Low-Cost Sensor Outlier Detection Framework For on-Line Monitoring of Particle Pollutants in Multiple Scenarios

Yinyue Xu (✉ 2019214086@tju.edu.cn)

Tianjin University <https://orcid.org/0000-0001-7661-0145>

Zhengwei Long

Tianjin University

Wuxuan Pan

Tianjin University

Yukun Wang

Tianjin University

Research Article

Keywords: Low-cost sensor, Multiple scenarios, Pollutant concentration, Outlier detection, Sensor performance, Time-series

Posted Date: February 25th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-218927/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Low-cost sensor outlier detection framework for on-line monitoring of particle pollutants in multiple scenarios

Yinyue Xu, Zhengwei Long, Wuxuan Pan and Yukun Wang*

Tianjin Key Laboratory of Indoor Air Environmental Quality Control, School of Environmental Science and Engineering, Tianjin University, Tianjin 300072, China

**Corresponding author: Zhengwei Long, School of Environmental Science and Engineering, Tianjin University, Tianjin, China. Email: longzw@tju.edu.cn.*

Abstract

Monitoring the concentration of particle pollutants is very important for industrial production control and workers' health protection. Low-cost sensors are widely used to reduce deployment costs. The outliers in the observed data of pollutant concentration can be eliminated by outlier detection algorithms. However, it is difficult to meet the actual needs of changing working conditions or scene migration in factories by building a single algorithm for specific scenarios. It is a feasible scheme to identify the changing characteristics of data and adaptively adjust the outlier detection algorithm. From the point of view of data characteristics, we creatively match typical data types with high performance algorithms. The framework proposed in this paper provides a general process including five basic tasks, and uses a modular structure to complete the outlier detection target. The actual pollutant data of the workshops are used to evaluate the performance of our framework. At last, we compare eight different strategies under this framework, and analyze the contribution of each step to outlier detection from the perspective of algorithm principle. The results show that low-cost sensors following the framework can meet the outlier detection requirements in the field of pollutant monitoring, thus greatly reducing the cost of algorithm selection and data adaptation.

Keywords

Low-cost sensor · Multiple scenarios · Pollutant concentration · Outlier detection · Sensor performance · Time-series

1. Introduction

Industrial development produces a large number of pollutants that can affect human and ecological communities. Epidemiology and animal studies have shown that long-term exposure to industrial pollutants may lead to laryngeal cancer, pneumonia and other respiratory diseases, which seriously endanger workers' health (Massin et al. 1996). In addition, the emission level of pollutants is different in various working conditions, which poses a potential threat to the health of workers exposed to pollutants

(Heyder et al. 1986). Due to their effect on our health, the follow-up of information on pollution is becoming more and more important (Sanchez-Lasheras et al. 2020). In principle, the limited measured concentrations can be used to estimate the likelihood exposure distribution (Hsu et al. 2012). The sensor monitoring system can provide us with automatic measurement of pollutant concentration. However, large-scale deployment and limited budget usually restrict the choice of sensor hardware, which usually reduces the accuracy and reliability (Barcelo-Ordinas et al. 2019; Pineiro Di Blasi et al. 2015). In the process of pollutant concentration monitoring in the industrial field, outliers caused by low-cost sensors (Ottosen and Kumar 2019; Zhang et al. 2020), which are widely used to reduce deployment costs, must be identified and eliminated to ensure data quality (Malings et al. 2019).

In view of the complexity and variability of the factory environment, the research on the low-cost sensor outlier detection algorithms is far from enough. Existing mature outlier detection algorithms were typically built for specific scenarios, and their optimization objectives and detection results depend on the characteristics of the monitoring data. However, in a more general industrial production process, operating conditions are constantly changing, and on-site monitoring concentration data will show obviously different characteristics. No single-outlier detection algorithm can achieve optimal results for all data types (Aggarwal and Sathe 2017; Chen et al. 2018). A typical example is that the outlier detection algorithm suitable for stable working conditions is difficult to adapt to the new features of the data after working conditions change, such as frequent oscillations and periodic fluctuations (Dong and Jia 2020). When the monitoring scene changes, the situation gets worse. Thus, a low-cost sensor matched with a particular single-outlier algorithm may effectively detect outliers in one scenario, but be entirely ineffective when the data characteristics change, thus the single-outlier algorithm cannot be in common use in different scenarios (Calikus et al. 2020; Yang et al. 2018). Besides, the selection of a matching algorithm and the migration of deployment scenarios require large amounts of expert advice and extensive algorithm testing (Hui et al. 2018). Therefore, detecting outliers in such continuously changing temporal data has received increasing attention from both the industry and the scientific community.

The above fact motivates the need for developing a combination of algorithms instead of seeking for the “one” that is suitable all the time. Wu et al. (2018) have achieved the automatic detection of different outlier types, but their study lacks experimentation with data sets for different scenarios. Calikus et al. (2020) have reviewed the basic tasks of outlier detection, and analyzed the algorithm combination strategies suitable for different outlier types from the perspective of algorithm principles. Their research inspires us to match high-performance outlier detection algorithm with various typical data types according to the data characteristics, and integrate the algorithms into a unified framework. Table 1 summarizes some previous studies on outlier detection algorithms for three typical data types. Data types 1-3 are usually processed by density-based, (Breunig et al. 2000; Sharma et al. 2011) statistics-based (Garoudja et al. 2017; Madakyaru et al. 2017) and prediction-based (Wang and Ahn 2020) methods, respectively. A sliding window can dynamically learn the stage characteristics of time series and update the model parameters (Wang et al. 2010). An algorithm based on machine learning evaluates abnormal behavior through regression and prediction, but the collection of labeled data sets is usually difficult in practical application (Campos et al. 2016; Chen-Chia Chen 2018). In this paper, a unified outlier detection framework includes the above-mentioned algorithms which are suitable for three typical data types. Low-cost sensors built with this framework can be migrated and commonly used in multiple scenarios.

Table 1 Studies on outlier detection for three typical data types are compared according to characteristics,

79 source, data description and algorithm.

Data type	Characteristic	Detection object	Algorithm
Datatype 1	Due to the interference of random factors, the time series has an operational trend of steady-abrupt alternation. The extreme range of the data set is large, but most points fluctuate near the mean value, and a few pulse peak points will not cause a large standard deviation.	Particulate pollutants (Ayadi et al. 2015)	Density-based LOF
		environmental pollutant (Breunig et al. 2000)	Density-based LOF
		Sudden abnormal events (Guan et al. 2015)	Density-based SLOF
		Ground observation data (Durre et al. 2010)	Statistic-based ZScore
Datatype 2	Time series fluctuates frequently and randomly within the whole fluctuation range, so it is difficult to predict its fluctuation law by establishing a model. The data set has a large standard deviation, but it follows a certain statistical distribution in the sliding window.	environmental pollutant (Wu et al. 2018)	Statistic-based ZScore
		Workshop oil mist (Long and Wang 2020)	Statistic-based Pauta
		environmental pollutant (Ottosen and Kumar 2019)	Statistic-based knn
Datatype 3	The trend of the time series is gentle, and it often has a periodic nature, which can be predicted by the established model. The data set fluctuates slowly up and down from the mean value, with large variance but low oscillation frequency, and only slight oscillation is noticeable on close examination.	Meteorological parameter monitoring (Arumugam and Saranya 2018)	Context-based ARIMA
		environmental pollutant (Aljoumani et al. 2018)	Context-based ARIMA
		Soil water content (Aljoumani et al. 2012)	Context-based ARIMA

80

81 In addition, some scholars have proposed a residual method based on the trend fitting of the original
82 data before outlier detection, and the residual of the observed value and the fitted value is taken as the

object of the outlier identification algorithm (Wu et al. 2018). Observations with large residuals are more prone to be marked as outlier candidates. The introduction of residuals can reduce false positives and missed detections in the process of outlier detection. Trend fitting methods are based on different algorithm, including signal decomposition (Li et al. 2020; Rhif et al. 2020), data prediction (Khalid et al. 2020; Valenzuela et al. 2008) and machine learning (Qiu et al. 2020; Yang et al. 2019). It is worth noting that in the process of outlier detection for some data types, the algorithm can be directly applied to the observed data to obtain better results. Therefore, this step in the framework is not mandatory, and depends on the comparative analysis of the specific data characteristics. An additional challenge is the automatic identification of the above three data types in practical application. This step is the key to matching high-performance outlier detection algorithms based on data characteristics. No scholar has studied the outlier detection algorithms from this point of view, in order to expand the application scenarios of low-cost sensors. This paper will provide some new ideas in this regard.

Starting from the actual needs of on-site monitoring in the industrial field, this paper reviews the literature of outlier detection algorithms for three typical data types, and develops a research on outlier detection framework applicable to low-cost sensors in multiple scenarios. Our framework provides a common process for three types of data, which can automatically identify data types and use a modular structure to complete anomaly detection tasks. The five relatively independent steps in the framework provide a higher degree of freedom for adjusting the algorithm combination strategy to adapt to different data characteristics. We have implemented 8 different algorithm combination strategies, and carried out extensive comparison and analysis on three types of real data sets to evaluate the applicability of the algorithm from the perspective of data characteristics.

The framework and algorithm are described in Section 2. Next, Section 3 introduces the validation of the model in real datasets, and the applicability of the selected algorithms is discussed. Finally, Section 4 presents the conclusions of this study and their significance.

2. Multi-scenario outlier detection framework

In this section, we introduce a framework of outlier detection which can be applied to three typical data types. It will extract the basic steps from the existing single outlier detection algorithm and combine it into a unified process as a general process that can be applied to multi-scene outlier detection including typical data types.

2.1 Overview of the framework

The framework of outlier detection proposed in this paper can be applied to the three typical data types in various scenarios, including five main steps, namely: feature extraction, type recognition, trend fitting, outlier recognition and gap filling. The first step, feature extraction, is to extract effective features from the original data and weaken the interference brought by outliers, thus becoming an important basis for data classification. The second step, type recognition, classifies the original sequence according to the feature difference, and prepares for matching the high-performance outlier detection algorithms. The third step, trend fitting, uses residual error to characterize the difference between the observed value and the fitted trend, which can adapt to the alternating or periodic changes of original data. This step is not compulsory. The fourth step, outlier recognition, is to measure each observation value and quantifies the

"inconsistency" of the observation point. Fifthly, gap filling, which replaces the points identified as abnormal with values suitable for the context, so as to improve the data quality.

The framework is designed based on the concept of "separation of concerns", in which the algorithm components of each step are independent and serve different purposes. In the process of outlier detection of three typical data types, according to the characteristics of differentiated data, the most suitable processing strategy is selected for each step, and the combination of high-performance algorithms under the framework of outlier detection is realized. When the framework is used in three data types in different scenarios, it only needs to reorganize the corresponding algorithm components according to the data characteristics, without changing other parts of the framework, such as the example in Section 3. The relatively independent settings of each step in the framework also provide a basis for the experiment and evaluation of more matching methods, so that the adaptability of each step's algorithm selection to the data features can be introduced from the perspective of algorithm principle and unique feature analysis, and then it is verified that the algorithm combination strategy of the three typical data types in this paper are optimal, as shown in Section 3.3.

2.2 Introduction of the algorithm

This section gives a high-level overview of the algorithm components in each of the core steps in the framework, and introduces the basic principles of the algorithm and the interface between the steps.

Step1 feature extraction

Generally speaking, feature extraction of data includes automatically converting observed data into representations or features, which can be effectively used in outlier detection tasks. Extraction of data features can capture important clues about context association of time series and key features related to outlier detection objects. The main purpose of this step is to extract the characteristic sequence which can reflect the trend of the original data and the quantitative index which can measure the periodicity and volatility of the data, so as to help distinguish the data types.

Outliers doped in raw data sets will have an impact on the overall characteristics. Furthermore, there are obvious differences in sequence fluctuation intervals in different usage scenarios (Quispe-Coica and Perez-Foguet 2020). These interference factors must be eliminated by pretreatment (Bin et al. 2012). In this paper, the discrete wavelet transform (DWT) is used to decompose the original sequence in the frequency domain, which can reflect the local characteristics of data (Zhang et al. 2017). The low-frequency component is regarded as the operational trend of the original data. Next, the operational trend is dimensionless processed by normalization, and then the characteristic sequence is obtained. The two indicators that quantitatively describe the data dispersion and volatility are standard deviation and oscillation coefficient. The calculation methods of them are given below. Pseudo-code 1 shows the logical expression of the algorithm in Step 1.

Definition1 Generate characteristic sequence

Let $X^N = \{x_1 = (c_1, t_1), x_2 = (c_2, t_2), \dots, x_N = (c_N, t_N)\}$ be the time series collected by the low-cost sensor, where N is the number of observation points, $x_i = (c_i, t_i)$ represents the observation value c_i corresponding to t_i . DWT is used to decompose the original signal, and obtain $X^N = D_1 + D_2 + \dots + D_n + A_n$. Here $D_1, D_2, \dots, D_n$ are the detailed components. Meanwhile $A_n = \{a_{n1}, a_{n2}, \dots, a_{nN}\}$ is the approximate component of the n th layer, which is the main part of the original sequence. The optimal decomposition level is selected by using singular spectrum analysis (SSA) (Tie

and Jie 2006). The specific process is introduced in Pseudo-code 1. The operational trend A_n is normalized by the mapminmax function (Li et al. 2020), and is used as a characteristic sequence for data-type identification:

$$a'_{ni} = \frac{a_{ni} - \min(a_n)}{\max(a_n) - \min(a_n)} \quad (1)$$

where a'_{ni} is the i th component of the characteristic sequence after n -level wavelet decomposition and normalization.

Indicator1 Standard deviation

In this paper, the standard deviation (std) of characteristic sequence is used to describe the discrete degree of time series, and the calculation formula is:

$$std = \sqrt{\frac{\sum_{i=1}^N (a'_{ni} - \bar{a}'_{ni})^2}{N}} \quad (2)$$

Indicator2 Oscillation coefficient

In this paper, the oscillation coefficient($oscl$) of characteristic sequence is used to describe the volatility of time series, which is defined as the number of times that the characteristic sequence crosses the mean value at a frequency of 1000.

The data feature extraction method introduced above can weaken the follow-up of operational trend to outliers, and prevent the abnormal features of outliers from causing the sensitivity of type recognition to drop. The normalization process can construct a characteristic sequence so that each type of data can be unified in the same range independently of the scenario. In this way, the combination of characteristic sequence and two quantitative indicators can characterize the overall characteristics of the original data, which is the basis of identifying three typical data types.

Pseudo-code1: Data Feature Extraction

```

Input: Raw data  $X^N = \{x_1, x_2, \dots, x_N\}$ ,
Singular Values' Number  $m$ ,
Output: Optimal Decomposition Level  $n$ ,
Characteristic Sequence  $A'_n$ ,
Standard Deviations  $std$ ,
Oscillation Coefficient  $oscl$ 
1:  $m \leftarrow 10$ 
2: for  $i = 1$  to  $5$  do:
3:    $D_i \leftarrow X^N = D_1 + \dots + D_i + A_i$  //Step 1
4:    $e_{1i} \geq \dots \geq e_{mi} \geq 0 \leftarrow SSA(D_i)$  //DWT
5:    $K_i \leftarrow e_{1i}/e_{mi}$ 
6:   Calculate  $th_K$  from (1)
7:   if  $K_i > th_K$  then:
8:      $n \leftarrow i$ 
9:     break()
10:  end if
11: end for
12:  $A_n = \{a_{n1}, a_{n2}, \dots, a_{nN}\} \leftarrow (3)$ 
13:  $A'_n = \{a'_{n1}, a'_{n2}, \dots, a'_{nN}\}$  //Normalize
14:  $std \leftarrow std(A'_n)$ 
15:  $oscl \leftarrow oscl(A'_n)$ 
16: return ( $A'_n, std, oscl$ )

```

Step2 Type recognition

When the application scenarios and working conditions of low-cost sensors change frequently, it is

necessary to identify the corresponding data type according to the constantly updated features. The core task of this step is to focus on how to accurately identify the three typical data types, ensure that they are classified correctly, and provide the premise and foundation for matching high-performance algorithm strategies.

Recognition of the three typical data types is essentially a classification problem. As a classical classification model, back propagation neural network (BP neural network) has been widely used. In this paper, a neural network model is established to identify data types. Sample data sets containing three typical data types are used as the basis of establishing the model. The training set and test set are divided according to the Pareto rule, and stratified sampling is strictly used for segmentation (Matei and Bruno 2015). According to step 1, the key feature (A'_n , std and $oscl$) are extracted from each sample in the data set as the input of the neural network. The output of the model is the data type, so that automatic classification can be realized. The process of type recognition is shown in Fig.1.

Fig.1 The framework of the type recognition step shows the input, output and structural settings of the neural network model.

The selection of parameters has a significant impact on the classification effect of neural network model. Parameters in the model include the number of hidden neurons, learning rate and iteration times. Besides, regularization parameters are introduced to prevent over-fitting. A grid search is performed to optimize parameter values in our model (Xu et al. 2020). The above four parameters are combined within a certain search range, and the optimal combination is determined by comparing the model effects under different combinations of parameters.

Step3 Trend fitting

Traditional outlier detection algorithm works directly on the observed values, which often identify the normal fluctuations of data as outliers. In addition, due to the influence of the change trend, the algorithm is not sensitive to real outliers in the stationary phase. In order to overcome this limitation, we have added a trend fitting step to the framework. By using the method of Definition 2 to generate the residual sequence, the observations with large residual are more easily marked as outliers. The two trend fitting algorithms included in the framework are algorithm 1 based on signal decomposition and algorithm 2 based on prediction.

Definition2 Calculate the residual sequence

$fit_C^N = \{fit_c_1, fit_c_2, \dots, fit_c_N\}$ is the operational trend of the original data fitted by Algorithm 1 or Algorithm 2, in which fit_c_i is the fitted value corresponding to the i th observation. The generation method of residual sequence $R^N = \{r_1, r_2, \dots, r_N\}$ is as follows:

$$r_i = c_i - fit_c_i, \quad i = 1, 2, \dots, N \quad (3)$$

Algorithm1 (Signal decomposition-based method)

The wavelet transform is a commonly used tool in the field of signal decomposition. This method obtains different modes of signals by constructing an orthogonal wavelet filter bank, which can adapt to data with non-stationary characteristics (Cekim 2020; Gilles 2013). The signal decomposition method is included in Definition 1.

Algorithm2 (Prediction-based method)

The prediction-based method uses the historical data of the sequence to make predictions. One advantage of this method is that it takes into account the contextual characteristics of the time series, and has a better fit for data with a certain periodicity or steady change (Arumugam and Saranya 2018). A widely used prediction-based algorithm is Autoregressive Integrated Moving Average (ARIMA), which is also used in this article. In the ARIMA algorithm, the best model parameters are selected by observing the characteristics of autocorrelation graph, partial autocorrelation graph truncation, and tailing, combined with the AIC criterion, taking into account the comprehensive optimal configuration of fitting accuracy and the number of unknown parameters (Soni et al. 2015).

It is worth noting that not all data need residual preprocessing, and this step is not mandatory. When the original data fluctuates frequently and does not have predictable periodic fluctuations, it is often difficult to fit its operation trend. In this case, applying the outlier recognition algorithm directly to observation data may produce better results.

Step4 Outlier recognition

Outlier recognition is to find the observed values that do not conform to normal behaviors in the time series (Pineiro Di Blasi et al. 2015; Sanchez-Lasheras et al. 2020). Therefore, the characterization of data inconsistency and the boundary of outlier determination are crucial. Definition 3 sums up the idea of outlier recognition in the framework. According to the literature review, the three outlier recognition algorithms included in the framework are density-based Algorithm 3, statistic-based Algorithm 4 and context-based Algorithm 5. Pseudo-code 2 gives the core idea of outlier detection.

Definition3 Characterization of inconsistency

The degree of abnormal is characterized by assigning parameters $char_i$ to each observation value c_i . The preset fluctuation interval of the characteristic parameter is $Bound^N = \{bound_1, bound_2, \dots, bound_N\}$, which is determined according to statistical methods or inconsistent distribution of characteristic parameters. The features of data are updated in the sliding window to adapt to the change of sequence over time. $bound_i = [bound_{i,down}, bound_{i,up}]$ indicate the lower and upper boundaries of the preset fluctuation range of the i th data. We introduce a threshold scaling factor z to represent the strictness of the boundary setting. When $char_i$ falls outside of $z \cdot Bound_i$, the corresponding observation value is judged as an outlier.

Maximizing the efficiency of outlier detection and improving the performance of the algorithm, while also meeting actual application requirements, require the selection of reasonably good values for the threshold scaling factor z , the calculation window length k and other parameters in the combined algorithm. A grid search is performed to optimize parameter values in our model (Xu et al. 2020). Parameter pairs (z, k) are combined freely within the search interval and search step, and the outlier detection results obtained with the model are compared under different parameter combinations and degrees of applicability of the model. The optimization goal is to achieve the highest possible accuracy under the promise of a false positive rate.

Algorithm3 (Density-based method)

The density-based method compares the density of the point to be detected with other points in the neighborhood, so as to adapt to the non-unique characteristic of the data distribution pattern. LOF

algorithm is one of the representatives. According to the density of adjacent points, it assigns an outlier factor lof_i to each observation. The upper and lower limits of the threshold boundaries are determined by the box plot.

Algorithm4 (Statistic-based method)

The statistic-based method is to identify the points with obvious abnormal statistical characteristics in the sequence to be detected as outliers. Zscore algorithm is one of the representatives. Here, smoothing factor smo is introduced to weaken the influence of points that have been identified as outliers on statistical characteristics. The mean value avg_{ik} and standard deviation std_{ik} of the calculation window $lag_{ik} = \{c_{i-k}, c_{i-k+1}, \dots, c_{i-1}\}$ are constantly updated with the sliding window (Wang and Ahn 2020). The method of data smoothing is as follows:

$$c_j = \begin{cases} c_j & c_j \text{ is normal} \\ smo * c_j + (1 - smo) * c_{j-1} & c_j \text{ is outlier} \end{cases} \quad (4)$$

Here, $char_i = c_i - avg_{ik}$ is used as a parameter in ZScore algorithm to characterize the inconsistency of detection points at time t_i , and its fluctuation interval is $bound_i = [-std_{ik}, std_{ik}]$.

Algorithm5 (Context-based method)

The context-based method can synthesize historical data characteristics to determine the degree of outlier. ARIMA algorithm is one of the representatives. This method can not only fit the trend of the original data, but also give the possible fluctuation range of the predicted value $bound_i = [upline_i, downline_i]$ based on statistics. Here, $char_i = c_i$ is a parameter that characterizes the inconsistency of the detection point at time t_i .

Pseudo-code2: Outlier Detection

```

Input: Raw sequence  $X^N = \{(t_1, c_1), (t_2, c_2), \dots, (t_N, c_N)\}$ .
        Size of sliding window  $k$ ,
        Threshold scaling factor  $z_t$ 
Output: Outlier_local (the location of outliers in  $X^N$ )
1:  $fit\_C^N \leftarrow$  residual preprocess model( $C^N$ ) //Step 3
2:  $R^N \leftarrow \{\}$ 
3: for  $i = 1$  to  $N$  do:
4:    $r_i \leftarrow (c_i - fit\_c_i)$  //get residual value
5: end for
6:  $Char^N\{char_1, char_2, \dots, char_N\} \leftarrow \{\}$ 
7:  $Bound^N\{bound_1, bound_2, \dots, bound_N\} \leftarrow \{\}$ 
8: Outlier_local  $\leftarrow \{\}$ 
9: for  $i = k$  to  $N$  do:
10:   $SW^k \leftarrow \{c_{i-k+1}, c_{i-k+2}, \dots, c_i\}$  //get sliding window
11:   $Char^N, Bound^N \leftarrow$  outlier judgment model( $SW^k$ ) //Step 4
12:  if  $char_i \in [z \cdot bound_{i,down}, z \cdot bound_{i,up}]$  then:
13:     $c_i$  is normal
14:  else:
15:     $c_i$  is outlier
16:    outlier_local  $\leftarrow t_i$ 
17:  end if
18: end for
19: return(outlier_local)

```

Step5 Gap filling

After the outliers in the original data are detected, the vacancies in the abnormal positions need to be smoothly filled. For short gap filling, linear interpolation exhibits the best performance (Junninen et al. 2004). Therefore, linear interpolation is used to fill the gaps, thereby improving data quality.

2.3 Strategy of algorithm combination

This paper reviews the outlier detection algorithms for three typical data types, and proposes the most effective algorithm combination strategy in Step 3 and Step 4, which is integrated into the outlier detection framework. According to the characteristics of sharp value of pulse in data type 1, wavelet transform based on signal decomposition can fit abrupt waveform, and outlier recognition algorithm based on density can detect local outliers effectively. According to the frequent oscillation of data type 2, the statistics- based outlier recognition algorithm is directly applied to the original sequence, and outliers with obvious different statistical characteristics are detected. In view of the stable periodic trend of data type 3, the prediction-based algorithm can fit the original data trend, and the statistic-based algorithm can detect outliers in details. The specific combination strategies in the framework are shown in Table 2. Low-cost sensors built with this framework and algorithm combination strategy can be migrated and commonly used in three typical data types. Fig.2 is the framework of outlier detection proposed in this paper.

Table 2 The outlier detection strategy suitable for three typical data types shows the combination of the algorithms of step 3 and step 4 under the framework.

Data type	Strategy	Step3	Step4
Datatype1	W-D	Signal decomposition-based	Density-based
Datatype2	N-S	No need for this step	Statistic-based
Datatype3	P-S	Prediction-based	Statistic-based

Fig.2 Multi-scene outlier detection framework suitable for three typical data types includes five core steps to automatically improve data quality.

3. Multi-scenario experiments and model verification

3.1 Preparation of data set

Low-cost sensors were deployed in a factory to collect real data sets for the above three typical data types. Sensors monitor the concentration of oil fume and oil mist in different production links and the concentration of PM 2.5 in the workshop. The difference of data characteristics is necessary to construct

an original data set which contains three typical data types in various scenarios.

In accordance with the actual requirements for monitoring in the production process, the sampling frequency of PM2.5 data in the air was once per minute, and that of oil mist data in the factory was set at once every ten minutes to reflect the changes in concentration. For the oil fume data, the concentration peaked immediately when the fume-producing equipment started running, and then decreased rapidly in the steady-state. Therefore, the sampling frequency needed to be increased to five times per minute to accurately reflect the instantaneous change in oil fume concentration. After visual analysis, a basic data set containing 150 sample sequences of 150,000 observation points was established, thus covering three typical data types comprehensively and evenly.

It is worth noting that although the basic data set is the data of oil fume, oil mist and PM2.5 concentration from real factories, in the subsequent data type recognition process, we did not pay attention to the source and practical significance of the data, but only analyzed the data from the perspective of signal decomposition and feature extraction, which can represent the characteristics of the three typical data types. Therefore, the model built in this paper based on the basic data set still has wide applicability in multiple scenarios.

3.2 Analysis of experimental results

Firstly, automatic classification model is established for three typical data types. According to the stratified sampling method, 150 marker sample sequences are divided into 80% training set and 20% test set. After feature extraction for each sample sequence, two characteristic indicators and a characteristic sequence with a dimension of 1000 can be obtained. The input and output dimensions of the neural network model can be determined as 1002×120 and 1002×3 , respectively. A neural network model with a hidden layer has been built. The model parameters determined by grid search are shown in Table 3. The accuracy of this model on the training set can reach 100%. After the cross-validation of the test set, the classification accuracy of this model can reach 93.33%, and it can effectively distinguish three typical data types.

Table 3 Optimal model parameters of neural network obtained by grid search

Parameters	Optimal value
Neurons	21
Lambda	0.1
Learnrate	0.007
Iteration	25000

The framework proposed in this paper is applied in a variety of industrial scenarios to process real oil fume, oil mist and PM2.5 data, and the model's outlier detection and quality improvement effects are verified. After the singular spectrum analysis of the original data, the optimal decomposition level of the wavelet transform for the three typical data types is determined to be level 4. The operational trend of the three types of data is shown in Fig.3.

Fig.3 Three types of original sequence operational trend extraction: **a** oil fume, **b** oil mist, **c** PM2.5.

There are obvious outliers in the original sequence of the three types of data, but the operational trend extracted by the wavelet transform has less followability, which can effectively eliminate the influence of outliers. The operational trend is normalized into characteristic sequence, and the result is shown in Fig.4. The two classification indicators are calculated as described in Step 1, and the result is shown in Table 4.

Fig.4 The feature sequence diagrams of three types of original data sets are used to demonstrate the effect of wavelet transform and normalization processing.

Table 4 Two calculated characteristic indicators for three typical data sets. The unit of the *oscl* is %.

Indicator	Oil fume	Oil mist	PM2.5
<i>std</i>	0.12	0.18	0.22
<i>oscl</i>	0.90	1.70	1.10

The extracted features are input into the BP neural network model, and the type recognition results of the above three groups of real data are as follows: oil fume data belongs to data type 1, oil mist data belongs to data type 2, and PM 2.5 data belongs to data type 3. Next, three typical data types are matched with high performance algorithms. The outlier detection results of three groups of data are shown in Fig.5.

Fig.5 Outlier detection results for three types of original data sets: **a** oil fume, **b** oil mist, **c** PM2.5.

It is clear that the combined model can be adapted successfully to the different characteristics of the three typical data types. The detected outliers are evenly distributed, which conforms to the actual outlier situation when low-cost sensors are used. For short gap filling, linear interpolation exhibits the best performance (Ottosen and Kumar 2019). The sequence after filling is shown in Fig.6. After processing, the data is relatively smooth, and there are no obvious outliers; this result meets the actual monitoring needs.

Fig.6 Gap-filling results for three types of original data sets: **a** oil fume, **b** oil mist, **c** PM2.5.

To verify the effectiveness of our model in multi-scenario applications, we needed to test the model's adaptability and quantitatively evaluate its performance. Validation of the model is achieved through quantitative evaluation of labeled data sets. The filled sequence is regarded as normal labeled data, and six groups of outliers with different deviation rates are randomly inserted as outlier labeled data to simulate small and gross errors that may occur in the actual monitoring process with low-cost sensors. Table 5 summarizes the deviation rate of 120 outliers inserted in each group. The detection results of outliers are represented by the confusion matrix shown in Table 6. Among them, false negative (FN) indicates that a data point whose label is an outlier is judged as a normal point, and false positive (FP) indicates that a data point whose label is normal is judged as an outlier.

Table 5 six groups of outliers with different deviation rates

Groups of outliers	1	2	3	4	5	6
Deviation rates (%)	10	-10	30	-30	50	-50

Table 6 Confusion matrix for outlier detection

Label	Detected as outlier	Detected as normal
Outlier	True Positive (TP)	False Negative (FN)
Normal	False Positive (FP)	True Negative (TN)

When the application framework performs outlier detection on the generated labeled data set, a grid search is used to determine the optimal parameters. The length of the calculation window is set to $k = \{5, 6, \dots, 20\}$, and the threshold scaling factor to $z = \{1, 1.5, \dots, 5\}$. Under the premise of setting the global false positive rate limit to 8-15%, the outlier detection results of the model in the three sequences

are shown in Fig.7, and the outlier detection accuracy under different deviation rates is shown in Table 7.

Fig.7 Outlier detection results for three types of labeled data sets: **a** oil fume, **b** oil mist, **c** PM2.5.

Table 7 The accuracy of outlier detection under different deviation rates.

Data type	1	2	3	4	5	6
Datatype 1	70%	84%	66%	90%	68%	96%
Datatype 2	55%	65%	60%	65%	70%	75%
Datatype 3	75%	85%	60%	80%	85%	90%

In the grid-search process, the correlation between the calculation window length k , the threshold scaling factor z and the detection result can be obtained through quantitative evaluation. Fig.8 and Fig.9 are the results of a grid search under outlier group 5 with Type 3 data as an example. An increase in z indicates that the boundary of outlier determination is relaxed, and the probability of an outlier's being correctly detected decreases. When k increases, the false positive rate has a downward trend, while the false negative rate increases. The above results confirm the importance of reasonable selection of parameters in the outlier detection model.

Fig.8 The 5th group of outliers of Type 3 data is shown here as an example of the variation in outlier detection accuracy rate with z under different k values in the grid-search process.

Fig.9 The 5th group of outliers of Type 3 data is shown here as an example of the variation in the FP and FN rates for outlier detection with k under different z values in the grid-search process.

Now we define four adaptability evaluation indicators: sensitivity, specificity, positive prediction value (PPV) and negative prediction value (NPV) (Liu et al. 2014; Yu et al. 2014), which represent, respectively, the proportion of correctly detected outliers, the proportion of correctly detected normal points, the proportion of true outliers detected among all outliers detected, and the proportion of true normal points detected among all normal points detected. Based on the above test results, the adaptability evaluation indicators for the three typical data sets are summarized in Table 8.

Table 8 Calculation results for the four adaptability evaluation indicators in our model for the three typical data sets.

Datatype	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Datatype1	79	92.1	34.4	98.8
Datatype2	65.8	85.1	35.8	94.1
Datatype3	79.2	85.1	42.0	96.8

The specificity indicator directly reflects the limit of the false positive rate in the grid-search process. Sensitivity characterizes the overall accuracy of outlier detection in the labeled data set we are using. The sensitivity of model detection for Type 1 and Type 3 data can be close to 80%, while the sensitivity for Type 2 data is lower. When the specific analysis of the detection accuracy of outliers is combined with the different deviation rates mentioned above, it can be seen that since the small outliers with a low deviation rate of $\pm 10\%$ in the Type 2 data are mixed with normal points, the ZScore algorithm can easily miss them, leading to an overall reduction in sensitivity. Meanwhile, the NPV indicator is close to 100% for all data types, indicating that the model is reliable in detecting outliers. The PPV was low overall because the model in this paper allows a certain level of false positives, which is acceptable in the actual production and application process. The labeled data set contains a large sample-point base, and few outliers were inserted, which leads to a sharp drop in the PPV under the existing false positive rate. The model uses linear interpolation to fill the gaps, and the interpolation can still be kept within a fluctuation range that is normal for the context. After filling, the time series is smoother, meeting the actual monitoring needs, and it can effectively detect outliers and improve the overall data quality. In addition, when the model is combined with experimental analysis, obvious outliers such as the interrupted zero value of the original data and the extremely high value of the sensor error can be detected correctly. Thus, the actual effectiveness of the model will be higher than that of the quantitative detection.

3.3 Comparison of strategies

To verify that the algorithm combination strategy proposed in this paper is optimal, we further compare the application effects of other strategies in three typical data types. Under the framework of outlier detection, the different algorithms in Step 3 and Step 4 are matched with each other, and 8 different combination strategies are obtained, which are listed in Table 9. Quantitative evaluation indicators are used to evaluate the applicability of each strategy, and the comparison of detection accuracy is shown in Fig.10. Our strategy is clearly efficient for the three typical data types.

Table 9 List of 8 combined strategies under the framework of outlier detection.

Strategy	Step3	Step4
----------	-------	-------

P-S	Prediction-based	Statistic-based
P-D	Prediction-based	Density-based
W-D	Signal decomposition-based	Density-based
W-S	Signal decomposition-based	Statistic-based
W-C	Signal decomposition-based	Context-based
N-D	No need for this step	Density-based
N-S	No need for this step	Statistic-based
N-C	No need for this step	Context-based

Fig.10 Comparison of outlier detection accuracy rates of 8 strategies in three typical data types

Considering that difficulty and diversity levels can vary greatly between datasets and evaluation metrics, it is unfair to directly compare quantitative evaluation indicators of different data types. Considering this, we introduce the concept of “relative performance”, which is calculated by taking the average difference between the absolute performance of this strategy and that of all the other strategies (Calikus et al. 2020). The relative performance calculation formula is as follows.

$$Rel_d^M = \frac{1}{|S_d|} \sum_{s \in S_d} \frac{\sum_{\hat{s} \in \hat{S}_d} s - \hat{s}}{\max(S_d \cup \hat{S}_d) - \min(S_d \cup \hat{S}_d)} \quad (5)$$

Where d is a target data set, S is the absolute performance list under method M , and $\hat{S}$ is the list of absolute performance of other methods. Table 10 shows the relative performance of these 8 strategies in the three typical data types.

Table 10 Calculate relative performance on three typical data types and compare algorithm combination strategies.

Strategy	Relative sensitivity	Relative specificity	Relative PPV	Relative NPV
1 - N - D	3.82	2.04	10.39	2.93
1 - N - C	-10.73	-7.25	-53.61	-3.88
1 - N - S	4.48	0.89	9.44	3.64
1 - P - D	-6.30	-0.09	-12.72	-2.14
1 - P - S	-4.54	0.60	-6.94	-1.65
1 - W - D	4.69	1.18	10.16	3.87

1 - W - C	-26.85	0.89	-43.54	-4.38
1 - W - S	1.85	1.09	5.43	1.26
2 - N - D	-7.50	1.03	-7.65	-2.75
2 - N - C	1.62	-2.85	-4.39	0.42
2 - N - S	6.95	-1.78	4.69	5.26
2 - P - D	-15.23	5.34	2.88	-3.39
2 - P - S	3.06	-2.85	-1.96	1.86
2 - W - D	-1.75	5.27	11.20	0.42
2 - W - C	-2.19	-3.59	-12.73	-2.32
2 - W - S	0.63	-1.87	-3.96	0.21
3 - N - D	-7.94	-0.58	-6.96	-4.26
3 - N - C	0.12	-2.69	-4.99	-0.59
3 - N - S	4.30	-2.44	-1.57	3.30
3 - P - D	-8.57	4.40	3.50	-3.48
3 - P - S	5.15	-2.44	-0.99	4.20
3 - W - D	-1.95	5.93	8.21	-0.40
3 - W - C	-1.60	-0.97	-3.54	-1.35
3 - W - S	3.18	-2.44	-2.42	2.21

We can see the advantages of the combination strategy adopted by this model for the three typical data types. Under the premise of global false positive rate limit, outlier detection pays more attention to sensitivity and negative predictive value, which have the highest scores under the strategy selected in this paper, and the corresponding specificity and positive predictive value also perform well. On the other hand, the data type identification step in the outlier detection framework is crucial for subsequent algorithm matching. The accuracy of the automatic classification model in this paper can reach 93.3%, and the decrease in the relative sensitivity of the model caused by the judgment error is also controllable, as shown in Table 11. This result strongly validates the effectiveness of this framework and the superiority of the strategy.

Table 11 The degree of decrease in the relative sensitivity of outlier detection caused by type recognition errors

Real datatype	Detected as datatype1	Detected as datatype2	Detected as datatype3
Datatype1	0.00%	-0.05%	-2.05%
Datatype2	-2.75%	0.00%	-1.23%
Datatype3	-3.52%	-0.43%	0.00%

According to the characteristics of data type 1, on the premise of keeping normal oscillation by using wavelet transform, outliers which are obviously different from neighboring points can be detected by using density-based algorithm of residual sequence. However, the trend fitting algorithm based on prediction can't reflect the sharp value in data type 1 because it depends on the trend of historical data, which leads to the normal peak point being judged as abnormal due to its large residual error. Data type 2 oscillates frequently and has no periodic characteristics, which makes it difficult to fit the trend of the original sequence. However, it often has consistent statistical characteristics in the sliding window, so it is suitable for directly applying the outlier recognition algorithm based on statistics, and updating outliers

smoothly in time. Prediction-based method can effectively fit the trend of data type 3, and ZScore algorithm can be used to detect outliers with obvious different statistical characteristics. However, the residual preprocessing strategy based on wavelet analysis retains some details of the original data, and weakens the abnormal degree of outliers. Therefore, combined with the algorithm principle and data characteristics, the outlier detection framework and algorithm combination strategy proposed in this paper have obvious advantages in dealing with three typical data types.

4. Conclusions

In this research, we creatively match data types with high-performance algorithms from the perspective of data characteristics. The framework we proposed can be applied to three typical data types to complete outlier detection tasks in multiple scenarios in the field of industrial pollutant monitoring. Five mutually independent core steps are used to solve the basic tasks in outlier detection, which can automatically identify data types and match the highest performance algorithm combination strategy for each data type. The quantitative evaluation of the model in this paper was carried out on the actual data of the factory, and the relative performance of the traditional single outlier detection algorithm and other combination strategies in the framework were compared in the data set. For the pollutant monitoring process that is not included in the three typical data types, it is necessary to further consider the most suitable outlier detection algorithm. The following conclusions can be drawn from our research:

(1) The characteristic sequence obtained by wavelet transform and the two indicators representing discreteness and volatility can form a feature vector to distinguish data types. The classification accuracy of the parameter-optimized BP neural network model can reach 93.3%, and the degradation of outlier recognition performance caused by classification errors can be controlled within 3.52%. The results shows that the data type identification step can cope with the scene migration and the change of working conditions in the process of pollutant monitoring, and provide reliable guarantee for the adaptive adjustment of outlier detection strategy.

(2) The combined outlier detection model adopts corresponding algorithms for different data types: the wavelet transform and the density-based LOF algorithm for Type 1 data, and a sliding window and the statistics-based ZScore algorithm for Type 2 data. On this basis, the LOF model combines with the context-based ARIMA algorithm to process Type 3 data. Under the premise of an 8–15% false positive rate limit, our model can increase the outlier detection accuracy of the three typical data types to 70–90%, which is better than the accuracy of other algorithm combination strategies in the framework.

(3) The outlier detection framework includes two main parameters: the threshold scaling factor z and the calculation window length k , which have an important impact on the data quality improvement in the actual pollutant monitoring process. A grid search can be used to determine the best values of these parameters, and the search target must be determined according to the false positive rate required by actual monitoring.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data and materials used during the current study are available from the corresponding author upon request.

Competing interests

The authors declare that they have no competing interests

Funding

This study was supported by National Science Foundation of China (Grant No.51878442)

Authors' contributions

ZL deployed low-cost sensors to monitor pollutant concentration and provided raw data sets. YX developed the outlier detection framework and processed the data. WP and YW contributed to the improvement of the paper. All co-authors have contributed to this research.

Notes and references

- Aggarwal CC, Sathe S (2017) Outlier Ensembles: An Introduction. Springer International Publishing, Cham, pp 1-34. doi:10.1007/978-3-319-54765-7_1
- Aljoumani B, Sanchez-Espigares JA, Canameras N, Josa R, Monserrat J (2012) Time series outlier and intervention analysis: Irrigation management influences on soil water content in silty loam soil. *Agricultural Water Management* 111:105-114 doi:10.1016/j.agwat.2012.05.008
- Aljoumani B, Sanchez-Espigares JA, Canameras N, Wessolek G, Josa R (2018) Transfer Function and Time Series Outlier Analysis: Modelling Soil Salinity in Loamy Sand Soil by Including the Influences of Irrigation Management and Soil Temperature. *Irrigation and Drainage* 67:282-294 doi:10.1002/ird.2187
- Arumugam P, Saranya R (2018) Outlier Detection and Missing Value in Seasonal ARIMA Model Using Rainfall Data. *Materials Today-Proceedings* 5:1791-1799 doi:10.1016/j.matpr.2017.11.277
- Ayadi H, Zouinkhi A, Boussaid B, Abdelkrim MN, Ieee (2015) A Machine Learning Methods: Outlier detection in WSN. In: 2015 16th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering. *International Conference on Sciences and Techniques of Automatic Control and Computer Engineering*. pp 722-727

- Barcelo-Ordinas JM, Doudou M, Garcia-Vidal J, Badache N (2019) Self-calibration methods for uncontrolled environments in sensor networks: A reference survey. *Ad Hoc Networks* 88:142-159 doi:<https://doi.org/10.1016/j.adhoc.2019.01.008>
- Bin GF, Gao JJ, Li XJ, Dhillon BS (2012) Early fault diagnosis of rotating machinery based on wavelet packets-Empirical mode decomposition feature extraction and neural network. *Mech Syst Signal Proc* 27:696-711 doi:10.1016/j.ymssp.2011.08.002
- Breunig MM, Kriegel HP, Ng RT, Sander J (2000) LOF: Identifying density-based local outliers. *Sigmod Rec* 29:93-104 doi:10.1145/335191.335388
- Calikus E, Nowaczyk S, Sant'Anna A, Dikmen O (2020) No free lunch but a cheaper supper: A general framework for streaming anomaly detection. *Expert Systems With Applications* 155
- Campos GO et al. (2016) On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Min Knowl Discov* 30:891-927 doi:10.1007/s10618-015-0444-8
- Cekim HO (2020) Forecasting PM10 concentrations using time series models: a case of the most polluted cities in Turkey. *Environmental Science and Pollution Research* 27:25612-25624 doi:10.1007/s11356-020-08164-x
- Chen-Chia Chen C-TK (2018) Calibration of Low-Cost Particle Sensors by Using Machine-Learning Method. *APCCAS*. <https://doi.org/10.1109/apccas.2018.8605619>
- Chen L-J, Ho Y-H, Hsieh H-H, Huang S-T, Lee H-C, Mahajan S (2018) ADF: An Anomaly Detection Framework for Large-Scale PM2.5 Sensing Systems. *Ieee Internet of Things Journal* 5:559-570 doi:10.1109/jiot.2017.2766085
- Dong Z, Jia H (2020) Outlier detection method for thermal process data based on EWT-LOF. *Chinese Journal of Scientific Instrument* 41:126-134 (in Chinese)
- Durre I, Menne MJ, Gleason BE, Houston TG, Vose RS (2010) Comprehensive Automated Quality Assurance of Daily Surface Observations. *Journal of Applied Meteorology and Climatology* 49:1615-1633 doi:10.1175/2010jamc2375.1
- Garoudja E, Harrou F, Sun Y, Kara K, Chouder A, Silvestre S (2017) Statistical fault detection in photovoltaic systems. *Sol Energy* 150:485-499 doi:10.1016/j.solener.2017.04.043
- Gilles J (2013) Empirical Wavelet Transform. *IEEE Trans Signal Process* 61:3999-4010 doi:10.1109/tsp.2013.2265222
- Guan H, Li Q, Yan Z, Wei W, Ieee (2015) SLOF: Identify Density-based Local Outliers in Big Data. 2015 12th Web Information System and Application Conference. doi:10.1109/wisa.2015.40
- Heyder J, Gebhart J, Rudolf G, Schiller CF, Stahlhofen W (1986) Deposition of particles in the human respiratory tract in the size range 0.005–15 μm . *Journal of Aerosol Science* 17:811-825 doi:[https://doi.org/10.1016/0021-8502\(86\)90035-2](https://doi.org/10.1016/0021-8502(86)90035-2)
- Hsu H-I et al. (2012) Assessing Long-Term Oil Mist Exposures for Workers in a Fastener Manufacturing Industry Using the Bayesian Decision Analysis Technique. *Aerosol and Air Quality Research* 12:834-842 doi:10.4209/aaqr.2012.04.0107
- Hui M et al. (2018) Detecting Anomalies in Time Series Data via a Meta-Feature Based Approach. *Ieee Access* 6:27760-27776 doi:10.1109/access.2018.2840086
- Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. *Atmospheric Environment* 38:2895-2907 doi:10.1016/j.atmosenv.2004.02.026
- Khalid R, Javaid N, Al-zahrani FA, Aurangzeb K, Qazi EUH, Ashfaq T (2020) Electricity Load and Price Forecasting Using Jaya-Long Short Term Memory (JLSTM) in Smart Grids. *Entropy* 22:21

doi:10.3390/e22010010

Li L-L, Chang Y-B, Tseng M-L, Liu J-Q, Lim MK (2020) Wind power prediction using a novel model on wavelet decomposition-support vector machines-improved atomic search algorithm. *Journal of Cleaner Production* 270 doi:10.1016/j.jclepro.2020.121817

Liu S, Che H, Smith K, Chen L (2014) Contamination event detection using multiple types of conventional water quality sensors in source water. *Environmental Science-Processes & Impacts* 16:2028-2038 doi:10.1039/c4em00188e

Long Z, Wang Y (2019) Monitoring and purification of oil mist particles in a machining workshop. *Heating Ventilating & Air Conditioning* 49:50-55 (in Chinese)

Madakyaru M, Harrou F, Sun Y (2017) Improved data-based fault detection strategy and application to distillation columns. *Process Saf Environ Protect* 107:22-34 doi:10.1016/j.psep.2017.01.017

Malings C, Tanzer R, Hauryliuk A, Saha PK, Robinson AL, Presto AA, Subramanian R (2019) Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation. *Aerosol Science and Technology* 54:160-174 doi:10.1080/02786826.2019.1623863

Massin N, Bohadana AB, Wild P, Goutet P, Kirstetter H, Toamain JP (1996) Airway responsiveness, respiratory symptoms, and exposures to soluble oil mist in mechanical workers. *Occupational and environmental medicine* 53:748-752 doi:10.1136/oem.53.11.748

Matei SA, Bruno RJ (2015) Pareto's 80/20 law and social differentiation: A social entropy perspective. *Public Relations Review* 41:178-186 doi:10.1016/j.pubrev.2014.11.006

Ottosen TB, Kumar P (2019) Outlier detection and gap filling methodologies for low-cost air quality measurements. *Environ Sci Process Impacts* 21:701-713 doi:10.1039/c8em00593a

Pineiro Di Blasi JI, Martinez Torres J, Garcia Nieto PJ, Alonso Fernandez JR, Diaz Muniz C, Taboada J (2015) Analysis and detection of functional outliers in water quality parameters from different automated monitoring stations in the Nalon River Basin (Northern Spain). *Environmental Science and Pollution Research* 22:387-396 doi:10.1007/s11356-014-3318-5

Qiu Y, Chang X, Qiu Q, Peng C, Su S (2020) Stream data anomaly detection method based on long short-term memory network and sliding window. *Journal of Computer Applications* 40:1335-1339

Quispe-Coica A, Perez-Foguet A (2020) Preprocessing alternatives for compositional data related to water, sanitation and hygiene. *The Science of the total environment* 743:140519 doi:10.1016/j.scitotenv.2020.140519

Rhif M, Ben Abbes A, Martinez B, Farah IR (2020) An improved trend vegetation analysis for non-stationary NDVI time series based on wavelet transform. *Environmental Science and Pollution Research* doi:10.1007/s11356-020-10867-0

Sanchez-Lasheras F, Ordonez-Galan C, Jose Garcia-Nieto P, Garcia-Gonzalo E (2020) Detection of outliers in pollutant emissions from the Soto de Ribera coal-fired power plant using functional data analysis: a case study in northern Spain. *Environmental Science and Pollution Research* 27:8-20 doi:10.1007/s11356-019-04435-4

Sharma KG, Ram A, Singh Y (2011) Efficient Density Based Outlier Handling Technique in Data Mining. In: Meghanathan N, Kaushik BK, Nagamalai D (eds) *Advances in Computer Science and Information Technology, Pt I*, vol 131. Communications in Computer and Information Science. Springer-Verlag Berlin, Berlin, pp 542-+

Soni K, Parmar KS, Kapoor S (2015) Time series model prediction and trend variability of aerosol optical depth over coal mines in India. *Environmental Science and Pollution Research* 22:3652-3671 doi:10.1007/s11356-014-3561-9

- Tie CAI, Jie ZHU (2006) Adaptive Selection of Optimal Decomposition Level in Threshold De-noising Algorithm Based on Wavelet. *Control and Decision* 21:217-220
- Valenzuela O et al. (2008) Hybridization of intelligent techniques and ARIMA models for time series prediction. *Fuzzy Sets and Systems* 159:821-845 doi:10.1016/j.fss.2007.11.003
- Wang B, Yang XC, Wang GR, Yu G (2010) Outlier Detection over Sliding Windows for Probabilistic Data Streams. *J Comput Sci Technol* 25:389-400 doi:10.1007/s11390-010-9332-2
- Wang XL, Ahn SH (2020) Real-time prediction and anomaly detection of electrical load in a residential community. *Appl Energy* 259:10 doi:10.1016/j.apenergy.2019.114145
- Wu H, Tang X, Wang Z, Wu L, Lu M, Wei L, Zhu J (2018) Probabilistic Automatic Outlier Detection for Surface Air Quality Measurements from the China National Environmental Monitoring Network. *Advances in Atmospheric Sciences* 35:1522-1532 doi:10.1007/s00376-018-8067-9
- Xu HT et al. (2020) AGA-SVR-based selection of feature subsets and optimization of parameter in regional soil salinization monitoring. *Int J Remote Sens* 41:4470-4495 doi:10.1080/01431161.2020.1718239
- Yang J, Wang J, Zheng Y, Lei M, Yang J, Wan X, Chen T (2018) Method for identifying outliers of soil heavy metal data. *Environmental Science and Pollution Research* 25:12868-12875 doi:10.1007/s11356-018-1555-8
- Yang X, Zhou W, Shu N, Zhang H (2019) A Fast and Efficient Local Outlier Detection in Data Streams. Paper presented at the Proceedings of the 2019 International Conference on Image, Video and Signal Processing, Shanghai, China,
- Yu Y, Zhu Y, Li S, Wan D (2014) Time Series Outlier Detection Based on Sliding Window Prediction. *Mathematical Problems in Engineering* 2014 doi:10.1155/2014/879736
- Zhang H, Zhang S, Pan W, Long Z (2020) Low-cost sensor system for monitoring the oil mist concentration in a workshop *Environmental science and pollution research international* doi:10.1007/s11356-020-11709-9
- Zhang J, Hou D, Wang K, Huang P, Zhang G, Loaiciga H (2017) Real-time detection of organic contamination events in water distribution systems by principal components analysis of ultraviolet spectral data *Environmental Science and Pollution Research* 24:12882-12898 doi:10.1007/s11356-017-8907-7

Figures

Figure 1

The framework of the type recognition step shows the input, output and structural settings of the neural network model.

Figure 2

Multi-scene outlier detection framework suitable for three typical data types includes five core steps to automatically improve data quality.

Figure 3

Three types of original sequence operational trend extraction: a oil fume, b oil mist, c PM2.5.

Figure 4

The feature sequence diagrams of three types of original data sets are used to demonstrate the effect of wavelet transform and normalization processing.

Figure 5

Outlier detection results for three types of original data sets: a oil fume, b oil mist, c PM2.5.

Figure 6

Gap-filling results for three types of original data sets: a oil fume, b oil mist, c PM2.5.

Figure 7

Outlier detection results for three types of labeled data sets: a) oil fume, b) oil mist, c) PM_{2.5}.

Figure 8

The 5th group of outliers of Type 3 data is shown here as an example of the variation in outlier detection accuracy rate with z under different k values in the grid-search process.

Figure 9

The 5th group of outliers of Type 3 data is shown here as an example of the variation in the FP and FN rates for outlier detection with k under different z values in the grid-search process.

Figure 10

Comparison of outlier detection accuracy rates of 8 strategies in three typical data types