Journal of Wetlands Research
Vol. 27, No.4, November 2025, pp.348-356

ISSN 1229-6031 (Print) / ISSN 2384-0056 (Online)
DOI https://doi.org/10.17663/JWR.2025.27.4.348

하천 수온 측정 자료의 결측 유형별 보간 기법 적용 특성 비교

김준오* · 박정수**,†

*국립한밭대학교 환경공학과
**국립한밭대학교 건설환경공학과

Comparative Analysis of the Characteristics of Imputation Methods for Different Types of Missing Data in River Water Temperature Measurements

Juneoh Kim* · Jungsu Park**,†

Department of Environmental Engineering, Hanbat National University, Korea
***Department of Civil and Environmental Engineering, Hanbat National University, Korea*

(Received : 21 July 2025, Revised : 30 September 2025, Accepted : 05 November 2025)

요 약

다양한 수질 측정자료가 오염원 추적 및 수질 환경 평가 등에 널리 활용되고 있으며, 이에 따라 현장 모니터링을 통한 수질자료 취득을 위한 노력이 계속되고 있다. 하지만 현장 수질 모니터링의 특성상 센서 오류, 고장 및 강우에 따른 재해 등 다양한 원인에 따른 결측이 발생할 수 있으며 수질 모니터링 결과의 신뢰도 확보를 위한 결측 관리의 중요성이 커지고 있다. 본 연구에서는 현장의 수질 특성을 확인할 수 있는 수질 환경 변수 중 하나인 수온에 대하여 4가지 유형의 결측을 생성하고, 2개의 통계기반 보간 기법인 선형 보간 (Linear)과 다항 보간 (Polynomial) 그리고 2개의 머신러닝 기반 모형인 K-Nearest Neighbors (KNN) 및 autoencoder (AE)를 적용한 총 4개의 보간 모형을 적용하여 성능을 비교하였다. 4개의 결측 유형은 단기 결측 (Case 1), 장기결측 (Case 2), 첨두 구간 전후의 급격한 수질변화 구간의 결측 (Case 3), 첨두 및 저점을 포함한 장기간의 수질변화 구간의 결측 (Case 4)으로 구분되었다. 분석결과 단기 결측이 발생되는 Case 1 및 3에서는 Linear 모형이 RSR 0.26 및 0.76으로 가장 우수한 보간 성능을 보였으며, 장기간의 결측을 포함하는 Case 2이 경우 AE가 RSR 0.63으로 가장 우수한 성능을 보이는 것을 확인하였다. Case 4는 KNN (k=3)의 RSR 이 0.66으로 가장 우수한 성능을 보였으며, AE의 RSR이 0.68로 KNN에 비해 다소 낮은 성능을 보였지만 그 차이는 크지 않았다. 본 연구를 통해 결측 유형에 따라 보간 모형의 성능에 차이가 있음을 확인할 수 있었다.

핵심용어 : 오토인코더, 보간 모형, 머신러닝, 결측 자료 보간, 수온

Abstract

Various water quality measurements are used to track pollution sources and assess water environments. As such, efforts to collect water quality data through field monitoring continue to expand. However, due to the nature of field monitoring, missing values are often observed as a result of sensor errors, equipment failures, and external factors such as rainfall or disasters. This highlights the growing importance of managing missing data to ensure the reliability of water quality monitoring results. This study generated four types of missing patterns for water temperature, a key indicator of field water quality conditions. Then, four imputation methods were applied. The methods included two traditional statistical approaches (linear interpolation and polynomial interpolation) and two machine learning models (K-nearest neighbors (KNN) and autoencoder (AE)). The four missing data scenarios were defined as follows: short-term missing (Case 1), long-term missing (Case 2), missing around peak values with rapid water quality change (Case 3), and extended missing periods including both peaks and troughs (Case 4). The results showed that the linear model achieved the best performance for Cases 1 and 3, with RSR values of 0.26 and 0.76, respectively. For Case 2, AE achieved the highest performance with an RSR of 0.63. In Case 4, KNN (k=3) showed the best result with an RSR of 0.66, followed closely by AE with an RSR of 0.68. These findings indicate that imputation performance varies depending on the missing data pattern.

Key words : autoencoder, imputation model, machine learning, missing data imputation, water temperature

†To whom correspondence should be addressed.
Department of Civil and Environmental Engineering, Hanbat National University
E-mail : parkjs@hanbat.ac.kr
• **Juneoh Kim** Hanbat National University, Korea/Ph.D. Student(juneohkim@edu.hanbat.ac.kr)
• **Jungsu Park** Hanbat National University, Korea/Associate professor(parkjs@hanbat.ac.kr)

1. 서 론

최근 수질 정보를 실시간으로 측정하고 분석하는 데이터 기반 접근의 중요성이 높아지고 있다. 하천 수질 측정자료와 같은 시계열 데이터는 환경 예측 모델 구축, 오염원 추적 등 수질관리 전반에 다양하게 활용되는 중요한 자료이다. 따라서 현장관리 및 연구 등 다양한 목적으로 수온, 수소이온농도, 전기전도도, 용존산소, 생화학적산소요구량 등 다양한 항목이 지속적으로 측정되며, 수질 변화 감지, 이상 징후의 조기 탐지 등 수질 평가 및 환경 관리에 폭넓게 활용되고 있다 (Fraga et al., 2020; Zhang et al., 2020).

통제된 공간이 아닌 하천 등 외부 현장에서 측정되는 환경 데이터의 특성상 센서의 고장, 자연재해, 유지보수 등의 다양한 원인으로 인해 결측치가 빈번하게 발생하며 (Ma et al. 2020; Zhang et al. 2019), 이러한 결측치는 분석 및 예측 결과의 왜곡 및 모형 성능 저하 등의 원인이 될 수 있다 (Fekade et al., 2017; Gao et al., 2018; Larson et al., 2023). 특히 시계열 데이터는 이전 시점의 값이 이후 값에 영향을 줄 수 있는 특성이 있어, 일부 구간에 결측이 생기면 전체 데이터의 흐름이나 패턴 분석에 영향을 미칠 수 있다. 따라서 결측값 보간은 환경 시계열 데이터의 연속성과 분석 신뢰성을 유지하기 위한 중요한 전처리 과정이다.

최근에는 이러한 시계열 자료의 결측치 보간에 다양한 머신러닝 기법을 적용하는 것에 대한 관심이 높아지고 있다. K-Nearest Neighbors (KNN), multilayer perceptron (MLP) 및 autoencoder 등 다양한 머신러닝 기반 모형을 시계열 자료의 결측치 보간에 적용하기 위한 연구가 계속되고 있으며, 이와 함께 선형 보간 (linear interpolation)과 다항 보간 (polynomial interpolation) 같은 전통적인 방법도 여전히 널리 활용되고 있다 (Park et al., 2023; Chen et al., 2022; Wang et al., 2024). Chen et al. (2022)은 시계열 대기 데이터에 대해 logistic regression 기반의 새로운 first five last three logistic regression imputation (FTLRI) 보간 기법을 제안하고 기존 기법 대비 성능을 비교 하였으며, Park et al. (2023)은 MLP 구조를 이용하여 장기간 결측 시계열 보간 성능을 평가하였다. 또한 Wang et al. (2024)은 수질 데이터 결측 패턴을 고려하여 2단계 보간 전략을 제안하고 성능을 분석하는 등 다양한 알고리즘을 이용한 결측치 보간을 위한 연구가 지속되고 있다.

데이터 기반 모형의 특성상 각 데이터 보간 기법의 성능은 대상이 되는 데이터의 특성과 결측 패턴에 따라 다르게 나타날 수 있다. 전통적인 선형 보간이나 다항 보간은 계산이 간단하고 적용이 용이하다는 장점이 있으나, 급격한 변화나 비선형적 패턴을 반영하는 것은 한계가 있고, KNN과 AE와 같은 비선형 접근법은 복잡한 패턴을 포착할 가능성을 제공하지만 계산 시간과 hyper parameter 설정 등에 따라 성능이 크게 달라질 수 있다.

다양한 원리에 기반한 보간 모형의 특성상, 고도화된 특정 기법이 항상 우수한 성능을 보이는 것은 아니며, 결측의 패턴과 기간에 따라 보간 결과가 달라질 수 있다. 특히 하천 수질자료의 경우, 장기간 일정한 값이 유지되는 구간뿐만 아니라 수질이 급변하는 홍수기 등에도 결측이 발생할 수 있어, 다양한 유형의 결측 패턴이 나타난다. 따라서 이러한 패턴을 고려한 적절한 보간 방법의 선택이 필요하다.

하지만 기존 연구들은 대부분 특정 보간 기법을 단독으로 적용하거나 제한된 조건에서 보간 성능을 비교한 사례가 많으며, 다양한 결측 패턴과 변수에 대해 여러 보간 기법을 체계적으로 적용한 연구는 상대적으로 제한적이다.

본 연구에서는 실제 하천 현장에서 측정된 수질 시계열 자료를 대상으로, 다양한 결측 패턴에 따른 데이터 보간 모형의 적용 특성을 분석하였다. 이를 위해 하천의 주요 수질 변수 중 하나인 수온을 대상 변수로 설정하고, 무작위 결측, 장기 결측, 피크 구간 결측, 수치 상승·하강 구간 결측의 네 가지 유형의 인위적 결측을 생성하였다. 이후 각 결측 유형에 대해 4가지 (선형 보간, 다항 보간, KNN, autoencoder) 보간 기법을 적용하여 보간 성능을 정량적으로 평가하고, 기법별 보간 결과의 특성을 분석하였다.

2. 연구방법 및 실험방법

2.1 연구대상 지역

대청호 하류 구간은 대청댐에서 방류되는 수량과 수질의 영향을 받는 지역으로 수온, 수소이온농도 (PH), 전기전도도 (EC), 용존산소 (DO), 생화학적산소요구량 (BOD)와 같은 주요 수질 항목들이 장기적이고, 연속적으로 측정되고 있는 지점이며, 하류 수계에 대한 영향성 및 수질 변화 모니터링을 위한 주요 지점이다 (Fig. 1). 본 연구에서는 환경부 국립환경과학원의 물 환경정보시스템에서 제공되는 총량측정망 금본G 지점(site No. 3010A20) 에서 2007년 1월부터 2023년 12월까지 측정된 총 801개의 주간 자료를 사용하였다 (NIER, 2024). 모형 구축에 사용된 원본 데이터는 모두 결측치가 없었다.

2.2 결측치 생성

수온은 녹조류의 성장과 번식 등에 직접적인 영향을 미치며 현장의 수질 상태를 확인할 수 있는 대표적인 수질 환경 변수이다 (Paerl and Huisman, 2008; Robarts and Zohary, 1987). 또한 수온은 센서를 이용한 실시간 측정이 비교적 용이하여, 수질 모니터링 현장에서 널리 활용되고 있다.

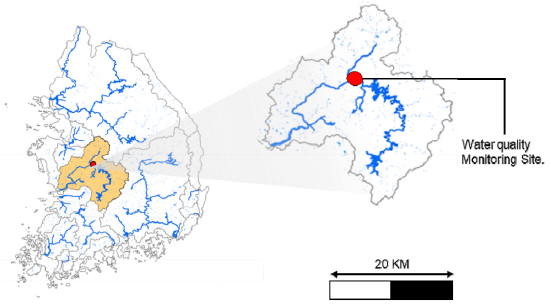
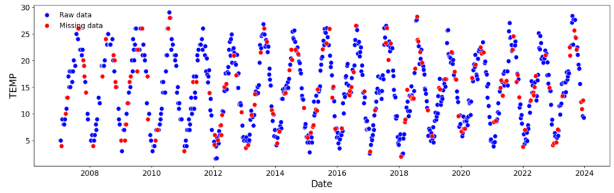


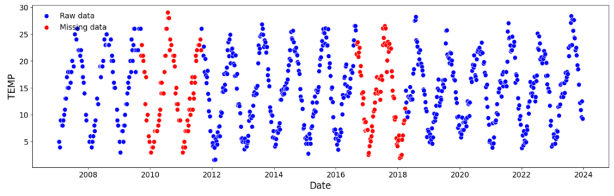
Fig. 1. Research site.

본 연구에서는 수온 (TEMP)을 분석 대상으로 사용하였으며, 다양한 형태의 수온 결측 유형에 따른 보간 특성을 비교하기 위해 센서의 오류 등 특정 시점의 자료가 불규칙적으로 누락되는 무작위 결측 (Case 1), 장비 고장이나 측정 공백 등 일정 기간 연속적으로 자료가 결측되는 장기 결측 (Case 2), 여름철 고수온기와 같은 극값 구간에서 센서 오작동 등으로 발생되는 첨두 (peak) 결측 (Case 3) 및 계절적 변화에 따라 수온이 상승하거나 하강하는 구간에서의 결측 (Case 4)의 4가지 유형의 결측치를 생성하였다 (Fig. 2).

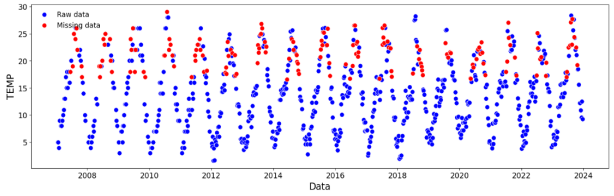
Case 1은 전체 자료에 대해 무작위로 160개의 결측치를 생성하여 구축된 자료이다. 불규칙적으로 발생하는 일반적인 결측 상황을 가정하고 특정 규칙 없이 무작위로 단기 결측을 생성하였다. Case 2는 연속적으로 데이터가 누락되는 장기 결측 상황을 가정하여 총 162개의 장기 결측치를 포함하도록 자료를 구성하였다. Case 3은 외부 요인 등 환경 변화로 수온이 급격히 변하는 상황을 가정하였으며, 급격히 수온이 상승하여 높아지는 첨두 (peak) 구간의 전후에 대해 160개의 결측을 생성하였다. Case 4는 수온이 상승하는 첨두 구간의 전후 및 수온이 낮아진 후 다시 상승하는 저점 전후에 162개의 결측을 생성하였다. Case 1-4의 결측치는 각각 160, 162, 160, 162개로 전체 데이터 801개 대비 약 20%의 결측치를 포함



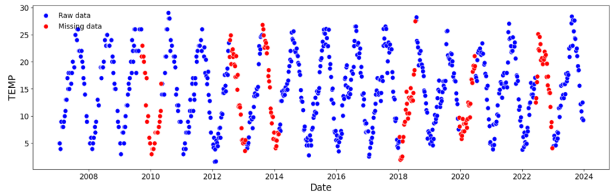
(a) random missing (Case 1)



(b) long-duration missing (Case 2)



(c) peak-period missing (Case 3)



(d) Missing values in increasing or decreasing phases (Case 4)

Fig. 2. Synthetic missing data.

하도록 구성되었다. 결측치의 생성을 위한 프로그램은 python 라이브러리 NumPy 및, Pandas 등을 이용하여 구축하였다 (Harris et al., 2020; McKinney, 2010).

2.3 보간 기법

본 연구에서는 생성된 네 가지 결측 유형에 대해 전통적인 통계적 보간 기법인 선형 보간 (Linear) 및 다항 보간 (Polynomial)의 2가지 방법과 머신러닝 보간 기법 중 거리 기반 보간 기법인 KNN과 신경망 기반 보간 기법인 autoencoder (AE)을 각각 적용하는 모형을 구축하여 데이터 보간을 수행하였다.

모형의 구축을 위해 결측치가 없는 원본 자료를 기반으로 인위적으로 결측치를 삽입한 자료를 생성하였으며, 결측치를 포함한 자료를 학습 및 보간에 활용하였다. 통계 기반 기법인 Linear는 별도의 학습 과정 없이 보간이 가능하며, Polynomial은 설정된 차수에 따라 다항 함수를 적합하는 방식으로 보간을 수행한다.

머신러닝 기반 모형은 보간 대상의 결측값의 보간을 위해 주변의 다른 항목의 값을 활용하여 학습할 수 있다. 본 연구에서는 머신러닝 보간 모형인 KNN과 AE 모형의 구축을 위한 입력자료로 PH, EC, DO 및 BOD를 함께 사용하였다.

Linear는 결측값을 인접한 두 관측값을 기준으로 직선 형태로 추정하는 방법이며, 보간식은 Eq. (1) 과 같다 (Meijering, 2002).

$$x(t) = x(t_1) + \frac{(t - t_1)}{(t_2 - t_1)} \times (x(t_2) - x(t_1)) \quad (1)$$

where,
t_1, t_2 : Time points with observed values
$x(t_1), x(t_2)$: Observed values at t_1 and t_2
t: Time point with missing values

Polynomial은 결측값을 포함하는 주변 시점의 데이터를 기반으로 n차 다항식을 적합하여 결측값을 추정하는 방법이다. 보간식은 Eq. (2) 와 같다. 보간을 위한 모형의 구축은 python 라이브러리 scipy를 사용하였으며 (Virtanen et al., 2020), Polynomial 모형의 다항식 차수는 2차 및 3차를 적용하였다.

$$x(t) = a_0 + a_1 \cdot t + a_2 \cdot t^2 + \cdots + a_n \cdot t^n \quad (2)$$

where,
t : Time
$a_0, a_1, a_2, \cdots, a_n$: Polynomial coefficients
n: Degree of the polynomial

KNN은 결측 시점의 값을 주변 관측치들과 유사한 특성을 가진 k개의 인접 데이터를 탐색한 후, 이들의 평균값을 이용해 결측값을 추정하는 방식이다. 일반적으로 유클리드 거리 등 거리 기반 유사도를 활용하여 인접 값을 선택하며 보간식은 Eq. (3)과 같다. 보간을 위한 모형의 구축은 python 라이브러리 scikit-learn를 사용하였으며(Pedregosa et al., 2011), KNN

모형의 k 값은 3, 5, 7로 설정하여 분석을 수행하였다.

$$x(t) = (\frac{1}{K}) \times \sum_{i=1}^{K} x_i \quad (3)$$

where,
t: time point with missing value
x_i: i-th nearest neighbor to the missing point
K: the number of nearest neighbors

AE는 입력 데이터를 압축한 후 복원 하는 방식으로, 인공신경망을 활용하여 결측값을 추정하는 비선형 보간 기법이다. 입력 시계열 데이터의 패턴과 구조를 학습한뒤 결측이 포함된 데이터를 입력하면 이를 기반으로 결측값을 복원한다. 인코더와 디코더로 구성되며, 보간식은 Eq. (4), (5)와 같다 (Vincent et al., 2008). 보간을 위한 모형의 구축은 python 라이브러리 Tensorflow를 사용하였으며 (Abadi et al., 2016), 모형에 적용된 주요 하이퍼파라미터를 Table 1에 제시하였다.

$$Z = f_{enc(X)} \quad (4)$$

$$\hat{X} = f_{dec(Z)} \quad (5)$$

where,
Z : Latent space
X : Input data with missing values
$\hat{X}$: Interpolated values by the decoder

2.4 보간 성능 평가 기준

각 보간 성능을 비교 하기 위해 root mean squared error (RMSE), Nash-sutcliffe efficiency (NSE) and RMSE-to-standard deviation ratio (RSR) 3개의 지표를 이용하여 성능을 비교하였다 (Eq. 6, 7, 8) RMSE가 작을수록 예측 성능이 좋음을 나타내고, NSE는 -∞에서 1 사이로, 1에 가까울수록 성능이 좋음을 나타낸다. RSR의 값은 0에서 ∞ 사이로, 값이 0에 가까울수록 성능이 좋음을 나타낸다 (Bennett et al., 2013, Moriasi et al., 2007).

$$\text{RMSE} = \sqrt{\left(1/n * \sum (y_i - \hat{y}_i)^2\right)} \quad (6)$$

$$\text{NSE} = 1 - \left(\sum (y_i - \hat{y}_i)^2\right) / \left(\sum (y_i - \bar{y})^2\right) \quad (7)$$

$$\text{RSR} = \sqrt{\left(1/n * \Sigma \left(y_i - \hat{y}_i\right)^2\right)} \; / \; \sqrt{\left(1/n * \Sigma \left(y_i - \bar{y}\right)^2\right)} \quad (8)$$

where,
y_i: Observation value
$\hat{y}_i$: Predicted value
$\bar{y}$: Mean of observations
n: Number of data points

Table 1. Hyperparameters of the AE imputation model.

	Layer	Nodes	Activation
Input	-	4	-
Encoder	Hidden layer 1	4	ReLU
	Hidden layer 2	3	ReLU
Decoder	Hidden layer 1	4	ReLU
	Out layer	4	Linear
Training setup	Learning late	0.001	-
	Batch size	32	-
	Epochs	100	-

3. 결과 및 고찰

3.1 Case 별 보간 결과 비교

본 연구에서는 TEMP 변수에 대한 4가지 유형 (Case 1-4)의 결측 자료에 대해 Linear, Polynomial, KNN 및 AE 모형을 적용하여 보간을 수행하고 그 결과를 비교하였다.

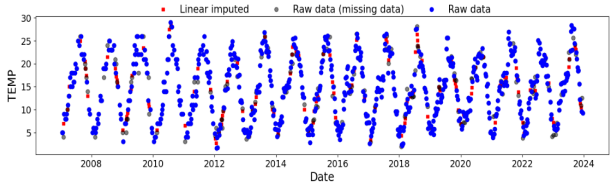
무작위로 결측치를 생성한 Case 1의 TEMP에 대한 보간 성능 분석 결과, Linear 방법이 가장 우수한 성능을 보이는 것으로 분석되었다. 선형 보간의 RMSE, NSE 그리고 RSR은 각각 1.72, 0.93, 0.26으로, 다른 보간 기법에 비해 상대적으로 높은 정확도를 보였다 (Table 2). Polynomial은 차수 2와 3의 RSR이 각각 0.32 및 0.33으로 분석되어 Linear에 이어 두 번째로 우수한 보간 성능을 보였다. KNN은 k 값에 따라 RSR이 0.68-0.70의 범위를 보이고 AE의 RSR은 0.69로 확인되어, 전통적인 통계 기반 기법에 비하여 머신러닝 알고리즘의 보간 성능이 낮은 것으로 분석되었다 (Fig. 3).

장기간의 결측을 포함하도록 구축된 Case 2의 경우 전체 결측 구간이 길어짐에 따라 상대적으로 단순한 모형인 Linear 및 Polynomial 모형의 보간 성능이 현저히 떨어지는 경향을 확인하였으며, 반면 머신러닝 기반 모형인 KNN 및 AE가 상대적으로 우수한 성능을 보이는 것을 확인할 수 있었다. AE의 RMSE, NSE 및 RSR값은 각각 4.45, 0.60 및 0.63으로 적용한 4가지 모형 중 가장 우수한 성능을 보이는 것으로 확인하였다 (Table 3). KNN은 k 값에 따라 다소 차이가 있으나 RSR 0.64-0.68로 AE에 이어 안정적인 성능을 보였다. 반면 Polynomial의 RSR은 6.75-7.55, Linear의 RSR은 1.69로 머신러닝 기법에 비해 현저히 감소한 성능을 보여 Case 1의 단기간의 무작위 결측값에 대한 보간결과와는 다른 경향을 보였다 (Fig. 4).

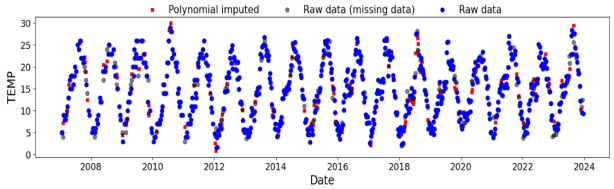
급격히 수온이 상승하여 높아지는 첨두 구간의 전후에 대한 결측을 포함하도록 구축된 Case 3의 경우 Case 1과 2에 비해 전체적으로 보간 성능이 저하되는 경향을 보였다. 모형별로는 Linear가 가장 양호한 성능을 보였으며, RMSE, NSE 및 RSR이 각각 2.15, 0.43 및 0.76으로 분석되었다 (Table 4). 반면 KNN과 AE는 Linear나 Polynomial에 비해서 성능이 매우 낮았으며 Case 1과 2에 비해서도 보간값의 정확도가 크게 저하되는 경향을 보였다 (Fig. 5).

Table 2. Imputation performance comparison for Case 1

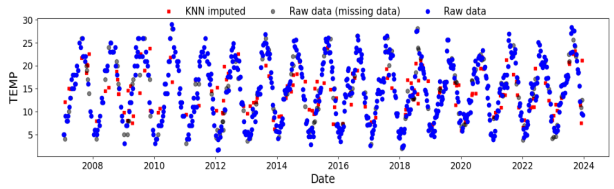
	Imputation method	RMSE	NSE	RSR
TEMP	Linear	1.72	0.93	0.26
	Polynomial (n = 2)	2.10	0.90	0.32
	Polynomial (n = 3)	2.14	0.89	0.33
	KNN (k = 3)	4.62	0.50	0.70
	KNN (k = 5)	4.47	0.54	0.68
	KNN (k = 7)	4.63	0.50	0.70
	AE	4.56	0.52	0.69



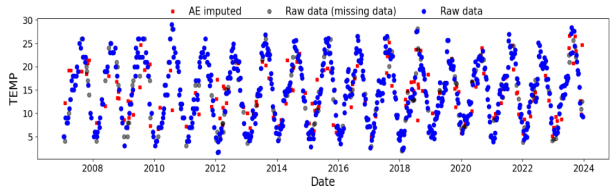
(a) Linear



(b) Polynomial



(c) KNN

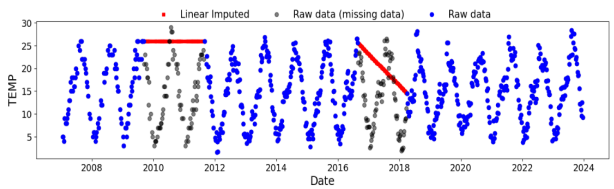


(d) AE

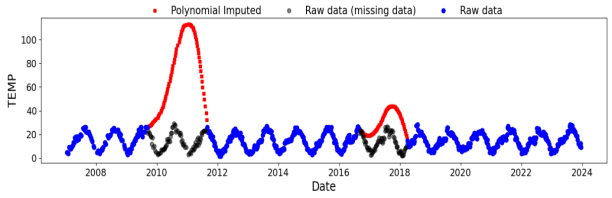
Fig. 3. Comparison of imputation performance by different methods in Case 1.

Table 3. Performance comparison for Case 2

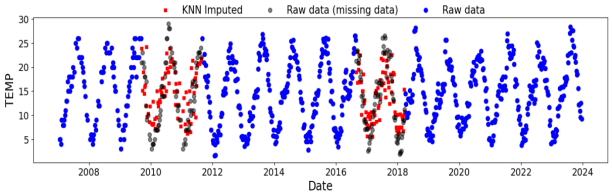
	Imputation method	RMSE	NSE	RSR
TEMP	Linear	11.90	−1.86	1.69
	Polynomial (n = 2)	47.53	−44.63	6.75
	Polynomial (n = 3)	53.12	−55.98	7.55
	KNN (k = 3)	4.48	0.59	0.64
	KNN (k = 5)	4.76	0.54	0.68
	KNN (k = 7)	4.75	0.54	0.68
	AE	4.45	0.60	0.63



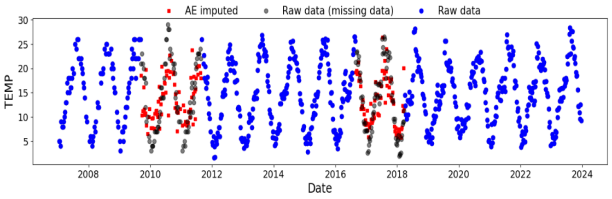
(a) Linear



(b) Polynomial



(c) KNN

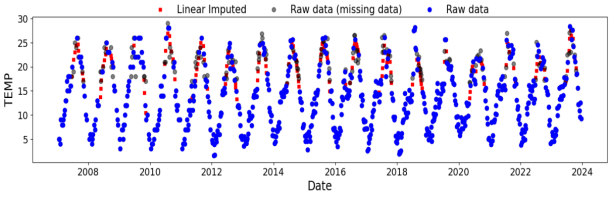


(d) AE

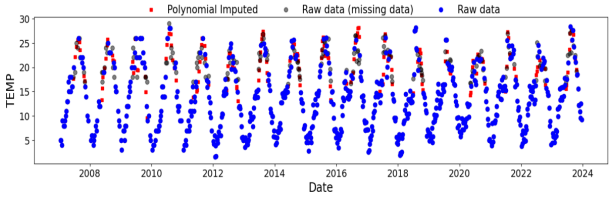
Fig. 4. Imputation performance for Case 2.

Table 4. Performance comparison for Case 3

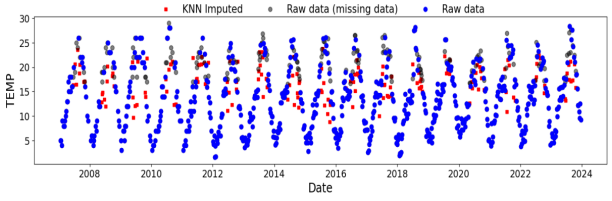
	Imputation method	RMSE	NSE	RSR
TEMP	Linear	2.15	0.43	0.76
	Polynomial (n = 2)	2.63	0.14	0.93
	Polynomial (n = 3)	2.69	0.10	0.95
	KNN (k = 3)	5.63	−2.92	1.98
	KNN (k = 5)	5.61	−2.89	1.97
	KNN (k = 7)	5.82	−3.20	2.05
	AE	5.21	−2.37	1.83



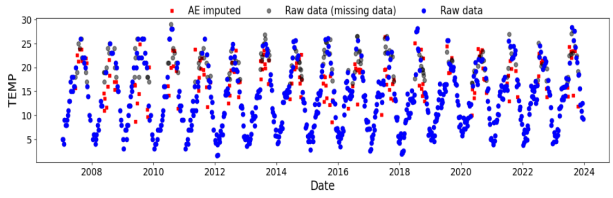
(a) Linear



(b) Polynomial



(c) KNN

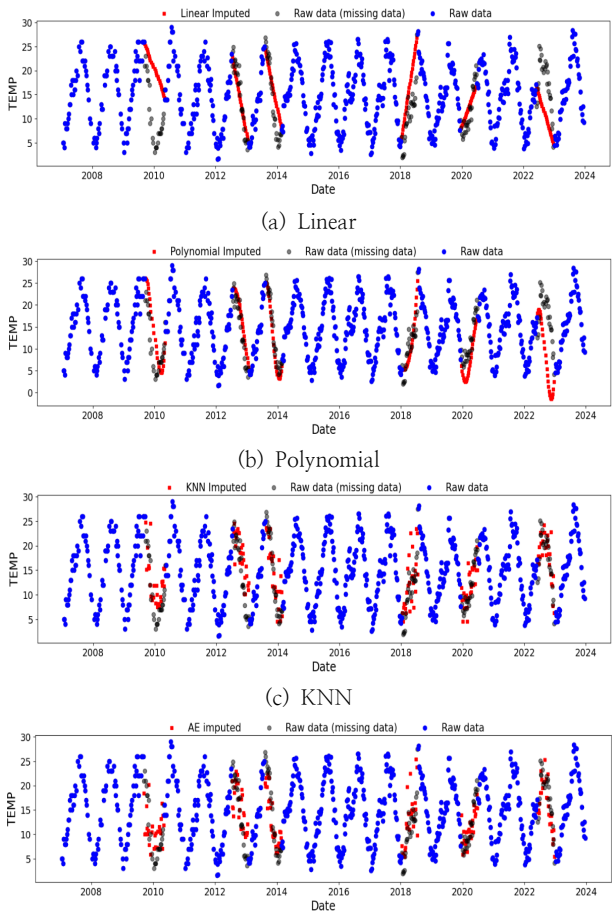


(d) AE

Fig. 5. Imputation performance for Case 3.

Table 5. Performance comparison for Case 4

	Imputation method	RMSE	NSE	RSR
TEMP	Linear	6.31	0.06	0.97
	Polynomial (n = 2)	6.11	0.12	0.94
	Polynomial (n = 3)	5.81	0.20	0.89
	KNN (k = 3)	4.28	0.57	0.66
	KNN (k = 5)	4.41	0.54	0.68
	KNN (k = 7)	4.44	0.54	0.68
	AE	4.44	0.54	0.68



(a) Linear

(b) Polynomial

(c) KNN

(d) AE

Fig. 6. Imputation performance for Case 4.

Case 4의 경우 KNN의 성능이 가장 양호한 것으로 나타났으며, k가 3인 경우의 RMSE, NSE 및 RSR이 각각 4.28, 0.57 및 0.66으로 가장 우수한 성능을 보였다 (Table 5). AE는 KNN보다 다소 낮으나 유사한 수준의 성능을 보여

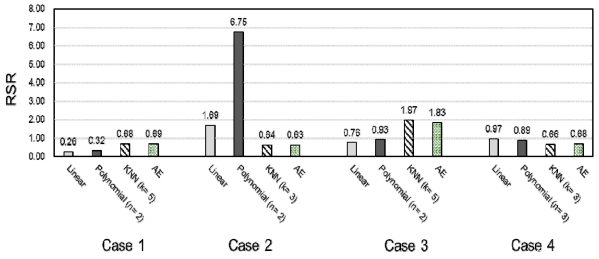


Fig. 7. Comparison of imputation methods by Case.

RMSE, NSE 및 RSR가 4.44, 0.54 및 0.68로 분석되었다. 반면 Linear는 상대적으로 낮은 성능을 보였으며 Polynomial의 경우 역시 일부 구간에서 과도한 곡선형 보간이 발생하는 등 왜곡이 나타나며 추세를 제대로 반영하지 못하는 것으로 분석되었다 (Fig. 6).

각 Case별 결측치 보간 모형의 RSR 값을 Fig. 7에 비교하였다. RSR 값이 0에 가까울수록 모형의 보간 성능이 우수하여 실측값을 잘 보간하였음을 의미하며 모형의 성능이 저하될수록 값이 커지게 된다. Linear와 Polynomial 기법은 가장 성능이 우수한 경우의 값만 포함하였다. 분석 결과, 단기간의 무작위 결측을 반영한 Case 1에서는 Linear 모형이 가장 우수한 보간 성능을 보이는 것으로 나타났다. 장기 결측을 반영한 Case 2에서는 KNN (k=3)과 AE의 RSR이 각각 0.64 및 0.63으로 우수한 성능을 보였으며, 첨두 지점의 결측을 고려한 Case 3에서는 Linear와 Polynomial이 상대적으로 좋은 성능을 나타냈다. 수온이 상승하거나 하강하는 구간인 Case 4에서는 KNN (k=3)이 RSR 0.66로 가장 우수한 성능을 보였으며, AE의 RSR이 0.68로 두 번째로 우수한 성능을 보여, 결측 유형별로 보간 모형의 성능이 다양하게 나타남을 확인 하였다.

3.2 보간 성능 세부 분석

본 연구에서는 4개의 결측 패턴에 대한 보간 성능을 정량적으로 분석하였으며, 보간 기법 간의 상대적 성능을 보다 명확히 비교하기 위해 일부 결측 구간을 확대하여 시각적으로 종합 분석하였다 (Fig. 8). 무작위로 생성된 결측을 포함한 Case 1의 경우 Linear 모형이 가장 우수한 보간 성능을 보였으며, 측정자료의 중간 중간 발생된 단기 결측에 대해 Linear 모형이 원본 데이터의 흐름을 비교적 자연스럽게 보간하고 있음을 확인할 수 있다 (Fig 8(a)). Polynomial 모형의 경우도 Linear와 유사한 수준의 성능을 보였으며, 일부 구간에서는 Linear에 비해 정확도가 다소 저하되는 양상을 보였다 (Fig 8(a)). 반면 KNN 및 AE 모형을 적용한 경우 일부 지점에서 지나치게 높거나 낮은 값으로 보간되는 등 상대적으로 낮은 정확도를 보이는 것을 시각적으로 확인할 수 있다 (Fig. 8(a)). 장기적인 결측을 포함한 Case 2의 경우 머신러닝 모형인 AE가 가장 우수한 성능을 보였으며, KNN이 두 번째로 우수한 성능을 보였고, Linear와 Polynomial의 성능이 상대적으로 낮은 것으로 분석되었다. Linear 모형은 결측 구간 전후의 측정값을 직선으로 연결하여 보간하는 알고리즘의 특성상 장기 결측 구간을 선형으로 연결하여 실질적인 추세를 반영하지 못하는

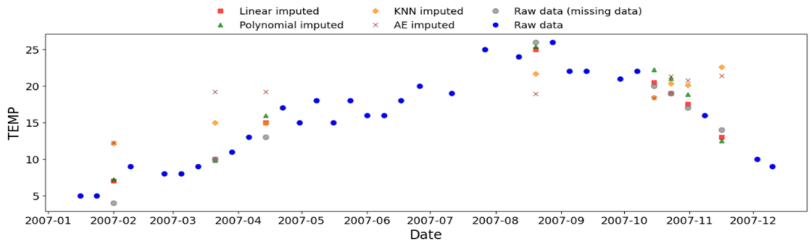
것을 확인할 수 있다 (Fig 8(b)). Polynomial의 경우 증감되는 추세 자체는 유사하나 보간값은 실측값과 큰 폭의 차이를 보였다. 반면 머신러닝 기반 모형인 KNN 및 AE는 전체적인 추세를 상대적으로 잘 반영하였다 (Fig. 8(b)). Case 3의 경우 첨두지점 전후의 결측값에 대해 Linear와 Polynomial은 원본 데이터와 경향을 상대적으로 잘 반영하였으나, KNN과 AE의 경우 과도하게 크거나 작은 값으로 보간하는 것을 확인할 수 있다 (Fig. 8(c)). Case 4는 상대적으로 장기간의 첨두 혹은 저점의 결측값에 대해 Linear와 Polynomial은 실질적인 변화 추세를 반영하지 못하였으며, KNN과 AE는 전체적인 수질변화 추세를 반영하여 안정적으로 보간을 수행한 것을 확인할 수 있다 (Fig. 8(d)).

3.3 보간기법에 따른 특성 및 향후 연구

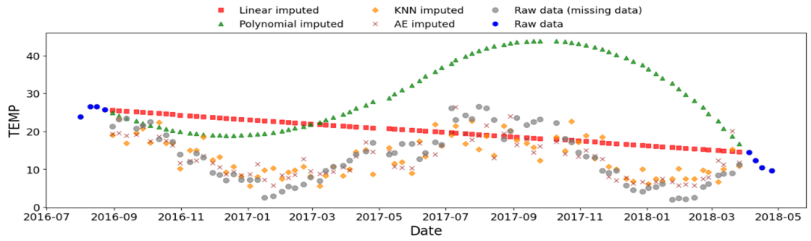
본 연구에서는 네 가지 유형의 결측을 포함한 자료에 대해 전통적인 통계 기법과 머신러닝 기반 보간 기법을 적용하여 결과를 비교하였으며, 결측 유형에 따라 모형의 보간 성능이 달라질 수 있음을 확인하였다.

Linear는 상대적으로 짧은 구간의 결측에서 안정적인 성능을 보였지만 장기간의 결측에 대해서는 성능이 크게 저하되는 특징을 보였다. Polynomial도 장기 결측에 대한 보간 시 모형의 성능이 크게 저하됨을 확인할 수 있었다.

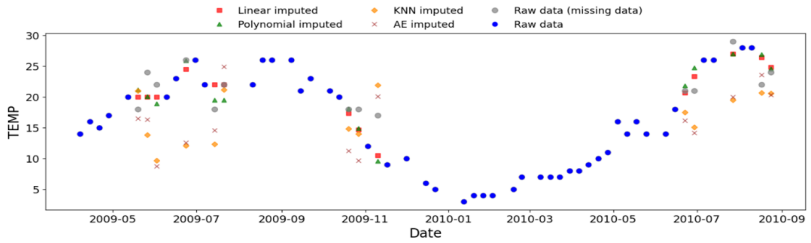
머신러닝 기반 모형인 KNN과 AE는 단기 결측에 대해서는 Linear 및 Polynomial 모형에 비해 오히려 성능이 저하되었으나 장기 결측에 대해서는 안정적이고 우수한 보간 성능을 보였다. KNN 및 AE는 단순히 결측이 발생된 자료만을 이용하여 보간을 수행하는 것이 아니라, 함께 측정된 다양한 다른 수질 항목을 분석에 포함하며, 이러한 모형의 특징이 장기간의 결측에 대해서 상대적으로 우수한 성능을 유지할 수 있게 하는 한 원인이 될 수 있을 것으로 판단된다.



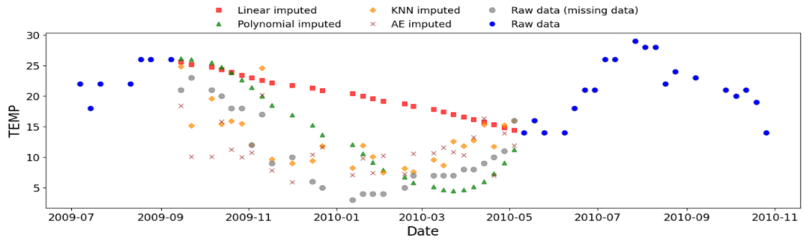
(a) Case 1



(b) Case 2



(C) Case 3



(d) Case 4

Fig. 8. Detailed visualization for missing data type.

현장 측정자료는 결측이 발생되는 구간의 특성 및 결측 기간 등에 따라 매우 다양한 형태의 결측을 포함할 수 있다. 데이터 기반 모형의 성능은 모형 구축에 사용된 데이터의 특성이 반영되며, 향후 여러 유역에서 다양한 측정 빈도와 항목으로 구성된 보다 다양한 유형의 결측 자료를 이용하여 결측 유형의 체계적 분류 및 분류된 결측 유형에 맞는 최적 모형의 적용을 위한 지속적인 연구로 현장 측정자료의 보간 정확도를 높이는데 기여할 수 있을 것으로 판단된다.

4. 결 론

본 연구에서는 수질 데이터에서 발생되는 4가지 유형의 (Case 1－4) 결측 자료에 전통적인 통계 기법인 선형 및 다항 보간과 KNN 및 autoencoder 알고리즘을 이용한 머신러닝 기반 보간 모형을 적용하여 보간 기법에 따른 결측 유형별 보간 성능을 비교하였다.

분석결과 단기간의 무작위 결측 자료를 포함한 Case 1의 경우 선형 보간이 가장 우수한 성능을 보였고, 장기간의 결측을 포함한 Case 2의 경우 autoencoder 알고리즘을 이용한 머신러닝 모형이 가장 우수한 성능을 보이는 것을 확인할 수 있었다. 한편 첨두 구간을 포함한 급격한 수질 변화가 발생하는 구간에 결측이 발생한 Case 3의 경우 선형 보간이 가장 우수한 성능을 보였다. 또한 첨두 구간 및 저점에서 상대적으로 장기간의 결측이 발생하는 Case 4의 경우 KNN 및 autoencoder 알고리즘을 이용한 모형이 우수한 성능을 보였다.

본 연구의 분석을 통해 결측 유형에 따라 보간 모형이 성능이 달라짐을 확인할 수 있었다. 따라서 결측값 보정을 위해 일률적으로 동일한 모형을 적용하는 것보다는 결측 유형에 맞는 적정 모형의 선정을 통해 보간 성능을 높일 수 있음을 확인할 수 있었다. 향후 결측 유형과 데이터의 구조적 특성을 고려한 최적 보간 모형의 구축을 위한 지속적인 연구를 통해 모형의 성능을 개선하고, 수질관리 효율과 신뢰도를 높이는데 기여할 수 있을 것으로 판단된다.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... and Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (pp. 265－283).

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., ... and Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling and Software, 40*, 1－20.

Chen, M., Zhu, H., Chen, Y., and Wang, Y. (2022). A novel missing data imputation approach for time series air quality data based on logistic regression. *Atmosphere, 13*(7), 1044.

Fekade, B., Maksymyuk, T., Kyryk, M., and Jo, M. (2017). Probabilistic recovery of incomplete sensed data in IoT. *IEEE Internet of Things Journal, 5*(4), 2282－2292.

Fraga, M. D. S., Reis, G. B., da Silva, D. D., Guedes, H. A. S., and Elesbon, A. A. A. (2020). Use of multivariate statistical methods to analyze the monitoring of surface water quality in the Doce River basin, Minas Gerais, Brazil. *Environmental Science and Pollution Research, 27*(28), 35303－35318.

Gao, Y., Merz, C., Lischeid, G., and Schneider, M. (2018). A review on missing hydrological data processing. *Environmental Earth Sciences, 77*(2), 47.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... and Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357－362.

Larson, D. M., Bungula, W., Lee, A., Stockdill, A., McKean, C., Miller, F. F., ... and Hlavacek, E. (2023). Reconstructing missing data by comparing interpolation techniques: Applications for long-term water quality data. *Limnology and Oceanography: Methods, 21*(7), 435－449.

Ma, J., Cheng, J. C., Ding, Y., Lin, C., Jiang, F., Wang, M., and Zhai, C. (2020). Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series. *Advanced Engineering Informatics, 44*, 101092.

Meijering, E. (2002). A chronology of interpolation: From ancient astronomy to modern signal and image processing. *Proceedings of the IEEE, 90*(3), 319－342.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE, 50*(3), 885－900.

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51－56.

NIER. (2024). Realtime water information system. National Institute of Environmental Research. https://water.nier.go.kr/web(accessed on March 24, 2025)

Park, J., Müller, J., Arora, B., et al. (2023). Long-term missing value imputation for time series data using deep neural networks. *Neural Computing and Applications, 35*(12), 9071－9091.

Paerl, H. W., and Huisman, J. (2008). Blooms like it hot. *Science, 320*(5872), 57－58.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(85), 2825 – 2830.

Robarts, R. D., and Zohary, T. (1987). Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom-forming cyanobacteria. *New Zealand Journal of Marine and Freshwater Research, 21*(3), 391 – 399.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 1096 – 1103).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... and van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261 – 272.

Wang, F., Cui, X., Gui, Y., and Qiao, Y. (2024). Two stage iterative approach for addressing missing values in small-scale water quality data. *Marine Development, 2*(1), 1 – 11.

Zhang, Y. F., Thorburn, P. J., Xiang, W., and Fitch, P. (2019). SSIM—A deep learning approach for recovering missing time series sensor data. *IEEE Internet of Things Journal, 6*(4), 6618 – 6628.

Zhang, Y. F., Fitch, P., and Thorburn, P. J. (2020). Predicting the trend of dissolved oxygen based on the kPCA-RNN model. *Water, 12*(2), 585.