RESEARCH

UDOT Keeping Utah Moving
Utah Department of Transportation - Research Division
4501 South 2700 West - P.O. Box 148410 - SLC,UT 84114-8410

Report No. UT-20.15

MULTI-STAGE ALGORITHM FOR DETECTION-ERROR IDENTIFICATION AND DATA SCREENING

Prepared For:

Utah Department of Transportation
Research & Innovation Division

Final Report
October 2020

DISCLAIMER

The authors alone are responsible for the preparation and accuracy of the information, data, analysis, discussions, recommendations, and conclusions presented herein. The contents do not necessarily reflect the views, opinions, endorsements, or policies of the Utah Department of Transportation or the U.S. Department of Transportation. The Utah Department of Transportation makes no representation or warranty of any kind, and assumes no liability therefore.

ACKNOWLEDGMENTS

The authors acknowledge the Utah Department of Transportation (UDOT) for funding this research, and the following individuals from UDOT on the Technical Advisory Committee (TAC) for helping to guide the research:

- Kelly Njord (Traffic Performance Engineer)
- Travis Jensen (Research Consultant Project Manager)
- Glenn Blackwelder (Safety Operations Engineer)
- Jamie Mackey (Freeway Operations Manager)
- Mark Taylor (Traffic Signal Operations Engineer)
- Mark Parry (Freeway Operations Engineer)

TECHNICAL REPORT ABSTRACT

1. Report No. UT-20.15	2. Government Accession No. N/A	3. Recipient's Catalog No. N/A

4. Title and Subtitle Multi-Stage Algorithm for Detection-Error Identification and Data Screening	5. Report Date October 2020 6. Performing Organization Code N/A
7. AuthorS Bahar Azin, Xianfeng Terry Yang	8. Performing Organization Report N/A
9. Performing Organization Name and Address University of Utah Department of Civil & Environmental Engineering 110 Central Campus Drive, Suite 2000 Salt Lake City, UT 84112	10. Work Unit No. 5537915D; 7283401D 11. Contract or Grant No. 19-8770
12. Sponsoring Agency Name and Address Utah Department of Transportation 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410	13. Type of Report & Period Covered Final Jan 2019 to Oct 2020 14. Sponsoring Agency Code UT18.301

15. Supplementary Notes

Prepared in cooperation with the Utah Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration

16. Abstract

During the past decades, roadside traffic detectors have been widely deployed in traffic management systems to help transportation agencies monitor and control traffic. However, detector data may contain some erroneous information caused by the malfunctioning of the detection system. Typically, these errors result from the lack of maintenance and the need for device recalibration, which can affect the decision-making processes that use detected data as their basis. To identify these errors in the database, a reliable screening algorithm is needed to examine the quality of recorded detector data, identify potential errors, and find the detection stations that require maintenance or recalibration. To fulfill such needs, this research project develops a multi-stage screening algorithm and fully considers the impacts of detector locations to the screening process as they would affect the data and the comparison source. First, the quality of data will be pre-evaluated and primary errors will be identified. Second, statistical analysis of data, according to their station locations, will be performed to confirm the pre-identified errors. Last, an in-depth review of stations will be carried out to certify the stations with potential errors. Notably, locating the detection stations with potential failures will help UDOT traffic engineers prioritize the detectors that need immediate attention, as timely actions on those detectors are essential to support the functioning of various traffic management tasks.

17. Key Words Traffic Management, Detector Error, Data Screening, Station Location, Average Effective Vehicle Length, Data Distribution.		18. Distribution Statement Not restricted. Available through: UDOT Research Division 4501 South 2700 West P.O. Box 148410 Salt Lake City, UT 84114-8410 www.udot.utah.gov/go/research	23. Registrant's Seal N/A
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 55	22. Price N/A

TABLE OF CONTENTS

LIST OF TABLES v

LIST OF FIGURES vi

LIST OF ACRONYMS vii

Executive Summary 1

1.0 INTRODUCTION 3

- 1.1 Problem Statement 3
- 1.2 Objectives 4
- 1.3 Scope 5
- 1.4 Outline of Report 5

2.0 LITERATURE REVIEW 6

- 2.1 Overview 6
- 2.2 Data Screening Algorithm 6

3.0 INVESTIGATION OF DATASETS AND ALGORITHM DEVELOPMENT 10

- 3.1 Overview 10
- 3.2 Potential Errors Within the PeMS Database 10
  - 3.2.1 Missing Data 12
  - 3.2.2 Large Variations Within the Data 12
  - 3.2.3 Out-of-Range Data 13
  - 3.2.4 Inconsistencies 13
- 3.3 Multi-Stage Data Screening Algorithm 14
  - 3.3.1 Stage 1: Primary Screening 14
  - 3.3.2 Stage 2: Piecewise Quality Check 15
  - 3.3.3 Stage 3: Continuous Quality Check 18
  - 3.3.4 Step 4: Further Investigation 19

4.0 APPLICATION OF THE DATA SCREENING ALGORITHM 22

- 4.1 Overview 22
- 4.2 Error-Identification Case Study 22
- 4.3 Error-Identification Results 23

5.0 IDENTIFICATION OF HIGH-SPEED SPOTS 32
5.1 Overview 32
5.2 Data Description 32
5.3 Using the Data Screening Algorithm to Validate Datasets 33
5.4 Speed Data Analysis and High-Speed Locations 34
6.0 CONCLUSIONS 41
6.1 Summary 41
6.2 Findings 41
6.3 Limitations and Challenges 42
REFERENCES 43

LIST OF TABLES

Table 4.1 Summary of data used for the case study 23

Table 4.2 Single variable thresholds..23

Table 4.3 Distribution of data errors based on station location and algorithm stage....................24

Table 4.4 Detector error summary based on station type 26

Table 4.5 Detector error summary based on the traffic flow rate 27

Table 4.6 Malfunctioning stations on I-15 northbound 28

Table 5.1 Counts of records obtained from detectors utilizing each approach..............................33

Table 5.2 One-hour maximum higher than 85 mph records for eastbound stations......................37

Table 5.3 One-hour maximum higher than 85 mph records for westbound stations....................38

LIST OF FIGURES

Figure 3.1 Traffic flow of on-ramp station vs. merged flow 16

Figure 3.2 Traffic flow of off-ramp station vs. merged flow..16

Figure 3.3 Detector map for spatial comparison..17

Figure 3.4 Detector comparisons with farther stations. 20

Figure 3.5 The data screening algorithm 21

Figure 4.1 I-15 northbound segment for case study (302.75-312.16) 22

Figure 4.2 Distribution of data points after algorithm implementation..24

Figure 4.3 Stacked distribution of error detection stage based on station type 25

Figure 4.4 Pie chart of errors by detector location..26

Figure 4.5 Distribution of flow level for various errors..27

Figure 4.6 Location of malfunctioning on-ramp station 1418...29

Figure 4.7 Location of malfunctioning fwy-fwy station 3422...29

Figure 4.8 Location of malfunctioning on-ramp station 1431 and off-ramp station 91431 30

Figure 4.9 Error rate of stations by locations...31

Figure 5.1 Geographic positioning of detectors on the I-80 corridor 32

Figure 5.2 Ratio of erroneous records by station location...33

Figure 5.3 Error type distribution for eastbound 34

Figure 5.4 Error type distribution for westbound 34

Figure 5.5 Spatial distribution of the 85th percentile speed for eastbound stations 35

Figure 5.6 Spatial distribution of the 85th percentile speed for westbound stations 36

Figure 5.7 Maximum speeds at eastbound stations 36

Figure 5.8 Maximum speeds at westbound stations 37

Figure 5.9 The percentage of records over the threshold on the eastbound approach...................39

Figure 5.10 The percentage of records over the threshold on the westbound approach................39

Figure 5.11 High-speed locations at I-80 freeway...40

LIST OF ACRONYMS

AEVL	Average Effective Vehicle Length
Caltrans	California Department of Transportation
CI	Confidence Interval
FHWA	Federal Highway Administration
HCM	Highway Capacity Manual
HOV	High-Occupancy Vehicle
ITS	Intelligent Transportation System
K-S	Kolmogorov-Smirnov
LOS	Level of Service
MCB	Multiple Comparison with the Best
NEMA	National Electrical Manufacturers Association
PeMS	Performance Measurement System
TGT	Technical Ground Truth
TOC	Traffic Operations Center
UDOT	Utah Department of Transportation
VDOT	Virginia Department of Transportation
VHT	Vehicle Hours Traveled
VMT	Vehicle Miles Traveled
VSL	Variable Speed Limit

EXECUTIVE SUMMARY

In traffic management systems, detectors are deployed to obtain real-time traffic information and support various traffic control functions. The most commonly used detectors, such as loop and radar detectors, can measure traffic flow, speed, and occupancy on major roads. In the state of Utah, freeway detector data are accessible through online platforms, such as UDOT's Performance Measurement System (PeMS). However, it should be noted that detectors' performance might become unstable over time due to lack of maintenance and recalibration. Impaired detectors will produce errors in the databases and could affect traffic control system decisions. Erroneous data usually can be found by exploring the historical data of a detector, in comparison to the rest of detectors in the same corridor. In the literature, preliminary screening algorithms usually first identify the missing data. Then, the overall data quality is evaluated based on traffic flow characteristics and the conservation law, where traffic flow patterns are assumed to be steady through time, and traffic states such as speed and flow rate should have reasonable values. However, such preliminary screenings can only find the obvious errors. Hence, statistical tests need to be conducted for an in-depth review of data distribution and to further identify the potential data errors that cannot be examined by the preliminary screening process. With the completion of the screening, a data quality evaluation report could be generated and those malfunctioning detectors would be found.

This research project uses UDOT's detector database, PeMS, to construct an error-identification algorithm. The algorithm uses a multi-stage scheme that pinpoints the potential errors in the database and marks the detectors that need to be corrected through maintenance or recalibration. First, the algorithm looks for missing data and irregular recorded parameters through the preliminary data screening. Then, the distribution of Average Effective Vehicle Length (AEVL) recorded by detectors is examined in two stages. In each stage, data from adjacent detectors are compared by using the Kolmogorov-Smirnov (K-S) test and Multiple Comparison with the Best (MCB) method. Notably, when a potential malfunctioning detector is identified, its data will be eliminated in further MCB tests as this method aims to compare the data from adjacent detectors. Results of the developed screening algorithm can help locate detectors that need to be prioritized for maintenance and recalibration. The algorithm can also

help traffic management systems to monitor detector performance and keep the recorded datasets validated.

For algorithm implementation, a freeway corridor in Utah is selected for case study. Using the validated datasets, the recorded data are analyzed with the algorithm and the corresponding data quality is evaluated. In addition, high-speed locations are also identified by investigating speed data after the screening process.

1.0 INTRODUCTION

1.1 Problem Statement

In traffic management systems, how to improve the efficiency of road traffic performance is always a vital issue due to increasing traffic demand. For example, implementing Intelligent Transportation Systems (ITS) can help achieve this goal by using real-time control to better utilize the capacity of existing road networks (Lawrence A. Klein et al., 2006). Particularly, traffic detection systems are one type of ITS facility that are designed to show the presence or passage of vehicles in a road segment, as defined by the National Electrical Manufacturers Association (NEMA). In the state of Utah, two types of detectors are commonly used to measure flow, speed, and occupancy of traffic: loop detectors and radar sensors. The collected data are accessible through online platforms such as the Freeway Performance Metrics (i.e., occurrence data) and the Performance Measurement System (PeMS). These platforms are maintained by staff at UDOT's Traffic Operation Center (TOC).

In practice, much decision-making in traffic operations tasks is inferred by measured data from detectors. Therefore, malfunctioning detectors can produce data errors or biases and affect control decisions. Accordingly, a reliable and efficient data screening algorithm is needed to prevent such situations by identifying potential errors in the database, evaluating the detector performance, and notifying the system managers of the need for detector maintenance. In the literature, commonly used screening algorithms often search for obvious errors based on traffic characteristics and flow conservation laws, considering that recorded values must be within reasonable ranges. As an extension, the value of the Average Effective Vehicle Length (AEVL) is used as a measurement to inspect the quality of the dataset with more in-depth analysis (Payne, Harold J., E. D. Helfenbein, 1976). For example, the comparison of AEVL distributions by adjacent detectors in the same road segment is applied to check for data consistency and identify potential errors (Lu et al., 2014).

The algorithm developed in this research project consists of three main stages that can evaluate the quality of PeMS data. The data include the 5-min aggregated mean speeds and flow rates, which are computed based on the occurrence data. In the first stage, the primary screening is conducted to identify missing data and to perform single and multiple variable threshold checks. Specific data errors are identified based on fundamental traffic concepts and missed data due to the system malfunctioning, e.g., deactivation of the detector.

In the second stage, statistical analysis is performed on each station with a comparison of its AEVL distribution to that of adjacent stations. This stage is a dual comparison based on the traffic flow theory, where the significant variation of AEVL at two successive stations, depending on the flow characteristics, may imply an error in the recorded data. Notably, the analysis method shall fully account for the impact of detection locations. For example, traffic from an on-ramp may affect the AEVL at the downstream detection site.

The last stage uses another advanced statistical method, named Multiple Comparison with the Best (MCB), to find all potential errors in a road segment that may not be caught by the previous two screening stages and to identify possible malfunctioning detectors. In statistics, the MCB method is designed to concurrently compare data in multiple sets and to identify the set that is most different from all others. In this stage, MCB is implemented to concurrently compare AEVLs at multiple adjacent stations. If the AEVL at one station is found to be statistically different from AEVLs at other stations, the detector at that station can be marked with a "malfunctioning" label and its data would be eliminated in future comparisons.

1.2 Objectives

The first main objective of this research project is to develop a multi-stage data screening tool to evaluate the data quality of UDOT's PeMS database and mark probable detection errors. After the preliminary screening, which is based on traffic flow/speed thresholds and traffic conservation laws, the advanced data screening is achieved by implementing statistical models to identify the potential errors in the datasets. Also, the algorithm will be used to perform an in-depth review of detector stations and to highlight the ones that need maintenance or re-calibration.

Using the data screened by the developed algorithm, another objective of this research project is to conduct an in-depth analysis of vehicle speed profiles. The purpose of this objective is to identify high-speed locations that pose high crash risks.

1.3 Scope

There are four tasks involved in this research. Task 1 is a literature review that focuses on reviewing papers and resources related to detector data screening methods. Task 2 involves algorithm development and the identification of potential detection errors within Utah's data. Task 3 demonstrates the algorithm by applying it to a real dataset in order to locate malfunctioning stations. Task 4 applies the algorithm to a group of datasets to validate them and remove invalid detector data. It then analyzes the resulting validated speed data.

1.4 Outline of Report

The remainder of this report is structured as follows:

- Literature review
- Investigation of datasets and algorithm development
- Application of data screening algorithm
- High-speed spot identification
- Conclusions and key findings

2.0 LITERATURE REVIEW

2.1 Overview

This chapter presents the research background and reviews existing studies that focus on detection error identification and data screening. According to the literature, many scholars have developed error identification algorithms to detect potential errors in detection databases. Also, they have utilized a variety of tools and approaches to evaluate data records in ITS applications, which can be useful depending on the scope and limitations of a given use case.

2.2 Data Screening Algorithm

Malfunctioning detectors, due to lack of maintenance or recalibration, are often the cause of errors in databases. In traffic management systems, such errors could negatively affect control decisions and pose risks of downgrading systems' performances. However, those erroneous data usually can be tracked over time. Hence, an effective data screening algorithm is necessary for evaluating the detector performance and notifying traffic engineers when maintenance is needed. The literature contains many studies that identify potential detection errors and develop data screening algorithms to determine the reasons for these failures. According to a Federal Highway Administration (FHWA) report, methods used to establish the criteria of detector data screening can be categorized into the following three groups (Turner, 2007):

- The variation range check of thresholds for both singular and combinations of variables,
- The consistency of traffic characteristics within the spatial and temporal recorded data, and
- An in-depth diagnosis including a supplementary estimation to be implemented

The earliest studies in this area mainly focused on tracking single variables and relied on basic traffic engineering regulations. In 1976, Payne et al. used aggregated volume, speed, and occupancy data to develop an incident detection algorithm and introduced a threshold for each variable to specify any inaccurate information in databases (Payne, Harold J., E. D. Helfenbein, 1976). As the traffic flow theory evolved, it became possible to improve the data screening algorithm. The basic traffic parameters and boundaries were used to evaluate data recorded by single loop detectors (Chen et al., 2019). Historical data and upstream/downstream data, with the combination of control parameters, were used to identify unusual records in paired-loop systems (Cleghorn et al., 1991). Turochy also used multiple variable thresholds based on traffic flow theory principles to mark the errors in existing datasets (Turochy and Smith, 2000). FHWA suggests that quick variations between variables in consecutive periods of time implies that the datapoints are inaccurate (Federal Highway Administration, 2012). The thresholds used by many scholars were determined to be dependent on where the lane detector is placed as well as the features of passing traffic flow (Hamad, 2015).

In early years, Chen et al. (1976) utilized time series data to develop an evaluation model based on linear regression with neighboring historical loop detectors. Vanajakshi and Rilett also used the concept of vehicle conservations through nonlinear optimization modeling to uncover detector errors (Vanajakshi and Rilett, 2004). Turochy and Smith classified the data collected from detectors using thresholds of plotting time-series data, which showed constant values over time or higher values compared to the adjacent detectors at specified times are inaccurate (Turochy and Smith, 2002).

Since occurrence data is commonly used to take measurements for traffic management and operations, it is also essential to validate the accuracy of these measurements. Using individual vehicle records, the AEVL concept was found to be an efficient way to monitor the occurrence data (Bullock and Achillides, 2004; J. Wells et al., 2008; Lu et al., 2014; Turochy and Smith, 2000). The value of AEVL and its criteria have been used in a case study at the Virginia Department of Transportation (VDOT) to screen and prepare the data for further use in freeway management (Turochy and Smith, 2002). Wells et al. devised a six-step procedure to analyze data using the AEVL concept alongside historical data (J. Wells et al., 2008). The findings proved that AEVL is an effective operational approach for managing online data collected by detectors. Yu and Zhijie used the AEVL for real-time data as well as parameter thresholds and the probability distribution of vehicle arrivals to screen traffic data (Yu and Zhijie, 2016). Along the same line, Lu et al. (2018) set a threshold for AEVL to remove anomalous records between 10 and 75 feet.

A recent study by Zhang et al. (2019), focusing on wrong-way driving hotspots, suggested a new screening algorithm for data validation. They made use of probability distribution with the t-test comparison method to determine if detector data was inaccurate. Furthermore, Lu et al. (2014) utilized data quality and error identification based on the AEVL distribution to compare lanes and detectors. They developed a multi-stage algorithm to first check for primary data errors including missing data and values. Then, with the implementation of the AEVL concept, the probable erroneous records were marked by comparing these detectors with inflow stations, and the malfunctioning detectors were identified. Combining AEVLs with the MCB method (Hsu and Nelson, 2003) created an efficient algorithm to identify detector errors. The MCB method was found to reduce the number of comparisons needed for cases with large sample sizes and employs the best performing data to analyze other samples' behavior (Hochberg and Tamhane, 1987; Horrace and Schmidt, 2000). Along the same track, some studies have shown that the AEVL varies in time and between lanes (Maghrour Zefreh et al., 2017; Zhanfeng Jia et al., 2001). Meanwhile, it should be noted that PeMS is an interface that helps planners and engineers access real-time traffic network data. It collects data from detectors all over the network and turns them into useful and understandable information that is accessible to users. More specifically, the data (speed and flow) are collected through detectors within a specific period of time and are converted to aggregated information that can be retrieved online (Chen, 2003). Therefore, the use of AEVL variations to identify potential detection errors in the PeMS database is applicable to the research being conducted in this study (Chen, 2003; Varaiya, 2004).

Recently, the popularity of machine learning methods has also yielded data screening algorithm tools. For example, a group of studies utilized the K-mean clustering method to remove outliers and identify anomalous data points (Lin et al., 2012; Megler et al., 2016). Fuzzy clustering, based on the relationships of three basic traffic parameters, is also used to screen and evaluate records (Ishak, 2003). The stochastic process in the Markov model also contributed to

ITS data screening methods. This method with two other supplemental modules is used to identify sensors' status instead of evaluating the measurements (Randeniya and Kim, 2013).

In a more recent study, a new technique called technical ground truth (TGT), is introduced to use an indicator to evaluate the qualities of real-time traffic data. Hubber et al. (2014) estimated the key traffic parameter, compared it to the one recorded by the detector, and used the error rate to measure the data quality. After using an index for data consistency evaluation, the obtained information was then analyzed both spatially and temporally to evaluate real-time traffic data (Park et al., 2015). Spatio-temporal data provided a basis for further studies using graphical modeling of the detector network to demonstrate the data quality (Wu et al., 2018). A quality check also can be accomplished through evaluation of data from various resources, as was performed by Ackaah et al. (2016) in their comparison of loop detector data with space-time traffic information using a variable speed limit (VSL) system. Predicting a parameter such as level of service (LOS) with data obtained from other resources can also help traffic system managers in evaluating detector data quality (Xiao et al., 2015).

3.0 INVESTIGATION OF DATASETS AND ALGORITHM DEVELOPMENT

3.1 Overview

At UDOT, many real-time traffic data are collected by detectors and are made available through the PeMS online platform. Inaccuracies caused by various factors can be identified in this database. This section classifies the potential errors into several types by exploring PeMS data records. Then, the corresponding error-identification method for each type is discussed.

3.2 Potential Errors Within the PeMS Database

PeMS is a web-based platform that gives users access to offline and online traffic data that have been recorded by all types of detectors in the state of Utah. Many analytical tools embedded in this platform can help transportation planners and engineers obtain the information needed for traffic analysis. Data provided by PeMS can also help TOC employees monitor traffic flow patterns and make decisions to establish a more efficient and safe traffic operational environment (Caltrans, 2002).

The PeMS system collects data from detectors all over the network and transforms them into useful and understandable information for users. It receives vehicle count and occupancy data at 20-second intervals. Then, it calculates the *f*-factor of each detector based on the *g*-factor to measure the speed (Choe et al., 2002). Notably, Caltrans defines the data process differently, in which the detector (i) senses vehicles through time and reports the vehicle count (Q_i), average occupancy (K_i), and average speed (V_i). Average speed is only measured directly from double loop detectors. However, for other types of detectors, it is measured from vehicle miles traveled (*VMT*) and vehicle hours traveled (*VHT*). *VMT* and *VHT* are defined by a period of time (t) on a segment of road ($i=1, 2, \ldots, n$). Therefore, these two variables are the sum of *VMT* and *VHT* at a specific segment of the route, as shown in the equations below:

$$VMT(t) = \sum_{i=1}^{n} VMT_i(t) \tag{3.1}$$

$$VHT(t) = \sum_{i=1}^{n} VHT_i(t) \tag{3.2}$$

on which VMT_i and VHT_i are calculated using the following equations:

$$VMT_i(t) = Q_i(t)l_i \tag{3.3}$$

$$VHT_i = \frac{Q_i(t)l_i}{V_i(t)} \tag{3.4}$$

and are measured at five-minute intervals. Here, l_i is the length of segment *i* between detectors *i*+1 and *i*-1 at location x_{i+1}; x_{i-1} and correspondingly equates to

$$l_i = \frac{1}{2}(x_{i+1} - x_{i-1}) \tag{3.5}$$

In the case of a double loop detector, the speed is calculated by the distance between two loops (*d*) divided by the time difference (Δ), when the front edge of the vehicle passes the edges of the two successive loops, as measured by the equation below:

$$Speed = \frac{d}{\Delta} \tag{3.6}$$

However, for the other types of detectors, the average speed would be measured using the following equation:

$$\hat{V}(t) = \frac{Q(t)}{K(t)} L(t) \tag{3.7}$$

where, *Q(t)* represents the average five-minute period of volume and *K(t)* denotes the average five-minute period of occupancy that is derived from the 30 seconds of raw data collected by the detectors. *L(t)* is the average vehicle length of period *t* at the location, which is usually continuous (e.g., 20 feet) but can be intermittent depending on the location and the time. When a vehicle's length changes through time and location, another PeMS algorithm is used to measure this varying element (Chen, 2003).

In this project, the online data from UDOT PeMS was analyzed to determine possible errors that may occur within the database. The preliminary data analysis showed four different types of errors in the one-month period of data collection on the I-15 and I-80 freeways: missing data, large variations, out-of-range data, and data inconsistencies. We analyzed speed, flow, and occupancy to explore any potential errors.

3.2.1 Missing Data

Based on the preliminary data screening, it has been shown that many records examined were missing one or more of the vehicle parameters for flow, occupancy, or speed. Some records had unavailable data for all three variables. For example, in one case, the corresponding flow and occupancy were quite high; however, no speed variable was shown for that particular station. Temporarily missing data may be the result of insufficient information due to low flow rates during the collection time. However, if no records were associated with a specific detector for an extended period of time, the detector was considered to be faulty. This situation may be related to detector breakdown such as a problem with its wiring.

3.2.2 Large Variations Within the Data

Another type of potential error may occur when neighboring stations show a considerable variation in recorded variables over consecutive periods. For instance, there might be two adjacent detectors that show more variations in traffic flows than can be accounted for by changes in the road segment between them. Such variations may be due to malfunctions in one or both of the detectors. Notably, data variations between adjacent lanes along the same direction may not be an indicator of detector malfunction due to the fact that various types of vehicles (i.e., passenger cars, trucks, buses, etc.) are not evenly distributed among lanes.

3.2.3 Out-of-Range Data

The Highway Capacity Manual (HCM) specifies maximum values for flow, speed, and occupancy depending on road characteristics (Highway Research Board, 2000). Each variation has a maximum level and the values shown by detectors are usually lower than the maximum limit. Hence, each variation must be higher than zero and less than the maximum level depending on the geometric conditions of the freeway. For example, the traffic flow on interstate freeways ranges from 2,000 to 2,200 vehicles per hour per lane, according to the HCM. Hence, the flow rate should range from 0 to 180 vehicles per lane in each time interval because study data are aggregated at 5-minute intervals in this project. However, some records were found to exceed these maximum values. Also, in some cases, the detectors showed a non-zero value for one variable and a zero value for others. Such cases may indicate failures in detector performance.

3.2.4 Inconsistencies

PeMS data are in the aggregated raw form, which takes the occurrence data (individual vehicle records) and aggregates them by a 20-second interval. However, some detectors showed inconsistencies between the PeMS data and the occurrence data. These inconsistencies may originate from aggregation errors within the PeMS database. Moreover, in other cases, data inconsistencies can be observed by comparing the data from adjacent detectors. For example, if there is no ramp between two adjacent detectors, flows and speeds produced by them should be close to each other. Significant differences in flows and speeds could indicate potential detection errors in one or both of them.

To identify the four types of data errors mentioned above, this research aims to leverage the characteristics of traffic flow theory to develop a multi-stage data screening algorithm. The data screening process can also help identify potential malfunctioning detectors when any of them are found to produce erroneous data constantly. The next subsection will introduce the details of the developed algorithm.

3.3 Multi-Stage Data Screening Algorithm

“Turbulence” in a traffic stream means that traffic flow patterns are not always consistent and might undergo changes over time (van Beinum et al., 2018a). Turbulence affects flow behaviors and traffic characteristics. This phenomenon occurs most commonly on highway ramps and when vehicles are unevenly distributed among lanes [e.g., the majority of heavy vehicles are in the right-hand lane] (Kwon et al., 2003). According to the AEVL definition that is related to flow rate, occupancy, and speed, these scenarios create an exception in previous data screening algorithms since the inconsistencies in traffic patterns cause the spatial comparisons among data records to be imprecise. To account for such considerations, the screening algorithm developed in this research project was tested in four major stages. The primary screening stage was a check for missing data as well as single and multiple variable thresholds. Then, the data quality was evaluated using the AEVL distribution and the corresponding statistical analysis method. Different scenarios with various types of detector stations were studied. Multiple new approaches are presented in this project report.

3.3.1 Stage 1: Primary Screening

At this stage, primary monitoring was applied to check for any apparent faults in the databases. The first step was to examine all stations for any missing variables. In addition, the single variable threshold check according to basic traffic concepts was conducted, which determined whether or not all variables were within the meaningful ranges. These ranges are related to the geometric characteristics of the road segment (Lu et al., 2014):

- Flow (q_i):

$$0 \le q_i \le q_{\max} \tag{3.8}$$

where q_i is the flow rate according to detector i in *vehs/lane/hr*, and the q_{max} is the maximum flow rate on the road. Notably, q_{max} is directly related to roadway conditions and differs based on the number of lanes and aggregated data from the station.

- Speed (v_i):

$$0 \le v_i \le v_{\max} \tag{3.9}$$

where v_i is the speed according to detector i in *miles/hr* and the v_{max} is the maximum speed limit based on road characteristics and station type.

- Occupancy (o_i):

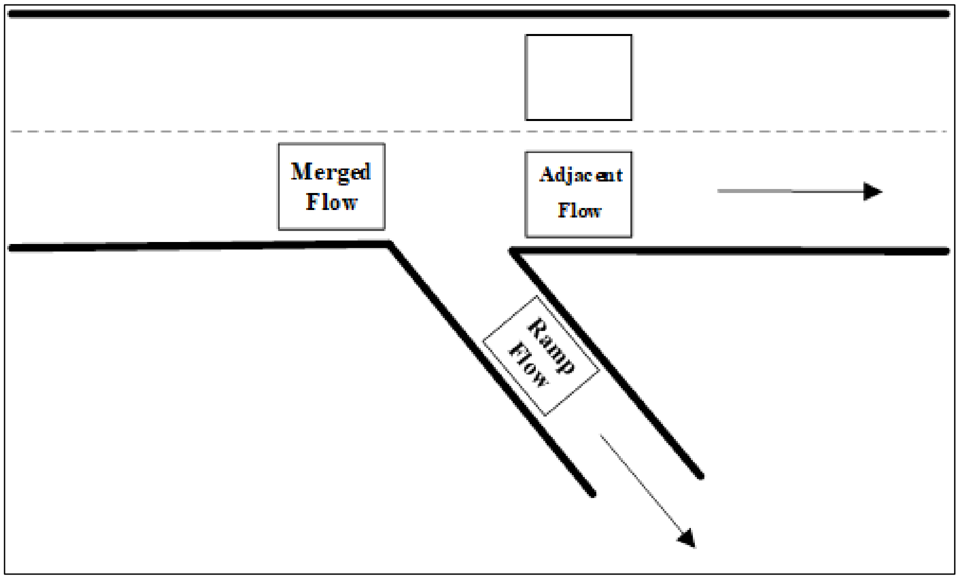
$$0 \le o_i \le o_{\max} \tag{3.10}$$

where o_i is the percentage of occupancy according to detector i and the o_{max} is the maximum occupancy rate. The effectiveness of the measurements should be tested according to the relationship among flow, speed, and occupancy (i.e., there should not be a record with zero value for a single variable and non-zero values for other variables). If any of the recorded data is missing or does not fall into the provided ranges, the corresponding detector is marked as faulty.

3.3.2 Stage 2: Piecewise Quality Check

At this stage, the AEVL distribution and Kolmogorov–Smirnov (K-S) test (Massey, 1951) were used to evaluate detector performance. This comparison was performed between two successive stations to check the consistency of data within partial road segments. More specifically, the AEVL distribution of a target detector was compared to its upstream or downstream detectors' values using the K-S test. Notably, this stage doesn’t compare different types of vehicles within the same station because typically vehicles are not evenly distributed among lanes. For example, we can often observe concentrations of heavy vehicles in the right-hand lanes, an accident in a specific lane, or a traffic barrier placed along the road in practice (Coifman, 2009).

Typically, there are several types of detection stations in the state of Utah: mainline, on-ramp, off-ramp, high-occupancy vehicle (HOV), and freeway-to-freeway. Ramp stations are placed in areas of heavy traffic stream changes, including the downstream and upstream stations [according to ramp type] (van Beinum et al., 2018b). As a result, these stations' traffic characteristics cannot be compared to those of adjacent lanes. Nevertheless, based on traffic flow conservation, the traffic characteristics of a ramp and those of its adjacent lane will be similar to nearby stations since these two traffic flow patterns should be similar to nearby station traffic flow. Figure 3.1 and Figure 3.2 show a schematic of ramp traffic flow for on-ramp and off-ramp stations compared to other stations nearby.



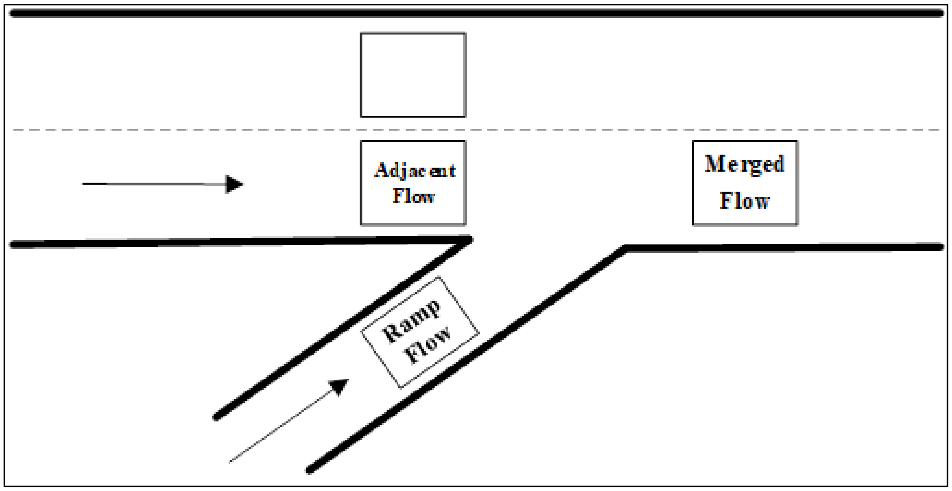


Figure 3.1 Traffic flow of on-ramp station vs. merged flow

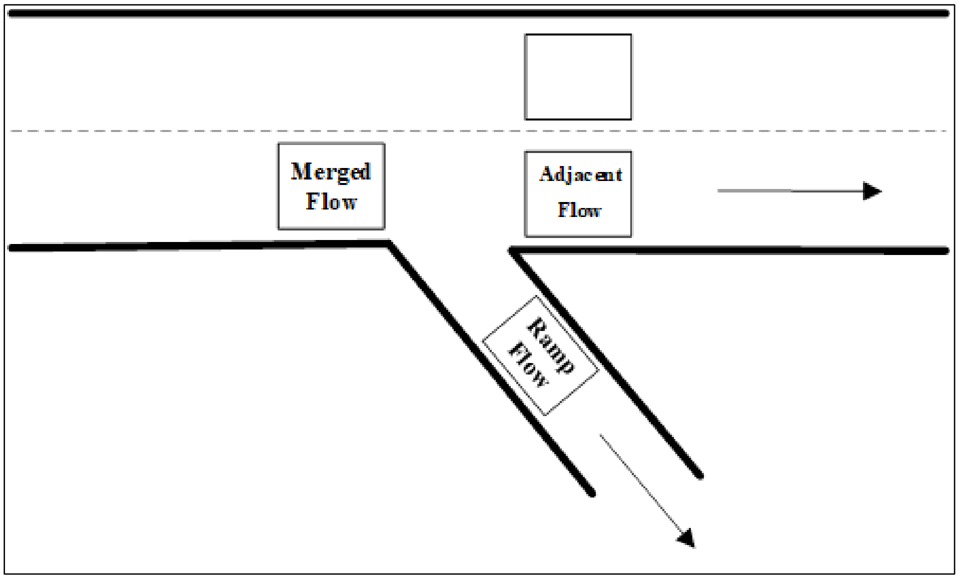


Figure 3.2 Traffic flow of off-ramp station vs. merged flow

As discussed in the literature, the AEVL is described as follows:

$$AEVL_i \approx \frac{3o_i v_i}{q_i} \tag{3.11}$$

where $AEVL_i$ is the average effective vehicle length of station i in *feet* (Jia et al., 2001). After measuring the AEVL distribution over the specified length of time (e.g., one hour), the distribution of each lane was compared to the distribution of its corresponding downstream or upstream stations, as shown in Figure 3.3. Using the K–S test, the estimated AEVL distribution of the target detector during the time granularity of recorded data should not be statistically different from the neighboring stations' AEVL distribution. Otherwise, further tests would be needed to evaluate the efficacy of the target detector.

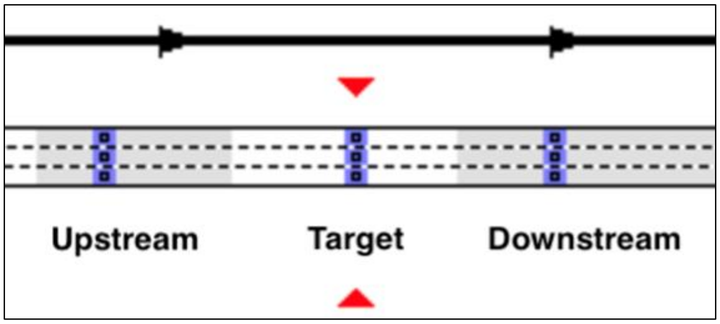


Figure 3.3 Detector map for spatial comparison

The purpose of the K-S test is to compare the sample data to a reference distribution and evaluate whether these two distributions are similar. According to the null hypothesis of the K-S test, the sample distribution belongs to the reference distribution if the maximum distances of two distributions are higher than the critical value. Then, the null hypothesis is rejected and shows that the sample is not part of the distribution (Massey, 1951), which indicates that there are potential errors in the recorded data. The critical value for the K-S test is as follows:

$$D_{m,n,0.05} = c(0.05)\sqrt{\frac{1}{m}+\frac{1}{n}} \tag{3.12}$$

where $D_{m,n,0.05}$ is the critical value for two distributions with a corresponding sample size of m and n, and $c(0.05)$ is the inverse of the Kolmogorov distribution table at 0.05 (Confidence Interval [CI]=95%), which depends on the sample size of the two distributions (Wolfe, 2012). The goal of this stage is to measure the exact changes in the data patterns and uncover any erroneous records.

3.3.3 Stage 3: Continuous Quality Check

To further test detector performance, the AEVL values of each lane at neighboring stations were compared using the MCB method. The ramp stations are exceptions at this stage due to the created weaving in traffic flows. The MCB compares the target lane to the inflow detectors and forms a CI for statistical assessment. If the CI is zero, this indicates no significant difference of AEVL at neighboring stations. The former K-S test method measures the data quality by comparing continuous road segment data to determine if the target detector produced erroneous records. In MCB, the inflow detectors include both the upstream and downstream data of the corresponding lane detector.

The MCB method requires detector data over a specific period of time (e.g., one week) to be grouped into equal segments (e.g., one hour), and the mean and variance of the group to be computed. The AEVL distribution for inflow stations can then be calculated and the CI is determined using the equation shown below:

$$CI = \begin{bmatrix} \min\left\{0, \left(AEVL_{i,ave} - \max(AEVL_{l \neq i,ave}) - \sqrt{2} \times T^{\alpha}_{k-1,k(n_i-1),0.5} \times \sqrt{\frac{S^2(n_i)}{n_i}} \right)\right\} \\ , \max\left\{0, \left(AEVL_{i,ave} - \max(AEVL_{l \neq i,ave}) + \sqrt{2} \times T^{\alpha}_{k-1,k(n_i-1),0.5} \times \sqrt{\frac{S^2(n_i)}{n_i}} \right)\right\} \end{bmatrix} \tag{3.13}$$

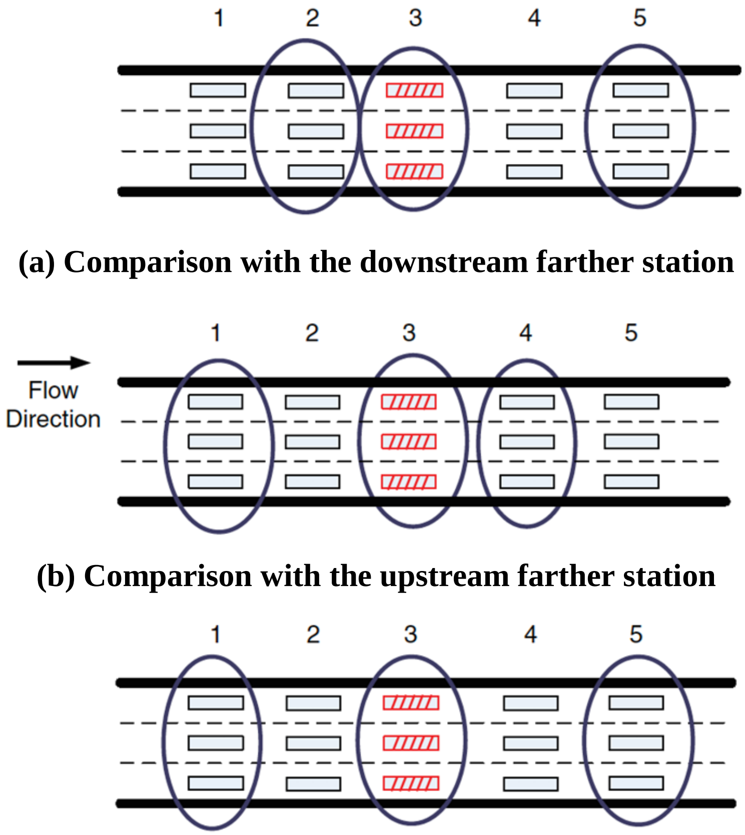
where i is the target detector, l represents the other inflow detectors, n_i is the sample size of detected data in the target lane, α symbolizes the significance level, $AEVL_{i,ave}$ shows the average *AEVL* of the target detector i of a specific time period, and S^2 represents the variance of group i. The value of parameter T depends on the sample size, the number of groups that are being compared (three in this case), and the significance level (Hochberg and Tamhane, 1987). Several scenarios were established for the detector status when the CI was calculated by using the MCB test and the results from previous steps. If the detectors show errors in both stages, there is a chance that there will be errors at all stations, which would require further investigation. More specifically, the following rules can be followed:

1. If detectors show no significant errors in both steps, there is a 95% CI that no other error is present.
2. If the target detector shows no errors during the continuous comparison test but does indicate some in the piecewise test, it can be concluded that there might be a change in traffic flow patterns such as a special event, or there might be a fixed error within the detector.

3. If the detector only fails the continuous comparison test, an additional investigation is needed due to two possible scenarios: a) the target detector may have an error, or b) the adjacent station detector may not be in proper working order.

3.3.4 Step 4: Further Investigation

When further investigation is needed, the AEVL of the target detector is compared to the neighboring stations (downstream and upstream). If the continuous comparison indicates significant data variations, the CI of that error will be at least 95%. Otherwise, its adjacent detector would be the faulty one. After a malfunctioning detector is found, it is not compared to other detectors during further performance evaluations. Instead, the nearest remaining detector is used for comparisons, depending on the location of the malfunctioning detector. Figure 3.4 shows the adjustment process when a malfunctioning detector is found.



(a) Comparison with the downstream farther station

(b) Comparison with the upstream farther station

(c) Comparison with both upstream and downstream farther stations

Figure 3.4 Detector comparisons with farther stations.

The developed data screening algorithm is presented in Figure 3.5 to illustrate the procedure presented in this study.

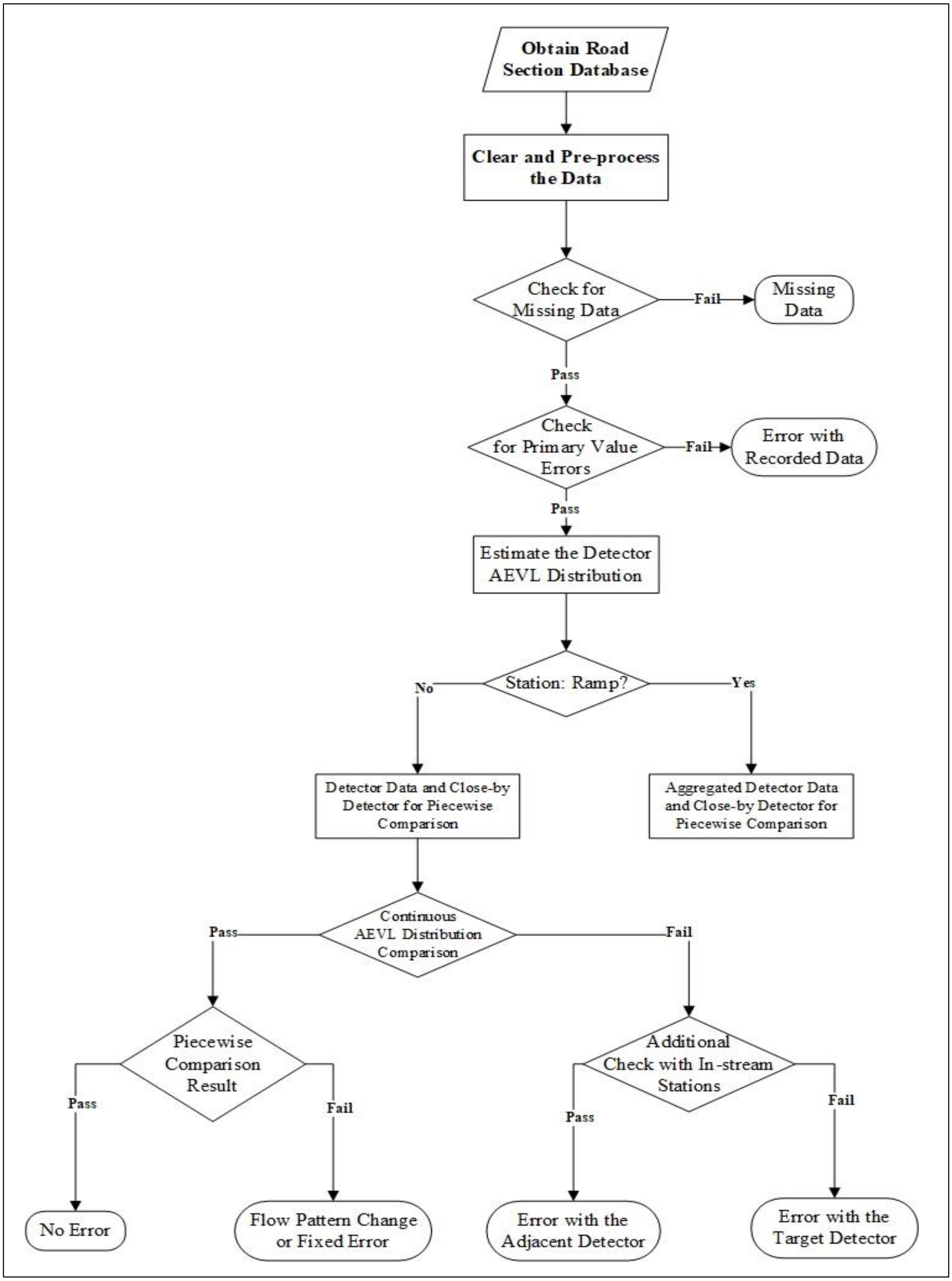

Obtain Road Section Database
Clear and Pre-process the Data
Check for Missing Data
Fail
Missing Data
Pass
Check for Primary Value Errors
Fail
Error with Recorded Data
Pass
Estimate the Detector AEVL Distribution
Station: Ramp?
No
Yes
Detector Data and Close-by Detector for Piecewise Comparison
Aggregated Detector Data and Close-by Detector for Piecewise Comparison
Continuous AEVL Distribution Comparison
Pass
Fail
Piecewise Comparison Result
Pass
Fail
Additional Check with In-stream Stations
Pass
Fail
No Error
Flow Pattern Change or Fixed Error
Error with the Adjacent Detector
Error with the Target Detector

Figure 3.5 The data screening algorithm

4.0 APPLICATION OF THE DATA SCREENING ALGORITHM

4.1 Overview

This chapter focuses on a case study in which the developed data screening algorithm was implemented. Specifically, the error identification tool was applied to a section of freeway to uncover any malfunctioning detectors. This chapter includes a brief summary of the case study and a discussion of the thresholds used in the algorithm. Then, some erroneous data were discovered by applying all stages of the algorithm.

4.2 Error-Identification Case Study

Algorithm performance was tested by conducting a case study on I-15. Data were collected from January 12-19, 2019 in the northbound direction between mileposts 302.75 and 312.16.

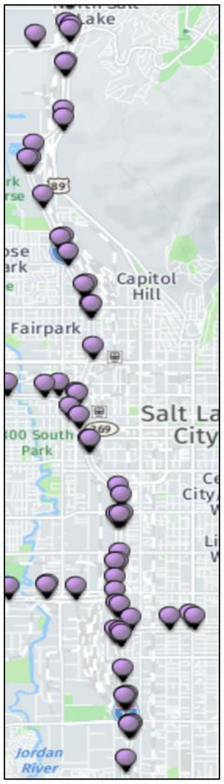


Figure 4.1 I-15 northbound segment for case study (302.75-312.16)

After the raw data were collected from the PeMS database, they were cleared and pre-processed for application of the screening algorithm. Table 4.1 shows a summary of the data that were used in the case study.

Table 4.1 Summary of data used for the case study

Detectors	Count	Percentage	Data Point	Percentage
Mainlines	20	32.79%	35,277	34.92%
HOV	19	31.15%	33,480	33.14%
Ramps	22	36.07%	32,277	31.95%
Total	61	100.00%	101,034	100%

The developed data screening algorithm was first used to seek out the problematic data. Then, the single variable thresholds were tested. Those thresholds are shown in Table 4.2.

Table 4.2 Single variable thresholds

Variable	Threshold
Flow	180 (Vehs/5mins/lane)
Speed	90 (mph)
Occupancy	95 (%)

4.3 Error-Identification Results

The algorithm identified some potential errors in the obtained dataset at various stages. These potential errors are shown in Table 4.3. The highest percentage of errors occurred in the missing data section, which may be due to problematic data transferring from the stations to the database. Also, the distribution of errors appeared to differ at each station location.

Table 4.3 Distribution of data errors based on station location and algorithm stage

Station Type	Ramps		HOV		Mainline		Summation	
	Total	Percent	Total	Percent	Total	Percent	Total	Percent
No Error	26,521	82.17%	27,470	82.05%	32,299	91.56%	86,290	85.41%
Missing Data	5,521	17.11%	4,368	13.05%	679	1.92%	10,568	10.46%
Primary Error	235	0.73%	420	1.25%	754	2.14%	1,409	1.39%
Piecewise Comparison	79	0.24%	1,448	4.32%	2,128	6.03%	3,655	3.62%
Continuous Comparison	-	-	1,222	3.65%	1,545	4.38%	2,767	2.74%

*Note: the total of each percentage column doesn't have to be 100% as some errors can be detected several times at different stages.

Mainline station errors can be discovered by comparing detector data (i.e., with piecewise and continuous comparisons) at different stages of the data screening process. By contrast, errors at other station types are often identified at earlier stages. The data quality distribution is shown in Figure 4.2.

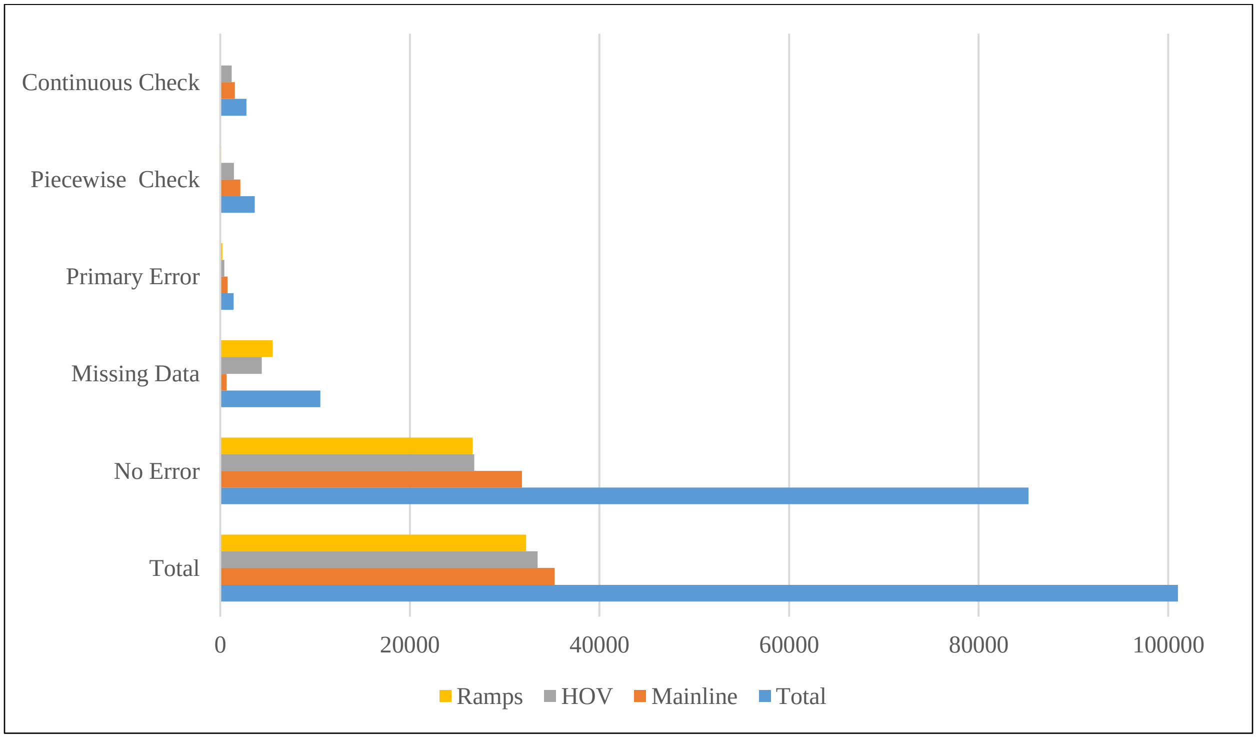


Figure 4.2 Distribution of data points after algorithm implementation

Analysis of the error distribution at various station locations can give decision makers a better sense of the type of errors that are likely to occur at a particular type of station. The error distribution based on the station type is presented in Figure 4.3.

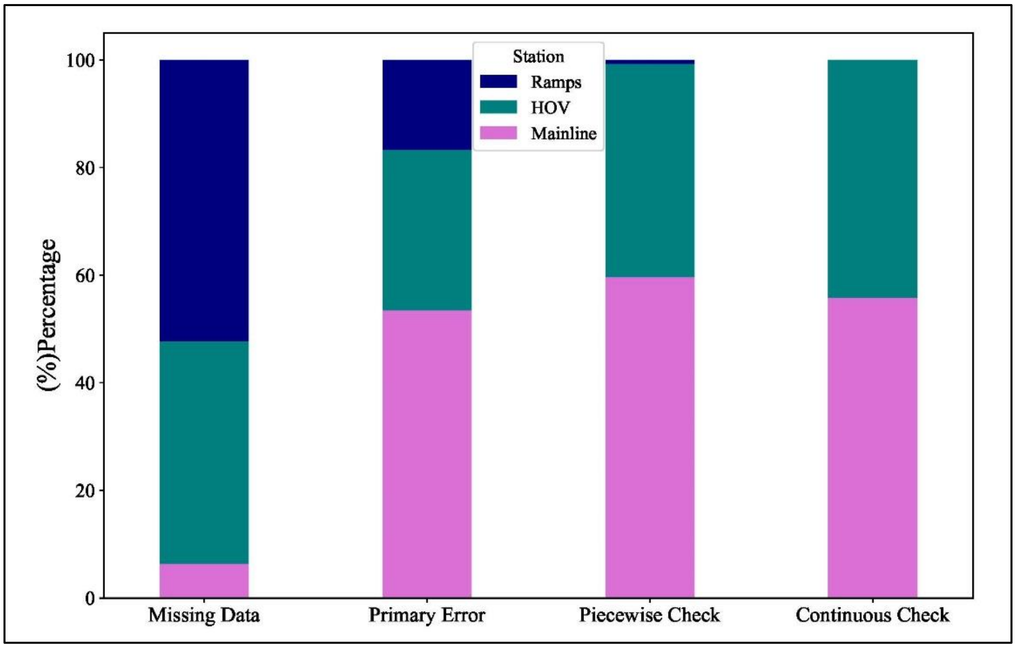


Figure 4.3 Stacked distribution of error detection stage based on station type

As shown in Figure 4.3, missing data is most likely to occur at ramp stations and least likely at mainline stations. The reason is that mainline stations are more regularly maintained and thus less likely to yield obvious errors such as missing data. More in-depth study is needed to identify malfunctioning detectors that yield inaccurate measurements.

Figure 4.4 shows that more errors were found at ramps and HOV stations than at mainline stations. This may be the result of a lack of detector maintenance.

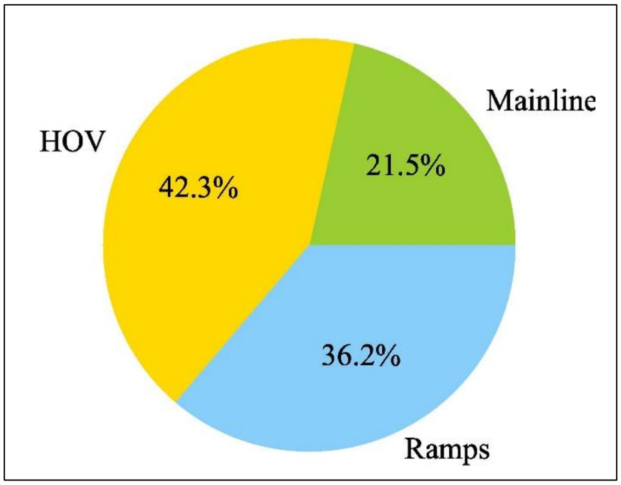


Figure 4.4 Pie chart of errors by detector location

The percentages of errors detected at various stages of the developed algorithm are summarized in Table 4.4 based on station type. The highest percentage of ramp station errors was due to missing data. Also, errors related to flow distribution (stages 2 and 3) accounted for the highest percentage of mainline station errors, which can indicate the high level of traffic fluctuations around these detectors. However, it must be noted that when these errors are observed less than 10% of the time, a given station should not be deemed as problematic.

Table 4.4 Detector error summary based on station type

Test Stage	**Station Type**			**Total**
	Mainline (%)	**HOV (%)**	**Ramps (%)**	
Missing Data	16.84	63.27	96.50	**63.44**
Primary Error	18.70	6.08	4.11	**8.46**
Piecewise Check	52.76	20.97	1.38	**21.94**
Continuous Check	38.31	17.70	0.00	**16.61**

Table 4.5 shows the distribution of errors based on flow levels. The missing data issues often happened at stations with lower flow exposures. As the flow level increased, the distribution progressively shifted to later stages of the algorithm process. This means that errors have been chiefly found in stages that look for consistency of flow distribution in more congested situations. Moreover, in free-flow conditions, errors are marked instead in primary stages.

Table 4.5 Detector error summary based on the traffic flow rate

Test Stage	Flow Level (veh/5mins)			
	<50 (%)	**50 – 100 (%)**	**100 – 200 (%)**	**>200 (%)**
Missing Data	65.35	52.29	28.15	0.22
Primary Error	5.48	3.81	2.05	29.79
Piecewise Check	13.09	28.84	55.28	54.47
Continuous Check	8.16	23.06	38.71	50.76

*Note: the total of each column doesn't have to be 100% as some errors can be detected several times at different stages.

The distribution displayed in Figure 4.5 indicates the flow level variation for each error type.

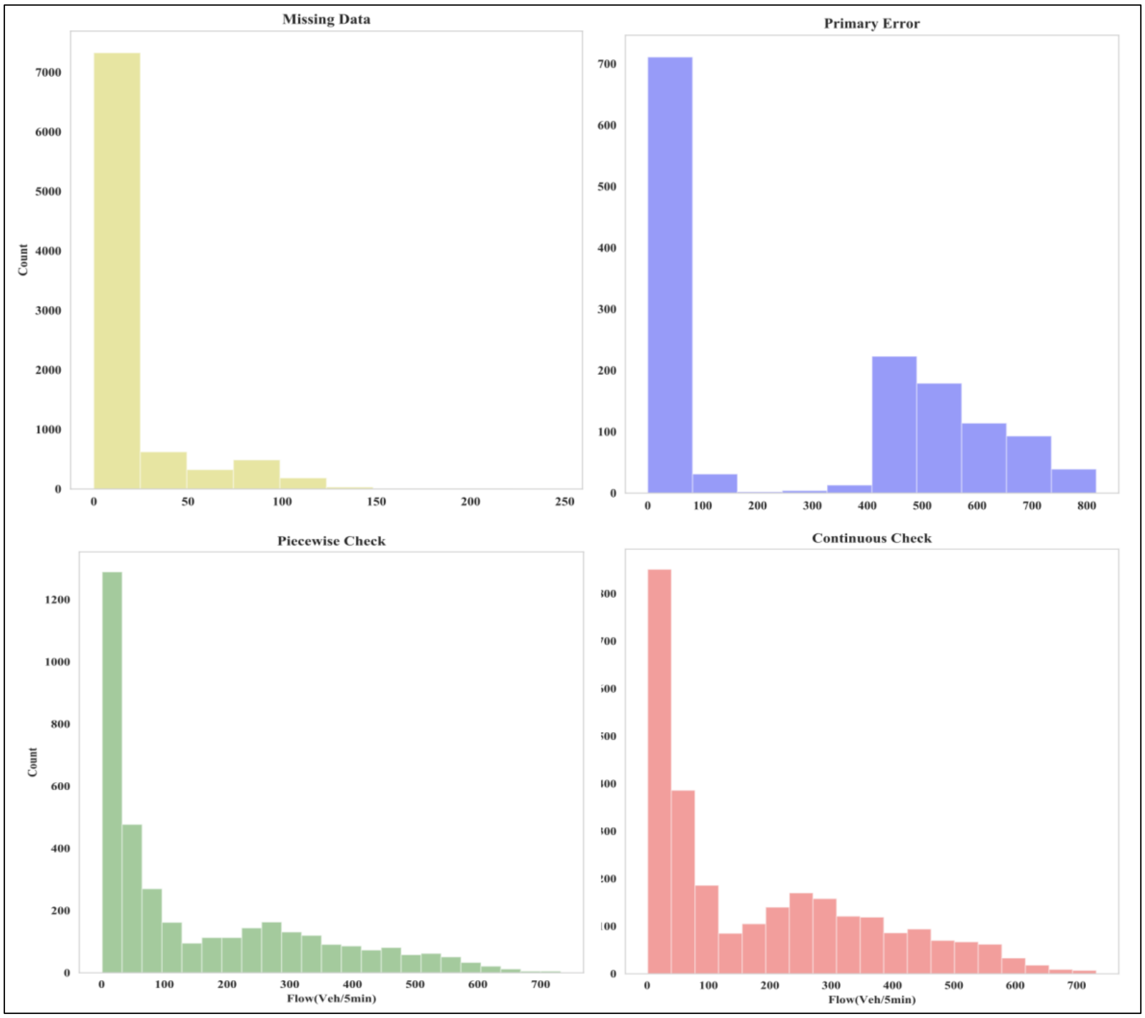


Figure 4.5 Distribution of flow level for various errors

After implementation of the developed algorithm, the next step was to identify the detectors that frequently produce inaccurate measurements. Detector stations with more than 50% inaccurate data were denoted as problematic, as shown in Table 4.6. Most malfunctioning detectors were either on ramps or in HOV lanes. The average rate of erroneous data points observed at mainline stations over the one-week study period was only 8%, which was satisfactory.

Table 4.6 Malfunctioning stations on I-15 northbound

ID	Milepost	Lanes	Type
1418	303.63	2	On Ramp
3422	304.53	2	Fwy-Fwy
1431	306.51	2	On Ramp
91431	306.51	1	Off Ramp

Our examination showed that all four of the stations mentioned in Table 4.6 experienced missing data issues more than 50% of the time. However, at station 3422, it only occurred in one lane. Moreover, although stations 1431 and 91431 are at the same location, data were only received from one detector. The specific locations of each malfunctioning station are shown in Figure 4.6 through Figure 4.8.

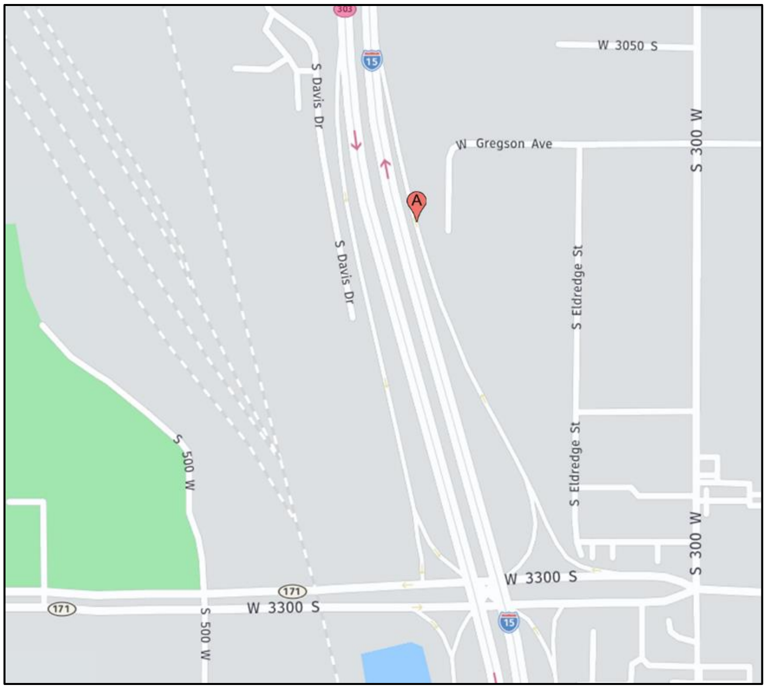


Figure 4.6 Location of malfunctioning on-ramp station 1418

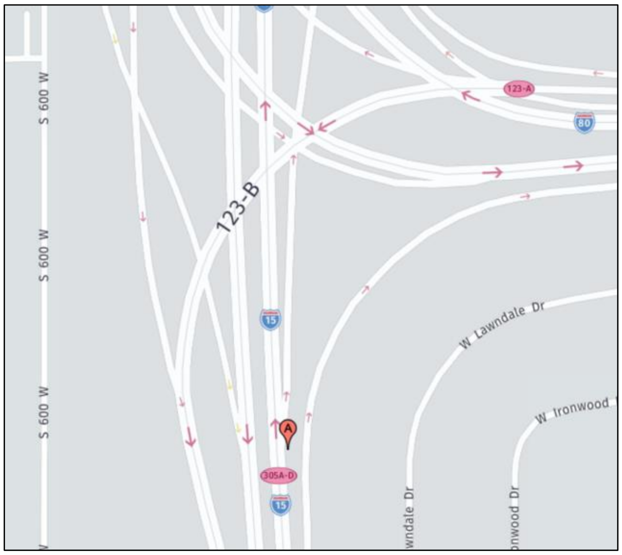


Figure 4.7 Location of malfunctioning fwy-fwy station 3422

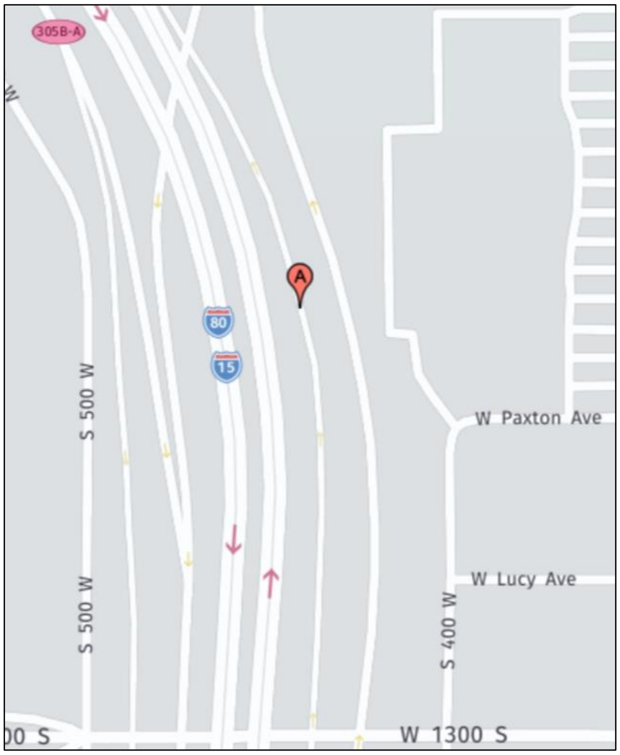


Figure 4.8 Location of malfunctioning on-ramp station 1431 and off-ramp station 91431

The rates of error types at each station location are shown in Figure 4.9. As stated earlier, the stations with a rate higher than 0.50 are defined as malfunctioning.

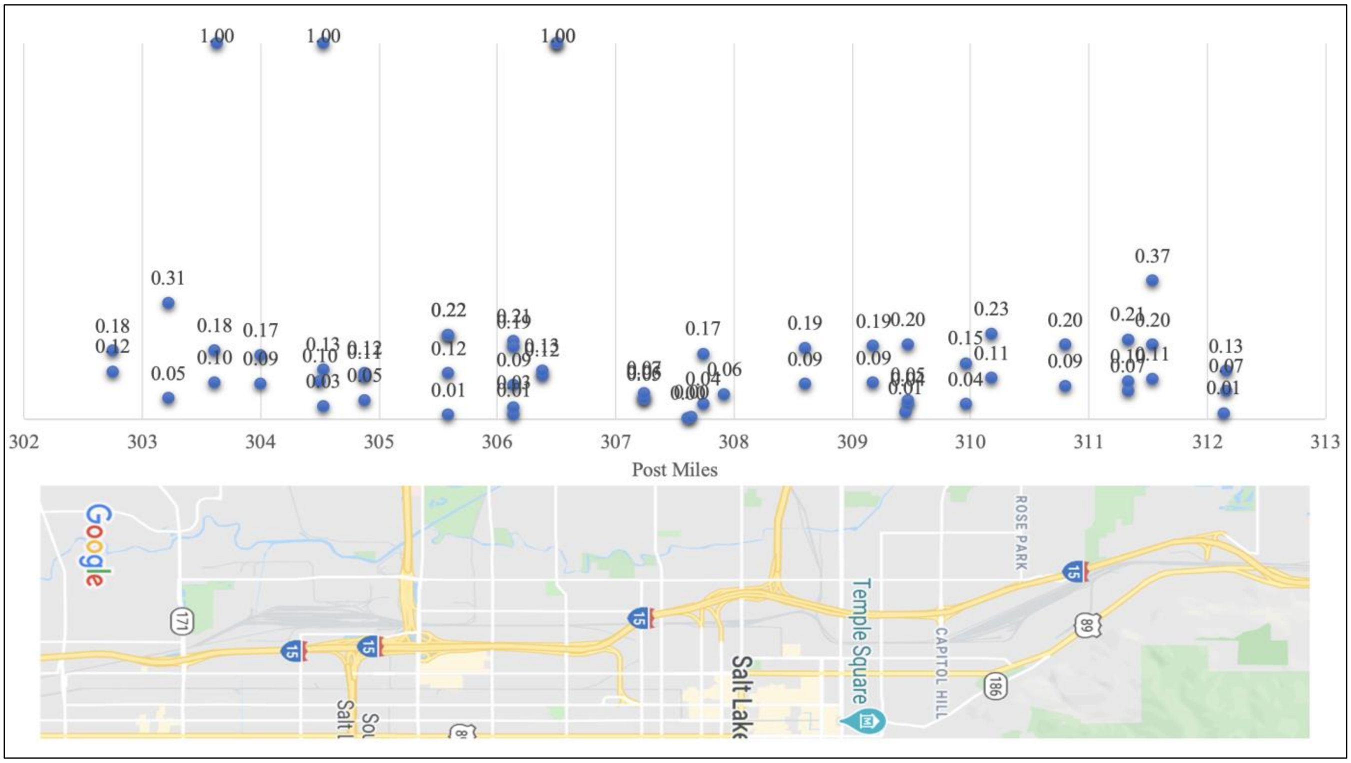


Figure 4.9 Error rate of stations by locations

5.0 IDENTIFICATION OF HIGH-SPEED SPOTS

5.1 Overview

The second objective of this project is to identify segments where traffic tends to exceed the speed limit by analyzing cleaned data. To accomplish that objective, we applied multiple evaluation criteria to the cleaned data.

5.2 Data Description

A segment of the I-80 freeway between mileposts 128.0 and 141.0 was chosen for this analysis of cleaned data. Data were recorded at 21 controllers in each direction (42 total detectors) from October 15th to 31st, 2019, along both directions. The geographic locations of these detectors are shown in Figure 5.1.

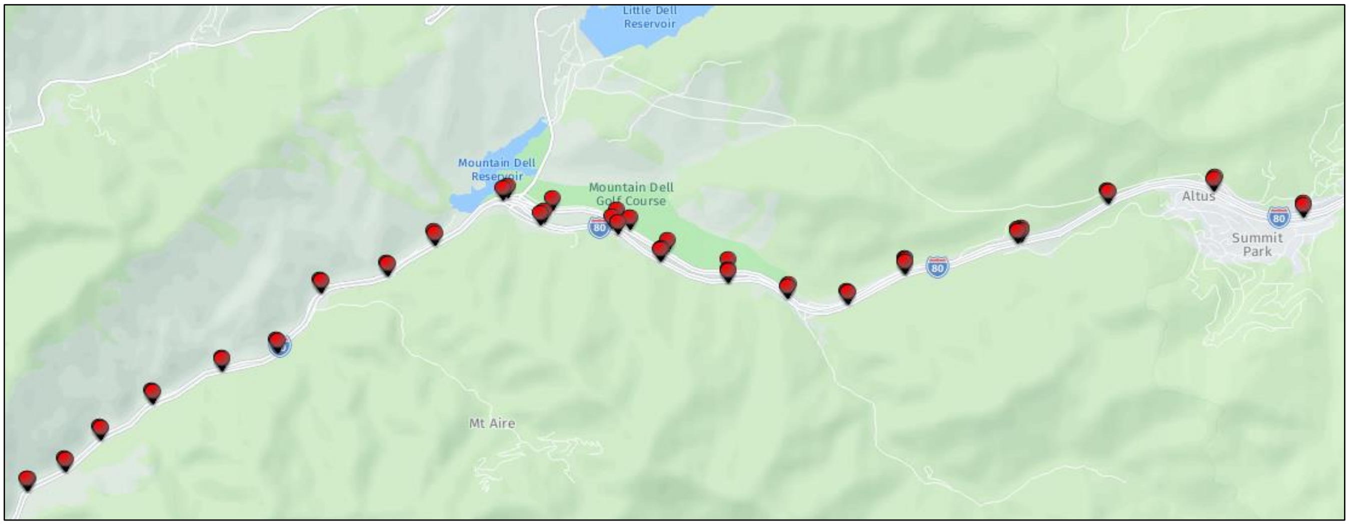


Figure 5.1 Geographic positioning of detectors on the I-80 corridor

Since some stations had insufficient data and were ignored in this study, the counts of useful records are shown in Table 5.1 for each direction of the corridor.

Table 5.1 Counts of records obtained from detectors utilizing each approach

Direction	# Detectors	Data Points	Data Time Series
Eastbound	18	5,028,558	Oct 15-31, 2019
Westbound	18	4,901,388	Oct 15-31, 2019

5.3 Using the Data Screening Algorithm to Validate Datasets

The obtained datasets included 5-minute average speeds and flow rates. Data points were evaluated using the screening algorithm and then the percentage of records with an error in the dataset was calculated. The location of each station contrasted with the ratio of erroneous data points is presented in Figure 5.2.

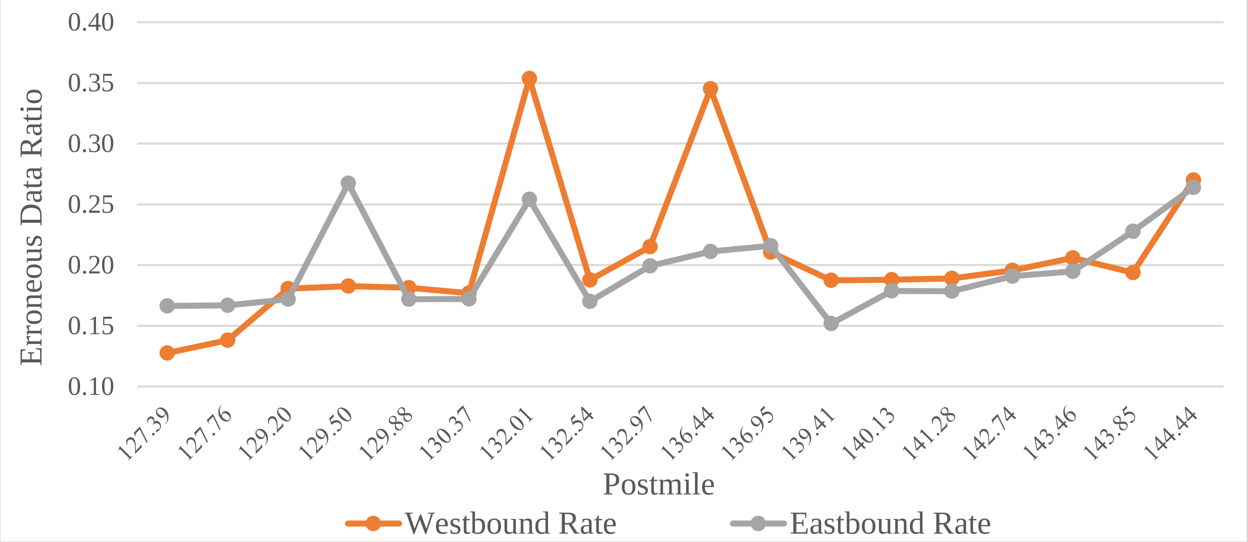


Figure 5.2 Ratio of erroneous records by station location

Some locations experienced an error rate greater than 20 percent. The speed analysis was only applied to stations with error rates that are smaller than 20%. A clear overview of the error types and their distributions is presented in Figure 5.3 and Figure 5.4. The majority of errors are related to missing data, which may be due to collecting an insufficient amount of information.

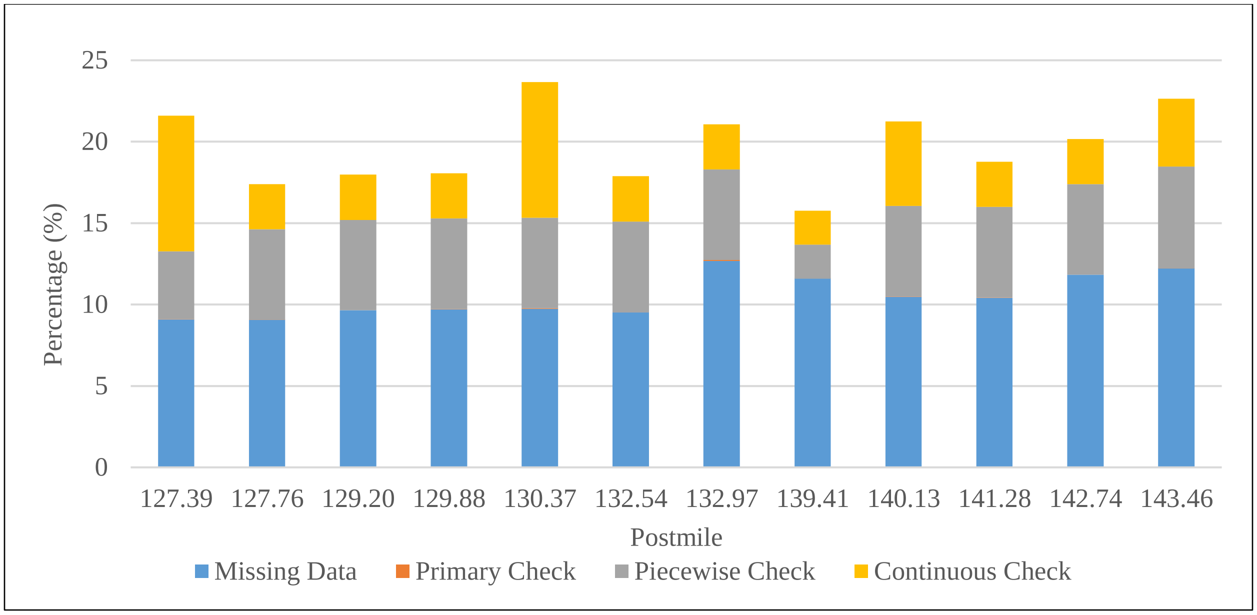


Figure 5.3 Error type distribution for eastbound

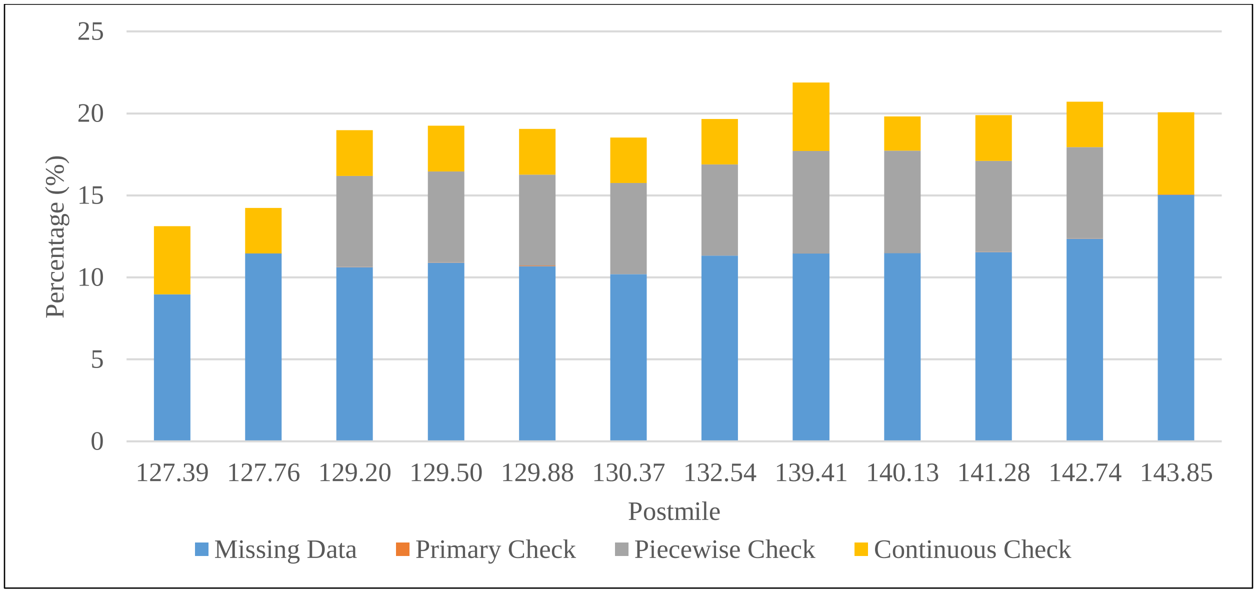


Figure 5.4 Error type distribution for westbound

5.4 **Speed Data Analysis and High-Speed Locations**

After validating the datasets, the next step was to analyze the speed profiles and identify high-speed locations. The 85th percentile of the average recorded speed was compared to the speed limit of 65 mph, as shown in Figure 5.5 and Figure 5.6. At all stations, the 85th percentile speeds were higher than the speed limit, which proves that people tend to drive faster than the speed limit. Moreover, as many existing studies (Abdel-Aty et al., 2006) indicated that the crash severity will be increased significantly if vehicles' speeds are 10 mph higher than the speed limit, this study also selects a second threshold of 75 mph, which is also shown in Figure 5.5 and Figure 5.6. Notably, the 85th percentile speeds at most stations were higher than the second threshold, which highlights the potential need of adding other speed enforcement countermeasures at those locations.

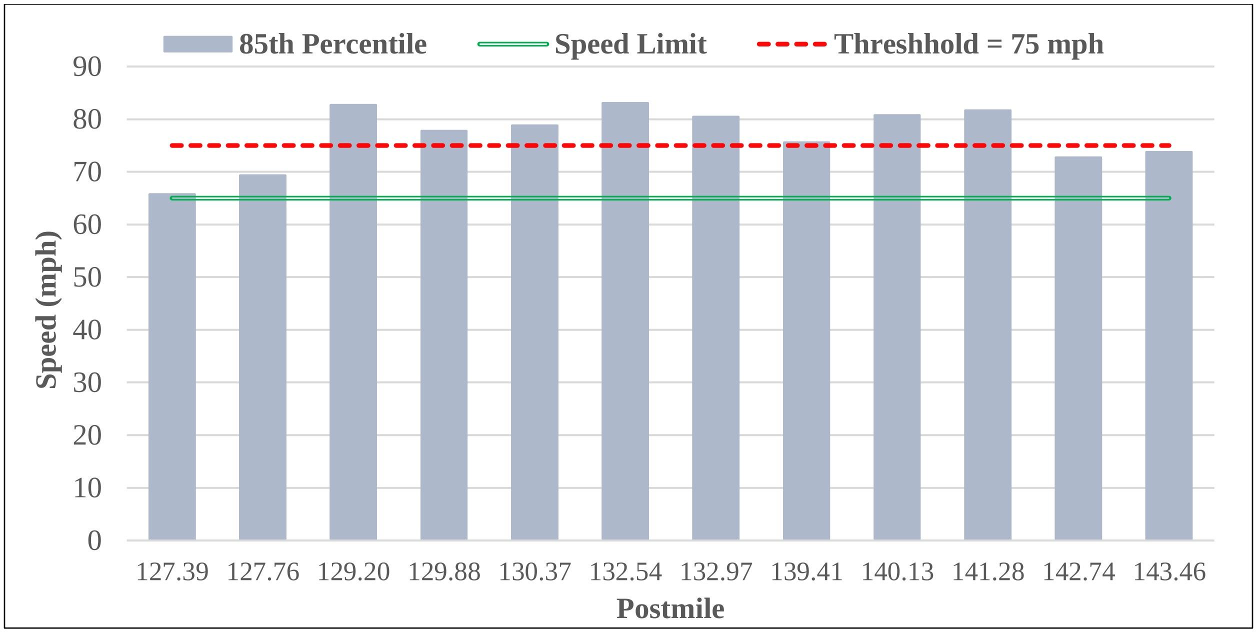


Figure 5.5 Spatial distribution of 85th percentile speed for eastbound stations

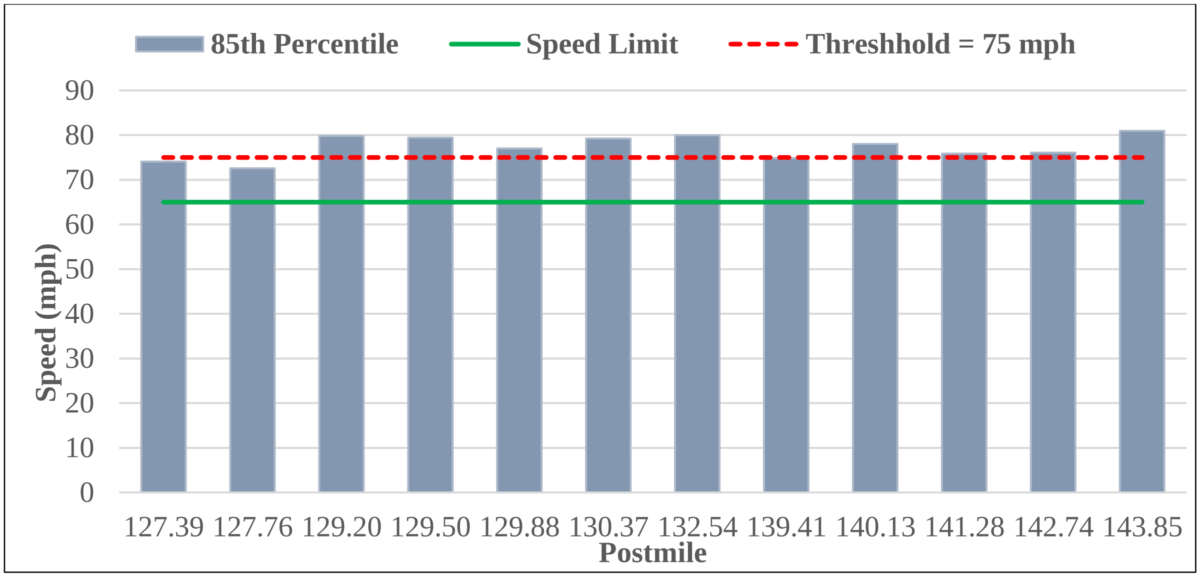


Figure 5.6 Spatial distribution of the 85th percentile speed for westbound stations

For a more detailed analysis at each station, we grouped data into four-hour intervals and attempted to calculate the maximum speed at each interval. Figure 5.7 and Figure 5.8 show these values throughout the timeframe. Eastbound stations 100618 and 100619 and westbound stations 100581 and 100169 were found to have the highest values at each interval (in excess of 120 mph), which make them good candidates for targeted speed enforcement.

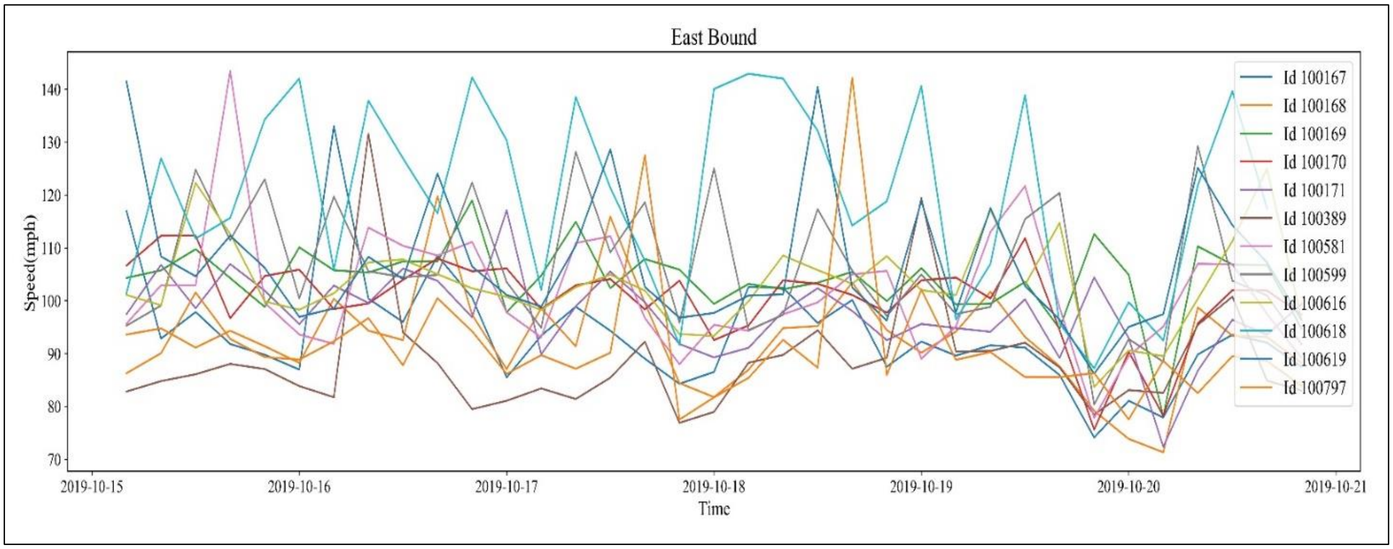


Figure 5.7 Maximum speeds at eastbound stations

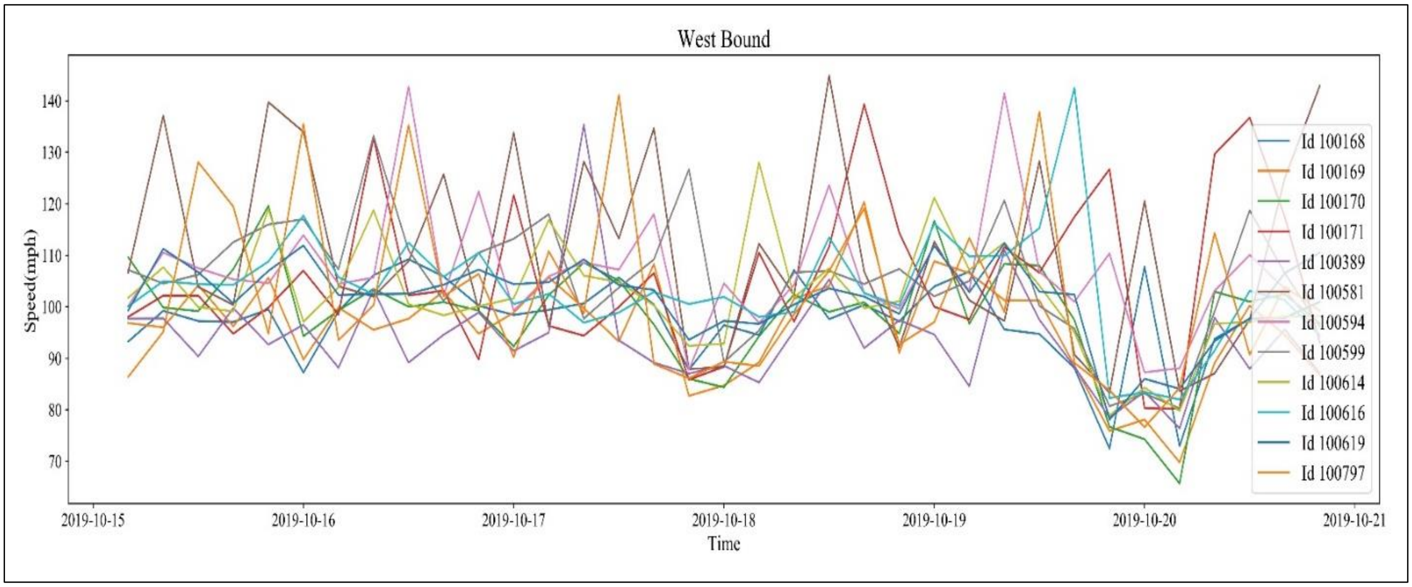


Figure 5.8 Maximum speeds at westbound stations

In the next step, we took a more in-depth look into data points and resampled the data for a one-hour interval. One-hour samples consist of the maximum speed records in one hour of collected data. Next, the ratio of one-hour samples greater than 85 mph over the whole period of collected data (resampled to one hour) are calculated. This ratio is presented for both directions of each station in Table 5.2 and Table 5.3. As seen below, some stations had at least one record higher than 85 mph in each one-hour interval for more than 90% of the range of collected data which is an indicator of high-speed hot spots in the corridor.

Table 5.2 One-hour maximum higher than 85 mph records for eastbound stations

Postmile	Internal ID	East Percentage
127.39	100389	23.44
127.76	100797	42.19
129.2	100619	97.40
129.88	100581	86.98
130.37	100599	92.19
132.54	100616	96.09
132.97	100618	92.71

139.41	100171	77.08
140.13	100170	89.06
141.28	100169	93.23
142.74	100168	69.01
143.46	100167	70.05

Table 5.3 One-hour maximum higher than 85 mph records for westbound stations

Postmile	Internal ID	West Percentage
127.39	100389	63.28
127.76	100797	63.54
129.2	100619	91.93
129.5	100594	91.41
129.88	100581	86.46
130.37	100599	89.84
132.54	100616	89.84
139.41	100171	79.95
140.13	100170	85.16
141.28	100169	76.82
142.74	100168	82.81
143.85	100614	94.01

The same comparison was made on the occurrence data for the speed limit and the threshold of 75 mph. The percentages of records over the speed limit and 75 mph are shown in Figure 5.9 and Figure 5.10. In the eastbound direction, the stations within mileposts 127.76 to 142.47 were shown to have higher percentages for both thresholds of 65 and 75 mph. For the westbound stations, over half the drivers were traveling at over 75 mph at all stations.

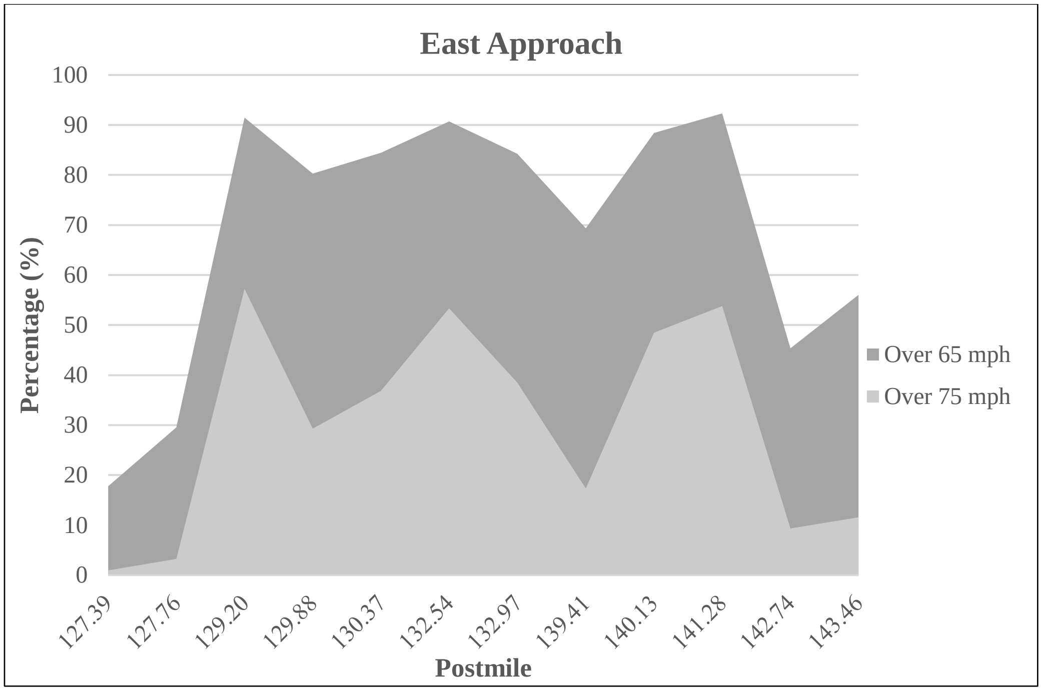


Figure 5.9 The percentage of records over the thresholds on the eastbound approach

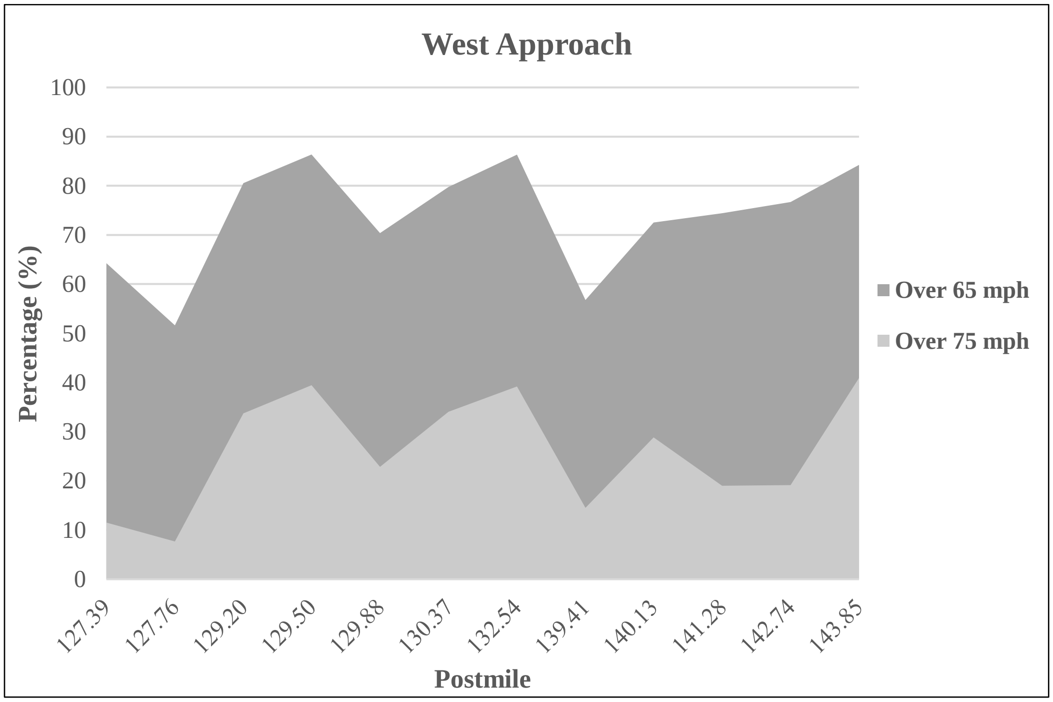


Figure 5.10 The percentage of records over the threshold on the westbound approach

Based on the speed analysis in this section, we concluded that stations located in the eastbound 129.2, 132.54, and 141.28 mileposts and the westbound 143.85 milepost are potentially high-speed zones. Each of these stations showed a large percentage of high-speed records according to the data analysis. Based on the records, the 85th percentile of average speed was higher than 80 mph, and the stations showed a record in the dataset that was higher than 100 mph every four hours. Furthermore, by examining the maximum individual speeds within hourly time intervals, these stations showed at least one record of 85 mph or higher more than 90 percent of the time. Finally, based on a speed comparison with the 65 and 75 mph thresholds, the ratio of records greater than 75 mph was above 0.5, which is another indicator of the presence of high-speed zones. The locations of those zones are shown in yellow in Figure 5.11.

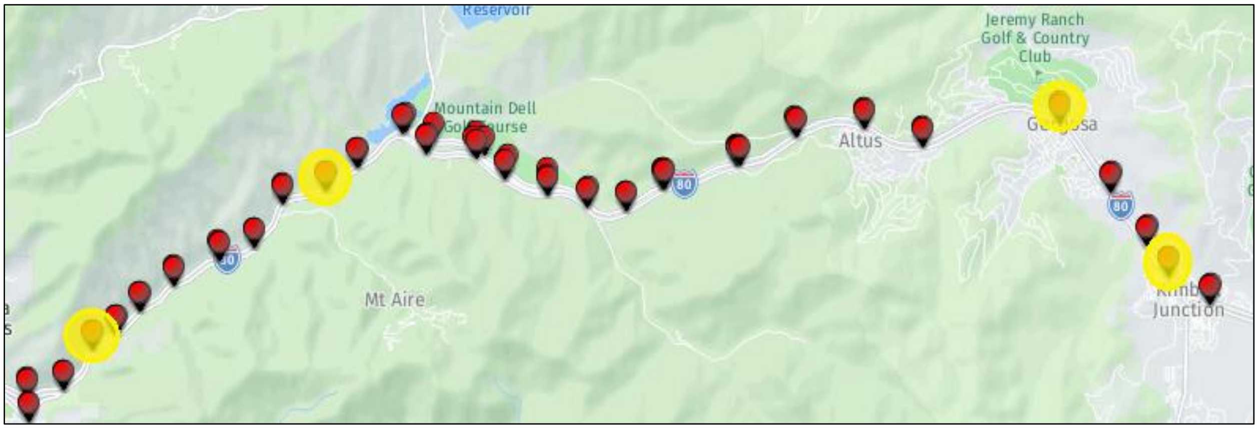


Figure 5.11 High speed locations at I-80 freeway

I-80 in the study area is generally uphill in the eastbound direction. However, the study station speed records show more spots in the eastbound direction with higher speed records. Even though the 85th percentile speed in the westbound direction is higher than the 75 mph threshold for most spots, they are limited to 80 mph. On the other hand, the eastbound sensors demonstrate spots with 85th percentile speeds higher than 80 mph as well as in excess of 75 mph at more than 50 percent of stations. The results show that despite the westbound direction being generally downhill, its speeds are more controlled, and the eastbound direction is experiencing spots with records further beyond the speed limit.

6.0 CONCLUSIONS

6.1 Summary

In this study, we developed a multi-stage data screening algorithm based on the types of potential errors in the PeMS database. Using the AEVL distribution as a key indicator, the data points were evaluated and statistically tested to identify inconsistency in traffic flow streams. The innovative aspect of this project was that we omitted the data comparison within stations due to uneven distribution of vehicle types among lanes. By applying the developed algorithm to the case studies, we were able to locate inaccurate records as well as potentially malfunctioning detectors.

6.2 Findings

The data screening algorithm mainly utilizes traffic flow regulation and flow conservations. Considering the variations within drivers' behaviors at particular locations, the use of this approach may result in inaccurate assessments due to the condition of the detectors. In the algorithm presented here, while keeping in mind that there will always be turbulence in traffic flow, the statistical tests were used to monitor data errors and faults. This technique uncovered any errors at stations and compared them to those at nearby stations. It also was used to assess road segment flow behaviors. Eliminating the in-station comparison will help UDOT engineers to make more reliable evaluations of the detector performance.

Furthermore, our results showed that the most common errors belong to the category of missing data, which implied that the detectors were not performing effectively for some periods or throughout the entire study period. Also, as shown in the case study, most detectors with inaccuracy issues are located at ramps. The reason might be that these detectors were placed on a lower priority list for maintenance compared to mainline detectors.

6.3 Limitations and Challenges

Some missing information in the PeMS database might be caused by data visualization functions of the website rather than problems with detectors. This can result in less precise evaluation results because the developed algorithm always classified them as missing data errors.

REFERENCES

Ackaah, W., Bogenberger, K., Bertini, R.L., Huber, G., 2016. Comparative Analysis of Real-time Traffic Information for Navigation and the Variable Speed Limit System. IFAC-PapersOnLine 49, 471–476. https://doi.org/10.1016/j.ifacol.2016.07.079

Bullock, D., Achillides, C., 2004. Performance Metrics for Freeway Sensors 125.

CalTrans, 2002. The Freeway Performance Measurement System (PeMS).

Chen, C., 2003. Freeway Performance Measurement System (PeMS), Institute of Transportation Studies, UC Berkeley, Institute of Transportation Studies, Research Reports, Working Papers, Proceedings.

Chen, C., Kwon, J., Rice, J., Skabardonis, A., 1976. Detecting Errors and Imputing Missing Data for Single-Loop Surveillance Systems 160–167.

Chen, Z., Qin, X., Schneider, E., Cheng, Y., Parker, S., Shaon, R.R., 2019. Designing a Comprehensive Procedure for Flagging Archived Traffic Data: A Case Study. Transportation Research Record 2673, 165–175. https://doi.org/10.1177/0361198119841286

Choe, T., Skabardonis, A., Varaiya, P., 2002. Freeway performance measurement system: Operational analysis tool. Transportation Research Record 67–75. https://doi.org/10.3141/1811-08

Cleghorn, D., Hall, F.L., Garbuio, D., 1991. Improved Data Screening Techniques for Freeway Traffic Management Systems. Transportation Research Board.

Coifman, B., 2009. Length based vehicle classification on freeways from single loop detectors.

Federal Highway Administration, 2012. Index - Methodologies to Measure and Quantify Transportation Management Center Benefits: Final Synthesis [WWW Document]. FHWA-HRT-12-054.URL https://www.fhwa.dot.gov/publications/research/operations/12054/007.cfm (accessed

9.2.20).

Hamad, K., 2015. QUALITY CONTROL OF ARCHIVED INTELLIGENT TRANSPORTATION SYSTEMS DATA. International Journal of Traffic and Transportation Engineering 5, 238–251. https://doi.org/10.7708/ijtte.2015.5(3).02

Highway Research Board, 2000. Highway Capacity Manual, Transportation Research Board, National Research Council, Washington, DC.

Hochberg, Y., Tamhane, A.C., 1987. Multiple comparison procedures. Wiley.

Horrace, W.C., Schmidt, P., 2000. Multiple comparisons with the best, with economic applications. Journal of Applied Economics 15, 1–26. https://doi.org/10.1002/(SICI)1099-1255(200001/02)15:1<1::AID-JAE551>3.0.CO;2-Y

Hsu, J.C., Nelson, B.L., 2003. Optimization over a finite number of system designs with one-stage sampling and multiple comparisons with the best 451–457. https://doi.org/10.1145/318123.318232

Huber, G., Bogenberger, K., Bertini, R., 2014. New Methods for Quality Assessment of Real Time Traffic Information.

Ishak, S., 2003. Fuzzy-Clustering Approach to Quantify Uncertainties of Freeway Detector Observations, in: Transportation Research Record. National Research Council, pp. 6–15. https://doi.org/10.3141/1856-02

J. Wells, T., J. Smaglik, E., Bullock, D., 2008. Health Monitoring Procedures for Freeway Traffic Sensors, Volume 1: Research Report, Joint Transportation Research Program.

Jia, Z., Chen, C., Coifman, B., Varaiya, P., 2001. The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors, in: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. pp. 536–541. https://doi.org/10.1109/itsc.2001.948715

Kwon, J., Varaiya, P., Skabardonis, A., 2003. Estimation of Truck Traffic Volume from Single Loop Detectors with Lane-to-Lane Speed Correlation. Transportation Research Record 106–117. https://doi.org/10.3141/1856-11

Lawrence A. Klein, Mills, M.K., Gibson, D.R.P., 2006. Traffic Detector Handbook: Third Edition—Volume I I, 288.

Lin, D.Y., Boyles, S., Valsaraj, V., Waller, S.T., 2012. Reliability assessment for traffic data. Journal of Chinese Institute of Engineers, A 35, 285–297. https://doi.org/10.1080/02533839.2012.655466

Lu, C., Dong, J., Sharma, A., Huang, T., Knickerbocker, S., 2018. Predicting Freeway Work Zone Capacity Distribution Based on Logistic Speed-Density Models. Journal of Advanced Transportation 2018, 1–15. https://doi.org/10.1155/2018/9614501

Lu, Y., Yang, X., Chang, G.-L., 2014. Algorithm for Detector-Error Screening on Basis of Temporal and Spatial Information. Transportation Research Record, Transportation Research Board 2443, 40–48. https://doi.org/10.3141/2443-05

Maghrour Zefreh, M., Török, Á., Mészáros, F., 2017. Average Vehicles Length in Two-lane Urban Roads: A Case Study in Budapest. Periodica Polytechnica Transportation Engineering 45, 218. https://doi.org/10.3311/PPtr.10744

Massey, F.J., 1951. The Kolmogorov-Smirnov Test for Goodness of Fit, Source: Journal of the American Statistical Association.

Megler, V.M., Tufte, K., Maier, D., 2016. Improving Data Quality in Intelligent Transportation Systems.

Park, H., Yoon, M., Kim, H., Oh, C., 2015. Development of a Novel Integrated Evaluation Index for Freeway Traffic Data. Journal of Korean Society of Transportation 33, 417–429. https://doi.org/10.7470/jkst.2015.33.4.417

Payne, Harold J., E. D. Helfenbein, and H.C.K., 1976. Development and testing of incident detection algorithms, volume 2: Research methodology and detailed results.

Randeniya, D., Kim, H.K., 2013. Estimation of its sensor operational states by analyzing measurements with errors using a Hidden Markov Model. KSCE Journal of Civil Engineering 17, 1740–1748. https://doi.org/10.1007/s12205-013-0284-2

Turochy, R.E., Smith, B.L., 2002. ALTERNATIVE APPROACHES TO CONDITION MONITORING IN FREEWAY MANAGEMENT SYSTEMS.

Turochy, R.E., Smith, B.L., 2000. New Procedure for Detector Data Screening in Traffic Management Systems. Transportation Research Record 1727, 127–131. https://doi.org/10.3141/1727-16

van Beinum, A., Hovenga, M., Knoop, V., Farah, H., Wegman, F., Hoogendoorn, S., 2018a. Macroscopic traffic flow changes around ramps. Transportmetrica A: Transport Science 14, 598–614. https://doi.org/10.1080/23249935.2017.1415997

van Beinum, A., Hovenga, M., Knoop, V., Farah, H., Wegman, F., Hoogendoorn, S., 2018b. Macroscopic traffic flow changes around ramps. Transportmetrica A: Transport Science 14, 598–614. https://doi.org/10.1080/23249935.2017.1415997

Vanajakshi, L., Rilett, L.R., 2004. Loop Detector Data Diagnostics Based on Conservation-of-Vehicles Principle. Transportation Research Record, Transportation Research Board 1870, 162–169. https://doi.org/10.3141/1870-21

Varaiya, P., 2004. Freeway Performance Measurement System (PeMS), Version 3 Phase II UCB-ITS-PR.

Wolfe, D.A., 2012. Nonparametrics : Statistical Methods Based on Ranks and Its Impact on the Field of Non parametric Statistics 1101–1110. https://doi.org/10.1007/978-1-4614-1412-4

Wu, L., Liu, C., Huang, T., Sharma, A., Sarkar, S., 2018. Traffic sensor health monitoring using spatiotemporal graphical modeling. International Journal of Prognostics and Health. Management 9.

Xiao, S., Liu, X.C., Wang, Y., 2015. Data-driven geospatial-enabled transportation platform for freeway performance analysis. IEEE Intelligent Transportation Systems Magazine 7, 10–21. https://doi.org/10.1109/MITS.2014.2388367

Yu, F., Zhijie, S., 2016. Methods of Real-Time Data Screening and Reconstruction for Dynamic Traffic Abnormal Data. Proceedings - 2015 6th International Conference on Intelligent Systems Design and Engineering Applications. ISDEA 2015 500–503. https://doi.org/10.1109/ISDEA.2015.130

Zhanfeng Jia, Chao Chen, Coifman, B., Varaiya, P., 2001. The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors, in: ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.01TH8585). IEEE, pp. 536–541. https://doi.org/10.1109/ITSC.2001.948715

Zhang, Z., Yang, X., Liu, C., Burns, K., Blackwelder, G., 2019. Data Screening Algorithm for Detecting Freeway Wrong Way Driving Hotspots, in: Transportation Research Board 98th Annual Meeting. Transportation Research Board 98th Annual Meeting.