Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 4, pp. 294-308, August 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.4.294

전이학습을 이용한 시계열 데이터의 결측치 대체와 예측 성능과의 상관성 분석

황희선[1] · 민대기[2†]

[1]이화여자대학교 빅데이터분석학 협동과정 / [2]이화여자대학교 경영학과

A Transfer Learning for Missing Value Imputation and Its Relationship with Prediction Performance in Time Series Data

Huiseon Hwang[1] · Daiki Min[2]

[1]Graduate School(Big Data Analytics), Ewha Womans University
[2]School of Business, Ewha Womans University

Missing values incur the lack of data availability and/or inaccurate predictions in the problem of time series prediction. We consider a transfer learning method for missing data imputation in time series data and test two research hypothesis; the first hypothesis is that the high similarity between two time series, one containing missing values and the other used for transfer learning, improves the imputation performance. Second, a better imputation performance results in a better prediction accuracy. Empirical analysis reveals that the transfer learning with high similarity in two time series improves the imputation performance. As known in the literature, we found a positive correlation between imputation performance and prediction accuracy. However, the correlation between imputation performance and prediction accuracy becomes insignificant when the time series has low volatility and a short length of consecutive missing data. It means that a simple method for missing data imputation is preferred to an expensive but effective method such as transfer learning if the time series is highly stable and predictable.

Keywords: Time Series Data, Missing Value Imputation, Transfer Learning, Prediction Accuracy, LSTM

1. 서 론

시계열 데이터는 규칙적인 시간 간격을 두고 시간 순으로 관측된 일련의 값들로 구성된 데이터를 의미한다. 관측 빈도와 대상 등에 따라 데이터의 규모와 차원이 높고 지속적인 데이터 갱신이 필요한 특성을 가지고 있다(Fu, 2011). 최근 데이터 수집 기술의 발전에 따라서 관측 빈도가 높아지면서 주식거래, 판매 내역, 심전도, 강수량, 온도 등 다양한 분야에서 대용량의 시계열 데이터를 수집할 수 있게 되었다.

예측은 시계열 데이터를 이용한 대표적인 분석 문제이다. 시계열 예측 문제는 결측치가 없는 완전한 데이터를 필요로 하지만 관측 장비 손상 등의 원인으로 시계열 데이터의 결측이 빈번하게 발생하고 있다. 데이터 결측이 존재하는 경우 결측치 발생 이전 데이터를 사용할 수 없어 데이터 양이 감소하거나, 정확한 예측 결과를 기대할 수 없는 문제가 발생한다. 이와 같은 문제를 방지하기 위하여 시계열 데이터의 결측이 발생하지 않도록 대비하는 것도 중요하지만, 이미 발생한 결측치를 적절한 값으로 대체하는 방안이 필요하다.

이 논문 또는 저서는 2022년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임 (NRF-2022S1A5A2A01038550).
† 연락저자 : 민대기 교수, 03760 서울특별시 서대문구 이화여대길 52 이화여자대학교 경영학과, Tel : 02-3277-3923, Fax : 02-3277-2835, E-mail : dmin@ewha.ac.kr
2023년 3월 3일 접수; 2023년 4월 14일 수정본 접수; 2023년 6월 20일 게재 확정.

시계열 데이터의 결측치 대체 방법은 전통적인 통계적 기법(statistical method)과 데이터 기반의 기계학습 기법(machine learning method)으로 구분할 수 있다. 시계열 데이터의 결측치 대체를 위한 전통적 기법인 통계적 방법으로 평균값/최빈값, 기댓값 최대화(Expectation maximization; EM), 선형 보간(Linear Inperpolation), 선형회귀(Linear Regression; LR) 등이 있다. 평균값/최빈값으로 결측치를 대체하는 방법은 낮은 성능에도 불구하고 단순하고 직관적이며 적용하기 쉬운 장점이 있다(Lin and Tsai, 2020). 기댓값 최대화 기법은 결측치가 포함된 데이터로부터 최대 우도추정치(maximum likelihood)를 Expectation 단계(E-step)와 Maximization 단계(M-step)를 반복적으로 수행하면서 구하는 방법으로 결측치 대체에 효율적 방법으로 알려져 있다(Park *et al.*, 2005). EM 기법은 단조 수렴성의 특성과 함께 다양한 분야에 적용할 수 있는 장점이 있으나 결측치가 포함된 데이터의 경우 계산량이 많아지는 단점이 존재한다(Rahman and Islam, 2011). 선형 보간법은 주변 값들을 연결하여 누락된 항목을 대체하는 기법이다(Zhang, 2016). 마지막으로 선형회귀는 시계열 데이터의 결측치 대체에 많이 사용되는 대표적인 통계적 기법이다(Weisberg, 2005). 다변량 데이터의 경우 선형회귀식을 구성하기 위하여 least squares, ordinary least squares, local least squares, sequential local least square, iterated local least square 등의 방법을 사용하고 있다(Pati and Das, 2017). 통계적 기법은 적용이 간단한 장점이 있으나 차원이 높거나 변동성이 높은 데이터의 경우 예측력이 매우 낮아지는 단점이 있다(Ramosaj *et al.*, 2022).

결측치 대체를 위한 대표적인 기계학습 기법으로 K-nearest neighbor(KNN), 의사결정나무(Decision Tree), Random Forest (RF) 등을 고려할 수 있다(Lin and Tsai, 2020). KNN은 결측치와 가장 가까운 다수의 관측값을 이용하여 추정한 값으로 결측치를 대체하는 방법이다(Huang *et al.*, 2017). 의사결정나무는 일반적으로 결과 해석의 용이성, 이산 및 연속형 데이터 처리, 변수의 자동 선택, 이상치에 대한 모형의 강건성, 대규모 데이터의 처리 가능성 등 많은 장점이 있는 방법론으로 알려져 있다(Murphy, 2012). RF는 의사결정나무를 보완하여 예측성능을 개선한 방법론으로 결측치 처리를 위해 이를 응용한 다양한 연구가 있다(Feng *et al.*, 2021; Xia *et al.*, 2017). 하지만 의사결정나무와 RF는 결측치 대체 문제에 있어 다른 기계학습 기법과 비교하여 결측치 예측 성능이 낮아 독립적으로 사용하기보다는 다른 기법의 보완적 목적으로 많이 사용되고 있다(Burgette and Reiter, 2010; Rahman and Islam, 2011).

최근 학습 데이터의 규모가 제한적이거나 기계학습 모형이 너무 복잡하고 계산 비용이 큰 경우 전이학습(Transfer Learning)을 이용하여 결측치를 대체하는 연구가 제시되고 있다(Ma *et al.*, 2019). 전이학습은 완결성이 높은 유사 데이터를 이용하여 결측치 대체를 위한 모형을 학습하고, 학습 결과를 결측치 예측이 필요한 데이터에 적용함으로써 결측치 대체 성능을 개선하는 방법이다. 특히, 시계열 데이터에 대하여 큰 규모의 결측, 연속형 결측, 무작위 결측 등 다양한 결측치 대체 문제에서 전이학습의 기본 학습모형으로 LSTM(Long Short-Term Memory)의 효과성을 확인한 연구가 다수 제시되고 있다(Chen *et al.*, 2021).

결측치 대체 기법의 유형과 우수성과 독립적으로 과거 많은 연구에서 결측치를 적절하게 대체함으로써 시계열 예측 성능을 개선하는 결과를 확인하였다(Cao *et al.*, 2018; Lin and Tsai, 2020; Pan *et al.*, 2015; Yu *et al.*, 2022). 따라서 모형의 복잡도와 계산 비용이 증가함에도 불구하고 결측치 대체 성능이 우수한 기계학습이나 전이학습을 활용한 연구가 최근 제시되고 있다(Pan *et al.*, 2015; Yu *et al.*, 2022). 하지만 결측치를 실제 값과 유사하게 예측하기 위하여 필요한 모형의 복잡도와 계산 시간을 고려할 때 결측치 대체 비용과 예측 정확도의 개선을 종합적으로 고려하여 결측치 대체 기법의 적정성에 대한 판단이 필요하다. Zhu *et al.*(2019)은 결측치 대체 비용(imputation cost)과 노이즈를 줄이기 위해 예측을 위한 충분한 정보를 이미 내포하고 있는 결측 샘플(absent sample) 및 간단한 방법으로 결측치의 추정이 가능한 샘플(predictable sample)을 포함하는 대체가 필요하지 않은 결측 샘플(neednot-impute sample)을 감지하는 모델을 제안하였다. 또한 Le Morvan *et al.*(2020)은 우수한 결측치 대체 결과가 언제나 예측 성능 개선에 도움이 되지 않음을 주장하였다.

Table 1. Summary of Literature: Research Theme

Reference	proposition and/or validation of algorithms	Validation of Positive effects of missing value imputation	Condition for positive effects of missing value imputation
This study	○	○	◎
Pati and Das(2017), Burgette and Reiter(2010), Rahman and Islam(2011), Ma *et al.*(2019), Chen *et al.*(2021)	◎	○	
Cao *et al.*(2018), Lin and Tsai(2020), Yu *et al.*(2022)	○	◎	
Zhu *et al.*(2019), Le Morvan *et al.*(2020)	◎		○

◎ main research theme, ○ research subtheme

<Table 1>에 제시한 바와 같이 새로운 결측치 대체 기법을 제안하거나 결측치 대체를 통한 예측 성능 개선효과 검증에 초점을 두고 있는 기존 연구와 다르게 본 연구에서는 시계열 예측 문제를 대상으로 결측치 대체와 관련하여 다음과 같은 두 가지 가설을 수치실험을 통하여 검증하는 것을 목적으로 한다. 첫째, 본 연구에서는 결측치가 시계열 데이터의 전 구간에 걸쳐서 발생함으로써 학습 데이터의 규모가 제한적이지만, 유사한 시계열 데이터가 다수 가용한 상황을 고려하고 있다. 따라서 기존 연구에서 고려한 결측치 대체 기법 중에서 자체 시계열을 이용한 결측치 대체 보다는 전이학습을 기본 방법으로 고려하였다. 이때, 전이학습을 이용한 결측치 대체 문제에서 결측치를 포함한 데이터와 유사도가 높은 데이터를 이용하여 결측치 예측 모형을 학습함으로써 결측치 대체 성능을 개선하는 것이 가능한지 검증한다. 둘째, 핵심 연구 주제로서 변동성을 중심으로 시계열 데이터의 특성에 따라서 결측치 대체 성능과 예측 성능 사이의 상관성을 분석함으로써 언제나 우수한 결측치 대체가 필요한지 검증한다.

본 논문의 구성은 다음과 같다. 제2장에서는 연구가설을 제시하고, 제3장에서는 연구가설을 검증하기 위한 연구 방법론 및 실험계획을 정리한다. 제4장에서는 수치실험을 통하여 제2장에서 제시한 가설을 검증하고, 마지막 제5장에서는 본 연구의 결론과 향후 연구주제를 제시한다.

3. 연구 가설

본 연구에서는 결측치가 시계열 데이터의 전 구간에 걸쳐서 발생함으로써 학습 데이터의 규모가 제한적이지만, 유사한 시계열 데이터가 다수 가용한 상황을 고려하였다. 따라서 전이학습을 기본 방법으로 고려하였으며, 이때 전이학습을 이용한 결측치 대체는 두 단계로 구성된다. 첫째, 결측치를 포함하는 불완전 시계열 데이터가 아닌 별도의 독립적인 유사 시계열 데이터를 이용하여 예측모형을 학습한다. 두 번째 단계에서는 학습된 예측모형을 불완전 시계열 데이터에 적용하여 결측치를 추정하고 추정 결과를 결측치의 대체값으로 사용한다. 전이학습은 일반적으로 예측모형 학습에 사용한 데이터와 결측치를 포함한 불완전한 데이터가 유사한 특성을 갖는다는 가정을 기반으로 하고 있다(Chen *et al.*, 2019). 따라서 결측치 대체 성능과 예측 성능 사이의 관계를 분석하기에 앞서 전이학습을 이용하는 경우 결측치를 포함한 시계열 데이터와 학습용 시계열 데이터의 유사도가 결측치 대체 성능에 미치는 영향을 분석하는 것이 필요하며, [연구가설 1]을 이용하여 이를 검증하도록 한다.

일반적으로 시계열 데이터의 안정성 또는 변동성은 예측성능에 매우 중요한 요소이다(Schnaars, 1984). 시계열 데이터 예측모형의 성능이 전이학습을 이용한 결측치 대체 성능을 결정하는 핵심 요소임을 고려할 때, 시계열 데이터의 변동성에 따라서 결측치 대체 성능에 변화가 있을 것으로 예상할 수 있다(연구가설 1-1).

결측치 유형(missing value pattern)은 결측치 대체 방법론의 선택과 성능을 결정하는 주요 요소이다(Emmanuel *et al.*, 2021). 시계열 데이터를 대상으로 하는 많은 연구에서 결측율(missing rate), 결측치 발생 위치(missing location), 연속적인 결측치 발생 길이 등 결측치 유형을 고려하고 있다(Yang *et al.*, 2022). 따라서 전이학습의 결측치 대체 성능과 관련한 연구가설 1을 검증하는데 있어 결측치 유형을 고려하는 것이 필요하다. [연구가설 1-2]에서는 결측율과 결측치 발생 유형(위치와 길이)에 의한 효과를 검증한다.

본 연구의 목적은 결측치 대체 성능과 예측 성능 사이의 상관성을 분석함으로써 언제나 우수한 결측치 대체가 필요한지 검증하는데 있다. [연구가설 2]에서는 결측치의 예측 정확도가 높아지는 경우 이를 이용한 예측 성능의 개선이 가능한가를 분석하도록 한다. [연구가설 1-1] 및 [연구가설 1-2]에서는 시계열 데이터의 특성(즉, 변동성)과 결측치 유형에 따라서 결측치 대체 성능에 변화가 있음을 확인하였다. 따라서 연구가설 2의 검증 과정에서도 변동성, 결측 유형 등 시계열 데이터의 특성에 따라서 결과에 변화가 있는지 살펴보는 것이 필요하며, 이를 위하여 [연구가설 2-1]과 [연구가설 2-2]를 정의하였다. 본 논문에서 검증하고자 하는 연구가설을 요약하여 <Table 2>에 제시하였다.

Table 2. Research Hypothesis

Hypothesis 1	The higher the similarity between the time series containing missing values and the time series for transfer learning, the better the missing value imputation performance.
Hypothesis 1-1	The high volatility in time series decreases the missing value imputation performance.
Hypothesis 1-2	As the missing rate and the length of consecutive missing data increase, the missing value imputation performance decreases.
Hypothesis 2	There is a positive correlation between the missing value imputation performance and the prediction accuracy.
Hypothesis 2-1	The positive correlation between the missing value imputation performance and the prediction accuracy becomes insignificant when the time series has high volatility.
Hypothesis 2-2	The positive correlation between the missing value imputation performance and the prediction accuracy becomes insignificant when the missing rate of time series and the length of consecutive missing data increase.

3. 연구 방법론 및 실험 계획

본 논문에서는 제2장에서 제시한 연구가설을 검증하기 위하여 2단계의 연구 절차를 구성하였다. 첫번째 단계에서는 [연구가설 1]을 검증하기 위하여, 전이학습을 이용한 결측치 추정과 예측 정확도를 평가한다. 전이학습에서는 먼저 결측치를 포함한 시계열 데이터와 결측치를 포함하지 않은 다른 완결한 시계열 데이터 사이의 유사도를 측정하고, 이를 기준으로 완결한 데이터를 선정하여 시계열 예측 모형을 학습한다. 이후 결측치를 포함한 시계열 데이터를 예측 모형에 적용하여 결측치를 추정하고 이를 결측치의 대체값으로 사용한다. 두번째 단계에서는 첫번째 단계에서 도출한 결측치 대체 결과를 이용하여, 예측 수행을 위한 별도의 시계열 예측 모형을 구성하고 예측 성능을 평가함으로써 [연구가설 2]를 검증한다(<Figure 1>).

<Figure 1>에 제시된 연구절차를 재정리하면 다음과 같다. f_z와 g_x는 각각 결측치 추정 모델과 시계열 예측 모델을 나타낸다. 소스 시계열 데이터 $Z=\{z_t, z_{t-1}, ..., 1\}$를 이용하여 결측치 추정 모델 f_z를 학습하였다. 즉, $f_z(z_{t-1}, ..., z_{t-8})=z_t$. 이후 결측치를 포함하고 있는 목표 시계열 데이터 X에 결측치 추정 모델 f_z를 적용함으로써 결측치 x_t를 $f_z(x_{t-1}, \cdots, x_{t-8})=x_t$와 같이 추정한다. 결측치 대체가 완료된 시계열 X를 이용하여 시계열 예측 모델 g_x를 $g_x(x_{t-1}, ..., x_{t-8})=x_t$와 같이 학습한다. 여기서 결측치 추정 모델 f_z와 시계열 예측 모델 g_x에 활용된 LSTM 모델은 모두 동일한 조건으로 학습 후 비교하였다.

3.1 시계열 데이터 예측 모형

<Figure 1>의 연구절차에서 제시한 바와 같이 본 논문에서는 시계열 데이터의 결측치 추정과 결측치 대체 결과를 이용한 시계열 예측을 목적으로 예측 모형을 구성한다. 시간 순서에 따른 순차적 관계가 중요한 시계열 데이터의 예측 문제에서는 관측치 사이의 시간 순서를 반영하지 못하는 인공신경망(ANN; Artificial Neural Network)의 한계를 고려하여 RNN(Recurrent Neural Network)을 예측을 위한 기본 모형으로 사용하고 있다(Chen *et al.*, 2021). 특히, LSTM은 RNN의 발전된 형태로써 긴 시계열 데이터의 예측에 우수한 성능을 보이는 것으로 알려져 있다(Tian *et al.*, 2019). 따라서 본 논문에서는 LSTM을 이용하여 결측치 예측과 시계열 예측을 위한 모형을 구성하며, 학습 성능을 평가하는 수치실험을 반복적으로 수행하여 최적의 LSTM 모형을 구성하였다.

앞서 제2장에서 제시한 바와 같이 본 연구의 핵심 목적은 시계열 데이터의 예측 성능 개선이 아닌 결측치 대체 성능과 예측 성능 사이의 관계를 분석하는데 있다. LSTM 시계열 예측 모형의 성능 최적화는 본 연구의 범위를 벗어나지만, 결측치 대체 과정에서 가장 우수한 성능을 갖는 추정 기법을 선정하고, LSTM 결측치 대체 모형의 타당성을 검증하기 위하여 LSTM과 세 개의 benchmark 모델인 Simple Moving Average(SMA), Hull Moving Average(HMA), XGBoost의 추정 성능을 비교하는 수치실험을 수행하였다. 추정 성능 비교에서는 다수의 시계열 데이터로 구성된 연구 데이터를 이용하여 전이학습 모형을 구성하고, 4개의 추정 모형에 동일하게 적용함으로써 전이학습에 가장 적합한 모형을 선정하는 절차로 진행하였다.

3.2 전이학습과 시계열 데이터의 유사도

다수의 결측치를 포함하고 있는 시계열 데이터를 추정 모형의 학습에 사용하는 경우 우수한 대체 성능을 기대하는 것이 어렵다. 이와 같은 문제를 해결하기 위하여 본 논문에서는 전이학습을 이용하였다. 전이학습은 적용 대상에 따라서 샘플 전이(sample transfer), 모형 전이(model transfer), 관계 전이(relation transfer) 등으로 구분된다. 예측모형의 학습 결과를 전달하는 모형 전이는 학습에 많은 시간이 소요되는 단점이 있으나 일반적으로 성능이 가장 우수한 것으로 알려져 있다(Ma *et al.*, 2020). 따라서 본 연구에는 모형 전이를 이용하여 완전한 시계열 데이터로부터 모형을 학습하는 방법을 이용하였다.

모형 전이를 이용한 전이학습은 도메인(domain)과 작업(task)으로 정의한다. 도메인 $D=\{F, P(X)\}$이며, 여기서 F와 X는 각각 특성 공간(feature space)과 관측치 $X=\{\chi_1, \cdots, \chi_m\}\in F$를 의미한다. 또한 $P(X)$는 X의 주변확률분포(marginal probability distribution)이다. 작업 T는 라벨 공간(label space) Y와 예측모형 $f(x)$로 구성된다: $T=\{Y, f(x)\}$. 전이학습에서는 소스 도메인 D_S와 학습 작업 T_S, 그리고 목표 도메인 $D_T(\neq D_S)$와 학습 작업 $T_T(\neq T_S)$가 주어졌을 때, D_S와 T_S로부터 얻은 지식을 사용하여 목표 예측 함수 $f_T(\bullet)$

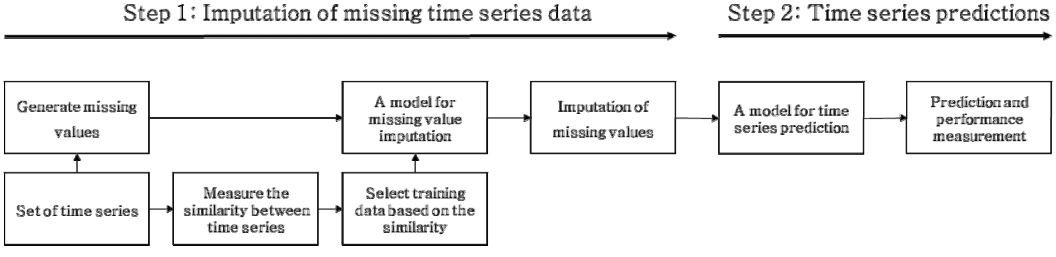

Figure 1. Research Framework

의 성능을 향상시키는 것을 목표로 한다(Tian *et al.*, 2019). 즉, 먼저 결측치를 포함하지 않은 완전한 시계열 데이터(소스 데이터)를 이용하여 예측모형을 학습하고, 학습결과를 결측치를 포함한 시계열 데이터(목표 데이터)를 예측하기 위한 모형으로 전달하여 예측을 수행하는 절차로 구성된다(Fawaz *et al.*, 2018).

전이학습 과정에서 소스 도메인(즉, 소스 데이터)과 학습을 적절하게 선정하는 것은 목표 예측 함수의 성능개선에 중요하다. 본 연구에서는 목표 데이터와 소스 데이터로 고려하는 시계열 데이터 사이의 유사도(similarity)를 기준으로 전이학습을 위한 소스 데이터를 선정하였다. 유사도에 의한 전이학습의 성능 효과를 분석하기 위하여 목표 데이터와 유사도가 가장 높은 데이터(Short distance), 유사도가 가장 낮은 데이터(Long distance) 그리고 유사도가 가장 높거나 낮은 두 데이터를 제외한 나머지 데이터 중 임의로 선택한 데이터(Random distance)를 대상으로 성능을 비교하였다. 시계열 데이터의 유사도 측정 기법으로 유클리드 거리(Euclidean distance)와 DTW(Dynamic Time Warping)를 대상으로 고려하였으며, 사전 수치실험 결과 40개의 시계열 데이터 중에서 유클리드 거리와 DTW를 이용하여 소스 데이터를 선정하였을 때 13개의 데이터에서 상이한 소스 데이터 선정 결과를 확인하였다. 13개 시계열 데이터를 대상으로 결측값 대체 성능을 비교한 결과, 유클리드 거리를 이용하여 전이학습을 수행한 경우 평균적으로 보다 더 우수한 결과를 나타냄을 확인하였다.

3.3 데이터와 결측치 유형

결측치 대체를 위한 전이학습 적용을 위하여 다수의 시계열 데이터로 구성된 뉴욕 311 서비스의 민원 접수 건수(311 service requests volume) 데이터를 사용하였다. 311 서비스는 미국과 캐나다의 시민들이 시(市)에서 제공하는 서비스와 정보에 좀 더 빠르고 쉽게 접근할 수 있도록 편의를 제공하기 위한 목적으로 운영하는 서비스이다. NYC 311 서비스는 뉴욕시에 관한 주요 정보를 제공함과 동시에 응급서비스를 제외한 모든 정부 서비스를 제공한다. 본 연구는 NYC Open Data(https://opendata.cityofnewyork.us/)에서 제공하는 데이터를 활용하였다. Ho *et al.*(2019)과 Sanneh *et al.*(2021)은 시계열 예측을 위한 새로운 모형을 개발하는데 NYC Open Data를 활용하였다. 본 연구에서는 동일한 데이터를 이용하되 서론에서 제시한 바와 같이 시계열 데이터의 결측치 대체 방법과 결측치 대체 효과를 검증하고자 한다.

본 연구에서는 2018년 1월 1일부터 2020년 3월 31일까지 26개월 동안의 민원 접수 이력을 사용하였으며, 본 논문에서는 가장 접수 건 수가 많은 상위 8개 부서만 실험에 사용하였다. 데이터의 원본은 접수 시간, 접수 부서(Agency), 지역(Borough) 등 총 41개의 변수로 구성되어 있지만, 전처리를 통해 총 4개의 변수(접수일자, 월, 요일, 접수 건수)를 포함하는 구조로 변형하였다. 그 결과 5개 지역과 8개 접수 부서 조합에 의해 총 40개의 일별 민원 접수 건수 데이터가 각각 4개의 변수를 포함하는 시계열 데이터로 재구성되었다. 40개의 데이터 모두를 순차적으로 결측치를 포함한 목표 데이터로 설정한 뒤, 설정된 목표 데이터를 제외한 나머지 39개의 완전한 시계열 데이터 중 하나(소스 데이터)를 유사도에 따라 선택하여 결측치 추정을 위한 LSTM 모델의 학습에 활용하였다. 소스 데이터는 24개월 기간의 일별 정보(총 730개 샘플)를 포함하고 있으며, 결측치를 포함한 데이터(목표 데이터)를 성능평가에 사용하였다. 예측 모형의 경우 2020년 1월 1일 이전 데이터를 학습에 사용하였다. 따라서 각 데이터 당 예측 모형의 학습에 사용된 샘플의 수는 결측치 예측 모형과 마찬가지로 730개이며, 예측 대상 샘플은 91개이다. 40개의 데이터 모두에 대해 실

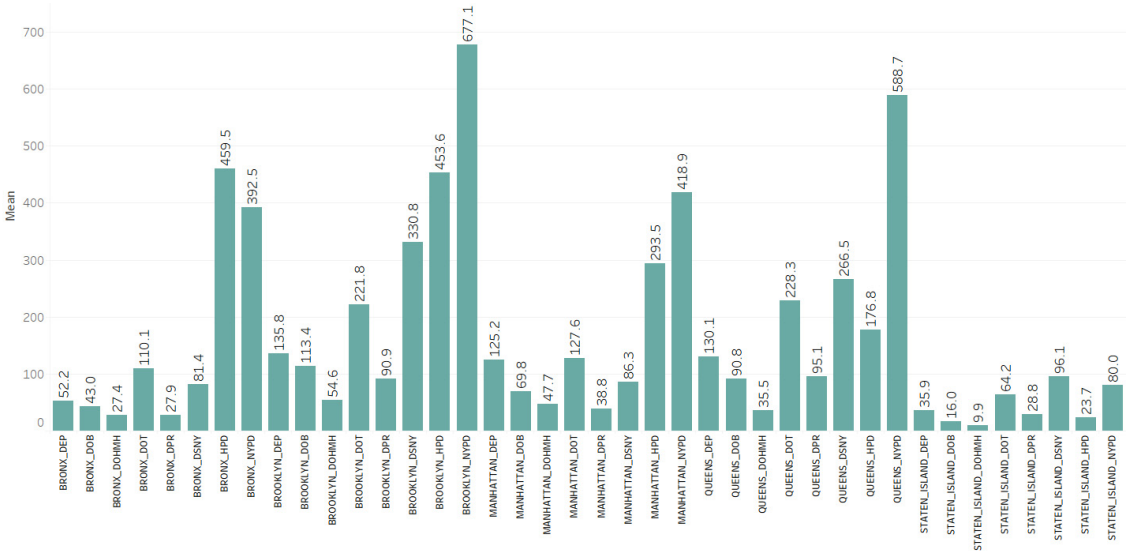

Figure 3. Average Daily Service Requests

Figure 4. Missing Points of Mixed Type(Missing Rate 30%)

험을 실시한 뒤 분석에 활용하였는데, 40개의 각 데이터의 규모가 다른 것을 <Figure 3>을 통해 알 수 있다. 따라서 본 실험에서 사용된 모든 평가 지표는 산출된 값을 각 데이터의 평균으로 나누어 표준화한 뒤 분석에 활용하였다.

시계열 데이터의 결측치 유형은 기존 문헌을 참고하여 결측 위치와 결측률을 고려하여 정의하였다. 먼저 결측 위치는 일반적으로 연속 결측(continuous missing)과 혼합 결측(mixed type of missing)으로 구분된다(Ma *et al.*, 2020). 연속 결측은 임의의 위치에서 연속적으로 결측이 발생하는 형태이며, 혼합 결측은 단일데이터의 결측과 연속 결측이 혼합된 유형을 의미한다. 연속 결측의 경우 임의의 결측 위치를 지정하고 결측률에 해당하는 데이터를 연속으로 결측 처리하였으며, 혼합 결측의 경우 결측치의 위치와 길이를 임의로 생성하였다. 문헌에 따라서 고려하고 있는 결측율의 범위는 10%에서 90%까지 상이하지만, 대부분 10%~50%의 범위를 고려하고 있다(Xia *et al.*, 2017; Zhu *et al.*, 2019; Ma *et al.*, 2020). 본 논문에서는 극단적으로 결측율이 높은 상황을 배제하고 데이터 결측 유형에 따라서 본 연구에서 고려하는 연구가설의 결과를 검증하기 위하여 10%에서 60%까지 10%p 단위로 결측률을 고려하였다.

결측률(6개 유형)과 결측치 위치(2개 유형)의 조합에 따라서 총 12개의 결측 유형을 실험에서 고려하였다. 예를 들어, 결측률이 30%이고 혼합 결측인 경우의 결측 샘플 생성 절차는 다음과 같다. 가장 먼저 40개의 일별 민원 접수 건수 시계열 데이터 중에서 목표 데이터를 한 개 선정한다. 이후, 선정한 시계열 데이터를 대상으로 설정한 실험 조건(결측률과 결측치 위치)에 따라 결측 샘플을 임의로 생성한다. 결측 샘플이 생성된 결과를 그림으로 표현한 것이 <Figure 4>이다. <Figure 4>에서 파란색 세로 실선은 혼합 결측일 때 무작위로 선정한 결측 샘플을 의미한다.

3.4 수치 실험 절차

제2장에서 제시한 연구가설을 검증하기 위한 수치 실험 절차는 <Table 4>와 같다. 총 40개의 시계열 데이터 중 한 개를 목표 데이터로 이용함으로써 수치 실험 조건 별로 총 40회의 교차 반복 실험을 수행하였다. LSTM 모형의 성능은 RMSE를 이용하였으며, 40회의 수치 실험 결과에 따른 RMSE 값들을 대상으로 통계적 유의성을 검정하였다(<Table 5>).

Table 4. Numerical Analysis Procedure

Research Hypothesis	Test Procedure
Hypothesis 1	• Select one of 40 time series data as a target and generate missing values according to the type of missing values • Measure the similarity between the target and the other 39 time series • Select source data according to the similarity(Short, Long, Random) and train the LSTM model • Impute missing values in target data using the LSTM model • Repeat the above procedure for 40 time series data and evaluate imputation performance
Hypothesis 2	• Select one of 40 time series data as a target and generate missing value according to the type of missing value • Measure the similarity between the target and the other 39 time series • Impute missing values in target data using the LSTM model that is trained with source data of a high similarity • Train another LSTM model for prediction with the time series containing imputed data, and then evaluate its prediction accuracy. • Repeat the above procedure for 40 time series data and evaluate prediction performance.

Table 5. Statistical Test Method

Research Hypothesis	Statistical Validation Method
Hypothesis 1	• Evaluation of imputation performance with respect to the similarity in time series(Short, Long, Random). • ANOVA test for the differences in mean RMSE with respect to the similarity
Hypothesis 1-1	• t-test for the difference in average RMSE between groups of time series data by the coefficient of variation of 0.6.
Hypothesis 1-2	• ANOVA test for differences in mean RMSE with respect to the type of missing values
Hypothesis 2	• Correlation analysis between the imputation performance and the prediction performance
Hypothesis 2-1	• Correlation analysis between the imputation and prediction performance for each set of times series classified by a coefficient of variation of 0.6
Hypothesis 2-2	• Correlation analysis between the imputation and prediction performance for each set of time series classified by the type of missing values

4. 수치 실험 결과

4.1 결측치 대체 모형

본 논문에서는 시계열 데이터의 결측치 추정을 위한 전이학습 기본 모형으로 LSTM 추정 모형을 사용하였다. 연구가설을 검정하기에 앞서 LSTM 결측치 추정 모형의 적정성을 평가하기 위하여 SMA, HMA, XGBoost 등 세 개의 추정 모형과 결측치 대체 성능을 비교하였다. 3.3절에서 제시한 연구 데이터와 결측치 유형을 대상으로 전이학습을 통한 결측치 대체를 수행하였으며, 결측치 대체 성능은 결측치 추정 모형을 통하여 예측한 결측치 대체 값과 실제 값의 차이를 세 가지 지표(MAE, MSE, RMSE)로 측정하여 비교하였다.

<Figure 6>과 <Figure 7>은 시계열 데이터의 결측 유형별로 40개의 목표 데이터를 대상으로 측정한 RMSE의 평균을 나타낸 그래프이다. 시계열 데이터의 결측 위치(연속, 혼합)와 결측 비율(10%~60%)에 따른 12가지 조건 모두에서 유사도가 높은 시계열 데이터(소스 데이터)를 이용하여 학습한 LSTM 모

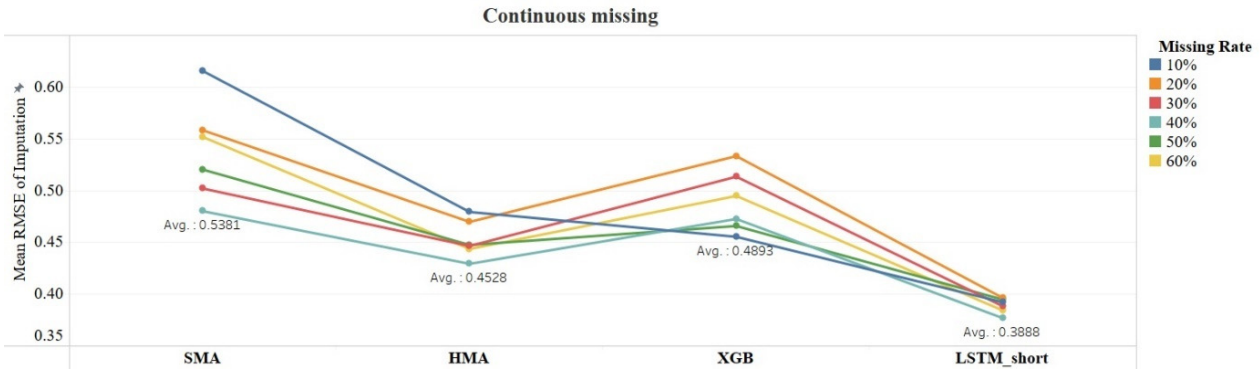

Figure 6. Imputation Performance of Benchmark Models(Continuous Missing)

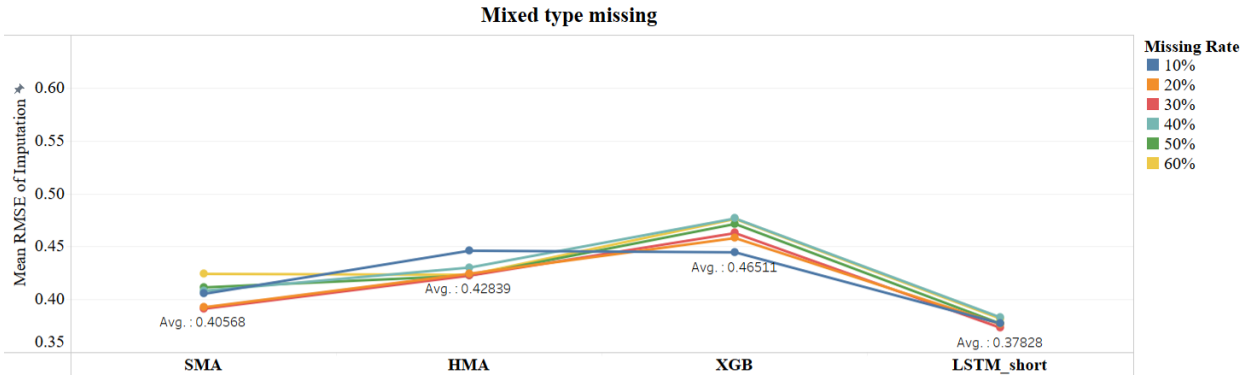

Figure 7. Imputation Performance of Benchmark Models(Mixed Type Missing)

형(LSTM_short; LSTM model with short distance)의 결측치 추정 성능이 가장 우수하였다. 반면에 유사도가 가장 낮은 소스 데이터를 이용하여 전이학습을 수행한 경우(LSTM model with long distance), 모든 실험 조건에서 결측치 추정 성능이 가장 낮은 것을 확인하였다. 이는 세 가지 성능 지표(MAE, MSE, RMSE) 모두에서 공통된 결과를 나타냈다.

이상의 수치실험 결과를 기반으로 본 논문에서는 연구가설을 검증하기 위한 기본 전이학습 모형으로 LSTM을 이용한 분석을 진행하였다. 최종적으로 본 연구에서 사용한LSTM모델의 구성은 다음과 같다. t 시점의 값을 예측하기 위하여 입력 데이터는 t-8부터 t-1 시점까지의 월(month), 요일(weekday), 민원접수 건수(incoming call count)을 사용하였다. 또한, 은닉층(Hidden layer)은 4개의 노드로 구성하였으며, adam optimizer, 휴버 손실함수(Huber loss function)로 모형을 구성하였다.

4.2 시계열 데이터의 유사도와 결측치 대체 성능: 연구가설 1

첫 번째 연구가설에서는 결측치를 포함하는 시계열 데이터와 전이학습에 사용한 시계열 데이터(소스 데이터) 사이의 유사도와 결측치 대체 성능의 상관성을 검증한다. <Table 6>은 결측 유형 별로 유사도에 따른 결측치 대체 성능(평균 RMSE)을 나타낸 결과이다. 먼저, 유사도가 높은(즉, 유클리드 거리가 짧은) 시계열 데이터로 학습한 LSTM 모형을 이용하여 결측치를 추정한 결과 평균 RMSE가 가장 낮았으며, 반대로 학습에 사용한 시계열 데이터의 유사도가 낮은 경우 높은 추정 오차가 발생하였다. 이와 같은 결과는 결측치 유형과 관계없이 모두 동일하게 발생하였으며, 결측 유형별로 ANOVA 검증 결과 평균 RMSE의 차이가 통계적으로 유의미함을 확인하였다. 따라서 결측치를 포함하는 시계열 데이터와 전이학습을 위한 시계열 데이터(소스 데이터)의 유사도가 높을수록 결측치 대체 성능이 개선될 것이라는 [연구가설 1]이 성립함을 확인하였다.

가장 유사도가 높은 소스 데이터로 학습한 LSTM 모델(LSTM_short)의 결측치 유형별 평균 RMSE의 차이에 대한 ANOVA 분석 결과 p-value=0. 9999로 통계적 유의성을 확인할 수 없었다. MSE(p-value= 0.7597)와 MAE(p-value= 0.9999)에 대한 동일한 분석도 마찬가지의 결과가 나타났다. 즉, 결측치 유형에 따라서 연구가설 1의 결과에 차이가 있을 것을 예상한 [연구가설 1-2]는 성립하지 않았다.

시계열 데이터의 변동성에 따라서 [연구가설 1]의 결과 차이를 예상한 [연구가설 1-1]을 검증하였다. <Table 7>은 결측치 유형별로 목표 시계열 데이터의 변동계수에 따른 결측치 추정 모델의 평균 RMSE 차이를 비교한 결과이다. 수치 실험 결과 전이학습을 활용한 결측치 추정 성능은 목표 시계열 데이터의 변동성이 증가할수록 감소(즉, 추정 오차 증가)하는 것을 확인하였다. [연구가설 1]에 대한 검증 결과를 <Table 8>에 요약하였다.

Table 6. Source Data Similarity and Missing Value Imputation Error(Mean RMSE)

Missing Value Type		Source Data Similarity Level (Euclidean Distance)		
Missing Location	Missing Rate	High (Short)	Random (Random)	Low (Long)
Continuous	10%	0.3923	0.6094	6.1310
Continuous	20%	0.3964	0.6659	5.7097
Continuous	30%	0.3884	0.7359	5.7376
Continuous	40%	0.3769	0.7688	6.0450
Continuous	50%	0.3948	0.7421	6.0301
Continuous	60%	0.3843	0.7201	6.0675
Mix	10%	0.3774	0.6465	1.9559
Mix	20%	0.3766	0.6421	1.9605
Mix	30%	0.3734	0.6411	1.9310
Mix	40%	0.3831	0.6506	1.9086
Mix	50%	0.3772	0.6498	1.9337
Mix	60%	0.3819	0.6568	1.9685

Table 7. Variability and Missing Value Imputation Errors (RMSE) in Time Series Data

Missing Value Type		Coefficient of Variation (CV)	Mean RMSE	T-Test (p-value)
Missing Location	Missing Rate			
Continuous	10%	< 0.6	0.3252	0.0057
		> 0.6	0.4929	
Continuous	20%	< 0.6	0.3079	0.0008
		> 0.6	0.5290	
Continuous	30%	< 0.6	0.2933	0.0000
		> 0.6	0.5310	
Continuous	40%	< 0.6	0.2830	0.0001
		> 0.6	0.5177	
Continuous	50%	< 0.6	0.3083	0.0001
		> 0.6	0.5245	
Continuous	60%	< 0.6	0.3055	0.0004
		> 0.6	0.5024	
Mix	10%	< 0.6	0.2893	0.0001
		> 0.6	0.5096	
Mix	20%	< 0.6	0.2908	0.0001
		> 0.6	0.5053	
Mix	30%	< 0.6	0.2901	0.0002
		> 0.6	0.4982	
Mix	40%	< 0.6	0.2933	0.0001
		> 0.6	0.5179	
Mix	50%	< 0.6	0.2900	0.0001
		> 0.6	0.5081	
Mix	60%	< 0.6	0.2955	0.0001
		> 0.6	0.5115	

Table 8. Test Results - Research Hypothesis 1

Research Hypothesis	Description	Result
Hypothesis 1	The higher the similarity between the time series containing missing values and the time series for transfer learning, the better the missing value imputation performance.	True
Hypothesis 1-1	The high volatility in time series decreases the missing value imputation performance.	True
Hypothesis 1-2	As the missing rate and the length of consecutive missing data increase, the missing value imputation performance decreases.	False

4.2 결측치 대체 성능과 예측 성능의 상관성: 연구가설 2

두 번째 연구가설에서는 결측치 대체 성능과 시계열 예측 성능의 상관성을 검증한다. 이를 위하여 결측치 대체 성능과 시계열 예측 성능을 나타내는 예측 정확도(RMSE) 사이의 상관분석을 수행하였다. <Table 9>는 결측치 유형과 시계열 데

Table 9. Correlation Analysis between Imputation Performance and Prediction Performance(RMSE)

Missing Value Type		Coefficient Of Variation	Pearson Correlation	Correlation Analysis
Missing Location	Missing Location	(CV)	Coefficient	(p-value)
Continuous	10%	All Data	0.4255	0.0062
		< 0.6	0.2240	0.2927
		> 0.6	0.2735	0.3054
Continuous	20%	All Data	0.5226	0.0005
		< 0.6	0.4506	0.0271
		> 0.6	0.2670	0.3174
Continuous	30%	All Data	0.5853	0.0001
		< 0.6	0.4586	0.0242
		> 0.6	0.4160	0.1090
Continuous	40%	All Data	0.6697	0.0000
		< 0.6	0.3837	0.0641
		> 0.6	0.6552	0.0059
Continuous	50%	All Data	0.5135	0.0007
		< 0.6	0.4003	0.0526
		> 0.6	0.6263	0.0094
Continuous	60%	All Data	0.6854	0.0000
		< 0.6	0.5599	0.0044
		> 0.6	0.6943	0.0028
Mix	10%	All Data	0.5913	0.0001
		< 0.6	0.1158	0.5899
		> 0.6	0.5405	0.0307
Mix	20%	All Data	0.5993	0.0000
		< 0.6	0.5707	0.0036
		> 0.6	0.3656	0.1638
Mix	30%	All Data	0.6442	0.0000
		< 0.6	0.4221	0.0399
		> 0.6	0.5962	0.0148
Mix	40%	All Data	0.5629	0.0002
		< 0.6	0.4232	0.0394
		> 0.6	0.5106	0.0433
Mix	50%	All Data	0.6600	0.0000
		< 0.6	0.5291	0.0078
		> 0.6	0.5975	0.0145
Mix	60%	All Data	0.5932	0.0001
		< 0.6	0.5643	0.0041
		> 0.6	0.4856	0.0565

Table 10. Test Results - Research Hypothesis 2

Research Hypothesis	Description	Result
Hypothesis 2	There is a positive correlation between the missing value imputation performance and the prediction accuracy.	True
Hypothesis 2-1	The positive correlation between the missing value imputation performance and the prediction accuracy becomes insignificant when the time series has high volatility.	True
Hypothesis 2-2	The positive correlation between the missing value imputation performance and the prediction accuracy becomes insignificant when the missing rate of time series and the length of consecutive missing data increase.	False

이터의 변동성에 따른 상관분석 결과를 나타낸다.

<Table 9>의 상관분석 결과를 살펴보면 결측치 유형에 관계없이 모든 경우에 있어 결측치 대체 성능이 우수할수록 예측 성능 또한 개선되는 양의 상관관계가 통계적으로 유의미하게 존재하는 것을 확인하였다(전체 데이터 결과). 즉, 기존 연구에서 제시한 결과와 동일하게 예측 성능을 개선하기 위해서는 결측치를 정확하게 추정하여 대체하는 것이 중요함을 확인하였다.

시계열 데이터의 변동성을 고려하지 않은 전체 데이터를 대상으로 하는 경우와는 다르게 변동성을 의미하는 변동계수에 따라서 시계열 데이터를 구분하는 경우 상관분석 결과에 차이가 발생하였다. 먼저, 전반적으로 연속 결측과 혼합 결측 모두 변동계수가 낮은 데이터보다 높은 데이터의 대체 성능과 예측 성능 사이에 실시한 피어슨 상관분석의 상관계수가 높은 경향이 있다. 변동계수가 0.6보다 작은 경우에는 연속 결측과 혼합 결측 모두 결측률 10%인 경우를 제외하고는 유의수준 10%에서 상관계수에 통계적 유의성을 가졌다. 반면 변동계수가 0.6보다 큰 경우 연속 결측 유형에서는 결측률이 낮을 때(10~30%)에는 상관계수에 통계적 유의성을 가지지 않았고 결측률이 높은 경우(40~60%)에만 유의미한 상관관계를 나타냈다. 연속 결측 유형과는 다르게 혼합 결측 유형 조건에서는 데이터의 변동성이 큰 경우 결측률 20%인 경우를 제외하고는 모든 결측 비율에서 통계적으로 유의미한 양의 상관관계를 보였다. 이와 같은 결과는 결측 위치가 혼합 유형으로 결측치의 연속 구간이 길지 않을 때에는 데이터의 변동성이 크더라도 결측 비율에 관계없이 결측치를 정교하게 추정하는 것이 예측 성능 개선에 도움을 준다는 것을 알려준다. 한편 결측치의 연속 구간이 긴 연속 결측 유형이면서 변동성이 큰 데이터의 경우 결측률이 작을 때에는 정확하게 결측치를 추정하는 것이 예측 성능 개선에 유의미한 영향을 주지 않지만, 연속 결측 유형인 경우에도 결측률이 큰 경우에는 정확한 결측치 대체가 예측 성능 개선에 도움이 된다는 것을 의미한다. [연구가설 2]에 대한 검증 결과를 <Table 10>에 요약하였다.

5. 결론

시계열 데이터의 예측 문제에서 데이터 완결성은 우수한 예측 성능을 확보하는데 매우 중요한 요소이다. 본 연구에서는 먼저 전이학습 방법을 이용하여 시계열 데이터의 결측치를 적절하게 추정하여 대체하는 것이 가능함을 제시하였다. 또한 정확한 결측치 대체를 통하여 예측 성능을 개선하는 것이 가능한지 결측치 대체 성능과 예측 성능 사이의 상관성에 대한 연구가설을 검증하였다.

본 논문의 분석결과 다음과 같은 흥미로운 결과를 확인하였다. 첫째, 전이학습을 통하여 결측치 대체 성능 개선이 가능하며, 이때 결측치 대체 성능 개선을 위해서는 유사도가 높은 시계열 데이터를 이용하는 것이 중요하다. 특히, 유사도가 낮은 경우 결측치 유형에 따라서 결측치 대체 성능에 차이가 발생하였으나, 유사도가 높은 시계열 데이터를 전이학습에 이용하는 경우 결측치 유형에 따른 결측치 대체 성능의 차이를 확인할 수 없었다. 둘째, 기존 연구와 동일하게 정확한 결측치 대체를 통하여 예측 성능을 개선하는 것이 가능함을 확인하였다. 하지만 시계열 데이터의 변동성이 낮고 결측 구간이 길지 않은 안정적 데이터의 경우 결측치 대체 성능과 예측 성능 사이에 유의미한 상관성이 존재하지 않았다. 따라서 시계열 데이터가 변동성이 작고 결측 비율이 낮은 안정적 특성을 갖는 경우 결측치 대체 성능을 개선하기 위하여 전이학습과 같이 계산 비용이 많이 소요되는 방법을 사용하는 것 보다는 SMA와 같이 결측치 대체 성능은 우수하지 않지만 계산이 간단한 방법론을 사용하는 것을 고려할 수 있다.

본 논문은 다음과 같은 몇 가지 추가 연구가 필요하다. 첫째, 다수의 상이한 특성 및 도메인을 가진 시계열 데이터를 대상으로 연구결과를 검증하는 것이 필요하다. 본 연구에서는 실제 데이터를 대상으로 수치실험을 수행하고 연구가설을 검증하였으나, 연구 결과의 타당성을 강화하기 위하여 다양한 도메인과 특성을 갖는 데이터를 대상으로 실험을 보완하는 것이 필요하다. 둘째, 시계열 데이터의 특성을 고려하여 LSTM과 함께 최신 시계열 예측 방법론을 고려하는 것이 필요하다. 예를 들어, 간헐적(intermittent) 특성을 갖는 시계열 데이터의 경우 LSTM과 비교하여 Croston 기법을 확장한 Deep Renewal Process 기법이 보다 우수한 예측 성능을 갖는 것으로 알려져 있다(Türkmen *et al.*, 2021). 마지막으로 본 논문에서는 기존 문헌을 참고하여 시계열 데이터의 결측치 유형과 변동성 등 실험 조건을 구성하였다. 하지만 실험 조건을 보다 더 세밀하게

구성함으로써 각 요인 별 효과를 정확하게 분석하는 것이 가능할 것이다.

참고문헌

Burgette, L. F. and Reiter, J. P. (2010), Multiple Imputation for Missing Data Via Sequential Regression Trees, *American Journal of Epidemiology*, **172**(9), 1070-1076.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., and Li, Y. (2018), Brits: Bidirectional Recurrent Imputation for Time Series, *Advances in Neural Information Processing Systems*, **31**.

Chen, Y., Wang, J., Huang, M., and Yu, H. (2019), Cross-position Activity Recognition with Stratified Transfer Learning, *Pervasive and Mobile Computing*, **57**, 1-13.

Chen, Z., Xu, H., Jiang, P., Yu, S., Lin, G., Bychkov, I., Hmelnov, A., Ruzhnikov, G., Zhu, N., and Liu, Z. (2021), A Transfer Learning-Based LSTM Strategy for Imputing Large-Scale Consecutive Missing Data and Its Application in a Water Quality Prediction System, *Journal of Hydrology*, **602**, 126573.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021), A Survey on Missing Data in Machine Learning, *Journal of Big Data*, **8**(1), 1-37.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2018), Transfer Learning for Time Series Classification, *2018 IEEE International Conference on Big Data (Big Data)*, 1367-1376.

Feng, R., Grana, D., and Balling, N. (2021), Imputation of Missing Well Log Data by Random Forest and Its Uncertainty Analysis, *Computers & Geosciences*, **152**, 104763.

Fu, T.-C. (2011), A Review on Time Series Data Mining, *Engineering Applications of Artificial Intelligence*, **24**(1), 164-181.

Ho, N., Vo, H., Vu, M., and Pedersen, T. B. (2019), Amic: An Adaptive Information Theoretic Method to Identify Multi-scale Temporal Correlations in Big Time Series Data, *IEEE Transactions on Big Data*, **7**(1), 128-146.

Huang, J., Keung, J. W., Sarro, F., Li, Y.-F., Yu, Y.-T., Chan, W., and H. Sun (2017), Cross-validation Based K Nearest Neighbor Imputation for Software Quality Datasets: An Empirical Study, *Journal of Systems and Software*, **132**, 226-252.

Le Morvan, M., Prost, N., Josse, J., Scornet, E., and Varoquaux, G. (2020), Linear Predictor on Linearly-generated Data with Missing Values: Non Consistency and Solutions, *International Conference on Artificial Intelligence and Statistics*, 3165-3174.

Lin, W.-C. and Tsai, C.-F. (2020), Missing Value Imputation: A Review and Analysis of the Literature (2006–2017), *Artificial Intelligence Review*, **53**(2), 1487-1509.

Ma, J., Cheng, J. C., Jiang, F., Chen, W., Wang, M., and Zhai, C. (2020), A Bi-directional Missing Data Imputation Scheme Based on LSTM and Transfer Learning for Building Energy Data, *Energy and Buildings*, **216**, 109941.

Ma, J., Cheng, J. C., Lin, C., Tan, Y., and Zhang, J. (2019), Improving Air Quality Prediction Accuracy at Larger Temporal Resolutions Using Deep Learning and Transfer Learning Techniques, *Atmospheric Environment*, **214**, 116885.

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.

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*, MIT press.

Pan, R., Yang, T., Cao, J., Lu, K., and Zhang, Z. (2015), Missing Data Imputation by K Nearest Neighbours Based on Grey Relational Structure and Mutual Information, *Applied Intelligence*, **43**(3), 614-632.

Park, J. H., Kang, M. S., Lee, J. O., and Kang, S. J. (2005), Handling Missing Data: What is the Most Effective Method?, *The Korean Journal of Physical Education*, **44**(1), 385-398.

Pati, S. K. and Das, A. K. (2017), Missing Value Estimation for Microarray Data Through Cluster Analysis, *Knowledge and Information Systems*, **52**(3), 709-750.

Rahman, G. and Islam, Z. (2011), A Decision Tree-based Missing Value Imputation Technique for Data Pre-processing, *paper presented at Proceedings of the Ninth Australasian Data Mining Conference-Volume* 121.

Ramosaj, B., Tulowietzki, J., and Pauly, M. (2022), On the Relation between Prediction and Imputation Accuracy under Missing Covariates, *Entropy*, **24**(3), 386.

Sanneh, J., Cohall, M., Lee, J., Wang, Y., Martínez García, D., and Keck, J. (2021), Spatiotemporal and Machine Learning-Based Time Series Assessment of Drinking Water Quality Complaints in New York City, *paper presented at World Environmental and Water Resources Congress* 2021.

Schnaars, S. P. (1984), Situational Factors Affecting Forecast Accuracy, *Journal of Marketing Research*, **21**(3), 290-297.

Tian, Y., Sehovac, L., and Grolinger, K. (2019), Similarity-based Chained Transfer Learning for Energy Forecasting with Big Data, *IEEE Access*, **7**, 139895-139908.

Türkmen, A. C., Januschowski, T., Wang, Y., and Cemgil, A. T. (2021), Forecasting Intermittent and Sparse Time Series: A Unified Probabilistic Framework Via Deep Renewal Processes, *Plos One*, **16**(11), e0259764.

Weisberg, S. (2005), *Applied Linear Regression*, John Wiley & Sons.

Xia, J., Zhang, S., Cai, G., Li, L., Pan, Q., Yan, J., and Ning, G. (2017), Adjusted Weight Voting Algorithm for Random Forests in Handling Missing Values, *Pattern Recognition*, **69**, 52-60.

Yang, J., Shao, Y., Li, C., and Wang, W. (2022), Multistage Large Segment Imputation Framework Based on Deep Learning and Statistic Metrics, arXiv preprint arXiv:2209.11766.

Yu, L., Zhou, R., Chen, R., and Lai, K. K. (2022), Missing Data Preprocessing in Credit Classification: One-hot Encoding or Imputation?, *Emerging Markets Finance and Trade*, **58**(2), 472-482.

Zhang, Z. (2016), Missing Data Imputation: Focusing on Single Imputation, *Annals of Translational Medicine*, **4**(1).

Zhu, X., Yang, J., Zhang, C., and Zhang, S. (2019), Efficient Utilization of Missing Data in Cost-sensitive Learning, *IEEE Transactions on Knowledge and Data Engineering*, **33**(6), 2425-2436.

<부록>

Table A-1. Data Dictionary of NYC 311 Service Requests

Column Name	Description	Example
Unique Key	Unique identifier of a Service Request (SR) in the open data set	56757534
Created Date	Date SR was created	02/08/2023 10:26:11 AM
Closed Date	Date SR was closed by responding agency	02/08/2023 01:48:52 PM
Agency	Acronym of responding City Government Agency	NYPD
Agency Name	Full Agency name of responding City Government Agency	New York City Police Department
Complaint Type	This is the fist level of a hierarchy identifying the topic of the incident or condition. Complaint Type may have a corresponding Descriptor (below) or may stand alone.	Noise - Residential
Descriptor	This is associated to the Complaint Type, and provides further detail on the incident or condition. Descriptor values are dependent on the Complaint Type, and are not always required in SR.	Loud Music/Party
Status	Status of SR submitted	Closed
Due Date	Date when responding agency is expected to update the SR. This is based on the Complaint Type and internal Service Level Agreements (SLAs).	02/09/2023 12:00:00 AM
Resolution Action Updated Date	Date when responding agency last updated the SR.	02/08/2023 01:48:56 PM
Resolution Description	Describes the last action taken on the SR by the responding agency. May describe next or future steps.	The Police Department responded and upon arrival those responsible for the condition were gone.
Location Type	Describes the type of location used in the address information	Residential Building/House
Incident Zip	Incident location zip code, provided by geo validation.	11429
Incident Address	House number of incident address provided by submitter.	217-32 100 AVENUE
Street Name	Street name of incident address provided by the submitter	100 AVENUE
Cross Street 1	First Cross street based on the geo validated incident location	217 STREET
Cross Street 2	Second Cross Street based on the geo validated incident location	99 AVENUE
Intersection Street 1	First intersecting street based on geo validated incident location	217 STREET
Intersection Street 2	Second intersecting street based on geo validated incident location	99 AVENUE
Address Type	Type of incident location information available.	ADDRESS
City	City of the incident location provided by geovalidation.	QUEENS VILLAGE
Landmark	If the incident location is identified as a Landmark the name of the landmark will display here	100 AVENUE
Facility Type	If available, this field describes the type of city facility associated to the SR	DSNY Garage
Community Board	Provided by geovalidation.	13 QUEENS
BBL	Borough Block and Lot, provided by geovalidation. Parcel number to identify the location of location of buildings and properties in NYC.	4107610020

Column Name	Description	Example
Borough	Provided by the submitter and confirmed by geovalidation.	QUEENS
X Coordinate (State Plane)	Geo validated, X coordinate of the incident location.	1,056,150
Y Coordinate (State Plane)	Geo validated, Y coordinate of the incident location.	200,029
Open_Data_Channel_Type	Indicates how the SR was submitted to 311. i.e. By Phone, Online, Mobile, Other or Unknown.	ONLINE
Latitude	Geo based Lat of the incident location	40.71541797382096
Longitude	Geo based Long of the incident location	-73.740635739316 13
Location	Combination of the geo based lat & long of the incident location	(40.71541797382096, -73.740635739316 13)
Park Facility Name	If the incident location is a Parks Dept facility, the Name of the facility will appear here	Unspecified
Park Borough	The borough of incident if it is a Parks Dept facility	QUEENS
Vehicle Type	If the incident is a taxi, this field describes the type of TLC vehicle.	Ambulette / Paratransit
Taxi Company Borough	If the incident is identified as a taxi, this field will display the borough of the taxi company.	BRONX
Taxi Pick Up Location	If the incident is identified as a taxi, this field displays the taxi pick up location	41-11 10 STREET, QUEENS (LONG ISLAND CITY), NY, 11101
Bridge Highway Name	If the incident is identified as a Bridge/Highway, the name will be displayed here.	1
Bridge Highway Direction	If the incident is identified as a Bridge/Highway, the direction where the issue took place would be displayed here.	1 Local Uptown & The Bronx
Road Ramp	If the incident location was Bridge/Highway this column differentiates if the issue was on the Road or the Ramp.	Grove St & W 4 St to Uptown & The Bronx
Bridge Highway Segment	Additional information on the section of the Bridge/Highway where the incident took place.	Entrance

Table A-2. Missing Value Imputation Error

Missing Value Type		Missing Value Imputation Error																	
		Mean MAE						Mean MSE						Mean RMSE					
Missing Location	Missing Rate	SMA	HMA	XGB	High (Short)	Random (Random)	Low (Long)	SMA	HMA	XGB	High (Short)	Random (Random)	Low (Long)	SMA	HMA	XGB	High (Short)	Random (Random)	Low (Long)
Continuous	10%	0.5270	0.3966	0.3607	0.3190	0.5191	6.0831	48.9972	30.9083	30.1422	18.7402	62.9862	2265.0754	0.6156	0.4795	0.4555	0.3923	0.6094	6.1310
Continuous	20%	0.4670	0.3853	0.4359	0.3190	0.5747	5.6494	49.4760	38.0272	52.4914	24.6409	77.0065	1980.1895	0.5584	0.4699	0.5334	0.3964	0.6659	5.7097
Continuous	30%	0.4187	0.3677	0.4235	0.3157	0.6462	5.6875	46.2358	38.4879	50.6809	24.1416	83.1748	2002.5660	0.5021	0.4466	0.5136	0.3884	0.7359	5.7376
Continuous	40%	0.4010	0.3529	0.3882	0.3068	0.6832	6.0012	40.6910	33.8833	42.3271	21.3262	79.8226	2226.8549	0.4802	0.4294	0.4727	0.3769	0.7688	6.0450
Continuous	50%	0.4412	0.3718	0.3836	0.3232	0.6560	5.9823	47.8410	35.7549	41.0901	21.8084	79.9655	2216.2513	0.5204	0.4476	0.4659	0.3948	0.7421	6.0301
Continuous	60%	0.4761	0.3707	0.4146	0.3170	0.6371	6.0224	57.1638	34.4309	46.8904	20.1617	76.6043	2243.4646	0.5519	0.4437	0.4950	0.3843	0.7201	6.0675
Mix	10%	0.3304	0.3671	0.3660	0.3074	0.5635	1.7041	25.8331	32.8153	35.5909	17.3746	75.4582	526.3943	0.4055	0.4463	0.4448	0.3774	0.6465	1.9559
Mix	20%	0.3205	0.3508	0.3853	0.3086	0.5644	1.7757	24.4747	28.9800	36.5630	16.6965	69.5694	545.4120	0.3928	0.4246	0.4583	0.3766	0.6421	1.9605
Mix	30%	0.3226	0.3517	0.3934	0.3078	0.5645	1.7721	24.3941	29.6884	38.1594	16.6641	70.1082	538.6101	0.3915	0.4228	0.4629	0.3734	0.6411	1.9310
Mix	40%	0.3338	0.3557	0.4047	0.3143	0.5702	1.7515	28.8496	32.6216	42.7243	20.0073	72.5802	533.0410	0.4083	0.4304	0.4768	0.3831	0.6506	1.9086
Mix	50%	0.3389	0.3509	0.4019	0.3109	0.5715	1.7861	30.1928	31.2089	42.0210	19.2481	72.9010	547.9158	0.4116	0.4228	0.4716	0.3772	0.6498	1.9337
Mix	60%	0.3504	0.3506	0.4055	0.3139	0.5779	1.8217	32.0117	32.4384	42.9395	20.1691	75.7637	568.7071	0.4244	0.4235	0.4763	0.3819	0.6568	1.9685

Table A-3. ANOVA Analysis of The Performance of All Imputation Models

Missing Value Type		MAE		MSE		RMSE	
Missing Location	Missing Rate	f_statistic	p-value	f_statistic	p-value	f_statistic	p-value
Continuous	10%	25.0128	0.0000	27.9872	0.0000	24.6027	0.0000
Continuous	20%	22.0974	0.0000	22.9941	0.0000	21.7716	0.0000
Continuous	30%	23.7743	0.0000	26.2377	0.0000	23.4046	0.0000
Continuous	40%	22.0914	0.0000	22.4340	0.0000	21.8136	0.0000
Continuous	50%	22.0374	0.0000	22.5229	0.0000	21.7844	0.0000
Continuous	60%	22.1086	0.0000	22.7179	0.0000	21.8610	0.0000
Mix	10%	8.3445	0.0000	7.8059	0.0000	10.1019	0.0000
Mix	20%	9.1461	0.0000	8.5481	0.0000	10.2299	0.0000
Mix	30%	9.0346	0.0000	8.1920	0.0000	9.8833	0.0000
Mix	40%	8.6878	0.0000	7.8567	0.0000	9.4778	0.0000
Mix	50%	9.0789	0.0000	8.3295	0.0000	9.7860	0.0000
Mix	60%	9.4543	0.0000	8.9945	0.0000	10.1367	0.0000

Table A-4. Variability and Missing Value Imputation Errors(MSE) in Time Series Data

Missing Value Type		Coefficient of Variation (CV)	Mean MSE	T-Test (p-value)
Missing Location	Missing Rate			
Continuous	10%	< 0.6	18.1018	0.7370
		> 0.6	19.6978	
Continuous	20%	< 0.6	14.9084	0.0372
		> 0.6	39.2397	
Continuous	30%	< 0.6	12.9343	0.0024
		> 0.6	40.9524	
Continuous	40%	< 0.6	12.0839	0.0013
		> 0.6	35.1896	
Continuous	50%	< 0.6	14.6538	0.0022
		> 0.6	32.5402	
Continuous	60%	< 0.6	14.2451	0.0025
		> 0.6	29.0366	
Mix	10%	< 0.6	12.5882	0.0037
		> 0.6	24.5543	
Mix	20%	< 0.6	12.6337	0.0023
		> 0.6	22.7906	
Mix	30%	< 0.6	12.6452	0.0023
		> 0.6	22.6924	
Mix	40%	< 0.6	12.9439	0.0016
		> 0.6	30.6024	
Mix	50%	< 0.6	12.6530	0.0007
		> 0.6	29.1408	
Mix	60%	< 0.6	13.3467	0.0004
		> 0.6	30.4029	

Table A-5. Variability and Missing Value Imputation Errors(MAE) in Time Series Data

Missing Value Type		Coefficient of Variation (CV)	Mean MAE	T-Test (p-value)
Missing Location	Missing Rate			
Continuous	10%	< 0.6	0.2630	0.0071
		> 0.6	0.4030	
Continuous	20%	< 0.6	0.2530	0.0010
		> 0.6	0.4180	
Continuous	30%	< 0.6	0.2421	0.0001
		> 0.6	0.4260	
Continuous	40%	< 0.6	0.2331	0.0003
		> 0.6	0.4173	
Continuous	50%	< 0.6	0.2553	0.0006
		> 0.6	0.4251	
Continuous	60%	< 0.6	0.2539	0.0013
		> 0.6	0.4116	
Mix	10%	< 0.6	0.2395	0.0004
		> 0.6	0.4094	
Mix	20%	< 0.6	0.2406	0.0005
		> 0.6	0.4104	
Mix	30%	< 0.6	0.2411	0.0006
		> 0.6	0.4079	
Mix	40%	< 0.6	0.2442	0.0003
		> 0.6	0.4193	
Mix	50%	< 0.6	0.2418	0.0004
		> 0.6	0.4146	
Mix	60%	< 0.6	0.2449	0.0003
		> 0.6	0.4175	

저자소개

황희선 : 창원대학교 독어독문학과, 경영학과에서 2013년 학사 학위를 취득하고 2023년 2월 이화여자대학교 빅데이터분석학 협동과정 석사학위를 취득하였다. 연구분야는 빅데이터 분석, 데이터 마이닝, 데이터 품질이다.

민대기 : 서울대학교 산업공학과에서1999년학사, 2001년석사 학위를 취득하고, 퍼듀대학교에서 산업공학 박사학위를 취득하였다. 2010년부터 이화여자대학교 경영대학에 재직 중이다. 연구분야는 Markov Decision Process, 강화학습, 비즈니스 애널리틱스, 에너지 시스템이다.