Article

Evaluating Signal Systems Using Automated Traffic Signal Performance Measures

Bangyu Wang, Grant G. Schultz *[ID], Gregory S. Macfarlane [ID] and Sabrina McCuen

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.); sabrinamccuen@gmail.com (S.M.)
* Correspondence: gschultz@byu.edu; Tel.: +1-801-422-6332

Abstract: Automated traffic signal performance measures (ATSPMs) are used to collect data concerning the current and historical performance of signalized intersections. However, transportation agencies are not using ATSPM data to the full extent of this "big data" resource, because the volume of information can overwhelm traditional identification and prioritization techniques. This paper presents a method that summarizes multiple dimensions of intersection- and corridor-level performance using ATSPM data and returns information that can be used for prioritization of intersections and corridors for further analysis. The method was developed and applied to analyze three signalized corridors in Utah, consisting of 20 total intersections. Four performance measures were used to develop threshold values for evaluation: platoon ratio, split failures, arrivals on green, and red-light violations. The performance measures were scaled and classified using *k*-means cluster analysis and expert input. The results of this analysis produced a score for each intersection and corridor determined from the average of the four measures, weighted by expert input. The methodology is presented as a prototype that can be developed with more performance measures and more extensive corridors for future studies.

Keywords: ATSPM; big data; *k*-means cluster analysis; performance measures; traffic signal

1. Introduction

Automated traffic signal performance measures (ATSPMs) are important and increasingly widespread tools for evaluating traffic signals. ATSPMs are constructed from high-resolution traffic signal controller data with a one-tenth of a second time resolution that aids traffic engineers and maintenance technicians in identifying hardware faults and improving traffic signal timing, coordination, operations, and maintenance. ATSPMs allow for the analysis of data passively collected 24 hours a day, 7 days a week, improving the accuracy, flexibility, and performance of signal equipment and the system as a whole [1–6].

The majority of existing ATSPM research is focused on the performance of individual movements or intersections and in developing detailed diagnostic tools such as the Purdue coordination diagram (PCD) [7]. However, given the large amount of data involved with ATSPM datasets—each month the Utah ATSPM database receives over 1 TB (equaling 1.47 million hours) of raw signal data drawn from 2040 signals [2]—it is difficult for traffic engineers, technicians, managers, and other operators to identify which signals might require investigation. Current ATSPMs can help diagnose and quantify problems at intersections, but operators still rely on their own experience or reports from the public to identify which intersections to examine with ATSPMs. A scoring or prioritization tool developed using ATSPM data, incorporating multiple time periods and measures and which can be compared across a set of intersections, is desirable. This paper presents an aggregation technique to combine multiple ATSPMs at a single signal and along a corridor into a composite score. This score was constructed from mapping each individual measure included in the ATSPMs onto thresholds determined through a clustering algorithm and

augmented by the literature and expert opinion. The score was applied to construct a high-level intersection and corridor prioritization scheme. The threshold development, scoring, and prioritization used ATSPMs from a set of 20 fully instrumented intersections on three corridors in Utah drawn throughout 2018.

The paper proceeds as follows: A literature review describes how ATSPMs were developed and how researchers and practitioners use ATSPM data to understand and gain insight into individual intersections and corridors. A methodology section describes the Utah Department of Transportation (UDOT) ATSPM dataset and presents a method to convert multiple individual measures into a composite intersection score using thresholds set by cluster analysis and expert review. The application section describes how threshold values for each measure are determined for specific intersections and provides an application of the threshold scores to rank and prioritize a set of intersections on the studied corridors. The discussion section describes the limitations of the research as well as associated opportunities for future research. The paper ends with a conclusion section that summarizes the findings and contributions of the research.

2. Literature Review

Traffic signals utilize various forms of signal detection to optimize operational efficiency by detecting traffic in real time. The idea of using signal detection to generate automated performance measures originated as early as the 1970s; however, the analog detection systems used at the outset were extremely expensive [8,9]. As detection equipment has improved and become more affordable, the array of possible measures generated from such equipment has expanded as well. For example, radar detector technology has more recently been used on full approaches at signalized intersections to detect vehicle arrivals [10]. Further, the ability to connect all of an agency's detectors into a single database has expanded the possibilities for developing rich and detailed performance measures. In addition, Marcianò et al. developed a systematic application for signal design on a road network during evacuation conditions to minimize evacuation times [11].

In 2010, Purdue University completed a National Cooperative Research Program (NCHRP) study that contained extensive documentation and drew multiple examples on the motivation, theory, and application of traffic signal performance measures [12,13]. In 2011, researchers at Purdue University and the Indiana Department of Transportation (INDOT) developed and defined an architecture for a centralized traffic signal management system that can be used on a large geographic scale by both maintenance and technical services staff. The architecture includes a visualization tool called the PCD [14]. Sturdevant et al. defined the enumerations used to encode performance measures events that occur on traffic signal controllers with high-resolution data loggers [15]. That code enumeration became the standard for developing ATSPM tools. In 2012, UDOT started development on ATSPMs with Purdue University and INDOT.

The Federal Highway Administration (FHWA) encourages the use of ATSPMs, because engineers and planners can measure the signal retiming efforts directly from actual performance rather than depending on software modeling or manually collected data. The open-source software development practices used in these projects has fostered collaboration and streamlined the creation of new performance measures [16]. As the number of performance measures increased, the need for engineers to navigate the ATSPM data required the development of aids such as the UDOT ATSPM website (https://udottraffic.utah.gov/ATSPM/ (accessed on 1 July 2022)) [2] and the Georgia Department of Transportation (GDOT) SigOps Metrics website (http://sigopsmetrics.com/main/ (accessed on 1 July 2022)) [17]. A 2020 FHWA publication estimated that UDOT's ATSPM system has saved taxpayers over USD 107 million in 10 years through reduced traffic delay [18].

The ATSPM website helps traffic engineers, technicians, planners, and researchers better understand how individual movements or intersections perform. For example, Day et al. used PCDs to optimize traffic signal offsets by visualizing existing offsets of coordinated signals [7]. In addition, Davis applied real-time ATSPM data to redirect

traffic and adjust signal timing [19]. At the same time, Lavrenz et al. used arrivals on green recorded in ATSPM data to identify how detector maintenance affected vehicle progression [20].

The deluge of information available from individual signals creates an additional problem, which is that engineers and planners still need to evaluate signals one by one. To improve efficiency and to help engineers and planners identify intersections or corridors that may be underperforming and/or operating incorrectly, there is a need to use ATSPMs not only as a diagnostic tool but to develop measures that would allow it to be a prioritization tool by "ranking" intersections according to performance measures, similar to the level of service in the Highway Capacity Manual (HCM) [21].

Day et al. provided an initial attempt at this by grouping interrelated aspects of performance (i.e., communication, detection, safety, capacity allocation, and progression) along a corridor, thus creating a corridor-level score. However, despite this method of corridor analysis, the entire corridor rating system is still in the preliminary stage, and the overall technique has room for improvement. For example, Day et al. only included five consecutive weekdays of data for evaluating the performance. Variation in traffic patterns across days or months might lead to a different prioritization scheme. Second, the overall prioritization score assigned to an intersection was determined by the lowest metric among all categories; this makes it difficult to compare different corridors if they are rated at the same level, but one can use different categories to establish the grade. Third, each performance measure was treated equally, which makes it difficult for engineers to identify priority tasks in which different performances may have varying outcomes [22]. The goal of this research was to consider other aggregation methods with additional performance measures that can be weighted and applied across a longer time period and larger scales.

3. Methodology

Developing a prioritization scheme for intersection performance first requires understanding which performance measures are available at each intersection and then developing a method to classify these measures along a uniform spectrum from "bad" to "good" and various points in between. The classified performance measures for each intersection can then be combined into a single intersection score, and intersections on a corridor can be "averaged" or "weighted" for a composite score as desired. The general workflow proposed for this methodology is described in Figure 1.
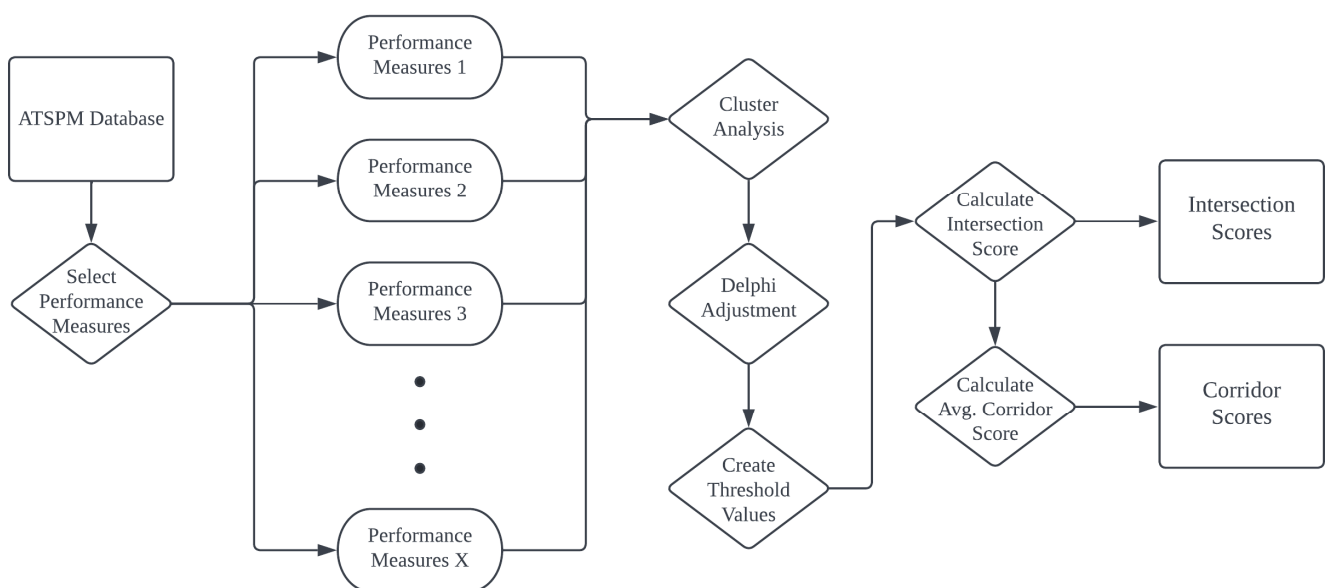

Figure 1. Threshold development and score evaluation workflow.

The following sections describe the intersections and corridors used in this analysis, the details of the UDOT ATSPM database and the performance measures available therein, a technique to classify performance measures using clustering algorithms supplemented by expert opinion, and a technique to combine these classified measures into intersections and corridors.

3.1. Study Data

The performance measure data available at a given signal were determined by the type of detection available at that intersection. The signals and corridors selected for this analysis were based on data availability and included five signalized intersections on 800 North, Orem, UT; five signalized intersections on State Street, Orem, UT; 10 signalized intersections on Fort Union Blvd., Cottonwood Heights, UT, as shown in Figure 2. The signal approaches for evaluation were the through movements on the major street for each signal. The reason for choosing the through movement was because some ATSPMs, such as platoon ratio and split failure, become difficult to calculate correctly for permitted or protected left-turn phases.

The chosen time periods for analysis were from 7:00 a.m. to 9:00 a.m. (AM peak) and 12:00 p.m. to 2:00 p.m. (mid-day) on Tuesdays and Wednesdays in March, July, and October of 2018. Tuesdays and Wednesdays were chosen as they historically have similar traffic patterns. Two days were chosen to increase the amount of data for analysis. This differs from the research conducted by Day et al., who only used two weekdays as opposed to five [22]. The motivation behind this is that days closer to the weekend are anticipated to have different traffic patterns than those closer to the middle of the week. Three separate months were chosen to account for changes in weather and traffic demand. The year 2018 was chosen because it was the most recent full year of data available at the time the research team began aggregating the data (summer of 2019). The PM peak was not selected for analysis because UDOT suggested that the research focus on AM peak and mid-day, theorizing that if the intersection performed poorly during the AM peak and mid-day time period, the PM peak performance will also be poor. Only the through movements for each signal were evaluated to simplify the interpretation of the performance measures.
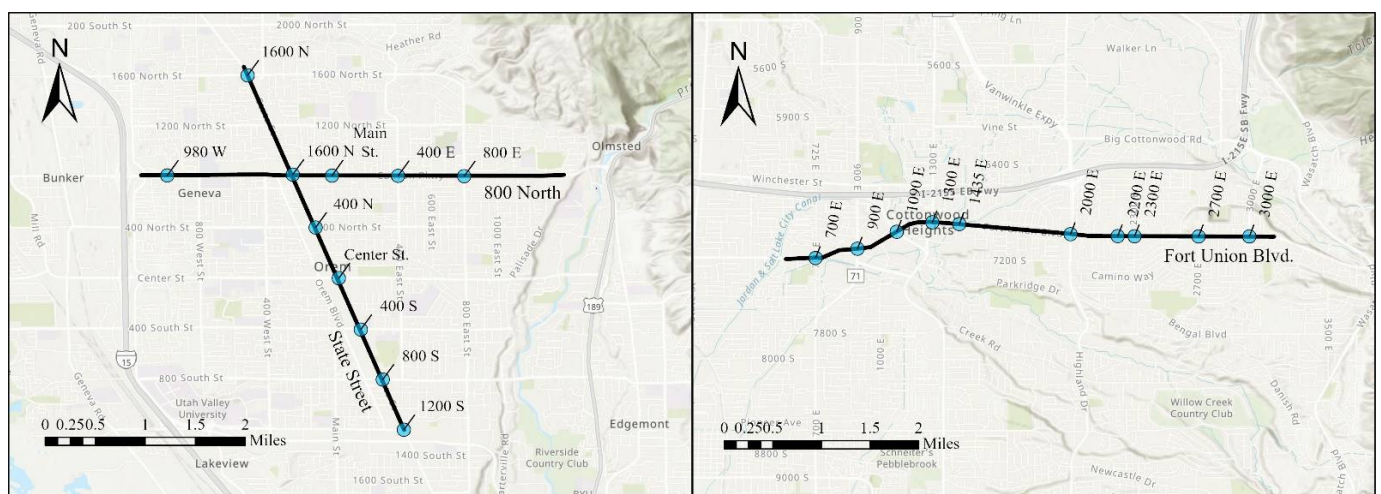
Figure 2. Study corridors.

3.2. UDOT ATSPM

ATSPM data were collected using existing signal infrastructure. In addition to the typical equipment required for a traffic signal, a high-resolution controller, data collection engine, and communication system or reporting engine are required for ATSPM analysis. An operator interface is also required so that the analyst can access the data [23]. Different detection types may be added to an intersection to collect various performance metrics.

A complete list of the performance measures and tools for evaluating performance that UDOT uses in its ATSPM system are listed in Table 1 [2]. After considering which of the measures from Table 1 were the most common based on the current intersection detection scheme in Utah, which of the measures were the most representative of signal performance from the literature, and which of the measures were perceived to be the most important to traffic engineers, technical planners, and state agencies, it was determined to focus the research on four performance measures: platoon ratio, percent arrivals on green, split failures, and red-light violations.

Table 1. UDOT ATSPM Metric and Performance Measures.

Metric/Tool	Definition	Performance Measure(s)
Purdue Phase Termination (PPT)	Evaluating performance measures by plotting the controller's phases and the reason the phase terminated	Force off, gap out, max out, and pedestrian activity
Split Monitor	Analysis generates separate plots for each phase of the controller indicating how much split time is being used for each phase	Programmed split, gap out, max out, force off, and pedestrian activity
Pedestrian Delay	Depicts the delay, in minutes, associated with each pedestrian actuation	Pedestrian actuations and delay time
Preemption Details	Identifies preemption events that might occur at a signal	Preempt request
Purdue Split Failure (PSF)	Calculates the percent of time that the stop bar detectors are occupied during the green phase and during the first 5 s of red	Split failure, green/red occupancy ratio, and percent failure
Yellow and Red Actuations	Plots vehicle arrivals during the yellow and red intervals of an intersection's phasing, where the speed of the vehicle is interpreted to be too fast to stop before entering the intersection	Red time, red clearance, yellow change, and detector activation
Turning Movement Counts	Generates traffic volume for each lane on an approach	Total volume by direction
Approach Volume	Uses advanced detection (generally 300 feet to 500 feet upstream of the stop bar) to count vehicles for the approach	Approach volume and D-factor
Approach Delay	Plot approach delay experienced by vehicles approaching and entering the intersection	Approach delay and approach delay per vehicle
Arrivals on Red	Plots both the volume and percentage of vehicles arriving on red for those phases where data are available	Arrivals on red and percent arrivals on red
Purdue Coordination Diagram (PCD)	Plots vehicle arrivals against the current movement (i.e., green, yellow, and red) and traffic flow in vehicles per hour using the percentage of vehicles that arrive on green and the platoon ratio	Volume per hour; change to green, yellow, and red; arrivals on green; green time; platoon ratio
Approach Speed	This metric tracks the speed of vehicles approaching a signalized intersection for those phases where data are available	Average MPH, 85th percentile, 15th percentile, and posted speed

The platoon ratio is a measure of how effectively an intersection is utilizing the green portion of a cycle, as outlined in Equation (1). The platoon ratio is also a measure of how well the traffic along a corridor is progressing [5]. UDOT places great importance on this measure, because it can quickly display whether a signal is performing well in terms of efficient vehicle throughput. A high platoon ratio signifies good performance, while a low platoon ratio signifies poor performance. Although there is no maximum value for a platoon ratio, any value higher than 1.5 is considered exceptional and any value lower than 0.5 is considered poor.

$$pr_{it} = \frac{PVG_{it}}{g_{it}/C_{it}} \tag{1}$$

where:

pr_{it} = platoon ratio;
PVG_{it} = percentage of vehicles arriving during the effective green;
g_{it} = effective green time;
C_{it} = cycle length.

The percent arrivals on green is a measure of individual phase progression that estimates the proportion of vehicles arriving on a green light versus the proportion that arrive on a red light [24]. Arrivals on green and arrivals on red were identified as performance measures that would be useful for this research. A high number of vehicles arriving on green is preferred to a low number of vehicles arriving on green, because these vehicles experience less delay, while the opposite is true for arrivals on red. To effectively compare the results between different signals, it was determined that these measures should be presented as a percent. Calculating the percent arrivals on green for a signal phase requires an additional data element: the total volume of the movement, as outlined in Equation (2).

$$aog_{it} = \frac{AOG_{it} + AOY_{it}}{Total\ Volume_{it}} \tag{2}$$

where:

aog_{it} = percent of vehicle arrivals on green in a 15 min bin;
AOG_{it} = number of vehicle arrivals on green in a 15 min bin;
AOY_{it} = number of vehicle arrivals on yellow in a 15 min bin;
$Total\ Volume_{it}$ = total number of vehicle arrivals in a 15 min bin.

Split failures measure the number of vehicles that take two or more cycles to execute their movement at an intersection [16]. Said another way, a split has "failed" if vehicles queued when a signal turns green remain in the queue when the signal turns red. As with arrivals on green and arrivals on red, to effectively compare the results between different signals, it was determined that this measure should be presented as a percent, that is, the share of cycles in a 15 min period that end in a split failure. This also required using the total volume of the movement in addition to the number of split failures, as outlined in Equation (3).

$$sf_{it} = \frac{Split\ Failures_{it}}{Total\ Cycles_{it}} \tag{3}$$

where:

sf_{it} = number of vehicles that failed to pass the intersection in each cycle;
$Split\ Failures_{it}$ = number of vehicles that failed to pass the intersection in a 15 min bin;
$Total\ Cycles_{it}$ = number of signal cycles in a 15 min bin.

Although the red-light violations performance measure has inconsistencies when comparing across intersections due to the fact of right turns on red as well as detection latency, it is still possible to determine if signal performance is worsening, staying the same, or improving over time using this performance measure by comparing the same

intersections longitudinally. The red-light violations performance measure was also the only measure related to safety. As such, red-light violations were included in the analysis.

3.3. Threshold Development

After the ATSPM data were collected for each study intersection, it was necessary to classify performance at the intersection. To do this, initial threshold values were determined for each performance measure using cluster analysis supplemented with expert opinion.

A k-means cluster analysis [25] was used in creating the initial threshold values. Clustering, in general, is a nonparametric machine learning technique used to classify data across multiple attributes. The specific k-means algorithm can be used when all attributes x_1, x_2, ... x_n of a particular data point p_i are continuous variables (i.e., there are no categorical or logical values). The k-means algorithm works as follows:

1. Select k random points in n-dimensional space as initial "mean points";
2. Calculate the "distance" between each data point and each mean point;
3. Calculate a new mean point as the average $\overline{x}_1$, $\overline{x}_2$, ..., $\overline{x}_n$ of the points closest to each existing mean;
4. Calculate the mean squared error for points associated with each new mean;
5. Iterate steps 2 through 4 until the change in mean squared error between iterations drops below a specified tolerance level.

The result of this algorithm is a set of "clusters" defining groups of points that are more alike to each other than those points in other clusters, compared along multiple dimensions. One important note is that each attribute x_i must be on effectively the same scale, or variables with wider ranges will exert more influence on the definition of the clusters. It is thus a common practice to rescale all attributes to the (0,1) range, a practice followed in this study. In this project, the k-means process informs a search for threshold values that can effectively distinguish intersections showing different performance characteristics across the four numeric performance measures. The research team created an interactive data visualizer using the Shiny application interface in R to apply the cluster analysis and visually investigate the threshold values [26,27].

The threshold values determined by the k-means cluster were subsequently adjusted by an expert panel. Accessing expert opinion in this way was by conducted by using the Delphi approach to decision making, which is " ... a qualitative, long-range forecasting technique, that elicits, refines, and draws upon the collective opinion and expertise of a panel of experts" [28]. Experts on the panel rely on past experiences relating to the topic being studied as well as on their own knowledge of the subject [29]. There have been many transportation-related studies that have used a Delphi approach to develop scoring systems, rank factors or qualities, and to predict future impacts. For example, Schultz and Jensen developed a scoring system for advanced warning signal systems in Utah [30]. For this study, the expert panel consisted of a team of traffic operations engineers, technicians, and engineering managers at UDOT, referred to as the Technical Advisory Committee (TAC). To aid the expert panel in developing thresholds, the research team developed an interactive graphical tool; the details of this tool are given in Schultz et al. [31]. The expert panel also considered thresholds previously determined for the platoon ratio in the HCM [21].

3.4. Combining Threshold Scores to Intersections and Corridors

The ultimate goal of this analysis was to aggregate the four selected performance measures into a score that can be compared across intersections and corridors. This required aggregation across two dimensions: first, the four threshold values in a given 15 min period needed to be combined into a single value; second, several periods needed to be combined in a way that described the intersection performance in a holistic way. To contain these two dimensions, Jansen provided the multi-attribute utility theory for decision making in steps [32]:

1. Defining performance measures and value-relevant attributes;
2. Evaluating each performance measure separately on each attribute;
3. Assigning relative weights to the performance measures;
4. Aggregating the weights of the attribute and single-attribute evaluations of the performance measures to obtain an overall evaluation of the performance measures;
5. Perform sensitivity analyses and make recommendations.

In the context of this study, a relevant question was the degree to which the four ATSPMs should be weighed against each other. Is platoon ratio twice as important as red-light adherence? Or do engineers have twice as much trust in the fidelity of this measure? Does changing the weights result in a different prioritization outcome? In this study, multiple different weighting schemes were developed and applied to the outcome. The weighted values for each of the performance measures were normalized so that the total of all performance measure weights summed to 1.0. The adjusted weight for each performance measure was calculated as outlined in Equation (4).

$$w'_{sf} = \frac{w_{sf}}{w_{pr} + w_{sf} + w_{aog} + w_{rl}} \tag{4}$$

where:

w'_{sf} = adjusted weight for split failures;

w_{pr} = weight for platoon ratio;
w_{aog} = weight for arrivals on green;
w_{sf} = weight for split failures;
w_{rl} = weight for red-light violations.

The overall score for an intersection in a 15 min period was the dot product of the numeric threshold scores for each period and the normalized weights as outlined in Equation (5).

$$S_{it} = pr_{it}w'_{pr} + sf_{it}w'_{sf} + aog_{it}w'_{aog} + rl_{it}w'_{rl} \tag{5}$$

where:

S_{it} = combined score for intersection i in period t;
pr_{it}, sf_{it}, aog_{it}, rl_{it} = threshold score for each individual measure included in the ATSPMs for intersection i in period t.

The second dimension of aggregation—combining S_i across the several time periods—could be performed numerous ways. Day et al. used the lowest period score, $S_i = \min(S_{it})$, as the score for the intersection under the logic that agencies should identify poor performing outliers [22]. A more forgiving or representative measure of an intersection's performance might be the arithmetic mean, $S_i = \sum_{i=1}^{I} S_{it}/I$, or some percentile of the distribution of S_{it}. Possibilities for this measure and the consequences of this decision are explored in the following section. A similar logic applied to aggregating the individual intersection scores to corridors.

4. Application

This section applies the methodologies described in Section 3 to identify the threshold values for each performance measure, incorporate threshold values into the scoring system at the intersection level, and aggregate intersection scoring to determine an overall corridor scoring.

4.1. Threshold Values

Threshold values to evaluate and compare different intersections were developed for each performance measure. These threshold values were derived from various sources including standardized manuals, the TAC, and the *k*-means cluster analysis. The threshold values for each performance measure corresponded with a score for the intersection per-

taining to the specific performance measure with scores ranging from 1 (low) to 5 (high). The values 1 to 5 were used based on the previous research that used five categories of scores ranging from A to E [22]. Rather than using letter-based scoring, it was determined that numerical scores would be easier to use in subsequent calculations and would not introduce confusion with HCM level of service measures. Because there were five scoring categories in this research, it was decided to use a *k*-means cluster analysis that divided the data into five groups; if performance measure data were unavailable for any reason in a 15 min bin, the algorithm was unable to determine a cluster, and those data were not assigned to a cluster. Table 2 summarizes the determined threshold values for each performance measure.

**Table 2.** Performance Measure Threshold Values.

Threshold for Level Score	Platoon Ratio	Percent Arrivals on Green	Percent Split Failure	Red-Light Violations
5 (Exceptional)	>1.50	>0.80	≤0.05	0
4 (Favorable)	1.15 ≤ 1.50	0.60 ≤ 0.80	0.05 ≤ 0.30	1.0–2.0
3 (Average)	0.85 ≤ 1.15	0.40 ≤ 0.60	0.30 ≤ 0.50	3.0–4.0
2 (Unfavorable)	0.50 ≤ 0.85	0.20 ≤ 0.40	0.50 ≤ 0.95	5.0–9.0
1 (Poor)	≤0.50	≤0.20	>0.95	≥10

Figure 3 displays histograms of the performance measures with the assigned cluster for each intersection in each 15 min aggregation bin for all corridors and intersections used in the analysis. Figure 3 also visually displays the threshold measures summarized in Table 2 for additional context. Figure 3a depicts the platoon ratio distribution and the assigned threshold values of 0.5, 0.85, 1.15, and 1.5. These values were chosen to separate the platoon ratio into five categories and were modeled after the thresholds found in the HCM [21]; the cluster analysis clearly grouped intersections with high platoon ratios into Cluster 2.

Figure 3b depicts the percent of arrivals on green distribution. These threshold values were set at 0.2, 0.4, 0.6, and 0.8, with the evenly spaced distribution reinforced by the rough boundary between Clusters 2 and 3 on the one hand and Cluster 4 on the other.

Figure 3c depicts the distribution of the percentage of split failures per 15 min; the threshold values were set at 0.05, 0.30, 0.50, and 0.95. Expert input from the TAC was used to decide to put signals with no split failures or all split failures in their own categories, given the high percentage of signals that fell in these two categories. The intermediate threshold values were informed by the left edge of Cluster 1.

Figure 3d depicts the distribution of red-light violations and the threshold values developed by the research team and shows this distribution for all corridors combined. All intersections with no red-light violations were noted by the highest score, all intersections with 1–2 red-light violations were second best, 3–4 violations corresponded with the third level, 5–9 violations corresponded with the fourth level, and any value 10 or greater was in the lowest scoring category. One important item to note is that the cluster analysis seemed to have placed intersections with multiple red-light violations into Cluster 1, which is the same cluster with a high percentage of split fails; this is a strong indication that these two measures are collinear, or that a large number of red-light violations occur during split fail conditions when motorists attempt to make it through the previous phase.
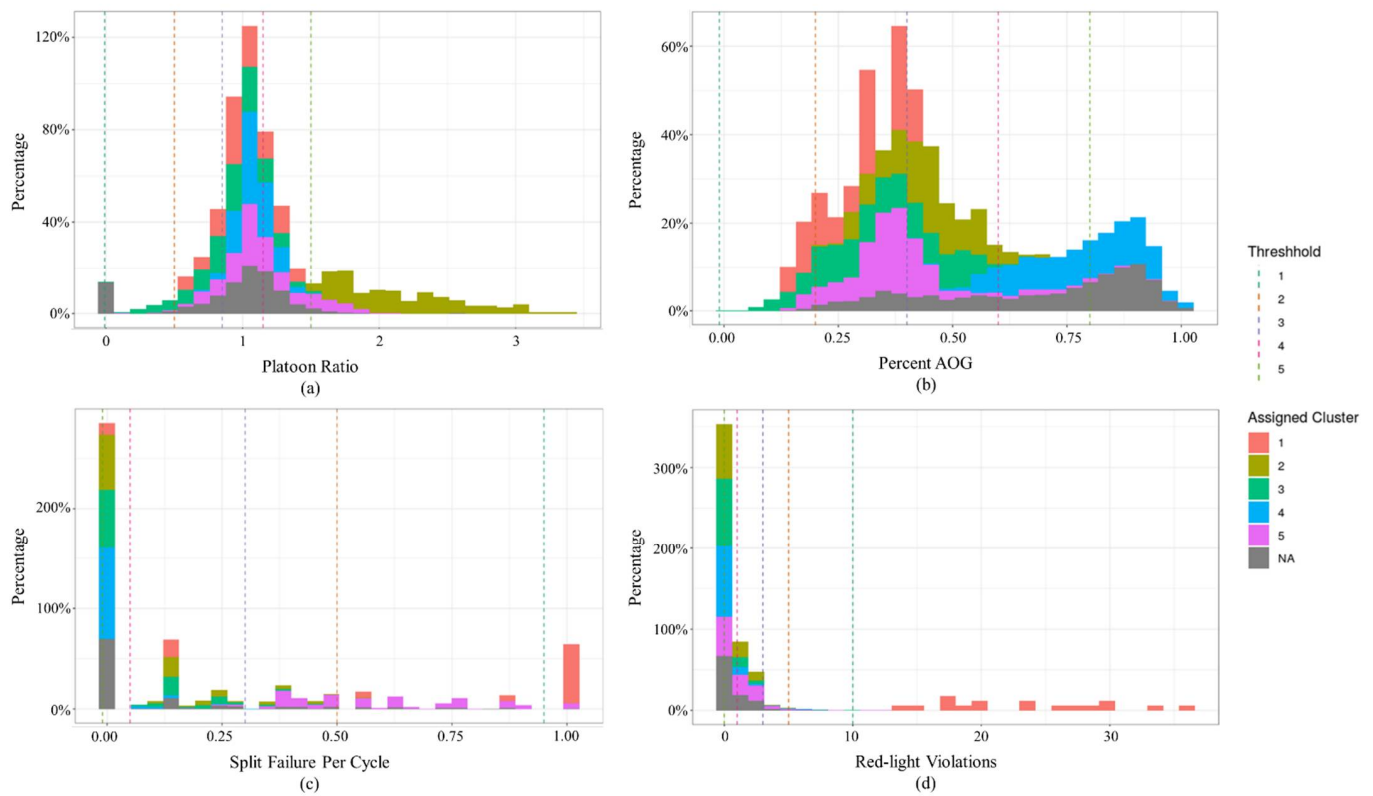
Figure 3. Clustering and threshold values: (**a**) platoon ratio: (**b**) percent arrivals on green (AOG); (**c**) split failure per cycle; (**d**) red-light violations.

4.2. Application to Intersections

To determine the effect of different weighting schemes on the intersection total score, the research team performed a sensitivity analysis. Figure 4 shows the empirical cumulative distribution of total score assigned to all 15 min bins for different weighting schemes for two intersections two times per day: Fort Union Blvd./1090 East and Fort Union Blvd./1300 East. The orange line displays a scheme where the weight for the platoon ratio was twice the value of the weights for the remaining measures. The green line displays a scheme where the weight for the red-light violations was twice the value of the weights for the remaining measures. The blue line displays a scheme where the weights for both the platoon ratio and red-light violations were twice the value of the weights for the arrivals on green and split failures. The purple line displays a scheme where the weight for the split failures is twice the value of the weights for the remaining measures.

The research team chose the two different intersections displayed in Figure 4 as representative examples of these plots for all intersections. The plot for the 1090 East intersection showed a wider distribution of scores, while the plot for the 1300 East intersection showed a narrower distribution of scores. The plot for the intersection of Fort Union Blvd./1090 East shows that the different weighting schemes only change the results slightly and do so in a relatively uniform manner. That is, the rank-ordering of the scores for the intersections would not change substantially were a different weighting scheme to be chosen. The plot for the intersection of Fort Union Blvd./1300 East shows that when the platoon ratio weight was higher than the weights for the other measures, the overall score was lower. The research team decided to adopt this scheme—the orange line in Figure 4—because it provides the most conservative scoring for the intersections while still representing the variation in scores. Because the signals behaved differently for the two time periods collected, 7:00 a.m. to 9:00 a.m. (AM peak) and 12:00 p.m. to 2:00 p.m. (mid-day), the results were separated by these time periods.
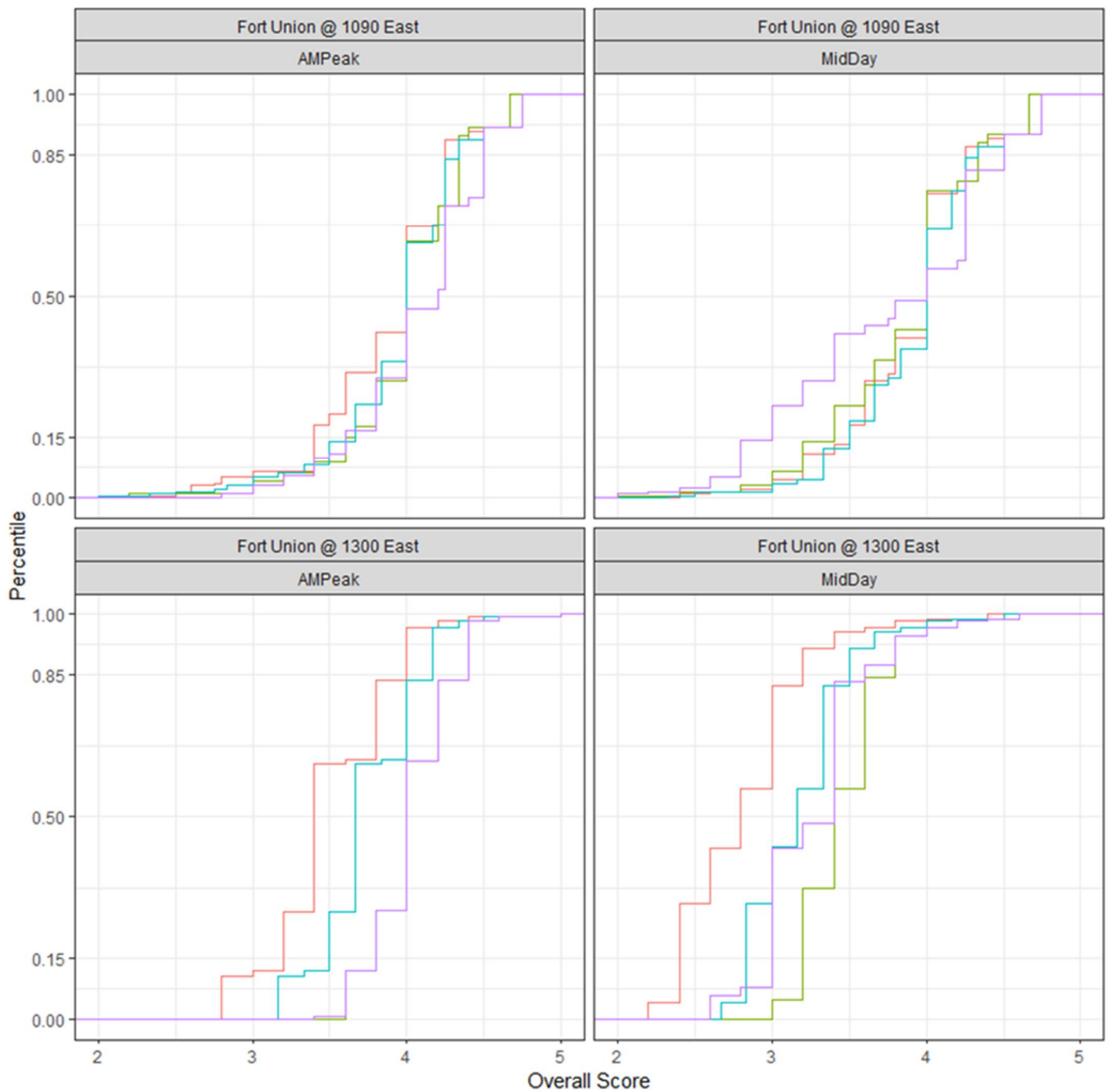
Figure 4. Empirical cumulative distribution for the different weighting schemes.

Figure 5 shows the overall score for the signalized intersections on Fort Union Blvd., 800 North, and State Street in both the AM peak and mid-day periods. The graphs were developed using the weighting scheme where the platoon ratio weight was two times higher than the other measures. The overall score was calculated for every 15 min bin in the selected time period by averaging the two major through movement signal phases for the appropriate bin. The overall scoring schemes, shown in Figure 5 on the *y*-axis, include the minimum, 15th percentile, median, mean, 85th percentile, and maximum. The overall scores were sorted from smallest to largest based on the mean value of the distribution of scores for that intersection.

Each colored line represents an individual intersection. The trend of each line shows that the intersections were scored and ranked differently at the various percentiles. This shows that only using the worst score to rank the intersection may not be the most repre-

sentative of the data. It is important to note that a steeper slope means less consistency in intersection performance and vice versa. As an example, for 400 North at State Street in the AM peak, the line was mostly horizontal, meaning that the signal performance was very consistent at the intersection. State Street at 800 North in the AM peak, on the other hand, was relatively steep meaning the performance changed considerably. In this case, it would be advisable for engineers, technicians, and planners to check the performance of that intersection.
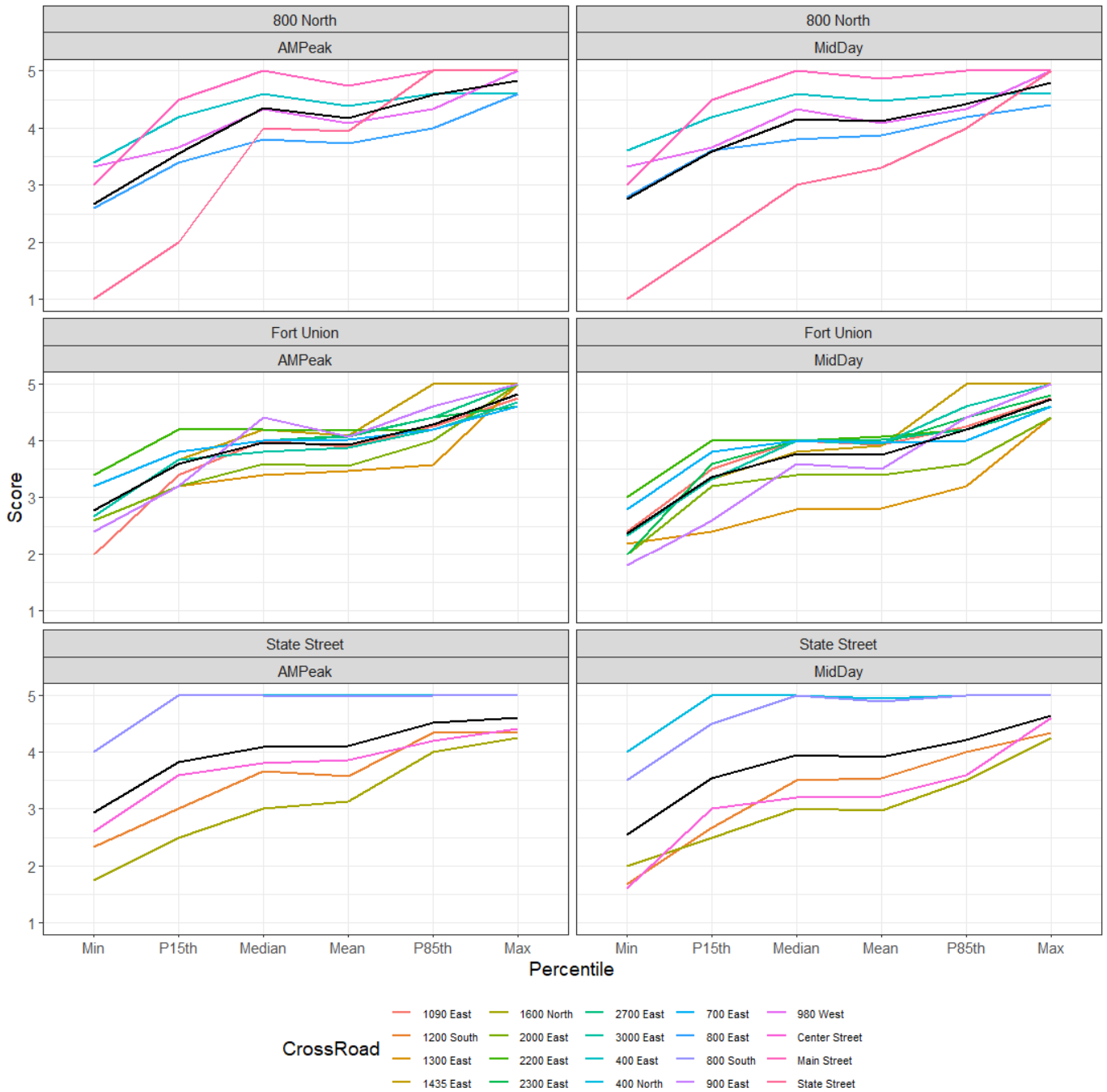
Figure 5. Overall score for the intersection and corridor levels.

4.3. Aggregation to Corridors

To calculate the overall score at the corridor level, each of the corridor scoring methods was performed for each intersection scoring method. The corridor scoring schemes were the same as the intersection including the minimum, 15th percentile, median, mean, 85th percentile, and maximum scoring methods. The averages for all the intersections along the corridor were used because it resulted in the ranking being the best to represent the overall performance for each corridor. The bold black lines in Figure 5 illustrate the corridor score at the different percentiles.

By comparing the corridor performance between the AM and mid-day peaks, several observations can be made. First, the Fort Union Blvd. and State Street corridors performed better in the AM peak than mid-day, while 800 North had similar results for both time periods. By comparing the different corridors, the intersections and corridor score showed that Fort Union Blvd. had the most consistent performance along the corridor. Using Day et al.'s methodology, all corridors would score less than three, but using the median or mean to represent the corridor performance, the results become more significant and the score closer to four [16]. The minimum score outlined in Figure 5 demonstrates that the Day et al. method may be overestimating poor performance along the corridors, which may not be as useful in assisting traffic engineers and planners in prioritizing intersections and corridors that may need to be addressed. Based on the mean score of the corridors, the prioritization scheme shows that Fort Union Blvd. had the lowest performance. Therefore, engineers, technicians, and planners should focus on this corridor first.

5. Discussion

The limitations of this research include using a small dataset with 20 signals and two time periods. The researchers made every attempt to ensure the selected data were representative of the expected conditions; however, there may be situations that were missed. The scoring and prioritization method represented the traffic condition only for the selected dataset; thus, it was hard to identify if the intersection's performance deteriorated or improved over time. Throughout the course of the research, it was determined that there were a limited number of ATSPMs that had sufficient data in the datasets for evaluation. Although the selected ATSPMs arrivals on green and platoon ratio may be somewhat correlated, thus emphasizing platoon progression in the research, it was determined that these ATSPMs would both be included in the analysis due to the limited number of ATSPMs available.

As mentioned in the methodology section, a *k*-means cluster analysis [25] was used in creating the initial threshold values. In this case, the four chosen ATSPMs were all continuously defined and easily normalized, which makes *k*-means a naturally applicable choice. Other clustering methods, including random forest [33] and hierarchical clustering [34], might be considered in future research, particularly when including potentially discontinuous ATSPMs.

A Delphi panel adjusted the thresholds resulting from the *k*-means analysis informed by their individual experience and findings in the professional literature. Another candidate technique could have been a more rigorously defined multicriteria decision-making (MCDM) approach, which has been successfully used in some transportation-related problems [35]. The number of potential criteria in this application was limited enough to make this unnecessary but might be attempted in future research.

For future research, evaluating signal performance longitudinally over time by comparing these scores would enable agencies to determine whether the performance of an intersection is worsening, staying the same, or improving. More corridors can be investigated to adjust the method to improve statistical significance. Finally, more performance measures could be included and brought into the evaluation method to make it more comprehensive and reduce the potential correlation between individual ATSPMs.

6. Conclusions

This study used a sample of high-resolution data in the UDOT ATSPM database to provide a method for evaluating the quality of signal operations and determining the measures that can be aggregated into higher level metrics at both the intersection- and corridor-level. This method will help to introduce and provide context for a large amount of performance measure data.

The aggregated tables from UDOT ATSPM database were combined using the R analysis tool. Charts and plots were then produced from the combined data utilizing the data visualizer created by the research team. This application provided a method for producing scores for each intersection using the performance measures of platoon ratio, split failures, arrivals on green, and red-light violations. These scores were then utilized to provide an overall corridor score for each corridor analyzed. The overall scores were visualized in Figure 5 for 20 for the intersections along the three corridors.

This paper provides a system-level method to evaluate the quality of signal operations throughout the UDOT ATSPM database. The methodology developed also includes much needed sensitivity analyses for performance weighting, and threshold development for ranking intersections and corridors. The ranking system developed for both intersection- and corridor-level performance could be used for evaluating all signalized intersections in Utah. Although this research did not take every signal in the state into account, the framework has been set-up; thus, this is now possible. This research used cluster analysis to assist in scoring the intersections that had not been previously in other research and provides a new method for applying ATSPMs to a traffic signal system. Because of this research, traffic engineers, technicians, and planners can better understand how intersections perform so that every effort can be made to prioritize the signals that need to be adjusted to improve traffic operations This research is an initial attempt to help ATSPM users understand and apply ATSPMs at a higher level. The scoring system presented here will assist agencies in maintaining and updating their planning process at the transportation network level rather than making changes at the individual traffic signal level.

References

1. Bullock, D.M.; Clayton, R.; Mackey, 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.
2. Utah Department of Transportation (UDOT). ATSPM Frequently Asked Questions. Available online: https://udottraffic.utah.gov/ATSPM/FAQs/Display (accessed on 1 July 2022).
3. Lattimer, C.R.; America, A.N. *Automated Traffic Signals Performance Measures*; Publication FWHA-HOP-20-002; U.S. Department of Transportation; Federal Highway Administration: Washington, DC, USA, 2020.
4. 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]
5. 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]
6. 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**, *24*, 28–43. [CrossRef]
7. Day, C.M.; Haseman, R.; Premachandra, H.; Brennan, T.M., Jr.; Wasson, J.S.; Sturdevant, J.R.; Bullock, D.M. Evaluation of Arterial Signal Coordination: Methodologies for Visualizing High-Resolution Event Data and Measuring Travel Time. *Transp. Res. Rec.* **2010**, *2192*, 37–49. [CrossRef]
8. Blase, J.H. Computer Aids to Large-Scale Traffic Signal Maintenance. *Traffic Eng. Control* **1979**, *20*, 341–347.
9. Parapar, S.M. Computerized Reporting for Traffic Signal Maintenance. *Traffic Eng.* **1973**, *43*, 47–51.
10. Day, C.M.; Helt, R.; Sines, D.; Emtenan, A.M.T. Leveraging Sensor-Based Vehicle Position and Speed Information in Traffic Signal Control with Existing Infrastructure. In Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 4049–4054.
11. Marcianò, F.A.; Musolino, G.; Vitetta, A. Signal Setting Design on a Road Network: Application of a System of Models in Evacuation Conditions. *WIT Trans. Inf. Commun. Technol.* **2010**, *43*, 443–454.
12. Day, C.M.; Bullock, D.M. *Arterial Performance Measures Volume 1: Performance Based Management of Arterial Traffic Signal Systems. Final Report, NCHRP 3-79A*; Transportation Research Board: Washington, DC, USA, 2011.
13. Haseman, R.; Day, C.M.; Bullock, D.M.; Using Performance Measures to Improve Signal System Performance. Indiana Local Technical Assistance Program (LTAP) Technical Reports. Paper 11. Available online: https://docs.lib.purdue.edu/inltaptechs/11 (accessed on 28 May 2020).
14. Day, C.M.; Brennan, T.M.; Ernst, J.M.; Sturdevant, J.R.; Bullock, D.M. *Procurement Procedures and Specifications for Performance Measure Capable Traffic Infrastructure Data Collection Systems*; Publication FHWA/IN/JTRP-2011/18, Joint Transportation Research Program; Indiana Department of Transportation and Purdue University: West Lafayette, IN, USA, 2012.
15. Sturdevant, J.R.; Overman, T.; Raamot, E.; Deer, R.; Miller, D.; Bullock, D.M.; Day, C.M.; Brennan, T.M.; Li, H.; Hainen, A.; et al. *Indiana Traffic Signal Hi Resolution Data Logger Enumerations*; Indiana Department of Transportation and Purdue University: West Lafayette, IN, USA, 2012.
16. Federal Highway Administration (FHWA). Automated Traffic Signal Performance Measures (ATSPM). Available online: https://ops.fhwa.dot.gov/publications/fhwahop20002/index.htm (accessed on 1 July 2022).
17. Georgia Department of Transportation (GDOT). SigOps Metrics. Available online: http://sigopsmetrics.com/main/ (accessed on 1 July 2022).
18. 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.
19. Davis, A. Focused Operations: Measuring Arterial Performance Using Automated Traffic Signal Performance Measures. I-95 Corridor Coalition. Available online: https://i95coalition.org/wp-content/uploads/2017/05/I-95CC-Signal_PM-Coalition_Slides-ver11-Compiled.pdf?x70560 (accessed on 8 February 2022).
20. Lavrenz, S.; Sturdevant, J.; Bullock, D.M. Strategic Methods for Modernizing Traffic Signal Maintenance Management and Quantifying the Impact of Maintenance Activities. *J. Infrastruct. Syst.* **2017**, *23*, 05017004. [CrossRef]
21. Highway Capacity Manual (HCM). *Transportation Research Board*; National Research Council: Washington, DC, USA, 2010; p. 1207.
22. Day, C.M.; Li, H.; Sturdevant, J.R.; Bullock, D.M. Data-driven Ranking of Coordinated Traffic Signal Systems for Maintenance and Retiming. *Transp. Res. Rec.* **2018**, *2672*, 167–178. [CrossRef]
23. Smaglik, E.J.; Bullock, D.M.; Sharma, A. Pilot Study on Real-time Calculation of Arrival Type for Assessment of Arterial Performance. *J. Transp. Eng.* **2007**, *133*, 415–422. [CrossRef]
24. Gettman, D.; Folk, E.; Curtis, E.; Ormand, K.K.D.; Mayer, M.; Flanigan, E. *Measures of Effectiveness and Validation Guidance for Adaptive Signal Control Technologies*; Publication FHWA-HOP-13-03; U.S. Department of Transportation; Federal Highway Administration: Washington, DC, USA, 2013.
25. Wong, M.A.; Hartigan, J.A. Algorithm AS 136: A K-means Clustering Algorithm. Journal of the Royal Statistical Society. *Ser. C Appl. Stat.* **1979**, *28*, 100–108.
26. Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. Shiny: Web Application Framework for R. R Package Version 1.4. 0. 2019. Available online: https://CRAN.R-project.org/package=shiny (accessed on 23 June 2022).

27. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available online: https://www.R-project.org (accessed on 27 February 2022).

28. Gupta, U.G.; Clarke, R.E. Theory and Applications of the Delphi Technique: A Bibliography (1975–1994). *Technol. Forecast. Soc. Change* **1996**, *53*, 185–211. [CrossRef]

29. Cavalli-Sforza, V.; Ortolano, L. Delphi Forecasts of Land Use: Transportation Interactions. *J. Transp. Eng.* **1984**, *110*, 324–339. [CrossRef]

30. Schultz, G.G.; Jensen, A.P. Evaluation and Installation Guidelines for Advance Warning Signal Systems. *IET Intell. Transp. Syst.* **2009**, *3*, 247–256. [CrossRef]

31. Schultz, G.G.; Macfarlane, G.S.; Wang, B.; McCuen, S. *Evaluating the Quality of Signal Operations Using Signal Performance Measures*; Report No. UT-20.08; UDOT Research & Innovation Division: Salt Lake City, UT, USA, 2020.

32. Jansen, S.J. The Multi-Attribute Utility Method. In *The Measurement and Analysis of Housing Preference and Choice*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 101–125.

33. Belgiu, M.; Drăguţ, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]

34. Sasirekha, K.; Baby, P. Agglomerative Hierarchical Clustering Algorithm-a. *Int. J. Sci. Res. Publ.* **2013**, *83*, 83.

35. Musolino, G.; Rindone, C.; Vitetta, A. Evaluation in Transport Planning: A Comparison between Data Envelopment Analysis and Multi-Criteria Decision Making Methods. In Proceedings of the 31st Annual European Simulation and Modelling Conference, Lisbon, Portugal, 25–27 October 2017.