Based on Correlation Analysis and K-Means: An Anomaly Detection Algorithm for Seasonal Time-Series Data

Xin Wang
School of Electronic Information and Artificial Intelligenceline
Shaanxi University of Science and Technology, China
wangxin@sust.edu.cn

Yingxue Yang
School of Electronic Information and Artificial Intelligenceline
Shaanxi University of Science and Technology, China
1551927946@qq.com

Xueshuang Ding
School of Electronic Information and Artificial Intelligenceline
Shaanxi University of Science and Technology, China
dingxueshuang999@163.com

Yantao Zhao
School of Electrical Engineering
Yanshan University, China
ysuzyt@ysu.edu.cn

Abstract: *Anomaly detection is widely used in fields like data processing, intrusion detection, and financial fraud prevention, helping to avoid potential accidents and economic losses. In time series anomaly detection, which deals with numerical sequences over time (e.g., urban temperatures, sales data, stock market trends), the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is an excellent choice. This paper presents an improved anomaly detection algorithm tailored for seasonal time series data. By combining autocorrelation coefficients with the K-means algorithm, precise clustering results down to the date level are provided, subsequently employing the DBSCAN algorithm for detection, the enhanced algorithm is capable of capturing a greater number of local anomalies. Experiment conducted on daily temperature data from Beijing and Sanya in 2023, the enhanced algorithm exhibited a respective increase of 11.6% and 78% in anomaly detection compared to the original algorithm, thus affirming the feasibility of the approach*.

Keywords: *Anomaly detection, seasonal data, time series, DBSCAN, K-means.*

Received June 8, 2024; accepted September 21, 2024
https://doi.org/10.34028/iajit/21/6/2

1. Introduction

Anomaly detection is primarily employed to identify anomalous data points that deviate from the normal distribution pattern within a dataset [1]. This technology is widely applied in various fields, such as big data processing [8], intrusion detection systems [27], credit card fraud [9], network traffic anomaly detection [26], and so on. Anomaly detection utilizes detected anomalies for data analysis to mitigate potential incidents and economic losses.

Hawkins [10] characterizes an outlier as an observation that garners attention due to its substantial deviation from other observations within the dataset. Thus, anomalies in time series data can be described as data points at particular time steps that exhibit behavior different from previous time step data points [17].

Anomaly detection methods can be classified into three categories: the first category is rule-based methods, such as fixed thresholds, dynamic thresholds, etc., although these methods can accurately detect anomalies that adhere to predefined rules, they are limited by the size of the rule repository; the second category is statistical-based detection methods, such as 3Sigma, moving average cost method, autoregressive integrated moving average models, etc., require assuming that the data follows a certain distribution.

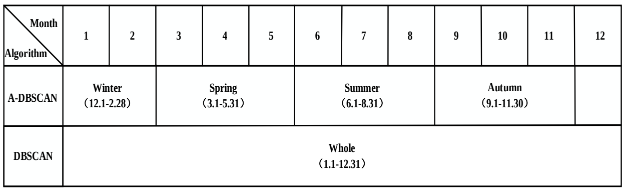
These methods are suitable for low-dimensional data; the third category is based on machine learning detection methods, such as clustering, autoencoders [18], random forests, etc. These methods have different constraints depending on the application context.

In recent years, the availability of time series data has grown exponentially [13]. Time series, also known as dynamic sequences, refers to numerical sequences where indicators of a phenomenon are arranged in chronological order, such as annual GDP, urban temperatures, population figures, stock trends, and so forth. The primary objective of studying time series data modeling is for data prediction, and to propose effective response strategies based on the predictive results.

For time series anomaly detection, Liu *et al*. [19] proposed Time-series Generative Adversarial Network (TimeGAN) to address the issue of insufficient abnormal data. The Time-GAN model can generate a large volume of reliable dataset based on a small amount of data, thereby reducing the time required for manual labeling of training dataset and enhancing the accuracy of fault diagnosis models. Lin *et al.* [20] proposed an unsupervised algorithm for detecting time series anomalies by combining Variational Auto-Encoders

(VAE) with Long Short-Term Memory (LSTM) networks. The model utilizes both a VAE module for forming robust local features over short windows and a LSTM module for estimating the long term correlation in the series on top of the features inferred from the VAE module. This detection algorithm is capable of identifying anomalies that span over multiple time scales. Li [21] proposed a rapid unsupervised anomaly detection framework addressing the limitation of the SPOT algorithm, which is sensitive only to extreme values in single-dimensional time series anomaly detection. By transforming non-extreme anomalies into extreme values, this framework achieves a significant improvement in detection accuracy.

The urban temperature dataset features low dimensionality, ease of processing, and suitability for seasonal classification. Based on this groundwork, employing the DBSCAN algorithm for anomaly detection is a commendable choice. Ghamkhar *et al.* [6] proposed an algorithm that relies on DBSCAN as the core, with Lempel-Ziv Complexity as a key feature, to overcome the curse of dimensionality and low data resolution. Feng *et al.* [14] proposed an anomaly detection method based on Variable-Scale Hypercube Accelerated Density-Based Spatial Clustering of Applications with Noise (VHCA-DBSCAN). Initially, they established the HCA-DBSCAN model based on Gaussian probability density estimation to effectively identify suspicious outliers. Then, they designed a traversal search strategy based on hypercube segmentation, where the length of edges varies according to the features of each dimension. Lastly, they utilized approximate upper bounds to propose a modified Local Outlier Factor (LOF) for assessing the degree of suspected outliers. This algorithm is applicable to dataset characterized by multidimensionality and high levels of noise. Dai *et al.* [3] proposed an improved DBSCAN algorithm, which combines KNN and Binning. This algorithm is able to reflect the distribution characteristics of the dataset itself, enabling adaptive parameter selection, and overcoming the limitations of traditional DBSCAN methods, which often struggle with parameter selection and exhibit a strong correlation between parameters and detection accuracy. Jain *et al.* [15] proposed the Attributes-DBSCAN (A-DBSCAN) algorithm by enhancing the classical DBSCAN algorithm [5]. Adding month labels to the dataset increased the fine-grained characterization of the data, thereby overcoming the drawback of DBSCAN, which is sensitive only to global anomalies in yearly cycles and insensitive to local anomalies. The algorithm performs well on specific seasonal time series dataset. However, due to the monthly partitioning of data still lacking granularity, therefore, for dataset with smooth fluctuations, the detected number of outliers will decrease. The classification results of A-DBSCAN and DBSCAN algorithms are shown in Figure 1 below.



Month / Algorithm	1	2	3	4	5	6	7	8	9	10	11	12
A-DBSCAN	Winter （12.1-2.28）		Spring （3.1-5.31）			Summer （6.1-8.31）			Autumn （9.1-11.30）			
DBSCAN	Whole （1.1-12.31）											

Figure 1. The classification results of the A-DBSCAN and DBSCAN algorithms.

Therefore, based on the work in [15], this paper further refines the classification of the dataset by combining autocorrelation coefficients with the K-means algorithm. Then, outlier detection is conducted using the DBSCAN algorithm. The final result is an anomaly detection algorithm, K-DBSCAN, which combines autocorrelation coefficients with the K-means algorithm. We conducted comparative experiments on temperature dataset from Beijing and Sanya in 2023. The results indicate that, compared to the DBSCAN algorithm, K-DBSCAN can detect local anomalies. Moreover, relative to the A-DBSCAN algorithm, K-DBSCAN is able to identify a greater number of local anomaly points unaffected by data variations. The main contributions of this paper are as follows:

- The design idea of this paper is to refine the clustering results (datasets) from months to days by fine-grained clustering of weather datasets, that is, the previous research is based on months, and the proposed method can achieve clustering in days, that is, more fine-grained depictions.
- The method adopted is to combine the autocorrelation analysis method with the K-means clustering method to solve the specific limitations of the traditional DBSCAN algorithm that the traditional DBSCAN is not sensitive enough to local anomalies through a fine-grained division. Firstly, the correlation analysis of the month is carried out through the autocorrelation matrix, and the correlation matrix of the size of 12*12 with the month as the basic unit is obtained, and the coarse classification of the annual temperature is obtained by setting different thresholds, and then the draw value of each class is set as the center value of the kmeans algorithm, and the clustering results in daily units are obtained through distance judgment, so as to realize the fine-grained division and improve the sensitivity to local anomalies.
- In terms of experiments, the detailed experimental settings are given on the temperature datasets in Beijing and Sanya, and the experimental results prove the feasibility and effectiveness of the proposed K-DBSCAN algorithm. The number of anomalies identified by the K-DBSCAN algorithm is 14.9% and 942.6% higher than those found by the DBSCAN algorithm in the Beijing and Sanya datasets, as well as 11.6% and 78.0% higher than those detected by the A-DBSCAN algorithm.

The remaining sections of this paper are as follows: Section 2 introduces the various methods employed in the study. Section 3 details the improved anomaly detection algorithm. Section 4 outlines the experimental setup and provides result analysis. Finally, section 5 concludes the paper, discussing its limitations and suggesting future research directions.

2. Overview and Background

2.1. Autocorrelation Coefficient

The autocorrelation coefficient is a statistical measure used in time series analysis to quantify the correlation between a time series and its own lagged values. Typically denoted by ρ, the autocorrelation coefficient ranges from [-1, 1]. Specifically, it assesses the correlation of a time series at different time points.

Assuming there is a time series $t=\{x_1, x_2\ldots, x_n\}$, the formula for computing the autocorrelation coefficient p with lag h is as follows,

$$\rho=\sum_{i=1}^{n-h}\frac{(x_i-\hat{\mu})(x_{i+h}-\hat{\mu})}{\sum_{i=1}^{n}(x_i-\hat{\mu})^2} \tag{1}$$

Where $\hat{\mu}$ represents the sample mean of the time series t.

Correlation analysis [4] is commonly employed to investigate the interrelationships between different attributes of objects. Common methods for correlation analysis include correlation coefficients, covariance, maximal mutual information, dynamic time warping algorithms [16], and similarity measures based on data dimensionality reduction.

This article considers the temperature data for each month as a whole and measures the data correlation between months using autocorrelation coefficients. This method enables the implementation of coarse-grained classification of the data and sets the stage for fine-grained classification using the K-means algorithm.

2.2. DBSCAN Algorithm

DBSCAN is a density-based clustering algorithm. The core idea of density-based clustering algorithms is to determine clusters based on the density of the neighboring area. If the density of the neighboring area of an object or data point exceeds a predefined threshold, it is added to clusters with similar density.

The DBSCAN [28] algorithm's advantages lie in its ability to discover clusters of arbitrary shapes and label outlier points, making it widely applicable in fields like data mining [22], image segmentation [23, 24, 25], geographic information systems [2], and signal processing [11, 29, 30].

DBSCAN can identify noise points and exhibits good robustness to outliers. However, compared to the K-means algorithm, DBSCAN requires more time to iterate to convergence. Additionally, DBSCAN has a decentralized nature, meaning there are no cluster centers. Jain *et al*. [15] points out the difficulty of DBSCAN in identifying local anomalies. Concurrently, time series data possesses a characteristic wherein the data within each cluster should demonstrate temporal continuity after clustering. Therefore, this paper combines the K-means algorithm with the DBSCAN algorithm. On one hand, it can obtain satisfactory clustering results through the K-means algorithm. On the other hand, conducting outlier detection separately on the results of K-means clustering can effectively reduce the execution time of DBSCAN [7].

The relevant definitions of the DBSCAN algorithm are as follows:

1. Eps-Neighborhood: objects within a radius of Eps from an object and can be represented by the relation,

$$N_{Eps}(p)=\{q\in D\mid Dist(p,q)\le Eps\} \tag{2}$$

Where, D represents the dataset; $Dist(p,q)$ signifies the distance between objects p and q; $N_{Eps}(p)$ encompasses all objects within dataset D that lie at a distance no greater than Eps from object p.

2. Neighborhood density threshold: Eps-neighborhood of an object containing at least *Minpts* of data points.
3. Core Object: given dataset D and a specified neighborhood density threshold *MinPt*s, if there exists an object p in D such that it satisfies in Equation (3), then object p is considered a core object,

$$|N_{Eps}(p)|\ge Minpts \tag{3}$$

Where $|N_{Eps}(p)|$ denotes the number of objects in the Epsilon neighborhood of object p.

The implementation process of the DBSCAN algorithm is as shown in Algorithm (1):

Algorithm 1: DBSCAN.

```
Input:
$Eps$: radius parameter
$Minpts$: neighborhood density threshold
$D$: dataset
Mark all samples in dataset $D$ as unvisited.
Do
  Randomly select an unvisited sample $p$.
  Mark $p$ as visited.
  Let $N_{Eps}(p)$ be the set of all neighboring points of point $p$.
  if $|N_{Eps}(p)| \ge Minpts$ then
    Create a new cluster $M$ and add $p$ to $M$.
    for every point $k$ in $N_{Eps}(p)$ do
      if the mark for $k$ is unvisited then
        Mark $k$ is visited.
        if $|N_{Eps}(p)| \ge Minpts$ then
          add these points to $N_{Eps}(p)$.
        end if
        if $k$ is not yet a member of any cluster then
          add $k$ to $M$.
        end if
      end if
    end for
  else mark $p$ as a noise point.
  end if
Until there are no objects marked as unvisited
Output: noise point
```

1. Set the sizes of parameters *Eps* and *Minpts*, and label all points as unvisited.
2. Randomly select a point p from the dataset and count the number of points within the *Eps* neighborhood of point p. If the number of points is greater than or equal to *Minpts*, label point p as a core point and create a new cluster. Starting from point p, search for points that are density-reachable from point p and find the maximum set of density-connected points. If the number of points in this set is less than *Minpts*, label point p as a noise point.
3. Select another point from the dataset and repeat 2, until all points are labeled as visited.

3. Anomaly Detection Algorithm Based on Correlation Analysis and K-Means

Kim *et al*. [16] divides the dataset into different categories based on seasonality, adding additional labels to overcome the limitation of DBSCAN in identifying local anomalies. However, the division based on months may not be applicable to all dataset. Taking city temperatures as an example, the distinction between seasons in many cities may not be clear-cut. Therefore, a more fine-grained division method based on historical data is needed. The implementation process of the K-DBSCAN algorithm is as shown in Algorithm (2).

Algorithm 2: K-DBSCAN.

```
Input:
$l$: calculation days threshold
$\tau$ : similarity threshold
$\sigma$: thresholds for fine-grained segmentation
$Eps$ and $Minpts$: parameters of the DBSCAN
Calculate the correlation coefficient $\rho_{ij}$ according to the (5), then we can get the autocorrelation matrix $P(\rho_{ij})_{12\times12}$.
Cluster the data according to the threshold $\tau$ , then get the clustering result $\{C_r\}_{r=1}^{k}$ . The autocorrelation coefficient between two months: $\rho_{ij} \geq \tau$ .
Calculate the average of the data in the clusters $C_1, C_2 \dots, C_k$, respectively, to obtain the corresponding k cluster centres $w_1, w_2 \dots w_k$.
for each cluster $r = 1$ to $k$ do
    $n \leftarrow 0$
    $h \leftarrow r + 1$
    if $r == k$ then
        $h \leftarrow 1$
    end if
    for data $x$ in $C_r$ and $C_h$ do
        if $dis(x, w_r) \geq dis(x, w_h)$ then
        $n \leftarrow n + 1$
        end if
        if $n \geq \sigma$ then
            All data after the $\sigma$th data are classified to $C_h$.
            break
        end if
    end for
end for
Get the final clustering result $\{C_r'\}_{r=1}^{k}$.
$N \leftarrow NULL$
for each cluster $r = 1$ to $k$ do
    $noise = DBSCAN(Eps, Minpts, C_r')$
    $N \leftarrow N \cup noise$
end for
Output: abnormal sample N
```

1. Let ρ_{ij} represent the autocorrelation coefficient, where parameters i and j denote the i-th and j-th months, respectively. Calculate the similarity matrix $P(\rho_{ij})_{12\times12}$. Simultaneously, in order to eliminate the influence of varying month lengths, data for all months is restricted to only the first *l* days. Thus, the time series for the ith month can be represented as,

$$t_i = \{x_{i1}, x_{i2}, \dots, x_{im}\} \tag{4}$$

Where x_{im} represents the temperature data for the *m*th day of the ith month, and x_{jm} represents the temperature data for the mth day of the jth month $(1 \leq m \leq l)$. Let $\hat{\mu}_{ij}$ denote the mean value of the data for the ith and jth months. Then, the formula for calculating ρ_{ij} is,

$$\rho_{ij} = \frac{\sum_{m=1}^{l}(x_{im} - \hat{\mu}_{ij})(x_{jm} - \hat{\mu}_{ij})}{\sum_{m=1}^{l}[(x_{im} - \hat{\mu}_{ij})^2 + (x_{jm} - \hat{\mu}_{ij})^2]} \tag{5}$$

Then, the similarity matrix *P* is obtained.

2. Given the similarity matrix *P*, clustering of the data are performed based on the threshold τ. Since the dataset consists of time series, data within the same cluster exhibit continuity in time. Additionally, the pairwise autocorrelation coefficient ρ_{ij} between each pair of months satisfies $\rho_{ij} \geq \tau$. The resulting clustering is represented as $\{C_r\}_{r=1}^{k}$.

This result, clustered by months, is coarse. To achieve a more precise clustering based on dates, it requires further refinement using the K-means algorithm. Calculate the mean of the data within each cluster C_1, C_2 …, C_k separately, Obtain the corresponding centroids w_1,w_2 …w_k for the *k* clusters accordingly.

Due to the specificity of the dataset, the data within the same cluster should be consecutive in time. For instance, data from May and September should not be grouped together as one cluster. Therefore, unlike the traditional K-means algorithm, the data in cluster C_r can only consider the adjacent clusters C_{r-1} and C_{r+1C_r}, Specifically, when *r*=*k*, the adjacent centroids for cluster C_r are C_{r-1} and C_1. Traverse the data in C_r and its adjacent center C_{r+1} sequentially according to the time label, until the number of data satisfying the condition $dis(x, w_r) \geq dis(x, w_{r+1})$ reaches the threshold σ. Then, the data from the σ-th onwards will be classified into C_{r+1}, then we can obtain clustering results accurate to the date, denoted as $\{C_r'\}_{r=1}^{k}$. In this context, the function dis() computes distances and can be chosen based on the dataset. The paper utilizes the L2 norm metric for purpose.

3. Finally, the DBSCAN algorithm is employed to identify outliers within each cluster.

4. Experiment

This section provides a comprehensive experimental setup and clear validation results. In the dataset section, the reasons for selecting the dataset are explained, and temperature variations for Beijing and Sanya from January 1, 2023, to December 31, 2023, are provided. In the algorithm configuration section, a detailed introduction of the comparison algorithms is given. In the experimental section, it is demonstrated through tests on two datasets that the enhanced algorithm can detect more local anomalies under the same parameters.

4.1. Experimental Setup

4.1.1. Dataset Setup

The dataset used in this study is from the China meteorological data service center [12]. Considering the availability of data, the temperature data for the typical Chinese cities Beijing and Sanya in the year 2023 were selected. Beijing belongs to the warm temperate climate zone, with distinct seasons and significant temperature fluctuations. It shares similarities with climates in cities such as New York and Washington; Sanya belongs to the tropical maritime climate zone, with consistently high temperatures throughout the year and minimal temperature variations between seasons. It features stable temperature fluctuations and less distinct boundaries. Climates similar to Sanya can be found in places like the Maldives and Hawaii. As shown in Figure 2, this is the annual temperature curve for Beijing and Sanya in 2023.

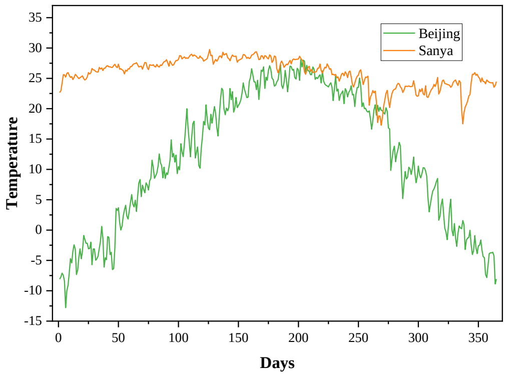


Figure 2. Beijing and Sanya's temperature variation curves in 2023.

To mitigate the impact of abrupt temperature changes on clustering results, the data from Beijing and Sanya for the years 2020 to 2023 were averaged, resulting in 365 data samples for each group (ignoring leap years) to be used for clustering.

4.1.2. Algorithm Configuration Description

In order to test the feasibility of the algorithms on the dataset, a total of three algorithms were selected for comparison in the experiment.

- **DBSCAN** algorithm does not require the prior determination of the number of clusters during clustering. It can discover clusters of arbitrary shapes and label outlier points. However, it is sensitive to global outliers while being less sensitive to local outliers.
- **A-DBSCAN** is an algorithm proposed in [16]. Its design principle involves grouping the dataset by adding extra attribute labels, overcoming the drawback of DBSCAN's inability to detect local anomalies. However, the algorithm lacks versatility since it requires manual label addition. Moreover, in situations where there is little variation in the data, there may be a problem of reduced grouping accuracy, thereby affecting the final results of anomaly detection.
- **In contrast, K-DBSCAN**, the algorithm proposed in this paper, aims to mitigate the issue of low data classification accuracy resulting from manual label addition. First, annotate and classify the dataset using autocorrelation coefficients; Then, combine with K-means to implement clustering on the dataset, and provide clustering results accurate to the day; Finally, Based on the clustering results, use the DBSCAN algorithm to identify outliers in the dataset. Through experimentation, it was found that K-DBSCAN can detect more local outliers compared to A-DBSCAN.

In the experiment, the parameter l=28 when calculating the autocorrelation coefficient ρ_{ij} for the data; The threshold for clustering based on the autocorrelation coefficient was set to τ=-0.4; The threshold for fine-grained partitioning using K-means is σ=8; The BSCAN algorithm employs two parameters, *Eps*=1.5 and *Minpts*=3.

4.2. Experimental Results

Since the DBSCAN algorithm treats the dataset as a whole without seasonal division, it does not appear in subsequent experiments. Only the total number of anomalies detected by the algorithm is analyzed in this section. The results of anomaly detection on temperature data for Beijing and Sanya in 2023 are shown in Figure 3.

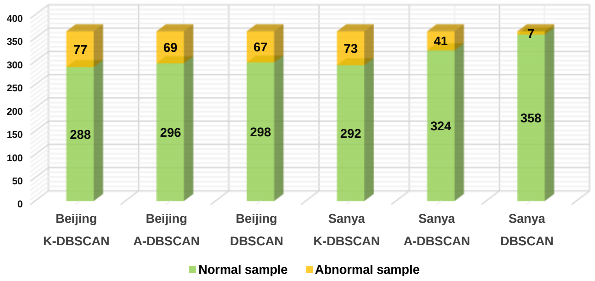
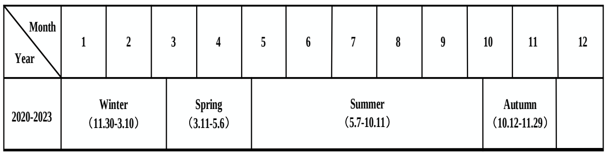


Figure 3. The total number of anomalies detected by the algorithm in the temperature dataset for Beijing and Sanya in 2023.

From Figure 3, it can be observed that in the temperature dataset for Beijing, K-DBSCAN detected 11.6% more anomalies compared to A-DBSCAN and 14.9% more anomalies compared to DBSCAN. In the dataset for Sanya, K-DBSCAN detected 78.0% more anomalies compared to A-DBSCAN and 942.6% more anomalies compared to DBSCAN. These two experiments demonstrate the importance of fine-grained partitioning of dataset and the feasibility of the improved algorithm.

4.2.1. Dataset with Volatile Temperature Fluctuations throughout the Entire Period

Beijing, located at 39°56'N and 116°20'E, experiences distinct seasons with brief springs and autumns but long winters and summers. Therefore, Beijing was chosen as a typical city with significant temperature fluctuations for algorithm validation in the experiment. Before conducting anomaly detection, it's necessary to cluster the dataset based on the historical data adjusted for mean values. After calculating the similarity matrix for the temperature data of Beijing, the data was clustered, and the clustering results are shown in Figure 4.



Month / Year	1	2	3	4	5	6	7	8	9	10	11	12
2020-2023	Winter (11.30-3.10)		Spring (3.11-5.6)		Summer (5.7-10.11)					Autumn (10.12-11.29)		

Figure 4. The clustering results of temperature data in Beijing.

The results show that the summer and winter seasons occupy 78.4% of the entire year, while spring and autumn account for 21.6%, which aligns well with the actual situation. Based on the clustering results, anomaly detection was conducted using algorithms, with A-DBSCAN and K-DBSCAN showing their respective detection outcomes in Figure 5.

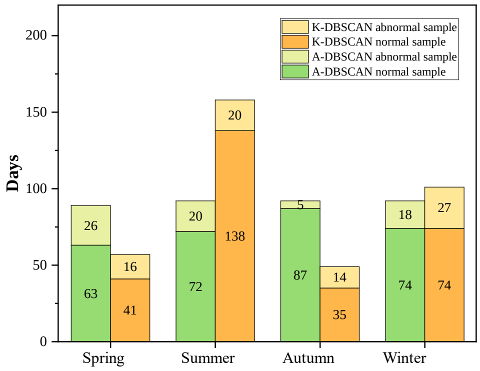


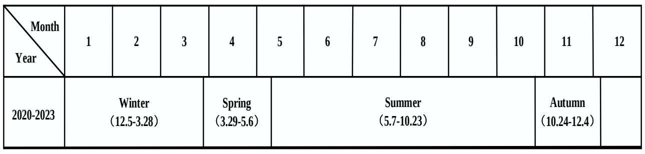
Figure 5. The results of anomaly detection on the temperature dataset of Beijing in 2023.

According to Figure 2, it can be observed that in 2023, the temperature fluctuations during spring, autumn, and winter in Beijing were more frequent and had larger amplitude compared to summer. Therefore, the proportion of abnormal temperature days during these three seasons is also higher than that of summer. In the detection results shown in Figure 5, the proportion of days with abnormal temperatures for each season is 29.2%, 21.7%, 5.4%, and 19.6% for A-DBSCAN, and 28.1%, 12.7%, 28.6%, and 26.7% for K-DBSCAN, respectively.

The comparison between Figures 1 and 4 reveals that the distinct difference between summer and autumn is due to A-DBSCAN categorizing summer temperatures into the autumn category during the clustering process. This coarse classification resulted in A-DBSCAN failing to detect some of the abnormal points in autumn.

4.2.2. Dataset with Gentle Temperature Fluctuations throughout the Entire Period

To validate the feasibility of the algorithm across different dataset, the experiment selected a city with climatic characteristics opposite to that of Beijing: Sanya. Sanya is located at 18°09'N and 108°56'E, with a climate characterized by perennial high temperatures and relatively gentle temperature fluctuations. As depicted in Figure 2, there is a stark contrast between the temperature curves of the two cities. The clustering results of temperature in Sanya in 2023 by K-DBSCAN are shown in Figure 6.



Month / Year	1	2	3	4	5	6	7	8	9	10	11	12
2020-2023	Winter (12.5-3.28)			Spring (3.29-5.6)	Summer (5.7-10.23)						Autumn (10.24-12.4)	

Figure 6. The clustered results of temperature data in Sanya.

Sanya's climate is characterized by perennial high temperatures, with long summers and winters and short spring and autumn seasons. From the results, it can be observed that the summer and winter seasons account for 77.8% of the entire year, while spring and autumn collectively occupy 22.2%. However, unlike Beijing, Sanya experiences smaller temperature variations between its four seasons, making it more challenging to manually add labels. According to the clustering results, anomaly detection was performed using different algorithms. The detection outcomes are illustrated in Figure 7.

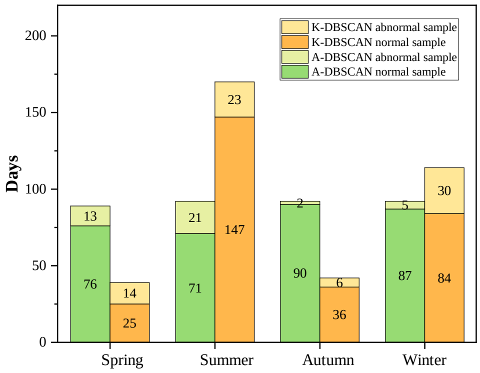


Figure 7. The results of anomaly detection on the temperature dataset of Sanya in 2023.

In 2023, similar to Beijing, Sanya experiences more frequent temperature fluctuations and larger amplitude during the spring, autumn, and winter seasons compared to the summer season, as shown in Figure 2. Consequently, the proportion of abnormal temperature days during these three seasons is also higher than in summer. The detection results show that summer accounts for 46.6% of the total number of days in the year. However, anomalies detected during the summer season only constitute 31.5% of the total anomalies detected throughout the year. In the detection results shown in Figure 7, the proportion of abnormal temperature days for A-DBSCAN across the four seasons are 14.6%, 22.8%, 2.2%, and 5.4%, respectively. For K-DBSCAN, the proportion of abnormal temperature days across the four seasons are 35.9%, 13.5%, 14.3%, and 26.3%, respectively. From the above results, it is evident that A-DBSCAN exhibits significant errors in detecting anomalies during the autumn and winter seasons.

The comparison between Figures 1 and 6 reveals that A-DBSCAN categorizes some summer weather as autumn and some spring temperatures as winter. Dividing data of different densities into the same category can indeed affect the detetion performance of DBSCAN. Comparing the two experiments, the anomalies detected by K-DBSCAN are consistently higher than those by A-DBSCAN and DBSCAN. The results further validate DBSCAN's deficiencies in detecting local anomalies, proving the necessity for finer data partitioning and demonstrating the feasibility of improved algorithms.

In summary, temperature data is a common use case for anomaly detection. This paper applies the proposed K-DBSCAN algorithm to actual temperature data from cities such as Beijing and Sanya, demonstrating its feasibility and providing support for related meteorological and climate research, with significant practical value.

5. Conclusions

DBSCAN is widely used in anomaly detection. Jain *et al*. [15] manually added extra labels to group the data, addressing DBSCAN's sensitivity to global anomalies and insensitivity to local anomalies. However, the applicability of this algorithm to dataset remains limited. Therefore, this study proposes enhancements based on this premise. Firstly, coarse-grained classification results based on months are obtained by calculating the autocorrelation coefficient. Next, the K-means algorithm is employed to achieve a finer-grained partition of the dataset, refining the clustering results from monthly to daily granularity. Finally, for each cluster, the DBSCAN algorithm is utilized for anomaly detection.

The improved algorithm shows a significant enhancement in anomaly detection performance, demonstrating the effectiveness of the proposed method. This could be valuable for researchers and practitioners working with time-series data where local anomalies are critical. For example, adjusting itineraries based on the frequency of unusual weather events, or studying the impact of the greenhouse effect on the frequency of such events, etc.

Building upon [15], this paper enhances the algorithm's applicability to additional dataset, but there are still directions for further research. For instance, when performing fine-grained clustering with K-means, the algorithm doesn't provide a clear method for selecting multiple thresholds; all thresholds are manually chosen. Selecting parameters like *Eps* and *Minpts* in DBSCAN for anomaly detection is vital. The algorithm should adaptively adjust clustering parameters based on the fluctuation patterns within the data. Matrix analysis [31] is a highly valuable method. We plan to apply autocorrelation matrices to financial datasets in the future to analyze market dynamics, thereby providing more precise insights for investment decisions and risk management.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant (No. 62471280). It is also supported by the Natural Science Basic Research Program of Shaanxi (No. 2020JM-509).

References

[1] Chandola V., Banerjee A., and Kumar V., “Anomaly Detection: A Survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009. https://doi.org/10.1145/1541880.154188

[2] Caudillo-Cos C., Montejano-Escamilla J., Tapia-McClung R., Ávila-Jiménez F., and Barrera-Alarcón I., “Defining Urban Boundaries through DBSCAN and Shannon's Entropy: The Case of the Mexican National Urban System,” *Cities*, vol. 149, pp. 104969, 2024. https://doi.org/10.1016/j.cities.2024.104969

[3] Dai Y., Sun S., and Che L., “Improved DBSCAN-Based Data Anomaly Detection Approach for Battery Energy Storage Stations,” *Journal of Physics: Conference Series*, vol. 2351, no. 1, pp. 012025, 2022. DOI:10.1088/1742-6596/2351/1/012025

[4] Ding X., Yu S., Wang M., Wang H., Gao H., and Yang D., “Anomaly Detection on Industrial Time Series Based on Correlation Analysis,” *Journal of Software*, vol. 31, no. 3, pp. 726-747, 2020. DOI:10.13328/j.cnki.jos.005907

[5] Ester M., Kriegel H., Sander J., and Xu X., “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” *in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portlands, pp. 226-231, 1996.

[6] Ghamkhar H., Ghazizadeh M., Mohajeri S., Moslehi I., and Khoshqalb E., “An Unsupervised Method to Exploit Low-Resolution Water Meter

Data for Detecting End-Users with Abnormal Consumption: Employing the DBSCAN and Time Series Complexity," *Sustainable Cities and Society*, vol. 94, pp. 104516, 2023. https://doi.org/10.1016/j.scs.2023.104516

[7] Gholizadeh N., Saadatfar H., and Hanafi N., "K-DBSCAN: An Improved DBSCAN Algorithm for Big Data," *The Journal of Supercomputing*, vol. 77, pp. 6214-6235, 2021.

[8] Habeeb R., Nasaruddin F., Gani A., Amanullah M., Hashem I., Ahmed E., and Imran M., "Clustering-Based Real-Time Anomaly Detection-A Breakthrough in Big Data Technologies," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 8, pp. 3647, 2019. DOI:10.1002/ett.3647

[9] Hilal W., Gadsden S., and Yawney J., "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," *Expert Systems with Applications*, vol. 193, pp. 116429, 2021. https://doi.org/10.1016/j.eswa.2021.116429

[10] Hawkins D., *Identification of Outliers*, Springer, 1980.

[11] Huang X., Wang Y., Li C., and Xu H., "Improved DBSCAN Algorithm Based Signal Recovery Technology in Coherent Optical Communication Systems," *Optics Communications*, vol. 521, pp. 128590, 2022. https://doi.org/10.1016/j.optcom.2022.128590

[12] He Q., Wang M., Liu K., Li K., and Jiang Z., "GPRChinaTemp1km: A High-Resolution Monthly Air Temperature Data Set for China (1951-2020) Based on Machine Learning," *Earth System Science Data*, vol. 14, no. 7, pp. 3273-3292, 2022. DOI:10.5194/essd-14-3273-2022

[13] Jain P., Quamer W., and Pamula R., "Electricity Consumption Forecasting Using Time Series Analysis," *in Proceedings of the International Conference on Advances in Computing and Data Sciences*, Dehradun, pp. 327-335, 2018. DOI:10.1007/978-981-13-1813-9_33

[14] Jin F., Wu H., Liu Y., Zhao J., and Wang W., "Varying-Scale HCA-DBSCAN-based Anomaly Detection Method for Multi-Dimensional Energy Data in Steel Industry," *Information Sciences*, vol. 647, pp. 119479, 2023. https://doi.org/10.1016/j.ins.2023.119479

[15] Jain P., Bajpai M., and Pamula R., "A Modified DBSCAN Algorithm for Anomaly Detection in Time-Series Data with Seasonality," *The International Arab Journal of Information Technology*, vol. 19, no. 1, pp. 23-28, 2022. https://doi.org/10.34028/iajit/19/1/3

[16] Kim S., Park S., and Chu W., "An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases," *in Proceedings of the 17th International Conference on Data Engineering*, Heidelberg, pp. 607-614, 2001. DOI:10.1109/ICDE.2001.914875

[17] Li G. and Jung J., "Deep Learning for Anomaly Detection in Multivariate Time Series: Approaches, Applications, and Challenges," *Information Fusion*, vol. 91, pp. 93-102, 2023. https://doi.org/10.1016/j.inffus.2022.10.008

[18] Li Z., Chen W., and Pei D., "Robust and Unsupervised KPI Anomaly Detection Based on Conditional Variational Autoencoder," *in Proceedings of the 37th International Performance Computing and Communications Conference*, Orlando, pp. 1-9, 2018. DOI:10.1109/PCCC.2018.8710885

[19] Liu W., Lei P., Xu D., and Zhu X., "Anomaly Recognition, Diagnosis and Prediction of Massive Data Flow Based on Time-GAN and DBSCAN for Power Dispatching Automation System," *Processes*, vol. 11, no. 9, pp. 2782-2799, 2023. https://doi.org/10.3390/pr11092782

[20] Lin S., Clark R., Birke R., Schonborn S., Trigoin N., and Roberts S., "Anomaly Detection for Time Series Using VAE-LSTM Hybrid Model," *International Conference on Acoustics, Speech and Signal Processing*, Barcelona, pp. 4322-4326, 2020. DOI:10.1109/ICASSP40776.2020.9053558

[21] Li J., Di S., Shen Y., and Chen L., "FluxEV: A Fast and Effective Unsupervised Framework for Time-Series Anomaly Detection," *in Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, New York, pp. 824-832, 2021. DOI:10.1145/3437963.3441823

[22] Liu H., Yao R., Cui C., and Zhao J., "A Data-Mining Interpretation Method of Pavement Dynamic Response Signal by Combining DBSCAN and Findpeaks Function," *Sensors*, vol. 24, no. 3, pp. 939, 2024. https://doi.org/10.3390/s24030939

[23] Loke S., MacDonald B., Parsons M., and Wunsche B., "Accelerated Superpixel Image Segmentation with a Parallelized DBSCAN Algorithm," *Journal of Real-Time Image Processing*, vol. 18, no. 6, pp. 2361-2376, 2021.

[24] Latha S., Samiappan D., Muthu P., and Kumar R., "Fully Automated Integrated Segmentation of Carotid Artery Ultrasound Images Using DBSCAN and Affinity Propagation," *Journal of Medical and Biological Engineering*, vol. 41, pp. 260-271, 2021.

[25] Mardani K. and Maghooli K., "Enhancing Retinal Blood Vessel Segmentation in Medical Images Using Combined Segmentation Modes Extracted by DBSCAN and Morphological Reconstruction," *Biomedical Signal Processing and Control*, vol. 69, pp. 102837, 2021. https://doi.org/10.1016/j.bspc.2021.102837

[26] Pei J., Zhong K., Jan M., and Li J., "Personalized Federated Learning Framework for Network Traffic Anomaly Detection," *Computer Networks*,

vol. 209, pp. 108906, 2022. https://doi.org/10.1016/j.comnet.2022.108906

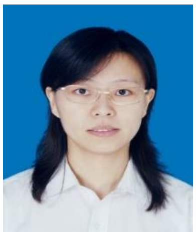
[27] Saba T., Rehman A., Sadad T., Hoshang K., and Bahaj S., "Anomaly-Based Intrusion Detection System for IoT Networks through Deep Learning Model," *Computers and Electrical Engineering*, vol. 99, no. C, pp. 107810, 2022. https://doi.org/10.1016/j.compeleceng.2022.107810

[28] Scitovski R. and Sabo K., "DBSCAN-Like Clustering Method for Various Data Densities," *Pattern Analysis and Applications*, vol. 23, no. 2, pp. 541-554, 2020.

[29] Su Y., Chen Z., Gong L., Xu X., and Yao Y., "An Improved Adaptive Radar Signal Sorting Algorithm Based on DBSCAN by a Novel CVI," *IEEE Access*, vol. 12, pp. 43139-43154, 2024. DOI: 10.1109/ACCESS.2024.3361221

[30] Wang H., He S., Liu T., Pang Y., Lin J., Liu Q., Han K., Wang J., and Jeon G.,"QRS Detection of ECG Signal Using U-Net and DBSCAN," *Multimedia Tools and Applications*, vol. 81, no. 10, pp. 13319-13333, 2022.

[31] Wang X., Guo Y., and Yang B., "Study on Suboptimal Diffusion Layer Based on Rotational-XOR Shifted Structure over Finite Domain$(F_2^8)^8$," *Journal of Shaanxi University of Science and Technology*, vol. 41, no. 4, pp. 188-194, 2023.



Xin Wang is an associate Professor at the School of Electronic Information and Artificial Intelligence at Shaanxi University of Science and Technology. She obtained her Doctor of Engineering Degree in the field of Communication and Information Systems from Xi'an University of Electronic Science and Technology. Her main research areas include Communication Coding and Data Analysis, as well as Cryptography and Privacy Protection.

Yingxue Yang is a Master's student at the School of Electronic Information and Artificial Intelligence at Shaanxi University of Science and Technology. Her major is Computer Technology, and her main research direction is Federated Learning and Anomaly Detection.

Xueshuang Ding is a Master's student at the School of Electronic Information and Artificial Intelligence at Shaanxi University of Science and Technology. Her major is Artificial Intelligence Technology, and her main research direction is the Optimization of Federated Learning Algorithms.

Yantao Zhao is an Associate Professor at the School of Electrical Engineering at Yanshan University. In 2022, he was a visiting scholar at Edinburgh Napier University in the United Kingdom. His main research areas include Process Industry Data Analysis and Processing, Intelligent Control, and Motion Analysis.