Modeling Severe Traffic Accidents with Spatial and Temporal Features

Soumya Sourav[1], Devashish Khulbe[2] and Vishal Verma[1]

[1] University of Texas at Dallas, Dallas, TX 75080, USA
[2] New York University, New York, NY 10003, USA
sxs180011@utdallas.edu
dk3596@nyu.edu
vxv180006@utdallas.edu

Abstract. We present an approach to estimate the severity of traffic related accidents in *aggregated* (area-level) and *disaggregated* (point level) data. Exploring spatial features, we measure 'complexity' of road networks using several area level variables. Also using temporal and other situational features from open data for New York City, we use Gradient Boosting models for inference and measuring feature importance along with Gaussian Processes to model spatial dependencies in the data. The results show significant importance of 'complexity' in aggregated model as well as as other features in prediction which may be helpful in framing policies and targeting interventions for preventing severe traffic related accidents and injuries.

Keywords: Accident involvement · Road Networks characteristics · Spatial modeling.

1 Introduction

Traffic related accidents contribute to the deaths of around 1.35 million people and injuries to around 30 million people worldwide. 93% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately 60% of the world's vehicles [1]. The above fact is informative regarding the possible link between street conditions and design and fatal accidents. Additionally, road traffic crashes cost most countries 3% of their gross domestic product, indicating that curbing traffic accidents is financially important. With around 60% of the world population predicted to live in cities by 2030, making urban areas safe for pedestrians and vehicles simultaneously is an important area to delve into. Further, assessing the role of multiple location and time factors in prediction can help the concerned authorities in deploying targeted interventions for public safety. With real-time large open data of accidents and information about urban network of streets, we argue that severity of accidents can be estimated using modern Machine Learning (ML) techniques. Recently, ML has proved to be an important tool for predicting traffic accidents and crash severity, with a variety of tools being used for accident risk prediction [2]. In this paper, we present approaches to infer injury-related and fatal accidents for area level and observation level data for New York City. We also introduce a new feature measuring the complexity of area-level street networks. We model the spatial autocorrelation in the data in the area level model which the regular model may not be able to learn. Interpretable techniques like Gradient Boosting are used to measure feature importance of our variables, the results of which show impact of several spatial and temporal features in inference.

This work contributes to the current research by introducing some new significant predictors and presenting a way to account for spatial dependencies in accidents data.

1.1 Related Work

Significant amount of literature can be attributed to modeling traffic accident and their severity using diverse set of variables. Features like curvature, road width, urban/rural area and gender of driver area have been shown to be significant in accident modeling [3]. Another work [4] describe models for predicting the expected number of accidents at urban junctions and road links as accurately as possible, explaining 60% of variation for road links. This is indicative of importance of location based features of streets and junctions in accident modeling. An approach [5] models severe accidents for area level predictions using linear model using features like intersection density, vehicle speed and number of households which turned out to be significant. This work also uses Geographic Weighted Regression (GWR) to account for spatial correlation. Deep learning approaches have also been proposed for modeling traffic accidents [9] in the past but we aim to build interpretable models in order to measure importance of features with this paper. Our work aims to further introduce and use new predictors to model and subsequently use non-linear models to predict severe accidents both for area-level and point-level data and also account for possible spatial dependencies for area-level data.

2 Data & Methods

Traffic Accidents We use open data for New York City maintained by New York City Police Department (NYPD). The data contains entries of motor vehicle related accidents, containing their coordinates, date and time of incident, type of vehicles involved and number of injuries and deaths. For the aggregated model, we aggregate the number of accidents on census tract level for the city. The model is thus a regression problem with the number of severe incidents as target. For the disaggregated model, the problem is essentially of binary classification type where we aim to classify a traffic accident as severe or non-severe. We define the severity as any incident where number of injuries or deaths are equal to greater than 1. We assign binary values as the target:

$$\begin{cases} 0, & injuries/deaths = 0 \\ 1, & otherwise \end{cases}$$

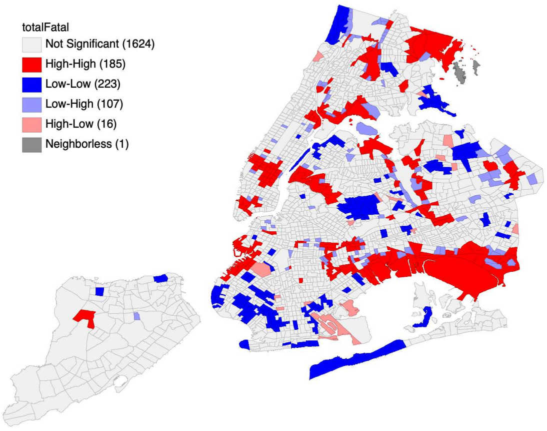
Considering the data from July 2011 till May 2019, this results in a total of around 1,000,000 non-severe accidents and around 200,000 severe accidents, indicating a problem of imbalance which we address further in this section.

Spatial autocorrelation Assuming that our data is not *i.i.d*, we measure local spatial autocorrelation for the number of severe accidents $\mathrm{y_j}$ for census-tract aggregated level through the *Local Moran's I* statistic calculated as:

$$I_j = (n-1)\frac{y_j - \overline{y}}{\sum_{k=1,k\neq j}^{n} w_{j,k}(y_k - \overline{y})^2} \sum_{k=1,k\neq j}^{n} w_{j,k}(y_k - \overline{y}) \qquad (1)$$

where n is the number of spatial units indexed by i and j; $\overline{y}$ represents the mean of $\mathrm{y_j}$ and w represents the spatial weight between the features j and k. The weights for neighbouring and non-neighbouring areas of each census tract j are taken as: $\mathrm{w_{j,\,k}} = 1$ if k is a queen contiguous neighbour of j and $\mathrm{w_{j,\,k}} = 0$ otherwise.

Fig. 1: Local spatial autocorrelation of total severe accidents in New York City ($p \leq 0.05$)



Complexity Introducing a new feature 'complexity' as a proxy for intricacies of street level networks, we define it as a multiple of number of intersections and circuity of a given network. We then measure complexity for our all networks of census tract areas. The circuity is defined as a ratio of network distance to Euclidean distance for a given street network. Considering that number of intersections has been show to be correlated to accidents in previous work [5], multiplying it with circuity account for further intricacies of the network. Thus, our new feature measures complexity of the networks by accounting for factors like number of turns, intersections and nodes.

Other Variables Other features we consider are average street width, vehicle types, average number of bike lanes, day of week and time of day of incident, for which only the hour value is taken. Subsequently, the census-tract level aggregated data set results in 2156 observations and the point level (disaggregated) data set contains around 1,200,000 observations.

Table 1: Aggregated (census-tract level) data set

Features (observations)	**Mean value**	**Std. Deviation**
Complexity (2156)	30.58	39.34
Avg. Street width in meters (2156)	34.13	5.85
Avg. Bike lanes (2156)	1.47	1.35
Avg. node degree (2156)	3.59	0.83

Data imbalance Imbalanced data set is the one which suffers from the problem of classes not being in proportion. This causes a machine learning model to generate fake accuracy reports with the imbalanced data set. With our model we have tried to evade this problem by the use of SMOTE, since our data is fairly imbalanced with positive class accounting for just about 23% of the total observations.

SMOTE (Synthetic Minority Over-Sampling Technique) - It is an oversampling method which can create synthetic samples from minor classes instead of just copying them. The algorithm selects two or more similar instances (using a distance measure) and changing an observation one feature at a time by a random amount within the difference to the neighboring data points. We select SMOTE for over-sampling as it is a relatively simpler method and give good results for data with small number of features [12].

3 Results

3.1 Spatial & Temporal Features

We observe that hour of day of the accident may be important as a predictor, with majority of incidents happening in the evening and night hours. Also, we consider day of week as a predictor, where we observe that weekdays have a slightly greater proportion of accidents as compared to weekends. Also, we decide to model the spatial autocorrelation observed in the data for the *aggregated model* since on observation, there seem to be spatial dependencies in the accidents with some areas having high proportion of incidents.

3.2 Aggregated model

For training and testing purposes we implement gradient boosting method with decision trees having depth of 5 as weak learners . The gradient boosting method

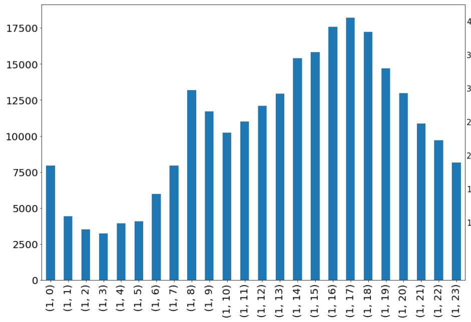


Fig. 2: Hour of day (x-axis) and number of severe accidents (y-axis)

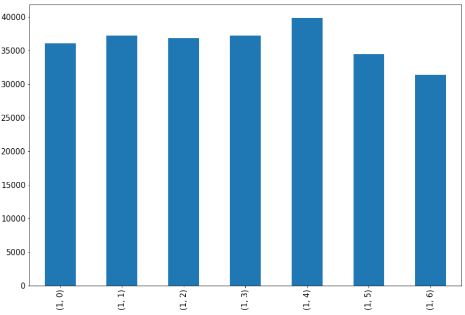


Fig. 3: Day of week (x-axis) and number of severe accidents (y-axis)

works on the principle of optimizing the cost function over a function in the space iteratively the function being a weak learner. The gradient of the weak learner is usually in the negative direction. During each iteration, the trees are added to the model to reduce the loss and this is done by parameterizing them. The gradient follows the direction which reduces the residual loss.

For the census-tract level model, we estimate the number of severe accidents as a function of spatial features s described in table 1 as $y = f(s) + \epsilon$, where ϵ is the error term. Further, we model the error ϵ using spatial context $\epsilon = g(x)$, where x are the centroids of the spatial area we are considering. We use Gradient Boosting (GB) model using 20 fold cross validation for the first part and Gaussian Processes with Radial Basis Function (RBF) kernel for modeling the residual error term. Results show that the first step explains around 34% variation (measured by the R^2 value) in the data and further 13% is explained by modeling the residual error term. Looking at the feature importance, complexity turns out to be the most important in prediction followed by average number of nodes, average street width and average number of bike lanes in the census tract.

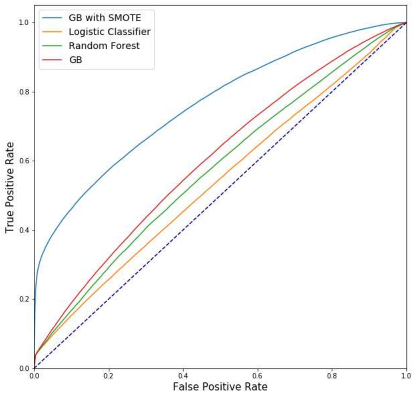
Table 2: Aggregated model

Step	**Model used**	R^2
$y = f(s) + \epsilon$	Gradient Boosting	0.338
$\epsilon = g(x)$	Gaussian Process	0.132

3.3 Disaggregated model

For the *disaggregated* data, the goal to classify each data point as severe or not severe based on the features. The data is thus trained as a binary classification

Fig. 4: Receiver Operator Characteristic (ROC) curve for classification models



problem with four classifiers with 10-fold cross validation. Oversampling the positive class (severe) points with SMOTE and then training the resulting data with Gradient Boosting results in highest AUC score of 0.72. Along with the temporal features like time, the spatial features are also used which are attributed to the census-tract in which the incident happened.

Table 3: Disaggregated (point classification) model

Model	**AUC**
Gradient Boosting with SMOTE	0.729
Gradient Boosting	0.604
Random Forests	0.575
Logistic Regression	0.539

The important features for classification include class of vehicle (specifically whether it was a two-wheeler or truck), complexity of the area and the hour of the incident.

4 Discussion

Feature Importance The importance of a node in a decision tree is computed as:

$$ni_{\mathrm{j}} = w_{\mathrm{j}}C_{\mathrm{j}} - w_{\mathrm{left(j)}}C_{\mathrm{left(j)}} - w_{\mathrm{right(j)}}C_{\mathrm{right(j)}} \tag{2}$$

where w_j is weighted number of samples in node *j*, C_j is impurity in this node, and *left(j)* and *right(j)* are its respective children nodes. Then, the feature importances are then averaged over all the trees.

We observe that spatial features which account for complexity like number of intersections, network nodes, circuity along with vehicle type are important features in prediction of severe accidents. Complexity turns out to be the most important predictor in the *aggregated* model while it it fairly good in the point level classification too. It is interesting to note that average number of bike lanes is not one of the most important predictors for injury and fatal accident classification, despite one of the research concluding that bike lanes make a route safer [6]. This may be because we took the average number of bike lanes in a neighborhood while making predictions. Maybe a more fine street level feature information on this can change our results, which may be a scope for future work in this work. We also notice that temporal features like day of week and hour of day when incident happened are important information about accidents in general. Most of the accidents take place over the weekdays and during the evening and late-night hours. Though, these temporal features do not contribute much in our predictive model, they are good predictors of vehicle collisions and accidents. The spatial features inform us from a road and street design perspective in a neighborhood while the temporal variables along with information about vehicle in the accidents can be important from a real-time emergency deployment viewpoint.

5 Conclusion

In this work, we presented an approach using Machine Learning techniques to model non-fatal (injury) and fatal (death) traffic accidents in urban environments using spatial and temporal variables. We found the importance of factors like street width, vehicle type, time of day and the new created feature 'complexity' of a street network in prediction of severe accidents both in the area level and point level data. These results indicate and validate that road design is a critical factor in severity of accidents. Additionally, we determined the importance of type of motor vehicles involved that can have an impact on the accident. This information can be critical in implementation of policies regarding construction and design of neighborhood streets, allowance of type of vehicles in an area or can help in effective operation of emergency services.
Future work in this domain can be extended to incorporate other socio-economic features in prediction and determination of most affected demographics in terms of road traffic incidents.

References

1. Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

2. Amirfarrokh Iranitalab, Aemal Khattak. Comparison of four statistical and machine learning methods for crash severity prediction, Accident Analysis & Prevention. Volume 108, 2017, Pages 27-36, ISSN 0001-4575, https://doi.org/10.1016/j.aap.2017.08.008.
3. Mohamed A. Abdel-Aty, A.Essam Radwan. Modeling traffic accident occurrence and involvement, Accident Analysis & Prevention. Volume 32, Issue 5, 2000, Pages 633-642, ISSN 0001-4575, https://doi.org/10.1016/S0001-4575(99)00094-9.
4. Poul Greibe. Accident prediction models for urban roads, Accident Analysis & Prevention, Volume 35, Issue 2, 2003, Pages 273-285, ISSN 0001-4575, https://doi.org/10.1016/S0001-4575(02)00005-2.
5. Hadayeghi, A., Shalaby, A. S., & Persaud, B. (2003). Macrolevel Accident Prediction Models for Evaluating Safety of Urban Transportation Systems. Transportation Research Record, 1840(1), 8795. https://doi.org/10.3141/1840-10
6. Harris MA, Reynolds CCO, Winters M, et al. Comparing the effects of infrastructure on bicycling injury at intersections and non-intersections using a casecrossover design. Injury Prevention 2013;19:303-310.
7. Persaud, B., Lord, D., & Palmisano, J. (2002). Calibration and Transferability of Accident Prediction Models for Urban Intersections. Transportation Research Record, 1784(1), 5764. https://doi.org/10.3141/1784-08
8. El-Basyouny, K., & Sayed, T. (2006). Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. Transportation Research Record, 1950(1), 916. https://doi.org/10.1177/0361198106195000102
9. Abdelwahab, H. T., & Abdel-Aty, M. A. (2001). Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. Transportation Research Record, 1746(1), 613. https://doi.org/10.3141/1746-02
10. Mohammed A. Quddus. Time series count data models: An empirical application to traffic accidents, Accident Analysis & Prevention. Volume 40, Issue 5, 2008, Pages 1732-1741, ISSN 0001-4575, https://doi.org/10.1016/j.aap.2008.06.011.
11. Z Sawalha and , T Sayed. Traffic accident modeling: some statistical issues Canadian Journal of Civil Engineering, 2006, 33(9): 1115-1124, https://doi.org/10.1139/l06-056
12. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique Journal of Artificial Intelligence Research. https://doi.org/10.1613/jair.953
13. Fei Wu, Xiao-Yuan Jing, Shiguang Shan, Wangmeng Zuo, Jing-Yu Yang. Multiset Feature Learning for Highly Imbalanced Data Classification. AAAI Publications, Thirty-First AAAI Conference on Artificial Intelligence
14. Xiao-Yuan Jing, Fei Wu, Xiwei Dong, Baowen Xu An Improved SDA Based Defect Prediction Framework for Both Within-Project and Cross-Project Class-Imbalance Problems IEEE Transactions on Software Engineering (Volume: 43 , Issue: 4 , April 1 2017)