



Article

A Methodology to Detect Traffic Data Anomalies in Automated Traffic Signal Performance Measures

Bangyu Wang¹, Grant G. Schultz^{1,*}, Gregory S. Macfarlane¹, Dennis L. Eggett² and Matthew C. Davis¹

¹ Department of Civil and Construction Engineering, Brigham Young University, 430 Engineering Building, Provo, UT 84602, USA; bwang3@byu.edu (B.W.); gregmacfarlane@byu.edu (G.S.M.); mcd597@byu.edu (M.C.D.)

² Department of Statistics, Brigham Young University, 2152 West View Building, Provo, UT 84602, USA; theeeg@byu.edu

* Correspondence: gschultz@byu.edu; Tel.: +1-801-422-6332

Abstract: Automated traffic signal performance measures (ATSPMs) have garnered significant attention for their ability to collect and evaluate real-time and historical data at signalized intersections. ATSPM data are widely utilized by traffic engineers, planners, and researchers in various application scenarios. In working with ATSPM data in Utah, it was discovered that five types of ATSPM data anomalies (data switching, data shifting, data missing under 6 months, data missing over 6 months, and irregular curves) were present in the data. To address the data issues, this paper presents a method that enables transportation agencies to automatically detect data anomalies in their ATSPM datasets. The proposed method utilizes the moving average and standard deviation of a moving window to calculate the z-score for traffic volume data points at each timestamp. Anomalies are flagged when the z-score exceeds 2, which is based on the data falling within two standard deviations of the mean. The results demonstrate that this method effectively identifies anomalies within ATSPM systems, thereby enhancing the usability of data for engineers, planners, and all ATSPM users. By employing this method, transportation agencies can improve the efficiency of their ATSPM systems, leading to more accurate and reliable data for analysis.



Citation: Wang, B.; Schultz, G.G.; Macfarlane, G.S.; Eggett, D.L.; Davis, M.C. A Methodology to Detect Traffic Data Anomalies in Automated Traffic Signal Performance Measures. *Future Transp.* **2023**, *3*, 1175–1194.
<https://doi.org/10.3390/futuretransp3040064>

Academic Editors: Chunhsing Ho and Li Zhao

Received: 20 June 2023

Revised: 31 August 2023

Accepted: 21 September 2023

Published: 2 October 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: ATSPM; big data; moving average and standard deviation; performance measures; traffic signals

1. Introduction

Automated traffic signal performance measures (ATSPMs) are an innovative technology that has garnered increasing attention in recent years due to their ability to collect and evaluate real-time and historical data at signalized intersections. This technology enables traffic engineers and planners to optimize mobility, manage traffic signal timing and maintenance, reduce congestion, save fuel costs, and improve safety through data that are passively collected 24 h a day, 7 days a week [1–5].

The Utah Department of Transportation (UDOT) is one agency in the United States that has actively engaged in collecting traffic data and evaluating traffic signal performance throughout the state. UDOT utilizes ATSPM data to evaluate the quality of traffic progression along corridors while displaying unused green time that may be available from opposing traffic movements. The information derived from ATSPM data helps to inform UDOT (and other government agencies using ATSPM data) of vehicle and pedestrian detector malfunctions while measuring vehicle delay and recording vehicle volume, speed, and travel time [6,7]. A 2020 Federal Highway Administration publication estimated that UDOT's ATSPM system saved taxpayers over USD 107 million in a decade by reducing traffic delays [8]. Additionally, an increasing number of states are either fully adopting ATSPMs or preparing to integrate them into their state transportation systems. Many

state agencies leverage ATSPMs to support their daily operations and maintenance, guide transportation policy making, and shape future transportation master planning efforts.

In 2020, a Brigham Young University (BYU) transportation research team developed an analysis procedure that establishes meaningful thresholds for various signal system performance measures and created a statistical scoring tool and interactive data visualization tool to evaluate intersections across a collection of individual performance measures. This scoring tool and interactive data visualization tool were used to provide context to the historic quality of signal system operations across the state of Utah [9]. In the process of developing the scoring and data visualization tools, the research team found several inconsistencies or anomalies in the data, such as data switching, data shifting, missing data, and irregular curves. These anomalies present in the ATSPM data were evaluated, and it was determined that they have the potential to produce inaccurate results for any ATSPM data analysis [10]. Transportation agencies that use ATSPM data for policy making and performance analysis without understanding and correcting for data anomalies may misunderstand traffic conditions and misallocate resources. Therefore, the purpose of this research is to produce a method to automatically identify data anomalies in the ATSPM data that will enable UDOT and other transportation agencies to better understand the data and guide their continuous efforts to improve the use of ATSPM data in the future.

The paper proceeds as follows: a literature review discusses previous ATSPM work from researchers and practitioners, the limitations and data anomalies associated with ATSPMs, the current methods for detecting data anomalies, and the practice of data detection in ATSPMs. The methodology section describes how data anomalies are identified from the UDOT ATSPM dataset and presents a method for automatically highlighting these anomalies for each intersection over time. The results section describes how moving average and standard deviation methods are used to determine data anomalies for performance measures. The discussion section describes the limitations of the research and associated opportunities for ATSPM data improvement and future research. Finally, the conclusion section summarizes the findings and contributions of the research.

2. Literature Review

Over the past several years, a BYU research team has been evaluating the quality of signal operations using ATSPM data in Utah. The primary objective of this research was to evaluate performance measurement data collected through the UDOT ATSPM database. First, suitable performance measures were identified to evaluate traffic signal maintenance and operations by determining which performance measures had appropriate data available for analysis. Next, threshold values for each selected performance measure were established. Finally, a process for overall evaluation of the historic quality of signal systems statewide was developed. The research findings are documented in the literature [9].

During the previous research, several limitations and challenges were encountered. The most significant challenge arose from the incompleteness of the ATSPM database, with missing or extreme values for various performance measures across multiple signals. This presented unexpected difficulties in data evaluation and analysis, as the researchers had to contend with incomplete data. The limited availability of usable data also posed constraints in constructing the scoring method [10].

Various factors may be responsible for causing the data anomalies in the ATSPM datasets. First, Chang et al. [11] found that properly installing microwave sensors to detect traffic volumes and speeds is paramount to data accuracy, while the improper installation of detectors at intersections may lead to inaccurate data. Second, the ATSPM controller event log could be inaccurate when compared with the aggregated performance measure data at various timestamps. Any discrepancies present between the raw data and the aggregated data may appear. Third, the methodology and coding for each performance measure may have errors in the calculations and aggregation performed [12].

Data anomalies and outliers in the ATSPM datasets refer to individual data points that do not behave the same as data points adjacent to the time series. These points are

generally isolated occurrences, but when multiple consecutive points behave in this manner, the pattern is called a subsequent outlier. Blázquez-García et al. [13] provided a review of methods used to detect these types of data anomalies in univariate and multivariate time series. Although various methods exist for addressing subsequent outliers, many rely on assumptions that are not present in the ATSPM data. For example, one of the most used methods to detect subsequent outliers in time series is called the HOT-SAX algorithm [14–16]. The algorithm was developed primarily to identify discords or the most unusual sequence in a time series, but the algorithm requires periodic data to discern an anomaly. Senin et al. [17] used the SAX algorithm, which can locate variable width subsequence discords, but this still relies on repeating data patterns to discern an anomaly. The ATSPM data provide non-periodic data with non-repeating patterns where the anomalies are present as discontinuous patterns. In these cases, methods used primarily for point outliers cannot be applied to detect subsequent outliers.

Multiple research projects have been conducted to visualize real data and compare data anomalies. The Georgia Department of Transportation (GDOT) SigOps metrics dashboard is one such tool that was created by GDOT to visualize ATSPM data longitudinally. Monthly, quarterly, and all-time summary tables are available with trending plots to compare changes in traffic-related performance measures over time [18]. The UDOT Watchdog report is a tool used by UDOT to evaluate signal controller data. The Watchdog report is an automatically generated summary of data errors found in traffic signal controllers over the previous 24 h, which are sent as an email to UDOT traffic engineers each weekday morning. These reports alert UDOT traffic engineers of detector issues that are manifested through inaccurate data [10]. Chamberlin and Fayyaz [19] compared continuous count stations (CCSs) data in Utah to evaluate turning movement count accuracy in ATSPM datasets. They found that the ATSPM volume data had many data anomalies present, mainly in the form of jump discontinuities, where volume data shifted up multiple hundred units for a time and then shifted back down to regular levels [19,20]. These research projects and applications can show when and where the data anomalies appeared but still need traffic engineering judgment to identify the data anomalies. The goal of this research was to create a method that can automatically identify data anomalies across all ATSPM data.

3. Methodology

The methodology section provides a step-by-step description of the research. It begins by selecting specific time periods, intersection movements, individual intersections, locations, and performance measures for analysis. The accuracy of the data is then assessed by comparing ATSPM data with the Watchdog report and CCS data. Various types of data anomalies encountered during the analysis are identified. Finally, the moving average and standard deviation method is employed to detect data anomalies to aid UDOT and other government agencies in automatically identifying data anomalies. The general workflow proposed for this methodology is illustrated in Figure 1.

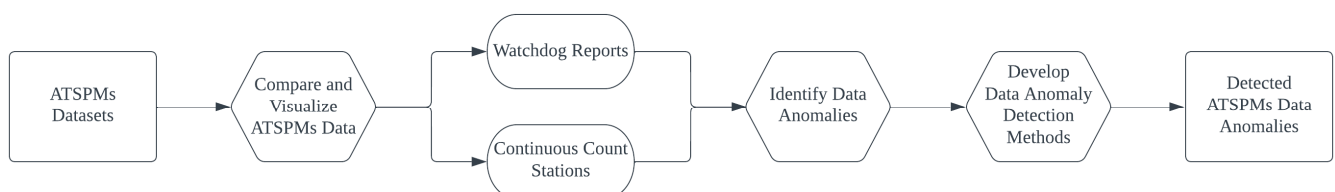


Figure 1. Methodology of ATSPMs data anomalies detection workflow.

3.1. ATSPM Datasets

The aggregated data for this research were acquired through the UDOT ATSPM database, which is located on a UDOT Traffic Operations Center server in the UDOT ATSPM Structured Query Language (SQL) database. This SQL server database has the same measures of effectiveness as the server used in previous research [9]. The UDOT ATSPM database contains traffic data collected at intersections across the state with information on

multiple performance measures. ATSPM data from 1 January 2018 to 31 December 2019 (the most recent aggregated data available at the time of this research) were organized in csv format for this research. Other agencies have similar databases that can be used for data collection.

To simplify and standardize the analysis, the AM peak, or the hours of 7:00 A.M. to 9:00 A.M., were selected as the study period based on common traffic patterns in Utah and as outlined in general in the *Traffic Monitoring Guide* [21]. The AM peak provided a time in which the presence of traffic volumes and split failures would be more representative of overall signal performance. The PM peak was not selected due to its higher traffic volumes compared to the AM peak, potentially leading to more data issues than those found in the AM peak. UDOT traffic engineers recommended that the research team focus on the AM peak as a benchmark because if issues arise during the AM peak, they are likely to be exacerbated during the PM peak [10].

Tuesdays, Wednesdays, and Thursdays were selected based on common traffic patterns in Utah and as outlined in general in the *Traffic Monitoring Guide* [21]. Traffic patterns on the weekends tend to differ from those on the days closer to the middle of the week. Selecting three days in the middle of the week provided a higher probability that the traffic would act in a more uniform manner across the data collection period.

The research team evaluated signal phases 2 and 6 (i.e., major street through movements at each intersection). The data for phases 2 and 6 generally contain the highest traffic volumes for each signal and, hence, the most data to analyze. Choosing phases 2 and 6 also streamlined the analysis by eliminating permitted and protected left turns that may complicate the interpretation of performance measure data by introducing multiple movements in the dataset. It was determined that once the methodology is developed and proven, turning movements could be added.

The initial signal selection was primarily based on the signals used in previous research, where 21 signals from three major corridors in the state of Utah, United States of America, were evaluated [9]. These three corridors were chosen because they had the type of detection required to capture the data needed to calculate the performance measures that would be analyzed. These corridors were 800 North and State Street in Orem, Utah, United States of America and Fort Union Blvd. in Cottonwood Heights and Midvale. In the previous research, the team stated that their research focused on ensuring that the ATSPM data were aggregated correctly, hence the low number of signals chosen. The research team indicated that it would be interesting to expand the number of signals in future research projects [9]. In addition to the 21 signals from the previous research, five signals were added from Fort Union Blvd. in Cottonwood Heights, six signals from 800 North in Orem, four from State Street in Orem, 26 from University Avenue in Provo, and 27 more signals from downtown Provo. Overall, the team added 68 new signals for a total of 89 signals.

The performance measure selected for this research was approach volume. Approach volume is a metric that measures the total number of vehicles passing a single traffic signal detector per 15 min bin over a specified unit of time. It is typically measured using advanced detection located approximately 400 feet before the stop bar [7]. Approach volume was chosen for this research study for multiple reasons. First, approach volume is essential to the calculation of many other performance measures, such as percent arrivals on green (AOG) and percent split failures, which use approach volume in the denominator to convert AOG and split failures to percentages. Therefore, ensuring the accuracy of the approach volumes used was critical. Second, approach volume can also be used to prioritize the importance of signal scores. In previous research, a signal scoring system was developed to give each signal a score based on the values of their performance measures. This scoring system then gave an aggregate score for the signal, and the priority for signal repair was determined by the aggregate score [9]. Although the signal score is crucial, the number of vehicles passing through the signal also provides valuable insight into prioritization. A minor intersection with a lower signal score may be a lower priority than a major intersection (with higher volume) with a slightly higher score. Including approach

volume allows this critical metric to be part of the signal servicing prioritization process for UDOT (or any other agency). Lastly, as a future goal for ATSPM analysis research in Utah is to create a longitudinal analysis of performance measures at signals over time, it is essential to consider whether the volume moving through an intersection has changed. If an increase in volume over time is observed, UDOT (or any other agency) can then investigate the reasons why the increase occurred, such as a holiday, a major event, weather, or a signal timing change. Including approach volume allows the research team and governing agency to observe and analyze increases or decreases over time.

3.2. Compare and Visualize ATSPM with Additional Data Sources

3.2.1. Watchdog Reports

To better understand the ATSPM data, the data from the Watchdog reports were visualized and evaluated using the statistical computing and graphics tool R. Figure 2 displays signals containing Watchdog alert categories and the day on which the Watchdog alerts were reported between July 2018 and December 2019. This figure uncovered potentially systemic issues in the ATSPM, and it was unclear whether the Watchdog reports could conclusively explain the inconsistencies found in the data.



Figure 2. Watchdog reports alert by signal from July 2018 to December 2019.

The most common type of alert in the Watchdog reports was low advanced detection counts. Under proper operating conditions, this alert is only reported when low numbers of vehicles are detected, which may occur at different times for each signal. The data showed that there were multiple days when nearly all 30 signals on the Watchdog reports had the low advanced detection counts report. This occurrence happened for one day in mid-November 2018 and 14 days between September 2019 and December 2019. The frequency and simultaneous occurrence of this alert suggest an internal issue in the ATSPM system. It is unclear whether this was due to an adjustment made in the system manually or an internal bug that simultaneously affected every signal, but it is too coincidental to suggest that every signal had the same issue on the same day across 30 signals.

The research team found that the Watchdog reports do not fully explain the discontinuities found in the data. The team compared the dates of Watchdog alerts to the dates when anomalies occurred in the ATSPM volume data and found that, while many of the alert dates were close to the anomaly dates, the two did not directly correlate, and no

clear patterns existed between the two datasets. Another issue discovered was that the “too few records” alert was reactionary to the data trends rather than explanatory about detector issues. While many of the alerts focused on detector or controller issues that could potentially affect the data, the “too few records” alert was generated directly when fewer than 500 records were made. This made it less useful for the research team, as it did not help explain why the volumes may have been low or irregular on a particular day. Additionally, many of the reports did not correlate with discontinuities in the data. While the “low advanced detection counts” alert was found across the system, it also occurred at unique positions for different signals rather than only at the same places for every signal. The research team assumed that it would be one of the most likely Watchdog alerts to influence the data, but there was not a strong correlation between the dates of the “low advanced detection counts” alerts and the discontinuities in the data. As a result, additional methods needed to be developed to identify data anomalies.

3.2.2. Continuous Count Stations

The research team used the UDOT CCS data as a baseline for validating the ATSPM volume data. Like the research conducted by Chamberlin and Fayyaz [19], the research team found CCS stations near ATSPM volume data collection sites and aligned the data graphically to compare the daily volumes. The CCS data were downloaded from the UDOT data portal [20]. Four CCS stations were selected due to their proximity to ATSPM-enabled signals and given names based on their relative locations: 800 East Orem station, 800 North Orem station, North University Avenue Provo station, and South University Avenue Provo station. Traffic volumes were selected for the AM Peak hours of 7:00 A.M. to 9:00 A.M. on Tuesday, Wednesday, and Thursday and for all through lanes within both signal phases 2 and 6 for the reasons outlined in Section 3.1. For simplicity, volumes for signal phases 2 and 6 were combined, and all individual lane volumes were combined separately.

Figure 3a,b displays the CCS volume data and ATSPM volume data comparisons for two signal locations during the AM peak hours on Tuesday, Wednesday, and Thursday. Figure 3a shows that while the ATSPM data are undercounted, supporting the findings from Chamberlin and Fayyaz [19], the trends in the volume data are very similar. In contrast, Figure 3b shows a problematic example where the ATSPM data do not match the CCS count data and where data anomalies may be present in the ATSPM data.

The research team faced a challenge in comparing ATSPM data to CCS data because not all selected intersections were in a corridor with CCS stations, and the accuracy of the comparison decreases as the distance between the intersection and the CCS station increases. This limitation made the method less comprehensive than originally anticipated due to the limited number of CCS stations available along the chosen corridors. As demonstrated by Chamberlin and Fayyaz [19], CCS data are useful for validating ATSPM volume counts, but in this case, they were not effective due to the proximity of the CCS stations to the study intersections and any future intersections analyzed. As a result, additional methods needed to be developed to identify data anomalies.

3.3. Identify Data Anomalies

To enable the research team to select signals that had data available for approach volumes, the research team needed to determine the percentage of available data at each signal selected for analysis. The research team used a count function in the statistical computing and graphics tool R to identify the number of “NA” values each signal had for its five performance measures and then determined which signals had “NA” data for approach volume. The “NA” values are those that were recorded from the traffic detectors but, for some reason, were not specified in the ATSPM data output. These data points are different from missing values, which are values that are not recorded at all. The team then narrowed the signal selection down to only the 32 signals that contained data for approach volume. This ended up being a reduction of 64 percent from the original 89 signals. The 32 signals selected had data for each performance measure, but the data were not always

complete. Each signal only had partial data for each measure to be selected. Figure 4 shows the 32 intersection locations selected for this research.

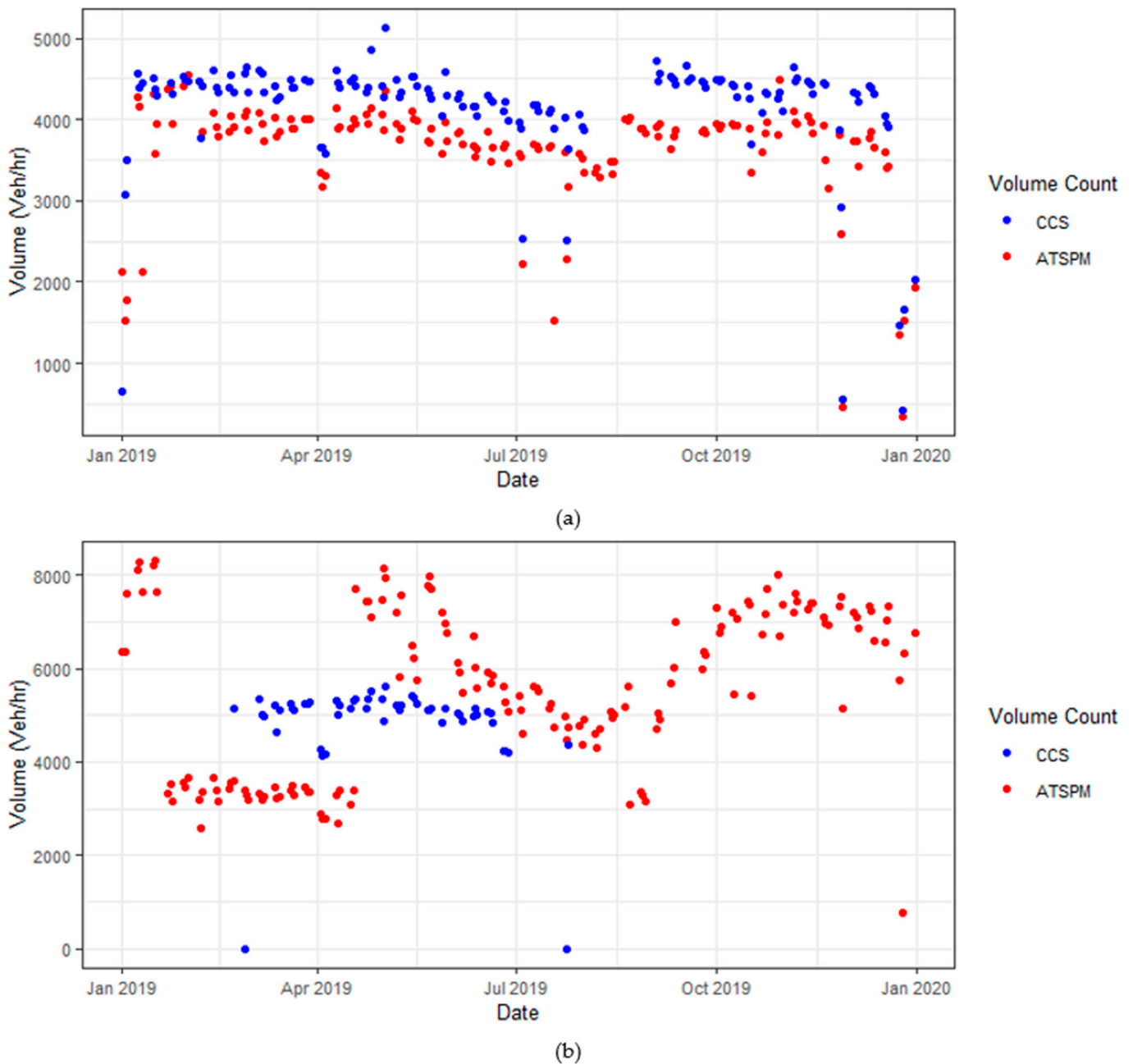


Figure 3. Comparison of traffic volume between ATSPM and CCS data for: (a) signal 6421 vs. North University Avenue CCS; (b) signal 6306 vs. 800 East CCS.

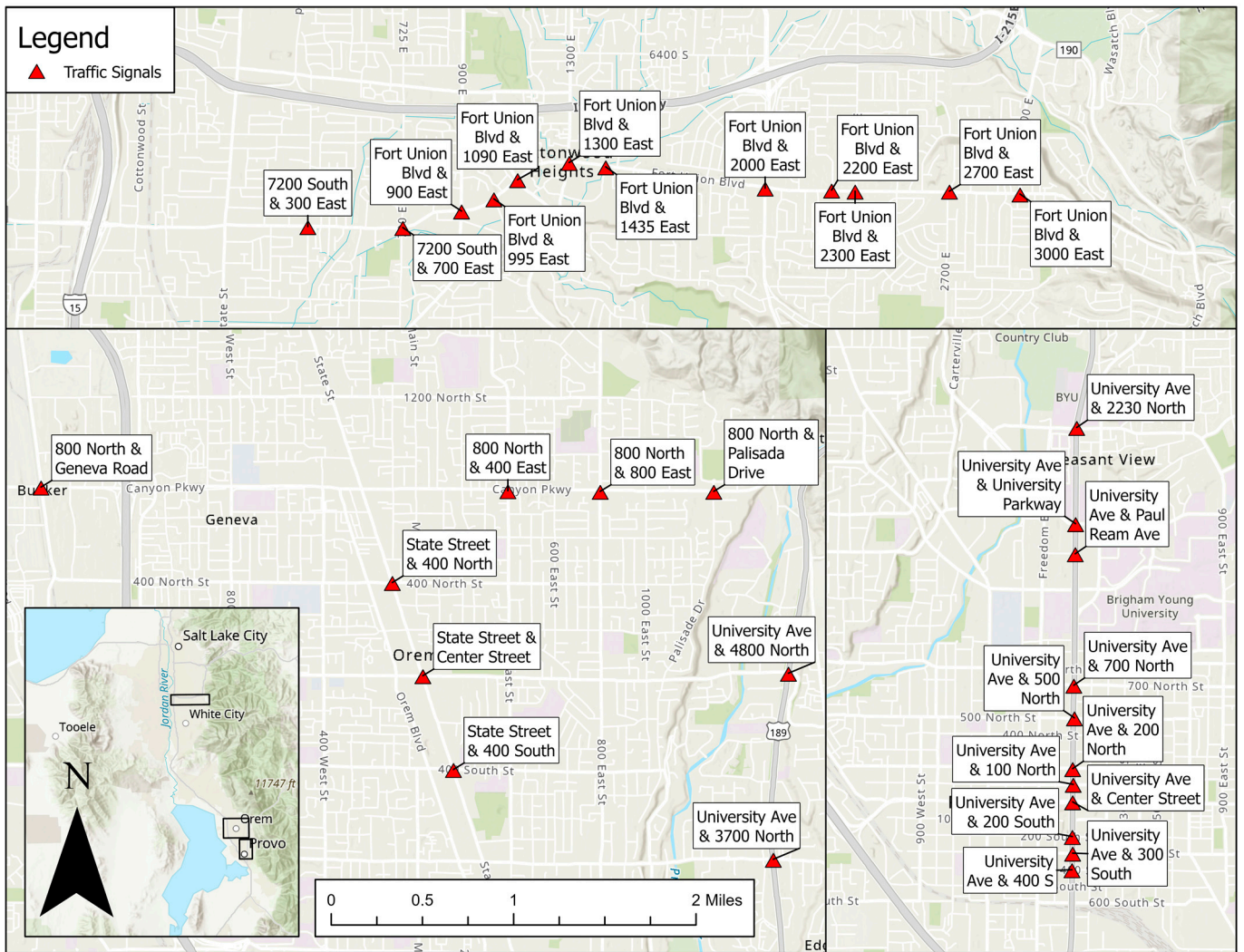


Figure 4. Study intersections and locations in Utah.

While data completeness was important to signal selection, data quality was also a concern for the research team. A future objective of ATSPM research in Utah is to measure the longitudinal trends found in signal data over time. However, if major anomalies are present in the data, the accuracy of the longitudinal trends could be distorted. To classify and properly analyze anomalies present in the signal data, five types of anomaly categories were identified: data switching, data shifting, missing data under 6 months, missing data over 6 months, and irregular curves. The definitions of each anomaly are provided in Table 1, while Figure 5 displays an example of each type of data anomaly. The five types of data anomalies are representative of all anomalies present in the data.

Table 1. Data Anomalies and Description.

Data Issue	Issue Description
Data Switching	Data switching is characterized by offset phases that periodically switch positions. The approach volumes were switching between Phase 2 and Phase 6 at signal 4029, Fort Union Blvd. and 700 East, but this issue was rare; it only appeared in 1 of the 32 signals (3%).
Data Shifting	Data shifting is characterized by the sudden and drastic upward shift of a phase or multiple phases, causing volume data to jump multiple hundred vehicles for a period of time and then drop back down to an average level. This issue was found in 6 of the 32 signals (19%). Potential causes could include changes made in the Traffic Operations Center, systemic errors, or temporary detector issues.

Table 1. Cont.

Data Issue	Issue Description
Data missing Under 6 Months	Missing data under 6 months is characterized by a gap of 1–6 months of missing data, which is categorized as either a missing data point presented as an “NA” value or a missing value presented as a “0”. Data missing for fewer than 6 months would not become a major problem for the longitudinal analysis. This issue was found in 7 of the 32 signals (22%), and the potential cause could be detector malfunction.
Data missing Over 6 Months	Missing data over 6 months is characterized by a gap of 6 or more months of missing data. This issue was found in 12 of the 32 signals (38%), and the potential cause could be detector malfunction.
Irregular Curves	Irregular curves are characterized by data that follow curves with continuous jumps that cannot be explained by regular traffic patterns in time or seasonality. Many of these curves involve large, sudden increases or decreases in values or large and irregular trends. This issue was found in 7 of the 32 signals (22%), and the potential cause could be long-term road construction.
Normal Operating Conditions	Any signals that do not have data switching, data shifting, or irregular curves will count as normal signals. These signals may have missing data but are characterized as normal because they do not exhibit any anomalies that shift or manipulate the input data. For 19 of the 32 signals (59%), no issues were found.

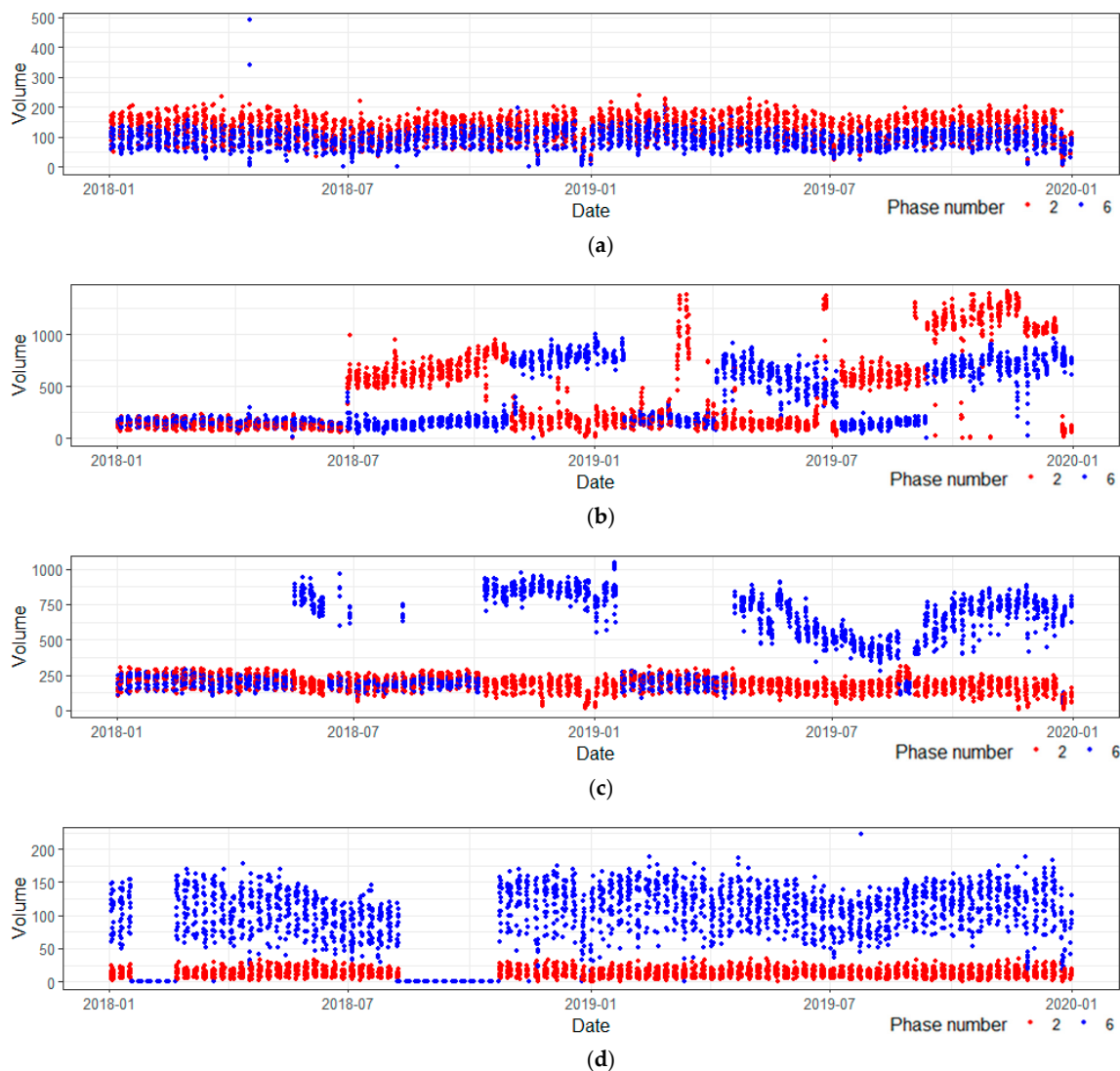


Figure 5. Cont.

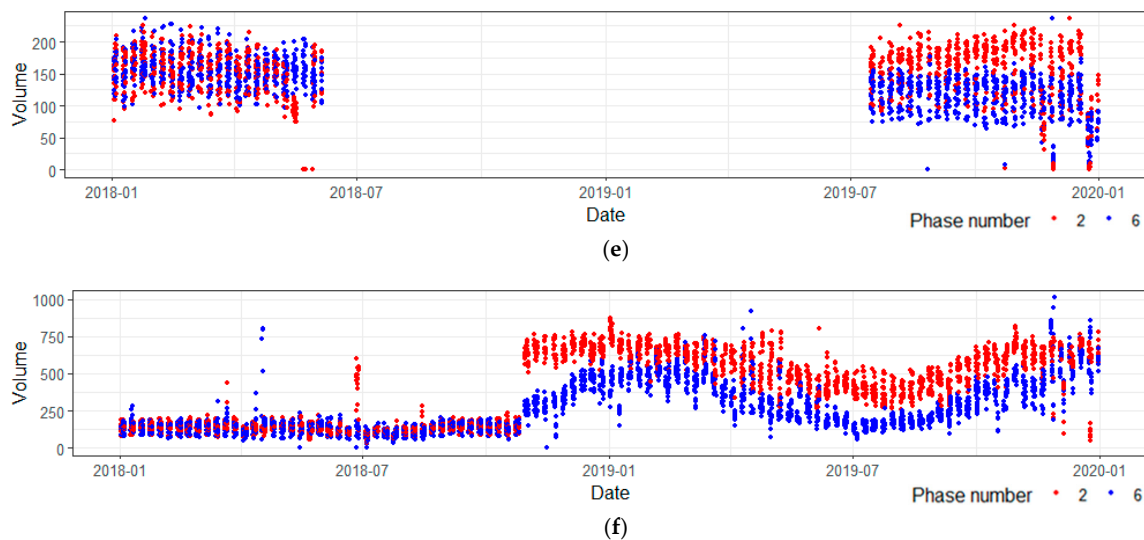


Figure 5. Example of types of data anomalies: (a) normal operating conditions; (b) data switching; (c) data shifting; (d) data missing under 6 months; (e) data missing over 6 months; (f) irregular curves.

3.4. Data Anomaly Detection Methods

Prediction model-based methods work with non-periodic data where normalcy in the data is not reflected in similar data patterns over time. These methods use a moving window of past data to predict what the future data should look like. Because these methods use a moving window that predicts the next point in the sequence iteratively, methods applied to point outliers can be applied to subsequent outliers. For example, Basu and Meckesheimer [22] calculated the median of the window of past data as a control to which the next data value is compared against. Zhang et al. [23] utilized an Auto-regressive Integrated Moving Average model and calculated the moving average of the window of past data as the control. These types of methods allow outlier detection to be performed on a non-periodic dataset where the generation of data is not easily modeled or predicted.

The research team initially explored two statistical and mathematical methods to create a data anomaly detection tool. First, the research team created a linear regression model to predict the general trends of the ATSPM data patterns over time [24]. It was run against the ATSPM volume data to create a regression of predicted volumes for a typical day without data anomalies. This predicted volume regression was then used by the team as a baseline to compare against the actual ATSPM volume data to identify data anomalies that diverged from the predicted volumes. While the linear regression model helped to identify significant data anomalies like data shifting or irregular curves, the method was not as sensitive to sudden, brief shifts in data values [10]. Second, the research team aggregated the 15 min bin volume data for each signal and created a histogram for each signal and phase. A normal signal without anomalies in the data was determined to follow a normal distribution curve. For each histogram, the research team calculated the mean volume and standard deviation for anomaly detection. The data were considered an anomaly if any volume data point was found to be two standard deviations away from the mean or outside of a 95 percent confidence interval. Although this provided a direct way to identify the outliers, there were several limitations to the procedure, particularly for intersections with high numbers of zero volumes in the dataset [10].

The third (and chosen) method explored was the moving average and standard deviation method. The research team utilized a moving average and moving standard deviation method to detect significant shifts in data values. It was hypothesized that this method would be more effective than assuming one blanket value of average and standard deviation for the entire signal and phase. The main advantage of this approach was the ability to adjust the moving window, or the number of points included in the calculation of the average and standard deviation. A larger window would result in a more generalized

average and standard deviation, while a smaller window would provide a more localized average and standard deviation.

The method developed used the moving window to identify the location of anomalies and to calibrate the model. The length of the moving window was used to calculate the moving average and standard deviation, which were then used to calculate a z-score for each data point, as defined in Equation (1). The z-score value was critical in determining when and where anomalies were occurring, as a high z-score value was likely indicative of a significant shift in data values. To ensure that the model was correctly adjusted, a sensitivity analysis was performed to optimize the length of the moving window and the z-score for analysis.

$$z = |x - \mu| / s \quad (1)$$

where:

- z = z-score,
- x = volume of observation,
- μ = moving average,
- s = moving standard deviation.

The research team conducted a sensitivity analysis on multiple window lengths to choose an appropriate moving window length. The volume data used for this method were 15 min bins of the AM peak hours from 7:00 A.M. to 9:00 A.M. on Tuesdays, Wednesdays, and Thursdays from January 2018 to December 2019. This resulted in eight data points for each day and approximately 96 data points per month. The team tested moving window lengths of 100, 200, and 300 data points, which represent approximately 1 month, 2 months, and 3 months, respectively.

For the sensitivity analysis, the volume data were plotted for each signal and phase as a time series in the statistical computation and graphics tool R, as shown in Figure 6. Figure 6 displays a sample from signal 4028 for both phases 2 and 6. The x -axis represents the study dates, including AM peaks on Tuesday, Wednesday, and Thursday from January 2018 to December 2019. The y -axis indicates the volume counts every 15 min during the AM peak hour. Each data point was colored based on its z-score, with blue representing 0 to 1.99 and red representing 2.0 and greater. The moving window was optimized so that when a jump anomaly occurred in the data, the data points directly before the shifted data points had a z-score of 2 or greater. A z-score of 2 was chosen as the cutoff because any points within two standard deviations are within a 95 percent confidence interval of the mean. This sensitivity analysis was performed graphically, and the results will be presented in the results section.

A moving average and standard deviation were chosen to identify jump discontinuities within the ATSPM data. The theory was that the standard deviation between points should be small when the signal operates with little to no data anomalies. However, the standard deviation between points would dramatically increase when a jump discontinuity occurred, providing a momentary spike identifying the location of an anomaly in the data.

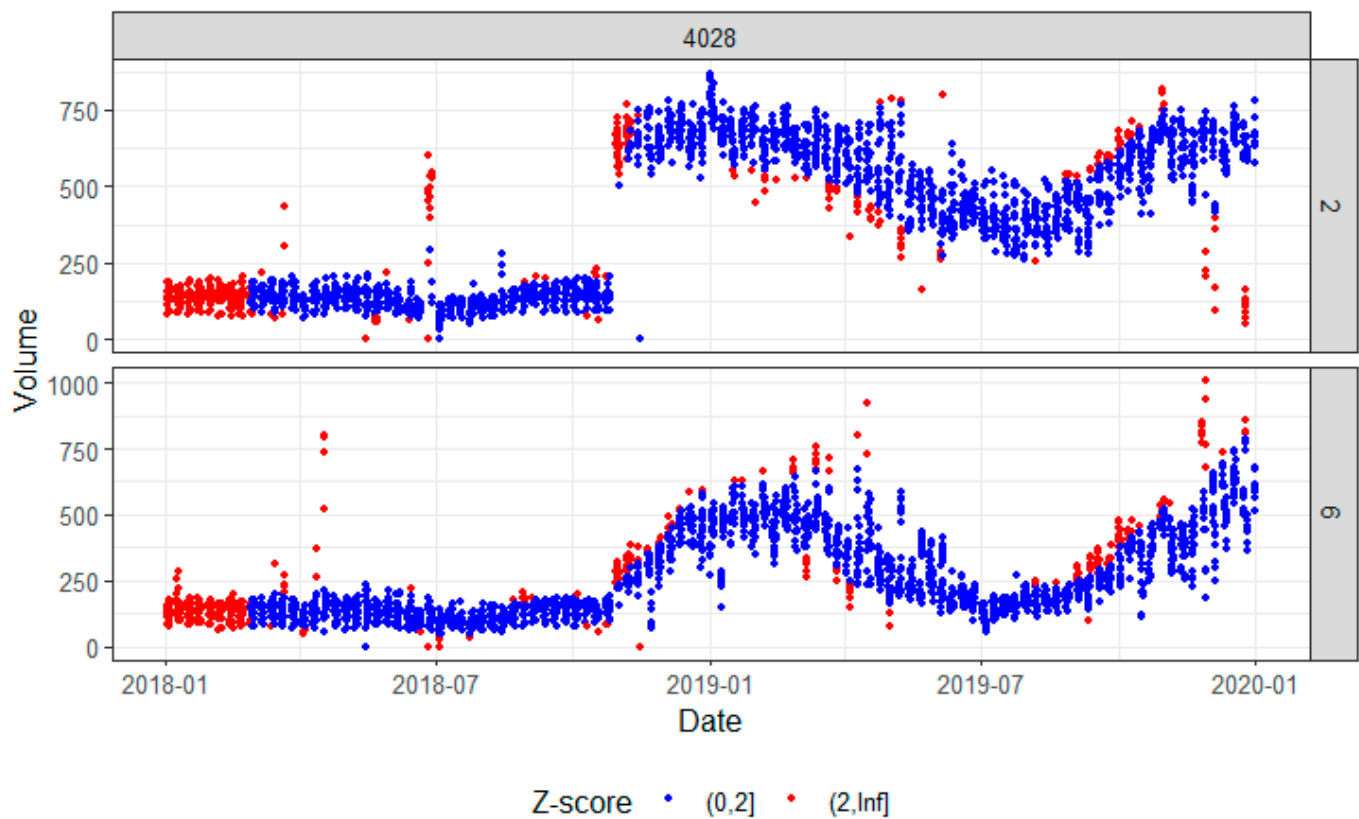


Figure 6. Example of color-coded z-score for data anomalies (Signal 4028).

4. Results

The objective of this research was to employ the moving average and standard deviation method to establish distinct patterns and to identify anomalies in ATSPM data. Figure 7 provides a visual comparison of signal results using different moving window sizes: 100 data points, 200 data points, and 300 data points for phase 2 at signal 4028. The first 100, 200, and 300 data points are shown in red because there is no previous data to calculate the moving average and standard deviation. The figure demonstrates that a moving window of 100 data points does not offer sufficient data for accurate outlier identification. Conversely, a moving window of 300 data points presents the disadvantage of incorporating excessive data, resulting in insufficient data for intersections and the potential generation of more outliers. After conducting an extensive sensitivity analysis on the 32 selected intersections, a moving window of 200 data points was determined to strike a balance, allowing the calculation of the moving average and standard deviation within the chosen window size.

Figure 8 presents the visualization of the z-scores for phases 2 and 6 at five sample signals (Signal 6408, 6409, 6410, 6411, and 6416) spanning a two-year period. In this graph, the blue dots represent z-scores less than 2, while the red dots represent z-scores greater than or equal to 2. After the initial 200 points, the graph displayed red dots whenever there was a significant shift in data volume. In a few instances, the first 200 points are shown in gray due to previous missing values. To facilitate the identification of data anomalies, Figure 9 was created for the same five sample signals, omitting the first 200 data points and the blue data points. The red dots, appearing only once or a few times in a short period, are considered noise and can be disregarded as the data quickly return to normal.

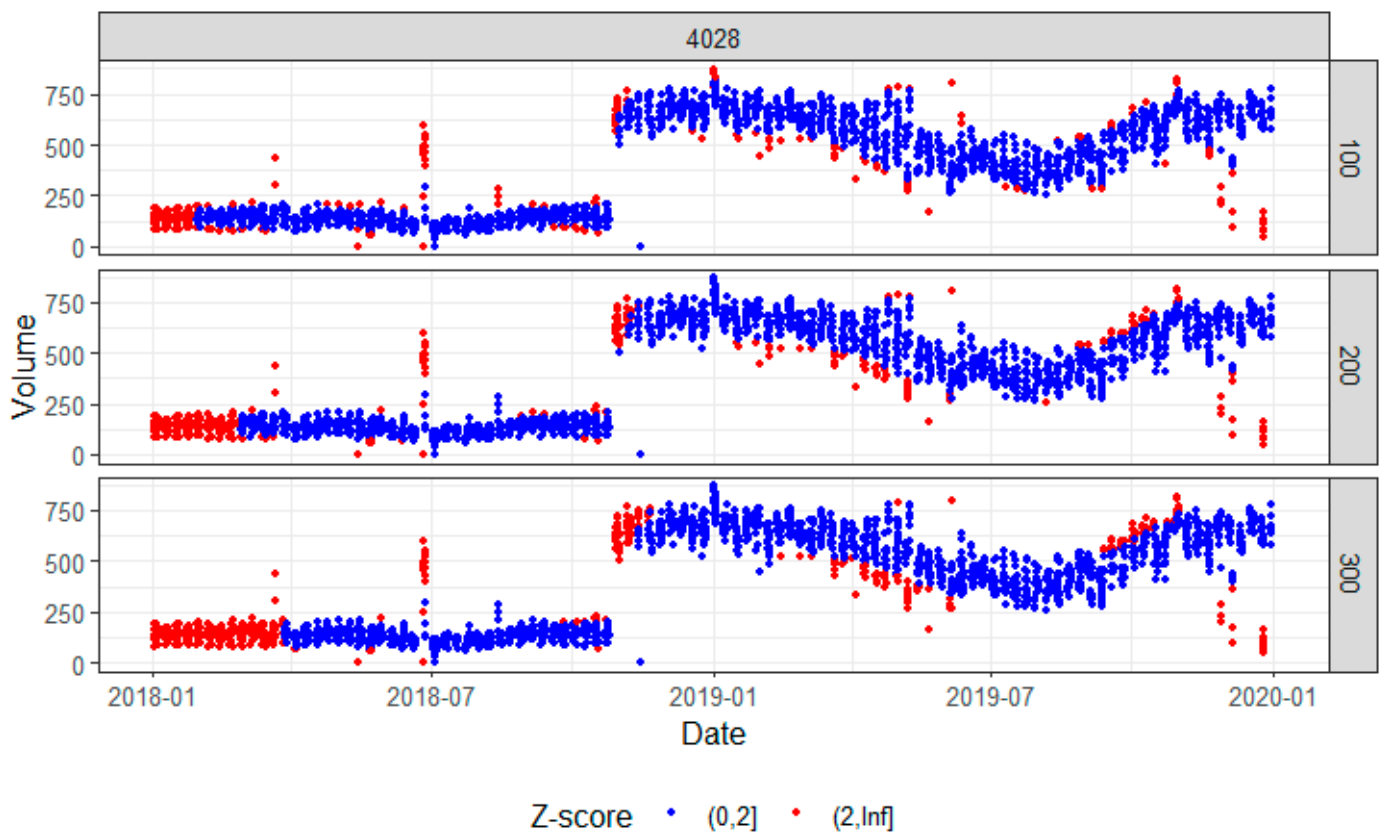


Figure 7. Moving average and standard deviation z-score visualization with 100, 200, and 300 data point moving windows (Signal 4028).

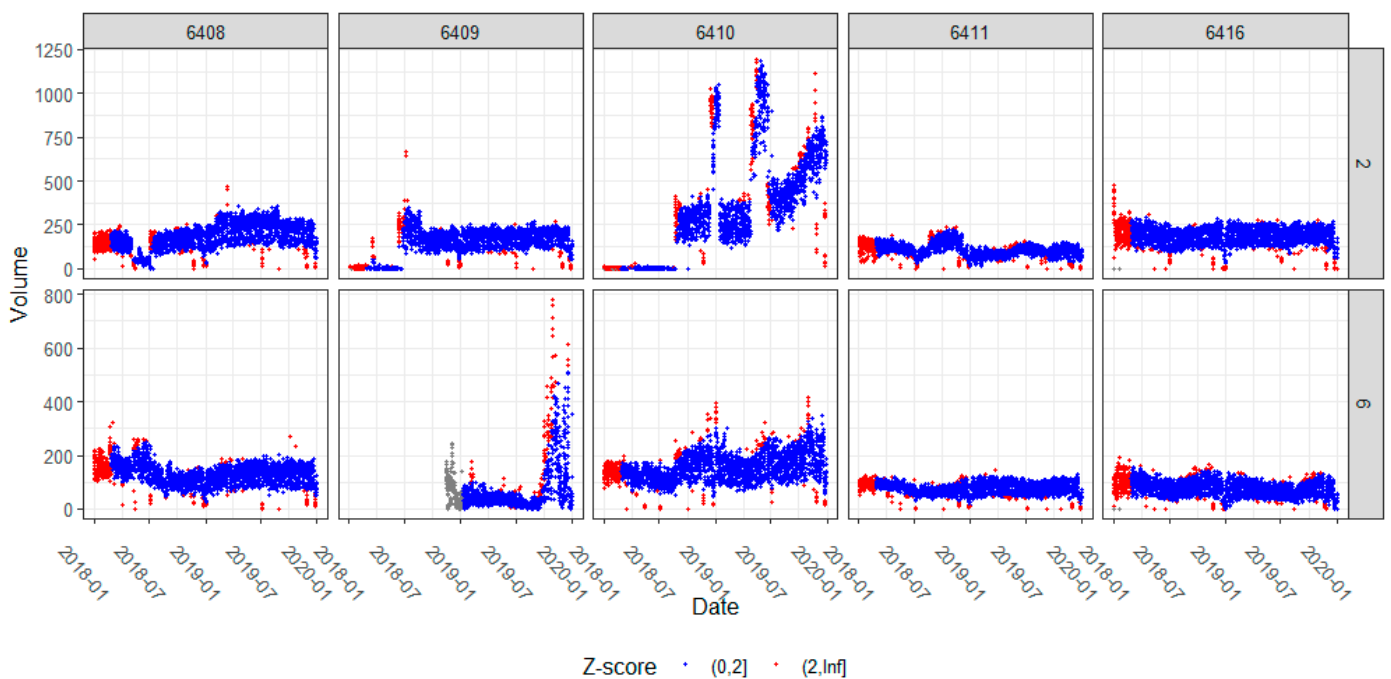


Figure 8. Moving average and standard deviation with z-score calculations (Signals 6408, 6409, 6410, 6411, 6416).

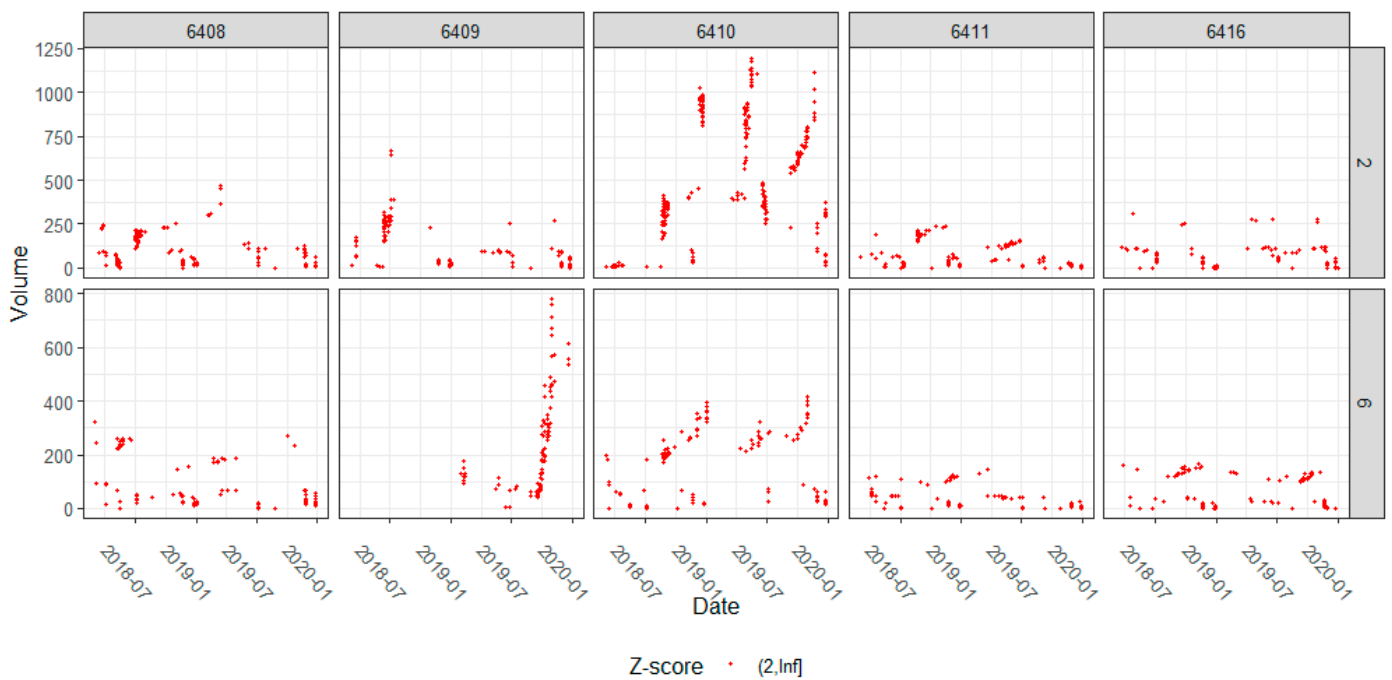


Figure 9. Detected data anomalies (Signals 6408, 6409, 6410, 6411, 6416).

To identify potential issues to be aware of, the research team established that any sequence of eight continuous red dots should be flagged. Eight data points would indicate anomalies persisting for more than one day. Hence, this method enables UDOT, or another governing agency, to automatically identify occurrences of data anomalies. The detailed results of the moving average and standard deviation evaluation for each signal can be found in Appendix A.

5. Discussion

ATSPM data are used to evaluate real-time and historical data at signalized intersections. The data are used for operational analyses, signal timing plans, congestion reduction, safety, and more. Although the benefits of ATSPM data in traffic analyses are proven, the data are not without issues related to data completeness and accuracy. For example, out of the initially selected 89 signals analyzed in this research project, 57 signals had missing data for at least one performance measure, rendering them unsuitable for analysis. This resulted in a reduced number of usable signals, totaling 32 or approximately one-third of the original selection.

Even among the signals with some data available for all performance measures, missing data remained a prevalent issue. For seven signals (nearly 22 percent of the 32 signals), missing values were present for less than 6 months of data and 12 signals exhibited over 6 months of missing data, posing significant obstacles to conducting a comprehensive analysis for almost half of the signals. Additionally, various other data anomalies were observed, including data shifting at six signals, which hindered the reliability and accuracy of simple linear regressions for trend predictions. Irregular curves were also common, with seven signals displaying data curves deviating from normal trends in approach volume data.

The data anomalies presented substantial challenges for analyzing the data, with 23 out of 32 signals affected and 10 of those exhibiting multiple anomalies. Consequently, detecting data anomalies became considerably more complex than initially anticipated. However, the moving average and standard deviation method demonstrated robustness in identifying different types of data anomalies for the selected signals. After considerable sensitivity analyses, the 200-point moving window was deemed appropriate for this research. For different sets of signals and study periods, further sensitivity analysis may be required to

determine the optimal moving window size, although it is expected that 200 points will be an acceptable starting point for any future analysis. At present, the data visualization tool stands as the most flexible and comprehensive method employed by the research team, enabling traffic engineers to visualize ATSPM data longitudinally and promptly identify occurrences of data anomalies.

Future research could explore data anomalies for other performance measures. Additionally, studying the impact of these anomalies on transportation agencies' decision making and policy formulation could provide valuable insights. Once UDOT enhances the accuracy of its ATSPM dataset, the research team intends to utilize the scoring system outlined in previous research [9] to reassess intersection performance longitudinally and conduct before–after studies to gauge potential improvements.

The primary recommendation from the research team was to investigate the root causes of data anomalies and missing data within the ATSPM database. Conducting a thorough analysis to identify the underlying reasons for these issues will contribute to obtaining more usable data. Notably, the discoveries made by the research team regarding missing and erroneous data prompted UDOT traffic engineers to seek assistance from consulting firms and coding technicians to rectify these data-related problems.

To enhance data accuracy and quality, several steps are suggested to aid in the data remediation process. This study identifies multiple actions that can be taken by transportation agencies to ensure consistent accuracy in their ATSPM data. First, intersections exhibiting unusually high numbers of data anomalies should undergo inspections of their data detectors. Inaccurate data can result from poorly installed detectors. Second, if no issues are detected at the data detector level, the controller event log should be compared with the aggregated performance measure data at the timestamps of data anomalies. This step helps identify any discrepancies between raw data and aggregated data. Third, the methodology and coding for each performance measure should undergo evaluation to ensure accurate calculations and proper aggregation of data. Taking these actions will enable agencies to identify and mitigate ATSPM issues and data anomalies.

6. Conclusions

This research was the first known study to detect traffic data anomalies and provide an automatic statistical method for detecting data anomalies in ATSPM data. Missing data and various anomalies posed significant obstacles in analyzing performance measures across the selected signals in Utah. However, the moving average and standard deviation method has proven effective and flexible in detecting anomalies, and the data visualization tool provides valuable insights for anomaly identification.

Future research should explore data anomalies longitudinally for different performance measures and investigate the impact of these anomalies on decision-making and policy formulation within transportation agencies. Furthermore, once data accuracy is improved, reassessing performance longitudinally and conducting before–after studies can help evaluate potential improvements.

The primary recommendation for transportation agencies is to investigate the causes of data anomalies and missing data within their respective databases. Implementing a thorough analysis to identify root causes will contribute to obtaining more usable and reliable data. Taking steps such as inspecting data detectors, comparing controller event logs with aggregated data, and evaluating methodology and coding for accurate calculations can enhance data accuracy and quality.

The importance of data accuracy cannot be overstated, as it directly influences the decision-making process for transportation engineers and policymakers. By striving for higher data accuracy and consistently monitoring anomalies, transportation agencies make more informed decisions, support policy development, and ultimately improve the efficiency and effectiveness of their systems.

In conclusion, addressing the limitations related to data completeness and accuracy is crucial for conducting robust traffic data analysis. By implementing the recommended steps

and refining data quality, transportation agencies harness the full potential of their data to inform decision-making, improve system performance, and enhance overall transportation operations.

Author Contributions: Conceptualization, B.W., G.G.S., G.S.M. and M.C.D.; Methodology, B.W., G.G.S., G.S.M. and D.L.E.; Validation, B.W. and M.C.D.; Formal Analysis, B.W. and M.C.D.; Investigation, B.W. and M.C.D.; Resources, B.W., G.G.S., G.S.M. and D.L.E.; Data Curation, B.W. and M.C.D.; Writing, Original Draft Preparation, B.W. and M.C.D.; Writing, Review and Editing, G.G.S., G.S.M. and D.L.E.; Visualization, B.W., G.G.S., G.S.M. and M.C.D.; Supervision, G.G.S., G.S.M. and D.L.E.; Project Administration, G.G.S.; Funding Acquisition, G.G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Utah Department of Transportation grant number 218405.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from the Utah Department of Transportation and is available from the authors with the permission of the Utah Department of Transportation.

Acknowledgments: The authors acknowledge the Utah Department of Transportation (UDOT) for funding this research and the following individuals on the UDOT Technical Advisory Committee for helping to guide the research: Mark Taylor, Adam Lough, Degen Lewis, Ivana Vladisavljevic, Michael Blanchette, W. Scott Jones, and Travis Jensen. 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 US Department of Transportation. The Utah Department of Transportation makes no representation or warranty of any kind and assumes no liability, therefore.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

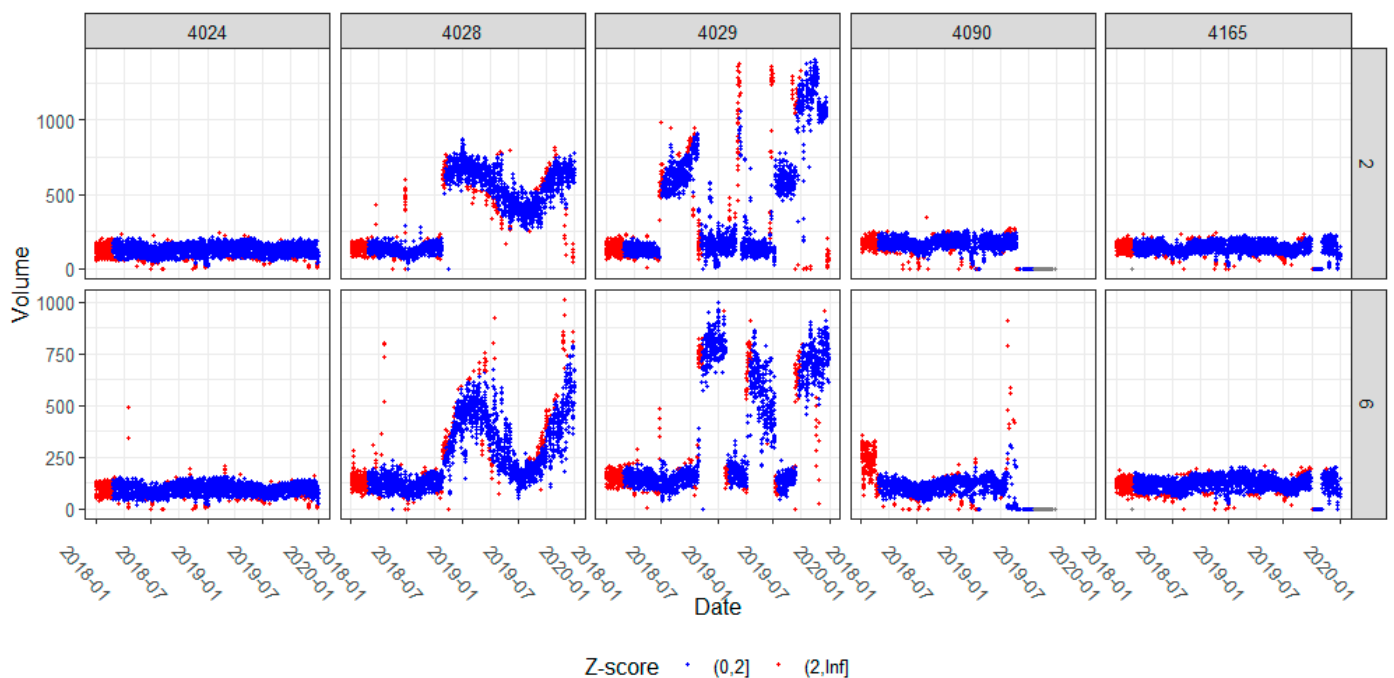


Figure A1. Moving average and standard deviation with z-score calculations (Signals 4024, 4028, 4029, 4090, 4165).

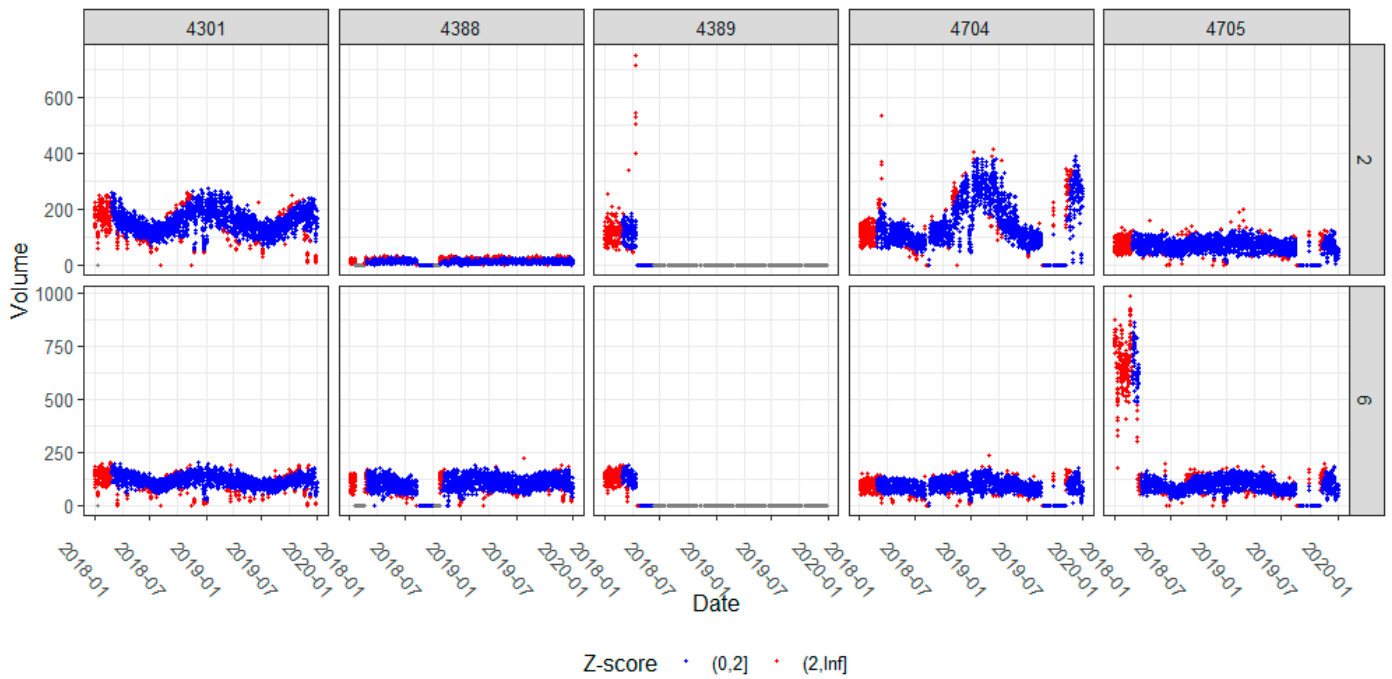


Figure A2. Moving average and standard deviation with z-score calculations (Signals 4301, 4388, 4389, 4704, 4705).

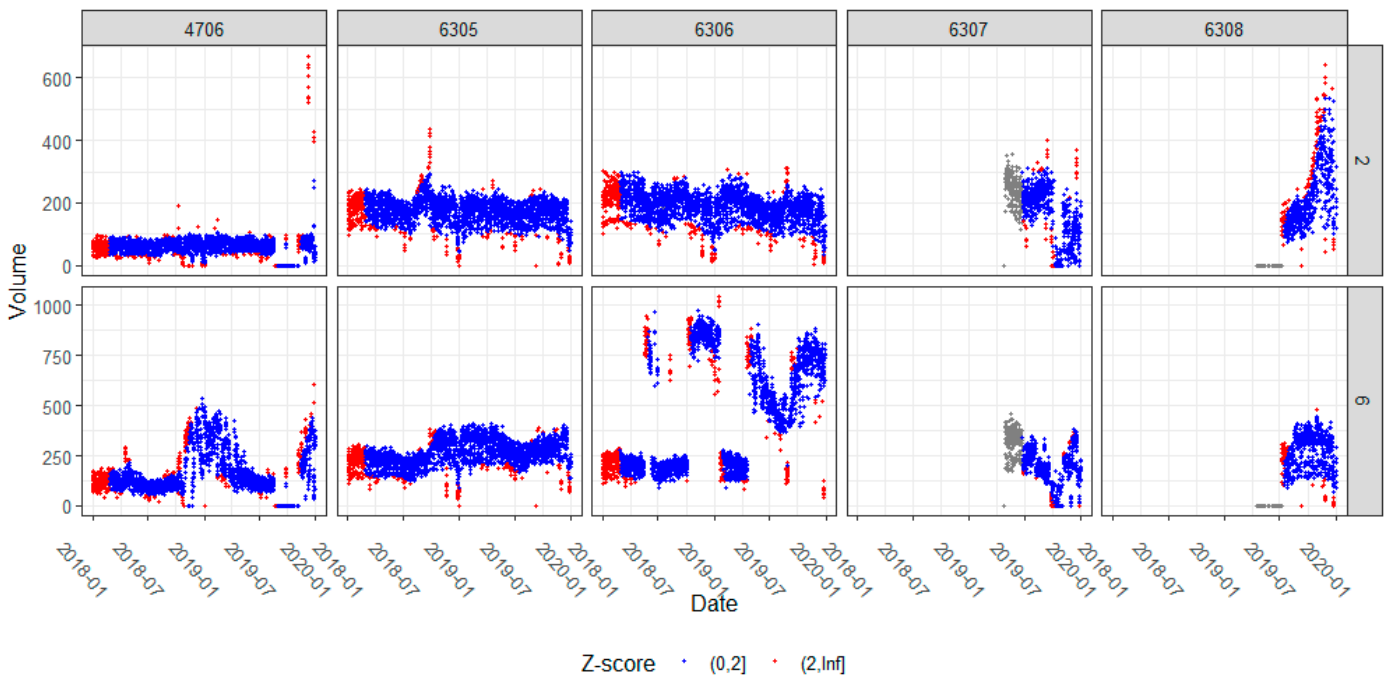


Figure A3. Moving average and standard deviation with z-score calculations (Signals 4706, 6305, 6306, 6307, 6308).

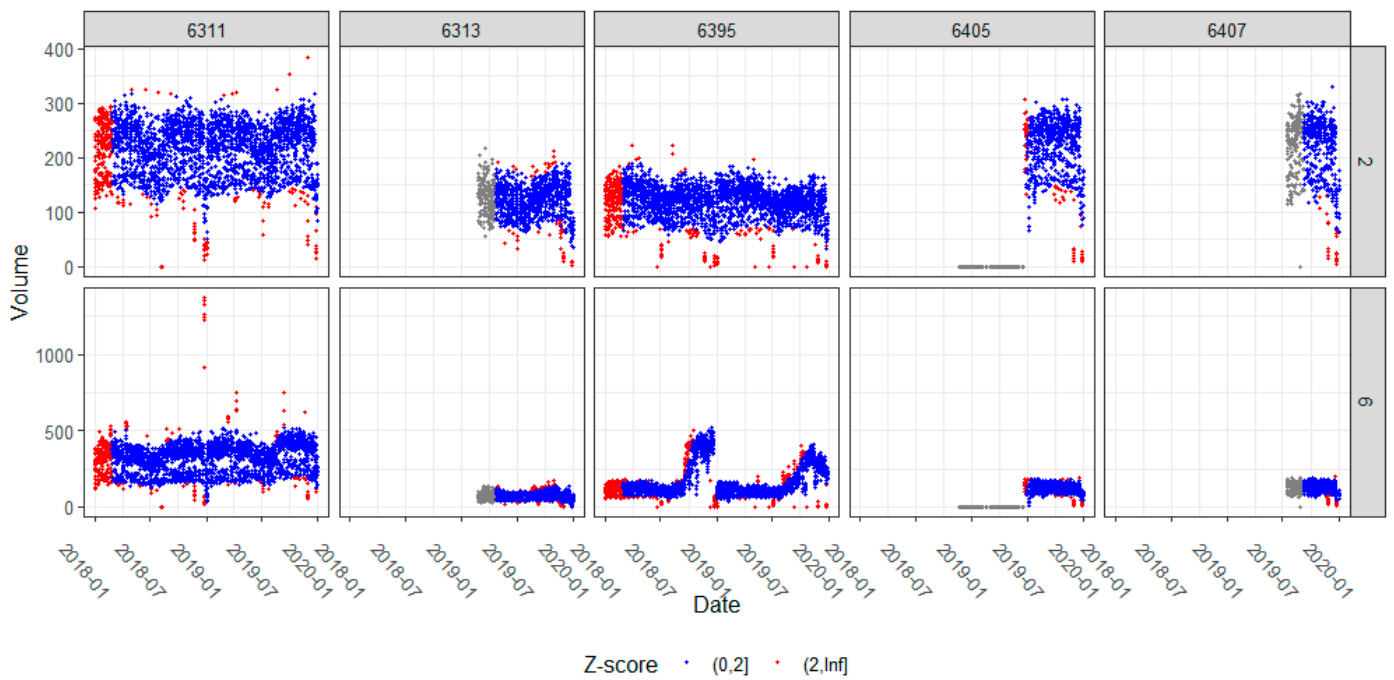


Figure A4. Moving average and standard deviation with z-score calculations (Signals 6311, 6313, 6395, 6405, 6407).

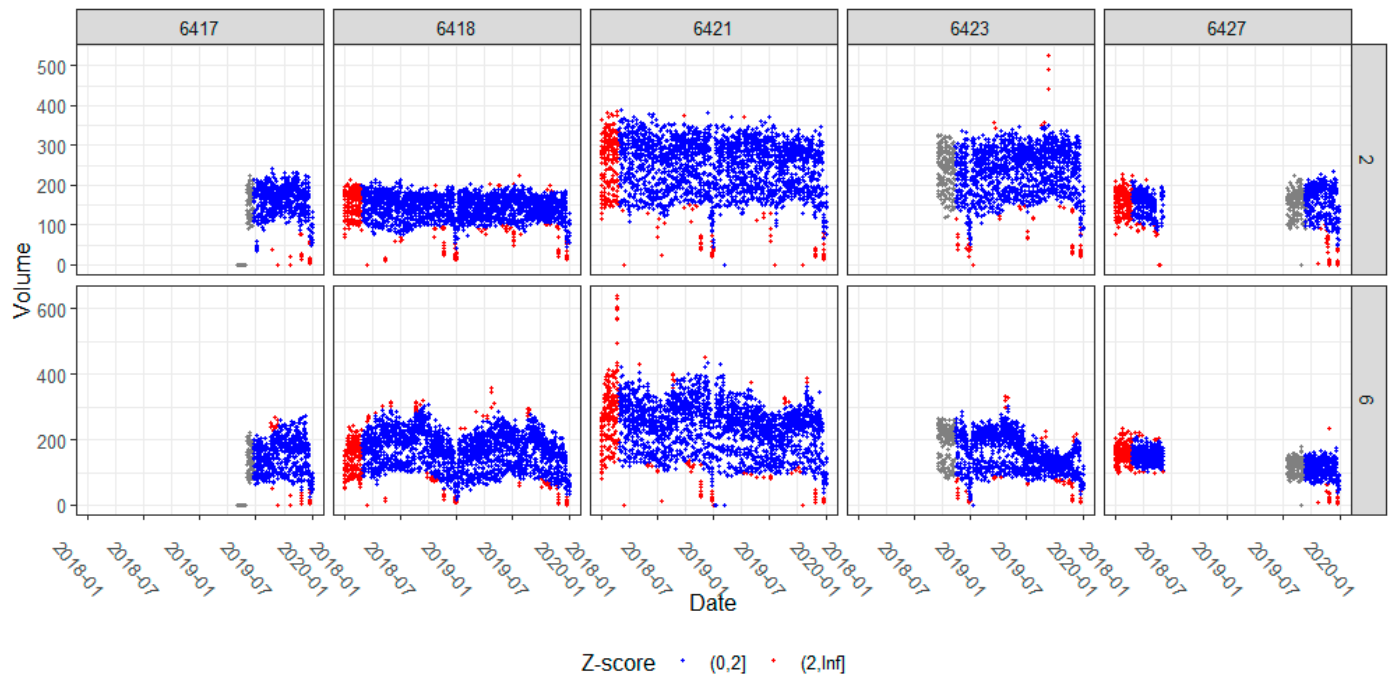


Figure A5. Moving average and standard deviation with z-score calculations (Signals 6417, 6418, 6421, 6423, 6427).

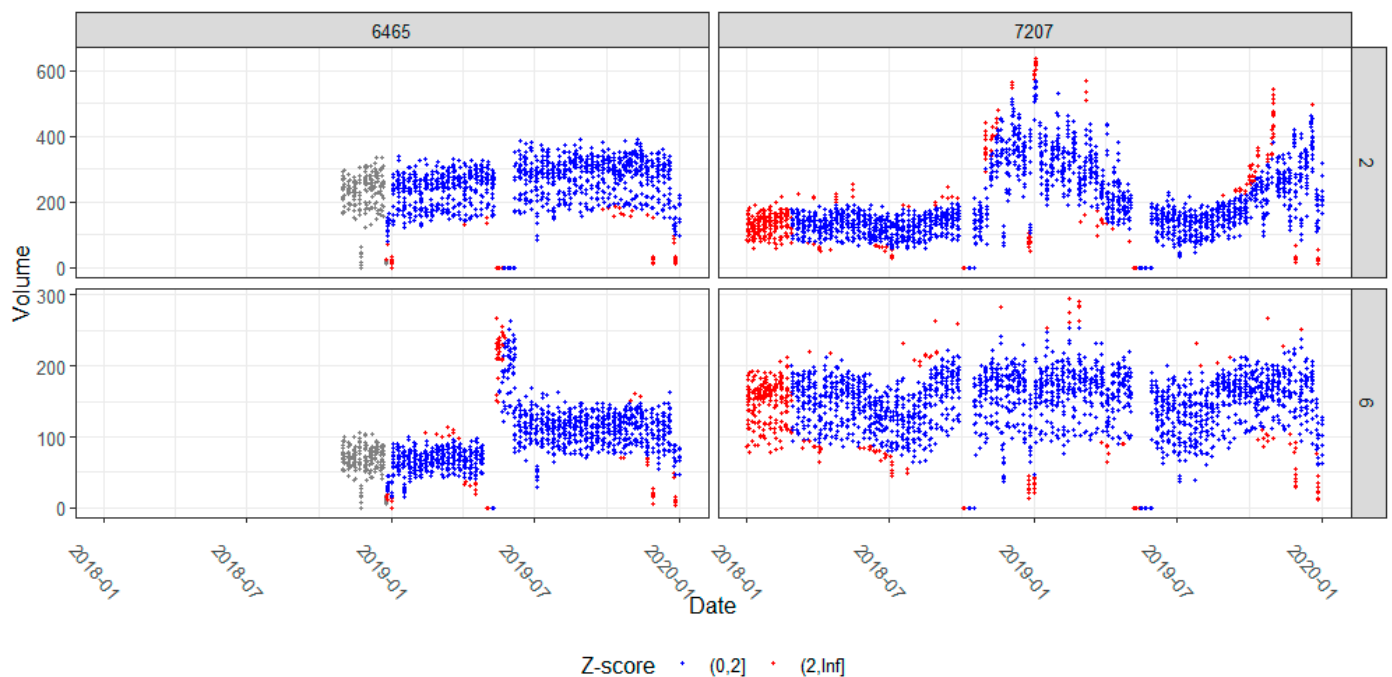


Figure A6. Moving average and standard deviation with z-score calculations (Signals 6465, 7207).

References

- Bullock, D.M.; Clayton, R.; MaSckey, J.; Misgen, S.; Stevens, A.L. Helping Traffic Engineers Manage Data to Make Better Decisions: Automated Traffic Signal Performance Measures. *ITE J.* **2014**, *84*, 33–39.
- Lattimer, C.R.; America, A.N. *Automated Traffic Signals Performance Measures*; Publication FHWA-HOP-20-002; U.S. Department of Transportation, Federal Highway Administration: Washington, DC, USA, 2020.
- Smaglik, E.J.; Sharma, A.; Bullock, D.M.; Sturdevant, J.R.; Duncan, G. Event-Based Data Collection for Generating Actuated Controller Performance Measures. *Transp. Res. Rec. J. Transp. Res. Board* **2007**, *2035*, 97–106. [[CrossRef](#)]
- Day, C.M.; Bullock, D.M.; Li, H.; Remias, S.M.; Hainen, A.M.; Freije, R.S.; Stevens, A.L.; Sturdevant, J.R.; Brennan, T.M. *Performance Measures for Traffic Signal Systems: An Outcome-Oriented Approach*; Purdue University: West Lafayette, IN, USA, 2014. [[CrossRef](#)]
- Wu, X.; Liu, H.X. Using High-Resolution Event-Based Data for Traffic Modeling and Control: An Overview. *Transp. Res. Part C: Emerg. Technol.* **2014**, *42*, 28–43. [[CrossRef](#)]
- Utah Department of Transportation (UDOT). ATSPM Frequently Asked Questions. Available online: <https://udottraffic.utah.gov/ATSPM/FAQs/Display> (accessed on 31 January 2022).
- Georgia Department of Transportation (GDOT). Automated Traffic Signal Performance Measures Component Details. Available online: https://traffic.dot.ga.gov/ATSPM/Images/ATSPM_Component_Details.pdf (accessed on 31 August 2022).
- Day, C.M.; O'Brien, P.; Stevanovic, A.; Hale, D.; Matout, N. *A Methodology and Case Study: Evaluating the Benefits and Costs of Implementing Automated Traffic Signal Performance*; Publication FHWA-HOP-20-003; FHWA and U.S. Department of Transportation, Federal Highway Administration: Washington, DC, USA, 2020.
- Wang, B.; Schultz, G.G.; Macfarlane, G.S.; McCuen, S. Evaluating Signal Systems Using Automated Traffic Signal Performance Measures. *Future Transp.* **2022**, *2*, 659–674. [[CrossRef](#)]
- Schultz, G.G.; Macfarlane, G.S.; Wang, B.; Davis, M.C. *Detecting Traffic Data Anomalies in Longitudinal Signal Performance Measures*; Report No. UT-22.21; Utah Department of Transportation, Research and Innovation: Salt Lake City, UT, USA, 2022.
- Chang, D.K.; Saito, M.; Schultz, G.G.; Eggett, D.L. How Accurate Are Turning Volume Counts Collected by Microwave Sensors? In *International Conference on Transportation and Development 2016*; American Society of Civil Engineers: Reston, VA, USA, 2016; pp. 945–956.
- Georgia Department of Transportation (GDOT). Automated Traffic Signal Performance Measures Reporting Details. Available online: https://traffic.dot.ga.gov/ATSPM/Images/ATSPM_Reporting_Details.pdf (accessed on 31 August 2022).
- Blázquez-García, A.; Conde, A.; Mori, U.; Lozano, J.A. A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–33. [[CrossRef](#)]
- Keogh, E.; Lin, J.; Fu, A. Hot Sax: Efficiently Finding the Most Unusual Time Series Subsequence. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05)*, Houston, TX, USA, 27–30 November 2005; IEEE: Piscataway, NJ, USA, 2005; p. 8.
- Keogh, E.; Lin, J.; Lee, S.H.; Herle, H.V. Finding the Most Unusual Time Series Subsequence: Algorithms and Applications. *Knowl. Inf. Syst.* **2007**, *11*, 1–27. [[CrossRef](#)]

16. Lin, J.; Keogh, E.; Fu, A.; Van Herle, H. Approximations to Magic: Finding Unusual Medical Time Series. In Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05), Dublin, Ireland, 23–24 June 2005; IEEE: Piscataway, NJ, USA, 2005; pp. 329–334.
17. Senin, P.; Lin, J.; Wang, X.; Oates, T.; Gandhi, S.; Boedihardjo, A.P.; Chen, C.; Frankenstein, S. Time Series Anomaly Discovery with Grammar-Based Compression. In Proceedings of the Edbt 2015, Brussels, Belgium, 23–27 March 2015; pp. 481–492.
18. Georgia Department of Transportation (GDOT). SigOpsMetrics. Available online: <http://sigopsmetrics.com/main/> (accessed on 7 July 2022).
19. Chamberlin, R.; Fayyaz, K. *Using ATSPM Data for Traffic Data Analytics*; Report No. UT-19.22; Utah Department of Transportation, Research and Innovation Division: Salt Lake City, UT, USA, 2019.
20. Utah Department of Transportation (UDOT). Continuous Count Station. Available online: <https://data-uplan.opendata.arcgis.com/datasets/uplan::continuous-count-stations> (accessed on 12 May 2022).
21. Federal Highway Administration. Federal Highway Administration Traffic Monitoring Guide. Available online: https://www.fhwa.dot.gov/policyinformation/tmguid/2022_TMG_Final_Report.pdf (accessed on 29 August 2023).
22. Basu, S.; Meckesheimer, M. Automatic Outlier Detection for Time Series: An Application to Sensor Data. *Knowl. Inf. Syst.* **2007**, *11*, 137–154. [[CrossRef](#)]
23. Zhang, Y.; Hamm, N.A.; Meratnia, N.; Stein, A.; Van de Voort, M.; Havinga, P.J. Statistics-based Outlier Detection for Wireless Sensor Networks. *Int. J. Geogr. Inf. Sci.* **2012**, *26*, 1373–1392. [[CrossRef](#)]
24. Brownlee, J. Linear Regression for Machine Learning. Machine Learning Mastery. Available online: <https://machinelearningmastery.com/linear-regression-for-machine-learning/> (accessed on 5 April 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.