Analyzing Traffic Accidents based on the Integration of Official and Crowdsourced Data

Salatiel Ribeiro dos Santos[1], Clodoveu Augusto Davis Jr.[1], Rodrigo Smarzaro[1]

Universidade Federal de Minas Gerais, Brazil
{salatiel.ribeiro,clodoveu,smarzaro}@dcc.ufmg.br

Abstract. Geographic information of public interest is routinely produced by several public agencies. At the same time, the use of smartphones and other mobile devices generates an increasing amount of unofficial georeferenced data. Although official data have usually higher reliability, it takes longer for governmental organizations to put together relevant datasets and make them available, while the opposite occurs with unofficial data. This work explores the potential for integrating data from official and unofficial sources, as part of a project that aims to verify possible roles for unofficial or crowdsourced data, as replacements or to complement official sources. The article presents a case study with two traffic accident datasets in the city of Belo Horizonte, Brazil. We compare official traffic accident data to unofficial data collected from Waze, a mobile app dedicated to helping users fight traffic congestion. We found that seven percent of accidents reported by official sources have also been reported by users of Waze. Accidents reported only by official sources are concentrated in the central region, while those recorded by Waze are mostly on some major roads all over the city. An analysis on the possible influence of weather is also presented, as well as the identification of accident hotspots from the integrated dataset.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Spatial databases and GIS; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: accidents, Belo Horizonte, spatial data integration, waze

1. INTRODUCTION

Amid their institutional responsibilities, several public organizations produce geographic data of general interest. However, due to operational or technological difficulties, part of these data does not become accessible to the public, or is published in formats that preclude their dynamic integration to other data sources, thereby making it harder to accomplish more elaborate analyses. In Brazil, in spite of the approval of the Law on Information Access in 2011[1], most governmental data producers still have no clear open data policy or practice, nor do they implement technological resources such as APIs and service-based spatial data infrastructures (SDI) to foster easy access to data that are relevant to the society at large.

On the other hand, the intensive use of smartphones and other mobile devices generates a significant volume of data, since wherever people go their trajectories can be recorded, and there are many ways for them to express themselves on the visited places [Zheng et al. 2009]. Simultaneously, the interest

[1] in Portuguese, Lei de Acesso à Informação (12.527/2011). Available at http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

This is an extended version of work presented at Geoinfo 2016. Along with many localized improvements, analyses on Section 5.3 were expanded and deepened, and Sections 5.4 and 5.5 were introduced.

Authors acknowledge the support of CNPq and FAPEMIG, Brazilian agencies in charge of fostering research and development. We also thank BHTrans for the official dataset, Patrick Prado for the Waze dataset, and professors Cláudia A. Soares Machado and José Alberto Quintanilha for their important contributions.

in geographic data and geographic applications grows, as they enable users to locate themselves and to find events and other people based on their location [Quercia and Capra 2009; Quercia et al. 2010]. In this scenario, users routinely issue comments and opinions through social networks (e.g., Twitter and Facebook), commenting on tourist attractions and commercial venues (Foursquare[2], Yelp[3]), publishing personal videos and photos (Instagram[4], Snapchat[5], Flickr[6]) or sharing real-time information on traffic (Waze[7]). With a large number of contributors, such applications become important, albeit unofficial, sources of information.

This work intends to explore the potential for integrating official and unofficial sources of data on urban events, in order to verify to which degree data generated by the usual work processes in public organizations can be confirmed or enhanced by data generated by volunteers, in crowdsourcing or crowdsensing systems. The possibilities for adding, expanding or replacing official data sources with unofficial ones is analyzed, with the additional challenge of treating the intrinsic heterogeneity of such data sources.

Therefore, the main objective of this work is to demonstrate the need and usefulness of integrating multiple and heterogeneous data sources to support analyses of urban phenomena, in order to support decision-making related to improvements in urban services. We present a case study that uses data on traffic accidents in the city of Belo Horizonte, Brazil, obtained from official sources and from Waze. Both sources contain geographic information on the location of accidents, but other descriptive attributes are quite distinct. We also present a brief analysis on the influence of weather in the number of accidents during the period covered by the data. Weather data were obtained from the Brazilian meteorological service (INMET) for rainfall in the city of Belo Horizonte, Brazil, in 2014.

This article is structured as follows. Section 2 describes related work. Section 3 presents a comparison between official and unofficial sources of traffic accident data. Section 4 describes our integration methodology. Results are presented and discussed in Section 5. Finally, Section 6 concludes the article and presents ideas for future work.

2. RELATED WORK

A study by the World Health Organization (WHO) [WHO 2015] indicates that traffic accidents are a serious problem worldwide. According to that study, in 2013 alone 1.25 million deaths occurred due to road accidents in 180 countries, especially in those with low income. Another alarming information is that traffic accidents represent the main cause of death among people aged between 15 and 29 years. Also, Brazil is ranked in position 56 among the countries with the highest number of deaths caused by traffic accidents, in spite of recent improvements in legislation, such as a strict law on alcohol consumption by drivers [Brasil 2008], and of permanent educational campaigning for traffic safety.

Reducing traffic accidents requires improvements in vehicles, road infrastructure and personal behavior, thus the responsibility of changes is shared among politicians, managers, road designers, vehicle manufacturers and citizens that use the road system. Professional and amateur drivers have access to several technologies for routing in streets, roads and highways, including real-time information on accidents and traffic events. On the other hand, as citizens we perceive the lack of objective and useful information that allows drivers to take instantaneous notice of problems and adopt measures against accidents in critical points along urban streets and highways.

[2] http://www.foursquare.com
[3] http://www.yelp.com
[4] http://www.instagram.com
[5] http://www.snapchat.com
[6] http://flickr.com
[7] http://www.waze.com

Data produced by governmental organizations, or data contributed voluntarily or unconsciously by common citizens, help in understanding city dynamics and user preferences in moving through urban space. Such information are instrumental in decision-making for improving infrastructure and can contribute to improvements in life quality [Smarzaro et al. 2017; Corsar et al. 2015; Wolf and Fry 2013]. In the current context, in which solving traffic and transportation problems in large urban centers becomes even more urgent, it is important to promote the use of alternative transit so that the number of vehicles in the streets can be reduced. However, it is necessary to provide minimal security conditions for people that opt for the alternative transportation methods. Machado et al. [2015] identify points in which traffic accidents are concentrated for the city of São Paulo (Brazil) and Rome (Italy). The focus of the study are accidents involving individuals traveling by non-motorized means (on foot or by bicycle). From the results of the study, users of alternative transportation can be extra careful or avoid completely such dangerous regions. The article invokes the discussion on traffic accidents in Brazil.

Even though there are public data on accidents in Brazil, the absence of a nationally integrated database is remarkable. Bezerra et al. [2015] present a number of difficulties inherent to Brazilian sources that have an impact on the establishment of such an integrated database. Among these difficulties, authors highlight the way federative agents are structured and the distribution of responsibilities among them. Brazil has a mixture of federal highways, state highways and urban thoroughfares in charge of local governments. In the first case, accidents are recorded and followed up by many organizations, in particular the national highways department (DNIT), the Ministry of Transportation, the Ministry of Cities and the Ministry of Health, along with the Federal highway police. In each state, there is a transit department, a state highway police, a military police, and administrative organizations in charge of transit and transportation. As to municipalities, while the largest ones maintain transit and transportation engineering companies, the less populous ones usually have no means to record and analyze accidents.

The diversity of traffic- and transit-related organizations would not be an important obstacle if there was a unified process for recording traffic events. Bezerra et al. [2015] highlight that there is not even a standardized police report form. Nevertheless, event reports are the largest source of information currently available. Undernotification of traffic accidents is also expected, since involved parties seek official reporting of the event mostly in case there are victims or the need to sustain insurance claims. Authors also indicate difficulties for data analysis in accident reports, due to incompleteness, coding errors, discontinuity and lack of elements with which to locate the accident. Highway accidents, for instance, tend to be reported with a linear reference, such as a distance from a kilometer marker. However, positioning in a linear referencing system tends to be hard to transpose to a geographic point, due to accumulating length distortions along the linear representation of the highway in a map [Scarponcini 1999].

In other countries, accident data are treated in a much different manner, leading to well-grounded analysis works. Morris et al. [2010] present the creation of a database on fatal traffic accidents in Europe. Many works use United States governmental sources to diagnose problems such as lack of attention while driving, hitting pedestrians, light vehicle crashes [Najm et al. 2003] and effects of driver population aging [Stamatiadis and Deacon 1995]. Such analyses are strongly hindered in Brazil due to the lack of a nation-wide system for recording traffic accidents.

Recent initiatives have attempted to obtain accident data from unofficial sources. Using Twitter data, Ribeiro Jr. et al. [2012] geolocate traffic-related events based on the content of posts. From that, traffic accidents, congestions and interruptions can be identified and mapped. Salazar et al. [2015] also propose a method for geocoding traffic-related events based on short Twitter texts, using a heuristic that is able to locate and to infer the nature of an event using the number of mentions to urban places or thoroughfares in the message. Also using unofficial data, Silva et al. [2013] detect traffic conditions in urban roads using Waze data. They also discuss limitations related to the source, including its coverage.

Integrating the various sources of official data and combining them with unofficial sources is an important problem for initiatives related to Brazilian traffic problems. Integration may help solving problems such as under-notification and characterization, and promoting unified access to official reports. A major goal is to obtain an integrated dataset that can be used in diagnosing and analyzing events such as those reported in the works listed in this section. The next section describes two datasets from the city of Belo Horizonte that are used in a case study for the integration of official and unofficial data. Following that, a methodology for integration is discussed.

3. DATASETS

3.1 Official Data

Accidents are usually reported to the authorities in charge of traffic or public security, who, in turn, record the event in police reports, incident reports or similar documents. Such information are kept in databases by the authorities at the federal, state or local levels. In Brazil, accidents in urban thoroughfares are recorded by municipal authorities. Accidents in state or municipal highways are recorded by the respective administrative levels. Accidents in federal highways can be recorded by authorities at any level, depending on existing administrative agreements. Accidents on segments of federal highways that cross urban areas are typically recorded by the state's military police.

This work uses accident data for the city of Belo Horizonte, Brazil, in 2014, as supplied by the municipal transit company, BHTrans. These data are well structured and reliable, since they originate in police reports, filed by the state's military police. However, these are the latest data available for analysis, since BHTrans is still unable to release the 2015 data at the time of this writing (April, 2017). The dataset contains 1,434 accident reports that took place between September 16 and November 11, 2014, in Belo Horizonte. The data contain, along with a precise location for each accident (based on the city's spatial database of addressing and thoroughfare network), a timestamp, and classification according to a set of accident types, and some additional descriptive attributes.

3.2 Unofficial Data

We classify as unofficial those data that come from social networks, active or passive crowdsourcing or crowdsensing applications [Mateveli et al. 2015] or any other source that is not connected to governmental institutions. In this work, we use data from Waze, 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.

The lack of an API for data collection is a serious shortcoming of Waze. Obtaining Waze data without an API requires monitoring the app's Web-based live map in small areas, and extracting relevant information from the underlying JSON files. As in the case of many crowdsourcing or volunteered geographic information (VGI) applications, Waze also suffers from lack of detail, questionable reliability of contributing users, and irregular spatial coverage. The validity of Waze information can partially be assessed by confirmation from other users.

The main advantages of Waze are the timely access to accident data, which allows users to plan trips that avoid congested areas. This strongly contrasts with data publication policies by official transit authorities. Waze is also expected to record accidents that are not officially communicated to transit or police authorities, which is the case of less serious incidents or accidents involving uninsured vehicles.

We collected Waze data on accidents and filtered our records to match the period from which official data was available, i.e., September 16 to November 11, 2014. Each record includes a timestamp, a location and a reported severity, as perceived by the contributor. Many accidents were reported by more than one contributor, thus requiring a consolidation step, described in more detail in the next section. After the consolidation, the dataset contained data on 1,543 accidents.

4. METHODS

The first step towards the implementation of this work is data acquisition. As mentioned, traffic accident reports for 2014 were provided by BHTrans. Besides geolocation, accident data includes date, time, type of accident and vehicles involved.

The unofficial data was obtained from Waze in 2014, as part of a data collection experiment in a project that intended to map frequently congested areas. Specifically, for this work we selected accident reports from the 2014 dataset. From each accident alert it is possible to obtain information that is similar to official BHTrans data, such as geolocation, date, time and type of accident.

Since Waze does not have an official API, data was collected through a GeoRSS file[1] generated by the Live Map at the application's Web site[2]. A JSON file is downloaded in regular intervals, containing real-time geographic features and locations of objects and, in this case, data on traffic congestion and alerts. Live Map can hide some alerts, depending on zoom settings. This is the main limitation on the data collection process, since it reduces data coverage. Furthermore, since JSON collection is re-executed frequently, the series of JSON files needs to be processed to eliminate duplicate incident reports.

The second step of the process is data processing: select data in the same time window (both between 2014-09-16 and 2014-11-06), group accidents reported more than once (by different users) on Waze and geocode records without geographical coordinates (some BHTrans records included the textual description of a location but not actual coordinates).

In order to consolidate accidents reported more than once on Waze, the average position of the reports is calculated and the number of contributions that each of the accidents received is noted. The resulting datasets contain 1,434 and 1,543 accident reports, respectively in BHTrans and Waze datasets.

Finally, the last step integrates data from the official and unofficial sources. In this step, data are checked for overlapping events. Furthermore, the characteristics of data coming specifically from either source are explored, thus verifying how complementary they might be. We established a set of matching criteria, by which two records, each one belonging to one of the data sources, refer to the same accident if they were reported within one hour of each other and occurred (1) within 50 meters of each other, or (2) within 150 meters on the same road (see Algorithm 1). At the end of the process, a single integrated dataset on accidents is created, containing annotations about the origin of each data item. We considered one hour a reasonable time interval for match criteria as there might be a delay between the time the accident actually happened (informed by the involved parties in a police report) and the time the accident was reported on Waze. The second criteria was adopted because there might be situations in which the Waze user is passing by the accident location and reports it in a position that is away from where the accident actually happened. In this case, if the records are at the same road, the distance limit is set to 150m, as it is very likely that the same accident has been reported. We do not assume this when the roads are different, and adopt a more restrictive distance of 50m for this case to prevent false positive matches.

[1] http://www.georss.org/

[2] https://www.waze.com/pt-BR/livemap

Algorithm 1 Integration Algorithm

```
function INTEGRATION
    $wazeAccidentsSet \leftarrow$ set of waze accidents
    $bhtransAccidentsSet \leftarrow$ set of BHTrans accidents
    $matchedAccidentsSet \leftarrow \emptyset$
    for each wazeAcc ∈ wazeAccidentsSet do
        for each bhAcc ∈ bhtransAccidentsSet do
            if DATETIMEDIFFERENCE($wazeAcc.DateTime$,$bhAcc.Datetime$)<=60 min then
                if DISTANCE($wazeAcc.Location$,$bhAcc.Location$)<=50 m then
                    $matchedAccidentsSet \leftarrow$ wazeAcc,bhAcc
                else
                    if DISTANCE($wazeAcc.Location$,$bhAcc.Location$)<=150 m and
                        wazeAcc.StreetName == bhAcc.StreetName then
                        $matchedAccidentsSet \leftarrow$ wazeAcc,bhAcc
                    end if
                end if
            end if
        end for
    end for
    return $matchedAccidentsSet$
end function
```

Table I. Number of dataset records

Dataset	# of records
BHTrans	1,333
Waze	1,442
Matches	101
Integrated	2,876

We evaluated the accidents reported in both datasets, the notifications contained only in the official data and the accidents contained only in the unofficial data. Next section presents and discusses the results of the matching process.

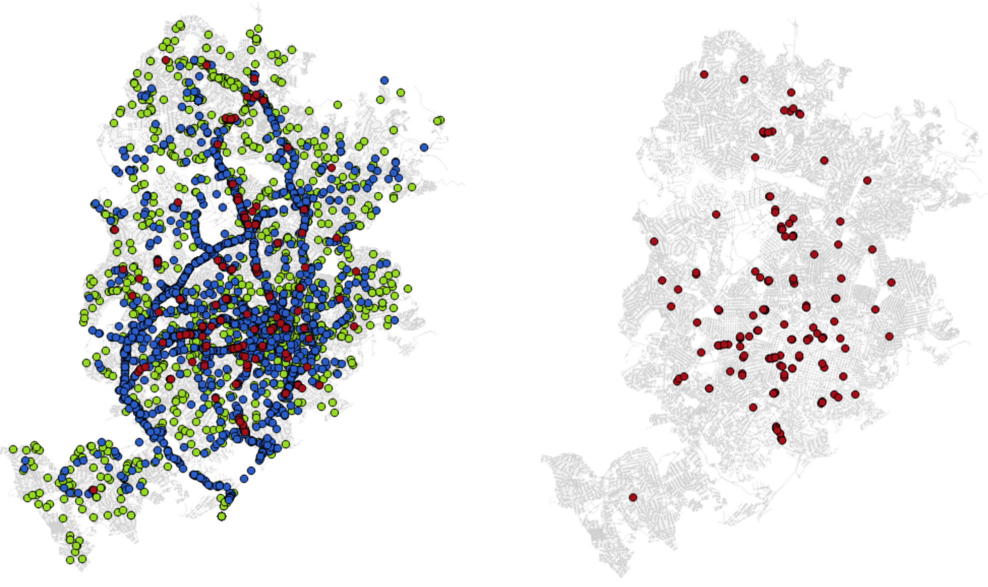
5. RESULTS AND DISCUSSION

After matching the two sources, an integrated dataset on accidents in Belo Horizonte was built. Table II shows its structure, indicating the attributes that were obtained from each individual source, plus an attribute that records the source of the original information. Table I shows the number of records in each dataset considered. There are 1,333 accidents that were reported exclusively to BHTrans, 1,442 gathered exclusively from Waze, and 101 that have been matched from both sources. Since the number of matching records is small, integrating Waze data to the official dataset represents a 100.5% increase in the overall number of accident reports. The integrated database describes 2,876 accidents in 52 days, or 55 accidents per day on average. The location of all accidents can be seen on Figure 1a while Figure 1b shows only the matched accidents from both sources.

This section presents our analyses on the integration process, regarding the characteristics of matched and unmatched records. We proceed with analyses on the severity of accidents, reported with different criteria by the two sources, on the temporal distribution of accidents, and on the influence of weather conditions. Lastly, we identify accident hotspots in the city, considering the integrated dataset. These analyses show the spatial and temporal distribution of accident spots, and seek to char-

Table II. Integrated accidents dataset

Integrated Attributes	Sources		Remarks
	BHTrans	Waze	
bhtrans id	police report id	—	—
waze id	—	alert id	—
date	date	date	—
street type	street type	—	—
street name	street name	street name	use BHTrans if available
number	number	—	—
neighborhood	neighborhood	—	—
region	region	—	—
city	—	city	—
country	—	country	—
geom	longitude, latitude	longitude, latitude	average position
type of accident	type of accident	—	—
severity	—	severity	—
number of victims	number of victims	—	—
number of deaths	number of deaths	—	—
vehicles involved	vehicles involved	—	number and type of vehicles
data source	—	—	BHTrans, Waze or both



a) Integrated dataset

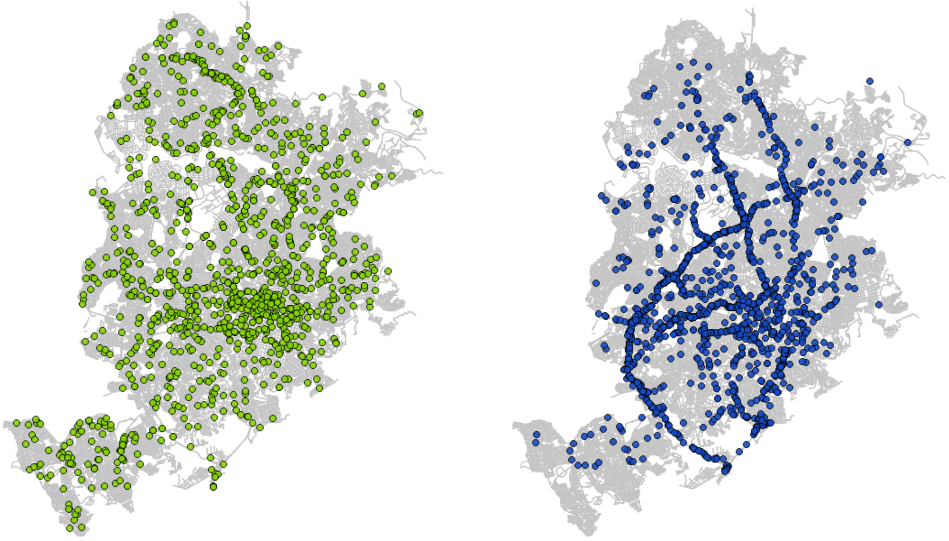
b) Matched records

Fig. 1. BHTrans and Waze accidents

acterize the types of accidents. Thereby, authoritative action can be decided based on existing data from multiple sources.

5.1 Matched Accidents

In order to verify which portion of Waze data corresponds to the data provided by BHTrans (that is, the number of accidents reported both officially and by Waze users), we apply comparisons following the criteria described on Section 4 (date/time frame and distance between reported positions). We



a) BHTrans b) Waze

Fig. 2. Unmatched accidents

found that only 7% of BHTrans reports matched accidents reported on Waze. Figure 1 shows the location of accidents reported on both official and unofficial datasets. Spatial distribution is sparse, with a few clusters of accidents in major thoroughfares.

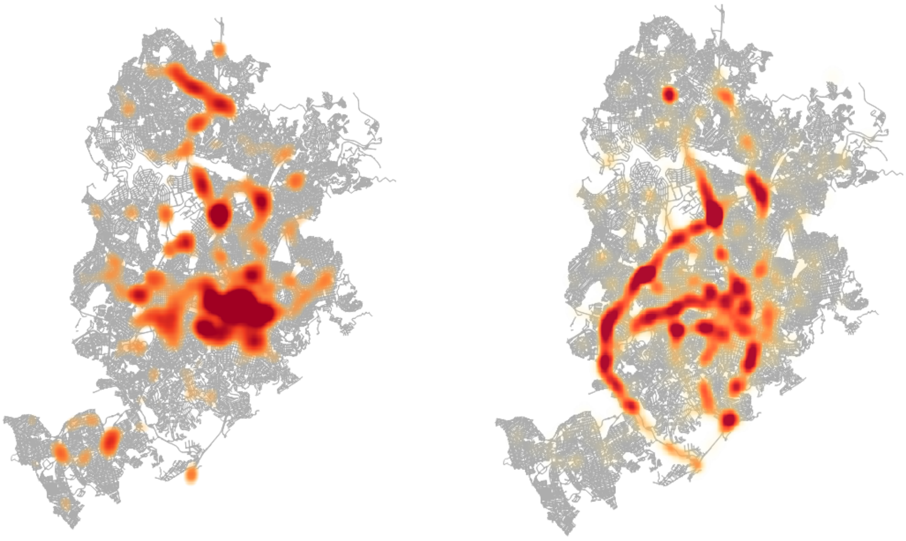
Since Waze allows a single event to be reported more than once, we consider only one of the multiple reports. Indeed, the consolidation step in the preparation of the Waze dataset shows that 38% of the accidents on Waze were reported by two or more users. A positive aspect on this repetition is that it makes the data more reliable, since many events are based on reports from more than one user. Notice, however, that Waze users that have already seen an accident report in their path through their smartphones will probably select another route or consciously avoid reporting that same accident again.

Official data on traffic accidents are usually recorded in police reports, which are mandatory only in cases where the parts involved aim some kind of material compensation for damages, directly from the responsible for the accident or through an insurance claim. Considering that, it is possible that part of the accidents with lower severity are not officially reported. We therefore expect that a share of the events recorded in Waze are not officially reported, and thus the datasets are expected to be complementary. Likewise, a share of the accidents reported officially may not be recorded by Waze users, especially if they take place in locations where the impact on traffic is small, or at times and places where the presence of Waze users is low. Notice that Waze usage naturally tends to be concentrated around rush hours, in which knowing about traffic problems along one's path is of greater concern. We now proceed to analyze accidents that have not been matched at either dataset.

5.2 Unmatched Accidents

Analyzing the reports that appear exclusively on either Waze or BHTrans datasets, we can see a different distribution from the one seen on Figure 1.

Accidents that appear only in the BHTrans dataset (Figure 2a) are found mostly in the central regions of the city, which is understandable since this area has a more intense traffic flow with a high concentration of economic activities. Secondary commercial areas in a northern and in a southwestern regions of the city also concentrate many accidents.



a) BHTrans

b) Waze

Fig. 3. Heat map of unmatched accidents

Events reported exclusively by Waze (Figure 2b), on the other hand, show a distinct pattern. Accidents are concentrated along some large city thoroughfares and urban segments of highways. Figure 3 shows heatmaps based on the concentration of unmatched accidents from either source.

5.3 Severity and Types of Accidents

Official records include the types of traffic accidents. Most frequent types are side collisions, collisions involving pedestrians, rollovers, rear-end collisions, and even accidents where people are thrown out of the vehicle. The existence of victims (injuries, fatalities) is also indicated. However, the records do not consistently inform on the severity of each accident. For example, a collision can either cause light injuries or more serious ones, requiring hospital treatment, or even leading to death. However, such details are absent from official records.

Waze records include less details than official ones. Data about an accident are informed by users when they are nearby, often driving, and it can be difficult for them to obtain details. Waze uses two severity classifications for accidents: major and minor. A minor accident is described as “fender benders with minor or no injuries, also no fatalities” while major represents “major damages to vehicle, major injuries and possible fatalities” [Waze 2016].

The most frequent types of accident found in official dataset are “side collision with victims” (39%), followed by “rear-end collision with victims” (18%). In the Waze dataset, 56.1% of the accidents are classified as minor, and 21% as major. The severity of the remaining 23% is not reported. Figure 4 shows the six most frequent types from official data and Figure 5 shows the distribution of severity classification from Waze data.

Comparing results from both datasets, most officially reported accidents can be understood as high severity (since they are classified as accidents with victims), while most accidents recorded by Waze are classified as minor severity. This discrepancy raises, at least, two hypotheses. First, it is difficult for Waze users to know detailed information about the accidents they report. If they witness the accident, they can overestimate its severity, while if they drive by the accident's location after some

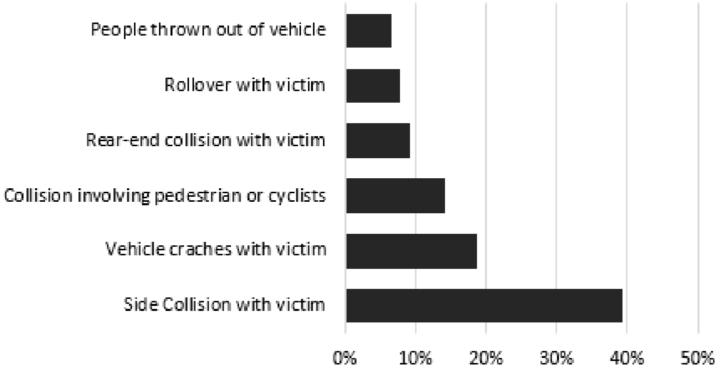


Fig. 4. Main types of accidents from official sources

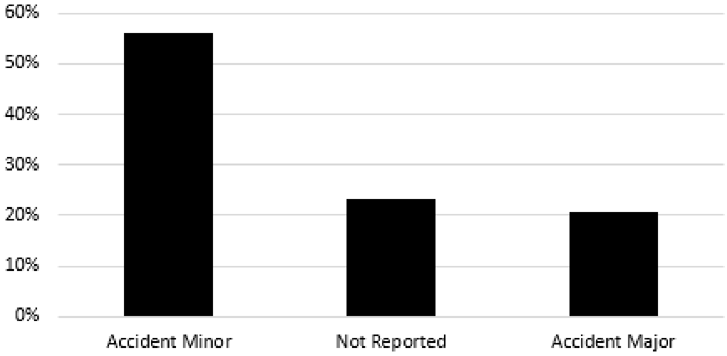


Fig. 5. Severity of accidents from Waze

time their assessment can be underestimated. Second, Waze users may not be able to pay much attention on the correct severity classification when informing an accident.

Among matched accidents, Waze users only tend to classify as major severity accidents those officially classified as collisions involving pedestrians and rear-end collisions with victims (Table III). In other cases, most accidents are deemed minor by Waze users. This indicates a semantic discrepancy between regular citizens (Waze users) and police or transit officials when an accident is reported in both sources, and this problem has to be investigated further. Also, for accidents reported by several different Waze users, the discrepancy among them must be assessed. However, as in any crowdsourcing process, accident classification by Waze users should be less reliable than official reports, since, as passers-by, Waze users do not have full access to the accident site.

Regarding the distribution of accidents along the day, Waze reports concentrate on rush hours, either in the morning or in the afternoon (Figure 6). This is expected, since at those hours the impact of accidents on traffic is the greatest. On the other hand, official data, while also recording a high number of accidents at rush hours, contain reports of accidents that took place all through the day, including times at which circulation is lower.

Taking the integrated dataset, Figure 6 shows the distribution of accidents throughout the day. Notice the peaks at rush hours in the morning (7:00 to 8:00 AM) and afternoon (4:00 to 7:00 PM). Even though the Waze and official datasets have been shown to be complementary, their behavior is similar. Notice, also, that no accidents were reported by both sources between 0:00 and 6:00 AM.

Table III. Classification of severity by Waze users for each type of accident from BHTrans

Accident Types (BHTrans)	Severity (Waze - values are %)		
	Major	Minor	Not Informed
Side collision with victims	21.6	56.8	21.6
Vehicle collision with victims	31.1	46.6	22.3
Collision involving pedestrians without fatality	54.1	37.5	8.3
Vehicle crashes with victims	47.3	42.1	10.5
Rollover with victims	28.6	42.8	28.6
People thrown out of the vehicle	37.5	50.0	12.5

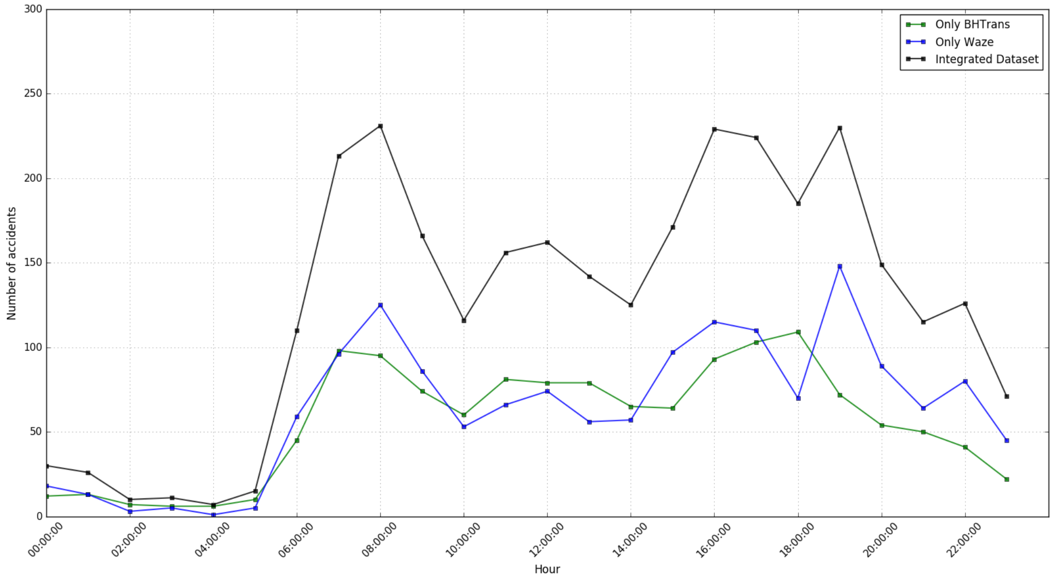


Fig. 6. Distribution of accidents throughout the day (complete dataset)

This confirms the tendency for Waze users to concentrate their contributions as to accidents at times in which accidents are likely to cause traffic disruptions. Accidents reported by the official source are distributed in a slightly more uniform fashion, also concentrating at rush hours. Report time variations may account for the drop in Waze accidents at 6:00 PM, as compared to BHTrans peak number at that time. Since the Waze peak occurs at 7:00 PM, that time slice may be receiving accidents that took place a little earlier.

Waze data also show that a large share of the accidents takes place at times in which other contributed events, such as traffic jam, vehicle stopped, police presence, and objects on the road, are also reported. From the total number of accidents reported in Waze, about 55% took place closer than 50m away and less than an hour before or after some other event. Figure 7 shows the correlation between accidents and other nearby events. The X axis indicates the time at which the accident occurred, and the Y axis shows the time at which the other event was reported. Blue triangles indicate accidents reported before the event, and red circles indicate accidents reported after the event. The number of circles is much greater than the number of triangles (about 65% to 35%), indicating that accidents are often reported after some other event. We then look at order in which events were reported. Figure 8 shows the distribution of which events were reported first. It can be seen that only in about 35% of the cases the accident was the first event to be reported. Traffic jam was reported first on approximate 45% of the cases. Police and hazards related events were reported first on 10% of cases, each. Our

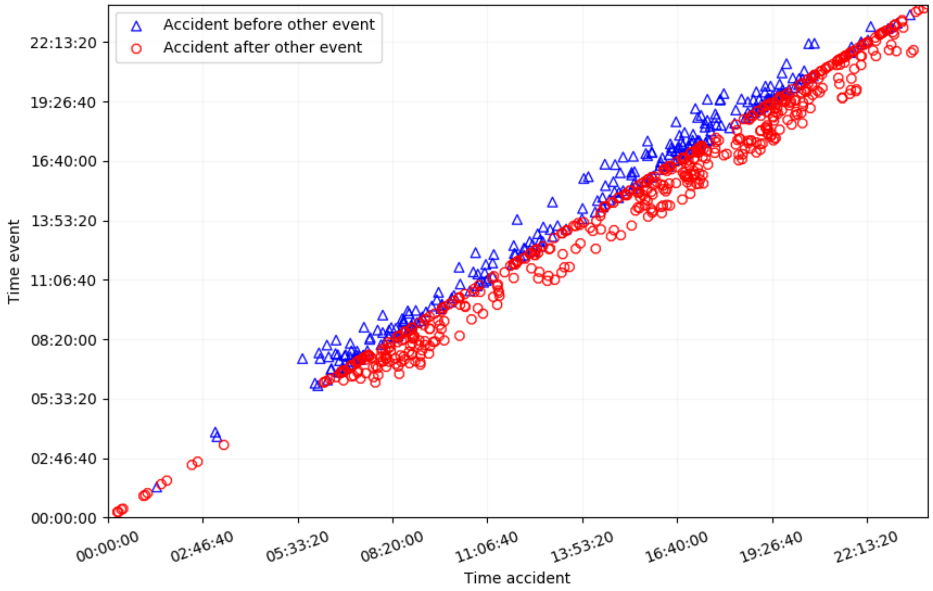


Fig. 7. Accidents and other nearby events.

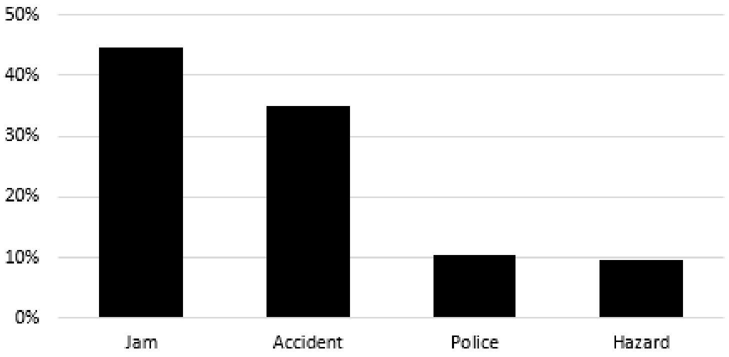


Fig. 8. Distribution of which events was reported first.

interpretation for the high rate of traffic jam reports registered before the accidents reports was that users usually experience first the effects of the accident event with the traffic becoming slow or even totally stopped and only sometime can understand that the cause was an accident when he can see what happened.

Notice that most of the times an accident is only reported after Waze users notice some disruption in traffic, such as a traffic jam. This reinforces the observation presented earlier in the sense that Waze users tend to report accidents that have a bigger impact on local traffic.

5.4 Influence of Weather

We obtained weather data for the city of Belo Horizonte in the period covered by the datasets, concentrating on rainfall, since it never snows in the city and temperatures are usually mild through-

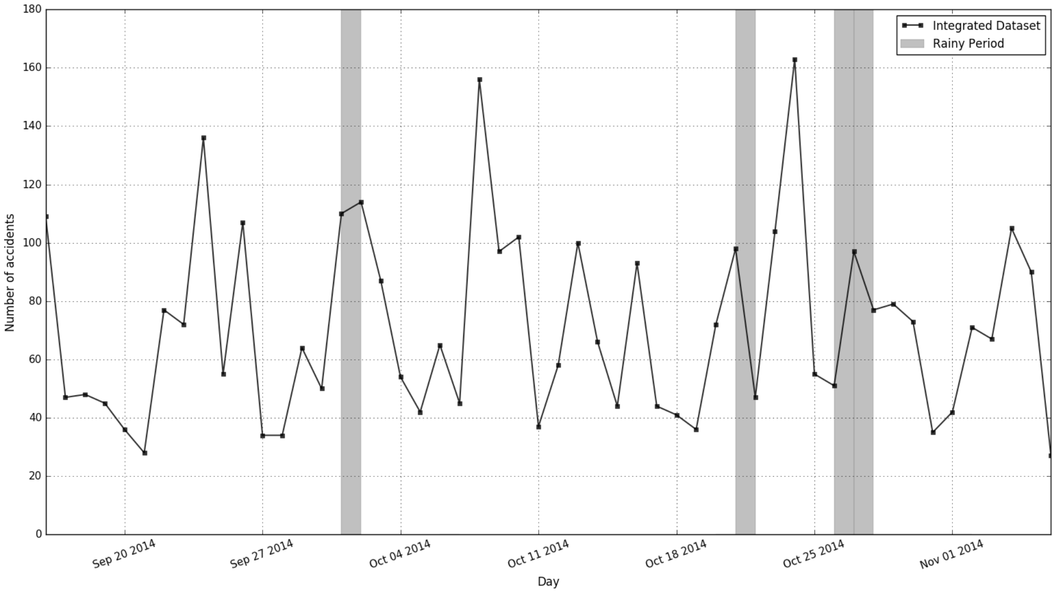


Fig. 9. Number of accidents per day

out the year. According to the Brazilian National Meteorology Institute (INMET)[8], Belo Horizonte receives an average of 1473mm of rainfall in a year, most of which between October and March. There are, on average, 116 rainy days a year in the city. As city residents, we observe that rainfall, isolated or on the first days of the rainy season, tends to cause a higher number of accidents, due to skidding and lower visibility.

However, the analysis of rainfall data against the accidents records in the period covered by our dataset found no correlation. Figure 9 shows the daily number of accidents in the integrated dataset per day, along with the indication of rainy days in the period. The year 2014 was atypically dry in Belo Horizonte, with only 944mm of rainfall, and a delayed start of the rainy season. In September 2014, Belo Horizonte had a single rainy day, with only 1.8mm of precipitation. In October, usually the first month of the rainy period, there was rain in 8 days, total precipitation of 69.4mm (8.6mm/day). Further analysis of the relationship between traffic accidents and rainfall data is therefore necessary, especially for the rainier months of the year in Belo Horizonte (usually December and January).

5.5 Accident hotspots

Considering the integrated dataset, we generated a list of accident hotspots in the city. These hotspots are concentrated in high-traffic street crossings, as expected. A normalization of the number of accidents by traffic flow measurements or estimates is not possible at this time, due to lack of data on traffic volumes, and is left for future work.

Nevertheless, Table IV describes the location of the top-10 accident hotspots, while Figure 10 shows the same hotspots on a map. Anel Rodoviário, a patchwork of road segments incrementally turned into an expressway that crosses the city and interconnects the main highways leading to and from it, is the site of many of the hotspots. The city grew around it, and now it receives a mix of inter-regional urban traffic and heavy-load cargo trucks. A major engineering project to improve traffic conditions at Anel Rodoviário has been under discussion for many years, was included in the Federal infrastructure budget back in 2010, but prioritization for execution has not been achieved so far. The

[8] 30-year climatological normals, available at http://www.inmet.gov.br/portal/index.php?r=clima/graficosClimaticos

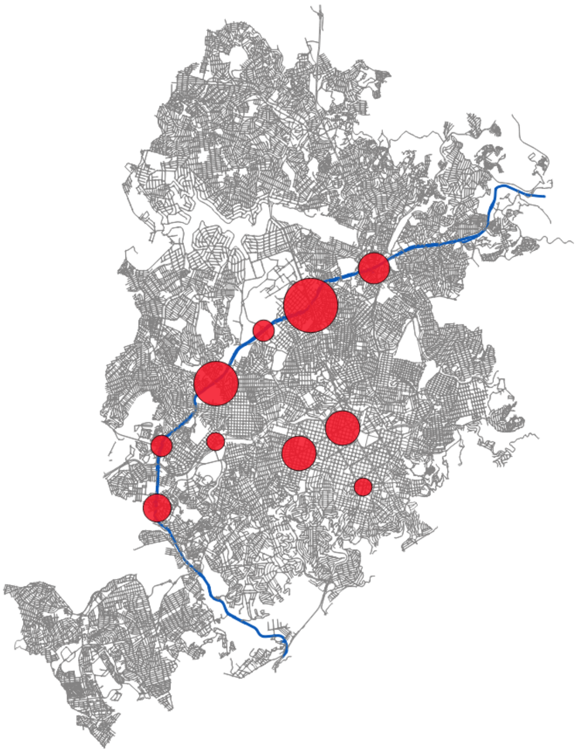


Fig. 10. Location of accident hotspots at Belo Horizonte. Anel Rodoviário was highlighted on the map

fact that Anel Rodoviário is actually the superposition of Federal highways under the responsibility of the National Transportation Infrastructure Department (DNIT), but usually managed by State and Municipal organizations, contributes to the budgeting difficulties. Anel Rodoviário presents choke points, lane reductions in overpasses, close presence of irregular housing, signaling deficiencies, and steep slopes in some spots. The downtown region concentrates other hotspots, some of which related to Amazonas Avenue, another high-traffic thoroughfare that connects to important highways, through Contagem and Betim, the more populous and industrialized cities in the metropolitan area.

If the same analysis were conducted from official data alone, only four hotspots presented on Table IV would remain: Anel Rodoviário at Antonio Carlos Avenue, Amazonas Avenue between Rua Padre Belchior and Afonso Pena Avenue, Amazonas Avenue at Contorno Avenue and Anel Rodoviário at Cristiano Machado Avenue. All of them are located around the downtown area. This fact reinforces our

Table IV. Accident hotspots

# of accidents	Location Description
151	Anel Rodoviário at Antonio Carlos Avenue
111	Anel Rodoviário at Pedro II Avenue
83	Amazonas Avenue between Rua Padre Belchior and Afonso Pena Avenue
81	Amazonas Avenue at Contorno Avenue
79	Anel Rodoviário at Cristiano Machado Avenue
65	Amazonas at BR-040 and BR-381 highways
60	Anel Rodoviário at Presidente Juscelino Kubitschek Avenue
51	Anel Rodoviário at Carlos Luz Avenue
45	Senhora do Carmo Avenue at Contorno Avenue
44	Tereza Cristina Avenue ate Craveiro Lopes Street

observation that both datasets are complementary, and therefore a more complete scenario regarding city traffic accidents arises when analyzing integrated data.

6. CONCLUSIONS

The growing use of mobile apps and social networks generates a considerable amount of data, which can be used for several purposes. Information from these sources can be used for identifying better routes, monitoring traffic conditions in real time and identifying areas with high vehicular accident rates. Besides that, data collected from this kind of source can complement the data extracted from official sources.

This work compared car accident records provided by BHTrans (official) and data collected from the mobile app Waze (unofficial). We found out that 7% of the car accidents officially reported were also reported through unofficial sources. Among these accidents, the main type was collision with victims, while most were classified as low severity according to Waze data. It is important to stress, however, that the classification used by Waze is not totally reliable, given that users usually do not have direct access to the accident site. Furthermore, the lack of an API introduces a significant difficulty for the collection of volunteer-produced Waze data.

The official data has the advantage of being more reliable and detailed. However, it is harder to access due to government-imposed limitations and time constraints. On the other hand, the unofficial registers are easy to access, but poor in details. We demonstrated how the data from these two sources can be integrated in order to obtain a dataset with broader coverage. This could compensate the deficiencies of each of the sources, taken individually. However, official data should be made available more readily, possibly using technologies such as spatial data infrastructures or APIs dedicated to the task of providing unrestricted and timely access to traffic accident data.

We verified that most of the accidents reported by Waze but not by BHTrans were classified as having lower severity. This happens possibly due to the fact that police reports are not mandatory, which probably implies that most of the lower severity accidents are not officially reported. Thus, Waze data can be used to fill the gap of under-notification in the case of traffic accidents, providing authorities with a broader view of the problem.

Accidents reported only by BHTrans were concentrated on the central region of Belo Horizonte, while the ones reported by Waze were mostly on highway segments, like Rodovia MG-10 and Anel Rodoviário. These patterns endorse the idea that the datasets are complementary, since coverages are quite distinct. After integrating the datasets, a view on the accident hotspots in the city arises, thereby generating a prioritization for official action. Accident-prevention actions may include signaling, physical presence of policemen, speed-control devices or educational campaigns. We showed that a hotspot list generated by only looking at official data would be quite different from the one produced from the integrated dataset. Therefore, decision-making based only on the official data could lead to biased and less efficient actions. In the future, a normalization of the number of accidents by the traffic volume can lead to other types of hotspots, with potentially lower global impact, but significant in a local perspective.

The next step for this work is to collect data from a wider range of official sources regarding traffic accidents, expanding the analysis to other cities besides Belo Horizonte. In Belo Horizonte, a Waze dataset covering a much wider period is already available, but official data remain difficult to obtain. Still regarding the official data, we aim to consider the Brazilian federal highways accidents database, provided by DNIT. We also intend to get data from other unofficial sources, such as Twitter. With multiple sources of heterogeneous data, integration methods need to evolve accordingly.

REFERENCES

Bezerra, B. S., Cunto, F. C., Barbosa, H. M., Davis, C., and Lança, J. F. d. A. Main stumbling blocks for a good traffic accident database system – evidences from Brazil. *Latin American J. Management for Sustainable Development* 2 (2): 112–123, 2015.

Brasil. *Lei N° 11.705, de 19 de Junho de 2008*. Brasília, 2008. Available: `http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11705.htm` [Accessed 15 August 2016].

Corsar, D., Markovic, M., Edwards, P., and Nelson, J. D. The Transport Disruption Ontology. In *International Semantic Web Conference*. Lecture Notes in Computer Science, vol. 9367. Springer International Publishing, Bethlehem, PA, USA, pp. 329–336, 2015.

Machado, C., Giannotti, M., Neto, F., Tripodi, A., Persia, L., and Quintanilha, J. Characterization of Black Spot Zones for Vulnerable Road Users in São Paulo (Brazil) and Rome (Italy). *ISPRS International Journal of Geo-Information* 4 (2): 858–882, 2015.

Mateveli, G. V., Machado, N. G., Moro, M. M., and Davis Jr., C. A. Taxonomia e desafios de recomendação para coleta de dados geográficos por cidadãos. In *Proceedings of the Brazilian Symposium on Databases*. Petrópolis, Brasil, pp. 105–110, 2015.

Morris, A., Brace, C., Reed, S., Fagerlind, H., Bjorkman, K., Jaensch, M., Otte, D., Vallet, G., Cant, L., Giustiniani, G., Parkkari, K., Verschragen, E., and Hoogvelt, B. The development of a european fatal accident database. *International Journal of Crashworthiness* 15 (2): 201–209, 2010.

Najm, W. G., Sen, B., Smith, J. D., and Campbell, B. N. Analysis of light vehicle crashes and pre-crash scenarios based on the 2000 general estimates system. Tech. rep., Volpe National Transportation Systems Center, 2003.

Quercia, D. and Capra, L. Friendsensing: Recommending friends using mobile phones. In *Proceedings of the Third ACM Conference on Recommender Systems*. New York, USA, pp. 273–276, 2009.

Quercia, D., Lathia, N., Calabrese, F., Di Lorenzo, G., and Crowcroft, J. Recommending Social Events from Mobile Phone Location Data. In *Proceedings of the IEEE International Conference on Data Mining*. Sydney, Australia, pp. 971–976, 2010.

Ribeiro Jr., S. S., Rennó, D., Gonçalves, T. S., Davis, C., Meira Jr., W., and Pappa, G. L. Observatório do Trânsito: sistema para detecção e localização de eventos de trânsito no Twitter. In *Proceedings of the Brazilian Symposium on Databases*. São Paulo, Brazil, pp. 81–88, 2012.

Salazar, J. C., Torres-Ruiz, M., Davis Jr., C. A., and Moreno-Ibarra, M. Geocoding of traffic-related events from Twitter. In *Proceedings of the Brazilian Symposium on GeoInformatics*. Campos do Jordão, SP, Brazil, pp. 14–25, 2015.

Scarponcini, P. Generalized model for linear referencing. In *Proceedings of the 7th ACM International Symposium on Advances in Geographic Information Systems*. New York, NY, USA, pp. 53–59, 1999.

Silva, T. H., de Melo, P. O. S. V., Viana, A. C., Almeida, J. M., Salles, J., and Loureiro, A. A. F. Traffic condition is more than colored lines on a map: Characterization of waze alerts. In *Social Informatics. SocInfo 2013*. Kyoto, Japan, pp. 309–318, 2013.

Smarzaro, R., de Lima, T. F. M., and Davis Jr., C. A. Could data from location-based social networks be used to support urban planning? In *Proceedings of the International World Wide Web Conferences*. Perth, Australia, pp. 1463–1468, 2017.

Stamatiadis, N. and Deacon, J. A. Trends in highway safety: Effects of an aging population on accident propensity. *Accident Analysis and Prevention* 27 (4): 443–459, 1995.

Waze. *Manual do Usuário versão 3.5*. Waze, 2016. Available: `https://wiki.waze.com/wiki/Como_Alertar` [Accessed 15 August 2016].

WHO. Global status report on road safety 2015. Tech. rep., World Health Organization, 2015. Available: `http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/` [Accessed 03 August 2016].

Wolf, K. and Fry, J. Benchmarking performance data. In *Beyond Transparency: Open Data and the Future of Civic Innocation*, B. Goldstein and L. Dyson (Eds.). Code for America Press, pp. 233–252, 2013.

Zheng, Y., Chen, Y., Xie, X., and Ma, W.-Y. GeoLife2.0: A location-based social networking service. In *International Conference on Mobile Data Management: Systems, Services and Middleware*. Taipei, Taiwan, pp. 357–358, 2009.