한국빅데이터학회지
제4권 제2호, 2019, pp. 105-110

미세먼지 자료에서의 결측치 대체 방법 비교

Comparision of Missing Imputaion Methods In fine dust data

김연진·박헌진†

인하대학교 통계학과

요 약

자료 분석에 있어서 결측치 대체는 큰 이슈중 하나이다. 결측치의 발생을 무시하고 분석을 진행하게 되면, bias가 발생하여 그에 따른 추정치에 대해 잘못된 결과를 줄 수 있다. 이 논문에서는 미세먼지자료에서 발생한 결측치를 적절한 대체 방법을 찾아 적용하자 한다. 이를 통해 시계열 자료에서 발생한 결측치를 R을 기반으로 한MICE, MissForest 등의 기존 방법과 시계열 기반 모델을 사용하여 여러 가지 상황에 대한 시뮬레이션을 설정해 비교해 밝히고자 하였다. 이 결과에 대해 각각을 변수 별로 비교하였을때 ImputeTS 패키지를 이용한 auto arima 모델의 kalman filter를 적용한 모형과 MissForest 모형이 미세먼지자료 결측치 대체에서는 좋은 결과를 주는 것으로 판단되었다.

■ 중심어: 결측치 대체, 미세먼지 자료, MICE, MissForest, ImputeTS

Abstract

Missing value replacement is one of the big issues in data analysis. If you ignore the occurrence of the missing value and proceed with the analysis, a bias can occur and give incorrect results for the estimate. In this paper, we need to find and apply an appropriate alternative to missing data from weather data. Through this, we attempted to clarify and compare the simulations for various situations using existing methods such as MICE and MissForest based on R and time series-based models. When comparing these results with each variable, it was determined that the kalman filter of the auto arima model using the ImputeTS package and the MissForest model gave good results in the weather data.

■ Keyword : Missing Imputataion, Climate Data, MICE, MissForest, ImputeTS

2019년 11월 29일 접수; 2019년 12월 10일 수정본 접수; 2019년 12월 31일 게재 확정.
† 교신저자 (hjpark@inha.ac.kr)

Ⅰ. 서 론

미세먼지 같은 역학 연구에서 결측치 처리는 주요 관심사 중 하나이다. 이러한 결측치 발생에 의한 불완전한 자료는 자료 분석 시 모델에서의 편향된 모수 추정등의 요소에서 문제가 발생하여 잘못된 결과를 초래할 수 있기 때문에, 적절한 결측치 처리는 분석을 할 때 중요한 요소라고 할 수 있다. 특히 미세먼지 자료의 가장 큰 특징 중 하나는 시계열 자료라는 것이다. 따라서, 이 논문에서 논의하고자 하는 내용은 시계열 자료에서의 결측치가 발생하였을 때 적절한 대체 방법을 찾고 논의 하는 것이다.[1]

결측치는 다양한 이유를 통해 발생할 수 있고, 자료 자료의 결측의 이유를 알 수 없는 경우가 많다. 또한 더 많은 자료 자료를 모으더라도 원래 자료의 결측치 발생으로 인한 문제점들을 해결할 수 없는 경우가 대부분이다. 따라서 자료 분석을 통한 모형을 적합 시에는 불완전한 자료 처리에 대한 적절한 전략 선택은 필수적이다. 이에 대한 논의는 이전부터 진행되어 왔는데, 이를 설명하기 위한 결측 자료 발생의 매커니즘은 Rubin(1976)을 통해 분류되었다.[2] 이후 결측치의 패턴을 분류하기 위해 군집분석을 수행하여 나누는 방식을 수행할 수 있다. Johannes Bauer et al.(2013).[3]

결측치 대체에 대한 접근 방법은 크게 2가지로 분류된다. 하나는 Likelihood 기반의 방법이고 다른 하나는 Imputation 기반의 방법이다. (Little and Rubin, 1989). Likelihood 기반의 결측치 대체 방법은 유연하게 작용할 수 있지만, 복잡한 계산을 필요로 한다. 그에 반해 Imputation 기반의 결측치 대체 방법은 계산이 더 단순하다. 이에 대한 논의는 이후에 이어서 진행하였다.[4]

이 논문에서는 시계열 자료에서 결측치 발생시 적절한 대체 방법을 적용하여 비교 연구를 하고자 한다. 비교를 위한 방법은 Mice[5], Amelia[6], MissForest[7], ImputeTS[8], DTWBI[9] 등의 결측치 대체 방법이다. 이를 위해 R에서 지원하는 Mice, Amelia, MissForest, ImputeTS, DTWBI 패키지를 활용하였다.

2장에서는 결측치 매커니즘과 패턴에 대한 분류 과정에 대해 논의한다. 3장에서는 MCAR test와 Little test에 대한 소개를 하고 이후 4장에서는 다중대체방법을 소개한 뒤, 5장에서는 미세먼지 자료에 기반한 시뮬레이션의 결측치 대체 방법 비교 결과를 파악한 뒤, 6장에서는 5장에서의 결과에 대해 논의한다.

Ⅱ. 결측치 패턴 분류

결측치 자료는 다양한 패턴으로 발생될 수 있다. 자료의 결측이 나타나는 패턴은 결측치의 처리 방법을 선택할 때 주의 깊게 고려해야할 요소 중 하나인데, 결측치 발생 패턴을 파악하면 그에 맞는 효과적인 결측치 대체 방법을 사용할 수 있다. 그래서 이번 장에서는 결측치 발생 패턴에 대해 언급하고자 한다.

2.1 표기법 정의

Missing 매커니즘에 대한 언급에 앞서 몇 가지 필요한 표기법에 대해 정의하고자 한다. 먼저 자료가 $X_{N\times P}$인 관측치가 N개, 변수의 수가 P개인 행렬로 구성이 되어있을 때, [1]우리의 자료는 알 수 없는 모수 θ의 분포 함수 $g(X|\theta)$로부터 생성된다. 그리고, 결측치가 나타난 위치에 대한 행렬을 M으로 정의한다. 이에 대해 다음과 같이 예시를 들 수 있다.

$X=\begin{bmatrix} x_{11} & NA & x_{13} \\ x_{21} & x_{22} & NA \\ NA & x_{32} & x_{33} \end{bmatrix}$이라는 3×3행렬이 주어져 있을 때, M은 다음과 같이 나타난다.

$$M=\begin{bmatrix}0\,1\,0\\0\,0\,1\\1\,0\,0\end{bmatrix}\text{ when data is missing } m_{ij}=0.$$

2.2 결측치 매커니즘

자료의 결측은 Missing Completely At Random (MCAR), Missing At Random(MAR), Missing Not At Random(MNAR)라는 3가지 매커니즘으로 분류된다. 우리가 가지고 있는 불완전한 데이터 셋 X는 $X=(X^{obs},X^{miss})$으로 나타낼 수 있다. 그리고, M을 결측치 패턴에 대한 행렬로 표현하면, $f(M|X,\phi)$를 통해 조건부 분포와 연관지을 수 있고 이때 ϕ는 unknown parameter이다. 이를 이용해 앞서 언급한 3가지 매커니즘에 대해 아래와 같이 설명할 수 있다.

2.1.1 Missing Completely At Random (MCAR)

자료의 결측 여부가 해당 값의 특징과는 무관하게 일어난 것으로 자료의 관찰여부와는 독립적으로 나타난 경우를 말한다.

$$f(M|X,\phi)=f(M|\phi).$$

2.1.2 Missing At Random (MAR)

자료의 결측 여부가 결측된 자료에서는 해당 값의 특징과는 무관하게 일어난 것으로 독립적으로 나타난 경우를 말한다.

$$f(M|X,\phi)=f(M|X^{obs},\phi).$$

2.1.3 Missing Not At Random (MNAR)

자료의 결측 여부가 해당 값의 특징과는 연관되어 일어난 것으로 자료에 의존하여 나타나는 경우를 말한다.

$$f(M|X,\phi)=f(M|X^{obs},X^{miss},\phi).$$

위의 3가지 결측치 매커니즘 중에 MCAR과 MAR은 ignorable 한 성질을 지닌다. Ignorability는 다음과 같이 설명할 수 있다.

가능도에 기반한 추정을 할 때, $L^{*}(\theta,\psi|X_i,W_i,y_i,r_i)\propto f(y_i,r_i|X_i,W_i\theta,\psi)$로 나타낼 수 있다. 여기서 ψ는 r_i에 대한 모수인데, r_i는 y_i가 관찰된 경우 1, 결측치가 발생한 경우 0을 나타내는 지시함수로 나타나는 변수이다. 그런데 자료분석을 할 때 관찰된 자료를 통해서만 진행이 되기 때문에, L^{*}를 L로 대체하여 다음과 같은 가능도 함수로 쓸 수 있다.

$L(\theta,\psi|X_i,W_i,y_i^{\circ},r_i)\propto f(y_i^{\circ},r_i|X_i,W_i\theta,\psi)$이고 [11] 이는 아래와 같이 쓸 수 있다. MAR 가정하에서 위의 밀도함수는 다음과 같이 쓸 수 있다.

$$\begin{aligned}&f(y_i^{\circ},r_i|\theta,\psi)\\&=\int f(y_i^{\circ},y_i^{m}|X_i,\theta)f(r_i|y_i^{\circ},y_i^{m},W_i,\psi)dy_i^{m}\\&=f(y_i^{\circ}|X_i,W_i,\theta)f(r_i|y_i^{\circ},W_i,\psi).\end{aligned}$$

과 같이 MAR 가정하의 확률밀도함수는 위와 같이 인수분해가 가능하다. 이러한 성질을 이용해 Multiple Imputation을 적용할 수 있다.

이에 반해 MNAR을 통한 결측치 대체방법은 매우 조심히 다루어야하며, MNAR을 통해 발생한 결측치의 경우 우리가 관찰한 자료들만으로는 제대로 된 결측치의 추정이 어려울 수 있다. 이런 경우엔 추가적인 자료의 수집 또는 도메인 전문가들의 통찰이 유용하게 작용할 수 있다고 알려져 있다. 이중 MCAR은 가정이 매우 강하고 성립하기 어렵지만, 많은 결측치 대체 방법은 Ignorability 성질을 기반으로 하고 있다. 이러한 성질을 만족하는지 확인하기 위해 다음 절에서는 MCAR에 대한 test 방법을 확인한다.

Ⅲ. MCAR test

대부분의 경우 주어진 자료에서 결측치가 어떤 매커니즘을 통해 발생하였는지에 대해 결정하기는 일반적으로 불가능하다. 그러나 다음에 나올 Little's test를 통해서 자료의 결측치 패턴이 MCAR 가정과 일치하는 지에 대한 가설을 검정해 볼 수 있다고 알려져있다. MCAR 가정을 명시적으로 검정하는 방법 중 하나는 기초통계량인 평균을 비교하는 것인데, 관찰된 X^{obs}의 평균과 X^{miss}의 평균의 차이를 비교하는 방법등을 예로 들 수 있다. 예를 들어 변수 v_1을 고려하였을 때 v_1에서 결측치가 발생한 경우에 해당하는 행 번호에서 다른 변수 $v_i(i \neq 1)$에 대한 평균 μ_{miss}과 v_1가 관찰된 경우에 해당하는 다른 변수 $v_i(i \neq 1)$에 대한 평균 μ_{obs}을 이용해 $H_0 : |\mu_{miss} - \mu_{obs}| \leq \epsilon$에 대한 평균 비교를 진행하여 확인해 볼 수 있다. 만약 v_1에 대한 결측치 패턴이 다른 관찰된 자료와 관련이 되어있다면, 우리는 위의 귀무가설을 기각해야한다. 그러나 이렇게 개별적으로 실행되는 t-검정은 해당 패턴과 결과를 해석하는데 있어서 어려움이 발생할 수 있다.[11] 이에 대한 어려움을 해결하기 위한 검정방법을 다음에서 소개한다.

3.1 Little's test

Little's test는 자료가 평균 벡터 μ와 공분산 행렬 Σ을 갖는 Multi Variate Normal(MVN)의 분포임을 가정한다.[11] 결측치의 패턴에 대해 $m = 1,\ldots,N_m$으로 나타내면 각 패턴이 갖는 값에 대해 $x_{obs,m}$으로 나타 낼 수 있다. 이를 이용하여 나타낸 로그 가능도 형태는 다음과 같다.

$$d^2 = \Sigma_m w_m (\overline{x_{obs,m}} - \mu_{obs,m}) \Sigma_m^{-1} (\overline{x_{obs,m}} - \mu_{obs,m})^T$$

여기서, w_m은 패턴 $m = 1,\ldots,N_m$이 나타난 수의 가중 비율이다.

Little에 의해 d^2은 χ^2검정과 같음이 밝혀졌고, MCAR 가정의 기각 여부는 d^2값에 의해 정해진다. 만약 d^2이 주는 p-value가 유의수준보다 낮다면, MCAR 가정을 기각하게 된다. 여기서 자료의 global covarinace matrix는 알 수 가 없는데, 이를 알기 위해 EM 알고리즘등의 방식을 적용할 수 있다.

Ⅳ. 다중 대체 방법

다중 대체 방법은 다음 순서로 진행된다.[5]

① Imputation : 결측 자료에 대한 Single Imputation을 N번 반복한다.
② Analysis : 각 N개의 Data Set을 분석하여 관심 있는 모수를 추정한다.
③ Pooling : N개의 결과를 결합한다.

이러한 과정의 Multiple Imputation의 목적은 관찰 및 관찰되지 않은 자료 f(X)의 분포 함수에 대한 적절한 근사를 하는 데 있다. 이것은 일반적으로 반복적인 메커니즘을 통해 달성되는데, f(X)를 통해 발생하는 다양한 결측 패턴에 대한 조건부 분포 함수에서 샘플링하여 결측치를 대체하는 방식으로 진행된다.

이 중 특정 프레임 워크를 연쇄 방정식으로 진행하는 것을 MICE 라고 한다. 이때 사용되는 연쇄 방정식은 자료 값과 매개 변수 값이 일련의 단계로 생성되는 반복 절차를 나타낸다. MICE에 대한 언급은 이후에 다루기로 한다.

다중 대체 방법에서 사용되는 일반적인 가정은 자료가 다변량 분포 함수 $p(X|\theta)$에서 생성된다는 것인데, 여기서 θ는 알려지지 않은 모수이다. 특정 경우에, 분포 함수 p는 특정 형태를 갖는 것으로 가정 될 수있다. 예를 들어, 자료가 Multi variate normal, 즉 $X \sim N(\mu, \Sigma)$이고 $\theta = (\mu, \Sigma)$를 따를 때 완전한 자료가 생성된다는 가정이다. 이를 통해

모든 분포 함수를 명시적으로 지정할 수 있으며 절차가 조금 더 명확해진다고 알려져있다.[5]

Ⅴ. 자료 분석

5.1 자료 설명

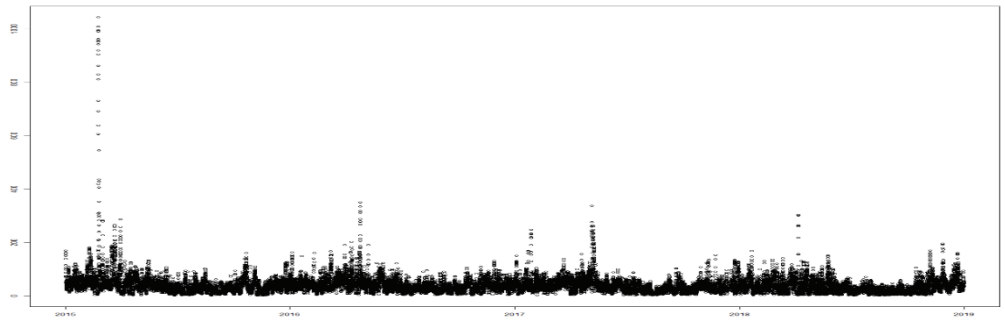
미세먼지청 서울 종로구에서 측정한 2015년 1월 1일부터 2018년 12월 31일에 해당하는 시간별 자료에 대한 11개의 미세먼지 자료를 수집하여 다음과 같은 자료 셋을 구성하였다.

이때 자료의 개수는 33952개였고, 수집한 변수는 미세먼지 농도(pm 10), 평균기온, 강수량, 풍속, 풍향, 습도, 증기압, 이슬점 온도, 기압, 적설량, 지면 온도였다. 이를 통해 시계열도를 분석해본 결과 총 4가지의 시계열 패턴으로 미세먼지자료가 나누어짐을 확인할 수 있었고 해당 그림은 다음과 같다.

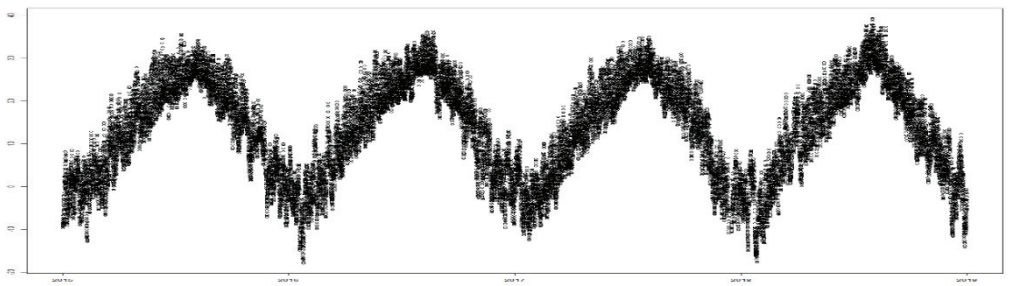
이때, 평균 기온의 경우엔 이슬점, 지면 온도등의 변수들과 비슷한 패턴을 나타내는 시계열도를 갖고 있음을 확인하였고, 나머지 강수량, 미세먼지 농도, 풍속에 대해서는 각각 고유한 다른 패턴이 진행되고 있음을 확인하였다. 이를 토대로 각각

〈표 1〉 2015년 1월 1일부터 2018년 12월 31일까지의 기상자료

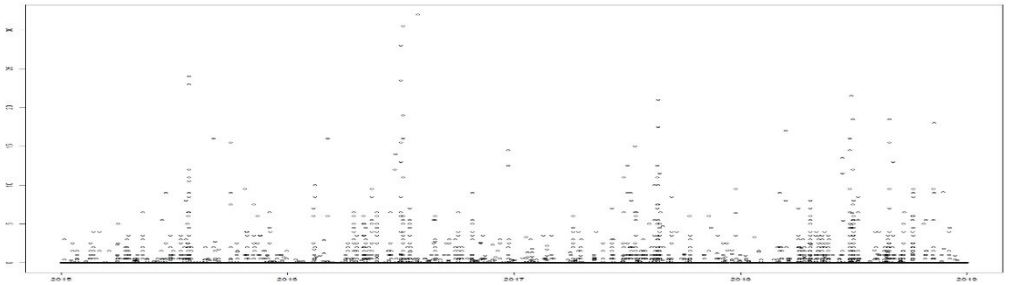
일시	미세먼지농도	평균기온	강수량	풍속		적설량
2015-01-01 01:00:00	51	-7.4	0	4.7	…	0
2015-01-01 02:00:00	72	-8.0	0	4.5		0
2015-01-01 03:00:00	88	-8.4	0	3.8		0
⋮						



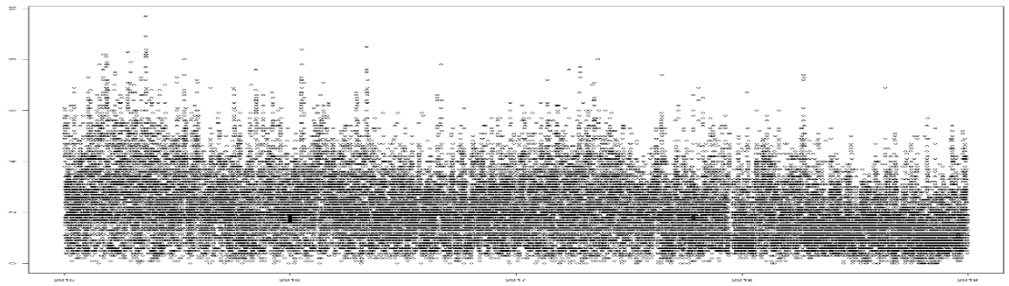
〈그림 1〉 시간대 별로 나타난 미세먼지 농도(pm 10)에 대한 시계열도.



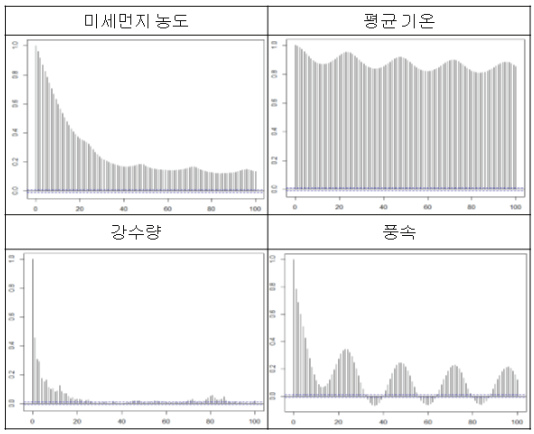
〈그림 2〉 시간대 별로 나타난 평균기온에 대한 시계열도.



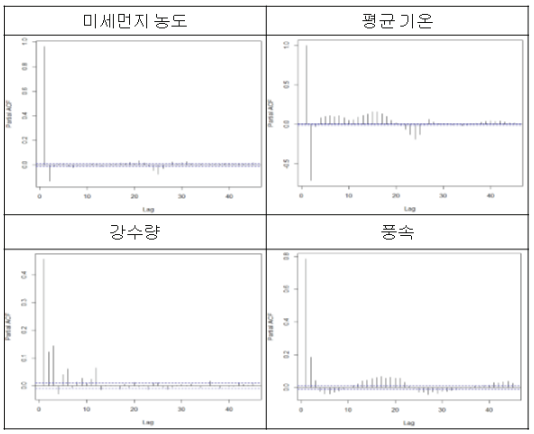
〈그림 3〉 시간대 별로 나타난 강수량에 대한 시계열도.



〈그림 4〉 시간대 별로 나타난 풍속에 대한 시계열도.



〈그림 5〉 각 변수에 대한 acf



〈그림 6〉 각 변수에 대한 pacf

선정한 미세먼지 농도, 평균기온, 강수량, 풍속 변수에 대한 acf와 pacf는 다음 그림 5, 그림 6과 같이 나타났다.

위의 그림 5, 그림 6을 통해 미세먼지 농도와 강수량의 경우엔 특별한 계절성을 지닌다고 보기 어려웠으나, 평균기온과 풍속의 경우엔 계절성이 존재한다고 판단하였다. 이러한 자료 탐색을 통해 각 상황에 대한 결과는 다음절에 명시해 놓았다.

5.2 시뮬레이션 자료 생성

미세먼지 자료에서 결측치가 발생하는 여러 가지 특징을 고려하여 3가지 경우에 해당하는 시뮬레이션 자료를 생성하였다. 그에 대한 설명은

다음과 같다.

시뮬레이션 ① : 총 결측치 비율을 전체 자료의 3%로 하여 각 변수에 대해 결측치를 랜덤하게 발생시킨 시뮬레이션 자료.

시뮬레이션 ② : 총 결측치 비율을 전체 자료의 20%로 하여 각 변수에 대해 결측치를 랜덤하게 발생시킨 시뮬레이션 자료.

시뮬레이션 ③ : 1개의 변수 A는 2017년 1월 1일부터 4월 30일까지의 결측치를 발생 시킨 뒤, 나머지 변수는 ②를 기반으로 하여 생성한 시뮬레이션 자료. (이때, 변수 A는 위에서 요약한 미세먼지 농도, 평균 기온, 강수량, 풍속으로 변경해가며 총 4개의 데이터 셋을 생성.)

위 3가지 종류의 시뮬레이션 데이터 셋은 다음의 표와 같다.

또한, 위 3가지 시뮬레이션 자료에 대해 Little test를 진행한 결과 p-value의 값이 0에 가까운 것을 확인하였고, 이를 통해모두 MCAR 가정이 깨져 있음을 확인할 수 있었다. 이는 시계열 자료라는 특징에 의해 당연한 결과로 나타났다고 판단할 수 있다.

미세먼지자료는 시계열적인 특징을 지니고 있어 시계열 결측치 대체 방법에 많이 이용되는 MICE, MissForest, Amelia, ImputeTS, DTWBI를 통해 결측치 대체를 진행하였다. 이때 DTWBI는 결측치 발생이 특정 긴 시간동안 이뤄진 경우에 사용되는 대체 방법인 이유로 시뮬레이션 ③에서만 활용하였다.

〈표 2〉 총 결측치 비율을 전체 자료의 3%로 하여 각 변수에 대해 결측치를 랜덤하게 발생시킨 시뮬레이션 자료

변수	미세먼지농도	기온	강수량	풍속	풍향
결측치 수	1133	1294	1228	1257	958
변수	**기압**	**이슬점**	**지면기압**	**적설량**	**지면온도**
결측치 수	971	957	739	847	861

〈표 3〉 총 결측치 비율을 전체 자료의 20%로 하여 각 변수에 대해 결측치를 랜덤하게 발생시킨 시뮬레이션 자료

변수	미세먼지농도	기온	강수량	풍속	풍향
결측치 수	8489	4881	8515	5685	8511
변수	**기압**	**이슬점**	**지면기압**	**적설량**	**지면온도**
결측치 수	4245	4275	8489	8518	6817

〈표 4〉 미세먼지 농도는 2017년 1월 1일부터 4월 30일까지의 결측치를 발생 시킨 뒤, 나머지 변수는 ②를 기반으로 하여 생성한 시뮬레이션 자료

변수	미세먼지농도	기온	강수량	풍속	풍향
결측치 수	2870	4881	8515	5685	8511
변수	**기압**	**이슬점**	**지면기압**	**적설량**	**지면온도**
결측치 수	4245	4275	8489	8518	6817

5.3 결과

결측치에 대한 평가지표로는 RMSE를 사용하였다. 아래는 시뮬레이션 결과표이다.

Ⅵ. 결론 및 토의

본 연구에서는 미세먼지자료에서 발생할 수 있는 결측치 발생에 대해 총 3가지 시뮬레이션을 진행하여 각각의 방법에 맞는 결측치 대체 방법들에 대한 비교를 통해 더 성능이 좋은 방법을 파악하였다.

시뮬레이션 결과를 통해 확인해보면 시뮬레이션 ①과 ②에서는 미세먼지 농도,강수량과 같이 다른 변수들을 통한 설명이 어려운 패턴을 지닌 경우 비모수적인 방법인 Random Forest를 기반으로한 MissForest 보다는 단일 시계열 모델을 사

〈표 5〉 총 결측치 비율을 전체 자료의 3%로 하여 각 변수에 대해 결측치를 랜덤하게 발생시킨 시뮬레이션 자료에 대한 RMSE

변수 / Model	미세먼지	평균기온	강수량	풍속	습도
MICE	29.3986	1.4972	0.6592	1.3045	4.3229
MissForest	10.5207	0.7474	0.5775	0.5282	2.6180
ImputeTS	6.5414	0.5748	0.6052	0.6446	3.1557
Amelia	28.5386	1.0124	0.7435	1.3126	5.5315
변수 / Model	**기압**	**이슬점**	**지면기압**	**적설량**	**지면온도**
MICE	0.9453	1.4510	5.0320	0.3753	4.5748
MissForest	0.4836	0.4683	0.8280	0.0324	3.7323
ImputeTS	0.5285	0.7205	0.2612	0.0155	1.5712
Amelia	2.6949	1.3681	0.6693	0.4035	4.9307

〈표 6〉 총 결측치 비율을 전체 자료의 20%로 하여 각 변수에 대해 결측치를 랜덤하게 발생시킨 시뮬레이션 자료에 대한 RMSE

변수 / Model	미세먼지	평균기온	강수량	풍속	습도
MICE	36.0023	2.3738	0.8074	1.2480	7.8090
MissForest	11.8472	0.7716	0.4869	0.5690	3.3018
ImputeTS	7.0549	0.5844	0.7140	0.6480	3.5238
Amelia	36.6005	1.5394	0.8487	1.2542	7.2888
변수 / Model	**기압**	**이슬점**	**지면기압**	**적설량**	**지면온도**
MICE	1.4336	2.3806	5.2370	0.3451	4.9467
MissForest	0.5201	0.5991	1.0701	0.0567	2.2612
ImputeTS	0.5990	0.8147	0.3032	0.0422	1.6746
Amelia	2.8998	2.0826	2.2377	0.3631	4.9411

〈표 7〉 1개 변수(예. -미세먼지)는 2017년 1월 1일부터 4월 30일까지의 결측치를 발생 시킨 뒤, 나머지 변수는 시뮬레이션 ②를 기반으로 하여 생성한 시뮬레이션 자료에 대한 RMSE

Set / Model	②-미세먼지	②-평균기온	②-강수량	②-풍속
MICE	29.4994	2.9654	0.4062	1.2444
MissForest	12.4685	0.5778	0.3227	1.1431
ImputeTS	23.7583	96.2022	1.3043	1.6317
Amelia	28.1133	2.6015	0.4979	1.2368
DTWBI	31.7217	6.3852	1.3634	1.7024

용하는 ImputeTS가 더 나은 성능을 보이는 것을 확인할 수 있었다. 그러나 시뮬레이션 ③을 통해 비교적 긴 시간에 해당하는 특정 기간이 빠져있는 경우에는 ImputeTS 보다는 MissForest가 확연히 나은 결과를 줄 수 있음을 파악할 수 있었다.

하지만, MissForest의 경우 초매개변수를 변경하며 최적의 모형을 찾는 결측치 대체과정에서 걸리는 시간이 다른 방법들에 비해 확연히 많이 걸리는 단점이 있을 수 있다.

위의 결과를 통해 확보한 자료에서의 결측치 발생 특징을 파악하여 알맞은 결측치 대체 방법을 활용할 필요성이 있다.

참 고 문 헌

[1] Eekhout, I., de Boer, R.M., Twisk, J.W.R., de Vet, H.C.W., Heymans, M.W. Missing data: a systematic review of how they are reported and handled. (2012).

[2] Rubin, D.B. Inference and missing data. (1976).

[3] Johannes Bauer, Orazio Angelini and Alexander Denev. Imputation of multivariate time series data performance benchmarks for multiple imputationand spectral techniques. (2013).

[4] Little, R.J.A. and Rubin, D.B. . Statistical Analysis with Missing Data. (1989).

[5] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. (2011).

[6] James Honaker, Garay King, and Matthew Blackwell. AMELIA II : A Program for Missing Data.

[7] Daniel J Stekhoven and Peter Buhlmann. MissForest non-parametric missing value imputation for mixed-type data. (2012).

[8] Seffen Moritz. imputeTS R Cran. (2019).

[9] Camille Dezecache, T.T.Hong Phan and Emilie Poisson-caillault. DTWBI R cran. (2018).

[10] Roderic J.A. Little. A test of Missing Completely at Random for Multivariate Data with Missing Values. (2019).

[11] Geert Molenberghs and Michael G.Kenward. Missing Data in Clinical studies. (2007).

저 자 소 개

김 연 진(YeonJin Kim)

·2011년~2018년 : 인하대학교 통계학과 학사

·2018년~현재 : 인하대 통계학과 석사

·관심분야: 빅데이터, 데이터마이닝

박 헌 진(Park HeonJin)
·1990년 Iowa State University 통계학과 (박사)
·1990년~1994년 SAS Institute Inc. Senior Research Statistian
·1994년~현재 : 인하대학교 통계학과 교수
·관심분야: 빅데이터, 데이터마이닝, 전산통계