Data in Brief 55 (2024) 110675

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A traffic accident dataset for Chattanooga, Tennessee

Andy Berres[a,b,1,2,3,*], Pablo Moriano[c,1,2,3], Haowen Xu[b,1,2,3], Sarah Tennille[b,1,2,3], Lee Smith[d,1,2,3], Jonathan Storey[d,1,2,3], Jibonananda Sanyal[a,1,2,3]

[a] *Energy Conversion and Storage Systems Center, National Renewable Energy Laboratory, 15301 Denver West Parkway, Mail Stop RSF 042, Golden, CO 80401, United States*
[b] *Computational Sciences and Engineering Division, Oak Ridge National Laboratory, PO Box 2008, MS-6085, Oak Ridge, TN 37830, United States*
[c] *Computer Science and Mathematics Division, Oak Ridge National Laboratory, PO Box 2008, MS-6013, Oak Ridge, TN 37830, United States*
[d] *Tennessee Department of Transportation, James K. Polk Bldg., Suite 700, 505 Deaderick Street, Nashville, TN 37243, United States*

ARTICLE INFO

Article history:
Received 20 December 2023
Revised 17 June 2024
Accepted 19 June 2024
Available online 4 July 2024

Dataset link: A Tagged Traffic Accident Dataset for Machine Learning (Original data)

ABSTRACT

This publication presents an annotated accident dataset which fuses traffic data from radar detection sensors, weather condition data, and light condition data with traffic accident data (as illustrated in Fig. 1) in a format that is easy to process using machine learning tools, databases, or data workflows. The purpose of this data is to analyze, predict, and detect traffic patterns when accidents occur. Each file contains a timeseries of traffic speeds, flows, and occupancies at the sensor nearest to the accident, as well as 5 neighboring sensors upstream and downstream. It also contains information about the accident type, date, and time. In addition to the accident data, we provide baseline data for

DOI of original article: 10.1016/j.eswa.2023.122813
* Corresponding author at: Energy Conversion and Storage Systems Center, National Renewable Energy Laboratory, 15301 Denver West Parkway, Mail Stop RSF 042, Golden, CO 80401, United States.
E-mail address: andy.berres@nrel.gov (A. Berres).
[1] @nrel
[2] @ornl
[3] @mytdot

https://doi.org/10.1016/j.dib.2024.110675

Keywords:
Radar sensor data
Weather conditions
Light conditions
Timeseries
Annotated data
Incident data
Machine learning
Transportation

typical traffic patterns during a given time of day. Overall, the dataset contains 6 months of annotated traffic data from November 2020 to April 2021. During this timeframe, and 361 accidents occurred in the monitored area around Chattanooga, Tennessee. This dataset served as the basis for a study on topology-aware automated accident detection for a companion publication [1].

Specifications Table

Subject	Transportation Management.
Specific subject area	Transportation Safety.
Type of data	Table, Analyzed, Filtered, Fused, Processed
Data collection	The accident data were acquired through TDOT's Enhanced Tennessee Roadway Information Management System (E-TRIMS) [3]. Data collection of this dataset was performed by local law enforcement officers and state patrol, and it was cleaned and reviewed to remove identifiable information by TDOT. The radar sensor data [4] were acquired from sensors maintained by the Tennessee Department of Transportation (TDOT). These sensors are placed along the highway system in Tennessee's metropolitan areas intervals of about half a mile. The weather data were acquired from NASA's Prediction Of Worldwide Energy Resources (POWER) [5]. The light condition data were acquired from Sunrise-Sunset.org [6].
Data source location	Chattanooga, Tennessee, USA
Data accessibility	Repository name: A Tagged Traffic Accident Dataset for Machine Learning Data identification number: 10.5281/zenodo.7964287 Direct URL to data: https://doi.org/10.5281/zenodo.7964287 Direct download from Zenodo
Related research article	P. Moriano, A. Berres, H. Xu, J. Sanyal, Spatiotemporal Features of Traffic Help Reduce Automatic Accident Detection Time, 2024. Expert Syst Appl. 244, 122813. https://doi.org/10.1016/j.eswa.2023.122813.

1. Value of the Data

- There is currently no Findable, Accessible, Interoperable and Reusable (FAIR) dataset for traffic accidents and traffic data. Accident data is generally not publicly available, and much less directly linked to traffic data. This FAIR dataset will serve as a baseline for comparable scientific results.
- The dataset is particularly useful to individuals developing machine learning methods to detect and analyze traffic behavior during accidents, as it provides tagged traffic data for accident scenarios and regular (non-accident) traffic.
- These data can be used to develop new accident detection or prediction techniques and compare different techniques with each other. They can furthermore be used to develop new metrics and visualizations.
- There are many open datasets for machine learning for topics from handwriting recognition and power consumption to stock market and medical datasets. However, there are no similar datasets for traffic safety. This publication aims to fill this gap.

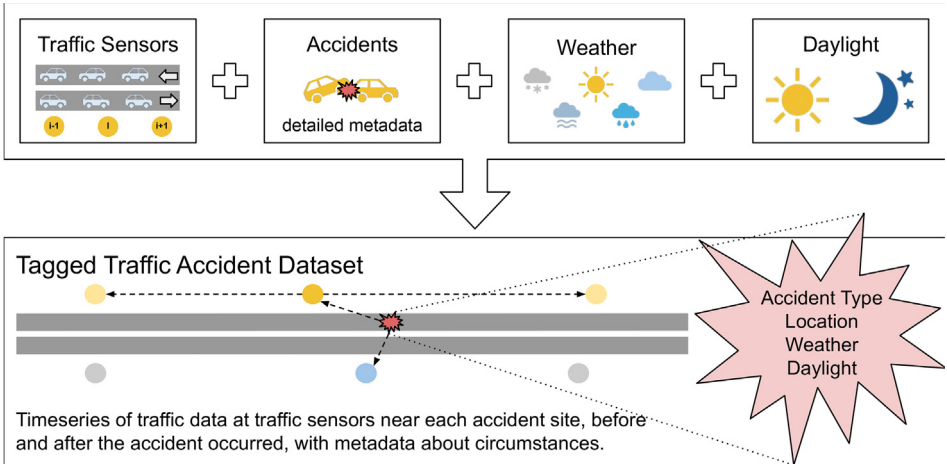

Fig. 1. This dataset combines data from traffic sensors, accident data, weather data, and daylight information into a comprehensive tagged traffic accident dataset. This dataset contains the accident information, as well as traffic data from the nearest sensor and its neighbors.

2. Background

When developing the methodology for our companion research article titled "Spatiotemporal features of traffic help reduce automatic accident detection time" [1], we wanted to directly compare the performance of our method with that from methods in scientific literature. However, we were unable to locate or get access to any of the datasets used for these works. This data publication is our response in trying to promote reproducibility of scientific work, and faithful comparisons between different methodologies. We were able to obtain both traffic data and accident data from TDOT, but they are separate datasets. Machine learning requires tagged data, which we provide with the presented dataset. We have furthermore enriched the data with weather and light conditions. The raw accident data from TDOT also included weather and light, but adding this data from a separate source enabled us to provide this information for non-accident data and use it for validation of accident times. The resulting dataset is published on zenodo.org [2] and described in detail in this manuscript.

3. Data Description

The zip folder ***annotatedData*** contains two subfolders: ***allData*** and ***bestData***. The *bestData* folder contains all data for which a full neighborhood of five sensors upstream and five sensors downstream is available, whereas *allData* includes everything from *bestData* as well as data with a smaller number of neighboring sensors. Each folder contains one subfolder called ***accidents*** and one subfolder called ***non-accidents***. The *accidents* folder contains one file per accident. The *non-accidents* folder contains files for the same location, day of the week and time as a corresponding accident, for each week during which there was no accident impact on the traffic.

The file names in both folders are formatted as follows: **yyyy-mm-dd-hhmm-rrrrrXaaa.a.csv**, consisting of date (yyyy-mm-dd), time (hhmm in 24-h format), and sensor name (rrrrrXaaa.a), which consists of road name (rrrrr; 5 alphanumerical characters), heading (X), and mile marker (aaa.a). For example, the file *2020-11-03-1611-00I24W182.8.csv* contains data for an accident which occurred at 4:11 p.m. on November 3, 2020, on I-24 Westbound near the radar sensor at mile marker 182.8. For more information on the radar data and its potential uses, please refer to [7–9].

Table 1
Columns of each timeseries file.

Column name	Interpretation
incident at sensor(i)	1 for yes (*accidents* folder), 0 for no (*non-accidents* folder).
road	Road name with heading, e.g., 00I24E.
mile	Mile marker of nearest radar sensor, e.g., 182.8.
type	Accident type, e.g., "Prop Damage (over)" for property damage exceeding a threshold of $400. For non-accidents, the type is given as "None".
date	Date of the data sample. For accidents, this is the date on which the accident occurred. For non-accidents, this is the date for which the non-accident data sample is collected.
incident_time	Time the reference accident was reported in hh:mm. This is the time which is provided in E-TRIMS as the time the 911 call was made.
incident_hour	Only the hour from the incident_time, in integer format.
data_time	Timestamp for the timeseries contained in the file in hh:mm:ss format. The timeseries consists of 30 s timesteps.
weather	Weather during *data_time*, based on data collected from NASA POWER. We used dry bulb temperature (°C), precipitation (mm/h), and wind speed (m/s) from the raw NASA POWER data to produce the classifications of *rain* (at least 1 mm precipitation and temperatures above 2 °C), *snow* (at least 1 mm precipitation and temperatures at or below 2 °C), and *wind* (wind speeds over 30 mph or 13.5 m/s). If there were no inclement weather conditions, we set the category to "–".
light	Light conditions during data_time. To produce this field, we collected sunrise, sunset, civil twilight start and civil twilight end times from https://sunrise-sunset.org, and derived the categories dawn, daylight, dusk, and dark using these start and end times.
Radar data	The last 33 columns contain radar data for the 11 sensors surrounding the accident or non-accident. For each sensor, we collected *speed* (mean over 30-s interval in miles per hour, or empty if no vehicles passed), *volume* (count of all vehicles passing during 30-s interval), and *occupancy* (mean% of occupancy over 30-s interval). These three variables are grouped in triples, of **speed (k), volume (k), occupancy (k)**, where k indicates the sensor number relative to the closest sensor i to the incident, $k < i$ indicate upstream sensors and $k > i$ indicate downstream sensors. For example, **speed (i-5)** refers to the mean speed at the sensor which is 5 hops upstream from the accident, and **volume(i + 1)** refers to the number of vehicles at the sensor immediately downstream from the accident.

Table 2
Summary of additional metadata provided.

Filename	Content
Accidents.csv	Cleaned-up accidents file with all accidents which happened on Chattanooga area highways between November 1, 2020, and April 29, 2021. We have removed accidents which happened on non-highway roads, and we have corrected the timestamps (which were in 12-h format but missing a.m./p.m. markers) by cross-referencing light and weather conditions.
WeatherDict.json	Dictionary containing the weather data synthesized from NASA POWER.
LightDict.json	Dictionary containing the light data synthesized from Sunrise-and-Sunset.
SensorTopology.csv	Neighborhood information for each radar sensor in the Chattanooga area.
SensorZones.geojson	Polygons used to determine the nearest radar sensor for each accident location. Each polygon is tagged with the corresponding radar sensor's name.

The content of each CSV file is a timeseries of radar data beginning 15 min prior to the reported incident and ending 15 min after the reported incident. It also contains metadata, such as the accident type, etc. Each CSV file contains the columns listed in Table 1.

The folder *metaData* contains the files described in Table 2.

4. Experimental Design, Materials and Methods

The published dataset depends on four other sources: accidents, traffic conditions, weather, and light. It furthermore includes additional useful data, such as the sensor topology. In the

Table 3
Accident data provided by E-TRIMS.

Column Name	Definition
County	Full county name.
Route	5-character road name, buffered with zeros between letter and number of the road.
Year of Crash	Year of Crash (yyyy).
Date of Crash	Date of Crash (m/d/yyyy).
Time of Crash	Time of Crash: hhmm (int); missing a.m./p.m. information.
Type of Crash	Identical to *type* in annotated data.
Relation to First Junction	Descriptor of road situation: acceleration/deceleration lane, intersection, driveway, rail grade crossing
Type and severity of crash	Most severe consequences of crash: Property damage, suspected minor/major injury, fatality.
Relation to First Roadway	Crash location with respect to road: On roadway, shoulder, median, parking lane, parking lot.
Total Killed	Number of deaths.
Total Inj	Number of individuals with injuries.
Total Incap Injuries	Number of individuals with incapacitating injuries.
Total Other Injuries	Number of individuals with other injuries.
Total veh	Number of vehicles involved in crash.
First Harmful Event	What was hit first? Various structures and living beings one can collide with, such as fire hydrant, snowbank, curb, deer; other incidents like cargo loss or shift.
Manner of First Collision	How did the collision happen? Sideswipe, rear-end, head-on, angle.
Weather Cond	Weather conditions during the crash, e.g., clear, cloudy, rain, fog, sleet/hail, snow, blowing snow, severe crosswinds, blowing sand/soil/dirt, smog/smoke, other, and unknown. Most accidents (1105/1593) were reported during clear weather. 247 occurred during rain, 87 under unknown or undocumented ("–") conditions, and 11 during other conditions (sleet/hail, snow, blowing debris, fog, or other).
Light Conditions	Light conditions during the crash, e.g. daylight, dusk, dawn, dark (lighted, not lighted, unknown lighting), other, and unknown. Most accidents were reported as daylight (990/1593), 467 occurred in the dark (200 in conditions with artificial lighting). 73 during twilight conditions, and 14 with unknown or other lighting conditions.
Locate type	How was the location entered? automatic (GPS) or manual.
Latitude Longitude	Geocoordinates.
HAZMAT Involved	Was hazardous material released? yes/no/unknown
Hit and Run	Yes/no (239 yes, 1354 no).
Hwy Const Zone	Highway construction, maintenance, other special conditions, "—".

following, we describe each of these datasets, and how we transformed them for this paper. Finally, we describe how we synthesized the final dataset from these data sources.

The **accident dataset** (Accidents.csv) is an anonymized extract of a richer dataset which TDOT maintain. This extract is updated weekly and contains all the relevant data for accident detection. It is tabular and contains numerous columns. Columns A-W contain original data whereas columns X-AJ contain augmented data. In the following, we focus on data description but for a deeper dive into the different parameters, please refer to [10].

The original traffic accident data has the columns defined in Table 3.

We augmented the columns defined in Table 4 to the original data.

In an accident detection scenario, only date and time, geolocation, and weather and lighting conditions are available, so these are the only ones we used.

Although this is a rich dataset, there were a few issues:

1. Weather conditions: Not all accidents were tagged with weather condition information. This is solved by using consistent synthesized weather data (based on NASA weather data) for accidents and non-accidents alike.

Table 4
Additional data provide, as described in this paper.

Column Name	Definition
Date	yyyy-mm-dd.
Year	Integer.
Month	
Day	
Hour	
Minute	
Time_orig	Original time as hh:mm (instead of integer) in 12 h format.
Time_fixed	Updated time to supplement a.m./p.m. information, in 24 h format using the methods described below (part of light data processing).
Light_orig_simple	Original light data simplified to fewer categories.
Light_synth	Synthesized light data based on original time.
Light_fixed	Updated light data using the methods described below.
Weather_orig_simple	Original weather data simplified to fewer categories.
Weather_fixed	Weather_fixed: synthesized weather data.

Table 5
Relevant data obtained from NASA's POWER dataset.

POWER Variable Name	Definition
Temperature (C)	Average air (dry bulb) temperature at 2 m above the surface.
Precipitation (mm/hour)	Average of total precipitation at the surface, including water content in snow.
Wind Speed (m/s) at 2 m	Wind speed at 2 m above the surface.
Surface Pressure (kPa)	Average pressure at the surface of the earth.

2. Lighting conditions: Not all accidents were tagged with lighting condition information. This is solved by using consistent synthesized light condition data (based on sunrise, sunset, and twilight times) for accidents and non-accidents alike.
3. Time: In Region 2 (i.e., the region of Tennessee surrounding Chattanooga), all timestamps in the studied time frame were in 12-h format, but missing a.m. and p.m. tags. We addressed this using the light condition data as described further below.

The **traffic condition data** are collected from radar sensors, which emit low-energy microwave radiation that is reflected by the vehicles (Xu et al., 2022), and captured by the sensor at lane resolution. At 30-s intervals, these sensors collect the following parameters:

- Sensor name and ID. The sensor name contains the road, heading, and mile marker in a fixed naming scheme.
- Lane name and ID.
- Average speed in miles per hour.
- Vehicle count.
- Occupancy.
- Vehicle class based on length (note that this feature is not functional for many sensors in this region, so we omitted it).

To create the accident dataset, we aggregated the data across lanes as tracking lane-level topology would have added a substantial level of additional complexity to an already complex dataset.

The **weather condition data** (WeatherDict.json) from the accident dataset contained many different weather conditions. In a first step, we simplified these categories to a smaller set which is reproducible using other data sources: rain, snow (snow, sleet/hail, blowing snow), wind (severe crosswinds, blowing sand/soil/dirt), and unknown (clear, cloudy, other, unknown). We then used POWER data to reproduce these categories from the hourly measurements. We used the variables described in Table 5 from this dataset.

Table 6
Weather conditions derived from POWER data.

Weather Condition	Definition
Precipitation	According to the United States Geological Survey (USGS), a "heavy (thick) drizzle" is defined as 1 mm of precipitation per hour, and it can impair visibility. We therefore used 1 mm as our threshold to determine precipitation, and we classify it as • Rain if the temperature was >2 °C, or • Snow if the temperature was ≤ 2 °C.
Wind	According to the National Weather Service, sustained winds of 30 mph or wind gusts of 45 mph can make it difficult to drive high-profile vehicles, and small objects may be blown around. Based on this definition, we used the category wind if wind speeds exceeded 13.5 m/s (30 mph).
(unknown)	If the category was neither rain, snow, nor wind, we define it as unknown.
When multiple categories were possible (e.g., blowing snow could be wind or snow in this scenario), we chose the category with a bigger traffic impact based on FHWA's report.	

Table 7
Definition of light conditions based on sunrise, sunset, and twilight times.

Light Condition	Definition
Dawn	Civil twilight start until sunrise,
Daylight	Sunrise until sunset,
Dusk	Sunset until civil twilight end, and
Dark	Civil twilight end until civil twilight start.

We used these basic measurements to synthesize weather conditions for the four categories, as outlined in Table 6.

We created a dictionary for efficient access by date and hour. This allowed us to quickly look up weather information for any given timestamp. For example, to access the weather conditions on November 12 at 1 p.m., one would use weather['2020-11-12']['13'].

The **lighting condition data** (LightDict.json) from the accident data contained various lighting conditions, including artificial lighting as well as natural daylight conditions. Working on the assumption that artificial lighting conditions did not change during the study period (i.e., no additional lights were added, and there were no widespread outages), we focused on only the natural daylight information. The sunrise-and-sunset (Sunrise Sunset) dataset has timestamps at minute resolution for

- Civil, nautical, and astronomical twilight start.
- Sunrise.
- Solar noon.
- Sunset, and
- Civil, nautical, and astronomical twilight end.

For this work, we chose civil twilight because it accounts for the local topography which can affect actual lighting conditions. We then aggregated the data into four light conditions, as listed in Table 7.

To prepare light data for more efficient access, we created a dictionary which lets us look up light conditions by date and timestamp. For example, to access the weather conditions on November 12 at 13:14 p.m., one would use light['2020-11-12']['13:14'].

For each timestamp, we check which lighting condition it corresponds to by using the mapping provided in the list above. In addition, this data allowed us to address the issue of missing a.m./p.m. information in the accident data. We cross-referenced the light conditions between the originally reported lighting at the time of accident and synthesized lighting data. If the information matched in the morning, we kept the time as a.m., whereas if it matched in the evening,

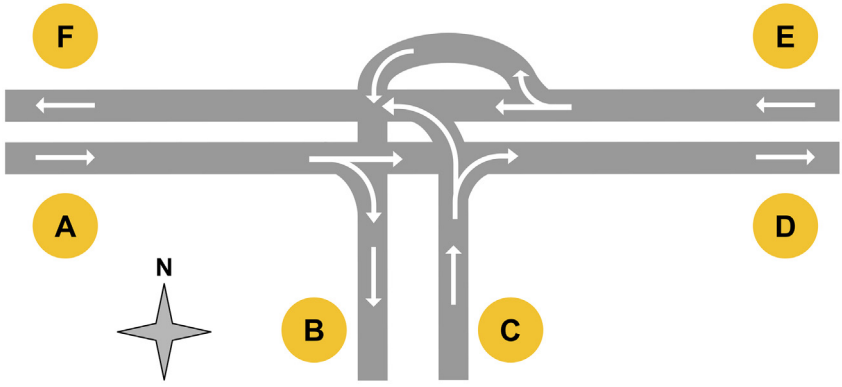

Fig. 2. Sketch of a highway junction and sensors A-F to illustrate neighborhood definitions.

we added 12 h to move the timestamp into the p.m. range. Note that this strategy still maintains a bias for a.m. time because there can be more or less than 12 daylight hours, but the majority of daytime accidents are captured correctly. All timestamps in the resulting dataset are provided in 24-h format to avoid future issues.

The **sensor topology** (SensorTopology.csv) information is needed twofold: we need a topology file that provides the previous (upstream) and next (downstream) sensors for each sensor, and we need a geometry file that contains polygons that correspond to each road segment.

To generate the topology file, we took advantage of the radar sensor names, which contain information about the road, heading, and mile marker. We know that mile markers increase in travel direction for Northbound and Eastbound highways, and decrease for Southbound and Westbound highways. Based this knowledge, we can automatically generate most of the sensor topology. For instance, if we have two Northbound sensors at mile markers 2.0 and 2.5, with no sensors in between, they are considered neighbors. As Northbound traffic has increasing mile markers, we add 2.0 as upstream neighbor for 2.5, and we add 2.5 as downstream sensor for 2.0. If these sensors were Southbound, the relationship would be reversed. There are two special cases: we can have no neighbor (end of detection range, or gap of more than 5 miles) and will leave the corresponding neighbor field empty. If there is more than one neighbor, as is the case at highway junctions, these neighborhoods were resolved manually. We determined one upstream neighbor, one downstream neighbor, and one neighbor in the opposite travel direction. For the sensors with two neighbors, we bias towards staying on the same road, or the direction with higher traffic volumes.

For example, consider the graphic in Fig. 2. Each sensor has either two visible upstream neighbors or two visible downstream neighbors. For sensors A, D, E, and F, one of the two neighbors is on the same road, which makes it a preferred neighbor.

- Sensor A has two downstream neighbors (B to the South, D to the East). As D is on the same road, it is a preferred neighbor.
- Sensor D has two upstream neighbors (A to the West, C to the South). As A is on the same road, it is a preferred neighbor.
- Sensor E has two downstream neighbors (B to the South, F to the West). As F is on the same road, it is a preferred neighbor.
- Sensor F has two upstream neighbors (C to the South, E to the East). As E is on the same road, it is a preferred neighbor.

For sensors B and C, both visible neighbors are associated with a different road, so the decision of preferred neighbor must be made based on traffic volumes.

Table 8
Description of the columns contained in the sensor topology CSV file.

Column Name	Definition
Road	5-character road name, buffered with leading zeros.
Mile	Nearest mile marker (0.1-mile precision).
Heading	Travel direction (single character for cardinal directions).
Name	Sensor name consisting of region code, road name, 5-character mile marker, and heading. E.g., R2G-00I75-000.2S is a sensor located at mile marker 0.2 on I-75S.
Previous	Name of upstream sensor.
Next	Name of downstream sensor.
Opposite	Name of nearest sensor for opposite heading.
PrevDist, NextDist, OppoDist	Distance from previous/next/opposite sensor (in miles).
Latitude, Longitude	Geocoordinates.

Table 9
Description of the properties associated with each road segment in the GeoJSON file.

Property	Definition
RDS_Sensor	Sensor name.
NBR_LANES	Number of lanes.
SPD_LMT	Speed limit (mph).
uniqueID	Unique ID for the polygon.

- Sensor B has two upstream neighbors (A from the West, E from the East).
- Sensor C has two downstream neighbors (D to the East, F to the West).

The sensor topology file contains the columns summarized in Table 8.

The **Sensor Zones** (SensorZones.geojson) are polygons outlining each sensor's section of road for fast assignment of crashes to sensors. We developed a GIS workflow which maps individual sensor IDs to their relevant road geometry, using GDAL and QGIS software, which are freely available open-source technologies. The workflow starts with the generic centerline geometry of individual highway segments and creates offset lines to represent two opposite driving directions. To further segment the space, we generate Voronoi polygons by using RDS sensor locations as the Voronoi nodes. Each Voronoi polygon is a region of a plane that is closer to its Voronoi node (the RDS sensor it corresponds to) than to any other RDS sensor. Using these spatially partitioned polygons, we can define the adjacent detection area of the RDS sensors while avoiding the creation of overlapping detection areas between nearby sensors. Next, we spatially clip unidirectional highway line segments by using the boundary of the sensors' Voronoi polygons, and we map the sensor IDs to these road segments. The major challenge presents itself when two adjacent sensors are placed at different elevations, such as elevated ramps and bridges (such as in the graphic above). We manually inspected and justified these spatial cases. The resulting geometry is saved as a GeoJSON file which contains the name of the corresponding sensor as one of its properties.

Each polygon in the GeoJSON file contains the properties listed in Table 9.

Finally, we produced the **tagged accident data** as detailed in the data description part in this manuscript. This data consists of one file per accident (or non-accident). Each file contains the following types of columns summarized in Table 10.

To create these files, we first iterated over all events. For each event, we created a file with the attributes listed above. The first few columns (list items 1–4) are fixed for all time steps. We determined the time at which the event occurred (e.g., 18:35), and we created a list of the required time steps from 15 min before the event to 15 min after the given time. Because we cannot know the second at which the event occurred, we used both 30-s time intervals for each

Table 10
Contents of tagged accident data files.

Column Type	Definition
Accident Status	A boolean signifies whether the event is an accident (i.e., 1) or a non-accident (i.e., 0).
Location Information	The road that the original accident occurred on and the mile marker it occurred at (e.g., 00I75S and 4.8 if the original accident occurred near mile marker 4.8 on I-75 southbound).
Accident Information	The type of event (e.g., Prop Damage [over] for an accident with property damage over a predefined threshold). If the event is a non-accident, we set this column to None.
Accident Time	The event's date, time (e.g., 18:35) and hour (e.g., 18). These columns have the date/time/hour of the accident recorded in E-TRIMS, and they remain the same value for the entire file.
Traffic Data	The sensor data's time. This column contains the timestamp of the sensor data contained in each row. Triplets of speed(i), volume(i), and occupancy(i) for each sensor from 5 sensors upstream to 5 sensors downstream (e.g., speed(i-5), volume(i-5), occupancy(i-5) ..., speed(i), volume(i), occupancy(i), ..., speed($i+5$), volume($i+5$), occupancy($i+5$)) from the data.
Environment	Weather and lighting conditions (e.g., Rain and Dusk).

minute, which resulted in 62 time steps (or rows) per file. For our example event, this means that we used time steps from 18:20:00 to 18:50:30.

For the remaining columns (list items 5–7), we collected information from multiple data products described in previous subsections. First, we determined which sensor is closest to the event location by using a point-in-polygon test with the polygons from the geometry file (SensorZones.geojson).

Next, we determined the list of relevant sensors near the event by using the sensor topology information to find each sensor's upstream and downstream neighbors up to 5 hops away. We sorted the sensors in the order the traffic passes them, starting at 5 sensors upstream, to the sensor nearest to the event, and continuing to 5 sensors downstream. This gave us a total of 11 sensors. For each of these sensors, we aggregated speeds, volumes, and occupancies across all lanes. For speed and occupancy, we used the mean speed and occupancy across all lanes. If there were no vehicles, and no speed/occupancy was recorded, then we set the value to NaN. For volume, we summed up the vehicle counts across all lanes. If there were fewer than 5 sensors, or the data had gaps, then the corresponding cells remain empty.

Finally, we fetched the current weather and lighting data for each time step.

Limitations

The raw traffic accident data had missing am/pm information, as well as gaps in weather and light condition information. We have imputed this missing data as described in the previous section, however, it is not ground truth as recorded by a human at the time of the accident.

Ethics Statement

We did not conduct human or animal studies and we had permission to use the NASA and E-TRIMS primary data per terms of use (NASA) and communication with the data owners (who are included in the authors list, for E-TRIMS) respectively.

CRediT Author Statement

Andy Berres: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization; **Pablo Moriano**: Validation, Formal Analysis, Writing - Review & Editing; **Haowen Xu**: Software, Data Curation, Writing - Review & Editing; **Sarah Tennille**: Conceptualization, Data Curation, Writing - Review & Editing; **Lee Smith**: Resources; **Jonathan Storey**: Resources; **Jibonananda Sanyal**: Conceptualization, Project administration, Funding acquisition, Writing - Review & Editing

Data Availability

A Tagged Traffic Accident Dataset for Machine Learning (Original data) (Zenodo)

Acknowledgments

The authors would like to thank all parties who provided the datasets this work is based on: The accident and traffic data were provided by Tennessee Department of Transportation. These weather data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program. Sunrise and sunset times were obtained from https://sunrise-sunset.org.

The authors would furthermore like to thank the Tennessee Department of Transportation and the Chattanooga Department of Transportation for their continued partnership and guidance.

Finally, the authors would like to thank the U.S. Department of Energy (DOE) Vehicle Technologies Office for funding this work.

This dataset has been prepared in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the DOE under Contract No. DE-AC36-08GO28308 and in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan)

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Moriano, A. Berres, H. Xu, J. Sanyal, Spatiotemporal features of traffic help reduce automatic accident detection time, Expert Syst. Appl. 244 (2024) 122813, doi:10.1016/j.eswa.2023.122813.
[2] [Dataset] A. Berres, P. Moriano, H. Xu, S. Tennille, L. Smith, J. Storey, J. Sanyal, A tagged traffic accident dataset for machine learning, Zenodo 1 (2023). v, doi:10.5281/zenodo.7964287.
[3] Tennessee Department of TransportationEnhanced Tennessee Roadway Information Management System (E-TRIMS), 2023. https://e-trims.tdot.tn.gov/.
[4] Tennessee Department of Transportation, Radar Detection Sensor Data, 2023. .
[5] The POWER Project, 2023. [Dataset] https://power.larc.nasa.gov.
[6] Sunrise – Sunset, Sunrise and Sunset Times in Chattanooga, TN, 2023. https://sunrise-sunset.org/us/chattanooga-tn.

[7] H. Xu, A. Berres, S.A. Tennille, S.K. Ravulaparthy, C. Wang, J. Sanyal, Continuous emulation and multiscale visualization of traffic flow using stationary roadside sensor data, IEEE Trans. Intell. Transp. Syst. 23 (8) (2022) 10530–10541, doi:10.1109/TITS.2021.3094808.
[8] H. Xu, A. Berres, S.B. Yoginath, H. Sorensen, P.J. Nugent, J. Severino, S.A. Tennille, A. Moore, W. Jones, J. Sanyal, Smart mobility in the cloud: enabling real-time situational awareness and cyber-physical control through a digital twin for traffic, IEEE Trans. Intell. Transp. Syst. 24 (3) (2023) 3145–3156, doi:10.1109/TITS.2022.3226746.
[9] A.S. Berres, T.J. LaClair, C.R. Wang, H. Xu, S. Ravulaparthy, A. Todd, S.A. Tennille, J. Sanyal, Multiscale and multivariate transportation system visualization for shopping district traffic and regional traffic, Transp. Res. Rec. 2675 (6) (2021) 23–37, doi:10.1177/0361198120970526.
[10] A.S. Berres, H. Xu, S.A. Tennille, J. Severino, S. Ravulaparthy, J. Sanyal, Explorative visualization for traffic safety using adaptive study areas, Transp. Res. Rec. 2675 (6) (2021) 51–69, doi:10.1177/0361198120981065.