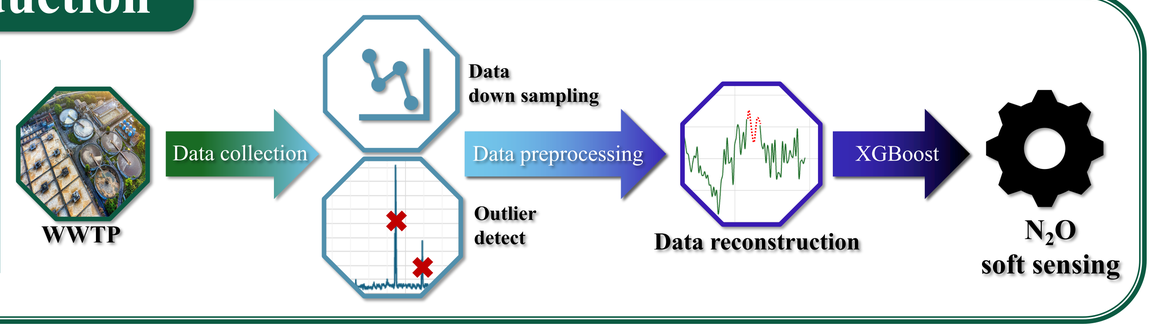
Water AI-ready data 기반 N_2O 소프트 센싱 알고리즘 개발

김선주[1], 허성구[1†]
[1]강원대학교 환경공학전공 †sungku.heo@kangwon.ac.kr

Introduction

- 하수처리장에서 배출되는 주요 온실가스 중 N_2O는 다른 온실가스 대비 GWP가 약 273배 높으므로 물 분야 탄소중립 달성을 위해서는 N_2O저감이 매우 중요함
- 현실적으로 국내 모든 하수처리장에 N_2O센서를 설치하는 것이 불가능 하기 때문에 본 연구에서는 이미 측정 가능한 수질 데이터를 활용하여 N_2O 배출량을 실시간으로 예측할 수 있는 소프트 센싱 알고리즘을 개발함



Materials & methods

Data collection

Originated from dataset reported by Daelman et al. (2015).

- 네덜란드 Kralingseveer 하수처리장의 유입량과 북쪽 Carrousel reactor에서 16개월 (2011.03.11 ~ 2012.01.19) 동안 측정된 9개 성분 농도 데이터를 확보함.

Data preprocessing

1. 균일하지 않은 시간 간격의 데이터를 1시간 간격으로 down sampling하여 time series resampling.
2. EWMA기반 1.5σ rule을 사용해 outlier detect.

Exponential weighted moving average (EWMA)

$$y_t = (1- \alpha)y_{t-1} + \alpha x_t$$

- α: specify smoothing factor between 0 and 1.
- y_t: EWMA of the estimated data at the time t.
- x_t: current measured data point at the time t.

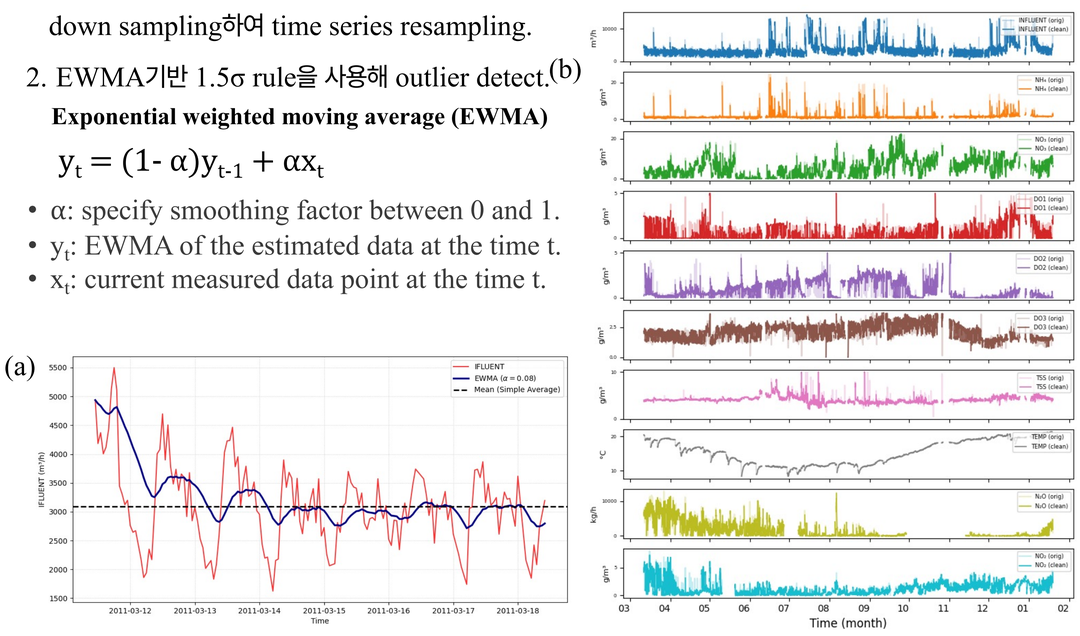


Fig. 1 (a) Comparison of EWMA and simple average for influent. (b) EWMA-based (α = 0.08) outlier detection using the 1.5σ rule.

Table. 1 Raw counts and outlier removal (%) per variable.

	influent	NH_4	DO1	DO2	DO3	NO_3	NO_2	TSS	temp	N_2O
# of raw	6,923	6,914	6,916	6,919	6,919	6,914	7,044	6,915	6,919	5,034
Cleaned %	9.24	8.06	8.88	9.24	8.38	7.88	8.05	8.59	8.41	8.66

Data reconstruction

Categorical boosting algorithm (Catboost algorithm)

- Catboost 알고리즘은 범주형 변수 처리에 특화되어 있어 높은 예측 성능을 위해 derived feature를 추가해 $R^2 \simeq 0.96$의 높은 정확도를 보임

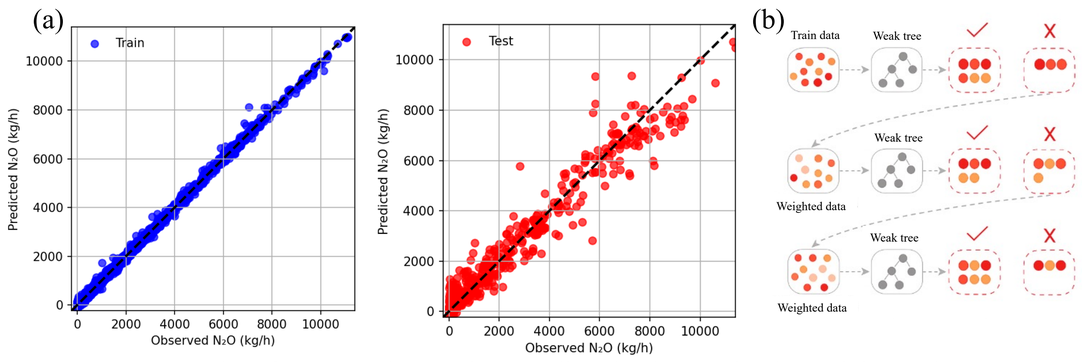


Fig. 2 (a) Q-Q plots for N_2O (observed vs predicted). (b) Mechanism of the Catboost algorithm.

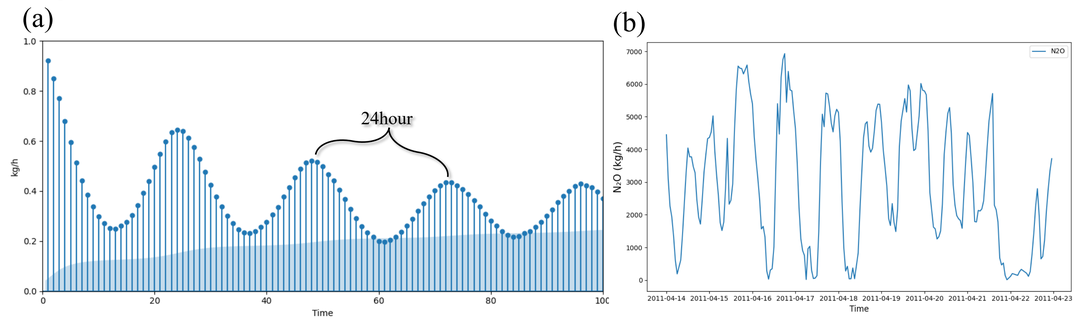


Fig. 3 (a) Autocorrelation of N_2O after reconstruction by model. (b) N_2O concentration (2011.04.14 ~ 2011.04.23) after data reconstruction

- 2011.06.11 ~ 2011.06.13 데이터의 8개 feature가 측정이 안되었기 때문에 2011.06.09까지의 데이터만 사용하여 reconstruction하였음.
- Data reconstruction 이후에도 뚜렷한 24시간 주기를 띄었기 때문에 기존 데이터의 주기성을 깨지 않고 data reconstruction하였음을 확인함.

Results & discussion

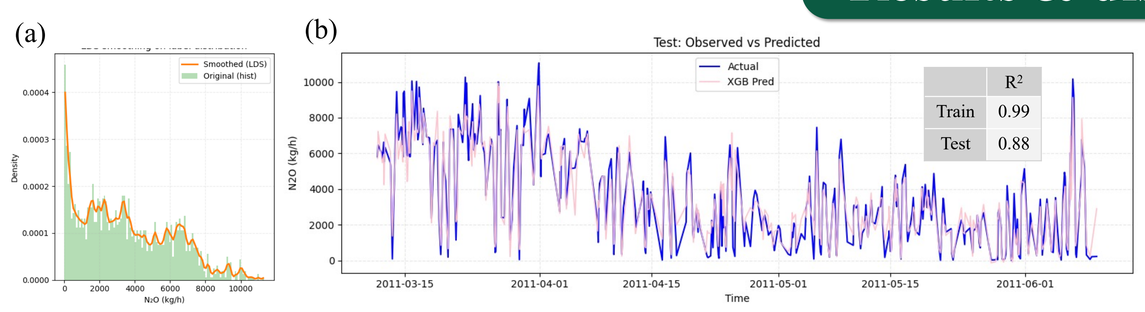


Fig. 4 (a) Comparison between the original and LDS-smoothed N_2O data. (b) line graph for N_2O (actual vs predicted) and R^2 value between observed and predicted N_2O by XGBoost algorithm.

- Label distribution smoothing (LDS) 및 extreme gradient boosting (XGBoost) 알고리즘을 사용하여 preprocessing된1시간 간격의 90일 ready-data로 N_2O 농도를 예측함.
- 약 3개월간의 데이터만으로 $R^2 \simeq 0.88$의 우수한 정확도를 보여 결측 데이터가 적은 data를 학습하여 reconstruction 및 학습하면 더 높은 정확도를 기대할 수 있음.
- Train set의 정확도는 100%에 근접한 것에 비해 test set의 정확도가 88%로, regularization을 통해 overfitting을 방지할 수 있을 것이라고 기대됨.

Conclusions

- 본 연구에서는 결측치가 존재하는 데이터도 imputation 및 reconstruction하여 target feature 예측에 사용할 수 있음을 보임.
- LDS 및 EWMA를 통해 정규화되지 않은 데이터의 분포를 smoothing하여 노이즈가 있는 데이터에도 유연하게 적용할 수 있도록 함.
- 이미 빅데이터화 된 하수처리장의 수질 ready-data를 사용하여 측정이 불가능했던 성분의 농도를 실시간으로 예측할 수 있는 가능성을 제시함.
- 더 많은 derived feature를 추가해 장기간 결측된 기간도 예측 가능하도록 하며 RNN 및 transformer알고리즘을 통하여 예측 정확도를 높이고자 함.

Acknowledgement 이 성과는 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-24852970)