Number: 1
Authors: Zhuoning Yuan, Xun Zhou and Tianbao Yang

Title: Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study

Venue: In Proceedings of 6th International Workshop on Urban Computing, Halifax, Nova Scotia, Canada, August 2017 (UrbComp-2017).

Year: 2017

Number Citations: 12

The Aim of the paper：
This paper, through a case study, presents explorations on effective techniques to address the accident prediction for better results. Specifically, they formulates the problem as a binary classification problem. For each road segment in each hour, they predicts whether an accident will occur. Big data including all the motor vehicle crashes from 2006 to 2013 in the state of Iowa, detailed road network, and various weather attributes at 1-hour granularity have been collected and map-matched. They evaluates four classification models, i.e., Support Vector Machine (SVM), Decision Tree, Random Forest, and Deep Neural Network (DNN). To tackle the imbalanced class problem, they performs an informative negative sampling approach. To tackle the spatial heterogeneity challenge, they incorporates Spatial Graph features through Eigen-analysis of the road network.

Conclusion：
They evaluates the models by using the following measures: accuracy, precision, recall, F-Score, AUC (Area Under Curve). Results show that employing informative sampling and incorporating the Spatial Graph features could effectively improve the performance of all the models. Random Forest and DNN generally perform better than other models.

How the work informs your project：
This paper offers a guideline for accident prediction problem, especially for feature selection part. Including Spatial Graph features can be novel in dealing this kind of problem. It informs us to use the spectral clustering to visualize the generated features based on the clustering results.

Number: 2
Authors: Ciro Caliendo and Maurizio Guida

Title: A crash-prediction model for multilane roads

Venue: Accident Analysis & Prevention 39(4):657-70 · August 2007

Year: 2006

Number Citations: 301

The Aim of the paper:
This paper fit crash-prediction models for a four-lane median-divided Italian motorway, which were set up on the basis of accident data observed during a 5-year monitoring period extending between 1999 and 2003. The Poisson, Negative Binomial and Negative Multinomial regression models, applied separately to tangents and curves, were used to model the frequency of accident occurrence. Model parameters were estimated by the Maximum Likelihood Method, and the Generalized Likelihood Ratio Test was applied to detect the significant variables to be included in the model equation. Goodness-of-fit was measured by means of both the explained fraction of total variation and the explained fraction of systematic variation.

Conclusion:
With reference to curves, the results show that crash frequency is positively associated with the length, the curvature and the presence of hazardous points such as junctions or tunnels. Also for tangents, crash frequency was found to be positively associated with length and tunnels variables. The influence of rain shows that with a wet pavement significant increases in the number of crashes are expected, more especially on curves compared to tangents.

How the work informs your project:
Due to accidents are related to many parameters, this paper reminds us to use Maximum Likelihood Method and the Generalized Likelihood Ratio Test to detect the significant variables. Besides that, involving curves and tunnels can make a difference in the further prediction.

Number: 3
Authors: Daniel Eisenberg

Title: The mixed effects of precipitation on traffic crashes

Venue: Accident Analysis & Prevention, Volume: 36, Issue Number: 4

Year: 2004

Number Citations: 274

The Aim of the paper：
This paper investigates the relationship between precipitation and traffic crashes in the US during the period 1975–2000. Traffic crashes represent the leading cause of death and injury for young adults in the US, and the ninth leading cause of death for the overall population. Prior studies have found that precipitation raises the risk of traffic crashes significantly.

Conclusion：
A negative binomial regression approach is employed. Two different units of analysis are examined: state–months and state–days. The sample includes all 48 contiguous states.A surprising negative and significant relationship between monthly precipitation and monthly fatal crashes is found. However, in the daily level analysis, a strong positive relationship is estimated, as in prior studies. The source of the contrasting results appears to be a substantial negative lagged effect of precipitation across days within a state–month. In other words, if it rained a lot yesterday, then on average, today there are fewer crashes. Additional analysis shows that the risk imposed by precipitation increases dramatically as the time since last precipitation increases.

How the work informs your project:
This paper gives us an idea that involving lag features into prediction models given this is a time-series prediction model. Except for time lagging features, we can also involve lagged effect of precipitation across days, this will bring up accuracy in our accidents prediction model.

Number: 4
Authors: Li YenChang and Wen ChiehChen

Title: Data mining of tree-based models to analyze freeway accident frequency

Venue: Journal of Safety Research, Volume 36, Issue 4, 2005, Pages 365-375.

Year: 2005

Number Citations: 271

The Aim of the paper:
This study collected the 2001–2002 accident data of National Freeway 1 in Taiwan. A Classification and Regression Tree (CART) model and a negative binomial regression model were developed to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors.

Conclusion:
The CART findings indicated that the average daily traffic volume and precipitation variables were the key determinants for freeway accident frequencies. By comparing the prediction performance between the CART and the negative binomial regression models, this study demonstrates that CART is a good alternative method for analyzing freeway accident frequencies. By comparing the prediction performance between the CART and the negative binomial regression models, this study demonstrates that CART is a good alternative method for analyzing freeway accident frequencies.

How the work informs your project:
This paper informs us a novel model to try predicting the road accidents by using Classification and Regression Tree. Since we are focusing on analyzing accident frequencies just like this paper states, it's a good try to implement this model than clustering methods. We tried k-d tree in the final project.

Number: 5
Authors: Minh X. Hoang, Yu Zheng, and Ambuj K. Singh.
Title: forecasting citywide crowd flows based on big data

Venue: In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems

Year: 2016

Number Citations: 42

The Aim of the paper:
In this paper, they proposes the novel problem of predicting two types of flows of crowds in every region of a city based on big data, including human mobility data, weather conditions, and road network data. To model the multiple complex factors affecting the crowd flows, we decompose the flows into three components: seasonal (periodic patterns), trend (changes in periodic patterns), and residual flows (instantaneous changes). The seasonal and trend models are built as intrinsic Gaussian Markov random fields, which can cope with noisy and missing data. Whereas, the residual model exploits the spatio-temporal dependence among different flows and regions, as well as the effect of weather.

Conclusion:
Their framework is robust to both noise and missing data. Experiments show that their approach is scalable and outperforms baselines (SARIMA, LmNei, STARMA, and HP-BC-MSI) significantly. For the parameters, holidays are clearly better than the one without holidays. The accuracy is further increased when trend is added.

How the work informs your project:
This paper reminds us adding the trend and holiday features, especially how to access Gaussian properties of trend and seasonal models. It's benefit for us to capture a general trend in the distribution of accidents and use appropriate method to predict corresponding frequency or number of accidents.

Number: 6
Authors: Afian Anwar, Till Nagel, Carlo Ratti
Title: Traffic Origins: A Simple Visualization Technique to Support Traffic Incident Analysis

Venue: *2014 IEEE Pacific Visualization Symposium*

Year: 2014

Number Citations: 28

The Aim of the paper
The paper attempts to present a simple visualization technique, called Traffic Origins, that highlights the effects of traffic incidents on congestion. The authors apply their methods to historical incident and traffic data and discuss the advantages and limitations of their technique. It also depicts the variations of traffic incidents over the course of a single day in a convenient and intuitive way.

Conclusion
The paper provided a visualization technique which draws attention to a road incident through the use of an expanding circle that reveals the state of the road network in the immediate vicinity of the incident, thereby enabling users to see the before and after effects of the incident. A proof of concept demonstration at a public exhibition showed that the visualization method was aesthetically pleasing and understandable by both transportation experts and members of the public.

How the work informs your project
One of our objectives for the project was to highlight specific areas on streets as being dangerous based on the number of accidents in its vicinity using techniques like k-means clustering, and the use of circles to denote the state of the road network was something we hoped to implement in our visualizations.

Number: 7
Authors: Alireza Pakgohar, Reza Sigari Tabrizi, Mohadeseh Khalili and Alireza Esmaeili

Title: The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach

Venue: Procedia Computer Science, Volume 3, 2011, Pages 764-769

Year: 2010

Number Citations: 63

The Aim of the paper:
The focus of this research study is to identify the role of human factor in incidence and severity of road crashes. For the purpose of this study mainly descriptive analysis, Logistic Regression, Classification and Regression Tree were employed. The data were used to carry out this research study were based on Database of Traffic Accidents of Iran's Police. Specifically, they focus on the impact of "Safety Belt" and "Driving License" on Severity of Road Crashes.

Conclusion:
The results proved that CART has higher accuracy than LR method. The results also revealed that "Driving License" and "Safety Belt" represent distinct relationship with the degree of injury in road crashes in Iran. The investigation also represented that gender of drivers had no significant relationship with the degree of injury. In addition, the study provided sufficient witnesses that age of the drivers have significant negative relationship with rate of deadly accidents in Iran.

How the work informs your project:
This paper involves the feature about human factor. That's a new feature we can implement on. Maybe adding human factor, especially for driving license and safety belt, our final web application can be made to prove for certain policy regarding to driving safety.

Number: 8
Authors: Ling Wang, Mohamed Abdel-Aty and Qi Shi

Title: Real-time crash prediction for expressway weaving segments

Venue: Transportation Research Part C: Emerging Technologies, Volume 61, December 2015, Pages 1-10

Year: 2015

Number Citations: 37

The Aim of the paper:
This study presents a multilevel Bayesian logistic regression model for crashes at expressway weaving segments using crash, geometric, Microwave Vehicle Detection System (MVDS) and weather data.

Conclusion:
The results show that the mainline speed at the beginning of the weaving segments, the speed difference between the beginning and the end of weaving segment, logarithm of volume have significant impacts on the crash risk of the following 5–10 min for weaving segments. The configuration is also an important factor. Weaving segment, in which there is no need for on- or off-ramp traffic to change lane, is with high crash risk because it has more traffic interactions and higher speed differences between weaving and non weaving traffic. Meanwhile, maximum length, which measures the distance at which weaving turbulence no longer has impact, is found to be positively related to the crash risk at the 95% confidence interval.

How the work informs your project:
For the real-time analysis, this paper informs us a good data source, Signal Four Analytics. It is an interactive, web-based system designed to support the crash mapping and analysis. Besides that, monitoring the speed difference and use t-test to measure the importance of variables can be helpful for future feature selection.

Number: 9
Authors: Do-Gyeong Kima, Yuhwa Lee and Simon Washington

Title: Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models

Venue: Accident Analysis & Prevention, Volume 39, Issue 1, January 2007, Pages 125-134

Year: 2007

Number Citations: 187

The Aim of the paper:
This study aims to identify factors that affect the probability that certain types of crashes will occur by exploiting the hierarchical structure of intersection crashes. The crashes themselves represent the lowest level of the hierarchy, while the intersection at which the crash occurred represents the higher-level hierarchy, or cluster. Since unobserved factors, crash frequencies and types observed at a particular location may be correlated, they overcomes this problem (when cluster level correlation exists) by applying multilevel binomial logistic models for predicting the probability of certain types of crashes.

Conclusion:
The findings of this study suggest that crashes at rural intersections in Georgia are indeed hierarchical in structure. The modeling results indicate that the effects of geometric characteristics and environmental factors can be modeled using multilevel modeling techniques. Crash data may consist of a hierarchical structure: driver's characteristics are nested within crashes, crash characteristics are nested within site characteristics, site characteristics are nested within regional characteristics, and so forth.

How the work informs your project:
This paper gives us an idea that how to solve correlation problem among crashes. This is quite novel to bring up since we have never noticed the hierarchical structure in our dataset. We only use t-test or likelihood to minimize the number of features. This offers us a different solution about using multilevel modeling techniques.

Number: 10
Authors: Helai Huang, Hanchu Zhou, Jie Wang and Fangrong Chang

Title: A multivariate spatial model of crash frequency by transportation modes for urban intersections

Venue: Analytic Methods in Accident Research,Volume 14, June 2017, Pages 10-21

Year: 2017

Number Citations: 38

The Aim of the paper:
This study proposes a multivariate spatial model to simultaneously analyze the occurrence of motor vehicle, bicycle and pedestrian crashes at urban intersections. The proposed model can account for both the correlation among different modes involved in crashes at individual intersections and spatial correlation between adjacent intersections.

Conclusion：
According to the results of the model comparison, multivariate spatial model outperforms the univariate spatial model and the multivariate model in the goodness-of-fit. The results confirm the highly correlated heterogeneous residuals in modeling crash risk among motor vehicles, bicycles and pedestrians. In regard to spatial correlation, the estimates of variance for spatial correlations of all three crash modes in the multivariate and univariate models are statistically significant.

How the work informs your project：
This paper also focus on the correlation problem in crash data. The difference is it studies the variance for spatial correlations of three crash modes. Since our dataset also have these three kinds of modes, it’s worth to notice the correlation between different modes of crashes, which will influence the prediction results.

Number: 11
Authors: Isabelle Thomas

Title: Road Traffic Accident Clustering With Categorical Attributes

Venue: Proceedings of the 83th Annual Meeting of the Transportation Research Board.

Year: 2004

Number Citations: 63

The Aim of the paper：
This paper develops an unsupervised categorical model-based accident clustering. This technique enables us firstly to take the typical categorical data aspect into account. Secondly, instead of employing a heuristic measuring the distance between the accidents – as prevailing cluster techniques would do – it uses a more appropriate density-based similarity to assign the accidents to the different clusters. Finally, using all the available data, the technique aims at an unbiased discovery of the data's inherent sub-structures. The method is applied to the road accident population observed in a Belgian suburban area.

Conclusion：
This accident examination at cluster level not only confirmed some existing findings but also generated new insights (or issues to get to the bottom of): the 'weekend accidents' are actually all-week accidents, the safety influence of passengers is subjected to weather conditions and the passenger formula, black zones consist mainly out of two accident types, confirmation of attribute relationships findings (e.g. age–gender) appears to be cluster dependent.

How the work informs your project：
This paper uses clustering method instead of regression or classification models. It provides us an another tool to predict the trend of accidents. Since our purpose is drawing the heatmap of accidents, using clustering method will be extremely helpful for us to access color palette to show the severity of accidents.

Number: 12
Authors: Madhar Taamneh, Salah Taamneh & Sharaf Alkhed

Title: Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks

Venue: International Journal of Injury Control and Safety Promotion, 24:3, 388-395

Year: 2016

Number Citations: 12

The Aim of the paper:
This paper aims to exploit hierarchical clustering method to build more accurate ANN classifiers for predicting the severity of road traffic crashes in Abu Dhabi, UAE. Using hierarchical clustering, accidents that are similar to each other can be grouped into separate clusters. A separate classification model can be later fitted for each cluster. This paper also uses WEKA (Waikato Environment for Knowledge Analysis) to investigate the impact of clustering road traffic accidents on the accuracy of classification models built to predict the severity of such accidents.

Conclusion:
The results show that the overall model prediction accuracy for the training set and percentage split options when running the MLP algorithm on the whole data set were 84.1% and 64.5%, respectively. The results also indicate that a considerable improvement was achieved for the training set option when accidents are clustered into groups. What's more, the results suggest that as the number of clusters increase the achieved accuracy also increases.

How the work informs your project:
This paper gives us an idea that using artificial neural network and hierarchical clustering can predict the accidents accuracy. It uses confusion matrix for a binary classifier to measure the performance. That's what we can borrow from for the further project given our main method is focusing on k-means clustering.

Number: 13
Authors: Ali Moslah Aljofey and Khalil Alwagih

Title: Analysis of Accident Times for Highway Locations Using K-Means Clustering and Decision Rules Extracted from Decision Trees.

Venue: International Journal of Computer Applications Technology and Research Volume 7- Issue 01, 001-011 , 2018, ISSN:-2319-8656

Year: 2018

Number Citations: 18

The Aim of the paper:
This study proposes a framework to analyze times of accident frequencies for highway locations. The proposal framework consists of clustering technique and classification trees. The k-means algorithm is applied to a set of frequencies of highway locations accidents within 24 hours to find out when and where accidents occur frequently. These frequencies were extracted from 358,448 accident records in Britain between 2013 and 2015.

Conclusion:
As a result of clustering technique, four clusters were ranked in descending order according to the accidents rate for location within the cluster. After that, the decision tree (DT) algorithm is applied to the resulting clusters to extract the decision rules as the cluster name represents the class value for all tuples contained. However, extracting decision rules (DRs) from the DT is restricted by the DT's structure, which can not extract more knowledge from a specific dataset. To overcome this problem, they develops an ensemble method to generate several DTs in order to extract more valid rules.

How the work informs your project:
This paper shows the work about how to combine clustering with decision trees. In this way, we can actually put the cluster results as input and output the decision rule. Since our framework is going to input a location and list the top 10 dangerous location near the input, this is extremely helpful after clustering work.

Number: 14
Authors: Tessa K

Title: Kernel density estimation and K-means clustering to profile road accident hotspots

Venue: Accident Analyssi & Prevention, Volume 41, Issue 3, May 2009, Pages 359-364.

Year: 2009

Number Citations: 395

The Aim of the paper:
This paper presents a methodology using Geographical Information Systems (GIS) and Kernel Density Estimation to study the spatial patterns of injury related road accidents in London, UK and a clustering methodology using environmental data and results from the first section in order to create a classification of road accident hotspots. A kernel density estimation map was created and subsequently disaggregated by cell density to create a basic spatial unit of an accident hotspot.

Conclusion:
The kernel density estimation tool enabled an overarching visualisation and manipulation of the accidents based on density which was used in turn to create the basic spatial unit for the hotspot clustering method. The classification of road accident hotspots in road safety still remains an important and yet under developed theme. These typologies provide a snapshot of the processes which are occurring at these sites and the people upon whom they impact. This information can lead road safety professionals to a better understanding, not only of the types of hotspots but their patterns across London. There are some evident potential policy implications for certain clusters.

How the work informs your project:
This paper combines the visualization tool to show the clustering results. It offers us some ideas for visualizing the clustering results, especially for the design of final dashboard.

Number: 15
Authors: V. Prasannakumara, H. Vijitha, R. Charuthaa and N. Geethaa

Title: Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment

Venue: International Conference: Spatial Thinking and Geographic Information Sciences 2011

Year: 2011

Number Citations: 134

The Aim of the paper:
This paper aims to evaluate and represent the accident hotspots in the Thiruvananthapuram city by modeling real accident location information in conjunction with various spatial attributes using spatial statistics in geographical information technology. The predictive zonation and analysis can help to identify vulnerable locations and zones that require remedial measures. Besides, the model will also help to delineate the safe road segments which in turn can be effectively used as models in the development of safer pathways.

Conclusion:
The Kernel density estimated from the derived hotspots (total accidents, monsoon, non-monsoon, near educational institutions and religious places), suggests spatial variability in the distribution pattern of accident hotspots and coldspots in the Thiruvananthapuram corporation. The advantages of such two dimensional hotspot surface representations, particularly of road accidents, can provide a more realistic continuous model of accident hotspot patterns, over space and time.

How the work informs your project:
This paper introduces Moran’s I method for spatial autocorrelation, which works for both feature locations and feature values. This can be used for feature analysis. What’s more, hotspot analysis is helpful for our future prediction. Implementing ArcGIS's Spatial Statistics tools would be a great idea to access these data.