Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5283283, 11 pages
https://doi.org/10.1155/2021/5283283

Research Article

Detecting Invalid Associations between Fare Machines and Metro Stations Using Smart Card Data

Pengfei Zhang,[1] Zhenliang Ma ,[2] and Xiaoxiong Weng[1]

[1]*School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510000, China*
[2]*Institute of Transport Studies, Department of Civil Engineering, Monash University, Clayton, VIC 3168, Australia*

Correspondence should be addressed to Zhenliang Ma; mike.ma@monash.edu

Received 24 April 2021; Accepted 31 May 2021; Published 11 June 2021

Academic Editor: Inhi Kim

Data quality is essential for its authentic usage in analysis and applications. The large volume of automated collection data inevidently suffers from data quality issues including data missing and invalidity. This paper deals with an invalid data problem in the automated fare collection (AFC) database caused by the erroneous association between the fare machines and metro stations, e.g., a fare machine located at Station A is wrongly associated with Station B in the AFC database. It could lead to inappropriate fare charges in a distance-based fare system and cause analysis bias for planning/operation practice. We propose a tensor decomposition and isolation forest-based approach to detect and correct the invalid associated fare machines in the system. The tensor decomposition extracts features of passenger flows and travel times passing through fare machines. The isolation forest coupled with a neural network (NN) takes these features as inputs to detect the wrongly associated fare machines and infer the correct association stations. Case studies using data from a metro system show that the proposed detection approach achieves over 90% accuracy in detecting the invalid associations for up to 35% invalid associations. The inferred association has a 90% accuracy even when the invalid association ratio reaches 40%. The proposed data-driven invalid data detection method is useful for large-scale data management in terms of data quality check and fix.

1. Introduction

Smart card data collected from the automatic fare collection (AFC) system (i.e., AFC data) enable many beneficial applications in the public transportation system such as collective and individual mobility analysis, system state monitoring, and operation planning and control [1]. The usefulness of these analysis applications is highly dependent on the data quality. The AFC data are collected online and in a large scale that may inevitably encounter data quality issues such as data missing and invalidity.

Data problems are prone to happen due to the following reasons:

(i) Human factors: in the AFC system, the transaction records may be missing if passengers fail to tap in/out properly.

(ii) Infrastructure failure: for example, AFC records are triggered when a passenger taps in/taps out through an entrance/exit fare machine. The malfunctioning of fare machines may lead to issues of missing data (machine fails to record or upload transactions) and invalid data (erroneous transactions).

(iii) Inadequate data management. Daily data management for transportation systems is a complex practice. Missing and invalid data may happen in the process of database merging, maintenance, or system update.

Among those data problems, missing and invalid data problems are the most critical and common ones. Figure 1 illustrates the characteristics of these two problems and also their difference. The missing data are cognizable and clearly identifiable via the data structure. For example, some AFC

AFC data

User ID	Origin	Tap-in time	Destination	Tap-out time
XXXX	A	01/01/2021 8:00	B	01/01/2021 8:10
YYYY	B	01/01/2021 8:05	F	01/01/2021 8:10

Missing data problem

User ID	Origin	Tap-in time	Destination	Tap-out time
XXXX	A	01/01/2021 8:00	B	01/01/2021 8:10
YYYY	B	01/01/2021 8:05	?	?

Invalid data problem

User ID	Origin	Tap-in time	Destination	Tap-out time
XXXX	A	01/01/2021 8:00	B	01/01/2021 8:10
YYYY	B	01/01/2021 8:05	D	01/01/2021 9:10

Figure 1: Missing and invalid data problems in the AFC dataset.

transactions may have missing data on tap-out records (empty cells). However, the invalid data are impossible to be directly recognised since the data structure is exactly the same as the valid data. Generally, the invalid data problem can be divided into two categories: data record and association errors. The data record error originates from the facility malfunctioning in the AFC system as mentioned above. The data association error occurs in the process of merging different sources of data (e.g., AFC fare machine records and station dictionaries). The data association error may come from the incomplete information inference and invalid information matching.

The paper deals with the invalid data problem to detect the hidden association errors of the complete and seemingly valid data. Specifically, it aims to detect the invalid association between fare machines and stations in the AFC data. For example, fare machine 001# is located in Metro Station A, but wrongly associated to Station B in the AFC database (Figure 2). The problem is prone to happen as the fare machines are frequently added, replaced, etc. in the metro systems, but the fare machine-station dictionary may fail to update timely. The consequences of invalid associations could be significant, e.g., under/over charging for a large amount of passengers. In addition, it is costly to fix this problem by manpower. One should manually check all the machines in metro stations to rebuild the correct association between fare machines and stations. Especially, it is impossible to manually detect such problem in the historical dataset since the fare machine distribution may not consistent with the current system.

We develop a data-driven approach, based on tensor decomposition and machine learning techniques, to automatically detect such invalid associations using AFC data, and also infer the correct association stations that a fare machine belongs to. The approach works in two steps: the tensor decomposition is utilized to extract the flow volume and travel time patterns of each fare machine. Then, the isolation tree technique and NN models are designed to detect the incorrect linked fare machines and infer their correct association stations based on the extracted features from tensor decomposition.

The remaining is organised as follows: Section 2 reviews the relevant studies on data quality issues, including overview of data quality problems, feature extraction techniques, and anomaly detection; the problem formulation and methodology are presented in Section 3; Section 4 reports the case study using the AFC data from a large metro system; the final section concludes the paper and discusses potential further studies and applications.

2. Related Work

2.1. Data Quality Problems. Data quality is one of the most important issues in big data area. Low or bad data quality is costly. For example, it is reported that bad data or poor data quality costs US businesses 600 billion dollars annually [2]. For metro systems, AFC systems collect massive transaction data of metro passengers. The literature has reported plenty of data quality problems related to AFC data. Robinson et al. [3] reported that the reasons of AFC data quality problems can be grouped into 4 categories: (1) software; (2) data; (3) hardware; (4) user. A recurrent information missing problem of the boarding station in Beijing Metro has been reported by Ma et al. [4]. Liu et al. [5] reported a time synchronisation problem of the AFC and AVL system, which causes the recorded boarding time information to be invalid in a large scale. Network, scheduling, fare table, etc. are important data stored in the AFC database. Errors in these data will lead to significant consequences. For example, the London Oyster smart card system crashed on Saturday 12th July 2008 due to erroneous data resulting in over 40,000 Oyster cards having to be replaced [6].

Although many studies deal with missing data in transportation, to the best of our knowledge, there is no study on detecting or fixing the association errors in transportation or other related areas, particularly the fare machine-station invalid association problem.

2.2. Feature Extraction Techniques. The key idea for a data-driven detection approach is to extract the passenger flow or/and travel time patterns between fare machines and stations.

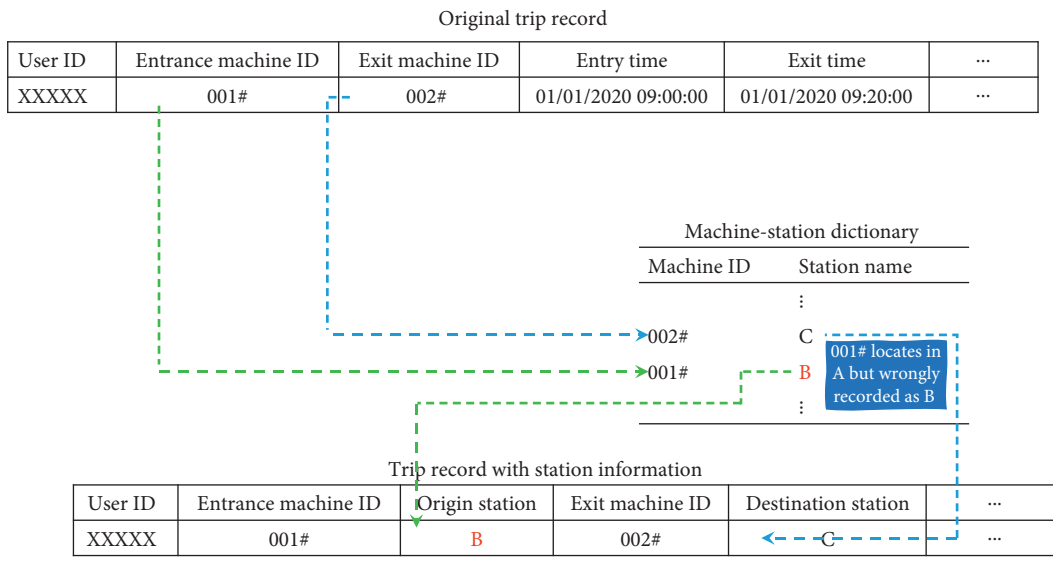

Figure 2: Invalid association problem between fare machine and metro station.

Feature extraction is one of the most important issues in the machine learning field. Feature extraction reduces the resources required to characterize a large set of data or/and a huge dimensions of input information. Plenty of methods are proposed in the machine learning community dealing with the feature extraction. These methods can be roughly divided into two parts: conventional statistical learning methods and deep learning-based method. Conventional statistical learning methods such as principle component analysis (PCA) [7], Isomap [8], and partial least squares (PLS) regression [9] mainly based on the statistical learning-based algorithms. The advantages of these methods are they are robust to small dataset, i.e., do not need large amount of samples to maintain the performance. However, the disadvantages are also critical. For example, they are not robust to noisy samples, and the feature extraction quality is highly dependent on specific tricks in different tasks, thus which are less generalized. Deep learning-based feature extraction methods become more and more popular recently. Variety forms of neural networks, e.g., convolutional neural network (CNN) [10] and long short-term memory (LSTM) [11] neural network. can be treated as feature extraction models. Different from the statistical learning-based algorithms, they extract the features in a latent, end-to-end manner. The advantage is that the extracted features are more representative and comprehensive. However, these models always require a large dataset in the training procedure; thus, they are not suitable in the few-shot scenario. In conclusion, there is no a generalized feature extraction method for all the tasks. Feature extraction methods should be designed based on the characteristics of the focused problem.

In our problem, passenger flow and travel time patterns are related to multiple modes, e.g., time and location. Tensor is a nature choice to represent and capture these patterns. Tensor is a multidimensional extension of matrix [12]. Tensor has been widely used in transportation area to deal with multidimension data. Tan et al. [13] utilized a tensor decomposition approach to capture the multimode correlations in traffic data and recover missing traffic data by reconstructing the traffic flow tensor. The results show that the proposed algorithm performs well even when the missing ratio is high. Chen et al. [14] proposed singular value decomposition (SVD)-combined tensor decomposition framework to complete the traffic data using traffic speed information. Sun and Axhausen [15] utilize a probabilistic tensor decomposition method to mine the urban mobility patterns. Mobility patterns of different passenger groups (e.g., students, adults, and elders) are explored. In our study, we also use tensor decomposition to extract the flow pattern related to each fare machine.

2.3. Anomaly Detection. The invalid associations (between fare machines and stations) are treated as anomalies. Anomaly detection is an important topic in data mining. The anomaly detection could be roughly divided into three categories, statistical, machine learning, and deep learning models.

(1) Statistical method: statistical methods are the early explorations of the anomaly detection. The methods in this category first make assumptions of the distribution of the studied dataset. The samples with low probabilities are treated as anomalies. Rousseeuw and Driessen [16] proposed an anomaly detection method based on the Gaussian assumption of the data. The performance of statistical anomaly detection methods highly depends on the fitting between the assumption and the reality, thus exhibiting limited performance.

(2) Machine learning-based methods: the most widely used anomaly detection methods are the machine learning-based methods, which generally have two categories: supervised and unsupervised methods. Supervised methods [17, 18] refer to the models applying to the dataset that the training data are

labeled with "nominal" or "anomaly." The models are trained with the labeled data and use to identify new instances. Unsupervised methods deal with the dataset without labels. These methods automatically detect the anomalies based on certain criteria. Popular unsupervised methods include LOF [19], DBSCAN [20], k-means [21], and the isolation forest [22] method.

(3) Deep learning-based methods: the emerging deep learning models bring new opportunities to better solve the anomaly detection problem. Hundman et al. [23] propose an LSTM network-based framework for anomaly detection; [24] utilized a generative adversarial network (GAN) to detect the anomalies in time series data. Nguyen et al. [25] detect the anomalies by constructing the model snapshot and outputting the ensembles of the NN models. Deep learning-based methods tend to have more a promising performance compared to other techniques. However, these methods require a large amount of training data to produce reasonable results. Its performance is low in scenarios with a small set of training data, e.g., the fare machine-station association problem studied in this paper.

3. Methodology

3.1. Problem Formulation. Let m be a fare machine, and $S_m, \widehat{S}_m \in \Delta$ its actual station and current association station in the AFC dataset, respectively, where $\Delta = \{S^1, S^2, \ldots, S^s\}$ contains all the stations in the metro system. Note that different fare machines could share the same station, i.e., located in the same station. If $S = \widehat{S}$, fare machine m is defined as valid association fare machine; if $S \neq \widehat{S}$, fare machine m is defined as invalid association fare machine. The fare machine-station association detection problem is defined as follows.

Given an AFC dataset **D** and a set of fare machines Φ recorded in **D**, detect invalid association fare machines and infer their associated stations for fare machines m in Φ.

Mathematically, the problem is defined as follows:

(i) *Invalid Association Detection*. Find $\phi \subset \Phi$, s.t. $\phi = \left\{m_{S \longrightarrow \widehat{S}} | S \neq \widehat{S}\right\}$, and $\Phi - \phi = \left\{m_{S \longrightarrow \widehat{S}} | S = \widehat{S}\right\}$

(ii) *Station Inference*. For each fare machine m in ϕ, find $\widehat{S}$ s.t. $S = \widehat{S}$

3.2. Fare Machine Features. For convenience, we define the concept of fare machine-related passenger flow (MRF). For an entrance fare machine, MRF refers to the passenger flow tapping in an entrance fare machine of the origin station and tapping out at a destination station (using any machine) during a certain time slot. For an exit fare machine, MRF represents the passenger flow tapping in at an origin station (using any machine) and tapping out at an exit fare machine during a certain time slot. MRF can be characterized using different features, such as flow volume and travel time. Indicators extracted from the MRF features can be used to characterize fare machines. The hypothesis is that MRF features share more similar patterns if the fare machines are located at the same station than at different stations.

The flow volume and travel time are selected to characterize the MRFs of fare machines. These two features reflect system dynamics from both the demand (mobility patterns) and supply (network and operations) points of view as well as their interactions. They provide complementary knowledge and therefore give a more comprehensive view of the MRF patterns. They are defined for entrance and exit fare machines separately:

(i) For entrance fare machines, MRF flow volume measures the number of passengers passing through each fare machine at an origin station and going to a destination station. For exit fare machines, it represents the number of passengers entering the metro system at an origin station and tapping out through an exit fare machine. MRF flow volume reflects the mobility behavior of passengers.

(ii) MRF travel time indicates the average travel time from a fare machine to a destination station for entrance fare machines and from an origin station to a fare machine for exit fare machines. It reflects the supply characteristics of the metro system, e.g., geographical relationship between stations and scheduling, but also demand characteristics of certain stations as it includes time waiting to board a train under capacity constraints.

Figure 3 shows the overview of the proposed framework. It consists of three modules: MRF feature extraction, invalid association detection, and associated station inference:

(i) MRF feature extraction module: it constructs the MRF flow volume and travel time tensors to characterize fare machines and extracts latent MRF flow and travel time features using the tensor decomposition technique.

(ii) Invalid association detection module: it detects the invalid associations (between fare machines and stations) in two steps. The valid and invalid associations are initially detected using the isolation forest method. Then, the invalid associations are reinspected (the feedback arrow) using neural networks (trained with the valid association data).

(iii) Association station inference: it infers the station that a fare machine (detected as invalid association) belongs to using the trained neural networks.

3.3. MRF Tensor Construction. For data representation, tensors are used to characterize the MRF flow volume and travel time. A tensor is a high-order generalization of a matrix. The multiway property of a tensor fits the nature of MRF features. For example, MRF flow volume can be characterized by "machine mode" (M), "time mode" (T), "day mode" (D), and "station mode" (S). For entrance fare machines, "machine mode" denotes the related fare machine ID, "time mode" represents the time interval of a day (e.g., 6:

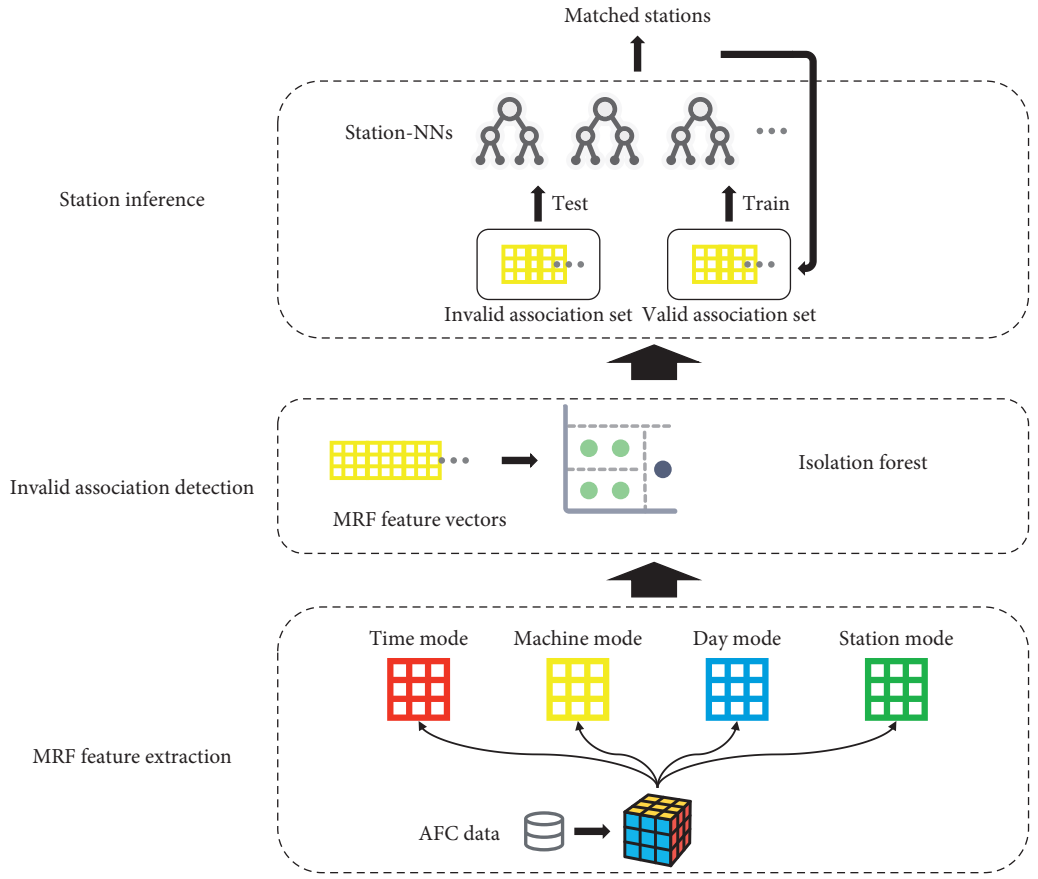

Figure 3: Overview of the proposed framework.

00 to 7:00 AM), "day mode" denotes the date, and "station mode" denotes a destination station ID. For exit fare machines, the definitions of tensor modes are the same with entrance fare machines, except for the "station mode." The "station mode" of an exit fare machine is the origin station ID. In this way, two 4-way tensors are used to represent the MRF flow volume of entrance and exit fare machines, respectively. For example, an entry: 50 at (A, 8:00 to 9:00 AM, January 1, B) of entrance machine tensor represents "the passenger flow volume passing through entrance machine A in the interval 8:00 to 9:00 AM on January 1 and exiting at Station B is 50 passengers." The methodology for fare machine-station association is the same for entrance and exit fare machines. Entrance fare machines are used to illustrate the proposed framework. Unless stated, the "fare machines" and "MRF tensors" refer to entrance fare machines and entrance MRF tensors, respectively.

To construct the MRF flow volume tensor, the mode variables above are transformed into numerical indices:

(i) Machine mode: the fare machines are labeled from 1 to M. Then, the machine IDs belong to a set $M = \{1, 2, \ldots, M\}$, where M represents the total number of fare machines.

(ii) Time mode: the hourly interval is used to represent the tap-in time $T = \{1, 2, \ldots, T\}$. Note that only the operating hours of the metro system are considered, where the i^{th} element in T denotes the i^{th} operating hour of the day.

(iii) Day mode: day mode represents the date, thus $D = \{1, 2, \ldots, D\}$ where 1 and D represent the first and the last day of the studied time span, respectively.

(iv) Station mode: the stations are labeled from 1 to S $S = \{1, 2, \ldots, S\}$, where S denotes the set of stations in the metro system.

The MRF flow volume is represented by a size $M \times T \times D \times S$ tensor $\mathcal{V}$. Figure 4 shows the structure of the MRF flow volume tensor. Each entry of $\mathcal{V}$, denoted as $\mathcal{V}_{mt\ ds}$, represents the MRF flow volume entering through fare machine m and exiting at destination station s during the t^{th} time interval of day d. For the exit fare machines, the tensor construction procedure is the same as the entrance machines. Accordingly, the entry $\mathcal{V}'_{mtds}$ denotes the MRF volume entering though station s and tapping out though fare machine m during the t^{th} time interval of day d.

Similarly, the MRF travel time tensor is denoted as $\mathcal{T} \in \mathbb{R}^{M \times T \times D \times S}$. An entry $\mathcal{T}_{mtds}$ in $\mathcal{T}$ represents the average travel time of all passengers entering through fare machine m and traveling to destination station s during the t^{th} time interval of day d.

The properties of MRF flow volume and travel time tensors are different, though they share the same structure. The difference stems from the tensor cells that have no AFC

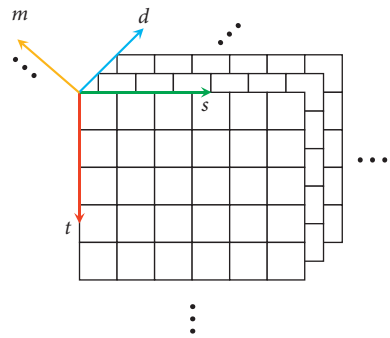

Figure 4: Structure of MRF flow volume tensor. MRF flow volume tensor consists of 4 modes, i.e., time (t) mode, day (d) mode, station (s) mode, and machine (m) mode.

observation. For the MRF flow volume tensor, the value of such cells is 0 since the MRF flow volume for the corresponding $[m, t, d, s]$ is 0 (no passenger flow). However, for the travel time tensor, cells having no observation cannot be directly filled with a zero. No observation in the MRF travel time tensor only means that no passengers traveled for the specified $[m, t, d, s]$ case. However, the corresponding travel time cannot be 0. An initial idea is filling these cells using the average travel time of such OD pairs in the historical data. Unfortunately, nonobservation cells always account for a large ratio of the MRF travel time tensor (e.g., 63.5% in the studied AFC dataset). Therefore, it is hard to estimate a reasonable average travel time for each cell based on limited information. Instead, "NaN" values are used to fill those cells to represent the unknown travel times.

3.4. Tensor Decomposition. Tensor decomposition is used to extract fare machine features from the MRF flow volume and travel time tensors. Given the different properties of these two tensors, different tensor decomposition methods are developed to extract the MRF flow volume and travel time features, respectively.

3.4.1. Tensor Decomposition of MRF Flow Volume. For MRF flow volume tensor $\mathcal{V}$, the CANDECOMP/PARAFAC (CP) decomposition [12] is used to extract the fare machine features. CP decomposition factorizes a tensor into a summation of a series of rank-1 tensors. A rank-1 tensor $\mathcal{V} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$ (I_i is the dimension of mode i) is an outer product of N vectors: $\mathcal{X} = \mathbf{a}^{(1)} \circ \mathbf{a}^{(2)} \circ \cdots \circ \mathbf{a}^{(n)}$, where $\mathcal{X}_{i_1 i_2 \ldots i_n} = a_{i_1}^{(1)} a_{i_2}^{(2)}, \ldots, a_{i_n}^{(n)}$, $\mathbf{a}^{(i)}$ denotes a vector, $a_k^{(i)}$ denotes the k^{th} element of $\mathbf{a}^{(i)}$, and the symbol $\circ$ denotes the outer product of vectors.

The CP decomposition of $\mathcal{V} \in \mathbb{R}^{M \times T \times D \times S}$ can be formulated as follows:

$$\widehat{\mathcal{V}} = \sum_{r=1}^{R} \mathbf{m}^r \circ \mathbf{t}^r \circ \mathbf{d}^r \circ \mathbf{s}^r, \tag{1}$$

where R represents the total number of components, $\mathbf{m}^r \in \mathbb{R}^M$, $\mathbf{t}^r \in \mathbb{R}^T$, $\mathbf{d}^r \in \mathbb{R}^D$, and $\mathbf{s}^r \in \mathbb{R}^S$ represent the component vector of the machine, time, day, and station modes, respectively. Figure 5 illustrates the process of CP decomposition of $\mathcal{V}$.

Computing the CP decomposition of $\mathcal{V}$ can be treated as an optimization problem. The goal is to find a CP decomposition $\widehat{\mathcal{V}} = \sum_{r=1}^{R} \mathbf{m}^r \circ \mathbf{t}^r \circ \mathbf{d}^r \circ \mathbf{s}^r$ with R components that could best approximate $\mathcal{V}$. The decomposition $\mathcal{V}$ is the solution of the following optimization problem, i.e., find

$$\widehat{\mathcal{V}}^* = \arg\min_{\widehat{\mathcal{V}}} \|\mathcal{V} - \widehat{\mathcal{V}}\|_F, \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This optimization problem can be solved using the alternating least squares (ALS) method [26]. Details of the solution procedure can be found in [12].

The feature matrix $\mathbf{M}_{\mathcal{V}} = [\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^R]$ is constructed utilizing all the component vectors $\mathbf{m}^r$ in $\widehat{\mathcal{V}}^*$. Since each entry in $\widehat{\mathcal{V}}^*$ is calculated as the outer product of all the 4 component vectors, $\mathbf{M}_{\mathcal{V}}$ could be treated as an indicator of the hidden information of all the other 3 modes. The entries in $\widehat{\mathcal{V}}^*$ that are related to the i^{th} fare machine are calculated only using the elements in i^{th} row of $\mathbf{M}_{\mathcal{V}}$. Therefore, each row of $\mathbf{M}_{\mathcal{V}}$ can be used as a latent feature vector to represent each fare machine's MRF flow volume pattern.

3.4.2. Tensor Decomposition of MRF Travel Time. CP decomposition cannot be applied directly to extract travel time features. This is because the travel time tensor has nonnumerical (i.e., NaN) entries, which makes the operation $\mathcal{T} - \widehat{\mathcal{T}}$ infeasible. A variation of CP decomposition, CP Weighted OPTimization (CP-WOPT) [27], is used to deal with the MRF travel time tensor decomposition. CP-WOPT is widely used to recover tensors with missing entries. CP-WOPT utilizes a weight tensor to indicate the location of NaN entries. The formulation is as follows:

$$\widehat{\mathcal{T}}^* = \arg\min_{\widehat{\mathcal{T}}} \|\mathcal{W} * (\mathcal{T} - \widehat{\mathcal{T}})\|_F. \tag{3}$$

The weight tensor $\mathcal{W} \in \mathbb{R}^{M \times T \times D \times S}$ has the same shape as $\mathcal{T}$ and is defined as

$$\mathcal{W}_{mtds} = \begin{cases} 0, & \text{if } \mathcal{T}_{mtds} \text{ is NaN}, \\ 1, & \text{otherwise}. \end{cases} \tag{4}$$

In the initialization phase, NaN cells are filled with random values. As these values are multiplied by 0 during the optimization, they do not influence the results of the optimization objective (optimal solution). After optimization, $\widehat{\mathcal{T}}^*$ can represent features of the observed travel time data well. As there exists strong relationship between the cells in $\mathcal{T}$, the features of the entries without observations can also be represented in the reconstructed tensor $\widehat{\mathcal{T}}^*$. A feature matrix $\mathbf{M}_{\mathcal{T}}$ is constructed using the machine mode component vectors in $\widehat{\mathcal{T}}^*$ to represent the multimode travel time features of fare machines. Details about the CP-WOPT method can be found in [27].

The MRF flow volume and travel time feature vectors of each fare machine are concatenated into one single vector to characterize the corresponding fare machine.

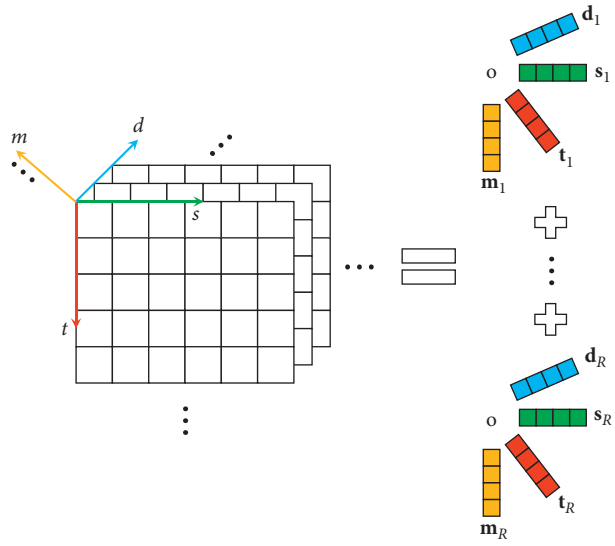

Figure 5: CP decomposition of the MRF flow volume tensor.

3.5. Fare Machine-Station Association. As fare machines at the same station share similar surrounding Point of Interests (POIs), the MRF features of these fare machines tend to be similar. Therefore, we should first extract the MRF feature of each station. Then, the MRF feature of each machine is compared to the station MRF feature. If a fare machine has a similar MRF feature with a station, then this station is likely to be the association station of the fare machine. We divide the inference process into two successive problems P1 and P2.

3.5.1. P1: Invalid Association Fare Machine Detection. To solve P1, we first give two assumptions: (1) the MRF features of the invalid associations are anomalies to their recorded stations. More formally, let $C(\cdot)$ be the count function, anomaly means $C(m_{S_1 \longrightarrow S_2}) \ll C(m_{S_2 \longrightarrow S_2})$ for $\forall S_1 \in \Phi, S_1 \neq S_2$. This indicates that the number of fare machines with association station S_1 but recorded station S_2 should be far less than the number of valid association fare machines in S_2. Note that this assumption does not mean the total number of invalid association fare machines of S_2 is less than the valid fare machines. We only restrict that fare machines recorded as S_2 but actually associated with S_1 should be minority to S_2. This assumption is reasonable since the error leads to fare machine-station invalid association tends to be random; for example, it is unlikely to have many fare machines located in the same station wrongly recorded as another station simultaneously. (2) The invalid associations happen randomly. This assumption indicates that for a fare machine m in station S, it experiences equal probability being wrongly associated to all the other stations in the system. This assumption is reasonable since the invalid associations mainly because of the inadequate data management in the process of database merging, maintenance, or system update.

Based on this assumption, the isolation forest method is adopted to solve P1. The isolation forest model is an unsupervised model for anomaly detection, which could be directly used for the contaminated dataset. The only requirement of this method is that the outlier should be few and different with the normal instances. This exactly fits the aforementioned assumption. The isolation forest detects the outliers using a special measurement: partitions. The isolation forest "isolates" observations by randomly selecting a dimension of the MRF feature vector and then randomly splitting the space between the maximum and minimum values of the selected dimension. Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate an MRF feature is equivalent to the path length from the root node to the terminating node. This path length, averaged over a forest of such random trees, is a measure of normality. Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular fare machines, they are highly likely to be anomalies [22].

Based on the results from the isolation forest, we can divide the fare machine MRF feature vectors into two parts:

$\mathbf{F}_\phi$ contains all the MRF feature vectors that are inferred as invalid (i.e., abnormal) by the isolation forest

$\widehat{\mathbf{F}}_\phi$ contains all the MRF feature vectors that inferred as valid (i.e., normal) by the isolation forest

The fare machines with their MRF features in $\widehat{\mathbf{F}}_\phi$ are detected as valid, while the fare machines in $\mathbf{F}_\phi$ are reinspected in the process of solving P2.

3.5.2. P2: Association Station Inference. In P2, a reinspection of the fare machines in $\mathbf{F}_\phi$ is conducted to refine the detection results from P1. The reinspection detects which associations are wrongly detected as invalid in $\mathbf{F}_\phi$. In practical applications, the inference provides a certain sense about the data quality in their AFC database. The model outputs the potential association stations of the detected invalid association fare machines, which facilities effective field investigation and reduces manpower.

Neural network (NN) is used to model the station MRF feature using the MRF features in $\widehat{\mathbf{F}}_\phi$ (detected as valid). As the number of samples (i.e., fare machines) are limited (e.g., 2000 fare machines in the studied network), the NN training may face underfitting issues. We built one shallow neural network for each station, which denotes as the station-NN. For a certain station-NN $\mathcal{N}_i$ of station S_i, we label the fare machines with the recorded station S_i in $\widehat{\mathbf{F}}_\phi$ as 1 and label other fare machines in $\widehat{\mathbf{F}}_\phi$ as 0. It is inadequate to directly train the station-NN with the labeled features. Since a metro system has many stations (e.g., 90 stations in the studied metro system), for one certain station, the number of positive samples (i.e., MRF features labeled as 1) is much less than the negative samples (MRF features labeled as 0), which will lead to the learning bias. We utilize the adaptive synthetic sampling (ADASYN) [28] approach to oversample the positive samples, ensuring that the number of the oversampled positive samples is similar with the number of negative ones. $\mathcal{N}_i$ is then trained with the oversampled MRF features and their corresponding labels. After $\mathcal{N}_i$ is well-

trained, the output of the network will be the probability that the input fare machine MRF feature belongs to this station.

For an MRF feature $\mathbf{v}$ in $\mathbf{F}_\phi$, we input it into all the well-trained station-NNs. Let $\mathbf{P} = [P_1, P_2, \ldots, P_S]$ denote the output probability from each station-NN, and $\mathbf{P}_\pi = [P_{\pi(1)}, P_{\pi(2)}, \ldots, P_{\pi(S)}]$ is the descend order permutation of $\mathbf{P}$, where $P_{\pi(i)} > P_{\pi(j)}$, given $i < j$. The top-k stations $\mathcal{K} = \{\pi(1), \pi(2), \ldots, \pi(k)\}$ would be the most possible association station of the corresponding fare machine of $\mathbf{v}$.

Using $\mathcal{K}$, the reinspection for P1 is conducted for the fare machine in $\mathbf{F}_\phi$ with the following rule: given a fare machine $m_{S\longrightarrow s} \in \mathbf{F}_\phi$, if $s \notin \mathcal{K}$, m is inferred as invalid, otherwise as valid. For the fare machines inferred as invalid after the reinspection, the top-k station $\mathcal{K}$ is treated as the potential association stations set. In the implementations, one can first check the stations in this set to find if this fare machine is there.

4. Case Study

We utilize AFC data from an urban metro system to evaluate the proposed detection and inference approach. The data cover 7 days from January 15 to 21 in 2018. The fare machine-station association information is carefully checked to ensure its validity for benchmarks. Figure 6 illustrates the statistic of the number of machines in the metro system during the studied time span.

4.1. Experimental Setup. We randomly select 1000 entrance fare machines and 1000 exit fare machines and collect the corresponding AFC transaction records to construct the experimental dataset. We randomly choose a set of fare machines and modify their associated stations (invalid associations). The proposed approach is validated with the ratio of invalid associated fare machines ranging from 5% to 40%. The approach runs 20 times per scenario to avoid random errors. Table 1 summarizes the model parameters used in the experiments.

4.2. Performance Evaluation. Table 2 shows the tabularised relations between truth/falseness of the detection and valid/invalid association.

A set of performance metrics is used to comprehensively evaluate the model performance, including accuracy (Accu), true positive rate (TPR), and false positive rate (FPR):

$$\text{Accu} = \frac{N_T}{N_{\text{PN}}}, \tag{5}$$

where N_{PN} is the total number of associations (or fare machines) and N_T the number of correctly detected associations (between fare machines and stations). The correctly detected fare machines include cases that are truly positive and negative:

$$\text{TPR} = \frac{N_{\text{TP}}}{N_P}, \tag{6}$$

where N_{TP} is the number of truthfully detected invalid association (correctly inferred an invalid association as invalid), and N_P is the number of invalid associations. TPR measures the model's sensitivity towards invalid associations:

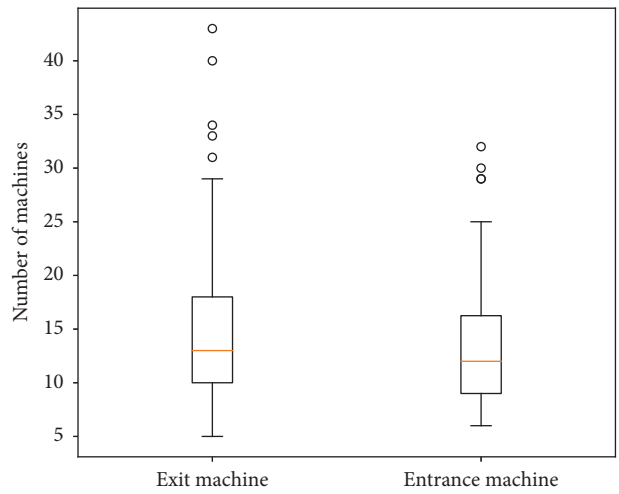

Figure 6: Number of entrance and exit fare machines in the studied metro system.

$$\text{FPR} = \frac{N_{\text{FP}}}{N_N}, \tag{7}$$

where N_{FP} is the number of falsely detected valid associations (falsely inferred a valid association as invalid) and N_N is the total number of valid associations. FPR measures the misjudgement rate of the valid associations.

4.2.1. Evaluation of Invalid Association Detection (P1). Figure 7 shows the detection results of associations with the invalid association ratio ranging from 5% to 40%. The results indicate that the isolation forest model is robust to the invalid associations when the invalid association ratio is less than 20% (the detection accuracy is over 96%). It can still achieve a detection accuracy of 87%, and even 40% of the fare machines are wrongly associated with stations in the data. The TPR is an essential characteristic of the detection of invalid associations in P1, since there is no reinspection of the invalid associations in $\widehat{\mathbf{F}}_\phi$ in the following procedures of the approach. That is, the wrongly associated fare machines in $\widehat{\mathbf{F}}_\phi$ will remain undetected which may eventually impact practical applications in reality. Also, it is favorable to detect more invalid associations to ensure a clean MRF feature set for each station, which benefits the correction of invalid associations in P2. The TPR is over 90% when the invalid association is less than 20%, which indicates the promising performance of the proposed approach in detecting the invalid associations. The falsely detected valid associations (FPR) is very low (less than 5%), and it decreases with the increase of the invalid association ratio as expected.

4.2.2. Evaluation of Association Inference (P2). For the P2 evaluation (rematching wrongly associated fare machines to stations), we quantify the model's capability to effectively

Table 1: Model parameters.

Tensor decomposition	Optimal value (potential values)
Number of components (R)	8 (1–15)
Optimization algorithm	ALS (ALS refers to the alternating least squares algorithm)
Error tolerance	$1e-6$ ($1e-3$–$1e-8$)
Maximum number of iterations	100 (10, 100, 500, 1000)
Isolation forest	Value
Threshold score (the threshold score is calculated with the *decision_function* in *sklearn.IsolationForest* package under *Python 3.7*)	0
Number of estimators	1000 (200, 500, 1000, 1500)
Station-NN	Value
The number of top stations in P1 reinspection	5
Number of hidden layers	2 (1–5)
Optimizer	Adam (Adam refers to the optimization algorithm proposed in [29])
Number of neurons	(16, 5)

Table 2: Confusion matrix of the valid association detection.

	Invalid association (positive)	Valid association (negative)
True detection (true)	True positive (TP)	True negative (TN)
False detection (false)	False positive (FP)	False negative (FN)

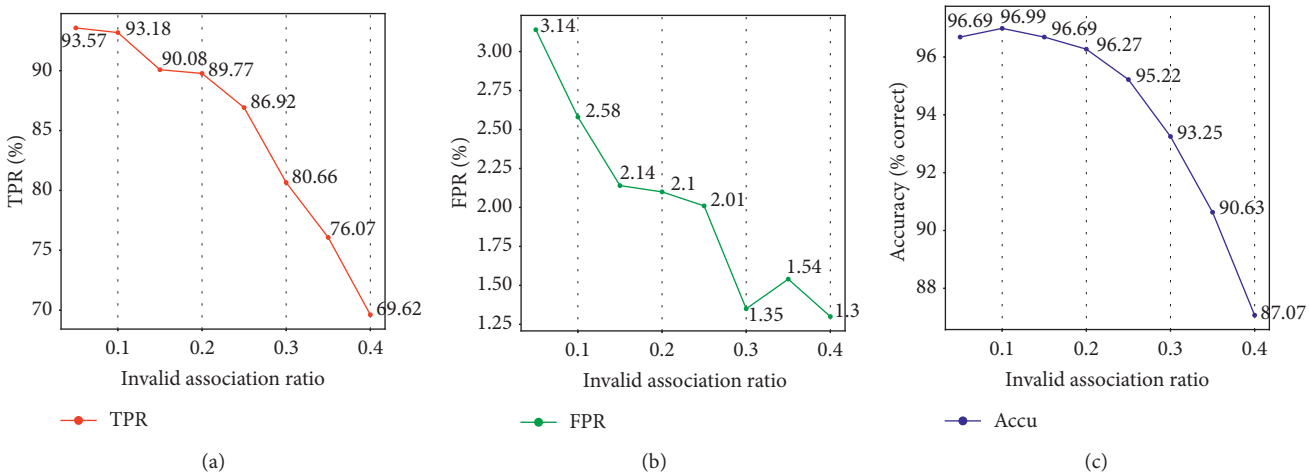

Figure 7: Model detection performance (a) TPR, (b) FPR, and (c) accuracy with invalid association ratio ranging from 0.05 to 0.6.

allocate large probabilities to the correctly matched stations. We use the top-k accuracy measure. Depending on different k values, it measures the probability that the inferred set of the top-k stations (ordered by probabilities) includes the actual associated station in reality:

$$\text{top} - k \text{ Accuracy} = \frac{N_c}{N_{\mathbf{F}_\phi}}, \tag{8}$$

where N_c is the number of fare machines in $\mathbf{F}_\phi^*$ with their matched station contained in $[\pi(1), \pi(2), \ldots, \pi(k)]$ and $N_{\mathbf{F}_\phi^*}$ is the number of fare machines in $\mathbf{F}_\phi^*$.

Table 3 summarizes the model performance of P2 with varied levels of invalid association ratios in the dataset.

Table 3: Model performance in rematching invalid associated fare machines.

	Invalid association ratio (%)							
	5	10	15	20	25	30	35	40
Top-1	76.4	77.8	77.1	78.9	75.7	74.1	70.9	69.1
Top-2	86.8	87.1	87.2	88.5	87.4	85.1	82.1	81.4
Top-3	90.8	91.1	91.7	91.8	91.7	89.9	88.3	87.7
Top-4	93.1	93.4	94.4	93.6	93.9	92.5	90.9	91.0
Top-5	94.1	95.0	95.9	95.3	95.1	94.1	92.9	93.3

The results show that the top-k accuracy exceed 90% when k is greater than 3, regardless the invalid association ratio. It indicates that the top 3 inferred stations from the

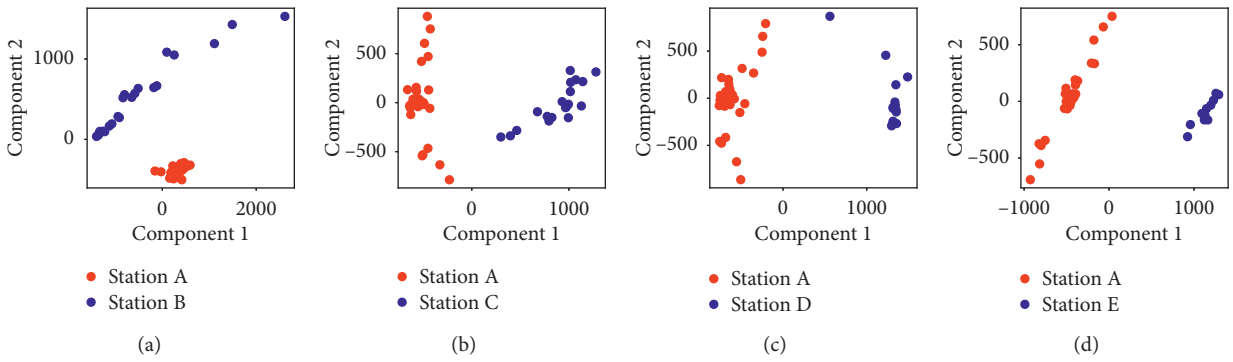

Figure 8: MRF feature vectors between Station A and (a) Station B, (b) Station C, (c) Station D, and (d) Station E.

model are highly likely to include the correctly associated station of the studied fare machine. This provides an important implication for further field investigations to these probable stations in practice, i.e., checking the most likely stations that the invalid associated fare machines may belong to.

4.3. Latent Feature Analysis. The foundation of the detection or inference model being effective is the quality of the MRF features. That is, the fare machines at different stations are preferable to have significantly different MRF features. To explore the feature quality, we utilize the principle component analysis (PCA) [7] to reduce the dimension of the MRF feature vector to two. We randomly choose 5 stations in the studied metro system, select one station as the reference station, and compare its MRF feature vector to that of the other 4 stations, respectively.

Figure 8 shows the MRF feature visualization results. The results show that the MRF features between stations exhibit significant differences, which indicates a high quality. This benefits the model to formulate relatively distinct MRF feature for each station, thus which is effective to detect the invalid associations and infer the associated station of the fare machines. For different stations, the MRF feature of fare machines appears different patterns. For example, the MRF features of machines in Station E (Figure 8(d)) are very similar to each other, while the MRF features of Station B (Figure 8(a)) appear a distributed manner. The reason partly lays in the different layout of the stations. For some large stations (e.g., transfer stations in the commercial center), there are many gates entering/exiting the stations, which may lead to variances in travel time between the same OD pairs. It would be the main reason for the miss and wrongly detection of the proposed model.

5. Conclusion

Ensuring data quality is essential for its effective use in practice. The paper proposes a model to detect the invalid data in the AFC dataset, caused by the erroneous association between fare machines and stations (e.g., due to delayed updating dictionaries or incorrect data merging). It combines tensor decomposition, isolation forest, and NN methods to detect the invalid associations in the recorded dataset and infer the correct association station that a fare machine belongs to.

The model is validated using the AFC data in a busy metro system. The experiment results show that the invalid association can be detected with more than 90% accuracy when the invalid association ratio is low. Also, the model is robust to invalid associations and it can still achieve 69.62% accuracy in the extreme case when the invalid association ratio is 55%. The association station inference results indicate that the top 3 inferred stations from the model are highly likely to include the correctly associated station of the studied fare machine (around 90%). This provides an important implication for further field investigations to these probable stations in practice.

The proposed model provides useful knowledge for the AFC data management in terms of data quality check and fixing invalid data. Though the study focuses on the invalid data detection problem, the model is general and can be generalized to inference applications, e.g., inferring the alighting stations for the bus system having only the boarding records. As the extracted MRF features are meaningful, further studies could focus on the analysis based on the MRF features, for example, analysing the different utilization of fare machines in different gates of the same station to improve the infrastructure efficiency.

Data Availability

The AFC data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. N. Koutsopoulos, Z. Ma, P. Noursalehi, and Y. Zhu, "Transit data analytics for planning, monitoring, control, and information," in *Mobility Patterns, Big Data and Transport Analytics*, pp. 229–261, Elsevier, Amsterdam, The Netherlands, 2019.

[2] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world,"

Proceedings of the VLDB Endowment, vol. 2, no. 1, pp. 562–573, 2009.

[3] S. Robinson, B. Narayanan, N. Toh, and F. Pereira, "Methods for pre-processing smartcard data to improve data quality," *Transportation Research Part C: Emerging Technologies*, vol. 49, pp. 43–58, 2014.

[4] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.

[5] Y. Liu, X. Weng, J. Wan, X. Yue, H. Song, and A. V. Vasilakos, "Exploring data validity in transportation systems for smart cities," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 26–33, 2017.

[6] D. Everett: 60000 Oyster Cards Corrupted, [EB/OL], 2008, https://www.yumpu.com/en/document/read/27117143/60000-oyster-cards-corrupted-smart-card-news.

[7] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[8] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[9] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[10] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, Vol. 1, MIT Press, Cambridge, UK, 2016.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[13] H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li, "A tensor-based method for missing traffic data completion," *Transportation Research Part C: Emerging Technologies*, vol. 28, pp. 15–27, 2013.

[14] X. Chen, Z. He, and J. Wang, "Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition," *Transportation Research Part C: Emerging Technologies*, vol. 86, pp. 59–77, 2018.

[15] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transportation Research Part B: Methodological*, vol. 91, pp. 511–524, 2016.

[16] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[17] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 427–438, Dallas, TX, USA, May 2000.

[18] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 15–27, Springer, Helsinki, Finland, August 2002.

[19] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, Dallas, TX, USA, May 2000.

[20] M. Çelik, F. Dadaşer-Çelik, and A. Ş. Dokuz, "Anomaly detection in temperature data using DBSCAN algorithm," in *Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 91–95, IEEE, Istanbul, Turkey, June 2011.

[21] G. Münz, S. Li, and G. Carle, "Traffic anomaly detection using k-means clustering," in *Proceedings of the GI/ITG Workshop MMBnet*, pp. 13-14, Hamburg, Germany, September 2007.

[22] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, IEEE, Pisa, Italy, December 2008.

[23] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 387–395, London, UK, August 2018.

[24] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "Mad-gan: multivariate anomaly detection for time series data with generative adversarial networks," in *Proceedings of the International Conference on Artificial Neural Networks and Machine Learning-ICANN 2019: Text and Time Series*, pp. 703–716, Springer, Munich, Germany, September 2019.

[25] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "Self: learning to filter noisy labels with self-ensembling," 2019, https://arxiv.org/abs/1910.01842.

[26] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[27] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41–56, 2011.

[28] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, IEEE, Hong Kong, China, June 2008.

[29] D. P. Kingma and B. J. Adam, "A method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.