Detecting spatial hotspots in active transportation safety: comparing kernel density estimation and local indicators of spatial associations (LISA) methods

# 1.0 Introduction

In road safety research, the term hotspot refers to specific locations in the road network that experience an unusually high number of crashes (Cheng and Washington, 2005; Morency and Cloutier, 2006). Hotspots are also often referred to as hot zones, black spots, accident-prone locations, and high risk locations, among others (Cheng and Washington, 2005). Regardless of terminology, detection of these locations can provide a basis to diagnose specific issues and implement countermeasures to treat identified problems.

Traffic crash hotspots are a fundamentally spatial phenomenon, in that they are realizations of underlying spatial processes that drive abnormal occurrences of crashes within the road network. Specifically, elements of the road environment, vehicles and road users interact in space and time to contribute causally to a given crash that occurs in the system. Spatial patterns of crashes that have occurred within a broader region and time period are therefore realizations of underlying spatial processes, in which crash risk factors related to road environment, road user and vehicles are spatially heterogeneous. The areas in which there are frequently occurring crashes represent areas of high crash *burden,* whereas *risky* areas are place where there are a higher probability of a crash occurring, relative to the number of opportunities for a crash to occur (i.e. road users interacting in that space) (Fuller and Morency, 2013). Measurements of the number of opportunities for a crash to occur is often referred to as “exposure” and are a fundamental data requirement in analysis of risk. Without exposure data, hotpots of crashes represent locations with a greater probability of a crash, but necessarily a higher risk for individual road users since they will not control for the amount of traffic.

Due to the geographic nature of traffic crashes, geographical approaches to identifying spatial clusters of events are well placed to detect traffic hotspots. Spatial analysis techniques such as kernel density estimation (KDE) and local indicators of spatial autocorrelation (LISA) are popular methods to detect hotspots of traffic crashes (Anderson, 2009; Benedek et al., 2016; Erdogan et al., 2008; Loo et al., 2011; Mohaymany et al., 2013; Nie et al., 2015; Pulugurtha et al., 2007; Thakali et al., 2015; Yu et al., 2014). These spatial methods are relatively simple, and are not as data intensive as gold-standard methods such as the Empirical Bayes (EB) or Full Bayes (FB) approaches (Lan and Persaud, 2011). Studies that seek to identify active transportation crashes using EB or FB approaches are rarely conducted (Kamel et al., 2020), making geographic methods an attractive option to researchers and practitioners seeking to improve active transportation safety.

The overarching goal of this research is to compare KDE and LISA approaches to identifying hotpots for active transportation research. To meet our goal, we have two objectives: (i) review commonly used methods for detecting spatial hotspots in road safety, including KDE and LISA approaches – without and with consideration of exposure data; and (ii) apply these techniques to a case study of bicycling safety data and compare the results.

# 2.0 Background: detecting spatial hotspots in Road safety

Geographical approaches to spatial hotspot detection in road safety can be delineated into two main categories: event-based and link-attribute-based (Loo and Yao, 2013; Yamada and Thill, 2010). In event-based analyses crashes are represented as points in geographic space and considered directly. In link-attribute based analyses crashes are aggregated to a spatial unit of analysis such as a census tract, or road segment, and spatial analyses are performed on the spatial unit of analysis and their attribute values.

## 2.1 Kernel Density Estimation

Kernel density estimators are an event-based method of hotspot detection where the spatial distribution of points is considered directly. Fundamentally, KDE’s convert discrete point data into a continuous surface that represents the intensity of the events. The form of the KDE in 2-D space is defined by:

(1)

Where, is the density at location , is bandwidth distance (search radius), is the distance from event to location , is the kernel function (typically a function of the ratio of to ) and is the attribute value at event . Essentially, each event that falls within the bandwidth distance , is weighted based on distance to location through a kernel function . The weights are then summed to determine .

The type of kernel defines precisely how each event within a bandwidth distance is weighted and comes in different forms such as Gaussian, Quartic, Conic, or Epanichnekov. In our case study we will use the Quartic kernel as it is computationally less burdensome approximation of the Gaussian kernel (Nelson and Boots, 2008; Xie and Yan, 2008). The form the KDE with a quartic kernel is given by:

(2)

The 2-D KDE has also been extended to 1-D linear network space, specifically to obtain density estimates for events that occur within network space (Okabe et al., 2009; Xie and Yan, 2008). A network-based KDE estimate is given by the following form (Xie and Yan, 2008):

(3)

Where, and are based on network distances rather than Euclidean distances. Using the quartic kernel, the form of the network-based KDE is:

(4)

For both planar and network versions of a KDE, the main parameter affecting output is the bandwidth , as the choice of kernel has been found to have very little impact on outputs (Nelson and Boots, 2008; Xie and Yan, 2008). Previous research has indicated that for both planar and network KDE’s, larger values of will result in fewer identified hotspots that each larger in size (Nelson and Boots, 2008; Xie and Yan, 2008). The appropriate size of the bandwidth is context specific and can depend on the size of the study area, the spatial dependence structure in the data, and the scale of the analysis (Nelson and Boots, 2008; Xie and Yan, 2008).

### 2.1.1 Integrating exposure and identifying hotspots

The use of KDE’s in traffic hotspot identification rarely integrates information on exposure, and therefore only highlights hotspots of greatest crash burden. To identify hotspots that represent greatest risk of a crash, information on the amount of traffic that occurs (exposure) is needed. Historically, spatially representative exposure data for active transportation users are difficult to obtain. In recent years, with the advent of fitness apps and smart phone technology, spatially continuous estimates of exposure are increasingly feasible to estimate across an entire road network (El Esawey et al., 2015; Roy et al., 2019; Strauss et al., 2015). Given spatially representative exposure counts are available for a given region, estimates of the density of crash risk can be obtained through a simple ratio:

(5)

Where, at a given location , is the density of crash risk, is the density of crashes, is the density of exposure and is a small constant to prevent instabilities through dividing by zero.

Once a KDE surface is created the next step is to identify the hotspots. While there is no widely accepted method for defining a hotspot, in most applications of KDE’s in road safety an arbitrary threshold is set such that values above that threshold are considered within a hotspot and values below are outside of hotspot (Anderson, 2009; Erdogan et al., 2008; Harirforoush and Bellalite, 2016; Pulugurtha et al., 2007; Thakali et al., 2015).

## 2.2 Local Indicators of Spatial Association

Spatial autocorrelation refers to the idea that near things are more related than things that are further away. Positive autocorrelation exists when there are there are similar values close to each other, while negative autocorrelation exists when values are dissimilar close to each other. The degree to which events are clustered in geographic space can be quantified using measures of spatial autocorrelation (Nelson and Boots, 2008; Yamada and Thill, 2010). Global measures of spatial autocorrelation assess the average level of spatial autocorrelation within a study area with a single statistic, while local measures assess the spatial autocorrelation for an individual spatial unit. Local measures can therefore be used to detect unusually high spatial autocorrelation at a local level (i.e a hotspot) (Nelson and Boots, 2008; Yamada and Thill, 2010).

Local measures of spatial autocorrelation are a link-attribute based analyses, where the events themselves are not directly analysed, rather, they are aggregated to a spatial unit of analysis and spatial analyses are conducted on the spatial units and the aggregated event values. A common local measure of spatial autocorrelation is the local Moran’s I statistic (Moons et al., 2009; Yamada and Thill, 2010). Local Moran’s I is given by the following:

(6)

Where, for spatial unit is the attribute value, is the mean attribute value over all , is the sample variance of the ’s, is the attribute value at neighbouring spatial unit and is the definition of spatial relationships amongst (e.g. which spatial units are neighbours with another). Positive values of indicates positive spatial autocorrelation for values higher or lower than the mean, while negative values indicate negative spatial autocorrelation for values higher or lower than the mean.

Depending on the spatial unit of analysis, spatial relationships () can be defined through Euclidian distances, number of nearest neighbours, or different adjacency metrics (Nelson and Boots, 2008). While Moran’s I (and other local measures) are more typically conducted when the spatial unit of analyses are areal units (polygons such as census tracts, or traffic analysis zones), they can easily be adapted to when the spatial units are network segments (such as road sections) by modifying to a linear network context (Yamada and Thill, 2010). In a linear network, spatial relationships may be defined by whether segment links share a common node, or are within a specified *network* distance of one another (Yamada and Thill, 2010).

Once the spatial relationships are defined and is computed for the each spatial unit , spatial hotspots can then be detected by combining information regarding the category of spatial autocorrelation with statistical testing (Nelson and Boots, 2008; Waller and Gotway, 2004). Categories of spatial autocorrelation can be delineated for each unit through a Moran’s scatterplot which compares a spatial units attribute value to a standardized average of its neighbours attribute values (Nelson and Boots, 2008). The types of spatial autocorrelation that can be identified include: high values surrounded by high values (high-high), high values surrounded by low values (high - low), low values surrounded by high values (low-high) and low-values surrounded by low values (low-low) (Nelson and Boots, 2008). Positive autocorrelation is represented by the categories high-high and low-low, while negative autocorrelation is represented by high-low and low-high.

Significance testing can then be used to identify which of these potential hotspots are significant. Due to the fact that the exact distribution of the local Moran’s I is not known, a randomization approach is often used to develop pseudo p-values to assess significance (Waller and Gotway, 2004; Yamada and Thill, 2010). Spatial units that are identified to have both high-high spatial autocorrelation and a significant value of can be considered as part of a spatial hotspot (Nelson and Boots, 2008).

### 2.2.1 Integrating exposure

To assess hotspots of crash risk we must be able to integrate exposure data. Given exposure data are available aggregated to the same spatial unit of analysis as crash data, one method would be to simply replace raw counts with an incidence rate. A more sophisticated method is suggested by Waller and Gotway (2004) which adapts the local Moran’s I equation based on a constant risk hypothesis. The constant risk hypothesis refers to the null hypothesis that location has no influence on risk of a crash, that any spatial patterns of counts are the result of differences in underlying exposure to risk (e.g. population size or traffic exposure) (Waller and Gotway, 2004). The better comparator of differences in risk between spatial units is the expected counts in that spatial unit, given the size of the underlying population, rather than average count over all units. Local Moran’s I for the constant risk hypothesis is given by:

(7)

Where, is the exposure data for spatial unit , is the event rate of the study area (total crashes/total exposure). The resulting value is interpreted in the same manner as before, but represents spatial autocorrelation in risk of an event at spatial unit . Hotspot detection also follows the same steps as before, but with slightly different approaches in significance testing through randomization (Besag and Newell, 1991; Waller and Gotway, 2004).

# 3.0 Case Study: Analysis of crowdsourced bicycling exposure and crash data in victoria, bC

For this case study we implement the kernel density and local measure of spatial association methods outlined above using crowdsourced data from Bikemaps.org and Strava Metro for bicycling incidents (crashes and near misses) and exposure, respectively.

## 3.1 Study area and data

While we conducted our analysis on the contiguous municipalities of Victoria, Saanich, Oak Bay and Esquimalt, we focus the results on the city of Victoria exclusively (Figure 1). This approach reduced the impact of edge effects on the results by considering outlying areas, as well as increased the clarity of differences in approaches by concentrating on a smaller geographic area.

A close up of a logo

Description automatically generated

Figure : Study area. Results are focused on the City of Victoria.

We use bicycling incident data from a crowdsourced website called BikeMaps.org. Bikemaps.org was launched in October 2014 and is a global web-mapping application that enables the collection of incidents bicyclist experience in traffic, including collisions, falls, and near misses (Nelson et al., 2015). Citizens are able to pin the location of a given incident by navigating to the location at which the incident occurred on a web map through a mobile application or web-browser (Ferster et al., 2018; Nelson et al., 2015).

Exposure data were collected from a popular GPS enabled fitness app: Strava. Previous research has shown it to be a significant predictor of overall bicycling counts within the City of Victoria (Jestico et al., 2016). In this case study we drew upon Strava counts from January 2016 to September 2017 as a proxy for average annual daily bicycling for every road section in the study area. We did not use bias-corrected counts (e.g Jestico et al., 2017; Roy et al., 2019) as the purpose of this study was to differentiate identified hotpots between different methods, rather than to identify true hotspots. For each road section we calculate the average daily number of Strava recorded trips over the 21 months period. We assume there are no road sections with 0 bicyclists and add a small constant to each.

## 3.2 Methods

We compare hotspots through planar KDE’s, network KDE’s and Moran’s I. First, we apply these methods without integrating exposure data to compare hotspots on the frequency of bicycling incidents. Second, we integrate exposure data and compare hotspots on the *risk* of an incident. All analyses were conducted in R version 3.6.3 (R Core Team, 2020).

For computation of planar KDE’s we use the density.ppp function from the spatstat package (Baddeley et al., 2015). We use a quartic kernel and test different bandwidth sizes () of 50, 100, 150, and 250m (eq. 2). For computation of network KDE’s we developed a function in R based on the algorithm outlined in Xie and Yan (2008) We use a quartic kernel and test the same bandwidth sizes () of 50, 100, 150, and 250m, though these are now based on network distances (eq. 4). To develop network and planar KDE’s of incident risk, we first compute KDE estimates based on Strava data with the same parameters. For each location we divide the density estimate of an incident by the density estimate of exposure (eq. 5). Hotspots were identified as locations that had a density value greater than the 99th percentile.

For Moran’s I we identified hotspots using a contiguity-based definition of spatial relationships using custom functions developed in R based on the spdep package (Bivand and Wong, 2018). Each road section has two nodes at either end. For each road section, we considered a neighbouring road section to be those that have a common node. We compared different neighbourhood sizes by adding neighbours of neighbours to the definition (lags). We calculated Moran’s I for first, second and third order lags. To detect hotspots, we categorized each road section through a Moran’s scatterplot and assessed significance through a conditional randomization procedure with 999 permutations a critical value of 0.01. Road sections that were categorized as either high-high or high-low and had a pseudo p-value <0.01 were identified as hotspots.

## 3.3 Results

* Bandwidth an important parameter (obvious a-priori). Results reflect different spatial processes for a given scale
  + Larger bandwidth results in larger hotspots that likely reflect neighbourhood level processes such as land use, connectivity that lead to greater density
  + Smaller bandwidth results in smaller hotspots, more reflective of site specific conditions such as geometric design etc.
* LISA stats don’t seem to work that well? Seems like there are a lot of false positive’s

## A close up of a map Description automatically generated

Figure : Planar kernel density estimate hotspots based on 99th percentile values.

## A picture containing text, map Description automatically generated

Figure 3: Network kernel density estimate hotspots based on 99th percentile values.

A picture containing text, map

Description automatically generated

Figure : Moran’s I hotspots based on pseudo significance of <0.01.

A close up of a map

Description automatically generated

Figure : Planar kernel density estimate hotspots of risk based on 99th percentile values.

A picture containing text, map

Description automatically generated

Figure : Network KDE estimates of risk based on 99th percentile values.

A picture containing text, map

Description automatically generated

Figure : Moran's I hotspots based on pseudo significance of <0.01, adjusting for exposure.

## 4.0 Discussion

* Integrating exposure requires spatially comprehensive exposure data. Bias corrected fitness app counts are one of the only means of obtaining reasonable estimates of bicycling exposure data for an entire study area.

## 5.0 References

Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accid. Anal. Prev. 41, 359–364. doi:10.1016/j.aap.2008.12.014

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London.

Benedek, J., Ciobanu, S.M., Man, T.C., 2016. Hotspots and social background of urban traffic crashes: A case study in Cluj-Napoca (Romania). Accid. Anal. Prev. 87, 117–126. doi:10.1016/j.aap.2015.11.026

Besag, J., Newell, J., 1991. The Detection of Clusters in Rare Diseases. J. R. Stat. Soc. Ser. A (Statistics Soc. 154, 143. doi:10.2307/2982708

Bivand, R.S., Wong, D.W.S., 2018. Comparing implementations of global and local indicators of spatial association. Test 27, 716–748. doi:10.1007/s11749-018-0599-x

Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accid. Anal. Prev. 37, 870–881. doi:10.1016/j.aap.2005.04.015

El Esawey, M., Lim, C., Sayed, T., 2015. Development of a cycling data model: City of Vancouver case study. Can. J. Civ. Eng. 42, 1000–1010. doi:10.1139/cjce-2015-0065

Erdogan, S., Yilmaz, I., Baybura, T., Gullu, M., 2008. Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. Accid. Anal. Prev. 40, 174–181. doi:10.1016/j.aap.2007.05.004

Ferster, C.J., Nelson, T., Robertson, C., Feick, R., 2018. Current Themes in Volunteered Geographic Information, in: Comprehensive Geographic Information Systems. Elsevier, pp. 26–41. doi:10.1016/b978-0-12-409548-9.09620-2

Fuller, D., Morency, P., 2013. A population approach to transportation planning: Reducing exposure to motor-vehicles. J. Environ. Public Health 2013, 2–7. doi:10.1155/2013/916460

Götschi, T., Garrard, J., Giles-Corti, B., 2016. Cycling as a Part of Daily Life: A Review of Health Perspectives. Transp. Rev. 36, 45–71. doi:10.1080/01441647.2015.1057877

Harirforoush, H., Bellalite, L., 2016. A new integrated GIS-based analysis to detect hotspots: A case study of the city of Sherbrooke. Accid. Anal. Prev. 130, 62–74. doi:10.1016/j.aap.2016.08.015

Heinen, E., van Wee, B., Maat, K., 2010. Commuting by Bicycle: An Overview of the Literature. Transp. Rev. 30, 59–96. doi:10.1080/01441640903187001

Jestico, B., Nelson, T., Winters, M., 2016. Mapping ridership using crowdsourced cycling data. J. Transp. Geogr. 52, 90–97. doi:10.1016/j.jtrangeo.2016.03.006

Jestico, B., Nelson, T.A., Potter, J., Winters, M., 2017. Multiuse trail intersection safety analysis: A crowdsourced data perspective. Accid. Anal. Prev. 103, 65–71. doi:10.1016/j.aap.2017.03.024

Kamel, M.B., Sayed, T., Bigazzi, A., 2020. A composite zonal index for biking attractiveness and safety. Accid. Anal. Prev. 137. doi:10.1016/j.aap.2020.105439

Lan, B., Persaud, B., 2011. Fully Bayesian Approach to Investigate and Evaluate Ranking Criteria for Black Spot Identification. Transp. Res. Rec. J. Transp. Res. Board 2237, 117–125. doi:10.3141/2237-13

Loo, B.P.Y., Yao, S., 2013. The identification of traffic crash hot zones under the link-attribute and event-based approaches in a network-constrained environment. Comput. Environ. Urban Syst. 41, 249–261. doi:10.1016/j.compenvurbsys.2013.07.001

Loo, B.P.Y., Yao, S., Wu, J., 2011. Spatial point analysis of road crashes in Shanghai: A GIS-based network kernel density method. Proc. - 2011 19th Int. Conf. Geoinformatics, Geoinformatics 2011. doi:10.1109/GeoInformatics.2011.5980938

Mohaymany, A.S., Shahri, M., Mirbagheri, B., 2013. GIS-based method for detecting high-crash-risk road segments using network kernel density estimation. Geo-Spatial Inf. Sci. 16, 113–119. doi:10.1080/10095020.2013.766396

Moons, E., Brijs, T., Wets, G., 2009. Improving Moran’s Index to Identify Hot Spots in Traffic Safety, in: Murgante, B., Borruso, G., Lapucci, A. (Eds.), Geocomputation and Urban Planning. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 117–132. doi:10.1007/978-3-540-89930-3\_7

Morency, P., Cloutier, M.-S., 2006. From targeted “black spots” to area-wide pedestrian safety. Inj. Prev. 12, 360–4. doi:10.1136/ip.2006.013326

Nelson, T.A., Boots, B., 2008. Detecting spatial hot spots in landscape ecology. Ecography (Cop.). 31, 556–566. doi:10.1111/j.0906-7590.2008.05548.x

Nelson, T.A., Denouden, T., Jestico, B., Laberee, K., Winters, M., 2015. BikeMaps.org: A Global Tool for Collision and Near Miss Mapping. Front. Public Heal. 3, 1–8. doi:10.3389/fpubh.2015.00053

Nie, K., Wang, Z., Du, Q., Ren, F., Tian, Q., 2015. A Network-Constrained Integrated Method for Detecting Spatial Cluster and Risk Location of Traffic Crash: A Case Study from Wuhan, China. Sustainability 7, 2662–2677. doi:10.3390/su7032662

Okabe, A., Satoh, T., Sugihara, K., 2009. A kernel density estimation method for networks, its computational method and a GIS-based tool. Int. J. Geogr. Inf. Sci. 23, 7–32. doi:10.1080/13658810802475491

Pulugurtha, S.S., Krishnakumar, V.K., Nambisan, S.S., 2007. New methods to identify and rank high pedestrian crash zones: An illustration. Accid. Anal. Prev. 39, 800–811. doi:10.1016/j.aap.2006.12.001

R Core Team, 2020. R: A Language and Environment for Statistical Computing.

Roy, A., Nelson, T.A., Fotheringham, A.S., Winters, M., 2019. Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. Urban Sci. 3, 62. doi:10.3390/urbansci3020062

Strauss, J., Miranda-Moreno, L.F., Morency, P., 2015. Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. Accid. Anal. Prev. 83, 132–142. doi:10.1016/j.aap.2015.07.014

Thakali, L., Kwon, T.J., Fu, L., 2015. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. J. Mod. Transp. 23, 93–106. doi:10.1007/s40534-015-0068-0

Waller, L., Gotway, C., 2004. Spatial Clustering of Health Events: Regional Count Data, in: Applied Spatial Statistics for Public Health Data. John Wiley & Sons, Hoboken, pp. 200–271.

Willis, D.P., Manaugh, K., El-Geneidy, A., 2015. Cycling Under Influence: Summarizing the Influence of Perceptions, Attitudes, Habits, and Social Environments on Cycling for Transportation. Int. J. Sustain. Transp. 9, 565–579. doi:10.1080/15568318.2013.827285

Woodcock, J., Edwards, P., Tonne, C., Armstrong, B.G., Ashiru, O., Banister, D., Beevers, S., Chalabi, Z., Chowdhury, Z., Cohen, A., Franco, O.H., Haines, A., Hickman, R., Lindsay, G., Mittal, I., Mohan, D., Tiwari, G., Woodward, A., Roberts, I., 2009. Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. Lancet 374, 1930–1943. doi:10.1016/S0140-6736(09)61714-1

Xie, Z., Yan, J., 2008. Kernel Density Estimation of traffic accidents in a network space. Comput. Environ. Urban Syst. 32, 396–406. doi:10.1016/j.compenvurbsys.2008.05.001

Yamada, I., Thill, J.C., 2010. Local Indicators of Network-Constrained Clusters in Spatial Patterns Represented by a Link Attribute. Ann. Assoc. Am. Geogr. 100, 269–285. doi:10.1080/00045600903550337

Yu, H., Liu, P., Chen, J., Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accid. Anal. Prev. 66, 80–88. doi:10.1016/j.aap.2014.01.017

*Increasing active transportation (walking and bicycling), at the expense of motor vehicle use, would result in better public health outcomes and reduced emissions from motor vehicles (Götschi et al., 2016; Woodcock et al., 2009). Concerns about crashes are a main deterrent to increased active transportation use, especially for bicyclists (Heinen et al., 2010; Willis et al., 2015). To encourage more walking and bicycling and to reduce associated harms, it is critical that jurisdictions provide safe and comfortable road environments. A foundational step in providing safe and comfortable road environments is to use available data to understand where and when crashes are occurring within a jurisdiction.*

Spatial analysis tehcniques approaches are designed to analyse spatial data such as crashes, in order to answer questions regarding where and when they occur.

Traffic crashes are a spatially explicit event in that they occur at specific ti