한국해양환경·에너지학회지
J. Korean Soc. Mar. Environ. Energy
Vol. 29, No. 1, 64-78, February 2026

https://doi.org/10.7846/JKOSMEE.2026.29.1.64
ISSN 2288-0089(Print) / ISSN 2288-081X(Online)

Original Article

해양환경측정망 자료의 결측대치기법 적용 및 성능 평가

김태균[1] · 조홍연[2] · 이기섭[3,†]
[1]과학기술연합대학원대학교 한국해양과학기술원 스쿨 UST 학생연구원
[2]한국해양과학기술원 해양빅데이터-AI 센터 책임연구원
[3]한국해양과학기술원 해양빅데이터-AI 센터 선임기술원

Application and Performance Evaluation on the Missing Imputation Methods of the Marine Environment Monitoring Network Data

Tae-Kyun Kim[1], Hong-Yeon Cho[2], and Gi-Seop Lee[3,†]
[1]UST Student, University of Science and Technology(Korea Institute of Ocean Science and Technology school), Daejeon 34113, Korea
[2]Principal Research Scientist, Marine Bigdata-AI Center, Korea Institute of Ocean Science and Technology, Busan 49111, Korea
[3]Senior Research Specialist, Marine Bigdata-AI Center, Korea Institute of Ocean Science and Technology, Busan 49111, Korea

요 약

본 연구는 해양환경측정망의 자료 활용성을 향상시키기 위해 다변량 결측 대치 기법의 성능을 정량적으로 평가하였다. 우선, 해양환경 주요 변수(수온, 염분, 영양염류 등)에 대해 대표적 다변량 대치 알고리즘인 Amelia와 MICE를 적용하였다. 알고리즘의 대치 성능을 검증하기 위해 실제 관측자료에 대해 결측률을 10%, 20%, 50%로 인위적으로 설정한 후, 관측치 대비 대치치의 정확도를 MAPE(mean absolute percentage error, %)로 평가하였다. 평가 결과, 결측률이 20% 이하일 때 두 방법 모두 대부분의 변수에서 MAPE가 약 5.0 이하로 유지되어 안정적인 성능을 보였으며, 예를 들어 WT_s의 경우 10% MR에서 MAPE가 Amelia 2.0, MICE 1.7 수준으로 나타났다. 반면, 50% 결측률에서는 변수 특성에 따라 성능 저하가 뚜렷하게 나타났으며, COD_s는 Amelia 38, MICE 35.6으로 최대 5-6배 이상의 오차 증가가 확인되었다. 전체 대치 자료와 원자료의 밀도 함수 비교에서는, MICE가 대부분의 변수에서 원자료의 분포를 가장 잘 복원한 반면, Amelia는 일부 영양염 변수에서 분산을 과대 추정하는 경향이 확인되었다. 이러한 결과는 Amelia와 MICE가 해양환경자료 결측 처리에서 상보적인 장점과 한계를 동시에 가진다는 점을 보여준다. 특히 간헐적·저비율 결측(≤20%) 상황에서는 다변량 대치 기법이 신뢰성 있게 적용 가능함을 실증하였다. 향후 연구에서는 계절성·자기상관 구조 등 시계열 특성을 반영한 고도화된 결측 대치 모델이 필요함을 제안한다.

Abstract – This study quantitatively evaluates the performance of multivariate imputation techniques to enhance the usability of data from the marine environmental monitoring network. Two representative multivariate imputation algorithms, Amelia and MICE, were applied to major marine environmental variables, including water temperature, salinity, and nutrients. To assess imputation performance, artificial missing rates of 10%, 20%, and 50% were imposed on the original observations, and the accuracy of the imputed values was evaluated using the mean absolute percentage error (MAPE, %) between observed and imputed data. The results indicate that when the missing rate was 20% or lower, both methods exhibited stable performance for most variables, with MAPE values generally remaining below approximately 5.0%. For example, for WT_s at a 10% missing rate, MAPE values were 2.0% for Amelia and 1.7% for MICE. In contrast, at a 50% missing rate, performance degradation became pronounced depending on variable characteristics; for COD_s, MAPE increased to 38.0% for Amelia and 35.6% for MICE, corresponding to an error increase of approximately five- to sixfold. Comparisons of kernel density estimates between the imputed and original datasets further showed that MICE more effectively reproduced the original distributions for most variables, whereas Amelia tended to overestimate variance for certain nutrient-related

†Corresponding author: freelgs7@kiost.ac.kr

variables. These findings demonstrate that Amelia and MICE possess complementary strengths and limitations in handling missing data in marine environmental observations. In particular, the results confirm that multivariate imputation techniques can be reliably applied under intermittent or low-level missingness (≤20%). Future research should develop more advanced imputation models that explicitly account for seasonality and autocorrelation structures inherent in marine time-series data.

Keywords: Marine environmental monitoring data(해양환경측정망), Missing data imputation(결측 대치), Multiple imputation(다중 결측 대치), Amelia(Amelia), MICE(MICE)

1. 서 론

해양환경 모니터링 자료는 장기적인 해양생태계의 변동성과 추세를 규명하는 데 핵심적인 역할을 수행한다. 그러나 자료에 결측이 존재할 경우, 시계열 분석 및 통계적 검정에서 편향된 결과를 초래할 수 있다. 특히 장기적인 해양환경 연구에서는 관측의 시간 해상도에 따라 달라지지만, 기후 연구와 같이 계절·연간·수년 규모의 변동성을 동시에 고려해야 하기 때문에 완전한(Complete) 시계열 자료 확보는 필수적이다(Afrifa-Yamoah *et al*.[2020]). 결측 자료의 존재는 회귀분석, 다변량 분석, 시계열 모형 검정 등 다양한 통계분석 기법 적용에 구조적 제약을 초래한다. 결국 이는 추세선의 기울기나 통계적 유의성이 편향되어 잘못된 정책적 판단으로 이어질 수 있으며, 실제 수질 지수(Water Quality Index)를 추정하는 과정에서도 결측이 편향된 결과를 나타낼 수 있음이 보고된 바 있다(Sierra-Porta[2024]).

다변량(Multi-Variate)의 해양환경 변수는 각 변수들 사이의 밀접한 상관성을 지니고 있다. 이에 결측을 단순히 무시하거나 삭제할 경우 분석 결과의 왜곡, 변수 간 관계성 해석의 오류, 유의성 계산의 불확실성 증가로 이어진다. 특히 타 변수에 영향을 많이 받는 용존산소 농도와 같이 변수 간 상관성이 높은 해양환경 자료에 대해 결측 대치의 필요성이 더욱 강조되고 있다(Chen and Xue[2023]; Rodriguez *et al*.[2021]; Wang *et al*.[2024]).

모델 입력 측면에서도 수치모형 및 기계학습 기반 분석에 요구되는 데이터는 결측이 없는 자료를 전제한다. 결측 대치 없이 불완전한 자료를 그대로 사용하면, 상호보완적 예측 모델의 비정상 작동, 학습 오류, 예측 불확실성의 증가 등의 문제를 유발한다(Lepot *et al*.[2017]). 또한 불연속하게 기록된 생지화학 변수(DO, pH, 유기탄소 등) 를 통합하려는 경우에는, 대치를 통해 시공간적 일관성을 확보한 후에야 종합 분석이 가능하다. 예를 들어 SPOTS(Synthesis Product for Ocean Time Series) 프로젝트는 다양한 시계열 프로그램의 자료를 통일된 구조로 통합함으로써 분석의 활용성을 극대화하였다(Lange *et al*.[2024]).

자료의 결측을 대치하는 방법은 전통적으로 평균값 대체나 선형보간 등과 같은 단변량(Single imputation) 방식이 주로 사용되어 왔다. 그러나 이러한 단일 대체 기법은 변수 간 상관성을 고려하지 못하고, 통계적 유의성의 왜곡과 분석 결과의 편향을 초래한다는 한계가 존재한다. 최근에는 이러한 문제를 극복하기 위해 다변량 결측 대치(Multiple Imputation) 기법이 활발히 사용되고 있으며, 특히 Amelia와 MICE(Multivariate Imputation by Chained Equations)가 대표적이다. Amelia는 expectation–maximization과 bootstrap을 결합하여 다변량 정규분포 가정 하에서 시계열과 단면 자료 모두에 적용 가능하다는 장점이 있으며, 데이터가 정규성에 가까울수록 안정적인 성능을 보인다(Honaker *et al*.[2011]). 반면 MICE는 각 변수를 다른 변수의 회귀식으로 반복 예측하는 chained equation 방식을 기반으로 하여, 다양한 분포 형태와 변수 타입에 유연하게 대응할 수 있다는 점에서 해양환경과 같이 이질적 변수가 혼재된 자료에 적합하다(Buuren and Groothuis-Oudshoorn[2011]).

국외에서는 결측이 있는 자료의 다변량 대치 기법이 활발히 적용되고 있다. 지하수 수질 자료를 대상으로 Amelia와 MICE를 적용하여 화학 변수 20개의(중금속, 음이온, 철, 규소, 인등 기본 수질변수) 다변량 대치의 성능을 비교한 연구가 있으며(Mahmood *et al*.[2024]), missForest와 같은 최신 랜덤포레스트 기반 기법 또한 많이 사용되고 있는 추세이다(Stekhoven and Buhlmann[2011]; Zhang *et al*.[2021]).

반면 국내에서는 주로 단변량 기반 기법에 국한되는 경우가 대부분이다. 울릉도 해상 부이 수온 관측자료를 대상으로 추세 성분과 잔차 성분을 합성하여 장기 결측 자료를 대치 하였고(Cho *et al*.[2021]), 인공지능 기반 BiRNN 모델과 통계적 MICE를 비교하여 단기 수온 대치에 대한 성능 차이를 분석한 사례도 존재한다(Sin *et al*.[2022]).

다변량 결측 대치가 단변량 기반 기법보다 통계적 신뢰도와 분석 정확성에서 우수하다는 점에도 불구하고, 국내에서는 다변량 대치 기법의 적용사례가 미흡한 실정이다.

해양환경측정망 자료는 1997년부터 현재까지 우리나라 연안 환경을 파악할 수 있는 대표적인 다변량 자료로서 해양환경공단에서 운영 및 관리되고 있다. 2025년 기준 연안에서 항만을 포함하여 425개 정점에 대한 해양환경자료를 제공하고 있지만 시간적으로는 매년 2월, 5월, 8월, 11월의 4회만 관측이 수행되고 있다. 이 자료는 관측시작 연도에 따라 정점 개수의 차이가 있다. 일부 정점은 1997년부터, 일부는 2017년 이후부터 자료가 존재하는 등 공간적으로도 동일한 정점 개수 및 관측 항목이 유지되지 않는 상황이다.

따라서 본 연구는 다수의 결측이 존재하는 해양환경측정망 자료를 결측이 없는 완전한 자료로 구축하는 것을 목표로 결측률·변수별 성능 비교를 통한 최적 기법을 제시하였다. 이를 위해 다변량 시계열 결측 대치 기법인 Amelia, MICE 모델을 활용하였으며, 각각의 기법에 대한 성능평가를 Leave-n-out 기법을 이용하여 수행하였다.

2. 재료 및 방법

2.1 해양환경측정망 자료

해양환경측정망은 1997년 관측 시작 후 2025년 기준 425개 정점을 관측하고 있는 우리나라의 유일무이한 장기 연안 정점 관측 자료이다. 시간해상도는 1년에 4회(2, 5, 8, 11월)를 관측하는 계절 관측 자료이며 공간 해상도는 5개 생태구(서해중부, 서남해역, 대한해협, 동해, 제주)를 구분하여 항만을 포함한 전국 연안이다. 항만 측정 정점은 1년에 2회(2월, 8월) 관측한다. 측정하는 변수는 수온, 염분, pH, DO 등 해양 물리환경 자료 및 수질환경 변수를 표-저층(_s, _b)으로 구분지어 관측하고 투명도를 포함하여 31개이다. 해양환경측정망의 원자료 변수 별 기초 통계량과 변수의 단위정보는 Table 1에 나타내었다. 원자료 변수 간 선형 상관관계는 다음 Fig. 1으로 나타내었다.

변수 중 변동범위가 3자리수 이상으로 변하는, 즉 천 단위의 자리수가 변할 정도로 큰 변수를 Table 1에 회색 음영으로 나타내었다. 규모 차이가 큰 항목은 원자료의 왜도 및 이분산성이 크게 나타나, 분석의 편향 및 불안정성을 초래할 수 있다. 이에 해당 변수에 로그 변환을 적용하여 분포 형태와 스케일을 조정하였다.

서론에서도 언급했듯이 해양환경측정망 자료는 정점 별, 변수 별로 관측 시점이 달라져 결측 현황을 우선적으로 파악해야 한다. 이를 위해 원 자료의 차원을 모두 맞추어주었는데 이는 A 정점이 2009년부터 관측을 시작했다면 1997-2008년 까지의 자료는 모두 NA로 채워주었다는 의미이다. 따라서 모든 정점에서 자료는 NA이더라도 1997-2024년 까지 채워진 자료의 결측 현황이다. 다만 425개의 정점, 28년의 관측 시간, 31개의 변수에 대한 결측 현황을 시각화하기엔 현실적으로 어려움이 있고 특정 시점을 기준으로 정점이나 변수가 추가되므로 시간에 대해 변수별 결측 현황을 비율로 확인하였다(Fig. 2).

해당 도표에서 확인 할수 있는 가장 큰 점은 시기에 따른 관측 항목의 변화이다. 결측현황에서 알 수 있듯이 9개 변수에서 관측 개시 초반에 높은 결측률을 나타내고 있으며 시간이 갈수록 정점

Table 1. Summary statistics of the KOEM data variables

	Unit	Min.	1st Qu.	Median	Mean	SD	3rd Qu.	Max.	NA's
WT_s	℃	-2.12	12.03	16.17	16.30	6.81	21.23	33.10	11782
WT_b		-2.16	10.47	15.00	14.72	6.36	18.23	33.36	11785
SAL_s	PSU	0.05	31.06	32.34	31.65	3.27	33.50	35.58	11781
SAL_b		0.05	31.51	32.89	32.36	2.30	33.85	90.95	11785
pH_s	-	6.12	8.03	8.14	8.12	0.18	8.22	9.83	11783
pH_b		6.06	8.00	8.10	8.09	0.18	8.19	9.98	11787
DO_s	mg/L	0.93	7.64	8.49	8.67	1.76	9.58	68.10	11781
DO_b		0.00	7.10	8.15	8.20	1.94	9.25	66.02	11785
COD_s		0.00	0.87	1.28	1.44	0.86	1.82	18.10	11781
COD_b		0.00	0.83	1.25	1.39	0.79	1.80	8.98	11787
NH4_s	μg/L	-0.10	5.59	14.10	36.99	92.44	37.46	8087.00	11782
NH4_b		-0.05	6.52	15.42	34.90	64.92	37.23	2590.00	11787
NO2_s		-0.12	1.70	4.22	8.20	12.78	9.30	328.93	11781
NO2_b		0.00	2.20	4.79	8.11	11.90	9.16	442.00	11785
NO3_s		0.00	14.00	57.39	107.17	174.42	124.00	3859.00	11781
NO3_b		0.00	22.28	67.13	97.51	125.05	127.00	3464.00	11786
DIN_s		0.18	32.79	89.29	152.41	229.02	174.01	8805.00	11781
DIN_b		0.00	48.33	101.07	140.52	163.18	174.00	3813.00	11785
TN_s		14.00	174.73	258.70	352.12	303.04	413.20	4873.00	16379
TN_b		3.86	186.03	261.12	329.56	244.57	384.91	4385.46	16382
DIP_s		-0.03	4.36	11.00	14.81	16.84	19.86	700.00	11782
DIP_b		-0.22	7.00	13.64	17.12	17.36	22.33	933.41	11786
TP_s		0.00	18.39	26.98	33.38	27.19	40.15	1182.08	16379
TP_b		0.00	20.77	29.26	36.23	34.86	42.56	2512.00	16382
SiO2_s		0.00	122.00	258.79	327.08	352.53	411.71	7908.01	18319
SiO2_b		0.00	173.91	303.02	355.51	300.96	458.67	5138.84	18322
SS_s	mg/L	0.00	4.30	7.60	12.79	19.21	13.90	623.50	11783
SS_b		0.00	5.00	8.90	15.70	26.44	16.80	1027.00	20029
CHLa_s	μg/L	0.00	0.90	1.80	3.41	5.66	3.76	208.00	14031
CHLa_b		0.00	0.82	1.61	2.85	4.12	3.36	182.44	20039
SD	m	0.00	1.60	3.00	4.15	3.50	6.00	47.00	12018

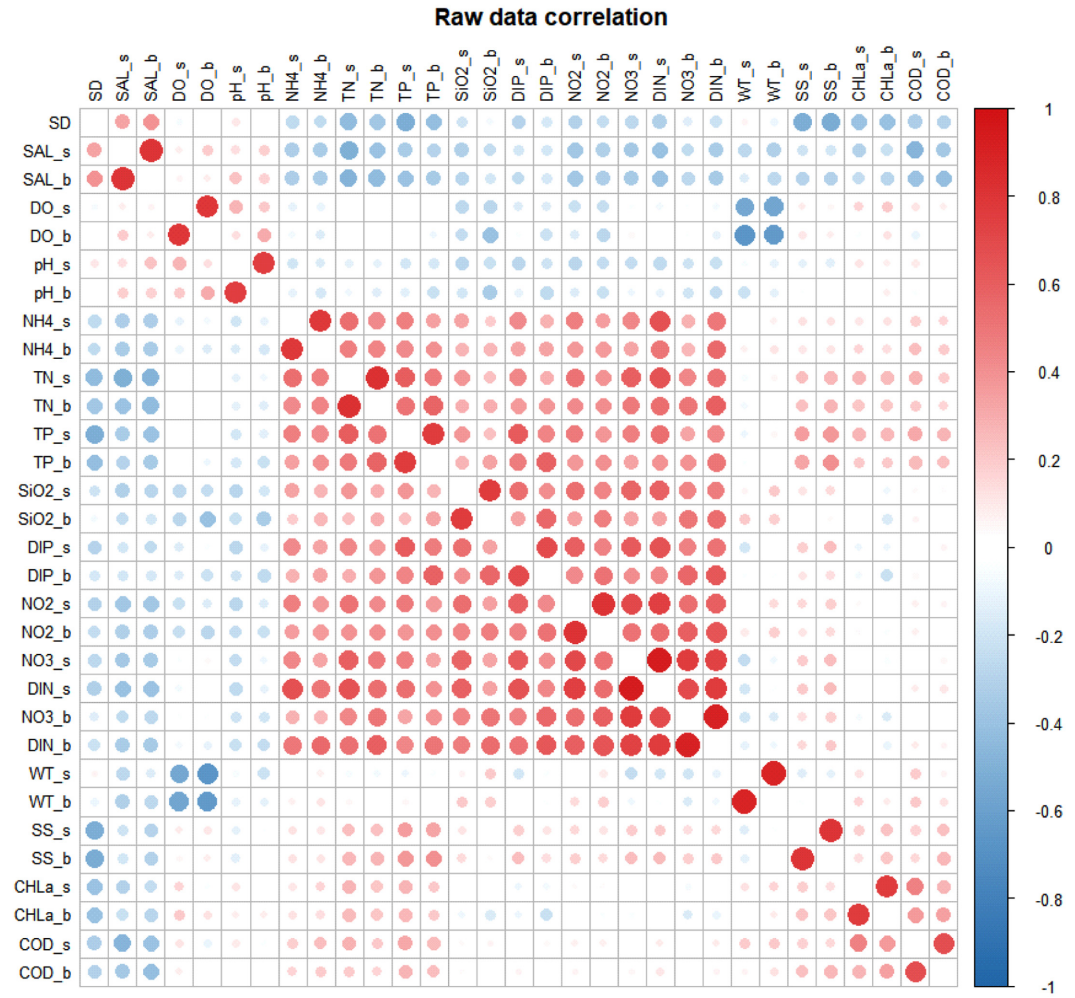

Fig. 1. Multivariate correlation analysis of raw data.

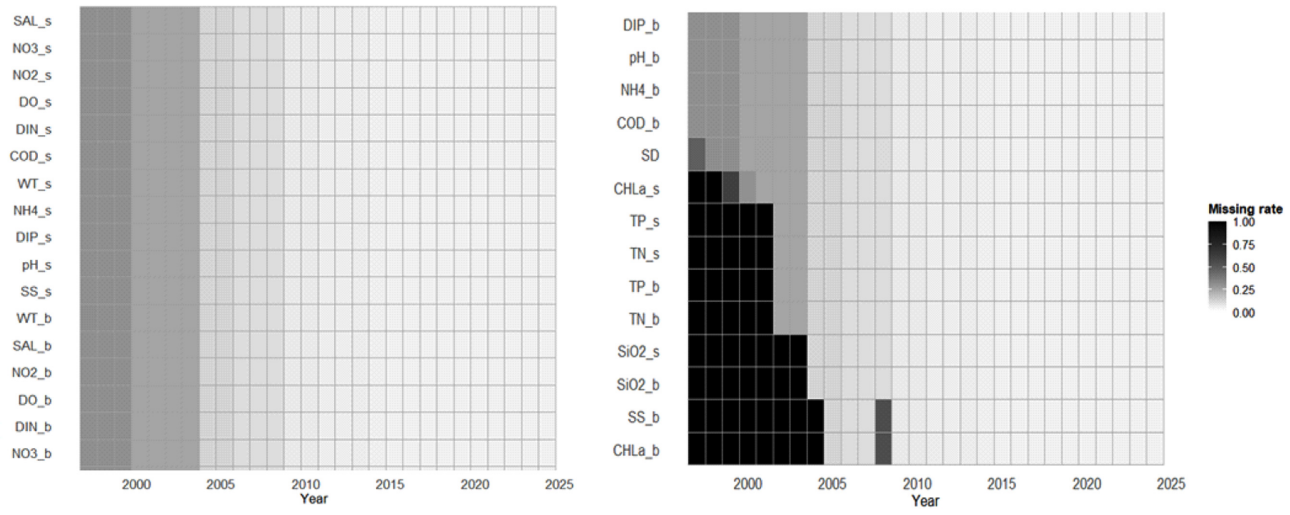

Fig. 2. Temporal variation of missing rates.

과 관측 항목이 늘면서 결측률이 낮아지는 모습을 볼 수 있다. 또한 자료 차원을 맞춘 전체 자료의 결측률은 평균 28%를 나타내었다. 본 연구에서는 해당하는 전체 자료(425개 정점, 28년 자료, 31개 변수)를 대상으로 결측치를 대치하되 대치에 대한 비교평가 방법은 2.3에서 자세히 서술하였다.

2.3 다변량 결측 대치 방법

본 연구에서는 해양환경관측자료의 다변량 결측치를 대치하기 위해 Amelia와 MICE(Multivariate Imputation by Chained Equations) 두 가지 대표적 알고리즘을 적용하였다. Amelia는 다변량 정규분포 가정을 기반으로 한 부트스트랩–EM(Expectation–Maximization) 절차를 이용하여 결측치를 반복적으로 추정하는 방식이며, MICE는 각 변수의 조건부 분포를 순차적으로 추정하는 fully conditional specification 구조를 통해 다중대치를 수행한다.

두 알고리즘 모두 결측치의 불확실성을 반영하기 위해 다중대치(multiple imputation)를 생성한다는 공통점을 가지며, 본 연구에서는 각 알고리즘에서 생성된 다중대치 결과의 마지막값을 최종 대치값으로 활용하였다. 알고리즘의 상세한 수식 전개, 사후분포 추정 과정, 변수별 조건부모형 구성 방식 등 기술적 세부 내용은 본문 논리 흐름을 유지하기 위해 Appendix A에 제시하였다.

2.4 알고리즘 성능평가 및 대치 자료의 비교 평가 방법

결측 대치 방법의 성능 평가는 다음과 같은 방법을 사용하였다. 먼저 해양환경측정망 자료의 결측 패턴은 특정 시점과 항목에 치중되어 있어 결측대치 기법에서 가정하는 무작위 결측 조건에는 부합되지 않지만, 현재로서는 무작위 결측 조건을 위배하는 경우에서의 다변량 결측 대치 기법은 매우 제한된 상황이다. 따라서 본 연구에서는 일단 현재의 결측 조건을 그대로 반영하여 결측 대치를 수행하였다. 기법에서 요구하는 결측 양상과 대상 자료의 결측 양상 차이로 인한 편향 등의 영향은 모델 성능 평가에서 검토하였다.

해양환경측정망 자료는 1997년 이후로 정점 및 관측 변수가 지속적으로 증가해왔으므로 1997년을 기준으로 관측 자료가 가장 많은 정점들을 우선적으로 선택하기 위해 1997년부터 관측을 시작한 정점 중(234개) 항만측정점(25개)을 제외한 209개 정점을 선정하였다.

이후 관측 변수 또한 자료가 가장 많은 변수들을 선택하여(1997년 기준 관측값이 전혀 없는 변수 및 상대적으로 결측률이 높은 투명도를 제외) Amelia와 MICE의 알고리즘 자체 성능을 평가하였다. 먼저 관측치가 있는 자료만을 대상으로 전체 자료중 10%, 20%, 50%를 랜덤하게 선택(Random seed = 1234)하여 임의적인 결측치(NA)를 생성하고 해당 결측을 두 가지 방법으로 대치하여 정답값 대 대치값을 산포도 그림(Scatter plot)으로 확인하였다. 단, 각 변수마다 단위가 상이하여 범위를 0~1로 조정해주는 표준화를 진행하였다. 이는 전체 자료에 대한 결측값 대치가 결측률 100%인 패턴을 보간한다는 가정하에 결측률이 증가할 수록 해당 알고리즘의 성능이 어느정도 나오는지 보증의 개념이라고 판단할 수 있다. 이후 MAPE(Mean Absolute Percentage Error)를 이용하여 오차를 정량화 해주었는데 이는 마찬가지로 단위가 달라 스케일의 의존적 문제를 해결하기 위해 오차를 비율로 환산해주는 MAPE를 선택하였다. MAPE의 계산과 해석은 다음과 같다.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

여기서 n은 표본 수, y_i는 정답값, $\hat{y}_i$ 는 추정값을 의미한다. MAPE 수치의 해석은 만약 MAPE가 0.05가 나왔다고 한다면, “평균적으로 정답값 대비 추정값의 오차가 5%이다”라고 해석할 수 있다.

완전하게 대치된 자료의 비교 평가는 밀도 분포 함수의 RMSE(Root Mean Squared Error)로 진행하였다. 우선적으로 대표 변수(염분, 용존산소, COD 등 10개)를 선정하여 표층과 저층을 평균하여 밀도 함수를 시각화 하였다. 시각화는 대치 전(원 자료), 대치 자료(추가한 자료)로 구분하였으며 각각 밀도 분포 함수의 RMSE를 통해 오차를 정량화 하였다. 여기서 RMSE는 원 자료의 밀도 분포 함수 대비 Amelia 대치 자료, MICE 대치 자료 각각의 밀도 분포 함수가 얼마나 차이 나는지 정량적으로 확인하기 위해 사용하였다. 이는 원 자료의 밀도 분포 함수 대비 Amelia 대치 자료 밀도 분포 함수의 RMSE를 100% 오차라고 가정하고 MICE 대치 자료 밀도 분포 함수의 오차를 상대적으로 나타내기 위해 비율 또한 나타내었다.

관측을 하지않은 정점, 관측을 하지않은 변수는 정답값이 없으므로 실제 오차를 내기엔 한계가 있으므로 가용하는 알고리즘의 성능을 우선적으로 평가하고, 해당 알고리즘을 통해 도출된 결측 대치 자료와 관측 자료의 밀도 분포 함수의 차이를 확인함으로써 관측 자료의 특성을 얼마나 반영하고 있는지를 확인하였다.

3. 결 과

3.1 성능 평가

본 연구에서는 결측률을 10%, 20%, 50%로 단계적으로 증가시킨 후 Amelia와 MICE 알고리즘을 적용하여 결측 대치 성능을 비교하였다(Fig. 3). 두 방법 모두 결측률이 증가함에 따라 관측값과 대치값 간의 일치도가 점차 저하되는 경향을 나타냈다. 결측률이 10% 수준에서는 대치값이 관측값의 분포를 비교적 잘 재현하였으나, 20% 이상에서는 분산이 커지며 불확실성이 증가하였다.

Amelia는 결측률이 낮을 때 안정적인 성능을 보였지만 결측률이 높아질수록 특정 변수(COD, NH_4, DIP 등)에서 외곽값이 크게 분포하고 편차가 확연히 증가하였다. 반면, MICE는 전반적으로 Amelia보다 대각선(1:1 Line) 근처에 데이터가 밀집하는 양상을 보이며 실제값의 분포 특성을 상대적으로 잘 반영하였다. 특히 결측률이 50%에 달하는 경우에도 Amelia보다 관측치의 구조를 더 잘 유지하는 경향을 확인할 수 있었다. 다만 MICE 역시 결측률이 매우 높은 상황에서는 일부 변수에서 분산이 과도하게 확대되거나 분포 왜곡이 발생하였다.

결측률이 증가함에 따라 Amelia와 MICE 모두 전반적으로 MAPE가 상승하는 경향을 보였다. 온도(WT)와 염분(SAL) 변수에서는 두 방법 모두 상대적으로 낮은 오차 수준을 유지하였으며, 예를 들어 WT_s는 10% MR에서 Amelia 2.0%, MICE 1.7%를 보였고, 결측률이 50%로 증가하더라도 각각 18.2%, 17.5% 수준으로 비교적 안정적인 패턴을 나타내었지만 50% 결측률에선 계절상관이 큰 수온의 경우 해당 상관을 반영하지 못한 결과로 해석된다. SAL 변수 역시 전 구간에서 낮은 오차 범위를 유지하며 결측률 증가에 따른

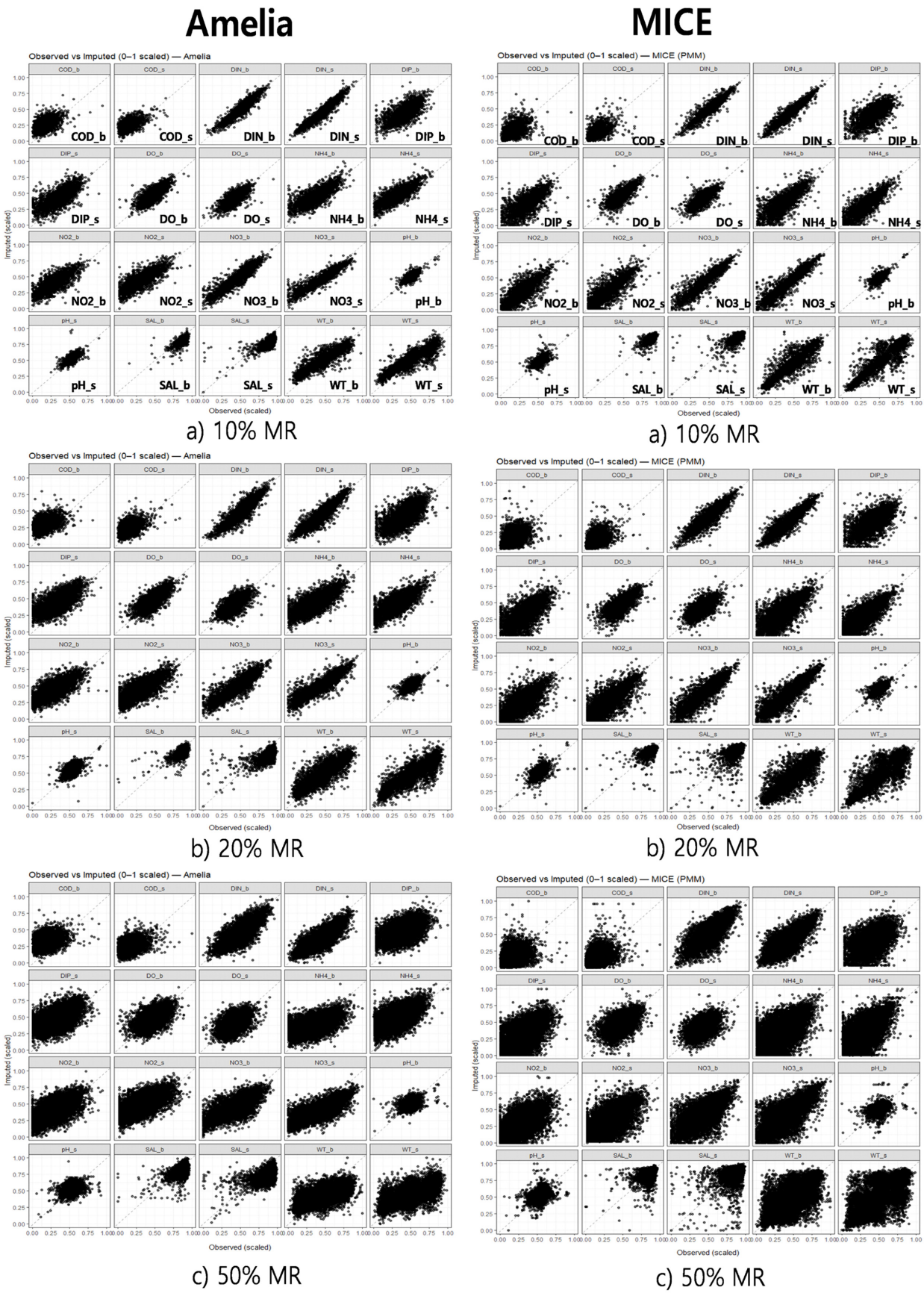

Fig. 3. Scatter plots of imputed versus observed values by imputation method at different missing rates (x–y axes scaled to 0–1).

Table 2. Result of MAPE estimation (unit: %)

	10% MR		20% MR		50% MR	
variable	MAPE_Amelia	MAPE_MICE	MAPE_Amelia	MAPE_MICE	MAPE_Amelia	MAPE_MICE
WT_s	2.0	1.7	4.8	4.1	18.2	17.5
WT_b	2.3	1.9	5.3	4.7	18.6	18.0
SAL_s	1.3	1.1	2.4	2.0	8.6	8.0
SAL_b	0.4	0.3	0.8	0.6	4.7	4.8
pH_s	0.1	0.1	0.3	0.3	0.9	0.8
pH_b	0.1	0.1	0.3	0.2	0.9	0.8
DO_s	1.1	1.0	2.3	2.2	7.2	6.8
DO_b	1.2	1.2	2.6	2.5	8.1	8.1
COD_s	6.7	5.8	14.7	11.9	38.0	35.6
COD_b	6.4	5.9	14.2	12.2	40.7	35.9
NH4_s	4.5	4.3	9.9	9.2	37.7	37.5
NH4_b	4.5	4.3	9.6	9.1	36.8	33.6
NO2_s	6.0	4.8	13.1	11.6	39.0	37.8
NO2_b	4.4	4.4	9.7	9.8	30.4	29.1
NO3_s	2.7	2.7	5.4	5.4	21.0	22.1
NO3_b	2.0	1.8	4.6	4.7	16.4	18.1
DIN_s	0.8	0.8	1.9	1.8	7.4	8.0
DIN_b	0.7	0.7	1.7	1.6	6.4	6.6
DIP_s	7.3	7.7	14.7	18.0	41.1	43.4
DIP_b	2.9	3.0	9.5	10.1	30.7	31.0

변동 폭이 제한적이었다.

반면, COD, NH4, NO2, DIP와 같은 영양염 및 유기물 지표에서는 명확하게 높은 MAPE가 관찰되었다. 특히 COD_s의 경우 10% MR에서 Amelia 6.7%, MICE 5.8%였으나, 50% MR에서는 각각 38.0%, 35.6%로 크게 증가하였다. NH4_s와 NO2_s 역시 50% MR에서 두 방법 모두 30% 이상의 높은 오차를 보이며 결측률 증가에 매우 민감한 특성을 나타냈다. DIP_s의 경우 MICE가 Amelia보다 상대적으로 안정적인 경향을 보였으며, 50% MR에서 Amelia는 41.1%, MICE는 43.4%로 다소 차이를 보였으나, 전반적으로 두 알고리즘 모두 높은 오차를 보이는 변수군에 속하는 것으로 나타났다.

pH 및 DO와 같은 기본 수질 변수는 전체적으로 가장 낮은 오차 범위를 보였다. pH_s는 모든 결측률 조건에서 0.1–0.3% 수준의 매우 낮은 MAPE를 유지하였고, DO 변수 또한 10% MR 기준 1.1–1.2%, 50% MR 기준 7.2–8.1%로 비교적 안정적인 성능을 보였다. 전반적으로 MICE는 Amelia와 비교하여 대부분의 변수에서 동등하거나 더 낮은 오차를 나타냈으며, 특히 SAL, COD, NH4, NO2 변수에서 더 일관된 성능을 보이는 경향이 확인되었다(Table 2).

대치기법간 MAPE 결과에 대해 각 결측률 별로 통계적으로 유의미한 차이가 있는 지 판단하기 위해 t-Test를 진행하였다. 10% MR에서는 p-value가 0.8, 20%는 0.85, 50%에서는 0.9로 MAPE 결과로 보았을 때 대치기법 간 성능에선 유의미한 차이가 없음을 나타내었다. 다만 MAPE 수치를 변수별로 절대적으로 판단 하였을 땐 MICE가 전반적으로 우수한 성능을 나타내었다.

3.2 상관분석

결측 대치에 따른 상관계수의 차이 결과는 다음과 같이 나타났다(Fig. 4). 먼저, 원 자료 상관분석에서는 먼저 표-저층 간 변수의 상관에서 대부분 1에 근사한 양의 상관을 나타내었다. 영양염 변수(*NH_4*, *TN*, *TP*, *SiO_2* 등)들 사이에서는 그림 중앙 부분에서 확인할 수 있듯 양의 상관을 뚜렷하게 나타내었다. 또한 용존산소 변수는 수온과의 강한 음의 상관을 나타냈으며 대부분의 변수에서 염분 변수와 음의 상관계수를 나타내었다.

다중대치 기법(Amelia, MICE) 적용 전후의 변수 간 상관성 변화를 비교한 결과, 결측 대치 수행 이후에도 상관계수의 전반적인 분포는 원자료와 큰 차이를 보이지 않았다. Amelia 대치 결과의 경우 원자료 대비 상관계수 차분 값의 분포가 대부분 –0.025~0.025 이내에 집중되어 있었으며 이는 상관구조가 원자료와 거의 유사하게 유지되었음을 의미한다. MICE 결과 또한 전반적으로 원자료와 유사한 상관성을 보였으나, 일부 변수쌍에서 Amelia에 비해 더 넓은 범위(–0.05~0.05)의 차이가 관찰되었지만 그 크기는 크지 않다.

따라서 두 대치 방법 모두 결측치 대치 과정에서 변수 간 상관성에 미치는 영향은 제한적이며 특히 Amelia는 상대적으로 더 안정적으로 원자료의 상관구조를 보존하는 경향을 확인할 수 있었다.

3.2 밀도 분포 함수 비교 및 평가

MICE, Amelia 기법의 성능평가는 분포 재현으로 수행하였으며 원자료의 분포를 기준으로 사용하였다. Fig. 5는 원 자료, Amelia, MICE로 대치한 자료의 밀도함수를 비교한 결과이다. 각 변수별로 전체적인 분포 형태는 유사하게 나타났으나, 분포의 중심과 꼬리

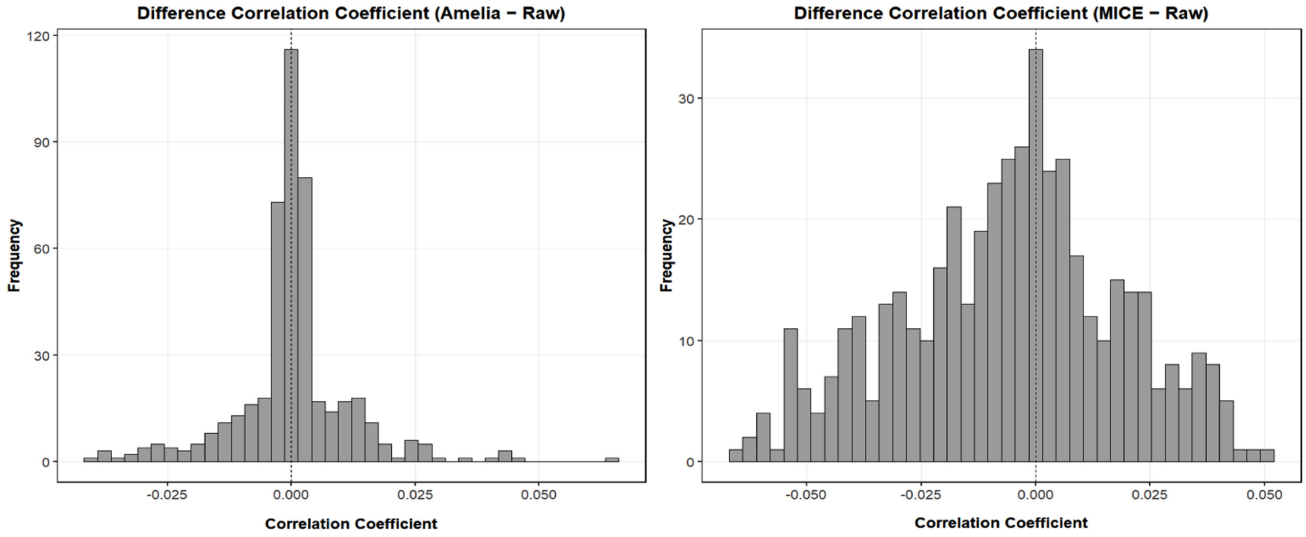

Fig. 4. Histogram of differences in correlation coefficients.

(tail) 부분에서 기법에 따라 차이가 관찰되었다. SAL과 DO의 경우 Amelia에 비해 MICE에서 원자료와 거의 동일한 분포를 보였으며, 특히 MICE는 분포의 첨두치(peak)를 원자료와 가장 가깝게 재현하였다. COD, DIN, DIP에서는 Amelia는 분산이 큰 형태를, MICE에서는 첨도가 더 큰 형태를 나타내었다. TN에서는 Amelia와 MICE 모두 원 자료에 비해 편향된 결과를 보였다.

SiO_2, SS, CHLa에서는 MICE가 Amelia 대비 원자료와 더욱 일치하는 분포를 나타냈다. 특히 CHLa에서는 Amelia가 분포 중심을 과대 추정하는 반면, MICE는 원자료와 거의 동일한 형태의 밀도 곡선을 재현하였다. 전반적으로 Amelia 기법이 정규분포 가정에 크게 제약되기 때문으로 판단된다. 반면 MICE는 원 자료가 가지는 분포와 유사한 비정규(non-normal)분포의 재현이 보다 우수하다.

RMSE 기반 정량 비교 결과, SAL, DO, SiO_2, SS, CHLa 등에서는 MICE가 Amelia 대비 낮은 RMSE를 보여 원자료 분포를 더 잘 보존하는 것으로 나타났다(Table 3). 해당 RMSE를 평가하기 위하여 Amelia의 RMSE를 기준(=100%)으로 두었을 때 MICE의 RMSE가 얼마인지 비율로 나타내어 Ratio(%)로 정의하여 평가하였다. SAL의 경우 Amelia 대비 MICE RMSE(RMSE Ratio in Table 3)는 약 38.8% 수준이었으며, DO, SiO_2, SS, CHLa 또한 각각 42.5%, 30.2%, 40.4%, 20.4%로 낮게 산정되어 MICE의 성능이 Amelia에 비해 상대적으로 좋음을 나타내었다. 반면 DIN, DIP에서는 MICE RMSE가 상대적으로 크게 나타나(214.6%, 125.0%), Amelia에 비해 분포 보존력이 떨어지는 것으로 평가되었다. COD, TN, TP의 경우 두 방법 간 차이는 크지 않았으며, MICE는 Amelia 대비 90~100% 수준의 유사도를 보였다. 해당 비율은 낮을 수록 원 자료 밀도 분포에 대한 MICE 결측 대치 방법이 Amelia 방법에 비해 상대적으로 원 자료의 분포를 잘 재현한다는 의미이다.

4. 토 의

4.1 시계열 자료 다중대치 방법의 한계

본 연구에서는 검증된 소프트웨어인 Amelia와 MICE를 활용하여 결측률을 인위적으로 조절한 모의실험을 수행하고, 실제 관측자료를 이용하여 대치 성능을 비교하였다. 간헐적 결측이 발생하는 경우에는 활용 가능한 주변 정보의 수가 충분하여 다변량 대치 기법이 대체로 우수한 성능을 보였다. 원 자료 자체의 변동성이 큰 수질 환경 변수의 대치 결과는 원 자료의 변동을 잘 나타냈으나 강한 계절성을 갖는 시계열 변수, 특히 수온 자료의 경우에는 대치 성능의

Table 3. RMSE results based on density distribution functions

Variables	RMSE_Amelia	RMSE_MICE	Ratio(%)(criteria : Amelia=100%)
SAL	0.0281	0.011	38.8
DO	0.0106	0.0045	42.5
COD	0.0315	0.0294	93.3
DIN	0.0321	0.0689	214.6
TN	0.1341	0.1305	97.3
DIP	0.032	0.04	125.0
TP	0.1067	0.01	93.3
SiO2	0.0739	0.0223	30.2
SS	0.0376	0.0152	40.4
CHLa	0.0657	0.0134	20.4

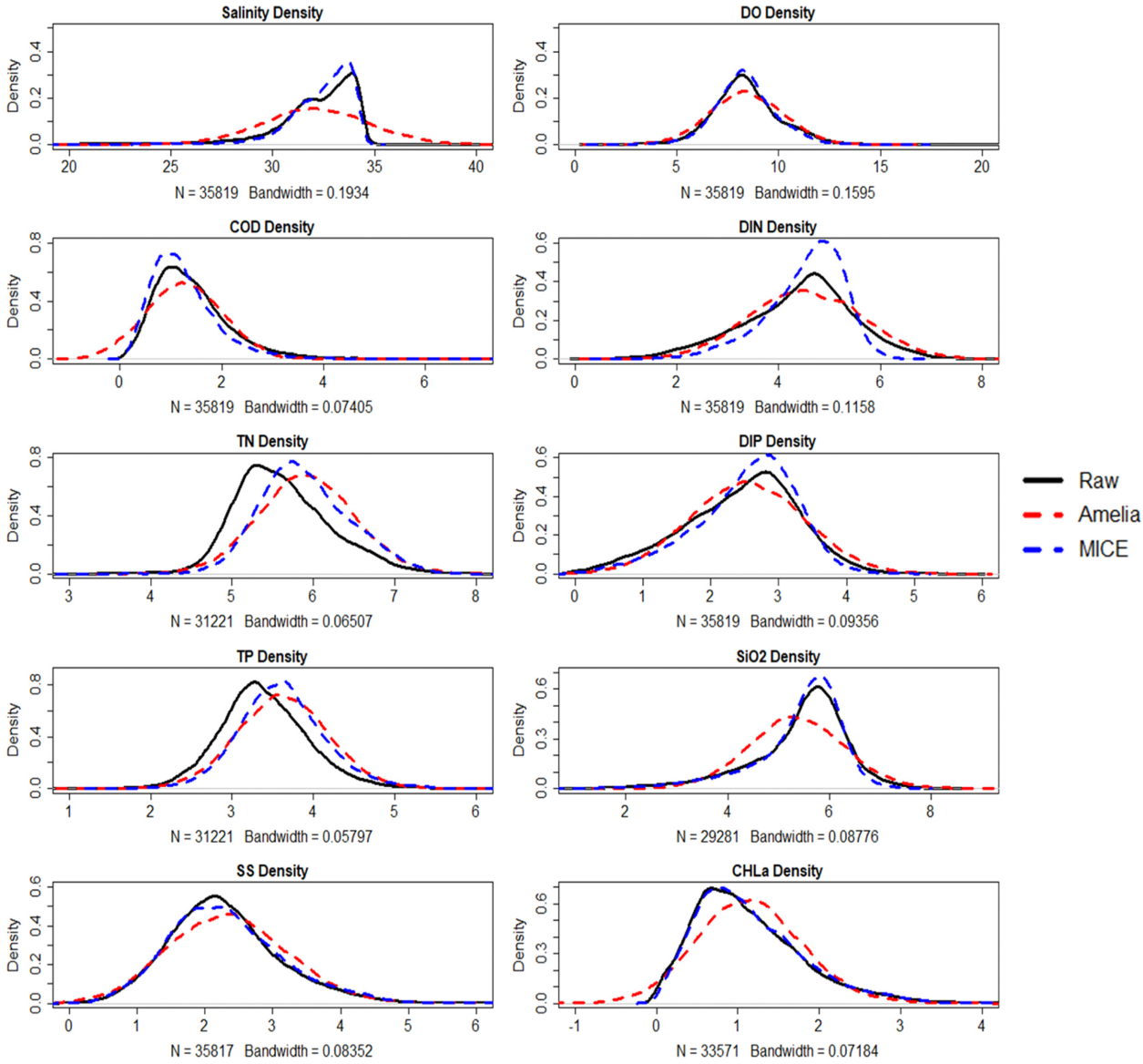

Fig. 5. Density function comparison: Raw vs. Amelia vs. MICE.

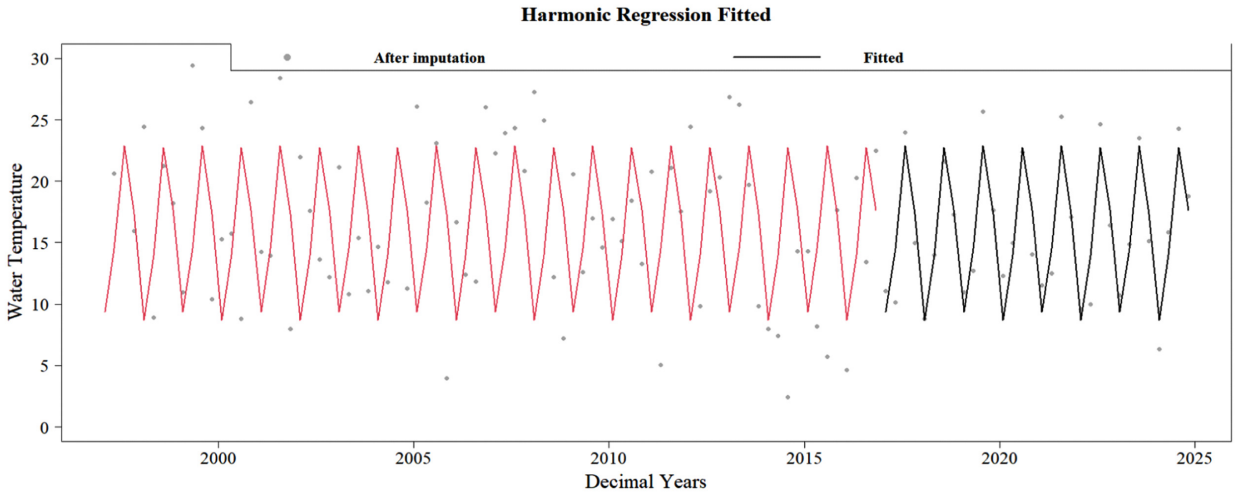

Fig. 6. Harmonic regression result (station : BE 1901).

저하가 뚜렷하게 나타났다(Fig. 3). 이는 단순히 분포 재현에 중점을 둔 기존 다변량 대치 방법이 시간에 따른 자기상관 구조와 계절적 변동성을 충분히 반영하지 못하기 때문에 발생한 현상으로 해석된다. 따라서 현재의 다변량 대치 모델 구조에는 이러한 시간 의존성을 구현할 수 있는 보완적 조치가 필요하다.

Junninen *et al.*(2004)는 대기질 시계열 자료에서 결측 비율과 패턴에 따라 대치 성능이 달라짐을 보여주었으며, 특히 연속적인 결측 구간에서는 성능 저하가 크게 나타남을 보고하였다. 또한 Moritz and Bartz-Beielstein(2017)은 Kalman smoothing, 계절분해 기반 보간 등 시계열 전용 알고리즘을 비교한 결과, 주기성이 강한 경우 자기상관과 계절성을 고려한 방법의 성능이 우수함을 제시하였다. Welch *et al.*(2014) 역시 시간 블록을 분리하여 독립적으로 결측을 대치하는 방법을 제안하였으나, 궁극적으로는 인접 시점의 정보를 활용하는 2-fold FCS(Fully Conditional Specification) 모델이 더 효과적임을 강조하였다.

따라서 수온과 같이 강한 주기성과 자기상관을 가지는 변수의 대치에서는 단순한 분포 기반의 다변량 대치만으로는 위상과 진폭 변화를 충분히 복원하기 어렵다는 점이 확인된다. 실제로 Amelia는 다변량 정규분포 가정을 기반으로 변수들 간의 상관구조를 잘 재현하였으나(Fig. 7~Fig. 8), 시간 축에서의 변동 양상(seasonal cycle, lagged memory)은 별도의 구조항을 반영하지 않는 경우 주기적 위상 정렬이 어긋나거나 극값 발생 시점을 재현하지 못하는 한계가 관찰되었다. 다만 Amelia는 시계열·패널 확장을 위한 polytime, intercs, lags/leads 등의 기능을 제공하므로 이를 활용한다면 이러한 한계를 완화할 수 있을 것으로 판단된다. MICE 또한 기본 구조는 시계열 전용이 아니지만, 대치 모형에 시간 변수나 외부 설명변수를 추가하는 방식으로 개선 가능성이 존재한다.

4.2 시계열 자료 대치 성능 개선 방안

앞서 확인된 한계점을 보완하기 위해, 시계열 자료의 구조적 특성을 대치 과정에 반영하는 개선 방안을 검토할 필요가 있다. 특히, 자료 구조상 자기회귀나 추출된 주기 성분을 소스코드 자체의 수정 없이 고려할 수 있는 방법으로 더미변수의 활용이 가능하다. 이를테면 주기 신호를 분별하기 위해 harmonic regression을 통해 추출

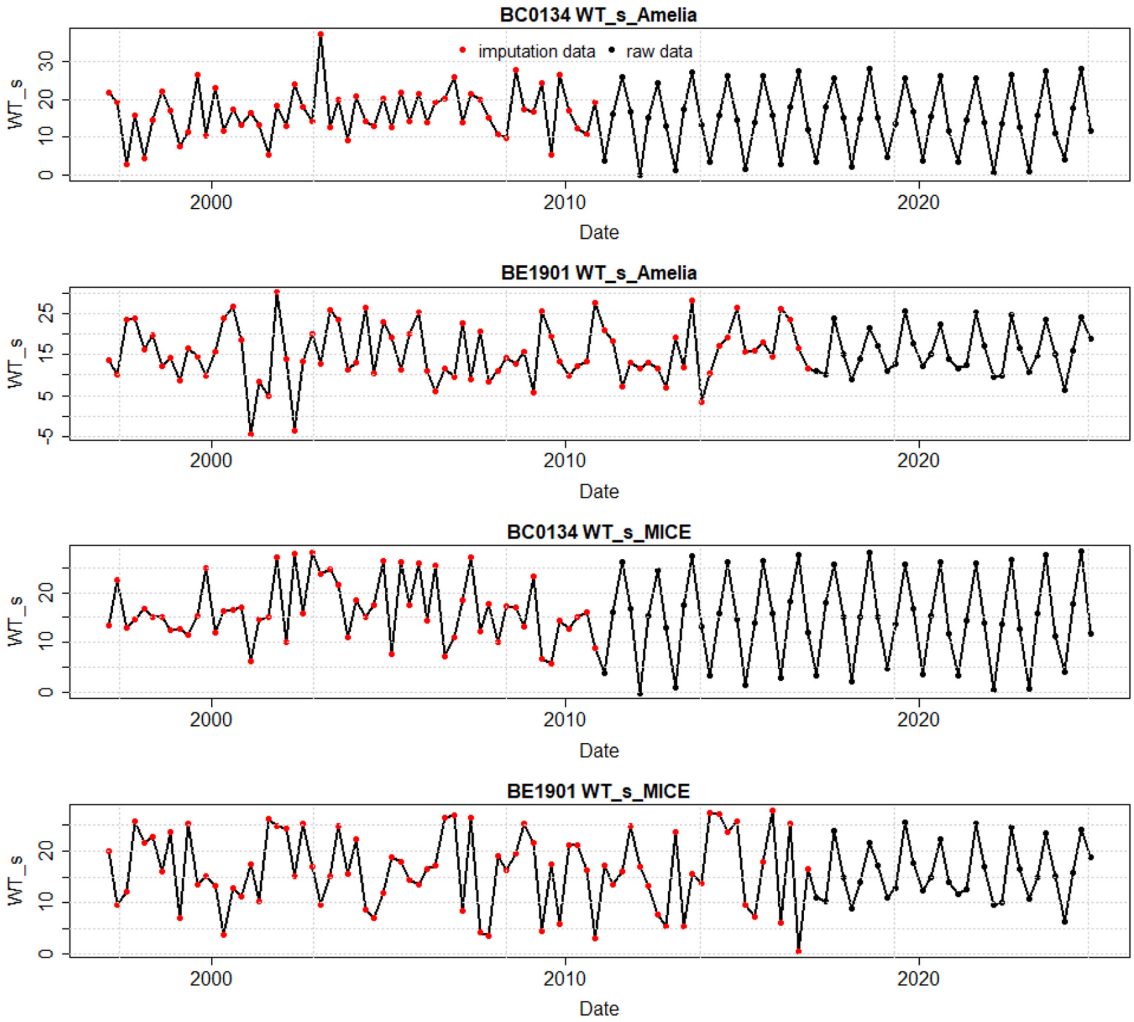

Fig. 7. Surface water temperature imputation results (Representative Stations: BC0134, BE1901).

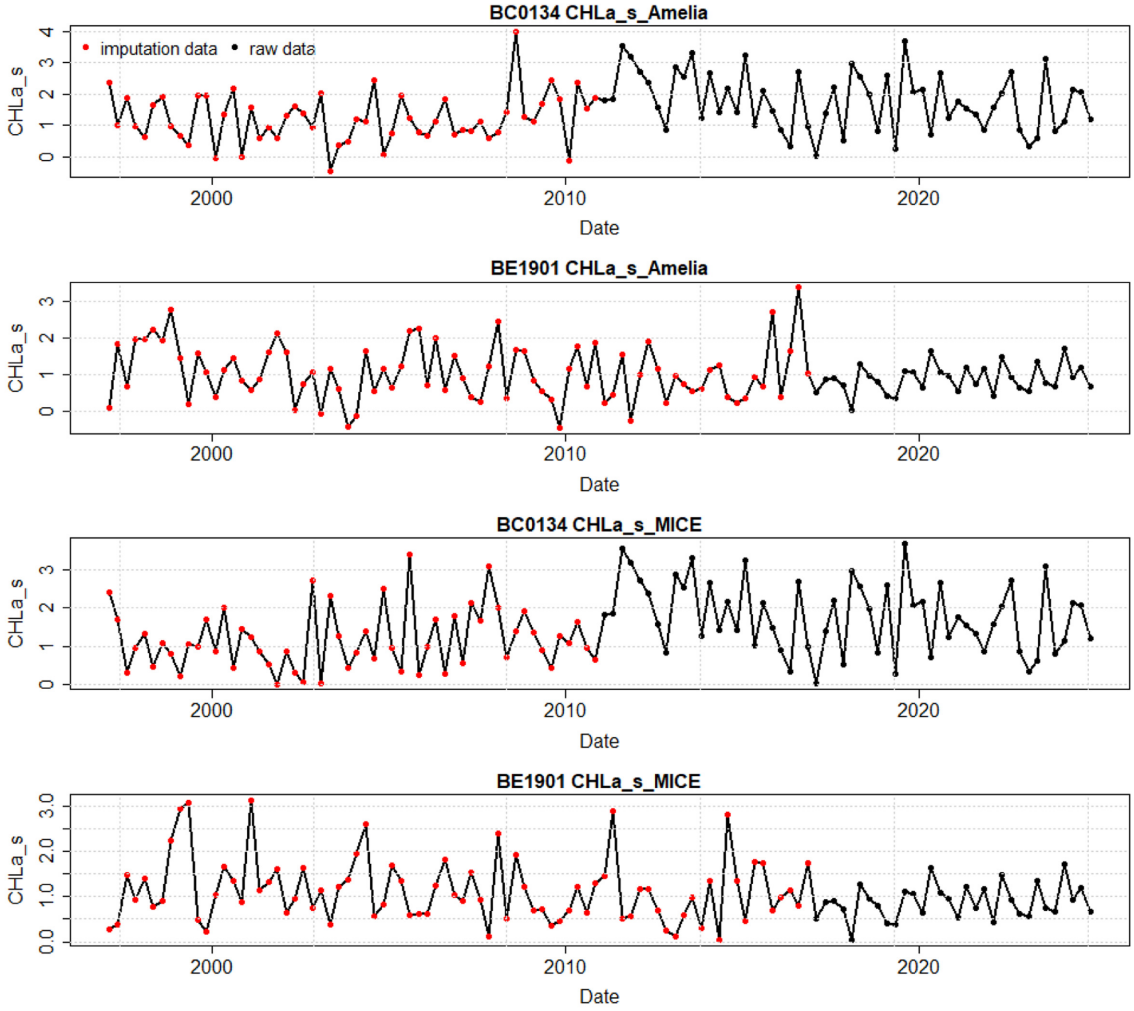

Fig. 8. Surface chlorophyll a imputation results (Representative Stations: BC0134, BE1901).

한 주기 항이나, lagging 기법을 활용하여 시차 변수를 도입함으로써 장기 블록 결측 구간에서도 대치 성능을 개선할 수 있을 것으로 기대된다. Fig. 6에서 조화분석을 통한 주기성분 추출의 예시로 기존 관측자료만을 사용하여 조화분석을 진행하고(Black line) 해당 결과를 통해 관측자료가 없던 과거 시간에 대해 조화회귀 결과(Red line)를 나타내었다.

Honaker and King (2010)은 Time Series Cross Section 자료에서 다중대치를 적용할 때, 시간 추세, 고정효과, 시차 등의 구조를 모형에 포함하는 것이 필수적이라고 지적한 바 있다. 이는 본 연구의 결과와도 일맥상통하며, 단순한 다변량 상관관계 기반의 접근을 넘어 시계열 고유의 동적 특성을 반영해야 함을 의미한다.

또한 대치 기법을 공간 차원으로 확장하는 방법론도 고려할 수 있다.

Alvera-Azcárate *et al.* (2005)는 EOF(Empirical Orthogonal Function) 기반의 공간적 상관성을 활용하여 구름으로 인한 위성 자료의 결측을 보완하였다. 그러나 본 연구는 전국 연안을 대상으로 하는 다수의 정점을 포함하고 있어, 공간적 균질성 가정을 적용한 선행 연구와는 다른 조건을 가진다. 따라서 공간 상관 구조를 적절히 구현할 수 있을지는 추가적인 검토가 필요하다.

본 연구는 주로 MAR(Missing At Random) 가정하에서 평가되었으나, 실제 해양 관측 환경에서는 계절적 유지보수, 센서 고장 등으로 인한 MNAR(Missing Not At Random) 발생 가능성이 높다. 태풍 등 돌발적인 자연재해 역시 연속적인 블록 결측을 유발할 수 있으며, 이 경우 고도화된 대치 모형을 적용하더라도 원본 정보를 충실히 복원하는 데에는 한계가 존재한다. 따라서 이러한 한계를 극복하기 위해서는 원격탐사 자료나 수치모델 기반의 재분석 자료를 보조변수로 포함하는 방안도 검토해 볼 수 있다.

4.3 결측 패턴에 대한 반응

본 연구에서 10, 20, 50%의 간헐적 결측을 인위적으로 생성해 비교한 결과, 결측률이 낮을수록(10–20%) 다변량 대치가 주변 시공간 공분산 구조를 잘 활용해 낮은 재현 오차를 보였고, 50% 수준의 결측 비율에서는 성능 저하가 뚜렷했다.

실험 결과는 약 20%의 오차까지는 비단조(nonmonotone) 결측

패턴에 대해서도 Amelia와 MICE 모델을 이용한 대치가 부분정보를 효율적으로 재현할 수 있음을 보여주었으나, 관측 시기가 늦은 정점과 같이 결측이 블록 형태로 길게 이어질수록 시계열 정보의 재구성이 어려워져 성능이 급격히 떨어졌다.

결측 비율과 관련하여, 분류모델을 구축하는 경우 Acuna and Rodriguez(2004)는 결측률에 따라 심각한 영향 없음(1% 미만), 관리 가능(1~5%), 정교한 방법론 필요(5~15%), 예측 모델의 심각한 성능 저하로 결측치 비율에 따른 결과 왜곡 현상이 있음(15% 이상)을 언급하고 있으며, 강필성(2012)은 결측치를 제외하는 방법보다 간단한 선형 대치 기법을 통해서 자료를 보완하는 것이 성능 저하가 적다고 언급했다.

다만, 앞서 언급되었듯이 단순히 결측률의 절대 수치가 아니라 “결측이 어떤 패턴으로 발생하는가”가 성능에 미치는 영향에 대한 연구는 상대적으로 미비한 편이며 본 연구에서 실험과 실제 자료와 같이 간헐적 결측에서는 대치가 비교적 안정적이지만 블록 패턴의 결측에서는 인접 정보의 활용도가 떨어져 재현력이 급격히 저하되는 특징을 지니기 때문에 결측 패턴 또한 중요하게 고려되어야 한다.

Welch *et al.*(2014)은 인접 시점의 데이터를 고려하는 것뿐만 아니라 임의로 생성한 최대 70%의 결측 정보를 대치하며 2-Fold FCS 모델의 성능이 비교적 안정적임을 보고하였다. 다만, 본 연구를 포함하여 해양환경에서 관측한 다양한 자료들은 많은 경우 MCAR, MAR 가정을 적용하기 어렵기 때문에, 실제 적용 시 한계는 50% 이하의 결측에 대해서만 효과적일 수 있으며, 같은 조건에서도 블록 형태의 편향된 결측은 인접 정보를 활용하는 다중대치 모델에 치명적일 수 있으므로, 결측률 뿐만 아니라 결측 패턴에 대한 영향을 동시에 고려한 정량적 평가가 필수적이다.

5. 결 론

본 연구는 해양환경측정망의 장기 시계열 자료를 대상으로 다변량 결측 대치 방법(Amelia, MICE)의 성능을 비교·평가 하고 실제 자료에 적용하였다. 실제 관측자료에 인위적으로 결측을 부여하고, 결측률(10%, 20%, 50%)에 따른 대치 결과를 검증함으로써 각각 알고리즘의 성능을 평가하였고, 해양환경 시계열 자료에서 다중대치 기법의 적용 가능성과 한계를 규명하였다.

첫째, 결측률 10%, 20%, 50%를 인위적으로 부여한 모의실험 결과, 두 방법 모두 결측률이 증가할수록 관측값과 대치값 간 일치도가 저하되었다. 10% 수준에서는 대치값이 관측분포를 비교적 잘 재현했으나, 20% 이상에서는 분산이 확대되고 불확실성이 커졌다. 특히 COD, NH_4, DIP와 같은 영양염 지표에서는 Amelia가 외곽값과 편차가 크게 발생한 반면, MICE는 대체로 실제값 분포를 안정적으로 보존하는 경향을 보였다. 반면 수온(WT)과 염분(SAL)은 결측률이 50%까지 증가하더라도 낮은 MAPE(0.4 이하)를 유지하여 상대적으로 안정적인 성능을 나타냈다.

둘째, 다변량 상관분석을 통해 결측 대치 전후의 상관구조 변화를 비교한 결과, Amelia와 MICE 모두 변수 간 상관성이 원자료와 큰 차이를 보이지 않았다. Amelia는 상관계수 차분 값이 대부분 –0.025~0.025 범위에 분포하며 원자료의 상관구조를 안정적으로 보존하는 경향을 보였다. MICE는 일부 변수쌍에서 Amelia보다 더 넓은 범위(–0.05~0.05)의 차이를 보였으나, 전체적으로는 원자료의 상관성을 잘 유지하였다.

셋째, 분포 비교에서 Amelia와 MICE 모두 원자료의 전반적 밀도 곡선을 재현했으나, 변수 특성에 따라 성능 차이가 나타났다. SAL과 DO에서는 MICE가 원자료 분포와 거의 일치하는 밀도를 보였으며, COD, DIN, DIP에서는 Amelia가 분산이 큰 형태, MICE는 첨도가 높은 형태를 나타냈다. SiO_2, SS, CHLa에서는 MICE가 원자료 분포를 더 잘 보존했으며, 특히 CHLa에서는 Amelia가 분포 중심을 과대 추정한 반면 MICE는 원자료와 거의 동일한 형태를 재현하였다.

넷째, 밀도 분포 함수에 대한 RMSE 기반 정량 비교 결과 SAL, DO, SiO_2, SS, CHLa 변수에서는 MICE가 Amelia 대비 낮은 RMSE(20–42% 수준)를 나타내 원자료 분포를 더 충실히 재현하였다. 반면 DIN과 DIP에서는 MICE가 Amelia 대비 높은 RMSE (125–215%)를 보여 분포 보존력이 떨어졌다. COD, TN, TP에서는 두 방법 간 차이가 크지 않았으며, 대체로 Amelia 대비 90–100% 수준의 유사도를 보였다. 이는 변수 특성에 따라 Amelia와 MICE의 성능 우위가 달라질 수 있음을 의미한다.

결론적으로 본 연구는 해양환경측정망과 같은 장기 시계열 자료에서 다중대치 기법이 결측 문제를 완화하는 유용한 방법임을 확인하였고, 동시에 자기상관·계절성과 같은 시간적 구조의 영향을 많이 받는 변수의 경우 성능 저하가 발생한다는 한계를 확인하였다. 따라서 간헐적 결측에서는 다변량 대치가 효과적이지만, 강한 계절성을 갖는 변수에서는 시간항을 포함한 구조화 대치가 필수적이다. 또한 Amelia/MICE의 표준 구현 위에 자기상관 및 계절 변동과 고정효과를 추가한 연구가 필요하며, 향후 인접 시공간 정보를 활용할 수 있는 n-fold FCS 모델 또는 다양한 시공간 변화양상을 고려할 수 있는 결측 대치 기법 모델로의 확장이 요구된다.

후 기

이 논문은 2022년도 정부(해양수산부)의 재원으로 해양수산과학기술진흥원-해양유해물질오염원 추적기법개발 사업 지원을 받아 수행된 연구입니다(RS2022-KS221655). 연구비 지원에 감사드립니다. 또한 해양환경측정망 자료를 제공해주신 해양환경공단에도 감사드립니다.

References

[1] Acuna, E. and Rodriguez, C., 2004, The treatment of missing values and its effect on classifier accuracy, in Classification, Clustering, and Data Mining Applications, Springer, Berlin, Heidelberg, 639-647.

[2] Afrifa-Yamoah, E., Mueller, U.A., Taylor, S.M. and Fisher, A.J., 2020, Missing data imputation of high-resolution temporal climate time series data, Meteorol. Appl., 27(1).

[3] Alvera-Azcárate, A., Barth, A., Rixen, M. and Beckers, J.M., 2005, Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: Application to the Adriatic Sea surface temperature, Ocean Model., 9(4), 325-346.

[4] Buuren, S. van and Groothuis-Oudshoorn, K., 2011, mice: Multivariate imputation by chained equations in R, J. Stat. Softw., 45(3), 1-67.

[5] Chen, C. and Xue, X., 2023, A novel coupling preprocessing approach for handling missing data in water quality prediction, J. Hydrol., 617.

[6] Cho, H.-Y., Lee, G.-S. and Lee, U.-J., 2021, Long-gap filling method for the coastal monitoring data, J. Korean Soc. Coast. Ocean Eng., 33(6), 333-344.

[7] Honaker, J. and King, G., 2010, What to do about missing values in time-series cross-section data, Am. J. Polit. Sci., 54(2), 561-581.

[8] James, H., Gary, K. and Matthew, B., 2011, Amelia II: A program for missing data, J. Stat. Softw., 45(7), 1-47.

[9] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M., 2004, Methods for imputation of missing values in air quality data sets, Atmos. Environ., 38(18), 2895-2907.

[10] MEIS(Marine Enviromental Information System), 2025, https://www.meis.go.kr/ Accessed at 2025.12.18.

[11] Kang, P., 2012, Missing value imputation based on locally linear reconstruction for improving classification performance, J. Korean Inst. Ind. Eng., 38(4), 276-284.

[12] Lange, N., Fiedler, B., Álvarez, M., Benoit-Cattin, A., Benway, H., Buttigieg, P. L. and Tanhua, T., 2023, Synthesis Product for Ocean Time-Series (SPOTS)–A ship-based biogeochemical pilot. Earth Syst. Sci. Data, 16, 1901-1931.

[13] Lepot, M., Aubin, J.-B. and Clemens, F., 2017, Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment, Water, 9(10), 796.

[14] Little, R.J.A. and Rubin, D.B., 2002, Statistical Analysis with Missing Data, Wiley-Interscience, New York.

[15] Mahmood, A.U., *et al*., 2024, Multiple data imputation methods advance risk analysis and treatability of co-occurring inorganic chemicals in groundwater, Environ. Sci. Technol., 58(46), 20513-20524.

[16] Moritz, S. and Bartz-Beielstein, T., 2017, imputeTS: Time series missing value imputation in R, R J., 9(1), 207-218.

[17] Rodríguez, R., *et al*., 2021, Water-quality data imputation with a high percentage of missing values: A machine learning approach, Sustainability, 13(11), 6140.

[18] Shin, Y., *et al*., 2022, Imputation of missing SST observation data using multivariate bidirectional RNN, J. Korean Soc. Coast. Ocean Eng., 34(4), 109-118.

[19] Sierra-Porta, D., 2024, Assessing the impact of missing data on water quality index estimation: A machine learning approach, Discover Water, 4(1).

[20] Stekhoven, D.J. and Buhlmann, P., 2012, MissForest—Nonparametric missing value imputation for mixed-type data, Bioinformatics, 28(1), 112-118.

[21] Wang, F., *et al*., 2024, Two-stage iterative approach for addressing missing values in small-scale water quality data, Mar. Dev., 2(1).

[22] Welch, C.A., Petersen, I., Bartlett, J.W., White, I.R., Marston, L., Morris, R.W., *et al*., 2014, Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data, Stat. Med., 33(21), 3725-3737.

[23] Zhang, S., *et al*., 2021, Imputation of GPS coordinate time series using MissForest, Remote Sens., 13(12), 2279.

Received 10 September 2025
1st Revised 22 October 2025, 2nd Revised 27 January 2026
Accepted 28 January 2026

Appendix. A

1) Amelia (Honaker *et al.*, 2011)

Amelia는 결측 자료 처리를 위해 완전한 자료는 "다변량 정규분포(Multivariate Normal)를 따른다"는 가정을 전제로 한다. 즉, 관측값과 결측값을 모두 포함하는 전체 데이터 셋이 정규성을 만족한다고 가정하여 $D \sim N_k(\mu, \Sigma)$ 이라 표현할 수 있다. 여기서 D는 완전한 데이터 셋이며, D가 평균 벡터인 μ와 공분산 행렬인 Σ로 이루어진 다변량 정규분포를 따른다는 의미이다. 다만 대치(imputation)의 본질적인 문제는 완전한 자료 D를 알 수 없고 오직 관측된 자료 D^{obs}만을 이용할 수 있다는 점이다. Amelia는 타 다중대치 기법과 마찬가지로 자료가 무작위로 결측됨(Missing At Random, MAR)을 가정한다. 이를 표현하기 위해 M을 결측 행렬로 정의하고 각 셀이 결측인지 여부를 나타낸다. MAR 가정은 다음과 같다.

$$p(M|D) = p(M|D^{obs})$$

만약 결측이 자료와 전혀 무관하게 발생한다면 완전 무작위 결측(Missing Completely At Random, MCAR)이라 부르며 Amelia는 다변량 정규성 가정 및 (완전)무작위 결측 가정을 모두 필요로 한다.

Amelia의 다중대치 알고리즘은 다음과 같다.

다중대치 에서는 완전한 자료의 모수(Parameter, $\theta = \mu, \Sigma$)를 얼마나 정확히 추정하는지가 중요하다. 실제로 관측하는 것은 D_{obs}와 결측행렬 M이므로 관측자료의 가능도(likeihood)는 $p(D^{obs}, M|\theta)$ 일 때, MAR 가정에 따라 다음과 같이 분해된다.

$$p(D^{obs}, M|\theta) = p(M|D^{obs})p(D^{obs}|\theta)$$

우리가 관심을 갖는 것은 완전자료의 모수 추정 이므로 가능도는 $L(\theta|D^{obs}) \propto (D^{obs}|\theta)$로 단순화 할 수 있으며 반복적 기대 법칙(law of expectations)에 의해 다음과 같이 표현된다.

$$p(D^{obs}|\theta) = \int p(D|\theta)dD^{mis}$$

이 가능도와 θ에 대한 평평한 사전분포(flat prior)를 결합하면 사후분포는 다음과 같다.

$$p(\theta|D^{obs}) \propto p(D^{obs}|\theta) = \int p(D|\theta)dD^{mis}$$

완전하지 않은 자료를 분석할 때 어려움은 사후분포로 부터 표본을 추출하는 과정이다. EM 알고리즘(Expectation-Maximization)은 사후분포의 최빈값(mode)을 찾기 위한 단순한 접근 방법이다(Fig. 3). Amelia는 EM 알고리즘에 부트스트랩 방법을 결합한 EMB 알고리즘을 사용하여 사후분포로 부터 표본을 추출한다. 각 표본에 대해 부트스트랩을 수행하여 불확실성을 모의하고 해당 부트스트랩 표본에 대해 EM 알고리즘을 수행하여 사후분포의 최빈값을 찾는다. 따라서 완전한 자료 모수의 사후분포로 부터 추출된 표본을 얻으면 이를 바탕으로 D_{obs}에 조건부인 D_{mis}의 분포로부터 결측값을 추출하여 다중대치를 생성하게 된다.

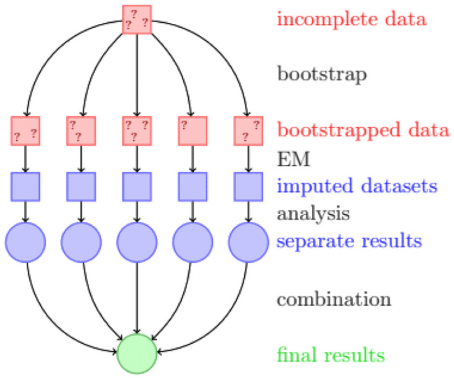

Fig. 9. A schematic diagram of multiple imputation with the EMB Algorithm (Honaker *et al.*, 2011).

여기서 다중대치된 데이터 셋이 m개 만들어 지는데 각각의 데이터 셋은 결측이 대치된 완전한 자료이지만 m개의 서로 다른 결과(추정치와 분산)가 생기게 된다. 이에 Rubin's Rules를 적용하여 m개의 추정 결과를 하나의 최종 추정 평균과 분산으로 합칠 수 있다. 여기서 결합된 최종 추정 평균($\bar{q}$)는 m개의 개별 추정치 $q_j (j = 1,2,...,m)$의 단순 평균으로 정의된다.

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j$$

최종 추정 분산($SE(q)^2$)은 m개의 데이터 셋에서 추정된 분산의 평균에 각 점추정된 표본분산을 더한 값으로 계산된다. 이때 m이 유한하기 때문에 생기는 편향(bias)을 보정하기 위해 표본 분산 항에 $1+1/m$의 보정계수를 곱한다. 즉, 데이터 셋 j에서 추정치의 분산을 $SE(q_j)^2 (j = 1,2,...,m)$이라 하고 $S = \sum_{j=1}^{m}(q_j - \bar{q})^2/(m-1)$을 m개의 표본분산이라 하면 최종 추정 분산은 다음과 같이 표현된다.

$$SE(q)^2 = \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + S\left(1 + \frac{1}{m}\right)$$

다만 실질적으로 대치된 데이터 셋을 사용하려면 추정된 m개 중 선택하여야 하는데 데이터셋 m개 모두 최종 추정된 평균과 분산을 따르는 분포에서 나온 "가능한 현실"이라는 의미이기 때문에 본 연구에서는 m번째 데이터셋을 사용하였으며 결측 대치 과정은 R 언어의 "Amelia"패키지의 Amelia 함수를 사용하였다.

2) MICE (Buuren and Groothuis-Oudshoorn, 2011)

MICE는 다변량 결측을 처리하기 위해 완전 조건부 대체법(Fully Conditional Specification, FCS)에 기반한 연쇄 방정식(Chained Equations) 접근을 사용한다. 이 방법은 각 변수별로 조건부 분포를 지정하여, 특정 변수를 제외한 나머지 변수들을 예측 변수로 활용해 결측값을 대치하는 방식으로 작동한다. 초기 대치를 통해 결측값을 임시로 채운 뒤, 각 변수에 대해 순차적으로 조건부 모형을 적합하고 결측을 다시 대치하는 과정을 반복하면서 전체 데이터셋을 점진적으로 갱신해 나간다. MICE 알고리즘에서 설명하는 다중 결측

대치의 초기 설정은 다음과 같다(Fig. 4).

$Y_j (j = 1,2,..., p)$를 p개의 불완전 변수 중 하나라고 가정한다. 여기서 $y = (Y_1, Y_2,..., Y_p)$일 때 관측된 부분과 결측된 부분은 Y_j^{obs} 와 Y_j^{mis} 로 나타낼 수 있다. 결측 대치 횟수 m은 1보다 큰 정수로 정의하고 h번째 대치된 데이터 셋은 $Y^{(h)} (h = 1,2,...,m)$으로 표기한다. 또한 $Y_{-j} = (Y_1,..., Y_{j-1}, Y_{j+1}, ...Y_p)$는 Y에서 Y_j를 제외한 나머지 p-1개의 변수를 의미한다. Q는 실제 연구자가 관심 있는 추정량을 의미하는데 대개 다변량 벡터(multivariate vector)인 경우가 많다.

다중대치의 세가지 주요 단계는 대치(imputation), 분석(analysis), 결합(pooling)이다. 컴퓨터 소프트웨어에서 각 단계를 특정 클래스 형태로 저장하는데 이는 mids, mira, mipo 클래스로 저장된다.

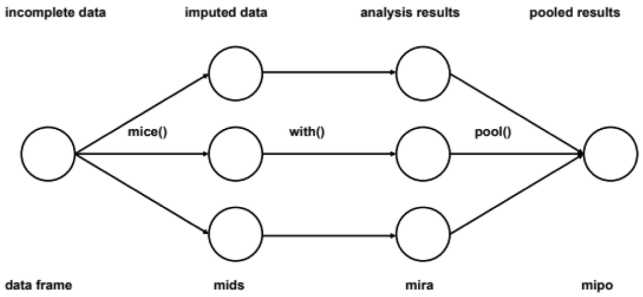

Fig. 10. Key steps of multiple imputation (Bauuren and Groothuis-Oudshoorn, 2011).

mids에서는 그럴듯한(plausible) 값으로 결측을 지정한 m개의 데이터 셋에 대해 대치해주고 mira는 각 대치 데이터 셋에서 Q를 추정하는 것이다. mipo에서는 Aemlia와 같이 Rubin's Rules를 적용하여 최종 추정 평균인 $\overline{Q}$ 와 그 분산을 추정한다.

MICE는 이러한 기반을 바탕으로 연쇄방정식 접근을 이용하여 변수별로 결측 대치 모형을 따로 지정하는 기법을 사용한다. 완전자료 Y가 모수 θ로 특정되는 p-변량 분포인 $P(Y|\theta)$에서 추출된 부분관측 자료라고 가정하고 θ의 다변량 분포를 암묵적으로 얻을 수 있다. MICE 알고리즘은 조건부 분포를 반복적으로 샘플링함으로써 θ의 사후분포를 도출한다.

$$P(Y_1|Y_{-1}, \theta_1), \ldots, P(Y_p|Y_{-p}, \theta_p)$$

여기서 $\theta_1,...,\theta_p$는 각각 조건부 밀도에 특화된 모수이며 반드시 실제 결합분포인 $P(Y|\theta)$의 인수분해로 도출되는 것은 아니다.

관측된 주변 분포로부터 단순하게 샘플링을 시작하여 t번째 반복에서 연쇄방정식은 Gibbs 샘플러를 이용한다.

$$\begin{aligned}
\theta_1^{*(t)} &\sim P(\theta_1|Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}) \\
Y_1^{*(t)} &\sim P(Y_1|Y_1^{obs}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_1^{*(t)}) \\
&\vdots \\
\theta_p^{*(t)} &\sim P(\theta_p|Y_p^{obs}, Y_1^{(t)}, \ldots, Y_{p-1}^{(t)}) \\
Y_p^{*(t)} &\sim P(Y_p|Y_p^{obs}, Y_1^{(t)}, \ldots, Y_p^{(t)}, \theta_1^{*(t)})
\end{aligned}$$

여기서 $Y_j^{(t)} = (T_j^{obs}, Y_j^{*(t)})$ 는 반복 t에서 j번째 대치 변수를 의미한다. 이전 반복의 대치값 $Y_j^{*(t-1)}$ 는 다른 변수들과의 관계를 통해서만 영향을 주므로 수렴이 상당히 빠르다는 특징이 있다. 일반적으로 반복횟수는 10-20 정도로 충분하며 MICE의 주요한 장점은 결합분포가 알려지지 않은 경우에도 모형을 지정할 수 있다는 점이다. MICE의 대표적인 회귀 모형으로는 PMM(Predictive Mean Matching) 방법이 있다. 이는 결측값이 있는 변수 Y_j를 선형회귀 등의 명시적 모형으로 예측값을 구한 후 관측 자료 중 예측값이 가장 가까운 n개의 케이스(기본적으로 5개)를 찾는다. 그 중 하나를 무작위로 선정하여 실제 관측된 값을 결측 값에 채워넣는 방법이다. PMM은 관측값을 가져오기 때문에 변수의 분포나 변동성을 매우 자연스럽게 유지할 수 있다. 따라서 본 연구에서는 R 언어의 MICE 패키지를 활용하여, 5개의 다중 대치 데이터셋을 만들고 연쇄방정식의 반복횟수는 10개로 지정하였다.