Applying the Colocation Quotient Index to Crash Severity Analyses

Pei-Fen Kuo*

Assistant Professor

Department of Geomatics

National Cheng Kung University, Taiwan

Tel. (886) - 06237-0876

Email: peifenkuo@gmail.com

Dominique Lord

Professor and A.P. and Florence Wiley Faculty Fellow

Zachry Dept. of Civil & Environmental Engineering

Texas A&M University, USA

November. 15, 2019

*Corresponding Author

ABSTRACT

Examining the spatial relationships among crashes of various severity levels is essential for gaining a better understanding of the severity distribution and potential contributing factors to collisions. However, relatively few scholars have focused on analyzing this type of data. Therefore, in this study, we utilized a new index, the colocation quotient, to measure the spatial associations among crashes of various severities that occurred in College Station, Texas. This new method has been widely used to define the colocation pattern of categorized data in various fields, but it has not yet been applied to crash severity data. According to our findings, (1) crashes tended to be at the same injury level as those of neighboring ones, which was most significant for fatal crashes and second most significant for non-injury crashes; (2) the colocation quotient matrix tended to be symmetrical in non-injury crashes versus injury crashes (minor injury, major injury, and fatal); and, (3) DWIs (driving while intoxicated) and hit-and runs did not show a strong pattern. These colocation quotient results could be helpful for predicting crash severity and by providing traffic engineers with more effective traffic safety measures.

Key words: colocation, crash severity, colocation quotient index, spatial correlation

1. Introduction

In order to improve traffic safety, we must define the spatial relationships among crashes of varying severity levels, which could lead to a better understanding of the severity distribution and potential contributing factors to crashes. However, most scholars who have studied this distribution have focused on predicting the number of crashes, or hotspot identification, rather than on severity analysis (Savolainen et al., 2011; Lord & Mannering, 2010). Furthermore, the majority of studies related to analyzing crash severity have employed advanced statistical models, such as the logit, the tobit, and ordered probit models (Kockelman and Kweon, 2002), multivariate Poisson–lognormal models (Park and Lord, 2007; Ye and Lord, 2014), Bayesian hierarchical analysis (Huang et al., 2008), and the gradient boosting data mining model (Zhang et al., 2018) to show the mathematical relationships among various severity levels and risk factors. However, relatively few scholars have examined the spatial dependence between crash severity levels (Chiou et al., 2013; Chiou & Fu, 2015; Castro et al., 2013; Anarkooli et al.., 2017; Liu et al.. 2019; Zeng, et al.., 2019). Chiou et al.. (2013), for example, have used aggregated crash data and macro models for this purpose. In other words, in the few existing studies, researchers have aggregated crash severity point data into polygon data and then conducted further

analyses of it instead of examining the "real" spatial relationships among crash points of varying severity levels (such as distance) and how they affect distribution in space. It must be noted that aggregating crash severity data for spatial analysis may affect the results depending on the extent of the study area (such as a county, zip code, or a specific intersection or segment of highway). Hence, we decided to conduct a study on disaggregated crash severity data in order to better understand severity patterns at the level of a highway or street network.

The new method used in this study, the colocation quotient (CLQ), can capture patterns by using original point data (Leslie and Kronenfeld, 2011). The method discussed here, also called co-occurrence, refers to the types of spatial associations (clustering, dispersion, or random tendencies) between two or more categories of a population. For example, if the CLQ of an "A" injury level[1] crash to a "B" level injury crash ($CLQ_{A \rightarrow B}$) is higher than one, an A-level crash is more likely to occur near a B-level type than other random severity level crashes. In addition, researchers use the CLQ to avoid the Modifiable Areal Unit Problem (MAUP), also known as the aggregation bias problem (Leslie and Kronenfeld, 2011; Wang et al.., 2017), which can have a significant impact on statistical hypothesis testing results (e.g., positive

[1]Often, crash severities are classified as follows: F for fatal injuries; Type A for incapacitating injuries; Type B for non-incapacitating injuries; Type C for possible injuries; and O for property damage only or PDO.

impacts to negative values) when various study scales or areas are used (Xu et al.., 2014). Because the CLQ is a point-based measurement of spatial phenomena, the researcher does not need to aggregate point data by districts. Furthermore, this method is more suitable for analyzing categorical data, such as crash severity (Leslie and Kronenfeld, 2011), rather than numerical data. It is mainly because crash severity data are grouped into five severity levels and recorded as categorical data as fatal (K), incapacitating injury (A), non-incapacitating injury (B), possible injury (C), property damage only (PDO or O).

Compared to the existing colocation index methodology, the CLQ provides more flexibly because it uses distance ranking and allows for an asymmetrical matrix (Leslie and Kronenfeld, 2011). Because distance rank is considered, there is no need for the researcher to obtain an accurate network distance, which can be particularly difficult to measure on a street network in a rural area or in a developing country without a detailed road map. In addition, because the CLQ matrix is asymmetrical, it can reveal correlations in various directions. For example, $CLQ_{C \rightarrow I}$ (e.g., how crashes are often clustered around an intersection) may not be equal to $CLQ_{I \rightarrow C}$ (e.g., how intersections cluster around crashes). Thus, this quotient can more accurately represent spatial patterns in the real world.

The CLQ works differently than other traditional indices that are commonly used to quantify spatial correlations, such as the bivariate Moran's I, the normalized cross-vario-gram, Ripley's k, Joint Statistics, and the Cross K function (Cliff and Ord, 1981; Vallejos 2008; Cromley et al.., 2014). While all measurements are somewhat helpful for conducting crash severity analyses, they all have significant limitations that prevent them from delineating a suitable index for identifying the appropriate relationships between crash severity levels (Leslie and Kronenfeld, 2011). For example, some methods discussed above, such as the Pearson correlation, Bivariate ordinary least square, and cross vario-gram can be employed to successfully analyze continuous rather than nominal variables, which are of primary concern in our study. Also, the Moran's I is more suitable to analyze autocorrelation one variable at the time. The Joint count statistics and Moran's I are area-based methods, which are applied in a polygon framework rather than as point data. As for the most similar measure to our proposed CLQ, the problems of the cross-k-function include (1) using a metrical distance instead of a topological distance to define its neighbor, (2) measuring the spatial association of two populations instead of one single population, and (3) the inability to control for population clusters (Leslie and Kronenfeld, 2011; Cromley et al.., 2014). Table 1 shows more details about the comparison between different methods. Only the CLQ can show the bidirectional relationship among

different crash severity data by using an asymmetric matrix, while the cross-k-function and joint statistics can only show a unidirectional relationship. Furthermore, the CLQ can solve the false positive related to the population cluster problem. Recently, Hu et al. (2018) applied both the global and local CLQ to define crash hotspots involving pedestrians and cyclists. Due to this pioneering work, researchers are able to determine how the colocation quotient can be used to define the relationships among crashes of different severities. Therefore, the purpose of this study was to utilize this method to define the colocation patterns of crashes of various severity levels on a street network.

Table 1 Comparisons of different spatial correlation methods

Methods	**Advantages**	**Disadvantages**
Pearson correlation and Bivariate ordinary least square	• Can estimate colocation of two or more variables	• For continuous rather than nominal variables • Area-based statistics (may have MAUP problem)
Moran's I	• Straightforward	• Most used in spatial autocorrelation for one variable • Area-based method which is applied in polygon framework rather than as point data
Joint Statistics	• Simple Estimator	• Cannot handle an asymmetric relationship • Areal-based method which is applied in polygon rather than point data
Cross-K-function	• Simple Estimator	• Uses metrical distance instead of topological neighborhood distance • Measures spatial association of two populations instead of one single population • Cannot control for the population clusters
CLQ	• Simple estimator • Easy interpretation • Bidirectional relationship (asymmetric matrix) • Deal with population cluster	• Datasets for a specific category with fewer than 10 observations might lower the statistical power of the CLQ test[2].

Note 2: Leslie and Kronenfeld (2011) have indicated that the statistical analysis of the comparison may become unreliable for a sample sizesmaller than 10. A possible reason may attribute to the CLQ calculation algorithm. As described in their paper, the CLQ is calculated in $O(n\log(n))$ time, where n is the number of points in the population. Each Monte Carlo simulation requires an $O(n)$ allocation of categories among the existing points but not for re-computing the nearest neighbors. Therefore, the overall computational efficiency is the larger value between $O(n\log(n))$ and $O(nm)$, where m is the number of simulations. In this case, if n is less than 10, and simulation time maybe reduce and unreliable.

2. Background

This section presents background information on previous analyses of the severity levels of crashes in general, highway safety, and studies that have used the colocation index.

2.1. Safety Studies Related to Crash Injury Levels (Traditional Relationship Methods)

This section describes recent studies that have attempted to established specific relationships among crash severity levels. For example, in a 2002 study, Kockelam and Kweon (2002) used an ordered probit model to define possible factors for estimating the severity of injuries after a crash. They found that the risk factors have more to do with the drivers, passengers, and vehicles than the spatial distribution of the crashes themselves. For example, they found that drivers of pickup trucks and sports cars on average are involved in a higher number of severe crashes than drivers of passenger cars, and in general, passengers are more severely injured than drivers. On the other hand, young and male drivers of new vehicles tend to be involved in less severe crashes and travel at lower speeds.

Park and Lord (2007) proposed a Poisson-lognormal model for predicting crash frequency for different crash severity levels. They separated the severity of crashes

into five levels: K (fatal), A (incapacitating injury), B (non-incapacitating injury), C (minor injury), and O (Property damage only crash). In addition to establishing optimal models for defining related factors associated with various severity levels of crashes, they uncovered the three noteworthy findings discussed below.

1) The predictive variables they used are similar in both their value/magnitude and direction (positive/negative) when associated with severe crashes, such as K and A levels. However, for minor injury crashes (B to O), their coefficient estimators are not consistent with the value and sign.

2) The above identified factors might be attributed to cofounding factors rather than real risk factors. For example, speed is a critical factor associated with fatal crashes, but since there was no speed limit data in their dataset, the lighting variable was chosen in their model because of the confounding effects.

3) Because the predictive variables might not be reliable, they also provided the covariance and correlation matrices of all severity levels. They found that crashes have higher correlation with those of the same injury level than with those of dissimilar injury levels.

Huang et al.. (2008) applied a Bayesian hierarchical model to define the relationship between crash severity and vehicle damage in Singapore. They found that

crashes that occurred during peak traffic times, in good lighting conditions, or that involved pedestrians tended to be less severe. However, when crashes occurred at night, at T/Y-shaped intersections, involved two-wheeled vehicles, or a youth or an elderly person, the driver or the person at fault tended to sustain more severe injuries. From this dataset, the researchers were able to define environmental risk factors such as vehicle type and driver characteristics. However, only a few geometric traits of roads were defined as risk factors in this crash severity analysis. For example, they discovered that horizontal curves and high speed traffic were highly correlated with fatal crashes.

A fairly new environmental variable, the presence of a red light running camera, was found to be correlated with higher crash frequencies. However, obviously, they were not the main cause. As a safety measure, these cameras are frequently placed at sites with very high numbers of traffic violations, crashes, and greater traffic flow. Along the same lines, several scholars have applied a similar logic and utilized advanced statistical models to examine how the factors and settings of crashes influence severity, such as if the crash occurs on a rural non-interstate roadway (Chen et al., 2016) versus on a rural two-lane road (Haghighi et al.., 2018) and if heavy trucks (Assemi & Hickman, 2018), and commercial trucks are involved (Zhang et al., 2018).

In an earlier study, Quddus (2008) incorporated spatial elements into crash severity modeling by proposing a spatial model to capture the relationship between crash correlation and heterogeneity. For this study, the author knew the value of choosing a spatial model that would be appropriate for the size of the study unit and area. For example, a classical spatial model would be more applicable for a large study unit, such as a state or county, while a Bayesian hierarchical model would be better suited for a small study unit, such as an intersection. Although Quddus successfully defined predictive variables for fatal, severe, and less severe crashes, it was unclear about how different severity crashes were linked with each other.

In summary, although several scholars have touched upon the issue of spatial dependency and its link to injury severity, the focus of these previous studies was about examining the relationship between environmental factors and crash severity, such as traffic characteristics and geometric configurations of roadways (Chiou et al., 2014; Chiou & Fu, 2015; Anarkooli et al.., 2017; Liu et al.. 2019; Zeng, et al.., 2019) rather than analyzing the direct spatial relationships among crashes of varying severity levels.

2.2. Application of the Colocation Index (Spatial Relationship Method)

As mentioned above, the CLQ has been widely applied in economics, ecology and other social science fields. For example, Leslie and Kronenfeld (2011) used the

quotient to measure everything from the spatial association between types of businesses in Phoenix to the health of trees on Barro Island, Colorado. For the first study, they separated the businesses into 16 categories, which included agricultural, manufacturing, wholesale, retail, transportation, and so on. According to the results, public administration buildings had the strongest CLQ, meaning that these public buildings were 14 times more likely to have another similar business close by compared to other random types of businesses. Agricultural, manufacturing, wholesale, transportation, and warehousing were found to have high location preferences with each other. In other words, they tended to be located close to one another. This kind of data could be used by urban planners, marketing executives, and land developers to create more effective clusters or campuses of related businesses.

The researchers then used the CLQ to categorize and analyze the health of trees in the forests of Barro Colorado Island in Panama. The eight health statuses, from the best (healthy) to worst (downed trees), were correlated to areas in the forest. Their results showed that same-category CLQs tended to be strong. Furthermore, leaning trees, broken trees, and dead fallen trees were strongly collocated with each other; however, standing dead trees were not collocated with any of the categories mentioned above. Hence, the authors determined that the deaths of standing trees

were most likely the result of problems specific to those trees (such as tree senescence or disease) instead of environmental factors (such as windthrow).

Wang and his colleagues (2017) also applied global and local CLQs to define the spatial relationship between crimes and public buildings in China. They investigated the crimes of robbery, burglary and motorcycle theft in order to analyze their colocations with schools, shops, and entertainment establishments (e.g., bars, restaurants, movie theaters, etc.). They found a strong colocation value between motorcycle thefts and entertainment venues. This information could be used by law enforcement agencies to target these specific locations to prevent this type of crime, or it could help managers realize that they must monitor their facilities (i.e. add video cameras). As mentioned above, Hu et al. (2018) applied these two CLQ indexes to identify hotspots for crashes involving pedestrians or bikes.

Based on the above examples, it is clear that CLQ could help researchers define spatial patterns, quantify the spatial interaction between variables, and identify possible contributing factors, especially when these relationships are not apparent on GIS or Map programs.

3. Methodology

The colocation quotient can be used to measure spatial associations among two or more categories of observations (type A points and type B points). The basic goal of

this study was to compare the observed percentage of level B crashes among level A's nearest neighbors (the numerator in Equation 1) to the expected percentage (the denominator in Equation 1). The equation for the colocation quotient is defined as follows:

$$CLQ_{A\to B} = \frac{C_{A\to B}/N_A}{N'_B/(N-1)} \quad (1)$$

where,

$CLQ_{A\to B}$ is the ratio of the observed to the expected percentages of point A that have B as its nearest neighbor (with the shortest distance from point A to point B);

$C_{A\to B}$ is the count of point A whose nearest neighbor is point B

N_A shows the population sizes of type A points;

N'_B is the population size of type B points or the population size of B minus 1 (if A=B); and,

N is the population size of the total points.

The numerator is the observed percentage of type B points that are the nearest neighbors of type A points, and the denominator estimates the expected percentage by chance. If the CLQ is higher than one, the results show that A point is the closest neighbor of B, which is much more significant than a random association.

Later, Cromley et al. (2014) proposed the local colocation quotient (LCLQ), an index that can be used to measure local spatial associations. The equation LCLQ is defined in Equation 2, shown below:

$$LCLQ_{A\to B} = \frac{C_{Ai\to B}}{N'_B/(N-1)}$$
$$C_{Ai\to B} = \sum_{j=1}^{N}\left(\frac{w_{ij} f_{ij}}{\sum_{j=1}^{N} w_{ij} f_{ij}}\right) \quad (2)$$
$$w_{ij} = \exp(-0.5\frac{d_{ij}^2}{d_{ib}^2})$$

where

$CLQ_{A\to B}$ is the ratio of the observed to the expected percentages of point A that have B as its nearest neighbor;

f_{ij}: if A's nearest neighbor is B, then f_{ij} is 1, otherwise f_{ij} is zero; and,

W_{ij} refers to the importance weight of object j to object A_i (Gaussian Kernel density here).

d_{ij} and d_{ib} refers to the distance of object i to object j and bandwidth distance.

It should be noted that we only used the global CLQ in this study, because the setting of the LCLQ's parameters (such as bandwidth and density function of weight) might have affected the results significantly. Also, we used the kernel map as our reference to define hotspots. It also has a good visual output as LCLQ.

4. Dataset and Descriptive Statistics

Our study area was College Station, a mid-sized college town located in Central Texas with a population of approximately 100,000 during the period of data collection between January 2005 and September 2010. However, it has recently increased. The crash data were provided by the College Station Police Department (CSPD). There were 14,710 crash reports containing data such as the location, date, time, and severity of the crash. The road network in College Station is relatively simple: there are two major urban arterial roads, Texas Avenue and University Drive, which intersect in the northern part of the city. There are no complex multi-level interchanges located inside the city limits. In addition, there is a bar area located to the north of the Texas A&M University campus. Figure 1 (a) shows a detailed roadmap of the city.

During the study period, there were 0.16% fatal crashes, 14.0% severe injury crashes, 56.1% less severe injury collisions, and 15.2% non-injury crashes (Table 1). Figure 1 shows the hotspot locations and concentration places for different crash injury levels. Figure 1b shows that most fatal crashes occurred near the bars and highway ramps (Hot spots are represented in red, and non-hot spots are represented in blue). Because the kernel density maps were generated for each crash injury level using the nature break method to classify data (each severity level has different break point), the different intensities in colors of each figure can only be compared for the

same severity levels (each figure, such as Figure 1 (a)), and not with each other.

According to Figure 1c, severe but not fatal crashes, were located near the main intersections of the arterial roads. Figure 1d shows that although minor injury crashes most frequently occurred near the main intersections, they were also more spread out over several nearby and adjacent intersections. Non-injury crashes were mostly concentrated near the bars (Figure 1e). Based on the above results, some spatial relationships are clear; however, as discussed below, the CLQ can provide additional information not readily apparent. In addition to crash severity, we also delineated the hotspots for crashes in relation to major traffic violations, such as Driving While Intoxicated (DWI) and hit-and-runs (HRs). As expected, these crashes were mostly concentrated within the bar area, but they also occurred to a lesser degree at major intersections (Figure 1f)..

Table 1

The crash percentages of various severity levels

	Numbers	Percentage
Fatal	23	0.16%
Major Injury	2060	14.00%
Minor Injury	8254	56.11%
Non-Injury	2233	15.18%
Hit and Run/Non-Injury	1630	11.08%
Hit and Run/Injury	42	0.29%
DWI/Major Injury	118	0.80%
DWI/Minor Severe Injury	350	2.38%
Total	14710	100%

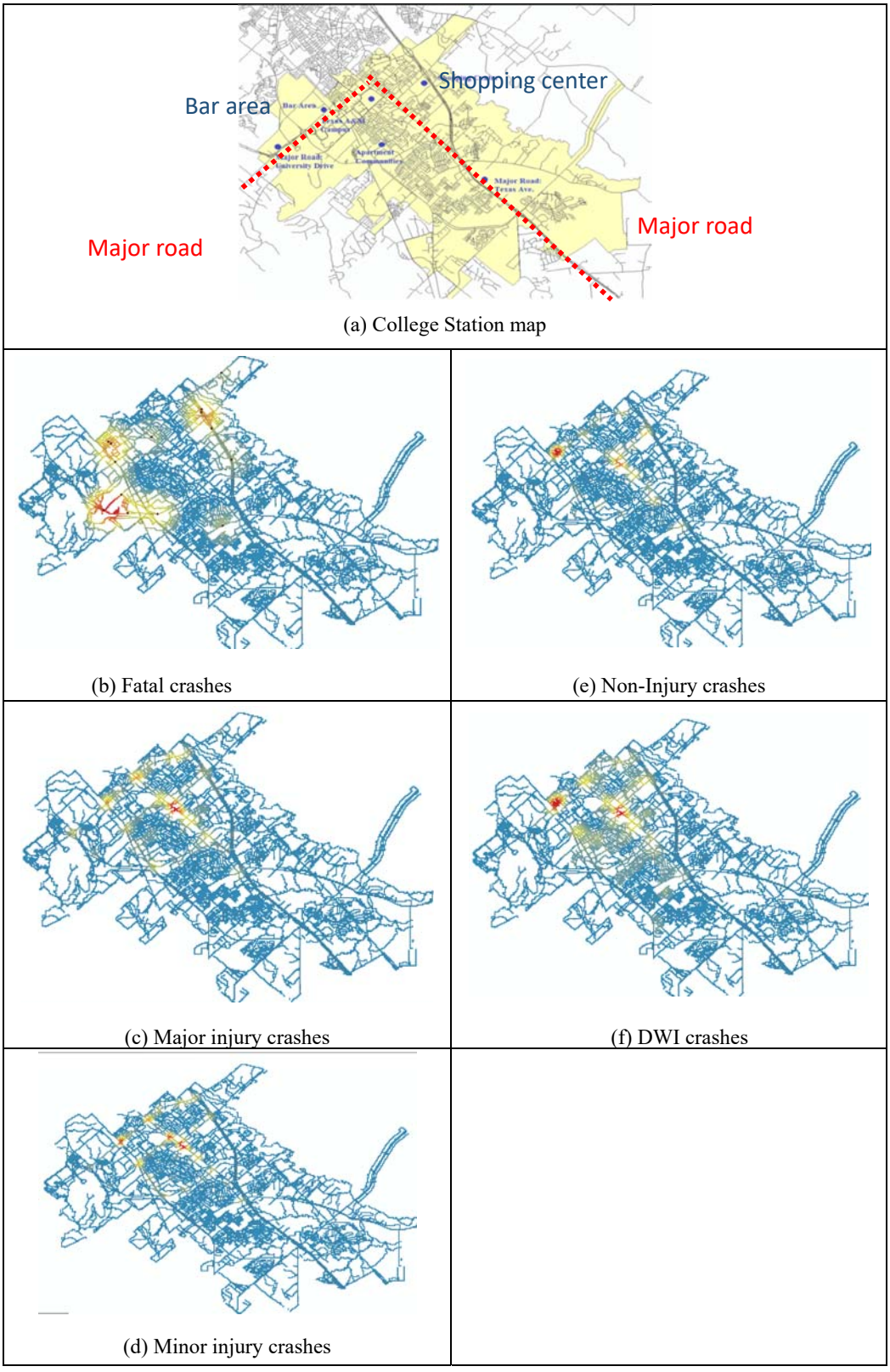

Figure 1. (a) a detailed road map of College Station, Texas and (b) to (f) the hotspot maps of different severity level crashes

5. Results and Discussion

As described above, we used the crash dataset for College Station from 2005 to 2010 in this study. Crash severity is categorized into eight different groups and four main levels of injury: Non-injury (NON), minor injury (MIN), major injury (MAJ), and fatal (FAT). Police officers also have four types of crashes related to violations of traffic laws, such as hit and runs (HIT/HIT INJ) and driving-while-intoxicated (DWI MIN INJ/ DWI MAJ INJ), a total of eight crash severity types. Table 2 shows the colocation quotients for all eight categories. Only statistically significant variables (with p-values less than 0.05) are featured in Table 2. It should be noted that the significance of CLQ here is calculated via a Monte Carlo simulation technique (to generate a sample distribution instead of making an assumption about the expected population distribution). Given this, the CLQ statistical test is estimated using 1,000 (or more) simulation runs. Hence, the results about the significance of the comparison is considered robust and are not overly dependent on the difference in sample sizes (as long as the sample size of the subgroup is 10 or greater). As mentioned above, the colocation quotient ($CLQ_{A \to B}$) provides the number of times an event type (type A) has its closest neighbor be of another specified type (Type B) than would be expected from a random distribution. If the colocation quotient is greater than one, this means

that point A is concentrated/clustered near point B. However, a colocation quotient of less than one means that point A is not correlated with point B.

Table 2

Colocation quotient values for different crash types and severity levels.

	NON	MIN	MAJ	FAT	HIT	HIT INJ	DWI MIN	DWI MAJ
NON	3.393	0.475	0.428		1.075	0.653	1.316	0.615
MIN	0.449	1.143	1.145		0.866	1.139		
MAJ	0.399	1.138	1.259		0.801		0.851	
FAT	0.713	1.011	1.397	9.039	0.805			
HIT		0.874	0.84		1.809		1.137	
HIT INJ	0.601							
DWI MIN	1.283	0.876	0.855		1.199		1.84	
DWI MAJ	0.483		1.301					

Note: A red value indicates a positive correlation, while a blank cell indicates an insignificant correlation.

In the same category, the colocation effect is strongest for fatal crashes (9.039). This means that a fatal crash is nine times more likely to have another fatal crash as its nearest neighbor. The colocation quotient effect is lowest for minor injury collisions (1.143), which means that this type of crash is just 1.14 times more likely to have another minor injury crash as its nearest neighbor.

When analyzing the different categories shown in the table above, most cells show a dispersed pattern between the two types based on their colocation quotient

values (17 cells are less than one), and only 11 cells have cluster colocation patterns (highlighted in red). Within these 11 colocation quotient values, the highest is only 1.397 (between FAT and MAJ type injuries), which means that the clustering effect between the two different type of crashes is relatively weak. The lowest colocation quotient is 0.399 for major injury crashes and non-injury crashes, which means that a major injury crash is only approximately 0.4 times more likely to have a non-injury crash as its nearest neighbor.

In summary, the key findings for this dataset are as follows:

a. For the same category, fatal crashes are mostly clustered together in the same areas, and non-injury crashes are the second-most clustered. In addition, although the colocation patterns of major injury and minor injury crashes are not as high, they are still significant. This result makes sense because police officers must make subjective decisions about the severity level of a crash and how to categorize it, meaning they are more likely to depend on cross-categorization in their reports. Furthermore, severity levels of crashes are also affected by driver characteristics, vehicle damage, and other possible factors that are unrelated to the network or road conditions. However, both fatal and non-injury crashes have strong colocation patterns with each other. The reason for this is because there are

more fatal collisions that occur in crash hotspots and fewer non-injury crashes are located in crash frequency hotspots. The former is affected by the speed limit while the latter is related to the density of road segments and intersections within a particular area. These results are consistent with previous studies (Chen et al., 2016; Hu et al., 2018). In addition, sample size might affect the colocation value. For example, if the sample size is too small, as in the case of hit-and run crashes that result in injuries, the results may be insignificant due to inflated p-values.

b. For the colocation between crashes with different severity levels, fatal crashes have a higher colocation with major injury collisions (1.397) than with minor injury and non-injury (1.011 and 0.713) crashes. Interestingly, the colocation factors between fatal crashes and other collisions that result in injuries are all greater than one; however, the correlation between fatal and non-injury crashes is less than one. In other words, the percentage of places associated with fatal crashes that have non-injury crashes as their nearest neighbors is only 70%, compared to the likelihood of having randomly distributed crashes of other severity levels. In contrast, the colocation quotients of non-injury crashes compared to other crash severity levels are all less than one. In sum, this suggests that 'safer' roads would result in less severe crashes, and 'riskier' roads would lead to a greater number of fatal crashes. In other words, common disturbance or

environmental factors that are correlated to injury crashes (minor and major crash), but not with PDO. The possible reason is that the traffic conditions, land use, and driver characteristics all affect crashes with different severity levels, but once again, our CLQ value can control the joint population clustering.

c. For crashes with traffic violations, the colocation quotient values of DWIs and hit-and-run crashes are both greater than one. The colocation quotient between DWIs and hit-and-run crashes is larger than one as well. A possible reason may be that people who drive under the influence are more likely to leave the scene of a collision than sober drivers in order to avoid being arrested. However, it might also suggest that drivers who tend to disobey traffic rules might live or travel in the same general areas.

d. For the mixture of crash severity and violations, the colocation quotient of non-injury crashes compared to other injury levels are all less than one except for hit-and-runs with no injuries and DWI crashes resulting in minor injuries. The possible reason for this finding is that the non-injury crashes are clustered around bars, which are also associated with DWI crashes. This pattern can be found in (d) and (f) of Figure 1. Several existing studies found a strong relationship between DWI crash and hit-and-run crashes, such as Solnick and Hemenway (1994, 1995), and MacLeod et al. (2012).

6. Summary and Conclusions

This paper has documented the application of the colocation quotient that can be used for estimating the spatial relationships among crash severity levels using disaggregated data. This method can show relationships that cannot be observed visually or through other correlation techniques, such as the bivariate Moran's I, normalized cross variogram, Ripley's k, and the Cross K function. For this reason, we applied the CLQ to a dataset collected in College Station, Texas between 2005 and 2010.

Several interesting results surfaced from this study: 1) crashes tend to be correlated with those of the same injury level as their nearest neighbors, which applies to fatal as well as for non-injury crashes; 2) the colocation quotient matrix tends to be symmetrical with non-injury crashes being separated from injury crashes (MIN, MAJ, and FAT); and, 3) there was not a correlation between DWI collisions and hit and run crashes but not a very strong pattern because of the small sample size, so most colocation correlations were removed due to high P-values. We found the same pattern for fatal crashes. With a larger dataset, we would probably observe more significant correlations. The effects of the sample size and randomness were addressed in details in Leslie and Kronenfeld (2011).

Based on the CLQ results, there are three potential applications that could be used in safety planning: (1) give higher priority to improve traffic safety for hotspots that are characterized by fatal crashes, because fatal crashes are mostly clustered together in the same areas; (2) separate all hotspots into crash fatality hotspots and non-fatal crash frequency hotspots, and then design more specific safety improvements tailored for each one separately; (3) provide more information about weighting social cost between major injury and minor injury crashes because their colocation patterns are relative low compared to fatal and PDO crashes (no clear difference between these two subgroups). Finally, since police officers tend to make subjective decisions about the severity level of a crash and how they are categorizing in the reports, the results of a CLQ analysis could be helpful in improving the classification of crash severity levels on a geographical basis.

In this analysis, we used the default definition of the colocation quotient calculation as the neighbor nearest the event point. We also utilized various definitions of neighbor, from 1st, 2nd and 5th nearest point, which yielded slightly different spatial pattern results that, although weak, maintain the same overall trends of those of nearest neighbor values. The result is same as previous studies, such as the one by Wang et al. (2017). In the future, scholars might explore different types of definitions of neighbor by considering changes in distance or non-adjacent points to

see if their results differ from those found in the present study. Although our CLQ matrix tended to be symmetrical, there were still some differences in the magnitudes of several cells. Therefore, future researchers may wish to examine if the direction of the CLQ of different crashes is an important factor to consider. For example, a DWI crash might be correlated with a hit-and run crash, but the relationship might not exist in reverse. It should be noted that because the CLQ is suitable for the colocation analysis of category data, future researchers might utilize it for understanding different types of collisions with regard to various driver characteristics, vehicles, or even traffic law violations. Another potential research topic is to examine the transferability of the CLQ method for crash severity analysis for different countries or geographical areas. Although the spatial pattern of different crash severity point data might change from place to place, especially for those with different traffic characteristics and road network, it would be interesting to examine if there are common colocation patterns between different datasets.

This research can be used to strengthen safety measures on the road. Since crash severity levels associated with extreme high or low values (fatal to non-injury) are strongly correlated spatially, this research can guide traffic safety officials to focus on similar severity levels of crashes because they are highly likely to recur in the same locations.

References

Anarkooli, A. J., Hosseinpour, M., & Kardar, A. (2017). Investigation of factors affecting the injury severity of single-vehicle rollover crashes: a random-effects generalized ordered probit model. Accident Analysis & Prevention, 106, 399-410.

Assemi, B., & Hickman, M. 2018. Relationship between heavy vehicle periodic inspections, crash contributing factors and crash severity. Transportation Research Part A: Policy and Practice, 113, 441-459.

Chen, C., Zhang, G., Huang, H., Wang, J., & Tarefder, R. A. 2016. Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. Accident Analysis & Prevention, 96, 79-87.

Chiou, Y. C., & Fu, C. 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accident Analysis & Prevention, 50, 73-82.

Chiou, Y. C., & Fu, C. 2015. Modeling crash frequency and severity with spatiotemporal dependence. Analytic Methods in Accident Research, 5, 43-58.

Cromley, R. G., Hanink, D. M., & Bentley, G. C. 2014. Geographically weighted colocation quotients: specification and application. The Professional Geographer, 66(1), 138-148.

Haghighi, N., Liu, X. C., Zhang, G., & Porter, R. J. 2018. Impact of roadway geometric features on crash severity on rural two-lane highways. Accident Analysis & Prevention, 111, 34-42.

Hu, Y., Zhang, Y., & Shelton, K. S. 2018. Where are the dangerous intersections for pedestrians and cyclists: A colocation-based approach. Transportation Research Part C: Emerging Technologies, 95, 431-441.

Huang, H., Chin, H. C., & Haque, M. M. 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis & Prevention, 40(1), 45-54.

Kockelman, K. M., & Kweon, Y. J. 2002 Driver injury severity: an application of ordered probit models. Accident Analysis & Prevention, 34(3), 313-321.

Leslie, T. F., & Kronenfeld, B. J. 2011 The Colocation Quotient: A New Measure of Spatial Association Between Categorical Subsets of Points. Geographical Analysis, 43(3), 306-326.

Liu, J., Hainen, A., Li, X., Nie, Q., & Nambisan, S. (2019). Pedestrian injury severity in motor vehicle crashes: an integrated spatio-temporal modeling approach. Accident Analysis & Prevention, 132, 105272.

Lord, D., & Mannering, F. 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation research part A: policy and practice, 44(5), 291-305.

MacLeod, K. E., Griswold, J. B., Arnold, L. S., & Ragland, D.R. (2012). Factors associated with hit-and-run pedestrian fatalities and driver identification. Accid Anal Prev, 45, 366-372.

Park, E., & Lord, D. 2007 Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record: Journal of the Transportation Research Board, (2019), 1-6.

Quddus, M. A. 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis & Prevention, 40(4), 1486-1497.

Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis & Prevention, 43(5), 1666-1676.

Solnick, S. J., & Hemenway, D. (1995). The hit-and-run in fatal pedestrian accidents: victims, circumstances and drivers. Accid Anal Prev, 27(5), 643-649

Solnick, S. J., & Hemenway, D. (1994). Hit the bottle and run: the role of alcohol in hit-and-run pedestrian fatalities. Journal of studies on alcohol, 55(6), 679-684.

Wang, F., Hu, Y., Wang, S., & Li, X. 2017. Local indicator of colocation quotient with a statistical significance test: examining spatial association of crime and facilities. The Professional Geographer, 69(1), 22-31.

Xu, P., Huang, H., Dong, N., & Abdel-Aty, M. 2014. Sensitivity analysis in the context of regional safety modeling: identifying and assessing the modifiable areal unit problem. Accident Analysis & Prevention, 70, 110-120.

Ye, F., & Lord, D. 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Analytic methods in accident research, 1, 72-85.

Zeng, Q., Gu, W., Zhang, X., Wen, H., Lee, J., & Hao, W. (2019). Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors. Accident Analysis & Prevention, 127, 87-95.

Zheng, Z., Lu, P., & Lantz, B. 2018. Commercial truck crash injury severity analysis using gradient boosting data mining model. Journal of safety research, 65, 115-124.