Clustering Analysis of Temporal Characteristics of Car Crashes in Polk and Scott County, Iowa Using Self-Organizing Maps

Shuyang Zhang, Arielle Wood, Gabriel Ibiassi Nambila, Luke Anderton, Sapir Carlo Dooley

Abstract - Vehicle crashes happen every minute in the United States, each crash having its own time and specific attributes such as road conditions and weather. The goal of this project is to find meaningful spatial and temporal patterns from the large datasets provided by state DOT's. Using a self-organizing map, data can be clustered into meaningful clusters based on time of year, month, or day in a way that is easy to read. This allows for a more in-depth analysis. Data can then be joined to point data in ArcMap to show how patterns relate spatially within certain attributes like road conditions, number of injuries or weather conditions. The significance of this is that it allows temporal data to be joined to spatial data. This allows for multifaceted in-depth analysis.

I.INTRODUCTION

Traffic injuries are one of the most severe public health problems, causing not only many deaths, injuries, and property damage, but also producing a significant economic loss (U.S. Department of Transportation, Bureau of Transportation Statistics, 2016; World Health Organization, 2015). It is believed that vehicle flow follows certain spatiotemporal patterns. The ability to deduce these patterns could prove very useful in adjusting traffic management and emergency response services. The data currently available to researchers is so extensive that proper examination of this data requires the use of data mining and machine learning in order to make any sort of useful conclusions. We have entered the world of big data, and high level analysis technologies are required to make sense of this data. In the era of "big data", massive volumes of connected data sources (e.g., traffic volume, road geometry, weather condition, land cover/land use, etc.) offer extensive opportunity for road safety researchers to discover new insights into the distribution and causes of crash occurrence based on a number of factors. In this project we will use Self-Organizing Maps (SOMs) to conduct our traffic analysis. We chose this method for two reasons: their ability to preserve topological relationships and their ability to be easily understood. This allows for the analysis to be shared with and understood by experts and non-experts in the field.

A case study is presented for Polk County and Scott County, IA, using crash data collected in 2013. In this paper, we will discuss how the application of SOMs as relates to crash data provided by the Iowa Department of Transportation (IOWA DOT) in both

Polk and Scott Counties, reveals certain temporal patterns. The next step is to overlay these patterns with their corresponding spatial attributes in ArcMap to bring to light any interesting patterns or rules. These can then be applied to the current traffic regulations to check for any gaps in regulation. There may be aspects of traffic that cause more accidents that are not being mitigated through regulation.

The remainder of this paper is organized as follows. In Section II, our main methodology, the SOM algorithm, is described with the basic ideas of our method. This method is applied to a real-world case study in Section III and results are reported and discussed in Section IV. Section V concludes the paper with orientation of future work.

II.METHODOLOGY

Self-Organizing Map (SOM) is one kind of neural network that can perform unsupervised learning. SOM maps a high-dimensional input vector to a typically two-dimensional space while preserving the topological relations of input vectors.

The essential SOM architecture is a two-layered network composed of an input layer and a competitive layer. Nodes in the competitive layer are arranged in two-dimensional rectangular (or hexagonal) lattice and fully connected with the input layer via weights $\omega_i = [\omega_{i1}, \omega_{i2}, \cdots, \omega_{im}]$, which assigns the d-dimensional input vector to the ith node in the competitive layer. The training process of SOM consists of iterative updating of the nodes' weights. Each time an input vector F (in this research, it is the flow volumes of 58 links in one time interval) is presented to SOM. The winner node c is calculated as the node closest to F in sense of Euclidian distance, i.e.

$$||F - \omega_c|| = \frac{min}{i}||F - \omega_i|| \quad (1)$$

Then the weights of all nodes are updated according to the learning rule:

$$\omega_i\,(s+1) = \omega_i(s) + \eta(t)h_{ci}F - \omega_i(s)] \quad (2)$$

where s indicates the iteration step, $\eta(s)$ is the learning rate, and $h_{ci}(s)$ is the neighborhood kernel around winner node c which decreases with the increasing distance of node i to node c. For the purpose of convergence, both the learning rate $\eta(s)$ and the neighborhood kernel's radius σ decrease as the training procedure processes:

$$\eta(s) = \eta(0)\ (1 - \frac{s}{K}) \quad (3)$$
$$\sigma(s+1) = \sigma(s)\ (1 - \frac{s}{K}) \quad (4)$$

where K is the maximum number of iterations. The training steps repeat until the maximum number of iterations is reached. Once the training procedure is completed, each data point F is clustered to the nearest node.

For better performance of SOM, several improvements are adopted in this research. The training process is performed in two phases. In the first phase (called rough learning), relatively large learning rate and neighborhood radius are used to tune the SOM approximately to the input space. In the second phase (called fine tuning), learning rate and neighborhood radius are smaller to avoid oscillation. We also adopt the conscience rule for biases updating in order to utilize all nodes uniformly.

III.CASE STUDY

This section details our case study including the study area, data resources. A case study was performed in Polk County, IA to examine the proposed approach. A comparison study was conducted using Scott County's data to assess the performance and repeatability of the proposed method.

3.1. Study area

Iowa has one of the largest road networks in the nation with numerous car collisions occurred every day. Among Iowa's counties, Polk County and Scott County are facing the most severe traffic safety problem. According to the 2010 census, Polk County has the largest population in Iowa with an area of 1,533 km^2. Scott County ranked third in population with an area of 1,212 km^2. In year of 2013, a total of 8,916 crashes occurred in Polk County and 3,545 in Scott County. And from 2011 to 2016, a total of 45,898 crashes occurred in Polk County and 17,895 in Scott County. These two counties are typical high-populated and crash-intense areas and more suitable for conducting traffic collision analysis. Fig. 4 plots five-year car-collision distribution in Polk County and Scott County.

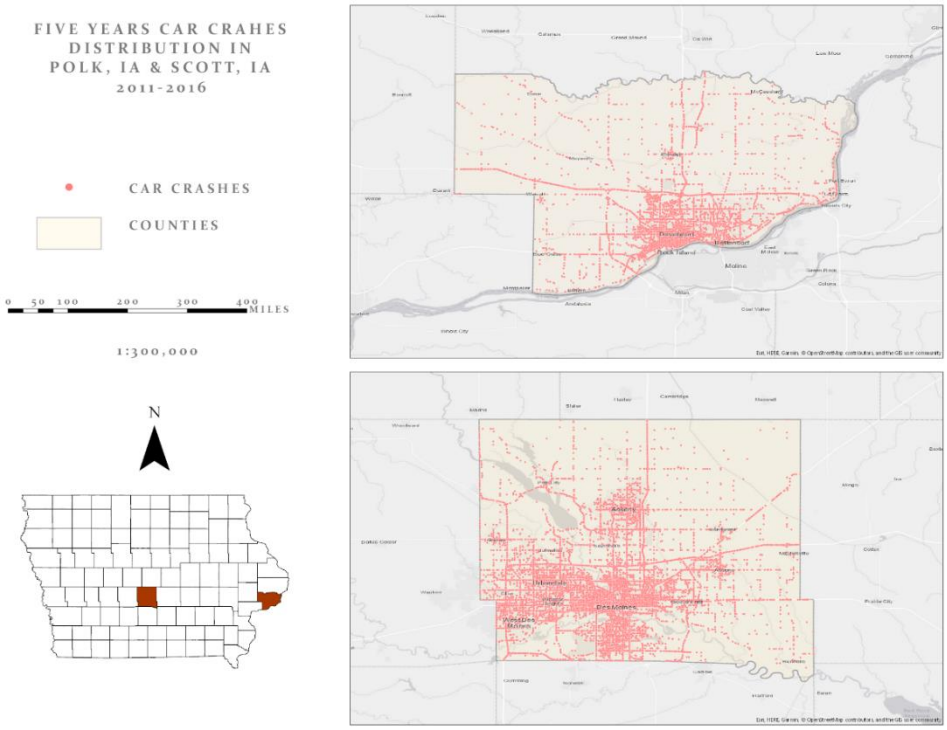


Fig. 1. Car crashes distribution in Polk County and Scott County, IA (2011-2016).

3.2. Data Sources

This study collected input data from four data sources from IOWADOT open data. It provides an open-access car crash dataset containing the statewide historical crashes and crash-related data. In this study, we extracted two counties' (Polk County and Scott County) year 2013 crash data from this dataset for examining the crashes' spatiotemporal concentrations.

IV.EXPERIMENTAL RESULTS

4.1. Temporal Analysis for Car Crashes

In this study, the temporal trend of car crashes was examined at different scales including annually, monthly, daily, and hourly, as shown in Fig. 2 and Fig. 3

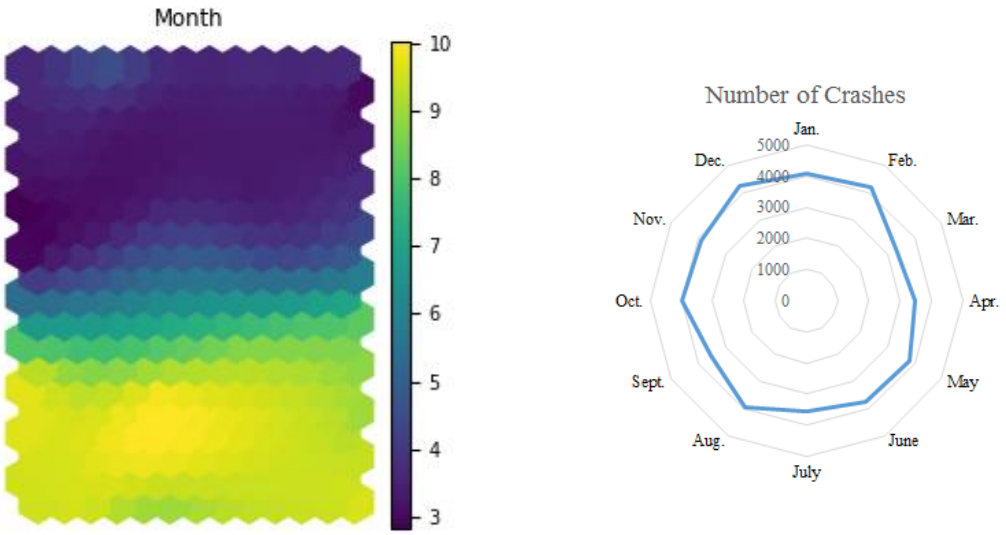
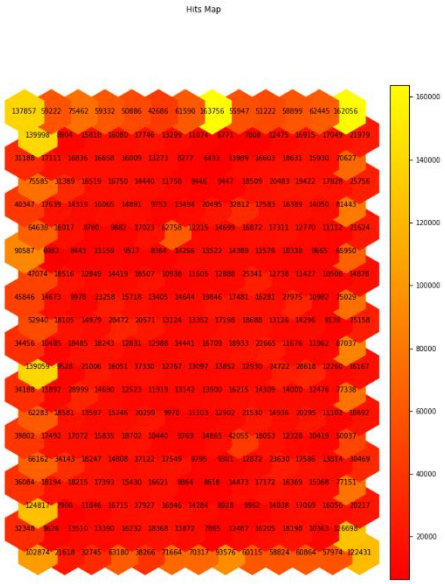


Fig. 2(a)Clustering result of car crashes monthly distribution in Polk County (2013); (b)Total car crashes data monthly distribution in Polk County (2011-2016)



(c)Hits map for clustered data in Polk County

Fig. 2(a) shows the clustering result of car crashes monthly distribution in Polk County. Fig. 2(c) shows the Hits map for clustered data in Polk County. It shows that the lighter the color of each grid, the more clustered data lying in this grid. Combining Fig. 2(a) and Fig. 2(c), we conclude that winter months (i.e., Dec, Jan, and Feb) presented in colors of yellow and dark blue take greater parts than other months. It

means that winter months have higher crash occurrences, which may be caused by poor weather conditions and fewer daylight hours.

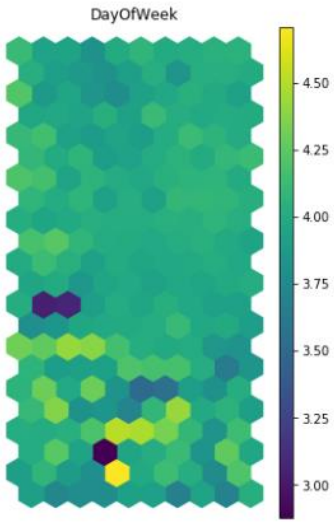
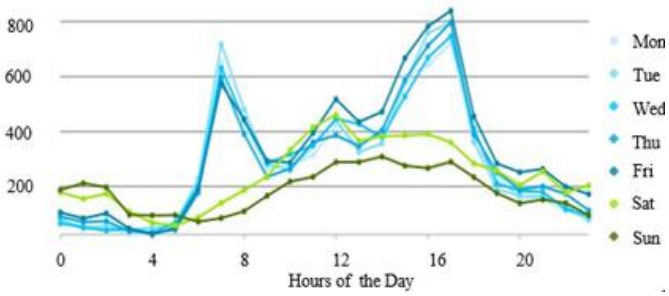


Fig. 3(a)Clustering result of car crashes daily distribution in Polk County (2013);



(b)Total car crashes data daily and hourly distribution in Polk County (2011-2016)

Fig. 3(a) shows the clustering result of car crashes daily distribution in Polk County. Combining Fig. 3(a) and Fig. 2(c), we conclude that weekends presented in colors of yellow take less parts than weekdays. It means that weekdays have more car collisions than weekends. The same car crashes daily distribution pattern is also shown in Fig. 3(b), that weekdays have more car collisions than weekends. And 7 am, 4 pm, and 5 pm are the top 3 high-risk hours for driving due to high traffic volume.

The same temporal pattern exists in Scott County, IA as well.

4.2. Spatial Analysis for Car Crashes

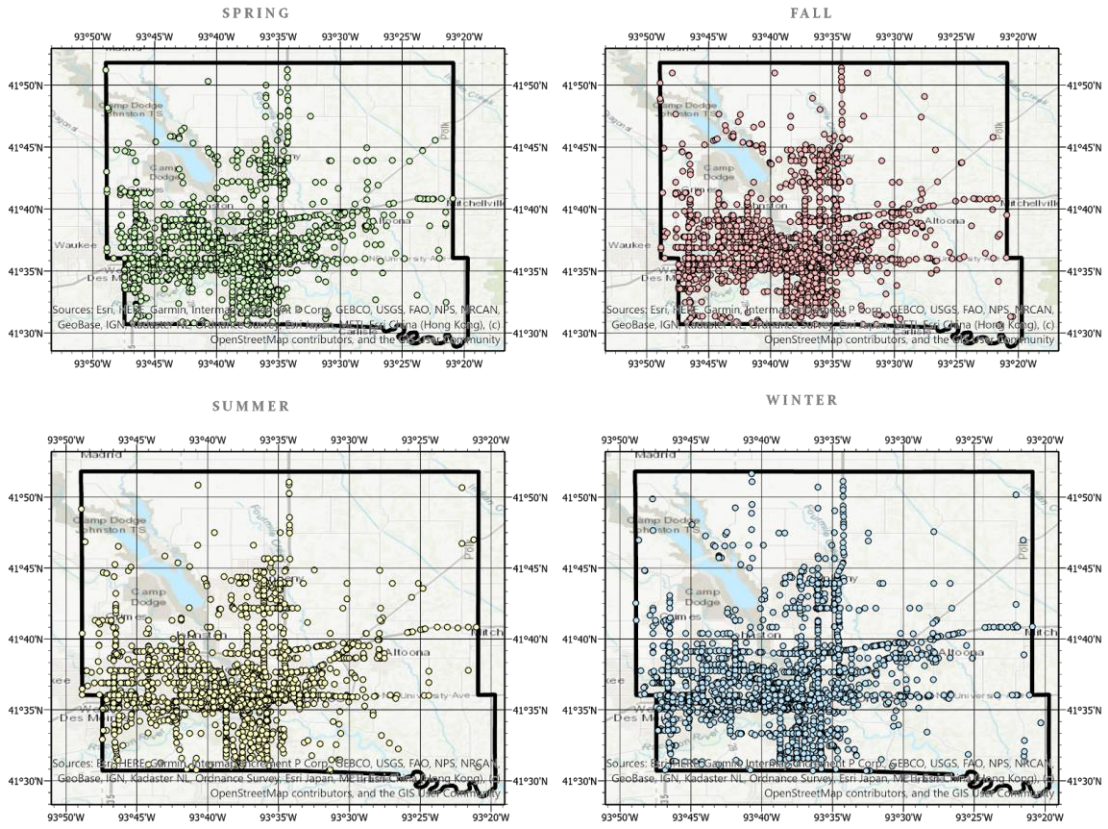


Fig. 4 Car crashes season distribution in Polk County (2013);

Fig. 4 shows how car crashes are distributed in different seasons in Polk County. Due to the massive data, we have found that it is hard to determine a suitable scale to generalize an overall spatial distribution pattern of car crashes. But it can still be noticed that in Fig. 4, the winter and spring car crash maps have more crash data points on the upright and bottom-right corner than other two season maps. So at least in these two areas in Polk county, there were more crashes occurred in winter and spring seasons. And from Fig.1, it is obvious that the area with car crash data points in Polk county map is larger than that in Scott county, meaning that Polk county has more car crashes than Scott county.

V.CONCLUSION AND FUTURE WORK

Traffic collisions cause severe public safety problems, leading to not only deaths and injuries, but also significant economic loss. In order to provide innovative insights on road safety researches, this study proposed a method with self-organizing maps neural networks to cluster traffic data and further make contribution to related research purposes. To evaluate the proposed method, we conducted a case study by utilizing car collisions in Polk County and Scott County, IA, using traffic data collected in the year of 2013 to analyze traffic distribution patterns through spatial and temporal aspects. The findings of this research are as follows: 1) The monthly variation trend of car collisions shows that crashes occur more frequently in the winter season (i.e., Dec, Jan, & Feb). It implies and confirms that weather conditions are highly associated with the crash occurrence; 2) Daily and hourly variation trend of car collisions shows that weekdays have more car collisions than weekends. And 7 am, 4 pm, and 5 pm are the top 3 high-risk hours for driving due to high traffic volume.3) The car collision spatial distribution shows that there is a difference existing among counties coming from county traffic construction while similar temporal patterns exist in both counties.

Many other conclusions can be made about the data gathered through this project. Patterns that stood out to us include the clustering of data points at peak driving hours and during the winter. What interests us more is the application of this model for future use in crash data modelling. At the basic level, we were able to cluster points based on time of day. The ability for this analysis to now be overlayed with accident specific attributes within a toolbox presents an interesting opportunity. By allowing the user to choose to filter by specific attributes, more in-depth analysis can be made. For example, should the user choose to filter by accidents occurring during high snowy conditions and are described as happening because a car was following too closely, one can determine which sections of roadway may require more markers for intersections or pinpoint intersections which may be blocked from view. Crashes happening at night might indicate roadways that require more lighting or have broken streetlights. A high occurrence of crashes happening on unpaved roadways may indicate high traffic areas that require pavement. A high occurrence of accidents happening where traffic is heading east in the morning or west at night may indicate crashes due to light reflection, which may be remedied through darker roads or road shading.

There are a myriad of conclusions that can be made through different combinations of data and these patterns can now be deduced because the large amount of data can be

accurately modelled through SOMs. However, as the spatial analysis results mentioned in section IV show that it is hard to summarize the traffic pattern spatial distribution from a general view and small scale. So research purposes with larger map scales would be more likely to achieve.

Reference

Anderson, T. (2007). Comparison of spatial methods for measuring road accident 'hotspots': a case study of London. *Journal of Maps*, *3*(1), 55-63.

Anderson, T.K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3), 359-364.

Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information & Management*, *37*(5), 271-281.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Plug, C., Xia, J. C., & Caulfield, C. (2011). Spatial and temporal visualisation techniques for crash analysis. *Accident Analysis & Prevention*, *43*(6), 1937-1946.

Prasannakumar, V., Vijith, H., Charutha, R., & Geetha, N. (2011). Spatio-temporal clustering of road accidents: GIS based analysis and assessment. *Procedia-Social and Behavioral Sciences*, *21*, 317-325.

Tang, L., Kan, Z., Zhang, X., Sun, F., Yang, X., & Li, Q. (2016). A network Kernel Density Estimation for linear features in space–time analysis of big trace data. *International Journal of Geographical Information Science*, *30*(9), 1717-1737.

Thakali, L., Kwon, T. J., & Fu, L. (2015). Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation*, 23(2), 93-106.