Journal of the Korea Academia-Industrial cooperation Society
Vol. 24, No. 1 pp. 570-580, 2023

https://doi.org/10.5762/KAIS.2023.24.1.570
ISSN 1975-4701 / eISSN 2288-4688

스마트 수도미터 계측 데이터 품질관리 기술 개발

이상호[1*], 이주영[1], 윤정환[2]
[1]국민대학교 건설시스템공학과, [2]주식회사 뉴컨

Development of Quality Management Technology for Data Measured by Smart Water Meters

Sangho Lee[1*], Juyoung Lee[1], Junghwan Yun[2]
[1]Department of Civil Engineering, Kookmin University
[2]Newcon Inc

요 약 최근 전 세계적으로 스마트시티가 추진되고 있으며 이를 위한 스마트 물관리 기술의 개발이 진행 중이다. 스마트 수도미터는 이러한 스마트시티 물관리를 위한 핵심적인 요소로서 실시간 물 사용량 데이터를 수집하여 제공함으로써 물 수요관리와 누수 탐지 등을 위한 정보를 확보할 수 있도록 한다. 그러나 스마트 수도미터에서 수집된 데이터는 각종 오류가 포함될 수 있어 이를 보정하기 위한 전처리가 필수적이다. 따라서, 본 연구에서는 스마트 수도미터에서 수집되는 데이터를 자동으로 전처리하여 품질을 향상시키고, 이를 활용하여 물 사용량 패턴을 분석하고 예측하기 위한 모델을 개발하고자 하였다. 연구를 위하여 실제 스마트 수도미터에서 수집된 원본 데이터를 활용하였으며, 오류 유형을 정의하고 각각에 대한 처리 방법을 개발하여 적용하였다. 물 사용량 패턴 분석을 위해서는 군집분석과 시계열 분해, LSTM(Iong Short Term Memory, 이하 LSTM)과 XGBoost (eXtreme Gradinet Boosting, 이하 XGBoost)모델 등을 적용한 후 그 타당성과 정확도를 분석하였다. 연구 결과 다양한 유형을 가지는 수집 데이터의 오류를 자동적으로 처리할 수 있는 알고리즘을 개발하여 물 사용량 패턴분석에 활용할 수 있는 품질의 데이터 셋을 만들 수 있었다. 물 사용량 패턴 분석의 경우 노이즈 성분을 제거하고 주기적 성분만 XGBoost 모델을 이용하여 학습시킨 경우가 가장 낮은 RMSE(Root Mean Squared Error, 이하 RMSE) 값을 나타내어 가장 성능이 좋은 것으로 판단되었다.

Abstract Smart cities are increasingly arising worldwide, with the ongoing development of smart water management technology. A key element for such smart city water management involves a smart water meter that collects and provides real-time water usage data to secure information for water demand management and water leakage detection. However, since data collected from smart water meters may contain various errors, preprocessing to correct them is essential. This study attempts to develop a model for automatically preprocessing data collected from smart water meters to improve the quality, and to analyze and predict water usage patterns. Original data collected from the actual smart water meters were used, the error types were defined, and the processing method for each was developed and applied in the research. To analyze the water usage pattern, cluster analysis, time series decomposition, LSTM, and XGBost models were applied, and their feasibilities and accuracies were analyzed. The results were applied to develop an algorithm that automatically handles errors in collection data of various types, to create a quality dataset that can be used for water usage pattern analysis. Water usage pattern analysis determined lowest RMSE value and best performance when the noise component was removed and only the periodic component was obtained using the XGBoost model.

Keywords : Smart City, Smart Water Meter, Data, Preprocessing, Quality Management, Deep Learning

본 논문은 환경부의 재원으로 한국수자원공사(K-water)의 연구과제(A-C-002)로 수행되었음.
*Corresponding Author : Sangho Lee(Kookmin University)
email: sanghlee@kookmin.ac.kr
Received November 9, 2022 Revised January 3, 2023
Accepted January 6, 2023 Published January 31, 2023

1. 서론

4차 산업혁명의 인한 기술혁신은 사람들의 생활방식과 공간을 빠르게 변화시키고 있으며, 이 중 대표적인 사례가 스마트시티(Smart city)라고 할 수 있다. 스마트시티는 전자통신 센서와 계측기 등을 사용하여 다양한 유형의 데이터를 수집하고 분석하고 자산과 자원을 효율적으로 관리하는 데 필요한 정보를 제공하는 도시로 정의된다 [1]. 수집된 데이터는 교통 및 운송 시스템, 전력공급 및 관리 시스템, 상하수도, 폐기물 관리, 정보화 시스템, 학교, 도서관, 병원 및 기타 커뮤니티 서비스를 모니터링하거나 관리하기 위해 처리·분석된다 [2]. 스마트시티에서는 데이터와 정보를 기반으로 하여 도시 운영 및 서비스의 효율성을 최적화하고 시민들과의 연결을 위한 네트워크를 제공한다 [3,4]. 따라서 스마트시티는 기존도시가 가지고 있는 여러 가지 문제점을 해결할 수 있는 대안으로서 주목받고 있으며, 전 세계적인 도시화 추세의 증가와 삶의 질 향상에 대한 시민들의 요구에 대응할 수 있을 것으로 기대를 모으고 있다 [4].

최근 국내 뿐 아니라 전 세계적으로 스마트시티 관련 프로젝트가 활발하게 추진되고 있다. 국외 시장조사기관인 Allied Market Research에 따르면 2020년 전 세계 스마트시티 시장규모는 약 6480억 달러이며, 연 평균 25.2 %로 성장하여 2030년에는 약 6조 달러에 이를 것으로 예상된다 [5]. 미국, 캐나다, 유럽, 싱가포르 등 선진국에서는 도시에 새로운 기능을 제공하기 위한 스마트시티 프로젝트가 추진되었거나 추진 중이며, 인도와 중국, 인도네시아 등의 개발도상국에서도 기존 도시의 문제를 해결하기 위한 여러 형태의 스마트시티 프로젝트가 추진되고 있다 [6,7].

스마트시티의 공공 인프라와 서비스는 기본인프라(Physiological infra), 안전 인프라(Safety infra), 사회적 서비스(Social service), 연결된 원활한 스마트 서비스(Connected seamless smart service) 등으로 구분할 수 있다 [3,4]. 이 중에서 상하수도와 같은 물관리는 기본 인프라에 해당하며, 스마트시티의 기반이 되는 핵심 구성요소이다 [5,7]. 따라서 스마트시티의 발전을 위해서는 도시에서의 스마트 물관리 기술의 발전과 적용이 필수적으로 요구된다 [8]. 이러한 스마트 물관리 기술 중 대표적인 것으로 스마트워터그리드(SWG: Smart Water Grid, 이하 SWG)가 있으며, 이는 상하수도 시설에 정보통신기술(ICT: Information and Communications Technology, 이하 ICT)를 접목하여 물 공급의 효율화와 운영의 최적화를 도모하는 스마트 도시 기술이다 [9].

스마트시티의 물관리를 위한 필수요소는 시설 및 장비(Smart devices), 데이터 분석 및 관리(Smart solutions), 정보제공 서비스 (Smart services) 등이다. 이를 위한 대표적인 하드웨어로는 센서와 계측 장치, 통신 장치, 서버/클라이언트 컴퓨터, 상하수도 시설 제어 장치 등이며, 소프트웨어로는 데이터 수집, 전처리, 저장, 분배, 분석, 실시간 제어, 의사결정 지원 소프트웨어 등이 있다 [8]. 이 중에서 특히 중요한 구성요소는 스마트 수도미터(Smart water meters)이다 [10]. 스마트 수도미터는 사용자의 실시간 물 사용량에 대한 데이터를 수집하여 제공하여 물 사용 패턴 등의 정보를 얻기 위한 것으로서 물 수요관리와 누수탐지 등을 위한 목적으로 활용될 수 있다 [9]. 또한 최근 IoT 네트워크의 적용 확대에 따라 IoT 기반의 스마트 수도미터가 적용되고 있어, 물 사용량 정보의 수집이 빠르게 증가하고 있는 추세이다 [11].

그러나 스마트 수도미터 등에서 수집된 데이터는 하드웨어의 한계와 일시적인 작동 오류에 따라 유효하지 않은 값을 포함하는 것이 일반적이며, 그 외에도 다양한 이유로 인하여 오류를 포함하게 된다 [10]. 이러한 낮은 품질의 데이터를 활용하는 경우 필요한 정보를 추출하여 활용하는 것이 불가능하므로 데이터의 분석을 위한 전처리가 필수적이다 [12]. 그러나 수집되는 데이터의 양과 수집속도가 증가하는 빅데이터의 시대에서 데이터의 품질관리와 패턴분석 등을 사람이 직접 할 수 없기 때문에 이를 위한 알고리즘과 모델, 소프트웨어의 개발이 시급하게 요구되고 있다. 따라서 본 연구에서는 스마트시티 등에 설치된 스마트 수도미터에서 수집되는 데이터를 자동으로 전처리하여 품질을 향상시키고, 이를 활용하여 물 사용량 패턴을 분석하고 예측하기 위한 모델을 개발하고자 하였다.

2. 연구개발 방법

2.1 연구대상 데이터

본 연구에서 사용되는 물 사용량 데이터는 I지역의 의 물 사용량 자료이다. 총 691개의 스마트 수도미터 로부터 2016년에서 2020년 사이에 계측된 시간별 누적 사용량을 원본 데이터로 활용하였다. 원본 데이터는 마이크로소프트 엑셀 파일 형태를 가지고 있으며, 누적 물 사용량을 시간별로 기록한 값을 포함하고 있다. 원본 데이터는 데이터 품질관리 알고리즘을 개발하기 위하여 사용되

었으며, 품질관리를 거친 데이터는 물 사용량 예측모델 개발을 위하여 사용되었다.

2.2 모델 개발

데이터 품질관리 모델과 분석모델은 파이썬으로 개발하였다. 데이터 분석 모델의 개발 및 검증을 위해서는 품질관리를 거친 후 충분한 품질을 가진 데이터 셋을 선택하여 활용하였다.

2.3 방법론 및 분석엔진의 선정

스마트 수도미터에서 수집된 데이터를 기반으로 하여 물 사용량 및 사용패턴을 예측하기 위하여 다음의 방법론을 적용하였다 [12,13].

2.3.1 군집분석

군집분석은 군집의 개수나 구조에 관한 특별한 가정 없이 개체들 사이의 유사성(similarity) 또는 거리(distance)에 근거하여 자연스러운 군집을 찾고 다음 단계의 분석을 꾀하는 탐색적인 통계분석 기법으로, 대상들의 특성을 분석하여 유사한 성질을 갖고 있는 대상들을 동일한 집단으로 분류하며 각 대상들이 갖고 있는 값을 거리로 계산하여 가까운 거리에 있는 대상들을 하나의 집단으로 묶는 방법이다. 본 연구에서는 개별 대상 간의 거리를 기준으로 나무모양의 계층구조를 상향식으로 형성해가는 방식의 계층적 군집분석을 이용하여 군집의 수를 정하였으며, 이때 거리 산정방법은 유클리디안 제곱거리를 이용하는 Ward방식으로 하여 군집의 수를 7개로 정하였다.

2.3.2 시계열 분해

시계열자료는 시간에 따라 변화하는 자연현상이나 사회현상을 기록한 데이터로서 시간파형 또는 신호라고 부르며, 개별적인 정보들의 혼합물이다. 시계열의 기본적인 표현법은 시간영역에서 시간의 함수로서 신호를 그려보는 것이지만, 신호를 주파수 영역이나 시간-주파수 영역으로 표현하는 방법이 신호특성의 근원을 파악하는데 더 유용한 것으로 알려져 있다. 따라서 신호를 후자의 영역에서 표현하기 위해 원래 시계열을 주파수 성분이나 시간-주파수 성분, 또는 시간-스케일 성분으로 분해하는 기법을 사용하며, 주파수 분석에는 푸리에변환, 시간-주파수 분석에는 단시간 푸리에변환, 시간-스케일 분석에는 웨이블릿 변환이 있다.

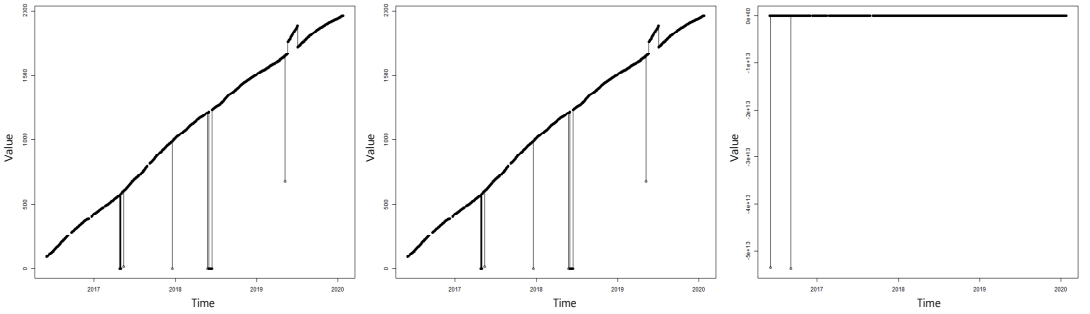
이 중에서 연속형 웨이블릿 및 이산형 웨이블릿 변환은 푸리에 변환 및 단시간 푸리에 변환을 통해 시간-주파수 분석을 수행 할 때 각각의 해상도를 동시에 만족할 수 없는 단점을 극복하기 위해, sine 및 cosine 함수 대신 기저함수를 이용하고, 창함수 대신 기저함수의 확장, 축소 및 이동을 통해 시간-주파수 분석을 수행한다. 이때, 각각의 주파수와 전이항의 값들이 연속적으로 적용될 경우 연속형 웨이블릿 변환이라고 정의하며, 이산적으로 적용될 경우를 이산형 웨이블릿변환이라고 정의한다. 데이터가 정상적(stationary)이라면 시간분석이나 주파수 분석만으로도 유용한 정보를 얻을 수 있으나 물 사용량 데이터는 비정상적 데이터이며 이산형 데이터이기 때문에 본 연구에서는 이산형 웨이블릿 방법을 사용하였다.

2.3.3 LSTM

순환신경망의 한 종류인 LSTM은 일시적으로 기억해야 할 것(short term memory)과 오랫동안 기억해야 할 것(long term memory)을 가지고 있어 순환신경망을 학습할 때 발생할 수 있는 기울기 소멸 문제를 해결해 줄 수 있기 때문에 대규모의 깊은 신경망에서 더 효율적으로 작업할 수 있게 만들어 준다. LSTM은 순환신경망 구조의 잊기 게이트(forget-gate), 입력게이트(input-gate), 출력게이트(output-gate)가 존재한다. 기존의 인공신경망에서는 장기의존성 문제로 기울기가 소멸하는 문제를 가지고 있지만, LSTM의 경우 잊기 게이트를 통해 일종의 컨베이어벨트 역할을 수행하여 기울기 소멸문제를 해결할 수 있어 시계열 데이터에서 순환신경망(RNN: Recurrent Neural Network, 이하 RNN)보다 효율적인 성능을 보인다. 따라서 본 연구에서는 장기 의존성 문제를 해결해 줄 수 있어 시계열 데이터를 학습하고 예측하는데 적합한 LSTM 모델을 이용하였다.

2.3.4 XGBoost (eXtreme Gradinet Boosting)

XGBoost는 선형 모델이나 트리 기반 모델의 과적합 문제를 해결하고, 규모가 큰 데이터셋의 안정성과 훈련 속도 향상의 목적으로 만들어졌다. XGBoost는 처리 속도가 빠르며 모델의 이전 결과를 활용하여 모델을 계속적으로 개선하고 훈련하는 등 성능이 뛰어나기 때문에, 최근 활용이 증가하고 있다. 본 연구에서는 모델의 이전 결과를 활용하여 계속적으로 모델의 정확도를 개선해 나갈 수 있는 XGBoost 모델을 추가적으로 적용하였으며, 앞서 LSTM 모델과 그 결과를 비교하였다.



(a) (b) (c)

Fig. 1. Examples of cumulative data by hour

2.4 예측 능력 평가

본 연구에서 예측력 평가의 지표로는 RMSE를 활용하였으며, 다음의 공식으로 계산하였다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2} \quad (1)$$

여기서 p_i는 예측된 값이고 y_i는 실제 측정된 값, n은 표본의 크기이다.

3. 결과 및 고찰

본 연구는 다음과 같은 과정으로 수행되었다. 먼저 수집된 데이터가 정상인지 비정상인지를 판단하기 위하여 데이터 오류 형태를 유형화하고 이를 검출하기 위한 알고리즘을 개발하였다. 다음으로 데이터 오류의 보정을 위한 물 사용패턴 예측모델을 개발하고자 수집·전처리된 데이터를 이용하여 딥러닝 모델을 학습하고 예측 정확도를 높이기 위한 방법을 개발하였다. 각 세부 단계에서의 연구결과는 다음과 같다.

3.1 데이터 오류 유형화 및 검출 기술 개발

3.1.1 계측 데이터 오류 유형 분석

본 연구에 사용된 원본 데이터는 현장에 설치된 다수의 스마트 수도미터에서 약 4년간 측정된 누적 물 사용량이며, 계측값의 총 개수는 약 1,065만 개였다. 그러나 이러한 계측된 데이터는 오류를 포함할 수 있기 때문에 이를 그대로 분석에 사용할 수는 없을 것으로 예상되었다. Fig. 1는 계측된 물 사용량 누적 데이터의 몇 가지 예를 제시하고 있다.

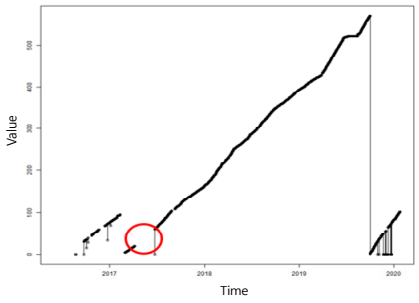
Fig. 1(a)의 경우 중간에 갑자기 계측이 제대로 되지 않거나 비정상적인 값이 나타나는 현상을 보여주고 있으며, Fig. 1(b)의 경우 누적 물 사용량이 중간에 초기화되는 현상을 보여준다. 이러한 계측데이터를 사용하기 위해서는 부적합한 계측 값을 삭제하고 초기화된 누적데이터를 보정해주는 작업 등이 필요할 것으로 판단되었다. 한편 Fig. 1(c)의 경우 측정 자체가 제대로 되지 않아서 시간에 따른 누적 물 사용량의 변화가 나타나지 않는데, 이러한 경우에는 해당 계측기에서 측정된 값을 수집 데이터로부터 제외할 필요가 있다고 판단되었다. 수집된 계측 데이터의 전처리를 위해서 먼저 대표적인 오류 유형을 정의하고 각각에 대한 보정 방법을 개발하였다. 가장 흔하게 나타나는 오류 중 하나는 원본 데이터에서 특정 날짜와 시간에서의 계측 값이 누락되는 것이었다. 이러한 형태의 오류를 N/A 오류로 정의하였다. Fig. 2(a)는 계측 데이터 안에 N/A 오류가 포함된 사례를 보여주고 있으며 Fig. 2(b)는 N/A 오류에 의하여 누적 물 사용량 데이터가 어떤 영향을 받는지를 보여주고 있다. N/A 오류는 데이터가 어느 순간에 수집되지 않는 경우이며, 장치의 오작동, 통신 과정에서의 손실 등에 기인하는 것으로 판단되었다. N/A 오류는 결측치이므로 이를 보정하기 위해서 앞뒤의 유효한 계측값을 이용하여 내삽(interpolation)하는 방식으로 보정을 하였다.

N/A가 발생할 경우 바로 뒤에 값의 경우 어느 시간대에 증가한 물 사용량인지 파악하기 어렵기 때문에 N/A 이후 첫 번째 값은 오류로 분류해야 한다. 예를 들어 N/A가 일주일동안 지속되었고 일주일 뒤에 100의 물 사용량이 증가했을 경우, 단순히 일주일 전의 값에서 N/A 이후 첫 번째 값을 뺀다면 한 시간에 100의 물 사용량이 증가하는 현상이 발생하며, 이는 원본 자료의 특성을 파악하는데 문제를 야기하여 추후 보정 및 예측 알고리즘

을 진행함에 있어 문제가 될 가능성이 있다. 따라서 보정 알고리즘으로서는 N/A 이후 첫 번째로 발생한 값은 모두 NA로 대체하는 방안을 적용하였다.

time	DVC_MTR	WRK_DT	WRK_HOR	BASE_ITGF
2016-05-31 23:00	133	2016-05-31	23	881.74
2016-06-01 0:00	133	2016-06-01	0	NA
2016-06-01 1:00	133	2016-06-01	1	NA
2016-06-01 2:00	133	2016-06-01	2	883.77
2016-06-01 3:00	133	2016-06-01	3	884.27
2016-06-01 4:00	133	2016-06-01	4	NA
2016-06-01 5:00	133	2016-06-01	5	885.3
2016-06-01 6:00	133	2016-06-01	6	NA
2016-06-01 7:00	133	2016-06-01	7	885.89
2016-06-01 8:00	133	2016-06-01	8	886.09
2016-06-01 9:00	133	2016-06-01	9	886.69
2016-06-01 10:00	133	2016-06-01	10	NA
2016-06-01 11:00	133	2016-06-01	11	892
2016-06-01 12:00	133	2016-06-01	12	NA
2016-06-01 13:00	133	2016-06-01	13	NA
2016-06-01 14:00	133	2016-06-01	14	898.85

(a)



(b)

Fig. 2. (a) Examples of N/A errors in the measured data (b) N/A errors in cumulative data (No. 639)

Fig. 3은 또 다른 오류 형태인 zero 오류를 보여주고 있다. 여기서는 물 사용량 자료가 누적값임에도 불구하고 값이 '0'으로 나타나는 오류를 말하며, 이는 N/A오류와 같은 오류로 볼 수도 있지만 1시간 단위 자료로 변환하는 과정에서 문제를 유발하기 때문에 다른 유형의 오류로 분류하였다. Zzero 오류는 N/A 오류와 같이 결측치로 취급하여 같은 방식으로 보정을 수행하였다.

time	DVC_M	WRK_DT	WRK_H	BASE_IT
2017-04-25 17:00	46	2017-04-25	17	573.593
2017-04-25 18:00	46	2017-04-25	18	0
2017-04-25 19:00	46	2017-04-25	19	0
2017-04-25 20:00	46	2017-04-25	20	573.682
2017-04-25 21:00	46	2017-04-25	21	0
2017-04-25 22:00	46	2017-04-25	22	573.739
2017-04-25 23:00	46	2017-04-25	23	0
2017-04-26 0:00	46	2017-04-26	0	573.787
2017-04-26 1:00	46	2017-04-26	1	573.787
2017-04-26 2:00	46	2017-04-26	2	573.797
2017-04-26 3:00	46	2017-04-26	3	573.808
2017-04-26 4:00	46	2017-04-26	4	0
2017-04-26 5:00	46	2017-04-26	5	0
2017-04-26 6:00	46	2017-04-26	6	573.81
2017-04-26 7:00	46	2017-04-26	7	574.014
2017-04-26 8:00	46	2017-04-26	8	574.038

Fig. 3. Examples of zero errors in the measured data

Fig. 4에 나타난 결과를 보면 정상 범위로 누적되던 물 사용량이 갑작스럽게 '0'이 아닌 감소된 값이 존재하는 것이 확인된다. 이는 시간 단위 물 사용량 자료로 변환하는 경우 문제가 되기 때문에 zero 오류의 경우와 마찬가지로 결측치로 취급한 후 내삽으로 보정하는 방법을 적용하였다.

2017-05-13 10:00	46	2017-05-13	10	605.38
2017-05-13 11:00	46	2017-05-13	11	605.567
2017-05-13 12:00	46	2017-05-13	12	605.61
2017-05-13 13:00	46	2017-05-13	13	605.619
2017-05-13 14:00	46	2017-05-13	14	605.671
2017-05-13 15:00	46	2017-05-13	15	14.538
2017-05-13 16:00	46	2017-05-13	16	605.7271
2017-05-13 17:00	46	2017-05-13	17	605.743
2017-05-13 18:00	46	2017-05-13	18	605.786
2017-05-13 19:00	46	2017-05-13	19	605.842
2017-05-13 20:00	46	2017-05-13	20	605.914
2017-05-13 21:00	46	2017-05-13	21	605.975
2017-05-13 22:00	46	2017-05-13	22	606.014
2017-05-13 23:00	46	2017-05-13	23	606.123
2017-05-14 0:00	46	2017-05-14	0	606.1461

Fig. 4. Examples of low-value errors in the measured data

Fig. 5에서는 Fig. 4와 반대로 정상 범위로 누적되던 물 사용량이 갑작스럽게 증가했다가 원래의 범위의 값으로 돌아오는 결과를 나타낸다. 이 경우에도 마찬가지로 오류로 판단하여 보정하도록 하였다.

2017-04-01 19:00	308	2017-04-01	19	536.298
2017-04-01 20:00	308	2017-04-01	20	536.33
2017-04-01 21:00	308	2017-04-01	21	536.343
2017-04-01 22:00	308	2017-04-01	22	536.3571
2017-04-01 23:00	308	2017-04-01	23	536.383
2017-04-02 0:00	308	2017-04-02	0	536.401
2017-04-02 1:00	308	2017-04-02	1	536.401
2017-04-02 2:00	308	2017-04-02	2	536.401
2017-04-02 3:00	308	2017-04-02	3	536.401
2017-04-02 4:00	308	2017-04-02	4	536.4041
2017-04-02 5:00	308	2017-04-02	5	536.525
2017-04-02 6:00	308	2017-04-02	6	536.64
2017-04-02 7:00	308	2017-04-02	7	784.75
2017-04-02 8:00	308	2017-04-02	8	536.8311
2017-04-02 9:00	308	2017-04-02	9	536.8311
2017-04-02 10:00	308	2017-04-02	10	536.8311
2017-04-02 11:00	308	2017-04-02	11	536.8311
2017-04-02 12:00	308	2017-04-02	12	536.8311
2017-04-02 13:00	308	2017-04-02	13	536.8311

Fig. 5. Examples of high-value errors in the measured data

Fig. 6은 결측이 발생한 후 계측이 재개되었을 때 누적 적산값임에도 불구하고 값이 초기화되거나 비정상적인 값을 보이는 경우의 예를 보여준다. 이러한 오류의 보정방법으로는 결측 발생 직전의 값으로 비정상적인 값을

대체하는 방법을 적용하였다.

2018-07-29 3:00	53	2018-07-29	3	63679.14
2018-07-29 4:00	53	2018-07-29	4	63679.14
2018-07-29 5:00	53	2018-07-29	5	63679.14
2018-07-29 6:00	53	2018-07-29	6	63679.14
2018-07-29 7:00	53	2018-07-29	7	63679.14
2018-07-29 8:00	53	2018-07-29	8	63679.14
2018-07-29 9:00	53	2018-07-29	9	63679.14
2018-07-29 10:00	53	2018-07-29	10	63679.14
2018-07-29 11:00	53	2018-07-29	11	NA
2018-07-29 12:00	53	2018-07-29	12	NA
2018-07-29 13:00	53	2018-07-29	13	NA
2018-10-17 23:00	NA	NA	NA	NA
2018-10-18 0:00	NA	NA	NA	NA
2018-10-18 1:00	NA	NA	NA	NA
2018-10-18 2:00	NA	NA	NA	NA
2018-10-18 3:00	NA	NA	NA	NA
2018-10-18 4:00	NA	NA	NA	NA
2018-10-18 5:00	53	2018-10-18	5	2558.49
2018-10-18 6:00	53	2018-10-18	6	2558.49
2018-10-18 7:00	53	2018-10-18	7	2558.49
2018-10-18 8:00	53	2018-10-18	8	2558.49

Fig. 6. Examples of initialization errors in the measured data

데이터의 계측이 제대로 수행되지 않으면 수집되어 저장된 데이터의 개수가 충분하지 않은 경우가 발생하며, 이러한 것을 개수 부족 오류라고 분류하였다. 데이터 개수가 부족하게 되면 이후 진행되는 머신러닝 혹은 딥러닝의 학습을 진행하는 것이 불가능하다고 판단되어 계측 데이터에서는 제외시키도록 하였다.

3.1.2 데이터 전처리 알고리즘 개발 및 적용

실시간으로 수집되는 물 사용량 데이터에서는 앞에서 언급하였던 오류들이 발생하며, 이러한 오류들은 딥러닝 혹은 머신러닝 기반의 물 사용량 보정 및 예측을 진행하는데 있어 학습을 방해하는 요소로 작용하게 된다. 따라서 오류들을 해결할 수 있는 데이터 품질관리를 위한 전처리 알고리즘을 개발하였으며, 물 사용량 데이터에 적용하여 이상치를 보정하는 작업을 진행하였다. 본 연구에서 도출한 데이터 전처리 알고리즘의 흐름은 Fig. 7과 같으며, 이에 의하면 데이터 전처리가 시작되면 먼저 데이터의 개수가 충분한지를 확인한 후 비정상적인 값이 나타나는지를 확인하고, N/A 오류와 zero 오류 등의 여부를 확인하고 다음으로 초기화 오류의 여부를 확인하고 다음으로 초기화 오류의 여부를 확인하게 된다.

각각의 경우에 대한 확인이 끝난 후 데이터는 앞서의 보정방법을 이용하여 처리되고 최종적으로는 시간별 물 사용 누적량의 형태로 결과가 만들어진다.

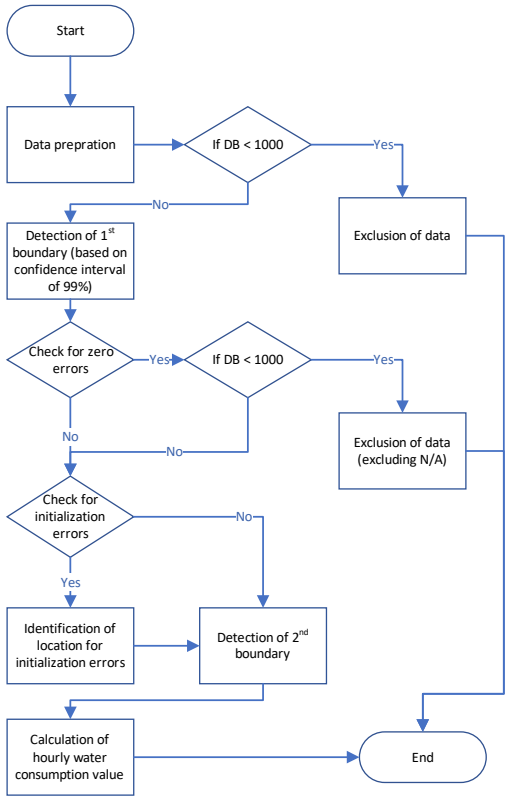


Fig. 7. Algorithms for data preprocessing

상기와 같이 물 사용량 보정 및 예측 모델을 개발하기 위해 물 사용량 데이터 분석과 데이터 품질관리 알고리즘을 적용하여 데이터 전처리 작업을 선행하였으나, 데이터 품질관리 알고리즘을 거치더라도 온전하게 거르지 못한 이상치 및 기계적 오류들이 포함되어 있는 것을 확인하였다. 따라서, 각 계측장비별 적산 물 사용량 그래프를 분석하여 본 연구에 사용될 데이터셋을 A급, B급, C급으로 분류하였다. 먼저 A급 데이터는 적산된 물 사용량의 데이터가 선형으로 적절하게 나타나 있고, 결측된 데이터의 개수가 적으며, 기울기가 급격하게 높아지거나 낮아지는 지점이 없는 데이터로 정의하여 각 장비별 계측치를 분류하였다. Fig. 8(a)와 Fig. 8(b)는 A급 데이터의 예를 보여준다.

한편, B급 데이터는 그래프가 어느 정도 선형을 따라가고, 결측된 개수가 많지 않은 데이터이며, C급 데이터는 그래프가 선형이지 않고 결측된 개수가 많으며 기울기가 급격하게 높아지거나 낮아지는 지점이 존재하는 데이터로 정의하여 이 기준에 따라 각 장비별 계측치를 분류하였다. 각각의 예는 Fig. 8(c)과 Fig. 8(d), Fig. 8(e)과 Fig. 8(f)에 나타내고 있다.

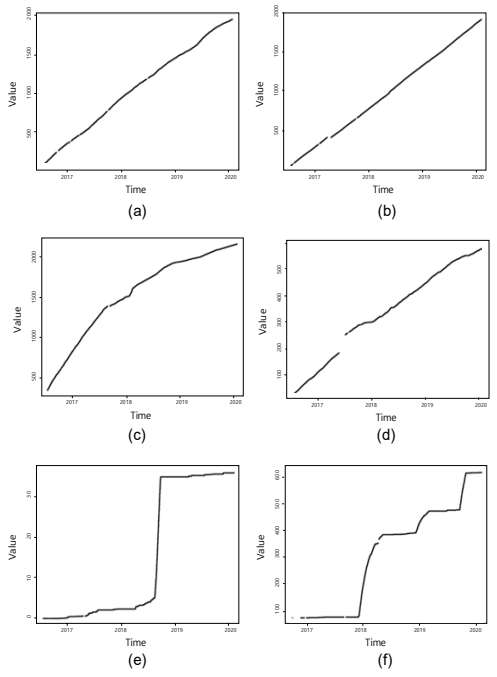


Fig. 8. Examples of data after quality management (a) Device No.46 (b) Device No.60 (c) Device No.300 (d) Device No.530 (e) Device No.797 (f) Device No.904

본 연구에서는 물 사용 패턴 분석 목적의 딥러닝 모델 개발을 위하여 각 계측장비별 A급 데이터(220셋)만을 사용하였으며, 전체 데이터셋 중 70 %는 학습데이터로, 30 %는 테스트데이터로 활용하였다.

3.2 데이터 품질관리를 위한 물 사용량 패턴 예측 모델 개발

3.2.1 군집화된 데이터를 이용한 예측 모델 개발

스마트 수도미터로부터 계측된 데이터가 결측되거나 수집된 값에 오류가 있을 때, 앞서 설명한 바와 같이 내삽 등의 방법을 통해서 전처리를 할 수 있다. 그러나 이러한 방법은 오·결측의 비율이 낮은 경우에만 적용이 가능하며 보정의 정확도가 낮은 한계점을 가진다. 따라서 이러한 한계를 극복하기 위해서는 물 사용량 패턴을 예측할 수 있는 딥러닝 모델을 개발하고 적용하는 것이 필요하며, 이를 위한 연구를 수행하였다.

먼저 앞서 확보된 A급 데이터에 대하여 각 계측장비별 물 사용량 최대값과, 평균값을 이용하여 계층적 군집분석을 수행하였다. 그 결과 Table 1에 나타난 바와 같이 총 7개의 군집으로 분류되었다. Fig. 9는 상기 7개 그룹 내 데이터간 묶인 순서와 거리를 시각화한 Dendrogram을 나타내고 있다. 본 연구에서는 우선적으로 상관계수가 가장 높은 Group 1의 데이터를 활용하여 LSTM 모델을 개발하고자 하였다. Group 1에 포함된 계측장비의 개수는 총 50개이었으며, 각각의 시간단위 물 사용량의 평균을 이용하였다.

Table 1. Classification of data collected from smart water meter devices to determine groups with correlation coefficients

Group	Correlation coefficient
Group 1	0.95
Group 2	0.92
Group 3	0.92
Group 4	0.92
Group 5	0.82
Group 6	0.94
Group 7	0.87

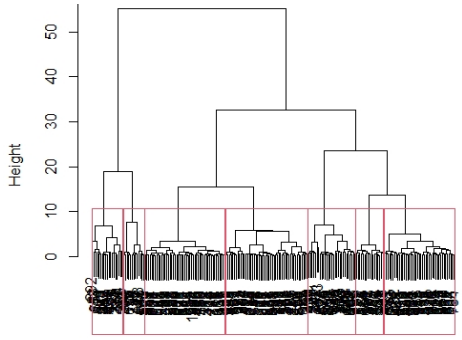


Fig. 9. Cluster dendrogram for classifying data collected from smart water meter devices

군집화 모형에 사용된 계수의 최적값은 Table 2에 정리되어 있다. LSTM 모델에서 활성화 함수는 대표적으로 시그모이드(Sigmoid), 쌍곡탄젠트(Tanh), ReLU(Rectified Linear Unit, 이하 ReLU) 등이 있으며, 그 중에서 가장 많이 사용되는 ReLU를 적용하였다. 최적화 기법으로는 Adam을 사용하고 학습률은 0.01로 하였으며, 훈련 반복횟수 (Epoch)는 16으로 진행하였다. 이때 손실은 MSE으로 계산하였으며, Group 1의 데이터에 대하여 적용한 결과는 Fig. 10에 제시되어 있다. 학습횟수를 13까지 했을때는 손실이 감소하는 경향을 보였으나 그 이상에서는 오히려 증가하는 경향을 보여 과대적합의 경향이 나타나는 것으로 판단되었다.

Table 2. Optimum parameters for group model

Parameter	Value
Activation	Relu
Epoch	15
Optimizer	Adam
Learning rate	0.01
Loss	Mean squared error

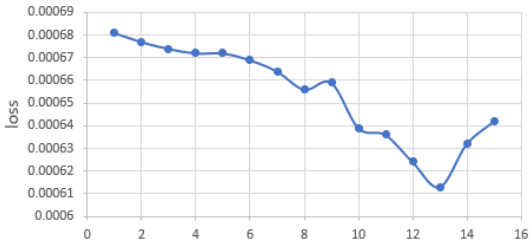


Fig. 10. Results on loss analysis for optimization learning range in group 1

Fig. 11은 Group 1의 데이터에 대하여 LSTM 모델을 적용한 결과를 보여준다. 파란 색으로 표시된 것이 실제 계측값이고 주황색으로 표시된 것이 모델의 예측값이다. 계측값과 예측값은 비슷한 경향을 보이기는 하지만 큰 차이를 보이는 것으로 나타났으며, 특히 극치 사상을 모델에서 잘 재현하지 못하는 것을 확인할 수 있었고 상관성은 25 %에 불과하였다.

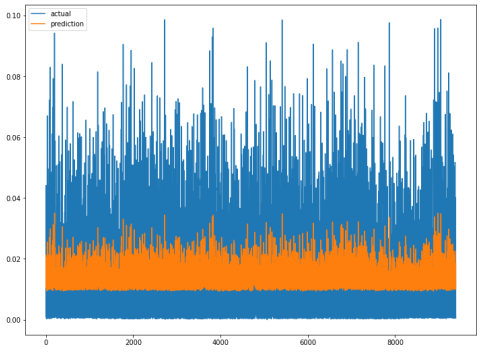


Fig. 11. Comparison of observed data and predictions by LSTM model for group 1 (blue line: observed data; orange line: model predictions)

이러한 결과가 나온 것은 군집화로 인하여 각 계측 데이터의 노이즈가 가중되었기 때문으로 추정되었으며, 따라서 Group별로 모델을 학습하여 적용하는 것은 어려울 것이라는 결론을 얻었다. 따라서 각 계측 장비별로 모델을 개발하는 것으로 방법을 바꾸어 적용하였다. 유효한 데이터를 얻은 모든 계측기에 대하여 개별적으로 모델을 적용하였으나 여기서는 그 예로서 몇 가지 경우에 대하여 다루고자 한다. Fig. 12는 스마트 수도미터 장치 51번에 대하여 모델 학습을 시켰을 때 학습횟수에 따른 손실 변화를 제시하고 있다. 여기서 LSTM 모델의 최적 매개변수는 Table 2에 나온 것과 동일하게 적용하였다. 그래프에 나타난 바와 같이 학습횟수에 따라 손실 값이 감소하는 경향을 보였으나 15회에 도달해도 여전히 높은 값을 보이는 것으로 나타났다. 해당 계측기의 수집 데이터에 대하여 LSTM 모델을 적용한 결과(Fig. 13)에서도 계측값(파란색)과 예측값(주황색)은 큰 차이를 보이는 것으로 나타났으며 상관성은 앞서의 경우보다 약간 증가한 28 %로 계산되었다.

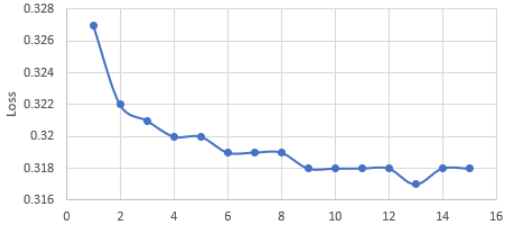


Fig. 12. Results on loss analysis for optimization learning range for device No.51

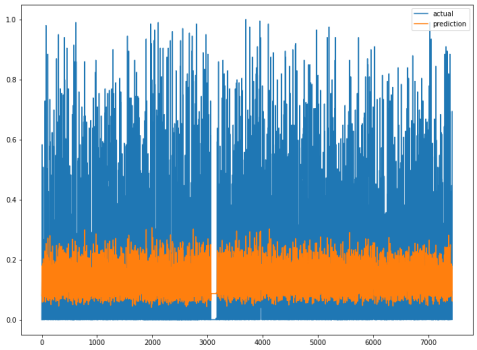


Fig. 13. Comparison of observed data and model predictions by LSTM model for device No.51 (blue line: observed data; orange line: model predictions)

3.2.2 웨이블릿을 활용한 장비별 예측 모델 고도화

앞에서 수행한 연구 결과 물 사용량 예측의 정확도를 향상시키기 위해서는 수집된 데이터에 대한 추가적인 처리가 필수적이라는 결론을 도출할 수 있었다. 따라서 여기서는 웨이블릿 변환을 활용하여 수집 데이터를 주기성 성분과 비주기성 성분으로 분해한 후 모델을 학습하고자

하였다. Fig. 14에 나타난 바와 같이 계측 데이터의 웨이블릿 변환 결과 8개의 세분화 성분(D1~D8)과 1개의 근사 성분(A8)으로 분류할 수 있었다. 각각의 특성을 분석한 결과 D1의 경우 주기성의 거의 없으며, D2의 경우 약한 주기성, D3의 경우 확연하게 주기성을 띄는 것을 확인할 수 있었다. 따라서 모델 학습에서는 D1을 제외한 주기성을 가지는 데이터(D2~A8)를 이용하였다.

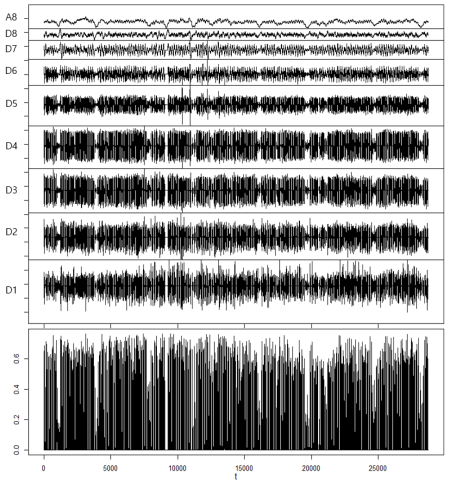


Fig. 14. Results on wavelet transformation for collected data from smart water meter No.46

웨이블릿 변환 후에는 얻은 주기성을 가진 데이터를 활용하여 LSTM 모델을 적용하였으며, 이 때 모델의 최적변수는 학습횟수(30)을 제외하고는 Table 2에 나온 조건을 적용하였다. Fig. 15에 제시된 바와 같이 학습횟수가 4회 이상인 경우 손실 값이 낮은 값으로 유지되는 것을 확인할 수 있었다. 또한 Fig. 16에 나타난 바와 같이 계측값과 예측값이 잘 부합하는 결과를 얻을 수 있었다. 특히 전반적인 경향성 뿐 아니라 극치 사상을 모델에서 잘 재현할 수 있는 것을 볼 수 있다.

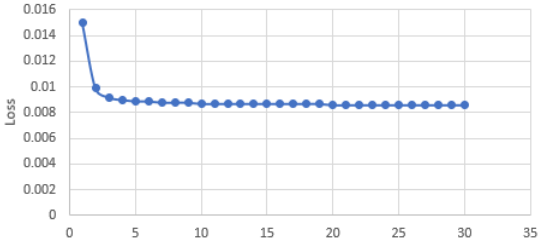


Fig. 15. Results on loss analysis on periodic components for device No.46

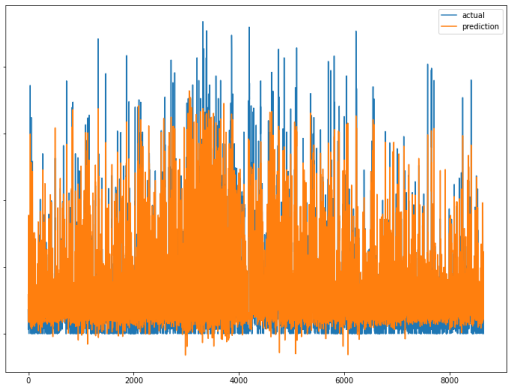


Fig. 16. Comparison of observed data and model predictions by LSTM model for device No.46 after wavelet transformation (blue line: observed data; orange line: model predictions)

LSTM 모델 외에 XGBoost 모델을 활용하여 데이터 분석을 수행하였으며 그 결과가 Fig. 17에 제시되어 있다. 여기서는 파란색으로 표시된 것이 계측값이며 검은색으로 표시된 것이 모델 계산값(예측값)이다. 앞서의 LSTM 모델과 유사하게 XGBoost 모델도 계측값에 대한 부합성이 높은 결과를 얻을 수 있었다.

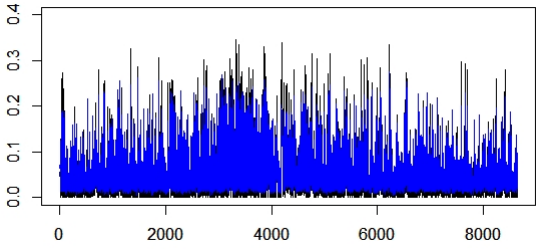


Fig. 17. Comparison of observed data and model predictions by XGBoost model for device No.46 after wavelet transformation (blue line: observed data; black line: model predictions)

3.2.3 LSTM 모델과 XGBoost 모델의 비교

LSTM 모델과 XGBoost 모델의 성능을 비교하기 위해서 다음의 3가지 경우에 대한 연구를 수행하였다.

① CASE 1: 웨이블릿 변환을 진행하지 않는 단일 계측장비 예측 모델
② CASE 2: 노이즈 성분과 주기성 성분을 각각 학습시킨 후 합산하여 예측하는 모델

③ CASE 3: 노이즈 성분을 제외하고 주기성 성분만을 학습시키는 예측 모델

Table 3과 Table 4는 각각 LSTM 모델과 XGBoost 모델을 적용하였을 때 위의 3가지 Case에 대한 RMSE 값을 나타낸 것이다. 두 가지 모델이 공통적으로 Case 1과 Case 2에 비해서 Case 3의 RMSE 값이 낮은 것으로 나타났으며, 노이즈 성분을 제외하고 주기성 성분만을 학습시키는 것이 예측 능력이 향상되는 것을 확인할 수 있다. 각 Case 별로 LSTM 모델과 XGBoost 모델을 비교해보면 성능 수준은 거의 비슷하였으나, XGBoost 모델의 RMSE가 낮은 것으로 나타났다.

Table 3. Verification results for LSTM models (case 1, case 2, and case 3)

Device No.	Case 1 RMSE	Case 2 RMSE	Case 3 RMSE
46	0.06199	0.06271	0.03457
51	0.04015	0.04163	0.02339
56	0.03551	0.03715	0.02011
58	0.02364	0.02450	0.01305
60	0.05605	0.05871	0.02898
62	0.05630	0.05935	0.03186
66	0.05285	0.05527	0.03219
67	0.09288	0.09708	0.05623

Table 4. Verification results for XGBoost models (case 1, case 2, and case 3)

Device No.	Case 1 RMSE	Case 2 RMSE	Case 3 RMSE
46	0.06322	0.06266	0.03236
51	0.04041	0.04063	0.02121
56	0.03556	0.03571	0.01853
58	0.02383	0.02383	0.01231
60	0.05447	0.05522	0.02872
62	0.05688	0.05790	0.03027
66	0.05316	0.05347	0.02832
67	0.08795	0.08952	0.04759

4. 결론

본 연구는 스마트 수도미터에서 수집되는 데이터를 자동으로 전처리하여 품질을 향상시키고, 이를 활용하여 물 사용량 패턴을 분석하고 예측하기 위한 모델을 개발하기 위하여 수행되었다. 실제 스마트 수도미터에서 수집된 원본 데이터를 분석하여 오류 유형을 구분한 후 각각에 대한 처리 방법을 개발하여 적용하였다.

물 사용량 분석과 예측을 위하여 LSTM 모델과 XGBoost 모델을 적용하였다. 그러나 수집한 데이터를 군집분석한 후에 바로 모델을 적용한 경우에는 상관성이 낮고 예측 능력이 부족한 결과를 얻었다. 따라서 이를 개선하기 위해서 웨이블릿을 적용하여 비주기성 노이즈를 제거한 데이터를 생성하였고 이를 활용한 결과 모델의 예측능력을 크게 향상시킬 수 있었다.

본 연구에서 개발된 스마트 수도미터 데이터 전처리 기법과 물 사용량 예측 모델은 수집된 데이터의 품질을 향상시키고 활용성을 높일수 있으므로 다양한 용도로 적용될 수 있을 것으로 판단된다. 특히 본 연구결과는 스마트시티에서 수집되는 물관리 데이터의 분석과 활용을 위하여 효과적으로 활용될 것으로 기대된다.

References

[1] Matt Hamblen, Just what IS a smart city?, Available From: https://www.computerworld.com (accessed Nov. 11, 2022)

[2] McLaren, Duncan; Agyeman, Julian, "Sharing Cities: A Case for Truly Smart and Sustainable Cities", MIT Press. 2015, ISBN 9780262029728.

[3] Boyd Cohen, The 3 Generations Of Smart Cities, Available From: https://fastcompany.com (accessed Nov. 11, 2022)

[4] Peris-Ortiz, Marta; Bennett, Dag R.; Yabar, Diana Perez-Bustamante, Sustainable Smart Cities: Creating Spaces for Technological, Social and Business Development, Springer, 2016, ISBN 9783319408958.

[5] Allied Market Research, Smart Cities Market by Component (Hardware, Software, and Service) and Functional Area (Smart Infrastructure, Smart Governance and Smart Education, Smart Energy, Smart Mobility, Smart Healthcare, Smart Buildings, and Others): Global Opportunity Analysis and Industry Forecast, 2021-2030, Market analysis report, 2022.

[6] M. H. Panahi Rizi, S.A. Senob, "A systematic review of technologies and solutions to improve security and privacy protection of citizens in the smart city", Internet of Things Volume 20, Nov. 2022, DOI: https://doi.org/10.1016/j.iot.2022.100584

[7] Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. J ournal of urban technology, 22(1), 3-21. Feb. 2015, DOI: https://doi.org/10.1080/10630732.2014.942092

[8] Y. Lee, “Evaluation of Smart Water Management System of Water Infrastructure in a Sustainable Smart City”, Ph.D. Thesis, Korea University, Feb. 2021.

[9] L. Fabbiano, G. Vacca, G. Dinardo, “Smart water grid: A smart methodology to detect leaks in water distribution networks” Measurement Volume 151, 107260, Feb. 2020,
DOI: https://doi.org/10.1016/j.measurement.2019.107260

[10] A. Cominola, M. Giuliani, D. Piga, A. Castelletti, A.E. Rizzoli, “Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review” Environmental Modelling & Software Volume 72, 198-214, Oct. 2015,
DOI: https://doi.org/10.1016/j.envsoft.2015.07.012

[11] C.D. Beal, J. Flynn, "Toward the digital water age: Survey and case studies of Australian water utility smart-metering programs," Utilities Policy, Vol. 32(C), 29-37. Jan. 2015,
DOI: https://doi.org/10.1016/j.jup.2014.12.006

[12] K. P. Wai, M. Y. Chia, C. H. Koo, Y. F. Huang, W. C. Chong, “Applications of deep learning in water quality management: A state-of-the-art review” Journal of Hydrology Volume 613, Part A, 128332, Oct. 2022,
DOI: https://doi.org/10.1016/j.jhydrol.2022.128332

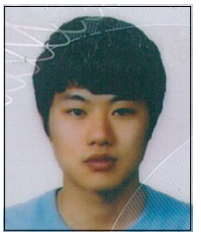
[13] G. Fu, Y. Jin, S. Sun, Z. Yuan, D. Butler, “The role of deep learning in urban water management: A critical review” Water Research Volume 223, 1, 118973, Sep. 2022
DOI: https://doi.org/10.1016/j.watres.2022.118973

이 상 호(Sangho Lee) [정회원]

- 1999년 2월 : 서울대학교 공업화학과 (공업화학박사)
- 1999년 3월 ~ 2003년 3월 : Northwester Univ. Senior Researcher
- 2003년 4월 ~ 2011년 2월 : 한국건설기술연구원 책임연구위원
- 2011년 3월 ~ 현재 : 국민대학교 교수

〈관심분야〉
막여과 공정, 스마트시티, 인공지능, 모델링

이 주 영(Juyoung Lee) [정회원]



- 2018년 2월 : 국민대학교 건설시스템공학과 (건설시스템공학학사)
- 2020년 2월 : 국민대학교 건설시스템공학과 (환경공학석사)
- 2021년 3월 ~ 현재 : 국민대학교 환경공학 박사과정

〈관심분야〉
막여과 공정, 스마트시티, 인공지능, 하폐수재이용

윤 정 환(Junghwan Yun) [정회원]

- 2002년 2월 : 호서대학교 물리학과 (물리학학사)
- 2015년 8월 ~ 2021년 2월 : 인하대학교 토목공학과 (토목공학석사)
- 2015년 4월 ~ 현재 : ㈜뉴컨 대표이사

〈관심분야〉
인공지능, 머신러닝, 빅데이터분석