딥러닝 기반 다변량 시계열 결측치 대체 모델을 이용한 대기질 결측치 보간

고현준, 김수현*
경북대학교

kgbg6632@knu.ac.kr, *suhyeonkim@knu.ac.kr

Deep Learning-based Multivariate Time Series Imputation for Air Quality Missing Data

Hyun Jun Ko, Suhyeon Kim*
Kyungpook National University

요 약

대기질 시계열 데이터를 정확하게 분석하는 것은 대기오염을 줄이기 위한 정책을 만들고 시행하는 데에 중요한 역할을 한다. 그러나 실제 환경에서 측정되는 대기질 시계열 데이터는 측정소 주변 조건 및 여러 원인들에 의해 불규칙적으로 수집될 수 있으며, 이는 미관측 데이터, 즉 결측치를 초래할 수 있다. 본 논문에서는 다변량 대기질 시계열 데이터에 존재하는 결측치를 효과적으로 처리하여 데이터의 활용도를 높이기 위해, 다양한 딥러닝 기반 다변량 시계열 결측치 대체 모델들을 적용 후 비교 분석하였다. 2021 년부터 2023 년까지의 대구 및 경북 지역에서 1 시간 단위로 측정한 센서 기반 대기질 및 기상 시계열 데이터를 수집하였으며, 전통적인 선형보간법, 스플라인 보간법 및 네 가지 딥러닝 기반 결측치 대체 모델을 이용하여 다변량 시계열 데이터에서의 결측치 대체 성능을 비교하였다. 실험 결과, 전통적인 보간법과 비교하여 딥러닝 기반 다변량 시계열 결측치 대체 모델들이 낮은 MAE, RMSE 를 보여주었으며, 그 중 Transformer 모델이 최소 오차값을 보여줌을 확인하였다.

Ⅰ. 서 론

대기오염은 전 세계적으로 심각한 문제로, 인간에게 다양한 직간접적인 위협을 가하고 있다. WHO에 따르면 약 24억 명이 등유, 바이오매스, 석탄을 연료로 사용하는 화덕이나 난로로 인해 위험한 대기오염에 노출되고 있으며, 매년 700만 명이 조기 사망하고 있다. 대기오염의 주요 원인은 주거용 에너지, 차량, 발전 등으로, 이러한 오염물질은 기상 조건 및 교통 상황 등과 상호작용하여 대기오염을 유발하며, 건강에 심각한 악영향을 미친다[1]. 따라서 대기환경의 정확도 높은 데이터 분석 및 대기질 관리는 공공 안전과 환경 보호를 위해 필수적이나, 다양한 요인으로 인해 발생할 수 있는 미관측 데이터(이하 결측치)는 데이터 편향을 초래하며 분석 정확도를 저하시킬 수 있다. 본 연구에서는 다변량 대기질 시계열 데이터의 결측치를 효과적으로 처리하여 데이터 활용성을 극대화하기 위해, 여러 딥러닝 기반의 다변량 시계열 결측치 대체 모델을 적용하고 그 성능을 비교 분석하였다.

Ⅱ. 데이터 및 분석 방법

본 연구에서는 AirKorea와 기상청으로부터 2021년부터 2023년까지 3년간 1시간 단위로 대구와 경북 지역에서 측정된 총 26,280개의 대기질 및 기상 시계열 데이터를 수집하였다. 대기질 데이터는 SO2, CO, O3, NO2, PM10, PM2.5의 변수들을, 기상 데이터는 온도, 습도, 강수량, 풍속, 풍향 변수들을 포함하며, 수집된 데이터에는 다양한 원인으로부터 발생한 결측값이 다수 존재한다. 표 1은 수집된 데이터에서의 대기질 변수별 결측치 비율을 표기한 것이다.

본 연구에서는 수집된 다변량 대기질 시계열 데이터를 8:1:1 비율로 훈련 데이터, 검증 데이터, 테스트 데이터로 분리 후 모델링을 진행하였다. 두 가지 전통적인 시계

열 결측치 대체 모형인 선형 보간법과 스플라인 보간법 및 네 가지 딥러닝 기반 결측치 대체 모델인 BRITS[2], TimesNet[3], SAITS[4], Transformer[5]를 이용하여 다변량 시계열 데이터에서의 결측치 대체 성능을 비교 분석하였다.

결측치 대체 성능을 측정하기 위해, 결측치 데이터 외의 일부 데이터를 함께 마스킹하여 정답 데이터와 대체 데이터 간 오차를 비교하였다. 대표적인 두 가지 오차 성능 지표인 평균 절대 오차(Mean Absolute Error; MAE)와 평균 제곱근 오차(Root Mean Squared Error; RMSE)를 계산하여 결측치 대체 모델의 성능을 최종적으로 비교 분석하였다.

표 1. 대기질 데이터 내 변수별 결측치 비율

Air Quality Variable	Missing Ratio (%)
SO2	4.1%
CO	4.6%
O3	3.7%
NO2	5.3%
PM10	4.6%
PM2.5	6.1%

Ⅲ. 성능 평가

표 2는 여섯 종류의 결측치 대체 모델의 성능을 비교한 결과이다. 이 중 Transformer 모델이 MAE(0.6146)와 RMSE(0.8198) 모두에서 가장 낮은 오차값을 기록하였다. 또한, 전통적인 보간법과 비교하여 딥러닝 기반 다변량 시계열 결측치 대체 모델들이 낮은 MAE, RMSE를 보여줌을 확인하였다.

표 2. 결측치 대체 모델 성능 평가

Model	MAE	RMSE
Linear Interpolation	6.2804	21.8904
Spline Interpolation	68.4128	334.3856
BRITS	0.6561	0.8765
TimesNet	0.6634	0.9040
SAITS	0.6648	0.8913
Transformer	**0.6146**	**0.8198**

그림 1은 실제 다변량 대기질 데이터의 결측치를 대체한 결과를 시각화하여 나타낸 것이다. 선형 보간법에 비해 Transformer 모델로 결측치를 대체한 경우 더 다양한 값의 분포로 미관측 값이 채워짐을 확인할 수 있다.

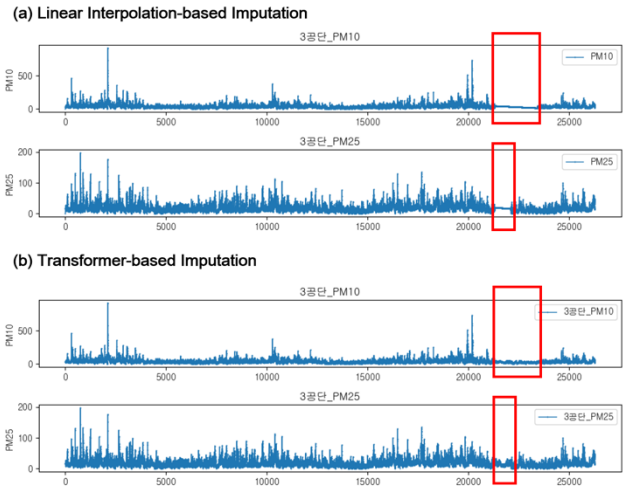

그림 1. 다변량 대기질 시계열 결측치 대체 결과

Ⅳ. 결론

본 연구에서 다변량 대기질 시계열 데이터의 결측치를 처리하기 위해 다양한 딥러닝 기반 모델과 전통적인 보간법을 비교한 결과, Transformer 모델이 가장 우수한 성능을 보였으며, 이는 딥러닝 모델이 전통적 방법에 비해 결측치 대체에 더 효과적임을 시사한다. 추후 공공 안전과 환경 보호를 위한 대기질 예측 모델 성능을 높이는 데에 정확하게 결측치가 처리된 데이터 활용이 중요한 기여를 할 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

The work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00242528) and by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2024-RS-2024-00437756) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

참 고 문 헌

[1] World Health Organization, "Air pollution," World Health Organization. [Online]. Available: https://www.who.int/health-topics/air-pollution#tab=tab_2. [Accessed: 27-Sep-2024].

[2] W. Cao, D. Wang, J. Li, H. Zhou, "BRITS: Bidirectional Recurrent Imputation for Time Series," *Advances in Neural Information Processing Systems*, vol. 31, pp. 6776-6786, 2018.

[3] Y. Zhou, T. Cao, Y. Shen, Q. Wu, "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis," in *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, Dec. 2022, pp. 1-14.

[4] Y. Duan, J. Chen, R. Liu, "SAITS: Self-Attention-based Imputation for Time Series," *arXiv preprint arXiv:2202.08626*, 2022.

[5] H. Wang, X. Yuan, Y. Chen, S. C. H. Hoi, "Transformer in Time Series: A Survey," *arXiv preprint arXiv:2201.10472*, 2022.