Term Project

ANALYSIS OF MONTANA VEHICLE CRASH DATA USING KERNEL DENSITY ESTIMATION AND K-MODES CLUSTERING

Matthew Campbell and Saidur Rahman

CSCI 550: DATA MINING | DECEMBER 7TH, 2018

Abstract

Within the US, vehicle crashes are the 13th most prevalent cause of death. Montana contributes significantly to this statistic and consistently ranks at or near the top of fatal vehicular crashes per capita within US states. In this project, we used Kernel Density Estimation (KDE) and K-modes clustering in an attempt to provide insight into both *where* and *why* these crashes are occurring. Although the dataset we used contained information on crashes throughout Montana, given scope and time constraints of this project we only performed this analysis for Gallatin, Madison, and Park counties. However, these same methods can be applied to other Montana counties in the future.

First, KDE was used to generate a density map of each county showing contoured areas where a significant amount, or density, of crashes were occurring, otherwise known as a crash "hotspot." Using this density map, we identified the two largest "hotspots" within each county and used querying to generate new datasets that only contained crashes from each "hotspot."

Next, K-modes clustering was performed on each "hotspot" dataset. We attempted several different combinations of crash attributes and number of clusters to identify an optimal clustering for each crash "hotspot." This same clustering was also applied to the three county datasets as a whole and the Montana vehicular crash dataset as a whole.

The results of this analysis indicate that most crashes at each hotspot occur in ideal roadway conditions, suggesting that other factors such as alcohol or distracted driving contributed to crashes at these locations. While showy weather and icy or snowy roads were attributes in about half of the clusters generated within the county and Montana datasets, they were only in about 12% of the clusters generated within the hotspot datasets, indicating that snow and ice actually influence crashes *less* at these "hotspots" then they do statewide. Dark lighting conditions also contributed to a surprising amount of crashes and were represented in about 30% of clusters generated.

Although this analysis was able to yield some interesting results, there is much more research that can be done into crashes in Montana. Possible future research directions include attempting other analysis strategies such as linear regression or analyzing more specific areas of crashes, such as within towns and cities, in greater detail.

Table of Contents

List of Tables

List of Figures

List of Equations

Introduction

Within the United States, automobile-related accidents rank as one of the leading causes of death each year. In 2016, approximately 37,000 people were killed in these accidents (NHTSA, 2017). Although there has been an overall downward trend over the last several decades–attributed to safer vehicles, better safety technology, and campaigns to increase safe habits such as seatbelt use– this trend has tapered in recent years thanks to factors such as cell phone use. With around 200 vehicle fatalities each year, Montana regularly ranks near or at the top of the list of fatal vehicle fatalities per capita within US states (IIHS & HLDI, 2017).

Although this ranking is somewhat expected due to the large number of highways and the overall vehicle-centric culture that exists within the state, it considered unacceptable by government officials. For this reason, the Montana Department of Transportation (MDT) launched a new initiative, called VisionZero, in 2014. This initiative has the goal of eventually achieving zero deaths or serious injuries on Montana highways. Thanks to VisionZero, efforts to improve the safety of roads and promote safe driving habits have received increased importance over the past 4 years. Although vehicle fatalities have been consistently dropping over the past several years, thanks at least in-part to VisionZero, there were still 1025 fatalities and serious injuries in Montana in 2016 (Montana Department of Transportation, 2017). Clearly, continuing the safety efforts under VisionZero while actively exploring additional methods to improve road safety is essential to ensure that this statistic continues to decrease all the way to 0.

For our project, we used Kernel Density Estimation (KDE) and k-modes clustering on Montana vehicle crash data to identify both significant spatial “hotspots” of crashes and common combinations of crash factors (i.e. icy roads, nighttime conditions, and presence of wildlife) that lead to crashes both throughout the state and specifically at these “hotspots.” Using the results of this analysis, MDT and other public entities should be able to target specific locations for road safety improvements that will theoretically prevent the greatest number of crashes.

The organization of this report is as follows. In Related Work, we begin by providing an overview of existing work into analyzing vehicular crash data using related data mining techniques. Then, in Background, we provide a detailed overview of current efforts to prevent vehicular crashes and the data used for our analysis. In The primary dataset we are using is an excel file that contains a comprehensive list of car crashes within the state of Montana. Each crash listing contains the county and city (if applicable) that the crash occurred in, its day and time, weather and lighting conditions, location on roadway (i.e. roadway, shoulder, off the pavement), type of collision (animal, rollover, rear-end, etc.), unique roadway ID (which can be used to reference additional roadway information), and crash’s GPS coordinates.

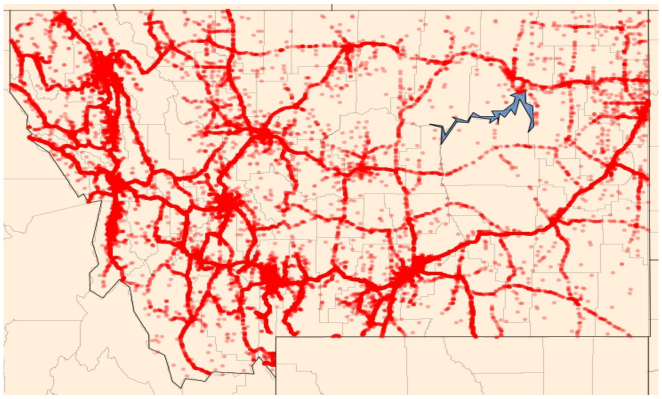
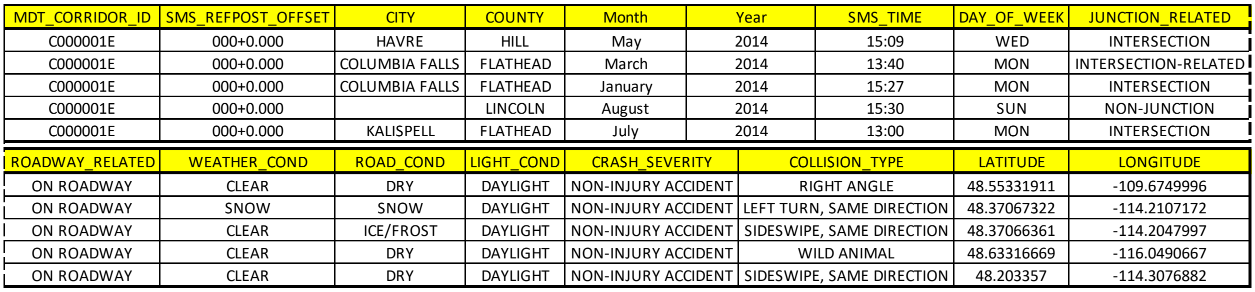


Figure 1: Plot of all crashes within the MDT Crash Dataset.

This dataset contains crash data for the last 5 years and is publicly available from the Montana Department of Transportation (MDT) webpage. Prior to applying data mining methods, we first refined and edited the Montana crash dataset to allow it to be more easily analyzed and imported into Python and R. Using excel, we omitted all data points that were missing any attribute except “city” (many crashes occurred outside city limits). We also removed redundant entries such as “X” and “Y” coordinates, which we assumed to be a different format of spatial location. Finally, the file was saved in a .csv format instead of an excel format. The first 5 rows of the refined dataset are shown in Table 1.

Table 1: First five columns of raw data from MDT.



MDT_CORRIDOR_ID	SMS_REFPOST_OFFSET	CITY	COUNTY	Month	Year	SMS_TIME	DAY_OF_WEEK	JUNCTION_RELATED
C000001E	000+0.000	HAVRE	HILL	May	2014	15:09	WED	INTERSECTION
C000001E	000+0.000	COLUMBIA FALLS	FLATHEAD	March	2014	13:40	MON	INTERSECTION-RELATED
C000001E	000+0.000	COLUMBIA FALLS	FLATHEAD	January	2014	15:27	MON	INTERSECTION
C000001E	000+0.000		LINCOLN	August	2014	15:30	SUN	NON-JUNCTION
C000001E	000+0.000	KALISPELL	FLATHEAD	July	2014	13:00	MON	INTERSECTION

ROADWAY_RELATED	WEATHER_COND	ROAD_COND	LIGHT_COND	CRASH_SEVERITY	COLLISION_TYPE	LATITUDE	LONGITUDE
ON ROADWAY	CLEAR	DRY	DAYLIGHT	NON-INJURY ACCIDENT	RIGHT ANGLE	48.55331911	-109.6749996
ON ROADWAY	SNOW	SNOW	DAYLIGHT	NON-INJURY ACCIDENT	LEFT TURN, SAME DIRECTION	48.37067322	-114.2107172
ON ROADWAY	CLEAR	ICE/FROST	DAYLIGHT	NON-INJURY ACCIDENT	SIDESWIPE, SAME DIRECTION	48.37066361	-114.2047997
ON ROADWAY	CLEAR	DRY	DAYLIGHT	NON-INJURY ACCIDENT	WILD ANIMAL	48.63316669	-116.0490667
ON ROADWAY	CLEAR	DRY	DAYLIGHT	NON-INJURY ACCIDENT	SIDESWIPE, SAME DIRECTION	48.203357	-114.3076882

Due to the large number of points within this dataset, nearly 1,000,000 points, efforts were taken to reduce its size for analysis. We ultimately decided to separate the dataset into counties, and only performed KDE analysis on Gallatin, Park, and Madison counties due to their proximity to Bozeman.

Methodology, we provide an overview of the algorithms and code used for both our KDE and k-modes analysis, in addition to changes in our methodology since submission of the Term Project Progress Report. In Experimental Evaluation, we provide a detailed breakdown of the results of this analysis. Finally, in Conclusion and Future Work, we provide an interpretation of these results, ways in which our methodology could be improved for future research, and proposed future directions for data mining research into vehicular crash data.

Related Work

Various data mining techniques have been used to analyze vehicular crash data throughout the world. To our knowledge, the only previous analysis of Montana vehicular crash data based on spatial information was performed by the law firm Ragain & Cook, P.C. This analysis simply analyzed crash data at intersections, and as far as we can tell only considered the sheer number of accidents that occurred at a given intersection (Ragain & Cook, 2017).

Within related work on the spatial analysis of vehicular crash data, two data mining techniques were most commonly used; Kernel Density Estimation (KDE) and spatial clustering. Sabel, Kingham, Nicholson, and Bartie used KDE on traffic accident data in conjunction with traffic volume data within New Zealand. Statistically significant traffic accident hotspots were identified through this analysis (Sabel, Kingham, Nicholson, & Bartie, 2005). Hashimoto et. all examined the relationship between traffic accident GIS (Geographic Information System) data and city characteristics (such as population and road features) to develop a model to predict traffic accident density. This model was then used to identify areas where future traffic calming and accident reduction projects could occur (Hashimoto et al., 2016). Xie and Yan used KDE in a 1-dimensional linear sense (treating a roadway as said 1D space). An overall roadway network was treated as a collection of 1D linear spaces, and analysis was performed to identify accident hotspots in this manner, differing from approaches described above that analyzed spatial data on a 2-dimensional plane (i.e. GPS coordinates) (Xie & Yan, 2008).

Several researched papers utilized other clustering techniques, often in addition to KDE, to enhance analysis. Prasannakumar et. all used the Morans I method of spatial autocorrelation in addition to KDE to identify clusters of vehicle accidents. The results of these two analysis methods were compared to gain additional insights such as the distribution of hotspots and accidents that occur outside of those significant hotspots (Prasannakumar, Vijith, Charutha, & Geetha, 2011). Anderson used K-means clustering in addition to KDE. KDE is first used in conjunction to GIS data from the UK, to identify significantly dense grouping of hotspots. Then, K-means is used to identify clusters of accidents within these hotspots that exhibit similar characteristics of roadway or environmental conditions (Anderson, 2009).

Several other studies have also analyzed categorical attributes of vehicular crash data using k modes. Kaur et. all used k modes on unspecified crash data to identify common attributes, with emphasis on optimize frequent “single things” and “people things” co-occurrence with them (Kaur, Luhach, & Pooja, 2017). Kumar et. all have also used k-modes clustering of vehicular crash data in an effort to improve later classification and correct for the biased nature of crash data towards less-serious crashes (as opposed to fatal or serious injury crashes) (Kumar, Semwal, Solanki, Tiwari, & Kalitin, 2017).

Although both data mining techniques we used in our project (KDE and k-means) have both been used in some form to analyze crash data, we do not believe that any previous study has implemented k-means on hotspots identified *by* KDE analysis as we do in this project. Additionally, none of the studies outlined in this section analyzed Montana-specific data or even crash data that is rural in nature with significant weather (i.e. winter conditions) and wildlife factors. We believe a major advantage of our project is that although we are not proposing any new data mining techniques per se, our analysis provides tangible results that can be immediately used in real-world applications throughout the state.

Background

As stated in the Introduction, vehicle fatalities are one of the primary causes of death within the US, and Montana regularly ranks near or at the top of the list of fatal vehicle fatalities per capita within US states (IIHS & HLDI, 2017). Vehicle crashes occur for a variety of reasons, and often a multitude of factors come into play when a crash occurs. However, all crashes can be attributed to a “critical event.” Reasons that cause these “critical events” can be classified into the following categories: errors attributable to the driver, the condition of the vehicle, failure of vehicle systems, adverse environmental conditions, and subpar roadway design (Highway Traffic Safety Administration, 2008).

Three of these reasons—driver error, environmental conditions, and roadway design—have spatial characteristics. Put another way, locations along a roadway or within a street network can be said to have higher or lower probabilities of crashes occurring because these three reasons. For example, a distracting billboard or confusing intersection at a specific location may be attributable to a higher rate of crashes due to driver error. A certain spot on a highway may have a higher probability of wildlife crossings or black ice patches, yielding a higher rate of crashes due to environmental conditions. Finally, an improperly designed highway curve or section of roadway prone to defects such as potholes can be associated with a higher rate of crashes due to subpar road design.

If these locations with higher rates of crashes attributable to one of the above reasons can be identified, then specific efforts such as re-construction or better safety outreach can be targeted to these locations to reduce crash rates. Currently, locations for new roadway construction in Montana are identified primarily analysis of a roadway’s condition. From our understanding, crash rates are not always taken into account, and when they are it is most often in the form of one or several high-profile crashes that receive noteworthy news attention.

We believe that a more methodological approach that supplements MDT’s current system of identifying locations for roadway improvements and outreach efforts can help ensure that taxpayer money and department resources are being put towards efforts to reduce vehicle-related fatalities and injuries in the most efficient manner possible. Applying KDE to the spatial components of MDT’s crash data (latitude & longitude) to identify hotspots of roadway crashes, then applying k-modes clustering to identify significant combinations of crash attributes at the hotspot, county, and statewide level serves as a promising method of providing this methodological approach.

Data Source

The primary dataset we are using is an excel file that contains a comprehensive list of car crashes within the state of Montana. Each crash listing contains the county and city (if applicable) that the crash occurred in, its day and time, weather and lighting conditions, location on roadway (i.e. roadway, shoulder, off the pavement), type of collision (animal, rollover, rear-end, etc.), unique roadway ID (which can be used to reference additional roadway information), and crash’s GPS coordinates.

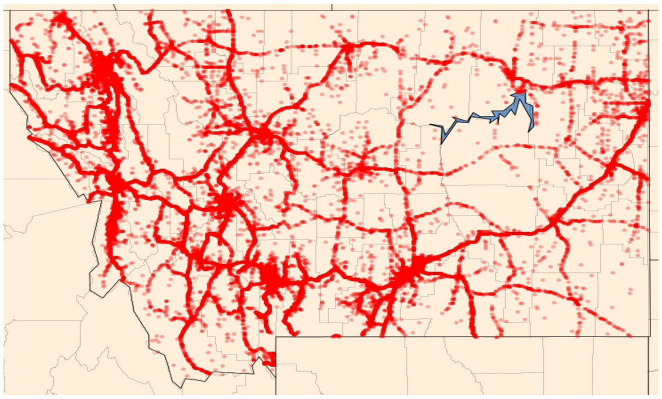
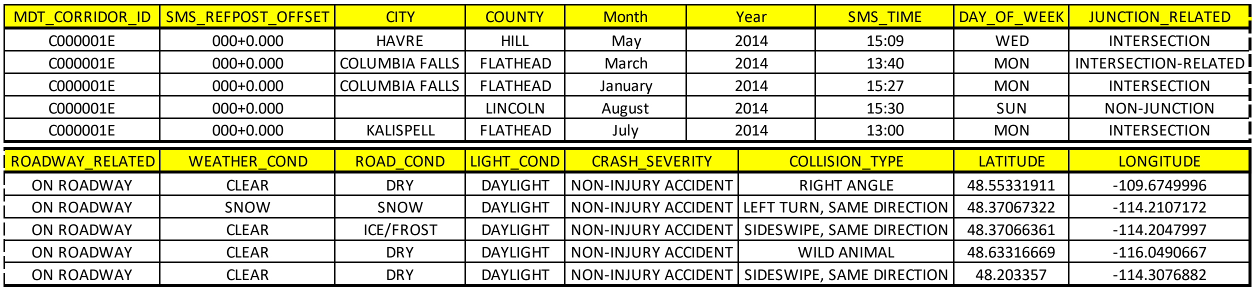


Figure 1: Plot of all crashes within the MDT Crash Dataset.

This dataset contains crash data for the last 5 years and is publicly available from the Montana Department of Transportation (MDT) webpage.[1] Prior to applying data mining methods, we first refined and edited the Montana crash dataset to allow it to be more easily analyzed and imported into Python and R. Using excel, we omitted all data points that were missing any attribute except "city" (many crashes occurred outside city limits). We also removed redundant entries such as "X" and "Y" coordinates, which we assumed to be a different format of spatial location. Finally, the file was saved in a .csv format instead of an excel format. The first 5 rows of the refined dataset are shown in Table 1.

Table 1: First five columns of raw data from MDT.



MDT_CORRIDOR_ID	SMS_REFPOST_OFFSET	CITY	COUNTY	Month	Year	SMS_TIME	DAY_OF_WEEK	JUNCTION_RELATED	ROADWAY_RELATED	WEATHER_COND	ROAD_COND	LIGHT_COND	CRASH_SEVERITY	COLLISION_TYPE	LATITUDE	LONGITUDE
C000001E	000+0.000	HAVRE	HILL	May	2014	15:09	WED	INTERSECTION	ON ROADWAY	CLEAR	DRY	DAYLIGHT	NON-INJURY ACCIDENT	RIGHT ANGLE	48.55331911	-109.6749996
C000001E	000+0.000	COLUMBIA FALLS	FLATHEAD	March	2014	13:40	MON	INTERSECTION-RELATED	ON ROADWAY	SNOW	SNOW	DAYLIGHT	NON-INJURY ACCIDENT	LEFT TURN, SAME DIRECTION	48.37067322	-114.2107172
C000001E	000+0.000	COLUMBIA FALLS	FLATHEAD	January	2014	15:27	MON	INTERSECTION	ON ROADWAY	CLEAR	ICE/FROST	DAYLIGHT	NON-INJURY ACCIDENT	SIDESWIPE, SAME DIRECTION	48.37066361	-114.2047997
C000001E	000+0.000		LINCOLN	August	2014	15:30	SUN	NON-JUNCTION	ON ROADWAY	CLEAR	DRY	DAYLIGHT	NON-INJURY ACCIDENT	WILD ANIMAL	48.63316669	-116.0490667
C000001E	000+0.000	KALISPELL	FLATHEAD	July	2014	13:00	MON	INTERSECTION	ON ROADWAY	CLEAR	DRY	DAYLIGHT	NON-INJURY ACCIDENT	SIDESWIPE, SAME DIRECTION	48.203357	-114.3076882

Due to the large number of points within this dataset, nearly 1,000,000 points, efforts were taken to reduce its size for analysis. We ultimately decided to separate the dataset into counties, and only performed KDE analysis on Gallatin, Park, and Madison counties due to their proximity to Bozeman.

Methodology

The two primary data mining techniques used within our project are Kernel Density Estimation (KDE) and k-modes clustering. KDE was implemented first, and k-modes was later used to analyze the "hotspots" determined through the KDE analysis. All code used for this project can be found at the referenced link.[2]

As stated previously, we only applied KDE on data within Gallatin, Park, and Madison counties in an effort to reduce the number of data points being analyzed. We decided that looking at the three most populous counties directly surrounding Bozeman would yield the most interesting/relevant results for a study based out of MSU.

Kernel Density Estimation (KDE)

KDE is a popular method for estimating the probability density function of a random variable. At each data point, a density estimate $\hat{f}(x)$ is calculated using Equation 1. *K* in the equation is the discrete kernel function. For our project, we used the *gaussian* kernel.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Equation 1: Kernel Density Estimation function.

The most significant parameter of the function is *h*, which is most often called the bandwidth. Essentially, this is smoothing parameter that determines how responsive the density estimate is to changes in a function's data. The smaller the value for *h*, the more sensitive the estimate will be to changes in density. However, if the value is too small the density estimate will over-fit the data, known as *undersmoothing.* The opposite occurs if a value for *h* is chosen that is too large. Although there are several different optimization parameters that can be used, often focusing on squared errors, given the scope and time constraints of this project we chose values for bandwidth that seemed to yield the best representation of crashes within a county.

[1] Data can be found at: https://www.mdt.mt.gov/publications/datastats/crashdata.shtml

[2] Code can be found at: https://github.com/rifathcsedu/Montana_Crash_Data

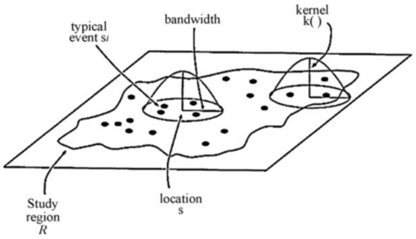


Figure 2: Visualization of KDE (Anderson, 2009).

Python was used to implement KDE. To apply the KDE function to our dataset, we used the **`KernelDensity`** function from the **`sklearn.neighbors`** library. This function calculated the value of the density function (specified in Equation 1) at a specified bandwidth for an array of points that covered the area we were analyzing (i.e. an array covering Gallatin, Park, or Madison county). Once we had density estimates for all points within the array, we used the **`Basemap`** function from the **`mpl_toolkits.basemap`** library to plot the results in the form of a density map. We used the database Postgres 11 (pgadmin4).

One issue we encountered when initially applying KDE to the county-level datasets is that given the resolution and small map scale we were analyzing data at (the base map is zoomed out to show the entire county, more than a million acres) the resulting density maps were simply showing hotspots that corresponded to major towns in that county. For example, as seen in Figure 3, the only "hotspots" identified corresponded to the locations of Bozeman/Belgrade, West Yellowstone, and Big Sky. Clearly, many crashes are happening in these towns, and these maps do not provide any useful information.

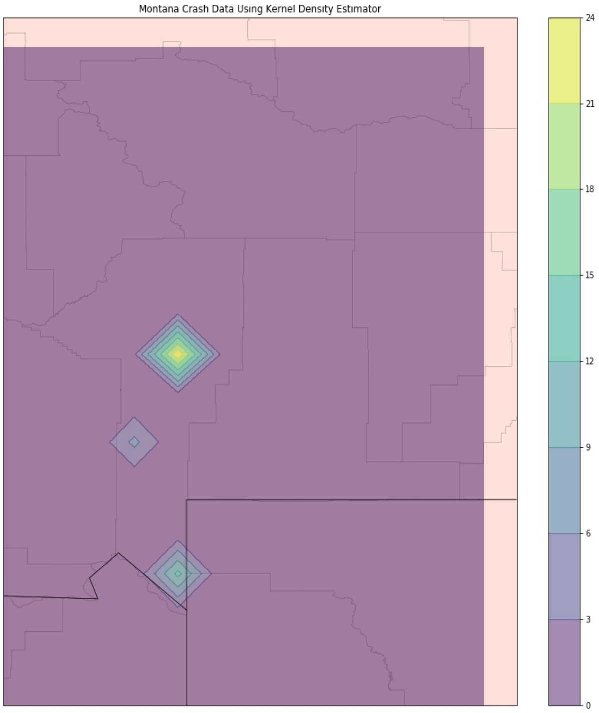


Figure 3: Initial KDE attempt for Gallatin County.

We decided that within the scope of this project, the most interesting results we could generate from KDE were *rural* highway locations that encountered a lot of crashes. Therefore, we omitted all data points that occurred within the established boundaries of major towns in each county; Bozeman/Belgrade, Manhattan, Three Forks, Big Sky, and West Yellowstone for Gallatin, Livingston and Gardiner for Park, and Ennis, Virginia City, and Twin Bridges for Madison. Therefore, our results shown in the next section show hotspots that are only influenced by rural crashes, ones that occur outside of towns.

Selecting Hotspot Data Points for K-modes Clustering

Once a proper density map had been generated, we then needed to establish a systematic method for identifying the main hotspots and querying crash data points that were associated with each hotspot. To determine the main hotspots, we identified the two largest "peaks" of densities for each county and labeled them as "Point 1" and "Point 2" respectively, as seen in Figure 4.

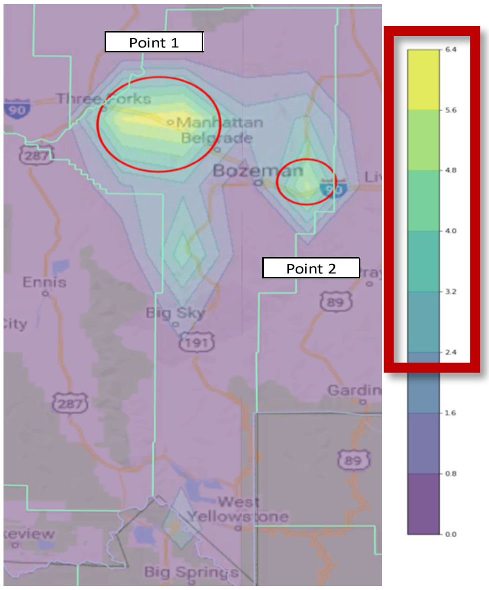


Figure 4: Visualization of circles drawn to represent hotspots for points 1 and 2 in Gallatin county.

Once the two largest hotspots were identified, we next developed a method to query the related points. We decided that any crash that occurred within the top 5 density "levels" determined by the `Basemap` function at one of our determined points would be considered to be associated with that hotspot. We were unable to identify the classification method `Basemap` uses to determine its 8 "levels" of density data (i.e. 8 colors representing different densities), but we assume either equal interval or quantile classification was used.

Due to limitations encountered with the codes and functions used, we determined that the most efficient way to proceed with selecting points using this method was to overlay the generated density map onto Google Maps and draw a circle around both points with a radius that best approximates the boundary of the 5th density "level" of the map (shown in Figure 4). Then, within Excel, the distance between every crash in the dataset and the center of both circles was calculated using the Haversine formula, shown in Equation 2. From there, we were able to simply query all crashes whose distance to a given circle center was less than its determined radius.

$$a = sin^2\left(\frac{\Delta\varphi}{2}\right) + cos\ \varphi 1 * cos\ \varphi 2 * sin^2(\Delta\lambda/2)$$
$$c = 2 * atan2(\sqrt{a}, \sqrt{1-a})$$
$$d = R * c$$

Equation 2: Haversine formula for great-circle distance between two GPS points.

Where:

φ = Latitude of each point,
λ = Longitude of each point,
a = Square of half of the chord length between points,
c = Angular distance between points in radians,
R = Earth's radius, and;
d = Great-circle distance between points.

We acknowledge that this is not a perfect method and will outline later in this report ways in which we would improve this method for future research.

K-modes Clustering

K-modes is a method of clustering categorical data that cannot be clustered using traditional methods such as k-means. K-modes functions similar to k-means with several key differences. First, for each iteration of the algorithm the mode of the cluster is calculated instead of the mean. Second, a *dissimilarity score* is calculated between a point and each cluster mode as opposed to calculating a more traditional distance (such as Euclidean). The specific steps are given below (Kar, 2017):

1. Randomly assign *k* number of clusters by selecting *k* random data points to serve as the cluster modes.
2. For each point:
    a. Calculate the *dissimilarity score* between it and all clusters.
    b. Assign to cluster with the lowest *dissimilarity score* (Equation 3).
3. Calculate the mode of each new cluster.
4. Repeat steps 2 and 3 until convergence, i.e. points no longer move between clusters.

$$Dissimilarity(X, Y) = \sum_{j=1}^{n} \delta(X_j, Y_j)$$

Equation 3: Dissimilarity score for K-modes clustering (Kar, 2017).

Where:

$$\delta(X_j, Y_j) = \begin{cases} 0 \; if \; X_j = Y_j \\ 1 \; if \; X_j \neq Y_j \end{cases}$$

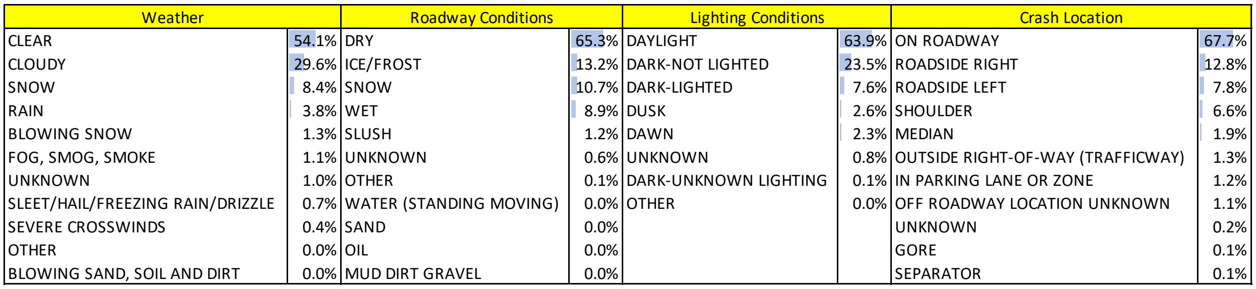
And *X* and *Y* are categorical objects with *n* attributes.

Similarly to KDE, Python was used to implement K-modes clustering. We used the **`Kmodes`** function from the **`kmodes.kmodes`** library and also used **`matplotlib.pyplot`** for initialization of data.

We were unable to find robust evaluation metrics for K-modes clustering similar to what can be used within K-means. The best metric we were able to implement was the percentage of data that was clustered for a given *k* number of clusters and number of attributes included within each cluster. We used a lower threshold of 2% within this analysis, i.e. if a series of clusters reported the smallest cluster as containing less than 2% of the dataset, that series of clusters was disregarded.

We chose four attributes to perform K-modes clustering on. They are shown below in Table 2. The percentages next to each attribute show what percentage of the crashes contain the given attribute.

Table 2: Attributes used in K-modes clustering.



Weather		Roadway Conditions		Lighting Conditions		Crash Location	
CLEAR	54.1%	DRY	65.3%	DAYLIGHT	63.9%	ON ROADWAY	67.7%
CLOUDY	29.6%	ICE/FROST	13.2%	DARK-NOT LIGHTED	23.5%	ROADSIDE RIGHT	12.8%
SNOW	8.4%	SNOW	10.7%	DARK-LIGHTED	7.6%	ROADSIDE LEFT	7.8%
RAIN	3.8%	WET	8.9%	DUSK	2.6%	SHOULDER	6.6%
BLOWING SNOW	1.3%	SLUSH	1.2%	DAWN	2.3%	MEDIAN	1.9%
FOG, SMOG, SMOKE	1.1%	UNKNOWN	0.6%	UNKNOWN	0.8%	OUTSIDE RIGHT-OF-WAY (TRAFFICWAY)	1.3%
UNKNOWN	1.0%	OTHER	0.1%	DARK-UNKNOWN LIGHTING	0.1%	IN PARKING LANE OR ZONE	1.2%
SLEET/HAIL/FREEZING RAIN/DRIZZLE	0.7%	WATER (STANDING MOVING)	0.0%	OTHER	0.0%	OFF ROADWAY LOCATION UNKNOWN	1.1%
SEVERE CROSSWINDS	0.4%	SAND	0.0%			UNKNOWN	0.2%
OTHER	0.0%	OIL	0.0%			GORE	0.1%
BLOWING SAND, SOIL AND DIRT	0.0%	MUD DIRT GRAVEL	0.0%			SEPARATOR	0.1%

Experimental Evaluation

The following sections will detail the results of our analysis.

KDE

For each county that KDE was implemented on, a density map was generated that shows significant areas of high-density crashes (hotspots). As stated previously, we labeled the largest "hotspot" as "Point 1" and the second-highest as "Point 2."

Gallatin county results are shown in Figure 5. Four hotspots were identified; three surrounding Bozeman on various highways and one just outside of West Yellowstone on US 20. All four hotspots make logical sense given the high volume of traffic and sometimes treacherous conditions that can occur on associated highway segments.

The top hotspot, labeled Point 1, occurs between Three Forks and Manhattan on I-90. This segment of roadway receives a large amount of traffic including both commuters to Three Forks or nearby towns and vehicles/trucks that are driving to other Montana towns such as Helena, Butte, and beyond.

The second highest hotspot, labeled Point 2, occurs just east of Bozeman on I-90. This segment of I-90 is the beginning of Bozeman Pass and can often be treacherous due to the tight highway curves through the mountains, the inclement conditions that can develop in the winter, and the high volume of traffic including commuters to Livingston and vehicles/trucks continuing to Billings.

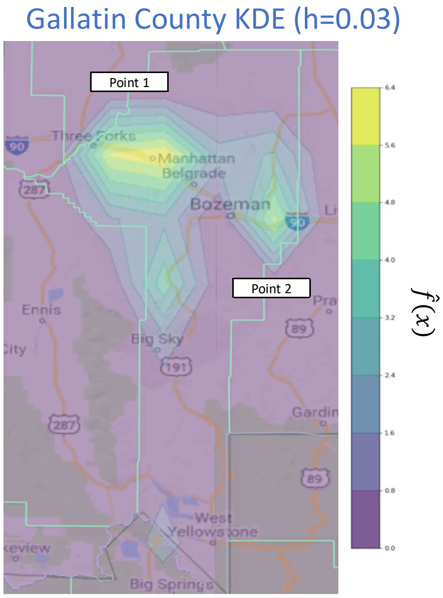


Figure 5: KDE results for Gallatin County.

Madison county results are shown in Figure 6. Four hotspots were identified throughout the county, with by far the largest hotspot occurring on the stretch of US 287 between McAllister and Ennis. Ennis is a popular town in the summer due to its quaint downtown and proximity to renowned fly-fishing and other outdoor activities. Many businesses and homes are scattered off the highway within this stretch, so many crashes are likely influenced by the high volume of turning traffic conflicting with thru vehicles in this area.

The second highest hotspot occurs between Twin Bridges and Silver Star on MT 41. This highway continues eventually intersects with I-90 near Whitehall and sees a fairly high volume of traffic associated with Twin Bridges and further south towns such as Dillon who are traveling to cities accessible from I-90 such as Bozeman.

One interesting hotspot identified is the one that occurs in Pony, MT. This is a small town that lies off any major highways, and therefore should not see a very high volume of traffic or crashes. One possible reason for this hotspot is its bar that is popular on weekends and frequented by many people from surrounding communities. Due to the lack of alternative transportation home if patrons are inebriated, many of these crashes may involve alcohol.

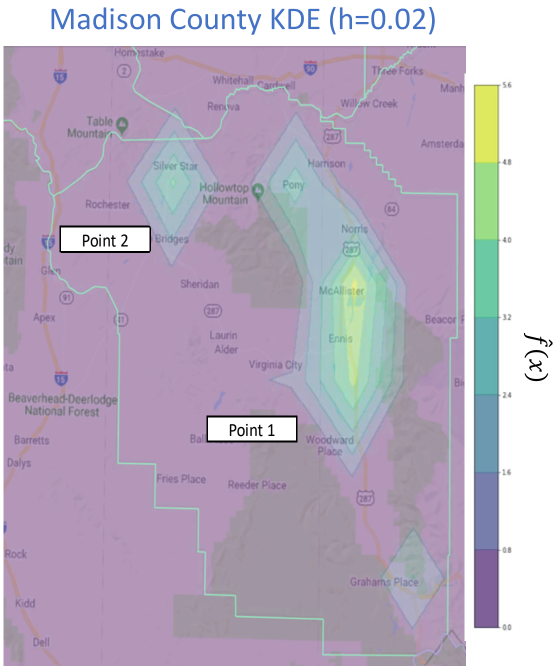


Figure 6: KDE results for Madison County.

Park county results are shown below in Figure 7. Three hotspots were identified; two along US 89 between Livingston and Gardiner and one on I-90 and US 89 just north of Livingston. The largest hotspot is north of Livingston and is likely due to a combination of tight highway turns along I-90 and inclement road conditions such as ice/snow and severe winds that can occur.

The second highest hotspot occurs on US 89 about at the location of a small town called Emigrant. Overall, Emigrant is a small town with just one main intersection on the main highway. Despite this, several cafes, bars, and other businesses are popular with tourists driving to or from Yellowstone National Park. The amount of traffic at this single, STOP-controlled intersection is likely a major contributor to crashes at this location.

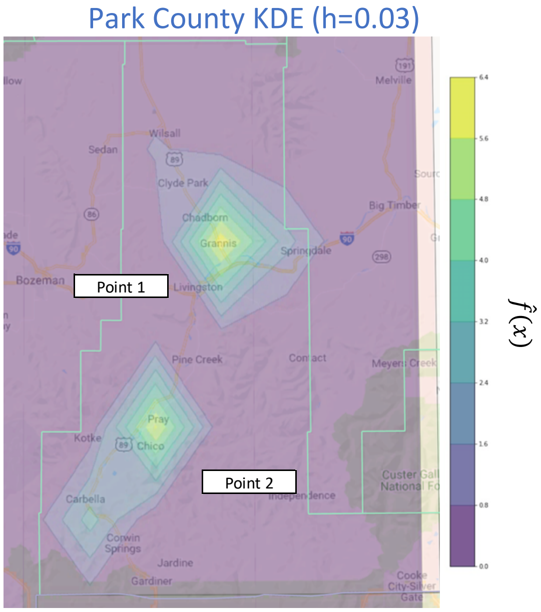


Figure 7: KDE results for Park County.

K-modes

For each of the two points for each county (six points total), K-modes clustering was applied. Two combinations of attributes were tested. The first combination was Weather, Roadway Conditions, and Lighting conditions. The second combination added a 4th attribute, Crash Location. 2-4 sets of clusters were attempted for both combinations. The results of this analysis are shown in Table 3, Table 4, and Table 5. For each cluster identified, the attributes associated with that cluster are given as well as the percentage of the data that is contained within that cluster, shown directly below. The optimal clustering for each hotspot point is highlighted in green.

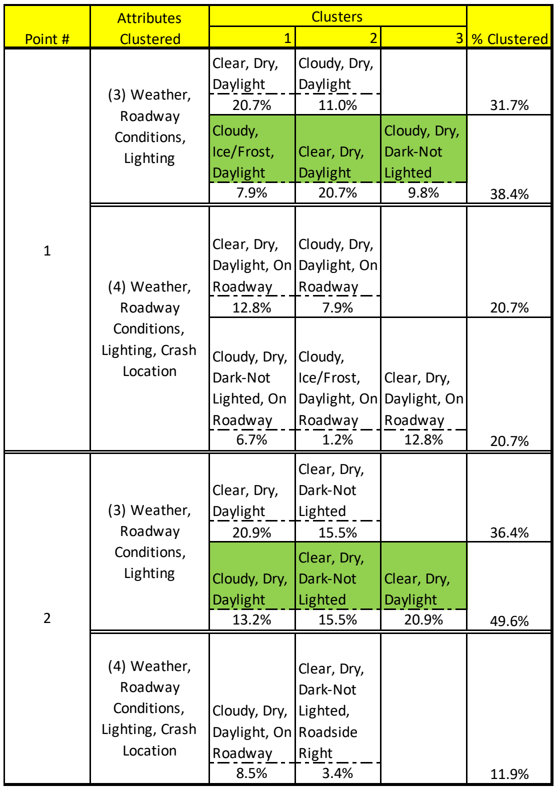
Table 3: K-modes clustering results for Gallatin county.

Point #	Attributes Clustered	Clusters				% Clustered
		1	2	3	4	
1	(3) Weather, Roadway Conditions, Lighting	Clear, Dry, Dark-Not Lighted 18.6%	Cloudy, Dry, Daylight 15.3%			33.9%
		Cloudy, Dry, Daylight 15.3%	Clear, Dry, Daylight 19.2%	Cloudy, Dry, Dark-Not Lighted 11.3%		45.8%
		Snow, Ice/Frost, Daylight 2.5%	Cloudy, Dry, Daylight 15.3%	Clear, Dry, Daylight 19.2%	Clear, Dry, Dark-Not Lighted 18.6%	55.6%
	(4) Weather, Roadway Conditions, Lighting, Crash Location	Clear, Dry, Dark-Not Lighted, Roadside Right 4.3%	Cloudy, Dry, Daylight, On Roadway 8.4%			12.7%
		Clear, Dry, Dark-Not Lighted, On Roadway 9.2%	Cloudy, Dry, Daylight, On Roadway 8.4%	Clear, Dry, Daylight, On Roadway 10.4%		28.0%
2	(3) Weather, Roadway Conditions, Lighting	Clear, Dry, Daylight 13.2%	Snow, Ice/Frost, Daylight 5.8%			19.0%
	(4) Weather, Roadway Conditions, Lighting, Crash Location	Cloudy, Dry, Daylight, On Roadway 4.7%	Snow, Ice/Frost, Daylight, Roadside Right 0.8%			5.5%

Table 4: K-modes clustering results for Madison county.

Point #	Attributes Clustered	Clusters		% Clustered
		1	2	
1	(3) Weather, Roadway Conditions, Lighting	Clear, Dry, Daylight 16.9%	Cloudy, Dry, Daylight 10.2%	27.1%
	(4) Weather, Roadway Conditions, Lighting, Crash Location	Clear, Dry, Daylight, On Roadway 8.5%	Cloudy, Ice/Frost, Dark-Not Lighted, Roadside Left 3.4%	11.9%
2	(3) Weather, Roadway Conditions, Lighting	Cloudy, Dry, Dark-Not Lighted 16.7%	Clear, Dry, Daylight 25.0%	41.7%

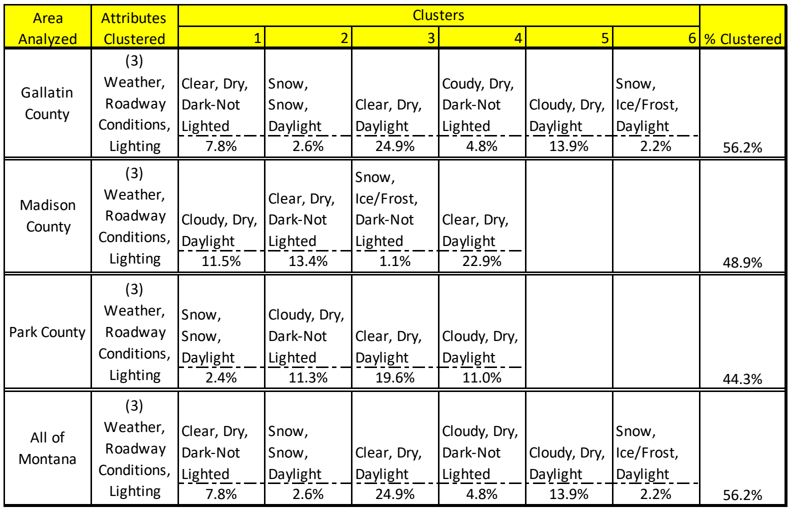
Table 5: K-modes clustering results for Park county.



Point #	Attributes Clustered	Clusters 1	2	3	% Clustered
1	(3) Weather, Roadway Conditions, Lighting	Clear, Dry, Daylight 20.7%	Cloudy, Dry, Daylight 11.0%		31.7%
		Cloudy, Ice/Frost, Daylight 7.9%	Clear, Dry, Daylight 20.7%	Cloudy, Dry, Dark-Not Lighted 9.8%	38.4%
	(4) Weather, Roadway Conditions, Lighting, Crash Location	Clear, Dry, Daylight, On Roadway 12.8%	Cloudy, Dry, Daylight, On Roadway 7.9%		20.7%
		Cloudy, Dry, Dark-Not Lighted, On Roadway 6.7%	Cloudy, Ice/Frost, Daylight, On Roadway 1.2%	Clear, Dry, Daylight, On Roadway 12.8%	20.7%
2	(3) Weather, Roadway Conditions, Lighting	Clear, Dry, Daylight 20.9%	Clear, Dry, Dark-Not Lighted 15.5%		36.4%
		Cloudy, Dry, Daylight 13.2%	Clear, Dry, Dark-Not Lighted 15.5%	Clear, Dry, Daylight 20.9%	49.6%
	(4) Weather, Roadway Conditions, Lighting, Crash Location	Cloudy, Dry, Daylight, On Roadway 8.5%	Clear, Dry, Dark-Not Lighted, Roadside Right 3.4%		11.9%

K-modes was also applied to the entirety of data for each of the three county datasets used for KDE analysis, in addition to the Montana dataset as a whole. Clustering was only performed using the first combination of attributes (Weather, Roadway Conditions, and Lighting). The optimal set of clusters for each dataset is shown below in Table 6.

Table 6: County and statewide K-modes clustering results.



Area Analyzed	Attributes Clustered	Clusters 1	2	3	4	5	6	% Clustered
Gallatin County	(3) Weather, Roadway Conditions, Lighting	Clear, Dry, Dark-Not Lighted 7.8%	Snow, Snow, Daylight 2.6%	Clear, Dry, Daylight 24.9%	Coudy, Dry, Dark-Not Lighted 4.8%	Cloudy, Dry, Daylight 13.9%	Snow, Ice/Frost, Daylight 2.2%	56.2%
Madison County	(3) Weather, Roadway Conditions, Lighting	Cloudy, Dry, Daylight 11.5%	Clear, Dry, Dark-Not Lighted 13.4%	Snow, Ice/Frost, Dark-Not Lighted 1.1%	Clear, Dry, Daylight 22.9%			48.9%
Park County	(3) Weather, Roadway Conditions, Lighting	Snow, Snow, Daylight 2.4%	Cloudy, Dry, Dark-Not Lighted 11.3%	Clear, Dry, Daylight 19.6%	Cloudy, Dry, Daylight 11.0%			44.3%
All of Montana	(3) Weather, Roadway Conditions, Lighting	Clear, Dry, Dark-Not Lighted 7.8%	Snow, Snow, Daylight 2.6%	Clear, Dry, Daylight 24.9%	Cloudy, Dry, Dark-Not Lighted 4.8%	Cloudy, Dry, Daylight 13.9%	Snow, Ice/Frost, Daylight 2.2%	56.2%

Overall Results

The top cluster using optimal clustering for each point is shown below in Figure 8. All 6 clusters contain the attributes "Clear" for Weather, "Dry" for Roadway Conditions, and "Daylight" for Lighting Conditions. Clearly, this indicates that other factors contributing to crashes also must come into play within these. Alcohol, distracted driving, poor roadway design, excessive speed, etc. are all possible additional factors that may contribute to crashes at these three hotspots.

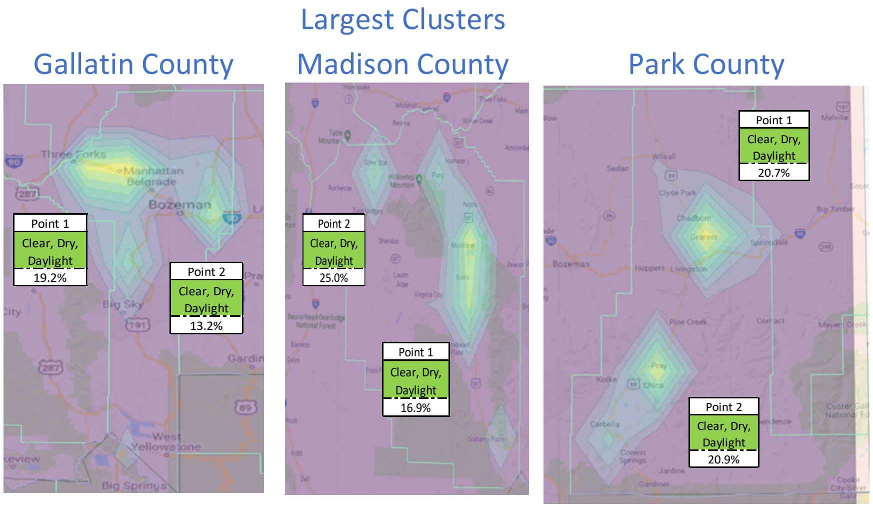


Figure 8: Largest clusters for each county hotspot.

A comparison of cluster attributes between the 6 hotspots, the three counties, and Montana as a whole are shown below in Figure 9. For example, with the "Snow" attribute, only 13% of the clusters (among optimal clusters of the 6 data points) contained this attribute. In contrast, "Snow" appears within 38% of the clusters of the three county datasets, and within 43% of the clusters of all of the Montana crash data. Therefore, snow is less of a factor at the six hotspots identified than it is within the state as a whole.

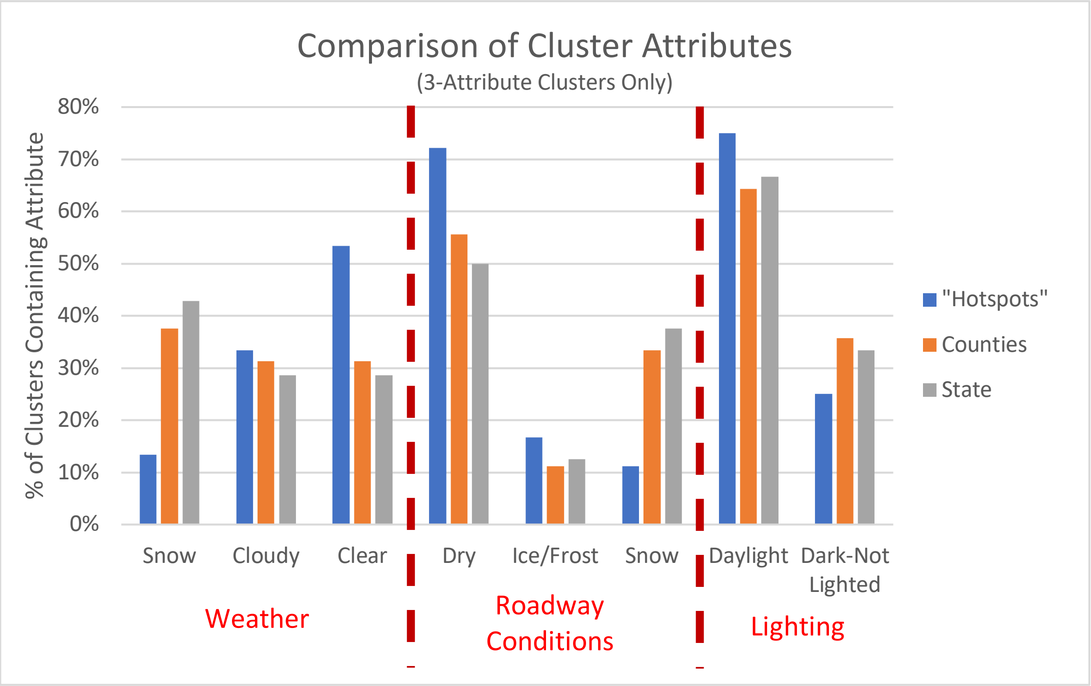


Figure 9: Comparison of cluster attributes.

Conclusion

Overall, the results of this project yielded several key insights into vehicle crashes within the three counties analyzed. Most crashes occurred with ideal roadway and lighting conditions, indicating that other factors came into play. These factors may include distracted driving, alcohol, excessive speed, poor roadway design, etc. although no definitive conclusions can be drawn as to the specific reasons these occurred.

In regard to KDE analysis, most hotspots identified were at expected locations. Highways such as I-90, US 191, US 89, US 287, and US 20 (all highways with one or more hotspots) all see high volumes of traffic and can become treacherous with inclement weather conditions. The only surprising location identified was the hotspot surrounding Pony, MT in Madison County. This hotspot is unexpected given the lack of a major highway with significant traffic. Overall, if efforts and programs to increase highway safety—programs including seatbelt campaigns, increased drunk driving enforcements, etc.—are targeted at the areas near these hotspots, this should ensure that they are as effective as possible. Additionally, efforts to identify and rectify the direct causes of the hotspot in Pony should also yield a significant reduction in accidents within Madison county.

Due to the wide variety of attributes within the Montana vehicular crash dataset—and the number of additional factors not included in the dataset that may come into play for a given crash—it is difficult to draw many conclusions from the clustering analysis given the relatively few number of attributes that could be included within the scope and time constraints of this project. One surprising insight is the number of clusters that included the “Dark-Not Lighted” attribute (about 30% of clusters). This number was higher than expected and indicates that lighting may play a considerable role in crashes. However, lighting may also simply be coincident to other factors without directly contributing to a crash (for example, alcohol-related crashes are significantly more likely to occur in the evening and in the dark).

One final insight into the clustering analysis is that while showy weather and icy or snowy roads were attributes in about half of the clusters generated within the county and Montana datasets, they were only in about 12% of the clusters generated within the hotspot datasets. This indicates that snow and ice actually influence crashes *less* at these areas then they do statewide. This was surprising given how many hotspots were located at highways where snow/ice can make them significantly more difficult to drive. One possible conclusion from this insight is that drivers who frequent these hotspots in the winter are better drivers (or at least drivers more familiar with winter conditions) than the statewide average, mitigating the increased risks of these highways in the wintertime.

Future Work

Given the number of vehicular crashes in Montana and the multitude of factors that contribute to these crashes, there is an almost never-ending amount of research that could be performed into this data and the causes of these crashes. Within this section, we will first outline the ways in which we would improve our data mining methodology prior to performing further analysis, then will outline several different possible future research directions.

Methodology Improvements

Although we were able to successfully implement KDE and K-modes within this project, we have identified several areas where we would improve methodology within both KDE and K-modes prior to performing additional analysis. Within KDE, improving the resolution that the KDE and base map is generated at would allow us to analyze hotspots in greater detail. This includes making the density map appear more contoured (most likely by increasing the number of points within the array calculated using our KDE algorithm) and adjusting the bandwidth to examine smaller hotspots that exist within the boundaries of towns (which we removed from analysis within this project).

If a density estimate were calculated for each crash point in the dataset (in addition to the array mentioned above), it would also allow us to avoid using the Google Maps workaround to query crashes within each hotspot. Instead of approximating a circle for each hotspot, we could simply query points that had density estimates above a certain threshold, then perform spatial k-means clustering or some other simple analysis to split the queried points into the identified hotspots.

By generating a density estimate for each crash point we could also employ other data mining/analysis techniques outside of K-modes. One possible method is to attempt linear regression, with density estimates serving as the response (Y) variable and other attributes serving as possible explanatory (X) variables. We could also encode categorical attributes into a series of arrays (using functions such as `OneHotEncoder` from Python), then perform k-means or other more traditional forms of clustering on the encoded crash points.

Future Research Directions

There is a large amount of future analysis that could be performed on crashes within Montana. First and foremost, the same analysis we employed in this project could be applied to the other 53 counties in Montana (or any other county or administrative region throughout the US). While performing this analysis, it would also be beneficial to attempt clustering using combinations that contain the other attributes within the Montana vehicular crash dataset.

Although examining the data using KDE at a county level yielded interesting insights into rural hotspots, if a smaller area were examined (i.e. a larger map scale), hotspots could be identified *within* towns and cities, yielding insight into both what attributes contribute to crashes within cities and how these crash attributes *differ* between urban and rural areas.

Within our analysis, we examined all severities of crashes (everything from fender benders to serious injury or death). Examining *just* crashes that resulted in serious injury or death may yield different combinations of attributes. Although it is beneficial to prevent all types of crashes, efforts by MDT (as outlined in VisionZero) are currently focused on preventing primarily serious injury or fatal crashes. Examining crashes that just fit this criteria may prove to be more useful and relevant for the state's current efforts.

Finally, there are a wide variety of factors that may contribute to crashes outside of what is included in the Montana vehicular crash dataset. Specific environmental factors (temperature, the *amount* of rain, etc.) and roadway features (such as the presence of a shoulder, 2-lanes versus 4-lanes, speed limit, etc.) may also play a part in vehicular crashes. Data on all of these factors specific to Montana is publicly available through sources such as NOAA and MDT. Future research that examines the effect of these factors—in addition to the ones within the Montana vehicular crash dataset—may yield additional insights.

Sharing Disclosure

Saidur Rahman and Matthew Campbell both consent to this report being used in future data mining classes.

Works Cited

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots, *41*, 359–364. https://doi.org/10.1016/j.aap.2008.12.014

Hashimoto, S., Yoshiki, S., Saeki, R., Mimura, Y., Ando, R., & Nanba, S. (2016). Development and application of traffic accident density estimation models using kernel density estimation. *Journal of Traffic and Transportation Engineering (English Edition)*, *3*(3), 262–270. https://doi.org/10.1016/j.jtte.2016.01.005

Highway Traffic Safety Administration, N. (2008). *National Motor Vehicle Crash Causation Survey: Report to Congress*. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059

IIHS, & HLDI. (2017). Fatal Crash Totals. Retrieved December 1, 2018, from https://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/state-by-state-overview/2016

Kar, A. (2017). Using K-modes for clustering categorical data. Retrieved December 5, 2018, from https://analyticsdefined.com/using-k-modes-clustering-categorical-data/

Kaur, I., Luhach, A. K., & Pooja. (2017). Mining Of Road Accident Data Using K- Mode Clustering And Improved Apriori. *International Journal of Computer Science and Information Security (IJCSIS)*, *15*(4), 235–249.

Kumar, S., Semwal, V. S., Solanki, V. K., Tiwari, P., & Kalitin, D. (2017). A Conjoint Analysis of Road Accident Data using K-modes Clustering and sayesian Networks (Road Accident Analysis using clustering and classification). *PTI*, 4. https://doi.org/10.15439/2017R44

Montana Department of Transportation. (2017). *Montana Annual Report for Federal Fiscal Year 2017*. Helena. Retrieved from http://www.mdt.mt.gov/visionzero/plans/safetyprg.shtml

NHTSA. (2017). USDOT Releases 2016 Fatal Traffic Crash Data. Retrieved December 1, 2018, from https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data

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. https://doi.org/10.1016/j.sbspro.2011.07.020

Ragain & Cook, P. . (2017). 2018 Most Dangerous Intersections in Montana. Retrieved December 1, 2018, from https://www.lawmontana.com/montana-intersection-study/

Sabel, C. E., Kingham, S., Nicholson, P. A., & Bartie, P. (2005). Road Traffic Accident Simulation Modelling - A Kernel Estimation Approach. *The 17th Annual Colloquium of the Spatial Information Research Centre*, 1–5.

Xie, Z., & Yan, J. (2008). Computers , Environment and Urban Systems Kernel Density Estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, *32*(5), 396–406. https://doi.org/10.1016/j.compenvurbsys.2008.05.001