THE INTEGRATED ANALYSIS OF ACCIDENT REPORTS AND TRAFFIC FLOW DATA SETS WITH EARLY TRAFFIC DISRUPTION DETECTION AND SEGMENTATION

Artur Grigorev
University of Technology Sydney, Ultimo, NSW 2007, Australia
Corresponding Author Artur.Grigorev@student.uts.edu.au

Adriana-Simona Mihăiţă
University of Technology Sydney, Ultimo, NSW 2007, Australia
adriana-simona.mihaita@uts.edu.au

Khaled Saleh
University of Technology Sydney, Ultimo, NSW 2007, Australia
khaled.aboufarw@uts.edu.au

Fang Chen
University of Technology Sydney, Ultimo, NSW 2007, Australia
fang.chen@uts.edu.au

Word Count: 6581 words + 1 table(s) $\times$ 250 = 6831 words

Submission Date: 01 August 2022

ABSTRACT

Traffic accidents are often miss-reported with regards to either the exact location or the start and end time of the disruption due to several external factors - communication delays or misreporting, delays in accident clearance and non reporting of exact number of lanes, etc. Misleading information can lead to wrong decisions being made and can also affect the accuracy of any data-driven model that is responsible to predict either the severity or the disruption length. Several studies so far are using the reported incident data logs as truth ground which may wrongly affect any model build on top of this. To address these issues, our paper presents a novel machine learning framework that can be used for the early detection and prediction of incident durations. Firstly, we start by mapping and fusing several data sets related to reported traffic incidents (flows, speed, incident locations and incident details). Secondly, we propose several mathematical metrics (among which the Wasserstein and the Chebyshev metrics) are fed into an early detection and disruption segmentation algorithm which allows us to choose the best performing metric for future model training. Thirdly, we proposed a modelling interpretation of the traffic speed distributions based on the winning metric to correctly identify between single versus cascade incidents. Last we train and predict using various machine learning models and show that by using our enhanced modelling approach we can reduce the RMSE by almost 41% - 45.6% to as compared to the traditional case of using only historical incident logs.

Keywords: traffic accident, incident duration prediction, machine learning, traffic management

INTRODUCTION

The number of vehicles has been substantially increasing during the past decades, which currently leads to an increase in the number of traffic accidents (*1*). The National Highway Traffic Safety Administration (NHTSA) reported more than 5 million traffic accidents happening in the United States during year 2013 (*2*). Traffic Managements Agencies usually rely on Traffic Incident Management Systems (TIMS) to collect data on traffic accidents, including information on various accident, traffic state and environmental conditions. Accurately predicting the total duration of an incident shortly after it is being finished, will help in improving the effectiveness of accident response by providing important information to decide the required resources to be allocated (response team size, equipment, traffic control measures) (*3*). Traffic accident is a rare event with stochastic nature. The effect of the accident can be observed as an anomalous state in the time series of traffic flow (*4*).

Challenges: The traffic accident analysis may be a challenging task due to incorrect or incomplete accident reports, including the set and the quality of the accident characteristics that have been reported. Accident reports can contain user-input errors related to the accident duration such as: 1) an approximate reporting of accident's start and end time 2) reporting of the accident start time could have been done after the incident finished in reality 3) a 'placeholder' accident duration reporting. In our previous research (*5, 6*) we found that timeline-related errors are present in accident reports across three different data sets from both Australia and the United States of America, which could be also the case with other data sets from around the world, due to multiple human and technical factors that can arise. To forecast the accident impact it is crucial to have an accurate and correct data regarding the observed disruption timeline. We emphasise that disruptions observed in a recorded traffic state can be automatically segmented and associated with a reported accident at the same time and location as when the accident occurred, which allows to eliminate user-input errors from reports and improve the accident duration prediction performance in many traffic management centres around the world. To help address this issue, in our paper we propose various methods for a correct traffic disruption segmentation, the method for an association between vehicle detector stations and accident reports.

Another important challenge is that many incident data sets around the world are private and not shared for public investigation; for those open data sets, there are several missing information fields, or even worse, incomplete information regarding the traffic conditions in the vicinity of the accidents. Even often publish crash data sets are limited in size as well and contain a very small number of records. This represents a tight constraint when testing one framework over multiple countries with different traffic rules and regulations. For our studies we have oriented our attention towards two big open data sets - CTADS (Countrywise traffic accident data set) which contains 1.5 million accident reports and the Caltrans Performance Measurement System (PeMS) which provides data on traffic flow, traffic occupancy and traffic speed across California. Despite both being extensive data sets, vehicle detector station readings from PeMS are not associated with traffic accident reports from CTADS either by time, location or coverage area. The lack of such association makes it impossible to analyse the relation between accidents and their effects on traffic flow and speed. To address this challenge, in our paper we introduce the following mapping algorithm which will secure several steps such as :

- an association of Vehicle Detection Stations (VDS) with reported accidents in their proximity,
- a segmentation of traffic speed disruptions from detector readings,

- an association of detector stations with reported accidents (we will further show that this step is necessary due to many detected user-input errors in accident reports).

As a result, we obtain traffic disruptions segmented by the traffic flow associated with reported accidents. This association makes it possible to perform various important tasks of the accident analysis: 1) prediction of the traffic accident impact on the traffic speed based on accident reports, 2) prediction of the traffic accident duration derived directly from the effect of disruption on the traffic speed (impact-based duration), 3) analysis of disruption propagation (each detected disruption can be studied for spatial-temporal impact within the traffic network). Through this work, we will focus on the prediction of the impact-based accident duration and lay the foundation for a further research. Overall, the main contributions (summarised in Figure 1) of our paper is as follow:

1. We conduct a fusion methodology of two large data sets (CTADS and PeMS) for a detailed traffic accident analysis. To the best of our knowledge, this is the first research study proposing the methodology for merging of two large data sets of such nature, which allows an association between observed disruptions in traffic flow and the reported accidents.

2. We propose a novel methodology for the disruption mining using a combination of different metrics: a) the Wesserstein metric, which allows us to measure the disruption severity and b) the Chebyshev metric, which provides a higher selectivity for the disruption mining and a rectangular shape of the disrupted segments, allowing an automated disruption segmentation. We detail all unique properties of both metrics utilized together to allow an accurate disruption segmentation.

3. We perform the estimation of traffic accident disruption duration from traffic speed via the above metrics which allows us to alleviate user-input errors in accident reports.

4. We evaluate multiple machine learning models by comparing both the reported and the estimated accident duration predictions extracted from traffic speed disruptions.

5. We introduce a new modelling approach which focuses on predicting the amount of the the disruption associated with an accident, rather than predicting via only historical logs. Our advanced disruption mining associated with traffic accident reports allows us to extract actual shapes of the disruptions, which allows a further analysis and modelling of the traffic flow impact properties. We also detect multiple accidents which produce secondary disruptions by using this approach.

Overall, this research forms the foundation for a new early traffic accident disruption detection, traffic flow disruption shape analysis and the use of observed traffic accident durations for correcting errors in user reports. Moreover, this work contributes to our ongoing objective to build a real-time platform for predicting traffic congestion and to evaluate the incident impact (see our previous works published in Mihaita et al. (*6*)-Shafiei et al. (*7*)-Mao et al. (*8*)).

The paper is further organised as follows: Section 3 discusses related works, Section 4 presents the data sources available for this study, Section 5 showcases the methodology, Section 6 presents the disruption segmentation results, showcases the result of data set fusion and Section 7 provides conclusions and future perspectives.

RELATED WORKS

Multiple studies rely on user-input-based incident reports from Traffic Management Centers (TMC) with different machine learning models to predict the traffic incident duration (*9*). The use of traffic flow features is found to be rare and mostly specific - incident detection and incident impact pre-

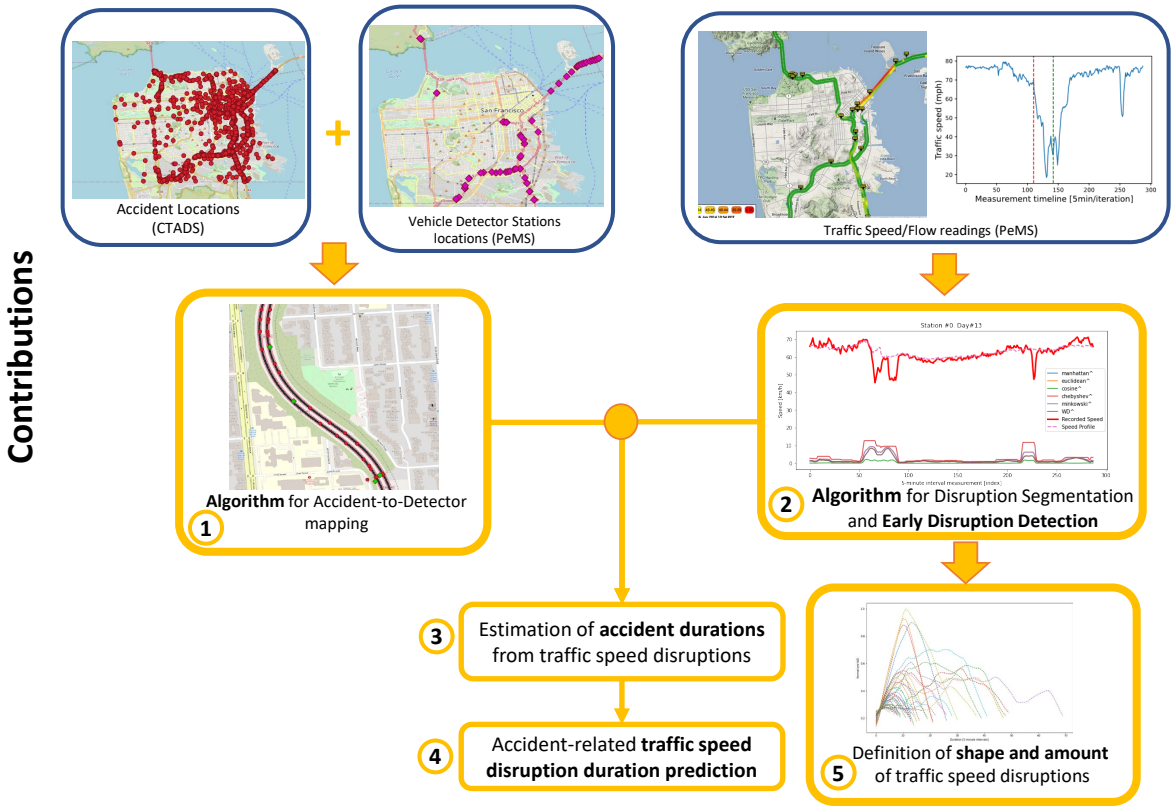


FIGURE 1 Contributions and data-flow schema for association of traffic speed readings with accident reports

diction by using traffic flow (*10*). In other words, traffic flow data is rarely combined with actual incident reports since it requires a higher system complexity and extensive data collection.

There were numerous studies related to accident detection from traffic flow using anomaly detection techniques (*11*). Various methods used for anomaly detection in time series are applicable for the task of traffic disruption detection. The ability to perform the detection of actual disruption, which should give us actual shapes of disruptions and time intervals allows in-depth analysis of usual accident statistics including the effect of the type of accident on the pattern of disruption in traffic flow. By integrating data on traffic state with accident reports we are able to further connect traffic flow disruption patterns to various accident characteristics (hour of the day, weather conditions, crash type, type of vehicle involved - truck/car (*12*), the effect of road pavement types (*13*), road design and road operation (*14*), etc).

Analysis of the effect of traffic incidents has been performed previously using Caltrans PeMS data, where the measure of incident impact was represented as a cumulative travel time delay (*15*), which is an aggregated value. However, traffic state recovery from disruptions is not necessarily following a single pattern - it may be slowly dissipating, we may observe secondary crashes, it may have a high or low impact, etc. Traffic accident duration prediction methodology relies on reported traffic accidents, but actual reports may contain user-input errors and be misaligned with the actual shape of disruption produced by the accident. Therefore, the approach for disruption segmentation may provide the accident duration estimated from the actual shape of

disruption in traffic flow.

Various machine learning models are used to solve the task of traffic accident duration prediction(*9*) including k-nearest neighbours and Bayesian networks (*16*), Recursive Boltsman Machines and Support Vector Machines (*17*), Random Forests and XGBoost (*6*).

The definition of traffic incident duration phases is provided in the Highway Capacity Manual (*18*) and includes the following time-intervals: 1) incident detection - the time interval between the incident occurrence and its reporting, 2) incident response - time between the incident reporting and the arrival of the response team, 3) incident clearance time between the arrival of the response team and the clearance of the incident, 4) incident recovery - the time between the clearance of the incident and the return of traffic state to normal conditions. In this research, we rely on total incident duration - the time between incident occurrence and return of the state to normal conditions. Also, we analyse the subset of traffic incidents - traffic accidents. As we found during the data investigation, traffic accident duration is reported at the time when the incident is cleared by the response team, which doesn't include the duration of the effect that the accident produces on traffic flow.

CASE STUDY

Before diving into the methodology, we provide a brief introduction into the data sets in use for showcasing our approach, which helps establishing the modelling base and understanding of the steps taken. We make the observation that the current methodology can be applied on any incident and traffic state data set which can contain a time component, and is not bounded to the chosen data sets for exemplification.

CTADS: Accident reports data set

We rely on accident reports from the "Countrywide Traffic Accident Dataset" (CTADS), recently released in 2021 (*19, 20*), which contains 1.5 million accident reports collected for almost 4.5 years since March 2016, each report containing 49 features obtained from MapQuest and Bing services. We select the area of San-Francisco, U.S.A and extract data for 9,275 accidents (see Figure 2).

FIGURE 2 CTADS reported accidents for San-Francisco

PeMS: Traffic speed and flow data set

We rely on Caltrans Performance Measurement System (PeMS) (*21*) to collect data on traffic flow and speed. This data set provides aggregated 5-minute measurements of traffic flow, speed and occupancy across California. We decided to extract the data for the area of San-Francisco (see 3a), which contains 83 Vehicle Detection Stations (VDS) placed in that area (see 3b), and we try to associate each traffic accident occurred with each of San-Francisco VDS in their 500m proximity using the algorithm detailed in the following section. In total, from 9,275 accidents in the area (extracted from CTADS) we have obtained 1,932 traffic incident reports which we were able to associate with the correct and complete traffic flow and speed readings from a VDS.

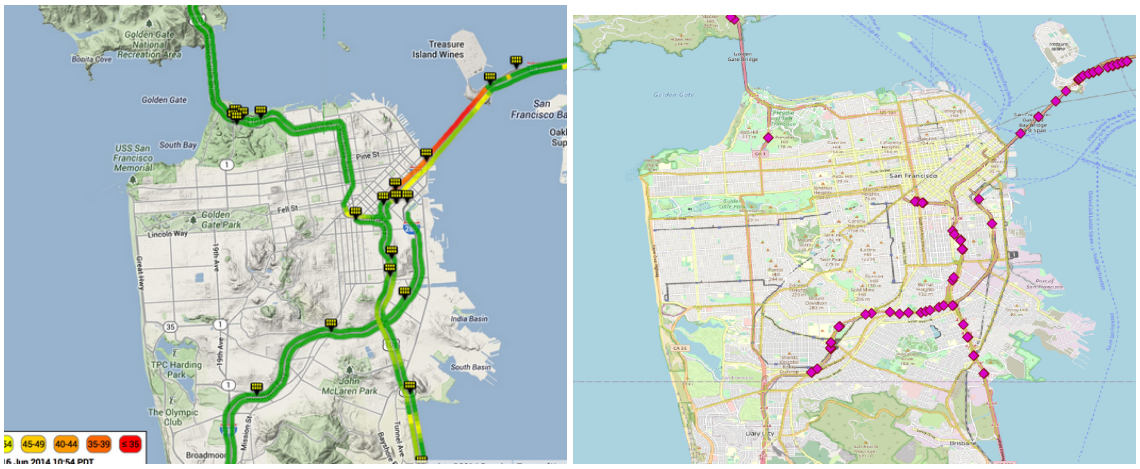


FIGURE 3 1) PeMS data set area coverage for San-Francisco 2) Mapping of the Vehicle Detection Stations from PeMS data set.

METHODOLOGY

The new framework we propose in this paper is represented in Figure 1 which we support across some initial definitions for our modelling approach (see next sub-section). First, we associate the road segments with their corresponding Vehicle Detector Stations (VDS) from the Caltrans PeMS data set, as well with the locations of reported accidents (see Algorithms 1 and 2 proposed in sub-section 5.2). The main outcome of this algorithm is that traffic accidents will get associated with the traffic flow, speed and occupancy readings from the VDS stations.

Second, we propose an new algorithm for an early disruption detection and segmentation, detailed in sub-section 5.3. By segmenting the disruptions occurred in time-space proximity of reported traffic accidents, we obtain the estimated traffic accident duration. This gives us much more information to include in the model training than just the simple accident duration: 1) the disruption shape in terms of modifications of speed data profiles from the standard patterns 2) the accident duration estimated from the impact on the traffic flow 3) the cumulative accident impact estimation.

Definitions

Speed difference estimation definitions

In the current study we compare the performance of multiple difference metrics that will help us to correctly estimate the impact of an accident and the deviation from the historical speed patterns.

These metrics are defined as follows:

a) the Chebyshev difference is expressed as:

$$D_{\text{Cheb}}(x,y) := \int \max_i(|x_i - y_i|) \tag{1}$$

b) the Wasserstein difference:

$$D_{\text{WD}}(u,v) = inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| \mathrm{d}\pi(x,y) \tag{2}$$

where $\Gamma(u,v)$ is the set of (probability) distributions on $\mathbb{R}\times\mathbb{R}$ whose marginals are u and v on the first and second factors respectively.

c) the Cosine difference:

$$D_{\text{C}}(u,v) = \frac{u \cdot v}{\|u\|_2 \|v\|_2}. \tag{3}$$

where $u \cdot v$ is the dot product of u and v.

d) the Euclidean difference between 1-D arrays u and v, is defined as

$$D_{\text{E}}(u,v) = \left(\sum (w_i |(u_i - v_i)|^2)\right)^{1/2} \tag{4}$$

e) the Minkowski difference between 1-D arrays u and v, is defined as

$$D_{\text{M}}(u,v) = \left(\sum |u_i - v_i|^p\right)^{1/p} . \left(\sum w_i (|(u_i - v_i)|^p)\right)^{1/p} . \tag{5}$$

5.1 Algorithm for vehicle detector station to accident association

In order to match correctly what traffic conditions reflect best the effects of each incident, we further define the association procedure between traffic accidents and VDS (A-to-VDS), for the San Francisco area, as shown in Figure 4. We observe that only a few traffic accidents have VDS stations in their proximity to allow a good traffic speed and flow extraction.

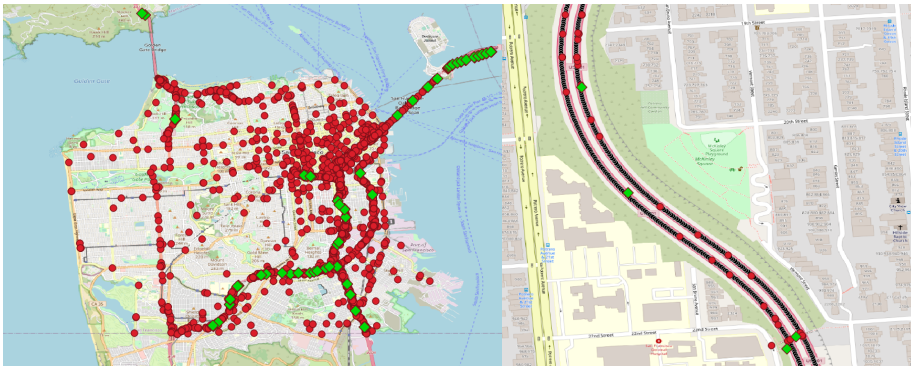


FIGURE 4 a) Traffic incidents [marked in red dots] and VDS stations [marked in green diamonds] across the San Francisco city area. b) A closer look into VDS and traffic incidents locations where black dots are road points.

In order to find the traffic incidents for which we can have traffic flow and speed data, we develop a mapping algorithm (A-to-VDS) which consists of two parts (see Algorithm 1-2), defined by the following steps:

1. We first extract the primary and secondary road lines from Open Street Maps.
2. Road segments are then transformed (segmented) into points at 2-meters equal distance.
3. Each VDS station and accident are mapped to the closest road point (up to 10m distance). Accidents that are not in the proximity of road points removed.

4. From this step we use Algorithm 1 to process the point-based representation of VDS, accidents and road segments. The *vdsPoints* array contains a tuple of the form (VDS ID, x and y coordinates); each point in *accidentPoints* contains an array *visitedBy* (initialized as empty) to maintain a list of stations in the proximity of the accident; *assignedVDS* is a final nearest VDS station close to the accident along the road.

The algorithm relies on a recursive function to implement the process of visiting road points (see Algorithm 2). The association part of the algorithm works as follows:

1. We select the current VDS station.
2. We move (jump by points) in all possible directions available from the starting and forthcoming points in a 3m radius. This radius allows us to move along the road jumping between road points. Movement in all possible directions allows to grasp the propagation of the traffic congestion associated with the accident. The maximum available distance is set to 500m (250 jumps) and allows to limit the observable impact distance.
3. By moving across points we collect traffic incidents in the 5m proximity of each point and associate them with the current VDS station.

Algorithm 1: A-to-VDS: Accident to VDS mapping algorithm

Input: *point*
Output: *None*
Access global arrays: $roadPoints, accidentPoints, vdsPoints$
Function $visitNearestPoints(VDSID, point, currentHops)$
$accidents := findNearestAccidents(point, accidentPoints, 10m)$
for $i = 0$ *to* $length(accidents)$ **do**
  $a := accidents[i]$
  $a.visitedBy.append([VDSID, currenthops])$
end
if $currenthops < 500/2$ **then**
  roadpoints:=findNearestRoadPoints(point, roadPoints, 3m) **for** $i = 0$ *to* $length(roadpoints)$ **do**
    $rp := roadpoints[i]$
    if *VDSID not in rp.visitedBy* **then**
      $rp.visitedBy.append(VDSID)$ $visitNearestPoints(point, currentHops + 1)$
    end
  end
else
  Return
end

The algorithm is recursive and relies on the list of visited points for each VDS. At the end of the algorithm, we have a subset of traffic accidents with their associated VDS which allows us to extract the traffic flow and speed in the vicinity of the accident. Ideally, all traffic accidents should have associated traffic flow but given their unavailability (due to detector coverage), we select accident reports which have associated traffic flow information currently available from the PeMS data set.

Algorithm 2: The recursive function for traveling across road points

Input: $roadPoints, accidentPoints, vdsPoints$
Output: $assignedAccidents$
for $i := 0$ **to** $length(vdsPoints)$ **do**
    vds: = vdsPoints[i]
    visitNearestPoints(vds, 0)
end
$assignedAccidents = []$
for $i := 0$ **to** $length(accidentPoints)$ **do**
    $accident := accidentPoints[i]$
    if $length(accident.visitedBy) > 0$ **then**
        $accident.assignedVDS = sort(accidetn.visitedBy, sortvalue = hops)[0]$
        $assignedAccidents.append(accident)$
    end
end
return $assignedAccdents$

5.2 Algorithm for automated disruption segmentation (ADS)

Once the accidents have been mapped and associated to their VDS stations which allows us to select the flow/speed that match the day of the incident, etc, we are using the extracted traffic state parameters to propose a new automated disruption segmentation (ADS) method. The algorithm for the segmentation of disruptions via traffic speed works as follows:

1. A time series pre-processing step prepares all the data for segmentation (see Alg. 3):
    (a) Calculate the average monthly profile for daily traffic speed measurements;
    (b) Iterate over the traffic speed time series using a moving window of 1-hour time interval (in total there are twelve measurements of 5-minute each)
    (c) On each iteration perform a comparison of a 12-unit window between the monthly profile and the current day of measurements. The resulting single value is added to the resulting time series sequence.
    (d) Calculated the time series differences (TS) choosing the above defined metrics will be then adjusted by selectivity (using the power function, which will keep values closer to one for the least affected by the function and minor values the most suppressed) and normalized to produce nTS and pTS arrays respectively.
2. The time-series segmentation step (see Alg. 4):
    (a) A first order derivative (dTS) is calculated for the resulting time series of the previous stage (nTS), which returns positive peaks when entering the disruption and negative peaks when exiting the disruption state.
    (b) We iteration over resulting derivative time series to record the opening and closing of each disruption in each time series. If two consecutive positive peaks (opening times) are observed then we choose the largest one between the two (we will further debate on this aspect in our future work plans). We repeat the same for consecutive negative peaks.
    (c) We then associate the detected disruptions with the accident reports: for each accident report, we extract the traffic speed time series on the day of the accident and

if both opening and closing times are recorded, we perform an association of the accident with these times and extract the actual time series sequence for further analysis.

Enhancing selectivity: We use the convolution with the kernel (1,1,1), which attributes to the morphological dilation operation, to facilitate the work of the segmentation algorithm. By applying this convolution we make multiple consequent differences to be accumulated ; for example, assuming we have a sequence of 0.3, 0.1, 0.1, 0.2 and 0.2 as differences for each 5-minute step, therefore a total of 0.9 change over 4 iterations. The convolution (1, 1, 1) will produce the values of 0.5, 0.4, and 0.5 by making a sequence of high values from the sequence of small changes (see Figure 5). The dilation operation is primarily used in computer vision tasks to make connected groups from closely placed scattered points to facilitate a further image analysis.

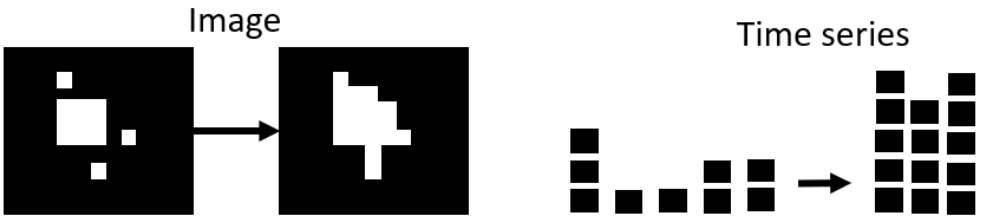


FIGURE 5 The application of dilation operation to an image and time series

To obtain the monthly profile, the traffic speed measurement sequence was obtained for a duration of 1 month from the VDS before the accident occurred, and was done separately for each accident. This sequence then gets reshaped into a matrix of the form $[number\_of\_days; 288]$, where columns contain the total number of measurements across an entire day (24*12=288). The monthly average was then calculated across axis 1 to obtain a vector with 288 values of measurements. This vector gets recalculated for a number of days of observations from each detector to be comparable with the VDS daily measurements.

As an observation, the constants *pThreshold* and *nThreshold* represent thresholds for change that observed in the time series of the metric derivative; they allow us to define a positive and negative change of the difference metric, the selectivity controls power function coefficient to suppress the non-significant and filter the most significant disruptions.

5.3 Modification of the algorithm for automated real-time early disruption detection

Since our proposed algorithm doesn't look into the future and calculates different metrics based on the currently observed traffic speed and a few measurements in the past (11 units in the current study), we can perform an early accident detection which will consist in calculating and comparing the FOD of Chebyshev metric based on the monthly profile. The detection of significant positive peaks (e.g. 0.3-0.5 of normalized difference metric) can identify the amount of disruption in real-time. The end of the disruption can be detected using the same approach in real-time as well by observing a significant negative peak.

Algorithm 3: Algorithm for automated disruption segmentation. Part 1

Input: $monthlyProfile, speedReadings$
Output: cTS
; `//Accident array contains a day number, starting and ending index for segmented traffic disruptions`
$step := 1$
$windowsize := 12$
$i := windowsize$
$lastDiff = 0$
$DS = []$
while $i < length(speedReadings)$ **do**
  $A := speedReadings[i - windowsize + 1 : i]$
  $B := monthlyProfile[i - windowsize + 1 : i]$
  $diff := metric(A, B)$
  $DS.append(diff)$
  $lastDiff = diff$
end
for $i = 0$ *to* $windowsize$ **do**
  $DS.append(lastDiff)$
end
$pTS = power(TS, selectivity)$
$nTS = normalize(pTS)$
$dTS = derivative(nTS)$
$cTS = convolution(dTS, [1, 1, 1])$
return cTS

5.4 Accident duration prediction definitions

Using all available data sets and the incident information, we first denote the matrix of traffic incident features as:

$$X = [x_{ij}]_{i=1..N_i}^{j=1..N_f} \tag{6}$$

where N_i is the total number of traffic incident records used in our modelling and N_f is the total number of features characterising the incident (accident severity, vehicles involved, number of lanes, etc) according to the accident report data set.

Traffic speed is represented as a vector with 5-minute averaged readings from Vehicle Detector Stations:

$$S = [s_i]_{i=1..N} \tag{7}$$

where N is the total amount of traffic speed readings.

As detailed in our methodological framework, the last step of our approach for improving the incident duration prediction is to assess the performance of various Machine Learning models on the tasks of predicting the reported and the estimated accident duration. This is mainly to understand if the approach we have proposed for an early accident detection is better or worse than by simply using reported incident logs from the traffic management centres. We define the task of the accident duration prediction as a regression problem.

Algorithm 4: Algorithm for automated disruption segmentation. Part 2

Input: $cTS, pThreshold, nThreshold, selectivity$
Output: $Accidents$
; `//Accidents array contains a day number, starting and ending index for segmented traffic disruptions`
$state := 0$
$Accidents = []$
for $i := 0$ **to** $length(cTS$ **do**
  if $cTS[i] > pThreshold$ **then**
    if $state <> +1$ **then**
      $state = +1$
      $enteridx = i$
    else
      if $cTS[i] > cTS[enteridx]$ **then** $enteridx = i$;
    end
  end
  if $cTS[i] < nThreshold$ **then**
    if $state <> -1$ **then**
      $state = -1$ $exitidx = i$
    else
      if $cTS[i] < cTS[enteridx]$ **then** $exitidx = i$;
    end
  end
  if $i \ mod \ 288 == 0$ *and* $i > 0$ **then**
    $state = 0$
    $Accident.append([i \ div \ 288, enteridx, exitidx])$
  end
end
return $Accidents$

The incident duration regression vector (Y_r) is represented as:

$$Y_r = [y_i^r]_{i \in 1..N}, \ y_i^r \in \mathbb{N} \tag{8}$$

and the regression task is to predict the traffic accident duration y_i^r based on the traffic incident features $x_{i,j}$. The regression models go via an 10-fold cross-validation procedure with intense hyper-parameter tuning. Several models have been trained and compared and their performance assessed in our results section.

The estimate the accident duration prediction performance we use the root mean squared error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t - F_t)^2} \tag{9}$$

where A_t represents the actual value of the incident duration logs (which can be either used raw from the incident logs or from our automated segmentation technique) and F_t the predicted value of the regression models.

RESULTS

Data exploration and setup

CTADS data set contains traffic accident reports, which after an initial data mining investigation, we found to contain several user-input errors; for example, a lot of traffic accident durations have been rounded to 30 or 360 minutes (see Fig. 6d)); or the incident start time which was reported is unrelated to any disruptions observed by the vehicle detector stations in the proximity - see Figure 6 in which we have provided two different examples of speed recorded during two different accidents A-1390 and A-7102; the red lines indicate the official reported start and end time of the accidents, while in reality the accidents have had a long lag in spreading across the network - see Fig. 6a) or were reported much later that the official speed drop was recorded - see Fig. 6b).

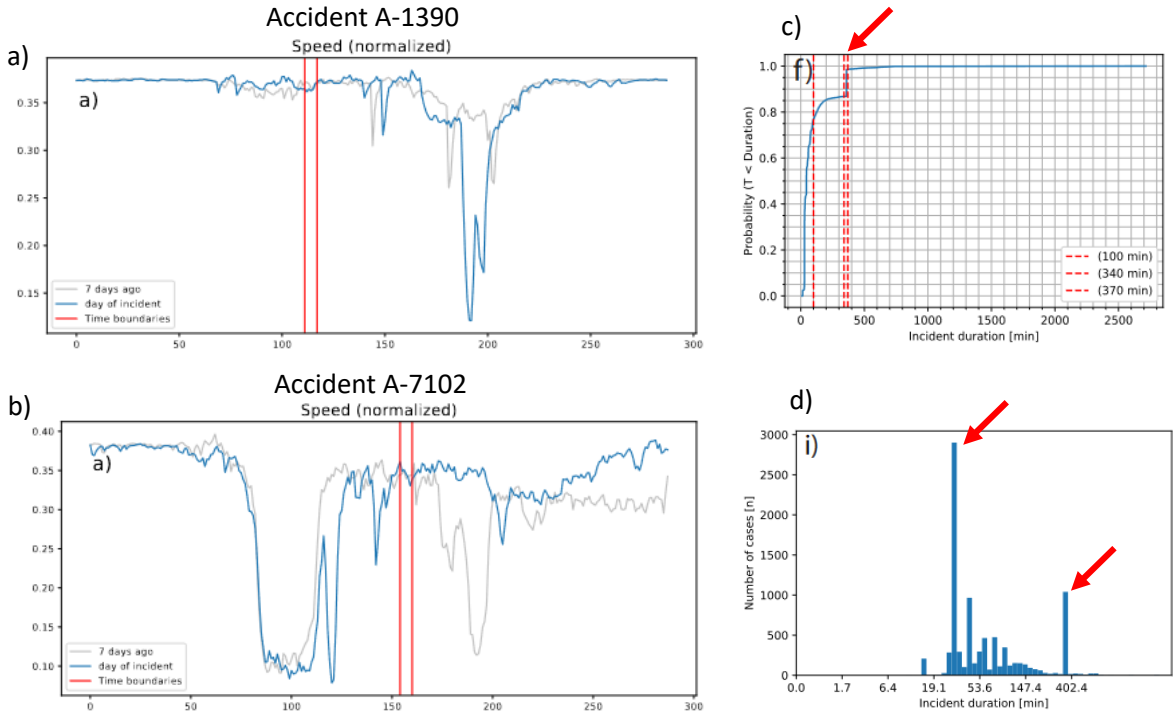
At this step we observed a significant amount of user-input errors in accident reports, which affect the accident duration/impact analysis: 1) accidents can be reported earlier or later than its occurrence (observable disruption misalignment in time) 2) a report can be filled with "placeholder" duration values not representing the actual accident duration 3) there may be no observable disruption in traffic speed despite the accident report (due to placement and management of the accident) (false positive) 4) there may be accident-related traffic disruptions not grasped by accident reports (false negative). Therefore, incorrect accident start time, duration and end time, unreported presence or abscence of disruption make it necessary to estimate accident duration characteristics from traffic state data instead of relying on user reports. In this paper our proposed methodology is really meant to solve the user-reporting issues related to traffic accidents and to be applied automatically on any data set, regardless of its nature or geo-location.

The use of PeMS data set allows to estimate the impact of accidents on the traffic states (flow, speed). For our scenarios, we choose the area of San-Francisco with accidents recorded from 2016 to 2020 in the CTADS data set. We then obtain Vehicle Detector Station locations from PeMS, the road network shape from OpenStreetMap and we perform an association of CTADS accident reports with VDS stations along the road within 500m proximity. We then try to segment the disruption time interval occurred on the day of an accident. Further, we associate observed disruptions in the traffic speed series with actual accident reports. The purpose of this step is to reduce user-input errors in accident reports and to enhance the modelling of traffic disruptions with an analysis of traffic speed.

Metric performance comparison

We apply the difference metrics detailed earlier in Section 5.1 to a monthly traffic speed/flow profile (monthly readings averaged to one day) and reading on the day of traffic accident. There are two approaches to applying the difference calculation: 1) a global difference - when we try to find the difference between monthly profile and traffic flow/speed readings on the day of accident; the global approach is too broad and will not allow the actual comparison between disruptions localized in time (very long but subtle disruptions can be measured as the same amount of disruption as abrupt but impactful one). We measure the amount of difference that occurred within a moving time window (we choose twelve 5-minute time intervals equivalent to one hour). Traffic speed/flow readings from the moving window are taken right before the currently observed value to ensure that the difference estimation algorithm is not looking into the future. It will be further necessary to perform the prediction of difference values.

To compare the metric performances we provide an example of speed readings from one of the detector stations. Each difference metric demonstrates its specifics as represented in Figure 7:



Misalignments of reported accident times (before or after accident happened, with "rounded" accident duration 30m) with readings from VDS stations in 500m proximity

"Rounded" accident durations (30 and 360 minutes) observed in CTADS data set.

FIGURE 6 User input errors located within the CTADS data set

1) the Chebyshev metric, which we define as the maximum difference between the monthly profile and the observed readings, produces a noticeably rectangular shape and demonstrates a higher selectivity towards major disruptions than other metrics; the Chebyshev metric will be further used for the automated accident segmentation; 2) the Cosine distance detects a change in the traffic flow state - speed fall and increase both represented as positive peak values, 3) the D_{WD} allows for smooth representation of the amount of disruption (conceptually, it measures the amount of work necessary to change one shape into another, which we can rephrase as the amount of work produced by an accident to deviate the traffic state from the normal operation), 4) the Minkowsky, Euclidean and Manhattan difference metrics show little to no difference to the Wesserstein distance; we choose to use the D_{WD} since its connection to physical interpretation.

Automated disruption segmentation results

Figure 8 presents the results obtained from our algorithm for the automated disruption segmentation. The segmentation line (dotted blue) represents the estimated disruption intervals represented as 0 and 1 to perform our visualisation investigation better. Figure 8a) shows that there may be multiple observed disruptions in a 300*5=1500 time interval. Due to errors in accident reports regarding the starting time and the duration of the accident, it is non-trivial to determine which disruption is associated with the accident. The situation may be easier in the case when only one disruption is observed during the day. According to our algorithm we select the largest disruption on the day the accident reported. Figures 8b) and 8c) highlight additional specific situations which

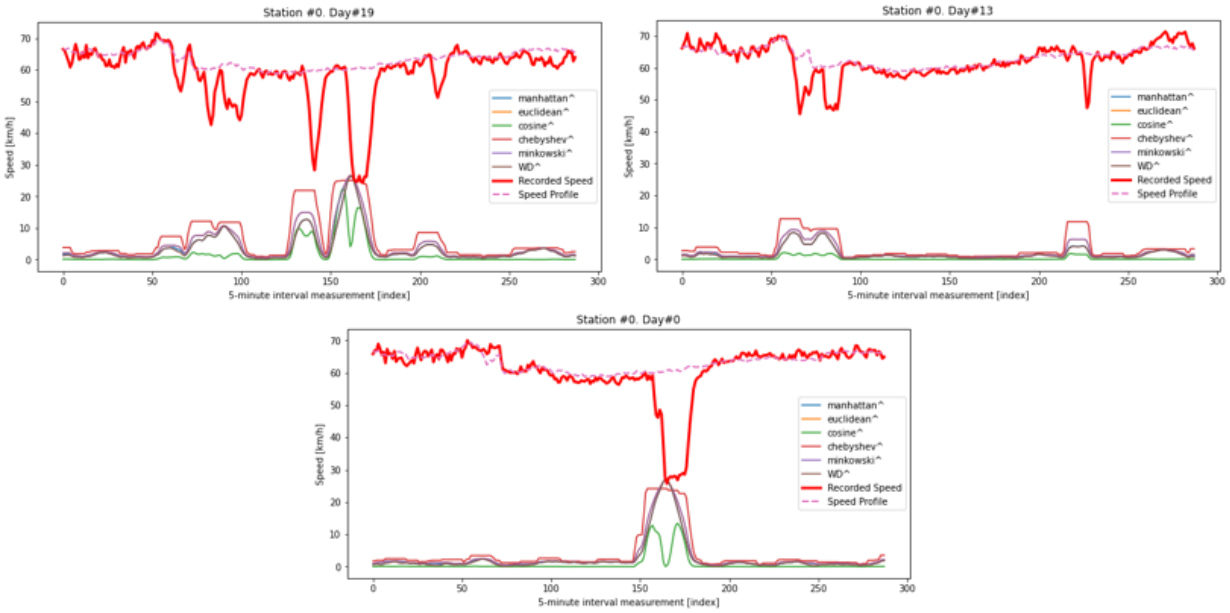


FIGURE 7 Various metrics applied to difference between recorded speed and speed profile

need to be considered: 1) higher traffic speed at the end of the day than observed from the monthly profile, 2) unstable traffic speed approaching normal traffic conditions with high frequency, 3) slight misalignment of disruption intervals with the visually observed disruption intervals. All these problems can be addressed by instead of using manual segmentation we can deploy several Deep Learning models since it has advanced computer vision methods proposed in recent years.

Comparison of our estimated durations versus the official reported accident durations

There is a significant difference between the estimated and the reported accident durations that we would like to highlight: 1) the reported accident durations contain a large amount of 30 and 360 minutes duration values (nearly 40% of data - see Figure 6d)), 2) the estimated accident durations using our approach have an average duration of 58 minutes, while the reported is 108 minutes (which is by assumption skewed due to 360 placeholder values), 3) the estimated accident durations are distributed between 29 and 69 minutes (0.10 and 0.90 quantiles correspondingly) (see Figure 9a)), while the reported durations are distributed between 29 and 360 minutes (see Figure 9b)), which highlights that disruptions observed from traffic speed are much shorter in reality than the ones reported in the original data set, 4) there is no noticeable correlation between the observed and reported durations - see Figure 10 in which we observe that the majority of correlation points are on the lower side of the Ox axis, indicating that in reality the accident reports are much longer that what we have succeeded to estimate via our proposed segmentation and early detection approach.

Extraction of disruption shapes

Previously we applied a Chebyshev metric to perform the segmentation of disruptions. By applying the Wasserstein difference between monthly speed profile and daily traffic speeds and extracting the corresponding disruption interval we are able to observe the shape of disruption impact. The Wasserstein distance (or metric), originally named an Earth Mover distance, has intuitive physical interpretation - the minimum "cost" of altering one pile of earth into the other, which is assumed

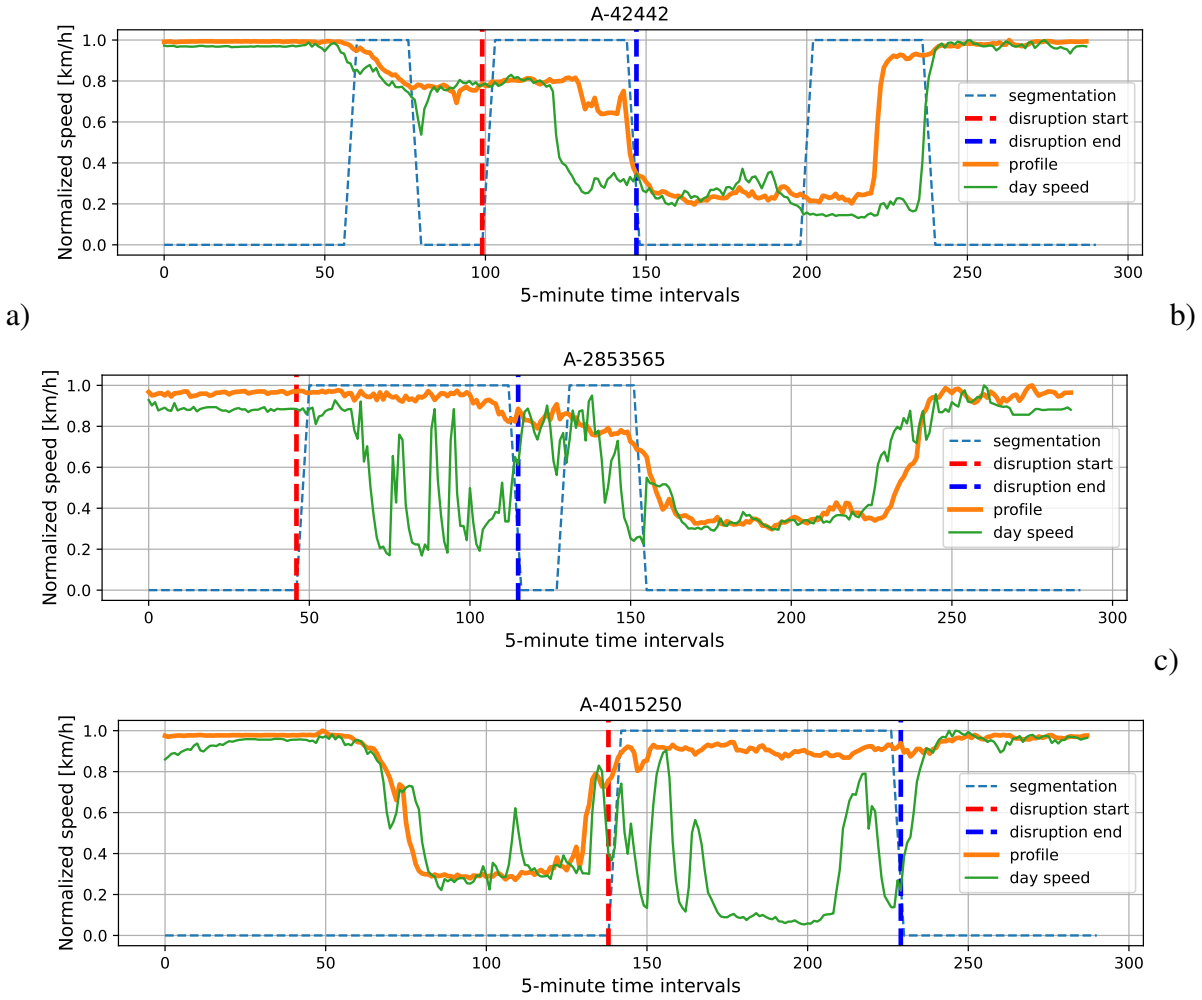


FIGURE 8 Results for automated disruption segmentation algorithm

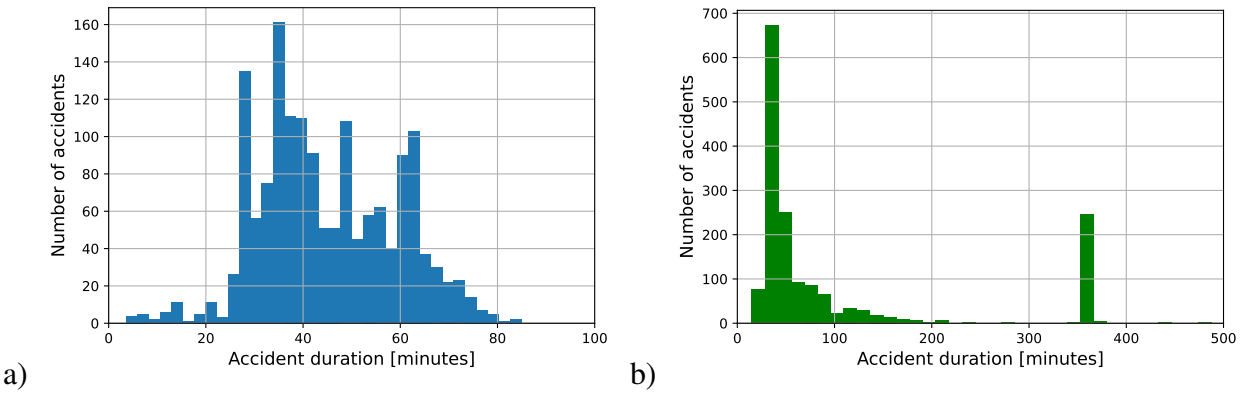


FIGURE 9 Distribution of accident durations for a) estimated and b) reported accident durations for the area of San Francisco

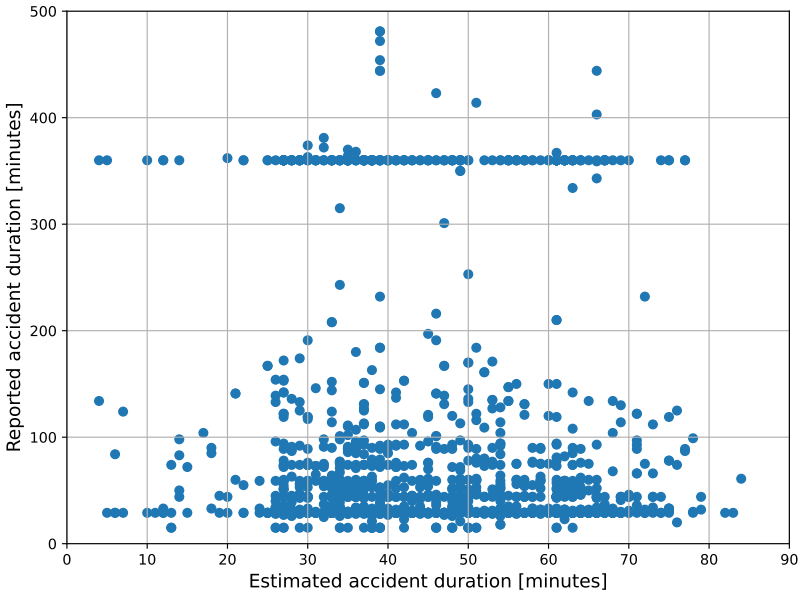


FIGURE 10 Scatter plot for a) estimated and b) reported accident durations for the area of San Francisco

to be the amount of earth that needs to be moved times the mean distance it has to be moved. In relation to the traffic state, it is the minimum amount of work necessary to alter the traffic state to a disrupted condition, or in other words - the amount of traffic disruption. We compare the normalized metric values since at every vehicle detector station there is a different average traffic speed. As in our proposed algorithm, we use a 12-units moving window (one hour) to estimate the Wesserstein difference between the traffic speed measurements and provide the plot for the first 40 segmented disruptions. It allows to conduct a shape analysis investigation in relation to the traffic disruption amount (see Figure 11). Firstly, we observe the similarity between multiple disruptions - they have a 'hill' shape wih a high slope that almost doubles or triples the initial Wasserstein distance (e.g. from 0.2 to 0.5-0.6). Secondly, there are secondary relapses of the Wasserstein distance (a double 'hill') which indicate long-lasting disruptions or cascading disruptions (a harder situation to predict in our future works). The observed shapes can be defined through the parametric equation to perform the classification of disruption effects and facilitate the prediction of the disruption impact timeline, since we observe that high-peak fast-ascending disruptions have a probability to end sooner than slowly ascending ones (which indicate a long term effect or possibly large scale accidents happening in the network).

Accident duration prediction results

We further compare a regression model prediction performance on the CTADS data set by using on the training data set both our estimated versus the reported accident durations (see Table 1). We report results of a 10-fold cross-validation over 1,792 accident reports for which we performed a Vehicle Detector Station association. Firstly, we need to consider that the performance using reported durations from CTADS can be affected because of the presence of user-input errors in the form of placeholder values. Secondly, the nature of estimated accident durations is different since accident response teams usually report the end of the accident at the moment they finished

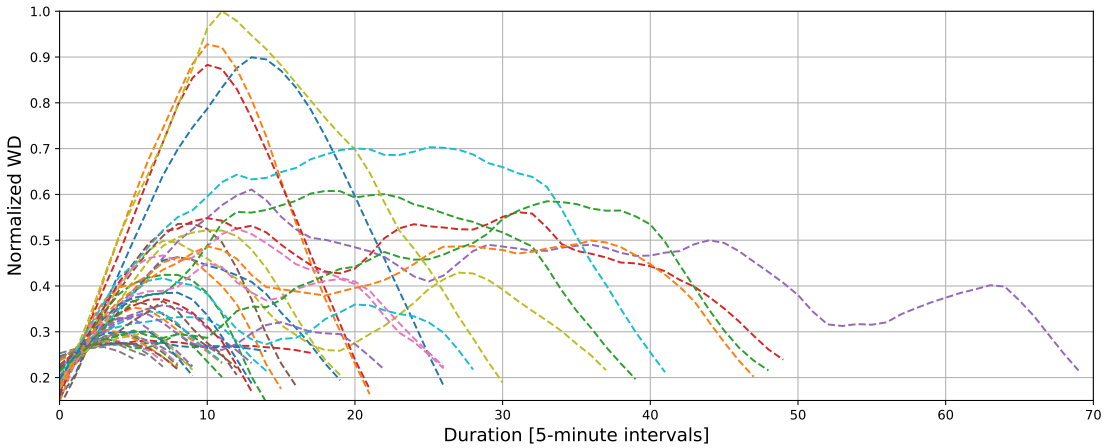


FIGURE 11 Normalized Wesserstein distance plot for disruption shapes extracted for segmented intervals

the accident clearence, without estimating the time for the traffic to return to a normal condition, which would require additional presence, calculations and access to measurements.

TABLE 1 Accident duration prediction results

Regression Model	$RMSE_{est}$	$RMSE_{rep}$	$MAPE_{est}$	$MAPE_{rep}$
KNN	**60.1**	85.7	46.0	**32.9**
RF	**57.8**	81.1	40.4	**28.6**
LR	**60.5**	129.6	**40.8**	72.7
SVM	**61.9**	113.8	**36.1**	50.3
GBDT	**59.3**	69.2	41.4	**31.1**

When we are using accident reports to predict the estimated accident duration, we obtain a better performance using the RMSE metric across all the regression models, which may be connected to the lower amount of long accident durations than reported (see Table 1 first column of $RMSE_{est}$ which is much lower than the $RMSE_{est}$ by almost 41% in the case of RF for example, and 45.6% in the case of SVM). When we compare the prediction results using the MAPE metric, we observe a significantly better performance only for the Linear Regression (LR) and the Support Vector Machines (SVM) - 48.87% and 28.2% improvement respectively, but for k-nearest neighbours (KNN), Random Forest (RF) and Gradient-Boosted Decision Trees (GBDT), results are lower, which can be explained by the presence of the high amount of 30-minute reported durations (30% of all data set), which may skew the model performance towards this value and gain an misleading performance improvement. Therefore the better model performance may be deceiving due to uncorrected user-input errors in accident reports. This makes the estimation of accident duration from traffic state important and necessary allowing an in-depth analysis to be undertaken.

CONCLUSION

The proposed methodology in this paper aims at detecting, segmenting and extracting the observed disruptions in the traffic speed which was modelled together with reported traffic accidents by traffic management centers. The approach is innovative in its distance metric approach for an automatic incident detection coupled with an incident segmentation which has shown to improve the incident prediction by almost 41% in RMSE across multiple machine learning models. By obtaining shapes of disruptions we lay the foundation for accident impact modelling. Many studies still rely on the modelling of reported accident durations and pre-defined parameters, while they can be estimated from traffic state measurements, which gives us more data than just aggregated variables (duration, start time, etc). By having information on how each accident affect the traffic flow, we can study the accident impact with precision.

Limitations of this work: The current modelling approach has been applied to a San Francisco data set due to its public availability and easiness to access. However, we would like to test the approach on multiple other countries and incident databases across the globe; the main challenge is the lack of both traffic states and traffic accidents logs to be released with synchronised timelines. **Future works:** We are currently modelling the cascading effect on traffic disruptions and how these can be automatically identified based on multiple incoming traffic state streams; the main challenge of detecting subsequent incidents lie in the time-span duration of the first incident which is normally stochastic in nature.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: A. Grigorev, A-S. Mihaita, F. Chen; data collection: A. Grigorev; analysis and interpretation of results: A. Grigorev, K. Saleh; draft manuscript preparation: A. Grigorev, A-S. Mihaita. All authors reviewed the results and approved the final version of the manuscript.

ACKNOWLEDGMENT

This work has been done as part of the ARC Linkage Project LP180100114. The authors are highly grateful for the support of Transport for NSW, Australia. This research is funded by iMOVE CRC and supported by the Cooperative Research Centres program, an Australian Government initiative."

REFERENCES

1. Organization, W. H., *Global status report on road safety 2015*. World Health Organization, 2015.
2. Administration, N. H. T. S., *Traffic safety facts 2013*. U.S. department of transportation, 2013.
3. Kim, W. and G.-L. Chang, Development of a Hybrid Prediction Model for Freeway Incident Duration: A Case Study in Maryland. *International Journal of Intelligent Transportation Systems Research*, Vol. 10, 2011.
4. Theofilatos, A., G. Yannis, P. Kopelias, and F. Papadimitriou, Predicting Road Accidents: A Rare-events Modeling Approach. *Transportation Research Procedia*, Vol. 14, 2016, pp. 3399–3405, transport Research Arena TRA2016.
5. Grigorev, A., A.-S. Mihaita, S. Lee, and F. Chen, Incident duration prediction using a bi-level machine learning framework with outlier removal and intra–extra joint optimisation. *Transportation Research Part C: Emerging Technologies*, Vol. 141, 2022, p. 103721.
6. Mihaita, A. S., Z. Liu, C. Cai, and M. Rizoiu, Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting. *CoRR*, Vol. abs/1905.12254, 2019.
7. Shafiei, S., A. Mihaita, H. Nguyen, C. Bentley, and C. Cai, Short-Term Traffic Prediction Under Non-Recurrent Incident Conditions Integrating Data-Driven Models and Traffic Simulation. In *Transportation Research Board 99th Annual Meeting*, 2020, pp. –.
8. Mao, T., A.-S. Mihăită, F. Chen, and H. L. Vu, Boosted Genetic Algorithm using Machine Learning for traffic control optimization. *IEEE Transactions on Intelligent Transportation Systems*, 2021.
9. Li, R., F. C. Pereira, and M. E. Ben-Akiva, Overview of traffic incident duration analysis and prediction. *European transport research review*, Vol. 10, No. 2, 2018, p. 22.
10. Fukuda, S., H. Uchida, H. Fujii, and T. Yamada, Short-term prediction of traffic flow under incident conditions using graph convolutional recurrent neural network and traffic simulation. *IET Intelligent Transport Systems*, Vol. 14, No. 8, 2020, pp. 936–946.
11. Parsa, A. B., H. Taghipour, S. Derrible, and A. K. Mohammadian, Real-time accident detection: coping with imbalanced data. *Accident Analysis & Prevention*, Vol. 129, 2019, pp. 202–210.
12. Eboli, L., C. Forciniti, and G. Mazzulla, Factors influencing accident severity: an analysis by road accident type. *Transportation Research Procedia*, Vol. 47, 2020, pp. 449–456, 22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18th – 20th September 2019, Barcelona, Spain.
13. Tsubota, T., C. Fernando, T. Yoshii, and H. Shirayanagi, Effect of Road Pavement Types and Ages on Traffic Accident Risks. *Transportation Research Procedia*, Vol. 34, 2018, pp. 211–218, international Symposium of Transport Simulation (ISTS'18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS'18)Emerging Transport Technologies for Next Generation Mobility.
14. Yannis, G., A. Dragomanovits, A. Laiou, T. Richter, S. Ruhl, F. La Torre, L. Domenichini, D. Graham, N. Karathodorou, and H. Li, Use of Accident Prediction Models in Road Safety Management – An International Inquiry. *Transportation Research Procedia*, Vol. 14, 2016, pp. 4257–4266, transport Research Arena TRA2016.
15. Miller, M. and C. Gupta, Mining traffic incidents to forecast impact. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, 2012, pp. 33–40.

16. Kuang, L., H. Yan, Y. Zhu, S. Tu, and X. Fan, Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor. *Journal of Intelligent Transportation Systems*, Vol. 23, No. 2, 2019, pp. 161–174.
17. Xiao, S., Traffic accident duration prediction based on natural language processing and a hybrid neural network architecture. In *2021 International Conference on Neural Networks, Information and Communication Engineering*, SPIE, 2021, Vol. 11933, pp. 194–202.
18. Alkaabi, A. M. S., D. Dissanayake, and R. Bird, Analyzing clearance time of urban traffic accidents in Abu Dhabi, United Arab Emirates, with hazard-based duration modeling method. *Transportation Research Record*, Vol. 2229, No. 1, 2011, pp. 46–54.
19. Moosavi, S., M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 33–42.
20. Moosavi, S., M. H. Samavatian, S. Parthasarathy, and R. Ramnath, A countrywide traffic accident dataset. *arXiv preprint arXiv:1906.05409*, 2019.
21. Chen, C., K. Petty, A. Skabardonis, P. Varaiya, and Z. Jia, Freeway performance measurement system: mining loop detector data. *Transportation Research Record*, Vol. 1748, No. 1, 2001, pp. 96–102.