2nd International Joint Conference on Water Distribution
Systems Analysis & Computing and Control in the Water Industry
Valencia (Spain), 18-22 July 2022
doi: https://doi.org/10.4995/WDSA-CCWI2022.2022.14152

MULTIDIMENSIONAL DATA GENERATION IN WATER DISTRIBUTION SYSTEMS USING THE CYCLE-GAN

Sehyeong Kim[1], Donghwi Jung[2]

[1]Master Student, School of Civil, Environmental and Architectural Engineering, Korea University, Anam-ro 145, Seongbuk-gu, Seoul, 02841, Republic of Korea

[2]Assistant Professor, School of Civil, Environmental and Architectural Engineering, Korea University, Anam-ro 145, Seongbuk-gu, Seoul, 02841, Republic of Korea

[1] kaiba0514@korea.ac.kr, [2] sunnyjung625@korea.ac.kr

Abstract

A water distribution system (WDS) consists of thousands of components interacting with each other. To analyze the status or to manage the operation of the WDSs, two main feature values have been mainly used: nodal pressure and pipe flow rate. However, insufficient data due to the malfunction of the sensors or economical limitations interrupts the collection of abundant information in many cases. Therefore, this study proposes a WDS data generation model based on the cycle-GAN (Generative Adversarial Networks). The proposed model learns the demand time series data and its corresponding time series data of pressure or flow. All training data is two-dimensionally constructed by considering the time series data for 24 hours of each component as a single row and arranging all row data vertically. After normalizing all the data to integers from 0 to 255, they become a greyscale image. Then, the cycle-GAN model consisting of two generators and two discriminators trains those image datasets, to translate the demand image data to the WDS feature image data (i.e., pressure or flow). Firstly, based on the random seeds, the first generator in the cycle-GAN model is trained to generate the demand image, and the second generator is trained to generate the WDS feature image. After the fundamental training for the generation of their own data, the second generator trained for the feature image data starts to use the synthesized demand image results from the first generator as its seeds, not using the random noise. This process makes the second generator have the ability to translate the demand images to the WDS feature images by using the convolutional and deconvolutional functions in the neural network layers. This model was demonstrated by applying to the Mays network, which is the benchmark network consisting of 13 nodes and 21 pipes with two reservoirs.

Keywords
Water distribution systems, Deep-learning-based hydraulic analysis, Multidimensional data generation, Image translation, Generative adversarial networks, Cycle-GAN.

1 INTRODUCTION

Improvements in monitoring technology have led it possible to set diverse sensors, including pressure and pipe flow rate, to measure the state of water distribution systems (WDSs). However, the cities have been also developed concentratively but extensively and the number of the critical points whose status data should be collected has been increased, making their WDSs more difficult to be completely analyzed. Even worse, the reliability of the sensors is not enough to observe the intact data if there are some malfunctions in the equipment or environmental fluctuations. Due to these limitations of the monitoring systems, the collectable data becomes more sparse and some studies have tried to supplement the missing values or empty space of the data that they cannot directly measure.

Some univariate and multivariate imputation methodologies using mean, median, and regression-based estimation were applied to the sparse WDS data and the sequential imputation that can consider the time variable was also utilized (Sankaranarayanan et al. 2019). Furthermore, machine learning techniques such as Kalman-filter-based nearest neighbor regressor or random forest regressor are also used to be assessed as statistical imputation models (Kabir et al. 2020, Rodriguez et al. 2021). These kinds of sparsity problems are not only in the WDS domain, and other fields like environmental engineering have widely used various techniques to construct the complete composition of the data-driven models (Zheng et al. 2016, Dumedah and Coulibaly 2011). For example, the subsurface environmental status data has been usually imputed using the techniques mentioned above, because it is not easy to monitor it in real-time. However, there is little effort to use the generative deep learning models for the imputation method although it has been appraised as one of the most important developments in the generation and the translation of the multidimensional data.

Therefore, this study suggests the multidimensional data translation model based on the cycle-GAN framework, especially transforming the demand time series data to the WDS feature data like nodal pressure or pipe flow rate. There are two generators in the cycle-GAN model, and the first generator is trained to synthesize the demand image data, while the second generator is trained to synthesize the WDS feature (i.e., pressure or flow) image data, based on the random seeds. After those two generators are fundamentally trained to synthesize their images in charge, the second generator starts to use the demand image outputs from the first generator as its seeds. Then, it can secure the ability to translate the demand data to the pressure or the flow data. As a result, this mechanism can provide the estimated WDS data if there is continuous missing data regardless of the cause.

2 METHODOLOGY

2.1 Generative Adversarial Networks (GAN)

Adversarial training is presented as a new way to train the generative models (Goodfellow et al. 2014). The generative adversarial networks (GAN) consist of two adversarial networks: generator and discriminator. The purpose of the generator G is to learn the existing data distribution to generate data as similar as possible, while the discriminator D is to distinguish whether the data given as input is synthesized by G or extracted from the real data (i.e., training data) distribution. The G constructs a mapping function that leads from a random noise distribution to a real data distribution in the form of a multi-layer perceptron. On the other hand, D receives the given input and derives the probability that the input came from the real data as a result. G and D are simultaneously learned. G learns to maximize the probability that the results generated from latent space (random noise) $\boldsymbol{z}$ used as an input is discriminated as real data by D. D learns to increase the discrimination performance between the real and the generated data itself. In other words, the result generated from the G is trained to minimize the probability that the result is determined as real data. It is in the form of a minimax game of two players, the generator and the discriminator, and it can be represented in equation (1):

$$\min_{G} \max_{D} V(D,G) = E_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + E_{\boldsymbol{z}\sim p_{data}(\boldsymbol{z})}[\log\{1 - D(G(\boldsymbol{z}))\}] \tag{1}$$

2.2 Multidimensional Data Translation: Cycle-GAN

As a baseline model for data generation through image translation, cycle-GAN (Zhu et al. 2017) is adopted in this paper. The purpose of training cycle-GAN is to learn two generators and two discriminators that can travel between two different domains. Let the two domains be X and Y, respectively, and the generator translating X to Y, be G_X, while the generator translating Y to X, be G_Y. The discriminator D_X of the X domain aims to better discriminate between the data $\boldsymbol{x}$ in the X domain, and $G_Y(\boldsymbol{x})$ translated through the G_Y in the Y domain. Conversely, the discriminator D_Y

in the Y domain aims to better discriminate between the data $\boldsymbol{y}$ in the Y domain and $G_1(\boldsymbol{x})$. These processes can be represented in equation (2) and (3):

$$\min_{G}\max_{D} V(D_X, G_Y) = E_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D_X(\boldsymbol{x})] + E_{\boldsymbol{y}\sim p_{data}(\boldsymbol{y})}[\log\{1 - D(G_Y(\boldsymbol{y}))\}] \tag{2}$$

$$\min_{G}\max_{D} V(D_Y, G_X) = E_{\boldsymbol{y}\sim p_{data}(\boldsymbol{y})}[\log D_Y(\boldsymbol{y})] + E_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log\{1 - D(G_X(\boldsymbol{x}))\}] \tag{3}$$

In the proposed model, X domain will be the demand image data distribution and Y domain will be the feature image data distribution. Therefore, the model $G_{\boldsymbol{x}}$ will be trained to mimic and synthesize the demand image, and the feature data will be synthesized by the model G_y, while their corresponding discriminators try to classify the real dataset images and the synthesized images from $G_{\boldsymbol{x}}$ or G_y.

3 STUDY NETWORK

3.1 WDS Information

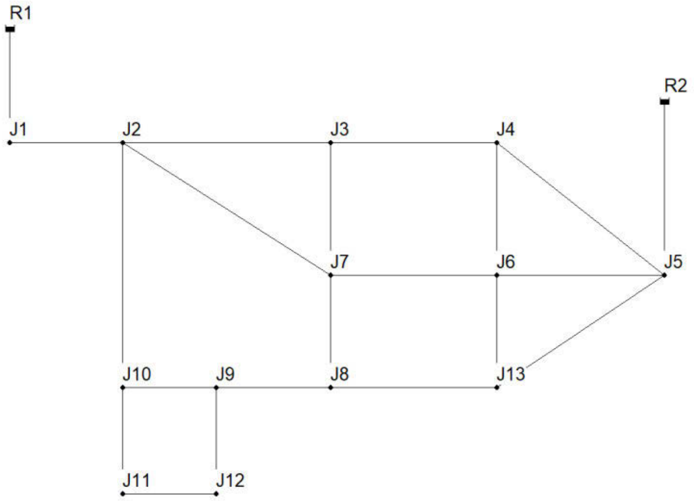

Figure 1. The configuration of Mays network

The proposed model is applied to the Mays network, a widely used benchmark network. Mays network consists of 13 nodes, 21 pipes, and two reservoirs. The configuration of the Mays network is represented in *Figure 1*, and the detailed information on the Mays network is the same as below.

- The total levels of the two reservoirs are both 60.96m.
- Extended period simulation (EPS) is possible by setting a demand pattern for 24 hours.
- The demand pattern has a sub-pattern with one-hour intervals to make a distinctive characteristic with the stripe pattern in the two-dimensional time series image.
- The uncertainty (i.e., variance) of the demand is limited with around 1/100 of each data value, to avoid the unnecessary confusion of the image patterns.

3.2 WDS Image Construction

The training image data is constructed in a form of a two-dimensional matrix for each data. One row of the image represents the 24-hour time series data of one component (i.e., a node or a pipe) with 5-minute time steps, and one column of the image represents the status of all components at a single time step. Therefore, according to the feature of the Mays network, the pressure image and demand image has 13 rows and the flow image has 21 rows, the same as the number of the nodes and pipes, respectively. All data are normalized within the integer from 0 to 255, to transform them into greyscale images. Similar stripe patterns in the pressure and flow images can be observed due to the one-hour pattern of the demand. Total of 10,000 images for each data are generated to be trained.

4 APPLICATION RESULTS

When the cycle-GAN model receives the image as training data, they fundamentally train to generate them based on the random noise from the latent space. In the proposed model, the generative function in layers of the neural network was the 2-D deconvolutional function. Their training performance can be represented by the losses of the discriminators (*Figure 2*). It can be observed that the loss of each discriminator started to fluctuate after around a fifth of the total 10,000 iterations. It means that the discriminators' performance started to be deteriorated due to the synthesized results from the generators having become realistic successfully.

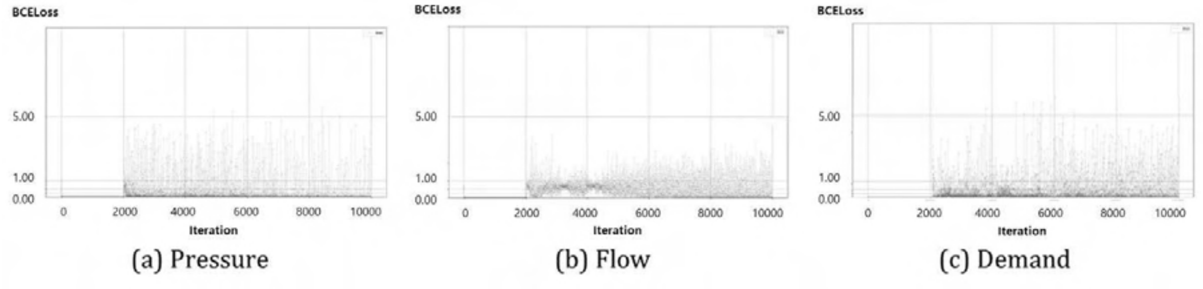

Figure 2. The loss graph of the discriminators based on the random noise in the latent space

Then, the data translation started after the fundamental training based on the random noise is completed. In principle, the cycle-GAN model can translate the data bidirectionally, but only demand-to-pressure and demand-to-flow cases were considered, taking into account the principle of the demand-based hydraulic analysis of the WDSs. After the training process for the image translation is finished, the model can generate the feature images based on the real demand images. The samples of the generated WDS feature values which were generated based on the real demand data are represented in *Figure 3*. It can be observed that the stripe patterns that can be found in the real image data used as the training datasets, although there was some unclearness.

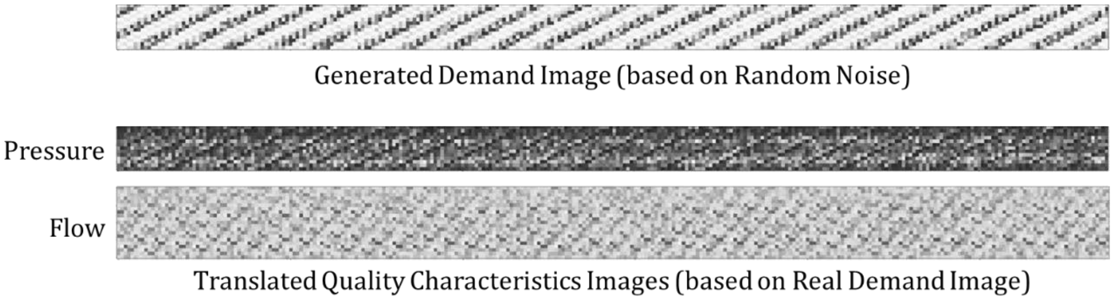
Generated Demand Image (based on Random Noise)
Pressure
Flow
Translated Quality Characteristics Images (based on Real Demand Image)

Figure 3. The generated demand image and the translated WDS feature images

To calculate the training performance of the translated results quantitatively, mean, median and mode with the class interval 10 were computed (Table 1). The differences between the original data and the generated data were not large, which means that the generators for each feature have good performance to synthesize the data. The mean values of the data had less than 3% of the normalized figures, and the median, and the mode values indicated the fine durability of the model representing stable distribution. Considering the results of this statistical analysis, it can be said that the suggested results showed that the proposed translation model can be used for data imputation in WDS, generating realistic data values with the original feature data.

Table 1. Statistically comparison of the normalized data values of the original and generated WDS features

Feature	**Mean**	**Median**	**Mode (Class interval: 10)**
Original Pressure	131.28	130	130-140
Generated Pressure	133.59	134	130-140
Original Flow	126.08	126	120-130
Generated Flow	124.89	123	120-130

5 CONCLUSIONS

This paper suggests the direction of utilizing multidimensional processing neural network technology as a translation tool for the hydraulic analysis of the WDSs, by using the cycle-GAN based on the demand and the WDS feature data. The proposed model learns the demand time series data and its corresponding time series data of pressure or flow. All training data is two-dimensionally constructed by considering the time series data for 24 hours of each component as a single row and arranging all row data vertically. After normalizing all the data to integers from 0 to 255, they become to a greyscale image. Then, the cycle-GAN model consisting of two generators and two discriminators trains those image datasets, to translate the demand image data to the feature image data (i.e., pressure or flow). The process of the suggested model makes the generator have the ability to translate the demand images to the feature images by using the convolutional and deconvolutional functions in the neural network layers. This model was demonstrated by applying to the small benchmark WDS, the Mays network. Finally, in the image data translation, the results indicated that the demand-to-flow translation is more successfully worked than the demand-to-pressure translation, showing the lesser RMSE errors than the demand-flow translation. The suggested results showed that the proposed translation model can be used for data imputation in WDS.

Besides, some possibilities for future studies were suggested based on the results of this study. Firstly, the bidirectional translation can be considered such as pressure(flow)-to-demand or the pressure-to-flow. As mentioned in the paper, the generators in the cycle-GAN model can travel between the two different distribution domains, but the proposed model only demonstrated one-side direction considering the analyzing approach. If this bidirectional translation is applied, data generation for the other feature estimations will be enabled. Additionally, the pressure-driven analysis (PDA) approach for the hydraulic data generation using the EPANET program can be considered. Due to the effectiveness of practical and realistic analysis of the WDSs, PDA has been intensively studied to be used more generally, especially in the simulation of abnormal circumstances. If this approach is applied to the proposed model, the various multidirectional translation and their corresponding data generation in the WDS data can be more valuable.

6 ACKNOWLEDGEMENT

This work was supported by the Korean Ministry of the Environment as a Subsurface Environment Management (SEM) project (No. 2021002170003)

7 REFERENCES

[1] Dumedah, G., & Coulibaly, P. (2011). Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. Journal of Hydrology, 400(1-2), 95-102.

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[3] Hu, P., Tong, J., Wang, J., Yang, Y., & de Oliveira Turci, L. (2019, June). A hybrid model based on CNN and Bi-LSTM for urban water demand prediction. In 2019 IEEE Congress on evolutionary computation (CEC) (pp. 1088-1094). IEEE.

[4] Jung, D., & Lansey, K. (2015). Water distribution system burst detection using a nonlinear Kalman filter. Journal of Water Resources Planning and Management, 141(5), 04014070.

[5] Jung, D., Yoo, D. G., Kang, D., & Kim, J. H. (2016). Linear model for estimating water distribution system reliability. Journal of Water Resources Planning and Management, 142(8), 04016022.

[6] Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2020). Handling incomplete and missing data in water network database using imputation methods. Sustainable and Resilient Infrastructure, 5(6), 365-377.

[7] Kang, D., & Lansey, K. (2010). Optimal meter placement for water distribution system state estimation. Journal of Water Resources Planning and Management, 136(3), 337-347.

[8] Kumar, S. M., Narasimhan, S., & Bhallamudi, S. M. (2008). State estimation in water distribution networks using graph-theoretic reduction strategy. Journal of Water Resources Planning and Management, 134(5), 395-403.

[9] Misiunas, D., Lambert, M., Simpson, A., & Olsson, G. (2005). Burst detection and location in water distribution networks. Water Science and Technology: Water Supply, 5(3-4), 71-80.

[10] Perelman, L., & Ostfeld, A. (2007). An adaptive heuristic cross-entropy algorithm for optimal design of water distribution systems. Engineering Optimization, 39(4), 413-428.

[11] Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., & Gorgoglione, A. (2021). Water-quality data imputation with a high percentage of missing values: A machine learning approach. Sustainability, 13(11), 6318.

[12] Romano, M., Kapelan, Z., & Savić, D. A. (2014). Automated detection of pipe bursts and other events in water distribution systems. Journal of Water Resources Planning and Management, 140(4), 457-467.

[13] Sankaranarayanan, S., Swaminathan, G., Radhakrishnan, T. K., & Sivakumaran, N. (2019). Missing data estimation and IoT-based flyby monitoring of a water distribution system: Conceptual and experimental validation. International Journal of Communication Systems, e4135.

[14] Shuang, Q., Liu, Y., Tang, Y., Liu, J., & Shuang, K. (2017). System reliability evaluation in water distribution networks with the impact of valves experiencing cascading failures. Water, 9(6), 413.

[15] Suribabu, C. R. (2010). Differential evolution algorithm for optimal design of water distribution networks. Journal of Hydroinformatics, 12(1), 66-82.

[16] Tabesh, M., Jamasb, M., & Moeini, R. (2011). Calibration of water distribution hydraulic models: A comparison between pressure dependent and demand driven analyses. Urban Water Journal, 8(2), 93-102.

[17] Tabesh, M., Shirzad, A., Arefkhani, V., & Mani, A. (2014). A comparative study between the modified and available demand driven based models for head driven analysis of water distribution networks. Urban Water Journal, 11(3), 221-230.

[18] Xing, L., & Sela, L. (2022). Graph Neural Networks for State Estimation in Water Distribution Systems: Application of Supervised and Semisupervised Learning. Journal of Water Resources Planning and Management, 148(5), 04022018.

[19] Ye, G., & Fenner, R. A. (2011). Kalman filtering of hydraulic measurements for burst detection in water distribution systems. Journal of pipeline systems engineering and practice, 2(1), 14-22.

[20] Yi, X., Zheng, Y., Zhang, J., & Li, T. (2016, June). ST-MVL: filling missing values in geo-sensory time series data. In Proceedings of the 25th International Joint Conference on Artificial Intelligence.

[21] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2223-2232).