Understanding Integrity of Time Series IoT Datasets through Local Outlier Detection with Steep Peak and Valley

Jungeun Yoon
ETRI
Daegu, KOREA
yje3058@etri.re.kr

Aekyeung Moon
ETRI
Daegu, KOREA
akmoon@etri.re.kr

Seung Woo Son
UMASS Lowell
Lowell, USA
seungwoo_son@uml.edu

ABSTRACT

With substantial advances in emerging and enabling technologies in IoT sensors, a vast amount of IoT-based environmental data allows preparation for adverse impacts by providing helpful information for predictive and precise services. However, data acquired by IoT sensors can be corrupted by external environmental factors, which can negatively affect the integrity of data interpretation. To address this problem, a prior study proposed outlier detection techniques using transform-based sparse profiles. However, it would lose its worth without an evaluation methodology for data integrity after probing datasets by outlier detection. In addition, it did not consider data with steep peaks or data that is dependent on other data, which is common in real-world scenarios such as soil moisture data used in this paper. Therefore, we propose a process of preprocessing defective soil moisture sensor data using local pattern-based outlier detection (LPOD) and evaluating the integrity of data after outlier detection. Our paper specifically aims to: 1) detect outliers of original soil IoT datasets to eliminate fault data possibly giving wrong decisions using local and global outlier detection (OD); 2) exploit the results of statistical evaluation to determine whether the outliers have been well eliminated; and 3) find the ground truth pattern of soil IoT datasets considering precipitation. Experiments using real-world soil moisture datasets show that the LPOD method outperforms other statistical outlier detection methods, suggesting that the preprocessed data can improve the integrity of IoT datasets.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

IoT data analytics, outliers, outlier detection, time-series data, data patterns.

ACM Reference Format:
Jungeun Yoon, Aekyeung Moon, and Seung Woo Son. 2023. Understanding Integrity of Time Series IoT Datasets through Local Outlier Detection with Steep Peak and Valley. In *2023 The 11th International Conference on Information Technology: IoT and Smart City (ICIT 2023), December 14–17, 2023, Kyoto, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3638985.3639007

ICIT 2023, December 14–17, 2023, Kyoto, Japan

ACM ISBN 979-8-4007-0904-3/23/12.
https://doi.org/10.1145/3638985.3639007

1 INTRODUCTION

The rapid development of Internet of Things (IoT) technology has led to the continuous collection of large amounts of data in various places, such as homes, offices, and even agricultural farms. Scientists have focused on data mining related to environment-related phenomena, which are known to impact agriculture substantially [3, 14, 18]. Effectively managing soil moisture data from sensors gives actionable knowledge, such as automation of irrigation, to help farmers [21]. Sensors typically acquire data and record them in a time order, thus constituting a time series. The proliferation of time series datasets generated by modern environment IoT sensors can help informative discoveries reach better and faster decisions if farming industries leverage the datasets correctly. However, it needs to have confidence that there are no silent data faults in the acquired real-time IoT datasets because data faults can adversely affect the integrity of data mining. Therefore, sensor data quality or outlier detection to improve data integrity plays a fundamental role in adopting corresponding IoT sensors [2, 12].

Our solution is to eliminate outliers and assess data fidelity using statistical methods. An efficient and effective outlier elimination empowers data fidelity by timely correcting anomaly situations [1, 13]. Outlier detection (OD) has become a data analytics field of interest for many researchers. It is now one of the main tasks of time series data analytics in wide-ranging domains [11]. To detect sensor failures or outliers (anomalies), identifying unusual instances that deviate significantly from the majority of data is underpinned. There are various outlier detection techniques, such as statistical-based, distance-based, clustering-based, and density-based, to identify and remove abnormal instances. In this case, statistical detection methods have the limitations of high computational cost. They might also suffer from the curse of dimensionality when applied to large datasets. To improve such a problem, a distance-based detection method has been proposed, which detects outliers by calculating the distance between all data objects. However, this method has the limitation of not being able to identify outliers properly when the data distribution is complex [19]. Due to the abovementioned problems, we designed an IoT sensor outlier detection elimination model to improve and reflect the data characteristics. Then, we evaluate the reliability of data eliminated outliers using soil moisture data, which has the statistical characteristic of peak and declining time series and is collected from April 29, 2023, to September 4, 2023, in two farm-land spots. The significant contributions of our approach presented in this paper are as follows:

- We propose a local pattern-based outlier detection (LPOD) algorithm that detects outliers based on local data. The soil moisture data in this study has steep peaks and valleys, and the values differ for each data interval. LPOD can reflect

these data characteristics, showing higher outlier detection performance than algorithms that detect outliers based on global datasets (z-score and transform-based).
- After detecting outliers and correcting or eliminating the data, we present a method to evaluate whether the data has been corrected accurately using statistical validation methods such as the Augmented Dickey–Fuller (ADF) test, Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, Auto Correlation Function (ACF), and Autoregressive Integrated Moving Average (ARIMA).
- To assess data reliability, we extensively evaluate our outliers elimination approach based on several outlier detection models, including Z-score, transform-based, and LPOD. Our results demonstrate that LPOD presents superior prediction accuracy measured in RMSE compared to statistical outlier detection algorithms based on z-score and transform-based OD.
- By directly integrating the environmental datasets collected from soil moisture and weather, we find a ground truth pattern for the outlier detection model.

2 PRELIMINARIES

2.1 Outliers of Soil Datasets

The soil moisture datasets acquired from real-world orchard sites can contain point and collective outliers, as depicted in Figure 1.

- **Point Outliers**: This type of outlier commonly occurs in a data point comparatively far from the whole dataset. For example, there are three-point outliers, P_1, P_2, and P_3, in Figure 1a. For illustrative purposes, point outliers in this figure show relatively high deviations from the original data points.
- **Collective Outliers**: When a subset of data points is abnormal to the entire dataset, those are called collective outliers. For example, Figure 1b contains one collective outlier period from C_1 to C_4. This type of outlier is defined as a sequence of data points making an outlier pattern [8].

Outlier detection (OD) is finding the patterns of an outlier or a fault in datasets whose behavior is not as expected [5, 9], like the example anomalous data points in Figure 1. We first explore the statistical techniques that form the fundamentals of OD. Then, we introduce signal transform-based OD and LPOD methods to detect anomalies based on time series data characteristics [15, 16].

- **Z-score**: The Z-score is a statistical measure of how many standard deviations away a given observation is from the mean. We use the Z-score to detect anomalous data points from a dataset's mean (μ) in terms of standard deviations (σ).

$$OD\_score(x_t) = (x_t - \mu)/\sigma, x = x_1, \ldots, x_n. \quad (1)$$

Given $OD\_score(x_t)$ at time t, one can detect an outlier if it is higher than a predefined threshold θ. If outlier data is detected more than θ times, we change detected outlier data into normal using interpolation. The choice of θ is critical because it determines which outliers are selected. For instance, using a specific θ can determine the range of outliers eliminated if a small θ can lead to the loss of valid data. Therefore, it needs to consider a trade-off in setting an appropriate θ to preserve normal data.

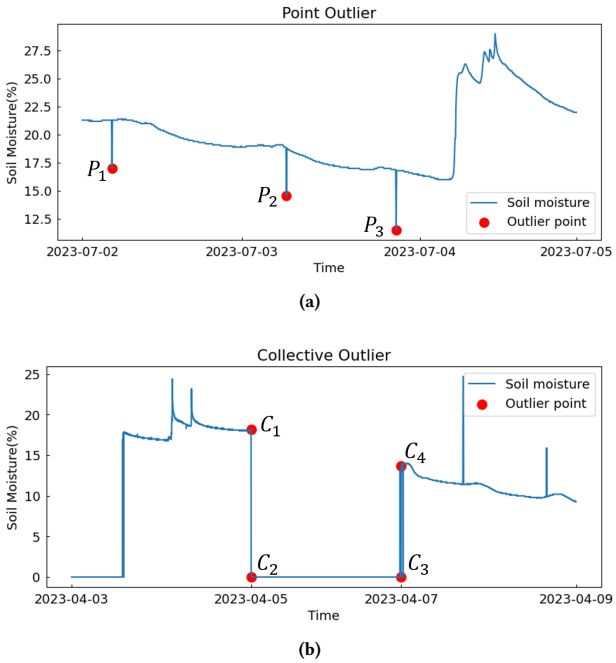

Figure 1: An illustration of two types of data outliers: (a) point and (b) collective outliers.

- **Transform-based**: The transform-based approach exploits spatial-temporal data characteristics to detect outlier patterns. This technique capitalizes on their overall patterns being spatiotemporally smooth in time-series datasets. In that case, transformation techniques can be more effective because the transformed data usually explicitly reveals the data's correlation [17, 23]. We apply the inverse transform to these selected coefficients to reconstruct data without outliers for outlier elimination.
- **Local Pattern-based OD**: The local pattern-based OD method detects outlier patterns by leveraging the characteristics of the soil moisture data used in this paper. The data has steep peaks and valleys, and the values in each data interval are different. The values of point outliers and collective outliers vary depending on the interval. Therefore, statistical techniques that consider the data locally are needed to detect outliers.

2.2 Time-Series Data Characteristics

This study aims to detect and correct sensor outliers and then statistically validate the accuracy of the data. The data collected by soil moisture sensors steep peaks and valleys, and the values differ for each interval of the data. It is essential to consider their

characteristics. In particular, the collected data may exhibit various characteristics such as trends and seasonality, which can be summarized as follows [22]:

- **Trend**: Soil moisture data collected directly in this study shows an increasing trend over time, as shown in Figure 2 declining points (V_1 to V_3). The declining points represent the time just before the next rainfall, which indicates that the soil moisture is increasing over time.
- **Seasonality**: Soil moisture data shows a pattern of increasing value at a specific point, such as the interval R in Figure 2, and then gradually decreasing. The cycle is the time from the start of the first precipitation to the start of the subsequent precipitation.

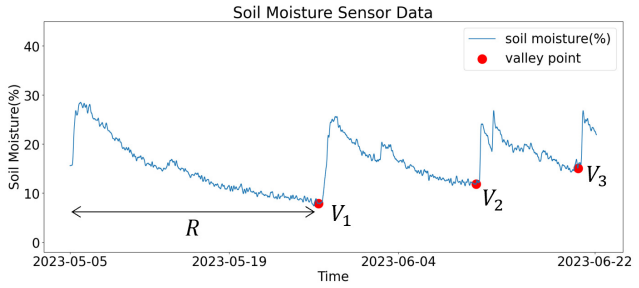

Figure 2: Characteristics of time series data from a soil moisture sensor.

Under these considerations of time series characteristics, the data must satisfy the stationarity. Stationarity implies that the data exhibit constant mean and variance over time and lack trends or seasonality. However, real-world data collected often exhibits trends and seasonality. This research involves preprocessing the data to transform it into a stationary form and then using statistical testing methods to validate this transformation. Converting the data into a stationary form eliminates trends and seasonality, making it suitable for applying statistical models. This process aims to enhance data reliability and enable more accurate analysis and modeling.

3 OUTLIER DETECTION APPROACH

Figure 3 shows an overall process for detecting outliers and assessing data validity, which consists of four steps. As shown in Figure 3, we first detect outliers in raw data using OD such as Z-score, Transform-based OD, and LPOD. After removing anomalies, we verify stationarity to fit the ARIMA model to the data. After fitting the ARIMA model, we compare the performance of each data processing method using the Root Mean Square Error(RMSE) value. This analysis outcome will suggest the most effective OD method and data validation method.

3.1 Transform-based OD

We adapt transform-based OD (proposed in [16]) in conjunction with three OD techniques in Section 2.1. In detail, the transform-based OD approach begins by transforming original datasets using DCT (Discrete Cosine Transform). Let us consider $\hat{x}$, which expresses the transformed components of the original datasets (x) after DCT to model the correlation between the transformed coefficients and energy (or information) represented among them.

Thus, each coefficient component has its energy coefficient defined as: $e(\hat{x}_{t,i})$. $EC(\hat{x}_{t,k})$ is formulated as the energy concentration (EC) contained in the number of coefficients components, denoted as k, of the entire transformed components ($\hat{x}_t$), which is calculated as:

$$EC(\hat{x}_k) = \frac{\sum_{n=1}^{k} e(\hat{x}_n)^2}{\sum_{n=1}^{N} e(\hat{x}_n)^2}, n = 1, 2, ..., N, k \leq N. \quad (2)$$

k refers to the number of dominant coefficients to represent related datasets [14]. When data is reconstructed even using only k-dominant coefficients, data fidelity is improved by deleting outliers.

3.2 Local pattern-based OD

The soil moisture data has steep peaks and valleys, and the values differ for each data interval, as shown in Figure 4. In this case, if outliers are removed by considering the entire data set, it has a limitation in that it cannot reflect the characteristics of each interval. Therefore, global algorithms such as Z-Score do not correctly remove outliers in data with such complex patterns and data distributions. We propose an LPOD algorithm that reconstructs data without outliers by investigating the difference between adjacent observations in a local of the data, not in the global data area.

LPOD is a method of reconstructing data without outliers by investigating the difference between adjacent observations in a subset of data, not the entire dataset. In this case, we use a sliding window technique to move the data interval to detect outliers in each interval [20]. The window size was set as a percentage of the total data set. At this time, we set the normal range for each interval, as shown in the bands of Figure 4, and we identify outliers as data that exceeds this range. We calculate the average and standard deviation of the data using statistical methods to construct the relationship between the data and to determine the data in the normal range. We also calculate the mean and standard deviation of the data, such as Equations 3 and 4, to construct the relationship between the data statistically and to identify normal data categories. Next, the upper and lower ranges are set with the average, standard deviation, and threshold values to set the normal range, as in Equation 5. Finally, data that exceeds the upper and lower ranges are detected as outliers, as in Equation 6. At this time, the data detected as outliers are removed and reconstructed into data without outliers through linear interpolation.

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i. \quad (3)$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2}. \quad (4)$$

$$upper, lower = \mu \pm threshold \cdot \sigma. \quad (5)$$

$$outliers = \{x_i \in X | x_i > upper \text{ or } x_i < lower\}. \quad (6)$$

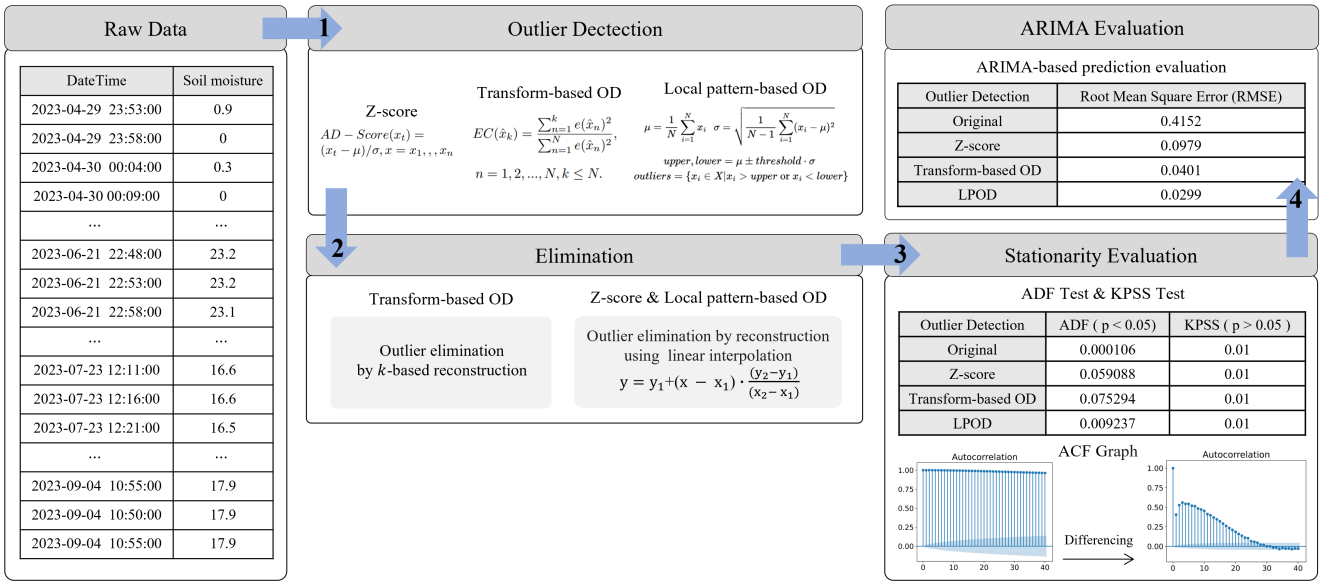

Figure 3: The process of detecting outliers and evaluating the validity of data.

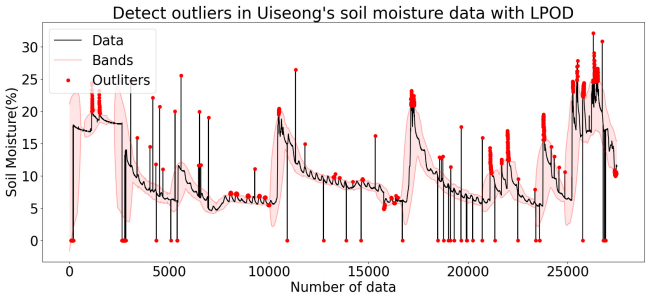

Figure 4: Outliers of soil moisture data in Uiseong detected by LPOD.

In summary, LPOD sets data intervals by moving a window over the entire data set to reflect the different characteristics of each interval. A normal range band is estimated for each interval through statistical calculations, as shown in Figure 4. Data that exceeds this range is detected as an outlier and is marked as a red point. The detected outliers are removed, and the data is reconstructed by interpolation using linear interpolation.

3.3 Stationarity Evaluation

After the OD process, we perform statistical analysis to evaluate the data's reliability. The data with eliminated outlier patterns is closer to stationarity because it has more normal patterns than the original data. Therefore, it is necessary to determine whether the data with the outliers eliminated is normal and compare it with the raw data regarding stationarity. To verify stationarity, we use ADF, KPSS, and ACF graphs.

The ADF test is a unit root test for time series. If a unit root exists, the time series is not stationary. The null hypothesis of the ADF test is that the time series has a unit root, and the alternative hypothesis is that the time series does not have a unit root. The null hypothesis can be rejected if the ADF test statistic is less than the significance level value. In other words, the stationarity of the time series can be evaluated [7]. In this study, the significance level value is set to 0.05, the commonly used level in statistics, and the probability of rejecting the null hypothesis is set to 5 percent. However, the ADF test can vary depending on the data, such as the sample size, so it is not ideal to use it alone. Therefore, it is desirable to judge stationarity by considering the ADF test results in conjunction with the KPSS test and the ACF graphs.

The KPSS test evaluates whether the variance of time series data is constant [6]. It can evaluate stationarity from a different perspective than the ADF test. The null hypothesis of KPSS is the opposite of the null hypothesis of the ADF test. Therefore, if the test statistic of the KPSS test is greater than the significance level, the time series is considered to be stationarity. An autocorrelation function (ACF) graph is a statistical graph that measures the autocorrelation of time series data. Autocorrelation is the correlation between data points at a given time interval or lag [10]. The ACF graph of data that satisfies stationarity should have autocorrelation values within the confidence interval (shown in sky blue in Figure 5) and converge to zero quickly.

The statistical validation process to determine the stationarity of data is as follows. First, if the ADF test statistic is less than 0.05 and the KPSS test statistic is more significant than 0.05, the time series is considered stationary at the beginning. Then, if the time series converges to a value close to 0 in the ACF graph, the data is finally considered stationarity.

3.4 ARIMA Evaluation

We use the ARIMA model to validate the reliability of soil moisture data for each outlier elimination method. Since the ARIMA model can evaluate the accuracy by predicting the data and comparing the predicted values with the observed values, we can assess data validity by evaluating the model accuracy of data that has stationarity. For example, in the case of original data with outliers, the specific patterns of the data may be distorted due to the outliers described in Section 2.1, making it difficult to fit the model correctly. In contrast, data with stationary characteristics of the time series is expected to show a more apparent pattern than the original data and fit the model better. As a result, the validity of the data can be evaluated by comparing the accuracy of the original data and the data with stationarity.

The data need to exhibit stationarity to fit data to the ARIMA model. However, data measured with soil moisture sensors represent non-stationary time series data. Therefore, before fitting the ARIMA model, it is crucial to determine whether the data is stationary using the stationarity evaluation Indicators in Section 3.3. Root Mean Square Error (RMSE) is used as the evaluation metric, and its formula is given in Equation 7 [4]. RMSE is a standard metric used to assess the difference between This metric is employed to evaluate the performance of ARIMA models, with lower values indicating more accurate predictions by the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})}. \quad (7)$$

4 RESULTS AND DISCUSSION

4.1 Datasets

To evaluate the proposed process to detect outliers in soil moisture sensors, we collect a real dataset from IoT stations installed in two farm-land spots in South Korea: Andong and Uiseong. The data used in this study are collected from soil moisture sensors and weather sensors. The weather sensor collects data on temperature, wind direction and speed, ground temperature, relative humidity, solar radiation, sunshine hours, and precipitation. The soil moisture data are a total of 35,946 data points collected at 5-minute intervals for 155 days from April 29, 2023, to September 4, 2023. The weather sensor data are 1,406 data points collected at 1-hour intervals for 84 days from May 1, 2023, to July 3, 2023. Each sensor is continuously monitored and collected at each measurement interval.

4.2 Stationarity Evaluation

Three statistical tests were performed to determine the stationarity of data collected from IoT sensors. The data used for the tests included the original data without preprocessing and the data with outliers eliminated using three OD techniques described in Section 2.1. Table 1 shows the results of the ADF and KPSS tests. We can make several observations from these results. In the case of ADF, if the statistical value is less than the significance level of 0.05, it is considered stationary. In the case of KPSS, if it is more significant than 0.05, it is considered stationary. However, all data showed non-stationarity in the KPSS test, so it was considered non-stationary.

Table 1: ADF and KPSS test results according to preprocessing method.

Outlier Detection	ADF (p < 0.05)	KPSS (p > 0.05)
Original	0.000106	0.01
Z-score	0.059088	0.01
Transform-based OD	0.075294	0.01
LPOD	0.009237	0.01

Most time series datasets measured and observed in reality are non-stationary, as shown above. Therefore, The differencing, which subtracts the previous value in a time series from the current value, was applied to reevaluate the stationarity of the data. This step keeps the mean of the series constant over time and reduces its time dependency, which makes it achieve stationarity. Table 2 shows the differencing results for all data sets. As we can see, the ADF and KPSS tests meet the significance level condition, thus being considered stationary.

Table 2: ADF and KPSS test results after differencing.

Outlier Detection	ADF (p < 0.05)	KPSS (p > 0.05)
Original	0	0.1
Z-score	0	0.1
Transform-based OD	0	0.1
LPOD	0	0.1

Finally, we visualize the stationarity of the data through the ACF graphs. In Figure 5, the x-axis of the graph, lag, represents previous data points from the current data point. For example, if the data is measured at 1-hour intervals and the lag is set to 2, the lag is 2 hours before the current data point. Therefore, the y-axis of the graph, ACF, compares the current data point with the data point 2 hours before. If the lag is 0, it is always 1 because it is the autocorrelation of the current data point with itself. Therefore, it is excluded, and the graph is analyzed.

In Figure 5, the cases of (5a) show significant deviations from the confidence interval around lag 10 and 20. Therefore, they are evaluated to be non-stationary data. The data detected as outliers using three preprocessing methods are converging stably based on a specific lag point (z-score(5b):8, transform-based OD(5c):27 LPOD(5d):6), as shown in Figure 5. Therefore, the data can be evaluated as stationary when using the three techniques of Z-Score, Transform-based OD, and LPOD. In Section 4.2, we conducted the same stationarity experiments using datasets from the Andong and Uiseong sites. The experimental results were similar, so we only included the experimental results using the datasets collected at Andong in this paper.

4.3 ARIMA Evaluation

After confirming the stationarity of the data, statistical analysis-based models can be used to assess model accuracy by comparing predicted values to observed values. The training and testing data are divided into an 8:2 ratio to employ the ARIMA model, with

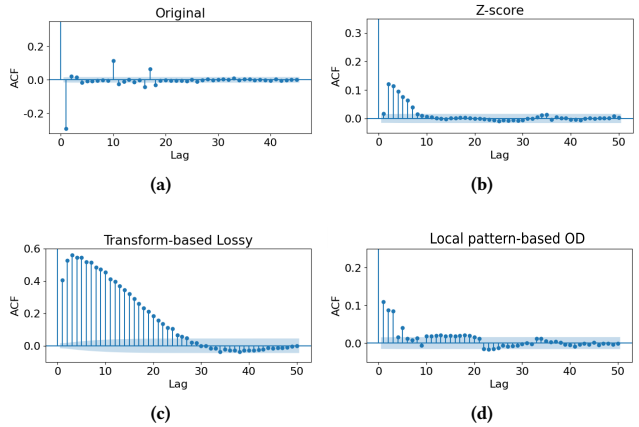

Figure 5: The ACF plot of the data processed by different outlier detection techniques: (a) original, (b) z-score, (c) transform based OD, and (d) LPOD.

RMSE used to gauge model performance. Table 3 and Figure 6 present the training outcomes with the ARIMA model. Figure 6 shows the overall patterns of the original data and the data after outlier data elimination and whether the model accurately predicts test values. We present the experimental results of the OD with the highest RMSE performance only in this paper. Verifying if the test and predicted curves match and reviewing the RMSE values in Table 3 is essential to assess this visually. In this case, decreasing the RMSE value signifies that the model makes precise predictions. Accurate predictions imply that the data exhibits a discernible pattern, affirming the validity of the data by minimizing the influence of trends and seasonality from a statistical modeling perspective.

Table 3: The ARIMA model's forecast results, with RMSE as the performance indicator.

Outlier Detection	Andong	Uiseong
	Root Mean Square Error (RMSE)	
Original	0.4152	0.7131
Z-score	0.0979	0.5593
Transform-based OD	0.0401	0.0645
LPOD	0.0299	0.0429

In both Andong and Uiseong, as shown in Table 3, the data with outliers removed using LPOD shows the best performance. The original data (6a, 6c) shows point and collective outliers throughout the data, so the test and predicted curves do not match. In LPOD (6b, 6d), both outlier phenomena are removed; we can confirm this since the two curves match. However, the part removed by the collective outlier in 6d looks unnatural because it is a simple linear interpolation algorithm, so there is room for improvement.

4.4 Finding Ground Truth Pattern

The current data shows outlier patterns, as shown in Figures 6a and 6c. Outlier patterns correspond to data instances with two different outliers, as described in Section 2.1. Removing segments that show these outliers makes it possible to identify the regular patterns or cycles in the data. This can be used as an important feature in classification or regression problems. In Figure 7, the ground trust pattern has very steep peaks and valleys, and the values differ for each interval, which can vary depending on external factors such as environmental impacts. We applied various outlier techniques to identify this pattern, and the data removed by LPOD showed the highest performance.

The data patterns with outliers removed can be analyzed by considering the associated data together. Precipitation, in particular, plays a central role in supplying water to the soil. Therefore, we can utilize the relationship between soil moisture and precipitation data. Figure 7 shows the results of mapping precipitation data to soil moisture data after removing outliers. The blue dots in Figure 7 represent cases of precipitation occurrence. When precipitation occurs, steep peaks and valleys appear in soil moisture data. In addition, the values of the peaks and valleys in soil moisture data vary depending on the frequency and amount of precipitation. Therefore, a positive relationship was observed between precipitation and soil moisture data, which can be seen as a ground trust pattern.

5 CONCLUSION

With the adoption of IoT sensors, the data collected can be utilized for various purposes, such as data mining, classification, and prediction. However, the collected data may have data defects due to the influence of external environments. Therefore, this study evaluated various outlier detection techniques (Z-score, Transform-based OD, LPOD) based on real-world data to detect and eliminate outliers. The soil moisture data has various characteristics, such as steep peaks and different peak and valley values in each interval. Therefore, we propose an LPOD algorithm that takes these characteristics into account. To assess the validity of the data with outliers removed, we evaluated the stationary of the data and analyzed the accuracy of the model by fitting it to an ARIMA model.

The results showed that LPOD was the most effective method for outlier detection and improved results over the original data set in all measures, including ADF, KPSS, ACF, and RMSE. In addition, we could find regular patterns in the data with outliers removed. We found that precipitation data are positively correlated with soil moisture data, allowing us to find the ground truth pattern.

ACKNOWLEDGMENTS

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) [23RD1100] and SME ICT convergence technologies project in Andong-si [23AD1100]. This material is, in part, based upon work supported by the National Science Foundation under Grant No. 2312982 and the NVIDIA hardware grant.

REFERENCES

[1] Eduardo Berrocal, Leonardo Bautista-Gomez, Sheng Di, Zhiling Lan, and Franck Cappello. 2015. Lightweight Silent Data Corruption Detection Based on Runtime Data Analysis for HPC Applications. In *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing(HPDC)*. https://doi.org/10.1145/2749246.2749253

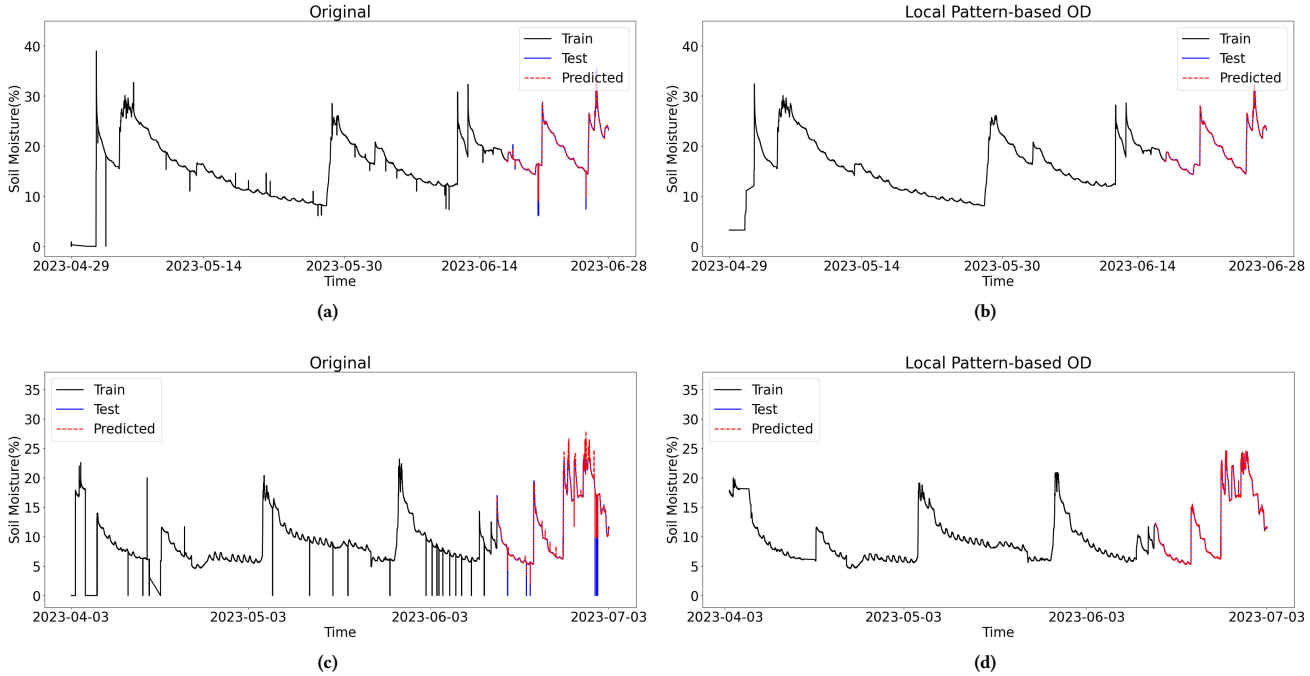

Figure 6: ARIMA-based prediction evaluation graph for (a) Andong Original, (b) Andong LPOD, (c) Uiseong Original, and (d) Uiseong LPOD.

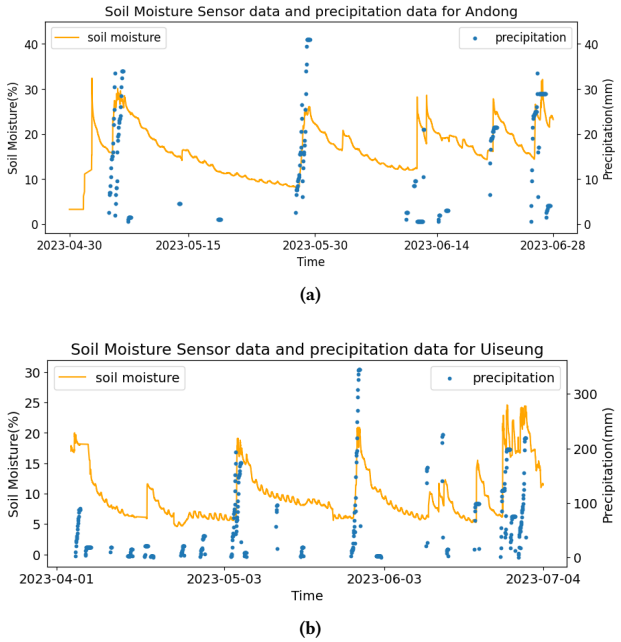

Figure 7: Ground truth pattern from soil moisture sensor: (a) Andong, and (b) Uiseong.

[2] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano. 2021. A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Comput. Surv.* 54, 3, Article 56 (apr 2021), 33 pages. https://doi.org/10.1145/3444690

[3] Y. Cai, W. Zheng, X. Zhang, L. Zhangzhong, and X. Xue. 2019. Research on soil moisture prediction model based on deep learning. In *PLOS ONE*, Vol. 14. https://doi.org/10.1371/journal.pone.0214508

[4] T. Chai and R. R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development (GMD)* (2014), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.* 41, 3, Article 15 (July 2009), 58 pages. https://doi.org/10.1145/1541880.1541882

[6] Peter Schmidt Denis Kwiatkowski, Peter C.B. Phillips and Yongcheol Shin. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* (1992), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

[7] David A. Dickey and Wayne A. Fuller. 1984. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *J. Amer. Statist. Assoc.* (1984), 427–431. https://doi.org/10.2307/2286348

[8] Alexander T M Fisch, Idris A Eckley, and Paul Fearnhead. 2019. Subset Multivariate Collective And Point Anomaly Detection . *ArXiv e-prints* (2019).

[9] Johan Florbäck. 2015. *Anomaly Detection in Logged Sensor Data.* Master's thesis. Chalmers University of Technology. Master's thesis in Complex Adaptive Systems.

[10] C. W. J. Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* (1969), 424–438. https://doi.org/10.2307/1912791

[11] Minqi Jiang, Songqiao Han, and Hailiang Huang. 2023. Anomaly Detection with Score Distribution Discrimination. In *KDD '23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

[12] Sumukh Marathe, Akshay Nambi, Manohar Swaminathan, and Ronak Sutaria. 2021. CurrentSense: A novel approach for fault and drift detection in environmental IoT sensors. In *IoTDI.*

[13] Luis Martí, Nayat Sanchez-Pi, José Manuel Molina, and Ana Cristina Bicharra Garcia. 2015. Anomaly Detection Based on Sensor Data in Petroleum Industry Applications. *Sensors* (2015).

[14] Aekyeung Moon, Jaeyoung Kim, Jialing Zhang, and Seung Woo Son. 2018. Evaluating Fidelity of Lossy Compression on Spatiotemporal Data from an IoT Enabled Smart Farm. *Computers and Electronics in Agriculture* 154 (Nov. 2018), 304–313.

[15] Aekyeung Moon, Minjun Kim, Jiaxi Chen, and Seung Woo Son. 2023. Anomaly Detection in Scientific Datasets using Sparse Representation. In *AI4Sys '23: Proceedings of the First Workshop on AI for Systems*.

[16] Aekyeung Moon, Xiaoyan Zhuo, Jialing Zhang, Seung Woo Son, and Yun Jeong Song. 2020. Anomaly Detection in Edge Nodes using Sparsity Profile. In *IEEE Big Data*.

[17] Aekyeung Moon, Xiaoyan Zhuo, Jialing Zhang, Seung Woo Son, and Yun Jeong Song. 2020. Anomaly Detection in Edge Nodes using Sparsity Profile. In *Proceedings of IEEE Big Data*. 1236–1245.

[18] José R. Rozante, Enver Ramirez Gutierrez, Pedro Leite da Silva Dias, Alex de Almeida Fernandes, Debora Souza Alvim, and Vinicius Matoso Silva. 2019. Development of an index for frost prediction: Technique and validation. *Meteorological Applications* (2019).

[19] Abir Smiti. 2020. A critical overview of outlier detection methods. *Computer Science Review* (2020). https://doi.org/10.1016/j.cosrev.2020.100306.

[20] Majid Vafaeipour, Omid Rahbari, Marc A. Rosen, Farivar Fazelpour, and Pooyandeh Ansarirad. 2014. Application of sliding window technique for prediction of wind velocity time series. *International Journal of Energy and Environmental Engineering* (2014), 1–7. https://doi.org/10.1007/s40095-014-0105-5

[21] Juan Vera, Wenceslao Conejero, Ana B. Mira-García, María R. Conesa, and M. Carmen Ruiz-Sánchez. 2021. Towards irrigation automation based on dielectric soil sensors. In *The Journal of Horticultural Science and Biotechnology*, Vol. 96. https://doi.org/10.1080/14620316.2021.1906761

[22] Kate Smith-Miles Xiaozhe Wang and Rob Hyndman. 2009. Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. *Neurocomputing* (2009), 2581–2594. https://doi.org/10.1016/j.neucom.2008.10.017

[23] Jialing Zhang, Xiaoyan Zhuo, Aekyeung Moon, Hang Liu, and Seung Woo Son. 2019. Efficient Encoding and Reconstruction of HPC Datasets for Checkpoint/Restart. In *Symposium on Mass Storage Systems and Technologies (MSST)*. https://doi.org/10.1109/MSST.2019.00-14