Large-Scale Data-driven Traffic Sensor Health Monitoring

Tongge Huang[a], Pranamesh Chakraborty[a], Anuj Sharma[a], Chinmay Hegde[b]

[a] *Civil, Construction, and Environmental Engineering Department, Iowa State University, Ames, Iowa, USA 50011*
[b] *Electrical and Computer Engineering Department, Iowa State University, Ames, Iowa, USA 50011*

Abstract

Accurate traffic data collection is essential for supporting advanced traffic management system operations. This study investigated a large-scale data-driven sequential traffic sensor health monitoring (TSHM) module that can be used to monitor sensor health conditions over large traffic networks. Our proposed module consists of three sequential steps for detecting different types of abnormal sensor issues. The first step detects sensors with abnormally high missing data rates, while the second step uses clustering anomaly detection to detect sensors reporting abnormal records. The final step introduces a novel Bayesian changepoint modelling technique to detect sensors reporting abnormal traffic data fluctuations by assuming a constant vehicle length distribution based on average effective vehicle length ($AEVL$). Our proposed method is then compared with two benchmark algorithms to show its efficacy. Results obtained by applying our method to the statewide traffic sensor data of Iowa show it can successfully detect different classes of sensor issues. This demonstrates that sequential TSHM modules can help transportation agencies determine traffic sensors' exact problems, thereby enabling them to take the required corrective steps.

Keywords: sensor health monitoring, anomaly detection, clustering analysis, bayesian changepoint modeling

1. Introduction

Intelligent Transport Systems (ITS) applications, such as for detection of traffic congestion or incidents (Chakraborty et al., 2018a,b) and for decision-making (Shi and Abdel-Aty, 2015; Ma et al., 2017; Mori et al., 2015) have shown great effectiveness for advanced traffic management. Implementation of these applications requires reliable, high-quality data collected from roadway sensors, such as radar sensors, loop detectors, and video detectors. However, inevitably these types of traffic sensors suffer from erroneous data due to communications loss and malfunctioning (Lee and Coifman, 2011). Therefore, it is important to determine sensor health conditions before data are used for real-time traffic operations purposes or planning/policy development.

Typically, traffic sensor health monitoring (TSHM) is done by matching sensor readings to predefined thresholds (Turochy and Smith, 2000; Chen et al., 2003). In these methods, thresholds are placed on the maximum volume or occupancy values observed by the sensors, number of sensors with zero volume and nonzero occupancy, average effective vehicle length, and other similar traffic statistics. Sensor health is determined based on the number of faulty records observed for each sensor. Other studies have also been performed where sensor health is determined at the network level by taking into consideration neighboring sensors (Sun et al., 2016; Lu et al., 2014).

However, thresholds determined by these methods are based on daily aggregate traffic data. Unfortunately, such high-level aggregation fails to capture sensors' frequent within-day temporal abnormalities, which can also indicate faulty sensor conditions, in part because detailed evaluation of sensors' temporal abnormalities requires the processing of large-scale traffic data, which is beyond the capabilities of traditional computation techniques. For example, data obtained from the 300 radar-based traffic sensors of Iowa account for 15 gigabytes monthly. In this study, we have therefore developed a data-driven, massively parallelizable TSHM module that can handle large-scale statewide traffic data sources to detect abnormal sensors based on overall and temporal abnormalities.

Our TSHM module utilizes three steps for detecting abnormal sensors. First, based on clustering analysis, sensors with an abnormally high missing data percentage are labeled as faulty sensors. The module's second and third steps are based on the average effective vehicle length ($AEVL$) statistic. $AEVL$, proposed by Turochy and Smith (2000) combines volume, occupancy, and speed records using traffic flow theory to estimate vehicle dimensions. The second step of our TSHM module uses each sensor's $AEVL$ distribution to determine abnormal sensors using clustering analysis. $AEVL$ values, being representative of vehicles' physical dimensions, are robust to exogenous factors, such as traffic incidents and weather conditions. Therefore, the third step of our proposed TSHM module utilizes the assumption of constant vehicle length to determine changepoints in the temporal time-series data for each sensor. The temporal matrix of the changepoints obtained are then processed to extract the sparse matrix of all sensor abnormalities detected. Sensors that show frequent abnormalities can then be classified as abnormal sensors.

The major contributions of this study are as follows:

- We propose a data-driven stepwise method of anomalous sensor identification based on clustering analysis. This helps identify thresholds automatically for anomalous sensor detection. Our proposed method is similar to sieving analysis, wherein each sieve/step can be used to identify anomalous sensors by their distinct characteristics. This can enable authorities managing traffic sensors to easily identify sensor issues and take steps accordingly.

- The third step of our TSHM module. Based on the constancy of the vehicle length assumption, we identify the changepoints in the temporal matrix of the sensor data based on Bayesian analysis. And our entire method is scalable using massively parallelizable techniques, making it feasible to apply at a statewide level.

The rest of the paper is organized as follows. Section 2 provides a brief description of the relevant literature on sensor health monitoring followed by the details of the methodology adopted in this study in Section 3. Section 4 provides details of the data used and Section 5 the results obtained. Finally, Section 6 provides a summary of our results and points to potential directions for future study.

2. Literature Review

Sensor health monitoring is a crucial component of ITS applications. Over the last several decades, a number of studies have investigated this area. Traditionally, anomalous sensors have been identified by comparing individual traffic parameters to predetermined thresholds. For example, Payne and Thompson (1997) used predetermined thresholds of volume, occupancy, and speed, comparing these with 30-second and 5-minute aggregated traffic data to detect abnormal sensors. However, the unlikely assumption that traffic parameters are independent of one another is a primary concern regarding any single-parameter threshold algorithm. Therefore, Jacobson et al. (1990) used the relationship between traffic volume (q) and density (k), introducing the volume-occupancy ratio to check for anomalous data. If observed data fell outside accepted $k-q$ boundaries, it was flagged as erroneous data. However, the $k-q$ ratio algorithm labeled the data as erroneous when both k and q were equal to zero although such situations can also arise when no vehicle passed by the sensor (Chen et al., 2003). Chen et al. (2003) have therefore suggested that the $k - q$ ratio region is sensitive to the threshold settings and developed a daily statistics algorithm (DSA) using density and volume time series data to generate four statistic factors to identify the anomalous loop detectors. Vanajakshi and Rilett (2004) have used the idea of conservation of vehicles for a series of sequential detectors, examining detector anomalies at the network instead of single-sensor level.

Most of the sensor anomaly detection algorithms listed above are mainly based on occupancy and volume, which can easily be obtained using single loop detectors. However, individual vehicle speed calculation requires paired loop detectors or advanced roadway sensors such as microwave radar sensors, and etc. Due to their low installation cost, high accuracy, and small size, such advanced sensors are now widely used for traffic data collection (Klein et al., 2006). Hence, algorithms have also been developed for faulty sensor detection using all three traffic parameters: speed, volume, and occupancy. Turochy and Smith (2000) have proposed average effective vehicle length ($AEVL$) as an approximate function of volume, occupancy, and speed. Their study showed that $AEVL$ can successfully capture a wide range of data anomaly types which single-parameter threshold-based methods cannot. Similar other studies using $AEVL$ for anomalous sensor identification have also shown its efficacy (Al-Deek et al., 2004; Wells et al., 2008).

Although $AEVL$ has proven a useful indicator for sensor health monitoring, most studies still use predetermined thresholds. However, different detectors such as dual loop detectors, laser sensors, microwave radar sensors, etc. record data in different ways, not all of which are well adapted for sensor analysis based on predetermined thresholds. For example, for sensors that can self-recover in the case of a temporal anomaly, threshold-based algorithms might be too sensitive. To eliminate these pre-defined $AEVL$ thresholds, Lu et al. (2014) have developed a temporal and spatial based sequential algorithm for sensor health screening. Their study proposed a Multiple-Comparisons-with-the-Best (MCB) model to compare $AEVL$ between adjacent lanes and stations to assess any target detector's data quality. However, such between-station comparisons might not work well if nearby stations also have data quality issues. Also, the sequential MCB algorithm cannot be applied to isolated sensors or sensors with only one lane in each direction since it requires both within-station and between-station comparisons. To overcome these limitations, this study's proposed TSHM module introduces a data-driven approach to identifying faulty sensors based on clustering analysis for large-scale statewide sensor data.

Currently, most sensor-screening algorithms use daily aggregated traffic data to determine abnormal sensors, thereby ignoring temporal anomalies in the data stream. For example, Wu et al. (2017) have recently introduced a spatiotemporal pattern network (STPN) algorithm that can capture the time-series features of volume and speed data by a symbolization process. Their D-Markov model, trained on systematically ordered stations, calculates a mutual information matrix to identify anomalous sensors having low mutual information values. However, the well-ordered systematic sensor information required for training their STPN model is difficult to obtain on a statewide scale. Additionally, if any traffic sensor is added or removed, retraining the model becomes difficult. Finally, considering the large scale of traffic data collected from sensors at the statewide level, it is well beyond the capabilities of traditional computation techniques to train a complex STPN model for use at this level. Therefore, in this study, we propose a massively parallelizable TSHM approach that can extract temporal anomalies from large-scale traffic data streams and identify anomalous sensors showing frequent temporal abnormalities.

3. Methodology

Sensor abnormalities can arise due to high missing data percentage or anomalous sensor readings. As explained earlier, according to traffic flow theory, $AEVL$ is an approximate function of speed, volume, and occupancy and is considered a useful indicator for monitoring traffic data quality. Therefore, in this study, we calculate $AEVL$ as follows to identify anomalous sensor readings:

$$AEVL = \frac{5280 \times S \times O}{V} \tag{1}$$

where, S and V denotes the average speed (miles/h) and the hourly flow rate (vehicles/h) during the time interval. O represents the occupancy, i.e., the fraction of the time the sensor is occupied with vehicles during a given time interval. The constant 5280 is the scalar conversion factor applied to the measurement unit (ft). Since $AEVL$ represents the physical dimensions of vehicles, it is robust to traffic anomalies such as incidents or bad weather. Further, Turochy and Smith (2000); Lu et al. (2014) have shown that AEVL can detect erroneous data records that cannot be detected using speed, occupancy, or volume individually. Therefore, our proposed TSHM module utilizes three steps to extract anomalous sensors from large-scale statewide $AEVL$ data based on missing data and erroneous sensor readings:

1. *Data Completeness Test:* Checks whether a traffic sensor has missing data by checking data readings for completeness.
2. *$AEVL$ Anomaly Test:* Identifies suspicious records by checking whether sensors have provided unreasonable volume, occupancy, or speed records.
3. *Temporal Pattern Anomaly Test:* Checks whether sensors have provided fluctuating $AEVL$ values over the daily time-series data.

Figure 1 shows the flowchart of our proposed large-scale data-driven TSHM screening algorithm.

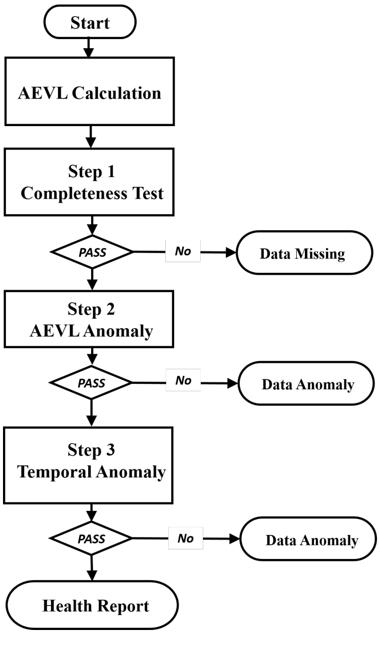


Figure 1: Workflow of proposed TSHM module

3.1. Setup

In the following subsections, N denotes the total number of sensors operating statewide and D the total number of days during the study period. In addition, the $AEVL$ for each sensor n on each day d per 20-

seconds interval (say t) with a total length of s is calculated as follows, with s being $3 \times 60 \times 24 = 4320$ for each sensor each day.

$$AEVL_n^d = \left(AEVL_n^{t_1,d}, AEVL_n^{t_2,d}, ..., AEVL_n^{t_s,d}\right) \tag{2}$$

3.2. Data Completeness Test

Long-time operation and various external reasons frequently lead to missing data issues in ITS traffic sensors (Turner et al., 2000). Therefore, our proposed TSHM module first attempts to detect sensors with an abnormally high missing data percentage, using completeness score (CS) as the metric for determining each sensor's missing data percentage Turner et al. (2000). CS is defined as the ratio of data readings received to the number expected. Thus, a lower CS score means a higher missing data percentage. The CS of each sensor for each day d was calculated with s the length of the time-series data:.

$$CS_{n,d} = \frac{s_{N_x,d}}{Max[s_{N_1,d}, s_{N_2,d}, ..., s_{N_x,d}]} \tag{3}$$

Then, the mean and standard deviation of the completeness score for a given sensor (n) over given days D was calculated and denoted as $S_n = (\mu_n^{CS}, \sigma_n^{CS})$. The features of the resulting 2-dimensional data were then used to determine anomalous sensors having abnormally high missing data percentages. This was accomplished via the unsupervised k-means clustering algorithm because of its computational efficiency (MacQueen et al., 1967; Berkhin, 2006). Here, $S \in \left\{\mu_n^{CS}, \sigma_n^{CS}\right\}_{n=1}^{N}$ defines the data points with length N, the total number of sensors. We first assign the data points to the K cluster centroids as $k_1, k_2, ..., k_j \in \mathbb{R}$. In this step, each data sample will be assigned to the cluster c_i by calculating $L2$ distance between the point ($S_{(n)}$) and cluster centroid (k_j) as:

$$c^{(i)} = \arg\min_j \| S_n - k_j \|^2 \tag{4}$$

Then, the centroids k_j are recomputed and updated by taking the mean of all the data samples which were assigned to the cluster centered by that centroid:

$$k_j = \frac{\sum_{n=1}^{x} 1\{c^{(i)} = j\} S_n}{\sum_{n=1}^{x} 1\{c^{(i)} = j\}} \tag{5}$$

The algorithm then iteratively computes $c^{(i)}$ and k_j until convergence. In this study, we used the elbow method (Kodinariya and Makwana, 2013) in which K is chosen by drawing the sum of squared distances to provide the appropriate data separation and determine the optimal number of clusters K. (Kodinariya and Makwana, 2013). Ideally, for each given day over the month, the μ_{N_i} and σ_{N_i} of a normal sensor are expected to be close to 1 and 0 respectively. In other words, normal sensors are expected to have a high completeness score (close to 1) with low variance. Therefore, after convergence of the k-means clustering, any cluster with a centroid close to $(1, 0)$ was labeled a normal sensor group. The rest of the clusters were classified as abnormal sensors and assigned to anomalous sensor groups based on their different levels of distance from the normal sensor group in terms of missing data. This was based on the assumption that anomalous sensors with high levels of missing data are rare in the population while the majority of sensors perform normally, a fundamental assumption of unsupervised anomaly detection (Chandola et al., 2009). The next steps of the proposed TSHM module, the AEVL anomaly test, and temporal anomaly test, were then applied to sensors labeled as belonging to the normal sensor group to find any further issues.

3.3. AEVL Anomaly Test

The first step of our proposed TSHM module attempts to detect sensors with high levels of missing data. However, sensors with a low missing data ratio (and therefore classified as normal sensors in the data completeness test) can also suffer from abnormal, corrupted sensor readings due to calibration issues, double-counting of vehicles changing lanes, and other noise and errors. Hence, the second step of our proposed TSHM module, the $AEVL$ anomaly test, attempts to detect such sensors with frequently occurring noisy/corrupt readings.

To reduce noise in the raw AEVL data collected at 20-second to 5-minute intervals, data were first aggregated by taking the average, similar to Yao et al. (2017). Thus, $T = \{t_1, ..., t_z\}$ denotes the time-series data in 5-minute intervals of length z for each sensor on a given day. Thus, the $AEVL$ for sensor n at z^{th} 5-minutes interval can be denoted as $AEVL_n^{t_z,d}$. We then represented the distribution of all $AEVL$ value for each sensor as a 2-dimensional feature vector $(\mu_n^{AEVL}, \sigma_n^{AEVL})$ using the aggregated AEVL records' mean and standard deviation.

The amount of data collected from state-wide traffic sensors is usually well beyond the capability of traditional computing techniques. For example, 500 MB of sensor data is collected daily in the state of Iowa, which aggregates to 15 GB monthly. A single conventional local machine cannot process such an enormous scale of data. To alleviate this issue, we used parallel computation techniques using Hadoop Distributed File System (HDFS) for storing the large-scale traffic data and Apache Pig Apache Pig (2018) for processing the data using MapReduce. While traditional single computer machines cannot process the monthly scale traffic data, usually the data processing can be completed in approximately 6 minutes. This enables abnormal sensor identification for statewide traffic data to be handled with ease.

Our methodology for detecting sensors with abnormal $AEVL$ records is based on the Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm because of its great ability to handle outliers. The DBSCAN algorithm can be used to discover clusters of an arbitrary shape with good efficiency on large dataset with minimal requirements of domain knowledge to define clustering parameters (Ester et al., 1996). DBSCAN has been used as an efficient clustering algorithm in different applications, including both classification and outlier/anomaly detection problems((Erman et al., 2006), (Nisa et al., 2014)). The fundamental assumption for DBSCAN in anomaly detection problems is that while normal data instances belong to clusters, anomalous data do not belong to any cluster (Chandola et al., 2009) and can therefore be attributed to noise or anomalies. The basic steps to identifying anomalous sensors based on the 2-dimensional feature vector $VL_n = (\mu_n^{AEVL}, \sigma_n^{AEVL})$ using DBSCAN algorithm is as follows (Schubert et al., 2017):

1. The algorithm starts with a random point $VL_n = (\mu_{AEVL,n}, \sigma_{AEVL,n})$ calculated from the $AEVL$ records of sensor n from N number of total sensors. If the point p_n is unclassified, the algorithm continues to find neighbors based on the two basic DBSCAN functions, namely the 'range query' function against other points defined as:

$$dist(P_i, P_j) = \sqrt{(\mu_{AEVL,P_i} - \mu_{AEVL,P_j})^2 + (\sigma_{AEVL,P_i} - \sigma_{AEVL,P_j})^2} \tag{6}$$

and the 'distance function' with the searching radius to find neighbors defined as:

$$Neighbors_\epsilon(P_i) = \{P_j \in DB \,|\, dist(P_i, P_j) \leq \epsilon\} \tag{7}$$

If the number of neighbors is greater than $minPts$, they will be defined as a cluster. If point p_n does not belong to any cluster, it will be labeled as noise.

2. After defining the distance function and the -neighborhood function, two input hyperparameters need to be chosen: (a) $minPts$ which defines the minimum number of neighbors to build a cluster, and (b) ϵ which defines the searching radius to find any neighbors. We determined these two input parameters based on the sorted k-dist graph technique. We first define the $minPts$ as k, and drew the sorted k-dist curve, with the first *valley* of the curve being considered the threshold to divide noise from other points (Ester et al., 1996). This allowed us to perform iterated search on a wide range of real-world data to determine the optimal k through reaching stability.

3.4. Temporal Pattern Anomaly Test

The $AEVL$ anomaly test captures anomalous sensors with suspicious data readings based on high-level aggregated $AEVL$ data distributions, represented by $\mu_{AEVL,n}$ and $\sigma_{AEVL,n}$. However, sensors reporting abrupt $AEVL$ fluctuations cannot be detected using aggregated distributions. Thus, it is necessary to consider each individual sensor's time-series data to detect this latter type of abnormal sensor. Therefore, after severely abnormal sensors had been captured through the $AEVL$ anomaly test, we used the temporal pattern test to examine the remaining sensors to flag any with frequent temporal abnormalities.

The core idea of the temporal pattern anomaly test is based on the constancy of vehicle length. Unlike raw 20-second-interval data, aggregated 5-minute-interval $AEVL$ moving averages reduce data readings' inherent noise, so that $AEVL$ values over a given time period are more stable (Wells et al., 2008). Thus, $AEVL$ data from a given sensor at times T_i and T_{i+1} should, based on traffic operations theory, fall into the same distribution. In contrast, if many abrupt changes in the $AEVL$ distribution exist over a given sensor's time-series data for a given day, it can be labelled as a potentially anomalous sensor.

Figure 7 shows the $AEVL$ time series plot of three consecutive sensors at a given day. The unanticipated "spikes" in the middle sensor (Figure 7b) compared to its upstream (Figure 7a) and downstream (Figure 7c) sensors indicate the middle sensor have recorded abnormal data, since its nearby sensors provide predictable data. The third step of our proposed TSHM module therefore attempts to detect abnormal sensors that report frequent abrupt fluctuations in $AEVL$ values. Its first task is to detect "spikes" (also referred to as changepoints) and then detect which sensors are reporting such frequent changepoints by extracting the temporal pattern of all changepoints detected.

3.4.1. Changepoint Detection

To detect abrupt $AEVL$ distribution changes over a given set of time-series data, we adopt changepoint detection, also known as time-series segmentation. Sensor faults in the real world are usually random and therefore unpredictable. Thus, we cannot know how many failures or changepoints will occur during a given time period. Therefore, abrupt $AEVL$ changepoint detection can be considered a multiple changepoint detection problem based on an exact Bayesian changepoint model (detailed mathematical description in (Fearnhead, 2006)).

In this study, we used the 5-minute $AEVL$ records ($AEVL_n^{t_z,d}$) determined from the step 2 of our TSHM module for change point detection. Let us assume that given time series has m changepoints with their positions referred to as $\tau_{1:m} = (\tau_1, \tau_2, ..., \tau_m)$ where $\tau_i < \tau_j$ for $i < j$ and the two change points at either end of the time series can be denoted $\tau_0 = 0$ and $\tau_{m+1} = z$. Note, $z = 288$ is the length of time series for a given day with 5-minute aggregated data. Further, the m changepoints will split the time series data into $m+1$ sub-segments and the observations of the j^{th} sub-segment will consist of $AEVL$ records from τ_{j-1} to τ_j. Then, an arbitrary prior distribution for the j^{th} sub-segment, which is mutual for all other sub-segments with a set of distribution parameters denoted as β. Further, the observations of any given subsegment being conditional independent of other subsegments' observations can also be assumed (Fearnhead and Liu, 2007). Therefore, the probability that two time points say (t and s) will belong to the same sub-segment can be represented as:

$$\begin{aligned}\mathrm{P(t, s)} &= \mathrm{Pr(AEVL_{t:s}\,|\,t, s\ in\ the\ same\ sub\text{-}segment)} \\ &= \int \prod_{i=t}^{s} f(AEVL_i\,|\beta)\pi(\beta)d\beta\end{aligned} \tag{8}$$

where, $\pi(\beta)$ denotes the prior with the parameter β. Then the marginal likelihood of the segment at $AEVL_{t:z}$ given a changepoint at $t-1$ can be defined as:

$$Q(t) = \begin{cases} \sum\limits_{s=t}^{z-1} P(t, s)\, Q(s+1)\, g(s+1-t) + P(t, z)\,(1 - G(z-t)), & \text{if } t = 2, ..., z, \\ \\ \sum\limits_{s=1}^{z-1} P(1, s)\, Q(s+1)\, g_0(s) + P(1, z)\,(1 - G_0(z-1)), & \text{if } t = 1 \end{cases} \tag{9}$$

where $g(t)$ is the probability mass function of the time interval between two successive changepoints, and $g_0(t)$ is the probability mass function of the first changepoint after 0. Thus, the $G(t) = \sum_{s=1}^{t} g(s)$ and $G_0(t) = \sum_{s=1}^{t} g_0(s)$ denotes the probability distribution respectively. Then the posterior probability distribution of the first changepoint can be derived as:

$$\begin{aligned}\mathrm{Pr}\,(\tau_1|\mathrm{AEVL_{1:z}}) &= \mathrm{Pr(AEVL_{1:z}}, \tau_1)/\mathrm{Pr(AEVL_{1:z})} \\ &= \mathrm{Pr}(\tau_1)\mathrm{Pr(AEVL}_{1:\tau_1}\,|\,\tau_1)\mathrm{Pr(AEVL}_{\tau_1+1:z}\,|\,\tau_1)\mathrm{Q(1)} \\ &= \mathrm{P}(1, \tau_1)\mathrm{Q}(\tau_1+1)\mathrm{g_0}(\tau_1)/\mathrm{Q(1)}\end{aligned} \tag{10}$$

Then based on τ_{j-1} for $\tau \in (1, z-1)$, the remaining changepoints τ_j can be derived as:

$$\mathrm{Pr}\left(\tau_j|\tau_{j-1}, \mathrm{AEVL_{1:z}}\right) = \mathrm{Pr}(\tau_{j-1}+1, \tau_j)\mathrm{Q}(\tau_j+1)\mathrm{g}(\tau_j - \tau_{j-1})/\mathrm{Q}(\tau_{j-1}+1) \tag{11}$$

In our case, the changepoints indicate changes in vehicle length distribution (i.e., a changing variance problem see (Fearnhead and Liu, 2007)). Specifically, for the normal upstream and downstream sensor in Figure 7, there is a higher chance to observe change points at around 5:30 AM and 11:30 PM which are recurrent. This observation follows standard traffic flow characteristics on freeways, where the traffic volume at night is lower than during the day. Therefore, the aggregated 5-minute interval vehicle length moving average has larger variance at night due to a smaller sample size. In contrast, changepoints present in abnormal sensors occur arbitrarily and erratically throughout the entire day.

This observation follows the traffic flow characteristics on the freeways, where the traffic volume during the night time is lower than during the day time. Therefore, the moving average for vehicle length over 5-minutes aggregation will have larger variance during the night times due to smaller sample size. On the other hand, the change points present in the abnormal sensor are arbitrary and erratic occurring throughout the entire day.

3.4.2. Abnormal sensor detection in the temporal changepoint matrix

To detect sensors reporting an abnormal temporal pattern in the changepoint matrix, we used changepoint probabilities calculated as described above to form changepoint temporal matrices as in Figure 8. Specifically, changepoint probabilities calculated at each 5-minute interval (t_z) for sensor n at a given day d

can be denoted as $CP_n^{t_z,d}$. We then accumulated the changepoint probabilities for each 5-minute interval for all days in the study period (each month of historical data) to obtain $CP_n^{t_z}$ as follows.

$$CP_n^{t_z} = \sum_{d=1}^{D} CP_n^{t_z,d} \tag{12}$$

This aggregated $CP_n^{t_z}$ was then used to create the following temporal matrix:

$$\begin{pmatrix} CP_1^{t_1} & CP_1^{t_2} & \dots & CP_1^{t_z} \\ CP_2^{t_1} & CP_2^{t_2} & \dots & CP_2^{t_z} \\ \dots & \dots & \ddots & \vdots \\ CP_n^{t_1} & CP_n^{t_2} & \dots & CP_n^{t_z} \end{pmatrix} \tag{13}$$

This change point matrix consists of two distinct data features comprised of (a) recurrent change point pattern and (b) abrupt presence of change points by anomalous sensors. Therefore, we need to extract enhanced anomalous data features that can be used to detect abnormal sensors reliably. This is done in this study using Robust Principle Component Analysis (RPCA) (Candès et al., 2011). RPCA has been successfully used in literature to detect moving objects from the video surveillance system (Bouwmans and Zahzah, 2014). RPCA attempts to decompose the changepoint matrix CP into a low-rank temporal matrix (L) and a sparse temporal matrix S as follows.

$$CP = L + S \tag{14}$$

The low-rank matrix reflecting the recurrent changepoint pattern (say the background) and the sparse temporal matrix reflecting the abrupt presence of changepoints (say the foreground) are calculated as follows:

$$\min_{L,S} \ \mathrm{rank}(L) + \lambda \parallel S \parallel_0 \quad s.t \ M - L - S = 0 \tag{15}$$

where $\lambda > 0$ is a balancing parameter. However, this is a non-convex problem for optimization and solving the *rank* and l_0-norm are NP-hard. The relaxation function with the convex envelop is obtained by replacing the $l_0 - norm$ by the $l_1 - norm$ $(\parallel . \parallel_1)$ and replacing *rank* with nuclear norm $(\parallel . \parallel_*)$:

$$\min_{L,S} \ \parallel L \parallel_* + \lambda \parallel S \parallel_1 \quad s.t \ M - L - S = 0 \tag{16}$$

where the $\lambda > 0$ is chosen by $\lambda = \frac{1}{\sqrt{\max(m,n)}}$. The RPCA algorithm to decompose the raw change point temporal matrix (CP) into the low-rank matrix and the sparse matrix. Then, we used the DBSCAN clustering algorithm described earlier to detect abnormal sensors. Clustering was done using the mean and standard deviation of the sparse temporal matrix values $(S_n^{t_z})$ for each sensor obtained from RPCA.

3.5. Baseline Comparison

To verify the feasibility and accuracy of our proposed TSHM module, we compared its algorithmic performance with 2 benchmark algorithms: 1) fixed-threshold based $AEVL$ control limit method (CLM) (Wells et al., 2008), and 2) the temporal and spatial comparison screening algorithm using multiple comparison with the best (MCB) technology (Lu et al., 2014).

The core idea of the CLM method is based on the 95% confidence interval calculated from the overall $AEVL$ distribution. The CLM method first identifies the experimental time period (31 days in our study's example month of July 2017) and then calculates $AEVL$ for all sites to form the $AEVL$ distribution. Then, the mean $AEVL$ is obtained from the distribution, denoted as μ_{cl}. The upper and lower control limit boundaries can be obtained by $\mu_{cl} + 2\sigma_{cl}$ and $\mu_{cl} - 2\sigma_{cl}$ respectively. If the average $AEVL$ value of any individual sensor falls outside the control limits, the sensor will be flagged as potentially anomalous for further study.

he MCB algorithm flags anomalous sensors on the basis of temporal and spatial information by comparing $AEVL$ in three steps: (1) Data aggregation: Individual $AEVL$ data points are grouped into 30-minute intervals for each sensor and each lane, and represented by the mean $AEVL$ and variance. (2) Within station comparison: MCB are performed between different lanes in the same station, using the confidence interval created by MCB to check if there is a statistically significant difference between the target lane and best max lane. (3) Between station comparison: MCB are performed between nearby stations uses the fuzzy logic decision tree to label potential errors when target lane/station data are significantly higher/lower than that of comparison lanes/stations.

4. Results

Traffic sensor data were collected from Iowa's 338 freeway radar sensors statewide from April–December 2017. These radar sensors use digital radar beams (virtual lines) to record passing vehicles' speed, occupancy, volume, vehicle type, etc. The volume of each month's sensor data aggregated at 20-second intervals is approximately 15 GB of distributed storage (via a Hadoop Distributed File System or HDFS). In this paper, we show sample results and test our proposed module against the two benchmark algorithms using sensor data from a sample month (July 2017), utilizing our remaining data for sensitivity and stability analysis. In real-world applications, however, our module could be implemented with a sliding window of 1 month to detect sensors functioning abnormally over the last month. We next discuss the results of testing our proposed algorithm on aggregated data for each month of the study period via our TSHM module's 3 steps: (a) data completeness check, (b) $AEVL$ anomaly check, and (c) temporal pattern anomaly check. We then compare our proposed algorithm against the benchmark algorithms described previously.

4.1. Data Completeness Test Results

The first step of our proposed TSHM module is the data completeness test to detect sensors with abnormally high missing data percentages. As discussed in Section 3.2, this uses K-means clustering based on each sensor's completeness score (CS) mean and standard deviation. For 8 of the 9 months of the study period (April–December 2017), the elbow method described in Section 3.2 found the optimum number of clusters to be 3, while for August 2017, it found the optimum number of clusters to be 2. Figure 2 shows the clusters from the sample month July 2017. These 3 clusters can be labelled based on missing data severity as: normal sensor, abnormal sensor level 1 and abnormal sensor Level 2 based on their severity of data missing problem. In other words, the cluster with the mean CS closest to 1 and standard deviation close to 0 (i.e., $S_n = (\mu_n, \sigma_n) \approx (1, 0)$) can be labelled as normal and the other two labelled as anomalous. Thus, in July 2017, Iowa had 299 operating freeway radar sensors, 36 of which were abnormal.

Figure 3 shows heatmaps of the 3 missing-data-percentage-based clusters from July 2017. As can be seen from Figure 3a, normal sensors rarely suffer missing data issues, while level 2 abnormal sensors have an exceptionally high percentage of missing data (Figure 3c), Although level 1 abnormal sensors' missing data rates are not excessive (Figure 3b), still they justify manual inspection and repairs. (It should be mentioned that although our proposed clustering-based method can automatically classify sensors as normal vs. abnormal, traffic operations agencies could also define anomalous sensors based on their individual requirements, choosing based on the cluster centers obtained the $S_n = (\mu_n, \sigma_n)$ for proposed method uses to identify abnormal sensors.)

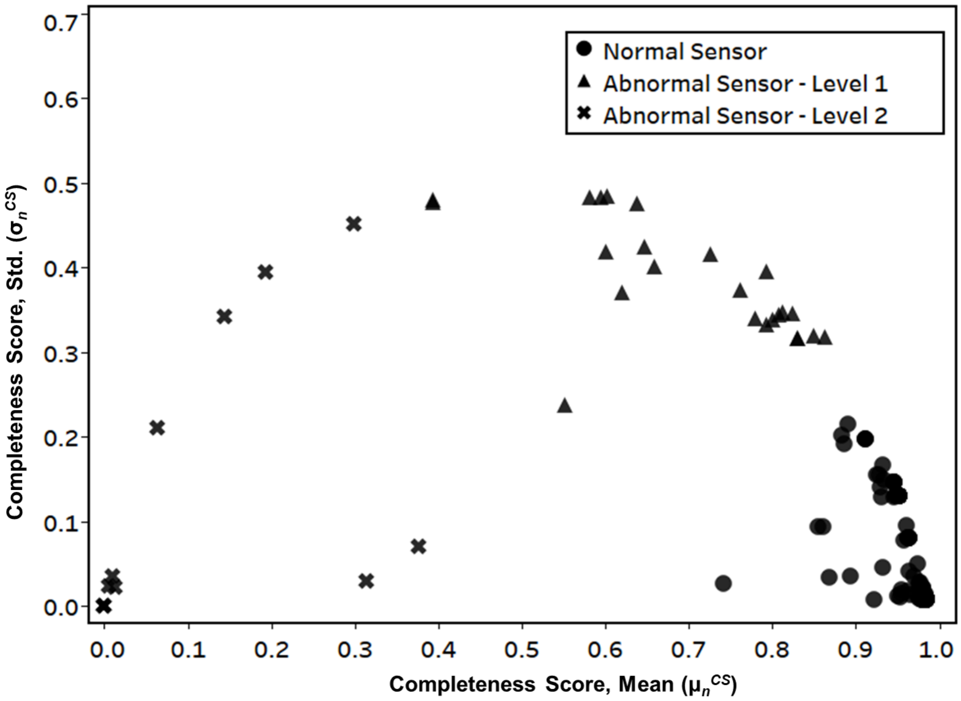


Figure 2: Data completeness test result on the sample month of July

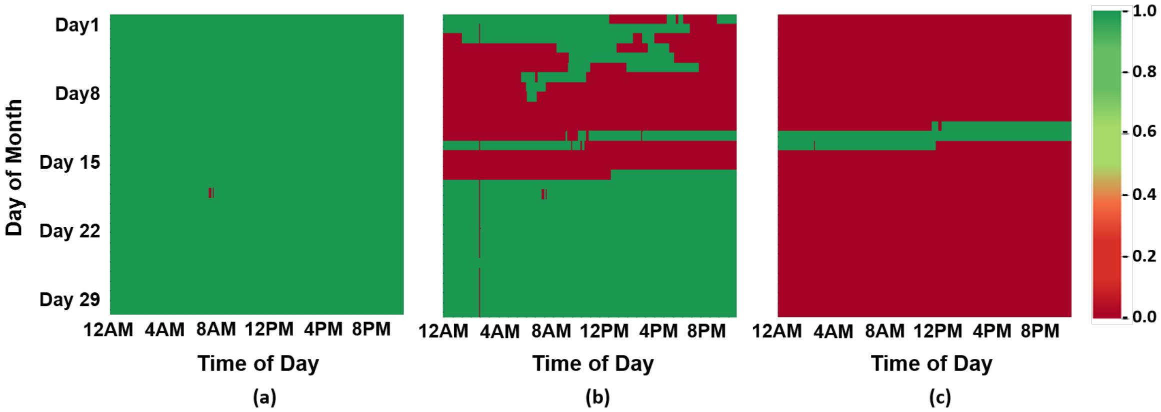


Figure 3: Sample missing data percentage heatmap of (a) Normal sensor, (b) Abnormal Sensor - Level 1, and (c) Abnormal Sensor - Level 2 from the data completeness test

It should be noted that sensors classified as normal in this data completeness test can nonetheless suffer from abnormal/atypical recorded values. Such "normal sensors" were therefore passed through step 2 of our proposed TSHM module, the AEVL anomaly test.

4.2. AEVL Anomaly Test Results

Sensors with no major missing data issues can still be abnormal by recording atypical sensor values. Therefore, the *AEVL* anomaly test, step 2 of our proposed TSHM module, uses the DBSCAN algorithm described in Section 3.3 to detect sensors recording abnormal $AEVL$ values. Figure 4 shows abnormal sensors detected in the sample month of July 2017 based on the 2-d $AEVL$ distribution $(\mu_n^{AEVL}, \sigma_n^{AEVL})$. Out of the 263 sensors classified as normal sensors based on data completeness test for July 2017, 13 were classified as abnormal based on the $AEVL$ anomaly test.

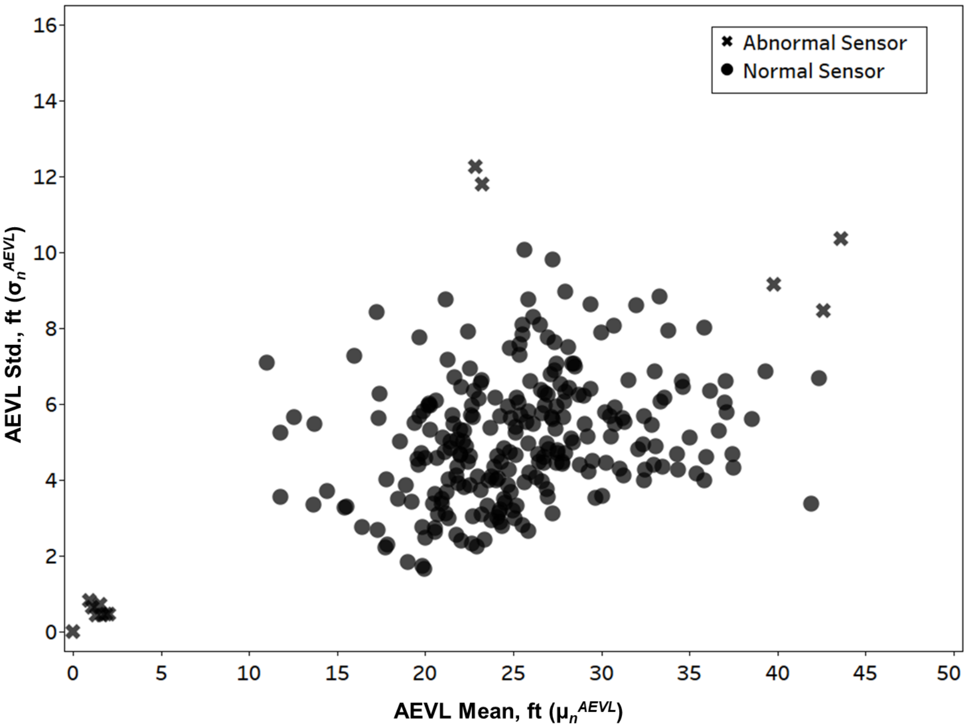


Figure 4: AEVL anomaly test result for the sample month of July, 2017

It can be seen in Figure 5, where we plot the 2-d completeness score $S_n = (\mu_n, \sigma_n)$ obtained from step 1's data completeness test with the sensor labels obtained from step 2's AEVL anomaly test, that the abnormal sensors detected in step 2 cannot be detected using step 1's CS feature vector S_n, since the sensors identified in step 2 didn't have any missing data issues. This demonstrates that both the data completeness and AEVL anomaly tests are required, since each identifies abnormal sensors having different issues that cannot be detected using a single test.

To illustrate the difference in $AEVL$ distributions observed in anomalous vs. normal sensors, Figure 6 shows the different types of $AEVL$ cumulative distribution function (CDF) plots reported by three sample anomalous sensors along with their adjacent upstream (u/s) and downstream (d/s) "normal sensors." In Figures 6a and 6b, the abnormal sensors reported lower and higher means, respectively, compared to their adjacent u/s and d/s sensors. In contrast, the abnormal sensor shown in Figure 6c reported higher variance compared to its adjacent sensors. Such CDF visualization of abnormal sensors helps justify our proposed method of detecting sensors reporting abnormal sensor readings by clustering $AEVL$ distributions $((\mu_n^{AEVL}, \sigma_n^{AEVL}))$

4.3. Temporal Pattern Anomaly Test Results

The temporal pattern anomaly test detects sensors showing abrupt fluctuations or "spikes" in $AEVL$ time-series values by using the Bayesian changepoint detection algorithm, as described in Section 3.4.1, and

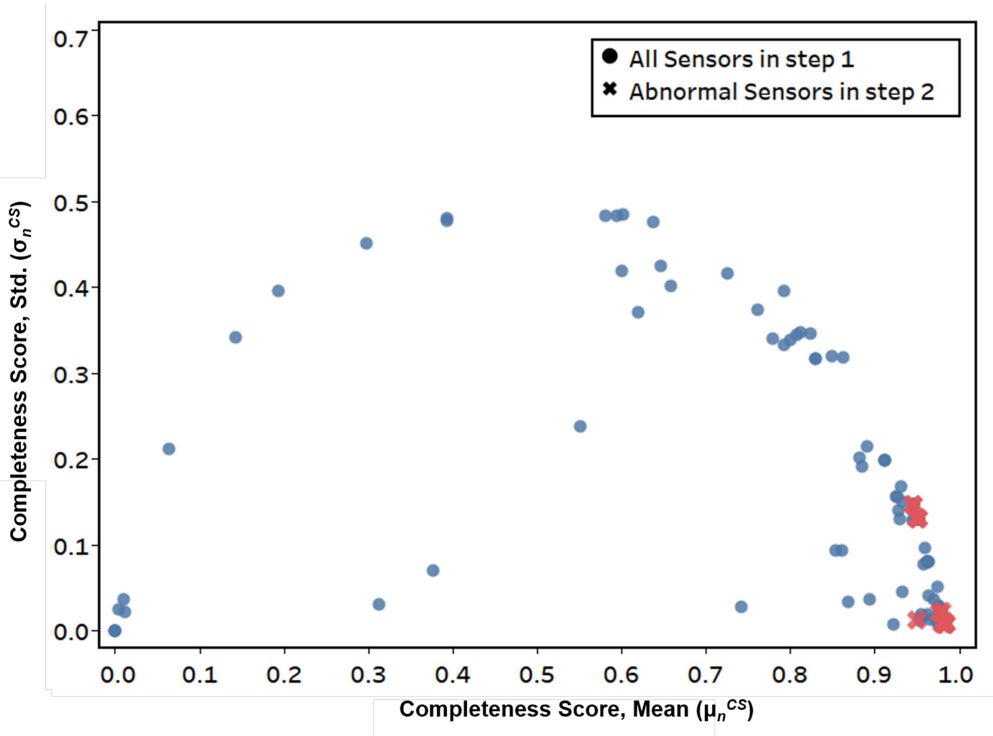
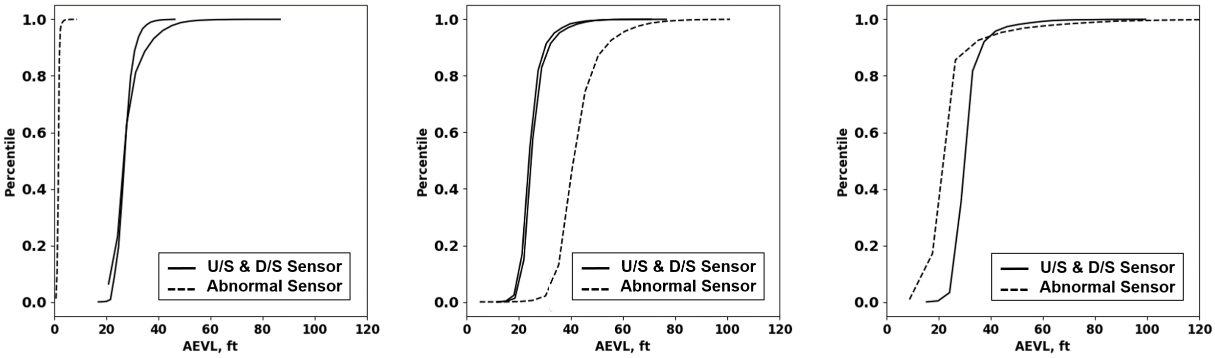


Figure 5: Completeness Score plots for all sensors in step 1 and abnormal sensors in step 2



(a)

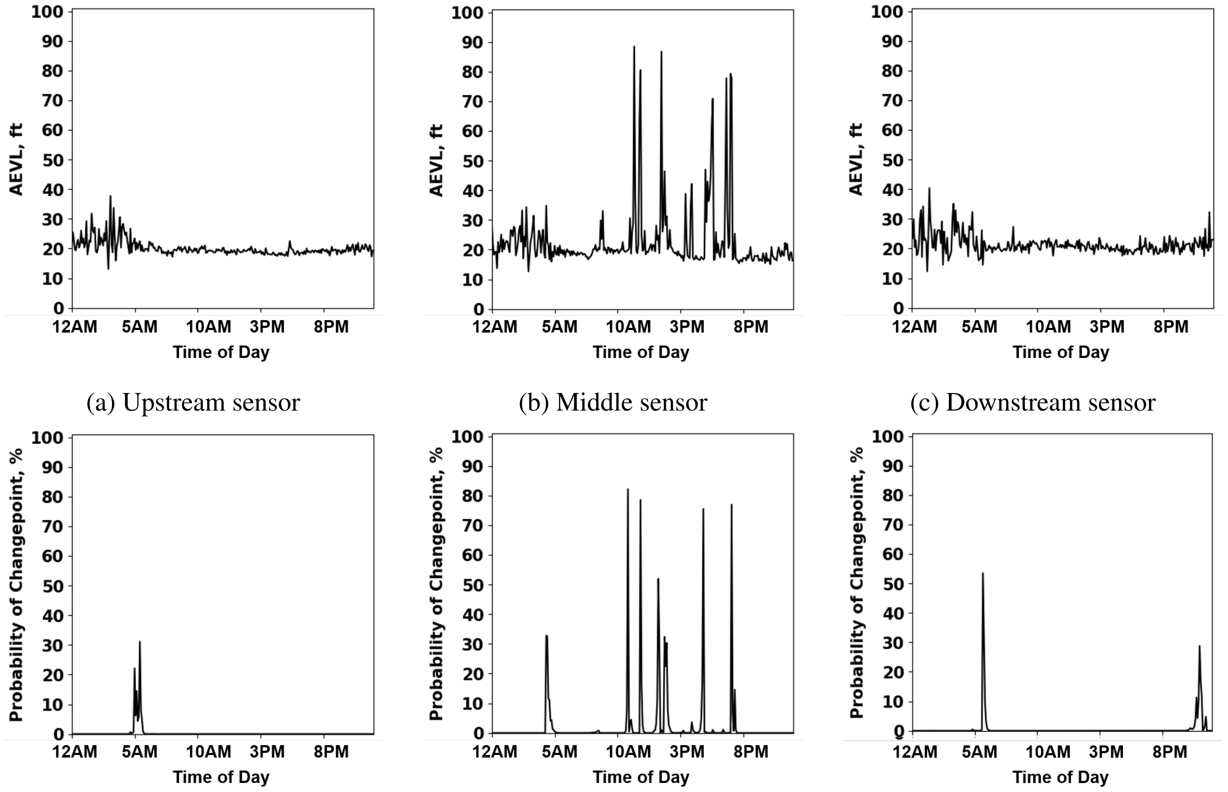
(b)

(c)

Figure 6: Sample CDF plots of abnormal sensor with their upstream(u/s) and downstream(d/s) sensors for abnormal sensor with: (a) low mean, (b) high mean, and (c) high variance

RPCA to denoise and enhance the changepoint matrix as described in Section 3.4.2. AEVL as a surrogate for vehicle length is not expected to be affected by time of day, unlike volume or speed. Therefore, frequent fluctuations in $AEVL$ values suggest sensor abnormality.

Figure 7 shows the raw $AEVL$ values and corresponding changepoint probability distributions of three consecutive sensors on a sample day. It can be seen that the middle sensor (Figure 7b) shows a significant number of spikes in $AEVL$ values, resulting in an increase in changepoint probabilities (Figure 7e) compared to it's u/s and d/s sensors. Each sensor's changepoint temporal matrix was formed by accumulating its changepoint probability at each 5-minute interval over the days of the study as described in Section 3.4.1.

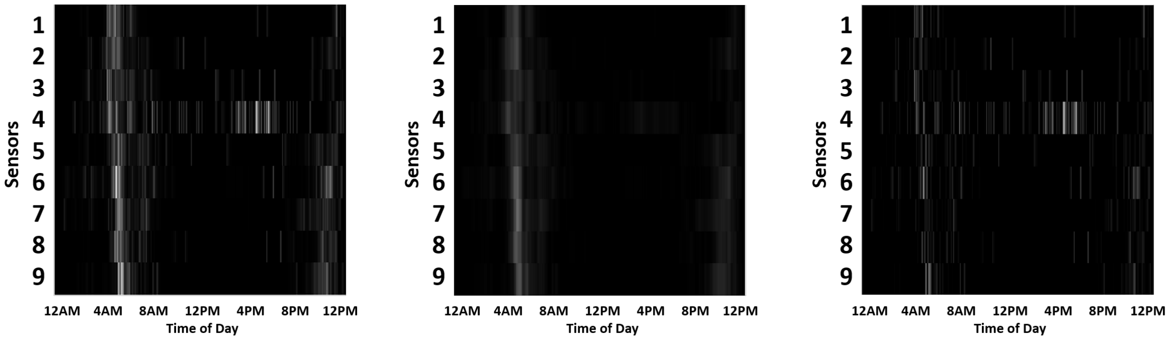


(a) Upstream sensor (b) Middle sensor (c) Downstream sensor

(d) Upstream sensor (e) Middle sensor (f) Downstream sensor

Figure 7: *AEVL* time series distribution of a sample day for (a) Upstream sensor, (b) Middle anomaly sensor, (c) Downstream sensor and changepoint probability distribution for the same day for the (d) Upstream sensor, (e) Middle anomaly sensor, (f) Downstream sensor

Figure 8 shows sensors in a sample plot following their actual spatial order on the freeway. In the sparse matrix shown in Figure 8c, we can see sensor 4 shows abrupt changes in *AEVL* over the study period, while its adjacent sensors are following the predictable temporal pattern.



(a) Raw CP matrix (b) Low rank CP matrix (c) Sparse CP matrix

Figure 8: Sample RPCA Decomposition: (*a*) Raw CP matrix, (*b*) Low rank CP matrix, (*c*) Sparse CP matrix

Then, similarly to the AEVL anomaly test, we used DBSCAN clustering algorithm to detect the anomaly sensors by calculating each sensor's sparse matrix mean and standard deviation as a 2-d feature vector. Figure 9 shows the July 2017 sample month's clustering results in which 8 sensors (2.04%) were classified as abnormal. In Figure 10, we visualize the different characteristics of one randomly selected anomalous sensor and one randomly selected normal sensor using a heatmap of their raw $AEVL$ data It can be seen that compared with the normal sensor (Figure 10a), the abnormal sensor (Figure 10b) shows more "spikes" or temporal fluctuations, resulting in it having been flagged as anomalous.

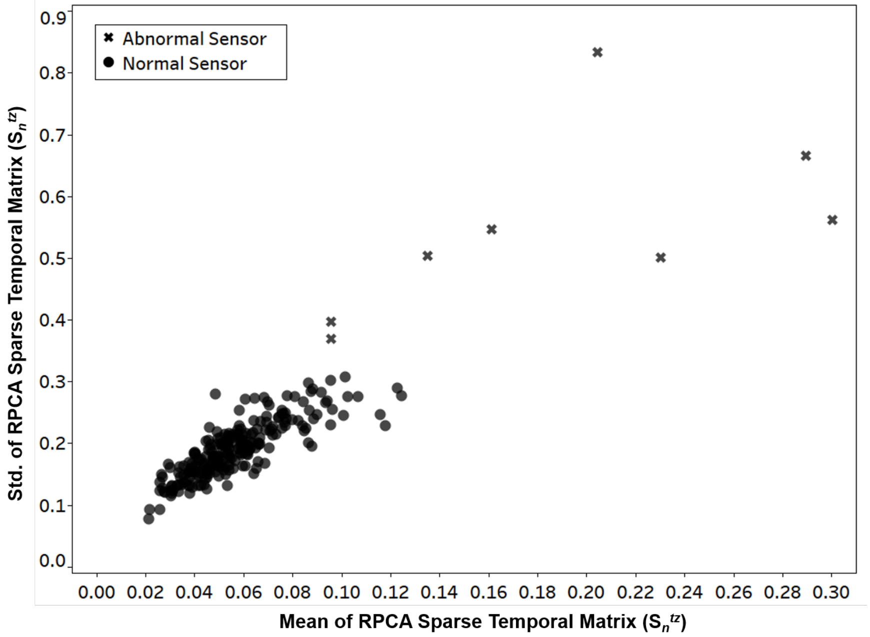


Figure 9: RPCA sparse temporal matrix clustering

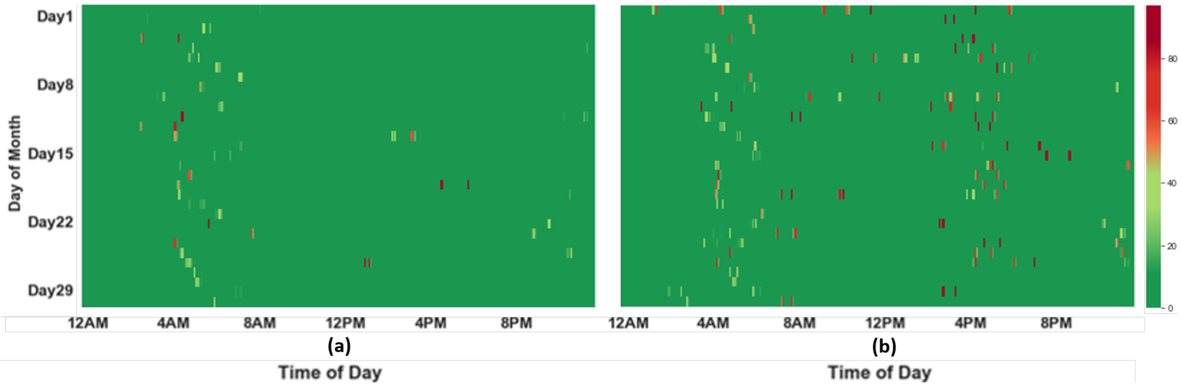


Figure 10: RPCA Sparse Matrix Clustering heatmap

Like before in Figure 5, in Figure 11 we verify the necessity of our TSHM modules's step 3 temporal pattern anomaly test by comparing its results with our step2 $AEVL$ anomaly test.Specifically, we plot the 2-d $AEVL$ distribution $(\mu_n^{AEVL}, \sigma_n^{AEVL})$ obtained from step 2 with the labels of the sensors obtained from step 3. Again, it can be seen that the abnormal sensors with frequent $AEVL$ fluctuations detected in step 3 cannot be identified in step 2, justifying that our temporal pattern anomaly test is required.

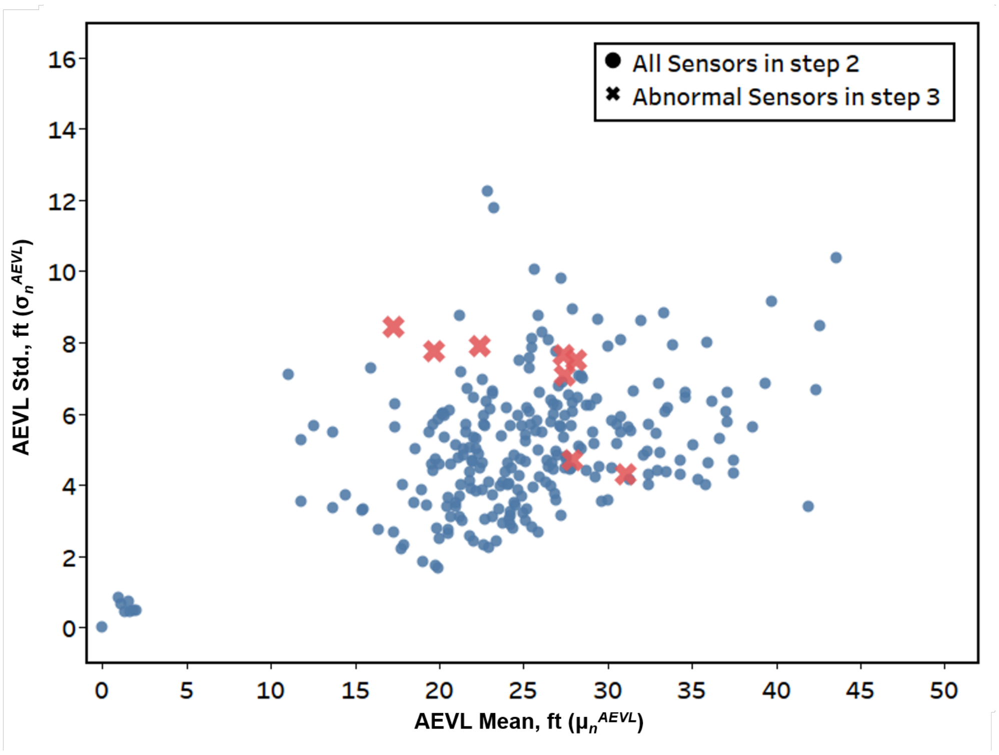


Figure 11: AEVL plots for all sensors in step 2 and abnormal sensors in step 3

4.4. *Baseline Comparison*

In this section, we compare our proposed TSHM module with the baseline CLM and MCB methods described in Section 3.5.

4.4.1. *CLM Comparison*

The CLM method attempts to detect abnormal sensors based on $AEVL$ values. Since CLM method doesn't deal with any missing data issues, so we don't use the results obtained from our step 1 (data completeness test) in this comparison and only rely on the remaining two steps ($AEVL$ anomaly test and temporal pattern anomaly test) since these also use $AEVL$ as the primary variable. Figure 12 shows the comparison's results for the sample month July 2017. All anomalous sensors detected using the CLM method (3.04% or 8 out of 263) were also labelled anomalous by our method, but our proposed method also labelled an additional 13 sensors (4.94%) as anomalous, 5 in step 2 and 8 in step 3.

Figure 13 shows heatmaps of the raw $AEVL$ values for the study month (July 2017) for a sample normal sensor and three different abnormal sensors. Figures 13a and 13b show heatmaps for sample sensors labeled normal and abnormal sensor, by both CLM and our proposed method. Figures 13c and 13d show $AEVL$ heatmaps of sample sensors detected only by our proposed TSHM module (in step 2 and 3) that reported either intermittent abnormal $AEVL$ values or frequent $AEVL$ fluctuations, thereby justifying the efficacy of the proposed method.

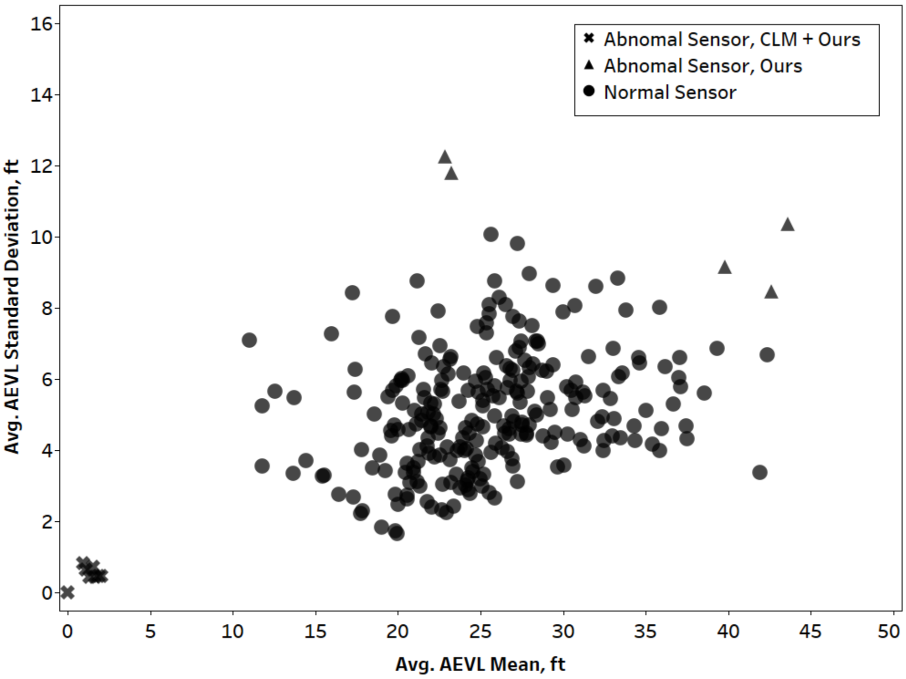


Figure 12: Clustering comparison with CLM method

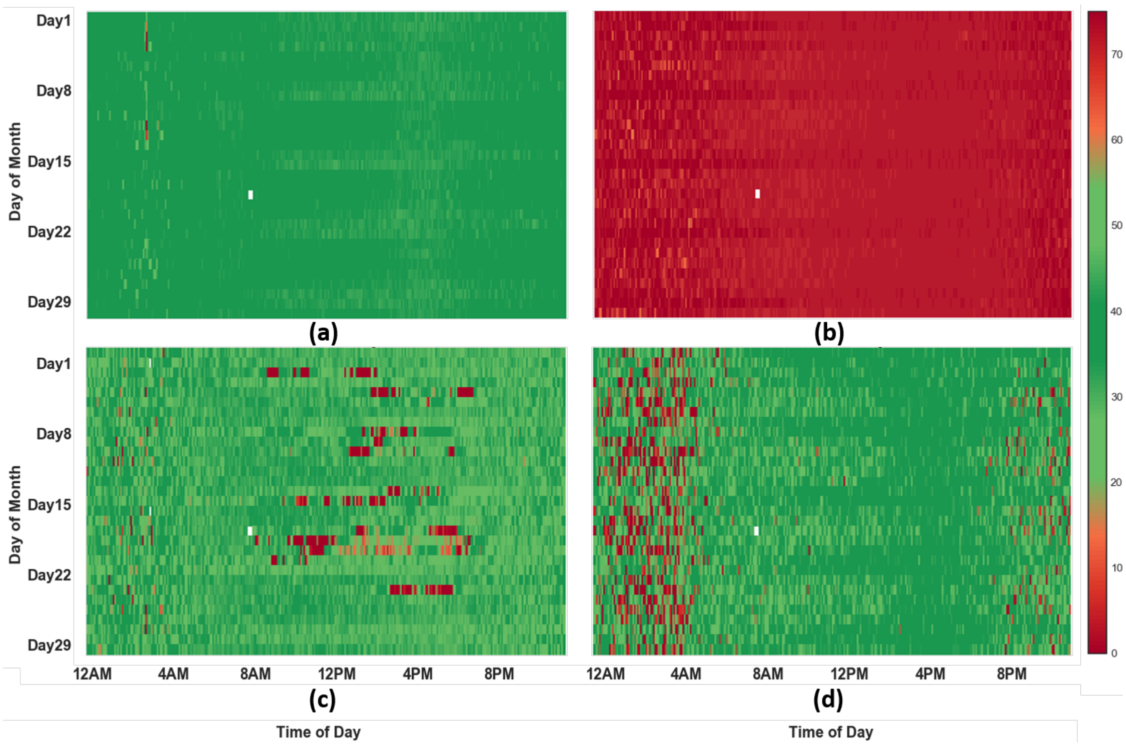


Figure 13: Sample monthly *AEVL* heatmap of July, 2017 for sample (a) normal sensor and (b-d) abnormal sensors

4.4.2. MCB Method

The MCB algorithm requires sensors to be spatially ordered. However, generating the accurate spatial ordering of all sensors in Iowa is time-consuming. Therefore, we selected 4 different routes for evaluation, namely I-235 EB (with the 6 sensors A1–A6), I-35 NB (with the 6 sensors B1–B6), I-80 EB (with the 7 sensors C1–C7), and I-74 NB (with the 7 sensors D1–D7). Each route's head and tail sensors were ignored in the evaluation, since they do not have the u/s and d/s sensors required for MCB comparison. Figure 14 shows the remaining 18 sensors' normal/abnormal status according to our proposed method vs. MCB. Our proposed method labels 7 of these as anomalous (A3, A5, C2, C5, D3, D4, and D5), whereas MCB labels only 4 (A5, B3, C2, and C5) that have considerably higher missing data and potential error percentages as anomalous (under our assumed threshold of 20%, since Lu et al. (2014) did not propose any definite threshold for labelling anomalous sensors based on error percentages).

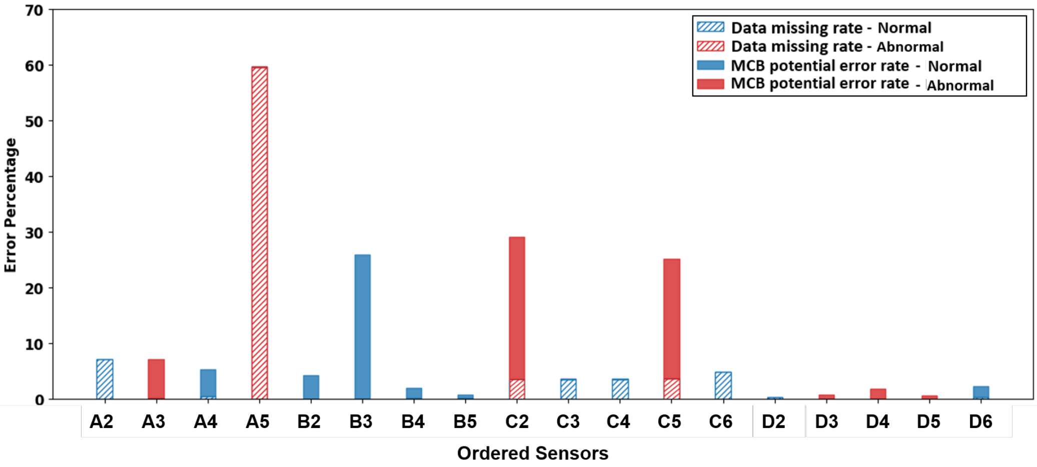


Figure 14: MCB algorithm based error percentages of sensors with labels from the proposed method

The greater efficacy of our proposed TSHM module compared to the MCB algorithm is additionally supported by Figure 15 raw *AEVL* heatmap in that:

- Both the MCB algorithm and our proposed method detect as abnormal the *AEVL* records for sensor A5, with its high missing data percentage observable in Figure 15a *AEVL* heatmap, as well as sensors C2 and C5 (cf. Figure 15c), with their substantially different *AEVL* both globally and with respect to their upstream (u/s) and downstream (d/s) sensors.

- The MCB algorithm labels sensor B3 as abnormal due to high potential error rate, but visual examination of B3's raw *AEVL* heatmap in Figure 15b shows no substantial *AEVL* abnormality. Further investigation reveals B3 and its nearest neighboring sensor B4 (2.1 miles away) to be located at two freeway interchanges where entering and exiting vehicles apparently affect vehicle composition and distribution (e.g., the B3 and B4 average hourly truck percentages were 9.34% and 16.12%, respectively). Therefore, the MCB method appears to be overly sensitive to vehicle composition varying between nearby locations.

- Our proposed TSHM module's step 3 temporal pattern anomaly test captures substantial temporal AEVL fluctuations (cf. Figure 15d raw AEVL plot for sensor D3), which led our method to report as abnormal the sensors D3, D4, and D5 that the MCB algorithm classifies as normal.

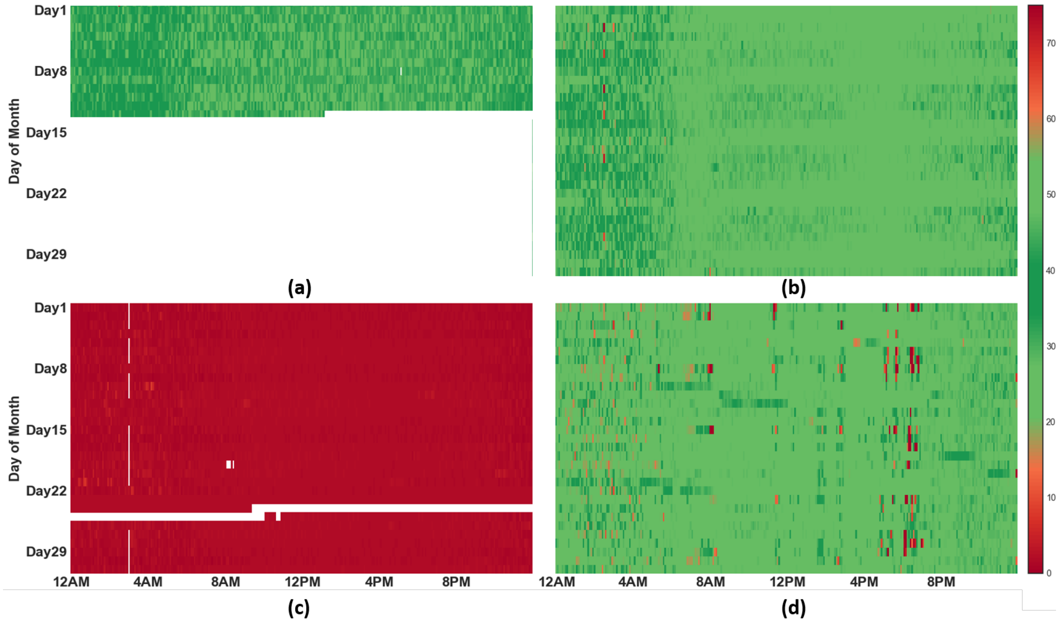


Figure 15: $AEVL$ heatmap of (a) A5, (b) B3, (c) C2, and (d) D3

5. Conclusion

This study proposes a large-scale, data-driven traffic sensor health monitoring (TSHM) module involving massively parallelizable data processing techniques that make it feasible to deploy over large traffic networks. Our proposed TSHM module can be compared with sieving analysis, where each step identifies distinct sensor abnormalities, enabling traffic management authorities to take the necessary steps to resolve. First, our module's data completeness test captures sensors with abnormally high missing data rates, providing a data completeness score (CS) that justifies assigning different levels of missing data severity. Second, reduced 2-d features from each sensor's aggregated $AEVL$ distribution are used to detect abnormal sensors based on DBSCAN clustering's anomaly detection logic. (We used the $AEVL$ metric since not only can it capture the variability of all three basic traffic variables (speed, density, and volume) simultaneously, but also it is a surrogate of vehicle length robust to daily or seasonal traffic variations and other external factors like inclement weather or traffic incidents.) Third, our novel temporal-pattern-based anomaly detection method utilizes the $AEVL$ assumption of constancy in the vehicle length distribution by introducing Bayesian changepoint detection in the temporal $AEVL$ matrix to detect sensors in the data stream reporting abnormally frequent spikes/fluctuations that suggest sensor problems requiring further attention.

One major challenge in abnormal sensor detection is the difficulty in obtaining groundtruth labels. Due to the absence of any explicit definition of abnormal sensors in the literature, this study has identified abnormal sensors by plotting sensor data along two different feature dimensions to identify points of agreement. For example, step 2 of our proposed TSHM module uses aggregated $AEVL$ records to detect abnormal sensors, but we also verify this cumulative feature vector's abnormality by comparing apparently abnormal sensors' CDF plots of with that of their adjacent u/s and d/s sensors. Similarly, we justify abnormal sensors identified by our step 3 temporal pattern anomaly test by demonstrating their raw $AEVL$ heatmaps also show frequent spikes. Finally, we compare our proposed module with two benchmark algorithms, the CLM and MCB. These baseline comparisons show our proposed method can successfully identify not only typical sensor error types, such as missing data or abnormal records, but also advanced error types such as frequent abrupt sensor data fluctuations. In addition, the efficacy of our proposed method is demonstrated in how, unlike the MCB algorithm, our method can successfully identify such abnormalities even for isolated or consecutive abnormal sensors.

However, our algorithm is an offline method that builds its model using historical traffic data. In future, our method can be extended to incorporate real-time sensor health monitoring to enable instant detection of abnormal sensors. Also, our proposed TSHM module's performance and reliability can likely be improved using traffic information from other sensor types (e.g., probe and camera data).

Acknowledgements

Our research results are based upon work jointly supported by the National Science Foundation Partnerships for Innovation: Building Innovation Capacity (PFI: BIC) program under Grant No. 1632116, National Science Foundation under Grants No. CNS-1464279, CCF-1566281, CCF-1750920, Award ATD-1830254 supported by NSF and the National Geospatial-Intelligence Agency, and Iowa DOT Office of Traffic Operations Support Grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Al-Deek, H.M., Venkata, C., Chandra, S.R., 2004. New algorithms for filtering and imputation of real-time and archived dual-loop detector data in i-4 data warehouse. Transportation research record 1867, 116–126.

Apache Pig, 2018. https://pig.apache.org/. Accessed July 20, 2018.

Berkhin, P., 2006. A survey of clustering data mining techniques, in: Grouping multidimensional data. Springer, pp. 25–71.

Bouwmans, T., Zahzah, E.H., 2014. Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance. Computer Vision and Image Understanding 122, 22–34.

Candès, E.J., Li, X., Ma, Y., Wright, J., 2011. Robust principal component analysis? Journal of the ACM (JACM) 58, 11.

Chakraborty, P., Adu-Gyamfi, Y.O., Poddar, S., Ahsani, V., Sharma, A., Sarkar, S., 2018a. Traffic congestion detection from camera images using deep convolution neural networks. Transportation Research Record: Journal of the Transportation Research Board 2672, 222–231. doi:10.1177/0361198118777631.

Chakraborty, P., Sharma, A., Hegde, C., 2018b. Freeway traffic incident detection from cameras: A semi-supervised learning approach, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 1840–1845.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR) 41, 15.

Chen, C., Kwon, J., Rice, J., Skabardonis, A., Varaiya, P., 2003. Detecting errors and imputing missing data for single-loop surveillance systems. Transportation Research Record 1855, 160–167.

Erman, J., Arlitt, M., Mahanti, A., 2006. Traffic classification using clustering algorithms, in: Proceedings of the 2006 SIGCOMM workshop on Mining network data, ACM. pp. 281–286.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, pp. 226–231.

Fearnhead, P., 2006. Exact and efficient bayesian inference for multiple changepoint problems. Statistics and computing 16, 203–213.

Fearnhead, P., Liu, Z., 2007. On-line inference for multiple changepoint problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69, 589–605.

Jacobson, L.N., Nihan, N.L., Bender, J.D., 1990. Detecting erroneous loop detector data in a freeway traffic management system. 1287.

Klein, L.A., Mills, M.K., Gibson, D.R., et al., 2006. Traffic detector handbook: Volume I. Technical Report. Turner-Fairbank Highway Research Center.

Kodinariya, T.M., Makwana, P.R., 2013. Review on determining number of cluster in k-means clustering. International Journal 1, 90–95.

Lee, H., Coifman, B., 2011. Quantifying loop detector sensitivity and correcting detection problems on freeways. Journal of Transportation Engineering 138, 871–881.

Lu, Y., Yang, X., Chang, G.L., 2014. Algorithm for detector-error screening on basis of temporal and spatial information. Transportation Research Record: Journal of the Transportation Research Board , 40–48.

Ma, D., Luo, X., Li, W., Jin, S., Guo, W., Wang, D., 2017. Traffic demand estimation for lane groups at signal-controlled intersections using travel times from video-imaging detectors. IET Intelligent Transport Systems 11, 222–229.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA. pp. 281–297.

Mori, U., Mendiburu, A., Álvarez, M., Lozano, J.A., 2015. A review of travel time estimation and forecasting for advanced traveller information systems. Transportmetrica A: Transport Science 11, 119–157.

Nisa, K.K., Andrianto, H.A., Mardhiyyah, R., 2014. Hotspot clustering using dbscan algorithm and shiny web framework, in: 2014 International Conference on Advanced Computer Science and Information System, IEEE. pp. 129–132.

Payne, H., Thompson, S., 1997. Malfunction detection and data repair for induction-loop sensors using i-880 data base. Transportation Research Record: Journal of the Transportation Research Board , 191–201.

Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X., 2017. Dbscan revisited, revisited: why and how you should (still) use dbscan. ACM Transactions on Database Systems (TODS) 42, 19.

Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transportation Research Part C: Emerging Technologies 58, 380–394.

Sun, Z., Jin, W.L., Ng, M., 2016. Network sensor health problem. Transportation Research Part C: Emerging Technologies 68, 300–310.

Turner, S., Albert, L., Gajewski, B., Eisele, W., 2000. Archived intelligent transportation system data quality: Preliminary analyses of san antonio transguide data. Transportation Research Record: Journal of the Transportation Research Board , 77–84.

Turochy, R.E., Smith, B.L., 2000. New procedure for detector data screening in traffic management systems. Transportation Research Record 1727, 127–131.

Vanajakshi, L., Rilett, L., 2004. Loop detector data diagnostics based on conservation-of-vehicles principle. Transportation research record 1870, 162–169.

Wells, T.J., Smaglik, E.J., Bullock, D.M., 2008. Implementation of station health monitoring procedures for its sensors, volume 1 .

Wu, L., Liu, C., Huang, T., Sharma, A., Sarkar, S., 2017. Traffic sensor health monitoring using spatiotemporal graphical modeling, in: Proceedings of the 2nd ACM SIGKDD Workshop on Machine Learning for Prognostics and Health Management, pp. 13–17.

Yao, B., Chen, C., Cao, Q., Jin, L., Zhang, M., Zhu, H., Yu, B., 2017. Short-term traffic speed prediction for an urban corridor. Computer-Aided Civil and Infrastructure Engineering 32, 154–169.