WTPP while precipitating

Forecasting model	R^2	RMSE
Original prediction method	0.665	19.58
DNN using derived variable	0.832	12.64

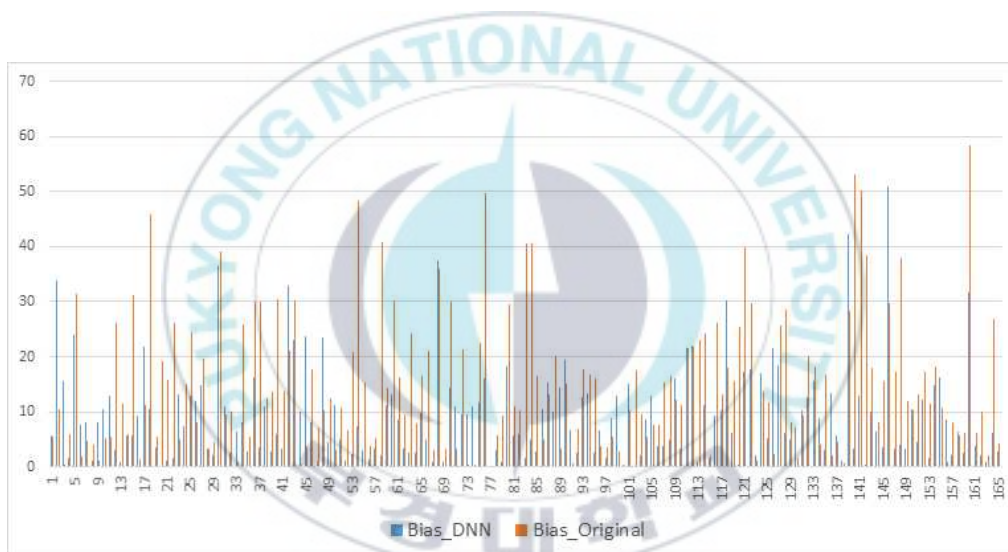


Fig. 23. Bias of predicted values in C WTPP while precipitating

마. 유입수 이상치 농도 예측력 평가

하수처리장 유입수의 BOD가 비정상적으로 높거나 낮은 경우가 발생할 수 있다. 이러한 이상치가 감지됐을 때 파생변수를 추가한 DNN 모델이 대응할 수 있는지 연구하였다. 본 연구는 관리도 기법(Control chart)을 적용해 3-시그마(표준편차) 범위 밖의 값을 이상치로 정의했다(Fig. 24). 관리도 기법이란, 1924년 벨연구소의 슈허트(W.A. Shewhart)가 공정이 안정 상태에 있는지를 조사하거나 공정을 안정 상태로 유지하기 위해 개발된 기법으로 평균에서 양쪽으로 3-시그마의 범위에 대부분의 값이 들어가기에 그 외의 값은 이상치로 간주한다. 이는 3-시그마 규칙(three-sigma rule)이라고도 한다.

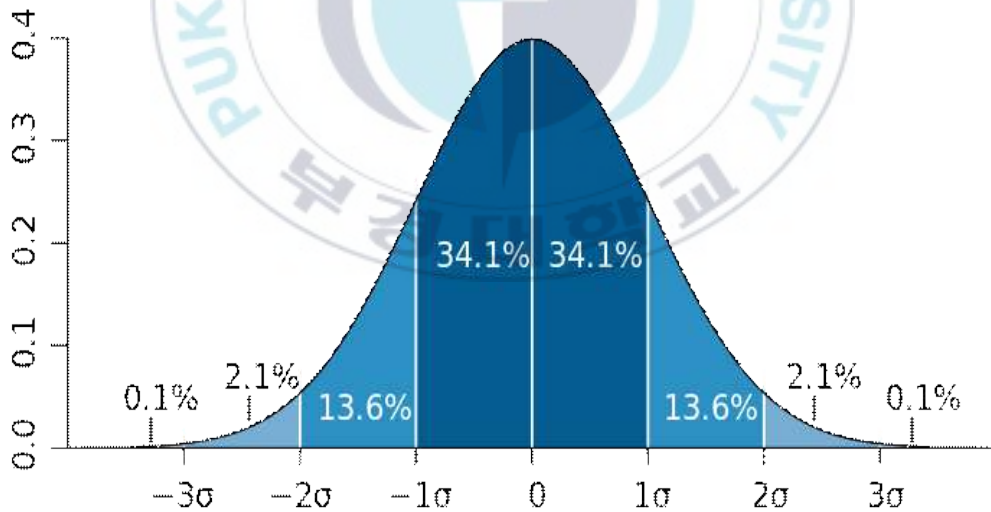


Fig. 24. Normal distribution with standard deviation

(1) A 하수처리장

A 하수처리장의 경우 BOD가 55.4 미만, 229.2를 초과하는 47개 데이터가 이상치로 관측되었고 이전 모델 학습 시, 검증 데이터 내 15개의 이상치 데이터가 포함된 것을 바탕으로 이를 바탕으로 이상치의 R^2 와 RMSE를 Table 23과 Fig. 25로 나타내었다. 연구 결과, 기존 하수처리장의 예측 방법보다 약 35.0% 우수한 예측력을 보였다. 하지만 모델의 결정계수에 비해 결정계수가 0.482로 떨어지는 성능을 보였다. 이는 3-시그마 범위 밖의 데이터 분포가 적기 때문에 해당 범위 학습이 제대로 이루어지지 않은 결과라 사료된다(Fig. 11). Fig. 25를 통해 예측 값이 실측값보다 낮은 값을 예측하는 것을 알 수 있는데 이는 데이터가 3-시그마 범위 밖 데이터보다 낮은 값에 집중적으로 학습되어 있기 때문이라고 판단된다.

(2) B 하수처리장

B 하수처리장의 경우 BOD 농도가 49.07 미만, 288.17을 초과하는 11개 데이터가 이상치로 관측되었고 이전 모델 학습 시, 검증 데이터 내 3개의 이상치 데이터가 포함된 것을 바탕으로 이상치의 R^2 와 RMSE를 Table 24와 Fig. 26로 나타내었다. 연구 결과, 기존 하수처리장의 예측방법보다 약 49.7% 우수한 예측력을 보였다. 하지만 모델의 결정계수에 비해 결정계수가 0.534로 떨어지는 성능을 보였다. 이는 3-시그마 범위 밖의 데이터 분포가 적기 때문에 해당 범위 학습이 제대로 이루어지지 않은 결과라 사료된다(Fig. 13). Fig. 26의 오차 값을 통해 예측값이 실측값보다 낮은 값을 예측하는 것을 알 수 있는데 이는 데이터가 3-시그마 범위 밖 데이터보다 낮은 값에 집중적으로 학습되어 있기 때문이라고 판단된다.

Table 23. Outlier prediction in A WTPP

Forecasting model	R^2	RMSE
Original prediction method	0.132	72.84
DNN using derived variable	0.482	38.69

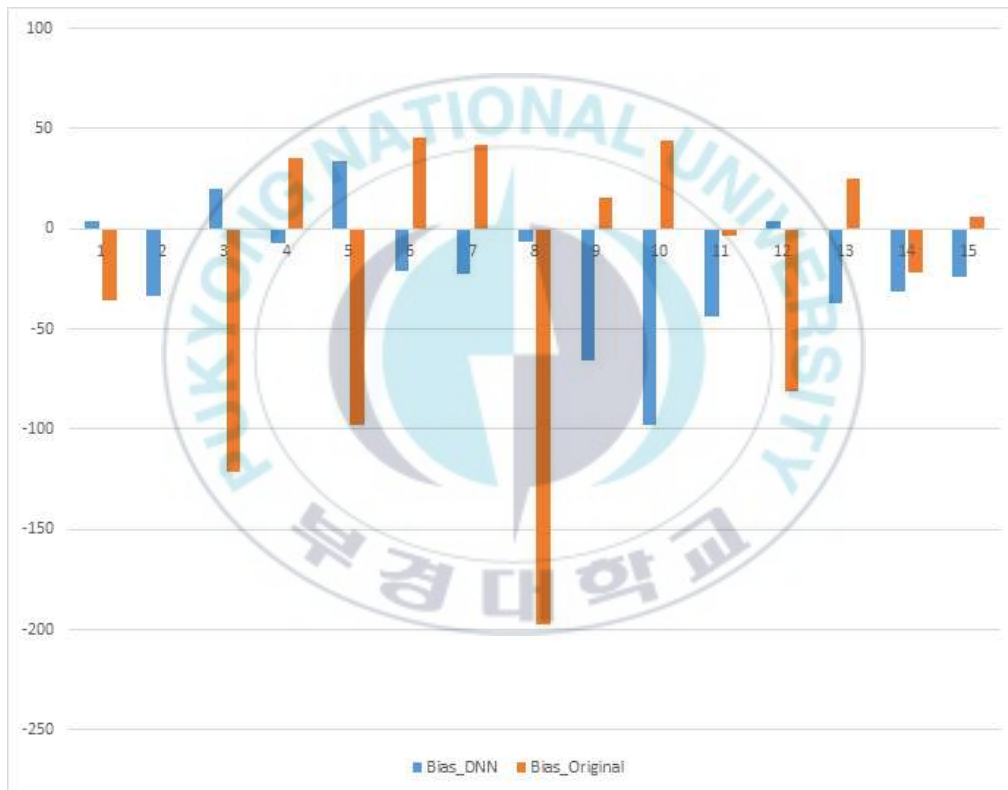


Fig. 25. Bias of predicted outlier values in A WTPP

Table 24. Outlier prediction in B WTPP

Forecasting model	R^2	RMSE
Original prediction method	0.037	55.00
DNN using derived variable	0.534	22.81

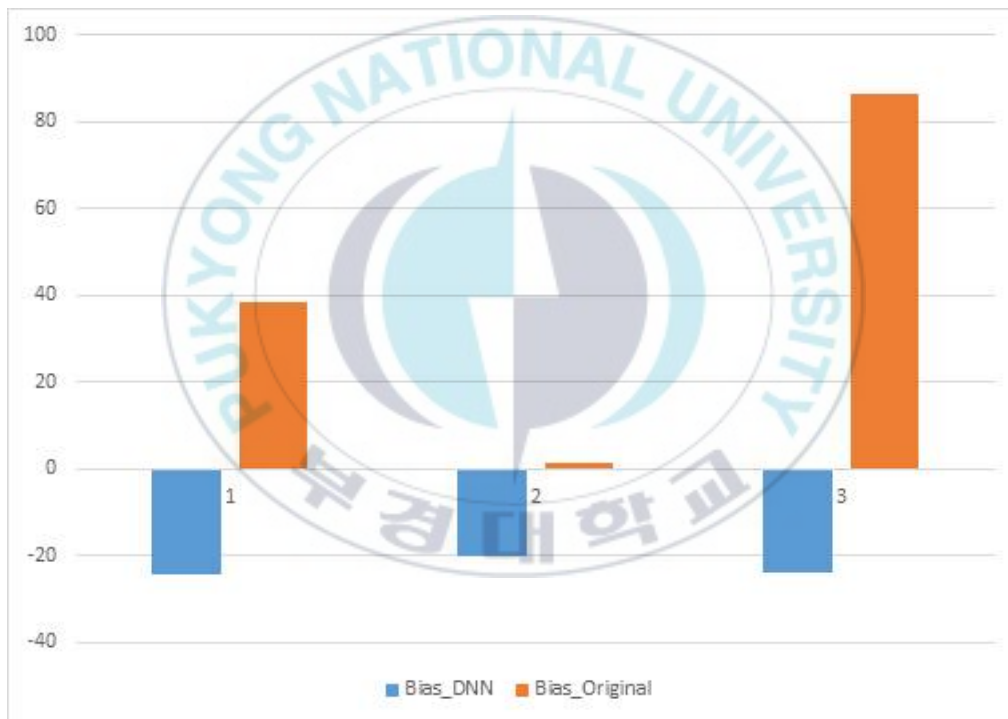


Fig. 26. Bias of predicted outlier values in B WTPP

V. 결론

본 연구에서는 유역 관리에 필요한 점오염원 관리대책으로 하수처리장을 대상으로 유입수의 BOD를 당일 예측하는 모델을 개발, 기존 하수처리장에서 사용하는 예측 방법보다 높은 예측력을 바탕으로 하수공정처리의 다양한 제어 전략을 사전에 제시하고 대응하여 공정과정의 효율성을 높이고 사후대처가 아닌 사전 감지 및 예방을 통한 법적 기준 초과 방류 문제를 해결할 수 있는 도구로 이용하고자 하였다. 이를 위해 2015년 1월 1일부터 2020년 9월 30일까지 B시 3개 하수처리장에서 측정된 2,100개의 일별데이터를 바탕으로 연구를 수행하였다. 유입수의 BOD를 결정짓는 유입 성상들의 복잡한 거동을 예측하기 위해 비선형함수로 이루어진 DNN 모델을 이용하였고 이를 검증하기 위해 단계 선택법과 주성분분석을 각각 적용한 MLR 모델과의 R^2 과 RMSE를 통한 모델 검정을 실시하였다. 또한, 유입수가 연속해서 들어오는 성질을 통해 직전 과거의 수질 성상이 당일 유입수의 BOD와 밀접한 상관관계가 있다고 판단하였다. 이에 따라 하루 전과 이틀 전의 수질 성상을 파생변수로 추가하여 모델에 적용하였고, 파생변수의 유효성을 검증하기 위해 파생변수를 추가하지 않은 데이터셋을 적용한 모델과의 예측력을 비교 및 평가하였다.

연구 결과, DNN 모델이 기존 하수처리장 예측방법 및 MLR 모델보다 뛰어난 성능을 보이는 것을 증명하여 하수처리장에 비선형 모델이 적용되었을 때 공정 효율을 높일 수 있다는 가능성을 제시했다. 또한, 세 하수처리장 모두 파생변수를 추가하였을 때가 추가하지 않았을 때보다 DNN, MLR 모델 모두 성능이 개선된 것을 확인하여 파생변수가 유효하다는 것을 검증했다. 특히 C 하수처리장은 파생변수를 추가하였을 때 DNN 모델이

32.8%, MLR 모델이 67.2% 성능 증가를 보이며 A, B 하수처리장보다 월등히 높은 예측력 변화를 보였는데 이는 C 하수처리장의 기존 설명변수가 종속변수와의 상관관계가 미미한 반면, 추가된 파생변수와는 매우 높은 상관성을 가졌기 때문이라 판단된다. 따라서 파생변수를 사용할 때 높은 예측력을 얻기 위해선 파생변수와 종속변수 간 상관관계를 확인하는 것이 가장 중요하다고 사료된다.

선행연구 사례를 통해 얻은 신경망의 예측력 검증을 본 연구 결과를 통해서도 확인할 수 있었으며, 그 대상인 하수처리장에서 사용하는 예측 방법 및 선형결합 모델을 대체할 수 있는 비선형 모델의 적용 가능성을 제시하였다.

또한, 강우 시에도 DNN 모델의 BOD 예측력이 뛰어난 것을 검증했다. 강우 값이 포함된 556개의 데이터를 바탕으로 기존 하수처리장의 예측 방법에 비해 A 하수처리장이 10.8%, B 하수처리장이 10.8%, C 하수처리장이 16.7%의 예측력 상승을 보이며 세 하수처리장 모두 DNN 모델이 높은 예측력을 보였고, 이를 통해 합류식 하수관거를 사용하는 하수처리장의 유입 BOD 예측을 통한 월류되는 유기물량을 예측할 수 있는 기초자료로 사용할 수 있는 근거를 제시하였다.

마지막으로 유입수의 이상 농도가 들어왔을 때 예측여부에 대한 연구를 진행하였다. 관리도 기법을 적용하여 데이터를 분석한 결과, C 하수처리장을 제외한 A, B 하수처리장에 각각 47개, 11개의 이상치 데이터가 존재하였다. 이를 바탕으로 모델을 검증한 결과, A 하수처리장과 B 하수처리장 모두 기존 하수처리장의 예측 방법보다 약 35.0%, 49.7% 우수한 예측력을 보였다. 하지만 모델의 결정계수는 각각 0.482, 0.534로 떨어지는 성능을 보였는데, 이는 3-시그마 범위 밖의 이상치 데이터 개수가 적기 때문에 충분한 학습이 이루어지지 않은 결과라 판단된다. 따라서 향후 추가적인 이상

데이터를 확보하여 학습 불균형 문제를 해결한다면 보다 나은 결과를 얻을 수 있을 거라 사료된다.

본 연구의 한계점도 존재했다. 본 연구는 자료 수집의 제한으로 정보공개 청구로 일별 데이터를 확보하였으며, 2,100개의 데이터를 이용했지만 실시간으로 변하는 수질 성상을 분석하는데 한계가 있었다. 또한 BOD 특성상 5일 뒤에 실측값이 나오기 때문에 현장 적용 시 과거 4일간의 BOD 실측값이 누락된다. 누락된 데이터는 보간법이나 파생변수를 추가하지 않은 DNN 모델을 통해 얻은 예측 BOD로 보완 가능하지만, 예측값을 통한 새로운 예측값은 정확도가 떨어지게 된다. 따라서 현장 적용 시, 수질 성상을 예측할 수 있는 다른 변수 연구와 함께 BOD₃을 사용하여 누락 일수를 줄이는 방향으로 파생변수 적용을 위한 추가적인 연구가 필요하다 사료된다.

그럼에도 불구하고 B 시 3개 하수처리장에 대해 DNN 모델로 예측 분석한 결과, 하수처리장의 기존 예측 방법 및 MLR 모델보다 높은 예측력을 제시했고 대부분의 수질 성상을 고려하여 설명변수로 활용하였으며, 유입수의 연속성을 고려한 과거 파생변수를 활용하여 유효성을 입증한 점에서 본 연구의 의의를 찾을 수 있다. 또한 본 연구는 세 하수처리장의 DNN 모델의 은닉층 구조를 동일하게 설정하여 연구를 진행하였으나, 데이터 수치값에 따라 모델 구조를 달리하였을 때 보다 높은 예측력을 얻을 수 있다. 따라서 향후 각 하수처리장의 데이터에 맞는 DNN 모델 구조를 구축한다면 보다 개선된 예측 결과를 얻을 수 있을 것이라 기대된다.

참고문헌

- An-jie Li, Xiao-yan Li, Han-qing Yu, 2011, "Effect of the food-to-microorganism(F/M) ratio on the formation and size of aerobic sludge granules, Process Biochemistry", Vol. 46, No. 12, pp. 2269-2276.
- Emrah Dogan, Asude Ates, Ece Ceren Yilmaz, Beytullah Eren, 2008, "Application of Artificial Neural Networks to Estimate Wastewater Treatment Plant Inlet Biochemical Oxygen Demand", Environmental Progress, Vol. 27, No. 4, pp. 439-446.
- Fatemeh Nargesian, Horst Samulowitz, Udayan Khurana, Elias B. Khalil, Deepak Turaga, 2017, "Learning Feature Engineering for Classification", Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, page. 2529-2535.
- Handan Camdevyren, Nilsun Demyr, Arzu Kanik, Syddyk Kesryn, 2005, "Use of principal component scores in multiple linear regressionmodels for prediction ofChlorophyll-ain reservoir", Ecological Modelling, Vol 181, No. 4, pp. 581-589.
- Hervé Abdi, Lynne J. Williams, 2010, "Principal component analysis", WIREs Computational Statistics, Vol. 2, No. 4, pp. 433-459.
- Jack V.Tu, 1996, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes", Vol. 49, No. 11, pp. 1225-1231.
- Jalili Ghazi Zade, M., Noori, R., 2008, "Prediction of Municipal Solid Waste Generation by Use of Artificial Neural Network: A Case

- Study of Mashhad”, International Journal of Environmental Research, Vol. 2, No. 1, pp. 13-22.
- Maged M. Hamed, Mona G. Khalafallah, Ezzat A. Hassanien, 2004, “Prediction of wastewater treatment plant performance using artificial neural networks”, Environmental Modelling & Software, Vol. 19, No. 10, pp. 919-928.
- Mahmoud S.Nasr, Medhat A.E.Moustafa, Hamdy A.E.Sief, GalalEl Kobrosy, 2012, “Application of Artificial Neural Network (ANN)for the prediction of EL-AGAMY wastewater treatmentplant performance-EGYPT”, Alexandria Engineering Journal, Vol. 51, No. 1, pp. 37-43.
- Minsky M. and Papert S., 1969, Perceptrons: An Introduction to Computational Geometry. MIT Press
- Mohammad Ali Abdoli, Maliheh Falah Nezhad, Reza Salehi Sede, Sadegh Behboudian, 2011, “Longterm Forecasting of Solid Waste Generation by the Artificial Neural Networks”, Environmental Progress & Sustainable Energy, Vol. 31, No. 4, pp. 628-636.
- Mohamed F.Hamoda, Ibrahim A.Al-Ghusain, Ahmed H.Hassan, 1999, “Integrated wastewater treatment plant performance evaluation using artificial neural networks”. Water Science and Technology, Vol. 40, No. 7, pp. 55-65.
- Nikolay Chumerin, Marc M. Van Hulle, 2006, “Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information”, IEEE Workshop on Machine Learning for Signal Processing, pp. 343-348.

- Ning Guo, Jian Zhang, Hui-Jun Xie, Lin-Rui Tan, Jie-Ning Luo, Ze-Yu Tao, Shu-Guang Wang, 2017, "Effects of the Food-to-Microorganism (F/M) Ratio on N₂O Emissions in Aerobic Granular Sludge Sequencing Batch Airlift Reactors", *Water*, Vol. 9, No. 7, pp. 477.
- Rania Ahmed Hamza, Zhiya Sheng, Oliver Terna Iorhemen, Mohamed Sherif Zaghloul, Joo Hwa Tay, 2018, "Impact of food-to-microorganisms ratio on the stability of aerobic granular sludge treating high-strength organic wastewater", *Water Research*, Vol. 147, No. 15, pp. 287-298.
- Richard D. De Veaux, Lyle H. Ungar, 1994, "Multicollinearity: A tale of two nonparametric regressions", *Selecting Models from Data*, pp. 393-402.
- R Noori, MA Abdoli, M Jalili Ghazizade, R Samieifard, 2009, "Comparison of Neural Network and Principal Component Regression Analysis to Predict the Solid Waste Generation in Tehran", *Iranian Journal of Public Health*, Vol. 38, No. 1, pp. 74-84.
- Samina Khalid, Tehmina Khalil, Shamila Nasreen, 2014, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning", *Science and Information Conference 2014*, pp. 372-378.
- Werbos P., 1974, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," PhD thesis, Harvard University

- Yi-Ju Wu, Liang-Ming Whang, Ming-Yu Chang, Toshikazu Fukushima, Ya-Chin Lee, Sheng-Shung Cheng, Shu-Fu Hsu, Cheng-Huey Chang, Wason Shen, Charn-Yi Yang, Ryan Fu, Tsair-Yuan Tsai, 2013, “Impact of food to microorganism (F/M) ratio and colloidal chemical oxygen demand on nitrification performance of a full-scale membrane bioreactor treating thin film transistor liquid crystal display wastewater”, *Bioresource Technology*, Vol. 141, pp. 35-40.
- Yin Liu, Haining Liu, Li Cui, Kaisong Zhang, 2012, “The ratio of food-to-microorganism (F/M) on membrane fouling of anaerobic membrane bioreactors treating low-strength wastewater”, *Desalination*, Vol. 297, No. 3, pp. 97-103.
- Yoon-Seok TimothyHong, Michael R.Rosen, Rao Bhamidimarri, 2003, “Analysis of a municipal wastewater treatment plant using a neural network based pattern analysis”, *Water Research*, Vol. 37, No. 7, pp. 1608-1618.
- 강승호, 정인선, 임형석, 2018, “실시간 공격 탐지를 위한 Pearson 상관계수 기반 특징 집합 선택 방법”, *융합보안논문지*, Vol. 18, No. 5, pp. 59-66.
- 김경준, 이수동, 전치혁, 박개명, 변상수, 2017, “연료 소비 패턴 발견을 위한 컨테이너선 운항데이터 분석의 통계적 절차”, *응용통계연구*, Vol. 30, No. 5, pp. 633-645.
- 김동균, 최정석, 2021, “신경망기법을 활용한 선박 가치평가 모델 개발”, *해양환경안전학회지*, Vol. 27, No. 1, pp. 13-21.
- 김동현, 김정옥, 곽재원, 아이미, 김종성, 김형수, 2020, “딥러닝을 활용한

- 산지습지 수위 예측 모형 개발”, 한국습지학회지, Vol. 22, No. 2, pp. 106-112.
- 김영선, 2019, “머신 러닝을 활용한 폐전자제품 발생량 예측 모델 연구”, 경기대학교 석사학위논문
- 김예진, 2006, “연속회분식 공정관리를 위한 추론모델 및 인공지능형 진단 알고리즘 개발”, 부산대학교 박사학위논문
- 김정환, 박지현, 최창현, 김형수, 2018, “호우피해자료에서의 고차원 자료 및 다중공선성 문제를 해소한 회귀모형 개발”, 대한토목학회 논문집, Vol. 38, No. 6, pp. 801-808.
- 류나현, 김형석, 강필성, 2016, “다중선형회귀모형에서의 변수선택기법 평가”, 대한산업공학회지, Vol. 42, No.5, pp. 314-326.
- 박영기, 김영일, 김슬기, 2006, “하·폐수 고도처리시 F/M비 변화에 따른 유기산 잔류 농도가 슬러지 침강성에 미치는 영향”, 대한환경공학회지, Vol. 28, No. 3, pp. 42-47.
- 사이토 고키, 밑바닥부터 시작하는 딥러닝, 2017, 한빛미디어
- 송기준, 고민진, 박찬미, 임길섭, 장양수, 김동기, 2005, “단계적 회귀분석을 이용한 혈중 지질농도에서 유전자의 상대적 효과 분석”, 대한심장학회, Vol. 35, No. 10, pp. 759-765.
- 송민구, 김선배, 2013, “빅데이터 분석방법을 이용한 예측모형의 신뢰도 향상에 관한 연구”, 디지털정책연구, Vol.11, No.6, pp.103-112.
- 서형준, 정상원, 박미옥, 이병렬, 2001, “승기천 유역의 오염물 유출특성 및 수질관리방안 연구”, 한국청정기술학회, Vol.7, No.4 pp. 251-263
- 안성만, 2016, “딥러닝의 모형과 응용사례”, 지능정보연구 Vol.22, No.2, pp.127-142.
- 안준수, 박옥근, 조정호, 2011, “공장폐수가 혼합된 하수처리장의 하수처

- 리 특성”, 한국산학기술학회논문지, Vol. 12, No. 7, pp. 3341-3352.
- 안철우, 박진식, 문추연, 2000, “기존 하수처리장에서의 처리효율 개선에 관한 연구”, 대한위생학회지, Vol. 15, No. 3, pp. 50-56.
- 이경훈, 문병석, 김태경, 양치열, 2002, “폐기물 예측인자의 특성과 신경망을 이용한 발생량 예측”, 한국물환경학회 공동 추계학술발표회, Vol. 2002, pp. 465-468.
- 이남정, 김성민, 정일주, 손석만, 이승철, 2020, 규칙기반과 딥러닝을 동시에 활용한 양상블 회전체 이상진단, 한국소음진동공학회, Vol.30, No.2, pp. 129-135.
- 이문희, 이종삼, 한상국, 2005, “영산강유역 생활하수처리장 방류수에서의 유기오염물질 분포특성에 관한 연구”, 이문희, 이종삼, 한상국, 대한환경공학회지 Vol. 27, No. 12, pp. 1332-1339.
- 이중원, 홍윤석, 서창원, 신항식, 2008, “인공 신경망을 이용한 대규모 BNR 하수처리장 장기간의 운전 성능 예측”, 한국물환경학회·대한상하수도학회 공동 추계학술발표회, Vol. 2008, No. 2, pp. 198-199.
- 정광욱, 윤춘경, 장재호, 김형철, 2006, “유역의 토지이용과 오염원 현황이 수질특성에 미치는 영향 분석”, 한국하천호수학회, Vol. 39, No. 1, pp.41-51.
- 정형석, 이상형, 신항식, 송의열, 2006, “대도시 하수종말처리장 유입 하수의 성상 평가와 인공신경망을 이용한 구성성분 농도 예측”, 대한환경공학회지, Vol. 28, No. 3, pp. 308-315.
- 최광미, 정유정, 2020, “딥러닝 합성곱 신경망을 이용한 효율적인 홍채인식”, 한국전자통신학회 논문집, Vol. 15, No. 3, pp. 521-526.
- 최성원, 김한준, 2019, “다층신경망을 이용한 장래 생활폐기물 발생량 예측 연구 : B시 사례연구”, 한국폐기물자원순환학회지, Vol. 37, No. 1,

pp. 37-43.

최정현, 박석순, 2006, “효과적인 수질관리를 위한 소유역 중심의 유역관리”, 대한환경공학회, Vol.28, No.8, pp. 820-825.

최재훈, 문태섭, 김예진, 김성희, 김효수, 김창원, 2007, “인공신경망을 이용한 하수처리장 유입수 및 공정성능 예측모델 개발”, 공동 추계학술발표회 Vol. 2007, No. 0, 116-126pp.

허남효, 최광수, 김창원, 김병균, 강호, 1999, “활성슬러지공정 최적화를 위한 호흡율측정에 기초한 제어 전략”, 한국물환경학회지 Vol.15, No.2, pp. 217-227.

홍준호, 오민지, 조용빈, 이경희, 조완섭, 2020, “다차원 데이터의 군집분석을 위한 차원축소 방법: 주성분분석 및 요인분석 비교”, 한국빅데이터학회지, Vol. 5, No. 2, pp. 135-143

하수도법, 2022