J. Korean Soc. Environ. Eng., 45(12), 625-642, 2023
https://doi.org/10.4491/KSEE.2023.45.12.625
ISSN 1225-5025, e-ISSN 2383-7810

Research Paper

Assessing the Impact of Temporal Resolution Using BSM1 on the Performance of Machine Learning

Wonki O[1] · SeoJin Ki[2,3] · Jin Mi Triolo[3] · Seung Gu Shin[3,4†]

[1]Digital Innovation Center, Korea Testing Laboratory, Republic of Korea
[2]Department of Environmental Engineering, Gyeongsang National University, Republi of Korea
[3]Future Convergence Technology Research Institute, Gyeongsang National University, Republic of Korea
[4]Department of Energy Engineering, Gyeongsang National University, Republic of Korea

(Received October 24, 2023; Revised December 14, 2023; Accepted December 15, 2023)

Objectives : This study aims to establish efficient strategies for data-driven operational management by examining the variations in machine learning modeling outcomes and data characteristics based on data acquisition intervals and methods.

Methods : The BSM1 was used to simulate wastewater treatment facilities and to generate influent and effluent water quality data at 15-minute intervals. The generated data was processed by volume reduction through down sampling and data characteristic observation via resampling techniques, including up sampling through interpolation. Subsequently, the study involved a comparative analysis of the performance of 30 machine learning models built with the down sampled data.

Results and Discussion : As data acquisition interval increased (i.e., down sampling progressed), R^2 decreased and RMSE increased. When using the mean value as a representation, data accuracy was high, and error loss was minimal. Utilizing the maximum value as a representation helped maintain data characteristics and reduce information loss. Simple interpolation methods did not yield improved data accuracy. Furthermore, with wider data acquisition intervals, the practical predictive performance of machine learning models decreased, and the models experienced a sharp decline in performance when data became insufficient.

Conclusion : For models requiring the ability to detect changes rather than accuracy, utilizing the maximum value over a specific period proves to be effective. The measurement interval of data emerges as a significant factor affecting the performance of machine learning models, with models developed under different measurement intervals often failing to demonstrate the expected performance. In this study, we have implemented all stages of data preprocessing, classification, training, and validation using LabVIEW, confirming the potential for integrating data analysis processes into LabVIEW, a widely used platform in the fields of control and measurement.

Keywords : BSM1, Temporal Resolution, LabVIEW, Machine Learning, Data Resampling

† Corresponding author
E-mail: sgshin@gnu.ac.kr
Tel: 055-772-3887 Fax: 055-772-3889

J. Korean Soc. Environ. Eng., 45(12), 625-642, 2023
https://doi.org/10.4491/KSEE.2023.45.12.625
ISSN 1225-5025, **e-ISSN** 2383-7810

연구논문

BSM1을 활용한 시간적 해상도가 기계학습모델의 성능에 미치는 영향에 대한 연구

오원기[1] · 기서진[2,3] · 박진미[3] · 신승구[3,4†]

[1]한국산업기술시험원 디지털혁신센터
[2]경상국립대학교 환경공학과
[3]경상국립대학교 미래융복합기술연구소
[4]경상국립대학교 에너지공학과

목적 : 데이터의 시간적 해상도의 변화에 따른 데이터의 특성 변화를 확인하고, 데이터의 시간적 해상도가 기계학습 모델의 성능에 미치는 영향을 확인한다.

방법 : 하수처리시설에 대하여 BSM1으로 시뮬레이션하여 활용해 15분 간격 한 유입수 및 유출수 수질데이터를 확보하였다. 데이터를 n개씩 묶는 down sampling으로 데이터 수를 줄이고, 보간법으로 데이터 늘리는 up sampling의 resampling으로 데이터의 특성 변화를 확인하고, down sampling된 데이터로 구축한 30개의 기계학습 모델들의 성능을 상호 비교하였다.

결과 및 토의 : down sampling을 진행할 수록 R^2는 낮아지고, RMSE는 증가하였다. 평균값을 대푯값으로 할 때 데이터의 정확성과 오차의 손실이 적고, 최대값을 대푯값으로 할 때 데이터 특성을 유지하여 정보의 손실을 감소시킬 수 있었다. 시간적 해상도가 다른 모델의 성능을 비교하기 위해서는 같은 수준의 데이터를 적용하여 모델간 성능을 비교할 필요가 있었다.

결론 : 사고경보 등 정확성보다는 변화를 감지하는 능력이 필요한 모델은 일정기간에 대한 최대값을 대푯값으로 활용하는 것이 효과적이다. 데이터의 측정 간격은 기계학습 모델의 성능에 영향을 주요 인자로 측정 간격이 다른 상태로 개발된 기계학습 모델이 제시된 성능을 발휘하지 못하는 주요 원인이 된다. 본 연구에서는 데이터의 전처리, 분류, 학습 및 검증의 모든 과정을 LabVIEW로 구현하여 향후 제어, 계측 분야에 널리 활용되고 있는 LabVIEW에 데이터 분석 과정이 포함된 통합개발환경의 구현의 가능성을 확인하였다.

주제어 : BSM1, LabVIEW, 기계학습, 데이터 재표본화

1. 서 론

하수처리시설은 가정에서 발생하는 생활하수나 공장이나 사업장의 폐수를 처리하여 수질을 개선하는 환경기초시설이다. 환경오염 사고를 사전에 예방하고, 수자원의 관리와 공중보건에 중요한 역할을 담당하는 동시에 대표적인 에너지 다소비시설로 분류되어 효율적이고, 경제적인 운영과 관리가 필요하다.[1,2,3] 수질원격감시체계(Tele-Monitoring System, TMS)를 통해 처리수의 수질을 실시간으로 수집 관리하고 있으며[4], 하수처리시설의 데이터는 일단위 데이터를 1개월 단위로 공공데이터포털(data.go.kr)에서 받아볼 수 있다. 수질원격감시체계의 경우 센서의 응답을 실시간으로 받는 센서형 측정기(SS, pH)의 데이터와 시료의 전처리 및 분석과정을 자동화하여 15분에서 60분의 간격으로 측정하는 분석형 측정기(TOC, COD, TN, TP 등)의 데이터를 5분 단위로 실시간 저장된 자료로 실시간으로 수집된 데이터 역시 데이터간 간격이 존재하는 이산적인(discrete) 데이터이다. 따라서 데이터 기반 연구를 수행할 때 분석대상인 데이터는 현실적 또는 기술적 원인으로 각 데이터간 시간적 해상도(temporal resolution)가 다르다.

샘플링 주기로 설명할 수 있는 데이터의 시간적 해상도는 데이터 기반 모델의 성능에 영향을 미친다.[5] 데이터의 시간적 해상도가 낮은 경우 변화된 환경을 제대로 반영할 수 없고, 수집환경과 방식에 따른 편향이 발생하게 된다.[6] 데이터의 시간적 해상도가 낮을수록 데이터 마이닝이나 기계학습의 성능이 저하되며, 높은 시간적 해상도의 데이터는 보다 자세하고 세분화된 통찰력을 제공하여 모델의 예측성능을 제시할 수 있다.[7] 데이터의 시간적 해상도를 높이기 위해서는 분석형 자동측정기의 시약 및 운영 비용이 증가하며[8], 분석대상 데이터

량이 증가하여 데이터의 처리 및 계산이 복잡해진다. 또한 다양한 변수가 서로 다른 시간간격으로 측정된 다변량시계열(Multi scale time series)에서는 데이터의 복잡성과 상호작용으로 새로운 특징과 패턴을 제시할 수 있다.[6] 데이터의 시간적 해상도를 향상시키는 것은 분석모델의 정확성과 비용의 trade off 관계인 것이다.[9]

데이터의 시간적 해상도에 대한 연구는 다양한 방식으로 진행되었다. 1분, 5분, 10분의 시간간격으로 혼합된 데이터에 대하여 시계열 분석을 통해 측정주기 단위로 데이터를 분해하여 혼합이전 데이터들의 특성을 파악하는 연구가 있었으며[10], 배수관망에서 급수관망에서 물소비량을 다른 수준의 시간적 해상도의 유량계로 예측하는 연구가 수행되었다.[11] 또한 센서형 수질데이터를 활용해 시간적 해상도가 낮은 측정 데이터를 보정하여 정확도를 향상시키는 것보다 시료채취 횟수를 증가시키는 것이 모델의 성능 향상에 도움이 된다는 연구의 결과도 있다.[12] 해당 연구들은 시간적 해상도가 다르게 수집된 자료를 기반으로 수행된 연구이다. 높은 수준의 시간적 해상도를 변화시킬 때 데이터의 특성 변화를 관찰하고, 해당 데이터로 기계학습을 수행하였을 때 생성되는 모델에 미치는 영향을 직접 확인하는 연구가 필요하다.

본 연구에서는 수처리 분야로 제한하여 데이터의 시간적 해상도에 따른 데이터의 특성 변화에 집중하고자 한다. IWA Task Group의 BSM1(Benchmark Simulation Model)[13]로 일정한 시간 해상도의 하수처리시설의 유입수 및 유출수 데이터를 확보하였다. 해당 데이터의 취득 주기를 감소시켜 시간당 데이터 수를 줄이거나(down sampling)[14], 반대로 측정 주기를 증가시켜 시간당 데이터 수를 늘리는(up sampling)는 데이터 재표본화(data resampling)[10,15]과정에서 데이터의 손실되는 정보를 시각화하여 관찰하였다. 이어서 다른 수준의 시간적 해상도로 기계학습을 수행할 때 학습된 기계학습 모델들의 성능변화를 확인하였다. 이때 다양한 기계학습 모델간의 특성과 성능에 따른 영향을 최소화하기 위하여 LabVIEW환경에서 제어할 수 있는 단순한 형태의 기계학습 모델을 적용하였다.

데이터 취득 이후 전처리, 기계학습 및 시각화의 전체 과정을 LabVIEW로 개발한 프로그램을 활용하였다. 그래픽기반의 개발환경인 LabVIEW의 이미지 형태로 구현된 코드는 프로그래밍에 대한 높은 접근성을 제공한다. LabVIEW는 각종 센서 데이터를 실시간으로 수집하고, 제어하는데 유용하여 하드웨어와 소프트웨어간 통합 모니터링 및 제어 계측분야에 널리 활용되고 있다.[16,17] LabVIEW의 데이터 구조인 map을 활용하면 key value 형태로 데이터를 관리하면서 클러스터 형태로 정리된 데이터 내부에 새로운 데이터셋을 계층화해 저장하고 호출할 수 있다.[18] 일정한 주기로 수집된 데이터를 down sampling하여 새로운 데이터셋을 구성하고, 다시 up sampling으로 생성된 데이터셋에서 필요한 데이터를 호출해 기계학습을 수행하는 프로그램을 LabVIEW로 구현하였다.

본 연구는 데이터의 시간 해상도에 따른 데이터의 특성 변화를 확인하고, 해당 데이터가 기계학습 성능에 미치는 영향을 확인하는 연구이다. 우선 데이터의 전처리와 재표본화 과정을 LabVIEW로 구현해 계층화된 데이터셋을 준비하는 과정을 자동화하였으며, 이후의 기계학습에서 LabVIEW 기반의 단순한 기계학습 모델을 적용해 준비된 down sampling 데이터셋과 연계하여 기계학습 모델들 간의 다양한 특성과 성능에 의한 영향을 최소화하였다.

2. 실험방법

2.1. 원시데이터의 획득 및 탐색적 데이터 분석

BSM1은 유입수의 특성, 처리조의 설계요소 및 운영조건, 활성슬러지의 특성 등을 고려해 처리시설의 효율 예측 및 운

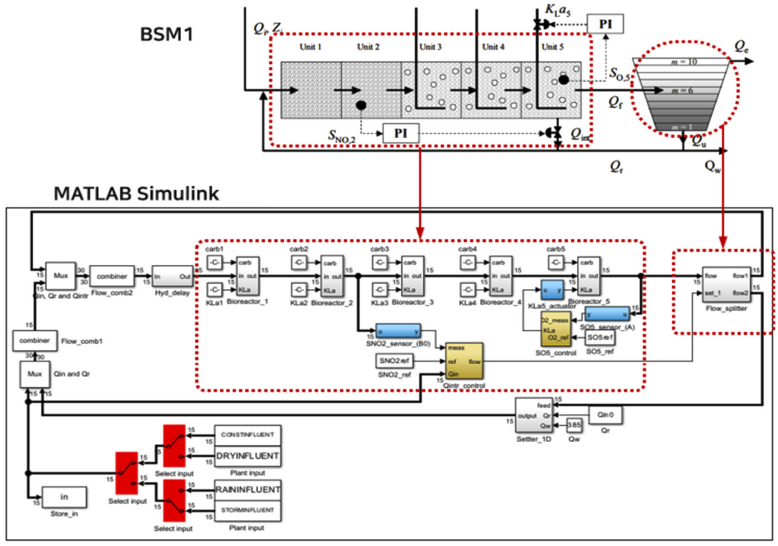


Fig. 1. Schematic overview of the BSM1 and the diagram of BSM1 embedded in MATLAB Simulink.

Table 1. List of input variables for BSM1.

Definition	Notation	Unit
Influent	Q*	$m^3.sec^{-1}$
Soluble inert organic matter	SI	g COD.m^{-3}
Readily biodegradable substrate	SS*	g COD.m^{-3}
Particulate inert organic matter	XI*	g COD.m^{-3}
Slowly biodegradable substrate	XS*	g COD.m^{-3}
Active heterotrophic biomass	XBH*	g COD.m^{-3}
Active autotrophic biomass	XBA	g COD.m^{-3}
Particulate products arising from biomass decay	XP	g COD.m^{-3}
Oxygen	SO	g(COD).m^{-3}
Nitrate and nitrite nitrogen	SNO	g N.m^{-3}
NH^{4+}+NH^3 nitrogen	SNH*	g N.m^{-3}
Soluble biodegradable organic nitrogen	SND	g N.m^{-3}
Particulate biodegradable organic nitrogen	XND	g N.m^{-3}
Alkalinity	SALK	mole.m^{-3}

*: used variable for predicting SS concentration of effluent.

영의 최적화를 위한 의사결정 지원도구로 널리 활용되고 있다. 유입폐수와 활성슬러지를 혼합하여 처리하는 무산소조 2개와 3개의 호기성조로 구성된 연속된 프로세스를 통해 유기성물질이 생물학적으로 처리되는 과정을 수학적으로 모델링한다(**Fig. 1**). 유입 유량(Q)을 포함한 14개 입력 변수는(**Table 1**)는 Dry, Rain(비), Storm(강우)의 세가지 시나리오로 15분 간격으로 총 2주간의 데이터이다.[19] 그 중 첫 주간의 데이터는 모델의 안정화를 위한 데이터이며, 다음 2주간 데이터가 시나리오별 특성이 반영된 데이터이다(**Fig. 2**). Rain 시나리오는 8일째부터 이틀간 연속해서 유입 유량이 증가하고, Storm 시나리오는 8일과 10일째 두 번의 피크유량을 확인할 수 있으며 Rain 시나리오에 비해 많은 60,330 m^3/d의 유량이 유입된다.

본 연구에서는 BSM1의 세가지 시나리오에 대하여 MATLAB (R2022b 버전)을 활용해 15분 간격의 2주간 데이터를 취득하였다. 각 시나리오별 입력데이터로 다른 시뮬레이션 결과를 얻을 수 있는데 다양한 패턴의 데이터 변동을 포함시키기 위해 Dry, Rain, Storm의 안정화 기간을 제외한 2 주째 데이터를 이어 붙인 총 3주간(2,106)의 입력과 출력변수 데이터를 분석에 활용하였다. BSM1 모델을 활용하면 하수처리시설 내 각 처리조와 반송 슬러지의 농도 및 유량변화를 15분 간격으로 확인할 수 있다. 최종침전조에서 유출수의 SS 농도(OutSS)를 목적변수(Y)로 선정하고, 목적변수와 입력변수들 간의 상관관계를 피어슨 상관계수와 R^2로 확인하였다(**Table 2**). 14개의 입력변수 중 SO, SNO, XP, XND의 경우 유출수 SS농도와 상관성을 확인할 수 없었다. 유입수 유량(Q)은 XS는 Dry, Rain, Storm의 모든 조건에서 양의 상관성을 나타냈고, SS, SNH, XI, XBH, XBA는 Dry 시나리오에서 양의 상관관계를, Rain 시나리오에서는 음의 상관성을 보였다. 한편 XI는 Rain 조건에서만 양의 상관성을 나타냈다. Q와 SI를 제외한 나머지 변수의 상관계수는 Dry 시나리오에서 0.72에서 0.96까지 비교적 높은 값을 유지하였지만, Rain과 Storm 시나리오에서 크게 낮아졌다.

BSM1의 입력변수들는 주어진 시나리오별로 유출수의 SS에 대해 서로 다른 상관성 특성을 나타냈다. 입력변수와 유출수의 SS농도에 대한 R^2를 계산하였을 때 유입수의 XS의 경우 Dry 시나리오에서 0.921로 높았지만, Rain과 Storm 시나리오에서는 0.002와 0.167로 크게 낮아졌다. SI를 제외한 다른 입력변수들의 R^2 역시 Dry 시나리오에서 0.518 보다 컸지만, Rain과 Storm 시나리오에서는 낮아졌다. 따라서 BSM1의 입력변수인 Q, SS, SNH, XI, XS, XBH 및 XBA는 날씨조건의 시나리오별로 유출수의 SS에 대해 서로 다른 형태의 상관관

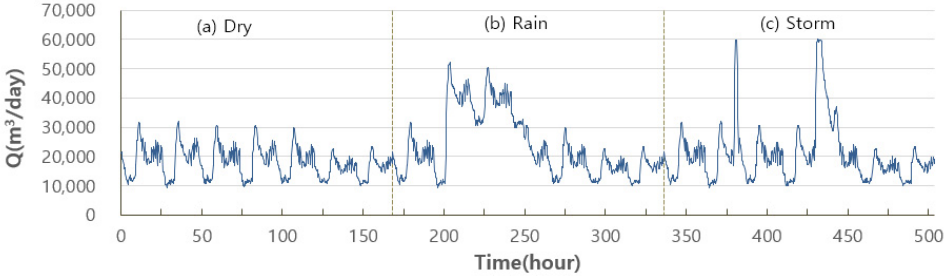


Fig. 2. Temporal variation in influent quantity(Q) of the BSM1.

Table 2. Pearson correlation coefficients between effluent SS concentration (target variable) and input variables (predictor variables).

		Q	SI	SS	SNH	XI	XS	XBH	XBA
Dry	Corr.	0.84	NaN	0.83	0.72	0.83	0.96	0.813	0.878
	R^2	0.705	NaN	0.689	0.518	0.689	0.921	0.661	0.772
Rain	Corr.	0.965	-0.829	-0.291	-0.406	-0.291	0.0479	-0.315	-0.228
	R^2	0.931	0.688	0.0847	0.165	0.0847	0.00229	0.0995	0.0518
Storm	Corr.	0.84	-0.821	-0.118	-0.234	-0.118	0.409	0.0131	0.159
	R^2	0.705	0.675	0.014	0.0546	0.014	0.167	0.000171	0.0253

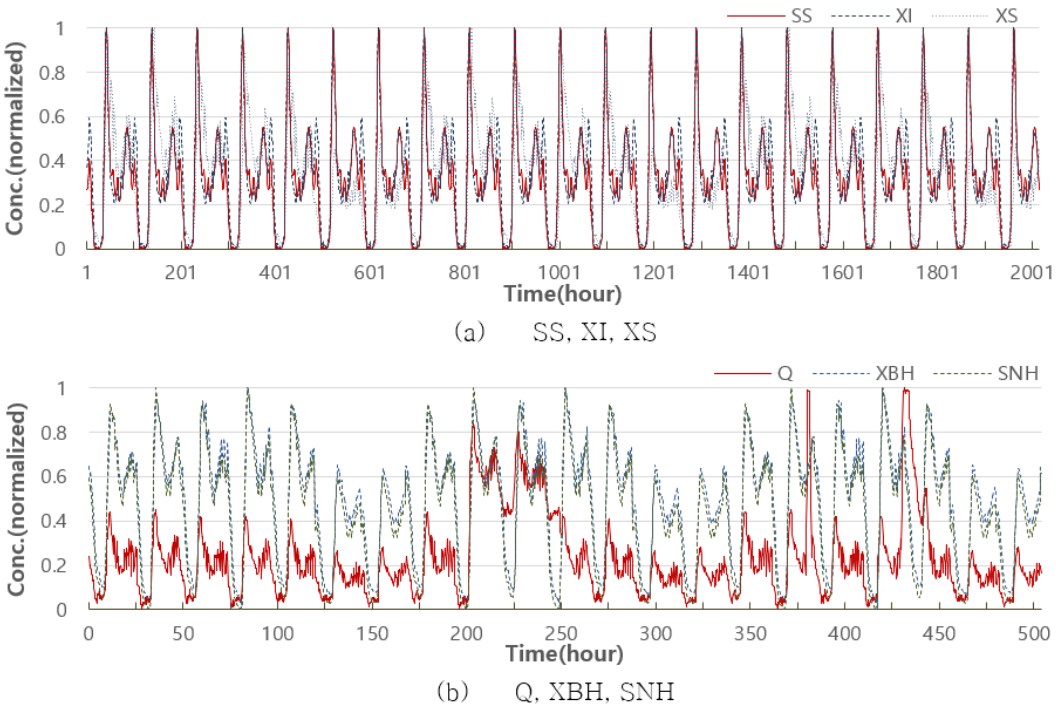


Fig. 3. Time-varying characteristics of BSM1 input variables used for BPN machine learning: (a) SS, XI, XS and (b) Q, XBH, SNH.

계와 설명력을 지니며, 입력변수 중에서 유입수 유량(Q)이 주요 변수임을 확인할 수 있다.

본 연구에서는 Dry 조건에서 유출수의 SS 농도와의 상관관계 연산에서 NaN 오류를 발생시킨 SI와 LabVIEW의 기계학습(BP Learn) 연산에서 NaN 오류를 발생시킨 XBA를 제외한 나머지 변수를(Q, SS, SNH, XI, XS, XBH) 기계학습의 입력변수로 정하였다. LabVIEW BP Learn은 역전파(back propagation)에 기반을 둔 모델로 학습률, 데이터의 스케일 또는 가중치 소실 등의 원인으로 학습과정 중에 NaN 오류를 발생시킬 수 있다. 시간 해상도의 변화에 따른 기계학습 모델의 성능 변화를 확인하는 본 연구에서는 NaN 오류의 원인을 확인하지 않고 입력변수 목록에서 제외하였다. 유출수의 SS 농도를 예측을 위한 기계학습모델의 입력 변수(Q, SS, SNH, XI, XS, XBH)에 대하여 시간에 따른 변화를 패턴별로 구분해 **Fig. 3**에 표시하였다. 15분 간격의 2016개 입력변수는 이후의 데이터 분석을 위해 Min-Max 정규화로 전처리하였다.

2.2. 데이터 재표본화 프로그램 개발

2.2.1. 프로그램 개발 개요

LabVIEW 기반 라이브러리와 드라이버를 활용하면 다양한 하드웨어 장치를 연계하여 각종 센서의 데이터를 실시간으로 수집하고, 제어할 수 있다.[17),16)] 데이터 흐름을 wire로 연결하는 방식의 그래픽 기반인 LabVIEW는 초보자도 쉽게 접근할 수 있는 친근한 개발환경을 제공한다. 본 연구에서는 무료로 사용할 수 있는 LabVIEW 2023 Q1 Community 버전을 활용해 BSM1 시뮬레이션 데이터를 재표본화(Resampling)하여 기계학습모델을 적용하는 프로그램을 개발하였다. 다만, LabVIEW의 경우 오픈소스 기반의 라이브러리와 패키지를 활용할 수 있는 R이나 Python에 비해 기계학습 모델의 개발분야에 활용도가 다소 낮기 때문에 R 기반 기계학습 모델과의 성능을 비교하는 과정을 연구에 포함시켰다.

데이터의 샘플링 주기를 의미하는 시간적 해상도를 변화시켜 하수처리시설의 데이터 정보가 변하는 것을 확인하는 것이 본 연구의 주요 목적이다. 데이터 resampling은 데이터의 샘플링 주기를 변경하는 과정을 모사하는 프로세스로 전체 데이터셋을 n개씩 묶어 ① 평균값, ② 최대값, ③ 최소값의 하나의 대푯값으로 표현하는 down sampling과 역으로 측정 간격을 나누고 선형보간법으로 데이터를 채워 넣는 up sampling 과정을 포함한다.[10)] 이미지의 픽셀 수를 줄여 작은 이미지를 생성하고(down sampling), 반대로 이미지 크기를 늘리기 위해 픽셀 수를 증가시켜(up sampling) 큰 이미지 자료를 얻는 것과 같은 과정이다. 측정 간격이 15분인 BSM1 모델의 변수로 down sampling을 진행하면 데이터의 측정 간격은 15 x n 분으로 증가하고, 데이터의 수는 1/n으로 줄어든다. 본 연구에서는 n을 30까지 늘려가면서 down sampling을 수행하여 15분에서 최대 450분 간격의 총 30개의 down sampling 데이터셋을 구분하여 저장하였다(**Fig. 4**).

Down sampling된 데이터셋은 역으로 N개씩 나누는 up sampling과정을 수행한다. 이 과정에서 LabVIEW의 map 구조를 활용하면 down sampling한 데이터셋이 존재하는 클러스터 내에 up sampling 데이터셋을 저장하여 계층화된 구조로

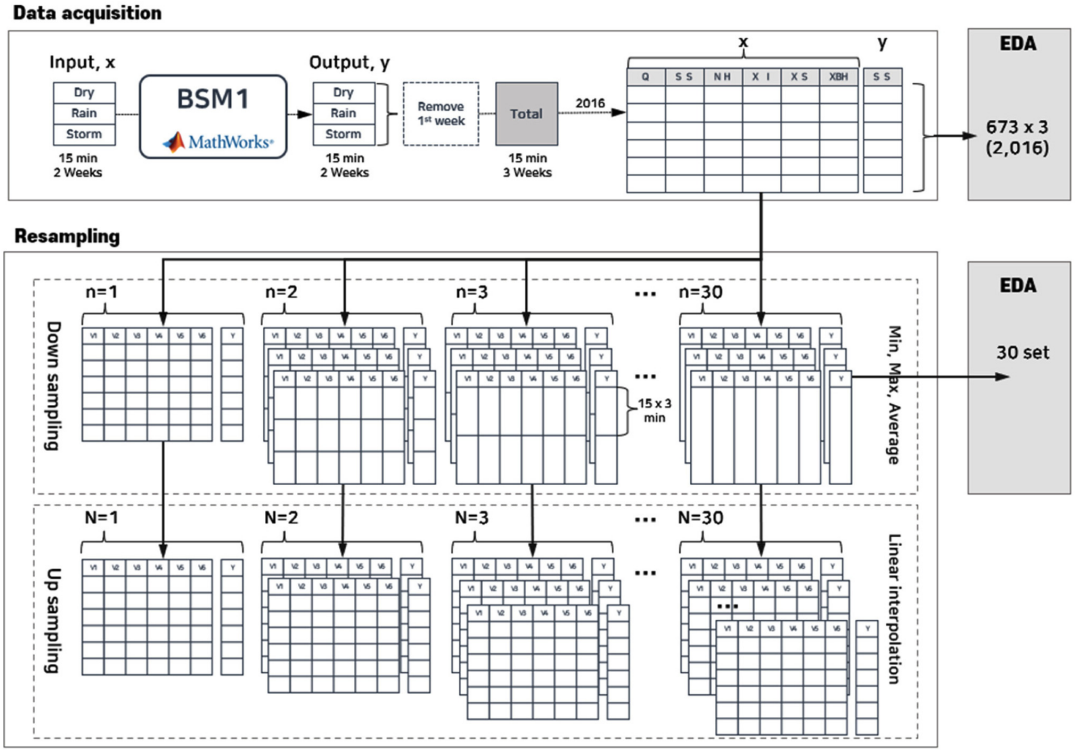


Fig. 4. Data acquisition and resampling process (down sampling and up sampling).

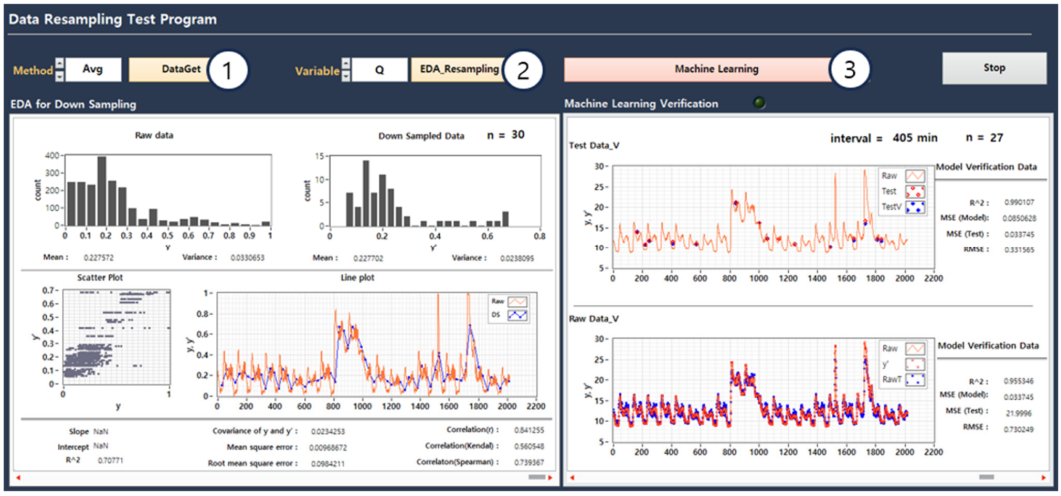


Fig. 5. User interface of the data analysis program developed for the study.

데이터를 체계화할 수 있다. Down sampling한 30개의 데이터셋에 해당 단계의 up sampling 데이터셋을 함께 관리하여 필요시 바로 호출하여 분석에 활용할 수 있도록 구성하였다.

2.2.2. LabVIEW 기반 데이터 분석 프로그램

최종적으로 개발된 데이터 분석 프로그램의 사용자인터페이스(user interface) 화면은 **Fig. 5**와 같다. 해당 화면에서 Min, Max, Average의 선택자를 선택하면 선택된 방법으로 down sampling을 수행하여 30개의 데이터셋을 만들고, 연속해서 up sampling을 수행하여 계층화된 데이터를 저장한다. 입력변수(Q, SS, SNH, XI, XS, XBH)를 선택하여 데이터분석(EDA_Reporting) 버튼을 활성화하면 Down sampling된 변수에 대해 최초데이터(n=1)와의 직선성, 상관계수를 표시하고, 데이터 분포에 대한 히스토그램을 상호 비교할 수 있도록 시각화하여 제시한다. 프로그램의 시각화를 위한 프로그램 코드는 **Fig. 6**와 같다.

기계학습(Machine Learning) 버튼을 활성화시키면 down sampling된 30개의 데이터셋으로 각각의 기계학습을 수행하고 검증한다. 각 단계의 down sampling 데이터를 훈련데이터(8)와 검증 데이터(2)로 나누어 단계별 훈련데이터를 기반으로

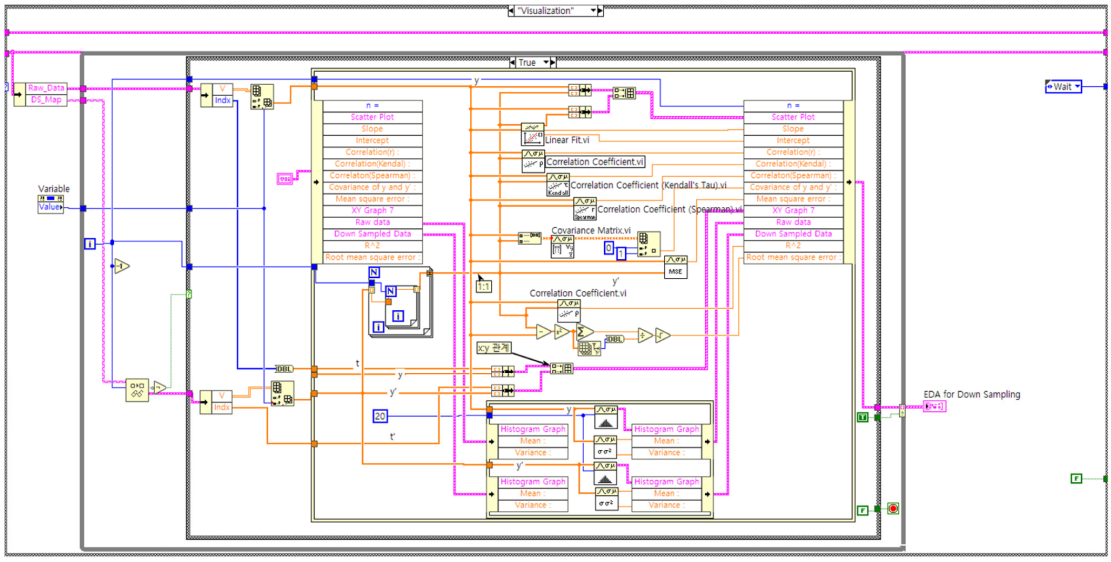


Fig. 6. Block diagram of LabVIEW to data visualization.

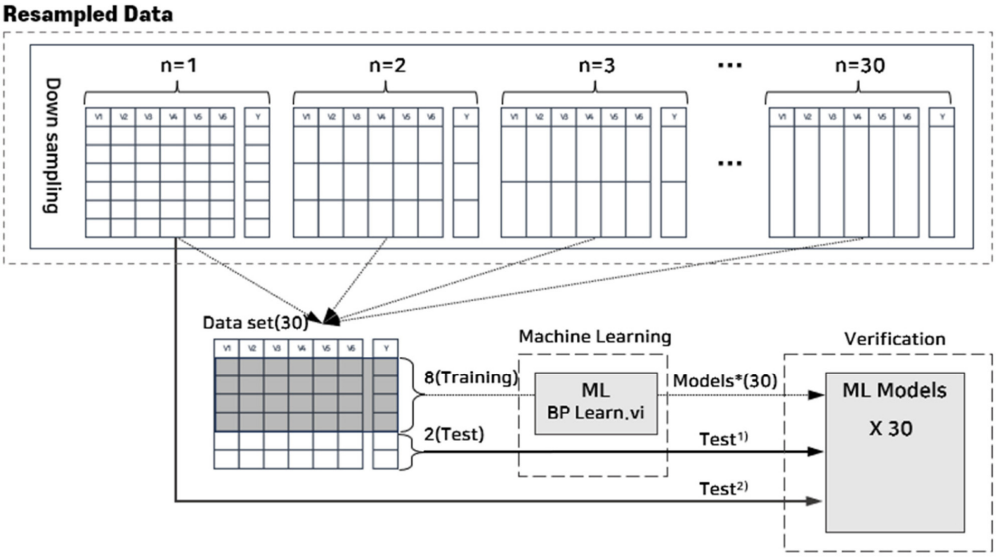


Fig. 7. Application of machine learning using down sampled data set in sequential steps.

하는 학습된 모델을 생성한다(**Fig. 7**). 데이터 down sampling 데이터의 정보가 변한다면 동일한 아키텍처를 가진 모델이지만 각 단계의 훈련데이터로 학습된 패턴을 기반으로 서로 다른 가중치의 새로운 학습모델이 생성될 것이다. 생성된 30개의 학습모델을 각 단계의 검증데이터로 모델의 성능을 평가하는 동시에 최초데이터(n=1)를 학습모델에 적용하여 단계별 학습모델들간 성능을 상호 비교할 수 있도록 하였다. 단계별 검증데이터의 결과(R^2, RMSE)를 화면에 표시하고, 시계열로 표현된 최초데이터 그래프에 검증데이터의 SS 농도(Y)와 학습모델의 예측값(Y')을 함께 시각화하였다. 본 연구에서 개발된 프로그램은 Windows 11 Pro, 16GB RAM의 시스템 환경에서 down sampling은 0.267 초, up sampling은 0.345 초, 기계학습에 10.197 초가 소요되었다.

2.3. 모델 평가 지표

Down sampling 데이터셋으로 생성된 학습 모델의 성능을 RMSE와 R^2로 평가하였다. RMSE(Root Mean Squared Error)는 회귀 분석이나 예측 모델의 성능을 측정하는 데 사용되는 평가 지표 중 하나로 모델이 예측한 값(Y')과 실제 관측값(Y) 간의 차이로 아래의 수식으로 산정한다.

R^2, 결정 계수(coefficient of determination)는 회귀 모델의 적합도(설명력)를 평가하는 지표 로 모델이 주어진 데이터에 얼마나 잘 적합되었는지를 나타낸다. 0과 1 사이의 범위에서 클수록 모델이 데이터를 잘 설명한다고 할 수 있다.

입력변수(x)와 목적변수(y)의 선형 상관관계를 피어슨 상관계수로(Pearson correlation coefficient)로 확인하였다. 피어슨 상관계수는 -1에서 1 사이의 값으로 0에 가까우면 상관관계가

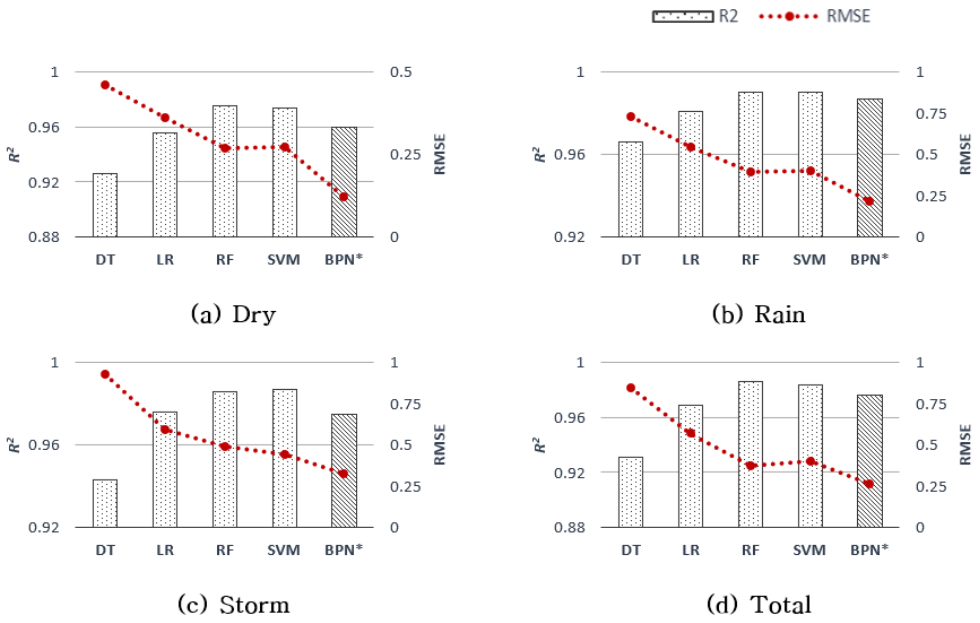


Fig. 8. Comparison of Machine Learning Results for BPN (LabVIEW) and DT, LR, RF, SVM(R) Under Dry, Rain, Storm, and Total Conditions.

없고, 1에 가까우면 양의 상관관계를 -1에 가까우면 음의 상관관계가 있음을 확인할 수 있다.

3. 결과 및 고찰

3.1. LabVIEW를 활용한 기계학습 모델의 성능 확인

본 연구에서는 데이터의 재표본화 이후의 기계학습 모델의 성능변화를 확인하는 과정에서 LabVIEW 기반 기계학습 모델을 적용하였다. LabVIEW 기반의 기계학습 모델을 적용하면 데이터의 재표본화 작업부터 기계학습 및 시각화까지 데이터의 형변환이나 연계를 위한 추가 작업 없이 프로그램을 개발할 수 있다. LabVIEW 라이브러리 중에서 Machine learning toolkit은 다양한 비지도학습 모델과 지도학습 모델을 포함하고 있는데 그 중 값을 예측하는 회귀모델은 역전파신경경망(Back propagation neural network)이 유일하다. Back propagation neural network(BPN)로 유출수의 SS 농도를 예측할 때 해당 모델의 적용 타당성을 확인하기 위하여 R 기반 기계학습 모델들과 성능을 상호 비교하였다. 구체적으로 4.2.3 버전의 R에 rpart library, random Forest library, e1071 library를 활용하여 Decision Tree(DT), Linear Regression(LR), Random Forest(RF), Support Vector Machine(SVM)으로 기계학습을 수행하였다. 기계학습에는 BSM1의 (a)Dry, (b)Rain, (c)Storm 날씨 시나리오와 Dry, Rain, Storm의 안정화 기간(1주)을 제외하여 2주째 데이터를 이어 붙인 (d)Total condition의 시뮬레이션 데이터를 활용하였다. 해당 데이터는 학습데이터(8)과 검증데이터(2)로 무작위 분류하였으며 이때 seed 값을 1로 설정하여 동일한 결과를 도출하도록 유도하였다.

LabVIEW의 BPN과 R의 DT, LR, RF, SVM의 성능을 비교하였다(**Fig. 8**). R^2로 확인한 모델의 정확성은 RF > SVM > BPN(LabVIEW) > LR > DT 순으로 LabVIEW의 BPN은 다섯 모델 중 중간 수준의 정확도를 나타냈다. RMSE로 확인한 예측오차는 DT > LR > SVM > RF > BPN(LabVIEW)의 순서로 낮아져 LabVIEW의 BPN을 적용하였을 때 오차가 가장 적었다. 이에 LabVIEW 기반의 BPN으로 기계학습을 수행하는데 문제가 없을 것으로 판단하고, 시간 해상도 변화에 따른 성능 변화를 확인하기 위한 기계학습 모델로 BPN으로 정하였다. LabVIEW기반 BPN을 적용하면 프로그램 내에서 데이터의 연계성이 높아져 오류발생을 줄일 수 있으며 기계학습 모델을 한가지 모델로 제한해 모델의 특성으로 인한 연구의 영향을 최소화할 것으로 판단하였다.

3.2. 데이터 재표본화에 따른 특성 변화

BSM1의 입력변수를 Dry, Rain, Storm의 안정화 기간을 제외하고 2주째 데이터를 이어 붙인 3주간의 데이터(Total)를 활용해 데이터의 재표본화에 따른 특성 변화를 관찰하였다. 우선 n을 30까지 늘리는 down sampling과정에서 n=1인 최초데이터를 기준으로 n을 증가시킬 때 산포도 그래프와 시계열 그래프로 데이터 분포의 특성 변화를 확인하고, 해당 데이터의 시간 해상도를 N까지 늘리는 up sampling을 수행할 때 데이터 변화를 관찰하였다.

3.2.1. Down sampling

BSM1 모델의 입력변수는 3가지 패턴으로 구분된다. 유입유량(Q)는 Dry, Rain, Storm의 날씨 시나리오의 특성이 직접

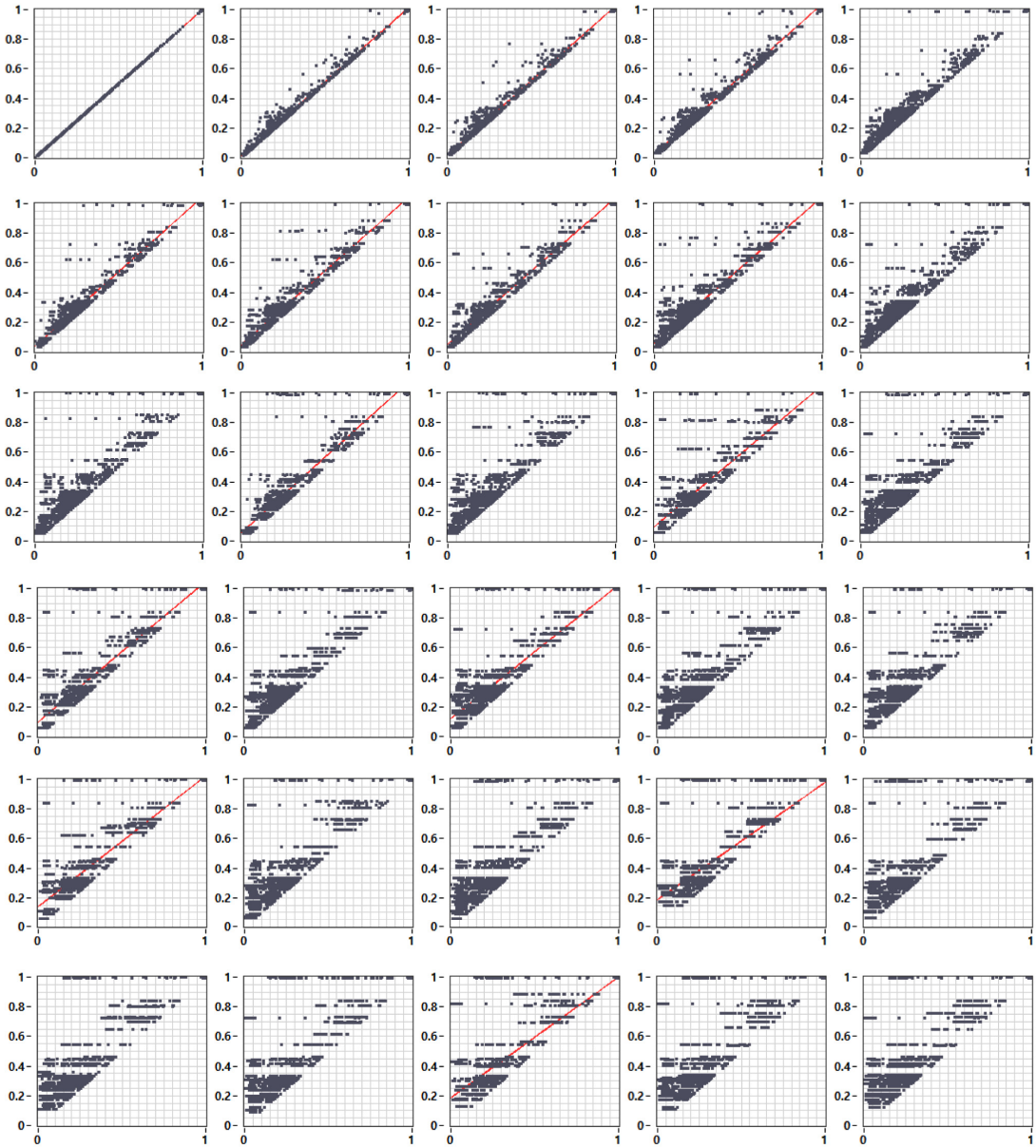


Fig. 9. Distribution characteristics change of influent quantity (Q) down sampled to the maximum value as n increases to 30(normalized by Min-Max).

적으로 반영된 변수이다(**Fig. 2**). XBH, SNH는(**Fig. 3(b)**)는 Q의 패턴이 반영된 형태이며, SS, XI, XS(**Fig. 3(a)**)는 날씨 및 Q와 관계없이 3주간 21번의 일정한 펄스로 반복되는 시계열 특성을 나타낸다. 우선 날씨 시나리오의 특성을 확인할 수 있는 Q에 대하여 down sampling을 수행하면서 평균, 최대, 최소의 방법으로 대푯값을 정한 산포도 그래프는 각각 **Fig. 9**, **Fig. 10**, **Fig. 11**과 같다. Down sampling으로 n이 증가할수록(시간해상도가 낮아질수록) 데이터의 횡방향으로 분포가 넓어지고, 데이터의 연속성이 단절되어 구간별로 데이터로 구분되는 현상을 확인할 수 있다. 최대값으로 데이터를 취하는 경우 1:1 기준선의 왼쪽편에 데이터가 분포하고, 그래프상에 붉은색으로 표현된 선형회귀선의 y축 절편은 증가하였다. 평균값으로 데이터를 취할 때 1:1 기준선의 양편으로 데이터가 퍼져 있으며, 선형회귀선은 오른쪽으로 기울고 y축의 절편이 상승하였다. 최소값으로 데이터를 취하면 1:1 기준선의 오른쪽 편에 데이터가 분포하면서, 선형회귀선의 기울기는 오른쪽으로 더욱 기울면서 y축 절편은 증가하지 않았다.

유입 유량(Q)에 대하여 down sampling을 수행하면서 평균, 최대, 최소의 방법으로 대푯값을 적용하였을 때 n이 증가시킬 때 단계별로 데이터 해상도가 낮아진 상태의 Q'가 새로 생성된다. 각 단계에서의 Q와 Q'의 평균, R^2와 RMSE에 대한 연속적인 변화는 **Fig. 12**와 같다. 대푯값을 평균값으로 선정하면 n이 증가시켜도 단계별 평균값은 변동이 없었지만(0.23), 최대값인 경우 단계별 평균은 지속적으로 증가하다가 n=30일

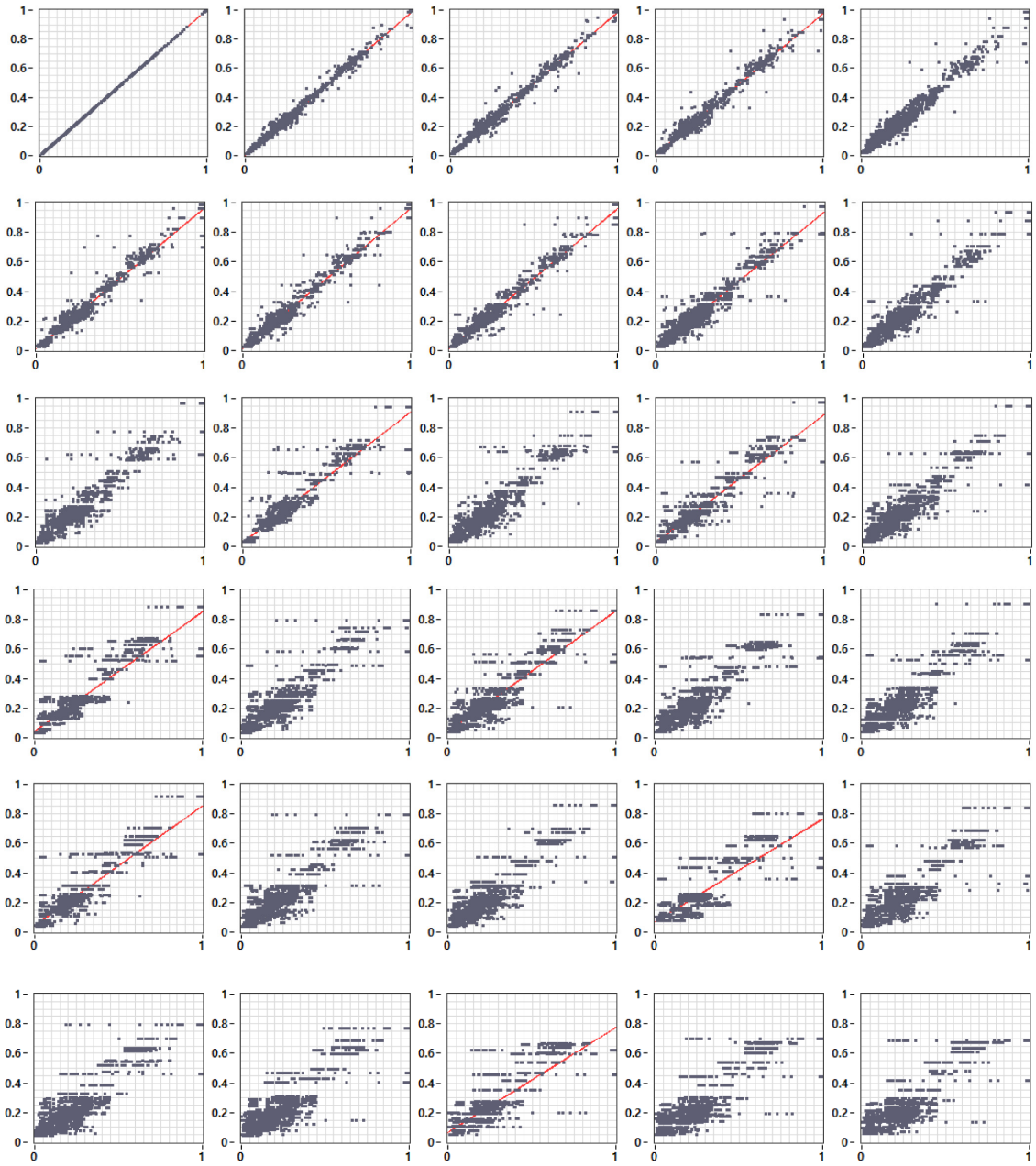


Fig. 10. Distribution characteristics change of influent quantity (Q) down sampled to the average value as n increases to 30(normalized by Min-Max).

때 0.38이 되었고, 최소값일 때는 반대로 0.12까지 감소하였다. n이 증가할수록 새로 생성되는 평균값(Q')은 더 넓은 구간에 대한 시간 해상도 구간을 대표하게 된다. Q의 분포에서 낮은 범위(0.5 이하)에 존재하는 데이터가 0.5 이상의 데이터보다 양적으로 많이 분포하는 것을 **Fig. 9**, **Fig. 10**, **Fig. 11**에서 확인할 수 있는데 이에 따라 전체 평균이 일정하게 유지하려면 n이 증가할 수록 선형회귀선의 기울기는 오른쪽으로 기울고 y 절편이 커지게 된다고 판단할 수 있다. 같은 이유로 최대값으로 데이터를 취하면 선형회귀선의 y축 절편은 크게 증가하지만 최소값인 경우 오른쪽으로 더욱 기울고 y 절편의 증가하지 않는 결과를 설명할 수 있을 것이다. 또한 n을 30까지 평균값을 대푯값으로 데이터를 취하면 R^2는 최종적으로 0.708까지 낮아졌지만 최대값(0.534)과 최소값(0.561)에 비하여 감소폭이 작았다. RMSE는 평균값일 때 0.0984까지 커졌지만 최대값(0.203)과 최소값(0.161)으로 데이터를 취할 때보다 낮았다. 따라서 데이터의 정확성과 오차를 우선적으로 고려한다면 일정 구간의 시간 해상도에 대하여 평균값으로 데이터를 취하면서 down sampling을 수행하는 것이 옳은 선택일 수 있다.

Down sampling에 따른 다른 입력변수들의 변화패턴을 확

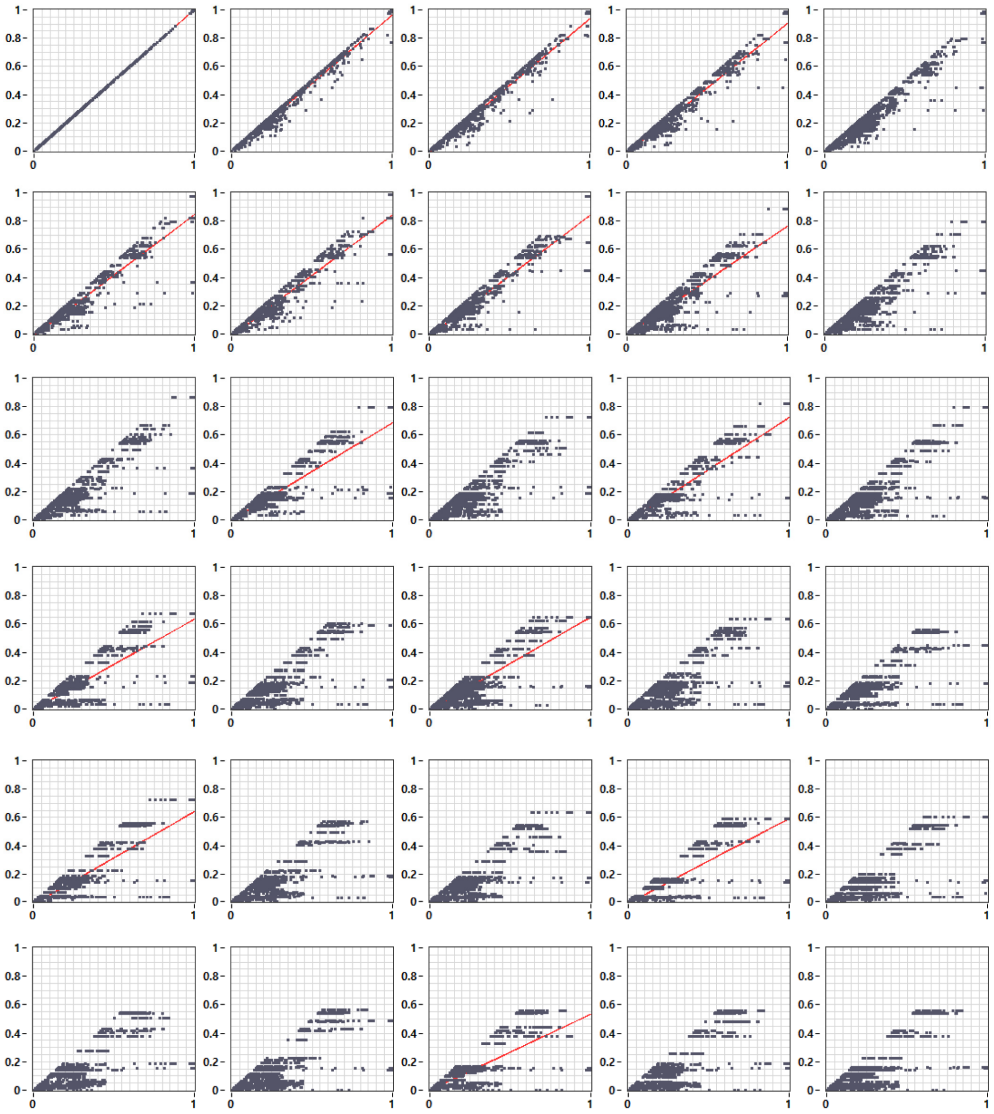


Fig. 11. Distribution characteristics change of influent quantity (Q) down sampled to the minimum value as n increases to 30(normalized by Min-Max).

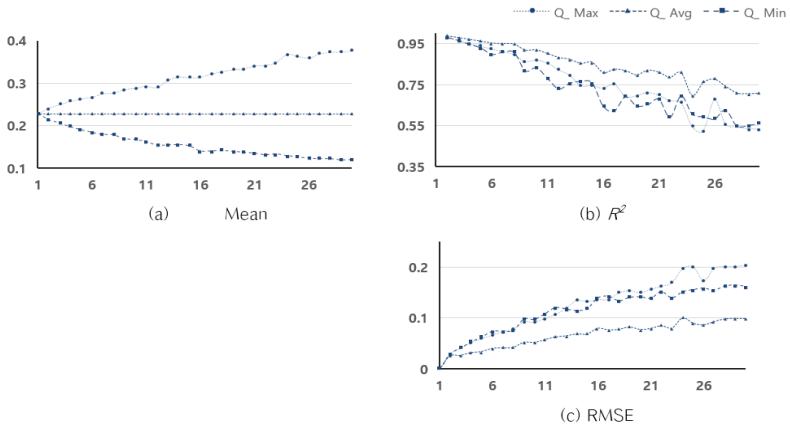


Fig. 12. Variations in averages, R^2, and RMSE for down sampling methods (n to 30) of Q (Max, Average, Min).

인하기 위해 n이 1, 10, 20, 30일 때 평균값을 대푯값으로 하는 각 변수들의 산포도 그래프를 **Fig. 13**에 정리하였다. n이 1, 10, 20, 30이면 데이터의 시간적 해상도는 15, 150, 300, 450 분에 해당하여 일정한 시간 간격(2.5 시간)의 데이터 변화를

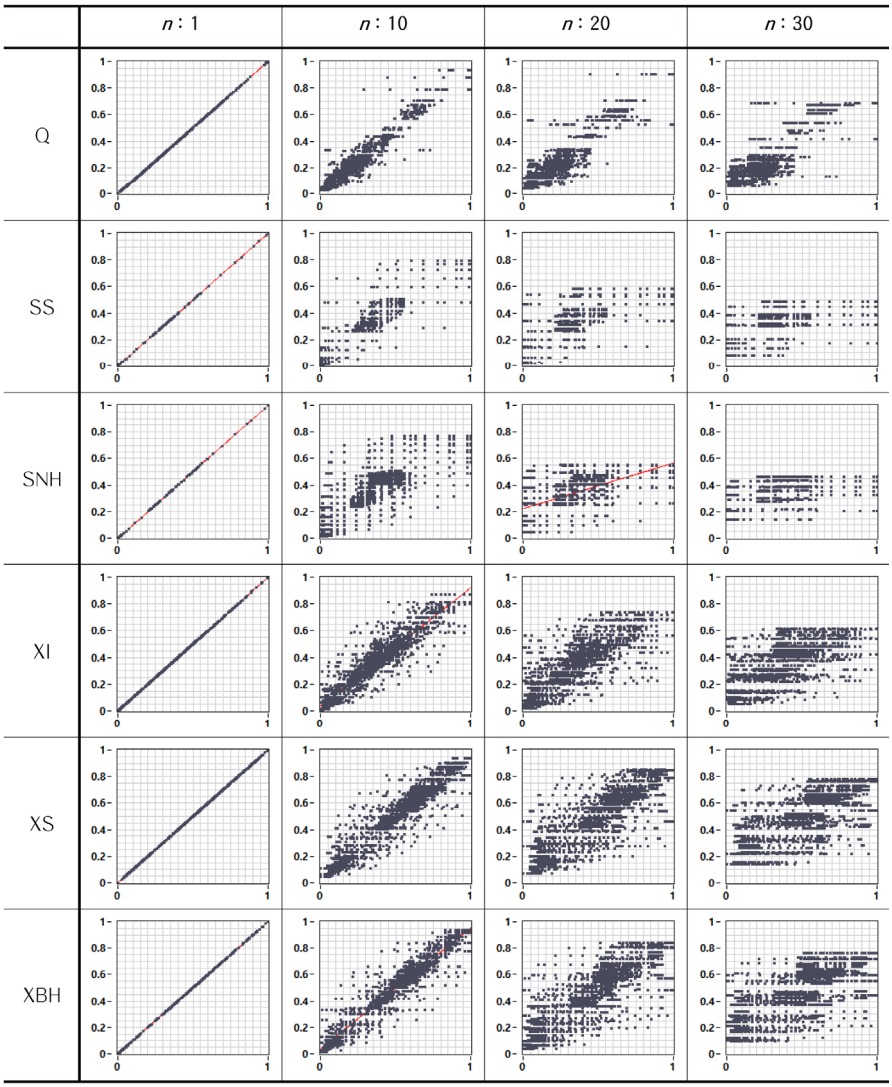


Fig. 13. Scatter Plot of Min-Max regularized raw data and down Sampled Data averaged at various temporal resolutions (n = 1, 10, 20, 30).

확인할 수 있다. 모든 변수에서 데이터의 시간 해상도가 낮아질 수록 데이터의 연속성은 단절되고, 구간별 데이터로 구분되면서 횡방향 분포가 넓어졌다. 또한 선형회귀선은 오른쪽으로 기울고, y축 절편이 증가하는 것을 확인할 수 있다.

Fig. 2는 BSM1 모델의 입력변수 중 유입 유량(Q)의 시계열 변화 특성을 확인할 수 있는 그래프이다. 해당 그래프에서 y축은 Q를 min-max 정규화한 값(max = 60,330 m^3/d)으로, x축은 단위시간 15분을 "1"로 단순화하여 표현하였다. 이후의 서술에서도 x축값을 언급할 때 그래프에 표현된 x축의 값으로 시간 표현을 대신한다. 이어서 Q의 시계열 패턴에서 확인할 수 있는 3개의 피크유량은 x축의 ① 800에서 1,000사이(Rain), ② 1,500(Strom1), ③ 1,700(Storm2)과 같다. 해당 피크유량의 지속시간은 ② Storm1 < ③ Storm2 < ① Rain의 순이며, 피크유량의 Q는 ① Rain에(52,163 m^3/d) < ② Storm2 (59,921 m^3/d) < ③ Storm2(60,330 m^3/d)로 정리할 수 있다.

Fig. 14에서 시간적 해상도의 구간별 대푯값을 산정하는 방법(Max, Average, Min)에 따른 데이터의 시계열적 변화과정을 확인할 수 있다. Down sampling을 할 때 n이 증가함에 따라 데이터의 산정 방법에 의존하여 Max일 때는 n=1인 기준데이터의 세부 피크들의 최대값을 연결한 값으로 데이터셋이 구성되며, 반대로 Min일 때는 피크들의 최솟값을 연결한 값들로 데이터셋을 구성한다. Average는 평균값 부근에서 데이터셋이 구성된다. n이 30이 되면 Min의 경우 ② Storm2와 ③ Storm1의 피크는 기준데이터(n=1)의 값을 추적할 수 없을 정도로 낮아졌고, ① Rain은 피크는 높이를 유지하였다. Average의 경우 n이 30일 때 ① Rain과 ③ Storm1은 어느 정도 피크의 높이를 반영하고 있지만 ② Storm2는 낮아졌다. Average의 경우를 down sampling 단계별로 보면 n=12일 때 ② Storm2의 피크는 최대높이의 절반까지 낮아져 ③ Storm1에 비하여 빠른 속도로 줄어드는 것을 확인할 수 있다. ②

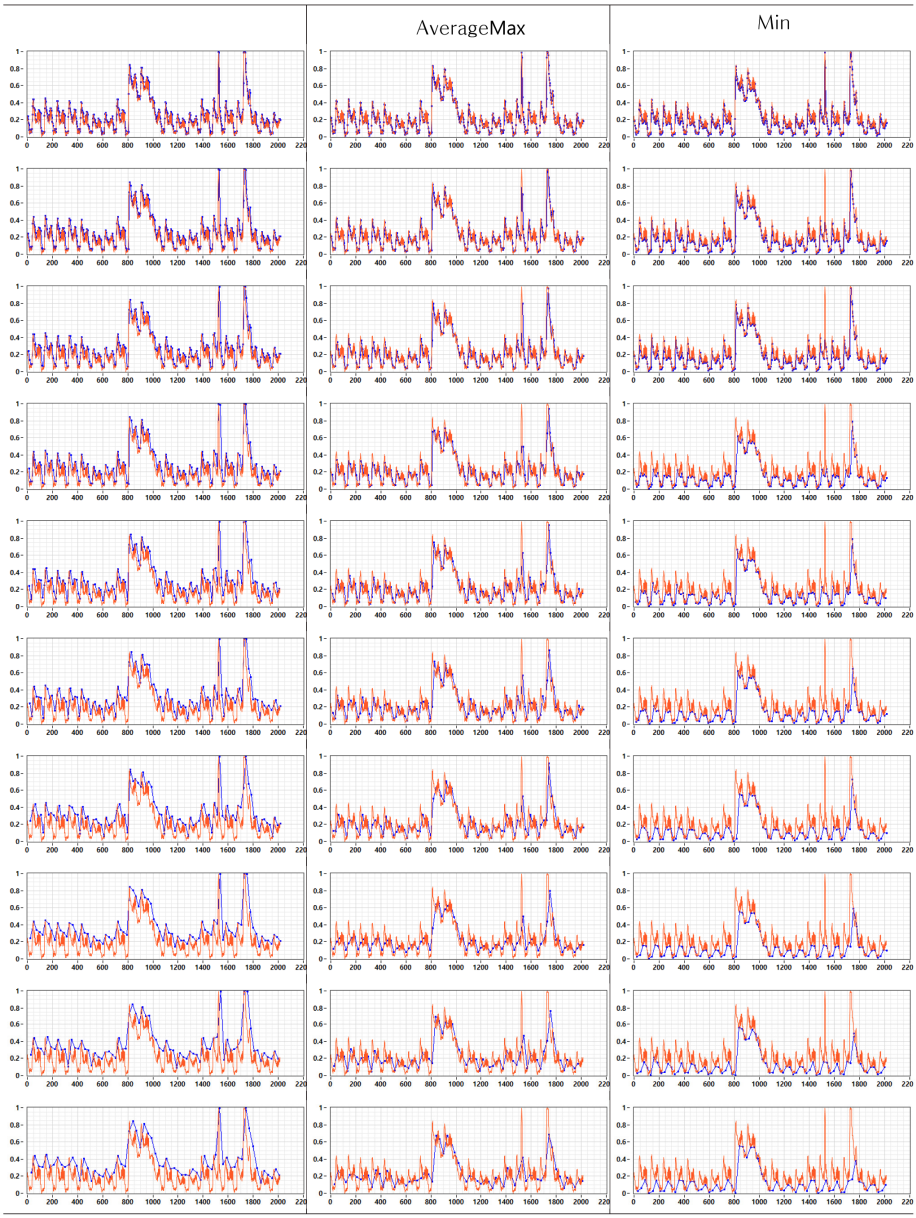


Fig. 14. Data changes due to down sampling for Q (Max, Average, Min) at different aggregation methods (in a vertical sequence, n=3, 6, 9, 12, 15, 18, 21, 24, 27, 30).

Storm2 (52,163 m³/d)와 ③ Storm1 (60,330 m³/d)의 기준데이터는 ① Rain (52,163 m³/d) 보다 크지만 피크 지속시간은 ① Rain이 가장 길고 ③ Storm1와 ② Storm2의 순서로 짧았다. n을 증가시킬 때 피크의 높이는 ② Storm2, ③ Storm1의 순서로 낮아졌다. 따라서 피크데이터의 감소 속도는 피크의 지속시간과 반비례 관계임을 확인할 수 있으며, down sampling으로 데이터의 시간 해상도가 낮아질 때 데이터의 값(크기) 보다는 지속시간에 의존하여 최초데이터의 정보를 보존할 가능성이 높다고 할 수 있다.

덧붙여 down sampling 기간에서 최대값을 대푯값으로 취하였을 때(**Fig. 14**. (a) Max) 세부 피크들의 최대값이 연결된 형태로 데이터셋이 구성되는 것을 확인할 수 있다. 데이터의 측정 간격이 넓어질 때 데이터의 특성과 정보의 손실을 막을 수 없지만, 일정 시간 간격의 데이터에서 최대값을 선택하면 원시데이터의 변화 패턴이 유지된 데이터를 얻을 수 있다. 따라서 사고 예방을 위한 경고 시스템과 같이 예측의 정확성 보다는 사건의 감지가 더 중요한 시스템인 경우 다른 데이터와의 시간적 연계성을 맞추거나, 데이터량을 조절하기 위하여 down sampling이 필요할 때 일정 시간 간격의 데이터에서 최대값을 취하는 것이 더 효과적일 것이다.

Guoliang은 Kalman filter의 잔차(residual)를 활용한 배수관망의 피크유량 예측에서 데이터의 임계값을 높이면 피크 유량의 예측 성능이 감소하고, 평균 유량의 예측 성능이 증가하였지만, 임계값을 낮추면 피크 유량의 예측 성능은 증가하고, 평

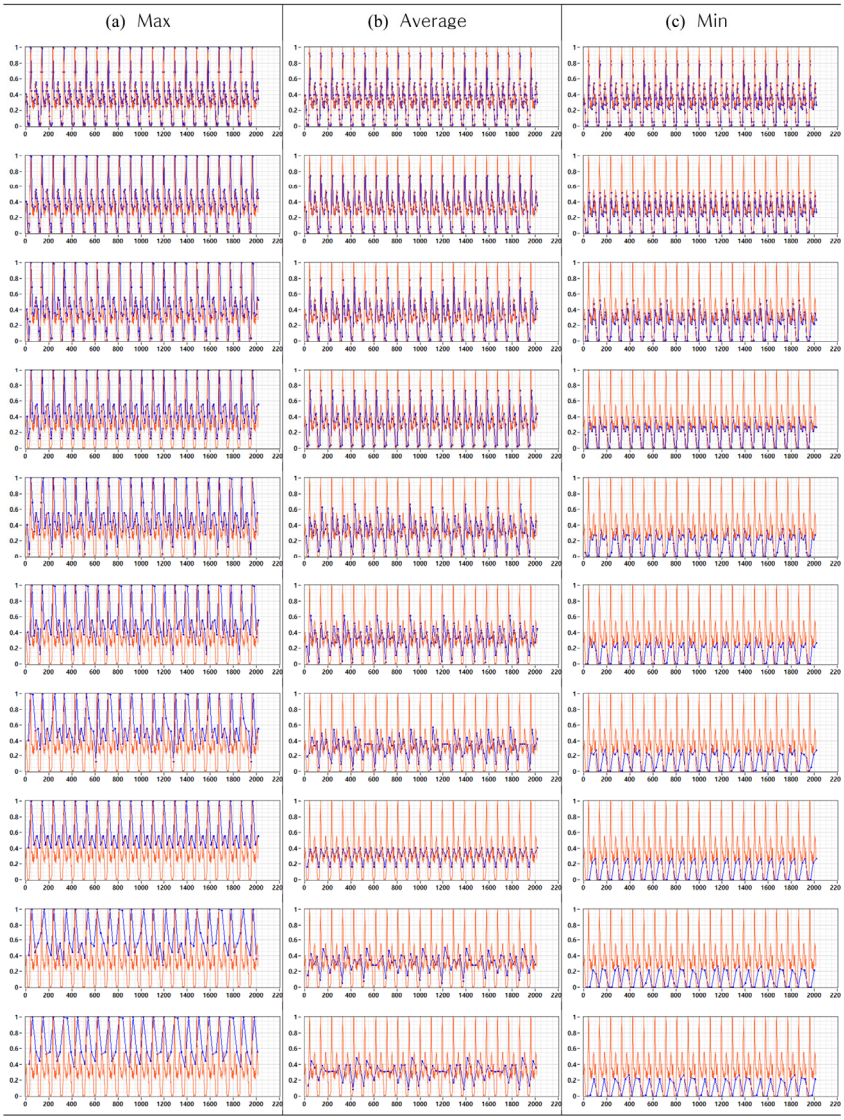


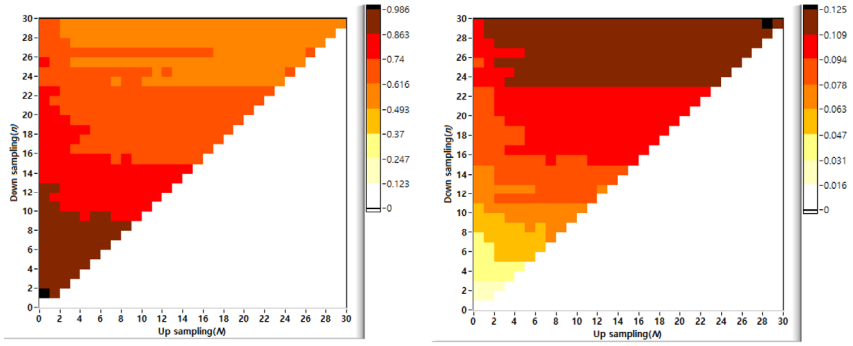
Fig. 15. Data changes due to down sampling for SS (Max, Average, Min) at different aggregation methods (in a vertical sequence, n=3, 6, 9, 12, 15, 18, 21, 24, 27, 30).

균 유량의 예측 성능이 감소하는 피크 유량과 평균유량의 예측 성능의 trade off 관계가 있음을 제시하였다.[22] Guoliang의 연구의 임계값과 같은 의미로 본 연구의 일정 시간간격에 대한 대푯값 선정 방법에 따라 Down sampling을 수행할 때 데이터의 정확성과 오차를 우선적으로 고려한다면 일정 구간의 시간 해상도에 대하여 평균값으로 데이터를 취하고, 경고 시스템의 경우 일정간격의 최대값을 취하는 것이 효과적이었다.

BSM1 모델의 입력변수 중 SS, XI, XS는(**Fig. 3(a)**)는 날씨 시나리오나 유입 유량(Q)의 패턴과 관계없이 일정한 펄스로(3주간 21번) 반복되는 시계열 특성을 나타낸다. 해당 데이터 중 유입수의 SS 농도에 대하여 시간적 해상도의 구간별 대푯값을 산정하는 방법(Max, Average, Min)에 따른 데이터의 시계열적 변화과정을 확인하였다(**Fig. 15**). 연속되는 반복 펄스 형태의 데이터의 경우에서도 Down sampling을 할 때 n이 증가함에 따라 최대값, 중간값, 최소값의 데이터 산정 방법에 의존하여 데이터셋이 구성되는 것을 확인할 수 있다.

3.2.2. Up sampling

시간적 해상도가 낮아진 상태의 데이터를 역으로 선형보간법을 적용해 해상도를 높이는 경우 데이터에 미치는 영향을 확인하기 위해 down sampling을 수행한 상태의 데이터셋을 대상으로 up sampling을 수행하였다. n을 30까지 늘리는 down sampling을 수행할 때 각 단계에서 n개의 up sampling 데이터셋이 생성되어 총 435개의 down sampling 데이터셋이 생성된다. 각 단계의 down sampling 데이터를 기준으로 해당 단계에서 생성된 down sampling 데이터와의 관계를 R^2와 RMSE으로 확인하여 해당 결과를 한 번에 확인할 수 있도록 heatmap으로 시각화하였다(**Fig. 16**).



(a) R^2 (b) RMSE

Fig. 16. Heat map of R^2 and RMSE after down sampling (n) and up sampling (N/n).

n이 3, 5, 9, 19일 때(down sampling) N을 2로 up sampling하였을 때 R^2는 각각 0.004, 0.001, 0.002, 0.005 만큼 증가하고, RMSE는 0.002, 0.001, 0.001, 0.001 만큼 감소하였다. 하지만 대부분의 결과에서 N이 증가할수록 데이터의 R^2값이 낮아지고 RMSE는 커지는 결과를 확인할 수 있다. 시간적 해상도가 낮은 데이터를 선형보간법을 적용해 해상도를 높일 때 긍정적인 효과를 기대할 수 없으며, 이는 해상도가 낮은 이미지를 보간법을 적용해 해상도를 높여도 이미지의 선명도가 높아지지 않는 것과 같은 결과로 해석할 수 있다. Abbaspour의 연구에서도 수질 예측에 있어 시간적 해상도가 낮은 데이터(12회/년)를 보정한 모델은 시간적 해상도가 높은 데이터(26회/년)의 성능에 미치지 못 하였고, 시간적 해상도가 낮은 데이터에 대한 보정에 더 많은 주의가 필요하며, 시간적 해상도를 높일 수 있는 센서를 활용하는 것이 모델의 불확실성을 낮추는 데 주요하다고 밝혔다.[7] 다만 이미지 분야에서 활용되고 있는 다양한 해상도 향상 기술을 데이터 시간적 해상도에 적용한다면

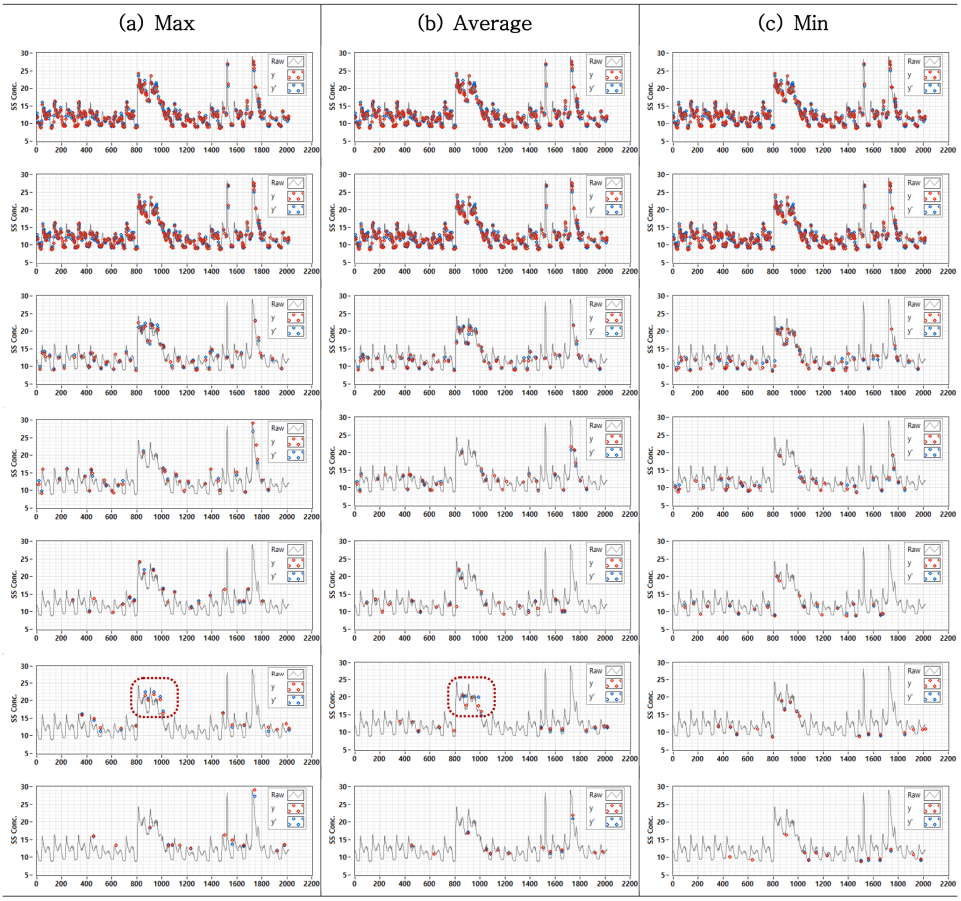


Fig. 17. Validation results of learning models applied to down sampled steps (n=0, 1, 6, 12, 18, 24, 30) for Q aggregation methods (Max, Average, Min) (Red: actual values, Blue: predicted values).

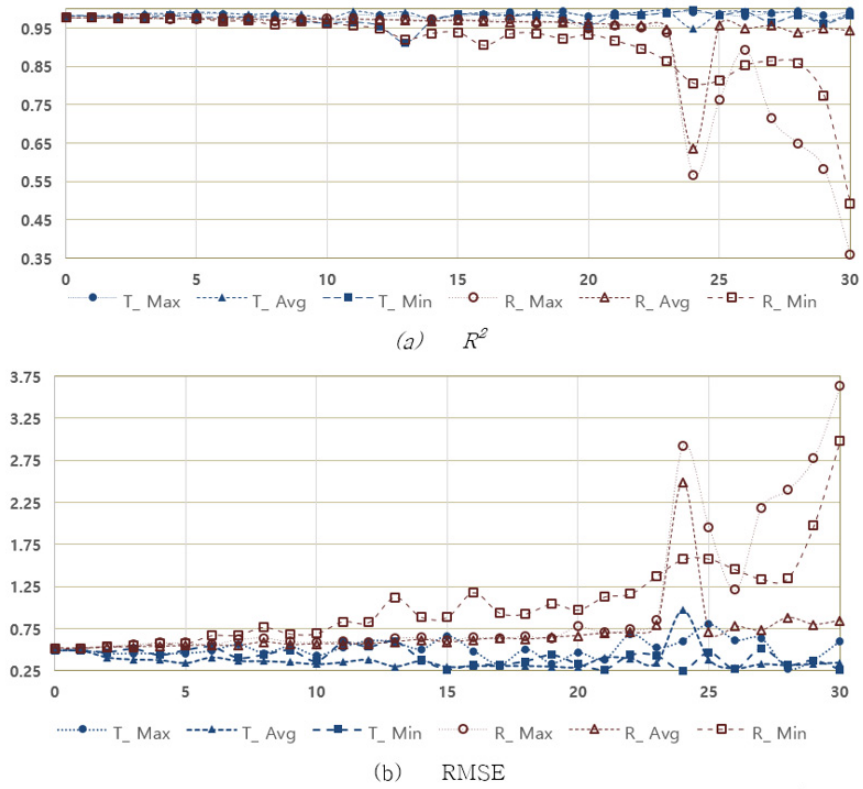


Fig. 18. Validation results of learning models applied to down sampled steps (n=0, 1, 6, 12, 18, 24, 30) for Q aggregation methods (Max, Average, Min) using validation data from each step (T_Max, T_Avg, T_Min) and Raw Data (R_Max, R_Avg, R_Min).

보다 긍정적인 효과를 기대할 수 있을 것이다.[20,21]

3.3. 데이터 측정 간격에 따른 기계학습 성능 비교

BSM1 모델의 유입수의 입력변수(Q, SS, SNH, XI, XS, XBH)를 활용해 처리수인 최종침전조의 부유물질 농도(OutSS)를 역전파신경망 모델로 예측하였다. Down sampling 방법(max, average, max) 별로 n이 30까지 증가하는 각 단계의 데이터셋으로 기계학습(BPN, LabVIEW)을 수행하여 총 90개의 학습모델을 구축하였다. 생성된 학습모델 중 n이 0, 1, 6, 12, 24, 30에 해당하는 그래프를 **Fig. 17**에 제시하였다. 해당 그래프는 down sampling 단계에서 생성된 데이터셋을 훈련데이터와 검증데이터를 구분하여 훈련데이터로 학습모델을 생성하고 해당 단계의 검증데이터로 학습모델을 검증한 결과이다. Down sampling을 진행할수록 데이터의 수가 감소하여 예측값(푸른색, y')과 검증데이터(붉은색, y)의 쌍이 줄어들고, n이 증가할수록 최초데이터(회색, Raw)에서 검증데이터가 이탈하는 현상이 관찰되었다.

Down sampling 데이터셋으로 생성된 단계별 학습모델들간 성능을 상호비교하기 위하여 down sampling 이전의 최초 데이터셋을 검증데이터로 적용한 R^2와 RMSE 결과를 **Fig. 18**에 제시하였다. Down sampling 단계의 검증데이터로 학습모델을 검증하였을 때 n이 증가하여도 R^2는 0.911에서 0.99까지 정확성을 유지하였고, RMSE는 최소 0.255에서 0.973의 범위에서 비교적 안정적인 결과를 보였다. 하지만 최초 데이터셋으로 검증하였을 때 n이 15일 때부터 R^2와 RMSE는 크게 변하였다. 최대값으로 down sampling한 경우 R^2는 0.360까지 최소값인 경우는 0.493까지 낮아졌고, 평균값인 경우 n이 24일 때 0.634로 낮아졌다가 이후 0.94 수준을 회복하였다. RMSE 역시 최대값으로 down sampling한 경우 3.632, 최소값인 경우 2.978까지 증가하였다. 평균값인 경우는 n이 24일 때 2.487로 최대였으나 이후 다시 낮아지는 경향을 보였다. Down sampling 데이터셋으로 생성된 단계별 학습모델들은 각 단계에 속하는 검증데이터로 검증할 때 높은 성능을 유지하였지만, 최초 데이터셋으로 검증할 때 n의 증가에 따른 성능감소를 확인할 수 있었다.

Down sampling 데이터셋의 단계별 검증데이터를 적용할 때 n이 증가하여도 학습모델의 성능이 유지되는 현상은 앞의 down sampling 과정에서 생성되는 데이터가 최초데이터에서 이탈되는 결과(**Fig. 14**, **Fig. 15**)와 함께 원인을 파악할 필요가 있다. n이 증가할수록 데이터의 측정 간격은 넓어진다(시간적 해상도가 낮아진다). 데이터의 시간적 해상도가 낮아지면서 기계학습을 위한 학습 및 검증데이터의 해상도 역시 같은 상태로 낮아지고, 경우에 따라 최초데이터 선에서 함께 이탈한다. 해상도가 낮은 학습데이터를 활용한 학습모델은 분해능

낮아진 상태에서 학습데이터와 같은 해상도의 검증데이터로 학습모델을 검증한다. 이에 따라 n이 증가하여도 모델의 R^2 및 RMSE의 지표는 높은 수준을 유지할 수 있다고 판단할 수 있다. E. Arandia는 물사용량을 SARIMA 모델로 예측하는 경우 단기 예측에는 시간적 해상도가 높은 데이터를 적용해 급격한 변화에 대응하고, 장기 예측에는 시간적 해상도가 낮은 데이터를 적용하여 시계열 데이터의 해상도에 따라 모델의 정확성이 달라지는 것을 밝혔다.[11] 따라서 낮은 시간적 해상도의 데이터로 준비된 기계학습 모델을 현장에 적용할 때 모델이 제시하는 성능에 미치지 못하는 경우, 데이터의 시간적 해상도가 그 원인 중 하나일 수 있다.

한편 down sampling 과정에서 n이 24일 때 최대값과 평균값으로 데이터를 취할 때 R^2와 RMSE의 이상치가 관찰되었으나 최소값인 경우는 상대적으로 영향이 적었다. 최대값으로 데이터를 취할 때 R^2는 0.568로 낮아졌다가 n=25일때 0.763으로 증가하였으며, 평균값일 때는 0.947로 낮아진 후 다시 0.988까지 증가하였다. 해당 단계에서 학습데이터와 검증데이터의 분류에 따른 영향인지 확인하기 위해 seed 값을 2로 변경하여 기계학습을 수행하였을 때 최대값의 R^2는 0.447까지 낮아졌다가 0.896까지 회복하였고, 평균값인 경우 0.900까지 낮아졌다가 0.962까지 증가하는 형태를 보였다. 본 연구에서는 down sampling 과정에서 n이 24 지점에서 발생하는 이상치에 대한 원인을 밝히지 못 하였다.

4. 결 론

본 연구의 목적은 데이터의 시간적 해상도 변화에 따른 데이터의 특성 변화를 확인하고, 시간적 해상도가 기계학습 모델에 미치는 영향을 확인하는 것이다. 우선, BSM1 모델로 dry, rain, storm 조건의 2주간 15분 간격의 데이터를 확보하였고, resampling으로 데이터의 해상도를 변화시킨 후 기계학습을 수행하는 LabVIEW 기반 프로그램을 개발해 연구에 활용하였다. LabVIEW 기반의 기계학습 모델인 BPN을 R의 DT, LR, RF, SVM과 예측성능을 비교한 후 down sampling과 up sampling의 과정에서 생성된 데이터셋으로 데이터의 특성변화를 확인하고, down sampling으로 시간적 해상도가 달라진 데이터셋으로 기계학습을 수행하면서 학습모델의 성능을 비교하여 다음의 결과를 얻었다.

1) 15분 간격의 데이터를 n개씩 묶어 데이터 측정 간격을 $15\times n$ 분으로 증가시킬 때 n이 증가함에 따라 R^2는 낮아지고, RMSE는 증가한다. 데이터의 연속성이 단절되어 구간별 데이터의 특성이 나타나며, n이 증가할수록 구간별 데이터의 분포가 넓어졌다(전반적인 추상화).

2) 구간별 대푯값 선정 기준으로 최대, 최소, 평균값 중 평균값을 적용할 때 데이터의 정확성과 오차의 손실이 가장 적고, 최대값을 취할 때 데이터의 특성을 보다 잘 유지할 수 있었다. 이러한 결과를 토대로 사고경보 등 정확성보다는 변화를 감지하는 능력이 중요한 모델을 개발할 때는 일정기간 동안의 최대값을 대푯값으로 활용하는 것이 효과적임을 알 수 있다.

3) Down sampling한 데이터를 보간법을 활용해 데이터 측정 간격을 $1/N$으로 나누어 줄이는 과정(up sampling)에서 대부분의 경우 N이 증가할수록 R^2가 낮아지는 경향을 확인하였다. 시간적 해상도가 낮은 데이터에 대해 선형보간법으로 해상도를 높이는 경우 데이터의 정확성 향상의 긍정적인 효과를 기대할 수 없으며 이미지 분야에서 활용되는 다양한 해상도 향상방법을 적용해보는 연구가 필요하다.

4) Down sampling으로 시간적 해상도를 낮추는 과정에서 생성된 30개의 데이터셋으로 구축한 학습모델에 같은 수준의 시간적 해상도의 검증데이터로 성능을 확인하는 경우 해상도가 낮아져도 모델의 R^2 및 RMSE의 성능지표는 높은 수준을 유지하였다. 한편 down sampling을 진행할수록 생성된 데이터셋은 최초 데이터에서 이탈하였고, 생성된 데이터셋의 학습데이터와 검증데이터는 같은 수준의 시간적 해상도를 갖는다. 시간적 해상도가 낮은 학습데이터 기반의 학습모델은 같은 수준의 검증데이터로 성능을 평가할 때 최초 데이터의 상태와 상관없이 높은 성능을 제시한다고 판단할 수 있다.

5) Down sampling 과정에서 생성된 30개 학습모델의 성능을 down sampling 이전의 최초 데이터를 검증데이터로 적용하면 시간적 해상도가 낮을수록 학습모델의 성능도 낮아졌다. n이 15일 때부터 R^2와 RMSE는 크게 변하여 R^2는 0.360(최대값), 0.493(최소값), 0.634(평균값)까지 낮아졌고, RMSE는 3.632(최대값), 2.978(최소값), 2.2487(평균값)까지 증가하였다. 시간적 해상도가 다른 모델의 성능을 상호 비교하기 위해서는 같은 수준의 데이터를 적용할 필요가 있다.

6) 본 연구에서 전체 데이터로 검증한 down sampling 단계별 기계학습에서 평균값으로 데이터를 취한 경우 n이 24인 지점에서 발견된 이상치에 대한 원인이나 학습모델의 성능 저하에 대한 명확한 상관관계를 밝히지 못 하였다.

Acknowledgement

본 연구는 한국연구재단(No. 2020R1C1C1013643, 2021H1D3A2A02039182)의 지원을 받아 수행되었습니다.

References

1. O W, Jang HN, Shin SG. Application of machine learning in water industry: A review, *J Korean Soc Water Wastewater*, 36(1), 9-21(2022).
2. Yoon Y, Park J-H, Kang J-H, Choi J-S, Park J, Kwak P-J, Analysis on the Utilization of Renewable Energy for Carbon Neutralization in Sewage Treatment Facilities, *J Korean Soc Env Eng*, 44(12)n 543-551(2022).

3. Niu K, Wu J, Qi L, Niu Q, Energy intensity of wastewater treatment plants and influencing factors in China, *Sci Total Environ*, 670j, 961-970(2019).
4. Korea Ministry of Environment, measurement equipments and accessories installation, *2022 water quality TMS handbook*, 3-24(2022).
5. Repp AC, Roberts DM, Slack DJ, Repp CF, Berkler MS, A comparision of frequency, interval, and time-sampling methods of data collection, 4(4), 501-508(1976).
6. Fettweis M, Riethmüller R, Van der Zande D, Desmit X, Sample based water quality monitoring of coastal seas: How significant is the information loss in patchy time series compared to continuous ones?, *Sci Total Environ*, 873, 162273(2023).
7. Abbaspour KC, Rouholahnejad E, Vaghefi S, Srinivasan R, Yang H, Kløve B, A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model, *J Hydrol*, 524(May), 733-752(2015).
8. Cominola A, Giuliani M, Castelletti A, Rosenberg DE, Abdallah AM, Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management, *Environ Model Softw*, 102, 199-212(2018).
9. Ching PML, So RHY, Morck T, Advances in soft sensors for wastewater treatment plants: A systematic review, *J Water Process Eng*, 102367, 44(2021).
10. Juo Yo Liao, Effects of Temporal Resolution on Data Mining and Machine Learning Algorithms in the Built Environment, Eindhoven university of technology, (2021).
11. Arandia E, Eck B, McKenna S, The effect of temporal resolution on the accuracy of forecasting models for total system demand, *Procedia Eng*, 89, 916-925(2014).
12. Piniewski M, Marcinkowski P, Koskiaho J, Tattari S, The effect of sampling frequency and strategy on water quality modelling driven by high-frequency monitoring data in a boreal catchment, *J Hydrol*, 579(May) (2019).
13. Alex J, Benedetti L, Copp J, et al., Benchmark Simulation Model no. 1 (BSM1), *Benchmark Simul Model*, 1(1BSM1), 1-58 (2008).
14. Ozdemir S., Data aggregation in wireless sensor networks, *RFID Sens Networks Archit Protoc Secur Integr*, 3(3), 297-322(2009).
15. Popescu AC, Farid H., Exposing digital forgeries by detecting traces of resampling, *IEEE Trans Signal Process*, 53(2 II), 758-767 (2005).
16. Kodosky J., LabVIEW, *Proc ACM Program Lang.*, 4(HOPL), (2020).
17. Whitley KN, Blackwell AF, Visual Programming in the Wild: A Survey of LabVIEW Programmers, *J Vis Lang Comput*, 12(4), 435-472 (2001).
18. GitHub PeteHorn/LV Sets Maps UKTAG, https://github.com/PeteHorn/LV-Sets-Maps-UKTAG, Accessed September 18 (2023).
19. Burdack J, Horst F, Giesselbach S, Hassan I, Daffner S, Schöllhorn WI, Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Performance of Gait Classifications Using Machine Learning, *Front Bioeng Biotechnol*, 8(April), 1-12(2020).
20. Menon S, Damian A, Hu S, Ravi N, Rudin C, PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models, *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2434-2442(2020).
21. Du J, Zhou H, Qian K, et al., RGB-IR cross input and sub-pixel upsampling network for infrared image super-resolution, *Sensors (Switzerland)*, 20(1), 1-20(2020).
22. Guoliang Ye, Richard Andrew Fenner, Study of Burst Alarming and Data Sampling Frequency in Water Distribution Networks, J. Water Resour. Plann. Manage, 140, (2014).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors and Contribution Statement

Wonki O

Digital Innovation Center, Korea Testing Laboratory, Senior Researcher, ORCID 0009-0007-9153-4284: Conceptualization, Modeling, Data analysis, Machine Learning, Programming, Visualization, Writing-original draft.

Seo Jin Ki

Department of Environmental Engineering, Gyeongsang National University, Associate professor, ORCID 0000-0001-7056-9217: Supervision, Conceptualization, Conceptualization, Writing-review and editing.

Jin Mi Triolo

Future Convergence Research Institute, Gyeongsang National University, Research professor, ORCID 0000-0002-1960-9823: Writing-review and editing.

Seung Gu Shin

Department of Energy Engineering, Gyeongsang National University, Associate professor, ORCID 0000-0002-6077-9576: Supervision, Project administration, Writing-review and editing, Final approval.