A Comprehensive Framework to Identify and Classify Traffic Accident Hotspots and Detect Contributing Risk Factors to the Formation of Hotspots

Zaniar Babaei

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Civil Engineering

Eastern Mediterranean University
July 2023
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy in Civil Engineering.

Assoc. Prof. Dr. Eriş Uygar
Chair, Department of Civil Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Doctor of Philosophy in Civil Engineering.

Assoc. Prof. Dr. Mehmet Metin Kunt
Supervisor

Examining Committee

1. Prof. Dr. Atakan Aksoy
2. Prof. Dr. Mustafa Ergil
3. Prof. Dr. Murat Karacasu
4. Assoc. Prof. Dr. Mehmet Metin Kunt
5. Asst. Prof. Dr. Hüseyin Sevay

ABSTRACT

Identifying roads' hazardous locations and solving their problems are the key measures in traffic safety management. However, since the traditional hotspot identification (HSID) rests on the yearly-aggregated crashes, two problems appear: the locations that become unsafe at specific short periods may remain unidentified as they may not show noticeable crash counts, and the results of the problem diagnosis analysis on hotspots' crashes potentially contain a great amount of uncertainty. Even though researchers have recently added the dimension of time and analyzed accidents spatio-temporally to obtain more insights, the mentioned problems have not been addressed fully. Hence, this study first suggests a new linear DBSCAN-based HSID method and demonstrates its acceptable performance by comparison with KDE+, the well-known clustering technique; second, employing the proposed technique, the study presents an algorithm for the spatial analysis of accidents through diverse time dimensions, which categorizes the risky locations based on their periodic reappearance. The tempo-categorization purpose is to enhance diagnosing causative risks by understanding their arising periods. The algorithm is tested using Allegheny highways crash data from 2014 to 2019. Results illustrate the contribution of the suggested method to the problem diagnosis and for detecting hidden unsafe points.

Keywords: traffic accidents, hotspot identification, DBSCAN clustering, spatio-temporal analysis, KDE+, safety problem diagnosis.

ÖZ

Yolların tehlikeli bölgelerini belirlemek ve sorunlarını çözmek, trafik güvenliği yönetiminde önemli önlemlerdendir. Ancak, geleneksel tehlike noktası tanımlama (HSID) yıllık toplam kazalara dayandığı için, iki sorun ortaya çıkar: belirli kısa dönemler için güvensiz hale gelen yerlerde, fark edilir kaza sayıları gösterilemeyebileceğinden tanınmayabilir ve tehlike noktalarının çözüm teşhisi analizinin sonuçları potansiyel olarak büyük bir belirsizlik içerebilir. Son zamanlarda araştırmacılar, zaman boyutunu ekleyerek kazaları mekansal ve zamansal olarak analiz etmiş olsalar da, bahsi geçen sorunlar tam olarak ele alınmamıştır. Bu nedenle, bu çalışma önce yeni bir lineer DBSCAN tabanlı HSID yöntemi önermekte ve iyi bilinen kümeleme tekniği KDE+ ile karşılaştırarak kabul edilebilir performansını ölçmekfedir. İkinci olarak, önerilen teknik kullanılarak, çeşitli zaman boyutları üzerinden kazaların mekansal analizini gerçekleştiren bir algoritmayu sunmaktadir. Bu algoritma, riskli bölgeleri periyodik tekrarlanmalarına göre kategorize eder. Tempo kategorizasyonu, neden olan risklerin ortaya çıkış dönemlerini anlayarak tanı koymayı arttırmak içindir. Algoritma, 2014-2019 Allegheny otoyolu kaza verileri kullanılarak test edilmiştir. Sonuçlar, önerilen yöntemin sorun teşhisi ve gizli tehlikeli noktaların tespiti konusundaki olumlu katkısını ortaya koymaktadir.

Anahtar Kelimeler: trafik kazaları, tehlike noktası tanımlama, DBSCAN kümeleme, mekansal-zamansal analiz, KDE+, güvenlik sorun teşhisi.

DEDICATION

This thesis work is dedicated to my parents who have always loved me unconditionally and encouraged me to pursue my dreams.

ACKNOWLEDGMENT

I would like to express my deepest appreciation to my supervisor, Assoc. Prof. Dr. Mehmet Metin Kunt for his guidance and consistent feedback throughout this project.

My appreciations also go to all staff members of the Civil Engineering Department for letting me be part of their reputable network and for providing me with a wonderful environment to study and work happily and confidently.

I would like to express my sincere gratitude to my dear parents, Habibollah Babaei and Fereshteh Hosseini, for all their unconditional support during my academic period.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

A_k	k^{th} Crash in A Cluster
C_j *(Lat)*	Latitude of the J^{th} Cluster Centre
C_j *(Lon)*	Longitude of the J^{th} Cluster Centre
C_x	Chord
d	Bandwidth
D(s)	Point-Event Density at Location *s*
d_x	Deflection Angle
e	Euler's Number
I (-d,d) (X)	Indicator Function of the (-D, D) Interval
k	Number of Crashes
L_i	Length of the i^{th} Segment
L_x	Network Distance
l	Length of Clusters
n	Crash Counts of Clusters
N_i	Total Crash Counts of the i^{th} Segment
r	Radius of the Searching Sphere (Referred to as Band-Width)
R	Radius
SD	Scaled Density of Clusters
t	Total Number of Crashes in a Cluster
w	Weight of an Object
x	distance from centre
α	Significance Level
$\mathcal{E}$	Epsilon, Radius of Searching

λ_i	Accident Mean Rate (Crash Counts Per 100 m) of the i^{th} Segment
1-D	One-Dimensional
2-D	Two-Dimensional
AADT	Annual Average Daily Traffic
AASHTO	American Association of State Highway and Transportation Officials
AV	Avenue
BB	Beta-Binomial
BL	Boulevard
CCO	Crash Count Only
D	Locations that Become Hotspot at Least in Daytime Off-Peak
DBSCAN	Density-Based Clustering of Application with Noise
DE	Daytime Off-Peak Emerging Hotspot
E	Locations that Become Hotspot at least at Evening
EB	Empirical Bayes
EE	Evening Emerging Hotspot
EX	Expressway
F	Locations that Become a Hotspot at least in Fall
FE	Fall Emerging Hotspot
GPS	Global Positioning System
HSID	Hotspot Identification
KDE	Kernel Density Estimation
M	Locations that Become Hotspot at least in the Morning
MCT	Method Consistency Test
ME	Morning Emerging Hotspot
MinPts	Minimum Points

N	Locations that Become a Hotspot at least at Night
NE	Night Emerging Hotspot
NKDE	Network Kernel Density Estimation
PKDE	Planar Kernel Density Estimation
RD	Road
S	Locations that become Hotspot in all Seasons
SCT	Site Consistency Test
Sp	Locations that Become a Hotspot at least in Spring
SpE	Spring Emerging Hotspot
SPF	Safety Performance Functions
Su	Locations that Become a Hotspot at least in Summer
SuE	Summer Emerging Hotspot
T	Locations that Become Hotspot in All Time Intervals of Day
TRDT	Total Rank Differences Test
TST	Total Score Test
U	Unbounded Hotspot
W	Locations that Become a Hotspot at least in Winter
WE	Winter Emerging Hotspot
Y	Yearly-Aggregated-Based Hotspot
YNP	Yearly Non-Periodic Hotspot

Chapter 1

INTRODUCTION

1.1 Background

Traffic accidents are a major public health and safety issue globally, with millions of people being injured or killed each year. There are approximately 1.35 million road traffic deaths annually, with an additional 20-50 million people being injured or disabled as a result of these accidents (WHO, 2018). Traffic accidents are the leading cause of the death of the 5-29-year age group (WHO, 2018). These accidents result in significant economic costs, both in terms of medical expenses and lost productivity.

There are several factors that contribute to traffic accidents, including driver behavior, vehicle design and maintenance, road design and maintenance, and environmental factors. Human factors, such as speeding, drunk driving, and distracted driving, are some of the most common causes of traffic accidents. In addition, vehicle design and maintenance can play a role in the likelihood of an accident, as well as the severity of the resulting injuries. Poor road design and maintenance can also increase the risk of accidents, particularly in developing countries where infrastructure may be lacking.

In recent years, there have been efforts to reduce the number of traffic accidents through various means, including the adoption of traffic safety laws and regulations, the improvement of infrastructure, and the promotion of safe driving behaviors. Many countries have implemented laws and regulations aimed at reducing the risk of

accidents, such as speed limits, seat belt and child restraint laws, and laws against drunk driving. Improving infrastructure, such as building better roads and improving public transportation systems, can also help to reduce the risk of accidents. Despite these efforts, traffic accidents continue to be a major public health and safety issue globally. Addressing this problem will require a multi-faceted approach that incorporates a range of strategies, including improving infrastructure, promoting safe driving behaviors, and investing in new technologies. The infrastructure improvement area is the one that this study focused on titled “traffic safety management”.

Traffic safety management is structured as a six-stage procedure (HSM, 2009). The first stage is “Network screening”. In this phase, the hazardous sections or points of road networks are identified by various approaches (as explained in section 2.2). Then, they are evaluated and ranked based on their overall criticality conditions (measured by multiple indexed) and prioritized for treatment. Following that phase, the “Diagnosis” phase is executed, which involves identifying the safety problems that make the locations unsafe. The diagnosis process is performed by field investigation (checking various items such as sight distance, geometric design elements, and markings and signs, etc.), reviewing the design and as-built drawings, and the analysis of the related crash data. Next, in the “Select countermeasures” phase, the appropriate safety improvement plans are selected or designed. The cost/benefit analysis is then conducted for the candidate countermeasures and compared one another to finalize the selected treatment in the “Economic appraisal” step. In the “Prioritize projects” phase, the economically justified improvements are evaluated at specific sites and across multiple sites to determine a group of improvement initiatives that can fulfill objectives like cost reduction, improved mobility, or minimized environmental impacts. The

projects are then implemented to correct those hazardous locations. The scale of the localization of engineering treatments and the weight of the actions may be limited to improvement of short-ranged environmental inferiorities at single or specific sites, or may be in larger scales like applying more general remedies at a road with relatively higher accident rate, or even in much greater scale, executing a range of treatments with a wide area coverage, and lastly, in the largest scale, as a massive action, implementation of known remedy in numerous sites on very large road networks that have similar accident problems (Szénási & Jankó, 2017) . The countermeasures implemented are one or a set of: engineering measures, which deal with road design, road equipment and road maintenance; Traffic control measures; and enforcement measures (e.g., speed control, training of road users, on-road checks).

As the final phase of the traffic safety management, "Safety effectiveness evaluation", a post-improvement study to evaluate the effect of treatment is required to assess the efficiency of the measures through a cost-benefit analysis. The findings gained in the hotspot safety works that may help in learning new lessons and generate new knowledge. In addition, as a result of systematic researches in this field, even the current standards (like geometry design standards) may turn out to need a revision (Kumar & Toshniwal, 2016; Montella, 2010; Benedek, et al., 2016).

The alternative traffic accident data to be analyzed in hotspot identification (HSID) might be crash frequency, fatal and injury crash frequency, or equivalent property damage only (Bandyopadhyaya & Mitra, 2015). Besides, the units to be analyzed may be road networks, either intersections or road segments, or both together, zonal geographical units like block groups, census tracts, traffic analysis zones, and ZIP areas (Xie, et al., 2017).

Among the aforementioned phases of traffic safety improvements, HSID (network screening) seems to be the dominant phase since the hotspots are naturally hard to be distinguished and usually, they are identified in subjective expertise attitudes, and on the other hand, the magnitude (number and extension) of hotspots affects the cost of the safety improvement actions, hugely. Therefore, in this study the major focus is on HSID.

1.2 Problem Statement

Traffic accidents are outcomes of complicated interactions between the non-locational (human-, vehicle- and environmental-related) factors and the location-dependent (road-related) factors (Edwards, 1998). Thus, identifying hazardous road sections or points and finding their causative factors (diagnosis) are key steps in traffic safety management (Hauer, et al., 2002; AASHTO, 2010). Locating hazardous road sections may be performed through two approaches: “Proactive” or “Reactive”. The proactive approach (non-accident based) is based on the road safety inspection in which the risk factors, which may potentially increase accident occurrence and severity are identified. That is an ordinary periodical verification of the characteristics and defects that require maintenance, which provides a quantification risk measure by calculating risk factor index for each segment (Ambros, et al., 2016). This approach is applied in numerous countries such as Austria, Ireland, Norway. The proactive road safety inspection should be carried out regularly in order to identify and address potential road hazards before they result in an actual accident. This is an advantage of proactive road safety assessment over the detection of traffic hotspots (Zahran, et al., 2019). The reactive approach for locating the unsafe road places is based on the analysis of the historical accident data, which is the approach around which this study revolves. The unsafe locations may be accident hotspots (also referred to as black spots or sites with

promise), which are places on roads with relatively high crash frequency compared with their similar sites due to risk factors related to road design, traffic control (Elvik, 2007), or local adverse environmental conditions. Alternatively, hazardous locations may not be necessarily hotspots if they don not experience abnormally great number of crashes.

1.2.1 Limitations of the Traditional HSID Regarding the Temporal Dimension

A potential problem related to the reactive approach is that, there may be certain hazardous areas not witnessing a significant number of accidents to be identified as high-risk zones. These locations are usually situated on roads with low traffic volumes. The traditional methods for identifying hazardous areas rely on the total number of accidents that occur within a year to pinpoint dangerous locations on roads. However, this approach overlooks the fact that, some areas may only pose a risk during certain time periods due to time-dependent factors such as weather, lighting, and traffic conditions. This study addresses this issue by exploring temporal patterns in accident occurrences that may be influenced by these factors. For instance, road sections with poor drainage or insufficient sunlight may become hazardous during rainfall or the freezing season, respectively. Temporary hazardous areas may also arise from uncommon situations that are specific to those locations and cannot be easily observed or recorded in crash data. For example, roads near woods that are infested with certain animals during certain seasons may cause distractions or loss of vehicle control for drivers, resulting in accidents. Identifying these hidden local risks that turn certain locations into hotspots may not be straightforward if hotspot identification is solely based on yearly data.

1.2.2 Limitations of the Existing HSID Tools Regarding the Spatial Dimension

Apart from the stated problem related to the temporal dimension of HSID approaches, the other identified limitation in the literature concerns the existing "tools and technique" used for detecting spatial patterns of crashes. The existing HSID techniques, including model-based approaches like the Empirical Bayes (EB) method, suffer from drawbacks such as resource-intensive requirements and limitations in identifying the exact extension of hazardous locations. Moreover, while the non-model-based methods like clustering techniques offer a promising alternative, they have not been extensively explored in the context of HSID. The current state-of-the-art non-model techniques, such as KDE, lack statistical measures to assess the significance of identified clusters and they are used for the area-based (2-D) clustering, overlooking the network-level exploration of hotspots. The recently introduced 1-D variants of KDE have solved the mentioned issues but they themselves show some limitations, as well, emphasizing the demand for a more modified method.

1.3 Research Objectives

To address the mentioned issue regarding the potential hotspot misidentification errors by the traditional HID approach, this study suggests executing HSID in narrower time segments, specifically the seasons and the different times of the day. The aim is differentiating risky locations with diverse lifecycles, referred to as "hotspot categorization" in this study, which could be advantageous as it may improve the problem diagnosis process and detect the potentially hidden unsafe locations on roads.

The other objective of this research is to establish a framework for the application of Density-Based Clustering of Application with Noise (DBSCAN) in the field of HSID. The objective includes developing a coded DBSCAN-based method (in Python) and

then, demonstrating the efficacy of the developed toll through a comparative analysis with other popular clustering methods, such as Kernel Density Estimation (KDE). By examining the potential advantages and performance of the developed tool in HSID, the study aims to contribute to a more comprehensive understanding of its applicability and suitability for identifying hazardous locations. The research will provide valuable insights to facilitate the wider adoption of DBSCAN in HSID and improve the accuracy and efficiency of safety improvement efforts.

1.4 Research Significance and Contribution

The research contributes to the field of traffic safety improvement by addressing gaps in previous studies related to identifying hazardous locations and diagnosing their causes. Previous spatio-temporal analyses were limited to 2-D, resulting in the inclusion of adjacent segments and difficulty pinpointing exact faulty points. This study introduces a 1-D spatio-temporal analysis approach that focuses on the reappearance of unsafe points at specific periods (e.g., a particular time of the day or a specific season) over multiple consecutive years. This approach enables the detection of hidden unsafe locations on specific roads and through understanding the problem-arising periods, it leads experts to more accurate problem diagnosis and an optimized resource allocation. The proposed DBSCAN-based 1-D HSID tool offers a new way of determining MinPts value and provides precise clustering performance without requiring additional data or specific thresholds. The developed approach and code are globally applicable and accessible free of charge, making it a valuable contribution to the field.

1.5 Research Plan

The objective of this research was accomplished by implementing the sequential study plans as the following:

1. Adapting the DBSCAN code to the planned road-by-road cluster analysis using Python programming language.
2. Acquiring a suitable crash data that satisfied the requisites of the developed model. Then, cleaning the data and prepare it for the analysis.
3. Obtaining the KDE+ tool and learn its usage instruction. Then, acquiring the needed shapefiles for executing the KDE+ in ArcGIS (ArcMap).
4. Segmenting the road network into homogeneous sections using Google Map. Then, repeating the segmentation task in the ArcMap environment, as needed for the comparison analysis.
5. Applying both competing methods, the DBSCAN-based tool and the KDE+, to the crash data. Then, compare their performance to validate the acceptable functioning of the developed tool.
6. Establishing the planned tempo-categorization algorithm (built on the developed DBSCAN-based tool), and then, applying it to the data.
7. Evaluating the final achievements of the proposed methodology.

1.6 Organization of the Thesis

This dissertation encompasses seven chapters. The first chapter contains a general introduction highlighting the scope of the research, statement of the problems, and suggested solutions for the stated problems.

Chapter 2 presents a summary of the conducted literature review to clarify the existing research landscape and identify its limitations and gaps.

In Chapter 3, in the first methodological part, the suggested DBSCAN-based HSID, is presented. Additionally, the competitor technique, the KDE+ tool, is overviewed.

The fourth chapter includes the second part of the methodology, the tempo-categorization procedure.

In Chapter 5, first, the case study used for validating the suggested methodology is described; then, results of the comparative study between the suggested DBSCAN-based HSID and the KDE+ tool are presented.

In Chapter 6, the significance of findings of the conducted research is discussed.

Finally in Chapter 7, a conclusion including a summary of the conducted research is presented.

Chapter 2

LITERATURE REVIEW

2.1 Spatio-temporal Analysis and Hazardous Locations categorization

The traditional HSID methods, which rely on yearly accumulated accident data, are unable to provide a comprehensive assessment of the temporal changes in hotspot characteristics. Therefore, the spatio-temporal analysis, which considers both the geographical position and timing of crashes, was introduced (Cheng & Lu, 2019). The spatial distribution refers to locational scattering in terms of geographical position (most often indicated by Global Positioning System, GPS), and temporal accumulation is about distribution based on time intervals. The Spatio-temporal analyses have been applied to general accidents (Pluga, et al., 2011; Dong, et al., 2016) or with a concentration on specific groups of road users such as pedestrians (Fox et al., 2015) or vulnerable road users (Ouni & Belloumi, 2018). From another viewpoint, spatio-temporal studies have been conducted to investigate the temporal variation in specific characteristics of the located hotspots, such as the drivers-related features of crashes of hotspots (Kaygisiz, et al., 2015) and the size of hotspots (Al Hamami & Matisziw, 2021).

Another feature of hotspots whose temporal trend has been studied to a greater extent of interest is the hotspot's stability (continuation extent). Most papers related to that topic completed their efforts by categorizing the identified hotspots to provide more

insights, as reviewed in this section. This proposed research falls within the scope of spatio-temporal analysis, but with a focus on a different characteristic: the continual reappearance of hotspots at certain periods. The scope of this study is refined as follows: firstly, it focuses on identifying hazardous points on road segments (not at intersections) at a linear or 1-dimensional spatial level (not regional or 2-dimensional); and secondly, it categorizes identified risky locations based on their reappearance periods (not crash density or continuity duration). These specifications are intended to connect the discovered temporal behavior of hotspots to the problem diagnosis domain. The divergences of previous studies from these specifications are also outlined at the end to highlight the novelty of this research. Therefore, the concluding sentences do not necessarily imply a limitation or drawback of those studies in all cases, but rather indicate the mismatch with the current study's objectives.

In their research, Cheng & Lu (2019) utilized the time-space cube method with a one-month time interval, along with spatial autocorrelation analysis, to demonstrate variations in the spatial distribution of crashes over time and to classify identified hotspots. However, the focus of their study was on intersections as the road element, and the study duration was restricted to one year, thus the results lacked validation of the identified hotspot locations and categories in subsequent years. Similarly, Wu, et al., (2021) employed the time-space cube methodology, albeit with a detailed process for selecting appropriate time and space intervals, and augmented with the cumulative frequency curve technique. They further grouped identified hotspots into various classes using the latent class analysis method to examine factors contributing to hotspot formation. Other spatio-temporal studies have also utilized the time-space cube technique, such as Hussain, et al., (2022) and Yoon & Lee (2021). However,

unlike the specific focus of this research, the spatial level in those studies (the last four) was two-dimensional, potentially covering multiple road segments, and all zones within the study area were analyzed simultaneously to detect crash clusters. Consequently, all locations were treated identically in terms of all features, whereas variations in road and traffic characteristics, such as road class and traffic volume, as well as facility density, do exist. That attitude seems biased when aiming to identify faulty facilities in areas with various exposures, especially in low-exposure zones. Another issue of those studies is that the categorization of the identified hotspots, performed by a toolbox in ArcGIS (emerging hotspot analysis), was based on two trends, the number of time intervals (months) the locations become a hotspot and the density of clusters, which differs from this study's categorization criterion (explained in Chapter 4).

Le, et al., (2019) utilized KDE method to examine the spatio-temporal patterns of crashes in a regional study. They segmented the time into daily and seasonal intervals and investigated the impact of crash severity on the patterns using the comap technique. The study categorized crash clusters based on crash density, similar to the aforementioned studies. For instance, the study indicated that winter had the highest number of "very high density" clusters when the severity weight was applied. However, 2-D level analyses such as those described do not suffice in pinpointing the precise location of hazardous points on each road, which is the objective of this study.

In contrast to the previously mentioned studies concerning hotspot categorization, Bíl, et al., (2019) focused on a linear HSID using KDE+ and limited the spatial dimension to one. They explored the consistency of hotspots over consecutive years and classified them into three groups: those that emerged recently, those that disappeared by the end

of the initial years, and those that remained stable throughout the entire study period. However, their categorization scheme was limited to these three levels, and they did not provide sub-categories for the stable hotspots. This research aims to address this gap by expanding the categorization scheme for stable hotspots.

To summarize, the described spatio-temporal papers mainly analyzed the temporal stability of hotspots by looking at their continuation over months or years. Moreover, those studies relied heavily on the visual assessment of the stability of hotspots, which may be imprecise. Furthermore, they did not provide a thorough analysis of the causal factors contributing to crashes in each hotspot category. This study aims to address those limitations by utilizing more precise methods to assess hotspot stability and exploring the causality of crashes in each hotspot category in greater detail.

In their study, Wang, et al., (2019) adopted a different temporal perspective by conducting a linear HSID through the Empirical Bayes (EB) method. That approach allowed them to analyze the effects of daily traffic variation on the spatial patterns and causation of crashes, leading them to demonstrate the dependence of hotspot locations on the time of day. Additionally, they revealed the limitation of the traditional year-wise HSID methods in identifying some hazardous sites. However, the study lacked a detailed discussion of the relation between hotspot categories and the diagnosis process, and the time segmentation was limited to daily intervals. Building upon their work, this study aims to expand the temporal scope by adding the season dimension to the categorization analysis; because the spatial patterns of crashes may be influenced by weather-related factors that vary seasonally as well. The analysis is through a clustering technique rather than the EB.

After conducting a thorough review of the literature, the objectives of this study regarding addressing the temporal-related limitations were determined as twofold: first, to identify hazardous locations with both high and low crash rates; and second, categorize these locations based on temporal factors in order to identify when safety risks may emerge. By categorizing these hazardous locations based on the time periods during which they exhibit higher crash rates, the study aims to improve the diagnosis procedure and help stakeholders implement more effective safety measures. The proposed approach includes considering multiple dimensions, such as space, time, and seasonality, to provide a comprehensive understanding of the spatio-temporal patterns of accidents.

2.2 Hotspot Identification Methods

In order to create the intended categorization algorithm, it was necessary to choose a suitable HSID method. The author understands that prioritizing identified hazardous sites for treatment is an essential step in managing safety improvements, which involves combining multiple risk measures (Bham, et al., 2017; Al-Ruzouq, et al., 2019; Afghari, et al., 2020; Wu, et al., 2021). Nonetheless, since this study's focus is restricted to the first stage of the traffic safety improvement, identifying hazardous locations, such a multi-criteria-based ranking is not considered here. Rather, the crash count only (CCO) risk index is used as the criterion in recognizing hotspots (Bandyopadhyaya & Mitra, 2015).

Two types of one-dimensional HSID approaches are available, which are categorized as model-based and non-model-based. The latter includes Naïve (e.g., crash frequency, crash rate), Sliding Window, and Clustering methods. The Naïve methods have demonstrated various limitations, including but not limited to the 'regression to the

mean bias', inadequate examination of crash dispersion, and the false assumption of a linear relationship between crash count and traffic volume (Hauer, 2005; Elvik, 2007; Afghari et al., 2020; Ghadi & Török, 2017). In the model-based methods, along with the potential influences of multi-collinearity among variables (Xu & Tao, 2018), splitting road segments into slices based on detailed geometrical attributes could result in too short sections with zero crash counts. Furthermore, setting arbitrary magnitudes as the range for identifying hotspots can hinder the efforts to identify only the unsafe portion of segments, leading to a misrepresentation of the size of the high-risk area (Ghadi & Török, 2017). It may cause misidentifying the causative factors or applying corrective measures to the safe parts in addition to the unsafe parts. The study suggests that certain risk factors may not impact a broad range of roads. Therefore, the terminology used to describe hazardous locations is altered, and the terms point, spot, or location are utilized instead of section or site. Additionally, partitioning roads into smaller segments may result in an error if a genuinely hazardous point falls on the boundary of two adjacent segments. This may lead to the split of the hazardous point, resulting in its failure to be detected as the two micro-segments may not meet the minimum density criterion (Ghadi & Török, 2017).

The state-of-the-art EB method (Hauer, et al., 1988; Montella, 2010; Bandyopadhyaya & Mitra, 2015; Cheng & Jia, 2015), which is a model-based HSID method, has demonstrated a comparatively high accuracy in detecting hotspots (Ghadi & Török, 2019). However, in addition to the drawbacks mentioned above for model-based HSID approaches, EB also require significant resources, including the measurement of multiple geometric parameters, the use of safety performance functions (SPF), crash modification factors, and calibration coefficients, which may not be readily available

in all countries. Additionally, when developing the prediction functions for EB, regression models are built based on data from road sections that are only similar in terms of geometric design and traffic volume, overlooking the potential variation of attributes such as vehicle composition and road users' characteristics across different roadways within the reference population. Dong et al. (2016) and Bandyopadhyaya & Mitra (2015) have also raised these concerns.

Given the aforementioned concerns, it is worth noting the significance of non-model-based methods that do not require intricate segmentation and entail lower expenses. In fact, clustering is that desired technique and, therefore, presents a promising approach for HSID analysis. Clustering is a technique in data analysis and unsupervised machine learning that involves grouping a set of objects or data points in such a way that, objects in the same group (called a cluster) are more similar to each other than to those in other groups or clusters. In other words, clustering is a process of dividing a dataset into groups or clusters based on similarities or patterns in the data. The goal of clustering is to identify inherent structures in the data and group similar objects together. One of the applications of clustering involves the identification of spatial patterns in distributed data for the purpose of identifying areas where data points are densely concentrated. Presently, Kernel Density Estimation (KDE) is considered one of the most commonly employed clustering techniques for analyzing HSID. To address a significant drawback of KDE, which is the absence of a statistical measure for assessing the significance of identified clusters, it is necessary to conduct an additional analysis techniques such as Getis–Ord Gi* (Gudes et al., 2017), Moran's (Al-Ruzouq et al., 2019; Cheng et al., 2018), point kernel density (Al-Ruzouq et al., 2019), and K-function (Ounio & Belloumi, 2018). In addition, as the conventional planar KDE

(PKDE) only calculates the planar distances between collision points, it is only suitable for performing area-based (2-D) clustering. Consequently, scholars have recommended an alternative version of KDE referred to as Network KDE (NKDE), such as KDE+ (Bíl, et al., 2013) and SANET (Xie & Yan, 2008), to overcome this limitation and measure distances linearly (1-D) to explore hotspots at the network-level (micro-level) rather than regional (macro-level). Density-Based Clustering of Application with Noise (DBSCAN) is a well-known clustering technique that differs from KDE in its approach to identifying clusters based on density rather than distance. Despite being widely used in other domains, DBSCAN has not commonly been employed for analyzing HSID due to certain perceived drawbacks. However, there has been little investigation into its performance in HSID when compared to other popular clustering methods such as KDE, and its potential advantages have not been fully explored. The application of DBSCAN in HSID has only been discussed briefly in a few papers, as outlined in the DBSCAN methodology section. Therefore, the primary objective of this study regarding addressing the spatial-related limitations is to establish a framework to facilitate the application of DBSCAN in HSID and to demonstrate its efficacy through a comparative analysis. This approach will allow a more comprehensive assessment of DBSCAN's potential in HSID, and potentially pave the way for its wider adoption in this domain.

2.3 Evaluation of HSID Methods

Numerous studies related to the comparison of different HSID methods are available in the existing literature (Cheng & Washington, 2008; Montella, 2010; Bandyopadhyaya & Mitra, 2015; Manepalli & Bham, 2016; Wang, et al., 2020). Nonetheless, comparison of DBSCAN and KDE in the traffic safety area is rather limited. To the best of our knowledge, the only relevant study in this area was

conducted by Chen, et al., (2021). They compared DBSCAN and KDE with two other spatial clustering algorithms, CFSFDP and K-Means, to assess their effectiveness in spatio-temporal analysis of vehicle trajectories. Their comparative analysis revealed that DBSCAN performed better than KDE in terms of accurately identifying non-spherical clusters, handling noise, creating heat maps with more accurate shapes, and faster runtime. However, KDE was found to be more accurate in classifying hotspots. Yet, the correctness of both in detecting correct hotspots was closely alike. However, in that study, the HSID was related to taxi trajectory instead of traffic crashes, and the spatial searching dimension was area-wise (2-D). The comparison of DBSCAN and KDE as HSID techniques via the specialized quantitative tests, namely the Site consistency test (SCT), Method consistency test (MCT), Total rank differences test (TRDT), and Total score test (TST) (Cheng & Washington, 2008; Montella, 2010; Wang et al., 2020) is lacking. Szénási & Jankó (2017) conducted the only evaluation of DBSCAN's performance using the tests mentioned above. In their study, they compared a DBSCAN-based approach against the Sliding Window technique and found moderately similar performance between the two methods. However, there is still a gap in the literature regarding the comparison of DBSCAN with other HSID methods using these specialized tests. Therefore, this study aims to fill this gap by comparing a proposed DBSCAN-based tool with KDE+. The stated tests are as follows:

- Site consistency test (SCT): to assess the HSID methods' ability to consistently identify hazardous areas over consecutive time periods. If there is a higher number of crashes in period i+1 at the locations identified as hotspots in period i, then that indicates the HSID method is more efficient.
- Method consistency test (MCT): to evaluate the accuracy of the HSID methods

in identifying hazardous sites by examining the consistency of hotspot locations over consecutive time periods. This can be achieved by counting the number of sites that have been identified as hotspots in both periods, where a higher count indicates better performance.

- Total rank differences test (TRDT): to determine the stability of the ranking of hotspots between consecutive periods identified by the HSID method. This is achieved by calculating the sum of differences in the ranks of hotspots in the two periods. A lower value in this test indicates better performance of the HSID method in terms of consistency of hotspot ranking.
- The Total score test (TST): to obtain a comprehensive comparison between the HSID techniques, the results of the three tests mentioned earlier are combined into a single aggregated measure. The weights assigned to the three tests are assumed to be equal. This measure yields a maximum value of 1 (or 100%), with a higher value indicating a better HSID method.

2.4 Safety Problem Diagnosis

Following identifying hazardous locations, the next phase of roadway safety management is diagnosing problems that caused or contributed to the safety deterioration at those places. In addition to the classic assessments such as collision diagramming (Kononov, et al., 2019), field visit, and plans review, a popular approach to performing the diagnosis process is the descriptive data analysis, suggested in Highway Safety Manual (HSM) (AASHTO, 2010). According to that method, the descriptive statistics of accident attributes, such as the dominant crash type, are evaluated to find any probable patterns, which are reliable bases upon which safety problems that caused the creation of hotspots could be tracked down and diagnosed. Even though some alternative methods such as direct diagnostics (Kononov & Janson,

2002), Beta-Binomial (BB) (HSM), and safety performance functions (SPFs) (Ivan et al., 2017) have been proposed, the descriptive statistic may be the only available option in many countries where the other methods are absent, while it is comparably reliable (Park & Sahaji, 2013). Thus, in the current work, the contribution of the categorization to the enhancement of the descriptive statistic diagnosis method is illustrated. A short description for each of the other mentioned methods is provided below for readers who interest in learning them:

- Direct diagnostics: in this method, accidents are viewed as random Bernoulli trials and the probability of k time occurrence (k= observed frequency) of each specific crash type at the targeted hotspot is calculated based on the historical data-driven Bernoulli probability (mean proportion) of the corresponding crash type at same road entity type. If the calculated probability is abnormally low for a crash type, the expert may conclude that the unknown safety problem tends to cause that crash type, which could be a valuable insight.
- Beta-Binomial (BB) test: it is a generalized version of the binomial test in the direct diagnostics method of Kononov & Janson (2002) to tackle overdispersed data (Park, & Sahaji, 2013).
- Diagnosis using the safety performance functions (SPFs): SPFs are prediction functions developed for different crash types separately, which are developed by applying regression models (e.g., negative binomial) to past crash data of collections of similar sites in a city or region. According to this method, if the observed frequency of a crash type is greater than its expected count calculated by SPFs, the analyst concludes that the crash type is overrepresented.

Chapter 3

CANDIDATE HSID METHODS FOR THE DESIGNED CATEGORIZATION

3.1 DBSCAN

DBSCAN is popular because of its ease of use, especially when handling large-sized data with noises and outliers, the ability to detect clusters with arbitrary shapes (Borah & Bhattacharyya, 2004), and its freedom from cluster-size inputting. To perform its function, DBSCAN requires a criterion to measure the density of the data. This criterion is the existence of a specified minimum number of data (MinPts) within a circular searching neighborhood with a given radius called Epsilon (Ɛ) (Ester et al., 1996). The recap of the DBSCAN algorithm functioning is that commencing from a random point, it verifies whether the collection of that point and its Ɛ-environment points satisfy the MinPts condition for forming a cluster. If the condition is met, the first cluster is formed (Figure 1); Otherwise, the first point is temporarily labeled as an outlier (noise). Similarly, the algorithm surveys all population points and detects clusters and outliers.

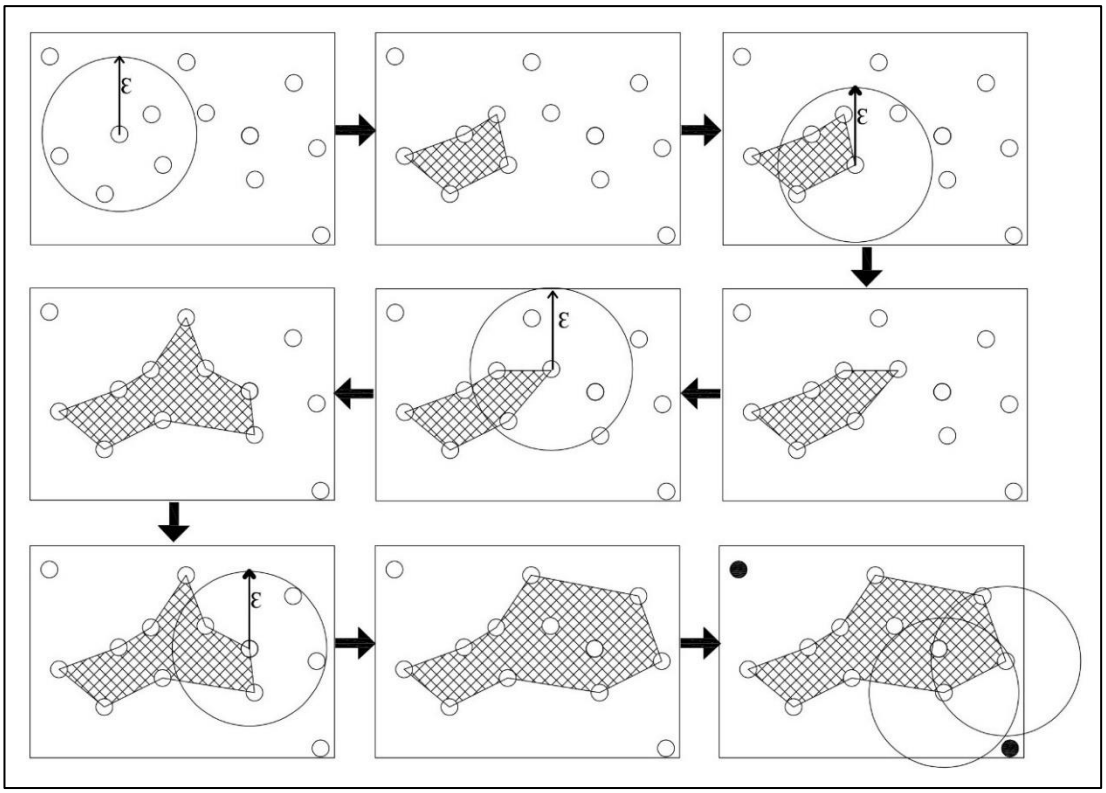

Figure 1: Cluster detection process by the DBSCAN algorithm; the black shaded circles are noises (Babaei & Kunt, 2023).

DBSCAN has seldom been applied in 1-D HSID due to few disadvantages, which some researchers tried to overcome; Szénási & Jankó (2017) aimed at reducing the computational time of DBSCAN. The scheme comprised subdividing the problem space into minor units and creating a spatial indexing matrix for the units to accelerate the ε-environment query. Qiu, et al., (2016) proposed a technique for selecting the appropriate quantities of ε and MinPts based on satisfying the condition of being shorter than a pre-defined cluster-lengths limit. The criterion for selecting the optimal parameter pair was the largest computed kurtosis of accident rate distribution for all clustered road sections. Zhang, et al., (2018) suggested Dijkstra's-DBSCAN, to overcome the potential imprecision of the Euclidean distance measure. That method was based on calculating the shortest distance between crash points (vertices) along

the given paths; thus, it needed the geographical layout of network paths (edges) as an input.

However, this study intended in enhancing the use of the original DBSCAN for HSID with an approach that would resolve the DBSCAN's limitations, specifically by a trick to bypass the Euclidean distance problem and a new way to determine the appropriate values for the input parameters. For the specific purpose of this study, hotspot categorization, long roads should be split into homogeneous segments, if needed, only based on the annual average daily traffic (AADT), lane number, and the median presence state so that each homogeneous segment will be designated as a reference population with all locations on them to be similar. That in turn will prevent the error of using a fixed crash density threshold (Yakar, 2021) (Figure 2). Subsequently, the quantification of the DBSCAN parameters and the cluster analysis is executed for each segment exclusively to prevent the placement of accident points from different roads (adjacent or grade-separated) in the same cluster. It is doable by just one run of a loop code; thus, it is not time-consuming.

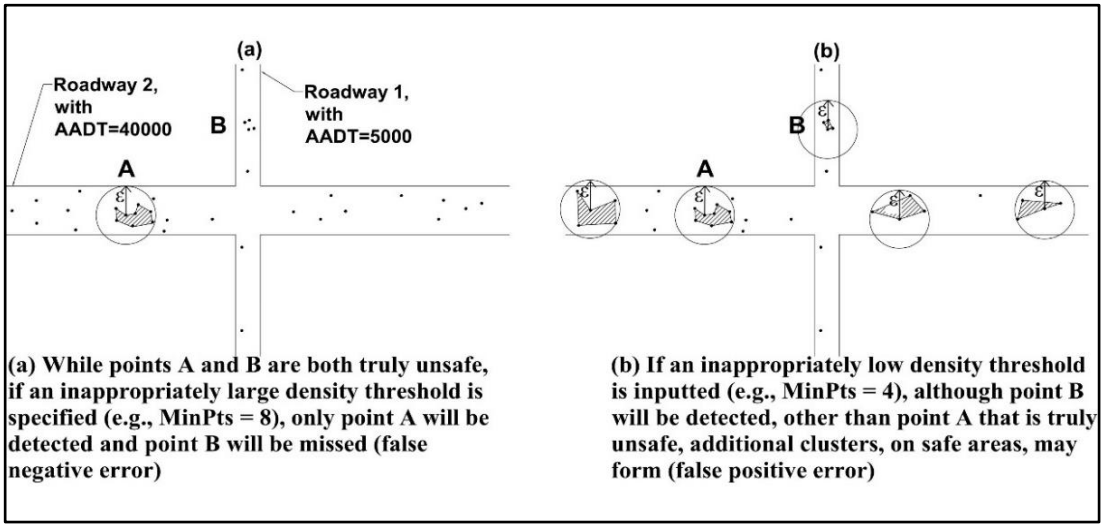

Figure 2: Disadvantages of using a fixed density threshold (Babaei & Kunt, 2023).

Determining the ε and MinPts values is the major challenge of DBSCAN; it needs two rules of thumb to be regarded. The first rule concerns the value of ε: it should be small enough such that during the search for neighboring data points, the difference between the Euclidean distance, which is the metric used by DBSCAN, and the network distance be as little as ignorable.

That constraint only applies to non-straight sections (i.e., curves); and the worst possible scenario is the presence of crashes on curves with too short radii of 150 meters as the minimum recommended radius for design speeds of 60 km/h by the American Association of State Highway and Transportation Officials standard (AASHTO, 2001). On such curves, when the chord length (which represents the Euclidean distance) is 50 meters, the curve length (network distance) equals 50.23 m if the maximum superelevation rate is 4% (or 50.48 m if e=12%), as shown in Figure 3 and expressed mathematically below and:

$$C_x = 2R \sin d_x = 50 \text{ m} \rightarrow 2 \times 150 \times \sin d_x = 50 \text{ m} \rightarrow d_x = 0.167 \text{rad}$$

$$d_x = \frac{L_x}{2R} \rightarrow 0.167 = \frac{L_x}{2 * 150} \rightarrow L_x = 50.23 \text{ m}$$

where C_x, R, d_x, and L_x are chord (Euclidean distance), radius, deflection angle, and network distance, respectively.

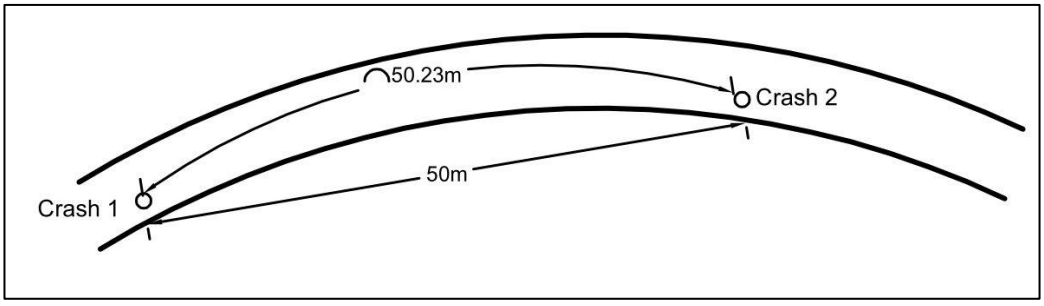

Figure 3: Difference between the Euclidean distance and the network distance.

Thus, by lowering the ε value to 50 m, the difference between the network distance and the Euclidean distance at any locations along highways becomes small enough to ignore, meaning that the potential error of the Euclidean distance will be avoided. The second rule relates to the MinPts parameter. Obviously, routes with dissimilar crash densities (due to the varied traffic volumes) should not be analyzed by a fixed MinPts value. The choice of the MinPts value rests on the defined threshold for cluster density. Using skewness, kurtosis, Shapiro-Wilk statistics, frequency plots, and the Chi-square test, Thomas (1996) demonstrated that, if roadways are segmented into sections of up to 300 meters, the distribution of accident counts per segment follows Poisson distribution (Figure 4).

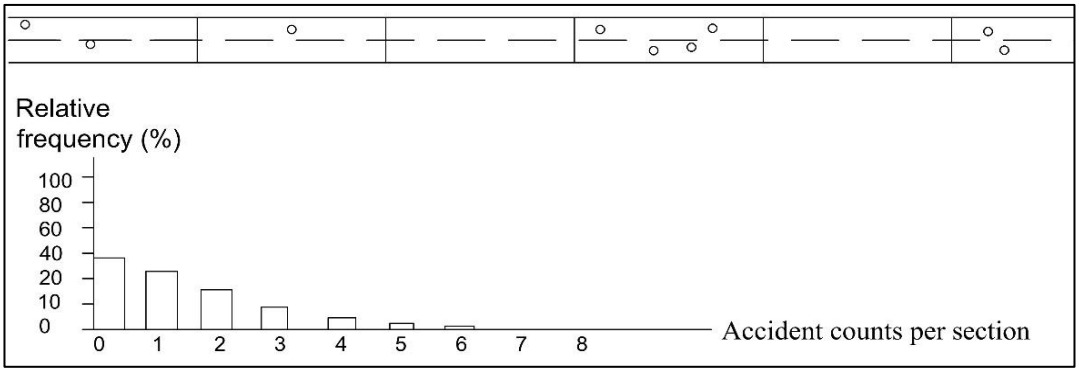

Figure 4: Distribution of accident counts per section for road sections up to 300 m follows Poisson (Thomas, 1996).

If we visualize a homogeneous segment as a set of connected 100-meter sub-segments, we may define a hotspot as the following: if a sub-segment experiences such a high crash counts, whose occurrence probability is lower than α (the significance level on the right tail of the Poisson probability density curve), it will be a hotspot with a 1-α degree of confidence as it rejects the null hypothesis that crashes are randomly distributed. Consequently, the MinPts value will be that described crash count. Determining the magnitude of α depends on the desired level of traffic safety; the value

of α was set as 0.1 in this study. The mathematical expression of the suggested threshold is as follows:

$$\text{MinPts} = k, \text{ if } P(X > k) = 1 - P(X \leq k) = 1 - \sum_{x=0}^{k} \frac{\lambda^x e^{-\lambda}}{x!} \leq \alpha = 0.1 \quad (1)$$

$$\lambda_i = (N_i/L_i) \times 100 \quad (2)$$

where k is the number of crashes; e is Euler's number; and λ_i, N_i, and L_i are accident mean rate (crash counts per 100 m), total crash counts, and the length of the i^{th} segment (in meters), respectively. The length of the imaginary sub-segments was set as 100 m because the selected value of the ε was 50 m, which means that the diameter of the circular neighborhood searching is 2×50 = 100 m. As an example, if the total crash count for a segment with 3000 m length during a period is 90, the mean value will be 3 crashes per sub-segment; thus, the crash count with a 10% or lower probability of occurrence, corresponding to the λ = 3, will be 5. So, the MinPts value for that segment, in that period, will be 5. The x values corresponding to any λ are either available or might be interpolated from the table of Poisson Cumulative Distribution (Figure 5).

λ=		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.2	1.4	1.6	1.8
x=	0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.3012	0.2466	0.2019	0.1653
	1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.6626	0.5918	0.5249	0.4628
	2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.8795	0.8335	0.7834	0.7306
	3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810	0.9662	0.9463	0.9212	0.8913
	4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963	0.9923	0.9857	0.9763	0.9636
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9985	0.9968	0.9940	0.9896
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994	0.9987	0.9974
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
λ=		2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.5	5.0	5.5
x=	0	0.1353	0.1108	0.0907	0.0743	0.0608	0.0498	0.0408	0.0334	0.0273	0.0224	0.0183	0.0111	0.0067	0.0041
	1	0.4060	0.3546	0.3084	0.2674	0.2311	0.1991	0.1712	0.1468	0.1257	0.1074	0.0916	0.0611	0.0404	0.0266
	2	0.6767	0.6227	0.5697	0.5184	0.4695	0.4232	0.3799	0.3397	0.3027	0.2689	0.2381	0.1736	0.1247	0.0884
	3	0.8571	0.8194	0.7787	0.7360	0.6919	0.6472	0.6025	0.5584	0.5152	0.4735	0.4335	0.3423	0.2650	0.2017
	4	0.9473	0.9275	0.9041	0.8774	0.8477	0.8153	0.7806	0.7442	0.7064	0.6678	0.6288	0.5321	0.4405	0.3575
	5	0.9834	0.9751	0.9643	0.9510	0.9349	0.9161	0.8946	0.8705	0.8441	0.8156	0.7851	0.7029	0.6160	0.5289
	6	0.9955	0.9925	0.9884	0.9828	0.9756	0.9665	0.9554	0.9421	0.9267	0.9091	0.8893	0.8311	0.7622	0.6860
	7	0.9989	0.9980	0.9967	0.9947	0.9919	0.9881	0.9832	0.9769	0.9692	0.9599	0.9489	0.9134	0.8666	0.8095

Figure 5: Table of Poisson cumulative distribution.

The explained MinPts quantification way allows identifying crash clusters on diverse road classes with different crash densities (Gregoriades & Chrystodoulides, 2018) (Figure 6).

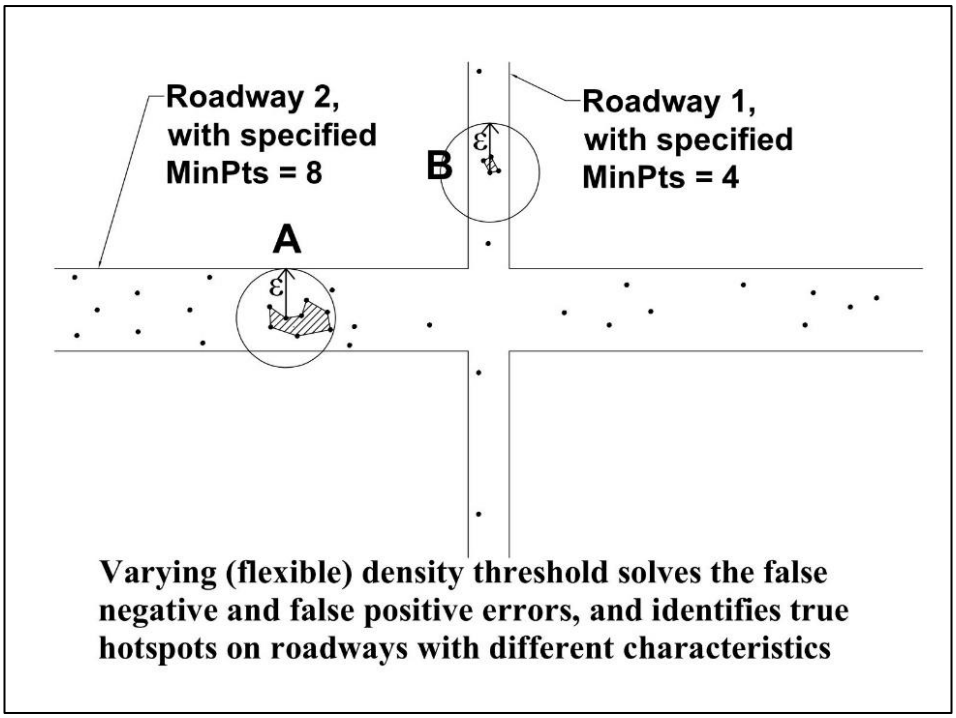

Figure 6: The advantage of using a varying density threshold (Babaei & Kunt, 2023).

Following the identification of the clusters, to provide a measure for ranking clusters, the equation below is suggested:

$$SD = \frac{n}{\log(l)} \tag{3}$$

where, SD, n, and l represent the scaled density, crash counts, and the length of the clusters. In this density index, the length is logarithmic (log to base ten) to avoid overweighting the length in the fraction. As a clarifying example, if clusters A and B with 40 m and 10 m lengths contain ten and three accidents, respectively, their non-scaled densities will be 0.25 and 0.3, which results in ranking B higher than A despite having much lower crash counts. But, using the logarithmic-scaled length, the cluster densities will be 6.24 and 3, more rationally ranking point A above B.

To solve the mentioned problem regarding missing crash clusters on segment boundaries, ‘transition segments’ are suggested in addition to the main segments. They are shared sections between two adjoined segments, with a length equal to 2 × ε and a MinPts equal to the average MinPts of the two segments (Figure 7). However, since that may result in finding some clusters twice, the duplicated clusters should be removed at the end, which is doable simply.

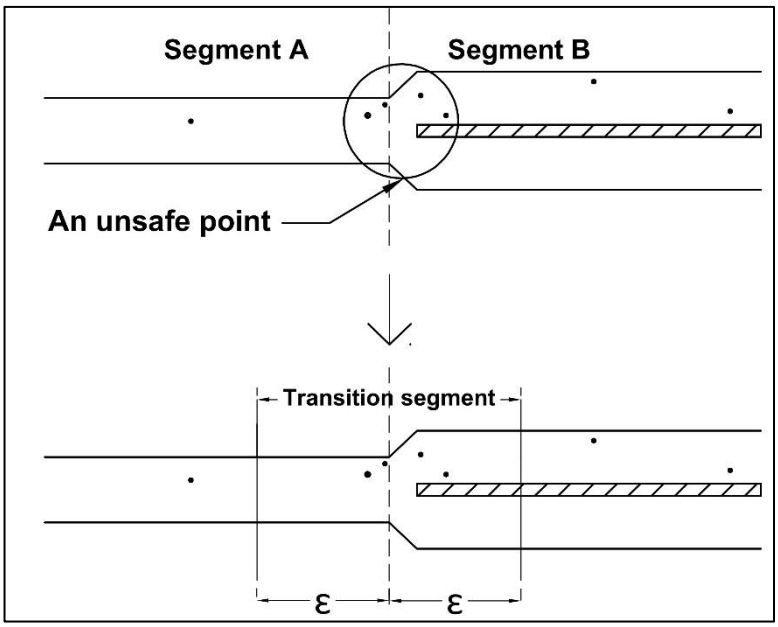

Figure 7: Transition (shared) segments.

To apply the DBSCAN algorithm to the crash data, the unit of Ɛ should be converted from meters to the global positioning system (GPS) coordinates format (WGS_1984, decimal degree format) to match the location of the accident points. Accordingly, 50 meters is approximately equivalent to 0.000513, the magnitude of the mean vector of the latitudinal and the longitudinal vectors.

3.2 KDE+

KDE is a popular technique applied to identify locations with high densities of point-events. It estimates unknown probability density functions using a kernel function (k).

The general equation of KDE is as the following (Xie & Yan, 2008):

$$D(s) = \sum_{i=1}^{n} \frac{1}{\pi r^2} w\left(\frac{d_{is}}{r}\right) \quad (4)$$

where D(s) denotes the point-events density at location s; r is the radius of the searching sphere (referred to as band-width), and w is the weight of the i^{th} object located in the d_{is} distance from point s (the center of the sphere).

KDE+ is a network-level (1-D) KDE, which uses a kernel function called Epanechnikov kernel as follows:

$$K_d(x) = \frac{3}{(4d)}\left(1 - \left(x/d\right)^2\right) I_{(-d,d)}(x) \quad (5)$$

where d denotes the bandwidth, x is distance from the center, and $I_{(-d,d)}(x)$ is the indicator function of the (-d, d) interval. KDE+ specifies a density threshold by Monte Carlo simulation to select the significant clusters. It provides the user with two additional benefits: a property called 'Strength' to rank clusters and a test called 'Stability' to filter out the clusters with unstable strengths. KDE+ has demonstrated acceptable performance in network event clustering and has been used widely in recent years (Bíl et al., 2019; Benedek, et al., 2016; Zahran et al., 2019). However, despite its excellent performance, KDE+ holds some limitations, as described in the results section. Moreover, its dependency on special software (ArcGIS) and files (shapefiles), which may not be available for some users, could be another motivation to seek a less expensive alternative. From the engineering point of view, a more efficient tool or technique is the one that provides an acceptable service in terms of quality, accuracy and precision, with a lower cost.

Chapter 4

HOTSPOTS CATEGORIZATION

4.1 Concept

The tempo-categorization approach we suggested was supposed to identify locations on roads that exhibit abnormally large crash counts, differentiated based on their corresponding arising periods. It is established upon two conditions: First, the reference relative to which a location is flagged as a hotspot should be the segment on which that location is situated (this condition has already been considered in the suggested DBSCAN-based approach); Second, a point is recognized as hotspot, if it keeps exhibiting a threshold-exceeding crash frequency consistently over consecutive periods whatever the period dimension is. Figure 8 depicts this categorization notion; It shows the crash counts of some distinct locations on a homogeneous segment during each season and time of day in one year, assuming that, the same pattern reoccurs in multiple consecutive years. The points where crash densities have exceeded the density threshold of the object periods are marked with a tick below them. In the time-segmentation step, the time-of-day intervals should be determined such that each of them would represent different traffic or environmental conditions (e.g., volume and lighting conditions) to enable us so as to examine the effects of such factors. Hence, they were defined as morning (6 to 9:59), daytime off-peak (10 to 15:59), evening (16 to 20:59), and night (21 to 5:59). The segmentation of the year into the four seasons follows the same logic, as the four seasons of winter, spring, summer, and fall reflect four different weather conditions resulting in varied road and driving circumstances.

As Figure 8 depicts, some points may become hazardous solely during a certain season or time of day, called periodic hotspots, such as Winter Emerging (WE) or Morning Emerging (ME) (e.g., A, B, C, E, G, I, and K). Some places may experience unsafety during more than one of the four defined intervals (e.g., H), or at every single season and time of day, called Unbounded hotspots (e.g., F and J). Some other locations may never become a hotspot at any particular period, while their cumulated crash counts may reach the annual threshold (MinPts), resulting in being recognized as a Yearly Non-Periodic (YNP) hotspot (e.g., D). Thus, each hotspot identified based on the yearly-accumulated crashes lies under one of the described classes. This categorization style yields two notable benefits, finding potentially hidden unsafe spots and enhancing the safety diagnosis process, as described below:

- **First benefit:** some hazardous sites may not be spotted by the yearly-based HSID manner, whereas the suggested algorithm in this study can detect them. These are locations where despite experiencing relatively low crash frequencies, show a steady temporal pattern and prove non-random crash occurrences.
- **Second benefit:** if the hotspot category is known, the analyst will have the chance to concentrate the diagnostic investigation on the appropriate subset of the crash records in the safety diagnosis process by descriptive data analysis method. Without performing the categorization, and merely by staying on the traditional approach, when hotspots are investigated individually by mining their recorded crash attributes, a dominant crash characteristic like crash type may not always be observable to track down the causation of the accidents. Even the attempt to find the period with the significantly higher crash frequency, which can yield a clue for diagnosing the safety problem, may fail

if crashes display a balanced distribution across the periods. The appearance of the temporal pattern of the under-investigation hotspots' crashes may even be deceptively in favor of a false critical period, and mislead the analyst. It happens when for example, the hotspot is actually Night Emerging, but due to the relatively low traffic volume during night, the number of nighttime crashes looks equal to or even lower than that of another interval, such as evening at which the crash frequency is normally higher along the entire road because of a higher traffic volume. However, the categorization overcomes such potential problems: being a Night Emerging hotspot means that, a threshold-exceeding crash frequency at that location only happens at night, meaning that the point becomes unsafe merely during the nighttime. Thus, the nighttime crashes should be focused on for exploring the problem, which is corroborated by the patterns of the crash attributes (e.g., dominant crash type). Knowing the hotspot category, even if the crash-type configuration of the problem-arising period is similar to that of the other periods, some other attributes like illumination, road condition, and vehicle-type may exhibit a divergent pattern, which can provide experts with insights that could not otherwise be collected.

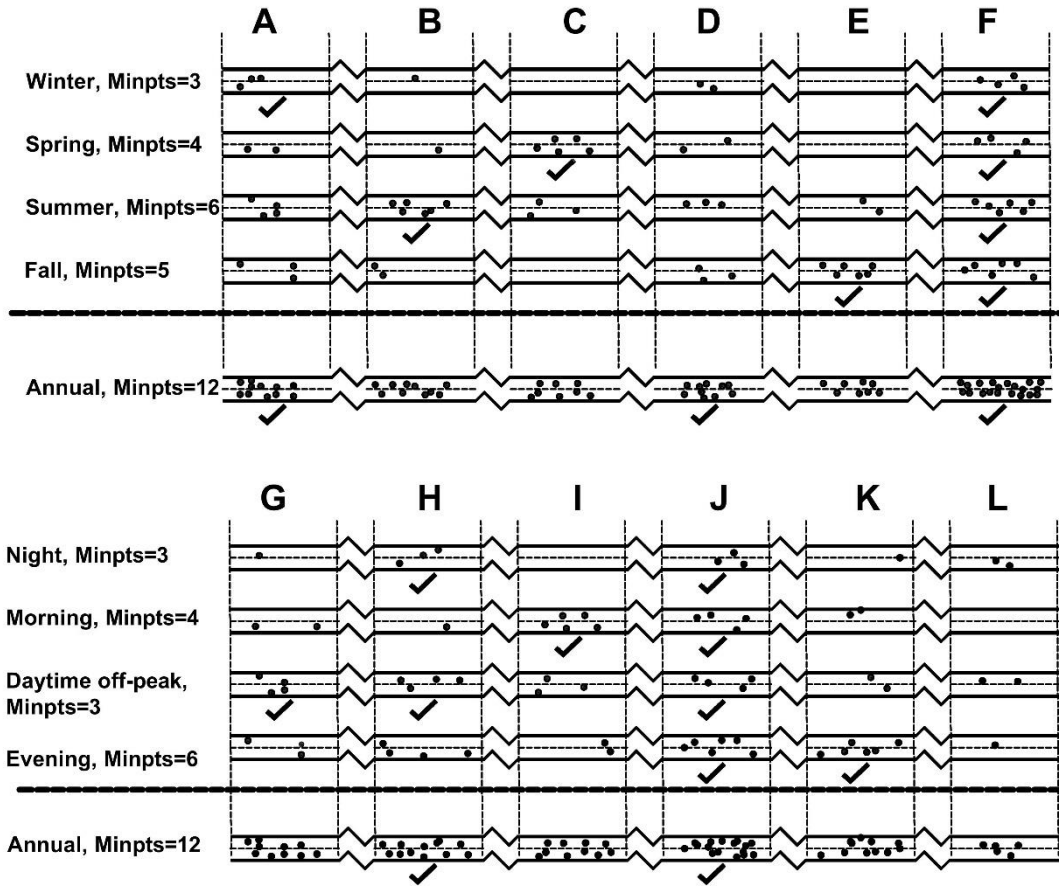

Figure 8: Points on a homogeneous segment with diverse temporal categories as they turn into hotspots at different periods (Babaei & Kunt, 2023).

The diagnosis process for the U and YNP hotspot includes eliminating the inapplicable risk factors. YNP hotspots are locations recognized as hotspots only based on the yearly-accumulated crashes, while they do not turn into hotspots at any of the defined periods. That property implies the insignificant effects of weather and traffic conditions in the formation of the YNP hotspots, but instead, the role of some geometrical faults. Unbounded hotspots are points that exhibit inferior safety performance in all environmental and traffic conditions. Thus, the causative safety problem is likely to be a full-time substantial engineering fault rather than traffic or environmental-related hazard.

4.2 Algorithm

Identifying hazardous locations with different categories is performed via a stepwise procedure, including clustering by DBSCAN and applying intersection and difference operations. The Python programming language (version 2.7., available at http://www.python.org) was employed for executing the algorithm. The required crash explanatory variables are date, time, latitude and longitude, and street names. The algorithm needs sorting data by the year, date, and time of the accidents to separate diverse subgroups of records related to the intended periods. The procedure is as follows:

- **Step 1:** for each period *i*, crash clusters and their centers (Equation 6) are identified for the segments separately, based on their corresponding MinPts values:

  $$C_j(Lat) = \frac{1}{n}\sum_{k=1}^{t} A_k\,(Lat)\ ,\ \ C_j(Lon) = \frac{1}{n}\sum_{k=1}^{t} A_k\,(Lon) \qquad (6)$$

  where C_j *(Lat)* and *Cj (Lon)* are the latitude and longitude of the j^{th} cluster-center; A_k represents crashes, and *n* is the total number of crashes in the cluster. Next, the cluster-centers of all segments in the period *i* are stored in a CSV file. Then, step1 is repeated for each of the remaining periods, *i*-1 and *i*-2. Since the number of crash points aggregated in one year seems too low for forming clusters by DBSCAN, especially in low-volume roads, two years were considered as the period length.

- **Step 2:** the overlapping hotspots between periods *i* and *i*-1 are identified such that, if the Euclidean distance between a cluster-center from period *i* and one from *i*-1 is shorter than 2×ε, these two clusters overlap (Equation 7); So, that area is a shared hotspot between periods *i* and *i*-1, which together with the other shared hotspots, creates a set named *i*, *i*-1_overlapped (Figure 9). Similarly,

sets *i*, *i*-1, *i*-2_overlapped and *i*, *i*-2_overlapped are created; then, the union of the three sets, instead of *i*, *i*-1, *i*-2_overlapped, will be considered as the set of hotspots stable over the three periods; that is to avoid the effects of the potential year-to-year variation in the crash frequency to forgive the probable deviation of crash counts of the truly hazardous points from the mean value in one or two years.

$$d(C_jC_k) = \sqrt[2]{(C_j(Lat) - C_k(Lat))^2 + (C_j(Lon) - C_k(Lon))^2} \quad (7)$$

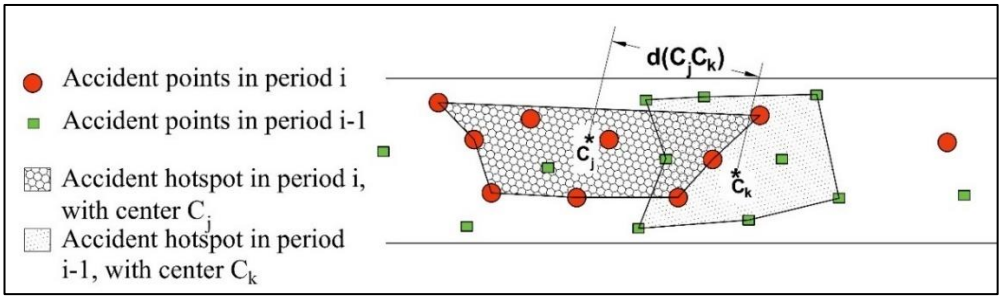

Figure 9: Overlapping clusters (crash clusters formed on the same location in two consecutive periods) (Babaei & Kunt, 2023).

The explicated two-step algorithm is applied to each season and time of day, for example, fall and night (Equations 8 and 9), and to the years (Equations 10).

$$F = [Falls_i \cap Falls_{i-1} \cap Falls_{i-2}] \quad (8)$$

$$N = [Nights_i \cap Nights_{i-1} \cap Nights_{i-2}] \quad (9)$$

$$Y = [Years_i \cap Years_{i-1} \cap Years_{i-2}] \quad (10)$$

In these equations, the intersection operator represents the satisfaction of the stated distance condition in step 2. Similarly, the overlapping hotspots between all four seasons (set S) and among all four time-of-day intervals (set T) are identified (Equations 11 and 12).

$$S = [F \cap Su \cap Sp \cap W] \quad (11)$$

$$T = [N \cap M \cap D \cap E] \quad (12)$$

where Su, Sp, W, M, D, and E stand for the sets of hotspots that form in summers, springs, winters, mornings, daytime off-peaks, and evenings, respectively, which are acquired similar to F and N.

- **Step 3:** The periodic hotspots are obtained; for instance, the Fall-Emerging hotspots, which are hotspots that only form in the fall seasons and not in the other seasons (Equation 13).

  $$FE = F - W - Sp - Su \quad (13)$$

- **Step 4:** Finally, the Unbounded hotspots (set U), which hold the highest rank in the treatment plan, are extracted from the previously retrieved hotspot groups. These are points that become hotspots in all seasons and all time-of-day or at least in three out of the four intervals of both the seasonal and time-of-day sections, obtained by the equation below:

  $$U = [S \cap T] \quad (14)$$

The overall framework of this study is depicted with the logical reasoning diagram in Figure 10.

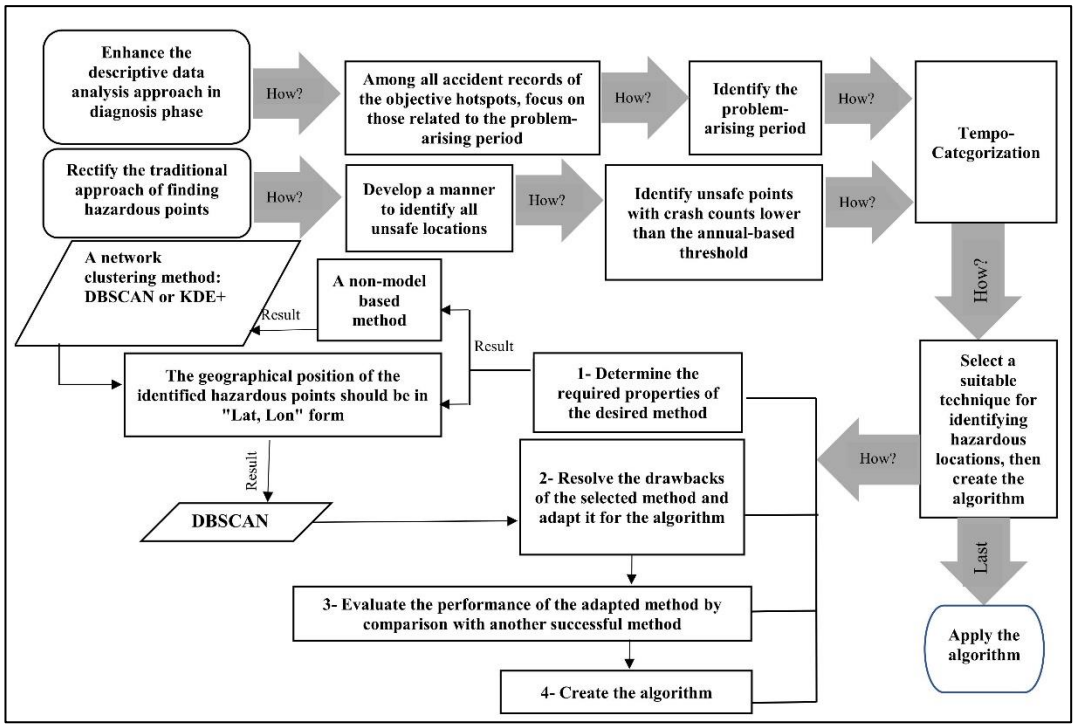

Figure 10: Logical reasoning diagram of this study (Babaei & Kunt, 2023).

Chapter 5

VALIDATION OF THE METHOD WITH A CASE STUDY

5.1 Study Area and Data

To test the proposed method, the accident data for the intra-city highway network of Allegheny County (Pennsylvania state, USA) was acquired from the Pennsylvania Open Data webpage (https://data.wprdc.org/dataset/allegheny-county-crash-data). The dataset contains the required crash attributes for the proposed method in this work. Six years, from 2014 to 2019, were selected as the study period assuming that the pavement conditions had not changed during this period. This time segment seems sufficient to indicate possible consistency of hotspots over time and to avoid the regression-to-the-means phenomenon, yet not too long to risk the immutability of the road geometry. As data cleaning, the records with the wrong or missing coordinates and those lacking street names were removed. Next, the intersection accidents were dropped as they were out of this study's scope. Afterward, the crash records of the opposite traffic directions of the divided highways were separated to analyze them distinctly since their characteristics may differ. Last, several roads with a total length of 570 km from different classes were randomly sampled and divided into 100 homogeneous segments, based on AADT, lane number, speed limit, and the opposing lane separation type. Table 1 illustrates an example of the measured characteristics for some segments. The traffic volume info was acquired from the webpage https://gis.penndot.gov/tire, and Google Maps was used to obtain the lane numbers and

verify the presence of medians. The total crash number of the sampled segments was 12229.

Table 1: Examples of the measured characteristics of the segments.

Street name details	Length (m)	AADT	Divided/ Undivided	Lane # in each direction	Speed limit	Start and end-point coordinates
ALLEGHENY RIVER BL (Lower)	4340	above 20k	U	1	45	40.48401, -79.90727 to 40.48516, -79.85691
ALLEGHENY RIVER BL (Middle)	2070	10k to 20k	U	1	45	40.4852, -79.8568 to 40.513966, -79.843215
ALLEGHENY VALLEY EX (Lower) W	22800	above 20k	D	2	55	40.4971, -79.9364 to 40.608490, -79.762676
ALLEGHENY VALLEY EX (Middle2) E	7000	above 20k	D	2	55	40.561625, -79.800184 to 40.608438, -79.762436
BABCOCK BL	6000	10k to 20k	U	1	35	40.506247, -79.990106 to 40.673616, -79.989552
BAUM BL	2630	5k to 10k	U	2	55	40.454004, -79.953291 to 40.460456, -79.925004
BIGELOW BL-E	3780	10k to 20k	D	2	35	40.441050, -79.993900 to 40.458703, -79.957737
BUTLER ST(North)	5050	10k to 20k	U	1	35	40.467554, -79.963762 to 40.486290, -79.913505

5.2 Comparative Analysis Results

To execute the comparative analysis between the proposed DBSCAN-based tool and KDE+, 12 segments with an overall length of 98 km were randomly chosen (Table A1). Next, using ArcGIS 10.7, the segments were stationed into 100-m length subsegments and were assigned identification numbers for addressing the location of the clusters. The two methods were applied to crashes on these segments in two periods

of 2014, 2015 and 2016, 2017, with 1362 and 1063 crashes, respectively (more details regarding the numerical outputs of the suggested DBSCAN-based method are presented through some tables in the next section). To count the crashes, solely the areas bound to the clusters were considered, whereas in the non-clustering methods, all sections, which are predetermined by specified start and end points, are considered. The cases in which clusters from the two periods did not overlap but were extremely close to each other (with lower than 50 m distance) were treated as overlapping in the MCT calculations. Furthermore, in calculating SCT, the crash points up to 30 m away from the boundaries of the past period clusters were assumed underlying. Figure 11 and Table 2 show snippets of the tests' computation corresponding to the Allegheny River BL (lower) segment, using ArcMap.

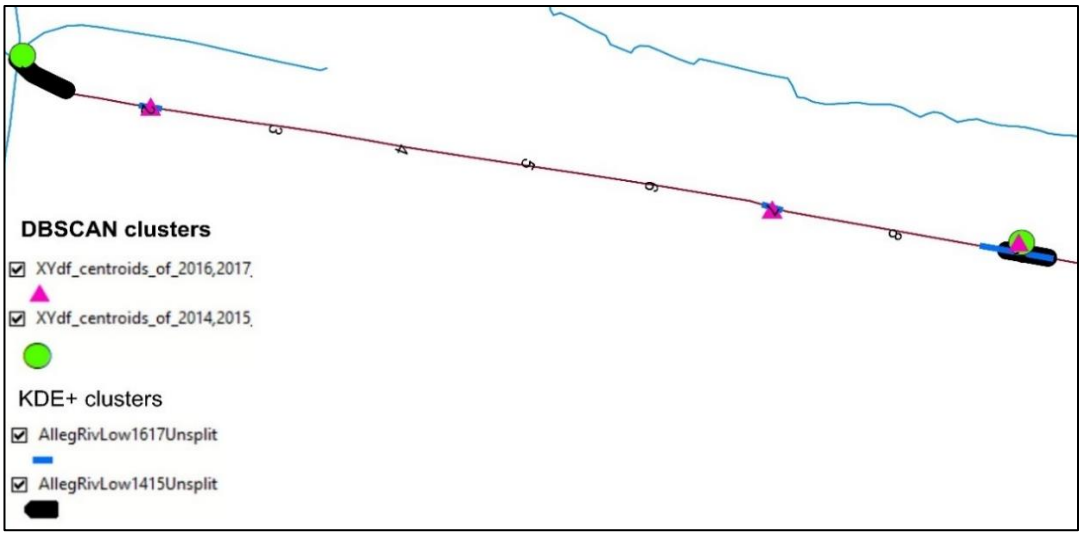

Figure 11: A segment with crash clusters detected by the DBSCAN and KDE+ in two consecutive periods.

Table 2: The tests' computations corresponding to the Allegheny River BL (lower) segment. The first two rows (clusters 1 and 2) are corresponding to the sippet shown in Figure 11.

DBSCAN							**KDE+**						
Cl. ID	Place	Crash # in P2	Repeat in P2	Rank in P1	Rank in P2	Rank difference (TRDT)	Cl. ID	Place	Crash # in P2	Repeat in P2	Rank in P1	Rank in P2	Rank difference (TRDT)
1	1	0					1	1	0				
2	8_9	3	*	6	3	3	2	8_9	3	*	4	3	1
3	26	1					3	26_27_28	9	*	2	2	0
4	27_28	8	*	2	1	1	4	32_33_34	2	*	3	8	-5
5	32_33	0							**14**	**3**			**-4**
6	33_34	2	*	4	4	0							
		14	**3**			**4**							

Table 3 presents the results of the four tests applied to DBSCAN and KDE+ (the details are provided in Table A2).

Table 3: Quantitative comparison between KDE+ and DBSCAN ability in identifying locations of hazardous points.

Method	**Total clusters in 2014,2015**	**SCT**	**MCT**	**TRDT**	**TST (%)**
DBSCAN	160	435	66	29	100
KDE+	148	424	64	61	80.7

As Table 3 suggests, DBSCAN slightly outperforms the KDE+ in the first three tests and resultantly in the TST. However, since the total length may seem insufficient to some readers, the observed outperformance of DBSCAN over KDE+ may not always be valid, and further evidence may be needed. Nonetheless, the overall result suggests a high similarity between the two methods in locating clusters, as 89% of the KDE+ clusters and 83% of the DBSCAN's were among the mutual clusters. The main perceived difference between the two methods' identified hotspots is that DBSCAN

occasionally detects more than one small-sized clusters in close vicinity separately, while KDE+ recognizes the collection of those separate clusters as one long continuous cluster (Figure 12). Consequently, the number of the DBSCAN clusters is higher than that of KDE+ (160 versus 148 in the 2014,2015 period); thus, MCT of DBSCAN is greater.

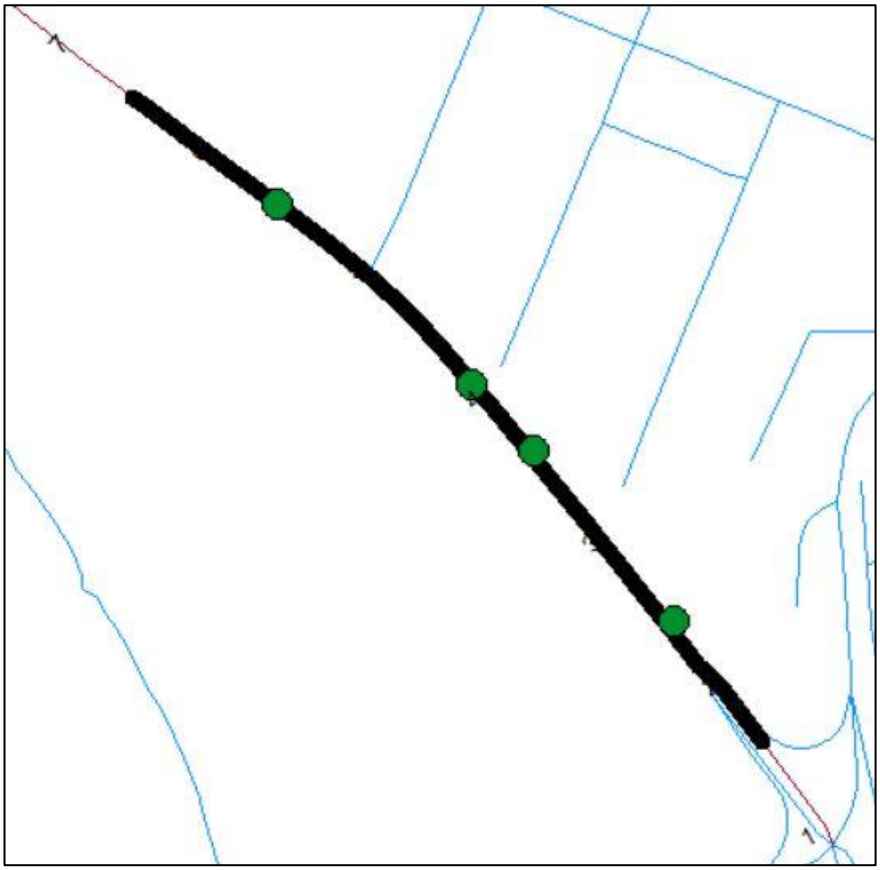

Figure 12: An example that shows how KDE+ merges adjacent clusters into one continuous cluster (the black line), while DBSCAN treat them as separate clusters (the circles) (Babaei & Kunt, 2023).

As an additional contrast between the two methods, Table 4 presents a general description of the characteristics of the two methods.

Table 4: Qualitative comparison between KDE+ and DBSCAN.

Method	Requirements (in addition to crash data)	Output	Graphical Display
DBSCAN	A programming language like Python (open-source)	A csv file presenting: center points of HSs (longitude & latitude), cluster length, number of crashes on clusters, scaled density, and rank of clusters.	Html (On Google Maps)
KDE+	ArcGIS, street shapefile, and crash data shapefile.	A table presenting: extension of HSs (station), cluster length, number of crashes on clusters, strength, density, and rank of clusters.	On ArcMap

The graphical displays of identified clusters by the two methods are shown in Figure 13.

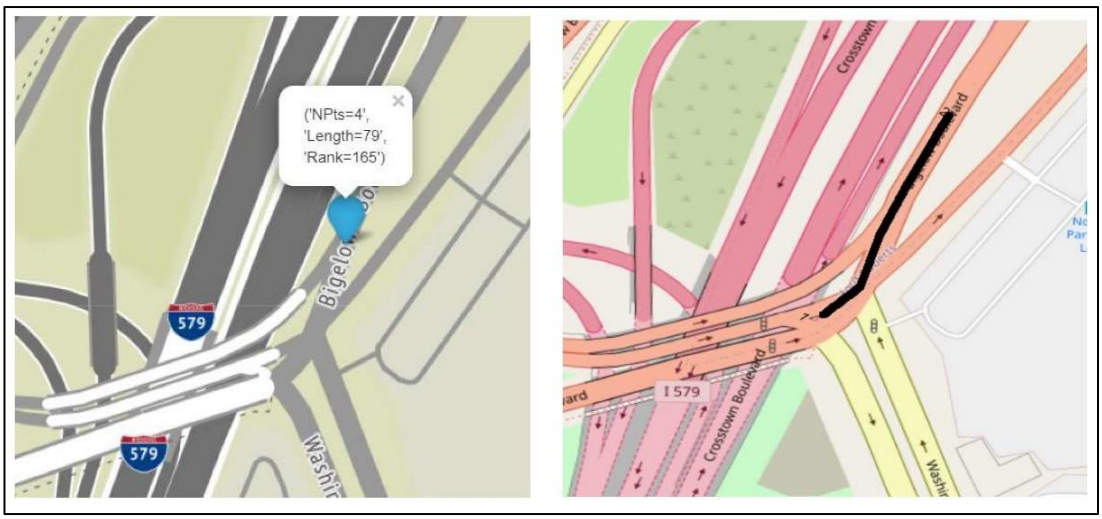

Figure 13: Graphical displays of the identified clusters by DBSCAN (left) and KDE+ (right) (Babaei & Kunt, 2023).

In addition to the two tables above, the methods' performances can be compared further. In this regard, some limitations of the KDE+ highlighted by Bíl, et al., (2016) were reviewed: first, if some crash points are located between adjacent routes due to the imprecisely recorded coordinates, they are susceptible to be assigned to the wrong

road. For the same reason, the two separate lines that represent a divided road in GIS, must be unified into one; that may impact the accuracy of the analysis. Second, any segments shorter than twice the bandwidth should be omitted from the analysis, while some of them may contain true hotspots. In addition to these drawbacks, the observations in this comparative study showed that in some cases where crash points are in close vicinity, the KDE+ assigns a length larger than the actual to the cluster extension (Figure 14).

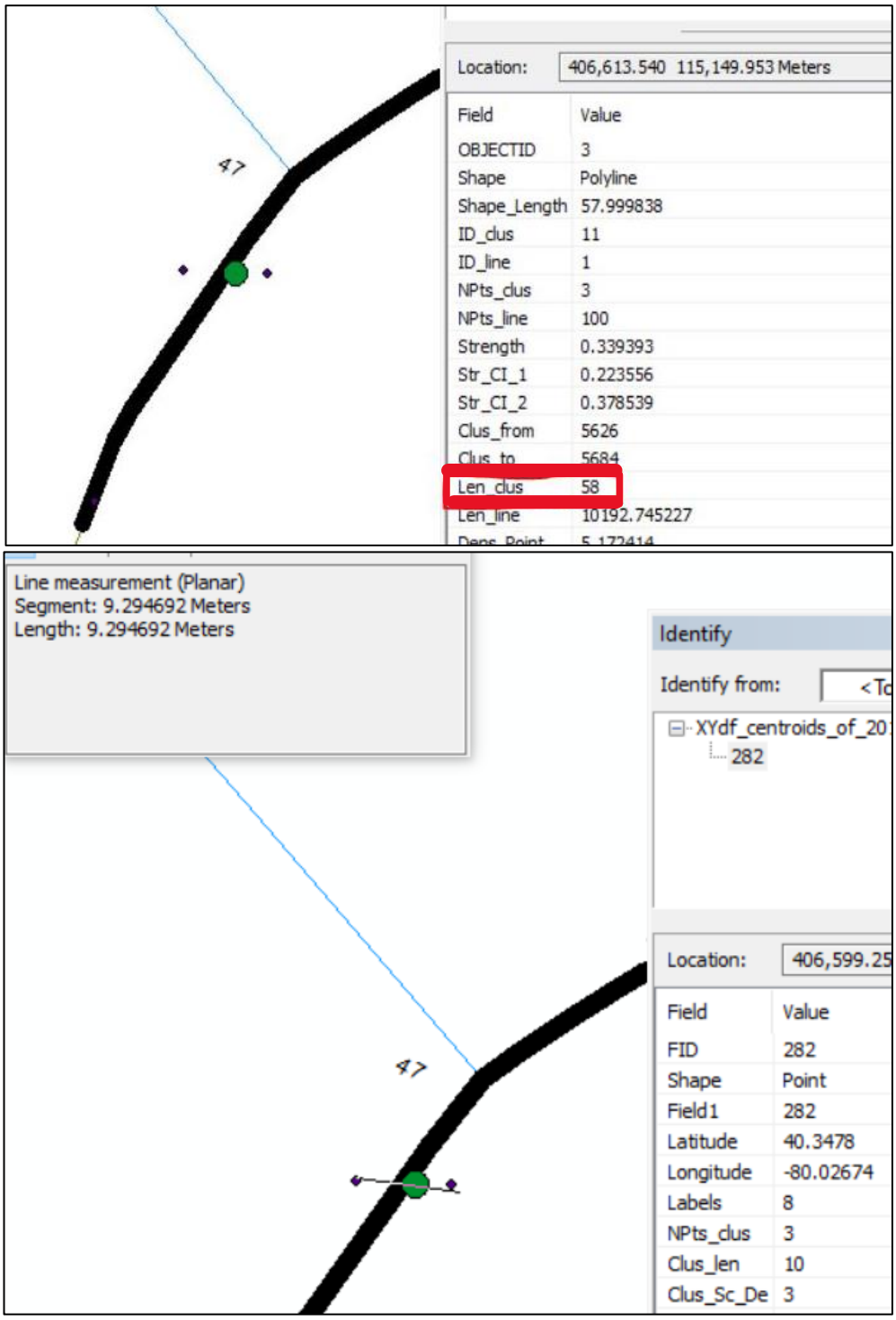

Figure 14: Top: KDE+ assigns a length larger than the actual; Bottom: DBSCAN measures the actual length of clusters correctly.

None of the mentioned problems are associated with the suggested DBSCAN-based procedure; hence, taking all mentioned advantages, in addition to its acceptable clustering performance, it may be admitted as a reliable HSID alternative. Another

advantage of this technique over KDE+ regards the categorization part in this study, where KDE+ is not apt enough as it does not retrieve the GPS coordinates of the hotspots. That inability complicates distinguishing the overlapping hotspots since it should be done merely by visiting streets one by one, visually, whereas the suggested method performs that with a code, fast and precisely.

5.3 Hotspot Categorization Results

The clustering analysis was performed. Table 5 shows an example of the generated MinPts values for the 2014, 2015 period where Figure 15 shows an example of the identified crash clusters on OHIO RIVER BL(Lower)-S road with 8 clusters and 13 noise points.

Table 5: An example of the generated MinPts values for the 2014, 2015 period.

Street	**Length**	**Counts**	**Mean**	**Minpts**
ARLINGTON AV	1800	31	1.7	3
BABCOCK BL	6000	133	2.2	4
BAKERSTOWN RD	15540	22	0.1	2
BANKSVILLE RD-N	4340	36	0.8	2
BANKSVILLE RD-S	4340	32	0.7	2
BAPTIST RD	5050	29	0.6	2
BAUM BL	2940	69	2.3	4
BEAVER GRADE RD (Lower)	9000	16	0.2	2
BEAVER GRADE RD (Middle)	420	8	1.9	4
BEAVER GRADE RD (Upper)	1740	21	1.2	3
…				

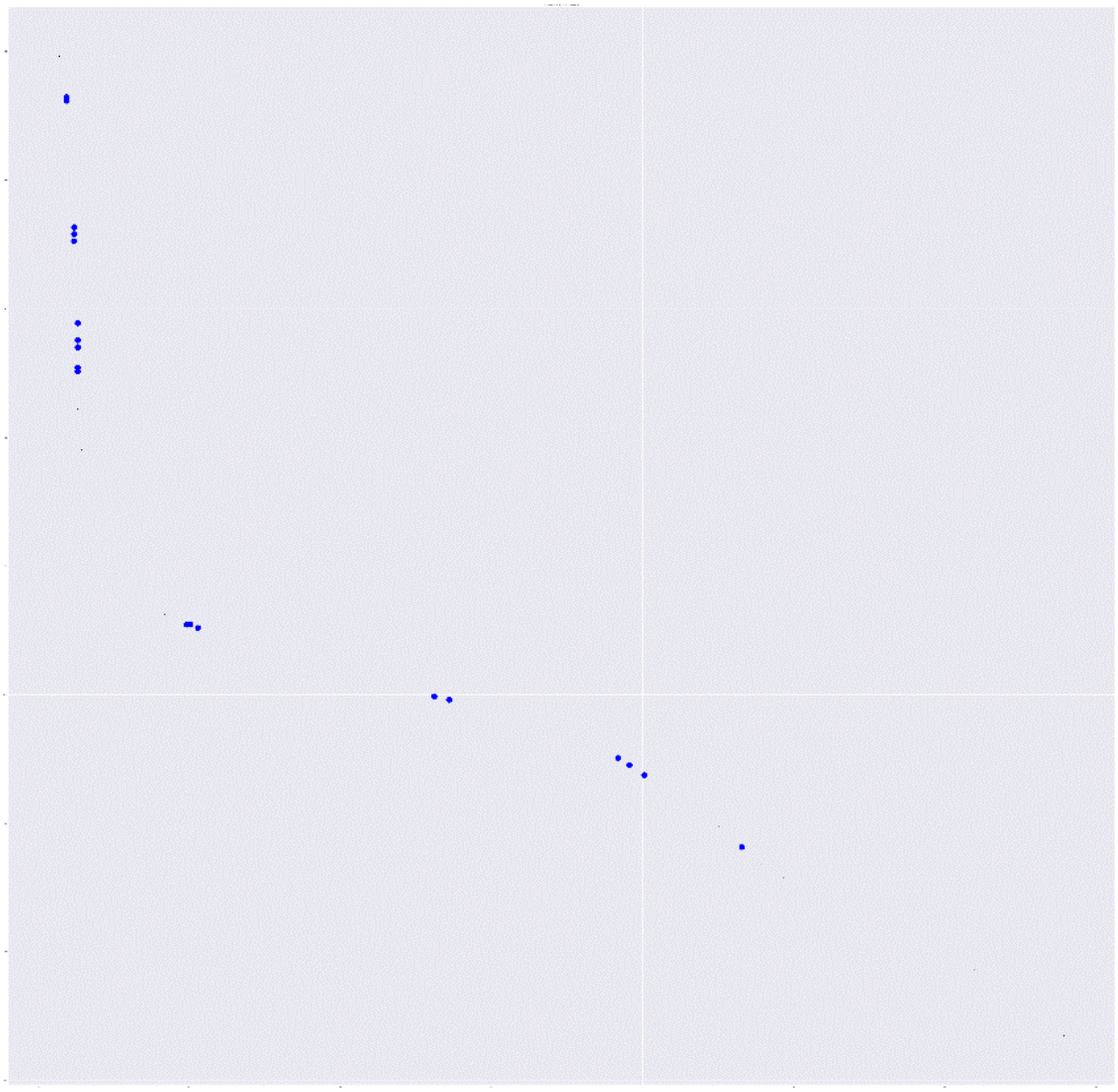

Figure 15: An example of the identified crash clusters by DBSCAN (output of Python).

The two groups of hotspots were obtained, those based on the yearly-aggregated crashes and those by the categorization algorithm (based on the eight defined periods), comprising the eight sets of periodic hotspots and the one set of U hotspots. Table 6 presents an example of the identified hotspots and their details, pertaining to the 2016, 2017 period.

Table 6: An example of the identified hotspots and their details.

Rank	Latitude	Longitude	Number of points in cluster	Cluster length	Cluster scaled density
1	40.49422	-79.9037	24	191	10.522
2	40.42745	-79.9525	21	283	8.565
3	40.41614	-80.0588	17	118	8.205
4	40.42943	-80.0285	19	273	7.799
5	40.41925	-80.0524	17	218	7.27
6	40.42909	-79.9443	18	306	7.241
7	40.45702	-79.8778	10	35	6.476

Out of the 415 yearly-based hotspots, 195 of them matched at least one periodic or U hotspot. The rest 220 did not overlap with any periodic hotshots; hence, they were categorized as YNP (Table 7).

Table 7: Counts of each hotspot category in the total 415 yearly hotspots.

Locations that became hotspots only at a specific season				Locations that became hotspots only at a specific time of day				Locations that became hotspots at more than one specific period	Locations that became hotspots at one specific season and one specific time-of-day	Locations that became hotspots without any temporal pattern	
WE	SpE	SuE	FE	ME	DE	EE	NE			Unbounded (U)	Yearly non-periodic (YNP)
23	14	7	10	11	22	33	9	36	24	6	220

That was done by projecting both mentioned hotspot groups to ArcMap (doable by Google Earth as well) and checking overlaps between them (Figure 16).

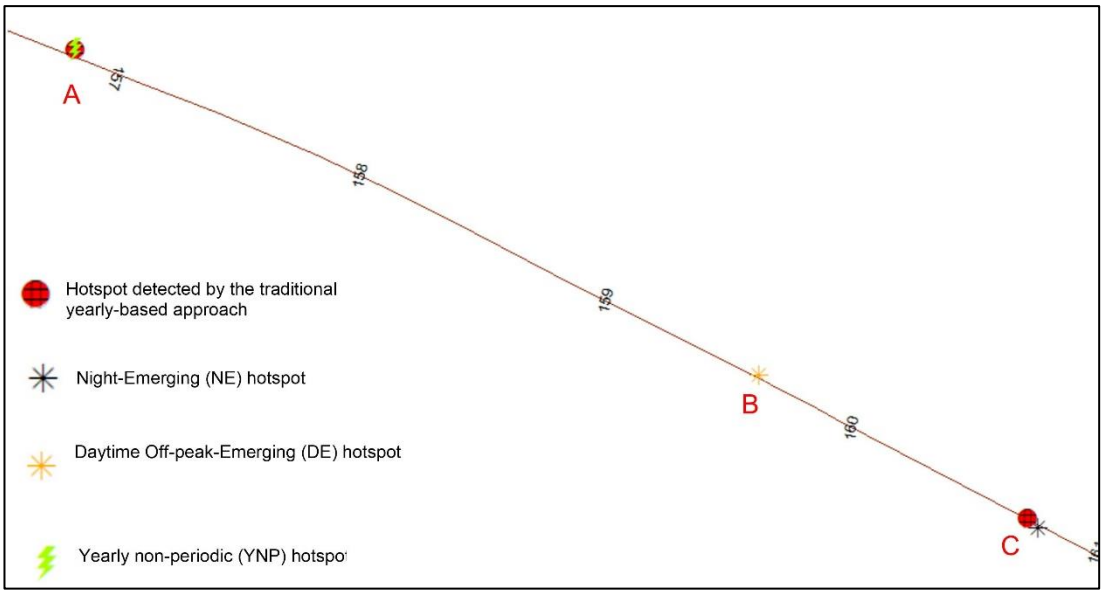

Figure 16: Point A is a YNP hotspot; point B is a DE hotspot, which failed to be detected by the traditional approach, and point C is a NE hotspot.

Table 8 illustrates a few examples of the categorized hotspots.

Table 8: A few examples of the categorized hotspots.

Longitude	Latitude	Time-of-day of emerging (if any)	Season of emerging (if any)	U or YNP
-79.7608	40.44108		WE	
-80.1221	40.54138	EE		
-79.9555	40.41103			YNP
-79.9019	40.4937	ME		
-79.9886	40.4287	NE		
-79.9359	40.4285	NE	FE	
-79.99	40.45662			YNP
-80.1928	40.46167		WE	
-80.0088	40.44082	NE		
-80.0641	40.49402			YNP
-79.9574	40.4303			U

The results reveal that, 30 periodic hazardous points, identified by the categorization algorithm, could not be detected by the yearly-based HSID. That finding challenges

the sufficiency of the yearly-based HSID approach. Figure 17 illustrates the summary of the method validation by the case study.

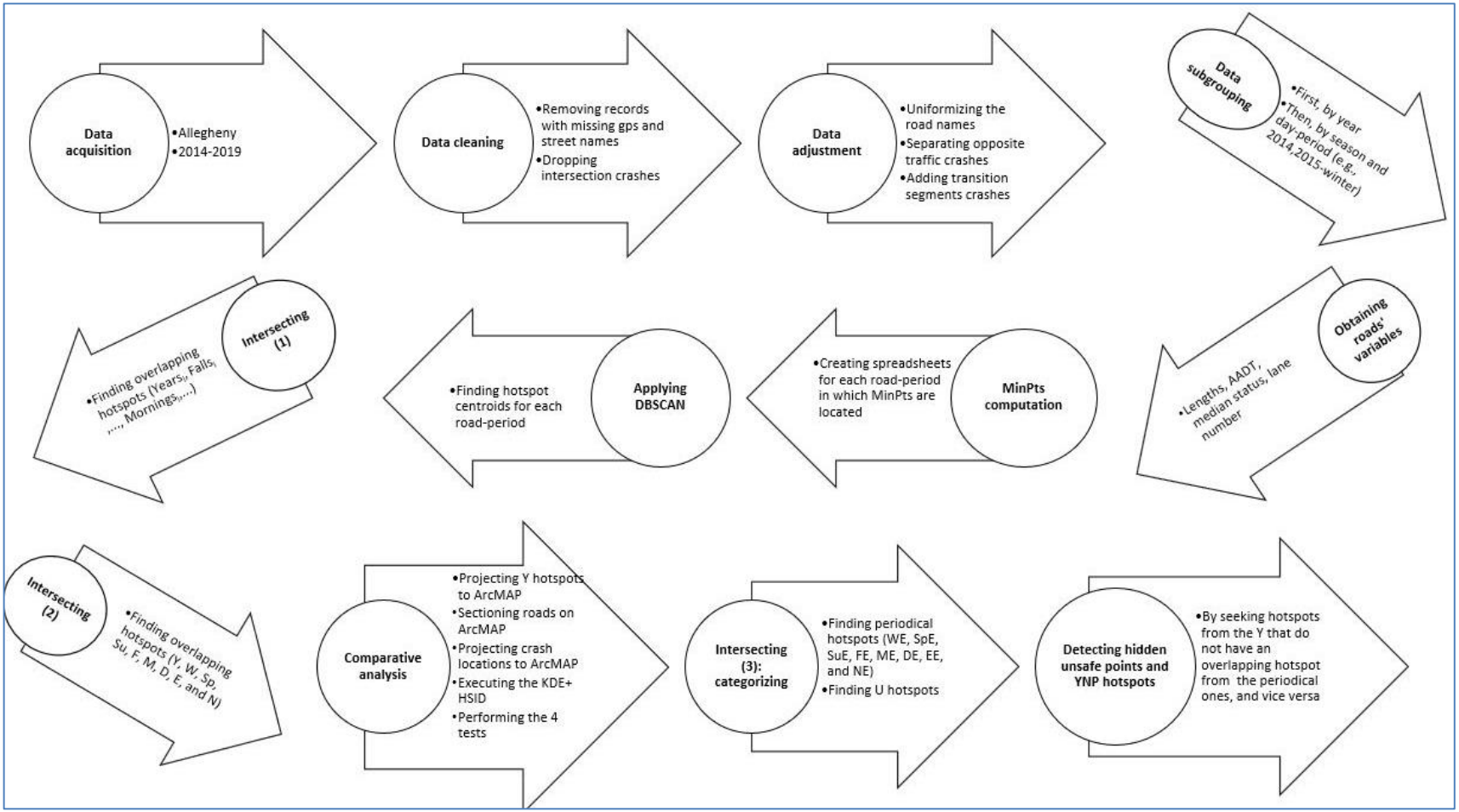

Figure 17: Summary of the method validation process by the case study.

Chapter 6

DISCUSSION

To illustrate how the hotspot categorization enhances the diagnosis process, the result of the proposed approach is compared with that of the previous studies': a yearly-based hotspot (not categorized yet) with center-point 40.4617, -80.1928 is chosen. According to the traditional diagnosis manner, applied in AASHTO (2010) and previous studies such as the paper of Park & Sahaji (2013), the collision type configurations related to all the 6-year accumulated crashes at this location were observed, and the dominant one were recognized as the problematic crash type for which safety measurements were implemented. The crash composition is 40% hit-fixed-object, 27% sideswipe (same direction), 20% rear-end, 6.5% angle, and 6.5% other or unknown. Among them, the hit-fixed-object crash type is the dominant; hence, it directs the safety expert to seek the problem in the list of the possible causes corresponding to that crash type. However, according to the suggested approach in this study, after applying the categorization algorithm, that point is classified as a Winter Emerging (WE) hotspot, meaning that it becomes unsafe merely during winters. Hence, the winter crashes were isolated from the other three seasons' crashes and the descriptive data analyses of the two groups were drawn separately. The results showed that, the overrepresented collision type for the winter crashes were sideswipe (same direction) with 57% proportion, while that of the other three seasons were hit-fixed-object (63%) (Figure 18). This observation implies that, the unknown hazard, which turns that location into a hotspot in winters, mostly causes the sideswipe crash rather than hit-fixed-object,

which differs from the result of the previous approach. Thereby, this new finding directed the diagnosis procedure into the right path. It is believed that this reasoning and its resulting findings, which could be generalized to the other hotspots, would possibly revise the diagnosis and pave the way for more efficient safety improvement. Because the new approach's results (the inferred problematic crash patterns) were directly deduced from the problem-arising period, thus, they were more meaningful, while the results of the traditional attitude potentially convey ambiguous info from a mixture of random and patterned crash types.

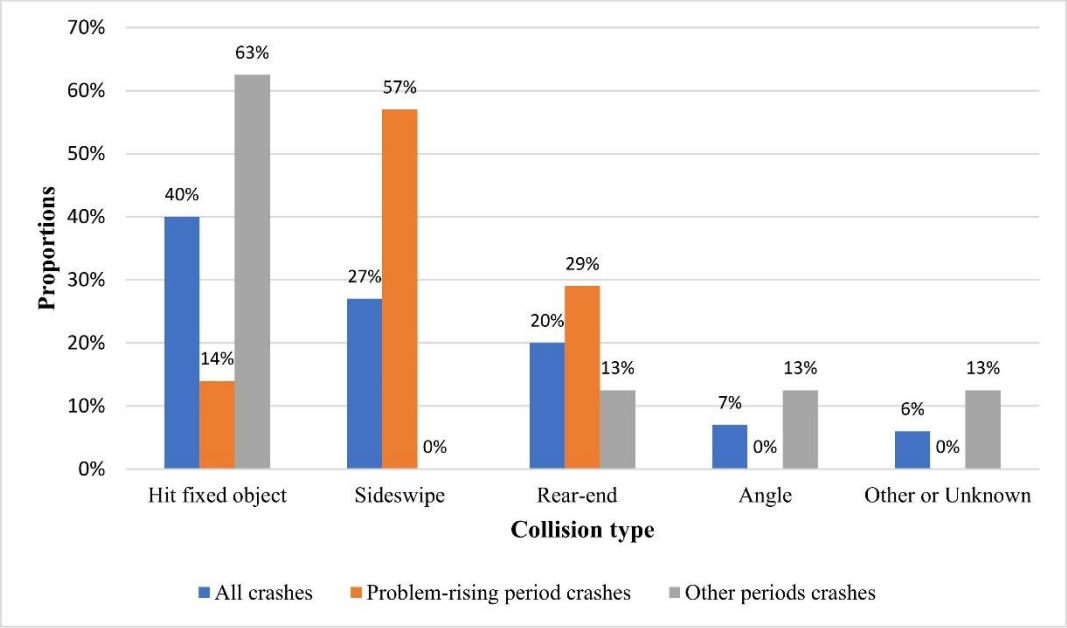

Figure 18: The disparity between collision patterns of the problem-arising period and other periods (Babaei & Kunt, 2023).

Further examples from the categorized hotspots in this study are exhibited in Table 9 to illustrate how categorizing can reveal hidden information and build more insights to enhance or even correct the diagnosis process.

Table 9: Examples for indicating the role of categorizing in drawing the correct inference about causative risk factor.

Hotspot location	Traditional approach: inference based on all aggregated crashes		The new approach: inference based on the category of the hotspot			
	Descriptive data analysis based on all crash records	**Inference about the safety problem: what crash patterns does it tend to cause?**	**Labeled Category**	**Descriptive data analysis for crash records of the problem-arising period**	**Descriptive data analysis for crashes of the other three periods**	**Inference about the safety problem: when does it emerge, and what crash patterns does it tend to cause?**
40.4547, -79.9461	Crash type configuration: Angle: 43% Rear-end: 33% Hit pedestrian: 10% Sideswipe (same dir.): 9% Head-on: 5%	Angle crash	WE	Rear-end: 50% Angle: 37.5% Sideswipe (same dir.): 12.5%	Angle: 46% Rear-end: 23% Hit pedestrian: 15% Head-on: 8% Sideswipe (same dir.): 8%	Winters; Rear-end crash
40.4287, -79.9886	Illumination configuration: Daylight: 56% Dark-street lights: 44%	At daylight condition	NE	Dark-street lights: 100%	Daylight: 100%	Nighttime, in dark-street lights condition.
40.3408, -79.9646	Involved vehicles configuration: Automobile: 60% Truck: 33% Bus: 7%	Threatening automobiles	SpE	Truck: 50% Bus: 25% Automobile: 25%	Automobile: 73% Truck: 27%	Springs; threatening heavy vehicles

To diagnose the problem, having the achieved information about the hotspot categories and the crash feature patterns in hand, a site inspection plan, necessarily in the problem-arising period, is indispensable. The reason is that, the available information in accident databases may not span the full spectrum of crash variables; also, the causes of the hotspot formation may be unique circumstances non-explainable in databases. Some imaginable examples for such scenarios are presented in Table 10. Additional or alternative policies, depending on the available resources, may include installing cameras equipped with scheduled activation programs to monitor the trajectory of vehicles at the targeted periods and analyse drivers' behaviours, particularly in case of near miss incidents as a surrogate safety measure, and installing warning signs to warn

the drivers about the hotspots' location and their hazard times. Table 11 shows some instances of hotspots where the actual problem-arising periods are not distinguishable.

Table 10: Imagined examples of localized environmental risks that arise at specific periods and make a location unsafe, which are unrecordable in data or are undistinguishable from descriptive statistics.

Category	Examples of possible contributing factors	Recordable in crash data (does it have any representative attribute in dataset)?	Distinguishable by descriptive statistics of crash data without knowing the category of the hotspot?	Distinguishable by site visit at periods other than the problem-arising period?
EE	Task-overload as the traffic volume is high	No	N/A	No
NE	An object/place near the road that becomes distractive only in nighttime	No	Not always	No
SuE	Hazards of farmers' activities at a farm near the road in harvest period in summers	No	N/A	No
FE	Hazards of a school zone at the beginning of the academic year	No	N/A	No

Table 11: Examples of hotspots where the problem-arising period is indistinguishable.

Hotspot location	Share of each period in total crashes	Apparent critical period	Category (actual critical period)
40.4285, -79.9368	winter 35% fall 29% spring 24% summer 12%	Winter	FE
40.3408, -79.96462	winter 40% spring 27% summer 27% fall 6%	Winter	SpE
40.44082, -80.00882	morning 44% night 22% evening 22% daytime off-peak 11%	Morning	NE

As another example, a non-periodic hotspot (Figure 19) is chosen. The most frequent crash type at this location is hit-fixed-object (64%). The potential contributing factors to the occurrence of that crash type, according to the Haddon matrix (HSM), are presented in Figure 20. That location is a U hotspot, which signifies its independence from the environmental and traffic conditions; thus, the contributing factors from those two factor groups (i.e., inadequate lighting and slippery pavement) should be rejected; that helps the diagnosis procedure as the list of the potential contributing factors becomes narrower. Resultantly, by the visual assessment of aerial photos, the diagnosed problem seems to be an inadequate clear distance due to the insufficient setback distance of the roadside cluttering objects (the trees).

Figure 19: A hotspot categorized as U type (Babaei & Kunt, 2023).

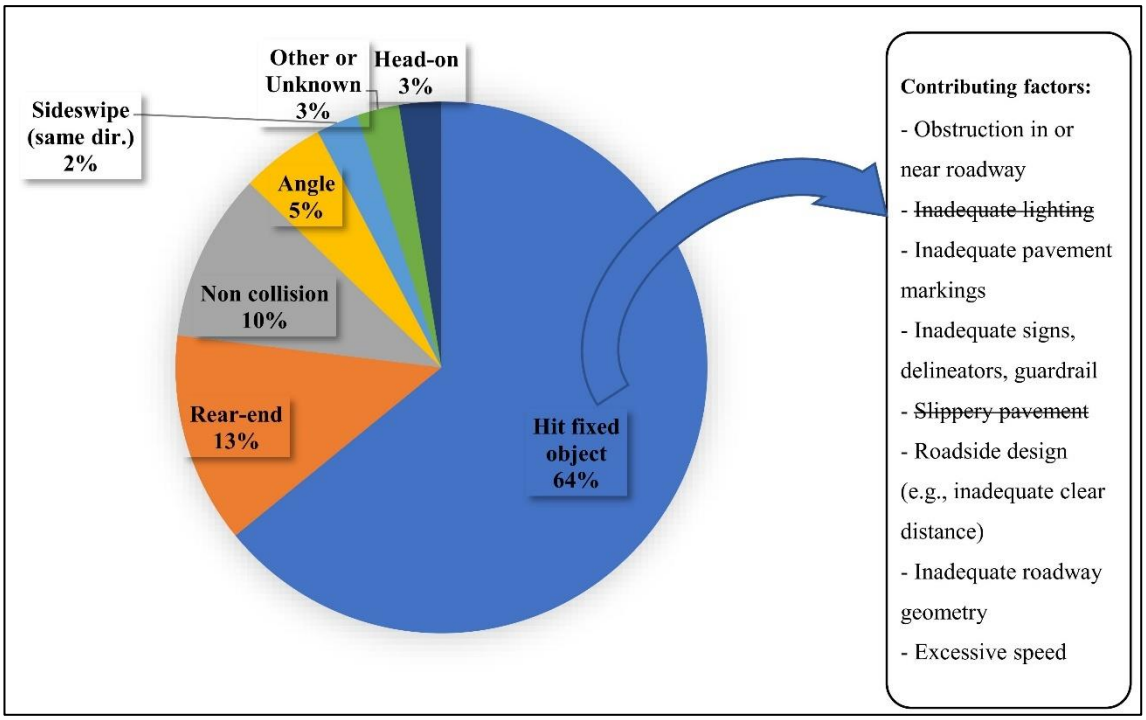

Figure 20: The crash type composition and the potential contributing factors for the dominant crash type, hit fixed object, (as per Haddon matrix) for the investigated U hotspot (Babaei & Kunt, 2023).

Chapter 7

CONCLUSION

This study was conducted to fill the gaps in the previous spatio-temporal studies to contribute to the two principal phases of traffic safety improvement: identifying hazardous locations and diagnosing the unsafety causes at those places. In most previous works, the spatial level was 2-D, potentially causing the inclusion of discrete adjacent segments in the identified hotspot, making it complicated to pinpoint the exact position of the faulty points. Few papers restricted the spatial level from 2-D to 1-D; however, either their hotspot categorization manner or their considered time dimensions were incompatible with the objective aimed in this work. The criterion of the proposed tempo-categorization was the reappearance of unsafe points at specific periods, for multiple consecutive years, from two time-segmentations perspective, time of day and season of year. In this type of analysis, as the HSID is applied to more restricted time segments, two benefits are achieved, as demonstrated: first, the potentially hidden unsafe locations are detected, and second, the problem-arising period of the hazardous points can be understood, which would contribute to a more accurately problem diagnosis progression as the reasoning will be over the relevant crashes. That may prevent implementing improvement plans on the wrongly diagnosed problems and instead, optimize the allocation of resources.

A prerequisite for the planned categorization sketch was a compatible linear HSID tool. Therefore, a DBSCAN-based 1-D HSID tool coded in Python was suggested. The

highlights of this tool include a new way of determining the MinPts value and the ability to analyze unlimited roads in a relatively short time without any observed errors. Compared with the previous DBSCAN modification ideas, this one does not need specific additional data like network layout or a threshold for clusters' length. Contrasted with KDE+, the acknowledged clustering tool, the precise clustering performance and other advantages of the suggested method was demonstrated; thus, it was found appropriate for the intended categorization action. The proposed approach and developed code are globally applicable with free-of-charge requirements.

A potential improvement for the proposed methodology concerns the criteria considered for ranking hotspots. The considered criterion was solely the crash counts. However, an additional criterion may be the severity level of accidents where in the ranking process, a weight (coefficient) would be assigned to each severity level (i.e., non-injury, moderate injury, severe injury and fatal).

REFERENCES

AASHTO. (2010). Highway Safety Manual, 1st ed. Washington, D.C., USA.: American Association of State Highway and Transportation Officials.

AASHTO. (2001). A policy on Geometric Design of Highways and Streets. American Association of State Highway and Transportation Officials (AASHTO).

Afghari, A. P., Haque, M. M., & Washington, S. (2020). Applying a joint model of crash count and crash severity to identify roadsegments with high risk of fatal and serious injury crashes. *Accident Analysis and Prevention, 144*. doi:https://doi.org/10.1016/j.aap.2020.105615.

Al Hamami, M., & Matisziw, T. C. (2021). Measuring the spatiotemporal evolution of accident hot spots. *Accident Analysis And Prevention, 157*. doi:10.1016/j.aap.2021.106133

Al-Ruzouq, R., Hamad, K., Abu Dabous, S., Zeiada, W., Khalil, M. A. (2019). Weighted multi-attribute framework to identify freeway incident hot. *Arabian Journal for Science and Engineering*, 44:8205–8223.

Ambros, J., Havránek, P., Valentová, V., Křivánková, Z., & Striegler, R. (2016). Identification of hazardous locations in regional road network – comparison of reactive and proactive approaches. *Transportation Research Procedia, 14*, 4209-4217.

Babaei, Z., & Kunt, M. M. (2023). Tempo-categorization of road accident hotspots to enhance the problem diagnosis process and detect hidden hazardous locations. *Journal of Transportation Safety & Security*. https://doi.org/10.1080/19439962.2023.2169800

Bandyopadhyaya, R., & Mitra, S. (2015). Fuzzy cluster–based method of hotspot detection with limited information. *Journal of Transportation Safety & Security, 7*, 307-323.

Benedek, J., Ciobanu, S. M., & Man, T. C. (2016). Hotspots and social background of urban traffic crashes: A case studyin Cluj-Napoca (Romania). *Accident Analysis and Prevention, 87,* 117–126.

Bham, G. H., Manepalli, U. R. R., & Samaranayke, V. A. (2017). A composite rank measure based on principal component analysis for hotspot identification on highways. *Journal of Transportation Safety & Security, 11*, 225:242.

Bíl, M., Andrásik, R., & Janoska, Z. (2013). Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. *Accident Analysis and Prevention, 55*, 265–273.

Bíl, M., Andrášik, R., & Sedoník, J. (2019). A detailed spatiotemporal analysis of traffic crash hotspots. *Applied Geography*, 107, 82-90.

Bíl, M., Andrasik, R., Svoboda, T., & Sedonık, J. (2016). The KDE+ software: a tool for effective identification and ranking of animal-vehicle collision hotspots

along networks. *Landscape Ecol, 31*, 231–237.

Borah, B., & Bhattacharyya, D. K. (2004). An improved sampling-based DBSCAN for large spatial databases. *International Conference on Intelligent Sensing and Information Processing*, (pp. 92-96).

Chen, H., Tao, F., Ma, P., Gao, L., & Zhou, T. (2021). Applicability evaluation of several spatial clustering methods in spatiotemporal data mining of floating car trajectory. *Geo-Information, 10*. doi:https://doi.org/10.3390/ijgi10030161

Cheng, W., & Jia, X. (2015). Exploring an alternative method of hazardous location identification: using accident count and accident reduction potential jointly. *Journal of Transportation Safety & Security*, *7*(1), 40-55.

Cheng, W., & Washington, S. (2008). New criteria for evaluating methods of identifying hot spots. *Transportation Research Record: Journal of the Transportation Research, 37*, 870-881.

Cheng, Z., Zu, Z., & Lu, J. (2019). Traffic crash evolution characteristic analysis and spatiotemporal hotspot identification of urban road intersections. *Sustainability, 11*. doi:10.3390/su11010160

Dong, N., Huang, H., Lee, J., Gao, M., & Abdel-Aty, M. (2016). Macroscopic hotspots identification: A Bayesian spatio-temporal interaction approach. *Accident Analysis and Prevention, 92*, 256-264.

Edwards, J. B. (1998). The relationship between road accident severity and recorded weather. *Journal of Safety Research, 29*, 249-262.

Elvik, R. (2007). State-of-the-art approaches to road accident black spot management and safety analysis of road networks. Oslo, Norway: Institute of Transport Economic (TOI).

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-based algorithm for discovering clusters in large spatial databases with noise. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, (pp. 226–231). Munchen.

Fox, L., Serre, M. L., Lippmann, S. J., Rodríguez, D. A., Bangdiwala, S. I., Gutiérrez, M. I., Escobar, G., Villaveces, A. (2015). Spatiotemporal approaches to analyzing pedestrian fatalities: the case of cali, colombia. *Traffic Injury Prevention, 16*, 571-577.

Ghadi, M., & Török, A. (2017). Comparison different black spot identification methods. *Transportation Research Procedia, 27*, 1105–1112.

Ghadi, M., & Török, A. (2019). A comparative analysis of black spot identification methods and road. *Accident Analysis and Prevention, 128*, 1-7.

Gregoriades, A., & Chrystodoulides, A. (2018). Extracting traffic safety knowledge from historical accident data. *14th International Conference on Location Based Services (LBS 2018)* (pp. 109 - 114). Zurich, Switzerland: ETH Library.

Gregoriades, A., & Mouskos, K. C. (2013). Black spots identification through a Bayesian Networks. *Transportation Research Part C*, 28-43.

Gudes, O., Varhol, R., Sun, Q., Meuleners, L. (2017). Investigating articulated heavy-vehicle crashes in Western Australia using a. *Accident Analysis and Prevention*, 243-253.

Hauer, E. (2005). Observational before/after studies in road safety: estimating the effect of highway and traffic engineering. Amsterdam: Pergamon.

Hauer, E., Kononov, J., Allery, B., & Griffith, M. S. (2002). Screening the road network for sites with promise. *Transportation Research Record Journal of the Transportation Research Board*, 27-32.

Hauer, E., NG, J. C. N., & Lovell, J. (1988). Estimation of safety at signalized intersections. *Transportation Research Record*.

Hussain, M. S., Goswami, A. K., & Gupta, A. (2022). Predicting pedestrian crash locations in urban India: An integrated GIS-based spatiotemporal HSID technique. *Journal of Transportation Safety & Security*. doi:10.1080/19439962.2022.2048759

Ivan, J. N., Persaud, B., Srinivasan, R., Abdel-Aty, M., Lyon, C., Mamun, S., Lee, J., Farid, A., Wang, J. H., Lan, B., Smith, S., Ravishanker, N., & Saleem, T. . (2017). Improved prediction models for crash types and crash severities. Washington, D.C.: Transportation Research Board of the National Academies.

Kaygisiz, Ö., Düzgün, S., Yildiz, A., & Senbil, M. (2015). Spatio-temporal accident analysis for accident prevention in relation to behavioral factors in driving: The case of South Anatolian Motorway. *Transportation Research Part F, 33*, 128–140.

Kononov, J., Allery, B., & Znamenacek, Z. (2019). Safety planning study of urban Freeways Proposed Methodology and Review of Case History. *Transportation Research Record: Journal of the Transportation Research Board*, 146-155. doi:10.3141/2019-18

Kononov, J., Williams, J., & Durso, C. (2020). A closer look at highway safety diagnostics and crash analysis. *Transportation Research Record: Journal of the Transportation Research Board, 2674*(5), 1-11.

Kononov. J., & Janson, B. N. (2002). Diagnostic methodology for the detection of safety problems at intersections. *Transportation Research Record: Journal of the Transportation Research Board, 184*, 51-56.

Kumar, S. & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. *Journal of Modern Transportation, 24*(1), 62-72.

Le, K. G., Liu, P., & Lin, L. T. (2019). Determining the road traffic accident hotspots using gis-based temporal-spatial statistical analytic techniques in hanoi, vietnam. *Geo-spatial Information Science*. doi:10.1080/10095020.2019.1683437.

Manepalli, U. R. R., & Bham, G. H. (2016). An evaluation of performance measures for hotspot identification. *Journal of Transportation Safety & Security, 8*, 327-345.

Montella, A. (2010). A comparative analysis of hotspot identification methods. *Accident Analysis and Prevention, 42*, 571–581.

Olsen, J. R., Mitchell, R. & Ogilvie, D. (2017). Effect of a new motorway on social-spatial patterning of road traffic accidents: A retrospective longitudinal natural experimental study. *PLoS ONE, 12*(9). doi:https://doi.org/10.1371/journal.

Ouni, F. & Belloumi, M. (2018). Spatio-temporal pattern of vulnerable road user's collisions hot spots and related risk factors for injury severity in Tunisia. *Transportation Research Part F, 56,* 477–495.

Park, P. Y., & Sahaji, R. (2013). Safety diagnosis: Are we doing a good job? *Accident Analysis and Prevention, 52,* 80-90.

Pluga, C., Xiab, J., Caulfield, C. (2011). Spatial and temporal visualisation techniques for crash analysis. *Accident Analysis and Prevention, 43,* 1937–1946.

Qiu, C., Xu, H., & Bao, Y. (2016). Modified-DBSCAN clustering for identifying traffic accident prone locations. *Traffic Management Research Institute of the Ministry of Public Security, Wuxi 214151, Jiangsu, China,* 99-105. doi:10.1007/978-3-319-46257-8_11.

Szénási, S., & Jankó, D. (2017). A method to identify black spot candidates in built-

up areas. *Journal of Transportation Safety & Security, 9*, 20-44.

Thomas, I. (1996). Spatial data aggregation: exploratory analysis of road accidents. *Accid. Anal. and Prev., 28,* 251-264.

Wang, K., Zhao, S., Ivan, J. N., Ahmed, I., & Jackson, E. (2020). Evaluation of hot spot identification methods for municipal roads. *Journal of Transportation Safety & Security, 12*(4), 463-481.

Wang, X., Qu. X., & Jin, S. (2019). Hotspot identification considering daily variability of traffic flow and crash record: A case study. *Journal of Transportation Safety & Security, 12*(2), 275-291. doi:10.1080/19439962.2018.1477893.

World Health Organization (WHO) (2018). Global status report on road safety.

Wu, P., Meng, X., & Song, L. (2021). Identification and spatiotemporal evolution analysis of high-risk crash spots in urban roads the microzone-level: Using the space-time Cube Method. *Journal of Transportation Safety & Security*. doi:10.1080/19439962.2021.1938323.

Xie, K., Ozbay, K., Kurkcu, A., & Yang, H. (2017). Analysis of Traffic crashes involving pedestrians using big data investigation of contributing factors and identification of hotspots. *Risk analysis, 37*(8), 1459-1476. doi:DOI: 10.1111/risa.12785

Xie, Z. & Yan, J. (2008). Kernel density estimation of traffic accidents in a network

space. *Computers, Environment and Urban Systems, 32,* 396–406.

Xu, Q. & Tao, G. (2018). Traffic Accident hotspots identification based on clustering ensemble model. *2018 4th IEE International Conference On Edge Computing And Scalable Cloud (Ieee Edgecom 2018)*, (pp. 1-4).

Yakar, F. (2021). A multicriteria decision making based methodology to identify accident prone road sections. *Journal of Transportation Safety & Security, 13*(2), 143-157.

Yoon, J. & Lee, S. (2021). Spatio-temporalpatterns in pedestrian crashes and their factors Application of a space-timecube analysis modeldetermining. *Accident Analysis and Prevention, 161*.

Zahran, E. M. M., Tan, S. J., Tan, E. H. A., Amirah, N., Mohamad, A, B., Putra, A., Yap, Y. H., & Abdul Rahman, E. K. (2019). Spatial analysis of road traffic accident hotspots evaluation and validation of recent approaches using road safety audit. *Journal of Transportation Safety & Security*. doi:10.1080/19439962.2019.1658673.

Zhang, Y., Han, L. D., & Kim, H. (2018). Dijkstra's-DBSCAN: fast, accurate, and routable density based clustering of traffic incidents on large road network. *Journal of Transportation Research Board, 2672,* 265-273.

APPENDIX

Table A1: Details of the chosen segments for the comparative analysis between the DBSCAN-based and the KDE+ methods

ON STREET NAME	length (m)	AADT	Divided/ Undiv	Lane # of each direction	Speed limit	Local/State/Interstate	Start and end point coordinates
ALLEGHENY RIVER BL (Lower)	4340	above 20k	U	1	45	S	40.484019, -79.907272 to 40.485161, -79.856914
ALLEGHENY RIVER BL (Middle)	2070	10k to 20k	U	1	45	S	40.4852, -79.8568 to 40.513966, -79.843215
ALLEGHENY VALLEY EX (Middle2) E	7000	above 20k	D	2	55		40.561625, -79.800184 to 40.608438, -79.762436
BABCOCK BL	6000	10k to 20k	U	1	35	S	40.506247, -79.990106 to 40.673616, -79.989552
BAUM BL	2630	5k to 10k	U	2	55	S	40.454004, -79.953291 to 40.460456, -79.925004
BIGELOW BL-E	3780	10k to 20k	D	2	35		40.441050, -79.993900 to 40.458703, -79.957737
BUTLER ST(North)	5050	10k to 20k	U	1	35	S	40.467554, -79.963762 to 40.486290, -79.913505
CLAIRTON BL	11650	10k to 20k	U	2	40	S	40.295999, -79.919002 to 40.376951, -79.985722
FRANKSTOWN RD (Upper)	3470	5k to 10k	U	2	55		40.460350, -79.842409 to 40.484137, -79.819273
FREEPORT RD	18030	10k to 20k	U	1	35		40.492700, -79.910092 to 40.54079, -79.80759
LIBRARY RD	10200	10k to 20k	U	1			40.311148, -80.032977 to 40.382044, -79.996461
PENN LINCOLN PY-Westbound	23410	above 20k	D	2	55		40.444554, -80.163539 to 40.426024, -79.896761

Table A2: Table for the comparative analysis

ALLEGHENY RIVER BL (Lower)													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	1	0					1	1	0				
2	8_9	3	*	6	3	3	2	8_9	3	*	4	3	1
3	26	1					3	26_27_28	9	*	2	2	0
4	27_28	8	*	2	1	1	4	32_33_34	2	*	3	8	-5
5	32_33	0							**14**	**3**			**-4**
6	33_34	2	*	4	4	0							
		14	**3**			**4**							

ALLEGHENY RIVER BL (Middle)													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	15_16	0					1	15_16	0				
2	24	0					2	24	2	*	2	6	-4
							3	26	0				
							4	33_34	1				
									3	**1**			**-4**

ALLEGHENY VALLEY EX (Middle2) E													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	1	0					1	1	0				
2	11	0					2	11	0				
3	16	1					3	16	1				
4	17	0					4	17	0				
5	19	0					5	19	0				
6	38	1					6	38	1				
7	46	3	*	4	1	3	7	46	1				
8	53	2	*	5	5	0	8	53	2	*	5	8	-3
9	59	1					9	58_59	1				
10	61	3	*	1	6	-5	10	61	3	*	1	1	0
11	63	2	*	6	2	4	11	63	2	*	7	6	1
		13	**4**			**2**			**11**	**3**			**-2**

BABCOCK BL													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	18	0					1	4_5	3	*	19	5	14
2	32_33	0					2	14	2	*	15	10	5
3	38	3					3	16	0				
4	62	2					4	17_18	3	*	2	4	-2
5	70	4	*	8	2	6	5	32_33	0				

6	91	0					6	38	3	*	12	9	3
		9	**1**			**6**	7	40	0				
							8	48	2	*	17	7	10
							9	61	1				
							10	62	2	*	4	12	-8
							11	65_66	1				
							12	70	5	*	8	1	7
							13	77-78	0				
							14	85-86	0				
							15	89_90	2	*	11	3	8
							16	91	0				
							17	93_94	2				
							18	99_100	1				
							19	119	0				
							20	179	0				
							21	181-182	0				
							22	207_208	1				
									28	**8**			**37**

BAUM BL													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	5	3	*	1	6	-5	1	4_5	4				
2	6_7	7	*	2	2	0	2	6_7	6	*	3	1	2
3	14_15	5	*	4	7	-3	3	12	5	*	4	3	1
4	18	3	*	3	1	2	4	14_15	5	*	5	2	3
5	20_21	3	*	5	4	1	5	18	3				
		21	**5**			**-5**			**23**	**3**			**6**

BIGELOW BL-E													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	1	1					1	1	1				
2	4_5	0					2	4_5	0				
3	28_29	3	*	1	2	-1	3	28_29	3	*	1	2	-1
4	30	4	*	5	1	4	4	30	4	*	6	1	5
5	34	1					5	34	1				
6	36	1					6	36	1				
		10	**2**			**3**			**10**	**2**			**4**

BUTLER ST(North)													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	6	3	*	10	4	6	1	11_12	1				
2	8	0					2	19	2				
3	9_10	1					3	20_21	1				
4	11_12	1					4	26	0				
5	19	2					5	34_35	0				
6	20_21	1					6	37	3	*	4	4	0
7	26	0					7	42_43	1				
8	34_35	0					8	49_50	7	*	1	2	-1
9	37	3	*	7	5	2	9	51_52	1				

10	42_43	1							**16**	**2**			**-1**
11	50	7	*	1	2	-1							
12	51	1											
		20	**3**			**7**							

CLAIRTON BL													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	3_4	1					1	3_4	1				
2	5	1					2	29_30	5	*	1	8	-7
3	21_22	0					3	36_37	1				
4	29_30	8	*	1	9	-8	4	38	2	*	15	12	3
5	36_37	1					5	39	2				
6	38	2					6	47_48_49	10	*	4	2	2
7	39	2					7	50_51	8	*	2	1	1
8	47_48_49	10	*	2	4	-2	8	52_53	2				
9	50_51	8	*	4	1	3	9	53_54_55	6	*	5	14	-9
10	52_53	2					10	60_61	4	*	3	6	-3
11	53_54	5	*	6	3	3	11	62	3	*	16	15	1
12	55_56	1					12	72_73	5	*	14	9	5
13	61	4	*	3	8	-5	13	78	1				
14	62	3	*	16	14	2	14	81_82	7	*	6	4	2
15	72_73	5	*	17	12	5	15	89_90	3				
16	74	2					16	93_94	1				
17	78	1					17	118_119	3	*	9	13	-4
18	81_82	7	*	5	5	0			**64**	**10**			**-9**
19	89_90	3											
20	93_94	1											

21	96_97	1				
22	108	0				
23	118_119	3	*	14	13	1
		71	**9**			**-1**

FRANKSTOWN RD (Upper)													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	3	1					1	3	1				
2	4_5	1					2	5	0				
3	13_14	1					3	13_14	1				
4	15_16	2					4	15	2				
5	18	4	*	7	4	3	5	18	4	*	6	2	4
6	21	1					6	21	1		5	1	4
7	22	4	*	4	1	3	7	22	4	*			
8	30_31	5	*	1	3	-2	8	31	4	*	1	5	-4
		19	**2**			**4**			**17**	**2**			**4**

FREEPORT RD													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	3	3	*	3	7	-4	1	3	3	*	8	8	0
2	10	3	*	10	6	4	2	9_10	3	*	13	7	6
3	16	2	*	10	5	5	3	16	2	*	14	15	-1

4	18	1					4	18	1				
5	19	0					5	19	0				
6	22_23	2	*	6	14	-8	6	22_23	2	*	4	19	-15
7	24	2	*	11	18	-7	7	24	2				
8	28	1					8	28	1				
9	39	1					9	39	1				
10	41	3	*	9	3	6	10	41	3	*	10	4	6
11	43	2	*	4	12	-8	11	42_43	3	*	3	11	-8
12	46_47	1	*	16	2	14	12	47	5	*	19	2	17
13	47	4	*	14	2	12	13	87	0				
14	54_55	1					14	88	0				
15	87	0					15	89_90	1				
16	88	0					16	96	3	*	9	10	-1
17	89_90	1					17	104	1				
18	96	3	*	3	9	-6	18	106_107	3	*	1	3	-2
19	104	1					19	109_110	2				
20	106	3	*	1	3	-2			**36**	9			**2**
21	109	1											
22	110	1											
		36	**11**			**6**							

LIBRARY RD													
DBSCAN							KDE+						
Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	1	0					1	14	0				
2	6_7	2					2	20	1				
3	13	4	*	18	10	8	3	26_27	4	*	5	8	-3
4	14	0					4		0				

5	20	1					5	52	0				
6	23	1					6	60	8	*	10	2	8
7	25_26	3	*	5	14	-9	7	65	10	*	14	3	11
8	26_27	4	*	6	8	-2	8	72_73	1				
9	28_29	1					9	78	5	*	2	1	1
10	42	0					10	80	7	*	1	1	0
11	46	0					11	88_89	9	*	8	6	2
12	49	2					12	96	2				
13	52	0					13	101	1				
14	60	8	*	4	9	-5	14	103	0				
15	61	2							**48**	**6**			**19**
16	63_64	9	*	17	5	12							
17	65	6	*	5	1	4							
18	72_73	1											
19	78	5	*	1	6	-5							
20	80	7	*	3	3	0							
21	82	0											
22	84_85	1											
23	88_89	9	*	8	2	6							
24	96-1	2											
25	96-2	0											
26	101	1											
27	103	0											
		69	**9**			**9**							

PENN LINCOLN PY-Westbound	
DBSCAN	KDE+

Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)	Cluster ID	Address	Crash# in P2 (SCT)	Repeated in P2 (MCT)	Rank in P1	Rank in P2	Rank difference (TRDT)
1	84_85	4	*	30	28	2	1	15	3	*	25	28	-3
2	92	0					2	82	0				
3	93_94	1					3	84_85	4	*	24	26	-2
4	95_96	5	*	7	25	-18	4	91_92	0				
5	98_99	2					5	93_94	1				
6	100	0					6	95_96_97	6	*	1	12	-11
7	101_102_103	6	*	2	16	-14	7	98_99	2				
8	126	0					8	100 to 103	6	*	8	8	0
9	127_128	4	*	13	27	-14	9	123	1				
10	132_133	2					10	126	0				
11	137	5	*	31	18	13	11	127_128	4	*	11	18	-7
12	149	2					12	132_133	2				
13	153_154	12	*	3	3	0	13	148_149	7	*	9	6	3
14	155_156	5	*	4	19	-15	14	153_154	12	*	7	3	4
15	160	10	*	21	7	14	15	155_156	7	*	2	21	-19
16	161_162	4					16	160	10	*	19	7	12
17	168	1					17	167_168	3				
18	170_171	1					18	170_171	1				
19	174	2					19	178	3				
20	176_177	0					20	195_196_197	8				
21	178	3					21	207	5	*	15	10	5
22	195	4					22	208 to 212	22	*	6	1	5
23	196_197	3	*	12	26	-14	23	214_215	5				
24	207	1					24	218	2	*	20	2	18
25	208 to 213	26	*	1	1	0	25	219_220	17	*	22	2	20
26	214_215	6	*	24	10	14	26	223_224	1				
27	218	2	*	15	2	13	27	237	5				
28	219_220	18	*	18	2	16	28	238_239	3	*	5	22	-17

29	223_224	4	*	21	29	-8	29	240_241	14	*	23	22	1
30	237	5	*	11	8	3			**154**	**15**			**9**
31	238_239	3	*	5	4	1							
32	240_241	12	*	14	13	1							
		153	**17**			**-6**							