Journal of The Korean Data Analysis Society (October 2025), 27(5), 1555-1566.
https://doi.org/10.37727/jkdas.2025.27.5.1555

서베이 응답 결측값 대체 방법에 관한 연구*

김유정[1], 권민수[2], 강창완[3], 최용석[4]

요 약

서베이 응답에서 발생하는 결측값(missing values)은 시간과 비용 측면에서 조사의 효율성을 저하한다. 이러한 결측값 문제를 적절히 처리하기 위해 여러 통계적 대체 기법을 적용할 것이다. 본 연구에서는 최근 연구 경향을 반영한 분석기법을 적용하여 여러 가지 방법으로 결측값 대체의 효과성을 평가하고자 한다. 특히, 결측값을 처리하기 위한 많은 연구에서 활용되고 있는 MICE, MissForest, Random Forest와 더불어, 최근 머신러닝 연구에서 활발히 다뤄지고 있는 XGBoost까지 총 4가지의 분석기법을 적용하려 한다. 이와 함께 향후 서베이에서 실질적으로 적용할 수 있도록 우수한 성능을 보이는 결측값 대체 방법을 파악하고자 하며, 결측이 발생하였을 때 통계적으로 대체할 수 있는 최적의 결측률을 도출하고자 한다. 성능 평가는 결측값과 실제값 간의 분포 시각화를 통한 비교, 정확도(accuracy) 및 NRMSE(normalized root mean squared error) 측정, 마지막으로 통계적 유의성을 검증하기 위한 대응 t-검정까지 총 세 단계로 수행한다. 세 가지의 실증 비교분석 결과, 우수한 성능을 보이는 분석기법에 차이를 나타내는 한 가지의 분석기법을 특정할 수 없었으며, 이는 데이터의 특성에 따라 차이가 있음을 시사한다.

주요 용어 : 결측값, MICE, MissForest, 랜덤 포레스트, XGBoost.

1. 서론

리서치에서 데이터의 품질은 분석의 신뢰성과 결정력에 중대한 영향을 미치고 있다. 통상적으로 리서치 회사와 실제 공공기관에서 적용하는 많은 통계적 분석은 완전한 데이터의 수집을 요구하고 있으며, 더불어 고품질의 데이터 수집은 높은 수준의 통찰력으로 직결된다. 그러나 현실의 자료수집 과정에서 필연적으로 발생하는 결측값(missing values)은 데이터의 보완이나 재조사법을 요구하고 있어, 시간과 비용 측면에서 전반적인 효율성 저하를 초래한다. 통계 패키지의 원활한 사용을 위해 결측이 포함된 데이터를 제거하여 완전한 데이터를 기반으로 분석하는 방법이 활용되곤 하지만, 이는 정보량의 손실로 분석 결과의 편의나 통계적 검정력의 문제를 일으키게 된다(Rubin, 1987). Ko, Tak(2016)는 실제 설문자료를 분석하여 결측값을 완전히 제거하고 분석하는 경우, 평균

*이 논문은 제1저자 김유정의 석사학위논문의 축약본입니다.
[1]46241 부산광역시 금정구 부산대학로 63 부산대학교 통계학과, 석사졸업. E-mail: yj_kim13@naver.com
[2]48059 부산광역시 해운대구 센텀북대로 60 ㈜에스티이노베이션, 대표이사. E-mail: ceo@stinnovation.co.kr
[3]47340 부산광역시 부산진구 가야동 산 24 동의대학교 산업경영빅데이터공학과, 교수.
E-mail: cwkang@deu.ac.kr
[4](교신저자) 46241 부산광역시 금정구 부산대학로 63 부산대학교 통계학과, 교수. E-mail: yschoi@pusan.ac.kr
[접수 2025년 8월 28일; 게재 확정 2025년 9월 11일]

및 분산과 같은 기술통계뿐만 아니라 상관관계나 회귀계수 같은 변수 간 관계에 대한 분석에도 영향을 미친다는 것을 확인했다. 또한, 이렇게 감소된 표본 수는 훈련 데이터의 감소로 직결되고 이는 머신러닝 알고리즘의 적합 과정에 영향을 줘 낮은 예측력의 원인이 될 가능성이 크다(Lim, 2022).

이러한 문제를 보다 효율적이고 정확한 방법으로 처리하기 위해서는 결측값 대체가 필수적이지만(Kim et al., 2017; Song, 2017; Song, 2021) 결측값을 처리하는 최적의 해법은 정해진 바 없으며(Berglund, Heeringa, 2014), 이를 어떻게 처리할지 결정하기는 쉽지 않다(Burgess et al., 2013). 결측값을 대체하기 위하여 많은 연구가 선행되고 있지만 실질적인 활용 방법에 관한 연구는 부족한 실정이고, 전통적인 방법으로 결측값을 대체하는 데 일부 한계점들이 발견되었다. 이에 본 연구에서는 기존의 결측값 대체 방법의 문제점을 보완하기 위해 최근 연구 경향을 반영한 분석기법을 적용하여 여러 가지 방법으로 결측값 대체의 효과성을 평가하고자 한다. 이를 통해 향후 실질적으로 분석에 적용할 수 있도록 우수한 성능을 보이는 기법에 초점을 맞추고자 한다.

본 연구에서는 실제 설문자료의 데이터를 이용하여 결측값을 발생시키고 이를 4가지의 분석 방법을 통해 결측값을 대체(imputation)할 것이다. 이때 사용되는 결측의 일반적인 구조와 대체 방법을 2절에서 설명한 뒤, 3절에서 본 연구에 적용되는 4가지 분석기법을 소개한다. 결측 대체 방법은 MICE, MissForest, Random Forest, XGBoost로 구성되어 있으며, 대체된 분포와 원자료 간의 분포 시각화를 통하여 대체 성능의 대략적인 개요를 파악한 뒤, 정확도(accuracy)와 NRMSE(normalized root mean squared error)를 측정하여 구체적인 성능지표를 평가한다. 최종적으로 통계적 검정(대응표본 t-검정)을 통해 우수한 대체 방법을 파악할 것이다. 이를 통해 향후 리서치에서 적용할 수 있는 효과적인 결측값 대체 방법을 도출하고자 하며, 이는 결측값 대체의 효율성과 정확성을 높이는 방향으로의 실질적인 기여를 기대한다.

2. 결측 구조와 대체 방법 소개

결측값은 자료에서 누락된 부분으로 검정력의 저하, 자료의 축소, 통계적 모형의 가정 위배 등 자료 분석 과정에서 여러 가지 문제를 초래한다(Little, Rubin. 2019, Kwon et al., 2024). 결측값을 포함하는 자료에 대한 분석은 사회과학 분야를 비롯하여 많은 분야에서 다뤄지고 있는데, 자료수집 과정에서 실험자의 실수, 조사 도구, 저장장치의 오류 등의 이유로도 발생하지만, 사람을 대상으로 하는 리서치의 특성상 응답자의 실수, 의도적인 응답 거부, 연구자의 실수나 의도에 의한 누락 등 자료수집 과정에서 결측값이 발생할 여지가 많다(Schafer, Graham, 2002).

결측값은 발생형태, 발생 메커니즘 등에 따라 여러 유형으로 나눌 수 있는데, 그중에서도 Little, Rubin(2019)은 발생 메커니즘에 따라 결측값 종류를 나눴다. 발생 메커니즘에 따른 결측값 유형 구분은 어떠한 요인이 결측값 발생에 영향을 미쳤는지에 따라 나뉘며 아래와 같이 세 가지 유형으로 나눠진다(Kim, 2021).

2.1. 결측 구조

결측 구조에 대한 자료행렬 $Y=\{y_{ij}\}$는 결측값을 가진 $N\times p$ 행렬이고, y_{ij}가 결측되었을 때는 1, 그렇지 않을 때는 0인 $N\times p$ 지시행렬을 $R=\{r_{ij}\}$라고 하자. $\Pr\{r_{ij}=0|y_{ij}\}=\Pr\{y_{ij}observed|y_{ij}\}$

$=p_{ij}$라고 정의할 때, 자료행렬 Y는 미지의 모수 θ에 대해 $P(Y|\theta)$를 가지고, R은 미지의 모수 ξ에 의해 영향을 받는 확률 분포 $P=(R|\xi, Y)$를 가진다고 할 수 있다. 또한 $Y=(Y_{obs}, Y_{mis})$로 표현되고, Y_{obs}는 관찰된 값이며, Y_{mis}는 결측값을 뜻한다. 이러한 가정하에 반응변수들과 결측 지시 변수들의 결합확률분포(joint probability distribution)는 다음과 같이 표현된다(Hong et al., 2010).

$$P(Y, R|\theta, \xi) = P(Y|\theta)P(R|\xi, Y)$$

여기에서 Little, Rubin(2019)은 조건부 분포인 $P(R|\xi, Y)$에 기초하여 결측 구조를 3가지로 범주화하였다. 첫 번째, 완전임의결측(missing completely at random, MCAR)은 결측의 발생 여부가 주어진 데이터에 전혀 의존하지 않는 것을 의미한다. 즉

$$P(R|\xi, (Y_{obs}, Y_{mis})) = P(R|\xi)$$

으로 표현할 수 있다. 두 번째로 결측의 발생 여부가 관찰된 값 Y_{obs}에만 의존할 때, 조건적 임의결측(missing at random, MAR)이라고 정의한다. 즉, 결측의 발생 여부가 Y의 결측값에 영향을 받지 않으나 관측값에는 영향을 받아 모든 Y_{mis}에 대하여

$$P(R|\xi, (Y_{obs}, Y_{mis})) = P(R|\xi, Y_{obs})$$

가 성립함을 뜻한다. 마지막으로 결측 발생이 관찰되지 않은 값 Y_{mis}에 의존할 경우를 NMAR(not missing at random)로 정의한다. 결측이 임의로 이루어지지도 않으며, 모형의 변수들을 가지고 특정한 변수의 값을 예측할 수 없는 경우를 뜻하며 MCAR과 MAR을 무시할 수 있는 결측 구조라고 하였다(Hong et al., 2010). MAR은 MCAR보다 일반적인 경우로서, 결측값의 발생이 오로지 데이터의 관측값 Y_{obs}의 영향만을 받으며, 결측값 Y_{mis}와는 무관하다고 알려져 있다.

하지만, 결측 발생의 종류를 세 가지로 나눌 수 있으나 이는 이론적인 개념으로 결측이 발생한 실제 자료에서 어떤 원리로 결측값이 발생했는지 확인하는 것은 매우 힘들다. Choi(2019)에 따르면, MAR 가정 상황 하에 결측값은 실제로 관측할 수 없는 값이기 때문에, 실제 데이터셋이 MAR 가정을 따르는지에 대한 검증에는 한계가 있다. Zhu(2014)와 Li(2013) 또한, 통계적 검정을 통해 결측이 MAR인지 NMAR인지 구분해 내기 어려우며 연구의 방향 등 상황에 맞는 적절한 가정을 진행한 채 연구를 수행해야 한다고 한다.

2.2. 결측값 대체 방법

전통적인 대체 방법(conventional imputation method)은 결측값을 포함한 케이스를 분석에서 제외하는 완전 제거법(listwise deletion), 어떤 변수의 결측값을 그 변수의 관측된 값의 평균으로 대체하는 평균 대체(mean substitution) 등이 있으며, 그 외에도 회귀분석을 이용한 대체(regression imputation), 최빈/중앙값 대체(mode/median imputation) 등이 있다(Graham, 2012).

한편, 다중 대체법(multiple imputation, MI)은 전통적인 대체 방법이 갖는 편의 및 변동성 문제들을 해결하기 위해 제안된 방법으로 결측치를 처리할 때 가장 빈번하게 사용되는 기법이다. 다중대체법은 대체될 변수에 대한 사전 분포 가정의 존재 여부에 따라 다변량 정규 대체(multivariate normal imputation, MVNI)와 연쇄 방정식에 의한 다중대체(multiple imputation with chained equations,

MICE) 방식이 있는데 MVNI는 모든 변수들이 정규분포를 따른다는 것을 가정하고 베이지안 접근(bayesian's approach)에 따라 정규분포에서 대체값을 획득하게 된다. 이러한 정규성 가정은 이항 변수나 범주형 변수에는 적용 가능성이 떨어질 수 있다는 지적이 있으나, Schafer(1999)는 정규성이 만족하지 않아도 MVNI 적용이 가능하다고 보았다. MVMI의 대표적인 방식은 몬테 카를로(markov chain monte carlo, MCMC)를 기반으로 한 방법이 있다.

MICE는 모든 변수들이 정규분포에 따른다는 가정이 없는 방식으로, 분포를 가정하지 않기 때문에, MVNI 방식에 비해 유연하게 사용할 수 있다. 결측값의 조건적인 분포가 다른 모든 변수들에 의해 결정되며, 분포 가정이 없으므로 서열 척도, 명목형 척도 등 다양한 변수에도 적용할 수 있다. 대표적인 방법으로는 완전 조건부 대체(fully conditional specification, FCS) 방식이 있다.

본 연구에서는 FCS 방법에 기반한 MICE 알고리즘을 4가지의 대체 방법 중 한 개로 택하여 실험을 진행하였다. FCS와 MCMC 방법의 가장 큰 차이점은, MCMC의 경우 분포 가정이 있는 한편, FCS에는 분포 가정이 없다는 것으로, 특히 수치형 변수와 범주형 변수가 혼합된 데이터에 대한 대체 방법을 제시하고자 하는 본 연구의 목적에 더 적합하다고 판단하였다.

본 연구에서는 일반적으로 결측값을 대체할 때 다중 대체법에서 가장 빈번하게 사용되는 MICE와, Random Forest 모델을 기반으로 데이터의 패턴을 학습하고 결측값을 예측하는 단일대체 방법인 MissForest, 그리고 분류 및 회귀분석을 위한 머신러닝 모델로 사용되지만 결정 트리의 앙상블로 구성된 모델로 결측값 대체에서도 활용되는 Random Forest까지 3가지의 분석기법을 적용하였다. 여기에 트리 기반의 앙상블 학습 알고리즘 모델인 XGBoost를 추가로 살펴보고자 한다. MICE, MissForest, Random Forest의 경우 기존의 결측값 대체 관련 연구에서 자주 사용되고 있는 한편, XGBoost는 2016년에 GBM(gradient boosting machine)을 바탕으로 새롭게 개발되어 모델의 성능이 더 우수한 것으로 알려져 있다. 따라서 본 연구에서는 해당 분석기법을 추가하여 총 4가지의 결측값 대체 방법을 살펴볼 것이다.

3. 사례분석

본 연구에서는 다음과 같은 실험을 진행한다. 먼저 결측값이 없는 데이터셋에 대하여 결측값을 임의로 발생시킨 후 각 대체 방법 간의 성능을 비교하기 위해 결측값이 없는 관측 데이터 중 일부를 실험 데이터로 사용하여 대체한다. 결측의 발생은 Python의 라이브러리를 활용하여 랜덤하게 생성했으며, 이는 MCAR 가정에 부합한다. 결측률은 전체 데이터셋 대비 10%, 20%, 30%씩 설정하여 결측 데이터셋을 구성하였다.

이어서 대체값과 실제값 간의 분포를 한눈에 파악하고자 MICE, MissForest, Random Forest, XGBoost를 통해 생성된 대체값들의 분포를 실제값의 분포와 시각적으로 비교하고 성능지표를 통해 평가한다. 대체값과 실제값의 차이를 나타내기 위한 성능지표는 정확도와 NRMSE를 사용한다.

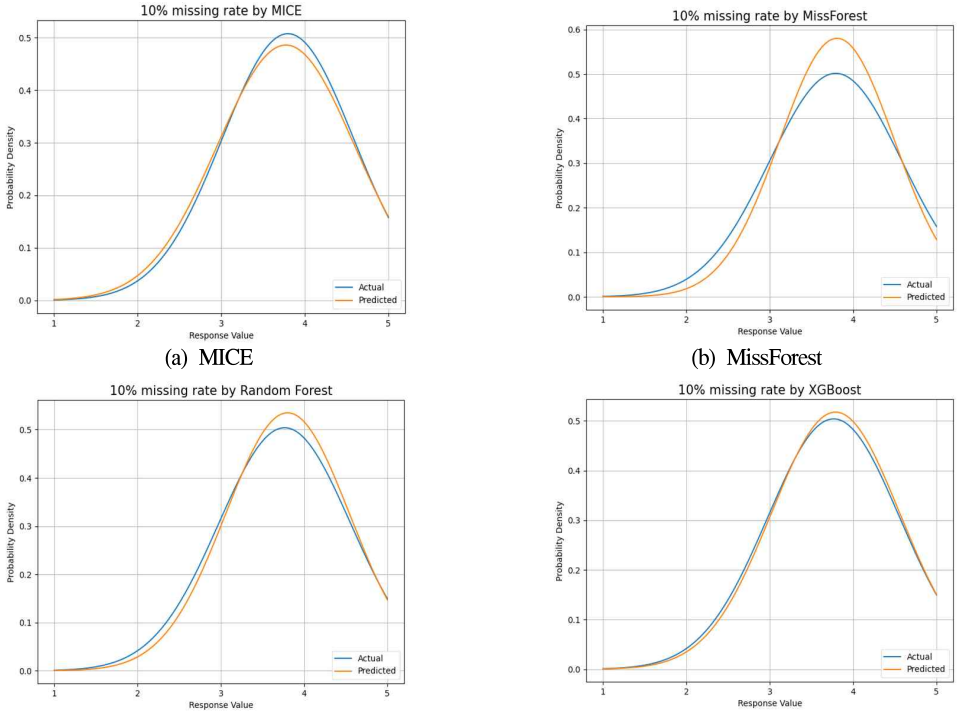
추가적으로 머신러닝 기법의 효율성을 입증하기 위해 결측값을 대체한 후 기존 데이터와 대체 후 데이터 간 통계적으로 유의한 차이를 나타내는지 파악하고자 대응표본 t-검정을 수행하여 결측값 처리의 효과를 평가한다. 실험에 사용한 데이터셋은 연속형과 범주형 변수가 혼합된 데이터로 총 4,000개의 샘플을 보유한 실제 서베이 데이터를 활용하는데 2가지의 인구 통계학적 특성(성, 연령대)과 만족도를 측정하는 28개의 문항으로 구성되어 있으며, 만족도 측정은 5점 리커트 척도

(likert scale)로 이루어진다. 본 연구에서는 전체 문항 중 전반적인 만족도를 나타내는 문항을 종속변수로 삼아 결측을 생성했으며 해당 변수에 영향을 미칠 것으로 예측되는 변수를 찾기 위해 Python의 Random Forest 함수를 활용하여 변수 중요도(importance)를 도출하였다. 이로부터 상위 5개 문항과 인구 통계학적 변인을 고려하여 결측값 대체를 진행한다.

3.1. 대체값과 실제값의 분포 비교

각 Figure 1, Figure 2와 Figure 3은 결측률 10%, 20%, 30%에 따라 MICE, MissForest, Random Forest, XGBoost를 통해 생성된 대체값의 분포를 실제값의 분포와 비교한 것으로, 각 방법의 대체값이 실제 분포를 대표할 수 있는지 시각적으로 확인한다. 그래프의 x축은 1점부터 5점까지 반응변수의 범주를 나타내고 있으며, y축은 각 범주의 발생 확률로 동일한 플롯 내에서 대체된 자료가 원자료의 분포와 얼마나 일치하는지 파악할 것이다. 이로부터 각 방법에 따라 결측 대체 성능을 한눈에 비교할 수 있다.

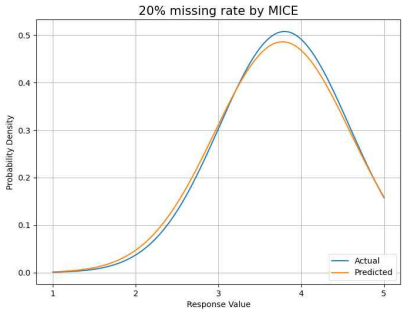
전반적인 분포를 살펴보면, 모든 대체 방법과 결측률에서 높은 예측률을 나타낸 것으로 파악되어 실제 분포와 대체값의 분포는 유사한 경향임을 볼 수 있었다. 다만, Figure 1, Figure 2, Figure 3을 통해 (d) XGBoost가 실제 분포와 시각적으로 가장 유사한 분포를 나타내는 것을 볼 수 있었다. 이어서 (a) MICE와 (c) Random Forest의 분포가 실제 분포와 유사한 경향으로 파악됐다. 한편, (b) MissForest에서는 타 분석기법 대비 분포 차이가 다소 있는 것으로 보인다. 보다, 구체적인 실제값과 대체값 간의 비교는 다음 절에서 성능지표인 정확도와 NRMSE를 통해서 살펴볼 것이다.



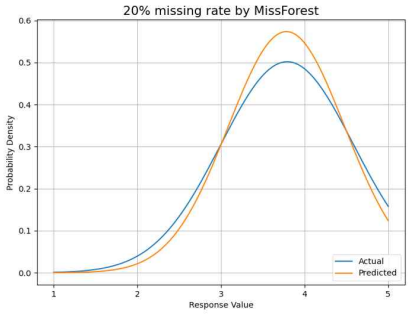
(a) MICE (b) MissForest

(C) Random Forest (d) XGBoost

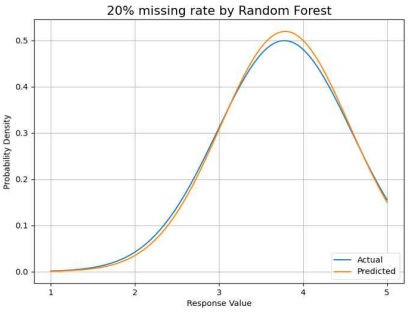
Figure 1. Comparison of the distributions of imputed and actual values at 10% missing rate



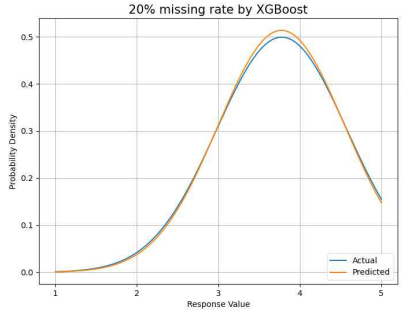
(a) MICE



(b) MissForest

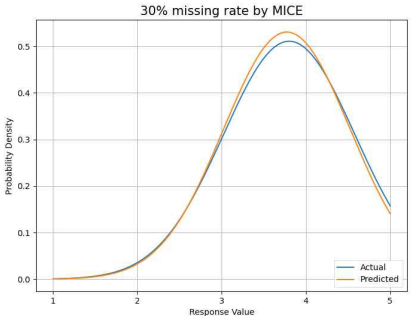


(C) Random Forest

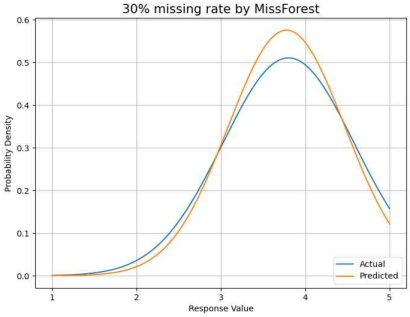


(d) XGBoost

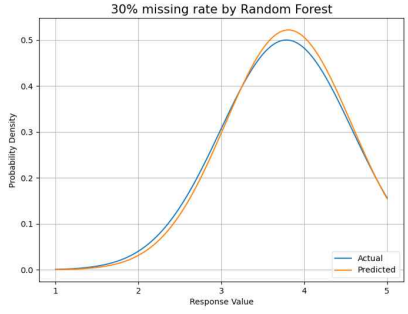
Figure 2. Comparison of the distributions of imputed and actual values at 20% missing rate



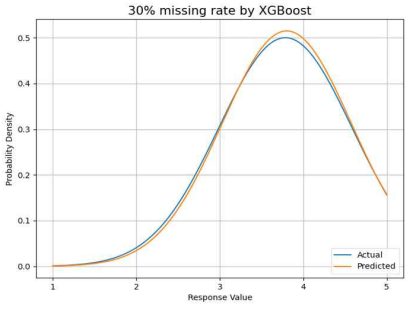
(a) MICE



(b) MissForest



(C) Random Forest



(d) XGBoost

Figure 3. Comparison of the distributions of imputed and actual values at 30% missing rate

Table 1. Confusion matrix

Classification Results	Correct Answer	
	True	False
True	*TP* (True positive)	*FP* (False positive)
False	*FN* (False negative)	*TN* (True negative)

3.2. 성능평가

본 절에서는 두 가지 성능지표인 정확도와 NRMSE를 통해 대체값과 실제값의 차이를 파악하고자 한다. Table 1과 식 (1)에 의하면 정확도는 올바르게 예측한 대체값의 비율로 높을수록 모델의 성능이 더 뛰어나다고 평가된다. 식 (2)의 NRMSE는 표준화된 평균 제곱근 오차로서 모델의 대체값과 실제값 간의 오차를 나타내는 품질 지표 중 하나로 낮을수록 모델의 성능이 더 뛰어나다고 평가한다. NRMSE는 대체된 값과 실제값 간의 오차를 데이터 스케일에 상대적으로 비교하기 위한 것이며, 0에 가까울수록 더 좋은 모델을 나타내고 0.2보다 작으면 모델의 예측이 실제값과 비교적 가까워 모델의 성능이 좋다고 평가된다.

$$ACCURACY = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$NRMSE = \frac{mean((X^{true} - X^{imp})^2)}{var(X^{true})} \tag{2}$$

1) 교차검증(cross validation)을 통한 정확도(accuracy)

대체 방법별 테스트 데이터의 정확도를 확인하여 해당 모델이 예측한 범주와 실제 범주가 얼마나 일치하는지 측정하고자 한다. 훈련 데이터의 정확도는 각 대체 방법별 모델이 훈련 데이터에 대해 얼마나 잘 맞추는지를 측정하는 것으로 일반적인 모델 평가에서는 훈련 데이터가 아닌 테스트 데이터의 정확도를 통해 모델의 성능을 평가한다. Table 2에서는 테스트 데이터의 성능을 평가함과 동시에 모델의 과적합 문제를 파악하기 위해 훈련 데이터의 정확도를 참고용으로 함께 확인한다.

Table 2는 교차검증을 통해 정확도를 산출한 결과이며, MissForest의 경우 R 패키지를 사용하였다. 단, 해당 기법의 경우 훈련 데이터의 정확도를 평가하거나 최적화하는 메커니즘을 보유하고 있지 않아 훈련 데이터의 정확도를 확인할 수 없었다.

이를 제외하고 테스트 데이터의 정확도를 계산했을 때, XGBoost > Random Forest > MissForest > MICE 순으로 모델의 정확도가 높아 성능이 높게 평가되었다. 구체적으로 살펴보면 결측률이 10%일 때 XGBoost를 통한 데이터 대체의 정확도가 75.9%로 가장 높게 나타났으며, 이는 모델이 예측한 값 중에서 실제로 맞춘 비율을 나타낸다. 이어서 Random Forest 74.2%, MissForest 66.4% 순으로 높은 정확도를 도출했고, MICE에서는 39.6%의 정확도가 계산되어 해당 대체 방법으로 결측치를 대체하는 것에는 주의가 필요함을 볼 수 있다. 이어서 결측률이 20%일 때도 마찬가지로 XGBoost의 정확도가 74.6%로 가장 높게 나타났으며, Random Forest 73.5%, MissForest 65.8% 순으로 뒤를 이었고, 결측률이 30%일 때도 유사한 경향성을 나타냈다. 결측률이 10%씩 상승하면서 테스트 데이터의 정확도가 하락하지만, 정확도 감소 폭이 낮게 도출되어 대체로 대체 방법의 성능이 우수하다고 짐작할 수 있었다.

Table 2. Accuracy comparison by performance of imputation methods

Imputation		Missing Rate		
Method	Split	10%	20%	30%
MICE	train	41.2%	41.5%	41.6%
	test	39.6%	39.1%	38.8%
MissForest	train	-	-	-
	test	66.4%	65.8%	63.2%
Random Forest	train	84.3%	84.5%	85.5%
	test	74.2%	73.5%	72.6%
XGBOOST	train	82.1%	82.3%	83.6%
	test	75.9%	74.6%	73.2%

Table 3. NRMSE comparison by performance of imputation methods

Imputation Method	Missing Rate		
	10%	20%	30%
MICE	0.1698	0.1660	0.1718
MissForest	0.1217	0.1245	0.1359
Random Forest	0.1491	0.1475	0.1508
XGBOOST	0.1496	0.1439	0.1491

2) NRMSE

Table 3은 식 (5)에서 소개한 바와 같이 평균 제곱 오차(mean square error, MSE)의 제곱근을 표준화한 값으로 모든 대체 방법에서 0.2 이하로 낮은 값이 계산됐다. 이는 종속변수가 척도형으로 되어 있어 연속형으로 간주했을 때 NRMSE의 성능이 높은 것으로 예측된다. 다만, 정확도와 달리 MissForest > XGBoost > Random Forest > MICE 순으로 모델의 성능이 높게 평가되었다.

구체적으로 살펴보면 결측률이 10%일 때 MissForest의 표준화된 평균 제곱근 오차가 0.1217로 가장 낮게 나타났다. 이는 모델의 예측이 상대적으로 정확하며, 실제값과 평균적으로 약 12.2%의 오차를 갖고 있다는 것을 의미한다. 이어서 Random Forest 0.1491, XGBoost 0.1496, MICE 0.1698 순으로 오차가 낮게 나타났다. 결측률이 20%일 때도 MissForest가 0.1245로 가장 낮았으나 XGBoost 0.1439, Random Forest 0.1475, MICE 0.1660 순으로 나타나 차 순위에서는 다소 차이가 있었고, 결측률이 30%도 20%일 때와 유사한 경향임을 볼 수 있었다.

본 결과에서는 대체 성능이 가장 우수한 방법도 차이가 존재했지만, 정확도와 달리 특정 방법(MICE)의 성능이 타 대체 방법 대비 큰 차이로 성능이 낮게 평가되지 않았다는 점도 함께 주목할 수 있어 척도에 따라 모델의 성능을 평가하는 데 유의해야 함을 알 수 있다.

각 평가지표는 결측 대체를 평가하기 위한 서로 다른 지표로서 정확도는 5점 척도로 구성된 범주형 데이터가 모델이 예측한 범주와 실제 범주가 얼마나 일치하는지 측정하며, NRMSE는 동일한 데이터를 연속형으로 간주해 연속 변수의 예측오차를 측정하여 실제값과 대체값 간의 오차를 스케일에 상대적으로 비교할 수 있게 한다. 종속변수의 범주화된 예측과 동시에 연속적인 평균 점수를 예측하기 위하여 두 지표를 함께 고려하여 모델을 평가하였다. 각 결과로부터 모델의 성능이 높게 평가된 기법이 다르게 도출된 것으로 보아, 서베이 데이터를 활용하여 대체 성능을 평가하기 위해 한 가지 평가지표가 아닌 복합적으로 평가하여 대체 모델을 고려해야 할 것으로 판단된다.

Table 4. Comparison of paired sample test results

Response Variable	imputation Method	Missing Rate					
		10%		20%		30%	
		$mean$	t ($p-$value)	$mean$	t ($p-$value)	$mean$	t ($p-$value)
satisfaction (mean=3.790)	MICE	3.795	-1.475 (.140)	3.785	0.992 (.321)	3.779	1.946 (.052)
	MissForest	**3.791**	-0.577 (.564)	3.788	0.475 (.635)	3.783	1.662 (.097)
	Random Forest	3.793	-0.181 (.856)	3.792	-0.098 (.922)	3.797	-0.377 (.706)
	XGBOOST	3.792	-0.112 (.911)	3.789	0.056 (.956)	3.794	-0.251 (.802)

3) 통계적 검정

끝으로 본 절에서는 결측값을 대체한 후, 결측값을 대체하여 처리하는 방법이 기존 데이터와 통계적으로 유의미한 차이를 가지는지 확인하기 위해 대체 데이터와 원자료 간의 대응표본 t-검정을 수행하여 결측값 처리의 효과를 평가할 것이다. 해당 검정을 통해 결측률과 대체 데이터셋의 개수에 따른 분석 결과를 비교하고자 한다.

대응표본 t-검정을 수행한 결과, 모든 대체 방법에서 원자료와 각 결측률 10%, 20%, 30%에 따라 대체된 데이터 간 평균에는 유의한 차이가 없는 것으로 나타나 유의수준 5% 하에 모든 가설을 기각할 수 없었다. 이로부터 결측 데이터 대체를 수행하기 위해서는 30%까지 결측이 발생할 경우, 해당 대체 방법으로 데이터를 대체할 수 있을 것으로 판단된다.

다만, 결측률이 높아질수록 평균 점수에 변동이 발생하므로 앞서 확인한 원자료와 대체 데이터 간 분포, 모델 평가지표 등을 함께 고려하여 결측 자료의 대체 방법을 고려해야 할 것으로 파악된다.

4. 결론

본 논문에서는 결측값으로 인한 데이터의 손실을 최소화하고자 하는 통계적 결측값 대체 기법을 다룬다. 해당 연구에서 사용된 데이터셋은 연속형과 범주형 변수가 혼합된 데이터로 독립변수와 종속변수는 5점 리커트 척도로 구성되어 있다. 실제 설문자료로 구성된 완전한 데이터에서 각 결측률(10%, 20%, 30%)에 따라 랜덤하게 결측을 발생시킨 후, 4가지의 대체 방법을 통해 해당 방법 간의 성능을 비교한다. 사용된 대체 방법은 기존의 결측 대체 관련 연구에서 빈번하게 사용되는 MICE, MissForest, Random Forest와 더불어 최근 머신러닝 연구에서 활발히 다루어지고 있는 XGBoost를 사용한다. 분석 결과는 다음과 같이 세 가지로 나누어 볼 수 있다.

첫째로, 대체값과 실제값 간의 분포를 비교하여 각 대체 방법 간의 성능을 시각적으로 평가한 결과, XGBoost에서 MICE와 Random Forest 대비 분포 유사도가 높은 것으로 짐작되었으며, MissForest의 분포 유사도가 가장 낮은 것으로 파악됐다. 둘째로, 종속변수의 범주화된 예측과 동시에 연속적인 평균 점수를 예측하기 위하여 정확도와 NRMSE를 통해 대체값과 실제값의 성능을 평가한 결과, XGBoost에서 높은 정확도가 나타난 한편, MissForest에서 NRMSE 성능이 우수한 것으

로 평가되었다. 셋째로, 결측값을 대체한 후 기존 데이터와 통계적으로 유의한 차이를 나타내는지 파악하고자 대응표본 t-검정을 수행한 결과, 모든 대체 방법에서 원자료와 유의한 평균 차이는 없는 것으로 나타났지만, 그중에서도 MissForest의 평균값이 실제값과 가장 유사한 것으로 나타났다.

종합적으로 살펴보면 대체 방법별로 우수한 성능을 보이는 지표에 차이를 보였는데, 이는 데이터의 특성에 따라 차이가 있음을 시사한다. 즉, 종속변수가 5점 척도로 구성되어 있으므로 범주형과 연속형으로 둘 다 간주할 수 있다. 여기서 자료를 범주형으로 간주했을 때는 XGBoost의 성능이 우수한 것으로 평가되나, 연속형으로 간주할 경우 MissForest의 성능이 우수한 것으로 요약된다. 위로부터 서베이 데이터에서 결측값 대체를 진행할 때, 특정 방법으로 단정하여 결측값을 대체하는 것 보다, 데이터의 형태, 유형 등을 데이터의 전반적인 특성을 고려하여 적절한 대체 방법을 선택해야 할 것으로 판단된다. 또한, 세 번째 결과로부터 결측이 30%까지 발생하였을 경우 해당 대체 방법의 결측 처리가 효과적이라고 기대할 수 있다. 다만, 결측률이 높아질수록 평균 점수에 변동이 발생하므로 앞서 고려한 세 가지의 방법을 모두 고려하는 것이 올바를 것으로 짐작된다.

이어서, 실무적으로 본 연구를 활용하기 위해 향후 더 방대한 데이터 유형에 대한 확장, 다른 결측 패턴에 대한 실험, 성능이 뛰어나다고 평가된 대체 방법의 파라미터 튜닝에 대한 탐구 등을 추가적인 연구 방향으로 제시할 수 있다. 특히, 본 연구에서 확인한 결측 구조는 라이브러리를 활용하여 랜덤하게 결측을 생성하였으므로 MCAR에 해당하는 것으로 파악되었으나, 이는 실제 설문조사에서 발생하는 결측 구조의 특성에 부합하지 않을 수 있으므로 일부 괴리가 생길 수 있다. 실제 설문조사에서 발생하는 결측의 종류는 다양하고 복잡하며, 이를 실무적으로 더 활용하기 위해서는 보다, 복잡한 MAR이나 NMAR 가정에 근거하여 후속 연구를 진행할 필요성이 있다. 이는 향후 실제 결측 메커니즘을 더 정확히 모델링하고 적절한 대체 전략을 개발할 수 있을 것으로 기대하며, 본 연구는 이러한 연구를 뒷받침하기 위한 발걸음이라 시사한다.

References

Ko, K., Tak, H. (2016). The Treatment of Missing Values using the Integrated Multiple Imputation and Callback Method, *Korean Journal of Public Administration,* 54(4), 291-319.(in Korean).

Kim, S. (2021). *Comparison of K-Nearest Neighbor Imputation and Full Information Maximum Likelihood for Handling Missing Data in Latent Profile Analysis*, master's thesis, Pusan University. (in Korean)

Kwon, M., Sung, J., Choi, S. B., Kang, C. (2024). Predicting Election Results through Non-Response Substitution in Polls, *Journal of The Korean Data Analysis Society*, 26(4), 1005-1013. (in Korean).

Lim, S. H. (2022). *A Study on Credit Data Analysis Method Using Missing Data Imputation*, master's thesis, Seoul University. (in Korean)

Choi, H. C. (2019). *Imputation Method based on a Voting Manner for Missing Data*, master's thesis, Hanyang University. (in Korean)

Hong, H., U., Choi, Y., Shin, S. M., Kang, C. (2010). A study on shape Variability in canonical correlation biplot with missing values, *The Korean Journal of applied Statistics*, 23(5), 955-966. (in Korean)

Berglund, P., Heeringa, S. G. (2014). *Multiple Imputation of Missing Data Using SAS*, SAS Institute.

Breiman, L. (2001). Random forests, *Machine Learning*, 45(1), 5–32.

Burgess, S., White, I. R., Resche Rigon, M., Wood, A. M. (2013). Combining multiple imputation and meta analysis with individual participant data, *Statistics in medicine*, 32(26), 4499-4514.

Chen, T. Guestrin, C. (2016). Xgboost: A scalable tree boosting system, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (KDD '16), Association for Computing Machinery, 785-794. https://doi.org/10.1145/2939672.2939785

Graham, J. W. (2012). Missing data analysis: making it work in the real world, *Annual review of psychology*, 60, 549-576.

Kim, J. H., Kim, J. Y., Kwon, E. H., Kim, H., Kang, H. C. (2017). The effects of non-response adjustment in election forecasting by public opinion poll, *Journal of The Korean Data Analysis Society*, 19(1), 107-117. (in Korean).

Li, C. (2013). Little's test of missing completely at random, *The Stata Journal*, 13, 795 – 809.

Little, R. J., Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, John Wiley & Sons.

Misztal, M. (2013). Some remarks on the data imputation using "MissForest" method, *Acta Universitatis Lodziensis. Folia Oeconomica*, 285, 169-179.

Natekin, A., Knoll A. (2013). *Gradient boosting machines*, a tutorial., Front. Neurorobot.

Pedersen, A. B., Mikkelsen E. M., Cronin-Fenton D., Kristensen N.R., Pham T.M., Pedersen L., and Petersen I. (2017). Missing data and multiple imputation in clinical epidemiological research, *Clinical epidemiology,* 9, 157 – 16.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.

Schafer, J. L., Graham, J. W. (2002). Missing data: our view of the state of the art, *Psychological methods*, 7(2), 147.

Schafer, J. L. (1999). Multiple Imputation:A Primer, *Statistical Methods in Medical Research*, 8(1), 3-15.

Song, J. (2017). The Effect of Nonresponse Rates on K-Means Cluster Analysis with Missing Data, *Journal of The Korean Data Analysis Society*, 19(3), 1273-1282. (in Korean). DOI: https://doi.org/10.37727/jkdas.2017.19.3.1273

Song, J. (2021). Imputation of missing values with incomplete measurements due to measurement errors, *Journal of The Korean Data Analysis Society*, 23(5), 2065-2075. (in Korean). DOI: https://doi.org/10.37727/jkdas.2023.25.4.1449

Stekhoven, D. J., Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data, *Bioinformatics*, 28(1), 112-118.

Van Buuren, S., Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software*, 45, 1-67.

Zhu, X. (2014). Comparison of four methods for handing missing data in longitudinal data analysis through a simulation study, *Open Journal of Statistics*, 4, 933 – 934.

A Study on Methods for Imputation of Missing Values in Survey Response*

You Jeong Kim[1], *MinSoo Kwon*[2], *Changwan Kang*[3], *Yong-Seok Choi*[4]

Abstract

Several statistical imputation methods can be applied to deal with the problem of missing values in survey responses, which reduce the efficiency of the survey research in terms of time and cost. In particular, this study intends to apply a total of four analysis techniques, including MICE, MissForest, and Random Forest, which are used in many studies to deal with missing values, and XGBoost, which has been actively dealt with in recent machine learning studies. In addition, we want to identify the imputation method for missing values with excellent performance so that they can be practically applied in the survey research in the future, and measure the optimal missing rate that can be statistically imputed when missing occurs. Performance evaluation is performed in three stages: comparison through visualization of distribution between missing and real values, measurement of accuracy and NRMSE, and finally, paired t-test to verify statistical significance. As a result of three empirical analyses, there was a difference in the analysis methods showing excellent performance, so one analysis method could not be specified, suggesting that there is a difference depending on the characteristics of the data

Keywords : missing value, MICE, MissForest, Random Forest, XGBoost.

---

*This Work was based on the master's thesis of first author(You Jeong Kim).

[1]Graduate student, Dept. of Statistics, Pusan National University, 2 Busandaehak-ro, 63beon-gil, Geumjeong-Gu, Busan, 46241, Korea. E-mail: yj_kim13@naver.com

[2]48059 Chief Executive Officer, Stinnovation, Busan, 48059, Korea. E-mail: ceo@stinnovation.co.kr

[3]Professor, Dept. of Industrial Management & Big Data Engineering, Dongeui University, Busan, 614-714, Korea. E-mail: cwkang@deu.ac.kr

[4](Corresponding author) Professor, Dept. of Statistics, Pusan National University, 2 Busandaehak-ro, 63beon-gil, Geumjeong-Gu, Busan, 46241, Korea. E-mail: yschoi@pusan.ac.kr

[Received 28 August 2025; Accepted 11 September 2025]