Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 416267, 12 pages
http://dx.doi.org/10.1155/2013/416267

Research Article

Detection of Outliers in a Time Series of Available Parking Spaces

Yanjie Ji,[1] Dounan Tang,[1] Weihong Guo,[2] Phil T. Blythe,[2] and Gang Ren[1]

[1] *School of Transportation, Southeast University, Nanjung 210096, China*
[2] *Transport Operations Research Group, School of Civil Engineering and Geosciences, Newcastle University, Newcastle upon Tyne NE1 7RU, UK*

Correspondence should be addressed to Yanjie Ji; jiyanjie@seu.edu.cn

Received 15 November 2012; Revised 4 February 2013; Accepted 18 February 2013

Academic Editor: Tsung-Chih Lin

With the provision of any source of real-time information, the timeliness and accuracy of the data provided are paramount to the effectiveness and success of the system and its acceptance by the users. In order to improve the accuracy and reliability of parking guidance systems (PGSs), the technique of outlier mining has been introduced for detecting and analysing outliers in available parking space (APS) datasets. To distinguish outlier features from the APS's overall periodic tendency, and to simultaneously identify the two types of outliers which naturally exist in APS datasets with intrinsically distinct statistical features, a two-phase detection method is proposed whereby an improved density-based detection algorithm named "local entropy based weighted outlier detection" (EWOD) is also incorporated. Real-world data from parking facilities in the City of Newcastle upon Tyne was used to test the hypothesis. Thereafter, experimental tests were carried out for a comparative study in which the outlier detection performances of the two-phase detection method, statistic-based method, and traditional density-based method were compared and contrasted. The results showed that the proposed method can identify two different kinds of outliers simultaneously and can give a high identifying accuracy of 100% and 92.7% for the first and second types of outliers, respectively.

1. Introduction

Along with the growth of urban populations, car ownership and car usage have continued to increase which has resulted not only in serious traffic congestion but also acute parking shortages. Cities have introduced intelligent transport systems (ITS) using a variety of different technologies and systems to address the imbalances between traffic demand and transport supply, among which, in many urban areas, parking guidance systems (PGSs) are an important component [1].

PGSs are mainly suitable for large- and medium-sized parking facilities and are widely used in government buildings, hospitals, railway stations, shopping malls, and other public parking lots. The main purpose of PGSs is to help drivers find spaces quickly, which can save time and energy as well as reducing congestion and emissions from cars searching for a parking space [2]. The guidance is usually based on real-time data but often augmented with a prediction of available parking space (APS) in future time slots which can provide a more flexible system for vehicles which receive information at the edges of an urban area and which may need to travel for some time to actually reach the desired parking destination [3]. The forecasting of unoccupied parking spaces has therefore attracted a strong research interest in recent years to improve the accuracy and utility of the PG information incorporating methods usually based on predicting the value of APS based on a series of past data samples at regular intervals, however, not all.

As a consequence of the increasing use of PGS in many cities, very large amounts of parking data are being gathered on a daily basis in many locations. Within these datasets, there is often incorrect or missing data due to factors, such as detector faults transmission distortion traffic accidents or a raft of other possible influencing factors, which may result in some of the data collected by the PGS being corrupted or generating abnormal data points that do not comply with the general behavior of the data model—these “outliers” are often mixed up with “normal” data. While variability and volatility are the main characteristics of the outliers, their existence will lead to low accuracy and efficiency of the APS prediction models that is more likely to arouse travelers’ doubts on the reliability of the PGS. Therefore, in order to improve the accuracy of parking guidance information, detecting outliers is a major priority in APS forecasting modelling. This can be decomposed into a time series outlier identifying problem, where one should analyse the overall periodic data features before performing some form of outlier detection. The most frequently used methodologies for this include the generalised extreme studentized deviate method (ESD), distance-based methods, and density-based methods, which can be classified into two main types: parametric and nonparametric methods.

Statistical parametric methods enjoy the advantages of being a low-time consuming and a relatively straightforward computational procedure. In 1983, Rosner proposed the ESD method to detect outliers in a univariate data set that follows an approximately normal distribution, where its primary limitation is that the suspected number of outliers, k, must be specified exactly which is not always practical [4]. Paul and Fung [5] improved the selection subjectivity of the parameter k in the ESD method and proposed the generalised extreme studentized deviate residual (GESR) method that assumes a known underlying distribution of the observations or is based on statistical estimates of unknown distribution parameters [5]. Zhang et al. [6] mentioned a fixed-sized time window-based outlier detection (FTWOD) technique. Each update step in this technique requires adding to the previous n measurements and then removing the oldest n measurements from the sliding window, where the outlier was detected using statistic values such as variance, first and third quartile [6]. However, they are unsuitable for high-dimensional datasets along with arbitrary datasets which lack prior knowledge of the underlying data distribution.

Exploring nonparametric outlier detection methods, Knorr and Ng proposed distance-based methods in the 1990s which were based on measures of local distance and were capable of handling large databases by considering data items as points in a high dimensional space [7] (Knorr). Outliers are defined as the distance is greater than a subjectively chosen threshold. However, distance-based methods have low efficiency for large datasets in high dimensional space and cannot recognise the local anomaly. In 1999, Breunig published the local outlier factor (LOF) algorithm which combines the distance between each record and the number of records in a given range [8, 9]. In this way, the concept of “density” was defined by which outliers could be identified. Later, a deviation-based outlier detection method was proposed by Sarawagi who introduced the discovery of drivable anomaly detection algorithms based on the offset by online analytical processing (OLAP) data cubes [10]. Evolutionary methods in outlier detection have been studied systematically in recent years. Banerjee presented a novel density-based distance measure and an outlier detection method using evolutionary search in his paper, where the methodology is tested on artificial datasets of varying sizes and dimensionalities [11]. Gupta further combined community matching with the evolutionary method, where experimental results on both synthetic and real datasets show that the proposed approach is highly effective in discovering interesting evolutionary community outliers [12].

Outlier detection is a very broad field and has been studied in the context of a large number of application domains where many detection methods have been applied according to the different data characteristics. Recently, there has been significant interest in detecting outliers in time series. Traditional time series literature defines two types of outliers (type I/additive and type II/innovative) based on the data associated with an individual object across time, ignoring the community aspect completely [13]. In 2001, Pena proposed that the detection of additive outliers played a more important role in time series prediction [14]. In this paper, we focus on the detection of additive outliers and further classify them into two types according to long-term observation of APS time series. Statistics-based methods are still conducted to detect outliers in time series because of their efficient structure. In 2012, Zhang developed an average-based methodology based on time-series analysis and geostatistics, which achieved satisfactory detection results in short snapshots [15] However, it will fail if the outliers gather together closely in the same short time slot. Hence recent researches have mainly focused on nonparametric outlier detection methods, such as Bayesian method and discrete wavelet transform (DWT). Frieda proposed a Bayesian approach to outlier modelling, approximating the posterior distribution of the model parameters by application of a componentwise Metropolis-Hastings algorithm [16]. Apart from the Bayesian method, discrete wavelet transform (DWT) in outlier detection has appeared in diverse application domains, such as manufacturing [17, 18], disease outbreak detection [19], and anomalies in computer networks [20, 21]. Their detections of outliers are mainly based on wavelet coefficients and setting up certain thresholds. For example, Bilen and Huzurbazar proposed an outlier detection procedure that used the discrete wavelet transform (DWT) of the original time series to detect jumps in the wavelet coefficients by using thresholds. Their method was compared with several parametric methods based on illusion data and proved to be more accurate [22]. Struzik presents a method of detecting and localising outliers in financial time series, which combined wavelet transform and multifractal formalism. If implemented on the wavelet transform modulus maxima tree, outliers can then be removed one by one with the possibility of dynamic verification of spectral properties [23]. Also based on wavelet transform, GranÈ and Veiga identified the outliers as those observations in the original series whose detail coefficients are greater than a certain threshold. They iterated the process of DWT and outlier correction until all detail coefficients are lower than the threshold. Based on real-world financial time series, their method achieved a lower average number of false outliers than Bilen and Huzurbazar’s [24]. In these works, thresholds are mainly set subjectively, which makes these methods inefficient and insensitive when there are several different kinds of outliers appearing in the same time series. Chaovalit suggested applying DWT in time series clustering processes to group similar time series data together into the same clusters and put dissimilar time series into different clusters, so that time series that contain outliers can be distinguished. However, Chaovalit has just reviewed several time series clustering methods and did not put their idea into practice [25].

In all these research and methods cited above, the majority of researchers have mentioned that density-based method cannot detect temporal changes. However, we believe density-based methods can play a powerful role in detecting outliers in time series by combining with wavelet transformation, the rationale, method, and experimental results will be described in the remaining part of this paper. To complete the landscape we introduce, the four issues associated with the PGS APS outliers problem which are seldom researched and merit further study.

(i) *Dealing with the Periodic Characteristics of APS Time Series*. Most previous research, especially that proposing nonparametric detection methods, has paid much attention to algorithm yet little to how to alter the optimal method for different kinds of dataset. When dealing with APS time series, directly applying detection method to the original dataset can cause false detection among peak and valley values.

(ii) *Allocating Reasonable Weights among Different Dimensions*. A method for dealing with multivariate datasets has been studied but as different dimensions may contain diverse quantities of information they need to be allocated different weights in a density-based method.

(iii) *Efficiently Identifying Outliers in Massive Datasets*. In today's APS data base, massive amounts of data are available for PGS study which may contain millions of data points. This results in the process of density-based detection being severely time consuming and inefficient because all data points must be visited to find the few outliers. Therefore, an efficient detection method should be developed to reduce process time by preliminarily locating outliers to a rather short time frame before conducting density-based method.

(iv) *Setting a Reasonable Threshold*. Thresholds are mainly set subjectively in formal studies, which means that many trials should be conducted until the optimal threshold is found. This is both time consuming and lacks theoretical basis.

With the aim of developing rational answers to the above questions, a number of proposed solutions have been examined and presented in this paper as follows. First, the changing features of APS time series and its outliers were analysed and a two-phase detection method was proposed by introducing discrete wavelet transform (DWT) to separate the detailed information, which contains the outlier properties, from the trend information which displays the periodic feature of the APS time series. Second, by mining statistical features of detailed signals, datasets were classified using two-step clustering algorithm into two classes: datasets with suspicious outliers and those without. The local entropy based weighted outlier detection algorithm (EWOD) that integrates all the transformed signals which are allocated with rational weight based on local entropy was then proposed to calculate outlier degrees for all data points in suspect datasets found in phase one. Finally, outliers were detected by clustering outlier degrees into several classes where the class far from the data centre is considered consisting of outliers.

To present this logically, the remainder of this paper has been organised as follows. Section 2 analyses the changing characteristics of APS and statistical features of its outliers. Thereafter, in Section 3 the DWT and EWOD algorithms are presented to conduct detail signals separation and calculation of outlier degrees while different signals have been allocated with reasonable weights. Subsequently, a case study of detecting both two kinds of outliers is presented based on real-world APS data. Following the case study, Section 5 presents a comparison of the EWOD method with the traditional parametric methods and nonparametric methods with the comparative analysis mainly quantifying levels of accuracy of the two methods. Finally, conclusions and recommendations for future investigation are discussed.

2. Features of the APS Dataset and Its Outliers

Before introducing the detection methodology, we first analysed the changing characteristics of APS. This is a crucial step with respect to which is the most appropriate detection algorithm to use; moreover, to judge this it is also important to have a suitable and representative datasets to test it with. With respect to the data set, we selected data from June 2011 to January 2012, which was collected from parking events recorded at the Eldon Multi-Story Car Park in Newcastle upon Tyne, UK, which forms part of a set of wider data sensing and collection projects underway in the city [26]. In the case of this research papers traffic flows and exit/entry events to and from the car park are recorded by the Tyne and Wear Urban Traffic Management and Control (UTMC) Centre which is located within Newcastle University. Observation of the car park operation and a review of the data derived from their PGS and parking management system indicated that data anomalies and outliers did exist in this data set and hence it was a reasonably representative sample of parking data that required additional processing techniques to improve it for PGI purposes. During the seven-month data-collection period, the data was recorded at an interval of 30 seconds and was collected during the opening hours of the car park, which was from 06:00 am to 22:00 pm. The capacity of this car park was 597 and the APS mentioned below represents the ratio of available parking space to capacity. The objectives of the analysis on this data were to identify and reveal the changing characteristic of APS time series and the statistical properties of its outliers.

2.1. Changing Characteristic of APS.

Choosing the data from July 28 (Thurs.), Aug. 1 (Mon.), Aug. 2 (Tues.), Aug. 3 (Wed.), and Aug. 4 (Thurs.) in 2011, the time interval was 30 seconds (to clarify, for some days we do not have overall APS data from 06:00 am to 22:00 pm, *e.g.*, for July 28 only the data

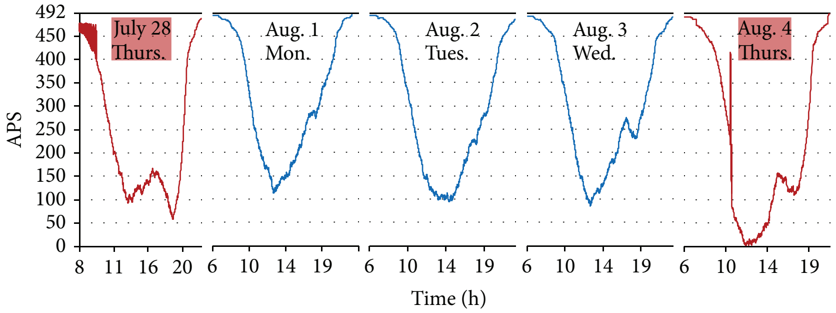


Figure 1: APS curve of July 28, Aug. 1, Aug. 2, Aug. 3, and Aug. 4 in 2011.

from 08:00 am to 22:00 pm are recorded). These five days are specifically chosen as not only are they successive days containing abundant data items for analyses, but they also contain outliers and normal data points at the meantime. The APS curve of these days is shown in Figure 1.

It can be clearly seen that during these days, the APS change tendencies are approximately similar while on some weekdays, such as on July 28, 2011, due to different social events taking place around Eldon Multi, the APS there shows a slight diversity with a largish APS around 19:00. Moreover, the rough curve also indicates a strong randomness along with natural variability especially when APS was low.

Because of the periodic changing characteristic, the outlier detection algorithm was able to extract the feature of outliers from the overall periodic data tendency and not falsely detect the peak or valley values as outliers. In addition, as a rather strong fluctuation occurs almost everywhere on the APS curve, the underpinning cause, "stationary noise," is also a prerequisite part in the detection algorithm.

2.2. Features of Outliers in APS Time Series. All real-life stochastic signals, including APS, are subject to contamination by noise. This can be relatively low level random noise or some systematic bias which can be observed on the blue curves in Figure 1. But there also exists a sudden jump from the current value of the regular information, which we call the first kind of outliers in this paper. In addition, outliers may also have high amplitudes and generally there will be relatively a few of them gathered together, which we call the second kind of outliers. The first and second kinds of outliers can be seen in Figure 1 on the red curve of July 28 and August 4. Their detailed information is shown in Figures 2(a) and 2(b) respectively, where the outliers are coloured red while blue represents normal data. Hence, the main difference between the stationary noise and an outlier is the inherently isolated and local character of the outlier. Therefore, an accurate method is able not only to recognise precisely the isolated feature of outliers from the normal data but also to correctly detect two different kinds of outliers simultaneously as well. However, it is a rather hard task as these outliers exhibit intrinsically distinct statistical features and they may occur in the same short time frame in another words are very close to each other. Thus statistical parametric methods may no longer be an effective way to identify outliers in APS time series and we should turn to nonparametric methods for help.

To sum up, the analysis of the features of APS time series and its outliers demonstrates that an accurate and efficient outlier detection method must have the following advantages: being able to distinguish outlier features from the overall periodic tendency, tolerate natural variability, and detect the two kinds of outliers simultaneously.

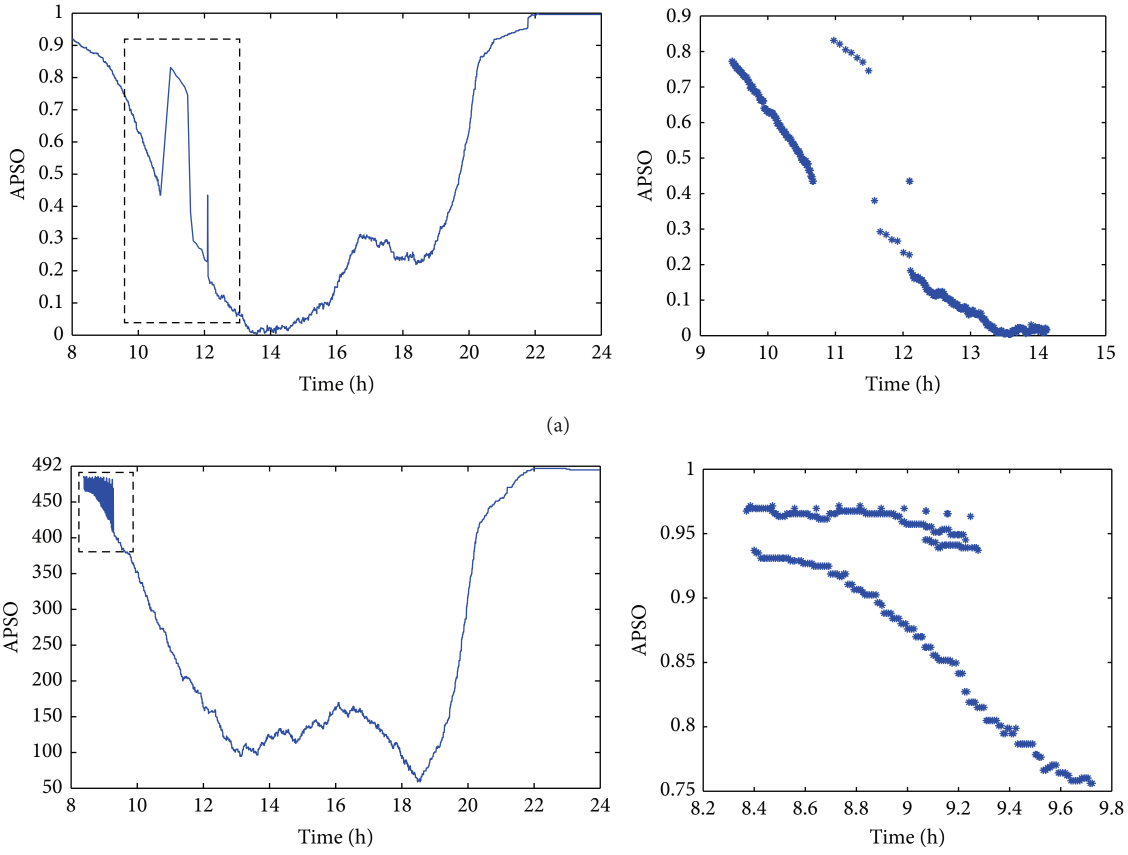
3. Methodology

After studying the features of APS time series and its outliers, we proposed the following solutions to problems described in Section 2 as follows: applying discrete wavelet transform (DWT) technology to extract outlier features from the periodic APS tendency and also utilizing a new nonparametric detection method named local entropy based weighted outlier detection algorithm (EWOD) to identify outliers—based on the outlier features extracted by DWT. The case study described towards the end of this paper shows that EWOD is reliably able to tolerate stationary noise and detect two kinds of outliers simultaneously.

In order to reduce the number of distance and density calculations between objects during the execution of our EWOD algorithm, a two-phase detection process was proposed and operated as follows.

Phase 1. The collected APS time series, containing several months' APS data, were transformed by DWT into approximate signals to reveal the overall periodic tendency and detailed signals, which includes outlier features. Thereafter, utilising a two-step clustering algorithm based on statistical features of everyday detail signals, and the analysed days were clustered into two categories: normal days, where their APS datasets did not contain outliers, and suspect days, where there is a high probability for outliers to occur. Suspect days needed further detection to locate the exact locations of outliers.

Phase 2. Considering all the detail signals, the EWOD algorithm was applied to these suspect datasets where several outlier factor vectors were calculated according to different k values in EWOD. Based on these outlier factor vectors, the two-step clustering algorithm was applied again to cluster



(a)

(b)

Figure 2: (a) APS curve of the first kind outliers (August 4, 2011). (b) APS curve of the second kind of outliers (July 7, 2011).

data points in the dataset into different categories consisting of points with the same outlier degree. Categories with rather high outlier degrees were considered as outlier categories and hence the exact locations of outliers could be detected by seeking every point in outlier categories.

In this section, two-phase outlier detection procedures are introduced including discrete wavelet transform algorithm and local entropy based weighted outlier detection algorithm that consists of local entropy theory and traditional density-based outlier detection method.

3.1. Discrete Wavelet Transform. Conceptually, the wavelet transformation [27] is a convolution product of the time series with the scaled and translated kernel—the wavelet function, usually an nth derivative of a smoothing kernel. Generally, there are two types of wavelet transformation: continuous wavelet transformation (CWT) and discrete wavelet transformation (DWT).

The CWT can be defined by

$$\mathrm{CWT}(a,\tau)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty}x(t)\,\psi^{*}\left(\frac{t-\tau}{a}\right)dt, \qquad (1)$$

where a represents the scale parameter, τ represents the translation parameter, ψ represents the "mother" wavelet, and ψ^{*} is the complex conjugate of ψ.

Unfortunately, as the CWT analysis technique is using different scales to analyse the time series, the time series must be calculated to obtain the wavelet coefficients with every possible scale. For this reason, the time series would generate a large quantity of computation wavelet coefficients. Therefore, the a and τ adopt the dyadic scale and translation to reduce computation wavelet coefficients and computation time. The dyadic method is the so-called DWT [28], which can be defined as

$$\mathrm{DWT}(a,\tau)=\frac{1}{\sqrt{2^{j}}}\int_{-\infty}^{\infty}x(t)\,\psi^{*}\left(\frac{t-2^{j}k}{2^{j}}\right)dt, \qquad (2)$$

where 2^{j} and $2^{j}k$ represent the scale parameter and the translation parameter, respectively.

The DWT can be depicted as a filter concept with complementary filters that contain a high-pass filter and a low-pass filter, which is high frequency (D_j-details) and low frequency (A_j-approximations) wavelet coefficients, respectively [29]. The analysis process can be iterated decomposition, and the reconstruction process can be assembled back into the original signal without loss of information, which consists of upsampling and reconstruction filters, so that the signal $x(t)$ can be expressed as

$$x(t) = A_j + \sum_{j<N} D_j, \tag{3}$$

where A_j and D_j represent the approximation and the detail coefficients of the Nth level [30].

By applying DWT to time series, the structure of singularities can be revealed in the analysis of reconstructed detail signals. In this paper, we have chosen the "db2" as the wavelet function which has optimal performance in singularities detection, in another word outliers mining. Nevertheless, although the DWT can improve the drawbacks of CWT, it is still hard to find the outliers by vision or classifier. Because of the reconstruction process, it will produce a large data quantity of the signal features which are difficult to classify by simply setting up threshold levels. For this reason, local entropy based weighted outlier detection methods are proposed to increase the accuracy of outlier identifying.

3.2. Local Entropy Based Weighted Outlier Detection Algorithm. The results of DWT with Nth level are one trend signal which reveals the overall periodic feature, and N detail signals which reveal the structure of singularities that provide information for outlier detections. However, in former studies which detail signal to choose is always a confusing problem where the choice is directly related to the accuracy of outlier detection. In this paper we combine all signals together into a multivariate time series (MTS) with $N+1$ dimension which consists of overall information in the former univariate time series. As different signals containing different quantities of information signals are weighted based on local entropy before they are applied with the density-based outlier detection algorithm.

3.2.1. Local Entropy. In information theory, entropy is a measure of the uncertainty associated with a random variable. In this context, the term refers to the Shannon entropy which quantifies the expected value of the information contained in a message [31]. In this paper, following Shannon's definition of entropy, local entropy of MTS is defined and applied to weight the DWT signals.

We first define a small neighborhood $N_k(p)$, the entropy of which is referred to here as local entropy. Let D be an MTS dataset, k a parameter, and $p \in D$. The k-neighborhood $N_k(p)$ of p is defined as

$$N_k(p) = \{x \in D \mid \operatorname{dist}(p, x) \le k\_\operatorname{dist}(p)\}, \tag{4}$$

where $k\_\operatorname{dist}(p)$ denotes the k-distance of p, which is defined as the distance $d(p, o)$ between object p and object o, such that o is the kth nearest neighbor point of p.

Hence, the local entropy of p among dimension A_i can be defined as

$$\begin{aligned} \mathrm{LEA}_{A_i}(p) = &-\sum_{q \in N_k(p)} \frac{\operatorname{dist}(F_{A_i}(p), F_{A_i}(q)) - d_{\min}}{d_{\max} - d_{\min}} \\ &\cdot \log_2\left(\frac{\operatorname{dist}(F_{A_i}(p), F_{A_i}(q)) - d_{\min}}{d_{\max} - d_{\min}}\right), \end{aligned} \tag{5}$$

where $p \in D$, $A_i \in A$, and $F_{A_i}(p)$ represents p's mapping on A_i in D. $d_{\max}$ and $d_{\min}$ are calculated as follows:

$$\begin{aligned} d_{\max} &= \max\{\operatorname{dist}(F_{A_i}(p), F_{A_i}(q)) \mid q \in N_k(p)\}, \\ d_{\min} &= \min\{\operatorname{dist}(F_{A_i}(p), F_{A_i}(q)) \mid q \in N_k(p)\}. \end{aligned} \tag{6}$$

Local entropy is related to the variance of signals in the neighborhood. From (5), we can see that the local entropy is larger for a heterogeneous region but smaller for a homogeneous neighborhood. Hence, the signals which contain information of outliers will have larger local entropy values than those that do not.

After calculating p's local entropy in all dimensions, a local entropy vector $\mathrm{LEA}(p) = \{\mathrm{LEA}_{A_1}(p), \mathrm{LEA}_{A_2}(p), \ldots, \mathrm{LEA}_{A_{N+1}}(p)\}$ is achieved and the weight vector $w(p) = \{w_{A_1}(p), w_{A_2}(p), \ldots, w_{A_{N+1}}(p)\}$ is based on it as

$$w(p) = \frac{\mathrm{LEA}(p)}{\sum_{i=1}^{N+1} \mathrm{LEA}_{A_i}(p)}. \tag{7}$$

Notably, for every point $p \in D$, $w(p)$ is not the same. Thus a new weighted MTS can be obtained by calculating every number in the new dataset DW as $\mathrm{FW}_{A_i}(p) = w(p) * F_{A_i}(p)$, where $\mathrm{FW}_{A_i}(p)$ represents p's mapping on A_i in DW.

3.2.2. Density-Based Outlier Detection. Following the definition of k-distance and k-neighborhood, the weighted k-neighborhood $\mathrm{NW}_k(p)$ is introduced [32] as

$$\mathrm{NW}_k(p) = \{x \in \mathrm{DW} \mid \operatorname{dist}(p, x) \le k\_\operatorname{dist}(p)\}, \tag{8}$$

where $k\_\operatorname{dist}(p)$ denotes the k-distance of p, which is defined as the distance $d(p, o)$ between object p and object o, such that o is the kth nearest neighbor point of p in weighted dataset DW.

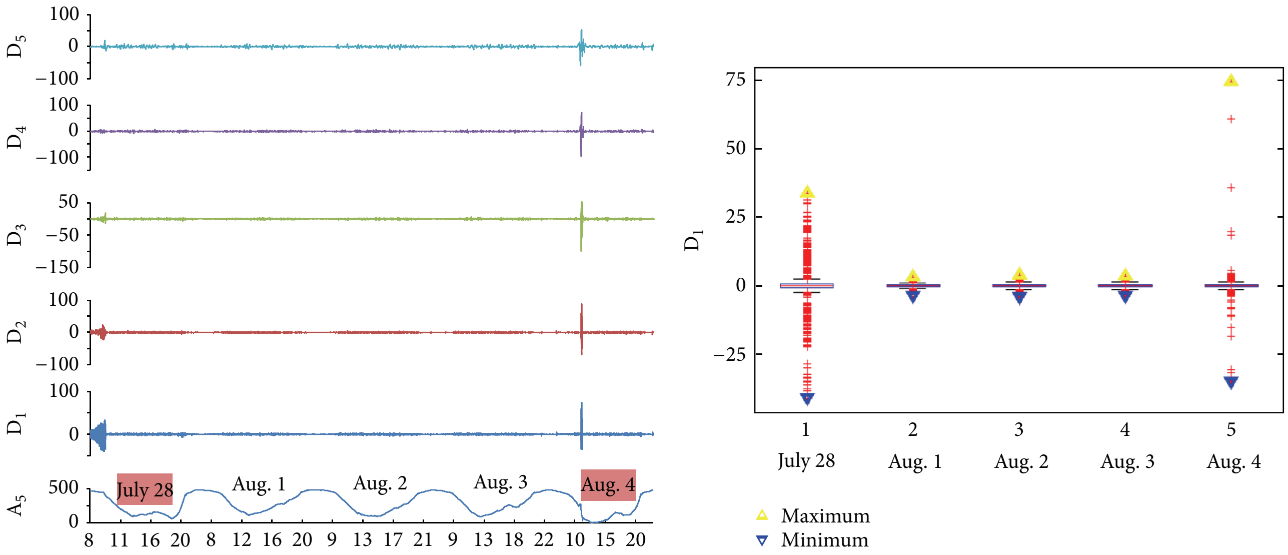
The local reachability distance of object p with respect to object $q \in D$ is defined as

$$\mathrm{reach}_{\mathrm{dist}_k(p,q)} = \max\{k_{\mathrm{dist(q)}}, \operatorname{dist}(p, q)\}. \tag{9}$$

The local reachability density of p is defined as the inverse of the reachability distance-based on the kth nearest neighborhood as

$$\mathrm{lrd}_k(p) = \frac{|\mathrm{NW}_k(p)|}{\sum_{q \in \mathrm{NW}_k(p)} \mathrm{reach\_dist}_k(p, q)}, \tag{10}$$

where $|\mathrm{NW}_k(p)|$ is the cardinality of $\mathrm{NW}_k(p)$.



(a)

(b)

Figure 3: (a) Approximate and detaild reconstructed signals of APS time series. (b) Boxplot of five days' D_1 signals.

The local outlier factor of p is defined as

$$\operatorname{lof}_k(p) = \frac{\sum_{q \in \mathrm{NW}_k(p)} \left(\operatorname{lrd}_k(q) / \operatorname{lrd}_k(p)\right)}{\left|\mathrm{NW}_k(p)\right|}. \tag{11}$$

As the average of the ratio of the reachability density of p and its kth nearest neighbors, the outlier factor of object p captures the degree of a point being an outlier. This notion assigns each object an outlier factor, in which being outlying is no longer a binary property.

4. Case Study

The objective of the case study was to examine the performance of the outlier detection method mentioned earlier to clean an APS time series by identifying erroneous or suspect observations that were likely caused by equipment failure or unusual traffic events.

4.1. Phase 1: Locating Outliers Roughly with DWT. Choosing APS datasets from July 28, 2011 and August 1, 2011 to August 4, 2011, which statistical features have been described in Section 2, the APS time series of these days were transformed into five levels with DWT technology. The reconstructed signals are shown in Figure 3(a).

It can be seen that all outliers had been removed in the A_5 signal which can be used to model short time APS predictions. With most detail information lost in A_5, we are still able to train the forecasting model for the detailed signal and predict error belonging to the same quantity order. However, simply applying DWT technology to clear outliers is an approximate way. An accurate clearing method can only be available before the locations of outliers are obtained. In the first phase, general locations of outliers are detected by analysing detail signals, such as $D_1, D_2, \ldots, D_5$. What can be directly seen in Figure 3(a) is that with the summation of the decomposition degree, the signs for the first kind of outliers are always clear while the signs for the second kind of outliers become less obvious from D_1 to D_5 because the second kind of outliers is regarded as noise in the DWT process. Thus, we choose the D_1 signal to roughly locate the outliers as the detail signal in this level already contains abundant information to detect both types of outliers.

In Phase 1, we analysed everyday D_1 signals by calculating their maxima, minima, and averages as shown in Figure 3(b). Then we applied a two-step clustering method to cluster these statistical values and classify APS datasets into two types: containing outliers or not containing outliers as shown in Figure 4(a).

Figure 4(a) shows that datasets with outliers present a dramatically distinct statistical feature from normal dataset, especially in aspects of maximal and minimal values. To illustrate a more practical situation, we selected 149 days' APS data by screening datasets from June 2011 to January 2012 and choosing the day when more than 2200 APS records were available. Based on their maximal D_1 values, minimal D_1 values and averages of D_1 signals' absolute values these 149 datasets were clustered into two classes: normal datasets coloured blue in Figure 4(a) and abnormal datasets that are suspected of containing outliers coloured green and red. To analyse the characteristics of outliers, abnormal datasets are further divided into two categories. Red datasets contain outliers with higher outlier degrees whose D_1 signals vary severely and have higher maximal as well as lower minimal values. Some green datasets do not even contain traditional outliers, such as APS in September 9, 2011 shown in Figure 4(b), where there is only a sudden APS increase around 16:00 as the data between 15:00 and 16:00 have been lost. This kind of abrupt changes in APS will lead to a severely variant signal in D_1 and also reveals instrument failures or

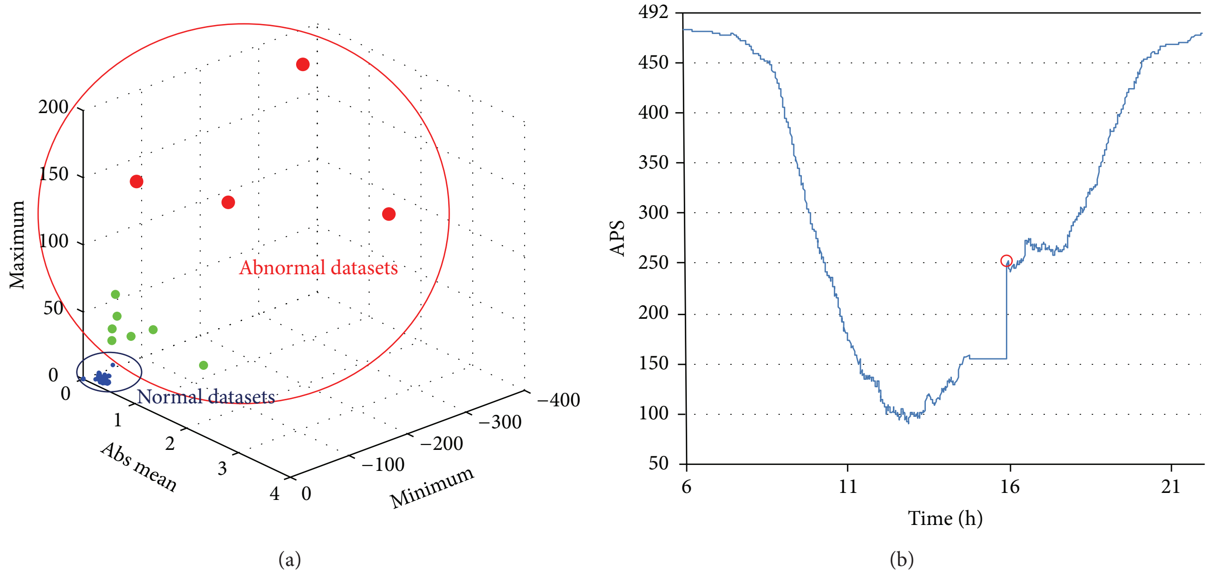


FIGURE 4: (a) Clustering result of 149 days' D_1 signals. (b) APS curve of september 9, 2011.

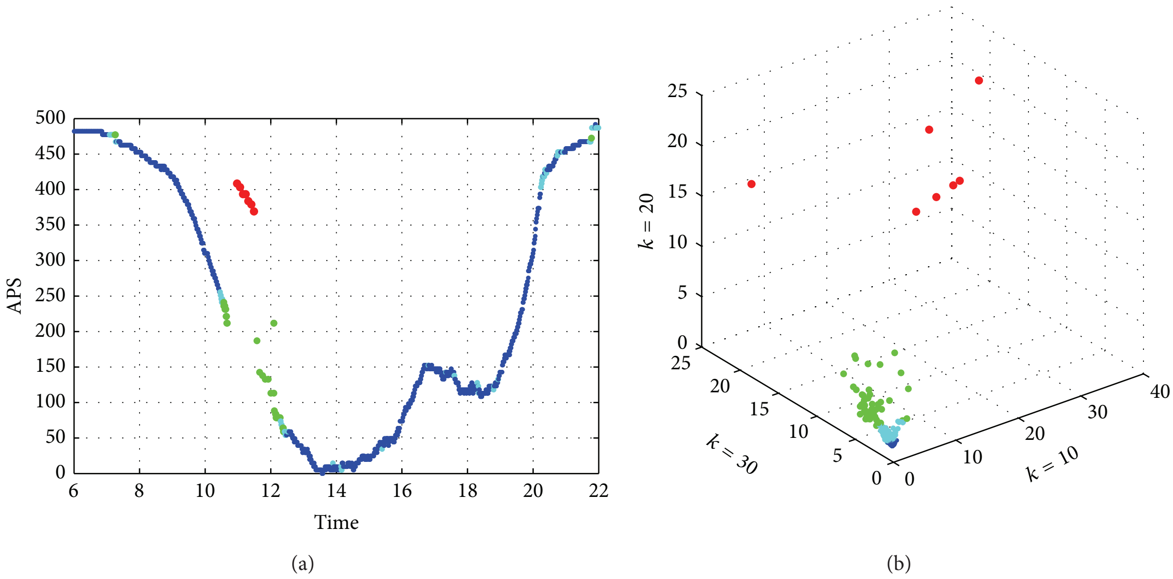


FIGURE 5: (a) Outlier detection result for August 4, 2011. (b) Outlier clustering result for August 4, 2011.

traffic incidences. Thus, we regarded them as outliers as well and classified them into first type of outliers.

4.2. Phase 2: Locating Outliers Exactly with EWOD. When Phase 1 was finished, the EWOD method was applied to suspicious datasets previously identified. Citing the dataset of August 4, 2011 as an example, outliers are detected after the APS time series is transformed by DWT as shown in Figure 3(b), the "Aug. 4" part. In DWT, one approximate signal A_5 and five detail signals $\{D_1, D_2, D_3, D_4, D_5\}$ constitute the final dataset $D = \{D_1, D_2, D_3, D_4, D_5\}$. Then the local outlier factor vector of this dataset can be calculated following the EWOD method.

In this procedure, different k values may lead to different detection results and to avoid this a two-step clustering methodology is put into use by classifying the dataset based on different k values' local outlier factor vectors. Specifically, the outliers of August 4, 2011's dataset are identified by first computing local outlier factor vectors when k equals 10, 20, ..., 50 and then clustering then into four types, found by experiment to be the optimal number of step to represent diverse outlier degrees as shown in Figure 5(a). To demonstrate vividly how datasets can be classified based on local outlier factor vectors, a three-dimensional image of local outlier factors when k equals 10, 20, and 30 along with the coloured clustering result is shown in Figure 5(b).

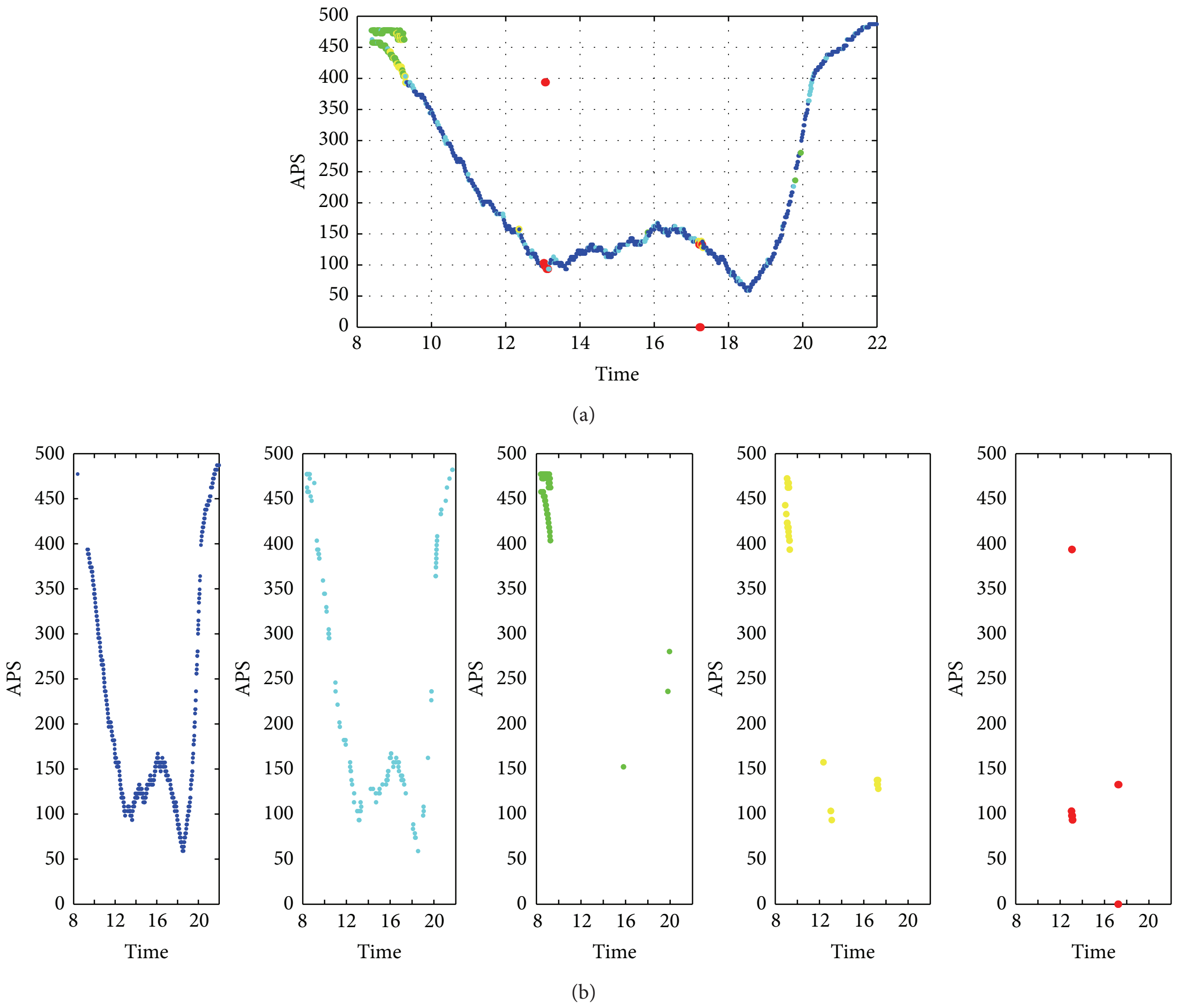


Figure 6: (a) Outlier detection result for processed August 4, 2011 dataset. (b) Outlier clustering result for processed August 4, 2011 dataset.

As illustrated in Figure 5(a), major outliers have been completely identified and coloured red; this pattern can also be observed in Figure 5(b) where these outliers are far away from other data items. There are also other outliers of the first kind, coloured green, detected as a result of the abrupt changes occurring around them, such as points near red outliers and points corresponding to ASPO sudden changes at approximately 7:00 am and 22:00 pm. Normal data are coloured in either dark or light blue with light blue showing a more deviated feature and often emerging near the inflection point of APS curve than the light blue.

To reveal the universality of our methodology, six artificial first kind outliers are added into the 13:00 pm and 17:15 pm data for July 28, 2011 to simulate the situation where both two kinds of outliers exist in the same dataset. DWT and EWOD are carried out on this processed dataset in succession after which the dataset is clustered into five classes basing on local outlier factor vectors where k equals 10, 20,. . ., 50. The outlier detection results are shown in Figure 6(a). Figure 6(b) is drawn to demonstrate the clustering results more clearly.

It can be observed from Figure 6 that first kind outliers have been clearly detected and coloured red while second kind of outliers are mainly included in the yellow and green classes where 92.7% (229/247) of second kind of outliers have been correctly identified. Yellow and green classes also illustrate abrupt changes in APS time serious while light blue class shows some other milder fluctuations.

5. Comparison Study

The case study revealed that our method, which combines DWT and EWOD in a two-phase procedure, gives an accurate and efficient performance in outlier detection and can provide detail outlier degree information for every data item. This section describes a comparison study between the statistical method, traditional density-based, and EWOD to show the superiority of the method proposed in this paper.

5.1. Statistical Methods Based on Sliding Time Window. The main idea of this statistical method is to use a sliding time window to extract the statistical characteristics of the data

within a certain time frame and use their statistic features to detect outliers [6]. Due to the application of the sliding time window, statistical characteristics can change with the fluctuation of data and thus reveal the periodic characteristic of APS time series. Specially, we counted out averages AVE along with standard deviations STD and set high and low thresholds as AVE + 2 ∗ STD and AVE − 2 ∗ STD, respectively. Data points with APS out of the threshold range were detected as outliers as shown in Figure 7.

Although a statistical method based on a sliding time window has the advantages of lower algorithm complexity and a simpler algorithm flow path, along with accurate performance in detecting first kind outliers, it fails to identify second kind outliers which gather closely in a certain length of time quantum for their outlier features cannot be directly made out by calculating the standard deviation in the time intervals they aggregated.

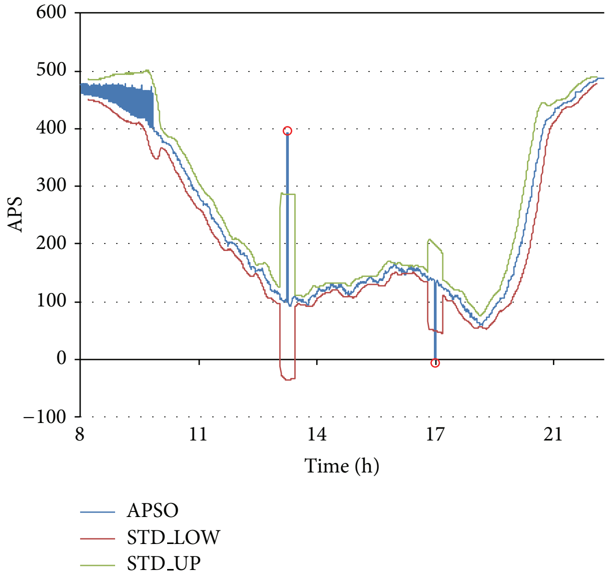


Figure 7: Outlier detection result of processed July 28, 2011 data using variance method based on sliding time window.

5.2. Traditional Density-Based Method and DWT-Based Method. Former studies applied the traditional density-based method or LOF algorithm [9], introduced in Section 3.2.2, directly to the original dataset and are not to be carried out upon time series for this method will mistake the date points of peak and valley values as outliers, which can be observed in Figure 8(a). This method can be slightly improved by first using DWT technique to decompose the original APS time series [24] and then carry out density-based outlier detection method on one of the detail signals whose detection result is displayed in Figure 8(b).

Experiments have shown that applying the density-based method on D_2 signal performs the best in outlier detection accuracy. However, it missed detecting one of the three first kind outliers around 17:15 pm and only identified 21.2% of the second kind of outliers. Although which detail signal to choose remains a conundrum, setting the threshold value is an even more difficult question in traditional density-based outlier detection methods. In conventional methods, as a data point's outlier degree increases when its local outlier factor increases, a threshold of outlier factor is set subjectively to judge whether a data point can be asserted as an outlier. However, this identifying method not only needs large numbers of trials to determine the final threshold, but it is also difficult to conduct when the dataset is mixed with both kinds of outliers because the threshold for these different kinds of outliers is inherently distinct, which can be solved by deploying this paper's methodology where outliers are automatically detected by a clustering algorithm.

6. Conclusion

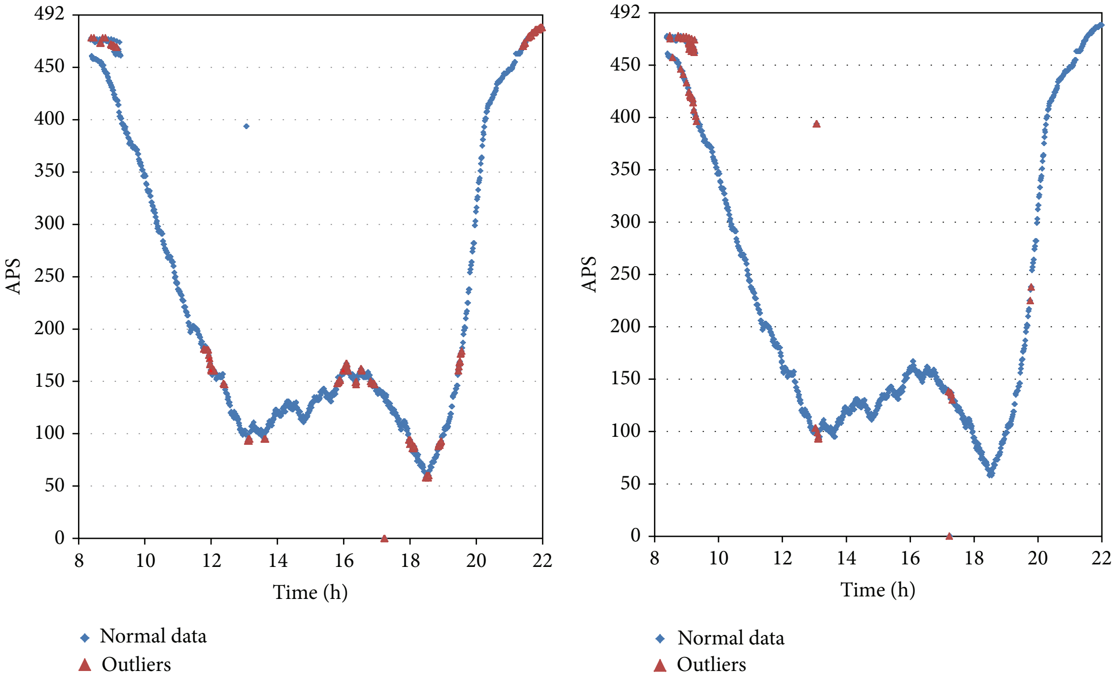
Outlier detection is a crucial research topic of parking guidance systems whose accuracy in identifying available spaces directly impacts the reliability of the forecasting process. In order to find an effective detection method for both two intrinsically different kinds of outliers in APS datasets, a two-phase detection method was developed in this paper and verified with observed APS data of Eldon Square Multi-Storey Car Park, Newcastle upon Tyne, UK made available through the Tyne and Wear UTMC Centre. In addition, its performance was compared with statistical-based and density-based outlier detection methods carried out using the same dataset. The main contributions of this paper are listed below.

(i) By conducting a two-phase outlier detection process, the efficiency of identifying outliers in massive data volumes has been significantly improved by locating outliers with DWT and exactly with the EWOD algorithm, as the time consumption of EWOD conducted in Phase 2 is obviously reduced for the time frame has been limited after Phase 1 was conducted.

(ii) In Phase 1 of the proposed two-phase method, outlier features are extracted out of the overall periodic tendency of APS time series by applying the DWT technique which solves the problem that peak and valley values are often false detected as outliers when directly carrying out outlier detection method on an original periodic time series dataset.

(iii) In Phase 2 of the proposed two-phase method, reasonable ways to weight different dimensions while detecting outliers in multivariate time series are proposed. Therefore, overall detail signals can be rationally used in the outlier detection procedure.

(iv) By calculating local outlier factor vectors for different k values, outliers can be clustered out based on these vectors instead of being judged by subjectively setting a threshold. Setting a threshold requires many trials and cannot readily take into account the distinct



(a)

(b)

Figure 8: (a) Outlier detection result when density-based method was applied to original APS time series. (b) Outlier detection result when density-based method was applied to D_2 signal.

features of two different kinds of outliers into consideration while simply identifying outliers by setting a threshold line.

The advantage of the two-phase outlier detection method has been verified by comparing its performance with statistical-based and traditional density-based method. When applied to the same dataset, the statistical-based method is not able to detect the second kind of outliers while the density-based method can detect both kinds of outliers simultaneously. However, identifying the accuracy of the traditional density-based method for the second kind of outliers is rather low, only 21.2%; moreover, how to decide the threshold of outlier degrees also remains a conundrum. Our method delivers a good performance in detecting outliers in APS time series as it enjoys an ability to identify two different kinds of outliers simultaneously, an efficient outlier distinguishing process, and an excellent accuracy of 100% and 92.7% for the first and second kind of outliers, respectively.

Moreover, because of the limitations of the survey data and the algorithm, several aspects of the research need to be improved in the future. When detecting the second kind of outliers, several normal data items have been falsely detected as outliers due to DWT's intrinsic sensitivity to singular points and the points around them may also be allocated high outlier factors. Moreover, the overall algorithm complexity can be reduced by improving the density-based outlier detection algorithm by introducing spatial index structures such as KD-trees, R-trees, or X-trees. Research is ongoing to address the above issues.

The Tyne and Wear Transport Authority is currently procuring a significant number of dynamic parking availability signs to integrate with its UTMC facility. An evaluation is underway to determine whether all or some of the data analysis techniques reported in this paper could be incorporated into the "real-world" scheme. This technique is also being considered for integration into the CVHS (Cooperative Vehicle and Highways Systems) project Compass4D which is funded by the European Commission under their 7th Framework Research Programme and includes Newcastle upon Tyne as a demonstration site.

Acknowledgments

The financial support from the National Key Basic Research Programme of China (no. 2012CB725402) and the National Natural Science Foundation of China (no. 50908051) is gratefully acknowledged. The authors would also like to thank the support received from the Transport Operations Research Group (School of Civil Engineering and Geosciences) at Newcastle University and from the Tyne and Wear Urban Traffic Management Control (UTMC) System.

References

[1] P. T. Blythe, "Using telematic tools to improve the management, operation and payment of car parks," in *Proceedings of the British Parking Association's Annual Congress*, Telford, UK, November 1997.

[2] F. Caicedo, C. Blazquez, and P. Miranda, "Prediction of parking space availability in real time," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7281–7290, 2012.

[3] Y. Ji, W. Guo, P. T. Blythe, W. Wanga, and W. Denga, "Design the parking guidance information for the drivers," in *Proceedings of the 18th World Congress on Intelligent Transport Systems*, Vienna, Austria, 2012.

[4] B. Rosner, "Percentage points for a generalized ESD many-outlier procedure," *Technometrics*, vol. 25, no. 2, pp. 165–172, 1983.

[5] S. R. Paul and K. Y. Fung, "A generalized extreme studentized residual multiple-outlier-detection procedure in linear regression," *Technometrics*, vol. 33, no. 3, pp. 339–348, 1991.

[6] Y. Zhang, N. Meratnia, and P. J. M. Havinga, "Ensuring high sensor data quality through use of online outlier detection techniques," *International Journal of Sensor Networks*, vol. 7, no. 3, pp. 141–151, 2010.

[7] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the 24rd International Conference on Very Large Data Bases(VLDB '98)*, New York, NY, USA, 1998.

[8] R. T. Ng and J. Han, "Efficient clustering methods for spatial data mining," in *Proceedings of the 20th International Conference on Very Large Data Bases (VLDP '94)*, pp. 144–155, 1994.

[9] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Optics of: Identifying Density-based Local Outliers," in *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge*, Springer, Prague, Czech Republic, 1999.

[10] H. Vuong, K. Shedden, Y. Liu, and D. M. Lubman, "Outlier-based differential expression analysis in proteomics studies," *Journal of Proteomics & Bioinformatics*, vol. 4, no. 6, pp. 116–122, 2011.

[11] A. Banerjee, "Density-based evolutionary outlier detection," in *Proceedings of the 14th International Conference on Genetic and Evolutionary Computation Conference*, pp. 651–652, 2012.

[12] M. Gupta, J. Gao, Y. Sun, and J. Han, "Integrating community matching and outlier detection for mining evolutionary community outliers," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 859–867, 2012.

[13] A. J. Fox, "Outliers in time series," *Journal of the Royal Statistical Society B*, vol. 34, pp. 350–363, 1972.

[14] F. J. Prieto and D. Pena, "Multivariate outlier detection and robust covariance matrix estimation," *Technometrics*, vol. 43, no. 3, pp. 286–310, 2001.

[15] Y. Zhanga, N. A. S. Hamm, N. Meratniaa, A. Steinb, M. van de Voorta, and P. J. M. Havingaa, "Statistics-based outlier detection for wireless sensor networks," *International Journal of Geographical Information Science*, vol. 26, no. 8, pp. 1373–1392, 2012.

[16] R. Frieda, I. Agueusopa, B. Bornkampb et al., "Bayesian outlier detection in INGARCH time series," Sonderforschungsbereich (SFB) 823, 2012.

[17] C.-S. Li, P. S. Yu, and V. Castelli, "MALM: a framework for mining sequence database at multiple abstraction levels," in *Proceedings of the 17th International Conference on Information and Knowledge Management*, Bethesda, Md, USA, 1998.

[18] Y. Yao, X. Li, and Z. Yuan, "Tool wear detection with fuzzy classification and wavelet fuzzy neural network," *International Journal of Machine Tools and Manufacture*, vol. 39, no. 10, pp. 1525–1538, 1999.

[19] W.-K. Wong, *Data Mining for Early Disease Outbreak Detection*, Carnegie Mellon University, Pittsburgh, Pa, USA, 2004.

[20] A. Magnaghi, T. Hamada, and T. Katsuyama, "A wavelet-based framework for proactive detection of network misconfigurations," in *Proceedings of ACM workshop on Network Troubleshooting: Research, Theory and Operations Practice Meet Malfunctioning Reality (SIGCOMM '04)*, pp. 253–258, Portland, Ore, USA, September 2004.

[21] P. Huang, A. Feldmann, and W. Willinger, "A non-intrusive, wavelet-based approach to detecting network performance problems," in *Proceedings of the 1st ACM SIGCOMM Internet Measurement Workshop (IMW '01)*, pp. 213–227, San Francisco, Calif, USA, November 2001.

[22] C. Bilen and S. Huzurbazar, "Wavelet-based detection of outliers in time series," *Journal of Computational and Graphical Statistics*, vol. 11, no. 2, pp. 311–327, 2002.

[23] Z. R. Struzik and A. P. J. M. Siebes, "Wavelet transform based multifractal formalism in outlier detection and localisation for financial time series," *Physica A*, vol. 309, no. 3-4, pp. 388–402, 2002.

[24] A. Grané and H. Veiga, "Wavelet-based detection of outliers in financial time series," *Computational Statistics & Data Analysis*, vol. 54, no. 11, pp. 2580–2593, 2010.

[25] P. Chaovalit, A. Gangopadhyay, G. Karabatis, and Z. Chen, "Discrete wavelet transform-based time series analysis and mining," *ACM Computing Surveys*, vol. 43, no. 2, pp. 1–37, 2011.

[26] S. Edwards, G. D. Evans, P. T. Blythe, D. Brennan, and K. Selvaraja, "Wireless technology applications to enhance traveller safety," *IET Intelligent Transport Systems*, vol. 6, no. 3, pp. 328–335, 2012.

[27] M. Holschneider, *Wavelets-An Analysis Tool*, Oxford Science Publications, Oxford, UK, 1995.

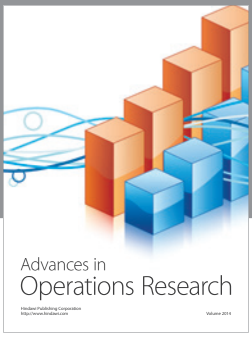
[28] S. G. Mallat, "A theory for multi-resolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

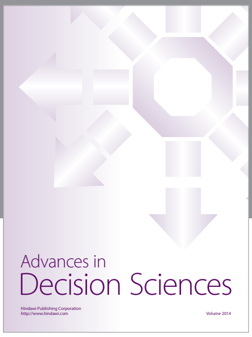
[29] C.-H. Hsia, J.-M. Guo, J.-S. Chiang, and C.-H. Lin, "A novel fast algorithm based on SMDWT for visual processing applications," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '09)*, pp. 762–765, May 2009.

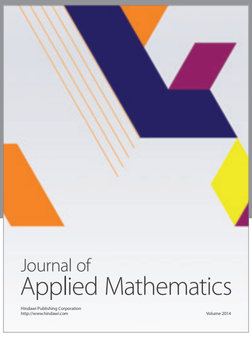
[30] W. S. Comanor, H. E. Frech III, and R. D. Miller Jr., "Is the United States an outlier in health care and health outcomes? A preliminary analysis," *International Journal of Health Care Finance and Economics*, vol. 6, no. 1, pp. 3–23, 2006.

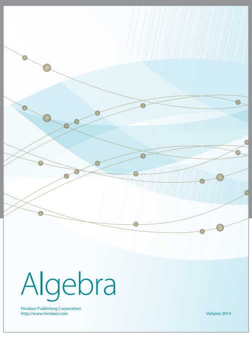
[31] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–324, 1948.

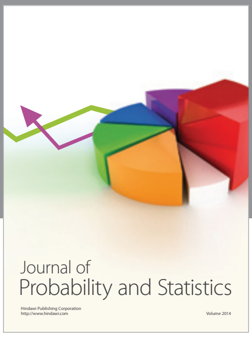
[32] X. J. Ban, Y. Li, and A. Skabardonis, "LOF: identifying density-based local outliers," in *Proceedings of the ACM SIGMOD international conference on management of data*, pp. 93–104, 2000.


Advances in Operations Research

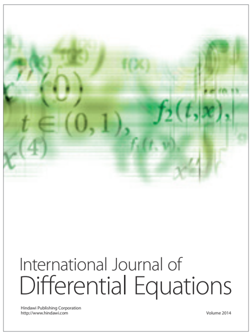

Advances in Decision Sciences


Journal of Applied Mathematics

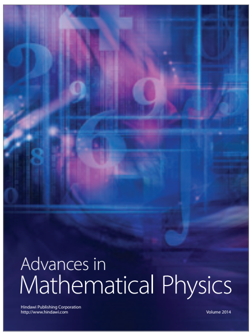

Algebra

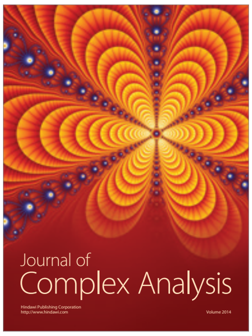

Journal of Probability and Statistics

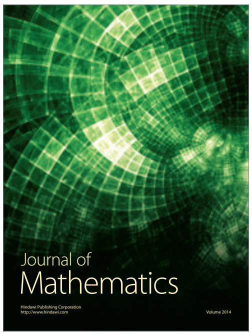
The Scientific World Journal

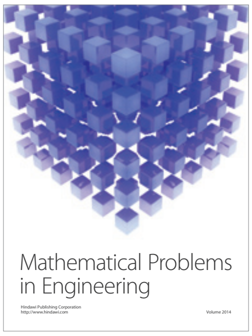

International Journal of Differential Equations

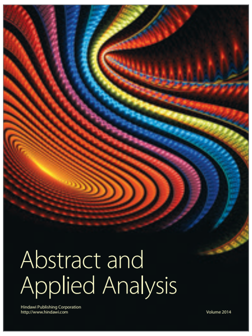
International Journal of Combinatorics

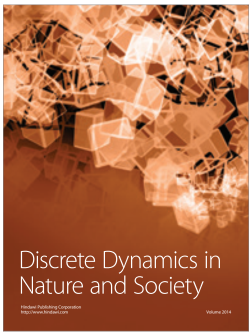

Advances in Mathematical Physics

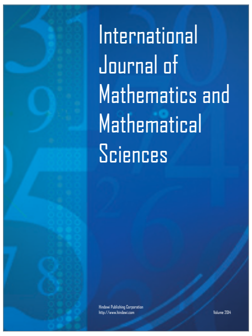

Journal of Complex Analysis


Journal of Mathematics


Mathematical Problems in Engineering

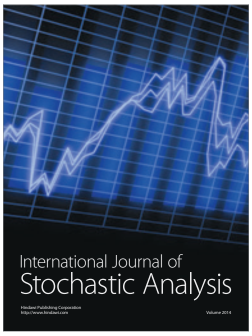

Abstract and Applied Analysis


Discrete Dynamics in Nature and Society


International Journal of Mathematics and Mathematical Sciences

Journal of Discrete Mathematics

Journal of Function Spaces


International Journal of Stochastic Analysis

Journal of Optimization