IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INNOVATION AND APPLICATION OF INTERNET OF THINGS AND EMERGING TECHNOLOGIES IN SMART SENSING

Received December 8, 2019, accepted December 20, 2019, date of publication December 31, 2019,
date of current version January 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963373

Central Prediction System for Time Series Comparison and Analysis of Water Usage Data

MINGEUN JI[1], GANGMAN YI[1], AND JAEHEE JUNG[2]
[1]Department of Multimedia Engineering, Dongguk University, Seoul 04620, South Korea
[2]Department of Information and Communication Engineering, Myongji University, Yongin 17058, South Korea

Corresponding author: Jaehee Jung (jhjung@mju.ac.kr)

This work was supported by the 2019 Research Fund of Myongji University.

ABSTRACT Revenue water flow is defined as the amount of water for which the water rate has been collected, against tap water production, whereas non-revenue water (NRW) is defined as water that has been produced, but for which payment cannot be charged. In South Korea, there are big differences in NRW among the regions, and the NRW ratio in urban areas is higher than in rural regions. To reduce regional differences and effectively manage the water system, a management system to lower the NRW ratio is required. In particular, the NRW ratio can be reduced through an automatic leakage detection and sensor-error automatic checking system for feed water pipes and piping in household, and through leakage detection of water supply and drainage pipes that transport large volumes of water. Therefore, this study develops a system that can generate automatic alarms whenever abnormal usage is predicted via analysis of household water flow rate. Linear regression, ARIMA model, and additive regression model are compared to find the best method with high accuracy. The proposed method can support efficient water system management to lower the NRW ratio.

INDEX TERMS Non-revenue water, time series, ARIMA, additive regression model, water leakage alert system.

I. INTRODUCTION

According to the waterworks statistics of South Korea, published in 2019, the sector with the highest water consumption, as of 2017, is the household sector, which accounts for 3,451 million m^2, or approximately 62%, of the total water consumption. The next highest consumer of water is the general business sector, at 1,606 million m^2, or approximately 29%, of the total. Other water consumers include public service, at 136 million m^2, or approximately 2% of the total, and other industries, for the remaining 7%. Thus, water usage for everyday living accounts for the largest share of water consumption, and that consumption amount has been exhibiting an increasing trend [1].

In the system of the Korea Water Resources Corporation, raw water is purified after being collected, filtered, and sterilized through the intake and mixing process at the intake station. Afterward, the processed water is fed to water tanks, and a stable supply is provided to each household.

The associate editor coordinating the review of this manuscript and approving it for publication was Honghao Gao.

The water supplied is divided into revenue water flow and Non-revenue water flow. In Seoul, the NRW ratio has been growing gradually because of the construction of water storage tanks and the replacement of aging pipes [2]. The NRW ratios of cities in the world, including Seoul, range from 5 to 50%, exhibiting large differences between countries and regions [3], [4].

According to the waterworks statistics of South Korea, published in 2019, before the water flows into households, many leakages occur through water supply pipes, drainage pipes, and feed water pipes because of aging and other environmental factors (Table 1 & Table2). The South Korean Ministry of Environment announced that 31.4% of all water supply pipes are older than 20 years, and as of 2014, the total amount of water leakage was 690 million tons, amounting to a total loss of 605.9 billion KRW annually [1]. When the NRW ratio increases because of physical leakages or inaccurate measurement, it not only affects the water quality, but also has a negative effect on the general water supply and service system. Therefore, a prediction system to lower the NRW ratio is required.

TABLE 1. Water leakage statistics for South Korea in 2017 (source: Waterworks Statistics 2019) [1].

Column name	Column type	Water supply pipes	Drainage pipes	Feed water pipes	Indoor
Number of leakage events (unit: leakage event)	Reports	495 (0.004%)	15,411 (0.1296%)	58,709 (0.4939%)	44,253 (0.3723%)
	Detections	153(0.0067 %)	5,319(50.2327 %)	15,878(0.6945 %)	1,512(0.095 %)
Number of leakage events (unit: m^2)	Reports	2,900,432	32,161,509	17,377,570	5,205,435
	Detections	990,925	16,906,874	12,451,636	275,370

TABLE 2. Leakage amounts and estimation from 2014-2017 (source: Waterworks Statistics 2019) [1].

Year	Leakage reports	Estimated report amount (m^2)	Total number of leakage detection	Estimated detection amount (m^2)
2014	116,560	99,922,298	22,938	32,826,566
2015	117,398	104,463,407	22,294	31,216,481
2016	132,242	77,391,016	22,641	35,916,067
2017	118,868	57,644,946	22,862	30,624,805

The revenue water ratio, which is defined as the ratio of the volume of water for which the water rate has been collected, against the total volume of tap water production, for South Korea, as of 2017, is approximately 85.2%, which is lower compared to those of other developed countries [5]. Moreover, the regional gaps are deepening. The revenue water ratios with respect to city size are 91.7% for special and metropolitan cities, 80.7% for cities, and 64.1% for counties [1]. Therefore, a comprehensive system for water leakage is required.

In this study, we analyze sensor input data for water usage per household and propose a system that can lower the NRW ratio based on a model for immediate prediction of households with potential water leakage. An example of such a model works by estimating the most efficient position-based service using machine learning [6]–[8].

Previous studies on water usage focused on the image processing, such as in the estimation of water level in rivers. To protect the flooding system of the Mekong River, Nguyen *et al.* [9] proposed a method for forecasting its daily water level using machine-learning models. On the other hand, [10] and [11] estimated the water level by employing image processing methods. These studies were not about prediction models that made use of various related information; rather, these studies were about general image estimation for flooding alert systems.

Jang and Choi [12], [13] predicted the NRW ratio using data on Incheon City via artificial neural network (ANN) and Z-score. They showed that the proposed method using ANN has a higher accuracy than that of the multiple regression analysis method but, in the end, proposed an optimally predictable model because the accuracy varied greatly depending on the number of layers of the hidden area.

Xu *et al.* [14] proposed the continuous deep belief echo state network (CDBESN), which uses a continuous deep belief network to extract features and uses the echo state network as a regression algorithm. The continuous deep belief network is used in various prediction areas and has the advantage of solving the overfitting problem. Thus, existing studies focused on water usage prediction models for entire regions, rather than a water usage alarm system for individual households [13]–[15]. On the other hand, [16] and [17] suggested real-time algorithms for leakage systems on pipeline prototypes based on the extended Kalman filter. Other previous studies detected the fault measurement of urban water networks [18]–[20].

Prediction models for either total revenue water or NRW have already been researched [21]–[23], but research on alarm systems for the NRW of each household is insufficient. In this study, a linear regression model and a time-series model are analyzed based on data on Osan City in Gyeonggi Province. For the linear regression model, the fitness of the model is verified, whereas for the time-series model, the differences between actual measured values and predicted values are determined using two different methods. A model for building an alarm system and a system method are then proposed. Finally, a central alarm system is developed to reduce the NRW ratio based on the proposed model.

II. MODELING

A. EXPERIMENTAL DATA

The data for this study were collected in Osan City, Gyeonggi Province, in South Korea between May 21, 2018, to June 5, 2019, from the sensor data of 869 households. The data were stored in electronic water meters and had the data format shown in Figure 1. For each generation, a total of 11 field values were transmitted wirelessly and stored in the database. The first field value, IDX, represents a unique key value transmitted to the database. The unique value for each household was created via the combination of RESOURCE_ID and WM_ID, as shown in Figure 1. WM_VALUE is a reading value sent from the electronic water meter. The water meter data were sent every eight hours and saved in the database. Thus, three sets of data can be collected per day from each water meter. The data storage time varied depending on the water meter, but data were collected once between 00:00 and 08:00 hours, once between 09:00 and 16:00 hours, and once between 17:00 and 24:00 hours. Table 3 outlines the meaning of each field. Among these data, we used DATE, RESOURCE_ID, WM, and WM_VALUE. Furthermore, the data were extracted for this experiment only if the value of WM_BATTERY was 02, which indicates a normal value, and also when the sensor data were stored with normal conditional values. MAX_FLOW,

```
IDX,RESOUCE_ID,WM_ID,WM_VALUE,WM_BATTERY,MAXIMUM_FLOW,INDOR_LEAK,BACKFLOW,UNUSED,NO_RESPONSE,CRC_ERR,DATE
3144005,140C5BFFFF13010A,17108781,267214,02,0,0,0,0,NULL,NULL,2018-12-03 16:58:32
3144006,140C5BFFFF1300B9,17100610,1066958,02,0,0,0,0,NULL,NULL,2018-12-03 17:01:10
3144007,140C5BFFFF1300F1,1309043,17027751,01,0,0,0,0,NULL,NULL,2018-12-03 17:04:17
3144008,140C5BFFFF1300F1,1309043,17027751,01,0,0,0,0,NULL,NULL,2018-12-03 17:04:23
3144009,140C5BFFFF1300F1,1309043,17027751,01,0,0,0,0,NULL,NULL,2018-12-03 17:04:30
3144010,140C5BFFFF1300FB,16018085,1313171,02,1,1,0,1,NULL,NULL,2018-12-03 17:12:22
3144011,140C5BFFFF130111,15351051,2298094,02,1,1,0,0,NULL,NULL,2018-12-03 17:22:30
```

FIGURE 1. Example of sensor data for each household.

TABLE 3. Data attribute.

Variable name	Data type	Description	Note
RESOURCE_ID WM_ID	VARCHAR(200)	Remote terminal ID	
WM_VALUE	INT	Meter reading (Cumulative water usage)	
WM_BATTERY	CHAR(2)	Meter battery state code	02 : Battery is good 01 : Battery is low 00 : No battery
MAXIMUM_FLOW	CHAR(1)	Whether or not the maximum flow is exceeded	0 : Normal 1 : Maximum flow exceeded
INDOR_LEAK	CHAR(1)	Indoor leak warning	0 : Normal 1 : Indoor leak warning
BACKFLOW	CHAR(1)	Backflow warning	0 : Normal 1 : Backflow warning
UNUSED	CHAR(1)	Unused warning	0 : Normal 1 : Unused warning
NO_RESPONSE	CHAR(1)	Meter no response code	0 : Normal 1 : No response
CRC_ERR	CHAR(1)	Meter data error code	NULL: Normal 1 : CRC error
DATE	DATETIME	Collected date and time	

INDOOR_LEAK, BACKFLOW, and UNUSED values are unstable, and thus these attributes were not considered. On the other hand, NULLs in NO_RESPONSE and CRC_ERROR were considered as normal values.

B. MODELS

Three models were analyzed using the data on Osan City.

First, regression analysis was used for prediction. Among the traditional prediction methods, regression analysis is used mainly when the component has a constant effect that does not depend on time. For the auto-correlation test of the error term for prediction, the fitness of the regression model was determined using the Durbin-Watson (DW) value.

Second, time-series analysis was also used for the prediction. Unlike the regression model, time-series analysis is used when time-dependent analysis is required. The ARIMA (autoregressive integrated moving average) model even considers the trend of past data and considers both autoregression and moving average. The difference between the ARIMA and ARMA (autoregressive moving average) models is that the ARIMA uses differences between observation values to describe the non-stationary property of the time series.

Lastly, the prediction accuracy of data was analyzed via the additive regression model. The additive regression model Prophet [24] used in this study is not time-dependent and can solve problems via curve fitting, whereas time-series models, such as ARIMA, are time-dependent. To build a more accurate model, two different options, one which considers the yearly trend and one which does not, were applied.

1) LINEAR REGRESSION MODEL

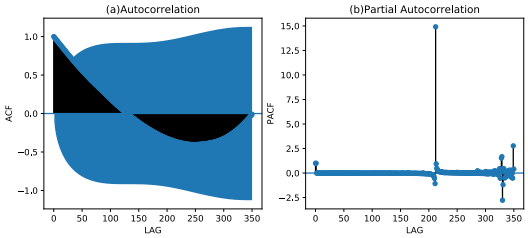
Linear regression model is a method of modeling correlations between one or more independent and dependent variables. Simple linear regression analysis is used when the definitions of dependent variables are analyzed according to one independent variable, whereas multiple linear regression analysis is used when multiple independent variables affect the dependent variables. In Equation (1), n denotes the number of independent variables.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots, n, \varepsilon_i \sim (0, \sigma^2) \tag{1}$$

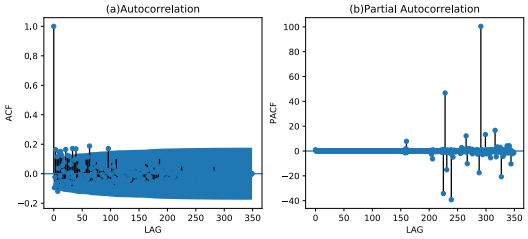
In the Osan City water data, the independent variable is time, and the dependent variable is the cumulative water usage, which is WM_VALUE. For the auto-correlation test of the error term for prediction, the DW value is used to determine whether the term is fit for the regression model. Equation (2) is an equation for obtaining the DW value, where ε_i signifies the i th residual, Y_i is the result obtained from actual sensor data, and $\hat{Y}_i$ is the value estimated by the least square line.

$$\frac{\sum_{i=1}^{n-1} \varepsilon_{i+1} - \varepsilon_i{}^2}{\sum_{i=1}^{n} \varepsilon_i{}^2}$$
$$\varepsilon_i = Y_i - \hat{Y}_i. \tag{2}$$

When DW value is close to 0, it indicates a positive correlation, whereas when DW value is close to 4, it indicates

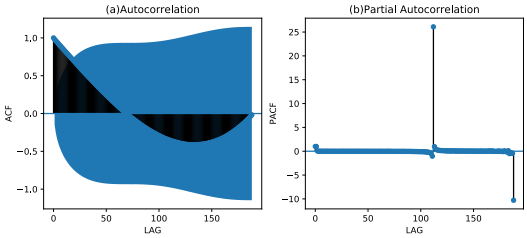


(a) The original data of ACF and PACF of 140C5BFFFF130105_16014537

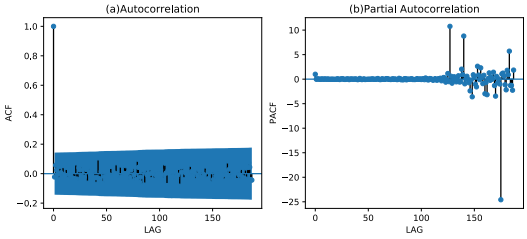


(b) The first-order difference of ACF and PACF of 140C5BFFFF130105_16014537

FIGURE 2. The ACF and PACF of 140C5BFFFF130105_16014537.



(a) The original data of ACF and PACF of 140C5BFFFF1301F5_15350104



(b) The first-order difference of ACF and PACF of 140C5BFFFF1301F5_15350104

FIGURE 3. The ACF and PACF of 140C5BFFFF1301F5_15350104.

a negative correlation. When DW value is closer to 0 or 4, it indicates that the regression model is unfit, because it has many residuals. By contrast, when this value is closer to 2, there is a static correlation. When the average DW value obtained using the data is 2, there is no auto-correlation. When it is between 0 and 2, there is a positive auto-correlation, and the data are general time-series data. If it is higher than 2 and less than 4, there is a negative auto-correlation, and the data are unusual time-series data. Therefore, the feasibility of the linear regression model is determined by the DW value.

In this experiment, the DW value of each household was calculated using the statsmodels Python module [25]. The average DW value of all the calculation data is 2.83E-05, which is close to zero. In other words, it is unfit for regression model. Thus, learning with additive regression model is proposed.

2) ARIMA MODEL

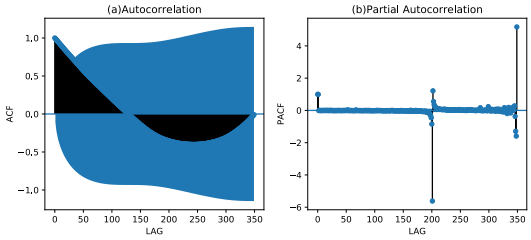
Regression analysis analyzes the correlation between dependent variable and independent variable but does not consider time. Thus, when a previous time affects the current modeling, time-series analysis is typically used. The ARIMA model is a representative time-series model that can encompass the AR (auto correlation) model, the MA (moving average) model, and the ARMA model. It can also reflect the trend of past data and is expressed as ARIMA(p, d, q), where p denotes the lag of the AR model, q denotes the lag of the MA model, and d denotes the difference. The auto-correlation function (ACF) and partial auto-correlation function (PACF) were calculated using the statsmodels python package [25] by selecting"140C5BFFFF130105+16014537", "140C5BFFFF1301F5+15350104", "140C5BFFFF1300CF +14019387", each of which represents one of many households [Figure 2(a), Figure 3(a) and Figure 4(a)].

To calculate the proper difference order, first-order difference was first performed, and then ACF and PACF were calculated again [Figure 2(b), Figure 3(b) and Figure 4(b)]. ACF shows that the auto-correlation shifts from positive to negative at around 150, whereas the PACF shows sharply decreasing trends near 220 and 339, excluding slight errors. Therefore, we assume that $p = 0$ and $q = 1$. The following graph shows first-order differentiated ACF and PACF. Because the time-series state is normal, the ARIMA(0,1,1) model is used. The ARIMA(0,1,1) model was adopted because other household data different from "140C5BFFFF130105+16014537" also exhibited similar results.

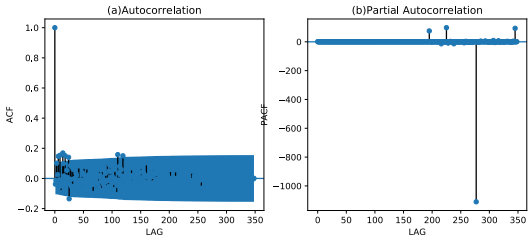
Figure 2, Figure 3 and Figure 4 visualize examples of randomly selected data that consist of the original data and the first-order difference of ACF and PACF. Graph patterns in these three different figures for the original data and the first-order difference between them are very similar.

3) ADDITIVE REGRESSION MODEL

Harvey and Peters [26] computed the maximum likelihood with the time domain, frequency domain, and expectation-maximization (EM) algorithm in the structural time-series model. Similar to the model of Harvey and Peters, the additive regression model follows $y(t) = g(t) + s(t) + h(t) + e$, where $g(t)$ is a trend with no repetitive element, $s(t)$ represents repetitive changes, such as day of the week or annual seasonality, $h(t)$ is an irregularly influencing factor, such as holidays, and e is a residual that follows normal distribution. However, holiday $h(t)$ was not considered because the current data set does not have a special and irregularly influencing factor such as holidays. To express via generalization, the s, g, and h functions can be expressed as function f, as follows. The additive model is an



(a) The original data of ACF and PACF of 140C5BFFFF1300CF_14019387



(b) The first-order difference of ACF and PACF of 140C5BFFFF1300CF_14019387

FIGURE 4. The ACF and PACF of 140C5BFFFF1300CF_14019387.

extension of multiple linear regression and can be expressed as follows:

$$\begin{aligned} Y_i &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + \cdots + f_p(x_{ip}) + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i \\ \mathbb{E}[\varepsilon = 0, \quad & Var(\varepsilon) = \sigma^2 \end{aligned} \tag{3}$$

A time-series model, such as ARIMA, has a time-dependent structure, but Prophet, which is the additive regression model used in this study, is not time-dependent and can solve problems via curve fitting. Two additive models were established. The first model set the day of the week, seasonality, and yearly periodicity as basic data. However, the currently used data have not been accumulated for more than one year. Thus, the second model was set with no yearly and seasonal components, and considered only time, day of the week, and monthly components using Fourier series.

This model is called an additive regression model because each x_i used as an element variable of the model is calculated separately, and their contributions are summed up. This model has the advantage of automatically modeling nonlinear relationships that can be defined by standard linear regression because a nonlinear function can be fit for each element.

III. SYSTEM

A. DATA PREPROCESSING

Figures 5 and 6 show the system model and used data. The input sensor data values are stored in the central database. The system performs preprocessing for each household ID by removing erroneous data and selecting data for learning. For real-time distributed processing of large data, Spark-based streaming was applied. Spark is an open-source distributed query and processing engine and provides scalability to MapReduce at a much faster speed. Scalability was considered to enable data analysis of several hundreds of thousands of homes in the future via real-time distributed processing of data stored in a database. According to the settings, data are imported from the database in Spark every N seconds, which is described as time slot in Figure 6. The time when Spark extracts data is marked in red. However, the original database does not store data sorted by household. The data of various households are mixed, as shown in the top-center part of the figure. Independent data with unique IDs by household are indispensable for the time-series analysis of each household.

Pre-processing was performed, wherein the ID was set by the combination of Resource ID and WM_ID among the fields input from the sensor data. Checks were first conducted on whether or not to perform a test for cases where the

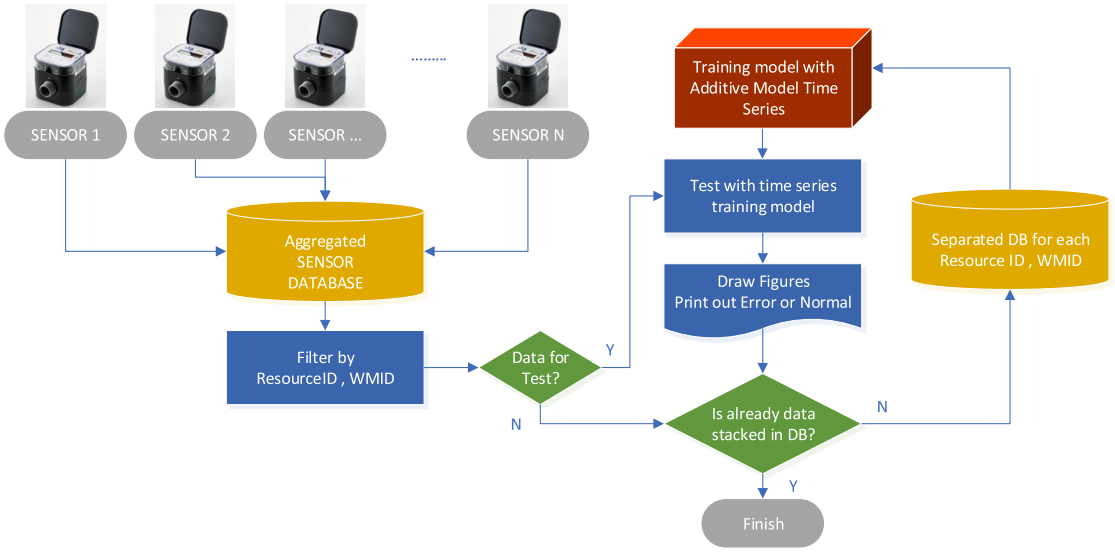


FIGURE 5. Overall system flow for training and testing with the sensor data.

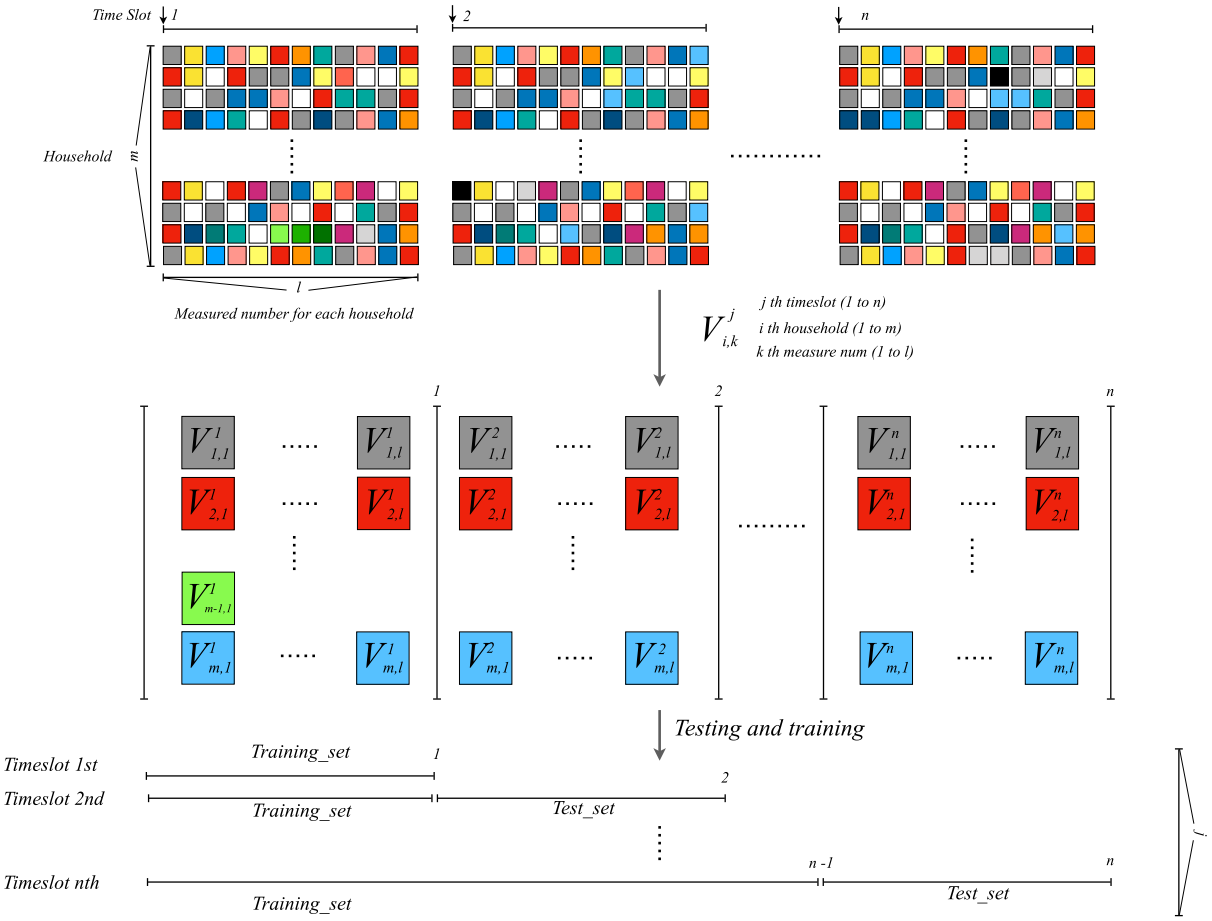


FIGURE 6. Pre-processing of training and testing sets. All data from the 1st to (*j*-*1*)th time slot are used as training data, whereas data in *j* th time slot are used as testing data.

preprocessed data repositioned in each household moved out of the predicted minimum to maximum range. In Figure 6, each V_{ik}^{j} is a part where a decision was made about data processing. If the preprocessed IDs had been stacked in the learned data, the preprocessed data were used for the test. However, if no data had been stacked, that is, if no water usage data had been stored for the ID, or if this model was started first, data for learning were stored separately for that ID.

In the case of error processing of duplicate processed data in the database, checks were also made on whether data for learning had been stacked. The analysis of trend with respect to day of the week for each household can be performed only if at least 42 rows of data have accumulated in the database. For trend analysis, cumulative data for at least two weeks are necessary, and because data are accumulated three times per day, 7 (days in a week) * 2 (number of weeks) * 3 (data per day) = 42 rows of data were set. In Figure 6, the area marked by a light green box is excluded from the criteria for accuracy because test data have not been collected sufficiently if the number of continuous accumulated rows of data is less than 42, although these are accumulated as learning data. The recent data preprocessed using Resource_ID+WM_ID for each household from the sensors were tested by the created model only if learned data already existed in the household. If there were no accumulated data, the sensor of the household was considered as having been newly installed. When the aforementioned 42 data had been accumulated, the data trends with respect to day of the week, month, and hour were modeled using the accumulated data.

For the preprocessed data for learning, whether the observed data went out of the minimum-to-maximum range predicted by the model using the observed data was determined through the process of "Test with time-series training model." If the observed data went out of the minimum-to-maximum range of the model's prediction, in the "Test with time-series training model." shown in Figure 5, a warning message was displayed, urging for the sensor error and other items to be checked because the predicted water usage exceeded the normal range. On the other hand, if water leak was doubted or if water used was less than normal, "Draw Figures; Print out error or normal" as shown in Figure 5, was performed.

After the test had terminated, whether the data have already been learned or not was determined, or the water usage values that had been used as learning data were accumulated for each ID in the database. The displayed data were sent in the format shown in Figure 7. The data are separated by tabs, where the first column denotes the ID of each household, the second

RESOURCE_ID+WM_ID	Date	Sensor value	Estimated Lower bound	Estimated Upper bound	ERROR
140C5BFFFF130181+16047892	2019-05-21 18:53:49	4919287	4906448.460084559	4934807.17031535	
140C5BFFFF13013C+17108800	2019-05-21 18:55:02	201406	201472.75256282662	202101.0888927272	ERROR
140C5BFFFF1301B7+16012645	2019-05-21 18:55:50	8524007	8508399.086775355	8543960.524630796	
140C5BFFFF1301E2+17250298	2019-05-21 18:53:54	7441862	7402628.78796547	7437125.664171245	ERROR
140C5BFFFF130163+15350152	2019-05-21 18:54:40	279213	278691.24812063866	279564.2868437001	
140C5BFFFF130167+17320254	2019-05-21 18:55:04	1197612	1189471.717723453	1194105.2812808168	ERROR
140C5BFFFF130112+18108504	2019-05-21 18:56:20	91824	90973.79040262818	91612.65005564902	ERROR

FIGURE 7. Example format of trained data sets. Columns represent the household number, date, measured sensor data, estimated minimum value, maximum value, and error status, respectively. The error status in the last column refers to out-of-range values, whereas a blank value indicates a normal condition.

TABLE 4. Accuracy measurements for the three different methods.

Method	boundMAE	minMAE	boundMSE	minMSE2	Average accurate rate
ARIMA	39,230.88	38,250.22	111,923,576,931.93	111,859,836,297.62	0.64
Fbprophet with considering yearly trend	17,635.15	13,580.30	31,978,695,924.27	23,677,996,349.83	0.46
Fbprophet with no considering yearly trend	23,566.54	15,727.04	68,072,978,330.04	34,345,343,071.18	0.65

column is the testing time, the third column is the actual value stored through the sensor, and the fourth and fifth columns signify the predicted minimum-to-maximum range. Lastly, the message ERROR was printed in the last column when the data went out of the minimum-to-maximum range. If the value in this column is ERROR, the system can display a warning message. If the error appears repeatedly for a certain period, water leak or sensor error checks can be performed through due diligence, in real time.

B. SYSTEM CONFIGURATION

The experiment was performed in a docker environment where Ubuntu, Spark, and Python were installed. As shown in Figure 5, the data obtained from the sensors vary by the number of measured sensors and by the time when the sensors were installed in the household. The measured test data per household are composed of 776 rows on average, which represent around 9 months of accumulated data. The measured data are stored in a temporary database, are received three times per day at different times, and are not yet sorted by household during storage. Whether there were data input through the sensors was continuously checked in real time by the docker-based Spark system. If new sensor data were stored into the database, the unsorted data were pre-processed to distinguish them by household, and the test module was run. After the optimization model was created based on the learning data that had been stacked for each household, the recently input data were tested. If the data exceeded or fell short of the predicted minimum-to-maximum range based on the model, there was likely a tentative error. If more than a few errors had accumulated, the warning system was activated, and physical inspection for water leak or sensor error could be performed for the corresponding machine.

For accuracy calculation, it was considered that the data input from sensors were stored into the database three times per day. The accuracy was calculated via testing at each time Spark took data from the database, according to the sequence shown in Figure 6. The total number of tested IDs was 869. After insignificant data were filtered out, because no sensor data were received, as shown in Figure 1, 289 IDs remained to be used for significant tests. The average number of tested data rows per ID was 670 in total. Because these data were recorded three times per day, the test used data for 670/3 = 225 days (approximately 7 months). The average error rate and accuracy were calculated using Equations (4) to (12), and the results are outlined in Table 4.

Algorithm 1 Testing and Making a Training With the Sensor Data

Input: Hourly Sensor data for each household

Output: Check the sensor data is over or under the estimated value

Initialization:
V_{ik}^{j} is i th household testing data.
TR_i is i th household trained data.
j is time slot for testing.
i is separated household number.
k is measured number for each household.

for $j = 1$ to n **do**
  for $i = 1$ to m **do**
    for $k = 1$ to l **do**
      $\widehat{Y_{ik}^{j}\_max}$ <- Calculate ith, jth and kth maximum range of water usage estimation
      $\widehat{Y_{ik}^{j}\_min}$ <- Calculate ith, jth and kth minimum range of water usage estimation
      // *check the number of training set is enough or not.*
      if num of $(TR_i) > 21$) **then**
        if $V_{ik}^{j} > (\widehat{Y_{ik}^{j}\_max})$ or $V_{ik}^{j} < (\widehat{Y_{ik}^{j}\_min})$ **then**
          Send the alert message to system.
          Save the alert figure.
        else
          Normal condition.
        end if
      end if
      Save $TR_i < -V_{ik}^{j}$
    end for
  end for
end for

To test the three models, five values were compared, as shown in Table 4. In the results shown in Figure 7, an ''ERROR'' message is displayed in the last tabbed separated column if the predicted value exceeds or falls short of the accurate value by a certain range. Therefore, the mean square error (MSE) was calculated in two ways. For performance evaluation, two equations, transformed from the general MSE Equation (4), have been suggested. For each household, if the measured water usage data value is out of range, i.e., not between the estimated minimum and maximum, the developed system sends an alert alarm. Otherwise, the system considers households where the estimated values are within the normal range as being in the normal status. Therefore, the equation was changed to Equations (5) and (6). Equation (5) represents the difference between the actual predicted value and the minimum/maximum predicted values at current time t, whereas Equation (6) signifies the minimum difference between the actual predicted value and the minimum or maximum predicted value. Thus, M is the number of tested households, among the total number of households, for which a valid amount of water sensor data has been accumulated. The value of M is 289. N is the number of input sensor data rows for each household, which were recorded three times per day. Because the numbers of stacked sensor data rows are different by household, N ranges widely from 42 to 809. $Y_{j\_t\_min}$ and $Y_{j\_t\_max}$ represent the minimum and maximum values, respectively, Y_{j_t} predicted at time t in the jth household, whereas Y_{j_t} denotes the actual consumption value recorded by the water meter as of time t in the j th household.

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(\hat{Y}_t - Y_t)^2 \tag{4}$$

$$boundMSE = \frac{1}{M}\sum_{j=1}^{M}\Big(\frac{1}{N}\sum_{t=1}^{N}(\widehat{Y_{j\_t\_min}} - Y_{j\_t})^2 + \frac{1}{N}\sum_{t=1}^{N}(\widehat{Y_{j\_t\_max}} - Y_{j\_t})^2\Big) \tag{5}$$

$$boundMSE = \frac{1}{M}\sum_{j=1}^{M}\Big(\frac{1}{N}\sum_{t=1}^{N}(min(\widehat{Y_{j\_t\_min}} - Y_{j\_t})^2, (\widehat{Y_{j\_t\_max}} - Y_{j\_t})^2)\Big) \tag{6}$$

Equation (7) expresses the mean absolute error (MAE), which, similar to MSE, is used often for evaluation, and the equation was additionally modified to Equations (8) and (9). boundMAE is the average value of the sum of the predicted values for each of the households and the absolute value of the difference between the minimum and maximum. In other words, the average for all households based on the average value for each household is boundMAE. Similarly, minMAE is the average value for all households based on the actual recorded values at time t and the average between the predicted minimum and maximum for each household. If the minimum-to-maximum range widens, a warning message is sent based on relaxed criteria, but the boundMSE, minMSE, boundMAE, and minMAE values will increase. Therefore, appropriate criteria are required.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{Y}_i - Y_i| \tag{7}$$

$$boundMAE = \frac{1}{M}\sum_{j=1}^{M}\Big(\frac{1}{N}\sum_{t=1}^{N}|\widehat{Y_{j\_t\_min}} - Y_{j\_t}| + \frac{1}{N}\sum_{t=1}^{N}|\widehat{Y_{j\_t\_max}} - Y_{j\_t}|\Big) \tag{8}$$

$$boundMAE = \frac{1}{M}\sum_{j=1}^{M}\Big(\frac{1}{N}\sum_{t=1}^{N}(min(|\widehat{Y_{j\_t\_min}} - Y_{j\_t}|, (|\widehat{Y_{j\_t\_max}} - Y_{j\_t}|))\Big) \tag{9}$$

$$Err\_cnt_{t\_j} = \begin{cases} 1 & \text{if } \widehat{Y_{j\_t\_max}} < Y_{j\_t} \\ & \text{or } \widehat{Y_{j\_t\_min}} > Y_{j\_t} \\ 0 & \text{if } \widehat{Y_{j\_t\_min}} < Y_{j\_t} < \widehat{Y_{j\_t\_max}} \end{cases} \tag{10}$$

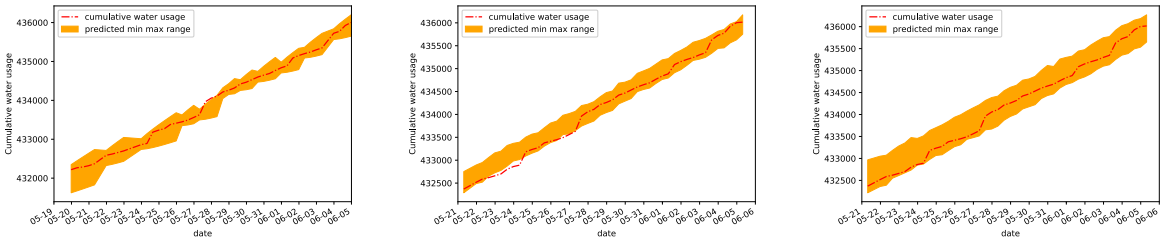
$$Total\_err\_cnt_j = \sum_{t=1}^{N} Err\_cnt_{t\_j} \tag{11}$$

$$Acc_j = 1 - \frac{Total\_err\_cnt_j}{N} \tag{12}$$

$$Avg\_acc\_rate = \sum_{j=1}^{M} Acc_j \tag{13}$$

The following equations are used to calculate the mean error. An error is defined as when the data Y_{j_t} measured at time t from an actual sensor in the jth household moves out of the range of the minimum predicted value $Y_{j\_t\_min}$ and the maximum predicted value $Y_{j\_t\_max}$. The error at time t in the jth household is defined by Equation 11, and all the error values of the j th household are accumulated using Equation 12. Equation 13 expresses the prediction accuracy of the jth household, which is 1 minus the error ratio for the total number N of tested data. Therefore, the average accurate rate can be calculated using Equation 10 to Equation 13. Table 4 shows the actual calculated values.

Figure 8 illustrates the actual usage data and the predicted data for the household ID ''140C5BFFFF130058+14001361'' for the limited period of 2019-05-19 to 2019-06-06. As shown in Figure 8(a), the water usage data were predicted using the ARIMA model. MAE and MSE are relatively high, and the average accuracy rate is also high. However, as shown in Figure 8(b), a yearly trend applied to the additive regression model exhibits a low average accuracy rate. In Figure 8(c), the minimum and maximum values are predicted to have a wider range than in Figure 8(b). Thus, the figure depicts a high average accuracy rate, and higher MAE and MSE than when yearly trend was applied.



(a) Minimum to maximum range predicted for the period from 2019-05-19 to 2019-06-06 and actual values using ARIMA method.

(b) Minimum to maximum range predicted for the period from 2019-05-19 to 2019-06-06 and actual values using fbprophet with considering the yearly trend.

(c) Minimum to maximum range predicted for the period from 2019-05-19 to 2019-06-06 and actual values using fbprophet with no considering the yearly trend.

FIGURE 8. Predictive range for each method.

IV. CONCLUSIONS

A model and a system for analyzing the water usage pattern of each household, based on actual water usage data, were proposed, and a system that can predict water leak risk was developed. The system was designed to enable real-time data analysis. Users can see the data-based analysis results for water leak risk through a Web-based analysis system, and current data can be provided through continuous data updates. The proposed system provides an opportunity to monitor in real time the water leaks and errors of the sensor systems, which otherwise increase non-revenue water in both rural and urban areas.

Furthermore, in this study, an additive regression model was used after it was first demonstrated that the data for each household do not follow linear regression. Because the water usage pattern can vary by household, we first determined whether a household followed the regression analysis model and examined the application of different models, depending on the result. In addition, we were able to analyze the accuracy of real errors in the warning system using real data, by adding monthly analysis elements, and verified an improved model using data accumulated for many years. The results of this analysis will enable real-time monitoring of leaks and sensor machine errors.

At present, all the accumulated data are used for learning. As the years pass, the learning time is expected to increase because of the increase in learning data. Therefore, in future studies, it will be necessary to train all cumulative periods for all purposes or to find an optimized learning period. Moreover, the water usages for each household according to time, e.g., daily, monthly, and seasonally patterns, need to be considered separately. We will focus on grouping or clustering related information on each household for an efficient water supply system as a future work.

REFERENCES

[1] *Statistics of Waterworks*, Ministry Environ., Sejong, South Korea, 2018.
[2] Y. Choi, J. Ahn, H. Im, and A. Koo, "Best management practices for water loss control in seoul," *Procedia Eng.*, vol. 89, pp. 1585–1593, 2014.
[3] B. A. R. Frauendorfer and R. Liemberger, *The Issues and Challenges of Reducing Non-Revenue Water*. Mandaluyong, Philippines: Asian Development Bank, 2010.
[4] B. A. K. S. Perera, H. Mallawaarachchi, K. S. Jayasanka, and R. R. P. N. Rathnayake, "A water management system for reducing non-revenue water in potable water lines: The case of Sri Lanka," *Engineer*, vol. 51, no. 2, p. 53, Jul. 2018.
[5] J. Cha and J. Kim, "Leakage reduction through establishment of block system in Jeju City," *J. Korean Soc. Water Wastewater*, vol. 26, no. 5, pp. 693–703, Oct. 2012.
[6] A. K. Sangaiah, D. V. Medhane, T. Han, M. S. Hossain, and G. Muhammad, "Enforcing position-based confidentiality with machine learning paradigm through mobile edge computing in real-time industrial informatics," *IEEE Trans. Ind. Inf.*, vol. 15, no. 7, pp. 4189–4196, Jul. 2019.
[7] A. K. Sangaiah, M. Sadeghilalimi, A. A. R. Hosseinabadi, and W. Zhang, "Energy consumption in point-coverage wireless sensor networks via bat algorithm," *IEEE Access*, vol. 7, pp. 180258–180269, 2019.
[8] A. K. Sangaiah, D. V. Medhane, G.-B. Bian, A. Ghoneim, M. Alrashoud, and M. S. Hossain, "Energy-aware green adversary model for cyber physical security in industrial system," *IEEE Trans. Ind. Inf.*, to be published.
[9] T.-T. Nguyen, Q. N. Huu, and M. J. Li, "Forecasting time series water levels on mekong river using machine learning models," in *Proc. 7th Int. Conf. Knowl. Syst. Eng. (KSE)*, Oct. 2015, pp. 292–297.
[10] K. Tanaka, Y. Fujihara, K. Hoshikawa, and H. Fujii, "Development of a flood water level estimation method using satellite images and a digital elevation model for the Mekong floodplain," *Hydrolog. Sci. J.*, vol. 64, no. 2, pp. 241–253, Jan. 2019.
[11] R. Hostache, X. Lai, J. Monnier, and C. Puech, "Assimilation of spatially distributed water levels into a shallow-water flood model. Part II: Use of a remote sensing image of Mosel River," *J. Hydrol.*, vol. 390, nos. 3–4, pp. 257–268, Sep. 2010.
[12] D. Jang and G. Choi, "Estimation of non-revenue water ratio for sustainable management using artificial neural network and Z-Score in Incheon, Republic of Korea," *Sustainability*, vol. 9, no. 11, p. 1933, Oct. 2017.
[13] D. Jang and G. Choi, "Estimation of non-revenue water ratio using MRA and ANN in water distribution networks," *Water*, vol. 10, no. 1, p. 2, Dec. 2017.
[14] Y. Xu, J. Zhang, Z. Long, H. Tang, and X. Zhang, "Hourly urban water demand forecasting using the continuous deep belief echo state network," *Water*, vol. 11, no. 2, p. 351, Feb. 2019.
[15] G. Romano, N. Salvati, and A. Guerrini, "Estimating the determinants of residential water demand in italy," *Water*, vol. 6, no. 10, pp. 2929–2945, Sep. 2014.
[16] A. Navarro, O. Begovich, G. Besancon, and J. Dulhoste, "Real-time leak isolation based on state estimation in a plastic pipeline," in *Proc. IEEE Int. Conf. Control Appl. (CCA)*, Sep. 2011, pp. 953–957.
[17] J. Delgado-Aguiñaga, G. Besançon, O. Begovich, and J. Carvajal, "Multi-leak diagnosis in pipelines based on Extended Kalman Filter," *Control Eng. Pract.*, vol. 49, pp. 139–148, Apr. 2016.
[18] J. Ragot and D. Maquin, "Fault measurement detection in an urban water supply network," *J. Process Control*, vol. 16, no. 9, pp. 887–902, Oct. 2006.
[19] S. Abbas, Y. Xuan, and R. Bailey, "Improving river flow simulation using a coupled surface-groundwater model for integrated water resources management," in *Proc. HIC 13th Int. Conf. Hydroinform.*, vol. 3. Palermo, Italy: EasyChair, 2018, pp. 1–9.

[20] S. Fellini, R. Vesipa, F. Boano, and L. Ridolfi, "Fault detection in level and flow rate sensors for safe and performant remote-control in a water supply system," *J. Hydroinform.*, vol. 6, 2019.
[21] A. Candelieri, "Clustering and support vector regression for water demand forecasting and anomaly detection," *Water*, vol. 9, no. 3, p. 224, Mar. 2017.
[22] Y. Bai, P. Wang, C. Li, J. Xie, and Y. Wang, "Dynamic forecast of daily urban water consumption using a variable-structure support vector regression model," *J. Water Resour. Planning Manage.*, vol. 141, no. 3, Mar. 2015, Art. no. 04014058.
[23] L. Cai, R. Wang, J. Ping, Y. Jing, and J. Sun, "Water supply network monitoring based on demand reverse deduction (DRD) technology," *Procedia Eng.*, vol. 119, pp. 19–27, 2015.
[24] S. J. Taylor and B. Letham, "Forecasting at scale," *Amer. Statistician*, vol. 72, no. 1, pp. 37–45, 2018.
[25] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," in *Proc. 9th Python Sci. Conf.*, Jan. 2010.
[26] A. C. Harvey and S. Peters, "Estimation procedures for structural time series models," *J. Forecasting*, vol. 9, no. 2, pp. 89–108, 1990.

MINGEUN JI received the bachelor's degree in computer science and engineering from Gangneung–Wonju National University, South Korea, in 2018. He is currently pursuing the master's degree with Dongguk University, Seoul, South Korea. His research area is the data science and computational biology.

GANGMAN YI received the master's and Ph.D. degrees in computer science from Texas A&M University, USA, in 2007 and 2011, respectively. In 2011, he joined the System Software Group, Samsung Electronics, Suwon, South Korea. He was with the Department of Computer Science and Engineering, Gangneung–Wonju National University, South Korea, in 2012. Since 2016, he has been with the Department of Multimedia Engineering, Dongguk University, Seoul, South Korea. He has been researched in an interdisciplinary field of researches. His research interests include the development of computational methods to improve understanding of biological systems and its big data. He actively serves as a Managing Editor and a Reviewer for international journals and the Chair of the international conferences and workshops.

JAEHEE JUNG received the master's degree in computer sciences from Korea University, South Korea, in 2002, and the Ph.D. degree in computer sciences from Texas A&M University, USA, in 2008. From March 2002 to June 2003, she was with the Mobile Software Group, LG Electronics, Seoul, South Korea. In February 2009, she joined the Mobile Application Platform Research and Development Team, Samsung Electronics, Suwon, South Korea. From 2014 and 2019, she was an Assistant Professor with the Department of General Education, Hongik University. Since 2019, she has been with the Department of Information and Communications Engineering, Myongji University, South Korea. Her data science research focuses especially on the development of computational methods to improve understanding of practical big data.