Exploring the Utilization of Crash Data from Waze and Twitter for Official National Crash Database

Panji BROTOISWORO[a], Varsolo SUNIO[b*]

[a] *Independent Consultant; E-mail: panji.p.broto@gmail.com*
[b] *Department of Transportation, Clark Freeport, Mabalacat City, Pampanga, Philippines 2009; Email: vars.sunio@gmail.com*

Abstract: We explore the use of crash data from Waze and Twitter for populating the national crash database of the Philippines. DRIVER, the official government database of crash data, suffers severe under-reporting. Other online sources, such as Waze and an official Twitter page by a traffic management agency, have higher reporting frequency. Nonetheless, data variables which are normally included in official crash reports are missing in Waze and Twitter reports. In this paper, we examine the temporal and spatial match rates of the three data sources, as well as their reporting frequency and the relevance of data elements they contain. In terms of match rates among the reports, we find that for temporal bin size of 3 hours and spatial buffer radius of 0.5km, we achieve match of 41% (DRIVER-MMDA), 34% (DRIVER-WAZE), 60% (MMDA-WAZE). We discuss ways in which unofficial sources can enhance the official crash database.

Keywords: road accident, crash national database, data integration, Waze

1. INTRODUCTION

Globally, an estimated 1.35 million people die and 20-50 million are injured each year due to road crashes, making it a major public health priority issue (World Health Organization, 2018). In the Philippines, data show that the number of deaths attributable to road crashes has been rising, from 8,023 in 2011 to 8,761 in 2013 and then 10,012 people in 2015.

One of the most common contributing factors in road crashes is speed (Afukaar, 2003; Aarts and van Schagen, 2006; Soole, Watson, and Fleiter, 2013; Theofilatos and Yannis, 2014). The higher the speed of a vehicle, the shorter the time a driver has to stop and avoid a crash. Setting and strong enforcement of speed limits appropriate to the function of roads is thus recommended as one of the most effective interventions to stem traffic crashes (World Health Organization, 2004; Soole, Watson, and Fleiter, 2013; Gargoum and El-Basyouny, 2016). Aside from speed, other factors affecting road safety are: congestion and road horizontal curvature (Wang, Quddus and Ison, 2013), weather/climate (Theofilatos and Yannis, 2014; Mussone, Bassani and Masci, 2017) and other meteorological factors (Gao *et al*, 2016).

In 2018, the Philippine government released a directive reinforcing the guidelines and standards for the classification of roads and setting of speed limits appropriate for the type of roads. Apart from this, it also directed the collection, monitoring and analysis of road crash data at the nationwide level.

Crash data collection is crucial for road safety improvement, since the quality of decision making in road safety is dependent on the quality of the data on which decisions are based

*Corresponding author

(Montella *et al*, 2012). WHO has thus urged countries to design and develop traffic accident information systems and databases (World Health Organization, 2013). Across the world, national road crashes databases have been established by several countries in order to build evidence base for road safety improvement (e.g. Crash Analysis System by New Zealand, and Community Road Accident Database in Europe). Having these databases can improve our understanding of crash incidents, which is essential for better planning to save lives and avoid the wasting of resources (Mohammadi, Ahmadi and Gharagozlu, 2016).

2. ONLINE CRASH DATA SOURCES IN THE PHILIPPINES

While police, traffic management and hospital reports are rich and reliable sources of crash data in the Philippines, they are mostly based on traditional logbooks. These logbooks and reports include the blotter book, incident report form and traffic accident investigation report by the police, road crash report by the traffic management agency, and the patient injury form by the hospital. In this paper, our interest is with online sources of crash data only: DRIVER, MMDA Twitter page and Waze.

2.1. DRIVER: Official National Crash Database

Data for Road Incident Visualization Evaluation and Reporting (DRIVER) is the central and nationwide repository of road crash data (Fig. 1). Covering the entire country, it is an open-source, web-based centralized system to report traffic incidents. Collection and encoding of road crash data for road crashes is the responsibility of local government units that have jurisdiction over areas where the incidents occurred. Data from DRIVER are open and publicly accessible.

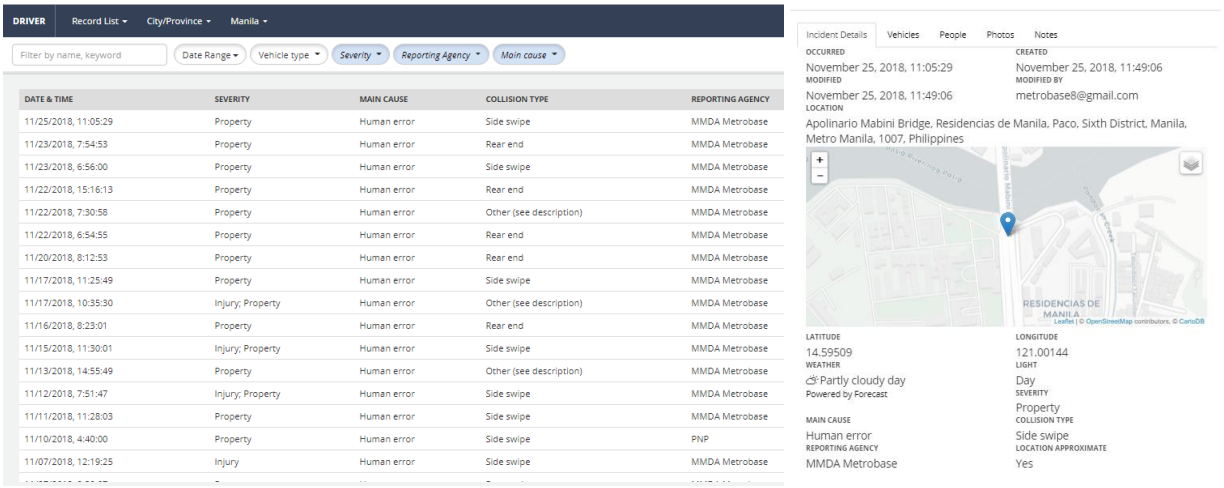


Figure 1. Crash-relevant information from DRIVER

It is well known in the literature that not all road crash incidents are reported to the proper authorities, which make official data incomplete and biased (Lopez *et al*, 2000; Amoros, Martin & Laumon, 2006; Watson, Watson, Vallmuur, 2015). In the case of DRIVER, heavy under-reporting has likewise been observed, since it also only relies on police and traffic management agencies of local government units for reporting crash data into DRIVER.

While road crash databases in other countries are linked to hospital data records (e.g. Watson, Watson, Vallmuur, 2015), the current version of DRIVER is not yet integrated with any hospital database system, such as the Online National Electronic Injury Surveillance System (ONEISS) of the Department of Health. Linking police, hospital and death records of road crash casualties can provide accurate outcome information for casualties in crashes reported to the police, as well as estimates of under reporting of crashes for different road user groups (Rosman, 2001). In other words, the use of multiple data sources will enable a more comprehensive understanding of the crash incidents (Watson, Watson, and Vallmuur, 2015).

2.2. Metro Manila Development Authority (MMDA) Official Twitter

The Metro Manila Development Authority (MMDA) is a government agency tasked to provide services within Metro Manila related to urban development plans, transport and traffic management, waste disposal, flood and sewer control, health and sanitation, pollution control, and public safety. It plays a big role in traffic management in the whole metropolis. Traffic-related incidents in Metro Manila are often reported to the MMDA and subsequently the MMDA Twitter page. This page serves as a traffic advisory channel of MMDA to the public (Fig. 2). Whereas DRIVER data cover the entire country, the MMDA data are only limited to Metro Manila. MMDA is one of the contributors into DRIVER of crash data that occurred in Metro Manila.

Figure 2. Sample of accident alert from the Official MMDA Twitter page. Elements in this alert will be explained below.

2.3. Waze Connected Citizens Program Platform

Waze is a GPS-based navigation application that is able to integrate data collected by users in order to guide others through traffic. Such collected data includes actively volunteered information on traffic congestion, police actions and accidents, as well as passively collected data on travel routes and speeds (Dos Santos, Davis, Smarzaro, 2017). Traffic data from Waze can be obtained as XML/JSON feed through its Connected Citizens Program Platform (CCP). Data from Waze CCP can also be shared publicly with proper attribution.

```
▼<item>
   <title>alert</title>
   <pubDate>Mon Dec 24 02:56:53 +0000 2018</pubDate>
   <georss:point>14.656633 121.02487</georss:point>
   <linqmap:uuid>357c3671-e159-3ab2-ae0f-d1e053885f22</linqmap:uuid>
   <linqmap:magvar>103</linqmap:magvar>
   <linqmap:type>ACCIDENT</linqmap:type>
   <linqmap:subtype>ACCIDENT_MINOR</linqmap:subtype>
   <linqmap:street>1: EDSA S</linqmap:street>
   <linqmap:city>Quezon City</linqmap:city>
   <linqmap:country>RP</linqmap:country>
   <linqmap:roadType>6</linqmap:roadType>
   <linqmap:reportRating>2</linqmap:reportRating>
   <nThumbsUp>0</nThumbsUp>
   <confidence>0</confidence>
   <linqmap:reliability>5</linqmap:reliability>
 </item>
```

Figure 3. Sample XML of Waze traffic data[1]

2.4. Summary

Table 1 enumerates the characteristics of each online crash data source in terms of reporting agent (government or crowd), reporting type (official or unofficial), elements (variables) and data access (public/open, restricted, private).

We highlight two points from the table. First, we classify the reporting of MMDA, a government agency, as unofficial. MMDA reporting is primarily for advising the public about possible heavy traffic due to a road crash incident; it is not principally for road safety improvement. Moreover, the reporting medium is an online news and social networking service. Second, examining the elements / variables included in the report, WAZE data include confidence/reliability and report rating scores which can be proxy for accuracy. DRIVER and MMDA data come from the government, so they can be reasonably considered accurate.

Table 1. Characteristics of the three online crash data sources

Crash Data Source	Agent	Reporting Type	Elements / Variables Available	Data Access
DRIVER	Government	Official	Date, Time, Severity, Cause, Collision Type, Location, Weather, Vehicles Involved, Persons Involved	Publicly available
MMDA	Government	Unofficial	Date, Time, Location, Direction, Accident Type, Vehicles Involved, Number of lanes occupied	Publicly available

[1] Refer to https://developers.google.com/waze/data-feed/incident-information for more information about the feed structure.

WAZE	Crowd	Unofficial	Location Coordinates, Unique System ID, Event Direction, Event Type (e.g. accident), Event Subtype (e.g. major/minor accident), Street, City, Country, Road type, Report Rating, Thumbs Up, Confidence, Reliability	Available to Waze CCP Partners only, but shareable to others with proper attribution

3. RESEARCH QUESTIONS AND OBJECTIVES

Following Amin-Naseri (2018), this paper aims to answer the following questions:

R_1: What are the characteristics of WAZE, MMDA and DRIVER data? How does the spatiotemporal coverage of each source compare to each other? How useful or relevant for official crash reporting are the data elements contained in WAZE and MMDA?
R_2: How does crowdsourced WAZE data compare to unofficial (i.e. MMDA tweet) and official (i.e. DRIVER) government data? What percentage of the recorded incidents by the government were detected by WAZE?
R_3: What is the estimated potential additional coverage that crowdsourced WAZE can provide to the official and unofficial government crash data? In areas where DRIVER and MMDA tweet have no coverage, can WAZE be trusted?

4. METHODOLOGY

4.1. Overview

Figure 4 shows the overview of the method used to answer the three research questions R_1, R_2, and R_3. Through the WAZE CCP, we generated an XML feed containing traffic alert data for the whole of Metro Manila[2]. There are at least seven WAZE alert types: accident, jam, weather hazard, hazard, miscellaneous, construction, and road closure. For this study, we only extract and save in a database all accident-related data using an XML Feed Parser, implemented on Python programming language that runs continuously at the background. We also create a Tweet2Map script that extracts tweets from the official MMDA Twitter, parses them, and saves them in a separate database. From these two databases, as well as the DRIVER database, we extract records covering the month of November 2018 only, and characterize them (this part of the methodology aims to answer R_1).

Since our objective is to examine if the crowdsourced WAZE records match well the government crash records, we specify a set of matching criteria: two records, each one belonging to one of the data sources, refer to the same crash incident if they are reported within a set time interval and distance. In other words, the matching algorithm first selects incidents in the temporal vicinity, then the geographic distance is examined. In Xavier *et al* (2016), a number of

[2] Data by Waze App. https://waze.com

similarity measures for geospatial matching of point data are enumerated, but in this paper we only consider temporal and spatial overlaps.

We perform temporal binning on the records from each dataset by grouping together crash incidents that occur within the same time window and placing them on the same time bin. Next we perform spatial matching. For same time bins, geo-coordinates of crash incidents belonging to one data source are compared to those belonging to another data source. Results of our matching procedure are then obtained and post-analyzed (this part of the methodology tackles R_2 and R_3). All source codes are available on GitHub[3].

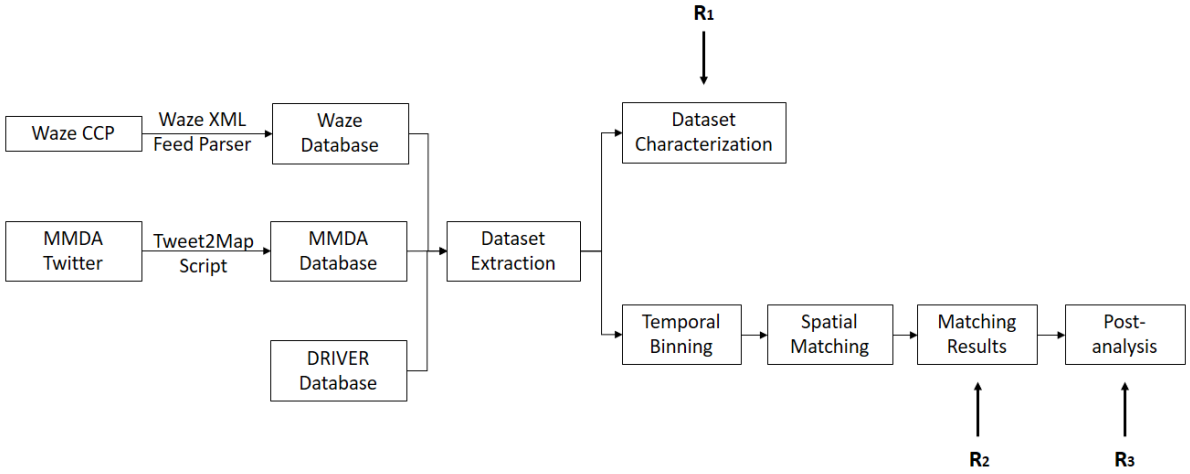


Figure 4. Overview of Methodology

4.2. Script: Waze Connected Citizens Platform XML Feed Parser and Tweet2Map

A python script was created to parse the elements – Timestamp, Unique ID, Title, City, Street, Latitude, Longitude, Type, Subtype, Confidence and Reliability – from the WAZE XML feed. Another script was made for the MMDA tweets. It is easy to notice that there is a specific structure in the tweets which can be seen in Figure 5. With this in mind we created a Python script called MMDA Tweet2Map which uses the Twitter API to automatically gather MMDA Tweets and saves the relevant information in the tweet to a CSV file.

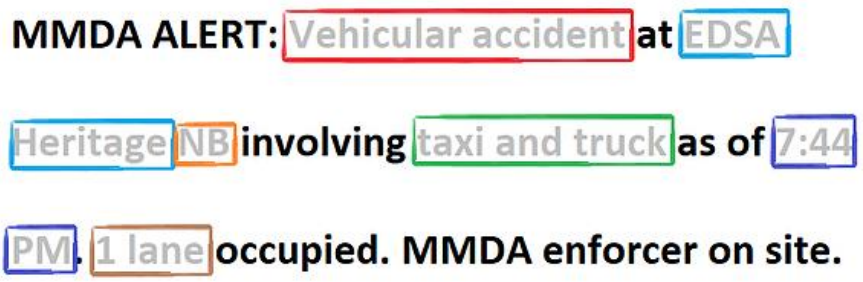


Figure 5. Structure of MMDA Alert Tweet. Red: Accident Type; Light Blue: Location; Orange: Direction; Green: Vehicles/road user involved; Dark Blue: Time; Brown: Lanes occupied.

[3] https://github.com/pbrotoisworo

Since the location of the alert post is in text format, we need to link it to a geographic location. For this purpose, a geocoding method was developed to predict the geospatial coordinates of latitude and longitude of the alert post (c.f. Gu, Qian and Chen, 2016; Melo and Martins, 2017). We build a database of locations that associate names with their respective latitude and longitude coordinates. This database is then used to retrieve coordinates that correspond to the name of the location. For example, a tweet that contains the location "EDSA Heritage" in Fig. 5 is annotated with its associated coordinates in decimal degrees, i.e. (14.536934, 120.993327). The coordinate location data are saved along with the tweet to the CSV.

4.3. Dataset Extraction and Characterization

From the three databases (DRIVER, MMDA, WAZE), we extracted datasets dating November 2018 on 3 December 2018[4]. We examine and characterize the dataset extracted in terms of (a) temporal and spatial coverage; and (b) relevance of the elements/variables contained in these datasets to crash data collection efforts (R_1). With regards to the latter, we recognize that minimum standard data requirements for crash incident reporting must be defined (c.f. Montella *et al*, 2012; Mohammadi, Ahmadi and Gharagozlu, 2016), and for this we use as reference the standards set in the Joint Memorandum Circular on "Guidelines and Standards for the Classification of Roads, Setting of Speed Limits Under Republic Act No. 4136, and Collection of Road Crash Data" by the Philippine government.

4.4. Temporal Binning

We constructed temporal bins of three sizes, 1 hour, 3 hours, and 24 hours. Hence, for the month of November, we have 30x24, 30x8, and 30x1 bins, respectively. Crash incidents occurring within the same time window are placed on the same bin. We vary the bin sizes to account for the possibility of time difference in reporting of the same incident by WAZE, MMDA and DRIVER. Consequently, we consider varying match levels (very close, close and very loose). Two crash incidents have "very close", "close" and "loose" temporal match if their respective timestamps are within 1-hour, 3-hour and 24-hour intervals, respectively.

4.5. Spatial Matching using ArcGIS Buffer Analysis

Spatial matching between records from two datasets is performed in two steps. After identifying which of the two datasets contain more records, we input the larger dataset into the ArcGIS and then create buffer polygons around input features to a specified distance of varying values (0.3km, 0.5km and 1km). The rationale is also to vary match levels (i.e. very close, close and very loose). We next load as points the dataset with fewer records. No buffer polygon is created on the second dataset. Next, we retrieve the points from the second dataset that lie within the buffer polygon created using the first dataset (Fig. 6). These are considered the records in the second dataset that have spatial and temporal match in the first dataset (R_2).

[4] Since the datasets used in this paper were extracted on 3 December 2018, incidents reported or inputted into DRIVER after this date – even though they occurred in November 2018 – were not included in the analysis. Extracting the datasets at a later time, e.g. February 2019, may result to a bigger dataset. This is a limitation of this paper.

Figure 6. Spatial matching by ArcGIS. We create buffer polygons around the points of the dataset with larger number of records (green circles with yellow center). Spatial match happens when points from the second dataset (represented by stars) lie within the buffer polygon.

4.6. Post-analysis

A post-analysis of a small sample of WAZE data that have spatiotemporal match with government data (i.e. MMDA tweet only) is carried out to extract their characteristics (R_3). By then assuming that only Waze data that possess these characteristics are trustworthy, we then estimate the extent of un-reporting in DRIVER and therefore the additional coverage that WAZE can provide.

5. RESULTS AND DISCUSSION

5.1. Reporting Frequency

We display the plot of the number/count of reports versus hour of the day for the month of November (Fig.7). Since WAZE relies on the crowd for reporting crash data, it has the highest reporting frequency, while DRIVER has the lowest – almost nil. We use two different axes for WAZE and MMDA/DRIVER.

Moreover, we observe that whereas WAZE and MMDA reports cover the entire day, DRIVER reports are only from 0400H-1500H.

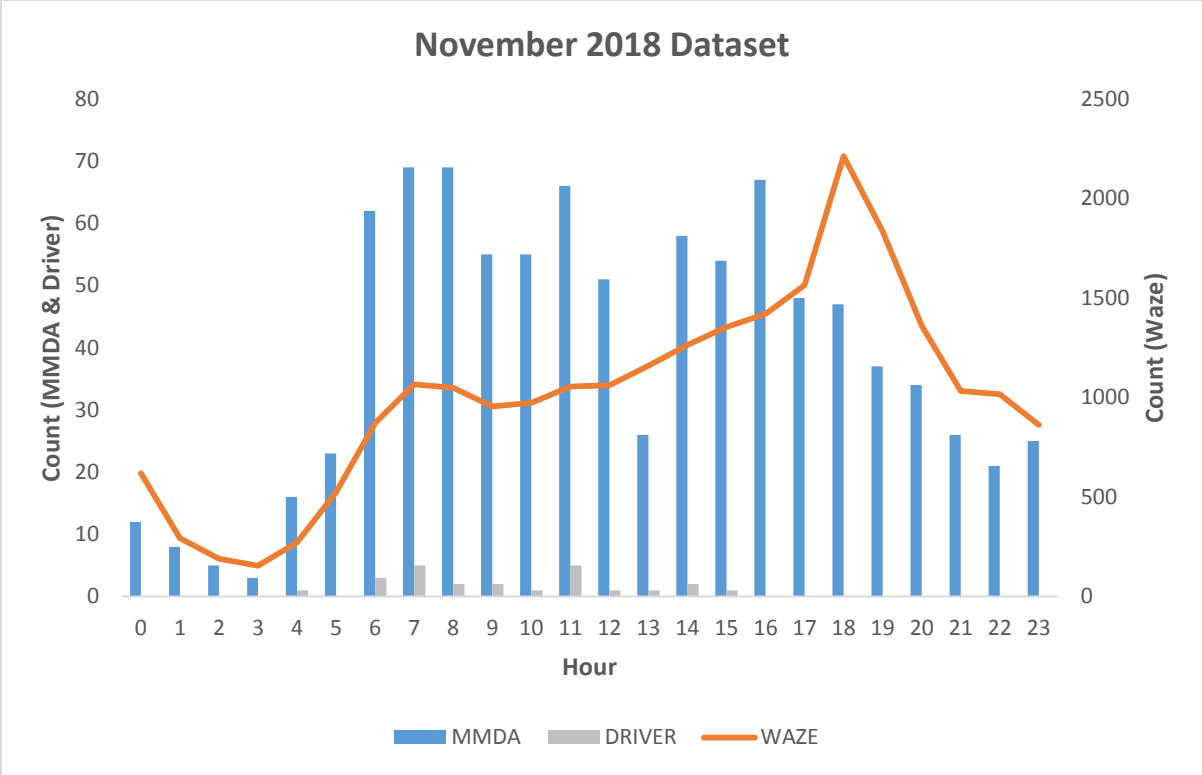


Figure 7. Number of reports by WAZE, MMDA and DRIVER in November 2018.

5.2. Spatial Coverage

In Figure 8, we observe that data points from WAZE cover the whole Metro Manila. MMDA data points are mostly along major thoroughfares (e.g. EDSA, C5, Quezon Avenue, Commonwealth Avenue; names of these roads are not shown in the map). This is explained by the fact that traffic enforcers from MMDA are stationed only along major roads, where there is a high volume of traffic. Data points from DRIVER have very limited geographical coverage.

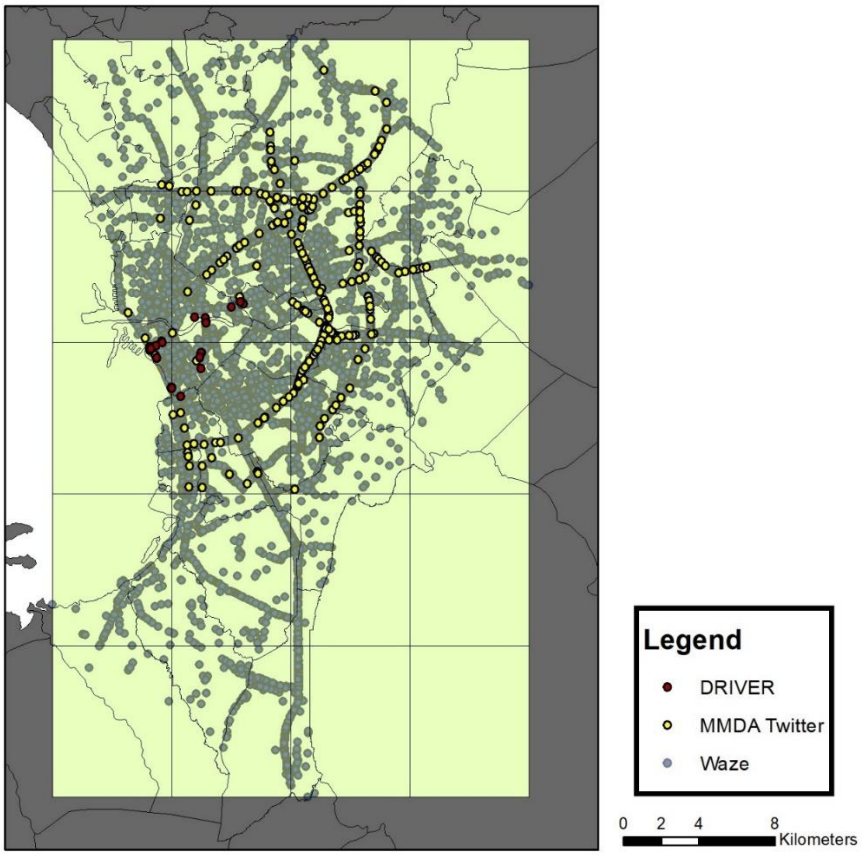


Figure 8. Spatial coverage of reports from the three datasets

5.3. Crash-relevant Data Elements

In the policy document by the Department of Transportation that was mentioned earlier, the government defines the minimum amount of information for crash incident reports. This is akin to the Minimum Uniform Crash Criteria Model (MMUCC) in the United States or Minimum Data Set proposed in Iran (Mohammadi, Ahmadi and Gharagozlu, 2016).

These crash-relevant data elements which the policy document specifies should be present are displayed in the first column in Table 2. By providing a minimum set of data elements of crash data, agencies are helped to collect reliable crash data. It must be noted that the standards set by the Department of Transportation conform to a certain extent to the minimum data requirements for crash reporting in other countries (c.f. Montella *et al*, 2012).

Nonetheless, other countries are also expanding the set of data elements to be included as standard into its Information Management System for Traffic Accidents. Iran, for instance, has proposed that the minimum data set for traffic police consists of 118 data elements; for trauma center 57 data elements; and for medical emergency center 64 data elements (Mohammadi, Ahmadi and Gharagozlu, 2016). In the table, DRIVER has complete information, whereas MMDA and Waze have significant crash detail deficiencies. This indicates that MMDA and Waze cannot be readily used for official crash reporting.

Table 2. Data elements in each data source and their value for official crash reporting

Data element	DRIVER	MMDA	WAZE
Date and Time	O	O	O
Location	O	O	O
Name of people involved	O		
Number of people involved	O		
Cause	O		
Collision type	O		
Severity (fatal, injury, property damage)	O		O
Vehicle or road user type (pedestrian, driver, etc.)	O	O	
Involvement of children (0-12)	O		

5.4. Matching Results

We apply the temporal and spatial match criteria to our three datasets. We compare two datasets at a time. By applying the matching criteria, we can retrieve the number of records in the dataset with fewer entries that have a temporal-spatial match with the larger dataset. Match rates are obtained by calculating the percentage of records from the smaller dataset with temporal-spatial match with records from the larger dataset. From Section 5.1, we observe that WAZE has the largest dataset, followed by MMDA, then DRIVER.

Table 3. Match rates of two data sources by time bin size and buffer radius with varying match levels (very close, close, very loose). % of total reports in (a) matched with (b)

Time Bin	Buffer radius	DRIVER[(a)]-MMDA[(b)]	DRIVER[(a)]-Waze[(b)]	MMDA[(a)]-Waze[(b)]
1 hour	0.3 km	33%	25%	34%
	0.5 km	44%	30%	42%
	1 km	44%	50%	57%
3 hours	0.3 km	32%	26%	49%
	0.5 km	41%	34%	60%
	1 km	41%	77%	77%
24 hours	0.3 km	38%	62%	78%
	0.5 km	46%	71%	86%
	1 km	46%	83%	92%

We first use a bin size of 1 hour and buffer radius of 0.3km. We relax this restriction and increase the sizes to 3 and 24 hours, and 0.5 and 1km. We do so because it is possible that a temporal lag exists in the reporting of the same incident to one database. Moreover, since there may be inaccuracy in the geo-tagging of incidents, we also vary the buffer radius. In the table, for a 3-hour time window and 0.3km buffer radius, 49% of MMDA data or 437 MMDA tweets in November 2018 have Waze match.

5.5. Post-analysis using Matched Waze Data

In WAZE, a user reports an incident, e.g. a car accident. The incident report becomes available on WAZE in real-time. An incident reliability score is calculated according to the WAZE users

experience level. The user experience levels are based on the map contributions providing an indication of user trustworthiness. As several users may simultaneously report the same incident, WAZE analyses and aggregates the data, contiguously providing the latest information. Reliability score ranges between 0 and 10, with 10 being the highest.[5]

Since crowdsourced data such as WAZE may not be accurate, we examine the reliability scores of WAZE data that have MMDA spatiotemporal match. Table 4 shows the reliability scores (mean and standard deviation) of Waze reports that are reported by MMDA tweet.

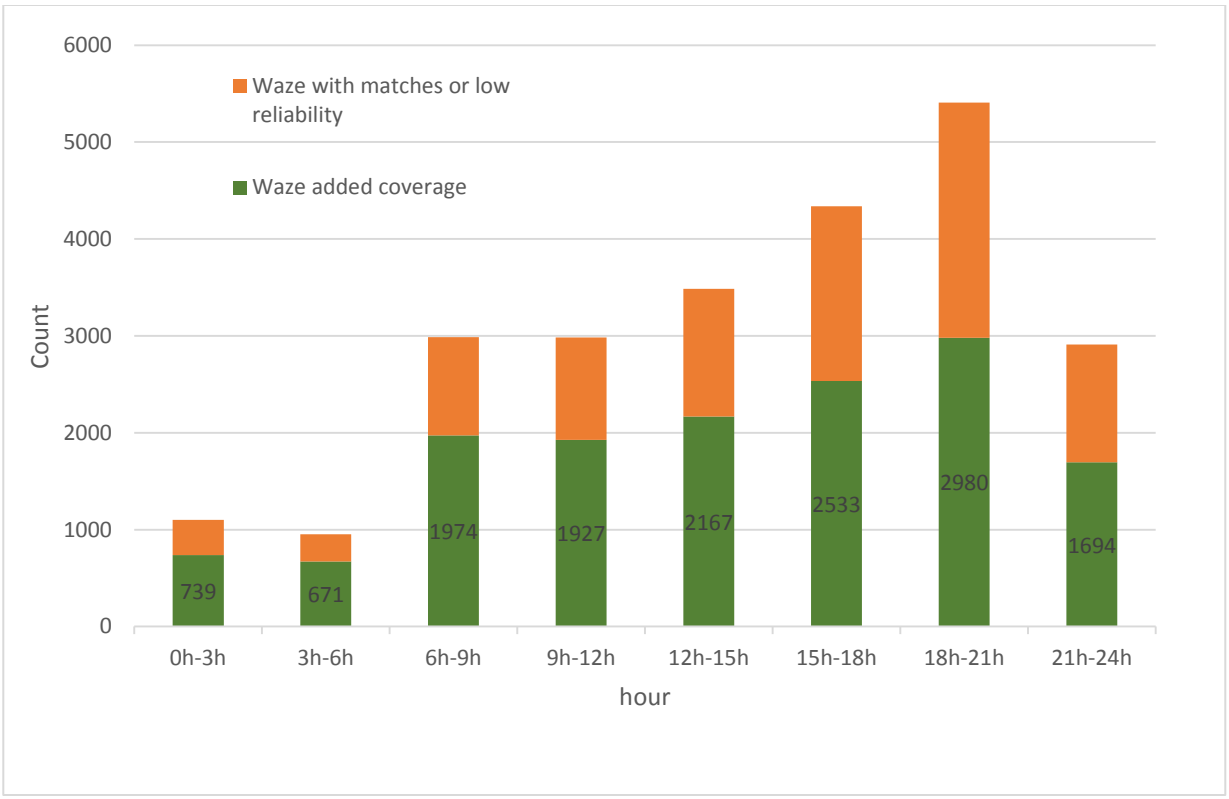
Table 4. Reliability scores of Waze reports matched with MMDA

Time Bin	**Buffer radius**	**Reliability (mean with standard deviation)**
1 hour	0.3 km	6.88 (1.90)
	0.5 km	6.86 (1.88)
	1 km	6.78 (1.85)
3 hours	0.3 km	6.84 (1.86)
	0.5 km	6.80 (1.85)
	1 km	6.73 (1.84)
24 hours	0.3 km	6.82 (1.87)
	0.5 km	6.76 (1.85)
	1 km	6.68 (1.81)

We observe an average reliability score of 6.68 and better. This suggests that WAZE crash reports with reliability score of 6 or more are valid. The Iowa Department of Transportation (DOT) also uses the same criteria: incidents with reliability>=6 are considered reliable (c.f. Amin-Naseri, 2018). Furthermore, we also notice in Table 4 that as we increase the time bin size and the buffer radius, the average reliability score decreases. This indicates that reports with higher reliability are more accurate.

Since one of the goals of utilizing crowdsourced data is to rely on them in locations where there are no other means for validation, we estimate using the preceding results the potential added coverage by WAZE. To determine this added coverage, we only retain WAZE crash reports that have no MMDA or DRIVER match as well those with reliability scores of 6 or more. Unlike the Iowa DOT, which apart from reliability score also uses report rating to evaluate how reliable the report is, in this paper we only consider the reliability score. Fig. 9 presents the added potential temporal and spatial coverage of Waze. In the figure, the green color corresponds to Waze-added coverage, while the orange color corresponds to Waze with DRIVER/MMDA matches or with lower reliability.

[5] https://support.google.com/waze/partners/answer/6324421?hl=en



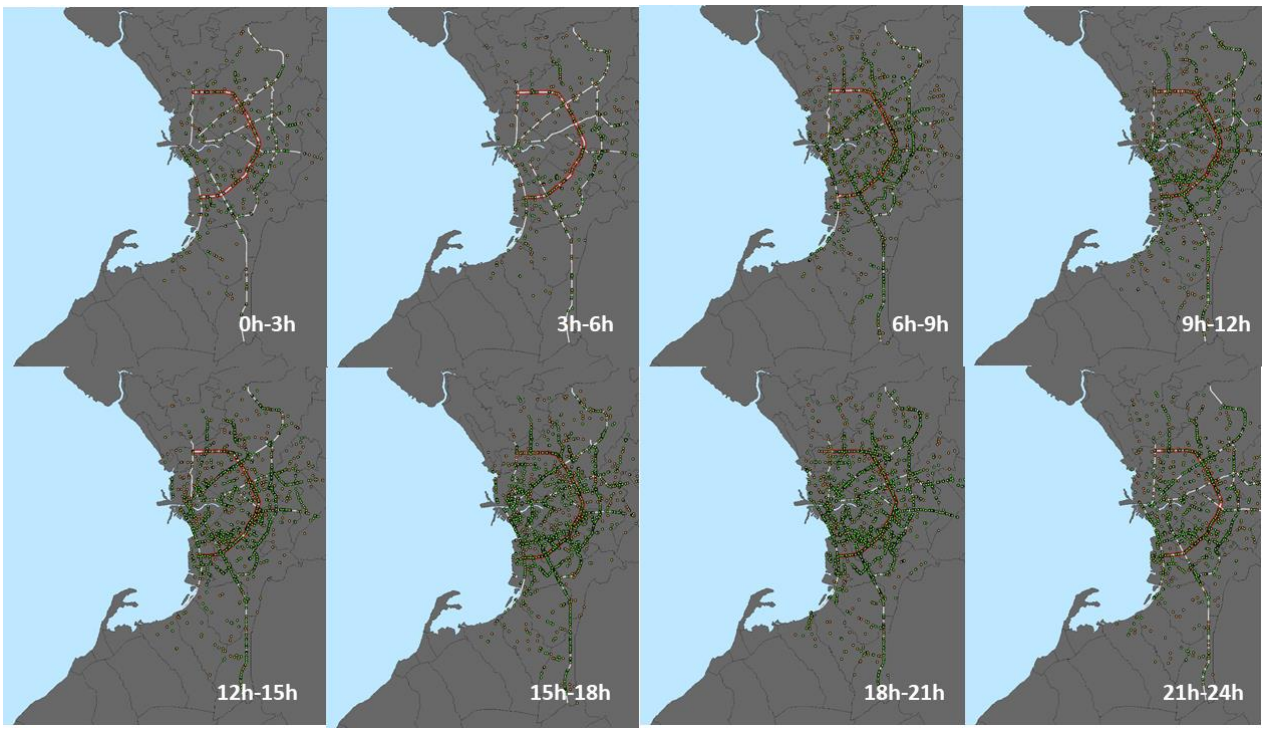


Figure 9. Added temporal and spatial coverage of Waze. The parameters are set to a window of 3 hours with a 500m buffer (total 1km). The green color corresponds to Waze-added coverage, while the orange color corresponds to Waze with DRIVER/MMDA matches or with lower reliability. The EDSA major road is marked with red border.

We note two things from the figure. First, the temporal coverage of the Waze-added contribution is over the entire 24 hours, and its spatial coverage is the whole of Metro Manila. This indicates that Waze is able to have a *unique* coverage over a broad range of time and space. Second, even in areas where we expect MMDA personnel to be present, Waze is able to detect crashes that are not recorded by either MMDA or DRIVER. We see this by zooming, for example, into EDSA, a major thoroughfare which is distinguished by red borders in the figure. Along EDSA, we observe the Waze-added contribution, even though MMDA has a significant presence on this road.

6. CONCLUSION AND RECOMMENDATION

This paper explores ways on how unofficial sources of crash data can complement official crash data collection efforts. Dos Santos, Davis and Smarzaro (2017) developed an algorithm to integrate official (police reports) and unofficial (WAZE) datasets in a city in Brazil. They demonstrated how the data from these two sources can be integrated in order to obtain a dataset with broader coverage. The only problem is that integration does not add any more information detail that is already contained in the datasets prior to integration.

In this study, we investigate the possibility of utilizing unofficial crash data as reported in WAZE and Twitter to enrich and populate the national crash database, DRIVER. Specifically, our objectives are:

- To understand the characteristics of WAZE, MMDA and DRIVER data, in terms of their spatiotemporal coverage and the relevance of the data elements contained in these data sources for official crash reporting (Research objective R_1)
- To compare the crash incidents as reported by WAZE, MMDA, and DRIVER and determine the percentage of the recorded incidents by the government that are detected by WAZE (Research objective R_2)
- To estimate the potential additional coverage that crowdsourced WAZE data can provide to the official and unofficial government crash data (Research objective R_3)

Since the main purpose of WAZE and the MMDA Twitter page is to alert drivers of congestion around the clock 24/7, this explains the high frequency of reporting observed in them over the entire day and night. Furthermore, in terms of spatial coverage, we observe that data points from WAZE cover the whole Metro Manila, while MMDA data points are mostly along major thoroughfares where there is a high volume of traffic. In contrast, DRIVER, the official crash database of the Philippines, is observed to suffer severe under-reporting, both in terms of geographical and temporal coverage. This is due to the fact that traditional crash incident reporting is labor-intensive.

Since WAZE, MMDA tweets and DRIVER data are collected for different purposes, this may also account for the fact that information about traffic alerts by WAZE and MMDA tweets does not contain much details normally found in official crash reports (e.g. persons involved). Consequently, as is, they cannot be integrated as data input to the official crash database, since DRIVER requires minimum set of data elements.

In terms of the detection rates by WAZE, we find that for 1-hour time bin and 0.3km buffer radius, 34% of MMDA reports and only 25% of DRIVER reports are detected by Waze. This percentage increases as we increase the size of the time bin and the buffer radius. For 24-hour time bin and 1 km buffer radius, 92% of MMDA and 83% of DRIVER reports are matched with

WAZE. We then characterize the WAZE reports that match MMDA and DRIVER reports, and find their average reliability score to be 6.68 and better. This suggests that WAZE crash reports with reliability score of 6 or more may be assumed to be valid. This also agrees with the criteria set by the Iowa Department of Transportation (DOT) for a reliable WAZE report (Amin-Naseri, 2018).

We also estimate the potential additional coverage that crowdsourced Waze to the government crash data. To determine this added coverage, we only retain WAZE crash reports that have no MMDA or DRIVER match as well those with reliability scores of 6 or more. We find that the temporal coverage of the Waze-added contribution is over the entire 24 hours, and its spatial coverage is the whole of Metro Manila. Moreover, even in areas where we expect the presence of MMDA personnel (e.g. EDSA), Waze is able to detect crashes that are not recorded by MMDA/DRIVER.

Given all these findings, we now evaluate the potential role of crowdsourced data (e.g. WAZE) and unofficial government data (e.g. MMDA Twitter) in official crash collection efforts.

6.1. Role of Crowdsourced Data (e.g. WAZE)

The purpose of WAZE is primarily to warn drivers of traffic-related incidents (such as jam, road closures, hazards, etc.) and not official crash data collection. Since it is crowdsourced, WAZE has a very broad temporal and spatial coverage: reports from the crowd are observed around the clock 24/7 and throughout the whole metropolitan area. While crowdsourced reports can be unreliable, WAZE associates reliability scores to every report – which makes it possible to filter out reports with lower reliability. Even if we are to drop WAZE reports with reliability lower than 6, the volume of remaining WAZE data is still considerable. Nonetheless, since its principal purpose is not crash data collection, data elements in WAZE related to crash are few.

All this suggests that Waze reports (with reliability of at least 6) can be used for timely detection of crashes, especially in areas or locations where police and traffic management officers are not normally deployed (e.g. secondary and tertiary roads), for dispatch of first emergency responders as well as collection of complete crash data.

6.2. Role of Unofficial Government data (e.g. MMDA Twitter Page)

Like Waze, the primary function of the MMDA Twitter page is also to warn drivers and not crash data collection. Reporting is done around the clock 24/7 but only along major thoroughfares in Metro Manila. Although they can be considered reliable, the data elements contained in these alerts that are related to crash are also few.

These alert posts are triggered when some netizens tag the MMDA about a possible traffic crash incident. The social media unit within MMDA then conducts further investigation and validates the incident using other sources such as the CCTV cameras or personnel deployed on the ground. Once confirmed, the MMDA tweets an official traffic alert post in its Twitter page.

Within MMDA, there is another unit, called Metrobase, which monitors and controls road activities throughout the whole metropolitan area, including collection and reporting of traffic incident data. Information about these incidents are sent to the social media unit for dissemination to the general public as well as reported to DRIVER[6]. However, examining the

[6] Metrobase of MMDA is one of the active contributors of crash data into DRIVER.

crash data retrieved from the official MMDA Twitter page and from the Metrobase (as reported in DRIVER) shows discrepancy: The Twitter page is observed to have generated more reports than Metrobase, which indicates the crash under-reporting by the latter.

All this suggests that the crash reports posted by MMDA Twitter page can be used by the Metrobase as one reference for the number of crash incidents that actually occurred along the major roads of Metro Manila. Since these MMDA Twitter reports do not contain complete crash information, Metrobase can fill up the missing entries –through further investigation via CCTV cameras or personnel deployed on the ground – so that they can be readily incorporated to official reporting database. Coordination between the social media unit, which maintains the Twitter page, and the Metrobase should not be a problem since both units belong to the same agency.

6.3. Recommendations

For future studies related to road crash and improvements of the existing database system, we recommend the following:

First, we suggest inclusion of more data elements defining a minimum set of requirements for comprehensive reporting. As of the moment, discussion is ongoing on integrating more than 200 data elements into DRIVER platform. This will entail the cooperation of several agencies of the government (e.g. Philippine National Police, Department of Health, MMDA, etc.) reporting crash incidents diligently into the same platform. Alternatively, if these agencies have their own databases, we recommend linking databases together to produce more useful crash-relevant information (e.g. Rosman, 2001). The use of multiple data sources will enable a more comprehensive understanding of the crash incidents (Watson, Watson, and Vallmuur, 2015).

Second, it is recommended that the MMDA Road Safety Unit upgrade into using DRIVER instead of MMARAS database (see footnote 3 where MMARAS is first mentioned). DRIVER enables the MMDA Road Safety Unit to standardize road crash reporting, plot crashes onto a map, automate analysis such as heatmaps and spatial autocorrelation techniques. This cannot be achieved through the MMARAS platform which is merely a spreadsheet. To avoid duplication, the MMDA as a whole should work together and decide which of the three offices within MMDA (namely, MMDA Metrobase, MMDA Road Safety Unit, and MMDA social media unit) will be the ones who will be recording crashes.

Third, for future study, it is recommended to investigate the demographic and characteristics of the users of Waze and those who record crash data in that platform. Although the dataset is rich, there might be inherent biases that can affect the analysis of road crashes. This can be in terms of space or time or other factors. For example, it might happen that a lot of reported crashes in Waze will gravitate in a certain corridor such as EDSA which then might over-represent EDSA compared to other roads which can be more dangerous but not well-known. Another example is that nighttime crashes can be underrepresented because Waze users mostly use the application during daytime. This has huge implications in terms of network planning in road safety.

ACKNOWLEDGEMENTS

We appreciate the help of a colleague who wishes to remain anonymous in the content development of the manuscript.

REFERENCES

Aarts, L., & Van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, *38*(2), 215-224.

Afukaar, F. K. (2003). Speed control in developing countries: issues, challenges and opportunities in reducing road traffic injuries. *Injury control and safety promotion*, 10(1-2), 77-81.

Amin-Naseri, M. (2018). Adopting and incorporating crowdsourced traffic data in advanced transportation management systems (Doctoral dissertation, Iowa State University, Ames, Iowa, United States of America). Retrieved from http://www.imse.iastate.edu/files/2018/07/Amin-NaseriMostafa-dissertation.pdf.

Amoros, E., Martin, J. L., & Laumon, B. (2006). Under-reporting of road crash casualties in France. *Accident Analysis & Prevention*, *38*(4), 627-635.

Dos Santos, S. R., Davis Jr, C. A., & Smarzaro, R. (2017). Analyzing Traffic Accidents based on the Integration of Official and Crowdsourced Data. *Journal of Information and Data Management*, 8(1), 67.

Gao, J., Chen, X., Woodward, A., Liu, X., Wu, H., Lu, Y., Li, L., & Liu, Q. (2016). The association between meteorological factors and road traffic injuries: a case analysis from Shantou city, China. *Scientific reports*, 6, 37300.

Gargoum, S. A., & El-Basyouny, K. (2016). Exploring the association between speed and safety: A path analysis approach. *Accident Analysis & Prevention*, 93, 32-40.

Gu, Y., Qian, Z. S., & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies*, 67, 321-342.

Lopez, D. G., Rosman, D. L., Jelinek, G. A., Wilkes, G. J., & Sprivulis, P. C. (2000). Complementing police road-crash records with trauma registry data—an initial evaluation. *Accident Analysis & Prevention*, *32*(6), 771-777.

Melo, F., & Martins, B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21(1), 3-38.

Mohammadi, A., Ahmadi, M., & Gharagozlu, A. (2016). Developing a minimum data set for an information management system to study traffic accidents in Iran. *Iranian Red Crescent Medical Journal*, 18(3).

Montella, A., Andreassen, D., Tarko, A. P., Turner, S., Mauriello, F., Imbriani, L. L., Romero, M. & Singh, R. (2012). Critical review of the international crash databases and proposals for improvement of the Italian national database. *Procedia-Social and Behavioral Sciences*, 53, 49-61.

Mussone, L., Bassani, M., & Masci, P. (2017). Analysis of factors affecting the severity of crashes in urban road intersections. *Accident Analysis & Prevention*, 103, 112-122.

Rosman, D. L. (2001). The Western Australian Road Injury Database (1987–1996): ten years of linked police, hospital and death records of road crashes and injuries. *Accident Analysis & Prevention*, 33(1), 81-88.

Soole, D. W., Watson, B. C., & Fleiter, J. J. (2013). Effects of average speed enforcement on speed compliance and crashes: A review of the literature. *Accident Analysis & Prevention*, 54, 46-56.

Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*, 72, 244-256.

Wang, C., Quddus, M. A., & Ison, S. G. (2013). The effect of traffic and road characteristics on road safety: A review and future research direction. *Safety science*, 57, 264-275.

Watson, A., Watson, B., & Vallmuur, K. (2015). Estimating under-reporting of road crash injuries to police using multiple linked data collections. *Accident Analysis & Prevention*, 83, 18-25.

World Health Organization. (2004). *World report on road traffic injury prevention*. Retrieved from https://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/en/.

World Health Organization. ((2013 .*WHO global status report on road safety 2013: supporting a decade of action*. Retrieved from http://www.who.int/iris/handle/10665/78256.

World Health Organization. (2018). *Road traffic injuries*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

Xavier, E., Ariza-López, F. J., & Urena-Camara, M. A. (2016). A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys (CSUR)*, 49(2), 39.