J. Korea Water Resour. Assoc. Vol. 56, No. 7 (2023), pp. 439-449
doi: 10.3741/JKWRA.2023.56.7.439
pISSN 2799-8746
eISSN 2799-8754

Multidimensional data generation of water distribution systems using adversarially trained autoencoder

Kim, Sehyeong[a] · Jun, Sanghoon[b] · Jung, Donghwi[c]*

[a]Master's Degree Researcher, Department of Civil, Environmental and Architectural Engineering, Korea University, Seoul, Korea
[b]Postdoctoral Research Associate, Hyper-converged Forensic Research Center for Infrastructure, Korea University, Seoul, Korea
[c]Associate Professor, School of Civil, Environmental and Architectural Engineering, Korea University, Seoul, Korea

Paper number: 23-024
Received: 19 May 2023; Revised: 20 July 2023; Accepted: 21 July 2023

Abstract

Recent advancements in data measuring technology have facilitated the installation of various sensors, such as pressure meters and flow meters, to effectively assess the real-time conditions of water distribution systems (WDSs). However, as cities expand extensively, the factors that impact the reliability of measurements have become increasingly diverse. In particular, demand data, one of the most significant hydraulic variable in WDS, is challenging to be measured directly and is prone to missing values, making the development of accurate data generation models more important. Therefore, this paper proposes an adversarially trained autoencoder (ATAE) model based on generative deep learning techniques to accurately estimate demand data in WDSs. The proposed model utilizes two neural networks: a generative network and a discriminative network. The generative network generates demand data using the information provided from the measured pressure data, while the discriminative network evaluates the generated demand outputs and provides feedback to the generator to learn the distinctive features of the data. To validate its performance, the ATAE model is applied to a real distribution system in Austin, Texas, USA. The study analyzes the impact of data uncertainty by calculating the accuracy of ATAE's prediction results for varying levels of uncertainty in the demand and the pressure time series data. Additionally, the model's performance is evaluated by comparing the results for different data collection periods (low, average, and high demand hours) to assess its ability to generate demand data based on water consumption levels.

Keywords: Water distribution systems, Multidimensional data generation, Generative adversarial networks, Adversarially trained autoencoder

적대적 학습 기반 오토인코더(ATAE)를 이용한 다차원 상수도관망 데이터 생성

김세형[a] · 전상훈[b] · 정동휘[c]*

[a]고려대학교 건축사회환경공학과 석사후연구원, [b]고려대학교 초융합건설포렌식연구센터 박사후연구원, [c]고려대학교 건축사회환경공학부 부교수

요 지

최근 계측 기술의 발전으로 압력계와 유량계 등 다양한 센서를 설치하여 상수도관망의 상태를 효과적으로 파악할 수 있게 되었으나, 도시가 광범위하게 개발됨에 따라 계측 신뢰도에 영향을 미치는 변수는 다양해지고 있다. 특히 상수도관망 분석에 중요한 영향력을 가지는 수요 데이터의 경우 직접 계측의 난이도가 높고 결측이 발생하기 쉬운 것으로 알려져 데이터 생성의 중요도가 증가하고 있다. 본 논문에서는 상수도관망에서 누락된 데이터를 정확하게 생성하기 위해 생성적 딥러닝 모델에 기반한 적대적 학습 기반 오토인코더(ATAE) 모델을 제안한다. 제안된 모델은 판별 신경망과 생성 신경망의 두 가지 신경망의 적대적 학습을 사용하여 압력 데이터로부터 수요 데이터를 생성한다. 학습이 완료된 ATAE 모델의 생성 신경망은 관망의 계측되는 압력 데이터가 존재하는 경우, 그로부터 추정된 관망 수요 데이터를 제공할 수 있다. ATAE 모델은 미국 텍사스주 오스틴의 실제 상수도망에 적용되어 성능이 검증되었다. 수요 및 압력 시계열 데이터의 불확실성 정도에 따른 ATAE 예측 결과의 정확도를 비교하여 데이터 불확실성의 영향을 분석하였으며, 또한 수요 수준에 따른 데이터 수집 기간별 생성 결과를 비교하여 이에 따른 데이터 생성 성능을 검토하였다.

핵심용어: 상수도관망, 다차원 데이터 생성, 생성적 적대 신경망, 적대적 학습 기반 오토인코더(ATAE)

***Corresponding Author.** Tel: +82-2-3290-4869
E-mail: sunnyjung625@korea.ac.kr (Jung, Donghwi)

© 2023 Korea Water Resources Association. All rights reserved.

1. 서 론

최근 계측 기술의 발전으로 압력계 및 유량계 등 다양한 센서를 설치하여 상수도관망의 상태를 효과적으로 파악할 수 있게 되었다. 그러나 도시가 집중적이고 광범위하게 개발되면서 상태 데이터를 확보해야 하는 중요 지점의 수가 증가했고, 이로 인해 상수도관망 분석의 난이도가 상승해왔다. 게다가 장비 오작동과 여러 환경 변수의 변화 등 센서 기반 계측 시스템의 신뢰도를 하락시키는 여러 원인들로 인해 온전한 데이터를 구축하는 것이 어려워지고 있다. 이러한 계측 시스템의 한계로 발생하는 상수도관망 가용 데이터의 결측값 문제를 해결하기 위해, 과거 여러 연구에서 데이터 보완(imputation) 등 다양한 기술을 위해 노력해왔다.

특히 수요 추정 분야에서는 위에 설명한 계측 난이도 및 신뢰도 문제로 인해 데이터 보완 연구가 최근 활발히 수행되고 있다(Bragalli *et al*., 2019; Zanfei *et al*., 2022). 주로 제한 계측 구역(Districted Metered Area, DMA)의 유입 유량에 대한 통계적 분석을 활용하여 생성에 사용하며, Brentan *et al*. (2018)의 경우 기후 데이터로부터 수요를 예측하기도 하였다. Jun *et al*. (2021)은 누락 데이터 보완에 가장 많이 사용되는 세 가지 방법인 분포 샘플링(distribution sampling), 0 치환(zero imputation), 과거 평균(historical mean)의 효과를 비교하였다. 또한 칼만 필터 기반의 최인접(nearest neighbor) 회귀 분석이나 랜덤 포레스트 기반 추정과 같은 머신러닝 기법도 통계적 추정 모델로써 활용되고 있다(Kabir *et al*., 2020; Rodríguez *et al*., 2021).

기존에 개발되었던 데이터 보완 기술의 대표적인 한계점은 변수 간 선형 관계를 가정하는 통계 기법에 의존하는 경우가 많다는 점이다. 그러나 현실의 대다수 관망 시스템은 선형 모델이 포착할 수 없는 복잡한 비선형 관계를 보이며, 특히 예측할 수 없는 방식으로 시스템 거동에 영향을 받는 상수도관망의 경우 그 특징이 두드러진다. 이러한 시스템의 누락 데이터를 정확도 높게 추론하기 위해서는 보다 정교한 비선형 모델이 필요하다. 생성적 딥러닝 모델은 다차원 데이터의 생성 및 변환에 있어 가장 중요한 발전 중 하나로 평가받고 있는 기계학습 기술로 상수도관망의 누락 데이터를 높은 정확도로 보완할 수 있다. 하지만 생성적 딥러닝 모델을 데이터 보완을 위한 전치 등에 활용한 연구는 부족한 것으로 파악된다.

생성적 딥러닝 모델은 최근 딥러닝 기술의 발전과 함께 적대적 학습(adversarial training)이 도입되면서 본격적으로 활용되기 시작했다. 적대적 학습은 2인 미니맥스 게임(two-player minimax game)의 원리에 기반해 동일한 환경에서 두 개의 신경망을 서로 대결시켜 훈련하는 방식으로, 생성적 적대 신경망(Generative Adversarial Network, GAN)(Goodfellow *et al*., 2014)을 통해 매우 사실적인 데이터를 생성할 수 있는 학습 방식임이 입증되었다. 이러한 모델은 최근 이미지 및 음성 인식부터 자연어 처리에 이르기까지 다양한 분야에 적용되고 있다(Shrivastava *et al*., 2017; Sun *et al*., 2018).

따라서 본 연구에서는 적대적으로 학습된 오토인코더(Adversarially Trained Autoencoder, ATAE)에 기반하여 절점의 압력 시계열 데이터로부터 정보를 학습하여 수요 시계열 데이터를 생성하는 다차원 상수도관망 데이터 변환 모델을 제안한다. 개발된 ATAE 모델은 미국 텍사스 오스틴에 있는 실제 상수도관망에 적용되어 그 성능을 검토하였다. 본 논문에서 검증을 위해 사용되는 데이터는 상수도관망 수리해석 프로그램인 EPANET을 사용하여 확보되나, 실제 시스템 적용 시의 현실성을 고려하기 위하여 데이터에 불확실성을 부여하였다. 수요량 및 압력 시계열 데이터의 불확실성 정도에 따른 ATAE의 예측 결과의 정확도를 비교하여 데이터 불확실성의 영향을 분석하였다. 또한, 관망 전체 수요 수준에 따라 데이터가 수집된 시간대별 결과를 비교함으로써 수요량 변화에 따른 ATAE의 성능 변화 추이까지 검토하였다.

2. 방법론

2.1 생성 신경망과 오토인코더

생성(generative) 신경망은 학습 데이터와 유사한 새로운 데이터를 생성할 수 있는 인공신경망(Artificial Neural Network, ANN)의 한 종류로, 입력되는 학습 데이터의 분포를 학습한 후 그 정보를 바탕으로 새로운 샘플을 생성한다.

생성 신경망의 핵심 구성 요소 중 하나는 인코딩과 디코딩이다. 인코딩 신경망은 입력된 데이터 포인트 x를 가져와 저차원의 잠재 공간(latent space) h에 시그모이드(sigmoid)나 하이퍼볼릭 탄젠트(hyperbolic tangent)와 같은 함수 s를 기반으로 가중치 W와 편향 b로 맵핑(mapping)한다(Eq. (1)). ReLU 및 SwiGLU 등의 활성함수는 항상 양수의 값만 출력할 수 있기 때문에 생성 연산에 제약이 발생할 가능성이 존재한다. 잠재공간 h에 맵핑된 x는 디코딩 신경망에 의해 다시 입력 데이터 공간의 점 $\hat{x}$로 맵핑된다(Eq. (2)). 인코더와 디코더를 함께 훈련함으로써, 모델은 입력 공간에서 잠재 공간으로 데이터를 맵핑하는 방법을 학습하여 입력 데이터의 기본 분포를 효과적으로 학습하게 된다.

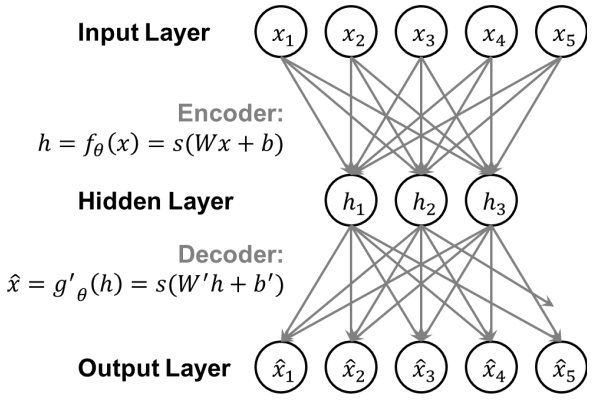


Fig. 1. Basic architecture of autoencoder model with bottleneck structure

$$h = f_\theta(x) = s(Wx + b) \tag{1}$$

$$\hat{x} = g'_\theta(h) = s(W'h + b') \tag{2}$$

Fig. 1에 도시된 것과 같은 오토인코더(autoencoder)는 입력과 출력이 동일한 형태의 인코더-디코더 구조를 갖는 신경망이다. 본래의 오토인코더는 출력 데이터를 입력 데이터와 동일하게 재생성하는 것이 목적으로 자기지도(self-supervised) 학습이 가능하여, 명시적인 레이블이 없이도 입력 데이터를 재구성하는 방법을 배울 수 있다. 해당 모델에서 파생되어 효과적인 성능을 입증한 생성 신경망 중 하나로 가변 오토인코더(Variational Autoencoder, VAE)가 존재하는데, 이는 확률론적 해석을 가진 잠재 표현을 학습하는 오토인코더이다. VAE는 잠재 공간에 대한 분포를 학습하여 작동하지만, 데이터 재구성시의 매개변수를 재매개화하면서 새로운 샘플을 생성할 수 있다. 본 연구에서 제안하는 ATAE 모델은 이러한 재매개화 과정의 학습을 적대적 학습으로 대체하여 활용한다고 볼 수 있다.

2.2 적대적 학습

적대적 학습(adversarial training)은 학습시키고자 하는 신경망과 또 다른 신경망을 대결시키면서 훈련을 진행하는 방식으로, 생성 신경망의 경우에는 판별(discriminative) 신경망과 훈련하게 된다. 해당 메커니즘은 2014년 Goodfellow *et al*. (2014)에서 GAN 모델을 통해 처음 제안되었으며, 생성 모델 학습 방식의 패러다임을 전환할 정도로 높은 성능을 인정받았다. GAN 모델은 생성 신경망과 판별 신경망으로 구성된 하나의 프레임워크로, 생성 신경망은 새로운 데이터 샘플을 생성하고 판별 신경망이 실제 샘플과 생성된 샘플을 구분하여 판단한다. GAN의 최종 목표는 판별 신경망이 제공하는 피드백을 이용해 실제 데이터와 유사한 샘플을 생성하도록 생성 신경망을 훈련시키는 것이다. 생성 신경망과 판별 신경망이 서로 대결함으로써 학습하기 때문에 해당 과정을 적대적 훈련이라 한다.

GAN의 적대적 훈련은 생성 신경망과 판별 신경망 간의 2인 미니맥스 게임 원리(Eq. (3))에 기반하여 진행된다. Eq. (3)에서, G는 생성 신경망, D는 판별 신경망, x는 실제 데이터 샘플, z는 생성 신경망의 입력 노이즈 벡터를 나타내며, $p_{data}(x)$는 실제 데이터의 분포, $p_z(z)$는 생성 신경망의 입력 노이즈 분포를 뜻한다. 생성 신경망은 판별 신경망을 속일 수 있는 데이터를 생성하려고 하는 반면, 판별 신경망은 실제 데이터와 생성 데이터를 구별하려 한다. 훈련 과정에서 생성 신경망과 판별 신경망 모두 업데이트되지만, 일정 수준 이상 훈련된 생성 신경망은 갈수록 실제 데이터와 유사한 생성 데이터를 만들어내며 판별 신경망의 성능을 저하시킨다. 적대적 훈련 과정은 판별 신경망이 실제 데이터와 생성 데이터를 구분하는 능력을 완전히 상실할 때까지 지속된다.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{x\sim p_z(z)}[\log(1 - D(G(x)))] \tag{3}$$

2.3 적대적 학습 기반 오토인코더(ATAE)

본 연구에서 개발한 ATAE 모델(Fig. 2)은 시계열 압력 이미지의 정보를 학습하여 시계열 수요 이미지를 생성한다. 해당 모델은 생성 신경망(G)에 해1당하는 오토인코더와 이를 판단하는 판별 신경망의 두 신경망 모델로 구성된다. 오토인코더는 합성곱(convolution)을 통해 입력된 압력 이미지의 특성을 추출하여 잠재 공간에 저장하고, 전치 합성곱(transposed convolution)을 이용해 차원 축소된 압력 이미지 특성을 기반으로 수요 이미지를 생성하는 병목 구조를 가진다. 판별 신경망(D)은 오토인코더의 출력으로써 생성된 수요 이미지를 평가하고 피드백을 제공하여 오토인코더의 매개변수를 업데이트하도록 한다.

제안된 모델은 적대적 학습 방법을 통해 입력 데이터와 출력 데이터 간의 강력하고 정확한 맵핑이 가능하도록 훈련된다. 오토인코더와 판별 신경망은 동시에 각각의 오차를 감소시키는 방향으로 학습한다. 판별 신경망은 실제 수요 이미지와 오토인코더가 생성한 수요 이미지를 구별하도록 훈련되며, 오토인코더는 판별 신경망을 속일 수 있는 수요 이미지를 생성하도록 훈련된다. 이 과정에서 오토인코더는 판별 신경망이 제시된 이미지를 보고 도출한 구별 결과에 따라 피드백을 제공받으며, 이러한 피드백에 따라 오토인코더 내 가중치와 편향 등의 매개변수가 조정된다.

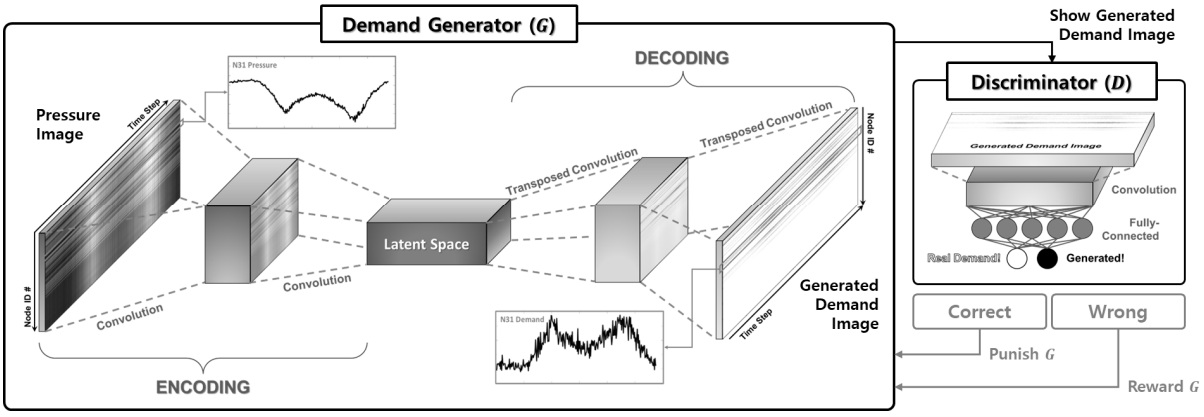


Fig. 2. Overview of adversarially trained autoencoder (ATAE) model

본 연구에서 제안하는 ATAE 모델은 기존의 GAN 모델과 유사하게 작동하나 이미지 생성 시드(seed)로써 정규 분포의 임의 난수 대신 압력 데이터를 활용하여 수요 데이터를 생성한다는 것에 차이점이 있다. GAN의 경우 크기가 작고 정규 분포 기반 임의 난수값으로 구성된 시드로부터 전치 합성곱 연산을 반복하여 적용해 차원을 단조 증가시키는 형태로 이미지를 생성한다. 반면 ATAE 모델의 경우 크기가 동일한 압력 데이터를 시드로써 사용하며, 이를 합성곱 연산을 통해 차원이 축소된 잠재 임베딩으로 변환하고 이를 다시 전치 합성곱 연산을 통해 차원을 재증가시켜 이를 수요 데이터로 변환할 수 있다. 이러한 원리를 통해 ATAE 모델은 기존 GAN 모델에서는 불가능한 이미지 변환 작업을 가능케 한다.

2.4 학습 데이터 구축

본 연구의 압력 및 수요에 대한 학습 데이터 구축 파이프라인은 Fig. 3과 같다. 각 패턴은 자정에 시작해 24시간(1,440분) 이후 자정에 종료된다(1-day pattern). 또한, 데이터를 계측할 경우 측정치 오류가 발생하는 실제 데이터의 특성을 반영하기 위해 데이터 불확실성을 고려하였다(Fig. 3(a)).

실제 수용가의 물 사용 데이터는 일별로 발생하는 변동성으로 인해 일정 수준의 불확실성을 가진다. 물 사용량이 큰 시점일수록 불확실성도 같이 증가하며, 반대의 경우 감소한다. 본 연구에서는 이를 구현하기 위해 역누적분포함수(inverse Cumulative Distribution Function, inverse-CDF) 샘플링을 이용하였다. 해당 방식은 다양한 상수도관망 분석 연구에서 데이터를 생성하는 경우 범용적으로 사용되었으며(Jung *et al*., 2010; Jang *et al*., 2022), 상수도관망과 유사하게 사용자의 불확실성을 고려해야 하는 네트워크인 전력망 데이터의 생성 시에도 사용되었다(Bagheri *et al*., 2015). 역누적분포함수 샘플링은 모든 확률 분포의 누적분포함수가 균등분포(uniform distribution) U를 따른다(Eq. (4))는 성질을 이용하여, U에 누적분포함수 F의 역함수를 취해 특정 확률 분포 $f(x)$를 따르는 확률변수 X를 샘플링하는 방법이다(Eq. (5)). 본 연구에서는 역누적분포함수 샘플링의 확률 분포 $f(x)$를 정규분포로 가정하였다. 각 절점에 실제로 삽입되는 시점별 수요 패턴값을 결정할 때, 해당 시점 패턴값의 평균을 평균 μ로, μ의 일정 비율(예, 10%) 값을 분산 σ로 갖는 정규분포 $N(\mu,\sigma)$를 이용하여 난수로 생성하였다.

$$F(X) \equiv U \sim Unif(0,1), \quad X \sim f(x) \tag{4}$$

$$F(X) \bullet F^{-1}(X) = X = F^{-1}(U) \tag{5}$$

생성된 관망 수요 데이터의 값은 일관된 기준을 통한 정량 비교를 위해 정규 분포 표준화(normal-distributed standardization) 이후 모두 흑백 이미지의 픽셀 값인 0 이상 255 이하의 값으로 최대-최소 정규화(min-max normalization)되어 사용된다. 이때 표준화 과정에서 사용되는 평균 및 표준편차는 이벤트 전체(구축된 이미지 데이터 장수와 동일)의 평균과 표준편차로, 모든 이미지를 표준화할 때 모두 동일한 값으로 적용한다. 이러한 전처리를 거친 이후에도 역누적분포함수 샘플링을 통해 부여한 불확실성을 유지하도록 하기 위해 일관된 통계치를 사용한 것이며, 이와 같은 방식으로 표준화된 데이터는 불확실성에 의해 가지는 상대적 크기를 유지한다. 다시 말해, 데이터의 값이 다른 경우 표준화하더라도 데이터가 속한 이미지에 관계 없이 해당 값들 간의 대소 관계는 유지된다.

위 과정을 거쳐 모든 절점의 수요 패턴에 대한 압력이 5분 간격 24시간(수요 패턴과 동일) 데이터로 구축된다(Fig. 3(b)). 이후 모든 절점의 24시간 수요 및 압력 데이터는 절점의 고유 번호에 따라 순차적으로 2차원 배열되며, 이를 단일 채널

(회색조) 히트맵으로 변환하여 ATAE 모델의 학습 데이터를 구축한다(Fig. 3(c)). 즉, 본 연구에서 활용된 Austin 관망 데이터의 경우 최종적으로 (총 절점 개수) × 288(24시간의 5분 단위 분할) × 1(단일 채널 회색조 히트맵) 크기의 데이터가 하나의 학습 데이터를 구성한다. 압력과 수요의 학습 데이터 예시는 Fig. 4에 도시하였다. 여기서 압력 히트맵(Fig. 4(a))은 시드(seed) 이미지로써, ATAE 모델이 해당 압력 이미지로부터 신경망을 통해 픽셀값을 변환하여 수요량 히트맵 이미지(Fig. 4(b))를 생성한다.

이때 기저 수요가 없는 단순 연결 절점의 데이터는 Fig. 4(b)에서와 같이 흰색 행으로 나타나게 된다. 관망의 수요를 예측함에 있어 연결 절점의 중요도는 상대적으로 낮지만, 본 모델의 학습 데이터를 구성할 때는 음성 샘플링(negative sampling)의 일종으로 활용하기 위해 포함하였다. 음성 샘플링이란 모델이 유용한(positive) 정보를 더 효과적으로 학습하도록 만들기 위해 불용한(negative) 정보를 의도적으로 제공하는 학습 방법론으로(Mikolov *et al*., 2013), 최근 이미지와 그래프 분석에서도 그 효과를 입증하였다(Wu *et al*., 2020; Yang *et al*., 2020). ATAE 모델은 압력 데이터를 입력받아 수요 데이터를 예측하여 출력하는데, 이때 0의 수요량을 가진 일정량의 데이터를 결과값으로 제공하게 되면 0이 아닌 수요량을 가진 데이터를 학습할 때 대조하게 되어 학습 성능을 향상시킬 수 있다. 물론 이때 최적의 학습을 위한 음성 샘플의 양은 그 수준을 조절하여 실험적으로 판단할 수 있으나, 본 모델의 주목적은 관망 전반의 수요 예측이므로 모든 절점의 데이터를 학습 데이터로 사용하였다.

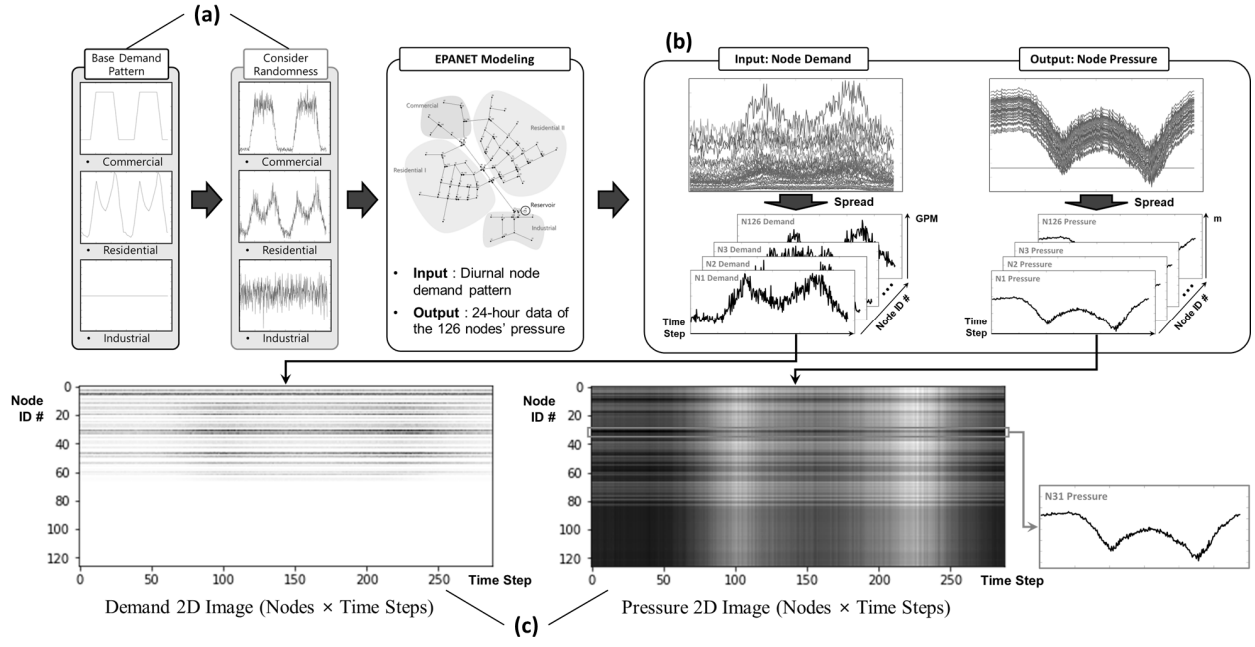
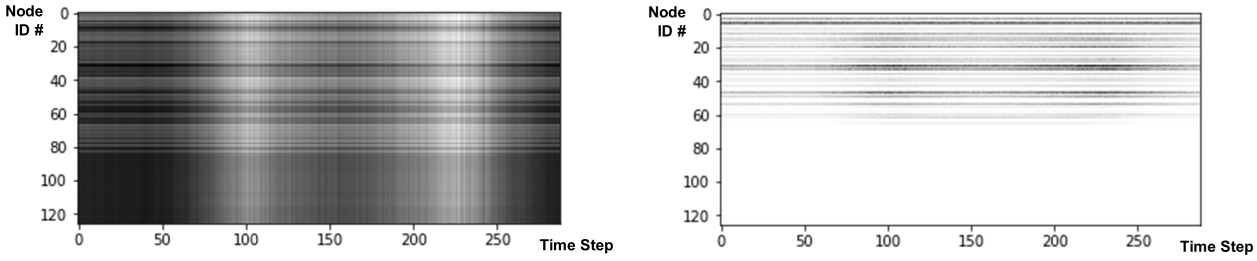


Fig. 3. Multidimensional water distribution system data synthesizing procedure



(a) Pressure image sample (b) Demand image sample

Fig. 4. Representative examples of (a) pressure and (b) demand training images

3. 대상 관망

본 연구에서 제안된 ATAE 모델은 미국 텍사스 주 북부에 위치한 실제 관망인 Austin 관망(Fig. 5)을 대상으로 적용되어 검증되었다. Austin 관망은 단일 수원지로부터 90개 상수관을 거쳐 126개의 절점으로 물이 공급되는 시스템이다. 2020년 기준 36,835 가구(110,506 명)가 해당 관망을 사용하며, 전체 시스템의 총 수요량은 약 726 L/sec, 관망 내 관의 총 길이 및 평균 직경은 각각 82.6 km, 424.4 mm이다.

EPANET 모형을 이용한 수요 기반 해석(Demand-Driven Analysis, DDA)하에 관망의 시계열 데이터를 확보하기 위해서는 관망의 모든 절점에 수요 패턴을 설정하여야 한다. 각 절점의 위치에 따라 상업, 주거, 및 공업 지역의 패턴을 적용하였다(Figs. 5(a)~5(c)).

Austin 관망이 존재하는 지역은 Fig. 5에서 확인할 수 있듯이, 중앙부 북서-남동으로 가로지르는 형태의 간선을 기준으로 상부는 상업 지역, 좌우는 주거 지역, 그리고 하부는 공업 지역으로 구성된다. 각 지역은 거주민 또는 이용자들의 생활 패턴에 따라 물 사용 패턴 또한 구분되는 특징을 가지고 있다. 예를 들어, Fig. 5(a)를 보면 주거 지역은 거주민들의 출근 시간 직전 및 퇴근 시간 직후 샤워나 세탁기 사용 등으로 인해 가장 높은 수요량을 기록한다. 상업 지역(Fig. 5(b))은 그와 반대로 인구가 출근 시간 직후부터 퇴근 시간 직전까지 가장 밀도 높게 존재하므로 해당 시간에 가장 높은 수요를 나타내며, 공업 지역(Fig. 5(c))은 인구의 이동과 관계없이 기계 공정이 24시간 일정 수준으로 작동한다고 가정하여 변동이 없는 패턴으로 가정한다.

이후 Austin 관망의 데이터를 ATAE 모델 학습에 적합한 다차원 데이터로 구성하기 위해 Fig. 3의 데이터 구축 프로세스를 적용하며, 총 10,000장의 학습 데이터가 구축된다.

4. 적용 및 결과 분석

제안된 ATAE 모델은 다양한 조건에서의 성능을 검증하기 위해 (1) 역누적분포함수의 분산 조정을 통한 수요량 데이터 불확실성 변동과 (2) 수요 수준(water consumption level)에 따른 시간 분할 조건에 따라 다른 데이터를 학습하고 결과를 출력하여 비교하였다. 불확실성의 경우 데이터 평균값의 10%를 기준으로, 5%, 20%인 경우의 데이터 학습 시 생성 능력의 차이를 확인하였다. 시간 분할의 경우 EPANET으로 제작한 전체 24시간 데이터를 기준으로, 전체 관망 수요량 수준에 따라 2시간 단위로 분할하였다. 수요량이 가장 낮은 시간은 02시부터 04시, 평균에 가장 근접한 시간은 11시부터 13시, 가장 높은 시간은 17시부터 19시였다. 이는 대상 관망 내 절점 중 70% 이상이 주거 지역에 속해 주거 지역 수용가의 물 사용 패턴과 다소 유사한 특징을 보인다고 판단하였다.

조작 변인 간 비교 기준이 되는 불확실성 10%, 24시간 데이터의 생성 결과는 Fig. 6에 도시하였다. ATAE 모델에 의해 불확실성 10%의 24시간 압력 이미지로부터 생성된 수요 이미지는 실제 이미지인 Fig. 4(b)와 비교하였을 때 색이 진하게 나타나는(수요량이 있는) 주요 절점들의 데이터를 시각적 측면에서 효과적으로 생성해내는 것으로 확인되었다. 하지만, 기저 수요가 0인 절점들의 백색 이미지를 온전히 재현해내지 못하고 다소 회색빛이 존재하도록(작은 수요량이 항상 있도

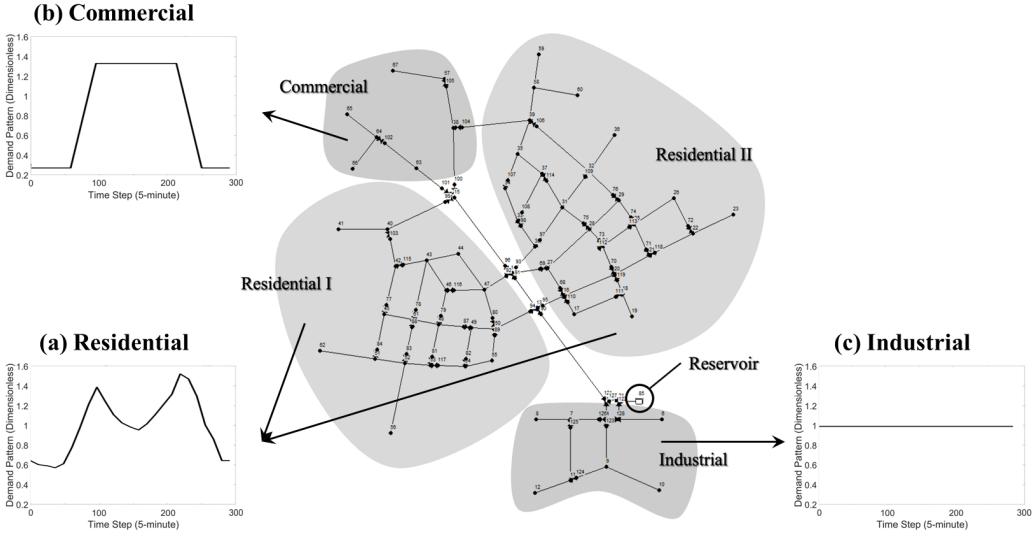


Fig. 5. Three types of users and their typical diurnal demand patterns in Austin network: (a) Residential, (b) Commercial, and (c) Industrial

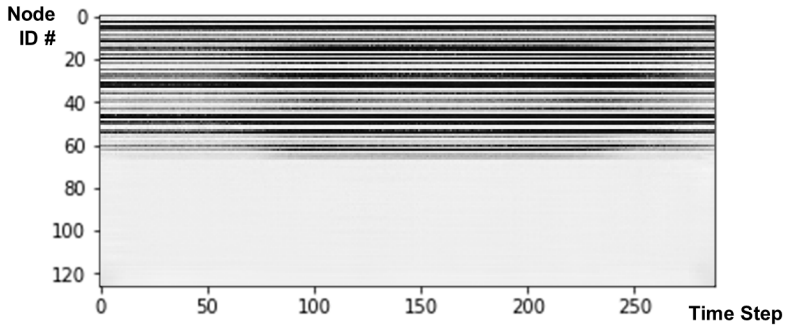
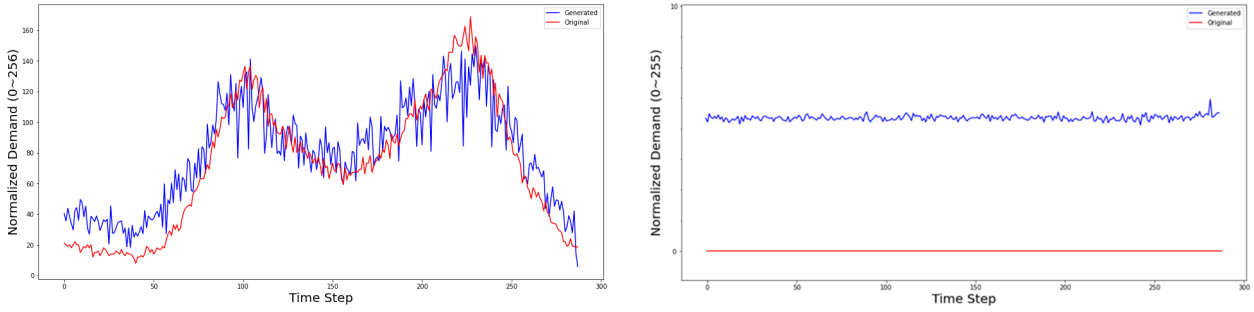


Fig. 6. Demand image sample transformed from original pressure image sample by ATAE model



(a) Generated demand of node 31 (b) Generated demand of node 88

Fig. 7. Examples of generated and original normalized demand with (a) high and (b) low similarity to the original demand

록) 표현하였다. 이를 정량적으로 비교하였을 경우 Fig. 7과 같이 표현할 수 있다.

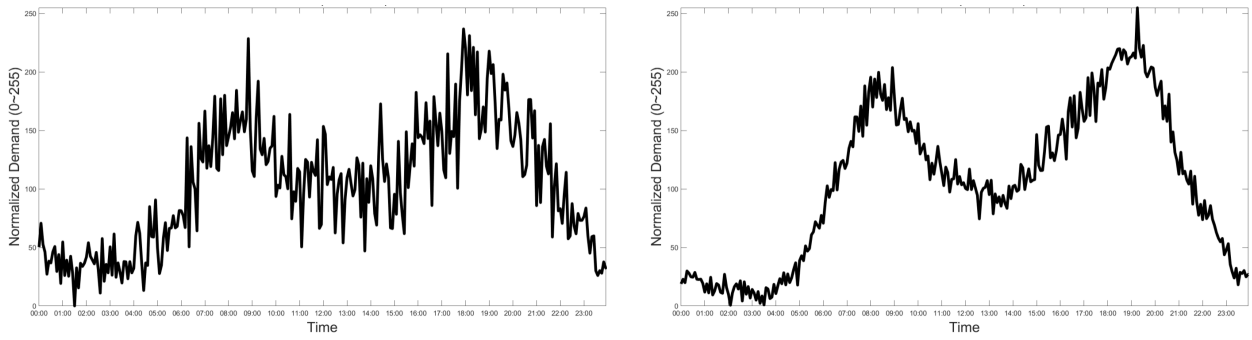
Fig. 7(a)에서는 주거 지역이자 실제 수요가 존재하는 31번 절점의 수요 패턴을 잘 모방하여 생성한 것을 확인하였다. 수요량 데이터의 불확실성을 정확하게 잡아내지는 못하지만 주거 지역의 수요 패턴의 경향을 파악하여 데이터를 생성한다. 다만 Fig. 7(b)에서는 동일한 주거 지역이나 실제 수요가 존재하지 않는 88번 절점의 경우 실제 0인 수요값에 수렴하지 못하고 약 6 정도에서 미세하게 진동하도록 생성된 모습을 확인할 수 있었다. 이러한 절점들 같은 경우에는 모든 학습 데이터에서 각 절점에 해당하는 행의 데이터가 전부 0인, 24시간동안 물의 수요가 없고 변동성이 없는 데이터이므로 ATAE 모델이 0에 매우 가까운 값을 생성할 것으로 예상되었다.

하지만 결과를 분석하였을 때, 생성된 수요량의 값이 0으로 점진적으로 수렴하지 않고 위 Fig. 7(b)에서 관찰되는 생성값 근처에서 지속적으로 진동하는 모습을 확인하였다. 수요량이 0인 다른 절점들을 확인한 결과 위와 유사한 값이 공통적으로 생성되었다. 이는 모델의 크기와 합성곱 연산의 선형 변환 원리로 인한 한계로 추정된다. ATAE 모델은 기존 오토인코더와 동일하게 합성곱을 통해 압력 데이터를 임베딩으로 변환하고, 전치 합성곱을 통해 임베딩을 다시 수요 데이터로 변환한다. 두 연산은 모두 다차원 선형 변환을 수행하며, 이 과정에서 수요량이 0인 데이터와 0이 아닌 데이터를 공통의 공유된 매개변수를 사용하여 처리하게 된다. 이로 인해 수요량이 0인 부분에서의 정확도가 감소하게 되며, 이를 해결하기 위해서는 더 많은 수의 매개변수를 사용하여 연산의 복잡도를 증가시켜야 할 것으로 예상된다. 이에 더해 모델의 크기가 확장되는 경우 수요량이 0인 데이터를 학습에 포함한 음성 샘플링의 효과도 증진될 것으로 기대된다.

ATAE 모델의 실적용 시 성능을 검증하기 위해, 다음 4.1절에서는 서로 다른 불확실성 조건에 따른 결과를 분석하며, 이후 4.2절에서는 수요 수준에 따른 시간 분할 조건에서의 결과를 분석한다.

4.1 불확실성 변동 조건 비교

위 장에서 분석한 기준 모델의 생성 결과 도출 시에는 학습 데이터 제작 시 역누적분포함수 샘플링에서 10%의 표준편차를 통해 불확실성 조건을 부여하였다. 이를 기준으로 불확실성 조건의 변동에 대한 ATAE 모델의 생성 성능을 비교하기 위하여 샘플링 시 표준편차가 5%인 경우와 20%인 경우를 모두 조사하였다. 각각의 불확실성에 대한 학습 데이터 예시는 Fig. 8에 도시하였다. 해당 데이터들로부터 생성된 결과들의



(a) 24-hour demand at node #31 when σ =20% (b) 24-hour demand at node #31 when σ =5%

Fig. 8. Examples of normalized demand with (a) high and (b)low uncertainty

Table 1. Quantitative comparison of 24-hour (00:00-24:00) true samples and generated samples from the ATAE model with regard to the standard deviation (σ) of the data (normalized in 0~255)

σ \ Sample		True	Generated (% error)
σ=10% (Standard)	Mean	107.70	104.36 (3.10%)
	Median	96.08	97.09 (-1.05%)
	Mode (Bin)	80-90	80-90
	RMSE	-	2.88
σ=20%	Mean	106.63	101.82 (4.51%)
	Median	98.78	100.66 (-1.90%)
	Mode (Bin)	80-90	90-100
	RMSE	-	5.98
σ=5%	Mean	105.31	104.20 (1.11%)
	Median	94.59	95.24 (0.07%)
	Mode (Bin)	80-90	80-90
	RMSE	-	2.54

평균, 중간값, 최빈값(구간) 등의 통계량과 평균제곱근오차(Root Mean Square Error, RMSE)에 대한 정량적 비교는 Table 1에 수행하였다.

먼저 불확실성이 20%로 기준보다 높아지는 경우, 생성된 데이터의 평균, 중간값, 최빈값의 오차가 증가하는 것으로 확인되었다. 더 자세하게는, 평균의 경우 양의 오차가 증가하고 중간값은 음의 오차 절댓값이 증가하는 방향으로, ATAE가 생성하는 샘플 분포와 실제 데이터 분포 간의 거리가 증가하는 것으로 해석할 수 있다. 다시 말해, 학습 데이터의 분산이 증가하면서 모델의 학습 성능이 저하되었다고 판단할 수 있다. 반면 학습 데이터의 불확실성이 5%로 감소하는 경우, 모든 통계량에서의 오차가 감소하는 것을 확인하였으며, 현재의 ATAE 모델은 학습 데이터의 분산(수요량 데이터의 불확실성 크기)에 다소 민감하게 반응한다는 결론을 도출하였다.

추가적으로, 오차가 증가함에 따라 평균 추정치는 더 작아지며 중간값 추정치는 더 커지는 모습을 확인할 수 있다. 이는 모델이 생성하는 데이터 샘플 분포의 편향이 감소하여 최고치 또는 최저치를 제대로 학습하지 못해 평균에 가까운 데이터를 더 많이 생성한다는 뜻으로도 해석할 수 있다. 이러한 분산 민감도는 향후 학습 데이터의 증가와 함께 모델의 크기 또한 증가시켜 잔차 연결(residual connection) 학습을 구현함으로써 해결할 수 있을 것으로 생각된다. 또한, 향후 해당 방식을 이용하여 모델의 개선 방향을 탐색하는 경우 학습 데이터에 다양한 분산의 데이터를 동시에 포함시키는 것으로 실적용 시 발생할 수 있는 여러 불확실성 문제에 대응할 수 있을 것으로 보인다.

4.2 수요 수준에 따른 시간 분할 조건 비교

불확실성 조건에 따른 생성 결과 변동과 함께, 학습 데이터로 사용한 24시간 단위 수요 중 수요량의 고저에 따라 2시간 단위로 추출된 특정 시간대에 대해서도 따로 생성 성능을 조사하였다. 수요량의 수준은 대상 관망 내 가장 많은 비율을 차지하는 주거 지역의 수요 패턴을 기준으로 선정하였다(Fig. 9). 먼저 (A) 낮은 수요량을 가진 시간대로는 수용가에서 물을 거의 이용할 일이 없는 02시부터 04시까지를 선정하였고, (B) 중간 수요량을 가진 시간대로는 점심 시간대인 오전 11시부터 오후 1시, (C) 높은 수요량의 시간대는 퇴근 시간 전후인 17시부터 19시로 결정하였다. 각 2시간 씩 추출된 시간대에 대해 2시간 단위의 수요 및 압력 이미지를 따로 제작하여 ATAE 모델로 학습시켰고, 해당 결과를 Table 2로 기록하였다.

24시간 단위의 이미지를 학습하고 생성했을 때와 비교하면, 먼저 샘플 데이터의 분포에 차이가 있다는 점을 확인할 수 있다. 기존의 학습 데이터는 평균이 가장 크고, 그 다음이 중간값, 마지막으로 최빈값이 따라오는 정적 편포의 형태이나, 시간대가 값에 따라 소단위로 분해됨에 따라 편향이 감소하였다. 그에 따라, ATAE 모델은 시간대에 관계 없이 2시간 단위의 학습 데이터는 상대적으로 오차가 작은 상태로 생성되었다

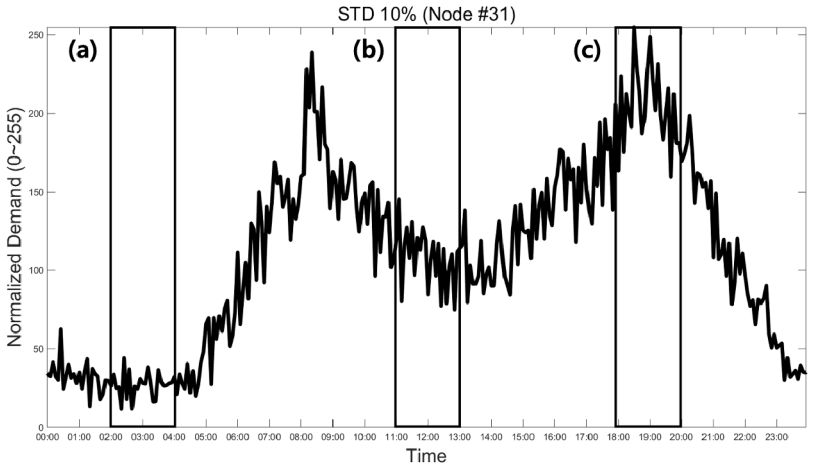


Fig. 9. Different 2-hour units with regard to demand level ((a) low level, 02:00-04:00; (b) medium level, 11:00-13:00; (c) high level, 18:00-20:00)

Table 2. Quantitative comparison of 2-hour true samples and generated samples from the ATAE model with regard to the demand level (STD 10%, normalized in 0~255)

Time \ Sample		True	Generated
Full Time (Standard; 00:00-24:00)	Mean	107.70	104.36 (3.10%)
	Median	96.08	98.09 (-1.05%)
	Mode (Bin)	80-90	80-90
	RMSE	-	2.88
(A) Low Demand (02:00-04:00)	Mean	85.38	85.89 (-0.60%)
	Median	85.12	83.78 (1.57%)
	Mode (Bin)	80-90	80-90
	RMSE	-	1.38
(B) Medium Demand (11:00-13:00)	Mean	103.10	102.30 (0.78%)
	Median	103.52	101.38 (2.07%)
	Mode (Bin)	100-110	100-110
	RMSE	-	1.40
(C) High Demand (18:00-20:00)	Mean	117.09	115.34 (1.49%)
	Median	115.93	113.34 (2.23%)
	Mode (Bin)	110-120	110-120
	RMSE	-	2.27

는 것을 알 수 있다.

세부적으로 비교해보면, 낮은 수요량 시간대인 (A)에서의 오차가 나머지 두 시간대보다 상대적으로 작은 것을 확인할 수 있다. 이는 역누적분포함수 샘플링을 통한 학습 데이터 생성에서 상대적으로 작은 값의 데이터에서 편차가 작은 데이터를 샘플링하기 때문으로 추측할 수 있다. 즉, 위 절에서 설명했던 불확실성에 따른 생성 성능 차이를 고려했을 때, 수요량이 작은 경우 그에 따른 분산도 작기 때문에 분포를 학습하고 모방하기 더 쉽기 때문으로 생각된다. 그에 따라, 점차 분산과 샘플 내 최대값이 큰 (C) 구간의 생성 결과의 오차가 상대적으로 더 큰 것을 확인할 수 있다.

이러한 결과를 통해 향후 ATAE 모델의 개선 방향으로서 수요량 수준에 따른 구간 별 학습이 가능한 다수의 하위 모델에 기반하여 하나의 앙상블 모델을 구축하는 것도 생각해 볼 수 있다. 또한 이 경우 각 구간의 데이터에 대한 정규화 또는 표준화 방법론을 차별화시켜 적용하여 전반적인 생성 성능의 향상을 도모할 수도 있다.

5. 결 론

본 연구에서는 상수도관망의 데이터를 시간과 절점 위치의 2차원 히트맵 이미지로 변환하여, 이를 학습하고 모방하여 데이터를 생성할 수 있는 적대적 학습 기반 오토인코더, ATAE 모델을 제안하였다. 해당 모델은 압력 이미지 데이터를 입력 자료로 받아 이를 잠재 공간 내에서 변환하여 수요 이미지 데이터로 출력하는 오토인코더를 생성 신경망으로, 생성된 이미지에 대한 진위 판별을 진행하는 합성곱 신경망을 판별 신경망으로 사용하는 적대적 훈련을 진행한다.

미 텍사스주에 실재하는 Austin 관망을 EPANET으로 구현한 모형에 ATAE 모델을 적용하여 데이터를 생성한 결과, 수요량이 아예 존재하지 않는 절점들의 데이터는 정확한 최저점을 0으로 모방하지는 못하지만, 관망 전체적으로 3% 내외의 오차를 발생하여 해당 생성 모델의 효과를 입증하였다.

본 연구에서 제안한 방법론의 추가적인 성능 향상을 위한 다양한 실험이 제안될 수 있다. 해당 모델의 불확실성 민감도 및 수요량 수준에 따른 생성 성능의 비교를 위하여 여러 조건하에 실험을 진행한 결과, 학습 데이터의 불확실성에는 다소 민감한 모습을 볼 수 있으나 이를 극복하기 위해 수요량 수준

등의 기준에 기반한 소구간 분류 등의 방안 또한 모색할 수 있었다. 향후 해당 모델의 발전을 위해서 잔차 연결 학습 기반의 대규모 모델 또는 구간 별 학습을 진행한 후 이를 종합하여 생성 결과를 도출하는 앙상블 모델 등을 시도할 수 있을 것으로 생각된다. 또한, 관망 내 전체 절점을 사용하는 현재의 방식은 실무에서의 사용을 고려했을 때 다소 비효율적일 수 있다. 따라서 관망의 거동에 있어 데이터가 민감하게 변동하는 주요 절점을 파악하여, 해당 절점들의 압력 데이터만을 활용해 전체 관망의 수요를 예측하는 방식으로 모델을 경량화하는 시도도 가능해질 것으로 예상된다. 위와 같은 방식으로 모델이 경량화될 경우 대규모 관망에의 적용 가능성도 높아질 수 있다. 이에 더해, 현재 데이터의 음성 샘플로 활용된(기저 수요가 없는) 연결 절점들 또한 제거 후 성능을 확인하는 방식으로 절제 연구(ablation study) 또한 수행할 수 있을 것으로 생각된다.

상수도관망의 데이터 생성을 위해 사용할 수 있는 다른 모델들과 성능을 비교하는 것 또한 좋은 연구가 될 수 있다. 본 연구에서는 기본적인 GAN의 학습 원리를 차용하여 ATAE 모델의 학습 파이프라인을 구성하였다. GAN 모델은 2014년 처음 발표된 이후 많은 응용 모델이 개발되었고, 그중 일부인 Conditional-GAN (Mirza and Osindero, 2014)나 Style-GAN (Karras *et al*., 2019) 등이 본 연구와 동일한 압력-수요 간 변환을 동일하게 수행할 수 있을 것으로 예상된다. 또한 현재 컴퓨터 비전 분야에서 가장 많이 사용되는 Diffusion (Rombach *et al*., 2022) 기반 모델 또한 비교의 대상으로 활용될 수 있다.

본 연구에서 제안한 상수도관망 수요 데이터 변환 방법론은 생성 모델에 기반하기 때문에, 이를 이용하여 활용성을 확장하는 경우 더욱 다양한 분야에서 유용하게 사용될 수 있다. 예를 들어, 현재 동시점의 압력 데이터를 이용하여 수요 데이터를 생성하는 방식에 시계열 예측 모델이 추가되는 경우에는 수요 생성 뿐 아니라 수요 예측까지 수행 가능한 모델이 될 수 있다. 게다가, 계측 시스템 오작동이나 관 파열 등 다양한 이상 상황에 대한 데이터를 학습 데이터로 확보할 수 있다면 이상 탐지에도 활용할 수 있으며, 단순 계측 오류와 실제 이상 상황을 분류하여 판단하는 작업에도 강건성을 높일 수 있다. 이후 단계에서는 현재 계측 시계열 데이터만을 활용하는 방법론을 넘어 관망 자체의 정보(관 직경, 유량계수, 마찰계수 등)를 입력으로 병렬 활용할 수도 있다. 이 경우 학습에 활용한 관망이 아닌 다른 관망 시스템의 데이터로도 추론이 가능할 것으로 보이며, 퓨-샷 러닝(few-shot learning) 등을 통해 실무에서 활용할 수 있는 가능성이 있다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1C1C1006481).

Conflicts of Interest

The authors declare no conflict of interest.

References

Bagheri, A., Monsef, H., and Lesani, H. (2015). “Integrated distribution network expansion planning incorporating distributed generation considering uncertainties, reliability, and operational conditions.” *International Journal of Electrical Power & Energy Systems*, Vol. 73, pp. 56-70.

Bragalli, C., Neri, M., and Toth, E. (2019). “Effectiveness of smart meter-based urban water loss assessment in a real network with synchronous and incomplete readings.” *Environmental Modelling & Software*, Vol. 112, pp. 128-142.

Brentan, B.M., Meirelles, G.L., Manzi, D., and Luvizotto, E. (2018). “Water demand time series generation for distribution network modeling and water demand forecasting.” *Urban Water Journal*, Vol. 15, No. 2, 150-158.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., and Bengio, Y. (2014). “Generative adversarial networks, 1-9.” *arXivpreprint*, arXiv:1406.2661.

Jang, H., Jung, D., and Jun, S. (2022). “Comparison of ANN model’s prediction performance according to the level of data uncertainty in water distribution network.” *Journal of Korea Water Resources Association*, Vol. 55, No. 12, pp. 1295-1303.

Jun, S., Jung, D., and Lansey, K.E. (2021). “Comparison of imputation methods for end-user demands in water distribution systems.” *Journal of Water Resources Planning and Management*, Vol. 147, No. 12, 04021080.

Jung, D., Chung, G., and Kim, J.H. (2010). *Optimal design of water distribution systems considering uncertainties in demands and roughness coefficients*. Water Distribution Systems Analysis 2010, In 12th Annual Conference, Tucson, AZ, U.S., pp. 1390-1399.

Kabir, G., Tesfamariam, S., Hemsing, J., and Sadiq, R. (2020). “Handling incomplete and missing data in water network database using imputation methods.” *Sustainable and Resilient Infrastructure*, Vol. 5, No. 6, pp. 365-377.

Karras, T., Laine, S., and Aila, T. (2019). “A style-based generator architecture for generative adversarial networks.” *In Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pp. 4401-4410.

Mirza, M., and Osindero, S. (2014). "Conditional generative adversarial nets." *arXivpreprint*, arXiv:1411.1784.

Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., and Gorgoglione, A. (2021). "Water-quality data imputation with a high percentage of missing values: A machine learning approach." *Sustainability*, Vol. 13, No. 11, 6318.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). "Highresolution image synthesis with latent diffusion models." *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, U.S., pp. 10684-10695.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). "Learning from simulated and unsupervised images through adversarial training." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, U.S., pp. 2107-2116.

Sun, S., Yeh, C.F., Hwang, M.Y., Ostendorf, M., and Xie, L. (2018). Domain adversarial training for accented speech recognition. *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Calgary, Canada, pp. 4854-4858.

Zanfei, A., Menapace, A., Brentan, B.M., and Righetti, M. (2022). "How does missing data imputation affect the forecasting of urban water demand?." *Journal of Water Resources Planning and Management*, Vol. 148, No. 11, 04022060.