# Road Accident Hotspot Identification: Literature Review

Saahitya E
*CSE Dept.*
*PES University*
Bangalore, India
saahitya.e@gmail.com

Samarth M
*CSE Dept.*
*PES University*
Bangalore, India
samarth.ms@gmail.com

Satyam Shivam Sundaram
*CSE Dept.*
*PES University*
Bangalore, India
satyspy007@gmail.com

***Abstract*—Identification of Road Accident Hotspots is an important problem, with improvements in the identification of these hotspots directly correlating to better road safety management. This paper explores previous work done in the road accident hotspot identification domain over the past few decades or so. We compare and critique the main approaches that have been used to identify road accident hotspots. We finally propose a different type of kernel function that might improve hotspot identification by being biased towards crash points with extreme attributes such as being a more severe accident or rainy road conditions, higher casualty rate, etc**

***Index Terms*—road accident hotspots, Kernel Density Estimation, Kriging**

## I. Introduction

The National Crime Records Bureau (NCRB) 2016 [1] report states there were 464,674 road accidents which caused 148,707 traffic-related deaths in India. This is more than death toll from all the wars fought by India put together. [2] Even though developed countries like UK show a strong trend of decrease in fatal accidents every year, in developing countries there is weak trend of fatal road accidents increasing.This further emphasized the need for a dedicated road safety program that uses scientific knowledge for effective decreasing road accidents. According to the WHO, road accidents are the leading injury-related cause of death among people aged 10-24. Even though human error and mechanical failure is the most likely cause of accidents, spatial factors like the location and the road condition and weather condition are highly underrated factors in accidents. Even more so, these spatial factors compound human error or mechanical failures like a drunk person driving on a wet road and hence is very important for road or highway agencies to consider while trying to make highways or roads safer.

Cataloguing of accident hotspots is the process of spotting geospatial areas that have a high occurrence of road accidents. Cataloguing accident hotspots can be an effective and low-cost tool to identify high-risk accident-prone areas. Proper identification of these accident hotspots or collision-prone areas can be used by city and highway or roadway authorities to take appropriate measures in these areas to reduce the accident rate after further case by case review and diagnosis. It is an interesting and important problem, because of the different approaches that have been used to solve it and the scope and use in real-world applications.

## II. Background

The approaches to solving cataloguing are broadly split into two techniques - model building and geostatistical techniques. Model building approach involves describing crashes as a function of attributes like weather, road conditions, etc. This model is then used to identify accident hotspots by clustering, regression, etc. On the other hand, the geostatistical based approach involves considering spatial correlation within geostatistical variables like latitude and longitude.

Originally a lot of work was done on model-based approaches. Early models were flawed as they assumed accidents was normally distributed Oppe [5]. Many other authors Hauer and Persaud [6] have used negative binomial regression models.

The geostatistical approaches are looked at in detail in the below subsections.We belive that these approaches are superior as they take into consider spatial autocorrelation that is important for identifying accident hotspots

### *A. Kernel Density Estimation*

Kernel Density Estimation or KDE is a non-parametric statistical tool to estimate the probability density function of a random variable. It is data smoothing problem where inference about population is made by looking at the finite data sample.

This technique differs from the previous approach by considering the effects of unmeasured confounding variables through the concept of spatial autocorrelation between the crashes event over a geographical space. The methodology to perform KDE is described below:

Kernel Density Estimation involves placing a symmetrical kernel function, which is a function of bandwidth on each crash point generating a smooth intensity surface. Here bandwidth is used as a parameter to increase or decrease the size of the intensity surface around a crash point. Typically a smaller bandwidth detects small irregularities, while a bigger bandwidth excessively smoothes the curve [4]. Hence the appropriate bandwidth for the right purpose must be chosen. Then, for a given point of interest, the crash intensity is a summation of the entire overlapping surface due to the crashpoint. There are many kernel functions. Some common kernel functions are normal, uniform, quartic, epanichnikov, and triangular.

$$f(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{d_i}{h})$$

where f(x, y) is th density estimate at location(x, y),
h is bandwidth,
K is kernel function,
$d_i$ is is distance between the points at $i^{th}$ observation and (x,y)

Lalita et al. [3] compares KDE and krigging. They used a geocoded dataset of crash data of accidents in Hennepin county in Minnesota.The authors are weighing points according to severity of crashes but do not mention details about how weighting for observations was performed.The authors also do not discuss how distance between two coordinates are measured since each observation has only latitude and longitude attribute.

Kernel Density Estimation has many advantages over model-based approaches in general and clustering techniques such as KNN. The main advantage is determining the spread of risk of an accident based on an area. This means that given a spatial unit area, it can be said if there is an increased likelihood of accidents around it or near it. This is useful because it allows a way to compare two arbitrary coordinates and compare the accident risk in that area. This also means that a topology or a way to quantify accident risk zones is possible.

Lalita et al. [3] describes one such way where the top 10% quintile of cells based on KDE are considered as hotspots and each quantile is given a category of risk.

### B. *Kriging*

Kriging, also known as Gaussian process regression is a method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariances. Kriging has been utilized widely across many different fields of studies which require spatial prediction. It works on a fundamental assumption of the existence of spatial autocorrelation in geostatistical methods.

The basic idea behind kriging is that the predicted outputs are the weighted average of sample data, and the weights are determined in such a way that they are unique to each predicted point and a function of the separation distance between the observed location and the location to be predicted. In other words, kriging provides estimates at unknown locations based on a set of available observations by characterizing and quantifying spatial variability of the area of interest.

Three variants of Kriging, Simple Kriging (SK) assumes a constant mean over the entire study area, Ordinary Kriging (OK) assumes constant mean over each local neighboring area, Universal Kriging (UK) is a hybrid method.

Ordinary Kriging is a simple yet powerful variant. Kriging is mathematically closely related to regression analysis. Both theories derive a best linear unbiased estimator (BLUE), the difference is that kriging is made for estimation of a single realization of a random field, while regression models are based on multiple observations of a multivariate dataset.

Lalita et al. [3] use Ordinary Kriging (OK) to identify hotspots on a highway, investigate for autocorrelation by modeling semivariograms based on available sampling. They also check whether the time of the day influences hotspot location with morning peak and evening peak hours. It is observed that the hotspots identified using kriging method were spread out and had a higher PAI value when compared to KDE.

Lalita et al. [3] conclude that kriging is the better option compared to KDE based on PAI, although kriging approach is not widely explored for the problem of hotspot identification

### C. *Performance Measure*

Lalita et al. [3]have used the Prediction accuracy index (PAI) as a performance measure to compare KDE and kriging.

$$PAI = \frac{\frac{n}{N} * 100}{\frac{m}{M} * 100}$$

where n is the number of crashes in hotspots
N is the total number of crashes
m is the length of the highway section in hotspots or area covered
M is the total length of highway section or total area covered.

This measure is a naive approach to compare the two methods and its value need not necessarily imply that one is better than the other as this measure tells about how grouped the crashes are but does not say if the predicted hotspot is actually an accident-prone zone or not.

## III. Data Description

The region of interest is the United Kingdom. The study is based on road accident data from 2005 to 2016, from across the country. The dataset is well structured and gives a lot of information about the accident. Some of the interesting information to look at would be the severity of the crash, weather at the time of crash, location, road condition etc. Around 20% of the accidents involve people of the 26-45 year age group, Fig. 1. Most of the accidents occurred, surprisingly, in conditions with fine-no-high-winds( 80.1% ), followed by rainy weather( 11.7% ), as from Fig. 2. The day of the week that the crash occurred is also known from the dataset. The majority of the accident's severity was Slight, followed by Serious and Fatal, as from the pie chart in Fig. 3. Fig. 4 shows the spread of fatal accidents across the UK. We can observe that the region towards London (South) is denser. We can also observe that plotting discrete points on the map is not enough to identify the hotspots, as it just gives us a vague idea, which is why methods like KDE, kriging are important which try give us a continuos smooth density over spatial area.

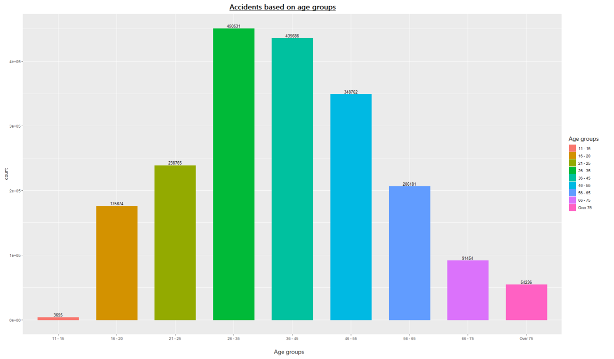

Fig. 1. Number of accidents for differnt driver age groups

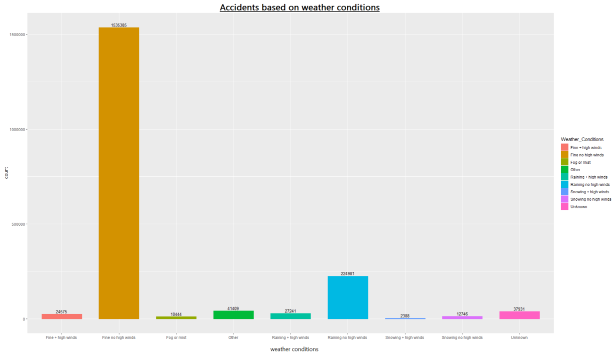

Fig. 2. Number of accidents under different weather conditions

## IV. Identification of Road Accident hotspots using Kernel Density Estimation

Some observations made from previous works are :-

- There has been no proper work on how to weigh extreme crash points in kernel functions for Kernel Density Estimation.
- There is no work that has considered different distance measure(especially since observations here are geo coordinates) in kernel functions for Kernel Density Estimation.

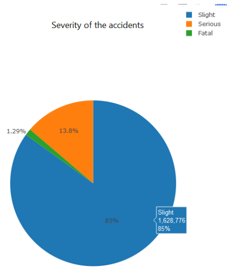

Fig. 3. Accident Severity Pie chart

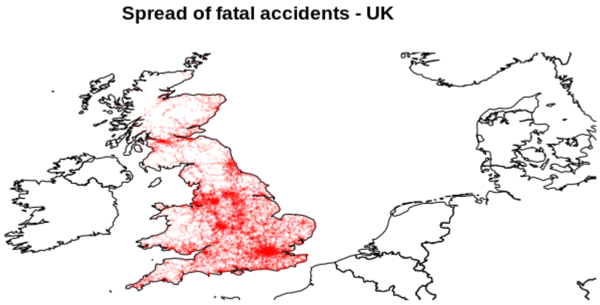

Fig. 4. Spread of Fatal road accidents across the United Kingdom

- Performance metrics have not been properly defined.

Even though Kernel Density Estimation has been widely used and implemented, There hasn't been any effort to weight observations in an appropriate manner before applying the kernel function. We plan to weigh extreme crash points on attributes such as Accident Severity, Weather Conditions, Road Conditions. Also as our application involves geo-coordinates, we plan to try measuring the distances in kernel functions with the appropriate distance measures such as haversine formula or spherical cosine formula. Finally, we plan to compare our modified approach with the traditional approach to see if our approach makes a difference.

## Acknowledgment

We would like to thank our Data Analytics professor Dr Gowri Srinivasan for giving us this opportunity to do this project and for her valuable input. We would also like to acknowledge the sources from which we obtained our data, the Kaggle user "Thanasis" for advancing the dataset and the website data.gov.uk for making this data publicly available.

## References

[1] http://www.indiaenvironmentportal.org.in/files/file/Road accidents in India 2016.pdf

[2] https://afro.who.int/sites/default/files/2017-06/vid_global_status_report_en.pdf

[3] Thakali, L., Kwon, T.J. Fu, L. J. Mod. Transport. (2015) 23: 93. https://doi.org/10.1007/s40534-015-0068-0

[4] Silverman, B., 1986. Density Estimation for Statistics and Data Analysis, 1st ed. Chap- man and Hall, London

[5] Oppe, S., 1979. The use of multiplicative models for analysis of road safety data.Accident Analysis and Prevention 11, 101115.

[6] Hauer, E., Persaud, B.N., 1987. How to estimate the safety of rail-highway grade crossing and the effects of warning devices. Transportation Research Record 1114,131140.