OUTLIER MINING BASED TRAFFIC INCIDENT DETECTION USING BIG DATA ANALYTICS

Pranamesh Chakraborty
PhD Student, Department of Civil, Construction and Environmental Engineering
Iowa State University, Ames 50014, Iowa
Tel:(515)-357-9076; Email: pranames@iastate.edu

Jacob Robert Hess
Freshmen, Department of Mechanical Engineering
Iowa State University, Ames 50014, Iowa
Tel:(612)-275-5354; Email: jrhess@iastate.edu

Anuj Sharma (Corresponding Author)
Associate Professor, Department of Civil, Construction and Environmental Engineering
Iowa State University, Ames 50014, Iowa
Tel:(765)-430-0023; Email: anujs@iastate.edu

Skylar Knickerbocker
Research Engineer, Institute of Transportation
Iowa State University, Ames 50014, Iowa
Tel:(515)-294-2238; Email: sknick@iastate.edu

Word Count: 5804 words of text + 6 figures/tables$\times$250 words (each) = 7304 words

Submission Date: November 14, 2016

ABSTRACT

Early detection of incidents is one of the key step to reduce incident-related congestion. With the increasing usage of GPS based navigation, promising data-scalable crowdsourced probe data is now available which can provide near-real time traffic speed information. This study utilizes such extensive historical datasets (approximately 500 GB) to gain useful insights on the normal traffic pattern of each segment. The insights come in the form of speed threshold for different time of the day and days of week for each segment. Thereafter, the anomalous traffic behaviour are classified as incidents. The dynamic thresholds developed for each segment simplifies the calibration steps that is often required when applying a model to a different dataset. Also, in this study, two alternatives of the traditional Standard Normal Deviate (SND) based incident detection algorithm are tested. The proposed algorithms can handle the masking effect of SND method where the outliers inflate the mean and standard deviation values and result in lower threshold values and in turn, lower detection rate. The high detection rate (94-97%) obtained by these algorithms compared to the SND method (83%) shows the efficacy of the models. Although higher false alarm rate (FAR) are observed for these models, but their values (4 false alarms/day) are quite lower than the acceptable FAR (10 false alarms/day) reported in previous literature (*1*).

Keywords: traffic incident detection, outlier mining, big data

INTRODUCTION

Traffic congestion has been defined by US Department of Transportation (USDOT) as "one of the single largest threats" to the economic prosperity of the nation (*2*). The cost of congestion in 2014 was calculated to be $160 billion for the top 471 urban areas of United States. This included 6.9 billion hours of wasted time and 3.1 billion gallons of wasted fuel (*3*). A major contributor to this congestion are traffic incidents. Schrank and Lomax (*4*) showed that implementation of improved incident management procedures in 272 out of 439 urban areas resulted in reduction of 143.3 million hours of incident-related congestion and $3.06 million in 2007.

Early detection of incident is one of the key step for improved incident management. Hence, significant efforts have been devoted in the past for development of accurate and fast automatic incident detection (AID) algorithms. Researchers have used pattern recognition algorithms, outlier mining methods, artificial neural networks, fuzzy set theory, genetic algorithms, wavelet transformation and other machine learning methods for traffic incident detection (*5*). However, a nationwide survey on deployment of AID algorithms in Traffic Management Centers (TMC) showed that 90% of survey respondents feel that the current AID algorithms are inappropriate for use either in present (70%) or in future (20%) (*1*). The two major reasons behind disabling of AID algorithms in TMCs are difficulty in algorithm calibrations and unacceptable false alarm rates when deployed in large scale. Thus, there is a significant need to revisit the AID algorithms and develop an algorithm which can address these major issues.

Automation of calibration process of AID algorithms can resolve one of the major hindrances of deployment of AID algorithms in TMCs. However, as pointed out by Castro-Neto et al. (*6*), development of an incident dataset with accurate start and end time of incidents is time-consuming and often requires manual investigation. This makes the calibration of AID algorithms even more difficult for TMC personnels. In this paper, the main goal is to develop an AID algorithm that can extract maximum information from the traffic data to generate the normal travel pattern of each segment. Thereafter, the anomalous behaviour can be classified as incidents and hence sidestep the need for algorithm training with incident dataset. In the era of big data, traffic parameters (e.g. speed, volume, etc.) are stored for each and every segment across 24×7 hours and 365 days. For example, in Iowa state, probe vehicle data of 23,000 segments spread across the entire state are archived every day in one minute interval. This results in generation of approximately five gigabytes of daily traffic data, which in turn produce around two terabytes of traffic data in an annual basis. And, for traffic incident detection, traffic data needs to be collected and processed continuously for each segment. With the cheap data storage technologies now available, it makes more sense to store the entire dataset and use it to gain useful insights on the performance of the road network. These insights can help in developing more efficient AID algorithms. Thus, incident detection turns out to be an important field in the area of transportation which can get direct benefits from the big data analytics.

This paper proposes detecting incidents considering them as outliers or anomalies in the continuous traffic data stream. The next section gives an overview of the past research done on AID algorithms and performance measures used to evaluate the algorithm. The third section provides description of the data used in this paper. Section 4 gives the details of the research methodology followed by the detailed results in Section 5. The final section provides a summary of the paper and outlines the future work.

BACKGROUND & RELATED WORK

Performance Measures

The following performance measures, used most commonly in AID studies (*5*) are also used in this paper.

Detection Rate (*DR*) is defined as the ratio of the total number of incidents detected to the total number of incidents actually occurred, given by Equation 1.

$$DR = \frac{\text{Total number of detected incidents}}{\text{Total number of actual incidents}} \times 100\% \tag{1}$$

False Alarm Rate (*FAR*) is defined as the ratio of the total number of false alarms to the total number of algorithm applications, given by Equation 2. The total number of algorithm applications implies the number of times the algorithm is applied during a given period. For example, if the traffic state is checked once every minute (as done in this paper) and five of them are reported as false alarms, then the *FAR* is 8.3%. It should be noted here that the *FAR* is computed over the entire system rather than the average for each segment and hence is a function of the number of road segments analysed.

$$FAR = \frac{\text{Total number of false alarm cases}}{\text{Total number of algorithm applications}} \times 100\% \tag{2}$$

In addition to the *FAR* given in Equation 2, *FAR* in terms of number of false alarms per day is also reported in this study. This is because as per the survey results of Williams and Guin (*1*), TMC personnels' perspective of definition of *FAR* is different from the traditional definition (given by Equation 2) and the maximum acceptable false-alarm rate is on an average ten false alarms per day.

Mean Time to Detect (*MTTD*) is defined as the ratio of the total time elapsed between detecting incidents to the number of incidents detected, given by Equation 3.

$$MTTD = \frac{\text{Total time used to detect incidents}}{\text{Total number of incidents detected}} \times 100\% \tag{3}$$

Related Work

Significant research efforts have been devoted since the last five decades for development of efficient AID algorithms. AID algorithms can be divided into two basic categories based on the type of traffic data collection: roadway-based algorithms and probe-based algorithms (*5*). Roadway-based algorithms use fixed detector data installed at specific points in the road segments whereas probe-based algorithms use probe vehicle data for detecting incidents. In this paper, probe vehicle data has been used for traffic incident detection. Hence, a detailed literature review on probe-based AID algorithms has been presented next. Summary of roadway-based AID algorithms can be found in Parkany and Xie (*5*) study.

AID algorithms can be further classified into two broad categories based on the methodology used to detect incidents (a) algorithms that compare the present traffic parameter values with the historical values observed under similar conditions (e.g., time of day, day of week) and (b) present traffic parameters are compared with the immediate previous *N* intervals to trigger an incident alarm. In either of these cases, the feature vectors are compared with a predetermined threshold for incident detection. Also, a persistence test is usually performed to confirm the preliminary detected incidents before triggering incident alarm (*7*). This is done to eliminate the false alarms caused due to sudden spurious traffic fluctuations. Probe-based AID algorithms that use historical traffic parameter values for incident detection are presented next followed by a discussion of the algorithms that utilizes sudden change of immediate traffic values during an incident to trigger an alarm.

Arterial traffic incident detection algorithms developed in the ADVANCE operational test by Sethi et al. (*8*) and Sermons and Koppelman (*9*) used discriminant analysis techniques for incident detection. Linear relationship of predictor variables were developed to distinguish incident conditions from incident-free ones. These algorithms use travel time and speed of a particular link and its immediate upstream link to trigger incident alarms. Balke et al. (*10*) considered traffic incidents as outliers in data stream and used the principle of standard normal deviates (SND) to indicate the confidence intervals for incident-free travel time conditions. Historical average travel time were computed for each link by time of day (in 15-min intervals) and day of week to denote normal travel conditions.

Algorithms were also developed to detect traffic incidents comparing the present conditions with the immdeiate past. For example, Parkany and Bernstein (*11*) algorithms were based on the principle that temporal and spatial discrepancies of travel time and headways and frequent lane switch maneuvers can be observed when traffic switches from incident-free to incident conditions. Waterloo algorithm proposed by Hellinga and Knapp (*12*) were based on the assumption that the travel time are log-normally distributed, rather than normally distributed as assumed by Balke et al. (*10*). And, the confidence limit in Waterloo algorithm were based on the travel time observed in previous N intervals instead of using the historical average travel time for the required interval of the day. The bivariate analysis model (BEAM) developed by Li and McDonald (*13*) use average travel time of probe vehicles and differences in travel time between adjacent time intervals to distinguish an incident condition from incident-free one. Zhu et al. (*14*) used speed differences between adjacent sections and adjacent time intervals as feature vector for mining incidents as outliers from non-incident conditions. Recently, Li et al. (*7*) extended the SND algorithm by introducing two modifications: (a) weighted average and standard deviates of the traffic parameter values are used based on the traffic flow, and (b) in order to eliminate the false alarms caused by acute fluctuations of SND values, if the coefficient of variation of the traffic parameter is below a predetermined threshold, the the SND value of the previous time interval is used to replace the SND of the current time interval.

In this paper, traffic incidents are considered as anomalies/outliers in continuous traffic data stream and are detected by comparing them with the historical averages. The basic reason behind adopting this technique is that it will allow to utilize the massive historical dataset to gain useful insights of the traffic pattern of each link thereby helping in detecting incidents. With the increasing usage of navigation applications installed in mobile phones, promising data-scalable crowdsourced probe data is now available which provide near real-time traffic speed information. Li et al. (*15*) used such crowdsourced probe data provided by INRIX (*16*) to identify shockwave boundaries while Park and Haghani (*17*) developed models for detecting secondary incidents utilizing same data source. So, it makes sense to also develop AID algorithms utilizing such extensive data source. In traditional AID algorithms, sample data are used for developing the models hoping that the model could be generalised and applied to every other segment. However, this makes the calibration and fine tuning of the model parameters even more difficult. Utilising data of each and every segment will help making the parameters dynamic and can be continuously trained from new incoming data.

Also, in this study, alternatives of the traditional SND algorithm are applied to detect outliers. A basic disadvantage of SND algorithm is that it is impacted heavily by the presence of outliers. So, in this study, two other outlier detection methods are applied and compared with the traditional SND algorithm to find out the efficacy of the proposed methods.

DESCRIPTION OF DATA

Probe vehicle speed data from 1st April, 2016 to 7th July, 2016 of Des Moines region, Iowa is used in this study. The study region comprises of the Interstates 35, 80 and 235 and is shown in Figure 1. The Des Moines region is the busiest region on Iowa roadways experiencing significant amount of congestion and incidents throughout the year. The details of traffic volume variation for each of these roads are shown later in Section 5. Besides this, video cameras are also installed in this region which helps in verification of incident data. Two hundred and fifty-four segments are located in this region covering 164 miles. The length of the segments vary from 0.2 miles to 1.5 miles.

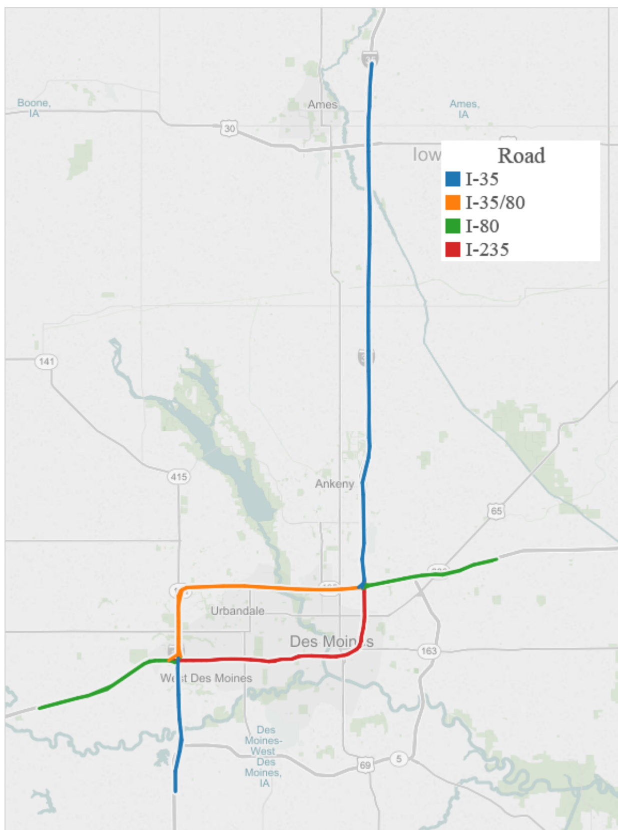

FIGURE 1 Location of the segments used

Speed data from 1st April to 30th June are used as the primary dataset to compute the threshold speed values for each segment (detailed procedure to obtain threshold speed values are given in Section 4). Approximately 500 GB of traffic data are analysed for determination of threshold speed values. Remaining dataset of 1st July to 7th July, 2016 is used as the validation set to verify incidents reported by proposed algorithm and incident dataset maintained by the local TMC. The incident database maintained by TMC records the location of incident, start and end time of incident and type of incident (e.g., accident, stalled vehicle, slow traffic, etc.). Apart from the incident database, each of the incidents detected by the proposed algorithm are also manually verified by video cameras installed in the study region. A total of 70 lane-blocking incidents causing disruption to traffic were reported in the study region during the one-week validation period. However, the incident database also has records of incidents which didn't caused any disruption to traffic (54% of the total incidents). Since AID algorithms relying solely on speed data cannot detect incidents which had no significant effect on traffic speed, these incidents were excluded from the incident dataset.

The probe-based speed data used in this study is provided by INRIX (*16*) with a reporting frequency of one-minute. Details of this cloud-based speed data can be found in Li et al. (*15*) study. Reliability of the speed data is dependent on the number of probe vehicles available, which in turn depends on the flow volume. Confidence score and C-value are two parameters provided by INRIX to indicate the data quality of the reported average speed of a particular segment. Confidence score of 30 indicates that the data is generated exclusively from real-time data sources while a score of 10 indicates that historical data is used to report the speed. When a mix of the two sources are used, a score of 20 is provided. The C-value is used to provide an additional degree of confidence to the real-time data. The C-value is reported only when the confidence score is equal to 30. In this paper, the reported speed data is considered to be reliable real-time speed data and used for further analysis only when the confidence score is equal to 30 and C-value is also greater than 30 (as suggested by Haghani et al. (*18*)).

METHODOLOGY

Traffic incidents have been often considered as outliers/anomalies in the continuous data stream. The common strategy applied to detect the anomalous traffic behaviour is using the SND algorithm. However, as stated earlier in Section 2, the SND algorithm is impacted heavily by the presence of outliers or incidents. This issue can be resolved by removing all incident-related data points before calculating the average and the standard deviation values. However, this will lead to application of semi-supervised learning instead of unsupervised learning which requires information of all incidents occurring in the study region over the entire study period. This is difficult because development of an accurate incident dataset is very time-consuming and cumbersome manual investigation is required in most cases (*6*). Particularly, information of the accurate start time and end time of incidents are often hard to get which makes the calibration process very difficult. However, alternate outlier analyses methods exist which can cater the affect of outliers for calculating the threshold. Detailed description of such outlier methods and their modifications to make them work as AID algorithms are discussed next.

Univariate Outlier Analysis

Univariate outlier analysis is the simplest method of detecting outliers where the output depends only on a single variable. Fundamentally, univariate outlier detection procedures involve selecting a reference value x_0 and a measure of variation ζ from the data sequence x_k (*19*). Then, data point x_k is said to be an outlier if it satisfies Equation 4,

$$|x_k - x_0| > t\zeta \tag{4}$$

where, t is the threshold parameter.

Different univariate outlier detection procedures exist depending on the choice of x_0 and ζ. The three most common techniques are given below:

1. *SND* rule: $x_0 = \bar{x}, \zeta = \widehat{\sigma}$;
2. *MAD* (Maximum Absolute Deviation) rule: $x_0 = x', \zeta = S$;
3. *IQD* (Inter-quartile distance) rule: $x_0 = x', \zeta = Q$;

where, $\bar{x}$ is the sample mean, x' is the sample median, $\widehat{\sigma}$ is the sample standard deviation, S is the *MAD* scale estimator and Q is the *IQD*.

The *MAD* and *IQD* are defined in Equations 5 and 6 respectively.

$$S = \frac{Median\{|x_k - x_0|\}}{0.6745} \tag{5}$$

$$Q = \frac{x_{\langle 0.75 \rangle} - x_{\langle 0.25 \rangle}}{1.35} \tag{6}$$

where, $x_{<0.75>}$ is the upper quartile i.e., 75$^{\text{th}}$ percentile and $x_{<0.25>}$ is the lower quartile i.e., 25$^{\text{th}}$ percentile. The factors 0.6745 and 1.35 are used to make the S and Q unbiased estimators of the standard deviation, ($\widehat{\sigma}$) (*19*).

Each of these methods have its own advantages and disadvantages. The basic disadvantage of the *SND* rule is that the x_0 and ζ parameters are influenced heavily by the presence of outliers, the phenomenon known as masking. This results in making the ζ parameter (i.e. $\widehat{\sigma}$ in this case) very high and thus making it hard to detect outliers. Or in other words, this results in having a low detection rate (*DR*) of incidents. The *MAD* and *IQD* methods do not suffer from this problem. However, both these methods suffer from a different phenomenon, namely swamping. In swamping, the ζ value becomes zero if more than 50% of the data values x_k have same value. This will lead to declaring any value different from the median as an outlier, irrespective of its distance from the median. For example, if the median speed value is 60 mph, the current speed value of 59 mph will also be declared as outlier since the ζ is zero in case of swamping. This will result in a very high value of *FAR* in the case of traffic incident detection. However, for AID algorithms, we can take advantage of the fact that an alarm should be triggered only in cases when congestion has occurred. As per FHWA guidelines, congested conditions is said to occur in freeways when the speed is less than 45 mph (*20*). So, typically alarm should not be triggered when speed is higher than 45 mph. Thus, it eliminates the false alarms which can trigger in swamping cases, where the ζ parameter is zero and the median speed value is quite high (greater than 45 mph).

Normal traffic condition for each segment varies depending on the time of day, day of week, weather conditions, etc. For univariate outlier analysis, the x_0 and ζ values are computed from historical speed data of each segment for each 15-min period for each day of the week (similar to Balke et al. (*10*) study). These are denoted by $x_{0,s}^{d,p}$ and $\zeta_{0,s}^{d,p}$ where, s denotes the segment, d denotes day of the week (e.g. Monday, Tuesday, etc.) and p denotes time period of the day divided in 15 minutes interval (e.g. 12:00 PM to 12:15 PM, 12:15 PM to 12:30 PM, etc.). Thus, for the SND edit rule, the x_0 and ζ are denoted as $\bar{x}_{0,s}^{d,p}$ and $\widehat{\sigma}_{0,s}^{d,p}$ respectively and can be determined as given in Equations 7 and 8 respectively.

$$\bar{x}_{0,s}^{d,p} = \frac{\sum_{\forall k} x_{k,s}^{d,p}}{\sum_{\forall k} k} \tag{7}$$

$$\widehat{\sigma}_{0,s}^{d,p} = \frac{1}{\sum_{\forall k} k} \sum_{\forall k \in (d,p,s)} \left(x_{k,s}^{d,p} - \bar{x}_{0,s}^{d,p} \right)^2 \tag{8}$$

Similarly, the x_0 and ζ for *MAD* and *IQD* methods are also calculated. In this paper, Apache Pig Latin is used for computation of these parameters for each segment from the respective historical data of 1st April, 2016 to 30th July, 2016. This required processing of approximately 500 GB of data which is not possible to process via traditional single CPU machines. For this reason, Pig Latin is used. It is a high level Map-Reduce (MR) language to run MR jobs on Hadoop cluster.

The next parameter to be determined for univariate outlier analysis is the threshold parameter, t. Usually, the threshold parameter is determined based on cross validation set, which in this case will be speed data observed during incidents. Extensive research has been done in the past for determination of threshold parameter from cross validation set e.g., $F1$ score, etc. However, determination of threshold parameter from cross validation data will mean that in order to apply proposed methodology to a new site, incident data will also be required for that site along with the traffic speed data. However, as discussed earlier, it is difficult to get incident dataset with accurate start and end time of incidents. Moreover, every segment in the proposed methodology has been treated separately and x_0 and ζ values for each segment are determined independently. However, it is never possible to expect each segment experiencing incidents which can be used to determine threshold parameter for it. For these reasons, the threshold values commonly used for outlier detection for the above mentioned three methods (*19*) are also used in this paper. The threshold values used for each of these three methods are given in Table 1. The threshold speed values obtained for each 15-min time period over all weekdays for each segment were used to trigger incident alarm. Alarm is triggered when input speed value is lower than the computed threshold speed value for consecutive three minutes for the same segment. This is done to reduce the alarms triggered due to sudden noise in incoming speed data.

TABLE 1 Threshold values used for outlier detection by each method

Method	Threshold value used (t)
SND	3
MAD	3
IQD	2

RESULTS

Figure 2 shows the variation of threshold speed values of a typical segment for Thursday. The regular congestion during AM peak hours (7 AM to 9 AM) and PM peak hours (4 PM to 6 PM) resulted in low threhsold speed values for those time period. Also, a nearby workzone scheduled during the night hours affected threshold speed value for night time (9 PM to 6 AM). The figure also shows that the *SND* method being more suspectible to outliers gives lower threshold values compared to the values obtained using the other two methods (i.e, *IQD* and *MAD* methods).

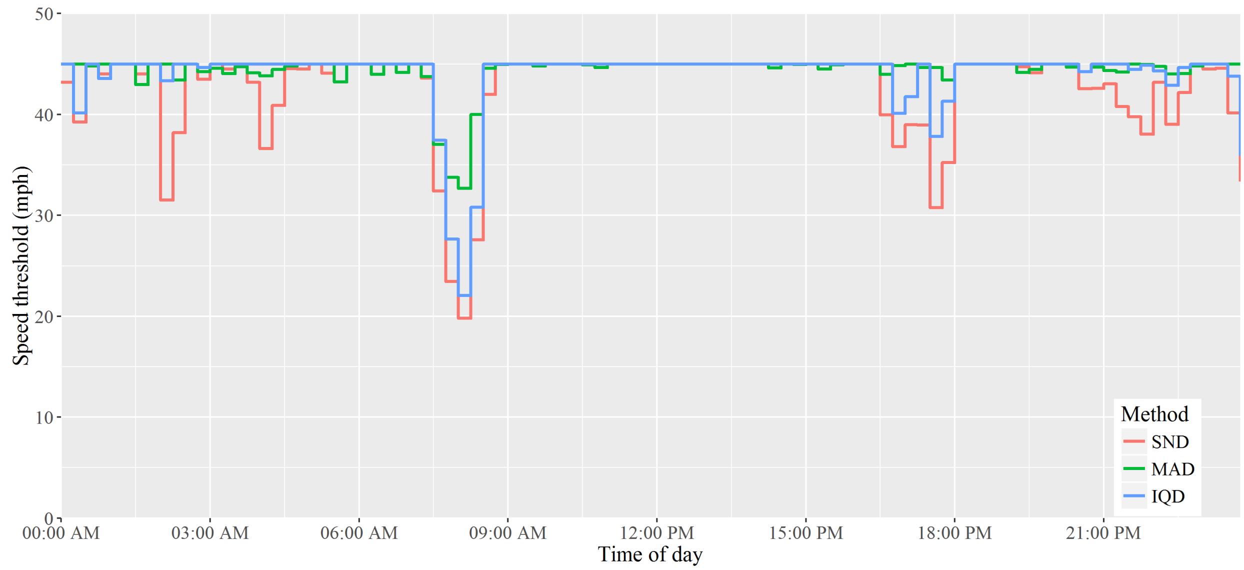

FIGURE 2 Speed Threshold for a typical segment

Figure 3 shows the average speed threshold variation for all the segments in the study region. It should be noted that threshold calculation is done for each segment individually and used for incident detection. However, for brevity, the average threshold variation for each road is shown here. Similar to Figure 2, threshold speed computed by *SND* method are lowest while that obtained by *MAD* method are highest. However, the *MAD* and *IQD* threshold values are often quite close to each other.

I-235 caters the heaviest traffic among all the four interstates covered in this study. An average daily traffic (ADT) of 109,472 vehicles was reported in I-235 during the study period. The downtown traffic of Des Moines produce heavy congestion during the weekdays peak hours in I-235 and I-35/80. Figure 4 shows the average hourly variation of traffic volume of the study region. The heavy traffic in I-235 and I-35/80 resulted in low threshold speed values (as shown in Figure 3) during the peak hours for the same. And low traffic volume in I-35 and I-80 resulted in speed threshold of 45 mph (which is taken as the speed threshold to detect congestion) for most of the time for *IQD* and *MAD* methods.

The threshold speed values obtained for each segment from the historical dataset is used for detecting incidents. The *DR*, *FAR* and $MTTD$ values obtained for each of the three methods are given in Table 2. Table 2 shows that the *IQD* and *MAD* methods achieve *DR* significantly higher compared to the conventional *SND* method. Even though the *FAR* is lowest in *SND*, however, the *FAR* obtained for all the three methods are quite lower than the acceptable false alarm rate stated in Williams and Guin (*1*) study, which is equal to ten false alarms per day. The $MTTD$ obtained from *MAD* is lowest while that obtained from *SND* method is highest. In this conetxt, it should be noted that the $MTTD$ obtained from each of the three algorithms are quite higher from those reported in previous literature (in order of two to seven minutes). (*21*). However, the study of Adu-Gyamfi et al. (*22*) showed that there is an average latency of eight minutes for INRIX freeway data. Also, a persistence test of three minutes is adopted in this study before triggering incident alarm. Taking these factors into consideration, the $MTTD$ values obtained can be said to be satisfactory.

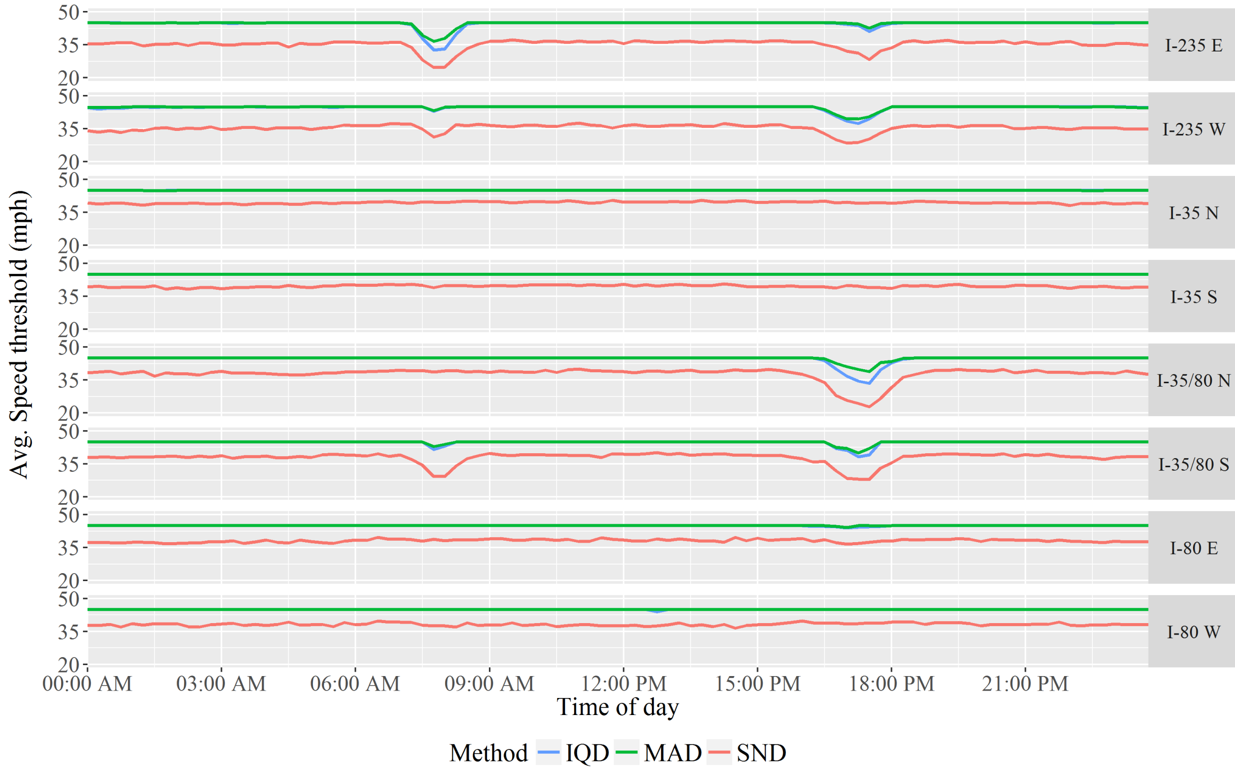

FIGURE 3 Average speed threshold variation for all segments

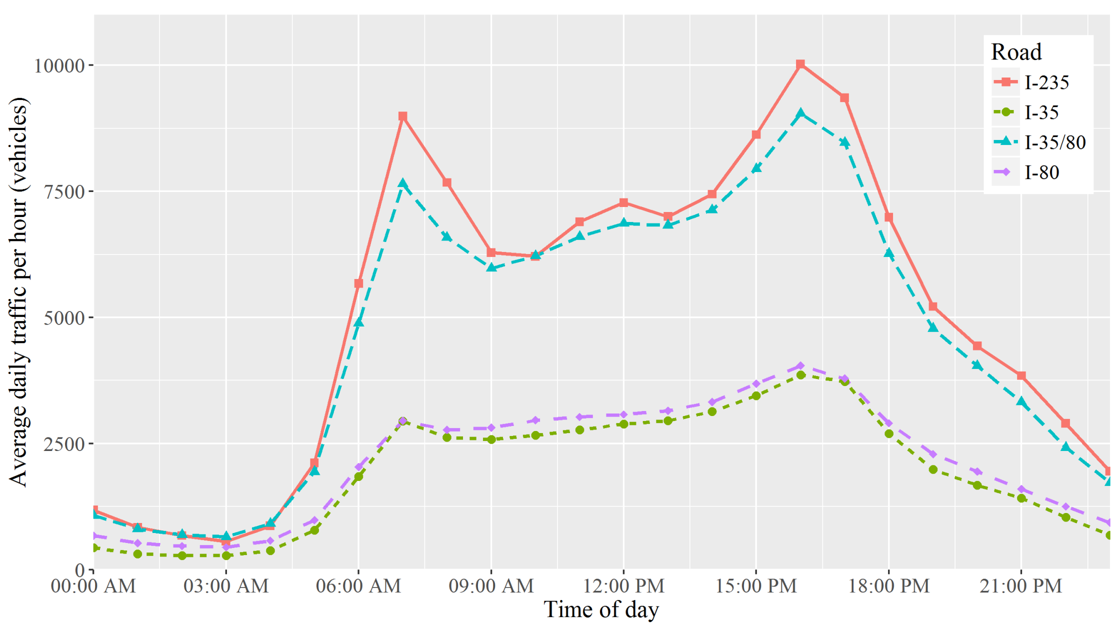

FIGURE 4 Hourly variation of traffic volume in the study region

CONCLUSIONS

In the big data era, traffic parameters are stored continuously for all freeways thereby resulting in generation of massive datasets. This paper uses Apache Pig Latin, a high level map-reduce lan-

TABLE 2 Validation results of the proposed algorithms

Method	DR (%)	FAR (%) [# of false alarms/day]	MTTD (mins)
IQD	97.1	4.84 [4.1]	12.4
MAD	94.3	6.56 [4.0]	10.1
SND	82.9	0.62 [1.0]	13.2

guage to analyse the extensive historical dataset (approximately 500 GB in this case) and obtain useful information about the performance of each road segment separately. This useful information comes in the form of threshold speed values for each segment over the time of the day and different days of the week. These threshold values are used to develop AID algorithms which treat traffic incidents as outliers or anomalies in the data stream. Two other variations of the traditional *SND* based AID algorithms are developed and tested in this study to cater the masking effect of *SND*. To sidestep the need of an incident database for calibrating AID algorithms which is often very time consuming, this study uses threshold values generally adopted for univariate outlier analysis. However, based on availability of incident data, these algorithms can be trained to determine the best threshold values. Nonetheless, the high detection rate (94-97%) and considerably low *FAR* (4 false alarms per day) achieved by the proposed algorithms show the efficacy of the methods used and the future prospects of using more efficient anomaly detection techniques for traffic incident detection. In future, weather data can also be included as a variable impacting traffic conditions and multivariate outlier detection can be applied for improved incident detection. Sensitivity analyses of 15-min aggregation level and 3-min persistence test can also be done. Combining the benefits of big data analytics and advanced anomaly detection algorithms in future can help in developing efficient AID algorithms with lesser calibration issues, low false alarm rates and hence wider applicability.

REFERENCES

[1] Williams, B. M. and A. Guin. Traffic management center use of incident detection algorithms: Findings of a nationwide survey. *IEEE Transactions on Intelligent Transportation Systems*. Vol. 8, No. 2, 2007. pp. 351–358.

[2] Owens, N., A. Armstrong, P. Sullivan, C. Mitchell, D. Newton, R. Brewster, and T. Trego. *Traffic Incident Management Handbook*. FHWA-HOP-10-013, FHWA, U.S. Department of Transportation, 2010.

[3] Schrank, D., B. Eisele, T. Lomax, and J. Bak. *2015 urban mobility scorecard*. Texas A&M Transportation Institute and INRIX, 2015.

[4] Schrank, D. L. and T. J. Lomax. *The 2007 urban mobility report*. Texas Transportation Institute, The Texas A&M University System, 2007.

[5] Parkany, E. and C. Xie. *A complete review of incident detection algorithms & their deployment: what works and what doesn't*. The New England Transportation Consortium, 2005.

[6] Castro-Neto, M., L. Han, Y.-S. Jeong, and M. Jeong. Toward Training-Free Automatic Detection of Freeway Incidents: Simple Algorithm with One Parameter. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2278*. Transportation Research Board of the National Academies, Washington, D.C., 2012. pp. 42–49.

[7] Li, X., W. H. Lam, and M. L. Tam. New automatic incident detection algorithm based on traffic data collected for journey time estimation. *Journal of Transportation Engineering*. Vol. 139, No. 8, 2013. pp. 840–847.

[8] Sethi, V., N. Bhandari, F. S. Koppelman, and J. L. Schofer. Arterial incident detection using fixed detector and probe vehicle data. *Transportation Research Part C: Emerging Technologies*. Vol. 3, No. 2, 1995. pp. 99–112.

[9] Sermons, M. W. and F. S. Koppelman. Use of vehicle positioning data for arterial incident detection. *Transportation Research Part C: Emerging Technologies*. Vol. 4, No. 2, 1996. pp. 87–96.

[10] Balke, K., C. Dudek, and C. Mountain. Using probe-measured travel times to detect major freeway incidents in Houston, Texas. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1554*. Transportation Research Board of the National Academies, Washington, D.C., 1996. pp. 213–220.

[11] Parkany, E. and D. Bernstein. Design of incident detection algorithms using vehicle-to-roadside communication sensors. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1494*. Transportation Research Board of the National Academies, Washington, D.C., 1995. pp. 67–74.

[12] Hellinga, B. and G. Knapp. Automatic vehicle identification technology-based freeway incident detection. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1727*. Transportation Research Board of the National Academies, Washington, D.C., 2000. pp. 142–153.

[13] Li, Y. and M. McDonald. Motorway incident detection using probe vehicles. In *Proceedings of the Institution of Civil Engineers-Transport*. Thomas Telford Ltd, 2005. Vol. 158, No. 1. pp. 11–15.

[14] Zhu, T., J. Wang, and W. Lv. Outlier mining based automatic incident detection on urban arterial road. In *Proceedings of the 6th International Conference on Mobile Technology, Application & Systems, Vol. 29*. ACM, 2009.

[15] Li, H., S. M. Remias, C. M. Day, M. M. Mekker, J. R. Sturdevant, and D. M. Bullock. Shock Wave Boundary Identification Using Cloud-Based Probe Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2526*. Transportation Research Board of the National Academies, Washington, D.C., 2015. pp. 51–60.

[16] *INRIX*. `http://inrix.com/`, accessed July 20, 2016.

[17] Park, H. and A. Haghani. Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C: Emerging Technologies*, 2015.

[18] Haghani, A., M. Hamedi, and K. F. Sadabadi. *I-95 Corridor coalition vehicle probe project: Validation of INRIX data*. I-95 Corridor Coalition, 2009.

[19] Pearson, R. K.. *Mining imperfect data: Dealing with contamination and incomplete records*. SIAM, 2005.

[20] Systematics, C.. *Traffic congestion and reliability: Trends and advanced strategies for congestion mitigation*. Vol. 6. Federal Highway Administration, 2005.

[21] Gakis, E., D. Kehagias, and D. Tzovaras. Mining traffic data for road incidents detection. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2014. pp. 930–935.

[22] Adu-Gyamfi, Y., A. Sharma, S. Knickerbocker, N. Hawkins, and M. Jackson. Reliability of Probe Speed Data for Detecting Congestion Trends. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 2015. pp. 2243–2249.