Assessing Trustworthiness of Crowdsourced Flood Incident Reports using Waze Data: A Norfolk, Virginia Case Study

Shraddha Praharaj
Graduate Research Assistant
Department of Engineering Systems & Environment
University of Virginia, Charlottesville, VA 22903
spraharaj@virginia.edu

Faria Tuz Zahura
Graduate Research Assistant
Department of Engineering Systems and Environment
University of Virginia, Charlottesville, VA 22903
fz7xb@virginia.edu

T. Donna Chen*
Assistant Professor
Department of Engineering Systems and Environment
University of Virginia, Charlottesville, VA 22903
tdchen@virginia.edu

Yawen Shen
Graduate Research Assistant
Department of Engineering Systems and Environment
University of Virginia, Charlottesville, VA 22903
ys5dv@virginia.edu

Luwei Zeng
Graduate Research Assistant
Department of Engineering Systems and Environment
University of Virginia, Charlottesville, VA 22903
lz6ct@virginia.edu

Jonathan L. Goodall
Professor
Department of Engineering, Systems and Environment
University of Virginia, Charlottesville, VA 22903
goodall@virginia.edu

* Corresponding Author

This is an author version of the paper published at the following reference. Please use this reference when citing the paper.

Praharaj, S., Zahura, F.T., Chen, T.D., Shen, Y., Zeng, L. and Goodall, J.L., 2021. Assessing Trustworthiness of Crowdsourced Flood Incident Reports Using Waze Data: A Norfolk, Virginia Case Study. Transportation Research Record, p.03611981211031212.
https://doi.org/10.1177/03611981211031212

ABSTRACT

Climate change and sea-level rise are increasingly leading to higher and prolonged high tides, which, in combination with the growing intensity ofrainfall and storm surges, and insufficient drainage infrastructure, result in frequent recurrent flooding in coastal cities. There is a pressing need to understand the occurrence of roadway flooding incidents in order to enact appropriate mitigation measures. Agency data for roadway flooding events are scarce and resource-intensive to collect. Crowdsourced data can provide a low-cost alternative for mapping roadway flood incidents in real time; however, the reliability is questionable. This research demonstrates a framework for asserting trustworthiness on crowdsourced flood incident data in a case study of Norfolk, Virginia. Publicly available (but spatially limited) flood incident data from the city in combination with different environmental and topographical factors are used to create a logistic regression model to predict the probability of roadway flooding at any location on the roadway network. The prediction accuracy of the model was found to be 90.5%. When applying this model to crowdsourced Waze flood incident data, 71.7% of the reports were predicted to be trustworthy. This study demonstrates the potential for using Waze incident report data for roadway flooding detection, providing a framework for cities to identify trustworthy reports in real-time to enable rapid situation assessment and mitigation to reduce incident impact.

Keywords: crowdsourced data, flooding, trustworthiness, logistic regression, incident management

INTRODUCTION

In recent years, crowdsourced data has emerged as a low-cost method for data collection in various fields. In the transportation domain, there are many areas of research which have insufficient or non-existent agency-provided data, where crowdsourced data shows promise to be a useful alternative resource for research and analysis in these domains, such as bicycle ridership ***(1)***, traffic analysis***, (2***), and accident reporting ***(3)***.

One of the domains with very limited agency data is incidences of roadway flooding. Recurrent flooding as a result of rainfall, high tides, or both is becoming more prevalent in coastal cities. These flood incidents cause inundation of roadways for up to several hours, which deteriorates mobility and accessibility of travelers. Over the past six decades, almost thirty coastal cities in the US have witnessed a spike in the number of annual flood days, with some cities witnessing as many as 50 extra flood days every year ***(4)***. NOAA estimated a 125% increase in the number of annual flood days in the southeast coast of the US and 75% in the northeast coast between the years 2010 and 2015 ***(5)***. Some city or state agencies may collect these flood incident data as a part of providing emergency management services on the roadway network. Preemptive knowledge of reliable flood locations could significantly reduce the duration of flooding and delays on the roadway network. However, most cities do not collect standardized comprehensive data of roadway flooding incidents, making it difficult to make data-driven decisions for traffic rerouting and flood mitigation measures. One promising source for crowdsourced flood incident data is Google Waze, a GPS navigation app that allows users to report on a multitude of traffic-related incidents, including roadway flooding. While crowdsourced data can be a powerful alternative to revolutionize data collection where agency data is lacking, such data comes with its own set of limitations. Since crowdsourced data is not regulated, there can be human error, technical error, wrongful reporting, among other issues ***(6).*** This study explores the trustworthiness of crowdsourced Waze flood incident data in a case study of Norfolk, Virginia. The City of Norfolk has been collecting limited roadway flooding data due to the increasing frequency of recurrent flooding. This study combines the limited city flood report (ground truth) data with publicly available topographical and environmental data to build a model to assess trustworthiness of the crowdsourced Waze flood incident reports.

BACKGROUND AND LITERATURE REVIEW

In the past decade, crowdsourced datasets have become increasingly popular in transportation applications where traditional data collection methods are cost prohibitive. Waze has emerged as a popular crowdsourced dataset for transportation research, as the app allows users to enter location-specific and time-stamped reports of all sorts of traffic-related incidents: road closures, hazards (including roadway flooding), traffic jams, police presence, crashes, and more. Several studies have examined the usability of Waze incident data in transportation applications, though none have examined flood incident data. For example, Hoseinzadeh et al. ***(7)*** used independently collected Bluetooth speed data as ground truth and assessed the quality of Waze speed data. Their models concluded that the Waze data was more accurate in peak traffic hours, and achieved a prediction accuracy of almost 85%. Amin-Naseri et al. (***8***) conducted an analysis to quantify the potential added coverage of traffic incidents through Waze data, in addition to police reported data. It was estimated that 34.1% of Waze-reported incidents provide additional information not covered by any other dataset. Flynn et al. ***(3)*** developed machine learning models with spatial and temporal data which estimated police crash reports with high accuracy, and concluded that Waze could be a potential data source to quickly identify crashes in real time, enabling faster police response.

Goodall and Lee ***(9***) compared Waze-reported crashes and disabled vehicle information with video footage on a roadway segment. They found that 80% of crashes and 50% of the disabled vehicles were captured by Waze data, implying the potential to leverage Waze incident data to enable faster response times, shorter incident durations, and better incident information dissipation to the public. Eriksson ***(10)*** proposed a methodology to integrate crowdsourced Waze incident and congestion data with official traffic data to reduce redundancy, improve reliability, and measure severity of incidents. As evidenced by these studies, crowdsourced Waze data has potential to greatly expand and improve existing agency provided transportation data. However, thus far, no study has examined the value or trustworthiness of Waze flood incident report data.

Using real-time crowdsourced data to analyze impacts of roadway flooding is particularly appealing since the alternative (installing sensor technology to gain accurate spatially disaggregate data on rainfall, tide, and flood levels for every roadway link in a city) is cost-prohibitive. However, as crowdsourced data is not regulated, there could be erroneous reporting due to misunderstanding, confusion, carelessness, incompetence, or even intent to deceive ***(11)***. Emerging studies in the computer science and electrical engineering domains have considered asserting trustworthiness to various crowdsourced datasets dealing with environmental conditions ***(12, 13)***. For example, Flanagin and Metzger ***(14)*** suggest collaborating geospatial and environmental knowledge with crowdsourced data to assert credibility of the data. Similarly, a few studies in the geography domain incorporate topological and environmental characteristics to assert credibility of crowdsourced incident detection datasets. For example, Ostermann and Spinsanti ***(15)*** used crowdsourced Twitter and Flickr data to conduct a context analysis to identify hotspots of forest fires in Spain, using forest cover, distance to nearest hotspot, and inhabitant density in the area as contextual variables. The study concluded that geographic features of crowdsourced location information (also known as Volunteered Geographic Information, or VGI) provides a useful approach to filter crowdsourced data. Hung et al. ***(16)*** assessed the credibility of crowdsourced flood incident data by using contextual topological data. A binary logistic regression model with variables such as elevation and distance-to-flood-risk-zones showed prediction accuracy of 90% and 80% for training and testing datasets, respectively.

The studies conducted by Hung et al ***(16)*** and Ostermann and Spinsanti ***(15)*** combined VGI datasets with other relevant event-specific topographical variables to assert credibility on crowdsourced datasets. Most of the VGI datasets used in previous studies are very location specific, and do not have a presence globally. This study uses a similar approach, by using potential flood-related explanatory variables (environmental and topographic) to assert trustworthiness on crowdsourced Waze flood report data. Waze data has a much higher global footprint, and with that, this methodology can be applied to any city with enough Waze users. Also, unlike the previous studies, the ground truth data set is built from credible but spatially and temporally limited agency data.

DATA SOURCES AND PRE-PROCESSING

To build the trustworthiness model, this study uses contextual parameters such as environmental, topographic, and roadway infrastructure variables to explain the occurrence of a flood incident. This section explains the different datasets used to build the model. The study period is restricted by the availability of flood incident data, which ranges from August 2017 to December 2018, with three weeks excluded due to loss of data.

Environmental data

The environmental data is composed of rainfall and tide level observations. Hourly tide levels referenced to the North American Vertical Datum (NAVD88) were obtained from National Oceanic and Atmospheric Administration's (NOAA) Sewell's Point station ***(17)***. Hourly rainfall data was collected from seven Hampton Roads Sanitation District (HRSD) observation sites. Both of these datasets are publicly available.

Topographic data

Three topographic features were used as model inputs: elevation, topographic wetness index (TWI) ***(18)***, and depth-to-water (DTW) ***(19)***. Elevation information at locations of the Waze flood incident reports is extracted from the United States Geological Survey (USGS) Digital Elevation Model (DEM), which has 1-meter horizontal resolution. The most recently published figure of absolute vertical accuracy of the 3D Elevation Program (3DEP) DEMs within the conterminous United States, in terms of the National Standard for Spatial Data Accuracy (NSSDA) at 95% confidence level, is 3.04 meters ***(27)***.

TWI and DTW were also derived using the DEM. TWI accounts for the tendency of any pixel (smallest grid in a raster file) in the topography to receive water from upstream and its tendency to drain that water. A high TWI value implies a high potential for accumulation of surface water runoff. TWI, defined by Beven and Kirkby ***(18)***, is a function of α, which is the upstream contributing area per unit contour length at a given pixel and $tan\,\beta$, which is the local slope at that pixel in the catchment, as shown in **Equation 1**:

$$TWI = ln\left(\frac{\alpha}{tan\,\beta}\right) \quad (1)$$

DTW, defined by Murphy et al. ***(19)***, is a relative measure of soil moisture conditions, which approximates the elevation difference between a pixel in the topography and the nearest surface water pixel along the least slope path. DTW is a function of $\frac{dz_i}{dx_i}$, which is the slope of a pixel i in the topography along the least-cost path to the nearest surface water pixel, a, which is either 1 or $2^{0.5}$ depending on whether the path crosses the pixel parallel to the pixel boundary or diagonally, and x_c, which is pixel size, as shown in **Equation 2**:

$$DTW\ (m) = \left[\sum \frac{dz_i}{dx_i} a_i\right] x_c \quad (2)$$

Topography pixels closer to surface water, in terms of both distance and elevation, tend to have smaller values of DTW, indicating wetter soil.

Predicted Surface Water Depth

Predicted street-level surface water depth was simulated using a physics-based hydrodynamic model (TUFLOW: Two-dimensional Unsteady Flow) model. The flood model solves 2D equations for shallow water and free surface flow to simulate overland flow, and it is coupled with 1D hydrodynamic network software ESTRY (***20***) to simulate pipe flow. The 1D pipe/2D overland hydrodynamic flood model described by Shen et al. ***(21, 22, 26)***, which provides details on the model construction, calibration, and evaluation process. The model used in this analysis covers roughly half the area of the city of Norfolk, VA (56.4 km^2). Surface flooding is simulated at one-hour time steps and at a spatial resolution of 2.5 m, which was then used to estimate water depth

on street segments. TUFLOW is a high-fidelity model which simulates realistic flood depth, however is very computationally intensive in nature.

Roadway characteristics data

Roadway characteristics consist of geometric design features like number of lanes, per lane capacity, intersection (binary variable, based on if the report falls at an intersection or not), and freeway (binary variable, based on if the report falls on a freeway link or lower functional classification link). These roadway properties are obtained from the Hampton Roads Regional Travel Demand Model (HRRTDM). The roadway functional classification categorizes major and minor freeways as freeways (binary variable = 1), and all other roadway links as non-freeways (binary variable = 0). For capacity of each roadway link, per lane capacity is multiplied with number of lanes, both obtained from the HRRTDM. For roads that are not covered in the HRRTDM network, the number of lanes is obtained from the City of Norfolk's streets shapefile, and multiplied with a default per-lane capacity of 650 vehicles per hour per lane (minimum per lane capacity recorded in the HRRTDM).

Agency-provided flood incident data

The flood incident data collected by City of Norfolk spanned from January 2017 to December 2018, using city employees' reported flood locations in a mobile phone application (System to Track, Organize, Record, and Map [STORM]). The app records the date and location of flooding. Furthermore, the app user can specify the flood location as an intersection, address, or block. Duplicate reports (reports occurring on the same day, within 50ft of each other) are eliminated. Then, several steps are followed in order to geo-tag the flood incident report to a specific location on the roadway network. In ArcGIS, the intersections named in the dataset are manually matched to the corresponding intersection, the addresses are relocated to the closest point on the roadway network, and the block locations are relocated to the lowest elevation point between the upstream and downstream intersections. These location-corrected reports are henceforth referred to as city reports. Due to the lack of a timestamp associated with the flood reports (only dates are recorded), the entire day is initially assumed to be flooded in this analysis. Then, these city reports are checked against TUFLOW model output to identify the flooded time periods with positive water depth, as explained in detail in the Methodology section.

Crowdsourced flood incident data

The mobile navigation application Waze contains a real-time information reporting tool, from which the crowdsourced flood incident reports are obtained. Waze provides user-reported incident data via its data sharing program (Waze for Cities), which is available to public entities worldwide. In this study, Waze incident reports (time-stamped and location-identified) related to roadway flooding in Norfolk (between August 2017 and December 2018, with three weeks excluded due to loss of data) are analyzed. Waze flood report data is only as comprehensive as the locations of road users reporting flooding; however, the spatial coverage is considerably greater than the agency data available through the City of Norfolk. Figure 1 shows an example of flood reports from both data sets in August 2018, to represent the coverage disparity between the City and Waze reporting of floods.

(a) (b)

Figure 1 Flooding reported by (a) City of Norfolk and (b) Waze in August 2018

Drainage characteristics data

A record of all the storm water structures (such as bridge drain, gutter basin, floor drain, manhole, etc.) is provided by the City of Norfolk in a GIS shapefile. This variable is regrouped from 18 structure types to 12 structure categories to combine similar drainage structure types on the roadway. Each flood report is characterized by the drainage infrastructure that is the closest feature by distance (in GIS) from the report location.

In addition to all the datasets described above, an additional parameter called "Proximity" is calculated. Proximity is a measure of closeness to other flood incident reports: the closer the other flood incident reports during the same time of day, the higher the proximity value. This score is assigned to each ground truth city flood report, and is calculated based on Waze- and city-reported flood incidents during the same time period (on the same date), as shown in **Equation 3**. Because Proximity can only be calculated for city reports when Waze flood report data is also available, the study period is defined as August 2017 to December 2018 (when both data sets overlap). Table 1 summarizes all the datasets used to build the ground truth model.

TABLE 1 Data inputs in predictive model

Variable (unit)	Explanation	Data Dimension	Data Source or Method of Derivation
Date	Date of flood report	Temporal	City of Norfolk or Waze.
Time period	Time period of flood report (1: 12:00 to 6:00 am, 2: 6:00 to 9:00 am, 3: 9:00 am to 3:00 pm, 4: 3:00 to 6:00 pm, 5: 6:00 pm to 12:00 am)	Temporal	Timestamp obtained from Waze, and aggregated to corresponding time period. No timestamp for City reports, thus they are initially assumed to apply for all 5 time periods on the report day.
Latitude and longitude	Location of flood report.	Spatial	Obtained from geo-located City of Norfolk data or directly from Waze.
Proximity score	Measure of closeness to other flood reports on the same date and during the same time period. Calculated as the sum of the squares of inverse distances between the current flood report and all the other flood reports.	Spatial and temporal	$Proximity_i = \sum_{j=1, j\neq i}^{J} \frac{1}{{d_{ij}}^2}$ (3) Where: i: flood report for which proximity score is being assigned j: other flood reports (city and Waze) on the same date and during the same time period as *i* d_{ij}: bird's eye distance between coordinates of *i* and *j*, in miles
Rainfall intensity (in/hr)	Collected across seven rain gauges in the city, and interpolated for each flood report location.	Spatial and temporal	Data obtained from HRSD; Interpolation done by Inverse Distance Weighting (IDW), a spatial analysis tool in ArcGIS.
Elevation (ft)	Elevation of each flood report location.	Spatial	Obtained from DEM of Norfolk (from USGS), and extracted for each flood report point.

Tide level (ft)	Maximum tide level recorded during the given time period at gauge at Sewell's Point.	Temporal	Obtained from NOAA Tides and Currents
Depth to water index (DTW)(m)	Elevation difference between the flood report location and closest water body based on least slope path	Spatial	Created from rasters of DEM and waterbodies ***(19)***, and extracted for each flood point
Topographic Wetness Index (TWI)	Measure of the tendency of an area to accumulate runoff: high TWI values imply a high potential for runoff accumulation	Spatial	Created from rasters of DEM and waterbodies ***(18)***, and extracted for each flood point
Intersection	Binary variable to identify if the flood report occurs at an intersection	Spatial	Manually obtained from ArcGIS
Total Capacity	Total capacity of the roadway segment	Spatial	Obtained from HRRTDM $Total\ Capacity = number\ of\ lanes \times per\ lane\ capacity$
Freeway	Binary variable to identify if the flood report occurs on a freeway segment	Spatial	Functional classification obtained from HRRTDM
Drainage	Different types of drainage structures found closest to the report location	Spatial	Obtained from the City of Norfolk

DATA PREPARATION & METHODOLOGY

Refining ground truth dataset

Due to the lack of a timestamp on city-reported flood events (and the initial assumption that the location is flooded for all five time periods of the day), the data undergoes another level of pre-processing before being included as a positive flood report in the ground truth dataset. The physics-based TUFLOW model is simulated on all the reported flood days to provide estimated water depth for each hour within the model boundary. The city report locations that are present within the TUFLOW model boundary are checked for maximum water depth within a 24ft buffer (width of the traveled way on a typical two-lane roadway) to ensure that time periods considered flooded have a predicted water depth greater than 0.1m ***(21)***. The TUFLOW model is not directly used to find the trustworthiness of Waze flood incident due to its computationally intensive nature, making such an approach infeasible for the end goal of using crowdsourced flood reports for real-time traffic management and flood mitigation. In the 16-month study period, 19 days incurred city flood reporting, which translated into 95 initial positive flood observations across five time-of-day-periods. When verified against TUFLOW water depth models, 70 ground truth positive flood observations remained.

The ground truth dataset also requires negative (non-flood) observations. Since any location on the roadway network during any time period that is not reported as flooded could be a potential non-flood observation, a procedure (outlined in Figure 2) was followed for random selection of negative observations at spatial and temporal scales.

Figure 2 Combining spatial and temporal dimensions of negative ground truth observations

For the random sampling of negative observations, locations within a 2000ft buffer of the positive ground truth city flood reports were removed. Variable thresholds are also considered in the random selection process of negative ground truth data to ensure sufficient variation in random sampling. The spatial parameters (TWI, DTW, elevation) were divided into quartiles, based on their range of values, shown in Table 2. 10% of the data is randomly selected from each quartile to have some representation of each level of spatial characteristics. For temporal selection, two levels of tide (≤ 0 ft, 0 ft<) and three levels of rainfall (0, 0< and ≤0.1, 0.1< in/hr) were chosen to have a balanced sample of tide and rainfall levels. The non-flood observations were then sampled to fill a 1:1 of true-to-false flood report ratio in the ground truth dataset for the model. The total number of negative ground truth observations per variable threshold is shown in Table 2 for the balanced 1:1 dataset.

TABLE 2 Negative ground truth observations by environmental and topographic variable thresholds

DTW	Range (index)	DTW <1089	1089≤ DTW <2171	2171≤ DTW < 3260	DTW≥3260
	# of Observations (1:1)	44	21	4	1
TWI	Range (index)	TWI < 3.8	3.8 ≤ TWI <8.9	8.9 ≤ TWI <14	TWI ≥ 14
	# of Observations (1:1)	16	42	10	2
Elev-ation	Range (m)	Elev <-1.5m	-1.5m≤ Elev <3.3m	3.3m≤ Elev <8.2m	Elev≥8.2m
	# of Observations (1:1)	0	45	25	0
Rainfall	Range (in/hour)	Rain = 0	0< Rain ≤0.1 in/hr	Rain > 0.1 in/hr	
	# of Observations (1:1)	60	3	7	
Tide Level	Range (m)	Tide ≤ 0 m	Tide > 0 m		
	# of Observations (1:1)	30	40		

Trustworthiness modeling

Roadway flooding events are closely related to environmental, topographic, and infrastructure conditions (as listed in Table 1). Given the same probability of roadway flooding, the probability of the reporting of a roadway flood event is closely tied to the traffic volume on that roadway segment during the time period of the flood event (approximated here by the time of day and total capacity variables, as explained in Table 1). Then, all of these variables can be used to predict the trustworthiness of a crowdsourced flood report, estimated as the probability of a flood incident report at a location given the environmental, topographic, roadway, drainage, and time-of-day characteristics, via a binary logistic regression model. Logistic regression is widely used to study the effects of explanatory variables on binary outcomes, and the probability of the event occurring is calculated as shown in **Equation 4**:

$$P\,(1|X_1, X_2, \ldots . X_n) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \quad (4)$$

where,

P Probability of occurrence of the event (flood report)

i Observation
α Constant
X_i Independent variables (as listed in Table 1)
β_i Corresponding coefficient

A random 70-30 split of the 140 ground truth observations are used, where 70% of the entire dataset is reserved for training, and remainder 30% for testing the dataset.. Random sampling of data subsets is performed on the training dataset to fit the samples into a prediction model, while reducing the total error in the model. Then, the regression model is used to calculate the probability of a Waze flood event explained by the independent variables.

Note that a classification tree model (which creates partitions in the dataset based on discrete characteristics) was also tested for this dataset. In the classification tree model, a parent node in the tree is divided based on an independent variable into two child nodes, such that each child node is more homogenous (or less impure) than the parent node. However, due to the large number of independent variables in the dataset, the classification tree model proved to be unstable (prediction accuracy and important variables varied widely when changing the observations in the training set). Hence, the logistic regression model is preferred for assessing trustworthiness.

Model Selection

Several different criteria are used to evaluate the fit of the various logistic regression models. These criteria and their definitions are listed in Table 3, including a confusion matrix and its associated criteria (true positive rate [TPR] (**Equation 5**), true negative rate [TNR] (**Equation 6**), false positive rate [FPR] (**Equation 7**), and false negative rate [FNR] (**Equation 8**)), Akaike Information Criterion (AIC), distance to corner (**Equation 9**), receiver-operating-characteristic (ROC) curves, and accuracy (**Equation 10**).

TABLE 3 Performance measures used for model selection

Performance Measure	Definition	Preferred value directionality	Equation (if applicable)
Akaike Information Criterion	Estimator of out-of-sample prediction error, or the relative amount of information lost by a given model	Smaller	
True Positive Rate (Sensitivity)	Ratio of true positives identified to all positive ground truth reports	Higher	$\frac{tp}{tp+fn}$ (5)
True Negative Rate (Specificity)	Ratio of true negatives identified to all ground truth negatives reports	Higher	$\frac{tn}{fp+tn}$ (6)
False Positive Rate	Ratio of false positives identified by the model to all truly negative reports in the ground truth	Lower	$\frac{fp}{fp+tn}$ (7)
False Negative Rate	Ratio of false negatives identified by the model to all truly positive	Lower	$\frac{fn}{tp+fn}$ (8)

	reports in the ground truth		
Receiver Operating Characteristic (ROC) Curves ***(23)***	Plot of Sensitivity vs 1-Specificity, helps to determine the diagnostic ability of the binary classifiers	Closer to the top left corner	
Distance to Corner ***(24, 25)***	Optimal threshold to minimize false positive and false negative rates	Lower	$Distance\ to\ Corner = \sqrt{(1-Sensitivity)^2 + (1-Specificity)^2}$ (9)
Accuracy	Shows the ratio of correctly identified reports to all reports.	Higher	$\frac{tp+tn}{all\ reports}$ (10)

RESULTS

Trustworthiness Model

A prediction model using the binary logistic regression structure was developed to analyze how well a ground truth report can be explained by the independent environmental, topographic, infrastructure, and temporal variables. The ground truth dataset was randomly split into 70% for model training, and the remainder 30% for model testing. First, all the continuous variables considered (as listed in Table 1) were tested for correlation. The highest correlation was found between elevation and TWI (-0.52), with all other correlation values less than 0.3. The correlation between elevation and TWI suggests that areas with lower elevation have a higher tendency to accumulate runoff (TWI), which is related to topography. Then, different binary logistic regression models with and without TWI were tested to examine the effect of inclusion of TWI on the tide level parameter estimate and overall model fit. Due to the relatively small ground truth report sample size, explanatory variables with p-values less than 0.3 (confidence level 70%) are retained. Among the different model specifications tested, elevation, TWI, and DTW are always found to be statistically significant at 95% confidence level. When roadway characteristics are considered, total capacity was the only variable to emerge as statistically significant ($p<0.1$ in all three different model specifications). However, inclusion of the total capacity variable reduced the accuracy of the model (from 90.5% to 88%) with a higher false positive rate in the model prediction. Thus, roadway capacity is excluded from the final preferred model. Tide level and rainfall are marginally significant ($0.25 <$ p-value < 0.3) in the preferred model specification. These variables are retained due to their effect on the discrete time period variables (which emerge as non-significant if tide level and rainfall are excluded). Despite the proximity variable being a byproduct of the number and closeness of other reports in the same time period, it shows a low correlation with tide level and rainfall. Proximity variable is retained in the preferred model due to its high statistical significance. The final preferred model is shown in Table 4.

TABLE 4 Preferred binary logistic regression model results

Variables	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	13.54	5.26	2.58	0.01(*)
Elevation	-4.99	1.60	-3.12	0.00(**)
Tide level	3.35	3.05	1.10	0.27
Rainfall	6.88	6.54	1.05	0.29
Period = 2 (6a – 9a)	-5.01	2.61	-1.92	0.05 (.)

Period = 3 (9a – 3p)	-5.45	3.18	-1.71	0.09 (.)
Period = 4 (3p – 6p)	-3.47	2.58	-1.34	0.18 (.)
Period = 5 (6p-12a)	-6.24	2.95	-2.11	0.03(*)
Proximity	1.30	0.59	2.22	0.03(*)
TWI	-0.52	0.24	-2.13	0.03(*)
DTW	0.00	0.00	2.31	0.02(*)

**: significant at 99% confidence level
* : significant at 95% confidence level
. : significant at 90% confidence level

In the final preferred model, all time periods are statistically significant at the 90% confidence level. In the absence of roadway characteristics in the final model, time period can be considered as a proxy exposure variable (as traffic volumes, and thereby active Waze users, are highly correlated to time-of-day). Proximity is also highly statistically significant in the preferred model. This indicates that the presence of other flood reports during the same time period (and physical proximity to those peer reports) is an important predictor variable. The Root Mean Square Error (RMSE) values for testing and training subsets of the ground truth datasets were found to be 0.27 and 0.29 respectively, with AIC value of 46.15.

The preferred binary logistic model is then used to find an optimum threshold to separate trustworthy and untrustworthy reports. To do this, TPR and TNR are calculated at different thresholds, as shown in Table 5. In Figure 3, additional performance measures are considered for these thresholds as *1 - sensitivity* and *1 – specificity* are plotted to display Receiver Operating Characteristic (ROC) curves ***(23)***. ROC curves help to determine the diagnostic ability of binary classifiers. An optimal threshold is defined as a point on the ROC curve with the value of *1 - sensitivity* and *1 - specificity* closest to 0 (i.e. when the FPR and the FNR are the lowest). Thus, the point on the curve with the shortest distance to the top-left corner of the plot, also known as distance-to-corner, corresponding to a threshold value of 0.8, is chosen as the threshold to differentiate trustworthy from untrustworthy crowdsourced reports. The corresponding confusion matrix for a trustworthiness threshold of 0.8 is shown in Table 6, which yields a model accuracy of 90.48% on the testing dataset.

TABLE 5 Performance measures of varying thresholds

Threshold Value	**0.60**	**0.70**	**0.80**	**0.85**	**0.90**	**0.95**
TPR (Sensitivity)	0.80	0.80	0.80	0.80	0.80	0.75
TNR (Specificity)	0.91	0.95	1.00	1.00	1.00	1.00
Distance to corner squared ($fpr^2 + fnr^2$)	0.0206	0.0055	0.0000	0.0000	0.0000	0.0000
Accuracy	85.71%	88.10%	90.48%	90.48%	90.48%	88.10%

Figure 3 ROC Curve

TABLE 6 Confusion matrix (threshold = 0.8)

N = 42	Predicted True	Predicted False
True Report	16 (tp)	4 (fn)

False Report	0 (fp)	22 (tn)

Trustworthiness of Waze data

Figure 4 shows the predicted probability of occurrence of all crowdsourced Waze flood reports between August 2017 and December 2018 in Norfolk, when the preferred trustworthiness model is applied. 502 of the 697 reports exceed the threshold value of 0.8, implying 71.7% of the Waze flood reports can be considered trustworthy based on their topographic, environmental, temporal, and peer reporting characteristics.

Figure 4 Waze report incident occurrence probabilities

Figure 5 Characteristics of trustworthy and untrustworthy reports

Density plots in Figure 5 show the distribution of independent variable values of trustworthy (cyan) and untrustworthy (red) reports. In terms of environmental conditions, untrustworthy flood incident reports tend to cluster around 0 rainfall. Furthermore, flood incident reports are more likely to be considered trustworthy in higher intensity rainfall periods. For tide level, untrustworthy reports generally follow the same density plot as trustworthy reports, with two high density spikes at 0 ft and 0.4 ft. Tide level is not a spatially disaggregate variable as it is only collected at one point in the city. However, plotting these report locations on a map, there is a cluster of trustworthy reports (Figure 6[a]) near the Chesapeake Bay coast line in the northern part of the city, and another cluster of reports near downtown Norfolk (a low elevation area) in the southwest part of the city. Untrustworthy reports (Figure 6[b]), on the other hand, show far fewer reports in these locations. Similar to rainfall, untrustworthy reports show a low density of occurrence at higher tide levels (greater than 0.5 ft). In terms of the topographical variables, trustworthy reports generally occur at lower elevation compared to untrustworthy reports. When considering TWI (tendency of a location to accumulate runoff), the distributions are very similar for trustworthy and untrustworthy reports. On the other hand, trustworthy reports are more likely to report higher DTW (proximity to the closest water body) values, with the reports at the highest DTW values related to intense rainfall events. This implies that despite being at locations with high DTW, flooding does occur in instances of heavy rain. On the other hand, when examining untrustworthy reports with high DTW values, they occur during time periods with no rainfall. The proximity score of reports, which is based on the quantity and proximity of peer flood incident reports, has a large range. Trustworthy reports' mean proximity value is 6077.32 (log value 2.08) and untrustworthy reports' mean proximity value is 2.15 (log value ~ 0), implying that the majority of untrustworthy reports are either sole reports during the time-of-day period, or had peer reports very far away, resulting in a very low proximity score.

(a) Untrustworthy (b) Trustworthy

Figure 6 Spatial distribution of untrustworthy and trustworthy reports

Table 7 examines the characteristics of peer flood reports for trustworthy and trustworthy flood incident reports. Untrustworthy reports are more temporally dispersed (195 reports over 117 unique time-of-day periods) than trustworthy reports (502 reports over 100 unique time-of-day periods). Additionally, the maximum number of peer flood reports in the same time-of-day period was found to be 6 for untrustworthy reports and 66 in trustworthy reports, implying the importance of peer reports in asserting trustworthiness in crowdsourced data.

TABLE 7 Observations on trustworthy and untrustworthy time periods from Waze

# of reports	Trustworthy Reports	Untrustworthy Reports
Total reports	502	195
Total affected time periods	100	117
Sole flood report (in the time period)	8.0%	39.4%
1 peer flood report (in same time period)	8.8%	20.5%
>1 peer flood report (in same time period)	83.2%	40.1%
Maximum number of peer reports in a single time period	66	6

DISCUSSION

The distinguishable characteristics in trustworthy reports were found to be high peer reporting, lower elevations, high rainfall intensity, and proximity to the coast. These characteristics would be intuitive for true flood events, and this study provides similar evidence. Additionally, the study also shows that peer reporting of true events is much higher, implying a higher confidence in quantity of reporting in Waze as well. This methodology can prove crucial in separating trustworthy and untrustworthy reports, thereby allowing local agencies to obtain valuable incident information across the entire city without deploying extensive manpower for incident reporting. Accessing trustworthy information in near-real time can significantly improve response times for local agencies to delegate emergency management services, thereby solving flooding related disruptions at a faster rate. In addition to this, the general framework adopted in this study can be used for different disruptions with contextual factors as well, enabling a wider range of applications.

CONCLUSION

Crowdsourced data has the potential to provide real-time transportation information without the cost of additional sensors, cameras, or other cost-prohibitive measures. However, since crowdsourced data is usually unchecked, verification of the data becomes challenging. This study presents a framework to assess the quality of a subset of Waze incident reporting data related to roadway flooding in Norfolk, Virginia. Roadway flooding occurs as a combination of environmental conditions and insufficient drainage infrastructure. Environmental, topographic, and infrastructure variables which potentially contribute to flooding and flood reporting are used in this study to build a logistic regression model to estimate the probability of occurrence of a flood incident report. While this methodology does not directly identify misreports or false reports, it provides a conservative approach to distinguish crowdsourced flood incident reports with a high level of trustworthiness. The preferred model developed in this study shows a prediction accuracy of 90.48% when applied to a subset of ground truth data, implying a high rate of correct identification of reports. When applying the model to crowdsourced Waze data over a 16-month study period, 71.7% of the user reported flood incidents were predicted to be trustworthy. Among the untrustworthy reports, the most notable characteristics included low occurrence of peer reporting, inland locations with lower tide levels, higher elevation, and lower rainfall intensity in the reported periods.

This study has limitations which should be addressed in future research. To start, the positive ground truth data set utilized in model development has a small sample size, and is biased towards higher intensity flood events. Within these events, the ground truth data is biased towards

tidal flooding compared to rainfall-induced flooding, due to the nature of city employees' data collection. In addition, the proximity variable defined in this study to account for peer reporting requires that crowdsourced datasets have high levels of user activity (and incident reporting), in order to achieve a wide range of proximity values. One of the assumptions used in the study was to use a default minimum capacity on smaller roads. The capacity variable was found insignificant in the regression model, and thus removed from the finalized model. However, since capacity of a roadway is an important characteristic of the roadway network, it would be prudent to test this variable when applied in different locations. Lastly, ground truth data (however limited) is still necessary to build a trustworthiness assessment model under this framework. In this study, the city flood reports (which are then validated by the physics-based TUFLOW model to confirm non-negligible flood depth) is used to build the trustworthiness model. TUFLOW model is computationally intensive, and not available for an entire city. Zahura et al. developed a random forest surrogate model ***(28)***, which replicates TUFLOW outputs in a fraction of the time. A search for agency provided flood incident data used in recent research only yielded a handful of cities in the US (including New York City, NY; Norfolk, VA; Charleston, SC; Miami, FL; Houston, TX; San Francisco, CA; and Tacoma, WA). This becomes a challenge in transferring the model framework to other coastal cities without agency flood incident data. Given the increasing research on crowdsourced flood incident data, there might be a potential for flood prone cities to invest in data collection during the periods in the year with heavy floods that can act as a ground truth for further Waze dataset usage. Alternatively, cities could also invest in strategically placing sensors throughout the city for a short period of time to collect ground truth data. For cities that lack means of collecting ground truth data, significant static variables from the current work such as topographic and roadway variables can be used for initial screening flood hotspot locations. Peer reporting on crowdsourced datasets can then be used for assigning priority to potentially more vulnerable locations.

Nonetheless, this study demonstrates the ability to assess trustworthiness of Waze flood incident reports with limited ground truth availability. This framework could eventually lead to identification of flooding hotspots in near-real time, allowing cities to deploy dynamic flood mitigation actions and ensure a faster recovery to normal conditions. The flooding hotspots identified through this methodology can be used to provide early improvements in addressing the long-term impacts of sea-level rise. Furthermore, the general methodology utilized in this study is not limited to assertion of trustworthiness of crowdsourced flood incidents, and can be used in other applications with available contextual and ground truth data sets.

ACKNOWLEDGEMENTS

The authors would also like to thank City of Norfolk and Waze for facilitating data acquisition. The authors would also like to thank Erin Robartes for giving valuable input in reviewing the manuscript. This work is supported by the National Science Foundation's Critical Resilient Interdependent Infrastructure Systems and Processes program (Award 1735587).

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: S. Praharaj, T.D. Chen, J. L. Goodall; data collection and processing: Y. Shen, S. Praharaj, F. Zahura; analysis and interpretation of results: S. Praharaj, T.D. Chen, L. Zeng; draft manuscript preparation: S. Praharaj, F. Zahura, T.D. Chen. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Medury, A., O. Grembek, A. Loukaitou-Sideris, and K. Shafizadeh. Investigating the underreporting of pedestrian and bicycle crashes in and around university campuses a crowdsourcing approach. *Accident Analysis & Prevention*, Volume: 130, 2019, pp. 99 – 107.

2. Nair, D.J., F. Gilles, S. Chand, N. Saxena, and V. Dixit. Characterizing multicity urban traffic conditions using crowdsourced data. PLoS ONE, 2019. Volume 14(3). https://doi.org/10.1371/journal.pone.0212845.

3. Flynn, D. F., M. M. Gilmore, and E. A. Sudderth. *Estimating Traffic Crash Counts Using Crowdsourced Data: Pilot analysis of 2017 Waze data and Police Accident Reports in Maryland*, 2018, Tech Report.

4. US EPA. *Climate Change Indicators in the United States: Coastal Flooding*, United States Environmental Protection Agency, 2016. www.epa.gov/climate-indicators. Accessed May 10, 2018.

5. Sweet, W., J. Park, J. Marra, C. Zervas, and S. Gill. *Sea Level Rise and Nuisance Flood Frequency Changes Around the United States.* Publication NOAA Technical Report NOS CO-OPS 073. National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 2016.

6. Sanchez, L., E. Rosas, and N. Hidalgo. Crowdsourcing Under Attack: Detecting Malicious Behaviors in Waze. In *Trust Management XII* (N. Gal-Oz and P. R. Lewis, eds.), Springer International Publishing, Cham, 2018, pp. 91–106.

7. Hoseinzadeh, N., Y. Liu, L. D. Han, C. Brakewood, and A. Mohammadnazar. Quality of location-based crowdsourced speed data on surface streets: A case study of Waze and Bluetooth speed data in Sevierville, TN. *Computers, Environment and Urban Systems*, 2020. Volume 83.

8. Amin-Naseri, M., P. Chakraborty, A. Sharma, S. B. Gilbert, and M. Hong. Evaluating the Reliability, Coverage, and Added Value of Crowdsourced Traffic Incident Reports from Waze. *Transportation Research Record*, 2018. Volume: 2672 (43), pp. 34–43.

9. Goodall, N., and E. Lee. Comparison of Waze crash and disabled vehicle records with video ground truth. *Transportation Research Interdisciplinary Perspectives*, 2019. Volume: 1.

10. Eriksson, I. Towards Integrating Crowdsourced and Official Traffic Data: A study on the integration of data from Waze in traffic management in Stockholm, Sweden, 2019.

11. Ouyang, R. W., M. Srivastava, A. Toniolo, and T. J. Norman. Truth Discovery in Crowdsourced Detection of Spatial Events. *IEEE Transactions on Knowledge and Data Engineering*, 2016. Volume: 28 (4), pp. 1047–1060.

12. Dong, X. L., L. Berti-Equille, and D. Srivastava. Truth Discovery and Copying Detection in a Dynamic World. *Proc. VLDB Endow*, 2009. Volume: 2 (1), pp. 562–573.

13. Prandi, C., S. Ferretti, S. Mirri, and P. Salomoni. Trustworthiness in crowd- sensed and sourced

georeferenced data. *IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 2015. pp. 402-407, doi: 10.1109/PERCOMW.2015.7134071.

14. Flanagin, A. J. and M. J. Metzger. The credibility of volunteered geographic information. *GeoJournal*, 2008. Volume: 72 (3), pp. 137–148.

15. Ostermann, F. and L. Spinsanti. Context analysis of volunteered geographic information from social media networks to support disaster management: a case study on forest fires. *International journal of information systems for crisis response and management*, 2012. Volume: 4 (4), pp. 16–37.

16. Hung, K.-C., M. Kalantari, and A. Rajabifard. Methods for assessing the credibility of volunteered geographic information in flood response: A case study in Brisbane, Australia. *Applied Geography*, 2016. Volume 68, pp. 37 – 47.

17. NOAA, 2018b. Sewells Point - Station Home Page - NOAA Tides & Currents [WWW Document]. URL https://tidesandcurrents.noaa.gov/waterlevels.html?id=8638610 (accessed 10.1.19).

18. Beven, K. J. and M. J. Kirkby, A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 1979. Volume: 24 (1), pp. 43–69.

19. Murphy, P. N. C., J. Ogilvie, K. Connor, and P. A. Arp. Mapping wetlands: A comparison of two different approaches for New Brunswick, Canada. *Wetlands*, 2007. Volume: 27 (4), pp. 846–854.

20. Syme, W. J. TUFLOW - Two one-dimensional Unsteady FLOW Software for Rivers, Estuaries and Coastal Waters, 2001.

21. Shen, Y., M. M. Morsy, C. Huxley, N. Tahvildari, and J. L. Goodall. Flood risk assessment and increased resilience for coastal urban watersheds under the combined impact of storm tide and heavy rainfall. *Journal of Hydrology*, 2019. Volume: 579.

22. Shen, Y. Flood Risk Assessment and Increased Flood Resilience for Civil Infrastructure in Coastal Regions Under Changing Climate. University of Virginia, 2019.

23. Zou, K. H., A. J. O'Malley, and L. Mauri. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation*, 2007. Volume 115 (5), pp. 654–657.

24. Berrar, D. Performance Measures for Binary Classification. *Academic Press*, Oxford, 2019, pp. 546 – 560.

25. Krzanowski, W. J., and D. J. Hand. *ROC Curves for Continuous Data.* CRC Press, 2009.

26. Shen, Y. *Flood Risk Assessment and Increased Flood Resilience for Civil Infrastructure in Coastal Regions Under Changing Climate.* Doctoral Dissertation. University of Virginia, Charlottesville, VA, 2020.

27. United State Geological Survey (USGS). What is the vertical accuracy of the 3D Elevation Program (3DEP) DEMs? *United States Geological Survey (USGS)*. https://www.usgs.gov/faqs/what-vertical-accuracy-3d-elevation-program-3dep-dems?qt-news_science_products=0#qt-news_science_products. Accessed May 27, 2021.

28. Zahura, F. T., Goodall, J. L., Sadler, J. M., Shen, Y., Morsy, M. M., & Behl, M. (2020). Training Machine Learning Surrogate Models from a High-Fidelity Physics-Based Model: Application for Real-Time Street-Scale Flood Prediction in an Urban Coastal Community. *Water Resources Research*, 2020. Volume 56(10).