Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 7191549, 12 pages
https://doi.org/10.1155/2018/7191549

Research Article

An Improved Robust Principal Component Analysis Model for Anomalies Detection of Subway Passenger Flow

Xuehui Wang,[1] **Yong Zhang**,[1] **Hao Liu**,[2] **Yang Wang**,[1] **Lichun Wang**,[1] **and Baocai Yin**[1]

[1] *Beijing Advanced Innovation Center for Future Internet Technology, Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*
[2] *Beijing Transportation Information Center, Beijing, China*

Correspondence should be addressed to Yong Zhang; zhangyong2010@bjut.edu.cn

Received 16 May 2018; Accepted 31 July 2018; Published 14 August 2018

Academic Editor: Darius Andriukaitis

Subway is an important transportation means for residents, since it is always on schedule. However, some temporal management policies or unpredicted events may change passenger flow and then affect passengers requirement for punctuality. Thus, detecting anomaly event, mining its propagation law, and revealing its potential impact are important and helpful for improving management strategy; e.g., subway emergency management can predict flow change under the condition of knowing specific policy and estimate traffic impact brought by some big events such as vocal concerts and ball games. In this paper, we propose a novel anomalies detection method of subway passenger flow. In this method, an improved robust principal component analysis model is presented to detect anomalies; then ST-DBSCAN algorithm is used to group the station-level anomaly data on space-time dimensions to reveal the propagation law and potential impact of different anomaly events. The real flow data of Beijing subway are used for experiments. The experimental results show that the proposed method is effective for detecting anomalies of subway passenger flow in practices.

1. Introduction

Owing to the high efficiency and the comfort, subway has generally become first choice for citizens' daily travel, and it directly facilitates the city's economic development and people's quality of life. For example, as one of the busiest subway systems in the world, the Beijing subway has the world's largest annual ridership with 3.03 billion trips delivered in 2016, averaging 8.26 million per day, with peak single-day ridership reaching 10.52 million. The public transportation in Beijing accounts for 45% of total traffic, in which the ridership of subway dominates nearly 40%.

Although bringing great convenience for residents, the subway system becomes more vulnerable at the same time, as the subway system is a large and complicated network running in a restricted time schedule. For example, there are 22 lines and 370 stations in Beijing subway, and more than 500 trains are running on the network with the minimal peak headway in 90 seconds. This will be more critical in the cases of encountering exceptional events, such as station accidents, major activities, and bad weathers. Once a station has an anomaly event, such as failure operation and chaos in station, the retention of passengers would happen, which would bring great loss with high security risks. Moreover, the bad situation would propagate through the urban subway system since it is a relatively closed and connected network. So the impact of anomaly event will not be restricted in a specific station, it may affect the traffic system in a large region, and the influence of abnormal events usually shows a certain space-time law. Thus, it is necessary to detect anomalies in urban subway transportation system and figure out its spreading rules, which can provide valuable proof for management to making strategy for dealing with abnormal events.

However, in the traditional road transportation system, many methods have been proposed for detecting transit anomalies, such as the Automatic Incident Detection (AID)

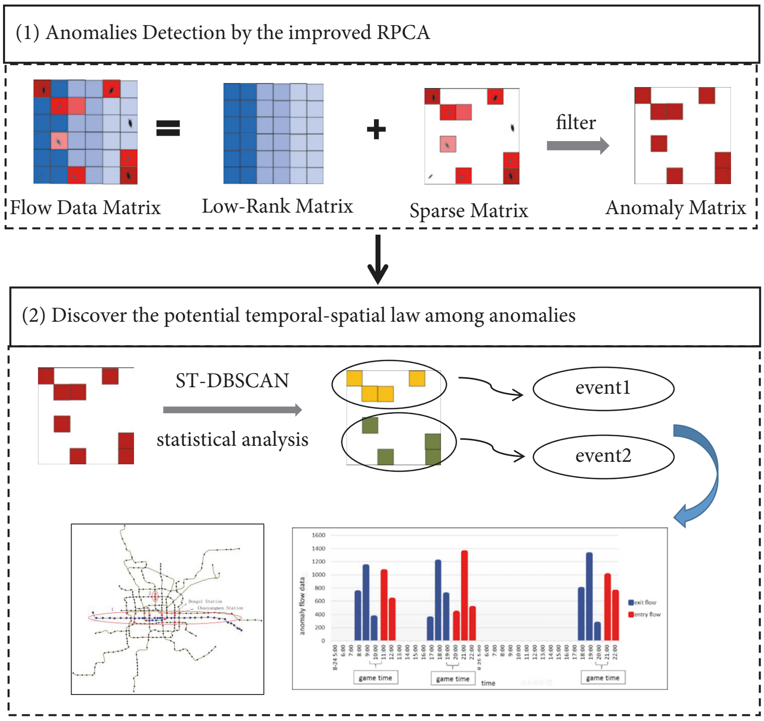


Figure 1: The whole procedure of anomalies detection by the improved RPCA. (1) Decompose the matrix of subway passenger flow into a low-rank matrix and a sparse matrix, filter the sparse matrix decomposed by the improved RPCA, and then acquire the anomaly matrix. (2) Mine the potential laws and discover the possible relations among anomalies through ST-DBSCAN algorithm and statistical analysis.

algorithms [1–3] based on comparison, statistics techniques, traffic flow model, and so on. These methods are mostly applied in freeway and urban roads, and they link the main regions of a city and try to find unexpected traffic flow between any two regions [4, 5]. As the subway is a different traffic system from the traditional road transportation, the above methods are difficult to introduce to subway system. At the present, few studies focus on the anomalies detection of passenger flow in subway transportation.

Since the subway anomaly events are always uncertain and sporadic, the anomalies show obvious sparse property among the whole subway traffic data. From this observation, based on the available subway passenger flow data collected by the AFC (Automatic Fare Collection) system, we propose a novel anomalies detection method based on Robust Principle Component Analysis (RPCA) model [6], which represents the temporal-spatial distribution of data and the sparsity of the anomalies by the low-rank and sparse regularization. Additionally, in order to reveal the propagation law and potential impact of different anomaly events, the ST-DBSCAN algorithm is adopted to group the station-level anomaly data on temporal-spatial dimensions. Thus the proposed method can not only detect anomaly of a single station but also find the relations among anomalies. Figure 1 shows the main structure of our model.

The main contributions of this paper are summarized as follows:

(i) A novel anomaly detection method of subway passenger flow based on RPCA is proposed, which utilizes low-rank nature of the passenger flow data and the sparsity of anomaly data. Experimental results demonstrate that our approach can achieve an accurate anomaly detection.

(ii) The ST-DBSCAN clustering algorithm is adopted to explore the temporal-spatial propagation law of anomalies, and the obtained expected results are verified by tweet data. The distribution law of anomalous flow caused by different anomaly events can provide prior information to cope with possible anomalies.

The rest of this paper is organized as follows. Related works are summarized in Section 2. Section 3 gives the methodology. Section 4 reports experimental results on real data and their visualization analysis. Finally, we conclude this paper in Section 5.

2. Related Works

In this section, we review the commonly used methods for anomalies detection in traffic systems and introduce RPCA model related to our work.

2.1. Anomalies Detection Methods of Passenger Flow. Most of the existing anomalies detection methods of traffic flow are in highways or urban roads scenarios, and the traffic data from fixed detectors is usually used for analysis. The typical approaches include the statistical methods, the comparison methods, and the traffic flow model based methods. The famous comparison algorithms are the California algorithm and its derivation [1], which discriminate the anomaly event by comparing traffic parameters between adjacent detectors. But they are not suitable for subway passenger flow because the relation between neighboring stations is not exactly similar with the relation between neighboring detectors. The statistical methods (like SND [2]) achieve traffic anomaly by judging change rate of the traffic parameters, and they adopt the threshold method [7] to identify unreasonable data values based on historical data. To these methods, the suitable thresholds are difficult to choose. The traffic flow model based methods (like McMaster algorithm [3]) define boundary between crowded traffic flow and noncrowded traffic flow to determine a speed threshold for distinguishing, which is not well for subway due to the difference of flow pattern while time and space scales change. In addition, the wavelet analysis [8] is used for detecting anomalous samples by separating the high frequency components and the low frequency components of traffic data.

For the transportation system of a city scale, the current studies on anomalies detection are mostly region-based. Pang and Linsey Xiaolin [4, 5] partition city into uniform grids and report anomalies if traffic volumes in neighboring cells are different, while Shekhar [9] focuses on detecting spatial outliers in graph structured datasets. Similarly, Liu [10] and Chawla [11] partition the city into disjoint regions linked by major roads and then find unexpected traffic flow between any two regions. However, the above methods are either road-based or region-based and the former cannot accurately identify location of events, and the latter may result in loss of information because of the coarse region partition.

There are few works concentrating on the anomalies detection of subway passenger flow. Some anomalies detections in subway system focus on the pedestrian abnormal activity inside the station [12, 13], and they generally adopt visual recognition techniques based on the video surveillance system in the station; thus the applied scale is valid only in the view of the cameras. Besides, the other studies on passenger flow data of subway mainly focus on passenger flow prediction and analysis [14, 15]. Differently, in this paper, we conduct anomalies detection of subway passenger flow and explore the temporal-spatial impact of anomaly events.

2.2. Robust Principal Component Analysis. Recently, due to the power of revealing the intrinsic structure or property underlying the data, the low-rank and sparse theory have been successfully applied in numerous areas such as image recovery and denoising [16], background modeling, and foreground object detection of video image [17]. RPCA is a typical model utilizing the low-rank and sparse matrix decomposition for data restoration and denoising. The basic idea is that the original data in form of a numerical matrix can be decomposed into a low-rank matrix and a sparse matrix as follows:

$$\begin{aligned} \min_{\mathbf{X},\mathbf{A}} \quad & \operatorname{rank}(\mathbf{X}) + \lambda_1 \|\mathbf{A}\|_0, \\ \text{s.t.} \quad & \mathbf{D} = \mathbf{X} + \mathbf{A} \end{aligned} \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{m\times n}$ is the raw data usually having noise, $\mathbf{X}$ represents the expected clean data which is assumed having low-rank property, and $\mathbf{A}$ represents the noise data or outlier which is considered being sparse. The target of RPCA in (1) is to estimate the unknown $\mathbf{X}$ and $\mathbf{A}$ given $\mathbf{D}$.

However, the optimization problem in (1) is a NP-hard problem [18] due to its nonconvexity and discontinuity. On one hand, the low-rank term should be processed properly. For this purpose, a widely used solving scheme is replacing rank ($\mathbf{X}$) by its convex envelope, nuclear norm $\|\cdot\|_*$ [6, 19], as nuclear norm minimization approaches can perform stably without knowing the target rank of the recovery matrix in advance. On the other hand, the nonconvexity and discreteness of the ℓ_0 penalty make it be not preferred. Considering that ℓ_1 is also good at modeling the sparse noise [6] and has high efficient solution, the ℓ_0 term in (1) is replaced with ℓ_1. Thus, (1) can be written as

$$\begin{aligned} \min_{\mathbf{X},\mathbf{A}} \quad & \|\mathbf{X}\|_* + \lambda_1 \|\mathbf{A}\|_1, \\ \text{s.t.} \quad & \mathbf{D} = \mathbf{X} + \mathbf{A} \end{aligned} \tag{2}$$

where $\|\mathbf{X}\|_* := \sum_i \sigma_i(\mathbf{X})$ denotes the nuclear norm, $\sigma_i(\mathbf{X})$ is the ith largest singular value of matrix $\mathbf{X}$, and $\|\mathbf{A}\|_1 := \sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|$, $a_{i,j}$ is the element of $\mathbf{A}$.

In this paper, we introduce RPCA into the anomaly detection of subway passenger flow. Moreover, the passenger flow data matrix has low-rank structure because it shows regular cycles with respect to day, week, month, and year. In addition, the real-world data is usually polluted by noise or outliers, and the outliers are considered anomalies for detection. So we adopt RPCA to represent the subway passenger flow and detect the anomalies by the sparse outliers. Additionally, we consider the temporal correlation among the data and propose an improved RPCA. The next section will give the improved RPCA in detail.

3. Methodology

In this section, we first represent the subway passenger flow as a matrix and give it decomposition for anomalies detection. Then the improved RPCA is applied to obtain preliminary abnormal flow information. Finally, the detected anomalies are grouped into several clusters for revealing the temporal-spatial laws.

3.1. Subway Passenger Flow Representation and Decomposition. The raw subway passenger riding data are collected from the subway AFC system; they include the boarding or alighting time at a station, the boarding line ID or alighting, and the boarding station ID or alighting. Based on the raw riding data, the subway passenger flow data are calculated in one hour interval, and then we obtain the subway passenger flow matrix $\mathbf{D}$, which is constructed with the row and column corresponds to the date and the time interval of each day, respectively. Therefore, each element in the matrix represents the passenger flow of a station at a certain time interval of a certain day.

As the passenger flow of a subway station shows similar varying degrees taking year, month, week, hour, or minute as a cycle, the temporal patterns of the passenger flow matrix **D** are typically a low-rank matrix [20]. Besides, passenger flow of adjacent stations also show certain similarity, which further supports the low-rank property of the passenger flow matrix. However when anomaly events happen, the low-rank property of the flow would be ruined by the outliers. So the matrix D can be considered as a combination of normal and outliers. Let $x_{e,t}$ and $a_{e,t}$ represent the expected flow component and the outlier interference of a station at time t on date e, so the measured passenger flow at time t can be expressed as $d_{e,t} = x_{e,t} + a_{e,t}$. Collecting n measurements and introducing matrices $\mathbf{X} := [x_{e,t}], \mathbf{A} := [a_{e,t}]$, the passenger flow matrix can be decomposed by

$$\mathbf{D} = \mathbf{X} + \mathbf{A} \tag{3}$$

By this decomposition, the subway passenger flow can be represented by two components: the expected flow **X** and the anomalous part **A**. The anomalous part **A** is explained as special events or special activities around the station; it is sporadic over time and may last for short periods relative to the (possibly long) measurement period **T**, which means that only a small fraction of the elements in observation traffic flow matrix **D** is supposed to be anomalous. Therefore, the anomaly matrix **A** would be sparse both in rows and columns.

From the above analysis, the subway passenger flow **D** completely has the RPCA model in (2) with the low-rank and sparse terms. In the following, we further exploit the temporal constraint for the model and propose our improved RPCA model for the anomalies detection of subway passenger flow data.

3.2. The Improved RPCA. For the subway passenger flow matrix **D**, the two adjacent rows of the same weekdays in different weeks are often approximately equal except some outliers, derived from the obvious day cycle of the passenger flow measurement. This property is conductively true for the corresponding expected flow **X**, while the current RPCA model has no specific description for this important property. So we propose a constraint to keep the consistence among rows of **X** by adding an item $\|\mathbf{HX}\|_1$ to the current RPCA model. The matrix **H** is defined as follows:

$$\mathbf{H} = \begin{pmatrix} 1 & -1 & 0\cdots 0 & 0 \\ 0 & 1 & -1\cdots 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0\cdots 1 & -1 \\ 0 & 0 & 0\cdots 0 & 1 \end{pmatrix}_{(m-1)\times m.} \tag{4}$$

The above temporal differential matrix is $\mathbf{H} = \text{Toeplitz}(0, 1, -1)$, in which the central diagonal is defined as ones and the first upper diagonal is defined as negative ones. The temporal constraint matrix intuitively expresses the fact that nominal passenger flow matrices at same time intervals for the same weekdays are usually similar. Actually, $\|\mathbf{HX}\|_1$ captures consistence between two adjacent rows of **X**. Moreover, compares with ℓ_2 norm, ℓ_1 norm is more inclusive and robust while considering temporal abrupt changes [6]. Thus we choose ℓ_1 norm to minimize **HX**, as it enforces the matrix **X** temporally stable [21]. Hence, we revise the RPCA model in (5) and obtain the following improved RPCA model:

$$\begin{aligned} \min_{\mathbf{X},\mathbf{A}} \quad & \|\mathbf{X}\|_* + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{HX}\|_1 , \\ \text{s.t.} \quad & \mathbf{D} = \mathbf{X} + \mathbf{A} \end{aligned} \tag{5}$$

where λ_2 controls weight of the term $\|\mathbf{HX}\|_1$.

To solve the improved model, we adopt the Alternating Direction Method of Multiplier (ADMM) [22], which is a popular algorithm for solving convex optimization problems. For this purpose, three auxiliary variables $\mathbf{L} \in \mathbb{R}^{m\times r}, \mathbf{Q} \in \mathbb{R}^{r\times n}$, and $\mathbf{S} \in \mathbb{R}^{(m-1)\times n}$ are introduced; let $\mathbf{X} = \mathbf{LQ}$ and $\mathbf{HX} = \mathbf{S}$, where **r** is the decomposition rank of **X**. Therefore (5) is rewritten as

$$\begin{aligned} \min_{\mathbf{X},\mathbf{A},\mathbf{S},\mathbf{L},\mathbf{Q}} \quad & \|\mathbf{LQ}\|_* + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{S}\|_1 , \\ \text{s.t.} \quad & \mathbf{D} = \mathbf{X} + \mathbf{A}, \\ & \mathbf{X} = \mathbf{LQ}, \\ & \mathbf{HX} = \mathbf{S} \end{aligned} \tag{6}$$

Remove the linear equality constraints in (6) with augmented Lagrangian method, and then we have the following objective function:

$$\begin{aligned} \mathbf{L}(\mathbf{X},\mathbf{A},\mathbf{S},\mathbf{L},\mathbf{Q}) = {} & \|\mathbf{L}\|_F^2 + \|\mathbf{Q}\|_F^2 + \lambda_1 \|\mathbf{A}\|_1 + \lambda_2 \|\mathbf{S}\|_1 \\ & + \langle \mathbf{Y}_1, \mathbf{D} - \mathbf{X} - \mathbf{A} \rangle \\ & + \frac{\mu}{2} \|\mathbf{D} - \mathbf{X} - \mathbf{A}\|_F^2 + \langle \mathbf{Y}_2, \mathbf{X} - \mathbf{LQ} \rangle \\ & + \frac{\mu}{2} \|\mathbf{X} - \mathbf{LQ}\|_F^2 + \langle \mathbf{Y}_3, \mathbf{HX} - \mathbf{S} \rangle \\ & + \frac{\mu}{2} \|\mathbf{HX} - \mathbf{S}\|_F^2 \end{aligned} \tag{7}$$

where $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ are Lagrange multipliers, $\mu > 0$ is adaptive penalty parameter, and $\langle \cdot, \cdot \rangle$ represents the standard trace inner product. We adopt an alternative iterations to solve this optimization as follows.

Update **X**. When **A**, **S**, **L**, and **Q** are fixed, (7) degenerates into a function with respect to **X**. So we solve **X** by the following optimization problem:

$$\begin{aligned} \mathbf{X}^{(i+1)} = \arg\min_{\mathbf{X}} & \frac{\mu^{(i)}}{2} \left\| \mathbf{D} - \mathbf{X} - \mathbf{A}^{(i)} + \frac{\mathbf{Y}_1^{(i)}}{\mu^{(i)}} \right\|_F^2 \\ & + \frac{\mu^{(i)}}{2} \left\| \mathbf{X} - \mathbf{L}^{(i)}\mathbf{Q}^{(i)} + \frac{\mathbf{Y}_2^{(i)}}{\mu^{(i)}} \right\|_F^2 \end{aligned}$$

$$+\frac{\mu^{(i)}}{2}\left\|\mathbf{HX}-\mathbf{S}^{(i)}+\frac{\mathbf{Y}_3^{(i)}}{\mu^{(i)}}\right\|_F^2 \tag{8}$$

Taking derivative of the objective function in (8) and setting it to 0, the closed-form solution is given by

$$\begin{aligned}\mathbf{X}^{(i+1)} &= \left(2\mu^{(i)}\mathbf{I}^{n\times n}+\mu^{(i)}\mathbf{H}^T\mathbf{H}\right)^{-1} \\ &\quad\cdot\left(\mu^{(i)}\left(\mathbf{D}-\mathbf{A}^{(i)}+\mathbf{L}^{(i)}\mathbf{Q}^{(i)}+\mathbf{H}^T\mathbf{S}^{(i)}\right)+\mathbf{Y}_1^{(i)}-\mathbf{Y}_2^{(i)} \\ &\quad+\mathbf{H}^T\mathbf{Y}_3^{(i)}\right)\end{aligned} \tag{9}$$

Update **A**. When others are fixed, in order to update **A**, one needs to solve the following ℓ_1 minimization problem:

$$\begin{aligned}\mathbf{A}^{(i+1)} &= \arg\min_{\mathbf{A}}\lambda_1\|\mathbf{A}\|_1 \\ &\quad+\frac{\mu^{(i)}}{2}\left\|\mathbf{A}-\left(\mathbf{D}-\mathbf{X}^{(i)}+\frac{\mathbf{Y}_1^{(i)}}{\mu^{(i)}}\right)\right\|_F^2\end{aligned} \tag{10}$$

whose solution is given by [23]:

$$\mathbf{A}^{(i+1)}=\delta\left(\mathbf{D}-\mathbf{X}^{(i)}+\frac{\mathbf{Y}_1^{(i)}}{\mu^{(i)}},\frac{\lambda_1}{\mu^{(i)}}\right) \tag{11}$$

where $\delta(a,b) = \operatorname{sgn}(a)(|a|-b)$ for $|a| \geq b$ and is zero otherwise.

Update **S**. In a similar way with updating **A**, the closed-form solution of S is given by

$$\mathbf{S}^{(i+1)}=\delta\left(\mathbf{HX}^{(i)}+\frac{\mathbf{Y}_3^{(i)}}{\mu^{(i)}},\frac{\lambda_2}{\mu^{(i)}}\right) \tag{12}$$

Update **L**, **Q**. In a similar way with updating **X**, the closed-form solutions of **L**, **Q** are given by

$$\mathbf{L}^{(i+1)}=\left(\mu^i\mathbf{X}^{(i)}+\mathbf{Y}_2^{(i)}\right)\mathbf{Q}^{(i)T}\left(2\mathbf{I}^{r\times r}+\mu^{(i)}\mathbf{Q}^{(i)}\mathbf{Q}^{(i)T}\right)^{-1} \tag{13}$$

$$\begin{aligned}&\mathbf{Q}^{(i+1)} \\ &\quad=\left(2\mathbf{I}^{r\times r}+\mu^{(i)}\mathbf{L}^{(i)T}\mathbf{L}^{(i)}\right)^{-1}\left(\mu^{(i)}\mathbf{L}^{(i)T}\mathbf{X}^{(i)}+\mathbf{L}^{(i)T}\mathbf{Y}_2^{(i)}\right)\end{aligned} \tag{14}$$

Update $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{Y}_3$, and μ. The Lagrangian multipliers $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ and penalty parameter μ could be updated as follows:

$$\begin{aligned}\mathbf{Y}_1^{(i+1)} &= \mathbf{Y}_1^{(i)}+\mu^{(i)}\left(\mathbf{D}-\mathbf{X}^{(i)}-\mathbf{A}^{(i)}\right), \\ \mathbf{Y}_2^{(i+1)} &= \mathbf{Y}_2^{(i)}+\mu^{(i)}\left(\mathbf{X}^{(i)}-\mathbf{L}^{(i)}\mathbf{Q}^{(i)}\right), \\ \mathbf{Y}_3^{(i+1)} &= \mathbf{Y}_3^{(i)}+\mu^{(i)}\left(\mathbf{HX}^{(i)}-\mathbf{S}^{(i)}\right), \\ \mu^{(i+1)} &= \min\left(\rho\mu^{(i)},\mu^{max}\right)\end{aligned} \tag{15}$$

where $\rho > 1$ is a constant and μ^{max} is the upper bound of μ.

```
Input: Data matrix $\mathbf{D}$, the parameters $\lambda_1 > 0, \lambda_2 > 0, r > 0$,
Initialize: $\mathbf{X}^{(0)} = \mathbf{A}^{(0)} = \mathbf{S}^{(0)} = 1 \in \mathbb{R}^{m\times n}$,
$\mathbf{Y}_1^{(0)} = \mathbf{Y}_2^{(0)} = 1 \in \mathbb{R}^{m\times n}, \mathbf{Y}_3^{(0)} = 1 \in \mathbb{R}^{m\times n}$,
$\mathbf{L}^{(0)} = 1 \in \mathbb{R}^{m\times r}, \mathbf{Q}^{(0)} = 1 \in \mathbb{R}^{r\times n}$,
$\mu^{(0)} = 10^{-6}, \rho = 1.1, \varepsilon = 10^{-6}, MaxIter = 1000, i = 0$.
while not converged and $i < MaxIter$ do
    Update $\mathbf{X}$ : via (9)
    Update $\mathbf{A}$ : via (11)
    Update $\mathbf{S}$ : via (12)
    Update $\mathbf{L}$ : via (13)
    Update $\mathbf{Q}$ : via (14)
    Update the multipliers: via (15)
    $i = i + 1$.
Output: Expected matrix $\mathbf{X}$, Sparse matrix $\mathbf{A}$.
```

Algorithm 1: Solving for the improved RPCA.

Convergence Conditions. The stopping criterion is measured by the following problem:

$$\max\left\{\begin{array}{l}\left\|\mathbf{X}^{(i+1)}-\mathbf{X}^{(i)}\right\|_\infty, \\ \left\|\mathbf{A}^{(i+1)}-\mathbf{A}^{(i)}\right\|_\infty, \\ \left\|\mathbf{S}^{(i+1)}-\mathbf{S}^{(i)}\right\|_\infty, \\ \left\|\mathbf{L}^{(i+1)}-\mathbf{L}^{(i)}\right\|_\infty, \\ \left\|\mathbf{Q}^{(i+1)}-\mathbf{Q}^{(i)}\right\|_\infty,\end{array}\right\} \leq \varepsilon. \tag{16}$$

where ε is tolerance error. If the convergence condition is met, the iteration terminates. The overall algorithm is summarized in Algorithm 1.

Once solving the improved RPCA, we obtain the expected flow **X** and the anomalous part **A**. To eliminate the interference of the noise, we use the three-sigma rule of thumb [24] to filter elements of **A**. ω_j is the standard deviation of $x_{1,j}, x_{2,j}, \ldots, x_{i,j}, \ldots, x_{m,j}$; if $-3\omega_j \leq a_{i,j} \leq +3\omega_j a_{i,j}$ is considered an allowable deviation and set as $a_{i,j} = 0$, then we get the anomaly flow matrix $\widehat{\mathbf{A}}$. Each element of $\widehat{\mathbf{A}}$ represents the abnormal amplitude of the space-time position; it may be positive or negative. The positive indicates that the passenger flow is higher than the expected flow and the negative indicates the passenger flow is lower than the expected flow.

3.3. Discovering the Potential Temporal-Spatial Laws among Anomalies.

Based on the improved RPCA, the anomalies of subway passenger flow are detected. To explore the potential laws of anomalies propagation, we group the anomalies into several clusters to identify anomalies in the region and their propagation laws. Because the detected anomalies have similar temporal-spatial characteristics, we use ST-DBSCAN algorithm [25] to cluster the station-level anomalies to find the anomaly in the region. We regard $p(lon, lan, t)$ as the feature of an anomaly data object, lon and lan are the longitude and the latitude of a station, and t denotes the time interval. The ST-DBSCAN algorithm requires three parameters: space radius R, time window ΔT, and density threshold $MinPts$; the first two parameters determine neighborhood on temporal-spatial dimension.

The algorithm starts with the earliest anomaly data object p and retrieves all neighbors of point p within spatiotemporal neighborhood. If the number of neighbors is greater than $MinPts$, a new cluster is created which has p as core of the cluster. Then, the algorithm iteratively collects neighbors beginning with another core point. The above procedure continues until all points have been processed.

4. Experiments

In this section, we evaluate the robustness of the improved RPCA by adding noise on a set of real subway passenger flow data, comparing with RPCA [6] the wavelet transform method [8] and the threshold method [7]. Then we apply our proposed framework on the real subway passenger flow data for anomalies detection and analysis; meanwhile the results are verified with traffic related tweet data.

4.1. Robustness Evaluation. The improved RPCA model is characterized by its robustness to noise, so we first validate the performance of our methods on noisy passenger flow data compared with the related methods.

First, we construct three real-world passenger flow datasets from three different geographical positions shown in Figure 2. By exploiting the strong weekly seasonality observed in the data, we convert hourly flow within one week into a row vector and stack 12 weeks vector to form the data matrix which contains much noise. To implement the verification experiment, it needs to know the ground truth. In the case of ground truth being unavailable, we have to estimate a relatively accurate ground truth. Here, we use 4 layers of wavelet to filter the small white noise and then take the average as ground truth value. As a result, we get three relatively clean and ideal ground truth datasets, denoted by $\mathbf{G}_i$ $(i = 1, 2, 3)$.

Next, we add sparse noise on the ground truth matrices to simulate the corresponding noise matrices $\mathbf{A}_i$ $(i = 1, 2, 3)$. The randomly corrupted proportion cp of these matrices varies from 0.06 to 0.50; the fluctuation range is ±80% of the average of $\mathbf{G}_i$ $(i = 1, 2, 3)$. So we obtain the noisy passenger flow matrices by mixing the ground truth matrices $\mathbf{G}_i$ and the produced noise matrices $\mathbf{A}_i$ by $\mathbf{D}_i = \mathbf{G}_i + \mathbf{A}_i$. These datasets will be used as the test datasets for anomalies detection and evaluating the robustness of the proposed method. The properties of the constructed datasets are summarized in Table 1.

Evaluation Criteria. To evaluate the performance of the improved RPCA algorithm, we use the precision rate pr in [21] to evaluate the recognition accuracy of anomalies, which are defined as follows:

$$pr = \frac{2 * precision * recall}{precision + recall} \qquad (17)$$

where $precision = a_{true}/a_{all}$, $recall = a_{true}/a$, a_{all} and a_{true} denote the number of anomalies recognized by our model and the number of true anomalies among them, respectively, and a represents the actual number of anomalies. pr is calculated by averaging results over 10 runs.

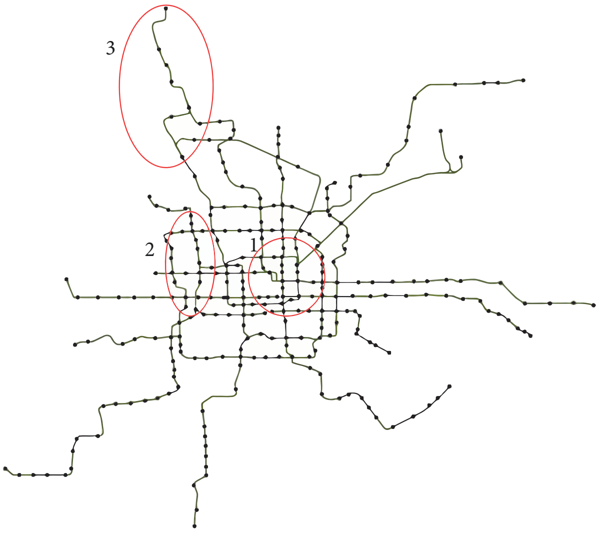


FIGURE 2: The selected stations in three areas of Beijing.

TABLE 1: Datasets description.

Dataset	Numbers of stations	Size
D1	20	240 × 126
D2	18	216 × 126
D3	12	144 × 126

Parameters Setting. The improved RPCA has three parameters λ_1, λ_2, and decomposition rank r, and they are important for the performance. Rank r needs to be as small as possible to minimize matrix sparsity and low-rank error. Here, we use singular value decomposition (SVD) [26] to estimate a superior rank for these three datasets. Figure 3 shows the distribution of the singular value of the three ground truth datasets. The x-axis presents the ith singular values and the y-axis presents the cumulative ratio of the first i singular values to the sum of all singular values. It can be found that the first 14 singular values almost dominate nearly 90% energy in all three datasets. To simplify, the rank r is set as 14 for all datasets.

For λ_1 and λ_2, we first change one parameter while fixing the other parameter in the model, and the parameter is gradually taken as $10^{-3}, 10^{-2}, \ldots, 10^{3}$. Then we achieve the relatively superior parameters. Next, we tune these parameters in a narrow range from 10^{-1} to 10 by step of 0.1. Finally we obtain the relatively optimal $\lambda_1$and λ_2. The setting of experimental parameters is shown in Table 2.

In our experiments, we apply 5 layers of discrete wavelet transform based on the wavelet of DB4. For the threshold method, the threshold value of flow at different time interval is different and we compute the mean value Av_j and standard deviation σ_j of $d_{1,j}, d_{2,j}, \ldots, d_{i,j}, \ldots, d_{m,j}$ and

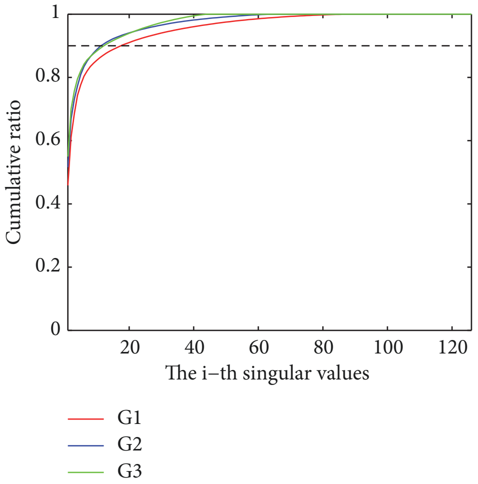


Figure 3: Cumulative ratio distribution of singular value in three ground truth datasets.

Table 2: Experiment parameters.

Dataset	The improved RPCA		RPCA
	λ_1	λ_2	λ
D1	1	0.4	0.1
D2	1.4	0.4	0.2
D3	1.1	0.5	0.1

set the confidence interval as $(Av_j - 3\sigma_j, Av + 3\sigma_j)$; it is judged to be an anomaly if $d_{i,j}$ is beyond the confidence interval.

Experiment Result Analysis. Figure 4 is the comparison results of the four methods. As can be seen, there is a downward trend of *pr* with the increasing of data corrupted proportion. The improved RPCA is superior to other methods, followed by RPCA. Notice that the threshold method has worse performance, because the noise data reduce the calculation accuracy of the threshold range. When the data corrupted proportion is low, the wavelet transform has a good detection result, but the local stationarity is destroyed in a high corrupted proportion, which results in the detection accuracy of a steady decline. The improved RPCA is more robust than RPCA even in high corrupted proportion, because the constraint item $\|\mathbf{HX}\|_1$ can capture the feature of abrupt changes of the time series data when the sparseness of anomalies becomes weak. Additionally, the improved RPCA performs well on different stations from different geographical positions. In a word, the improved RPCA is more suitable for anomalies detection of subway passenger flow.

4.2. Anomalies Detection and Verification. In order to demonstrate the practicability and the authenticity of the improved RPCA, we conduct anomalies detection experiments on real-world datasets and verify the results with tweet data. Figure 5 shows the decomposition results of the exit flow of Xidan station. The low-rank expected flow matrix $\mathbf{X}$ represents the weekly pattern and the anomaly matrix $\widehat{\mathbf{A}}$ successfully captures multiple outliers.

To further analyze and verify the anomalies, we collect tweet data which contain a wide variety of information and retrieve events information through natural language processing method. There are four explanatory anomaly regions, highlighted by ellipses in Figure 5. They correspond with the following events 1 ~ 3 extracted from tweet data, as shown in Figure 6, and specific analysis as follows:

(i) Event1: The ellipse region 1 in Figure 5 shows the increasing of the exit flow lasting about three hours in the evening. It is because many large shopping malls near Xidan station held sales for Chinese Valentine's Day, which attracted massive customers and led a rise in exit flow.

(ii) Event2: In ellipse regions 2 and 3, the flow was declined. It is because Xidan station was closed for facilitating celebration parade for the 70*th* anniversary victory of the anti-Japanese war.

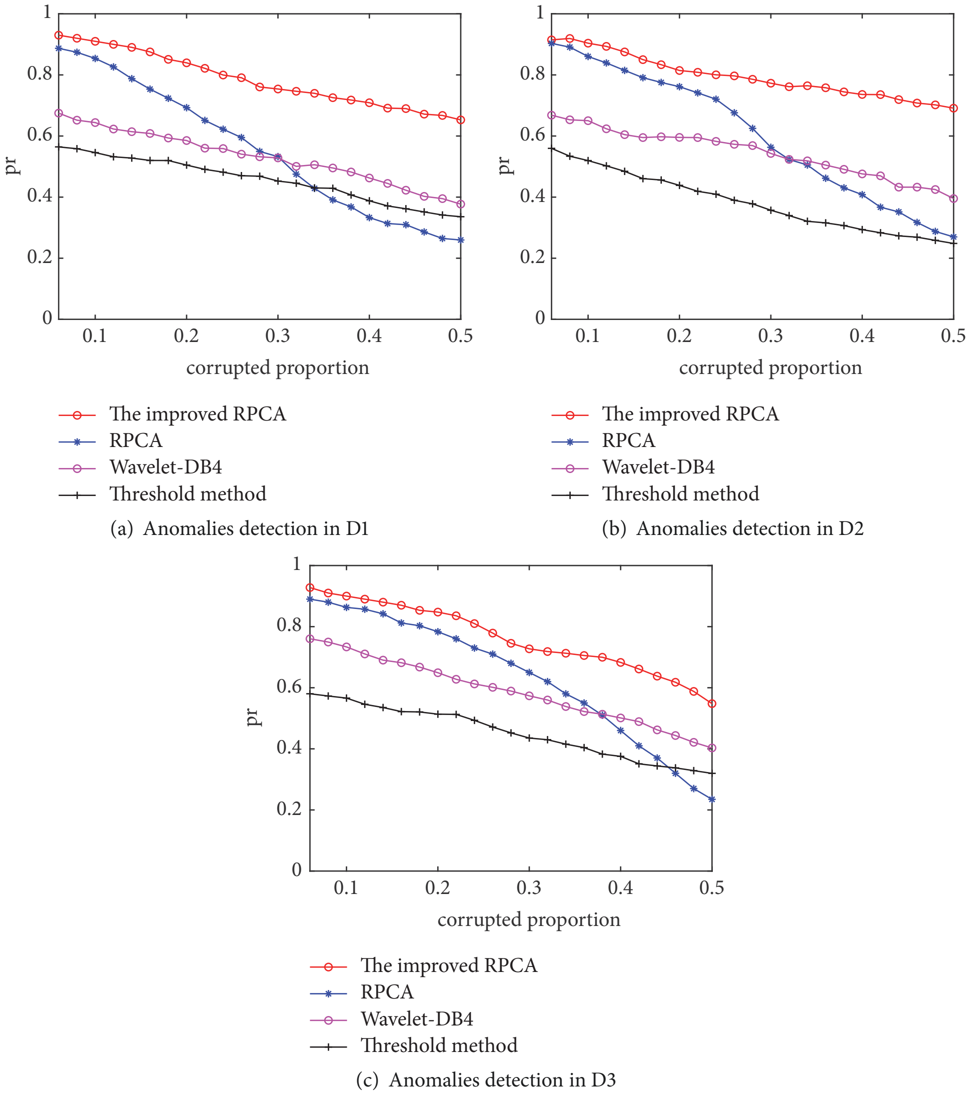
(iii) Event3: In ellipse region 4, the exit flow was higher than usual. Because it was a commuter day due to legal exchanging holiday, therefore the flow was increasing and consistent with the flow of a working day.

The improved RPCA can not only identify anomalies at the station level but also accurately detect anomalies. These anomalies could be used for a reference for real-time alerting.

4.3. Discovering the Potential Temporal-Spatial Laws among Anomalies. An isolated anomaly may affect neighbored stations consecutively, so anomalies among some stations have strong temporal-spatial correlations. Grouping several anomalies along temporal-spatial dimensions may reveal the evolution or the impact of the isolated anomaly; hence we adapt ST-DBSCAN clustering algorithm to group the anomalies to analyze the propagation feature of the anomalies.

In experiments, space radius $R = 0.03$ (Euclidean distance of latitude and longitude between two adjacent stations), time window $\Delta T = 1$ hour interval, and density threshold *MinPts* = 3 work well. We cluster the anomalies of all stations in one week and name each cluster as an anomaly event.

In Figure 7, we use ellipses to highlight four anomaly events grouped by ST-DBSCAN, and these clustered results are easier to be verified by tweet data and visually analyzed. The ellipse region 1 in Figure 7(a) shows the entry flow decrease of the stations on the same line. It was lasting for five hours and induced by the closure strategy. Meanwhile, it also led to the flow increase of the nearby stations. In particular, the transfer stations such as Dongsi station and Chaoyangmen station had an obvious flow increase. The ellipse region 2 in Figure 7(a) shows the surges of exit flow as attendees traveled to Bird's Nest stadium for the opening ceremony of IAAF World Championships. In



(a) Anomalies detection in D1

(b) Anomalies detection in D2

(c) Anomalies detection in D3

Figure 4: Performance of anomalies detection in three ground truth datasets.

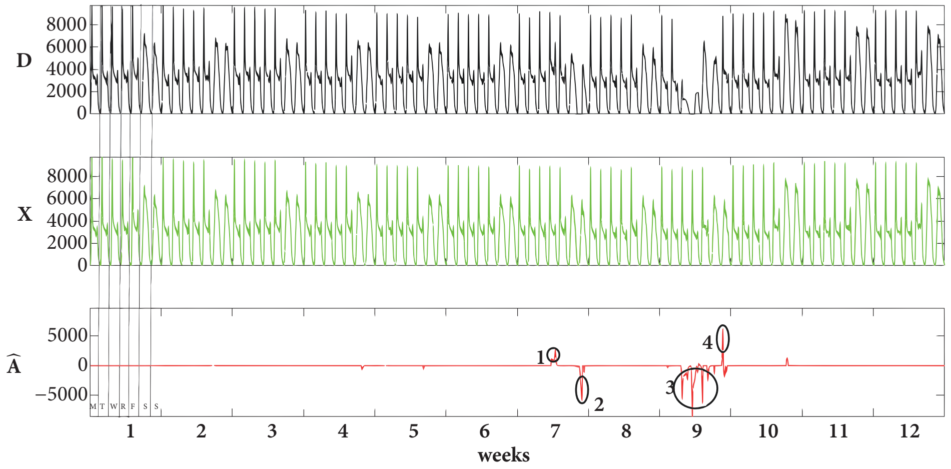
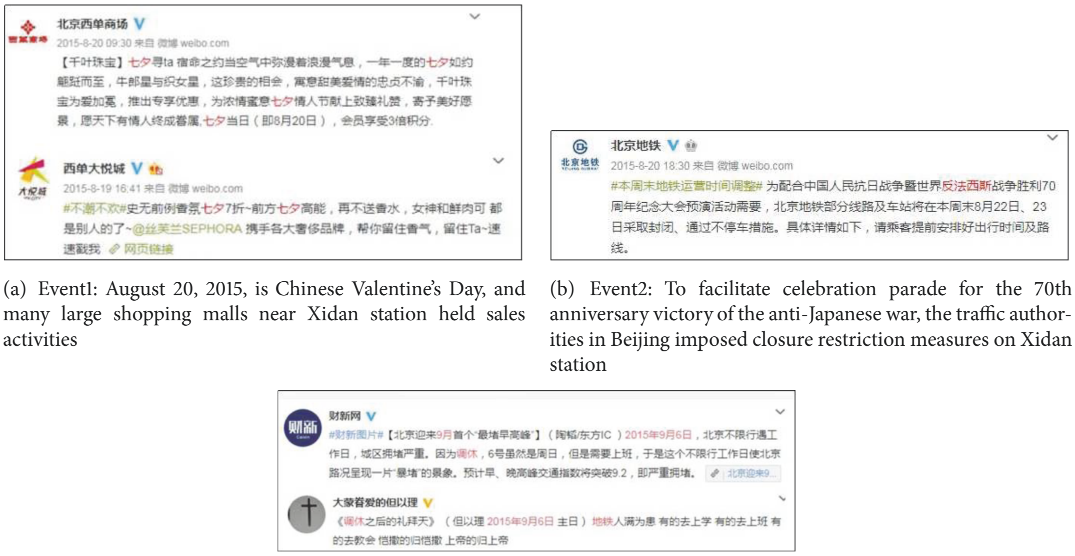


Figure 5: Component decomposition of exit flow of Xidan station by the improved RPCA. The horizontal axis is time interval covering 12 weeks, and the vertical axis is flow value.


北京西单商场 V
2015-8-20 09:30 来自 微博 weibo.com
【千叶珠宝】七夕寻ta 宿命之约当空气中弥漫着浪漫气息，一年一度的七夕如约翩跹而至，牛郎星与织女星，这珍贵的相会，寓意甜美爱情的忠贞不渝，千叶珠宝为爱加冕，推出专享优惠，为浓情蜜意七夕情人节献上致臻礼赞，寄予美好愿景，愿天下有情人终成眷属.七夕当日（即8月20日），会员享受3倍积分.

西单大悦城 V
2015-8-19 16:41 来自 微博 weibo.com
#不潮不欢#史无前例香氛七夕7折~前方七夕高能，再不送香水，女神和鲜肉可 都是别人的了~@丝芙兰SEPHORA 携手各大奢侈品牌，帮你留住香气，留住Ta~速速戳我 网页链接

(a) Event1: August 20, 2015, is Chinese Valentine's Day, and many large shopping malls near Xidan station held sales activities

北京地铁 V
2015-8-20 18:30 来自 微博 weibo.com
#本周末地铁运营时间调整# 为配合中国人民抗日战争暨世界反法西斯战争胜利70周年纪念大会预演活动需要，北京地铁部分线路及车站将在本周末8月22日、23日采取封闭、通过不停车措施。具体详情如下，请乘客提前安排好出行时间及路线。

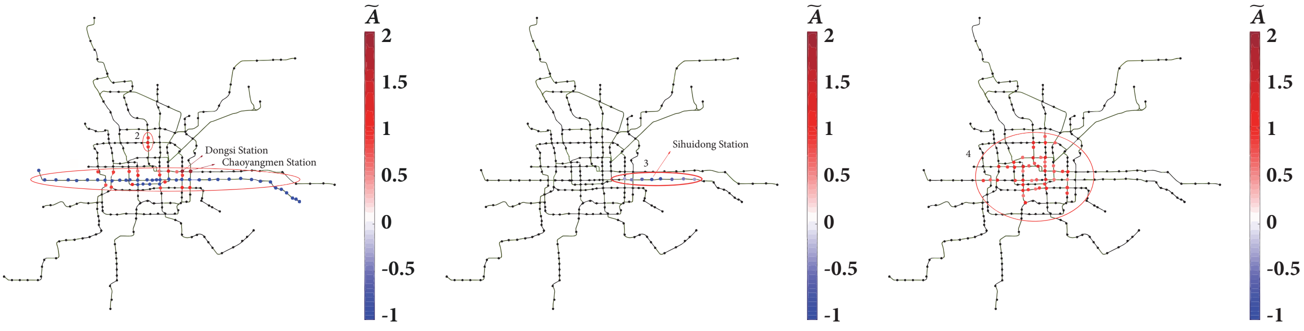
(b) Event2: To facilitate celebration parade for the 70th anniversary victory of the anti-Japanese war, the traffic authorities in Beijing imposed closure restriction measures on Xidan station

财新网 V
#财新图片#【北京迎来9月首个"最堵早高峰"】（陶韬/东方IC ）2015年9月6日，北京不限行遇工作日，城区拥堵严重。因为调休，6号虽然是周日，但是需要上班，于是这个不限行工作日使北京路况呈现一片"暴堵"的景象。预计早、晚高峰交通指数将突破9.2，即严重拥堵。 北京迎来9...

大蒙眷爱的但以理 V
《调休之后的礼拜天》（但以理 2015年9月6日 主日）地铁人满为患 有的去上学 有的去上班 有的去教会 恺撒的归恺撒 上帝的归上帝

(c) Event3: September 6, 2015, is Sunday but exchanged with the working day because of legal exchanging holiday

Figure 6: The events information sent by tweet users.



(a) Events in Aug. 23

(b) Event in Aug. 19

(c) Event in Aug. 20

Figure 7: Anomaly events (labeled with ellipse). Red points mean traffic flow increasing and blue points mean traffic flow declining. We normalize anomaly $\mathbf{A}$ to get $\widetilde{\mathbf{A}} = \mathbf{A}/\mathbf{X}$ and values of $\widetilde{\mathbf{A}}$ denotes deviation from the expectation.

Figure 7(b), ellipse region 3 shows the spreading of the anomaly caused by an hour's breakdown of the train on Sihuidong station. In Figure 7(c), ellipse region 4 shows the entry flow increase of many stations for one day, since the traffic control of city roads led more people to choose the subway.

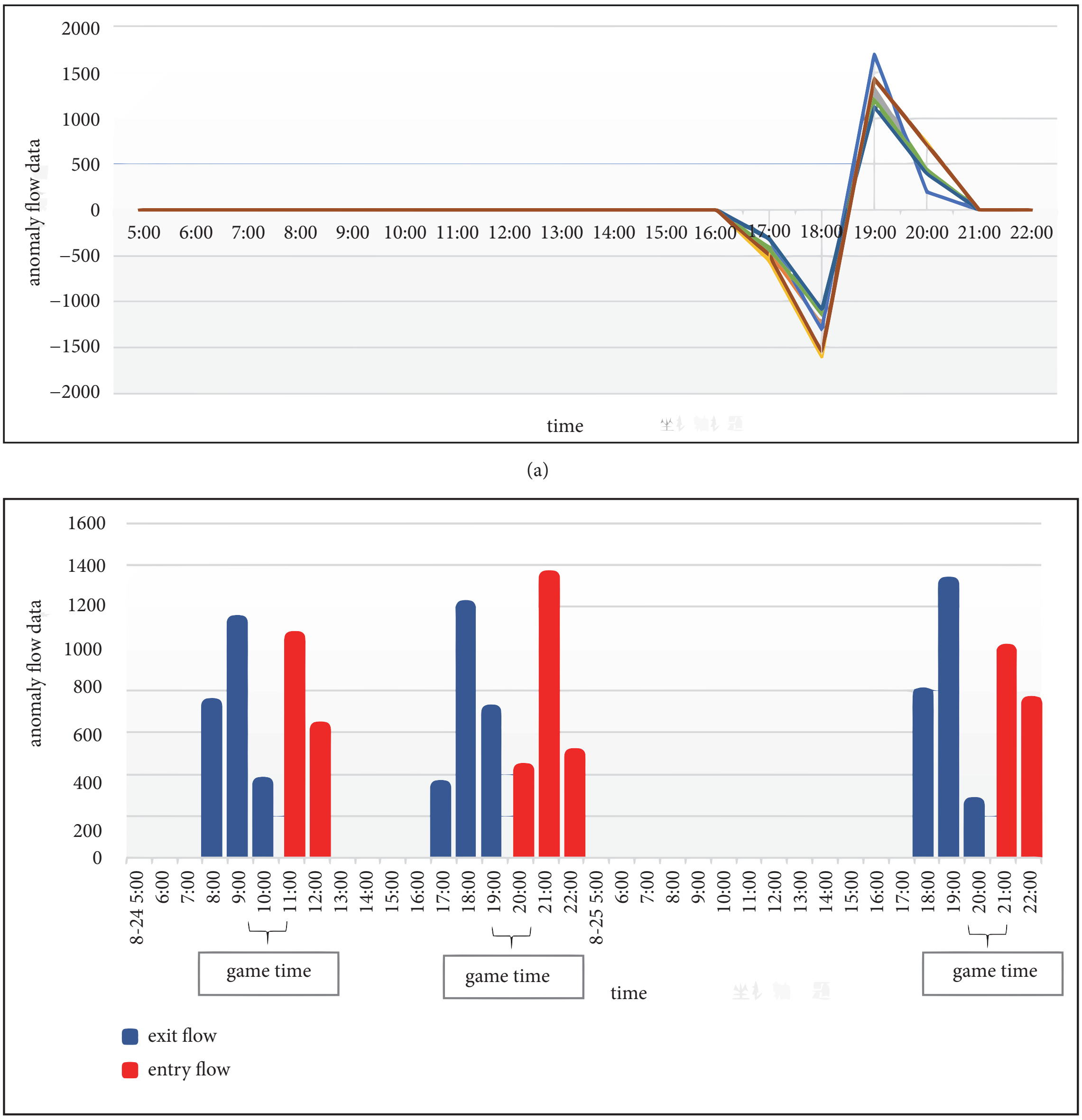
After the clustering and verification process, we can discover the potential temporal-spatial laws among anomalies from the following three aspects:

(i) Distribution and spreading of anomalies along time and space: how many stations? And how long are they affected? The center of anomalies and the range of spreading from some destine anomaly. As shown in Figure 7(c), ellipse region 4 shows the stations affected by traffic control measures. Ellipse region 3 shows the anomaly in Sihuidong station spread to the adjacent stations.

(ii) The serious degree of anomalies: Values in $\widetilde{A}$ reflect the serious degree of anomalies. In Figure 7, affected stations labeled with the red color having different levels reveal the degrees of impaction. The heavier the color, the more severe the anomaly impact.

(iii) The potential impact caused by events: Some anomaly events not only affect the corresponding stations, but also cause the potential impact on the surrounding stations. As shown in ellipse region 1 of Figure 7(a), the closure strategy resulted in a potential increasing flow of the surrounding stations.

Furthermore, we apply the statistical analysis to get some rules shown in Figure 8. The detected events are classified into two categories: some are predefined such as traffic control and vocal concerts and some are emergencies such as subway device failure and a sudden heavy rain. For both



(a)

(b)

Figure 8: In the two figures, the horizontal axis is time interval and the vertical axis is anomaly flow. (a) shows the exit anomaly flow of multiple stations on August 7; lines represent different stations. From 17:00 to 20:00, there are continuous negative and positive anomalies, because there was a sudden heavy rain at 17:00 that resulted in the passenger flow peak moving back. (b) shows the anomaly flow of Olympic Sports Center Subway Station near the Bird's Nest stadium. It is obvious that the exit flow surged before the beginning of game and the entry flow surged after the end of game.

two categories events, the above analysis with our method can provide a beneficial suggestion for subway managers:

(i) For emergency events, our framework provides distribution laws of anomaly events, and these can be used for estimating anomalies' propagation and impact on adjacent stations. As shown in Figure 8(a), a sudden heavy rain caused a delay of the evening rush hour, so that managers can further push announcement timely to remind passengers and take emergency measures. This would prevent subway station from chaos and hazard spreading and also save the travel time of passengers.

(ii) For predefined events, our framework indicates detailed rules along spatial and temporal dimension, so that subway managers can obtain prior information and make sufficient preparations to cope with possible anomalies. As shown in Figure 8(b), the exit flow of Olympic Sports Center Station surged in the two hours before the beginning of one game, and the entry flow surged in the two hours after the end of this game. These anomalies rules can help to estimate the impact of anomaly flow involved to major urban events and then take mitigation strategies in advance.

5. Conclusion

In this paper, the improved RPCA is suggested to detect station-level anomalies in subway, and ST-DBSCAN algorithm is used to group the detected station-level anomalies into clusters named as anomaly events. This framework can not only precisely locate anomalies in temporal dimension but also find the distribution and spreading in temporal and spatial dimension. With the detection results and impact analysis of events, subway managers can estimate traffic flow impact involved to predicted events and then take corresponding measures. Besides, they can push announcement timely for unpredicted events through decomposing the realtime data.

In future, we shall improve our work in three aspects. First, we shall extend our model to anomalies prediction as well as anomalies propagation process. Second, we shall consider temporal-spatial distribution by extracting comprehensive temporal and spatial information, e.g., OD flow data. Third, we shall propose more efficient ADMM algorithm for solving the proposed model and propose convergence analysis of the algorithm.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by Beijing Municipal Science & Technology Commission (Grant nos. Z151100002115040, Z171100000517003, and Z171100000517004), Project of Beijing Municipal Education Commission (Grant nos. KM201510005025, KM201610005033), and Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality (Grant no. IDHT20150504).

References

[1] H. J. Payne and S. C. Tignor, "Freeway Incident-Detection Algorithms Based on Decision Trees with States," *Transportation Research Record*, no. 682, pp. 30–37, 1978.

[2] C. L. Dudek, C. J. Messer, and N. B. Nuckles, "Incident detection on urban freeways," *Transportation Research Record*, 1974.

[3] B. N. Persaud and F. L. Hall, "Catastrophe theory and patterns in 30-second freeway traffic data- Implications for incident detection," *Transportation Research Part A: General*, vol. 23, no. 2, pp. 103–113, 1989.

[4] L. X. Pang, S. Chawla, W. Liu et al., "On mining anomalous patterns in road traffic streams," in *Proceedings of the International Conference on Advanced Data Mining and Applications*, pp. 237–251, 2011.

[5] L. X. Pang, S. Chawla, W. Liu, and Y. Zheng, "On detection of emerging anomalous traffic patterns using gps data," *Data & Knowledge Engineering*, vol. 87, no. 9, pp. 357–373, 2013.

[6] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, article 11, 2011.

[7] S. M. Turner, "Guidelines for developing its data archiving systems," *Data Collection*, 2001.

[8] L. Zhi Min, Y. Liang You, P. Xue et al., "Short-term traffic flow detection based on wavelet," *'Application Research of Computers*, vol. 28, no. 5, pp. 1677-1678, 2011.

[9] S. Shekhar, C.-T Lu, and P. Zhang, "Detecting graph-based spatial outliers: algorithms and applications (a summary of results," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 451–468, 2001.

[10] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xie, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1010–1018, August 2011.

[11] S. Chawla, Y. Zheng, and J. Hu, "Inferring the root cause in road traffic anomalies," in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM '12)*, pp. 141–150, December 2012.

[12] D. Jin and S. Zhu, "Spatio-temporal feature based anomaly event recognition," in *Proceedings of the 35th Chinese Control Conference, CCC 2016*, pp. 3818–3822, China, July 2016.

[13] X. Sun, S. Zhu, and Y. Cheng, "Temporal-spatial coherence based abnormal behavior detection," in *Proceedings of the 29th Chinese Control and Decision Conference, CCDC 2017*, pp. 1997–2001, China, May 2017.

[14] P. Wang, C. Wu, and X. Gao, "Research on subway passenger flow combination prediction model based on RBF neural networks and LSSVM," in *Proceedings of the 28th Chinese Control and Decision Conference, CCDC 2016*, pp. 6064–6068, China, May 2016.

[15] D. Y. Zhang and H. N. Yang, "Passenger Flow Analysis in Subway Using a Kind of Neural Network," *Applied Mechanics and Materials*, vol. 713-715, pp. 2284–2287, 2015.

[16] Z. L. Zhang and Y. F. Yan, "Image denoising based on recovering low dimensional manifolds in patch space," *Computer Engineering & Design*, vol. 34, no. 2, pp. 579–583, 2013.

[17] B. Y. Wang and X. G. Wang, "Research on background modeling based on low-rank matrix recovery," *Applied Mechanics and Materials*, vol. 635-637, pp. 1056–1059, 2014.

[18] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[19] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 597–610, 2013.

[20] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 219–230, 2004.

[21] W. Ye, L. Chen, G. Yang, H. Dai, and F. Xiao, "Anomaly-tolerant traffic matrix estimation via prior information guided matrix completion," *IEEE Access*, vol. 5, pp. 3172–3182, 2017.

[22] E. Wei and A. Ozdaglar, “Distributed alternating direction method of multipliers,” *Decision and Control*, pp. 5445–5450, 2012.

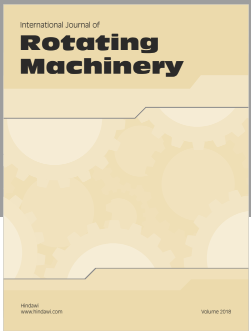
[23] A. M. Bruckstein, D. L. Donoho, and M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images,” *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

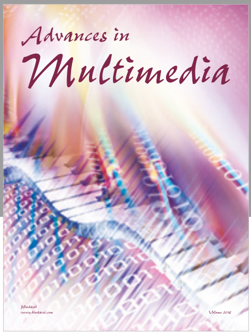
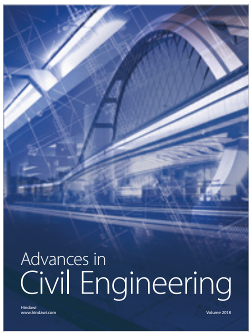
[24] M. Zhang and Y. Hui, “The pata criterion and rejecting the abnormal value,” *Journal of Zhengzhou University of Technology*, vol. 18, no. 1, pp. 84–88, 1997.

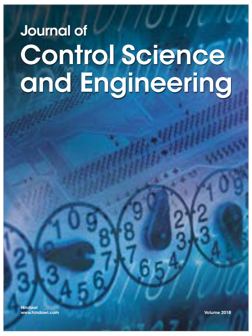
[25] D. Birant and A. Kut, “St-dbscan: An algorithm for clustering spatial¿ctemporal data,” *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[26] D. Kalman, “A Singularly Valuable Decomposition: The SVD of a Matrix,” *The College Mathematics Journal*, vol. 27, no. 1, pp. 2–23, 1996.

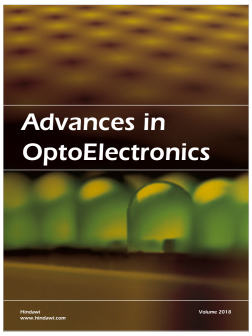
Journal of
Engineering

The Scientific World Journal

International Journal of
Rotating Machinery

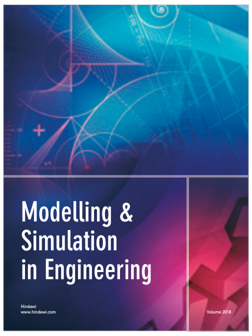
Journal of
Sensors

Advances in
Multimedia

Advances in
Civil Engineering

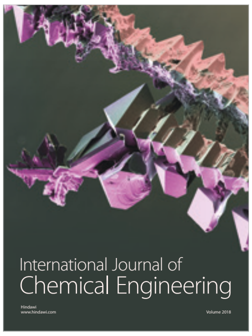
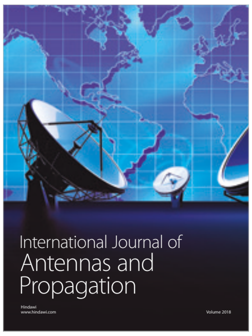

Journal of
Control Science and Engineering

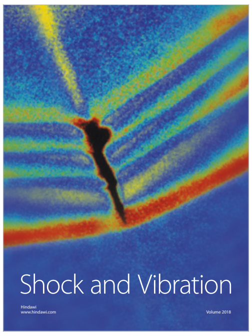
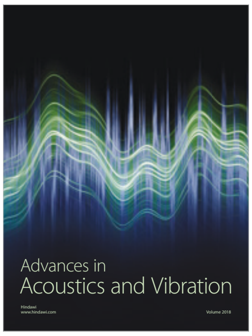
Journal of
Robotics

Journal of
Electrical and Computer Engineering

Advances in
OptoElectronics

VLSI Design

International Journal of
Navigation and Observation

Modelling & Simulation in Engineering

International Journal of
Aerospace Engineering

International Journal of
Chemical Engineering

International Journal of
Antennas and Propagation

Active and Passive Electronic Components

Shock and Vibration

Advances in
Acoustics and Vibration