Large-Scale Loop Detector Troubleshooting Using Clustering and Association Rule Mining

Amin Ariannezhad[1] and Yao-Jan Wu[2]

Abstract: The archived data from traffic sensors are used in a wide range of traffic management applications. However, missing or invalid data are becoming an important concern. This study proposes a systematic approach to identify and characterize data error patterns to facilitate large-scale loop detector troubleshooting. Data were collected from loop detectors in Phoenix. A set of quality control criteria was applied on daily 20-s data to find the error percentage for each loop detector. A fuzzy *c*-means clustering method was implemented on the data quality check results and preliminary clusters were identified. To discover the most frequent rules within the clusters, an association rule mining method was applied to the clusters' data subsets. Loop detector stations with different error patterns were visited in the field to verify the clustering and association rule mining results, identify potential causes, and recommend appropriate solutions. The analysis identified four key patterns, indicating that the proposed approach successfully found the relationships in the data errors. The findings of this study help traffic engineers to more easily diagnose and troubleshoot large-scale loop detector errors. **DOI: 10.1061/JTEPBS.0000387.** © *2020 American Society of Civil Engineers.*

Author keywords: Loop detectors; Data quality check; Fuzzy *c*-means clustering; Association rule mining; Error patterns.

Introduction

With a growing number of intelligent transportation system (ITS) sensors or detectors deployed across the nation's roadway facilities, the amount of data agencies are dealing with is increasing dramatically. Among various types of traffic detectors, inductive loop detectors (ILDs) can produce volume, occupancy, and average speed of vehicles along the freeway networks. Therefore, they are widely used in the US as the primary source of real-time freeway traffic data (Klein 2001; Klein et al. 2006). The data obtained from these sensors are archived by traffic operations centers and used for a great many applications in traffic management, including transportation planning, incident and congestion monitoring, performance measurement, and travel time analysis. A survey conducted by the Virginia Transportation Research Council indicated that 87% of the data in traffic management systems comes from ILDs (Turochy and Smith 2000).

A study on nearly 30 US cities by the mobility monitoring program of the Federal Highway Administration (FHWA) (Turner 2004) indicated that the archived data from ILDs suffered from invalid and missing data records in almost every city. Traffic data from ILDs are used as inputs for many data-driven applications. Therefore, incomplete or invalid data collected from these sensors can lead to unreliable analyses, results, and conclusions. The fundamental role of traffic data in different components of traffic management systems reveals the importance of performing systematic data quality checks.

Because of the contributions of previous studies, transportation agencies are now expected to be able to flag the erroneous data using data quality control criteria. The flagged data points as well as the missing data can be imputed using different techniques. However, even with the best practices in data imputation, it would still not be possible to impute erroneous data with 100% accuracy. Therefore, it is extremely important to identify and troubleshoot data errors at the source. Detecting and fixing all ILD data errors is a very time-consuming and costly process for traffic engineers because an individual ILD might report several different types of errors during the course of a single day.

This study aims to develop a systematic approach for investigating the relationship between different errors in ILDs to ease their troubleshooting and repair process. Troubleshooting the ILD errors could be a complex process: not only is it possible for a specific malfunction source to cause several different types of errors, each type of ILD error could be caused by several malfunction sources. Finding the possible relationships between these errors and causes is thus often quite challenging. The majority of previous studies have tried to identify ILD errors and possible causes using techniques either at macroscopic or microscopic levels; few have investigated the possible relationship between the errors by observing the data. This study, by utilizing clustering and association rule mining techniques, identifies error patterns, relationships between errors, and the specific source of the malfunction for each pattern to help traffic engineers troubleshoot and correct ILD errors. Each pattern includes either an error that occurs frequently, independent from others, or multiple errors that are frequently observed together, thus representing a state of loop detectors that was caused by a specific malfunction source or several malfunction sources.

The main contribution of this study is to help engineers troubleshoot ILD errors by highlighting the relationships between the errors, linking those that are similar, and identifying the source of the malfunction. This approach will enable traffic engineers to focus on a few error patterns in which errors are caused by a similar source, instead of looking into several individual errors. Traffic agencies

[1]Ph.D. Candidate, Dept. of Civil and Architectural Engineering and Mechanics, Univ. of Arizona, 1209 E 2nd St., Room 324G1, Tucson, AZ 85721 (corresponding author). ORCID: https://orcid.org/0000-0001-6679-7428. Email: ariannezhad@email.arizona.edu

[2]Associate Professor, Dept. of Civil and Architectural Engineering and Mechanics, Univ. of Arizona, 1209 E 2nd St. Room 324F, Tucson, AZ 85721. Email: yaojan@email.arizona.edu

Note. This manuscript was submitted on June 30, 2019; approved on February 19, 2020; published online on May 12, 2020. Discussion period open until October 12, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Transportation Engineering, Part A: Systems***, © ASCE, ISSN 2473-2907.

could allocate their limited resources to fix loop detectors more efficiently by prioritizing the patterns based on their importance, the frequency with which they occur, and the complexity of their repair process.

Literature Review

An ever-increasing volume of traffic data is being collected by traffic agencies along the networks of freeways and arterials. This could is collected from several different sources, including (1) sensors, such as video detectors, radars, and loop detectors; and (2) probe data from floating cars, GPS devices, and mobile applications. However, the data sets from all these sources include missing or invalid records. Extensive efforts have been made by previous researchers to impute missing or invalid traffic speed (Rodrigues et al. 2019), traffic volume (Chen et al. 2020), and travel time (Karimpour et al. 2019) data records.

Ma et al. (2017) used data collected from taxis equipped with GPS devices to predict traffic speed by using spatiotemporal adjacent records to impute missing values in the data. Chen et al. (2019) used a Bayesian probabilistic imputation framework to impute spatiotemporal traffic speed data using large-scale traffic speed data collected from a smartphone navigation application in Guangzhou, China. Missing traffic speed data records have been imputed in several more studies using various other data sources, such as crowdsourced traffic data (Rodrigues et al. 2019) and remote traffic microwave sensors (RTMSs) (Bae et al. 2018).

Chen et al. (2020) proposed utilizing synthetic and real traffic data in parallel to impute traffic flow data using a whole year of 5-min volume data from 147 loop detector stations in District 5, California. In a similar study, Duan et al. (2016) proposed a deep learning model named denoising stacked autoencoders to impute traffic volume data. This proposed model was evaluated considering temporal and spatial factors using loop detector data from California. Ran et al. (2016) used 4 weeks of data from 11 loop detectors stations in California to impute the missing traffic volume data using a low-rank tensor completion algorithm. Laña et al. (2018) used two machine learning techniques to impute 15-min aggregated traffic flow data. The data were collected from 3,600 automatic traffic recorders (ATRs) in Madrid, Spain.

Among different data sources, ILDs are one of the most commonly used sources for traffic data forecasting (i.e., speed, volume, occupancy, and travel time) (Vlahogianni et al. 2014). Therefore, it is important to diagnose the sources of erroneous and missing data in ILDs for further troubleshooting. Several types of malfunctions in ILDs could cause invalid or incomplete data. Bickel et al. (2007) found that communication errors or hardware breakdowns could both cause missing data in ILDs, while Payne et al. (1976) identified several detector faults responsible for invalid data, including hanging on or off, sensor stuck on or off, chattering, pulse breakup, and intermittent malfunctions. The definitions of these faults are as follows:

- Hanging on or off: The detector produces a very long (hang-on) or very short (hang-off) pulse. A hang-on detector causes high occupancy, while a hang-off detector causes low occupancy.
- Sensor stuck on or off: The detector is on or off for too long.
- Chattering: The detector produces very rapid and short pulses when a vehicle passes. A vehicle could be counted several times; therefore, the loop detector reports high volume.
- Pulse breakup: Instead of one pulse, the detector produces two or more pulses as a vehicle passes; therefore, the loop detector reports high volume.

Researchers have devoted considerable efforts to detect errors reported by ILDs, investigate the possible causes, and fix the malfunctions. Methods have been developed to detect ILD errors at both the macroscopic and microscopic levels (Lu et al. 2008). Macroscopic tests evaluate the aggregated ILD data for specific periods, for example, 20-s, 1-min, or 5-min, while microscopic checks seek to identify the errors based on the raw loop pulse signal collected from the loop card or loop on/off time instant. At a macroscopic level, several studies have sought to check the validity of data based on either univariate or multivariate range checks (Chen et al. 2003; Hu et al. 2001; Ishimaru 1998; Jacobson et al. 1990; Schmoyer et al. 2001; Turner 2004, 2007; Turochy and Smith 2000). Jacobson et al. (1990) proposed several validity criteria for single ILDs in Washington State using a two-stage process. The first part of their algorithm utilized 20-s data checks to detect errors caused by chattering detectors and intermittent ILD failures such as short pulses, and the second part used 5-min data to identify severe hanging-on malfunctions, which they defined as an occupancy value above 90%.

Ishimaru (1998) defined five acceptable ranges of volume and occupancy values for 20-s data collected from single ILDs. Turochy and Smith (2000) developed several tests based on threshold values and traffic flow theory. They defined feasible ranges for minimum and maximum average effective vehicle length (AEVL) to capture unreasonable volume and speed values. Chen et al. (2003) proposed a detector diagnostics algorithm to identify the bad ILDs in the California freeway Performance Measurement System (PeMS) data archive. The main difference between their approach and those proposed in previous studies was the use of a sequence of samples from each detector over a whole day, instead of just single samples. They checked the validity of each 20-s data batch based on four criteria developed for samples collected between 5:00 a.m. and 10:00 p.m. ILDs were then classified as either good or bad at the end of the day, depending on the number of invalid data rows that exceeded specific thresholds over the course of the day. They found that the data errors could be attributable to various ILD malfunctions such as detector stuck on or off and a loop hanging on or off, suggesting that an additional criterion for checking the high volume counts would be beneficial in detecting pulse breakup malfunctions.

Chen et al. (2019c) used one loop detector station in Wisconsin to design a procedure for flagging invalid traffic data records (volume, speed, and occupancy). They considered basic validity checks, multivariate range checks, and temporal consistency checks as potential tests for flagging traffic data. These validity checks were combined with a user survey to incorporate the users' preferences. Finally, they proposed a flagging procedure consisting of all types of tests to prioritize validity tests in a sequential manner.

At the microscopic level, several researchers, mostly in California and Washington State, have opted to analyze event-based data at the controller level because, as Lu et al. (2008) pointed out, analyzing data at a macroscopic level can only diagnose loop faults indirectly. A number of different ILD faults, including cross talk (Cheevarunothai et al. 2005; Coifman 1999), improper card sensitivity (Cheevarunothai 2006; Cheevarunothai et al. 2005), pulse breakup (Coifman and Lee 2011; Lee and Coifman 2011), splash over (Coifman and Lee 2011), and chattering (Lu et al. 2010) have therefore been investigated at a microscopic level.

Framework of the Proposed Approach

The framework of the proposed approach is presented in Fig. 1. The general idea of this paper is to place different ILD errors into different clusters using a clustering method, then apply an association

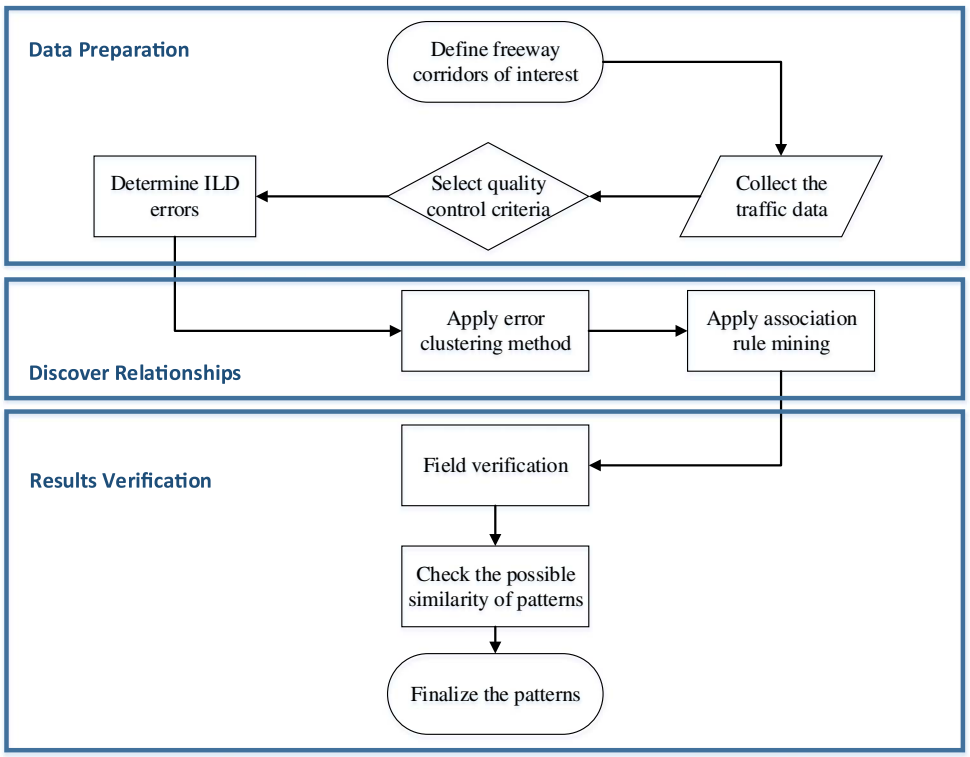

Fig. 1. Framework of the proposed approach.

rule mining technique on errors in different clusters to discover their relationship. The reasons for implementing association rule mining on data subsets resulted from the clustering method are (1) applying association rule mining on the original data set with all errors will result in a very high number of rules and thus will be hard to interpret, and (2) the clustering method partitions the data based on the value of errors in each observation while association rule mining only considers what errors exist in each observation without considering the error percentage. The following is a description of each step in the framework.

1. Data preparation:
   - Define freeway corridors of interest: Because the approach in this study is scalable, one or multiple corridors in the ITS database could be selected as long as these corridors are geospatially close.
   - Collect the traffic data: 1 month of high-resolution data (i.e., 20-s data) is favorable because a monthly analysis is recommended to summarize the error patterns.
   - Select quality control criteria: A panel of traffic engineers who have hands-on experience with the data should be recruited to select appropriate quality control criteria.
   - Determine ILD errors: Determine the percentage of daily errors based on the criteria for each ILD.
2. Discover relationships:
   - Apply error clustering method: This is a major step of the proposed approach, allowing the ILD error patterns to be identified using a clustering method.
   - Apply association rule mining method: If the relationship between the errors in any of the patterns is unknown, association rule mining is utilized to unveil the relationship between the variables.
3. Results verification:
   - Field verification: In an unsupervised learning algorithm, it would not be possible to validate the results with a test data set. Therefore, visiting the representative detector stations for each pattern can help to verify the errors in each cluster and identify the source of a malfunction. Once the underlying cause has been identified, an appropriate solution can be developed and implemented in other similar ILDs.
   - Finalize the patterns: Fine tune the error patterns based on the findings from the monthly analysis of error patterns and field visits. Once the most common error patterns have been identified, troubleshooting time in the future should be significantly reduced, although because error patterns do vary with time, the entire process should be regularly revisited by continuing to collect new data sets on a monthly basis.

Data Description

Dual-loop detector data collected by the Arizona Department of Transportation (ADOT) were utilized in this study. A dual-loop detector consists of two single-loop detectors placed several meters away from each other. Unlike single-loop detectors, dual-loop detectors are able to calculate a vehicle's speed fairly accurately based on the distance between their two single loops and the time duration between when a vehicle traverses from the first loop to the second one (Wang and Nihan 2003). The ADOT Traffic Operations Center (TOC) archives the 20-s data from all the dual-loop detectors located along freeways in the Phoenix area, a total of 377 stations, every day. Each station covers several mainlines and a high-occupancy vehicle (HOV) lane; each lane has a dual-loop detector that incorporates two loops: leading and lagging. Each individual loop in a dual-loop detector is referred to as a slot. A total of 102 ILD stations along the I-10 and I-17 corridors were selected for this study because these are major corridors carrying the inbound and outbound traffic for the Phoenix central business district (CBD). Along these two corridors, a total of 496 ILDs, equivalent to

Table 1. Quality control criteria

Criteria	Description	Possible reason
1	Volume = 0 and speed > 0	—
2	Occupancy >95%	Detectors stuck on
3	Volume = speed = 0 and occupancy >0%	—
4	Volume >0 and speed = 0	Speed trap not functioning properly
5	Speed > 90 mi/h	—
6	Speed < 5 mi/h	—
7	Occupancy = 0% and volume > $\text{Volume}_{(\max)}$[a]	Software truncates or rounds to integer value
8	Volume > 17	—
9	Density[b] >220 vehicles/mi	Improbable combinations of volume and speed
10	Min AEVL[c] (AEVL < 9 ft)	—
11	Max AEVL (AEVL > 60 ft)	—
12	Missing data	Communication interruption or breakdown

[a]$\text{Volume}(\max) = [(2.932 \times \text{speed} \times \text{elapsed time})/600$; based on maximum possible volume when occupancy value is truncated to zero.
[b]$\text{Density} = [\text{volume} \times (3,600/\text{elapsed time})/\text{speed}]$.
[c]$\text{AEVL} = 10 \times \text{speed} \times \text{occupancy}/\text{hourly equivalent flow rate (vehicles/h/lane)}$.

992 slots, are installed on the main lines and HOV lanes. To conduct a thorough check, data collected for the entire month of May 2016 were used for a total of 30,752 slot-days, each with 4,320 records of 20-s data. The reasons for selecting a month of loop detector data with 30,752 data records were to ensure that (1) any combination of errors that may be observed in a loop detector was included in the data, and (2) a sufficient number of samples was utilized for training the models while a computationally efficient model was also obtained.

Selecting Quality Control Criteria

The set of quality control criteria for the ILD data was selected based on two considerations: (1) quality control criteria documented in the literature, and (2) errors found in the preliminary analysis of data.

Table 1 lists the quality control criteria implemented in this study and their possible causes. The screening process consisted of 12 criteria, most of which were recommended by the FHWA's Mobility Monitoring Program (Turner et al. 2004) and previous studies (Turner 2007; Turochy and Smith 2000). These criteria include both threshold value checks and traffic flow theory–based checks.

Several criteria related to single-variable threshold checks were selected to test the validity of the ranges chosen for speed, volume, and occupancy records. Based on the traffic flow characteristics, it would be deemed unreasonable for an ILD to report an occupancy greater than 95%, a volume count greater than 17, and a speed outside the range of 5–90 mi/h. Four criteria were selected to check the relationship between the volume, occupancy, and speed values. These tests have the properties of both threshold value checks and traffic flow theory–based checks. Infeasible combinations of speed, volume, and occupancy records were identified through these four criteria for use in several traffic management systems because of their straightforward logic. A check based on traffic flow theory was selected for evaluating the density calculated based on volume and speed values. Improper combinations of speed and volume records result in infeasible density values.

Criteria 10 and 11 were chosen from a study by Turochy and Smith (2000) to check the feasibility of AEVL based on traffic flow theory. AEVL is a function of occupancy, volume, and speed values and is calculated from the data using traffic flow theory principles (see Table 1 for the AEVL formula). The final test checks for missing 20-s data records. Turner (2004) suggested completeness as a data quality measure because this measures the number of available data values relative to the number of total possible values. Therefore, missing data records were considered an error in this study.

The criteria used in this study were the known errors in ILD data; either the out-of-range values or impossible combinations of the occupancy, volume, and speed values. There might be other types of errors in the data that could not be captured by these criteria. To detect any unknown errors, the real traffic data (including vehicles occupancy, volume, and average speed) for these ILD stations need to be collected, which is out of the scope of this study.

Data Quality Check Findings

The selected criteria were applied to all 30,752 slot-days (a month of data in May 2016). The analysis showed that 13.33% of slot-days reported erroneous data, of which 22.16% met more than one criterion. These slot-days were considered as faulty slots. The criteria were then ranked based on the frequency with which they were observed (Table 2).

These findings indicate that the most common error was missing values because 62.33% of all the faulty slots suffered from incomplete data. Three other types of errors, namely, occupancy greater than 95%, occupancy greater than zero when volume and speed values are zero, and speed equals zero when volume count is greater than zero, were observed several times in the data. The first two of these errors were usually observed at the same time in the ILDs, indicating that occupancy was greater than 95% when the other error (occupancy greater than zero while speed and volume are zero) also occurred. Other types of errors were found comparatively less frequently in the data.

Methodology

Fuzzy c-Means Clustering

Fuzzy c-means clustering is an unsupervised learning method based on fuzzy logic, an extension of k-means clustering (Dunn 1973). The advantage of this method for k-means is that it does not consider sharp boundaries between the clusters, so individual data points can belong to two or more clusters. The degree of membership of each data point to a cluster depends on its distance to the cluster center, meaning that a data point could be assigned to the cluster with the highest membership level but still retain a lower degree of membership in one or more other clusters. For each data

Table 2. Data quality criteria ranking based on their frequency of occurrence

Criteria	Description	Frequency observed in the faulty slot-days (%)	Frequency observed in all slot-days (%)
1	Missing data	62.33	8.30
2	Volume > 0 and speed = 0	16.89	2.25
3	Occupancy >95%	16.15	2.15
4	Volume = speed = 0 and occupancy >0%	15.30	2.04
5	Min AEVL (AEVL < 9 ft)	8.90	1.19
6	Max AEVL (AEVL > 60 ft)	4.33	0.58
7	Speed > 90 mi/h	3.07	0.41
8	Volume > 17	1.96	0.26
9	Speed < 5 mi/h	1.90	0.25
10	Occupancy = 0% and volume > volume (max)	0.78	0.10
11	Density > 220 vehicles/mile	0.63	0.08
12	Volume = 0 and speed > 0	0.00	0.00

point, the degrees of membership to all the clusters sum to 1. The fuzzy c-means method is based on the minimization of the following objective function (Bezdek et al. 1984; Dunn 1973; Thambusamy 2014):

$$J_m = \sum_{i-1}^{N}\sum_{j-1}^{C} u_{ij}^m \|x_i - C_j\|^2, 1 \le m < \infty \quad (1)$$

where u_{ij} = degree of membership of x_i in the cluster j; m = any real number greater than 1; $c_j = d$-dimension center of the cluster ($d = 12$, the number of data quality checks in this study); $x_i = i$th d-dimensional measured data; and $\|x_i - C_j\|$ = Euclidean distance between the ith data and jth cluster center. Each data point, x_i, in this study is a slot and the daily error percentage for different criteria are the features for this data point. Fuzzy partitioning is based on the k iterative optimization of the objective function in Eq. (1). Degree of membership and cluster centers are updated in each iteration through Eqs. (2) and (3), respectively

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - C_j\|}{\|x_i - C_k\|}\right)^{2/(m-1)}} \quad (2)$$

$$C_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \quad (3)$$

The iteration will stop when $\max_{ij}\{|u_{ij}^{k+1} - u_{ij}^k|\} < \beta$, where β is a termination criterion between 0 and 1. The steps used in this method are

1. Initialize $U = [u_{ij}]$ matrix, $U^{(0)}$;
2. At k step, calculate the clusters centers with $U^{(k)}$;
3. Update $U^{(k)}$, $U^{(k+1)}$; and
4. Repeat Steps 2 and 3 until $\|U^{(k+1)} - U^{(k)}\| < \beta$.

Association Rule Mining

Association rule mining is a data mining technique for discovering local patterns and relations between variables in a data set. This technique discovers strong rules in a database using some measures of interestingness (Ghosh and Nath 2004). By using the concept of $X \Rightarrow Y$ (if X then likely Y), association rule mining could be described as follows (Agrawal et al. 1993):

Given

- $I = \{i_1, i_2, \ldots, i_n\}$ is a set of n attributes (ILD data errors in this study), which is called items; and
- $D = \{t_1, t_2, \ldots, t_m\}$ is a set of n observations (slot-days in this study), which is called the database

then each slot-day in the database D is unique and has a subset of attributes (errors) in I. In this study, each association rule is composed of two sets of attributes, also called error sets, and is defined with the form $X \Rightarrow Y$, where X is called the antecedent or left-hand side (LHS), Y is called the consequent or right-hand side (RHS), $X, Y \subseteq I$, and $X \bigcap Y = \emptyset$.

In association rule mining, measures of interestingness are used to find interesting patterns in the data. Interestingness measures how interesting the rules are. Two common measures of interestingness are support and confidence. Support of a rule is defined as the percentage of the observations (slot-days) with both error sets of X and Y to the total number of observations in the database. Confidence of a rule is defined as the percentage of the observations with both error sets of X and Y to the total number of observations that contain Error Set X. Therefore, association rule mining is the process of discovering all the interesting rules with support and confidence values higher than a user-specified minimum.

The Apriori algorithm is the most common method for discovering the association rules in a database. Although it is a robust, clear, and simple algorithm, it has its own limitations. The approach with which the Apriori algorithm searches in the database to find frequent patterns can take excessive amounts of time. Also, this algorithm scans the database several times to find the patterns, which could be inefficient and computationally expensive when the memory is limited (Al-Maolegi and Arkok 2014). Despite these limitations, considering the number of data points used in this study, this algorithm was utilized to find the frequent rules in the error sets. The details of this algorithm could be found in the study by Agrawal and Srikant (1994).

Clustering Results

Determining the number of clusters is an important step in clustering methods. The elbow method is the most well-known approach to determine the optimal number of clusters, by evaluating the homogeneity within the clusters (Hothorn and Everitt 2014). Fig. 2 illustrates the curve where total within-groups sum of squares (WSS) were plotted for different numbers of clusters. As depicted in the figure, the curve's slope is steep in the beginning; as the number of clusters increases, the slope becomes shallow. The optimal number of clusters is the point after which the WSS does not change significantly. Based on Fig. 2, six is the optimal number of clusters in this study, thus the errors were preliminary categorized into six clusters for further analysis.

The percentage of daily errors based on all criteria was found for each slot in the data quality check section. The fuzzy c-means

J. Transp. Eng., Part A: Systems, 2020, 146(7): 04020064

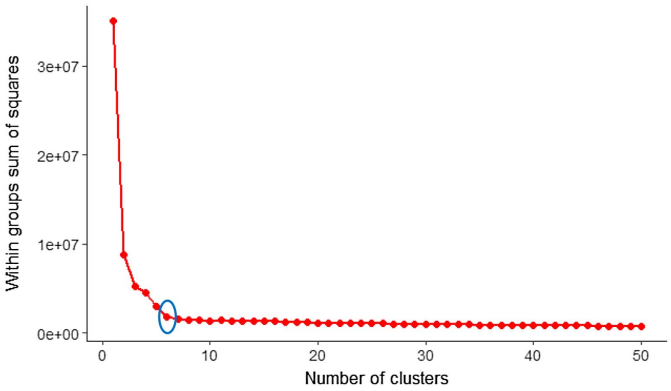

Fig. 2. Within-groups sum of squares for different numbers of clusters.

clustering method was implemented on the data quality check results to reveal six patterns in the data errors. The most similar pattern to each pattern was also found through the degree of membership of each observation for all the clusters. Table 3 presents the six patterns identified in the data errors.

Pattern 1 includes two major types of errors: (1) volume and speed equal to zero and occupancy greater than zero, and (2) occupancy greater than 95%. The erroneous data based on these two criteria were less than 70%. The error percentages for these two criteria were almost the same; when speed and volume were zero and occupancy was greater than zero in a 20-s data record, in most cases the occupancy in that record was greater than 95%, generally 100%. The correlation between these two criteria was found to be 99%. This error is usually observed because of a stuck detector. When the detector is on too long, it reports too many data samples with high occupancy.

The types of errors exhibiting Pattern 2 are similar to those with Pattern 1 but here the percentage error based on the two criteria was more than 70% (usually 100%). Also, no other error was observed in this pattern. As with Pattern 1, the most likely reason for the errors with this pattern would be a detector that was stuck in the on position. However, the important difference between the two is that the detector malfunction for Pattern 1 is intermittent, while for Pattern 2 it is systematic. The results found by the fuzzy *c*-means clustering method indicated that the nearest cluster for the observations in Pattern 1 was Pattern 2 and vice versa, showing the possible relationship and similarity of these two patterns.

Patterns 3 and 4 are similar in terms of error type. The only error found in both of these patterns is missing data records. Pattern 3 had 100% missing data in a day, while Pattern 4 had some percentage of missing data records less than 100%, with no other error. The fuzzy *c*-means analysis revealed that each pattern was the nearest cluster for observations in the other pattern, strongly suggesting the theoretical similarity of these two patterns. The reason for reporting no data or incomplete data during a day is likely to be a communication failure, which could occur at any point between the loop and the traffic operations center. The different sources of failure in communication could be the difference between Patterns 3 and 4. It would be possible that the malfunctioning source in a loop detector, causing a 100% missing data error, is different from the malfunctioning source, causing a partially missing data error. Therefore, because of similarities and differences of these two patterns, further analysis must be done to scrutinize their details and investigate whether these patterns would be better combined in a single pattern.

Pattern 5 includes most of the errors occurring randomly in the data, with no apparent specific relationship. Errors included in this pattern are often related to traffic flow theory discrepancies such as density- and AEVL-related errors with a percentage less than 20%. This pattern shows intermittent malfunctions, which could be caused by chattering, cross talk, or stuck-on detectors. Pattern 6 includes only one major error: volume greater than zero with zero speed. No other error was observed in this pattern. The important point is that the zero percentage of error is based on a volume greater than 17. This means that usually the volume counts are reasonable here and the problem could not be related to a high traffic volume (greater than 17) and were caused by another malfunction.

Association Rule Mining Results

Among all the patterns found by the clustering algorithm, Pattern 5 had different types of errors with low percentage and no obvious relationship between them. Therefore, only the data for Pattern 5 were separated from the original errors data set; then an Apriori algorithm was implemented on these data to investigate if any association rule exists between these errors. As mentioned in the "Methodology" section, interesting rules could be discovered by this method by setting a minimum acceptable limit for the support and confidence measures. To find all the possible rules in the errors, no limit was set at first; thus, 101 rules were found in the errors. All of these rules are in the form of $X \Rightarrow Y$. Both X and Y could

Table 3. Error patterns from fuzzy *c*-means

Pattern	Error criteria	Description	Possible reasons	Pattern with highest similarity
1	More than 20% and less than 70% error based on two criteria: 1. Volume = speed = 0 and occupancy >0% 2. Occupancy >95%	Error percentage based on both criteria are the same meaning when volume = speed = 0 and occupancy >0%, then occupancy >95%	Detector may be stuck on Loop is hanging on	Pattern 2
2	More than 70% error based on two criteria: 1. Volume = speed = 0 and occupancy >0 2. Occupancy >95%	Error percentage based on both criteria are the same, meaning when volume = speed = 0 and occupancy >0%, then occupancy >95% No error based on other criteria	Detector may be stuck on Loop is hanging on	Pattern 1
3	100% missing data	—	Communication down	Pattern 4
4	Incomplete data	—	Communication down	Pattern 3
5	Low error percentage randomly based on different criteria	No specific relationship between the errors	Intermittent malfunctions	Pattern 1
6	Any percent of error based on volume > 0 and speed = 0	—	Speed trap not functioning properly	Pattern 5

J. Transp. Eng., Part A: Systems, 2020, 146(7): 04020064

Table 4. Top five association rules found in the errors by the Apriori algorithm

Rule	LHS	RHS	Support	Confidence
1	Volume = speed = 0 and occupancy >0%	Occupancy >95%	38.0	1.0
2	Min AEVL (AEVL < 9)	Volume > 0 and speed = 0	12.4	0.6
3	Max AEVL (AEVL >59)	Occupancy >95%	7.4	0.5
4	Speed <5	Min AEVL (AEVL <9)	6.7	1.0
5	Max AEVL (AEVL >59), volume = speed = 0 and occupancy >0%	Occupancy >%95	6.6	1.0

include either one or more errors. Therefore, because there was no limit for the support and confidence value, these 101 rules cover all different combinations of errors in this pattern that were observed together, even if only once. However, because we are only interested in rules with high support values, the rules were ordered by their support value and only the top five rules were selected (Table 4). The rest of the rules reported low support values, meaning that they were not interesting rules for this analysis because the frequency with which they were observed was low. The corresponding confidence value for each rule is also given in the table. The results indicate that the top rule has the form $X \Rightarrow Y$, were X is volume and speed equal to zero while occupancy greater than zero, and Y is occupancy greater than 95%.

Fig. 3 is the graph-based representation of the top five rules. The following observations could be made from this figure:

1. Two independent groups of rules were created where the rules in each group are associated with each other.
2. When speed is less than 5 mi/h, the minimum AEVL error is likely to occur. Also, when the minimum AEVL error occurs, it is very probable that the error with zero speed and nonzero volume is observed. These results indicate that these three errors are highly related to each other and could result from the same malfunction source.
3. Similar to Observation 2, the three errors including maximum AEVL, occupancy greater than zero, and volume and speed equal to zero while occupancy greater than zero were frequently observed together. Findings suggest that these three errors are highly related and occur interchangeably, meaning that the same malfunction source could cause all these errors.

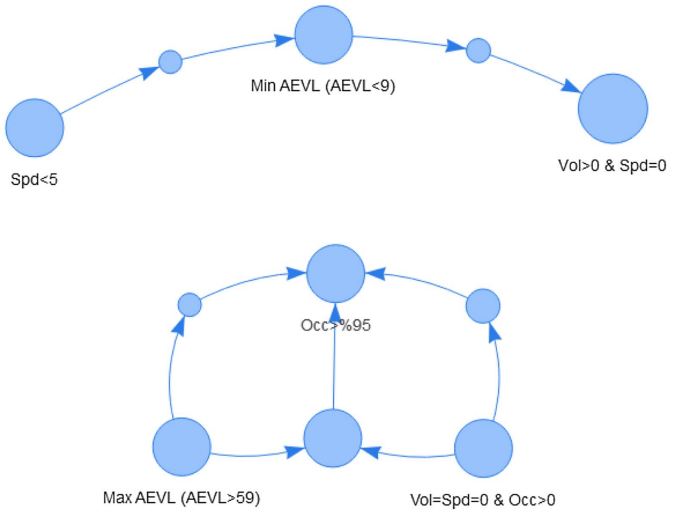

Fig. 3. Graph-based representation of association rules.

Field Verification

To verify the results found in our analysis, three representative stations with different patterns of errors were selected for a field visit on May 11, 2016, to diagnose the cause of their malfunctions. To investigate the source of errors for each loop station, the cabinet was opened by ADOT engineers to check the controller and loop cards. One loop card is installed corresponding to each slot in a dual-loop detector (e.g., 12 loop cards for six dual-loop detectors in six lanes), where the sensitivity level and frequency settings of the loops are adjustable through these loop cards. The first station was reporting erroneous data every day in only two lanes. In both of these lanes, one of the slots in the dual-loop detector was reporting Pattern 1 erroneous data, while the other slot was reporting a Pattern 6 error. Upon investigation, the source of the erroneous data was a defect in the loop cards. When the loop card was restarted, the system returned to normal operation.

The second station was reporting erroneous data every day, but only in a single lane. Here, one of the slots in the dual-loop detector was exhibiting a Pattern 2 error, while the other was displaying Pattern 6. The reason for reporting erroneous data in this case was once again found to be a defect in the loop card; the light on the loop card was continuously blinking, indicating a failure. When the loop card was replaced with a new one, the system returned to normal service. The third station was producing Pattern 5 erroneous data in all of the lanes; the errors in all the lanes and slots occurred on random days without any specific sequence. Visiting the cabinet at this station revealed that the loop cards were functioning correctly so other possible sources caused the errors.

The data reported by these three stations were then reanalyzed. The errors for the first two stations had disappeared after the field visit, but the third station was still reporting Pattern 5 erroneous data. Errors in this pattern can be quite complex, arising because of either mechanical wire defects (loop wires in the roadway or the lead-in wire) or nonmechanical malfunctions (e.g., chattering, pulse breakup, intermittent malfunctions), so failure sources with this pattern require further scrutiny by a traffic engineer.

Final Error Patterns

The patterns in the ILD errors could be finalized based on the following considerations: (1) errors in a pattern are highly associated with each other and are observed together in the data; and (2) the errors in a pattern are found to have the same malfunction while visiting in the field. Therefore, considering the findings from clustering and association rule mining of the data as well as the field visit findings, errors could be categorized into final patterns.

The final patterns with their detailed descriptions are presented in Table 5. Six initial patterns were modified and combined to create four primary patterns. The findings from the field visit troubleshooting revealed that when a slot reports erroneous data with Pattern 6, the other slot in the same dual-loop detector will report

Table 5. Final error patterns and detailed descriptions

Pattern	Pattern in errors	Relationship between errors in cluster	Relationship with adjacent lanes	Relationship with adjacent stations	Temporal trend	Possible reasons
1, 2, 6 (combined together)	More than 20% and less than 70% error based on two criteria: 1. Volume = speed = 0 and occupancy >0% 2. Occupancy >95% Any percent of error based on volume >0 and speed = 0	Error percentage based on first two criteria are the same, meaning when volume = speed = 0 and occupancy >0%, then occupancy >95% Volume >0 and speed = 0 usually occur separate from other two errors in a different slot of a dual-loop detector	No relationship between adjacent lanes	No relationship between adjacent stations	Happens intermittently on random days	Loop card defects: should be restarted
	More than 70% error based on two criteria: 1. Volume = speed = 0 and occupancy >0% 2. Occupancy >95%		No relationship between adjacent lanes	No relationship between adjacent stations	Continued for several consecutive days	Loop card could be totally failed and should be swapped Detector may be stuck on
3	100% missing data	—	All the lanes at the station have this pattern	No relationship between adjacent stations	Continued for several consecutive days	Communication is totally down Controller is down
4	Incomplete data	—	All the lanes at the station have this pattern	All the stations at study corridors have this pattern	Happens only on one day (not continued)	Data feed is unstable
				No relationship between adjacent stations		Communication is temporarily down
5	Low error percentage randomly based on different criteria	Errors are all associated with each other	No relationship between adjacent lanes	No relationship between adjacent stations	Happens intermittently at random days	Intermittent malfunctions caused because of chattering, cross talk, or pulse breakup

errors with Patterns 1 or 2. Therefore, Patterns 1, 2, and 6 in the initial clustering results were combined to form a single pattern that can then be further categorized into two subpatterns based on the error percentage because the cause of malfunction could be either temporary or permanent defects affecting the loop card. When the loop card defect is temporary, the errors based on the two criteria in this pattern are observed intermittently on random days, with a low percentage of errors. On the other hand, when the loop card defect is more severe, the errors are not only observed with very high percentage (mostly close to 100%) but also continue for several consecutive days.

Two patterns with no data and incomplete data remained separate. For the pattern with no data, it was observed that when a loop detector was reporting no data, all the detectors connected to the same controller were also reporting no data on consecutive days. The reason for this pattern was found to be a controller breakdown or communication faults. The pattern with incomplete data was also categorized into two subpatterns because it covers two distinct situations: (1) all the slots in all the stations are reporting incomplete data only for 1 day, and (2) all the slots for just one station are reporting incomplete data. The reasons here were found to be an unstable feed to the traffic operation center and a temporary breakdown in communications, respectively.

The final error pattern encompasses multiple errors that occur infrequently and apparently randomly throughout the month. Intermittent malfunctions caused by chattering, cross talk, or pulse breakup could be the underlying source for this error pattern, but the precise reason was not diagnosed during the field visit. Importantly, these errors were found to be closely associated with each other, and thus resulted from a similar malfunction source. Therefore, diagnosing and correcting any of these errors would result in correcting other errors in this pattern.

Because the amount of data is increasing, manually analyzing the data errors and discovering their underlying relationship is a time- and labor-intensive process. Therefore, the framework in this study could help improve the understanding of the relationships between different errors and failures. The three errors in the first pattern are among the most frequent errors observed in the ILD data. These errors were observed with different percentages during the day, where they could last for several consecutive days or only happen randomly. Using the data mining techniques in this study followed by the field visit could help to link these three errors together as well as identify their malfunctioning sources when occurring with either low or high percentages. Although the last pattern remained the most challenging, the association rule mining technique could find two separate rules in the errors that will ultimately lead technical specialists to identify the source of the malfunction.

The findings of this systematic framework could be used by agencies to save engineers' time and fix faulty ILDs in a more timely and efficient way. Different DOTs could utilize the same approach to identify different patterns in the ILD data. Managers could assign engineers possessing a detailed knowledge of loop detectors to visit different ILD stations for each error pattern and identify the source of the malfunction. According to the engineers' diagnosis in the field, the error patterns could also be revised and put into subcategories with the same error source. The maintenance personnel at DOTs could then be provided with the list of error patterns and the causes of malfunctions for further troubleshooting and necessary fixes. With this approach, maintenance personnel at DOTs will be able to fix the ILDs in the city using fewer distinctive error patterns and to assign their time and effort more efficiently.

Conclusion

This study proposed a systematic approach to ease the troubleshooting process for ILD errors by identifying patterns in daily ILD errors and linking them with their likely cause. The fuzzy *c*-means clustering method was implemented on the daily errors to identify six significant patterns initially. The most similar pattern to each pattern was also found by fuzzy *c*-means through the degree of membership of each observation to all the clusters. The clustering results were followed by an association rule mining technique to find the most frequent rules in the data, where the clustering method was unable to link the errors. By analyzing the findings from the clustering and association rule mining as well as the field visit, these six preliminary patterns were modified and combined to create four final patterns. Pattern 1, which includes three types of errors, was found to be caused by a defect in the loop card. Patterns 2 and 3, which represent 100% missing data and incomplete data, respectively, were found to arise because of communication error or controller breakdown in most cases, although an unstable data feed to the traffic operation center was another reason for incomplete data. Pattern 4 consists largely of traffic flow theory–based errors that occur infrequently and apparently randomly throughout the month. These intermittent malfunctions are likely to be attributable to chattering, cross talk, or pulse breakup and are the most difficult to diagnose. It should, however, be considered for further scrutiny by traffic engineers because the results from the association rule mining indicated that it might be possible that errors in this pattern are caused by different malfunctions; thus, this pattern could be further separated into subpatterns.

The amount of data archived from freeway loop detectors is increasing every day, and an individual ILD may suffer from several types of data error issues. Considering the frequency of data errors, it is almost impossible to find the relationships between them or link them manually by only looking at data errors. The primary contribution of this study is to help engineers troubleshoot ILD errors by highlighting the relationships between the errors and linking those that are similar. This approach will help agencies to address ILD malfunctions more efficiently by dealing with defined patterns of errors instead of several individual ones. The traffic engineers at ADOT are now able to send maintenance crews into the field to repair malfunctioning ILDs based on their observed error patterns, especially those with Patterns 1, 2, and 6, rather than individual errors. Because Patterns 1, 2, and 6 can often be fixed by simply installing a new loop card or restarting it, the engineers can come prepared, thus saving time and making the best use of the city's scarce resources.

In this study, not all of the representative stations for all other error patterns were visited because of limited human resources and the project timeline. In the future, for a similar analysis by different agencies, visiting more stations, especially the ones with more complicated patterns, will help to verify the sources of malfunctions for each pattern. In addition, during the field visit, only controllers and loop cards were checked at the loop detector station cabinets. Using technical engineers who are familiar with the loops wiring systems of the ILDs could help identify the exact source of the intermittent errors. Furthermore, the clustering and association rule mining methods were implemented using an offline computer program. The process could be streamlined by implementing the proposed systematic approach in an online computer program to identify error patterns more efficiently. In addition, the data in this study were only collected from the loop detector stations along the two corridors of I-10 and I-17. Other corridors in the Phoenix area could also be considered for data collection. Therefore, their data

could be utilized for the comparison and verification of the final error patterns in this study.

Data Availability Statement

All data, models, or code generated or used during the study are confidential in nature. All these items are part of a funded project by the Arizona Department of Transportation, so they are not allowed to be shared.

Acknowledgments

The authors would like to thank the Arizona Department of Transportation (ADOT) for funding and data support. We also acknowledge Mr. Vahid N. Goftar, Mr. David Riley, and Mr. Reza Karimvand for their professional advice and project coordination. The authors wish to extend their thanks to Mr. Robert Kluger and Ms. Jan Szechi for valuable comments and proofreading.

References

Agrawal, R., T. Imieliński, A. Swami, R. Agrawal, T. Imieliński, and A. Swami. 1993. "Mining association rules between sets of items in large databases." In *Proc., 1993 ACM SIGMOD Int. Conf. on Management of Data—SIGMOD '93*, 207–216. New York: ACM Press.

Agrawal, R., and R. Srikant. 1994. "Fast algorithms for mining association rules." In *Proc., 20th Int. Conf. on Very Large Data Bases, VLDB*, 487–499. New York: Association for Computing Machinery.

Al-Maolegi, M., and B. Arkok. 2014. "An improved Apriori algorithm for association rules." *Int. J. Nat. Lang. Comput.* 3 (1): 21–29. https://doi.org/10.5121/ijnlc.2014.3103.

Bae, B., H. Kim, H. Lim, Y. Liu, L. D. Han, and P. B. Freeze. 2018. "Missing data imputation for traffic flow speed using spatio-temporal cokriging." *Transp. Res. Part C: Emerging Technol.* 88 (Mar): 124–139. https://doi.org/10.1016/j.trc.2018.01.015.

Bezdek, J. C., R. Ehrlich, and W. Full. 1984. "FCM: The fuzzy *c*-means clustering algorithm." *Comput. Geosci.* 10 (2–3): 191–203. https://doi.org/10.1016/0098-3004(84)90020-7.

Bickel, P. J., C. Chen, J. Kwon, J. Rice, E. Van Zwet, and P. Varaiya. 2007. "Measuring traffic." *Stat. Sci.* 22 (4): 581–597. https://doi.org/10.1214/07-STS238.

Cheevarunothai, P. 2006. "Identification and correction of dual-loop sensitivity problems." *Transp. Res. Rec.* 1945 (1): 73–81. https://doi.org/10.1177/0361198106194500110.

Cheevarunothai, P., Y. Wang, and N. Nihan. 2005. "Development of an advanced loop event data analyzer (ALEDA) system for dual-loop detector malfunction detection and investigation." In *Proc., 12th World Congress on Intelligent Transport Systems*. Washington, DC: Transportation Research Borad.

Chen, C., J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya. 2003. "Detecting errors and imputing missing data for single-loop surveillance systems." *Transp. Res. Rec.* 1855 (1): 160–167. https://doi.org/10.3141/1855-20.

Chen, X., Z. He, and L. Sun. 2019. "A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation." *Transp. Res. Part C: Emerging Technol.* 98 (Jan): 73–84. https://doi.org/10.1016/j.trc.2018.11.003.

Chen, Y., Y. Lv, and F.-Y. Wang. 2020. "Traffic flow imputation using parallel data and generative adversarial networks." *IEEE Trans. Intell. Transp. Syst.* 21 (4): 1624–1630. https://doi.org/10.1109/TITS.2019.2910295.

Chen, Z., X. Qin, E. Schneider, Y. Cheng, S. Parker, and R. R. Shaon. 2019c. "Designing a comprehensive procedure for flagging archived traffic data: A case study." *Transp. Res. Rec.* 2673 (6): 165–175. https://doi.org/10.1177/0361198119841286.

Coifman, B. 1999. "Using dual loop speed traps to identify detector errors." *Transp. Res. Rec.* 1683 (1): 47–58. https://doi.org/10.3141/1683-07.

Coifman, B., and H. Lee. 2011. *Diagnosing chronic errors in freeway loop detectors from existing field hardware*. Washington, DC: Transportation Research Borad.

Duan, Y., Y. Lv, Y.-L. Liu, and F.-Y. Wang. 2016. "An efficient realization of deep learning for traffic data imputation." *Transp. Res. Part C: Emerging Technol.* 72 (Nov): 168–181. https://doi.org/10.1016/j.trc.2016.09.015.

Dunn, J. C. 1973. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." *J. Cybern.* 3 (3): 32–57. https://doi.org/10.1080/01969727308546046.

Ghosh, A., and B. Nath. 2004. "Multi-objective rule mining using genetic algorithms." *Inf. Sci.* 163 (1–3): 123–133. https://doi.org/10.1016/j.ins.2003.03.021.

Hothorn, T., and B. S. Everitt. 2014. *A handbook of statistical analyses using R*. Boca Raton, FL: CRC Press.

Hu, P., R. Goeltz, and R. Schmoyer. 2001. *Proof of concept of ITS as an alternative data resource: A demonstration project of Florida and New York data*. Oak Ridge, TN: Oak Ridge National Laboratory.

Ishimaru, J. 1998. *CDR user's guide*. Seattle: Univ. of Washington.

Jacobson, L. N., N. L. Nihan, and J. D. Bender. 1990. "Detecting erroneous loop detector data in a freeway traffic management system." *Transp. Res. Rec.* 1287: 151–166.

Karimpour, A., A. Ariannezhad, and Y.-J. Wu. 2019. "Hybrid data-driven approach for truck travel time imputation." *IET Intel. Transp. Syst.* 13 (10): 1518–1524. https://doi.org/10.1049/iet-its.2018.5469.

Klein, L. A. 2001. *Sensor technologies and data requirements for ITS*. Boston: Artech House.

Klein, L. A., M. K. Mills, and D. R. P. Gibson. 2006. *Traffic detector handbook*. 3rd ed. McLean, VA: USDOT.

Laña, I., I. I. Olabarrieta, M. Vélez, and J. Del Ser. 2018. "On the imputation of missing data for road traffic forecasting: New insights and novel techniques." *Transp. Res. Part C: Emerging Technol.* 90 (May): 18–33. https://doi.org/10.1016/j.trc.2018.02.021.

Lee, H., and B. Coifman. 2011. "Identifying and correcting pulse-breakup errors from freeway loop detectors." *Transp. Res. Rec.* 2256 (1): 68–78. https://doi.org/10.3141/2256-09.

Lu, X., Z. Kim, M. Cao, P. Varaiya, and R. Horowitz. 2010. *Deliver a set of tools for resolving bad inductive loops and correcting bad data*. Berkeley, CA: Univ. of California.

Lu, X.-Y., P. Varaiya, R. Horowitz, and J. Palen. 2008. "Faulty loop data analysis/correction and loop fault detection." In *Proc., 15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting*. Washington, DC: ITS America.

Ma, X., Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang. 2017. "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction." *Sensors* 17 (4): 818. https://doi.org/10.3390/s17040818.

Payne, H. J., E. D. Helfenbein, and H. C. Knobel. 1976. Vol. 2 of *Development and testing of incident detection algorithms.* Olympia, WA: Federal Highway Administration.

Ran, B., H. Tan, Y. Wu, and P. J. Jin. 2016. "Tensor based missing traffic data completion with spatial–temporal correlation." *Physica A* 446 (Mar): 54–63. https://doi.org/10.1016/j.physa.2015.09.105.

Rodrigues, F., K. Henrickson, and F. C. Pereira. 2019. "Multi-output Gaussian processes for crowdsourced traffic data imputation." *IEEE Trans. Intell. Transp. Syst.* 20 (2): 594–603. https://doi.org/10.1109/TITS.2018.2817879.

Schmoyer, R., P. S. Hu, and R. T. Goeltz. 2001. "Statistical data filtering and aggregation to hour totals of intelligent transportation system 30-s and 5-min vehicle counts." *Transp. Res. Rec.* 1769 (1): 79–86. https://doi.org/10.3141/1769-10.

Thambusamy, V. 2014. "Performance based analysis between k-means and fuzzy C-means clustering algorithms for connection oriented telecommunication data." *Appl. Soft Comput.* 19 (Jun): 134–146. https://doi.org/10.1016/j.asoc.2014.02.011.

Turner, S. 2004. "Defining and measuring traffic data quality: White paper on recommended approaches." *Transp. Res. Rec.* 1870 (1): 62–69. https://doi.org/10.3141/1870-08.

Turner, S. 2007. *Quality control procedures for archived operations traffic data: Synthesis of practice and recommendations*. Washington, DC: Federal Highway Administration.

Turner, S., R. Margiotta, and T. Lomax. 2004. *Monitoring urban freeways in 2003: Current conditions and trends from archived operations data.* College Station, TX: Texas Transportation Institute.

Turochy, R. E., and B. L. Smith. 2000. "New procedure for detector data screening in traffic management systems." *Transp. Res. Rec.* 1727 (1): 127–131. https://doi.org/10.3141/1727-16.

Vlahogianni, E. I., M. G. Karlaftis, and J. C. Golias. 2014. "Short-term traffic forecasting: Where we are and where we're going." *Transp. Res. Part C: Emerging Technol.* 43 (Jun): 3–19. https://doi.org/10.1016/j.trc.2014.01.005.

Wang, Y., and N. L. Nihan. 2003. "Can single-loop detectors do the work of dual-loop detectors?" *J. Transp. Eng.* 129 (2): 169–176. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:2(169).