**이상치: 기존 값과 다른 값**

**성능 개선: 기법의 정확도 높여서 이상치를 판단하자**

**신뢰성 일단 제외**

**TVOC 1개 device**

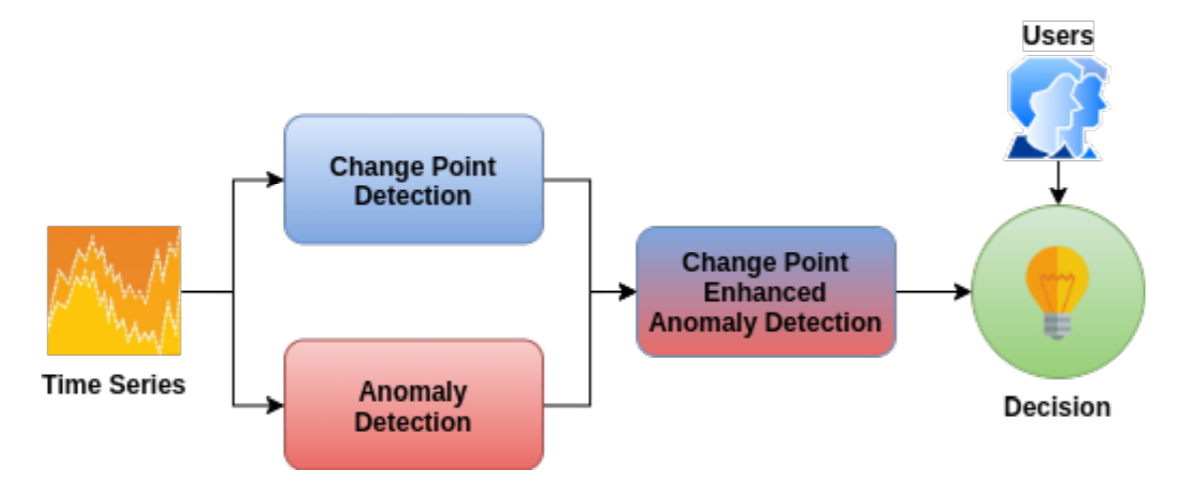
**이상치 탐지 왜? 예지 정비, 사용자에게 알람, 신뢰성**

**모델? 시계열 예측(ARIMA, prophet), 차원축소(PCA, AE),**

**Change Point Enhanced Anomaly Detection for IoT Time Series Data**

**MDPI, Water**

**-> Anomaly Detection + Change point detection**



**1. 정의**

**Anormal Detection 정의:**

An outlier or an anomaly is a **data point that significantly differs from other observations in a time series.**

**Outliers can appear** **due to an experimental error or an anomaly in the measurement**. Such suspicious points in the time series data must be identified and interpreted separately in order not to interfere with the analysis step and lead to wrong conclusions.

**Anomaly (or outlier) detection is the identification of events, items, or observations that do not correspond to normal behavior.**

Along with anomaly detection, **change point detection** plays an important part in time series analysis. It indicates an unexpected and significant change in the analyzed time series stream that represents a normal behavior and not an anomaly.

Anomaly detection is an important part of time series analysis, improving the quality of further analysis, such as prediction and forecasting. Thus, detecting sudden change points with normal behavior and using them to discriminate between abnormal behavior, i.e., outliers, is a crucial step used to minimize the false positive rate and to build accurate machine learning models for prediction and forecasting. In this paper, we propose a rule-based decision system that enhances anomaly detection in multivariate time series using change point detection.

**2. dataset**

시계열: xt = mt + st + yt

(i) mt : the trend component represents variations of low frequency and can be determined by the moving averages or spectral smoothing methods;

(ii) st : the seasonal component is a function that represents normal fluctuations that are more or less stable after a known period (or lag) h;

(iii) yt : the noise (residual) are used to check if a analysis model has correctly determined the information in the data points and can help to predict future values

For our experiment, we used a dataset containing customer water consumption information collected by smart meters. **In total, we had available 119 multivariate time series, each collected over a period of 6 to 24 months at intervals of 5 s. The dimensions were as follows: flow, pressure, outside humidity, outside temperature, rainwater quantity**

Anomaly detection in time series is a complex task that requires large datasets to accurately discriminate between points that present a normal and abnormal behavior. Thus, the experiments need to be conducted on real data to create accurate models**. In our experiments, we used a real-world dataset containing 119 multivariate time series, each collected over a period of 6 to 24 months at intervals of 5 s. The total size of the dataset is over 4.8 GB.** Using such a large dataset, we manage to fine-tune our models.

**3. 메커니즘**

we propose a novel rule-based decision system that **enhances anomaly detection using change point detection** for IoT time series.

Furthermore, many anomaly detection algorithms have high false positive rates, as they tend to label the change points as outliers. Thus, a method that distinguishes between them is required to increase the accuracy of anomaly detection systems and alleviate the decision process. By automating the matching process between anomalies and change points, the required human interaction and interpretation can be reduced only to rare cases. This process can be improved by using a rule-based decision system that manages to discriminate between a change point and an anomaly with high confidence.

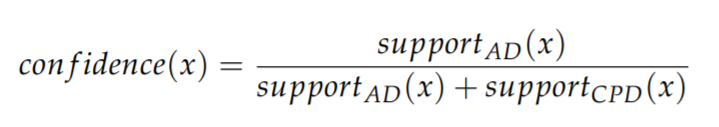
**-> Anomaly Detection 과 Change point Detection을 구분하는 것이 주 목적**

We use five anomaly detection algorithms: **Gaussian Distribution (GD), K-Means, Isolation Forest (IF), OC-SVM (OneClass Support Vector Machine), and Autoencoders (AEs).** For change point detection, we also use five algorithms: window-based segmentation (WinSeg), binary segmentation (BinSeg), bottom-up segmentation (BottonUp), Pruned Exact Linear Time (PELT), and exact segmentation dynamic programming model (OPT).

**-> 5개 Anomaly Detection 기법과, 5개 change point detection 기법 사용**

**-> PCA 사용하여 차원의 저주 없애고 차원 축소**

-> The Change Point Enhanced Anomaly Detection Module computes the confidence for a data point to be an outlier.



When a data point is both labeled as an anomaly and a change point, the confidence(x) function can remove the obvious false positives introduced by change point.

A1 When the support for anomaly detection is high, and there is no or low support for change point detection, then the system automatically marks the point as an anomaly.

A2 When the support for change point detection is high, and there is no or low support for anomaly detection, the system marks the point as a change point and not an anomaly.

A3 When the support for change point detection is low, and there is no support for anomaly detection, the system marks the point as a change point and not an anomaly. Human intervention is still required for the following very rare cases:

H1 When the anomaly support is low and no change point is detected or the change point support is low, regardless if supportAD(x) > supportCPD(x) or supportAD(x) < supportCPD(x).

H2 When both the anomaly support and change point support are high, regardless if supportAD(x) > supportCPD(x) or supportAD(x) < supportCPD(x).

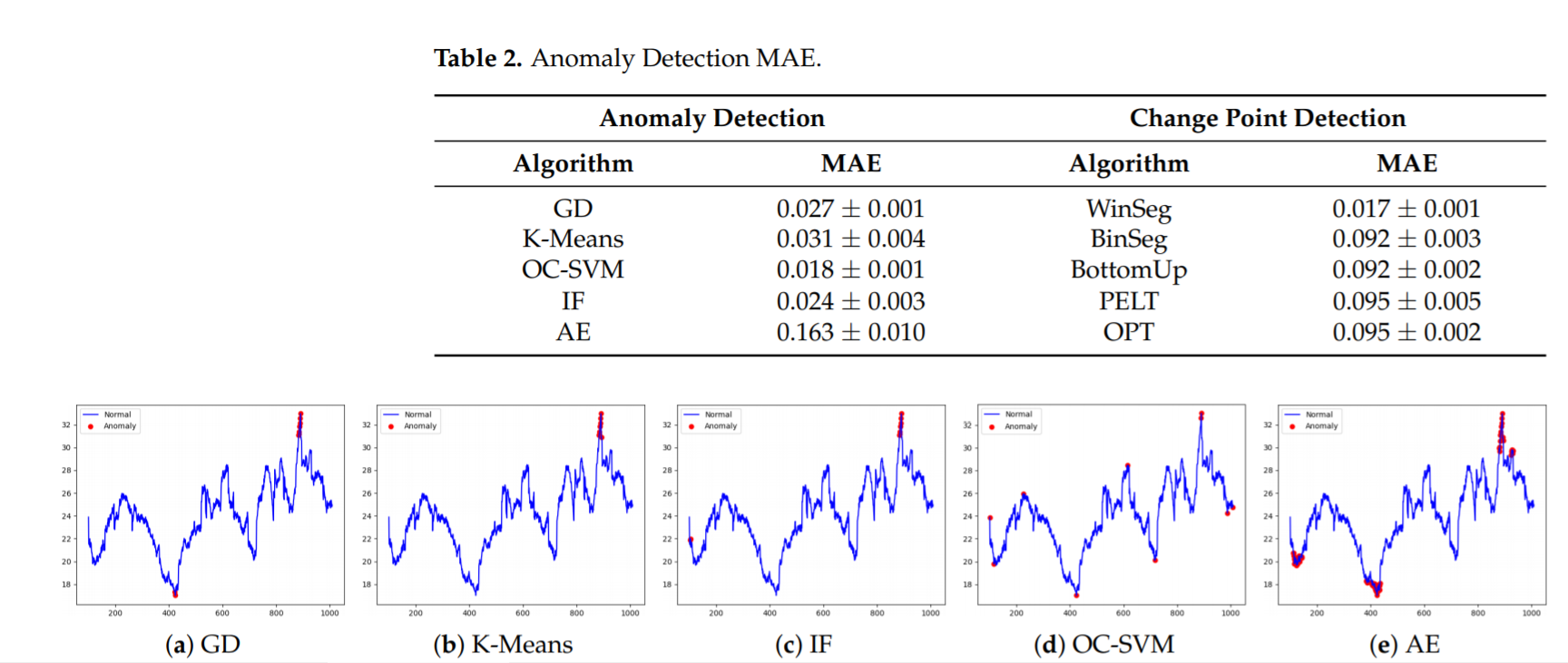
H3 When the anomaly support is equal to the change point support, regardless if both are high or low.

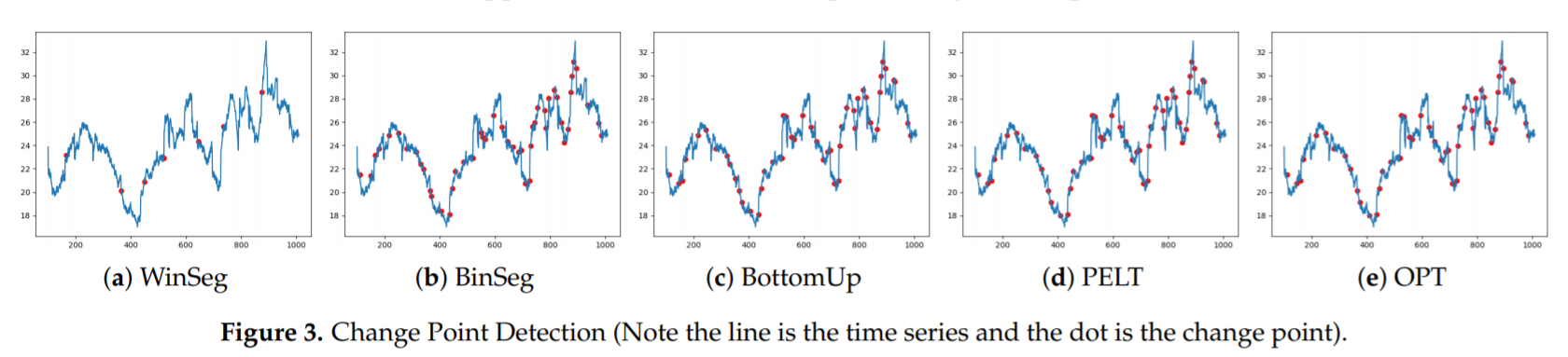
**4. 평가방식**

-**MAE** (Mean Absolute Error) Mean Absolute Error MAE = 1 /T ∑ T t=1 |yt − yˆt | to **determine the accuracy of the algorithms,** where yt is the label and yˆt is the predicted label.

Then we split the dataset into a training and a testing set using a **80–20%** train-test ration

Thus, in the training set, we kept the first 80% of the time series data points, maintaining the sequence order, and in the test set, we kept the last 20% of time series data points, again maintaining the sequence order





테이블이(가) 표시된 사진

자동 생성된 설명

**5. 기여점**

**Predictive Maintenance:**

the proposed system alerts end-users in real-time of any changes in the water distribution network in order to employ effective procedures that enable predictive [1] and proactive maintenance [2]. Predictive and proactive maintenance is a solution used to alleviate the costs related to loss of resources when water distribution networks encounter defects and the equipment breaks down. Predictive maintenance is used to make assessments regarding the well functioning of the water distribution network and consumption using real-time data to detect anomalies and change points and historical data to predict future failures. Proactive maintenance concentrates on monitoring and correcting the root causes of failures within the network using real-time anomaly detection.

Using the experimental validation, we can infer that the proposed decision rules stood. Thus, the system could automatically determine if there was a need to alert a human operator or take a decision by default. These actions led to efficient predictive and proactive maintenance of the water distribution network.

**6. Related Work**

비지도: 라벨링 X, 대부분의 데이터를 normal로 가정

- Autoencoder ensemble

- Feed Forward Neural Network

- RNN, LSTM RNN, LSTM- AE

**Environment Monitoring for Anomaly Detection System Using Smartphones**

**Sensors, MDPI**

1. 정의

2. dataset

3. 메커니즘

4. 평가방식

5. 기여점

**Improved Interpolation and Anomaly Detection for Personal PM2.5 Measurement**

**applied science, MDPI**

**1. 정의**

In this paper, we defined **anomalies as data that showed significant changes in values within a very short time**

For instance, if PM2.5 concentration, measured every 10 seconds**, showed significant drops or jumps during the 10-s period (for instance, observations that fall below Q1-1.5IQR or above Q3+1.5IQR in the box-and-whisker plot), we considered the value to be abnormal,** because the amount of change in PM2.5 concentration is assumed to stay stable or similar within a very short time.

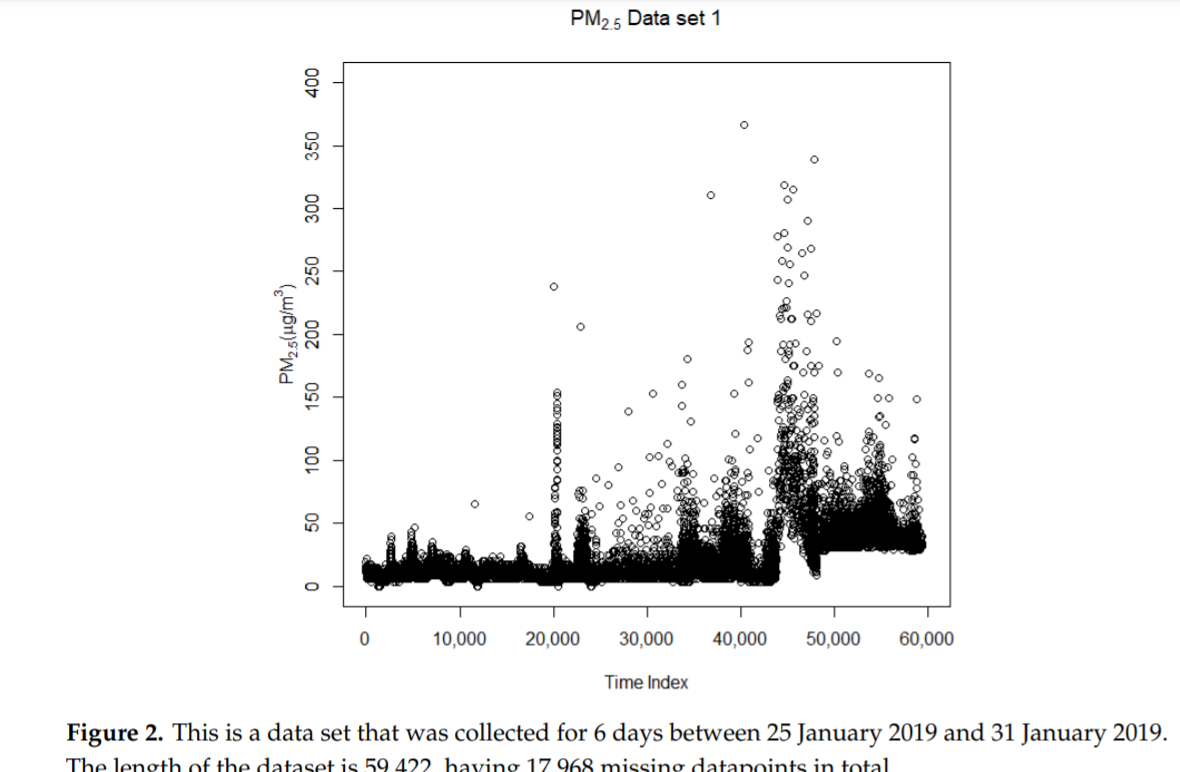
**2. dataset**

we verified the effectiveness of the proposed interpolation method. In order to verify its validity, we (1) **randomly removed some of the actual data, (2) interpolated the removed data based on our algorithm**, and then (3) compared them with each other in terms of certain performance criteria, including a comparison of applying results using other known methods

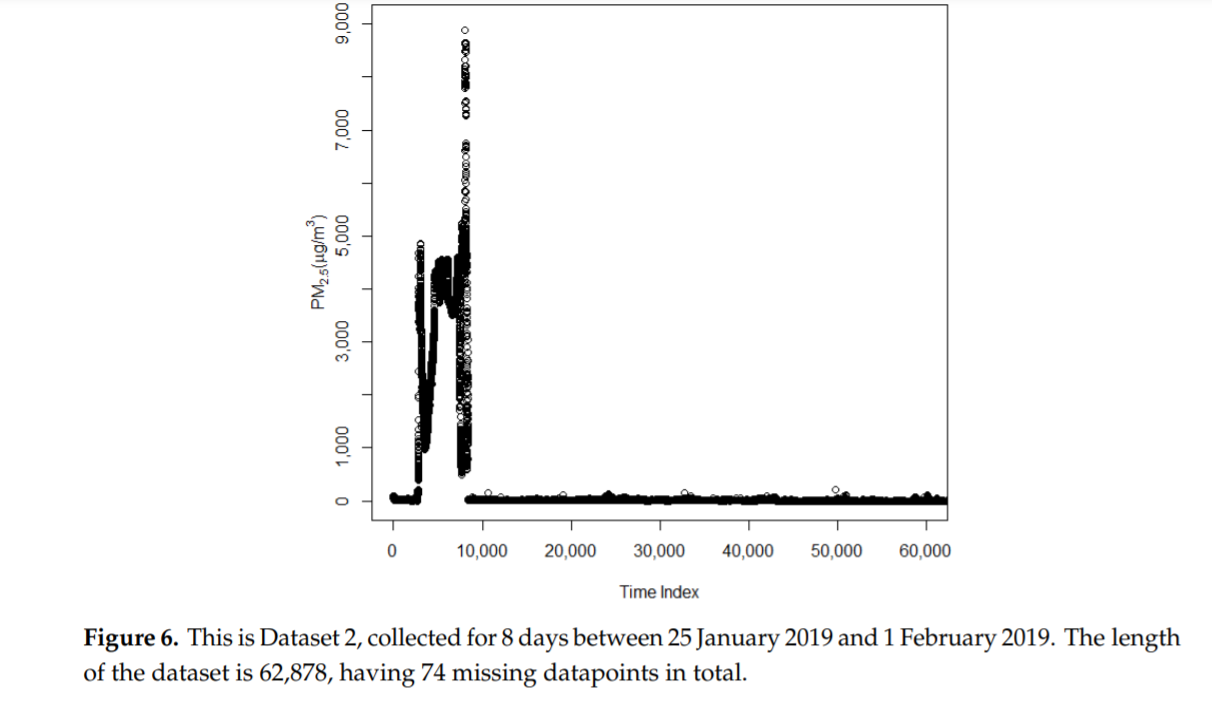
In our bootstrapping test, **40, 60, 80, and 100 datapoints were randomly deleted to create missing data**, and the interpolation results on the deleted data were evaluated compared to those of the original data in terms of RMSE (root mean squared error).

These PM2.5 data were collected at **10-s intervals** from portable personal PM2.5 monitors attached to the subjects. The data were measured between **25 January 2019 and 1 February 2019.**

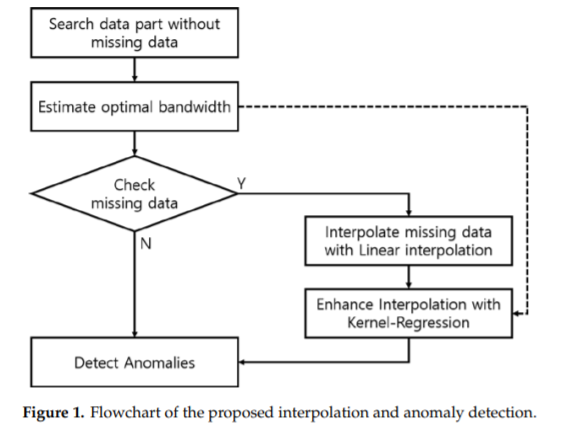
As shown in Figure 2, the level of PM2.5 was stable, implying that the subject was relatively calm with minimal abrupt activity changes. The length of the dataset was **59,422 at 10-s intervals, but it had 17,968 missing datapoints** in total. Before we went on to detect the anomalies, we first selected an optimal bandwidth for the dataset as we did previously.



This dataset was collected for 8 days between 25 January 2019 and 1 February 2019. The length of the dataset is **62,878, having 74 missing datapoints** in total. Dataset 1 used in the previous experimental test was very stable, because the distribution of PM data was mostly less than 100 µg/m3 during the data acquisition period. However, the PM data in Dataset 2 (Figure 6) reached up to 2000 to 8000 µg/m3 with 10-s intervals and showed a more dynamic change in the distribution of PM data, implying that the subject had various activities or was exposed to many different environmental conditions containing all the data-distribution patterns of rising, falling, and stable PM2.5 concentrations



**3. 메커니즘**



As the next step, for the entire dataset, we detected anomalies in the dataset, based on the method described in the method section**. When the difference of adjacent PM2.5 values is above a certain threshold (in this experiment, 200), we considered them to be anomalies.**

**4. 평가방식**

- interpolation할 때, RMSE

**5. 기여점**

**We think out findings have contributed greatly to overcoming the incompleteness of environmental data obtained from individual sensors and to providing an academic basis for more reliable data analysis. If the proposed algorithm is further improved, it will contribute a lot to advancing personalized healthcare and preventive medicine research.**

**Relevance of Drift Components and Unit-to-Unit Variability in the Predictive Maintenance of Low-Cost Electrochemical Sensor Systems in Air Quality Monitoring**

**sensors, MDPI**

**-> calibration에 대해 강조**

In reality, every low-cost sensor has a different baseline (zero), responds differently to the gases and environmental factors, and adds noise to the signal, thereby requiring individual calibration

**1. 정의**

Detecting unusual behavior, e.g., system malfunction, is a common ML task in which normal and non-normal operation states have to be distinguished—a problem known as **anomaly detection**

**2. dataset**

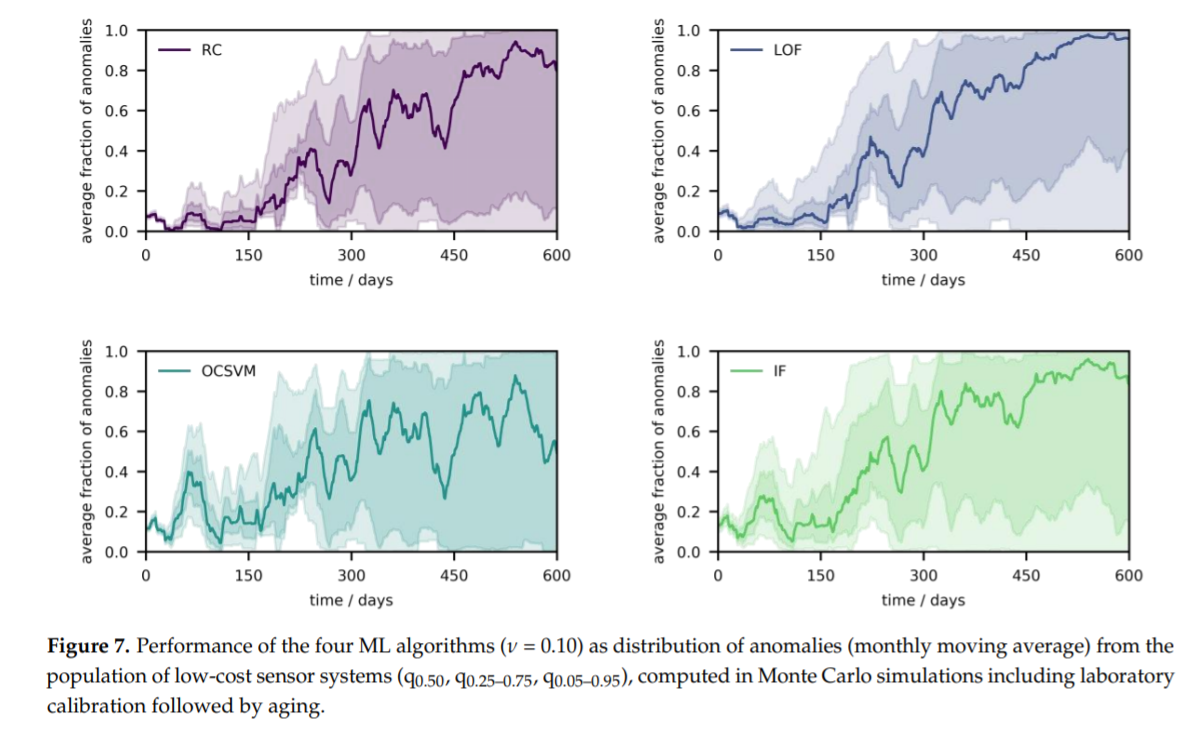
Field data from Zimmerman et al., were used for the analysis, which were collected at an urban background site [10]. The data consist, among other things, of measurements from several devices with built-in Alphasense gas sensors for **carbon monoxide (CO), carbon dioxide (CO2), nitrogen dioxide (NO2), ozone (O3/NO2combined), and sulfur dioxide (SO2), as well as temperature and relative humidity sensors, in four data points per hour over a period of six months** (August to February).

**The analysis was centered on the three pollutants CO, NO2, and O3** (CO2 and SO2 have been omitted). The properties of the corresponding sensors are listed in Table 1 [20–22]. Before analysis, the data set was preprocessed by imputing missing values with averages for each time trace, followed by scaling values to the range between zero and one.

테이블이(가) 표시된 사진

자동 생성된 설명

**3. 메커니즘**



**4. 평가방식**

**5. 기여점**

**The performed experiments have shown that devices that are calibrated on the field can yield untrustworthy measurement results over time and might require shorter recalibration times due to concept drift, which superimposes with drift due to sensor aging.**

Nonetheless, anomaly scores can disclose the reliability of the measurement results. On one hand, by combining laboratory calibration with drift compensation methods, concept drift could be removed from the software side. On the other hand, reduction in unit-to-unit variability and sensor drifts should be prioritized from the hardware side since it would enable population calibration without any further recalibration.

Moreover, the four discussed methods from unsupervised anomaly detection recognize drifts most of the time, but simulation results suggest that they might fail for a moderate portion of low-cost sensor systems, as sensor variability and joint dynamics of the drift components and atmosphere complicate the approach. In the selection of a suitable solution, interactions between all potential drift components and unit-to-unit variability have to be considered. Therefore, the obtained results can guide the development of novel algorithms and the establishment of methodologies that are adapted to the physics of low-cost electrochemical sensors and the calibration approach. In particular, network calibration techniques for electrochemical sensor systems could be explored in future studies

on both real-world data and simulations such as the ones proposed.

**A Temporal Forecasting Driven Approach Using Facebook’s Prophet Method for Anomaly Detection in Sewer Air Temperature Sensor System**

**2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)**

**1. 정의**

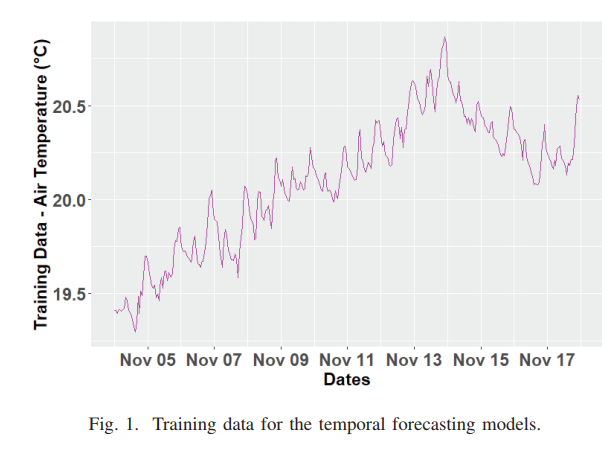
The water utilities spend millions of dollars each year to repair and rehabilitate the pipes affected by concrete corrosion [3], [4]. If the water utilities fail to address the corrosion problem, it can result in sewer infrastructure breakdown.

This predictive analytics based corrosion estimation needs longterm sensor inputs. **However, sensors can produce random anomaly or a continuous stream of anomalies in sewer environmental conditions** [20], [21]. **Hence, it is important to have a diagnostic tool to automatically detect anomalies in sensors such as sewer air temperature sensor, which provides crucial data inputs to the models predicting corrosion**

Time series or temporal forecasting models are widely used to develop anomaly detection approaches. Those approaches highly rely on the accuracy of the forecast data to statistically detect an individual anomaly or a group of anomalies.

**2. dataset**

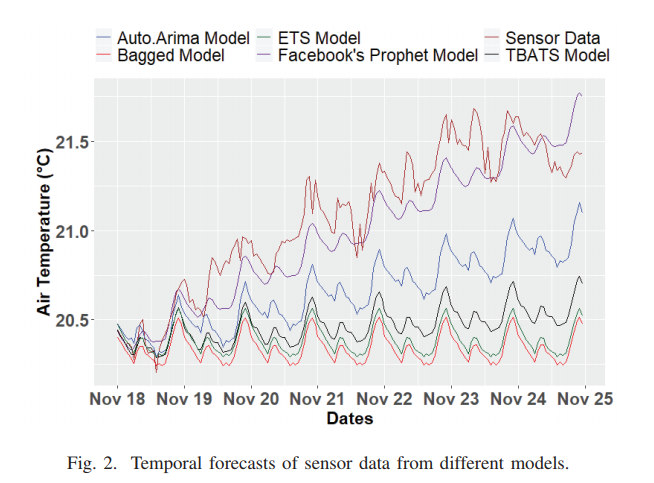
Each model was trained using the sewer air temperature sensor data from **4th November 2016 to 17th November 2016. The training data contains 336 sensor measurement values**. Figure 1 shows the training data plot. By using the two weeks training data, each model forecasted sensor measurements for one week from 18th November 2016 to 24th November 2016.



**3. 메커니즘**

**1) A. Performance Evaluation of Temporal Forecasting Models**

This section evaluates the forecasting performance of time series models by comparing the forecasts of Facebook’s **Prophet** method with the forecasts of other time series models such as **Auto.Arima model, TBATS model, ETS model, and Bagged model.**

테이블이(가) 표시된 사진

자동 생성된 설명

This shows that the temporal forecasting performance of **Facebook’s Prophet method is better than the other compared models.**

테이블이(가) 표시된 사진

자동 생성된 설명테이블이(가) 표시된 사진

자동 생성된 설명

**-> 날짜별로 MAE RMSE 분석, 첫 날에 가장 성능 좋음**

Therefore, from Table II and Table III, it can be concluded that the MAE and RMSE increase when the number of forecasting days increases.

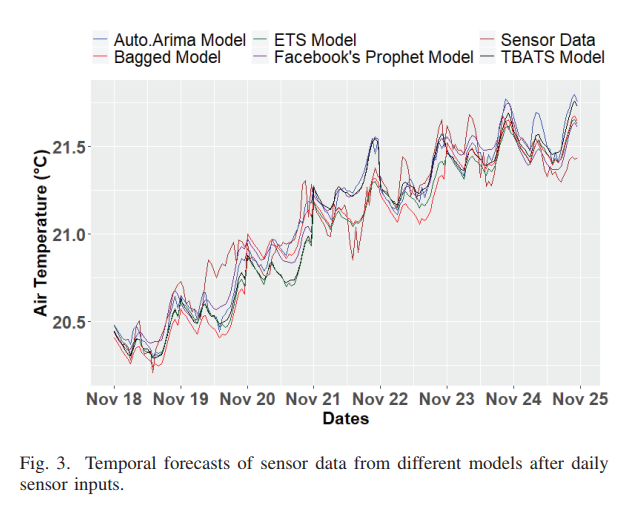
Therefore, Facebook’s **Prophet method is more suitable for forecasting short-term** (daily) sewer air temperature sensor data than forecasting longterm (weekly).

**2) B. Temporal Forecasting Performance Evaluation with Daily Feedback**

**This section evaluates the temporal forecasting performance of each model by forecasting one day.**

Figure 3 shows the forecasts for each model, where it can be observed that all the models follow the **same trend as the actual sewer air temperature sensor data.**

To analyse statistically, MAE and RMSE were computed for one week and tabulated in Table IV, where it can be noticed that the forecasts of **Facebook’s Prophet method have lowest MAE and RMSE** when compared with the other temporal forecasting models. However, there is no significant variation in MAE and RMSE between the models.



테이블이(가) 표시된 사진

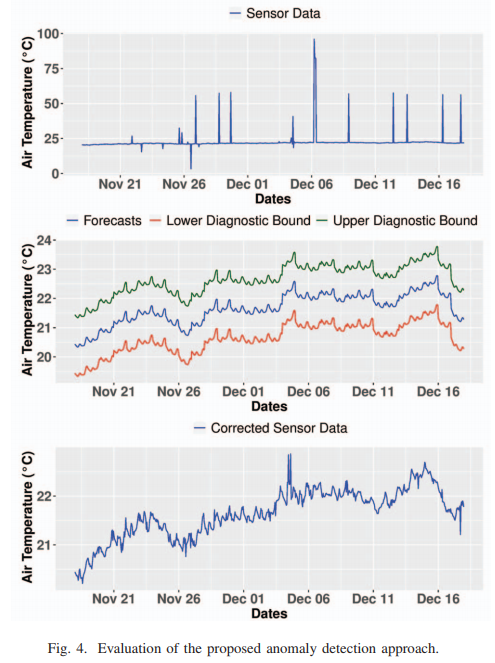
자동 생성된 설명테이블이(가) 표시된 사진

자동 생성된 설명

**3) C. Performance Evaluation of the Temporal Forecasting Driven Approach for Anomaly Detection**

This section evaluates the developed temporal forecasting driven approach using Facebook’s Prophet method for anomaly detection of sewer air temperature sensor measurements. **The anomaly detection approach was trained by using the sewer air temperature sensor data from 4th November 2016 to 17th November 2016.**

During the laboratory testing of the sewer air temperature sensor system, the sensor worked abruptly and produced a stream of anomalies. In this experimentation, we have manually injected those anomalies produced at the time of lab testing along with the anomalies produced by the sensor during the field testing inside the sewer pipe.



**There were a total of 25 anomalies**

Otherwise, if the sensor data is not within the diagnostic bounds, then the sensor measurement is treated as an anomaly. The detected anomaly is corrected with the respective forecasted data. Then, all the **24 data points are stacked into the training data set to perform forecasting for 19th November 2016.**

The proposed anomaly detection approach has **detected 23 out of 25 anomalies,** which shows a reasonable performance of the proposed approach.

**4. 평가방식**

To analyse this trend comprehensively, **MAE and RMSE** were computed by comparing the sensor data of each day with the forecast values. Table II tabulates the computed MAE for each day and Table III tabulates the computed each day RMSE values.

텍스트이(가) 표시된 사진

자동 생성된 설명

5. 기여점