Road Traffic Accident Hotspot Identification: A Literature Survey

Harshita Vidapanakal
CSE Department
PES University
Bengaluru, India
harshita.vidapanakal@gmail.com

Utkarsh Gupta
CSE Department
PES University
Bengaluru, India
utkarsh348@gmail.com

Varun MK
CSE Department
PES University
Bengaluru, India
mkvarun2001@gmail.com

***Abstract*—Identifying traffic crash hotspots is an important step in trying to improve conditions of infrastructure and mitigate risks for better road safety. This paper surveys previous work done in this domain to understand and critique how hotspots are identified and compared and finally propose a modified approach to map hotspots that weighs attributes of choice over a period of time to better identify the appearance and evolution of hotspots.**

***Keywords*—** *GIS, hotspots, spatio-temporal, KDE, kriging*

Introduction

With about 4.5 lakh crashes every year, 1.5 lakh people in India lose their lives in traffic accidents[1]. This is more concerning considering the fact that though India only accounts for 1% of all the world's vehicles, it accounts for 11% of all the traffic deaths. According to a WHO report[2], these crashes cost a country 3% of their GDP and are the leading cause of death in the age group of 5-29. Road traffic analysis is the seventh leading cause of death in the world. It is the highest risk system which people must confront everyday. Although, there is a visible trend of decreasing road accidents in countries like the UK with strong infrastructure, it becomes important to understand what spatial factors, environmental states etc, and their combination can not only cause dangerous incidents but can also aggravate common uncontrollable issues such as people driving under the influence, human errors or system failures in the car.

This is why it is important for us to analyse where and when these accidents occur frequently. The locations, which are identified by a high accident occurrence compared with the other locations, are known as hotspots or black spots [3]. Occurrences of accidents are not random in space and time. They depend on factors such as geometric design, severe weather conditions, time of the day, etc. Plotting, spotting and cataloguing these hotspots helps us locate the most accident prone areas and where the focused efforts, energy and resources of the concerned authorities is required to diagnose the core issues and help counter the said issues. Simply plotting the individual crash sites on a map does not work either as this does show a strong concentration of incidents at certain locations but is vague in the information it provides, giving no density information.

This also becomes more obviously helpful than just plainly surveying the data to understand the causes as different locations have different factors at play and a broader approach cannot prove as helpful. Furthermore, analysis like this proves helpful for construction companies and other corporations responsible for planning and building infrastructure around traffic in specific locations.

Background and literature survey

Many traditional statistical methods have been used to detect hotspots [4,5]. In these models, spatial characteristics of hotspots were modeled as a constant for a given period of time which is not true[6]. This dataset provides a means to analyse the surrounding environment, severe weather conditions which can further be used to calculate severity index or simply take these factors into account during model building which shows a significant improvement in ranking/detection of hotspots.[5][7].

A. Kernel Density Estimation(KDE)

KDE is a non parametric spatial statistical method, which is one of the most commonly used techniques in spatial analysis towards crash and accident hotspot identification, and uses a dataset to estimate probabilities, probability density functions for random variables. It is essentially used to solve a smoothening problem to get inferences from the data. It is preferred also due to the fact that

KDE depends on two parameters, the bandwidth for the kernel function and the cell size. In this KDE function(which is a symmetrical kernel function and is a function of bandwidth) is applied to each crash point, giving a smooth intensity surface and the summation of the overlapping density surface for multiple crashes gives the density for each cell. The bandwidth becomes the smoothening parameter and represents the size of the search radius for the kernel function.

Having a larger grid cell size saves time in processing but this results in a loss of information as the information is averaged over a larger area[8]. Trying to achieve an extremely high granularity on the other hand is not only computationally expensive but unnecessary considering the fact that measures taken to avoid accidents in the future will aim to benefit larger areas at once and keeping buffer for some errors in the study. Bandwidth selection is subjective to the case. A smaller bandwidth portrays more variability

among the different terrains and the characteristics they portray, making the differentiation finer in its form[8].

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

where f(x,y) is the density estimate at the location (x,y)

n is the number of observations
h is the bandwidth
K is the kernel function
d_i is the distance between the location (x,y) and the ith observation
W_i is the intensity of the observation[8].

Another advantage that KDE presents is the estimation of growth rate around crash risk, determined by the spatial analysis. Also, defining a spatial analysis unit for the whole study proves useful in comparison. The common kernel functions used are normal, quartic, triangular, uniform, and epanechnikov.

B. K-means clustering

The k-means clustering is one of the simplest clustering methods which uses simple geometric distance calculations. Advantages of using K-means is the good scalability and its ability to cluster large datasets with moderate to high dimensions at reduced computational expense. However, in our case this method also proves to be very computationally expensive as we're dealing with millions of entries in our dataset which increases the memory usage a lot.

In more recent studies we see the use of computational statistics and machine learning methods such as KDE and Kriging for mapping and categorizing of hotspots[9].

K-means can only be performed on variables having numeric values so that their Euclidean distance from the centroid can be found. This includes variables such as:

- Longitude information
- Latitude information
- Number of casualties
- Number of vehicles
- Age of driver

The aim of K-means clustering is to separate n data points into k clusters. It looks for patterns in the dataset with no prior knowledge of the dataset. Initially, we generate k random centroids in the given space. Every datapoint is grouped with one of these centroids based on the lowest euclidean distance from these centroids to that datapoint. The means of each new cluster are used to recalibrate(recalculate) the centroid and update the process. Thus, the cluster gradually changes with each iteration. The algorithm ends when the centroids reach stability(no further change) or when given iterations have been reached.

Due to the random nature of initialization of K-means it doesn't guarantee finding the optimal solution. Therefore, we run the test several times with different initial input and then average out the centroid coordinates that result.

We use the elbow method to calculate the optimum value of k(number of clusters) for a given number of variables. It plots the number of clusters against mean square error of all data points from their nearest cluster centroids.This process is repeated with increasing number of clusters till the line chart looks like an arm, then the elbow(point of inflection on the curve) is chosen as the new value of k. A strong inflection point indicates that the underlying model fits the best at that point.[10]

C. Kriging

Kriging is one of the many geostatistical techniques which is aimed at identifying hotspots. Like KDE, it considers the effect of unmeasured confounding variables through the concept of spatial autocorrelation between the crash events over a geographical space. [4]. It uses interpolation methods to provide a best linear unbiased estimator(BLUE) for variables that have a tendency to vary over space. Predicted outputs are a weighted average of the same 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.

Let x and x_i be the location vectors for the estimation point and a set of observations of known locations, respectively, with i = 1,2,….n. Based on the n number of available crash frequencies, we are interested in estimating a number of crashes in any given location , denoted by $\widehat{Z}(x)$. The expression of a general kriging model is as follows[11]:

$$\widehat{Z}(x) = m(x) + \sum_{i-1}^{n} \lambda_i [Z(x_i) - m(x_i)]$$

where m(x) and m(x_i) are expected values of the random variables $\widehat{Z}(x)$ and $Z(x_i)$;
λ_i is a kriging weight assigned to datum $Z(x_i)$ for estimation of a crash frequency at any location x.

The random field, $Z(x)$, can be decomposed into two components namely residual component $R(x)$ and a trend component $m(x)$, and expressed as $Z(x) = R(x)$ $m(x)$. Each of three main variants of kriging namely simple kriging (SK), ordinary kriging (OK), and universal kriging (UK) can be distinguished according to the model considered for the trend component, $m(x)$.

D. Performance metric

Choosing between kriging and KDE for hotspot selection can be done using the Prediction Accuracy Index(PAI) developed by Chainey et al.[12]. It is a performance measure used to compare the two proposed methods.

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

where n is the no. of crashes in hotspots
N is the total number of crashes
M is the length of the highway section/total area covered.

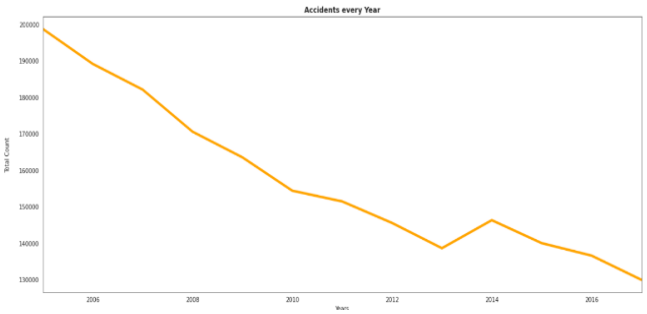

Figure 1: Accidents every year in the UK from 2005-2017

PAI index is used to locate a high number of potential crashes in a small area. (Note: This measure is a naive approach to compare the two methods and a higher value doesn't imply that one model is better than the other, this method doesn't prove if the predicted hotspot is actually accident-prone or not).

E. Comparison between KDE and Kriging

On comparing the 2 geostatistical methods with the region of interest being Hennepin County, Minnesota comparing the expected collision frequency of individual road sections and identifying crash hotspots in a highway network[1]. Kriging, which is not as prominent as KDE in hotspot estimation studies, proved to be a promising alternative. PAI indices show that the average matching rate of 65% indicates that the outcomes of the two tests vary significantly. The maximum overlap occurs when a bandwidth of 400m is considered KDE for MP crash groups.

Thus, Kriging seems to offer better performance than kriging. However, the reliability of and credibility of the PAI index can be questioned. One way to prove this hypothesis could be to use other methods of performance measures to further prove this assertion.

F. Moran's I

Moran's I is a correlation coefficient that measures the overall spatial autocorrelation of our dataset. For a given set of features and an associated attribute it evaluates whether the pattern expressed is clustered, dispersed, or random. It compares the value of the variable at any one location with the value at all other locations. [7]

$$I = \frac{N\sum_i\sum_j Wi.j(X_i-\overline{X})(X_j-\overline{X})}{(\sum_i\sum_j Wi.j)\sum_i(X_i-\overline{X})(X_j-\overline{X})^2}$$

where N is the number of cases
Xi is the variable value at a particular location
Xj is the variable value at another location
Wij is the Distance based weight matrix.

DATA DESCRIPTION

The region of interest is the United Kingdom. The study is based on road accidents that occured in 2005-2016 in the UK. The data shows us a lot of interesting information about the accident that we can choose to visualize such as the number of accidents, the conditions and status of the environment at the time, severity of these accidents etc.

One such observation is the decrease in the number of road accidents in the UK between 2005-2017[Fig.1], with most accidents including drivers in the age range of 26-35(little more than 20%). [Fig.2]

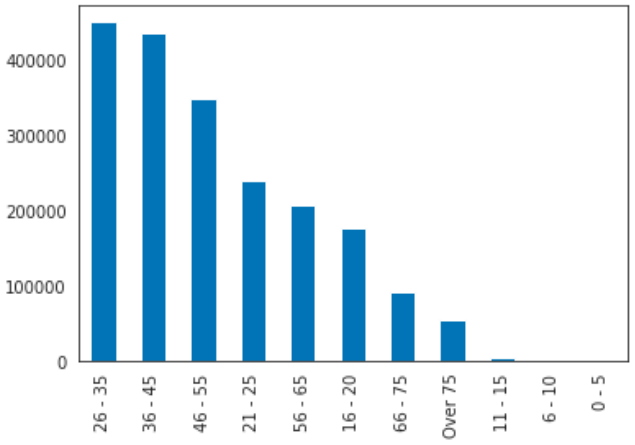

Figure 2: Number of accidents involving each age group

Unexpectedly, the number of accidents at different times of the day, segmented into morning, afternoon and night were more or less the same[Fig.3]. While not so surprising was that the number of fatal accidents formed a smaller portion of the total number of accidents[Fig.4], what was still interesting to observe was that the number of fatal accidents in the

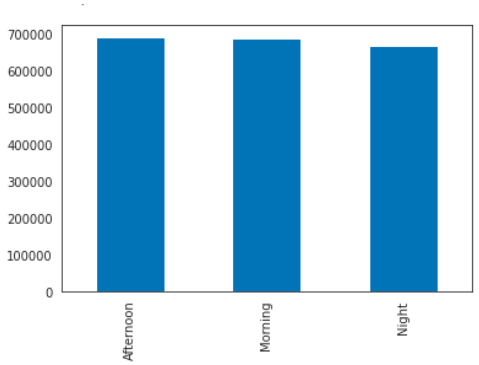

Figure 3: Grouping accidents based on the time of the day.

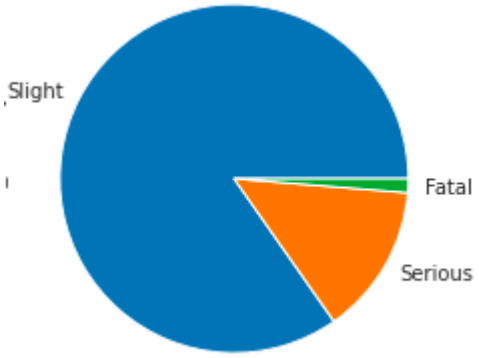

Figure 4: Proportions of severity in accidents

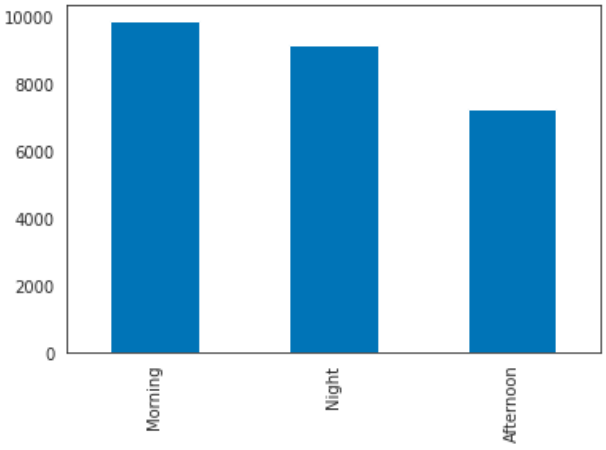

Figure 5: Number of fatal accidents at each time of the day

morning and night were more than the ones that occurred in the afternoon, arguably at peak traffic.
Weather seems to have a weaker effect on the number of accidents, where we would expect harsher conditions to cause more incidents but quite the opposite, with 80.11% of all accidents occurring when the weather was “fine with no high winds”[Fig.6].

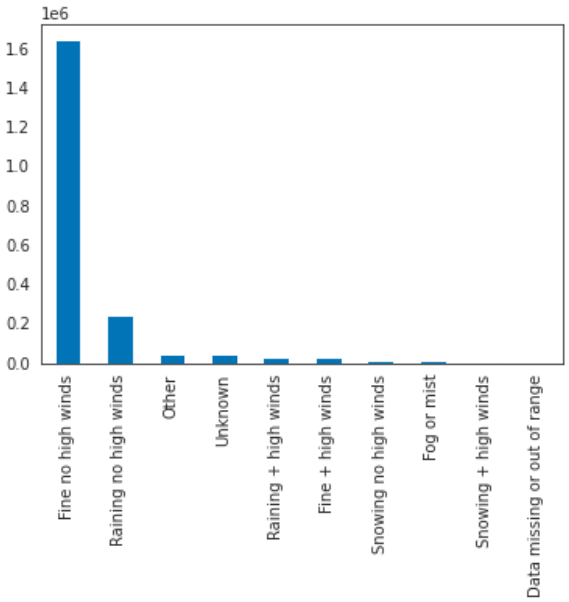

Figure 6: Plot of weather condition vs no. of accidents

Hotspot identification with spatiotemporal data weighed against accident parameters

Some inferences that we made from previous work done:
The temporal evolution of hotspots has been explored in three possible ways to gain insights as to what has changed in those particular locations over time[13].

- Type 1 - There is a dissipation in hotspots at certain locations either due to change in infrastructure or applications of new safety protocols.
- Type 2 - Indicates hotspots that have showed no change in their characteristics over time and continue to identify dangerous locations
- Type 3 - Appearance of new hotspots or increase in density of ones present already, suggesting worsening conditions.

On the other hand, steps have been taken to rank hotspots identified from spatiotemporal data from Road Traffic Accident(RTA) data from Hanoi, Vietnam[14] using a severity index to gain more information from each hotspot identified at particular points in time to better understand the conditions of the environment at the time of the accident and clearly, the severity of these accidents at these times. However, steps weren’t taken to analyse the evolution of the hotspots created with the severity index in mind. Therefore, plan on experimenting with KDE+[13] creating hotspots weighed against severity(or any attribute that may give more knowledge regarding identified hotspots) over the temporal data to understand more about the dangerous locations that have been marked. We plan to compare kriging and KDE if possible to understand their benefits against each other.

Acknowledgment

We would like to thank our professor Dr. Gowri Srinivasan for giving us the opportunity to pursue this project and for her guidance through the execution of the same and express our gratitude to the Data Analytics team for guiding us throughout. We would also like to express our gratitude to the user “Thanasis” and the UK Department of Transport on Kaggle for making the extensive data of road accidents publicly available for study and analysis.

References

[1] India accounts for 11 per cent of global death in road accidents, BusinessLine, https://www.thehindubusinessline.com/news/india-accounts-for-11-per-cent-of-global-death-in-road-accidents-world-bank/article33834556.ece

[2] WHO/T. Pietrasik, Road traffic injuries, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[3] Dereli, Mehmet & Erdoğan, Saffet. (2017). A new model for determining the traffic accident black spots using GIS-aided spatial statistical methods. Transportation Research Part A: Policy and Practice. 103. 106-117. 10.1016/j.tra.2017.05.031.

[4] Isabelle Thomas, Spatial data aggregation: Exploratory analysis of road accidents, Accident Analysis & Prevention, Volume 28, Issue 2,

1996, Pages 251-264, ISSN 0001-4575,
https://doi.org/10.1016/0001-4575(95)00067-4.

[5] Choudhary, Jayvant & Ohri, Anurag & Kumar, Brind. (2015). Spatial and statistical analysis of road accidents hot spots using GIS

[6] Srikanth, Lakshmi & Srikanth, Ishwarya & Arockiasamy, M.. (2019). Identification of Traffic Accident Hotspots using Geographical Information System (GIS). International Journal of Engineering and Advanced Technology. 9. 4429-4438. 10.35940/ijeat.B3848.129219.

[7] V. Prasannakumar, H. Vijith, R. Charutha, N. Geetha, Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment, Procedia - Social and Behavioral Sciences, Volume 21, 2011, Pages 317-325, ISSN 1877-0428,
https://doi.org/10.1016/j.sbspro.2011.07.020.

[8] Thakali, L., Kwon, T.J. & Fu, L. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. J. Mod. Transport. 23, 93–106 (2015). https://doi.org/10.1007/s40534-015-0068-0

[9] Sinclair, C and Das, S (2021) Traffic accidents analytics in UK urban areas using k-means clustering for geospatial mapping. https://doi.org/10.1109/sefet48154.2021.9375817

[10] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," J, vol. 2, no. 2, pp. 226–235, Jun. 2019.

[11] Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

[12] Chainey, Spencer & Tompson, Lisa & Uhlig, Sebastian. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. Security Journal. 21. 4-28. 10.1057/palgrave.sj.8350066.

[13] Michal Bíl, Richard Andrášik, Jiří Sedoník, A detailed spatiotemporal analysis of traffic crash hotspots, Applied Geography, Volume 107, 2019, Pages 82-90, ISSN 0143-6228, https://doi.org/10.1016/j.apgeog.2019.04.008.

[14] Khanh Giang Le, Pei Liu & Liang-Tay Lin (2020) Determining the road traffic accident hotspots using GIS-based temporal-spatial statistical analytic techniques in Hanoi, Vietnam, Geo-spatial Information Science, 23:2, 153-164, DOI: 10.1080/10095020.2019.1683437