A COMPARATIVE STUDY OF THE APPLICATION OF THE STANDARD KERNEL DENSITY ESTIMATION AND NETWORK KERNEL DENSITY ESTIMATION IN CRASH HOTSPOT IDENTIFICATION

Yue Tang  
Graduate Research Assistant  
Department of Civil and Environmental Engineering  
University of Massachusetts Amherst  
139 Marston Hall  
Amherst, MA 01003  
E-mail: yuet@engin.umass.edu

Michael A. Knodler, Jr., Ph.D.  
Associate Professor  
Department of Civil and Environmental Engineering  
University of Massachusetts Amherst  
216 Marston Hall  
Amherst, MA 01003  
Phone: 413.545.0228  
E-mail: mknodler@ecs.umass.edu

Mi-Hyun Park, Ph.D.  
Assistant Professor  
Department of Civil and Environmental Engineering  
University of Massachusetts Amherst  
16D Marston Hall  
Amherst, MA 01003  
Phone: 413.545.5390  
E-mail: mpark@ecs.umass.edu

ABSTRACT  
Despite a growing number of studies have claimed the network Kernel Density Estimation (network KDE) a more advanced method for crash hotspot identification than the planar Kernel Density Estimation (planar KDE), few conducted comprehensive study to examine their accuracy and practicality on a large-scale basis (i.e. municipal and county). This research attempted to fill the gap by conducting a comparative study of planar KDE and network KDE using the crash data of Hampden County, Massachusetts from 2009 to 2011. A two-tier planar KDE and a network KDE were implemented using the Kernel Density tool in ESRI ArcGIS 10 and SANET 4.1 developed at University of Tokyo respectively. Results showed that (1) Planar KDE is
computationally inexpensive and easily accessed. (2) Both methods yielded virtually similar hotspot patterns but with different rankings of the high crash locations. (3) In identifying specific hotspot locations, network KDE could achieve more accurate results and was more timesaving, although multiple runs of planar KDE identified specific locations as well. Accordingly, several suggestions were made for crash hotspot analysis: (1) Since KDE takes the interrelationship among crashes into consideration, it is a more statistically sound approach than traditional methods in crash hotspot identification and can be widely adopted by state and local agencies for initiating safety improvement projects. (2) Planar KDE is recommended to identify general hotspot patterns on large-scale basis for its practicality and efficiency. (3) Network KDE is recommended to identify specific intersections and roadway segments for accuracy.

1 INTRODUCTION AND BACKGROUND

In 2010, there were an estimated 5,419,000 police-reported traffic crashes, in which 32,885 people were killed and 2,239,000 people were injured; 3,847,000 crashes involved property damage only. The economic cost of traffic crashes in the same year was $230.6 billion (reported and unreported crashes) (Traffic safety facts 2010, 2012). Moreover, injuries resulting from motor vehicle crashes are the leading cause of death for age 4 and every age 11 through 27 (based on 2009 data). Thus, reduce deaths, injuries, and economic losses from motor vehicle crashes via roadway safety projects have long been the efforts of all parties.

Crash hotspot identification is the primary step of initiating safety improvement projects. Traditionally, high crash locations were identified when police reports piled up, citizens voiced their concerns, or government agencies sorted through the data. While these methods were generally effective, government agencies acted passively in conducting high crash location inspections, and more sufficient techniques were in need to identify the locations that require further examinations. With the aid of Geographical Information System (GIS), GIS users were able to visualize traffic accident data to identify hotspots on highways and urban networks. However, GIS software was often used inefficiently. Typical inefficient use was visualizing the physical locations of crash events by georeferencing tabular data and visually seeking for crash hotspots. Prior to the adoption of kernel density estimation (KDE), the development of spatial statistics and GIS had brought robust approaches to identify crash hotspots. For instance, repeatability analysis employing Poisson distribution was widely used to calculate crash rate (Erdogan, 2007). Similarly, Getis-Ord Gi* (Gi*) statistic looks at each feature within the context of neighboring features and allows each crash to be characterized by an attribute. As progressive as these statistical approaches are, however, they have been proved to be in insufficient in quantifying the geographical relationship between crashes and other environmental variables.

Estimating the density of points on a network is the most prevalent method nowadays in high crash location identification for it considers the relationship between crashes and other environmental variables (Okabe, 2009). In 2003, Flahaut et al. developed a planar KDE method on a 59 km long numbered Belgian road to determine the concentrations of accidents (Flahaut, 2003). Since this method estimated both network and the region it was embedded in, it was inaccurate for crash hotspot identification as crashes only happen on roadway network (Okabe, 2009). Yamada et al. compared accuracy of standard KDE and network KDE for traffic crash...
analysis and proved the accuracy of network KDE (Yamada, 2004). Christopher M. Monsere et al. compared different identification methodologies for top speeding-related crash locations and examined how these methodologies resulted in different priorities for safety improvements (Monsere, 2006). Xie developed a network KDE approach to estimate the density of spatial point events (Xie, 2008). Borruso and Porta et al. started dealing with density estimation on roadway networks with the aid of GIS (Borruso, Network density estimation: a GIS approach for analyzing point patterns in a network space, 2008). Kuo et al. have proved the advantages of using network KDE to analyze crash and crime data and demonstrated its application (Kuo, Guidelines for choosing hot-spot analysis tools based on data characteristics, network restrictions, and time distributions., 2011). In their study, Okabe et al. elaborated the computational method of a network KDE and developed its implementation SANET for ArcGIS (Okabe, 2009), which significantly increased the ease of use of network KDE.

In all the above studies, network KDE was proved theoretically more accurate than other statistical methods in crash hotspot identification. Nevertheless, two concerns are still left to be addressed. First, few existing studies have dealt with the practicality of implementing network KDE over planar KDE. Depending on circumstance, the priority of a safety analysis could be to identify the general hotspot pattern rather than result accuracy. This is especially true for hotspot analysis on large-scale basis. Second, the case studies in existing research often played an auxiliary role in testing the statistical soundness of network KDE over other statistical methods, thus they were usually on small geographical scales using simple data. For example, Okabe et al. introduced the network KDE and its application by conducting a case study on Kashiwa, Japan, which is a small city with an area of 114.9 sq km. The study contained only 35,235 links and 25,146 nodes, and the total length was 2190km. To address these two concerns, this study conducted a comparative study of planar KDE and network KDE on a larger-scale basis using the crash data of Hampden County, Massachusetts from 2009 to 2011.

1.1 RESEARCH OBJECTIVES
The overall objective of this study is to build upon the existing research to conduct a comparative study on planar and network KDE on large-scale basis (i.e. county level), thus to provide a decision support tool for state agencies and traffic engineers when targeting traffic improvement programs. Specifically, this study compares planar KDE and network KDE in the perspectives of practicality and accuracy.

2 METHODOLOGIES
A case study was done using the crash data in Hampden County, Massachusetts from 2009 to 2011. Hampden County is the most urban area in Western Massachusetts that has a total area of 1,642 sq km. The georeferenced crash data used for the analysis were obtained from various agencies through the UMassSafe Traffic Safety Data Warehouse, which is a multidisciplinary traffic safety research program housed in the University of Massachusetts Transportation Center at the University of Massachusetts Amherst. The Georeferenced data claimed an accuracy of approximately 85 percent. Crash data for the three years ranging from 2009 to 2011 were extracted and combined together to refine the analysis results. During 2009 and 2011, there were
13,710 crashes in total, which include fatal, non-fatal, and non-injury crashes. In the analysis, the roadway network contained 479,944 polylines and 13,710 points. Figure 1 shows the base map used in this study which contains three layers, i.e. Hampden County, crashes during 2009 and 2011, and roadway network.

Figure 1 Base Map of Crash Hotspot Analysis in Hampden County, MA, 2009-2011

In previous section we introduced the development of spatial analysis in crash hotspot identification. Kernel density calculates the density of point features around each output raster cell. Conceptually, a smoothly curved surface is fitted over each point. The surface value is highest at the location of the point and diminishes with increasing distance from the point, reaching zero at the assigned search radius distance from the point. The density at each output raster cell is calculated by adding the values of all the kernel surfaces where they overlay the raster cell center (Equation 1) (ArcGIS Resource Center Desktop 10: How Kernel Density Works.). The $\pi$ in the equation suggests its 2-D nature. Conventionally, this KDE method is called planar KDE to indicate its 2-D nature and distinguish from network KDE.

$$K(i) = \sum_{d<\tau} \frac{3}{\pi \tau^2} (1 - \frac{d^2}{\tau^2})^2$$

Where,
- $K$: Kernel density;
- $d$: The distance from event; and,
- $\tau$: Bandwidth

A planar KDE was implemented to the base map using the Kernel Density tool in ESRI ArcGIS 10. Due to the 2-D nature seen from Equation 1, the estimation identified the high crash areas across the entire county. To identify the specific high crash intersections or roadway segments, a second-tier estimation was implemented to the hotspots identified in the first-tier estimation respectively. To simplify the process, the three hotspots with highest kernel density
from the first-tier estimation were chosen for the second-tier estimation. Preparing for the second-tier estimation, the point data and polyline shapefiles of the three hotspots were clipped off from the base map.

However, since traffic crashes occur inside a roadway network that is usually a 1-D linear space, a planar KDE might not be the best approach to estimate the density of crashes. As an alternative, computing density on the network (network KDE) has been developed to address this issue for the network-constrained nature of some classes of point events such as crime occurrences or motor vehicle crashes (Produit, 2010). Recently, more and more studies have focused on network KDE, which is represented with basic linear units of equal network length, termed *lixel* (linear pixel), and related network topology (Xie, 2008). While a planar KDE aims to produce a smooth density surface of spatial point events over a 2-D geographic space, the network KDE can generate more accurate estimation for spatial point events such as traffic crashes. Xie utilized the network KDE approach to characterize the spatial patterns of crashes on roadways and recognized the limitations of applying standard 2-D planar KDE methods in a network space (Xie, 2008); Yamada *et al.* conducted a comparative study of planar KDE and network KDE based on Monte Carlo simulation using traffic crash data in the Buffalo, NY area in 1997. The results demonstrated the benefits of using a network KDE as the planar KDE entails a chance of over-detecting clustered patterns (Yamada, 2004). In a recent study, Kuo *et al.* have proposed to applying appropriate tools based on data characteristics and network restrictions. The network KDE is calculated as in Equation 2 (Kuo, Using Geographical Information Systems to effectively organize police patrol routes by grouping hot spots of crash and crime data, 2011). Compared to Equation 1, Equation 2 does not have the $\pi$, which indicates its 1-D nature. The selection of bandwidth is critical in determining computation accuracy and time. Smaller bandwidth selection yields sharper density curves while larger bandwidth yields smoother density curves, i.e. smaller bandwidth selection can achieve more accurate result but takes longer computation time, and vice versa. Thus, depending on the requirement for accuracy and user’s computation ability, a moderate bandwidth is critical for effective and efficient use of the network KDE.

\[
K(i) = \sum_{d < \tau} \frac{3}{\tau^2} (1 - \frac{d^2}{\tau^2})^2
\]

Where,
- $K$: Kernel density value;
- $d$: The distance from event; and,
- $\tau$: Bandwidth

In their recent studies, Okabe *et al.* extended the framework of network KDE method to three classes of kernel functions and explained each in detail. The three classes of kernel functions are the class of ‘similar’ shape kernel functions, the class of equal-split kernel functions, and the class of equal-split continuous kernel functions (Equation (3)-(5)). The class of ‘similar’ shape kernel functions was proved to be a biased estimator thus was not suitable for hotspot analysis. While the other two classes are both unbiased estimators, the class of
continuous kernel functions is slightly superior in accuracy. Nevertheless, the authors concluded the class of equal-split kernel functions is the most practical method due to its computational easiness (Okabe, 2009). Okabe also led a team at University of Tokyo and developed the GIS plug-in tool SANET. This study adopted the class of equal-split kernel functions in SANET 4.1 for network KDE.

\[
K_y(x) = k(x - y) \text{ for } |y| \geq h(l_1)
\]

\[
K_y(x) = \begin{cases} 
  \frac{k(d(x, y))}{(n_1 - 1)(n_2 - 1) \cdots (n_s - 1)} & \text{for } 0 \leq d(y, x) \leq h \\
  0, & \text{for } d(y, x) > h.
\end{cases}
\]

\[
K_y(x) = \begin{cases} 
  k(x) & \text{for } -h \leq x \leq 2d - h, \\
  k(x) - \frac{n-2}{n} k(2d - x) & \text{for } 2d - h \leq x \leq d \\
  \frac{2}{n} k(x) & \text{for } d \leq x \leq h.
\end{cases}
\]

3 RESULTS AND FINDINGS

Figure 2 shows the result of the first-tier planar KDE. The result yielded a general pattern of the crash hotspots that recognized the hotspot areas. The colors ranging from white to red represent kernel density from low to high. The figure shows that high crash areas concentrated on the urban area along the Connecticut River (colored in blue). The top three crash hotspots are circled in dark blue on the map, which are downtown Holyoke, Holyoke mall, and South End Bridge on the border of West Springfield and Springfield. The second-tier planar KDE was conducted to the three areas respectively. Figure 3 shows the KDE results of the three areas with roadway network displayed, which identified specific intersections or roadway segments. The specific hotspot locations with highest kernel density of the three areas were the intersection of Cabot Street and Main Street in Downtown Holyoke, the multi-lane divided roadway on Holyoke Street before the traffic-controlled intersection at the Holyoke Mall, and the middle of the westbound on the South End Bridge respectively.
Figure 2 Result of the first-Tier Planar KDE
Figure 3 Results of the Second-Tier Planar KDE
Figure 4 shows the result of the network KDE. The 1-D nature of network KDE assigned a k-value to each roadway segment rather than a planar area. The colors ranging from yellow to purple represent the kernel density from low to high. Compared to the results from the first-tier planar KDE, the network KDE yielded similar general patterns in that the hotspots mainly concentrated on the urban areas along the Connecticut River. However, the network KDE yielded different rankings of top hotspots. For instance, the top 3 locations with highest kernel density were all at the overpass on Interstate 91 northwest of the Holyoke Mall, rather than the three locations identified in Figure 3.

In terms of ease of use, planar KDE is relatively superior for it is embedded in ESRI ArcGIS software, while the network KDE is realized using SANET 4.1 that requires a license key from the SANET team at University of Tokyo. The network KDE is computationally expensive as it involves with more calculations, thus the implementation time is significantly longer than a single run of planar KDE. When the goal is to identify the specific high crash locations, however, implementing network KDE is in fact more timesaving with the preparation time for the second-tier planar KDE counted in.
4 CONCLUSION
Compare to the traditional “pin” map method, both planar and network KDE are more effective in high crash location identification on larger scales (i.e. county level). Comparing these two KDE methods, both are effective in identifying the general pattern of crash hotspots. Planar KDE is a more efficient method for its availability to users and timesaving nature. Nevertheless, network KDE is proven to be more accurate and timesaving with moderate bandwidth selection when the goal is to identify the specific roadway segments on large-scale basis.

5 FUTURE STUDY
Aside from the geographical relationships between crashes and other environmental variables, crash severity also needs to be considered to identify the roadway segments that cause the most loss. Thus, the next step of this research is to develop a severity-based network KDE method by extending the current KDE function. A potential method is to assign a weight of crash severity to each crash event. A popular standard in use is to give weights of 12, 3, and 1 to fatal, injury, and property damage only crashes respectively.

BIBLIOGRAPHY


