Development of an On-line Scalable Approach for Identifying Secondary Crashes

Hong Yang¹, Kaan Ozbay², Ender Faruk Morgul³, Bekir Bartin⁴, Kun Xie⁵

¹ (Corresponding author)
Post-Doctoral Fellow
Department of Civil & Urban Engineering, NYU-Poly
Center for Urban Science + Progress (CUSP), New York University (NYU)
One MetroTech Center, 19th Floor
Brooklyn, NY 11201
Tel: +1-(646)-997-0548
Fax: +1-(646)-997-0560
Email: hong.yang@nyu.edu

² Professor,
Department of Civil & Urban Engineering, NYU-Poly
Center for Urban Science + Progress (CUSP), New York University (NYU)
One MetroTech Center, 19th Floor
Brooklyn, NY 11201
Tel: +1-(646)-997-0552
Fax: +1-(646)-997-0560
Email: kaan.ozbay@nyu.edu

⁴ Assistant Professor,
Department of Civil Engineering,
Istanbul Kemerburgaz University,
Mahmutbey Dilmenler Caddesi No: 26,
Bagcilar, Istanbul, Turkey,
Tel: 00 1 (212) 604 01 00 Ext. 1411
Email: bbartin@gmail.com

³,⁵ Graduate Assistant,
Department of Civil & Urban Engineering
Polytechnic Institute of New York University (NYU-Poly)
One MetroTech Center, 19th Floor
Brooklyn, NY 11201
Email: efm279@nyu.edu
Phone: +1-(646)-997-0531
Email: kun.xie@nyu.edu
Phone: +1-(646)-997-0547

Word count: 5076 Texts + 1 Table + 9 Figures = 7576
Abstract: 246
Submission Date: August 1, 2013
Resubmission Date: November 15, 2013

Paper submitted for Publication in the
Transportation Research Record, Journal of Transportation Research Board after being presented
Transportation Research Board’s 93rd Annual Meeting, Washington, D.C., 2014
ABSTRACT

Secondary crashes are one of the most critical incidents occurring on highways. They can induce extra traffic delays and affect highway safety performance. Transportation agencies are interested in understanding the mechanism of the secondary crash occurrence and implementing appropriate countermeasures. However, there is no well-established procedure to identify secondary crashes, which in turn impedes the possibility of investigating the underlying mechanism of their occurrence. This study intend to develop an on-line scalable approach to help identify secondary crashes for a large number of highways that have insufficient traffic surveillance units collecting continuous traffic data required to classify secondary crash accurately. The developed approach consists of two major components: (a) acquisition of open source traffic data and (b) identification of secondary crashes through the use of these data. Unlike existing approaches based on static thresholds, queuing models or infrastructure-based sensor data, the developed approach takes advantage of various open sources data to identify traffic conditions in the presence of incidents. In this study, we propose to develop virtual sensors collecting traffic data from private traffic information providers such as Bing Maps, Google Maps and MapQuest. The availability of such data greatly expands our ability to cover more highways without installing infrastructure sensors. The virtual sensor output provides the basic input to run the developed automatic identification algorithm for identifying secondary crashes. The algorithm is described in a step-by-step manner to provide a readily deployable approach for transportation agencies interested in identifying secondary crashes on their highway networks.
INTRODUCTION
Traffic incidents affect the operational performance and safety of transportation systems. The National Traffic Incident Management Coalition (NTIMC) estimated that traffic incidents account for about one-quarter of all congestion on US roadways (1, 2), and increase the risk of secondary crashes (3). Previous research showed that the risk of crash occurrence increase more than six times if a prior crash occurred (4). One additional minute increase in the incident clearance time causes the possibility of secondary crashes increase by about three percent (5). Secondary crashes can account for twenty percent of all crashes and eighteen percent of all fatalities on US freeways (1, 6). Moreover, secondary crashes lead to additional traffic delays as more recovery time needed to clear their impact. Thus, the prevention of secondary crashes has been recognized as a high priority in traffic incident management (TIM) (6).

Transportation agencies attempted to develop TIM programs to improve safety for travelers by reducing the risk of secondary crashes (7, 8). In order to initiate appropriate countermeasures, the underlying mechanisms of secondary crash occurrences need to be known. Prior to exploring the casual relationship between secondary crashes and possible explanatory variables, one key step is to first identify the secondary crashes, as these crashes are not labeled as secondary crashes in the police crash reports. When reporting the crashes, each crash is independently documented, and the potential link between a pair of crashes is lost. To overcome the issue, previous research developed a number of approaches based on static thresholds (9, 10), queuing models (11, 12) and sensor data (13, 14) to establish the link between a potential secondary crash with the primary one. Previous studies provided important procedures to highlight the secondary crashes. However, the practical use of these approaches is inherently limited due to a number of shortcomings, such as subjective thresholds and lack of traffic data.

To expand the capability of existing identification approaches, the objective of this paper is to develop an on-line scalable approach for automating the classification of secondary crashes on more number of highways. The proposed approach consists of two major components including data acquisition and identification algorithm. The first component intends to develop a framework for utilizing third-party open source traffic data. The second component automates the detection of potential primary-secondary crash pairs using the open source traffic data. The proposed approach is described in a manner that practitioners can use it as a readily deployable tool for secondary crash identification in large-scale highway networks.

LITERATURE REVIEW
A number of studies have conducted research on identifying secondary crashes. TABLE 1 presents a summary of these studies on secondary crash identification and analysis. The majority of them focused on the static approach that defined the relationship of primary-secondary crashes by predefined temporal and spatial thresholds (5, 9, 15-19). Similarly, secondary crashes that occurred as a result of rubbernecking in another direction of traffic were also commonly determined by the fixed spatiotemporal thresholds (10, 20-22). As shown in TABLE 1, there is no agreement on the selection of these spatiotemporal thresholds. These static approaches relied only on incident (crash) information, and the subjectively defined thresholds greatly affected the reliability of results. Obviously, different primary crashes occurred under various traffic conditions can cause different spatiotemporal impact. In addition, not all crashes occurred within the pre-defined thresholds are dependent. As a result, a large threshold will overestimate the frequency of secondary crashes whereas a short one will underestimate it. The lack of accurate information on the incident duration and queue length makes it difficult to define appropriate thresholds.

In contrast to the static approach, several studies as shown in TABLE 1 developed queuing models to describe the dynamic impact area of primary crashes (11, 12, 23, 24). Dynamic thresholds were usually modeled as a function of a set of explanatory variables such as primary incident duration and the number of lanes blocked. The use of queuing models greatly contributed to defining better thresholds for the identification of secondary crashes. However, there were still several issues associated with the use of the theoretical queuing models. For instance, it was assumed that the impact area of a primary incident reached its maximum at the clearance time (24) or the spatial impact only presented within the incident duration (12). These assumptions were frequently violated in reality. For instance, the maximum queue
might build up after the incident was cleared given that the upstream traffic arrival rate exceeded the
downstream dissipation rate (25). More importantly, the performance of these queuing models was
constrained by the availability of incident information. To develop more reliable models, accurate
incident information such duration, clearance time and the number of lanes blocked were usually required
(11, 12, 23, 24). However, detailed incident records are usually very limited.

<table>
<thead>
<tr>
<th>Study</th>
<th>Type</th>
<th>Major Data Needs</th>
<th>Identifying Secondary Crash</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5, 9, 15)</td>
<td>static incident data</td>
<td>&lt; clearance time+15 minutes, &lt; 1 mile</td>
<td></td>
</tr>
<tr>
<td>(16)</td>
<td>static incident</td>
<td>&lt; clearance time+15 minutes, &lt; 3 miles</td>
<td></td>
</tr>
<tr>
<td>(17)</td>
<td>static crash data</td>
<td>&lt; 2 hours, &lt; 2 miles</td>
<td></td>
</tr>
<tr>
<td>(18)</td>
<td>static incident</td>
<td>&lt; clearance time+15 minutes, &lt; 2 miles, lane closure</td>
<td></td>
</tr>
<tr>
<td>(19)</td>
<td>static incident</td>
<td>&lt; actual duration, &lt; 1 mile upstream</td>
<td></td>
</tr>
<tr>
<td>(10, 21)</td>
<td>static incident data</td>
<td>&lt; 2 hours, &lt; 2 miles (both directions)</td>
<td></td>
</tr>
<tr>
<td>(20)</td>
<td>static incident data</td>
<td>&lt; 2 hours, &lt; 2 miles; &lt; 0.5 hour, &lt; 0.5 mile (other direction)</td>
<td></td>
</tr>
<tr>
<td>(22)</td>
<td>static crash data</td>
<td>&lt; 80 minutes, &lt; 6,000ft; &lt; 1,000ft (other direction)</td>
<td></td>
</tr>
<tr>
<td>(23)</td>
<td>dynamic incident data</td>
<td>maximum queuing model</td>
<td></td>
</tr>
<tr>
<td>(11, 24)</td>
<td>dynamic incident data</td>
<td>incident progression curves</td>
<td></td>
</tr>
<tr>
<td>(12)</td>
<td>dynamic incident data</td>
<td>deterministic queuing model</td>
<td></td>
</tr>
<tr>
<td>(26-27)</td>
<td>dynamic incident + simulated traffic data</td>
<td>determine impact area based on simulated speed contour map</td>
<td></td>
</tr>
<tr>
<td>(28-30)</td>
<td>dynamic incident + monitor + sensor data</td>
<td>identify influential area by ASDA model</td>
<td></td>
</tr>
<tr>
<td>(13, 31)</td>
<td>dynamic crash + sensor data</td>
<td>determine spatiotemporal impact by speed contour map</td>
<td></td>
</tr>
<tr>
<td>(14)</td>
<td>dynamic crash + sensor data</td>
<td>determine crash impact region by speed contour map</td>
<td></td>
</tr>
</tbody>
</table>

To address the issues associated with static approach and queuing model, several recent studies
attempted to describe the influential area by aggregating incident/crash information and traffic data (13,
14, 26-31). For instance, Chou and Miller-Hooks (27) and Haghani et al. (26) simulated the impact of
primary incidents to identify secondary crashes. Vlahogianni et al. (28, 30) and Orfanou et al. (29) used
the recognition and tracking model by Kerner et al. (32) based on traffic sensor data to automatically track
the propagation of moving traffic jams induced by primary incidents. Yang et al. (13, 31, 33, 34) and
Chung (14) developed speed contour maps and identification algorithms based traffic sensor data to
capture the impact of primary crashes. These data-driven approaches greatly improved the description of
spatiotemporal impact of incidents because they have taken into account the prevailing traffic conditions
associated with the incidents. However, a large-scale use of these approaches could not be processed
because of the unavailability of sensor data on many highways. Although many types of traffic sensors
are currently in use, widespread deployment of these sensor systems has been limited due to high costs
associated with sensor installation, maintenance, data processing and storage (35). As an alternative, the
present study attempted to extend the data-driven approaches by taking advantages of more accessible
third-party traffic data sources.

DEVELOPMENT OF AN ON-LINE SCALABLE APPROACH

Overview of Proposed Approach

The development of a scalable approach for identifying secondary crashes involves two major tasks: (a)
large-scale traffic data acquisition and (b) development and implementation of a new identification
algorithm. Instead of using data from infrastructure-based sensors such as loop detectors, this study
proposes utilizing traffic information from open sources such as Bing Maps™, Google Maps™ and
MapQuest™. Combining the obtained traffic information and the identification algorithm enables
practitioners to identify secondary crashes faster and at a larger scale. FIGURE 1 shows the
methodological framework of the proposed approach. The following sub-sections describe the proposed
approach in detail.
Open Source Data Collection

A number of third-party map services such as Google Maps\textsuperscript{TM}, Bing Maps\textsuperscript{TM} and MapQuest\textsuperscript{TM} provide publicly available real-time traffic information. Travelers can now easily find out about prevailing traffic conditions and incidents that might disrupt their journeys. These data come from a variety of sources, including government agencies and private data providers. More recently, much of the traffic data used by the map service providers are sourced from thousands of commuters from their GPS feature enabled smartphones while traveling in their cars. For instance, Microsoft\textsuperscript{TM} has indicated that their Bing Maps now uses traffic information offered by Nokia (HERE) Maps\textsuperscript{TM} (36). This allows Bing Maps to acquire traffic data from many locations around the world as Nokia has access to millions of devices powered by their Symbian platform. Therefore, Bing Maps expanded their traffic coverage to even a larger number of roadways. For example, traffic information in the United States that was already available on Bing Maps now includes local roadways. Similarly, Google Maps tap into GPS data received from Android handset users through traffic crowdsourcing (37). Any user that has enabled "My Location" with Google Maps on their device contributes anonymous GPS data and help Google capture and display live traffic.

Most of these third-party map service providers provide the Application Programming Interface (API) for developers to (partially) access their live traffic information on maps. For instance, Bing Maps REST Services API offers a Representational State Transfer (REST) interfaces to allow users to perform tasks such as creating a route and geocoding an address (38). Particularly, its traffic API provides the access to real-time information such as incidents, congestion and construction sites along the suggested road. Similarly, Google Maps, MapQuest and many others also provide APIs to access their traffic information (39, 40). More detailed discussion on these data sources can be find in our study (41).

In this paper we propose the idea of creating a number of virtual sensors to capture the necessary traffic data, through customized API. This idea is depicted in FIGURE 2. For example, on a highway segment of AB, eight virtual sensors are created. These virtual sensors are essentially a number of user-defined pushpins with coordinate information. The distance between two virtual sensors is then calculated by their coordinates. Depending on user needs, a pair of the two adjacent virtual sensors can be placed to collect traffic information for a short link or a long route, e.g. 0.1 mile, 0.5 mile or 1 mile. Similar to regular query on these maps, one can easily obtain travel time information in Extensible Markup Language (XML) format between two adjacent sensors based on the programmed query function in the API. The obtained travel time information is then converted into speed based on the section length.
FIGURE 2 Obtaining real-time speed data by deploying virtual sensors on maps.

Rather than relying on limited infrastructure-based sensor data on a single road, users can create a large number of virtual sensors to cover the entirety of the highway network. This approach in turn enables us to obtain more traffic data to support the identification of secondary crashes for networks without infrastructure-based sensors.

Traffic Data Validation
In order to use the derived speed measurements from virtual sensors, first, the data quality has to be examined. The derived speed measurements from virtual sensors have to be consistent with the other known reference data such as the infrastructure-based sensor output. As a case study, multiple remote traffic microwave sensor (RTMS) stations deployed on the New Jersey Turnpike were selected and the corresponding virtual sensors were created on the Bing Maps. A customized API was developed to derive speed measurements from virtual sensors on the Bing Maps. We should mention that other data sources also can be considered given the access to their APIs for traffic information. The speed estimations from the virtual sensors were then compared with the outputs from the RTMS. Comparisons were conducted under different weekdays, incident and work zone conditions. FIGURE 3 illustrates the speed data collected by both types of sensors. The Wilcoxon signed-rank test with a significance level of 0.05 was applied to examine the two-paired samples (The null hypothesis $H_0$ is that the two data sets are equivalent). The corresponding p-values for all the scenarios were greater than 0.05. Thus it was concluded that the speed measurements derived from the virtual sensors were consistent with the measurements from infrastructure-based sensors. Like the RTMS data, the virtual sensor measurements can be used to identify the events that disrupt the normal traffic.
Despite the relatively acceptable performance of virtual sensors, users need to recognize that the accuracy of their information is often affected by the penetration rate of smartphone users with their GPS enabled. For instance, the penetration rate might be too low to generate reliable speed estimates for highways in a rural area. In addition, the performance of the undisclosed algorithms used by the map service providers also affects the quality of the results. With increasing penetration rate of smartphone users with their GPSs enabled and more open source data, the quality of the virtual sensor measurements is expected to improve.

**Automatic Secondary Crash Identification Algorithm**

A secondary crash is the one that occurs within the impact area of a previous traffic incident. As an example, FIGURE 4 shows a potential primary-secondary crash pair on New Jersey's 511 interactive map on July 19, 2013. Crash A had occurred earlier on southbound I-95 and induced congestion due to left lane blockage. Then crash B had occurred at the upstream just 9 minutes after the first crash A. In order to identify whether or not crash B is a secondary crash, the major task is to examine the impact area of the prior incident A. The impact area is described by temporal and spatial constraints. If a simple static threshold 1 shown in FIGURE 4 was used as the spatiotemporal criterion, crash B will not be identified as a secondary crash. However, if another static threshold 2 with longer spatial and temporal limits was used, crash B will be classified as a secondary crash. The difficulty is to determine an appropriate threshold to capture the progression of the actual impact of crash A. This is because each crash will have different impact on traffic. FIGURE 5 shows an example of the impact progression of a severe crash on July 30, 2013 occurred the New Jersey Turnpike. It can be seen the queue length reached as long as 10 miles and queue presented more than 3 hours. Thus any static thresholds cannot capture the queuing evolution process.
FIGURE 4 Example of potential primary-secondary crashes shown on 511NJ map.

(a) 30 minutes after accident; As of 8:30am, there was an accident and a tractor trailer fire on the northbound I-95. All lanes blocked (07-30-2013)

(b) As of 10:04am, right lane blocked; 10-mile delay

(c) As of 10:52am, accident was cleared; all lanes open

(d) As of 11:25am, queue at downstream was released

(e) As of 11:43am, queue at upstream was released; queue move to downstream

(f) As of 11:54am, major queue was cleared

FIGURE 5 A crash with significant impact (10-mile maximum queue; queue presented 3+ hours).
Instead of using fixed thresholds or simplified queuing models, this study develops an algorithm to dynamically capture the impact of each incident based on the traffic information from available open sources. The algorithm consists of three major steps described below.

**Step 1:** Construct speed contour Maps (SCM). The speed measurements derived from the virtual sensors are extracted to develop SCM similar to the one shown in FIGURE 6. Each cell in the figure represents a speed measurement \( V(S_i, T_s) \) from \( m \) virtual sensor \( S \) on the studied highway at the \( n \)th time interval \( T \), \( \forall m=1,2,\ldots,M-1,M \) and \( \forall n=1,2,\ldots,N-1,N \). The speed measurement of each cell is coded by a continuous range of spectral colors shown in FIGURE 6. The colored SCM helps highlight the congested area. For example, there are three types of congested areas shown in the figure. The left one actually represents the recurrent congestion between 11:00 and 16:00. The middle one represents congestion induced by two crashes. The right one shows the congestion caused by a long-term work zone that was activated every day between 22:00 and 23:00.

**Step 2:** Develop a representative speed contour map (RSCM). In order to automatically detect the crash-caused congestion area shown in FIGURE 6, the SCM has to be compared with the normal traffic condition. Therefore, a RSCM that represents daily incident-free conditions on a highway is developed. Such RSCM can be constructed by the representative speed measurements. We propose to use the \( \rho \)th percentile speed \( \tilde{V}(S_i, T_s) \) of the historical incident-free virtual sensor speed measurements of each cell as the representative speed. The detailed procedure for obtaining the representative speed measurements is described in our previous papers (13, 31). FIGURE 7 represents the typical normal traffic conditions of the same highway shown in FIGURE 6. The recurrent congestion in the afternoon and the long-term work zone congestion at night are still visible on the RSCM. However, as the crashes are random and rare events, the highway section between sensor 25 and sensor 110 has a relatively smooth traffic condition between 17:00 and 19:00 in most of the days. Therefore, we can highlight the crash-induced congestions by comparing SCM with RSCM in the next step.

**Step 3:** Construct a binary speed contour map (BSCM). This step checks the original virtual sensor speed measurement \( V(S_i, T_s) \) of SCM in step 1 and compares it with the corresponding representative speed \( \tilde{V}(S_i, T_s) \) of RSCM generated in step 2. If \( \tilde{V}(S_i, T_s) - V(S_i, T_s) > \Delta V \), the speed measurement \( V(S_i, T_s) \) in the original SCM is converted into \( V(S_i, T_s) = 1 \). Otherwise, it is denoted as \( V(S_i, T_s) = 0 \). The constraint \( \tilde{V}(S_i, T_s) - V(S_i, T_s) > \Delta V \) classifies an abnormal traffic condition if the virtual sensor speed is \( \Delta V \) mph less than the normal traffic condition, where \( \Delta V \) is a user-defined threshold that detect congestion occurrence. A larger \( \Delta V \), for instance, 25 mph, indicates a more conservative criterion to differentiate abnormal traffic condition from the normal ones. A smaller \( \Delta V \), for instance 10 mph, provides a more aggressive threshold to determine congestion. A reasonable \( \Delta V \) can be defined based on the highway speed limits and transportation agencies’ practices for speed reductions under incident conditions. In general, transportation agencies can consider their specific definitions of congestion for different facilities, for example, 30 percent or more reduction in speed on a freeway (or equivalently, speed decrease from 65 mph to 45 mph or lower). FIGURE 8 shows an example of converting the SCM generated in step 1 into a BSCM. Such BSCM together with the incident information provide the basis to identify secondary crashes. As shown in FIGURE 8, instead of manually reviewing each BSCM and incident coordinates, our goal is to automatically detect whether crash \( B \) is a secondary crash of the prior crash \( A \) given the BSCM, occurrence time and location of each crash.
FIGURE 6 Example of a speed contour map based on virtual sensor measurements.

FIGURE 7 Example of a RSCM based on historical virtual sensor measurements.

FIGURE 8 Example of a binary speed contour map.
Step 4: Identify a secondary crash. The constructed BSCM provides the basis for the identification of secondary crashes. By overlapping the incident information (time and location) on the BSCM, we can visually see the relationship between a pair of crashes. Using FIGURE 9 as an example, assuming there were two crashes B and C that occurred at a later time at the upstream of the first crash A. If the static approach with a spatial threshold larger than \(\max(|S_a - S_b|, |S_a - S_c|)\) and a temporal threshold larger than \(\max(|T_c - T_d|, |T_d - T_b|)\) are used, both crashes B and C will be classified as secondary crashes. However, we can see that crash B was a secondary crash as it occurred within the (marked) impact area of the previous crash A whereas crash C was not a secondary crash because it did not occur within the impact area. Crash A and crash C obviously have to be classified as independent crashes. In order to identify secondary crashes along more highways, we cannot manually review these BSC and crash information. Therefore, an automatic identification algorithm has been developed to efficiently detect the potential secondary crashes. The algorithm is described below as a step-by-step procedure. The key idea is to detect whether later crashes (i.e., crash B & crash C) is located in the impact area of the prior crash (i.e., crash A).

Sub-step 4.1: Build line equation. For any two crashes \(E(S_{i,j}, T_{i,j})\) and \(L(S_{i,j}, T_{i,j})\), assuming that \(S_{i} > S_{i+1}\) and \(T_{i} < T_{i+1}\). This means crash \(L\) occurred later at the upstream of crash \(E\). Estimate the equation for the line connecting \(E\) and \(L\): \(S_{i} = \frac{S_{i}-S_{i+1}}{T_{i}-T_{i+1}} \times T_{i} - T_{i+1} + S_{i+1}\). For example, we can estimate the line equation for \(AB\) and \(AC\) in FIGURE 9.

Sub-step 4.2: Predict the coordination of special points on the line built in Step 4.1. These special points represent the intersection points between the line and the edges of each cell of BSCM. Assuming that the line is divided into \(K\) segments by \(K\) cells. The first segment and the last segment are determined by an intersection point and the coordinates of the crash. The remaining segments are determined by two adjacent intersection points, for example, points \((s_{j,t})\) and \((s_{j+1,t_{j+1}})\) shown in FIGURE 9.

Sub-step 4.3: Detect the \(K\) cells that contain the short line segments obtained in Sep 4.2. Given the intersection points \((s_{j,t})\) and \((s_{j+1,t_{j+1}})\), find out a cell \((S_{i}, T_{i})\) so that the following two criteria are satisfied:

- Criterion (a): find the minimum \(S_{i}\) such that \(S_{i} \geq s_{j+1}, \forall S_{i} \in \{S_{i}, S_{i+1}, ... S_{M-1}, S_{M}\}\);  
- Criterion (b): find the minimum \(T_{i}\) such that \(T_{i} \geq t_{j+1}, \forall T_{i} \in \{T_{i}, T_{i+1}, ... T_{M-1}, T_{M}\}\);
Sub-step 4.4: Check the binary speed measurement of each cell \((S_q,T_q)\) identified in Step 4.3. Find the summation of these speed measurements as 
\[
\sum_{s=1}^{K} V_s(S_q,T_q) = Q.
\]
If \(Q < K\) then this means that not all the short line segments are contained in the cells with speed \(V_s(S_q,T_q) = 1\). If \(Q = K\) then this means each short line segment is contained in a cell with speed \(V_s(S_q,T_q) = 1\). In other words, \(Q = K\) suggests that the built line between two crashes is located in the impact area of the primary crash, for example, line \(AB\) shown in FIGURE 9. Crash \(B\) is then classified as a secondary crash of the primary crash \(A\). Otherwise, the built line is not in the impact area, for example, line \(AC\) shown in FIGURE 9 and the corresponding crash \(C\) was not classified as a secondary crash.

The above algorithm was modified based on our previous identification algorithm (13) to improve the calculation efficiency. It serves as one of the key components of the identification approach. Other than the described steps, it is advised to follow the arrows in FIGURE 9 which will help understand the logic of the proposed approach in a simplified manner.

The developed identification approach enables users to take advantage of the third-party open source traffic data to explore secondary crashes in a larger scale. With the increased penetration of smartphones and GPS use among travelers, the quality of these third-party data is also expected to increase. This will provide more reliable traffic data sources for the virtual sensors based approach described in this paper. These traffic data in turn can support extensive investigation of secondary crashes on highways with limited or no infrastructure-based sensors. With the developed algorithm, the identification process will be automated and the identification performance will be improved as the real-time traffic information becomes ubiquitous. In practice, many existing traffic and transportation information systems can extend their functions to conduct real-time identification of secondary crashes using the developed approach. For example, the New Jersey’s 511 system shown in FIGURE 4 already displays live traffic speed and traffic incidents on its interactive map. In that figure, the first crash occurs at 11:29 am causes notable congestion and then another crash occurs in the queue at 11:38 am. Consequently, these two crashes are independently recorded in the police reports and shown on the map. If the system had implemented the proposed identification approach based on the live traffic speed and crash information from the interactive map, it would have detected such a primary-secondary crash pair in real time and alerted the travelers.

CONCLUSIONS

Recent research improved the traditional secondary crash identification approaches by using traffic data acquired from infrastructure-based sensors (13, 14, 31). These studies greatly improved the classification performance by mapping the prevailing traffic conditions as a result of primary incidents. However, the practical use of these approaches is still largely limited by the available infrastructure-based sensors on highways. Transportation agencies can only use these sensor-based approaches to examine secondary crashes on their highways if they have instrumented traffic sensors collecting continuous traffic data.

In order to address this issue, this study developed an on-line scalable approach to identify secondary crashes without relying on infrastructure-based sensors only. The developed approach has two major components: (a) collection of traffic data from third-party open source data and (b) development and implementation of an automatic "secondary crash" identification algorithm. This first component is an innovative framework designed to address the issue of traffic data availability. The concept of virtual sensors was proposed to acquire traffic data from various third-party traffic map services providers such as Bing Maps, Google Maps and MapQuest. The deployment of virtual sensors helps engineers easily collect traffic data over a large-scale sparsely instrumented transportation network. In our analysis, it was shown that the information quality of the virtual sensor output was acceptable in comparison with data from roadside traffic sensors. The second component of the approach proposes an automatic "secondary crash" identification algorithm based on the virtual sensor data. Instead of using fixed thresholds or simple queuing models to represent the impact area of crashes, the proposed algorithm dynamically models the actual progression of the impact area through a binary speed contour map. An automatic
detection procedure was then developed to examine the relationship between crashes that are likely to be
associated. The identification algorithm is described in a step-by-step manner so that the transportation
agencies/users can adopt it in practice.

For wider use of the developed approach, an extensive comparison between the infrastructure-
based sensor data and the virtual sensor output is necessary. This can help reveal the possible capabilities
and limitations of using virtual sensor data from private sector companies. For example, the output from
virtual sensor might not be as sensitive as the data from a loop detector in a rural area (because of low
penetration rate of GPS users). In addition, the potential of integrating the developed approach with many
existing traffic information systems to report secondary crashes in real-time deserves more attention. Last
but not least, the applications of the proposed approach in larger network analyses are advised to get more
tests. Its performances and sensitivity compared to the conventional methods deserve more exploration.

ACKNOWLEDGEMENTS
The contents of this paper only reflect views of the authors who are responsible for the facts and accuracy
of the data and results presented herein. The contents of the paper do not necessarily reflect the official
views or policies of sponsoring agencies. The authors appreciate the Region 2 University Transportation
Research Center (UTRC) for partially funding this study. The authors also thank for the insightful and
constructive comments and suggestions by the anonymous reviewers to help enhance this paper.

REFERENCES
Highway Administration, Office of Transportation Operations.
Success. Federal Highway Administration, Office of Operations.
Model to Assess the Safety Impacts of Implementing IVHS User Services. In: Proceedings of the
4th Annual Meeting of IVHS America, Atlanta, Georgia, pp. 343–352.
Traffic Management: An Investigation of Secondary Crash Causes. Journal of Intelligent
Operations: Top Five Issues". In: Proceedings of the National Conference on Traffic Incident
MNDOT0452.00. Minnesota Department of Transportation.
Transportation Quarterly 51(3), pp. 93–104.
Transportation Research Symposium, Ames, Iowa.


