

# A Countrywide Traffic Accident Dataset\*

Sobhan Moosavi

The Ohio State University  
Department of Computer Science and Engineering  
Columbus, Ohio  
moosavi.3@osu.edu

Srinivasan Parthasarathy

The Ohio State University  
Department of Computer Science and Engineering  
Columbus, Ohio  
srini@cse.ohio-state.edu

Mohammad Hossein Samavatian

The Ohio State University  
Department of Computer Science and Engineering  
Columbus, Ohio  
samavatian.1@osu.edu

Rajiv Ramnath

The Ohio State University  
Department of Computer Science and Engineering  
Columbus, Ohio  
ramnath@cse.ohio-state.edu

## ABSTRACT

Reducing traffic accidents is an important public safety challenge. However, the majority of studies on traffic accident analysis and prediction have used small-scale datasets with limited coverage, which limits their impact and applicability; and existing large-scale datasets are either private, old, or do not include important contextual information such as environmental stimuli (weather, points-of-interest, etc.). In order to help the research community address these shortcomings we have - through a comprehensive process of data collection, integration, and augmentation - created a large-scale publicly available database of accident information named *US-Accidents*. *US-Accidents* currently contains data about 2.25 million instances of traffic accidents that took place within the contiguous United States, and over the last three years. Each accident record consists of a variety of intrinsic and contextual attributes such as *location*, *time*, *natural language description*, *weather*, *period-of-day*, and *points-of-interest*. We present this dataset in this paper, along with a wide range of insights gleaned from this dataset with respect to the spatiotemporal characteristics of accidents. The dataset is publicly available at [https://smoosavi.org/datasets/us\\_accidents](https://smoosavi.org/datasets/us_accidents).

## KEYWORDS

US-Accidents, Traffic Accident Dataset, Traffic Accident Prediction

## 1 INTRODUCTION

Reducing traffic accidents is an important public safety challenge around the world. A global status report on traffic safety [19], notes that there were 1.25 million traffic deaths in 2013 alone, with deaths increasing in 68 countries when compared to 2010. Accident prediction is important for optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure, all in order to make the roads safer. Given its significance, accident analysis and prediction has been a topic of much research in the past few decades. While a large body of research

has been focused on small-scaled datasets with limited coverage (e.g. a small number of road-segments, or just one city) [4, 5, 11, 26], the value and impact of predictive solutions may be better studied when using large-scale data. Although some studies conducted their work based on large-scale motor-vehicle crash datasets, their data is usually private or poses strict rules to be shared with outside researchers, which makes their framework and results unproducible [7, 8, 15, 28]. While there are still a few publicly available large-scale accident datasets, their data is either old, limited to one state or one city, or incomprehensive (regarding data attributes or average reports per year) [9, 13, 16, 20, 24].

In order to mitigate these challenges and to provide a context for future research on traffic accident analysis and prediction, we present a new dataset, we name it *US-Accidents*, which includes about 2.25 million instances of traffic accidents took place within the contiguous United States<sup>1</sup> between February 2016 and March 2019. Unlike some of the available large-scale accident datasets (such as [16]), *US-Accidents* offers a wide range of data attributes to describe each accident record including *location data*, *time data*, *natural language description of event*, *weather data*, *period-of-day information*, and *relevant points-of-interest data* (traffic signal, stop sign, etc.). Very importantly, we also present our *process* for creating the above dataset from streaming traffic reports and heterogeneous contextual data (weather, points-of-interests, etc.), so that the community can validate it, and with the belief that this process can itself serve as a model for dataset creation.

Using *US-Accidents*, we performed a variety of data analysis and profiling to derive a wide-range of insights. Our analyses demonstrated that about 40% of accidents took place on or near high-speed roadways (highways, interstates, etc.) and about 32% on or near local roads (streets, avenues, etc.). We also derived various insights with respect to the correlation of accidents with time, points-of-interest, and weather conditions.

We summarize the contributions of this paper as follows:

- A unique dataset, made publicly available at [https://smoosavi.org/datasets/us\\_accidents](https://smoosavi.org/datasets/us_accidents). This dataset has been collected for the

\*All rights reserved to the authors, and The Ohio State University (2019).

<sup>1</sup>The contiguous United States excludes Alaska and Hawaii, and considers District of Columbia (DC) as a separate state.

contiguous United States over three years, and contains about 2.25 million traffic accident records. Further, the raw accident records have been augmented by *map-matching*, and contextual information such as *weather*, *period-of-day*, and *points-of-interest*.

- A new methodology for the heterogeneous data collection, cleansing, and augmentation; needed to prepare a unique large-scale dataset of traffic accidents.
- A variety of insights gleaned through analyses of accident hot-spot locations, time, weather, and points-of-interest correlations with the accident data; that may directly be utilized for applications such as urban planning, exploring flaws in transportation infrastructure design, traffic management and prediction, and personalized insurance.

The rest of the paper is organized as follows. Section 2 provides an overview of related work, followed by definitions and preliminaries in Section 3. The process of creating the accident dataset is presented in Section 4, and analyses and insights are discussed in Section 5. Finally, Section 6 concludes the paper.

## 2 RELATED WORK

Accident analysis and prediction has an active research topic during the past few decades, with a large body of research has been focused on using small-scale datasets with limited coverage of a few road-segments or one city [1, 4–6, 10, 11, 26]. Chang et al. [5] used information such as road geometry, annual average daily traffic, and weather data to predict frequency of accidents for a highway road using a neural network model. Kumar et al. [10] applied data mining techniques to extract association rules to perform causality analysis using a small-scale dataset. Likewise, Wenqi et al. [26] applied a convolutional neural network model to perform accident prediction on a road-segment. Although the insights and findings look interesting, the employed datasets are of limited scale; hence, the applicability and generalizability of results might be questionable.

There are, of course, numerous studies that have used larger-scale datasets [7, 8, 15, 21, 27, 28]; however, the datasets have been either private or not easily accessible. Eisenberg [8] conducted a thorough analysis on the impact of precipitation on road accidents, using a large-scale dataset of about 456,000 crashes collected from 1975 to 2000 for 48 states of the US. Recent studies by Yuan et al. [27] and Najjar et al. [15] have also employed very large-scale accident datasets to perform real-time traffic accident prediction. However, in neither study have details been shared regarding how the data used may be obtained by others in order to reproduce results for wider use.

Finally, and speaking of datasets alone, there are several publicly available motor vehicle crash datasets; however, they suffer from the limited coverage (e.g., one city or one state) [2, 9, 13, 16, 20], or their data is old [24], or the provided attributes are not comprehensive enough (missing location, time, or weather data) [9, 16]. To address these challenge, we propose a new process to collect and build a new large-scale accident dataset, with countrywide coverage, and comprehensive data attributes including location, time, weather,

period-of-day, and points-of-interest annotations (e.g., intersections, junctions, and traffic signals).

## 3 TERMINOLOGY

In this section we provide a set of definitions.

**Definition 3.1** (Traffic Event). We define a traffic event  $e$  by  $e = \langle lat, lng, time, type, desc \rangle$ , where  $lat$  and  $lng$  are GPS latitude and longitude,  $type$  is the type of the event, and  $desc$  provides a natural language description of the event. A traffic event is of one of the following types: *accident*, *broken-vehicle*<sup>2</sup>, *congestion*<sup>3</sup>, *construction*<sup>4</sup>, *event*<sup>5</sup>, *lane-blocked*<sup>6</sup>, or *flow-incident*<sup>7</sup>.

**Definition 3.2** (Weather Observation Record). A weather observation  $w$  is defined by  $w = \langle lat, lng, time, temperature, humidity, pressure, visibility, wind-speed, precip, rain, snow, fog, hail \rangle$ . Here  $lat$  and  $lng$  represent the GPS coordinates of the weather station which reported  $w$ ; *precip* is the precipitation amount (if any); and rain, snow, fog, and hail<sup>8</sup> are binary indicators of these events.

**Definition 3.3** (Point-of-Interest). A point-of-interest  $p$  is defined by  $p = \langle lat, lng, type \rangle$ . Here,  $lat$  and  $lng$  show the GPS latitude and longitude coordinates, and available types for  $p$  are described in Table 1. Note that several of definitions in this table are adopted from <https://wiki.openstreetmap.org>.

**Table 1: Definition of Point-Of-Interest (POI) annotation tags based on Open Street Map (OSM).**

Type	Description
Amenity	Refers to particular places such as restaurant, library, college, bar, etc.
Bump	Refers to speed bump or hump to reduce the speed.
Crossing	Refers to any crossing across roads for pedestrians, cyclists, etc.
Give-way	A sign on road which shows priority of passing.
Junction	Refers to any highway ramp, exit, or entrance.
No-exit	Indicates there is no possibility to travel further by any transport mode along a formal path or route.
Railway	Indicates the presence of railways.
Roundabout	Refers to a circular road junction.
Station	Refers to public transportation station (bus, metro, etc.).
Stop	Refers to stop sign.
Traffic Calming	Refers to any means for slowing down traffic speed.
Traffic Signal	Refers to traffic signal on intersections.
Turning Loop	Indicates a widened area of a highway with a non-traversable island for turning around.

<sup>2</sup>Refers to the situation when there is one (or more) disabled vehicle(s) in a road.

<sup>3</sup>Refers to the situation when the speed of traffic is slower than the expected speed.

<sup>4</sup>An on-going construction or maintenance project on a road.

<sup>5</sup>Situations such as *sports event*, *concerts*, or *demonstrations*, that could potentially impact traffic flow.

<sup>6</sup>Refers to the cases when we have blocked lane(s) due to traffic or weather condition.

<sup>7</sup>Refers to all other types of traffic events. Examples are *broken traffic light* and *animal in the road*.

<sup>8</sup>The case of having solid precipitation including ice pellets and hail.

## 4 DATASET CREATION PROCESS

An overview of the dataset creation process is shown in Figure 1, with the following sub-sections provide detailed descriptions of each step.

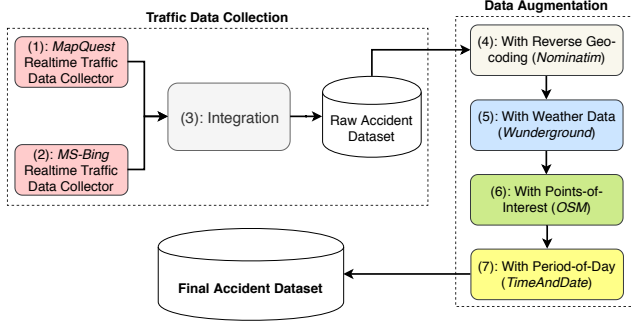


Figure 1: Process of Creating Traffic Accident Dataset

### 4.1 Traffic Data Collection

**4.1.1 Realtime Traffic Data Collection.** We collected streaming traffic data using two real-time data providers, namely “MapQuest Traffic” [12] and “Microsoft Bing Map Traffic” [3], whose APIs broadcast traffic events (accident, congestion, etc.) captured by a variety of entities - the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. We pulled data every 90 seconds from 6am to 11pm, and every 150 seconds from 11pm to 6am. In total, we collected 2.27 million cases of traffic accidents between February 2016 and March 2019; 1.7 million cases were pulled from MapQuest, and 0.54 million cases from Bing.

**4.1.2 Integration.** Integration of the data consisted of removing cases duplicated across the two sources and building a unified dataset. We considered two events as duplicates if their Haversine distance and their recorded times of occurrence were both below a heuristic threshold (set empirically at 250 meters and 10 minutes, respectively). We believe these settings to be conservative, but we settled on them in order to ensure a very low possibility of duplicates. Using these settings, we found about 24,600 duplicated accident records, or about 1% of all data. The final dataset after removing the duplicated cases comprised 2.25 million accidents.

### 4.2 Data Augmentation

**4.2.1 Augmenting with Reverse Geo-Coding.** Raw traffic accident records contain only GPS data. We employed the *Nominatim* tool [17] to perform reverse geocoding to translate GPS coordinates to addresses, each consisting of a *street number*, *street name*, *relative side* (left/right), *city*, *county*, *state*, *country*, and *zip-code*. This process is same as *point-wise map-matching*.

**4.2.2 Augmenting with Weather Data.** Weather information provides important context for traffic accidents. Thus, we employed

*Weather Underground* API [25] to obtain weather information for each accident. Raw weather data was collected from 1,977 weather stations located in airports all around the United States. The raw data comes in the form of observation records, where each record consists of several attributes such as *temperature*, *humidity*, *wind speed*, *pressure*, *precipitation* (in millimeters), and *condition*<sup>9</sup>. For each weather station, we collected several data records per day, each of which was reported upon any significant change in any of the measured weather attributes.

Each traffic event  $e$  was augmented with weather data as follows. First the closest weather station  $s$  was identified. Then, of the weather observation records which are reported from  $s$ , we looked for the weather observation record  $w$  whose reported time was closest to the start time of  $e$ , and augmented it with weather data. In our integrated accident dataset, the average difference in report time for an accident record and its paired weather observation record was about 15 minutes; and the 50<sup>th</sup>, 75<sup>th</sup>, and 90<sup>th</sup> percentiles on time difference distribution were about 11, 20, and 26 minutes, respectively.

**4.2.3 Augmenting with Points-Of-Interest.** Points-of-interest (POI) are locations annotated on a map as *amenities*, *traffic signals*, *crossings*, etc. These annotations are associated with *nodes* on a road-network. A node can be associated with a variety of POI types, however, in this work we only use a subset of 13 types described in Table 1. We obtain these annotations from Open Street Map (OSM) [18] for the United States, using its most recently released dataset (extracted on April 2019). The applicable POI annotations for a traffic accident  $a$  are based on the actual POI located within a distance threshold  $\tau$  from  $a$ . We determine this threshold by evaluating different threshold values to find the value that is best able to associate a POI with an accident. Essentially, the objective is to find the best distance for which a POI annotation can be identified as a relevant to an accident record. Therefore, we need a mechanism to measure the relevancy. To begin with, we note that the natural language descriptions of traffic accidents follow a set of regular expression patterns, and that a few of these patterns may be used to identify and use as an annotation for the location type (e.g., intersection or junction) of the accident.

**Regular Expression Patterns.** Given the description of traffic events of type accident, we were able to identify 27 regular expression patterns; 16 of them were extracted based on MapQuest data, and 11 from Bing data. Among the MapQuest patterns, the following expression corresponds to *junctions* (see Table 1): “... **on** ... **at exit** ...”, and the following pattern mostly<sup>10</sup> determines an *intersection*: “... **on** ... **at** ...”. We consider a location an *intersection* if it is associated with at-least one of the following annotations (see Table 1): *crossing*, *stop*, or *traffic signal*. Among Bing regular expression patterns, two of them identify junctions: “**at** ... **exit** ...” and “**ramp to** ...”. Table 2 shows several examples of accidents,

<sup>9</sup>Possible values are *clear*, *snow*, *rain*, *fog*, *hail*, and *thunderstorm*.

<sup>10</sup>Regarding the distribution of data and using 200 random cases which were manually checked on a map, about 78% of matches using this pattern were actually happened on intersections.

**Table 2: Examples of traffic accidents with their *annotation type* assigned using their natural language description by regular expression patterns.**

Source	Description	Type
MapQuest	Serious accident <b>on</b> 4th Ave <b>at</b> McCullaugh Rd.	Intersection
MapQuest	Accident <b>on</b> NE-370 Gruenther Rd <b>at</b> 216th St.	Intersection
MapQuest	Accident <b>on</b> I-80 <b>at</b> Exit 4A Treasure Is.	Junction
MapQuest	Accident <b>on</b> I-87 I-287 Southbound <b>at</b> Exit 9 I-287.	Junction
Bing	<b>At</b> Porter Ave/ <b>Exit</b> 9 - Accident. Left lane blocked.	Junction
Bing	<b>At</b> IL-43/Harlem Ave/ <b>Exit</b> 21B - Accident.	Junction
Bing	<b>Ramp</b> to I-15/Ontario Fwy/Cherry Ave - Accident.	Junction
Bing	<b>Ramp</b> to Q St - Accident. Right lane blocked.	Junction

where the regular expression pattern (in bold face) identifies the correct POI type<sup>11</sup>.

---

**Algorithm 1:** Find Annotation Correlation

---

- 1: Input: a dataset of traffic accidents  $\mathcal{A}$ , a database of points-of-interest  $\mathcal{P}$ , and a distance threshold  $\tau$ .
  - 2: Extract and create a set of regular expression patterns  $RE$  to identify a specific POI  $v$ .
  - 3: Create set  $S_1$ : for each traffic accident  $a \in \mathcal{A}$ , we add it to  $S_1$  if its natural language description  $a.desc$  can be matched with at least one regular expression in set  $RE$ .
  - 4: Create set  $S_2$ : for each traffic accident  $a \in \mathcal{A}$ , we add it to  $S_2$  if there is at least one POI  $p \in \mathcal{P}$  of type  $v$ , where  $haversine\_distance(a, p) \leq \tau$ .
  - 5: Output: Return  $Jaccard(S_1, S_2)$ .
- 

The essential idea is to find a threshold value that maximizes the correlation between annotations from POI and annotations derived using regular expression patterns. Thus, for a set of accident records, we annotate their location based on both regular expression patterns as well as OSM-based POI annotations (using a specific distance threshold). Then, we measure the correlation between the annotations derived from these methods to find which threshold value provides the highest correlation (i.e., the best choice). Note that we employ the regular expression patterns as *pseudo* ground truth labels, to evaluate OSM-based POI annotations using different threshold values. We propose Algorithm 1 to find the best distance threshold. We use a sample of 100,000 accidents as set  $\mathcal{A}$  (step 1). For step 2, we consider either “intersection” or “junction”, and use the set of relevant regular expressions (see Table 2) in terms of  $RE$ . Next we create set  $S_1$  by annotating each traffic accident  $a \in \mathcal{A}$  by using the regular expression patterns in  $RE$  (step 3). Then we annotate each traffic accident  $a \in \mathcal{A}$  based on points-of-interests in  $\mathcal{P}$ , using the distance threshold  $\tau$  to create  $S_2$  (step 4). Finally, we calculate the Jaccard similarity score using Equation 1 (step 5):

$$Jaccard(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \quad (1)$$

<sup>11</sup>These cases were also manually checked on a map to ensure the correctness of the annotation.

We examined the following candidate set to find the optimal threshold value (all values in meters): {5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 125, 150, 200, 250, 300, 400, 500}. We separately studied samples from Bing and MapQuest, and employed corresponding regular expression patterns for “intersection” and “junction”. Figure 2 shows the results for each data source and each annotation type. From Figure 2a, we see that the maximum correlation for intersections is obtained for a threshold value of 30 meters. Figures 2b and 2c show that 100 meters is an appropriate distance threshold for annotating a junction.

Thresholds for the other types of available annotations in Table 1 are derived from the thresholds for junction and intersection as described below:

- **Junction-based threshold.** Given the definition of a junction (i.e., a highway ramp, exit, or entrance), we used the same threshold for the following types: amenity and no-exit.
- **Intersection-based threshold.** Given the definition of an intersection, we used the same threshold for the following annotation types: bump, crossing, give-way, railway, roundabout, station, stop, traffic calming, traffic signal, and turning loop.

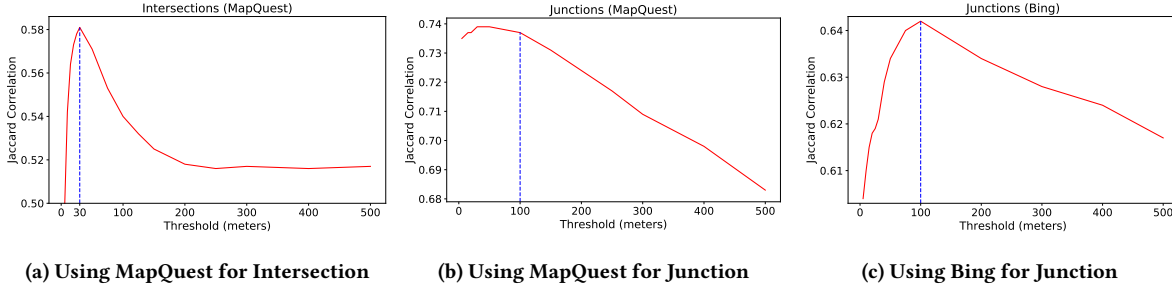
Using these thresholds, we augmented each accident record with points-of-interest. In summary, 27.5% of accident records were augmented with at least one of the available POI types in Table 1. Further discussion on annotation results are presented in Section 5.

**4.2.4 Augmenting with Period-of-Day.** Given the start time of an accident record, we used “TimeAndDate” API [22] to label it as *day* or *night*. We assign this label based on four different daylight systems, namely *Sunrise/Sunset*, *Civil Twilight*, *Nautical Twilight*, and *Astronomical Twilight*. Note that these systems are defined based on the position of the sun with respect to the horizon<sup>12</sup>.

## 5 US-ACCIDENTS DATASET

Using the process described in Section 4, we created a country-wide dataset of traffic accidents, which we name *US-Accidents*. US-Accident contains about 2.25 million cases of traffic accidents that took place within the contiguous United States from February 2016 to March 2019. Table 3 shows the important details of US-Accidents. Also, Figure 3 provides more details on characteristics of the dataset. Figure 3-(a) shows the daily distribution of traffic accidents, where significantly **more accidents were observed during the weekdays**. Based on parts (b) and (c) of Figure 3, it can be observed that the hourly distribution during weekdays has two peaks (**8am and 5pm**), while the weekend distribution shows a single peak (**1pm**). Figure 3-(d) **demonstrates that most of the accidents took place near junctions or intersections (crossing, traffic signal, and stop)**. MapQuest tends to report more accidents near intersections, while Bing reported more cases near junctions. This shows the complementary behavior of these APIs and the comprehensiveness of our dataset. Figure 3-(e) describes distribution of road types, extracted from the map-matching results (i.e., street names). **Here we note** that about 32% of accidents happened on or near local roads (e.g., streets,

<sup>12</sup>See <https://en.wikipedia.org/wiki/Twilight> for more details.



**Figure 2: Correlation study between regular-expression and OSM-based extracted annotations to find the best distance threshold values.**

avenues, and boulevards), and about **40% took place on or near high-speed roads** (e.g., highways, interstates, and state roads). We also note that Bing reported more cases on high-speed roads. Finally, the period of day data shows that about **73% of accidents happened after sunrise (or during the day)**.

**Table 3: US-Accidents: details as of March 2019.**

Total Attributes	45
Traffic Attributes (10)	id, source, TMC [23], severity, start_time, end_time, start_point, end_point, distance, and description
Address Attributes (8)	number, street, side (left/right), city, county, state, zip-code, country
Weather Attributes (10)	time, temperature, wind_chill, humidity, pressure, visibility, wind_direction, wind_speed, precipitation, and condition (e.g., rain, snow, etc.)
POI Attributes (13)	All cases in Table 1
Period-of-Day (4)	Sunrise/Sunset, Civil Twilight, Nautical Twilight, and Astronomical Twilight
Total Accidents	2,243,939
# MapQuest Accidents	1,702,565 (75.9%)
# Bing Accidents	516,762 (23%)
# Reported by Both	24,612 (1.1%)
Top States	California (485K), Texas (238K), Florida (177K), North Carolina (109K), New York (106K)

To further compare the US-Accidents dataset with the other publicly available sources, Table 4 provides some details in this regard. Regarding the size of data, US-Accidents is by far the largest available set. UK Accidents [24] is the only publicly available countrywide dataset, and its yearly reports are of about 100K accidents<sup>13</sup>. US-Accidents, however, contains about 750K accidents for each year. US-Accidents also provides many more details for each accident record than (say) New York Accidents [16].

### 5.1 Applications of the Dataset

US-Accidents may be used for applications such as real-time accident prediction; studying accident hotspot locations; casualty analysis (extracting cause and effect rules to predict accidents); or studying the impact of precipitation or other environmental stimuli

<sup>13</sup>Based on [24], there is no data reported for 2008.

**Table 4: Comparing publicly available, large-scale accident datasets**

Dataset	State	Country	Time	Size	Source
UK Accidents [24]	–	UK	2000–2016	1.6 M	Police Reported
Seattle Crash Report [20]	WA	–	2004–2018	208 K	Police Reported
Iowa Accidents [9]	IA	–	2008–2018	557 K	Iowa DOT
New York Accidents [16]	NY	–	2014–2016	1.65 M	NYS DMV
Maryland Accidents [13]	MD	–	2015–2018	400 K	Police Reported
US-Accidents	–	US	2016–2019	2.25 M	Streaming Data

on accident occurrence. Given the scale of data, researchers may utilize this dataset to derive a variety of insights which can benefit applications such urban planning and improving transportation infrastructures. In our own recent study, we employed the US-Accidents dataset along with the other traffic and weather events to perform pattern discovery over large-scale geo-spatiotemporal data, and revealed a variety of insights in terms of *propagation* and *influential* patterns [14].

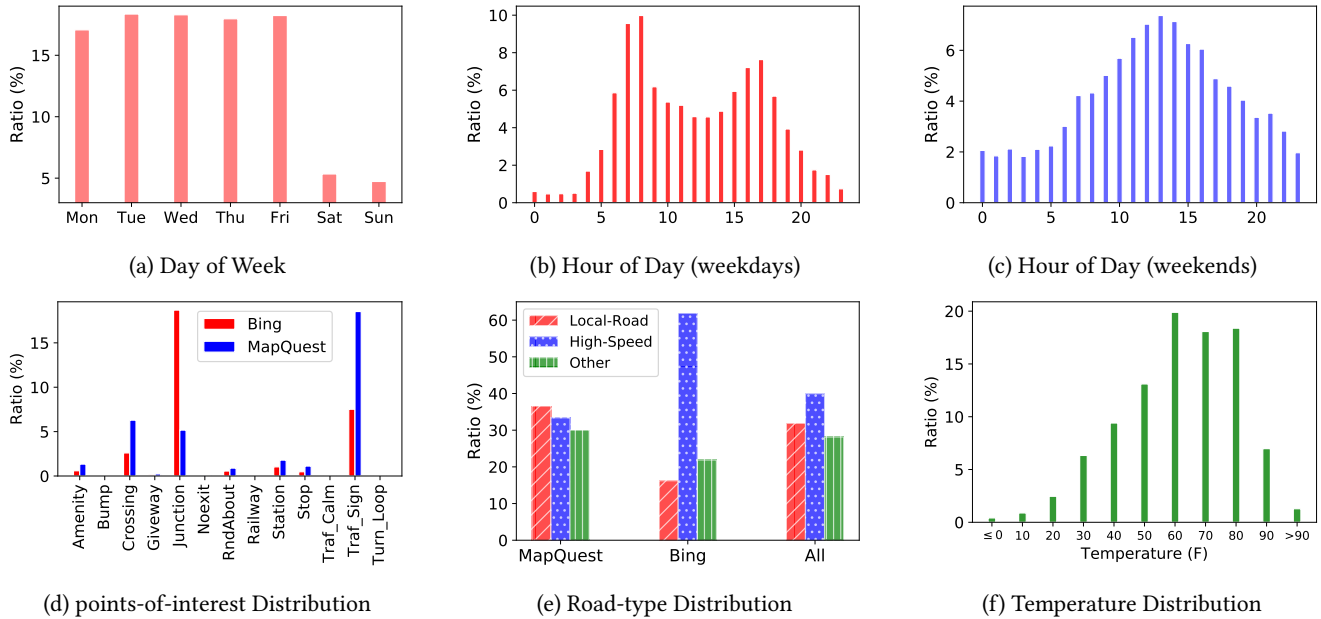
## 6 CONCLUSION AND FUTURE WORK

This paper describes US-Accidents, a unique, publicly available motor vehicle accident dataset, and its process of creation – that includes several important steps such as real-time traffic data collection, data integration, and multistage data augmentations using map-matching, weather, period-of-day, and points-of-interest data. To the best of our knowledge, US-Accidents is the first countrywide dataset of this scale, containing about 2.25 million traffic accident records collected for the contiguous United States over three years. From this dataset, we were able to derive a variety of insights with respect to the location, time, weather, and points-of-interest of an accident. We believe that US-Accidents provides a context for future research on traffic accident analysis and prediction. In terms of our own future work, we plan to employ this dataset to perform real-time traffic accident prediction.

## ACKNOWLEDGMENT

This work is supported by a grant from the Ohio Supercomputer Center (PAS0536).





**Figure 3: Characteristics of US-Accidents dataset, in terms of time analysis (a)–(c), points-of-interest-based augmentation distribution analysis (d), map-matching-based road type coverage analysis (e), and temperature analysis (f).**

## REFERENCES

- [1] Joaquín Abellán, Griselda López, and Juan De Oña. 2013. Analysis of traffic accident severity using decision rules via decision trees. *Expert Systems with Applications* 40, 15 (2013), 6047–6054.
- [2] Baton Rouge City (LA) Accidents. 2019. <https://data.brla.gov/Transportation-and-Infrastructure/Baton-Rouge-Traffic-Incidents/2tu5-7kif>. (2019). Accessed: 2019-05-05.
- [3] Bing Map Traffic API. 2019. <https://www.bingmapsportal.com/>. (2019). Accessed: 2019-05-05.
- [4] Ciro Caliendo, Maurizio Guida, and Alessandra Parisi. 2007. A crash-prediction model for multilane roads. *Accident Analysis & Prevention* 39, 4 (2007), 657–670.
- [5] Li-Yen Chang. 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety science* 43, 8 (2005), 541–557.
- [6] Li-Yen Chang and Wen-Chieh Chen. 2005. Data mining of tree-based models to analyze freeway accident frequency. *Journal of safety research* 36, 4 (2005), 365–375.
- [7] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- [8] Daniel Eisenberg. 2004. The mixed effects of precipitation on traffic crashes. *Accident analysis & prevention* 36, 4 (2004), 637–647.
- [9] IOWA Traffic Accidents. 2019. [http://data.iowadot.gov/datasets/cbd84abf01894f4a8404d6990ad2eb2e\\_0](http://data.iowadot.gov/datasets/cbd84abf01894f4a8404d6990ad2eb2e_0). (2019). Accessed: 2019-05-05.
- [10] Sachin Kumar and Durga Toshniwal. 2015. A data mining framework to analyze road accident data. *Journal of Big Data* 2, 1 (2015), 26.
- [11] Lei Lin, Qian Wang, and Adel W Sadek. 2015. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. *Transportation Research Part C: Emerging Technologies* 55 (2015), 444–459.
- [12] MapQuest Traffic API. 2019. <https://www.mapquest.com/>. (2019). Accessed: 2019-05-05.
- [13] Maryland Traffic Accidents. 2019. <https://catalog.data.gov/dataset/maryland-statewide-vehicle-crashes-cy2017-quarter-1-c6361>. (2019). Accessed: 2019-05-05.
- [14] Sobhan Moosavi, Mohammad Hossein Samavatian, Arnab Nandi, Srinivasan Parthasarathy, and Rajiv Ramnath. 2019. Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.
- [15] Alameen Najjar, Shun-ichi Kaneko, and Yoshikazu Miyayaga. 2017. Combining satellite imagery and open data to map road safety. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- [16] New York Traffic Accidents. 2019. <https://data.ny.gov/Transportation/Motor-Vehicle-Crashes-Vehicle-Information-Three-Year-24f>. (2019). Accessed: 2019-05-05.
- [17] Nominatim Tool. 2019. <https://wiki.openstreetmap.org/wiki/Nominatim>. (2019). Accessed: 2019-05-05.
- [18] Open Street Map (OSM). 2019. <https://www.openstreetmap.org/>. (2019). Accessed: 2019-05-05.
- [19] World Health Organization. 2015. *Global status report on road safety 2015*. World Health Organization.
- [20] Seattle (WA) Traffic Accidents. 2019. [http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab\\_0](http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0). (2019). Accessed: 2019-05-05.
- [21] JD Tamerius, X Zhou, R Mantilla, and T Greenfield-Huitt. 2016. Precipitation effects on motor vehicle crashes vary by space, time, and environmental conditions. *Weather, Climate, and Society* 8, 4 (2016), 399–407.
- [22] Time And Date website. 2019. <https://www.timeanddate.com/>. (2019). Accessed: 2019-05-05.
- [23] Traffic Message Channel (TMC) Code. 2019. [https://wiki.openstreetmap.org/wiki/TMC/Event\\_Code\\_List](https://wiki.openstreetmap.org/wiki/TMC/Event_Code_List). (2019). Accessed: 2019-05-05.
- [24] United Kingdom Traffic Accidents. 2019. <https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales>. (2019). Accessed: 2019-05-05.
- [25] Weather Underground. 2014-2019. <https://www.wunderground.com/>. (2014-2019). Accessed: 2019-05-05.
- [26] Lu Wenqi, Luo Dongyu, and Yan Menghua. 2017. A model of traffic accident prediction based on convolutional neural network. In *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*. IEEE, 198–202.
- [27] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 984–992.
- [28] Zhuoning Yuan, Xun Zhou, Tianbao Yang, James Tamerius, and Ricardo Mantilla. 2017. Predicting traffic accidents through heterogeneous urban data: A case study. In *Proceedings of the 6th International Workshop on Urban Computing (UrbComp 2017)*, Halifax, NS, Canada, Vol. 14.