IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Examining Network Segmentation for Traffic Safety Analysis with Data-Driven Spectral Analysis

Xi Zhao[1,2], Dominique Lord[3], and Yichuan Peng[4]
[1]Wuhan University of Technology, Wuhan, Hubei 430070, China
[2]Clemson University, Clemson, SC 29630, USA
[3]Texas A&M University, College Station, TX 77840, USA
[4]Tongji University, Shanghai 201804, China

Corresponding author: Yichuan Peng (E-mail: muderx@gmail.com).

This work was supported in part by the Center for Connected Multimodal Mobility Grant, the Fundamental Research Funds for the Central Universities (WUT: 2019IVA046), and the National Natural Science Foundation of China (71601143).

ABSTRACT Network segmentation is foundational and critical to traffic safety analysis. Existing approaches to conduct segmentation require engineering judgement and are subject to a lack of standard metrics for assessing segmentation performance. This paper presents a novel methodology for data-driven analytics of crash distribution, crash aggregation, and network segmentation. It provides general solutions to determine optimal segment lengths for rigorous safety analysis, and extends the knowledge of crash distribution and aggregation for innovating segment-based safety analysis. The methodology is based on a redesigned spectral analysis of crash density in the spatial frequency domain (SFD) in which frequency components represent the natural patterns how crashes occur along roadways. By proposing the one-dimensional spatial frequency domain analysis (SFDA), this paper reveals the characteristic of power spectral concentration within the low frequency band for crash distribution. Based on this finding, this paper further proposes the power spectral segment length (PSSL) for determining optimal segment lengths and the power spectral percentage (PSP) for assessing the segmentation performance. Based on those new concepts and inferences, the paper proposes the low-pass filtering (LPF) method that outperforms the sliding window (SW) method, and the improved wavelet-based method that identifies high-risk segments properly. Those new techniques are easy to implement and ready for practical application. This research illustrates that interdisciplinary and innovative analytics combined with high-quality data collected by intelligent transportation infrastructure can reshape the fundamental knowledge and conventional paradigms in traffic safety.

INDEX TERMS Traffic safety, safety analysis, crash distribution, segmentation, segment length, spectral analysis

I. INTRODUCTION

Motor vehicle crashes are among the leading causes of death in the United States [1]. To improve safety performance of highway facilities, researchers conduct safety analysis based on crash and roadway data for highway planning, design, operations, and maintenance. From the perspective of data science, deep understanding of those safety data is always foundational to conducting successful safety research and practice.

Segment-based safety analysis approaches have been widely implemented in the safety realm based on the commonly accepted postulate that crashes are not fully randomly distributed but are highly dependent on the geometric attributes and traffic condition [2]. By segmenting the network into either homogeneous or heterogeneous segments and establishing the relationship between crash data and roadway characteristics, researchers and engineers can trivially conduct safety analysis tasks, such as network screening, diagnosis, countermeasure selection, crash predication, and the development and application of crash modification factors (CMFs) [3]. Different segmentation approaches (such as fixed-length and variable-length) have their own advantages, but they all descriptively and discretely quantify the spatial distribution of crashes with specified safety measurements (such as crash frequency, crash rate, or equivalent property damage only). Improper segmentation can degrade the analysis on spatial characteristics of crashes, and may lead to misleading conclusions.

Determination of appropriate segmentation strategies is one of the long-standing problems researchers and practitioners frequently confront in traffic safety analysis. Thus, several critical questions arise: "Which segmentation method is effective for safety analysis tasks?" "What is the optimal segment length?" "How should crash data be aggregated?" Either short segments or long segments can cause bias towards certain measurements that lead to flawed outcomes of safety analysis. Professionals are often required to use engineering judgment to conduct segmentation, and many of them arbitrarily choose empirical methods and segment lengths or simply refer to case study in the Highway Safety Manual (HSM) [3].

In recent years, the Model Minimum Uniform Crash Criteria (MMUCC) was developed to provide a guideline for standardizing the minimum set of uniform data elements of motor vehicle crash data [4]; and the Model Inventory of Roadway Elements (MIRE) was developed to provide a guideline for standardizing the structure of roadway inventory data [5]. Meanwhile, technical advances in automated crash data collection, reporting and processing have significantly improved accuracy of crash data and prevented several systematic errors and erroneous inputs [6]. With the advent of rich standardized safety data that contains comprehensive and reliable information, sophisticated data-driven technologies become feasible to explore safety issues and support reasonable decision-making.

This paper examines the knowledge and techniques of network segmentation using interdisciplinary concepts and analytics which are driven by reliable crash and roadway data. The research explores the spatial distribution of accurately geocoded crashes on continuous linear features of roadways in the one-dimensional (1-D) spatial domain and the corresponding spatial frequency characteristics in the spatial frequency domain (SFD). Based on the natural characteristics of crash distributions revealed by experiments, this paper proposes several new concepts and inferences for improving segmentation methods to support rigorous safety analysis. With those new tools, the paper further proposes several techniques that outperform or improve conventional techniques in traffic safety analysis.

II. BACKGROUND

A. SEGMENTATION APPROACHES

The American Association of State Highway Transportation Officials (AASHTO) published the HSM, which provided analytical tools and techniques to guide safety decision making [3]. Most existing segmentation approaches can be categorized as fixed-length, variable-length, or dynamic segmentation.

Fixed-length segmentation weights the crash information aggregated in each segment in a spatially equivalent manner by dividing the network into predetermined lengths. This approach is straightforward to implement, especially for many traditionally aggregated datasets [7]. The predetermined segment length significantly affects subsequent safety analysis. A short segment length is necessary for capturing fine details of spatial information but may cause measurements to be oversensitive while a long segment length may alleviate over-sensitivity and redundancy but could filter out many fine details.

Variable-length segmentation divides the network into homogeneous segments based on their characteristics (such as number of lanes) or geometry (such as intersections or ramps). Essentially, it classifies segments according to characteristics of roadways and then processes different classifications in different manners. This approach is highly dependent on the selected variables, which are considered to be critical to safety performance. Homogeneous segmentation is preferred by HSM [3] and many studies [7]–[9], not only because it is helpful to identify the causal relationship between roadway characteristics and safety performance (i.e., development of crash prediction models), but also because homogeneous segmentation is the foundation of CMFs and related paradigms. However, this approach is inevitably limited by the availability of roadway datasets that have a sufficient inventory. Koorey [7] proposed a set of criteria to determine a rational method for aggregating the crash data into logical road segments. The results suggested that analysts should use variable-length methods and should avoid short segments. Cafiso el al. [9] compared the performance of several segmentation methods in estimating safety performance functions (SPF). Interestingly, the results showed that the fixed-length segmentation method outperformed the homogeneous segmentation method based on the goodness-of-fit (GOF) of the models.

Dynamic segmentation is a technique that computes the map locations of the segments stored and managed in an event table using a linear referencing system (e.g., mile-point). Its concept is to store the organized and detailed information so the segments can be dynamically located. "Dynamic" refers to the fact that segmentation occurs on the fly whenever the underlying shape feature changes or any of the event features

changes. The implementation of this approach usually requires that the geographic information system (GIS) being used supports dynamic segmentation [10]. Besides this GIS-based technique, researchers also use the term "dynamic segmentation" when referring to multi-scale segmentation, which divides the network into segments with varying lengths. Boroujerdian et al. [11] presented a model that identified lengths and locations of high-risk segments using the "Mexican hat" wavelet. The length of each high-risk segment identified by the model could be different depending on the crash distribution.

B. SEGMENT LENGTH

Segment length has been a long-standing concern of researchers, and varying lengths have been used in traffic safety [12]. California Department of Transportation (Caltrans) has used a segment length of 0.2 miles; Utah DOT (UDOT) has used a segment length of 1 mile; Washington State DOT (WSDOT) has used a segment length of 0.1 miles or less depending on the highway type; and New York State DOT (NYSDOT) has used a segment length of 0.3 miles. A segment length of 0.125 miles was used in New Zealand [7]. Although HSM does not recommend a specific segment length, a minimum length of 0.1 miles is popular in its case studies, and a window size of 0.3 miles with an increment of 0.1 miles is demonstrated in one of HSM's examples [3] for the sliding window (SW) method as part of the screening program.

Many researchers suggest that short segment or window size should be avoided for reliable and meaningful results of safety analysis. By comparing four regression models, Miaou and Lum [13] indicated that short segment length (less than or equal to 0.05 miles) caused linear regression models to make questionable probabilistic statements, but Poisson regression models possessed most of the desirable statistical properties. In the following research, Miaou [14] indicated that negative binomial regression models were also sensitive to inclusion of short segments by comparing the performance of Poisson and negative binomial regression models. Resende and Benekohal [15] analyzed effects of roadway segment lengths on accident modeling and suggested a length of 0.5 miles for reliable accident prediction models. Ogle et al. [16] empirically indicated that short segments length (less than 0.1 mile) lead to unreliable results in safety analysis. Cook et al. [17] conducted a sensitivity analysis by comparing the effect of segment lengths of 0.5, 1, and 2 miles on safety analysis for two-lane rural primary roads and secondary low-volume rural roads. The results recommended a short segment length (0.5 miles) for two-lane rural primary roads and using all crash severities rather than just fatal and major injury crashes for secondary low-volume rural roads. Qin and Wellner [2] conducted a sensitivity analysis to quantify the similarity and discrepancy of varying window sizes for the sliding window method. The empirical results indicated that a window size of 0.5 miles might create more candidates for further review than a window size of 1 or 2 miles. Bahar and Hauer [8] recommended to regroup sub-segments shorter than 0.1 miles to a length of 0.1 miles, as a minimum, and calculated a combined average CMF.

III. SPATIAL FREQUENCY DOMAIN ANALYSIS

A. SYSTEMATIC AND RANDOM COMPONENTS

Both deterministic factors and stochastic factors of transportation systems can lead to safety issues that may cause or contribute to crashes. For example, the topology and design of civil infrastructure are deterministic due to geography, demography, history, etc.; while vehicles' movement and drivers' behavior are stochastic due to environment, vehicle status, human factors, etc. As a result, the crash data containing information about the safety performance of transportation systems should contain two types of components: systematic components and random components. Peng et al. [18] presented a similar concept that separates observed variability into three parts: randomness, proneness, and liability. Although the probable cause of each crash is usually recorded in the report, it is not reasonable to ascribe each individual crash to a single cause because transportation systems are so complex that various factors may contribute to a single crash indirectly. An appropriate paradigm for conducting safety analysis based on crash data requires distinguishing between the representations of systematic and random components, as well as their meanings. To comprehensively analyze the spatial distribution of crashes for proper segmentation in the spatial domain, a spatial frequency domain analysis (SFDA) in the spatial frequency domain (SFD) is needed.

B. One Dimensional Space of Linear Feature

Crashes are originally distributed in three-dimensional space, and two-dimensional maps are commonly used for intuitively locating crashes in projected coordinate systems. Most crashes are naturally distributed on or close to roadways and associated with geometry and properties of the network. Thus, projecting crash data into a 1-D space provides a feasible and practical form to aggregate crash data on continuous linear features of concerned roadways. For each crash instance C_i in original space V, if the distance (analogous to the offset distance) from its location c_i to the 1-D space L (finite space determined by the linear feature) is less than a certain value ε, then C_i is projected to Y_i at location y_i in L. The threshold ε can be determined by the width of the roadway and the accuracy of the crash location in the police reports [6], and it is also dependent on safety analysis tasks.

C. QUANTIFYING CRASH DENSITY

Varying quantifications can be used for safety analysis, and each of them have their own advantages as well as biases. This research uses the number of crashes and equally weights all types and all severities of crashes to explore the spatial distribution of crashes.

Histogram is a simple but effective non-parametric density estimation (NPDE) method. The crash density estimate at location y in the interval B_i on the linear feature can be determined by

$$\widehat{d_h}(y) = \frac{1}{h}\sum_{i=1}^{n}\sum_{j=1}^{m} I(y_i \in B_j) I(y \in B_j) \quad (1)$$

where h is the width of intervals, n is the total number of

crashes, and m is the total number of intervals. Interval width determines the sampling rate in quantifying crash density in the 1-D space. The discontinuities of the estimate are artifacts of interval locations which have little effect on SFDA unless m is an extremely small number.

Kernel density estimation (KDE) is another NPDE method that conducts inference for an unknown population based on finite samples by smoothing data with a specified kernel [19]. Constructing a kernel $K(*)$ at the location y_i of each crash instance i in the 1-D space and then integrating all kernels can achieve crash density estimate $\widehat{d_h}(y)$ by:

$$\widehat{d_h}(y) = \sum_{i=1}^{n} K_h(y - y_i) = \frac{1}{h}\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right) \quad (2)$$

where h is the bandwidth of the kernel $K(*)$. Varying kernel functions can be applied in estimating crash density depending on how the data should be smoothed and weighted. Boroujerdian et al. [11] used a linear interpolation method to map the number of crashes to corresponding grid points in a wavelet-based segmentation method, which was mathematically equivalent to a KDE process using a tent kernel. Besides the selection of kernel functions, the selection of bandwidth, which leads to the trade-off between the bias of the estimator and its variance, is also critical to KDE. Since the quantification in this research focuses on spatial frequency characteristics of crash distribution, it requires a kernel with a very small bandwidth for preserving high frequency components of the crash data. Thus, the Kronecker delta function (or unit impulse), is selected to avoid the loss of high frequency components from smoothing. A KDE using the Kronecker delta function with a discretion processing can be mathematically equal to histogram.

The output from the used quantification of the crash data projected in 1-D space is crash density—a series of numbers which represent the number of crashes aggregated in the corresponding interval (or at the sampling locations). Fig. 1 (a, c, and e) shows the example of outputs at different sampling rates. Similar to a time domain signal that shows how signal changes over time or a raster image that has only one row or column of pixels, the output is a spatial domain sequence in a 1-D space that shows how crash density changes over distance along the linear feature. This "crash density sequence" (CDS) can be applicable for methodologies in signal and image processing, but the meanings of results need to be interpreted carefully. It is worth noting that the concept of sampling rate in this paper is easily confused with the upper limit of a frequency band that is used to determine segment length for some readers. In this research, sampling rate is dependent on the accuracy of crash location, the segment length that the raw data uses to aggregate records (if the raw data is maintained in the traditional form), and the interval used to quantify the measurement (number of crashes in this paper).

D. SPATIAL FREQUENCY DOMAIN

Spatial frequency is a measurement for events that are periodic across location in space [20]. Whereas a CDS (or other spatial domain representations) shows how crash density changes over distance, an SFD representation of the CDS shows how crash density lies within each given spatial frequency band over a range of spatial frequencies. More specifically, it shows how frequent or in which periodic pattern crashes occur along the linear feature, which is critical to safety analysis that focuses on the spatial distribution of crashes. In this paper, the unit of spatial frequency is defined as samples per mile (samples/mile) and denoted as S-Hz (spatial hertz).

The discrete Fourier transform (DFT) is employed to convert CDSs from spatial domain to SFD. Mathematically, the DFT converts a CDS x_n into a sequence of complex numbers X_k by

$$X_k = \mathcal{F}[x] = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} = \sum_{n=0}^{N-1} x_n[\cos(2\pi kn/N) - i\sin(2\pi kn/N)] \quad (3)$$

where N is the total number of samples of x_n, and it is also the total number of samples of X_k. The magnitudes of X_k indicate how many spatial frequency events lie at each spatial frequency, or more specifically, how many crashes periodically occur at each spatial frequency along the linear feature. Although X_k does not directly present spatial distribution of crashes on the linear feature, it reveals the correlation of the number of crashes distributed in consecutive intervals (or sampling locations) to each other along the linear feature. The spatial correlation of crash distribution is essentially governed by both deterministic factors and stochastic factors of the whole transportation system.

E. ENERGY AND POWER OF CRASH DENSITY

To explore how the energy of a CDS is distributed over spatial frequency, the periodogram, a non-parametric spectral density estimation (SDE) method, is employed to estimate the power spectral density (PSD) of CDSs in this research. Mathematically, the PSD $\hat{P}(f)$ of a CDS x_n at each spatial frequency f can be computed by

$$\hat{P}(f) = \frac{\Delta t}{N}\left|\sum_{n=0}^{N-1} x_n e^{-i2\pi fn}\right|^2, -\frac{1}{2\Delta t} < f < \frac{1}{2\Delta t} \quad (4)$$

where Δt is sampling period. The range of f is determined by the Nyquist frequency [21]. In this research, the unit of PSD is crash^2/S-Hz since CDS is defined as number of crashes/sample (or crashes/interval).

Sampling rate of a CDS determines the effective band of its PSD according to the Nyquist–Shannon sampling theorem [21]. As shown in Fig. 1 (b, d, and f), sampling rates of 4, 10, and 20 S-Hz lead to bands of 0–2, 0–5, and 0–10 S-Hz, respectively. Different sampling rates present identical power distributions (they have exactly the same shape but are shrunk in scale) within the low frequency band they share. High sampling rate can provide extra information for the power distribution within the high frequency band.

F. A HYPOTHESIS FOR LOW FREQUENCY COMPONENTS

As a type of safety performance presentation of transportation systems, a CDS should contain both systematic and random components. Systematic components indicate the major pattern in which crashes are distributed along the linear feature, which is presented as the trend and basic structure of a CDS. In other words, as a part of CDS, systematic components naturally correspond to low frequency components in the SFD of which the events are deterministically associated with locations in the spatial domain. On the other hand, random

components indicate the uncertainty or randomness of occurrence of crashes, which is presented as fine details and randomness of a CDS. In other words, random components correspond to high frequency components of which the events are independent of location in the spatial domain. Extremely high frequency components are theoretically the representation of noise in the CDS.

The gap between "high" and "low" frequency components varies depending on cases. Fig. 1 (b, d, and f) shows the PSD of the corridor of SC 146 in Greenville County with the 2013 crash data at sampling rates of 5 S-Hz, 10 S-Hz, and 20 S-Hz, respectively. Most of the power of this CDS is concentrated within the band of 0–0.3 S-Hz, and the power at frequencies higher than 1.5 S-Hz is insignificant. This PSD pattern indicates that the spatial distribution of crashes along the linear feature is mainly determined by the low frequency components while the influence of high frequency components is very limited. A hypothesis can be established that the CDS of a corridor has major power concentrated in low frequency bands. This hypothesis is justified in the following sections.

IV. EXPERIMENTS

A. DATA COLLECTION

The development of MMUCC and MIRE, and the advent of map-based crash geocoding systems that have been deployed by law enforcement agencies in many states has greatly improved the quality of crash data. South Carolina began their deployment of one such system: the South Carolina Collision and Ticket Tracking System (SCCATTS) in 2010. SCCATTS provides law enforcement officers the tools to identify the approximate crash location using global positioning system (GPS), and then accurately locate (or pin map) the crash at the precise location it occurred on the map display. Fig. 2 (a and b) shows 2013 crash data along a section of US 146 in Greenville, South Carolina using the new map-based system (SCCATTS). Compared with previous datasets, the crash data collected by SCCATTS has significantly improved quality and reduced errors [6].

B. DESIGN OF EXPERIMENTS

Experiments were designed to explore the spatial frequency characteristics of crash data using proposed SFDA. Implementation was based on ArcMap 10.2 and MATLAB R2016a. The implementation included the following the steps:

1. Projecting crash instances recorded in crash data into the 1-D space determined by the linear shapefile feature of studied roadway data.
2. Quantifying the projected crash data to generate the CDS of that linear feature. ε was set to 30 feet [6].
3. Performing the DFT and SDE (periodogram) for the Fourier transform and the PSD of the CDS, respectively.
4. Computing concerned statistics and metrics for further analysis.

Besides PSD plots, three percentage metrics were used to quantitatively identify the proportion of power concentrated within low frequency bands (see Table I): the power within the band of 0–0.5/0–2/0–3 S-Hz over the power within the band of 0–2/0–5/0–5 S-Hz.

A group of crash and roadway data were studied in the experiments. The roadway data included linear features from nine corridors (see Table I) which were selected from among the top eleven risky corridors that were identified as dangerous and representative in a project sponsored by SCDOT [6].

V. RESULTS AND DISCUSSION

A. COMPUTATION RESULTS

The CDSs of studied corridors are shown in Fig. 3; and corresponding PSDs are shown in Fig. 4. All PSD plots illustrate that the major power of a CDS is concentrated within a relatively low frequency band to some extent.

The results of three percentage metrics are listed in Table I. The first metric indicates that 7 corridors have major power (73.65%–86.60%) concentrated in the lowest 1/4 bandwidth (0–0.5 S-Hz/0–2 S-Hz), and 2 corridors have over half (54.04% and 57.53%) power concentrated in the lowest 1/4 bandwidth. The second metric indicates that 8 corridors have major power (73.23%–84.47%) concentrated in the lowest 2/5 bandwidth (0–2 S-Hz/0–5 S-Hz), and only 1 corridor has over half (65.94%) power concentrated in the lowest 2/5 bandwidth. The third metric indicates that all 9 corridors have major power (79.15%–92.22%) concentrated in the lowest 3/5 bandwidth (0–3 S-Hz/0–5 S-Hz). Among all 9 corridors, only the corridor of US 17 in Berkeley County had considerable proportion of power within midrange frequency bands. The reason was that it had significantly fewer crashes than other corridors.

B. HYPOTHESIS AND FINDINGS

PSD plots and all three percentage metrics justifies the hypothesis that the CDS of a corridor has major power concentrated in low frequency bands. Although the extents of power spectral concentration vary for different corridors, all CDSs present the similar PSD patterns in the SFD.

This characteristic of power spectral concentration within low frequency bands indicates that spatial distribution of crashes on corridors (top-risk corridors in South Carolina) is mainly determined by low frequency components (corresponding to systematic components) of which the events are deterministically associated with locations in the spatial domain. On the other hand, the influence of high frequency components on spatial distribution of crashes is very limited because of minor power distributed in the corresponding bands. The meaning of high frequency components makes them even less impacting in analyzing crash distribution because higher frequency components more likely represent randomness of occurrence of crashes as if noise existing regardless of locations in the spatial domain.

The findings imply basic trends in large regions have much more impacts on the overall safety performance of transportation systems than individual high-risk spots, and it theoretically shows that major proportion of crashes are associated with specific locations. Thus, adequately filtering out high frequency components can remove fine details and randomness while preserving the primary information and needed detailed information about crash distribution. Similarly, using a low sampling rate to quantify CDSs may cause loss of fine details and randomness that contain very limited power.

C. INFERENCE 1—POWER SPECTRAL SEGMENT LENGTH

By distinguishing the primary components containing major power of the CDS from high frequency components containing minor power, it is possible to capture primary information and a certain level of detailed information about crash distributions that can be sufficient and effective for specified safety analysis tasks.

The characteristic of power spectral concentration within low frequency bands indicates that a proper segment length is capable of capturing low frequency components containing major power of the CDS. According to the Nyquist–Shannon sampling theorem [21], the effective maximum segment length L can be determined by a threshold frequency f which is the upper limit of a band (its lower limit is zero) containing major power in the PSD. Mathematically,

$$L = \frac{1}{2f} \tag{5}$$

For example, the CDS of the corridor of SC 146 in Greenville County (see Fig. 4 (a)) has a large proportion of power concentrated within the band of 0–0.3 S-Hz. If 0.3 S-Hz is chosen as the threshold frequency, the effective maximum segment length should be 1.67 miles, which is capable of capturing the information contained by low frequency components within the band of 0–0.3 S-Hz. If 2 S-Hz is chosen as the threshold frequency for finer detailed information (considering the power within the band 0.3–2 S-Hz), the effective maximum segment length should be 0.25 miles. This will allow additional information contained by the components within the increased band to be captured.

The selection of threshold frequency is dependent on the PSD of the concerned dataset as well as the extent to which safety analysis tasks demand detailed information about crash distributions. For corridors that have a considerable proportion of power within midrange frequency bands (like the corridor of US 17 in Berkeley County), a carefully chosen threshold frequency can still lead to a useful maximum segment length.

The authors defined the power spectral segment length (PSSL) as the maximum segment length capable of capturing spatial information about crash distributions within a specified frequency band. An optimal PSSL should be the maximum length that can capture primary components within low frequency bands to satisfy the demands of safety analysis tasks. Any segment length shorter than the optimal PSSL can capture finer details because of wider bandwidth. However, the improvement could be very limited if the optimal PSSL is determined by a carefully chosen threshold frequency.

D. INFERENCE 2—POWER SPECTRAL PERCENTAGE

A more informative description of the PSSL can be in a form referring to a baseline (a popular length like 1, 0.25, or 0.1 miles) for intuitive understanding and comparison. Thus, the authors propose the power spectral percentage (PSP) as the percent of power a PSSL can preserve compared with a baseline. In the example of the corridor of SC 146 in Greenville County, the segment length of 1 mile can preserve 86.60% (see Table I) power compared with a baseline of 0.25 mile, which can be represented as $PSP_{0.25\text{mi}}^{1\text{mi}}$=86.60%. The corresponding PSSL can be represented as $PSSL_{0.25\text{mi}}^{86.60\%}$=1 mi. Similarly, for the corridor of US 1 in Lexington County, a $PSSL_{0.1\text{mi}}^{83.26\%}$ of 0.25 miles would imply a $PSP_{0.1\text{mi}}^{0.25\text{mi}}$ of 83.26%.

The PSP can potentially be a metric for assessing the performance of segment lengths. Currently, there is no standard or commonly accepted metric for network segmentation, which is one of the reasons why engineering judgment is required in much of the research and practice. Almost every research used different assessing methods. Cook et al. [17] conducted a sensitivity analysis, which employed ranking list shifts of segments ranked with the Iowa DOT scoring method. The output of ranking list shifts was not intuitive for interpreting the results. Qin and Wellner [2] conducted a sensitivity analysis, which calculated the miles of high-risk highway that can be identified with different segment lengths. This method could barely provide sufficient information for making practical decision. Boroujerdian et al. [11] employed three different methods: a group of statistics of crash density, the percentage of accidents covered by the highest ranked segments occupying 15% of total length of all segments, and a graph of relationship between aggregate percentages of accidents and the percentage of studied length. The first two methods relied on preset default conditions; while the third method was intuitive for comparison but could not provide numerical descriptions for performance.

The PSP can quantitatively describe the extent to which the segment length can capture primary information about crash distributions. It can be practical as a metric for several reasons: 1) the form is concise, 2) the output is easy to interpret for assessment and comparison, and 3) it is applicable and extensible to various crash data and networks. However, it requires special attention since PSP is nonlinear in several aspects: 1) the segment length is not linear to threshold frequency as shown in (5), 2) the power would not be evenly distributed in the SFD as shown in Fig. 4, and 3) components at different frequencies represent different meanings.

E. OPTIMAL SEGMENT LENGTH

In practical safety analysis, there is not a universally-best segment length. The optimal segment lengths are dependent on the studied dataset and the study objectives. For top-risk corridors in South Carolina, a PSSL of 0.25 miles can achieve a $PSP_{0.1\text{mi}}^{0.25\text{mi}}$ of 73.23%–84.47%, which refers to the capability of capturing the primary information about the crash distribution with a moderate level of detail; a PSSL of 0.1667 miles can capture most of crash distribution information and achieve a $PSP_{0.1\text{mi}}^{0.1667\text{mi}}$ of 83.50%–92.22%.

VI. SPECTRAL-BASED SEGMENTATION

A. LOW-PASS FILTERING

The SW method uses a window to conceptually move along the road segments from beginning to end [3]. Theoretically, it is a smoothing process using moving average filter, which is a type of low-pass filter. The size of the window is functionally equivalent to a PSSL, which is another form of threshold frequency. Given that the SW method is essentially a filtering process, the authors propose the low-pass filtering (LPF) method as a general solution for fix-length segmentation.

The basic concept of the LPF method is to determine how the crashes are aggregated and weighted in fix-length segmentation

according to the primary information about crash distribution in order to avoid impacts of uncertainty and randomness on safety analysis. Its implementation is straightforward: 1) use the SFDA to determine the threshold frequency (or the PSSL) based on natural characteristics of crash distributions of studied data and the demands of safety analysis tasks, and 2) use a properly designed low-pass filter to filter out unwanted high frequency components of CDS.

For readers who do not have relevant background knowledge, a simple explanation is that the LPF method is a generalized SW method that has flexible window size, uneven weights, infinitely small increment, and other features. Compared with the SW method, the LPF method has several major advantages:

1. *Flexibility*. The LPF method allows PSSL to be a rational number multiple of sampling/interval. On the contrary, the SW method can only use a window size of integer multiple of sample/interval. Combined with aforementioned features, the LPF method is capable of conducting segmentation flexibly within an extensive range.
2. *Distinguishability*. The LPF method can process components at different frequencies more effectively than the SW method. It is capable of distinguishing desired information about crash distribution and unwanted randomness (see the following discussion).
3. *Superiority in identifying risky segments*. The capabilities of identifying optimal PSSL and conducting flexible segmentation in an efficient manner enable the LPF method superior performance in most segment-based analysis. As a result, the LPF method outperforms the SW method in identifying top-risk regions even if they use exactly same length (see the following discussion).

An example based on the corridor of SC 9 in Spartanburg County is shown in Fig. 5. Fig. 5 (c and d) shows the result of implementing a SW method on the CDS with a window of 0.3 miles and increment of 0.1 miles. Fig. 5 (e and f) shows the result of implementing a LPF with a threshold frequency of 1.67 S-Hz (the authors designed and used a Butterworth low-pass filter with a pass band of 0–1.67 S-Hz). According to (5), a window size of 0.3 miles is functionally equivalent to a threshold frequency of 1.67 S-Hz. It is clear that the "windowed" CDS achieves similar but worse results than the "filtered" CDS, even though they are results from the equivalent PSSL. Obviously, the SW method preserves a part of unwanted details while loses a part of desired details. The reason is that moving average filter is a very poor low-pass filter due to its slow roll-off and poor stopband attenuation. On the contrary, the Butterworth LPF provides clean results because it adequately filters out unwanted components and preserves desired components.

To compare the performance of the LPF and SW methods in an intuitive (or traditional) way, the authors employed a metric which is the percentage of aggregated crashes within studied length of top-risk samples/intervals sorted by the number of crashes (from the largest to the lowest). This metric is similar to the graph method used by Boroujerdian et al. [11]. Several SW and LFP segmentations were evaluated (see Fig. 6 and Table II).

Unsurprisingly, the LPF method with a threshold frequency of 3 S-Hz (equivalent to a PSSL of 0.167 mi) can aggregate crashes in a more efficient manner than other two methods. Fig. 6 clearly shows the comparison. In additional to the graph, other metrics also provide intuitive evaluation about their performance (see Table II).

Interestingly, the LPF method with a threshold frequency of 1.67 S-Hz can aggregate more or approximate amount of crashes within the top-risk samples/intervals than the equivalent SW method with a window of 0.3 miles. Specifically, using those two methods to rank and identify risky regions, the top-risk regions identified by the LPF method have more or approximate amount of crashes than those identified by the SW method, even though they use equivalent PSSL. However, for mid-risk regions that have relatively fewer crashes, the SW method shows higher aggregate rates of crashes than the equivalent LPF method. Those two methods have similar performance for low-risk regions that have very few crashes.

For example, regarding the corridor of SC 146 in Greenville County, the LPF method can better identify top 26.09% risky regions than the equivalent SW method, and those regions have 78.08% crashes of all crashes on the corridor (see Table II).

Experimental results (see Table II) indicate that the LPF method can identify top-risky regions better than the equivalent SW method for five of those corridors; and they are very close in identification of top-risky regions for the remaining four corridors. The reason for this phenomenon is that the LPF method better preserves primary components of crash distribution which enable it to identify the most impacting regions; while the SW method fails to eliminate a part of high frequency information which enable it can better identify less risky regions. The comparison indicates that the LPF method is a better choice than the SW method in network screening, even if they use exactly equivalent PSSL.

B. IMPROVED WAVELET-BASED METHOD

Wavelet-based methods use a sum of scaled and translated copies of the mother wavelet to represent the crash distribution along the linear feature. Those methods are capable of dividing segments as well as locating high-risk segments [11]. Corresponding to the parameter of frequency in Fourier transform, the parameter of scale determines the extent to which details of crash distribution can be captured by wavelets. Large scale (corresponding to low frequency) components indicate the trend and basic structure of crash distribution; while small scale (corresponding to high frequency) components indicate fine details and randomness. A rigorous wavelet-based segmentation requires correct understanding and appropriate interpretation of spatial wavelet analysis (SWA), which is lacking in existing research [11].

Different to the short-time Fourier transform (STFT), which has uniform frequency resolution and spatial resolution, the wavelet transform has bad frequency resolution at high frequencies and bad spatial resolution at low frequencies [22]. The pseudo-frequency f corresponding to a scale α is determined by the center frequency f_c of the chosen wavelet function and the sampling period Δt of the CDS, as

$$f = \frac{f_c}{a\Delta t} \tag{6}$$

The upper limit of the frequency band is determined by the sampling rate and chosen wavelet function. An insufficient sampling rate or an improper wavelet function could lead to too much loss of details. Thus, a proper upper limit of the frequency band is necessary for capturing sufficient details while avoiding high frequency randomness.

On the other hand, the lower limit of the frequency band is determined by an arbitrarily chosen upper limit of scale range. The components at very large scales do not contain useful details due to averaging and bad spatial resolution. Especially in wavelet analyses using scalogram, overlarge scales can cause overweighting of the amplified energy contained by components at low frequencies as well as the aliasing of them. Thus, in addition to a proper upper limit of the frequency band, a proper upper limit of scale range is also necessary for appropriate interpretation of SWA.

An example based on the corridor of SC 9 in Spartanburg County is shown in Fig. 7. By implementing a wavelet transform with a Mexican hat wavelet (0.25 S-Hz center frequency) and a scale range of 1–32, high-risk segments can be identified based on positive coefficients of wavelets within the chosen scale range of 1–32 (see Fig. 7 (e)) or the corresponding frequency band of 0.08–2.5 S-Hz (see Fig. 7 (c)). The wavelets' energy distribution in terms of frequency has the same pattern as the PSD of the CDS: major energy is concentrated within the low frequency band (or large scale range) (see Fig. 7 (b, d, and f)). In this SWA, the capability of capturing details of crash distribution is limited to within the frequency band of 0.08–2.5 S-Hz. To achieve the capability of capturing finer details represented by components at frequencies higher than 2.5 S-Hz, a higher sampling rate or a different wavelet function having a higher center frequency is required.

Dividing segments and identifying high-risk segments are critically dependent on appropriate interpretation of the wavelet transform of CDSs. Boroujerdian et al. [11] proposed a model using components at a large scale to identify long segments and using components at a small scale to identify short segments. They arbitrarily defined "long/short" and "large/small" in their method without explicit explanation. Based on wavelet transform and Fourier transform, components at each scale or frequency can only represent the crash distribution information at the corresponding scale or frequency (depending on frequency resolution). In other words, a high-risk segment identified with components at a certain scale can only indicate that more crashes occurred within the segment at the corresponding de-noised and de-trended level of details. For example, as shown in Fig. 8 (a, d, g, and j), the coefficients of components at the scale of 2 can indicate which segments are risky at a more detailed level than those at the scales of 4, 8, and 16. Components at a single scale or frequency are insufficient to lead to the conclusion that identified segments are really riskier because they do neither include more basic trends contained by larger scale components nor finer details contained by smaller scale components.

Based on the SFDA and the SWA discussed above, the authors propose an improved method to divide and identify high-risk segments by integrating the components within a proper scale range (or frequency band) for more comprehensive crash distribution information. The improved method requires the following points:

1. A proper lower limit of the scale range is required depending on the extent to which the details are needed (see Fig. 8 (b and e) for the results using different lower limit of scale range for integrating components).
2. A proper upper limit of the scale range is required to avoid overweighting and aliasing the basic trend (see Fig. 8 (h and k) for the results that overweight components at large scales).
3. A normalization process is required to allow components at different scales to be weighted equally (or using specified weighting strategies if safety analysis tasks demand). The normalization can be computed by dividing all coefficients by the corresponding scales.

The improved method can provide intuitive and comprehensive information for dividing and identifying high-risk segments (see Fig. 8 (c, f, i, and l) for the results using different scale ranges). It is worth mentioning that using a scale range that is very small can lead to results similar to using a single scale, however the normalization process can avoid biases towards overweighed components. The improved method can be implemented using various scale range for flexibility and effective decision making depending on the demands of safety analysis tasks.

VII. CONCLUSIONS

Proper network segmentation is essential to rigorous segment-based safety analysis. Traditional approaches to conduct network segmentation and crash aggregation require engineering judgment and are subject to a lack of standard metrics to assess their performance. Alternatively, this paper proposes a novel methodology for interdisciplinary analytics of crash distribution, crash aggregation, and network segmentation in traffic safety realm. This methodology reveals the natural characteristics of crash distributions, and it provides general solutions for determining optimal segment lengths and conducting effective network segmentation.

This research demonstrates the feasibility of exploring natural characteristics of crashes using redesigned spectral analysis. It also demonstrates that data-driven analytics from non-traditional perspective is capable of solving long-standing problems and innovating traditional segment-based safety analysis. The paper extends fundamental knowledge in traffic safety and provides new paradigms for safety analysis.

Applying the proposed methodology in practical safety analysis is not as complex as the theoretical elaboration in this paper. PSSLs and corresponding PSPs, the LPF method, or the improved wavelet-based method can be implemented with a few lines of code in MATLAB. Its application is subject to the availability of accurate spatial information about crashes and roadways. MATLAB examples are available if readers contact the authors.

This paper selects the number of crashes, which is directly associated with the exposure of traffic, for quantification. In future research, the potentials of other measurements that consider the exposure of traffic or the severity of crashes will be explored, including the segmentation used for developing predictive crash-count models [23]. Due to both uniform

frequency and uniform spatial resolutions, the STFT is potentially a better route than the wavelet transform for dividing and identifying high-risk segments. In future research, a new approach based on it will be studied in order to reveal the nature of crash incidence in transportation systems.

ACKNOWLEDGMENT

The authors thank Wayne Sarasua and Kweku Brown for providing data. The authors thank Stephen Fry for literary input. The authors thank Douglas Dawson for constructive comments.

REFERENCES

[1] Centers for Disease Control and Prevention, "WISQARS (Web-based Injury Statistics Query and Reporting System)|Injury Center|CDC," 2018. [Online]. Available: https://www.cdc.gov/injury/wisqars/. [Accessed: 15-Nov-2018].

[2] X. Qin and A. Wellner, "Segment Length Impact on Highway Safety Screening Analysis," in *Transportation Research Board 91st Annual Meeting*, 2012, pp. 12-0644.

[3] National Research Council (US), *Highway Safety Manual*. Washington, DC: American Association of State Highway and Transportation Officials, 2010.

[4] "MMUCC Guideline: Model Minimum Uniform Crash Criteria," 2017.

[5] N. Lefler, Y. Zhou, D. Carter, H. McGee, D. Harkey, and F. Council, "Model Inventory of Roadway Elements – MIRE 2.0," 2010.

[6] W. A. Sarasua, J. H. Ogle, M. Chowdhury, N. Huynh, and W. J. Davis, "Support for the Development and Implementation of an Access Management Program Through Research and Analysis of Collision Data," 2015.

[7] G. Koorey, "Road Data Aggregation and Sectioning Considerations for Crash Analysis," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2103, pp. 61–68, 2009.

[8] G. Bahar and E. Hauer, "User's Guide to Develop Highway Safety Manual Safety Performance Function Calibration Factors," Washington, DC, 2014.

[9] S. Cafiso, C. D'Agostino, and B. Persaud, "Investigating the influence of segmentation in estimating safety performance functions for roadway sections," *J. Traffic Transp. Eng. (English Ed.*, vol. 5, no. 2, pp. 129–136, 2018.

[10] K. J. Dueker and R. Vrana, "Dynamic Segmentation Revisited: A Milepoint Linear Data Model," *J. Urban Reg. Inf. Syst. Assoc.*, vol. 4, no. 2, pp. 94–105, 1992.

[11] A. M. Boroujerdian, M. Saffarzadeh, H. Yousefi, and H. Ghassemian, "A model to identify high crash road segments with the dynamic segmentation method," *Accid. Anal. Prev.*, vol. 73, pp. 274–287, Dec. 2014.

[12] J. Geyer, E. Lankina, C.-Y. Chan, D. Ragland, T. Pham, and A. Sharafsaleh, "Methods for Identifying High Collision Concentration Locations for Potential Safety Improvements," 2008.

[13] S.-P. Miaou and H. Lum, "Modeling vehicle accidents and highway geometric design relationships," *Accid. Anal. Prev.*, vol. 25, no. 6, pp. 689–709, Dec. 1993.

[14] S.-P. Miaou, "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions," *Accid. Anal. Prev.*, vol. 26, no. 4, pp. 471–482, Aug. 1994.

[15] P. T. V. Resende and R. F. Benekohal, "Effects of Roadway Section Length on Accident Moeling," in *Traffic Congestion and Traffic Safety in the 21st Century*, 1997, pp. 403–409.

[16] J. H. Ogle, P. Alluri, and W. A. Sarasua, "Model Minimum Uniform Crash Criteria and Minimum Inventory Roadway Elements: Role of Segmentation in Safety Analysis," in *Transportation Research Board 90th Annual Meeting*, 2011.

[17] D. J. Cook, R. R. Souleyrette, and J. Jackson, "Effect of Road Segmentation on Highway Safety Analysis," in *Transportation Research Board 90th Annual Meeting*, 2011.

[18] Y. Peng, D. Lord, and Y. Zou, "Applying the Generalized Waring model for investigating sources of variance in motor vehicle crash analysis," *Accid. Anal. Prev.*, vol. 73, pp. 20–26, Dec. 2014.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)*. Springer, 2009.

[20] G. D. Boreman, *Modulation Transfer Function in Optical and Electro-Optical Systems*. Bellingham, WA: SPIE Press, 2001.

[21] C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[22] R. X. Gao and R. Yan, *Wavelets: Theory and Applications for Manufacturing*. Springer, 2011.

[23] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transp. Res. Part A Policy Pract.*, vol. 44, no. 5, pp. 291–305, Jun. 2010.

XI ZHAO received his Ph.D. degree in transportation engineering from Clemson University. He is currently a lecturer in the department of Automation at Wuhan University of Technology. His research interest includes transportation safety, traffic simulation, data science, and deep learning.

DOMINIQUE LORD received his B.Eng. in civil engineering from McGill University and his M.A.Sc. and Ph.D., also in civil engineering, from the University of Toronto. He is currently a Professor and A.P. and Florence Wiley Faculty Fellow in the Zachry Department of Civil Engineering at Texas A&M University. His work focuses on conducting fundamental research in highway safety and crash data analyses.

YICHUAN PENG received the Ph.D. degree in Transportation engineering from Texas A&M University, USA, in 2013.
He is currently an assistant professor in Transportation college of Tongji University, CHINA. His research interests include the area of traffic safety, including crash prediction modeling and human behavior. He is currently focusing on crash prediction in adverse weather, real-time detection of aggressive driving behavior.

TABLE I
CORRIDORS STUDIED IN EXPERIMENTS

Corridor	Length (mi)	Crash Count (crash)	Average Density (crash/mi)	Description	Power% (0–0.5 S-Hz /0–2 S-Hz) (1 mi /0.25 mi)	Power% (0–2 S-Hz /0–5 S-Hz) (0.25 mi/0.1 mi)	Power% (0–3 S-Hz /0–5 S-Hz) (0.1667 mi/0.1 mi)
SC 146 Greenville	11.5	917	79.74	suburban	86.60%	84.47%	89.24%
SC 9 Spartanburg	17.5	421	24.06	suburban	77.85%	77.34%	87.79%
US 1 Lexington	29.5	637	21.60	rural / suburban	73.65%	83.26%	92.22%
US 1 Richland	13.5	797	59.04	suburban	84.09%	78.29%	86.47%
US 17 Berkeley	33.0	195	5.91	rural / suburban	57.53%	65.94%	79.15%
US 17 Horry (1)	18.5	515	27.84	suburban / urban	84.56%	82.95%	89.70%
US 17 Horry (2)	21.5	486	22.60	rural / suburban	54.04%	74.31%	85.01%
US 25 Greenville	53.5	1017	19.01	rural / suburban	77.71%	73.23%	86.65%
US 52 Florence	30.0	295	9.83	rural / suburban	78.08%	73.81%	83.60%

TABLE II
PERFORMANCE EVALUATION BASED ON PERCENTAGES OF AGGREGATED CRASHES AND POWER SPECTRAL PERCENTAGE

Corridor	Aggregated Length (LPF won)	Aggregated Length (close)	Aggregated Crashes	PSP (0.3 mi window)	PSP (1.67 S-Hz low-pass)	PSP (3 S-Hz low-pass)
SC 146 Greenville	26.09%	–	78.08%	80.26%	84.84%	90.13%
SC 9 Spartanburg	–	12.57%	50.83%	74.71%	77.93%	89.45%
US 1 Lexington	–	17.29%	68.45%	78.38%	81.76%	92.80%
US 1 Richland	8.89%	–	25.09%	73.69%	73.30%	87.32%
US 17 Berkeley	3.64%	–	32.82%	59.85%	78.94%	88.00%
US 17 Horry (1)	–	9.73%	32.82%	76.32%	79.22%	91.03%
US 17 Horry (2)	12.09%	–	49.38%	65.40%	73.96%	87.46%
US 25 Greenville	–	3.93%	30.29%	67.28%	66.15%	86.87%
US 52 Florence	12.67%	–	50.85%	69.33%	70.45%	84.61%

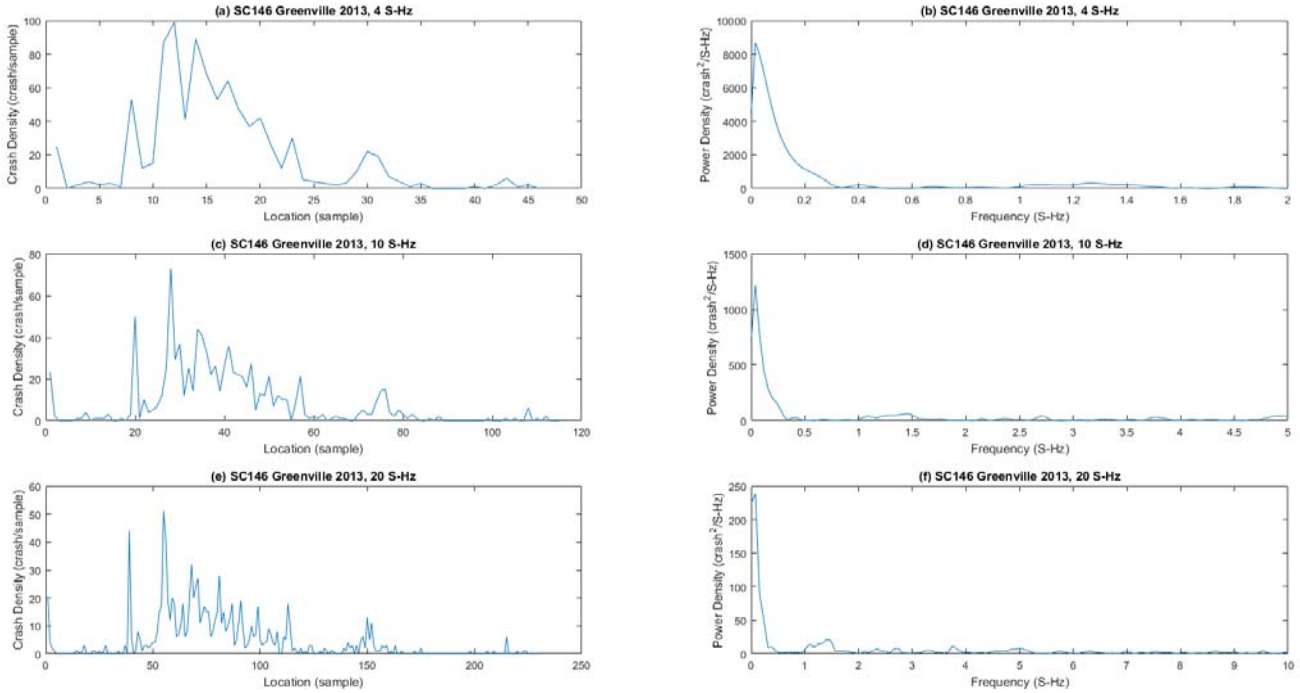
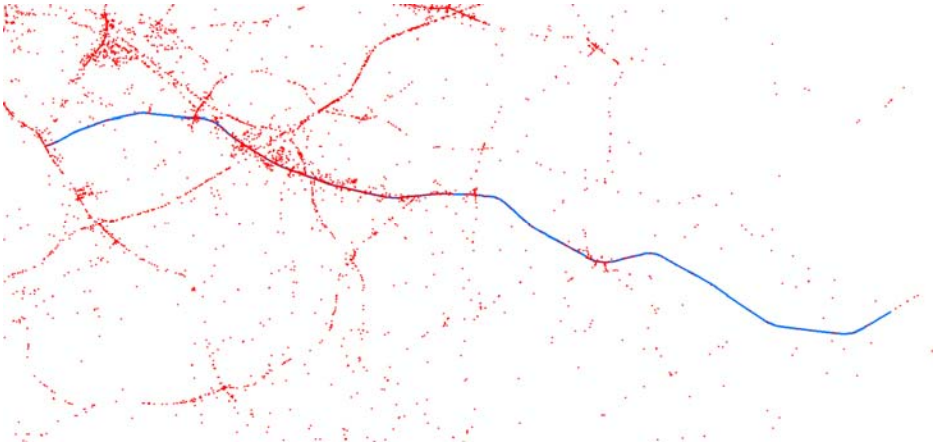


FIGURE 1 The CDS and corresponding PSD of the 2013 crash data on the corridor of SC 146 in Greenville County at different sampling rates



(a)

(b)

FIGURE 2 The corridor of SC 146 in Greenville County with the 2013 crash data: (a) zoomed out; (b) zoomed in.

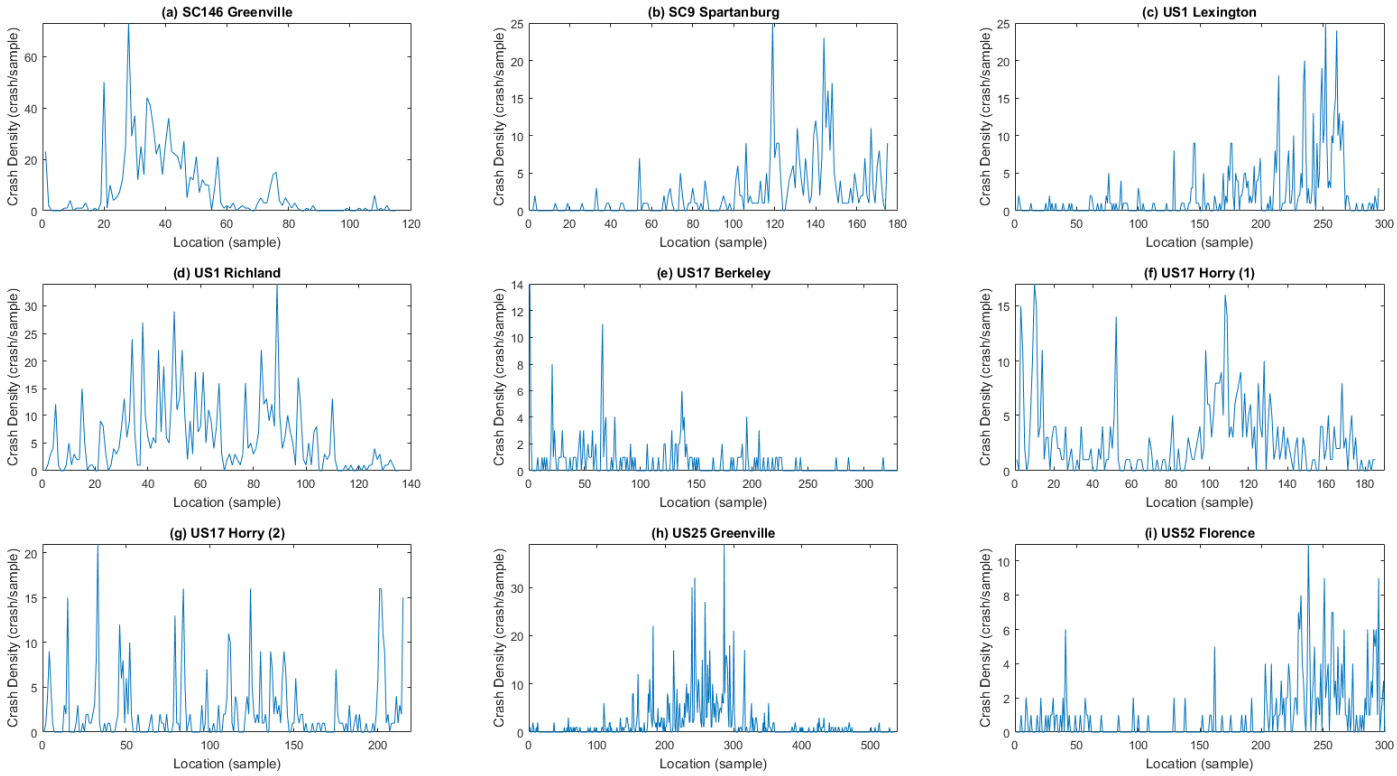


FIGURE 3 CDSs of studied corridors at the sampling rate of 10 S-Hz

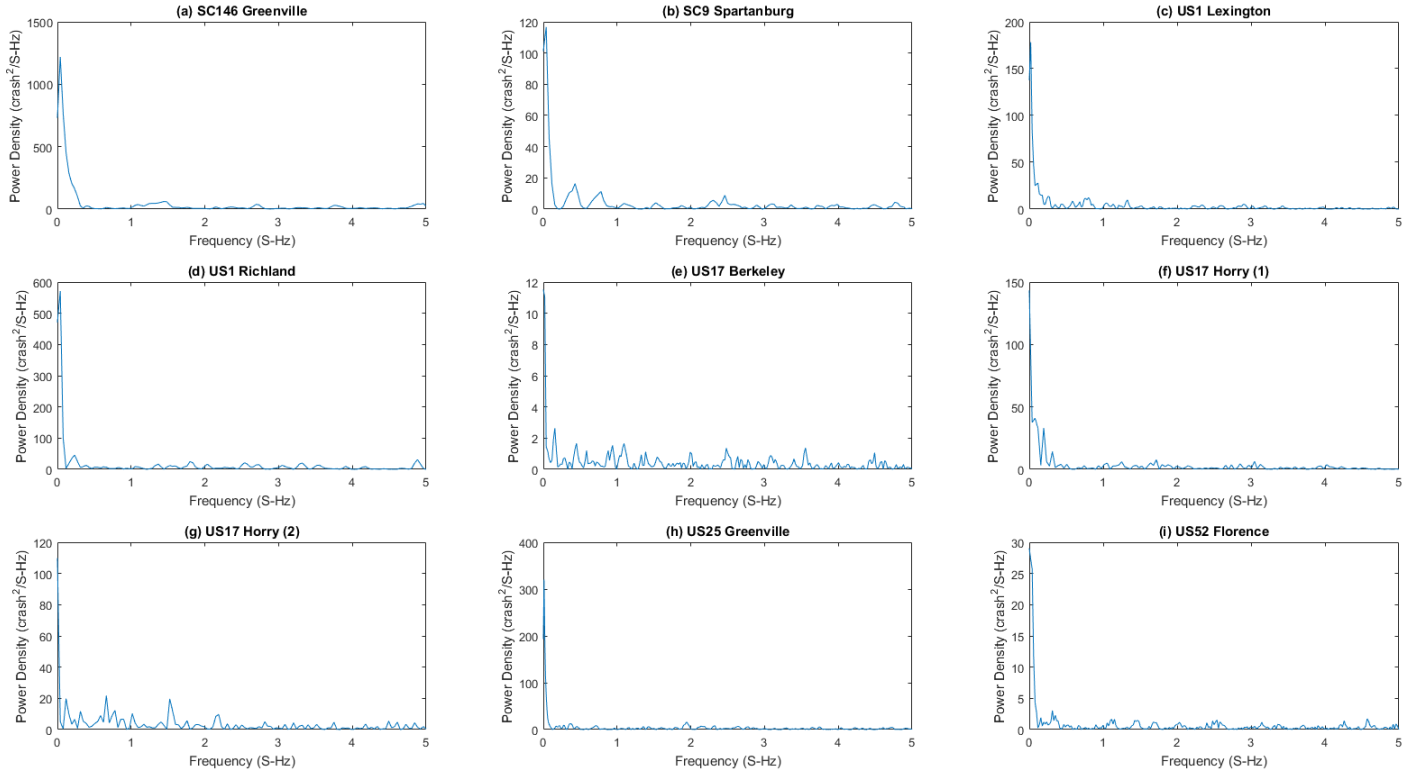


FIGURE 4 PSDs corresponding to CDSs in Figure 3

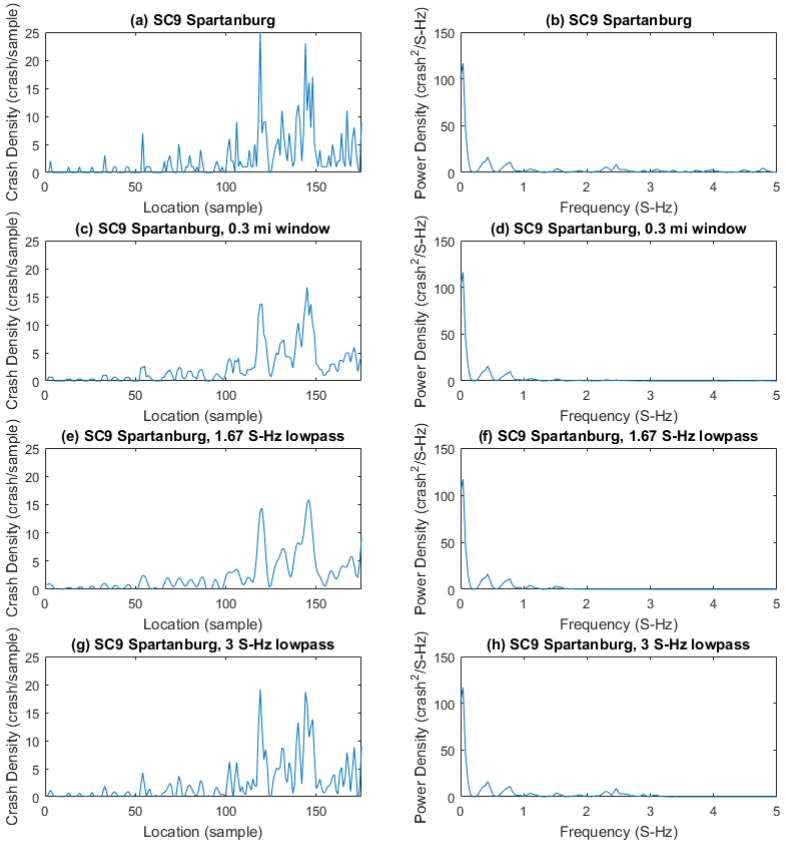


FIGURE 5 The corridor of SC 9 in Spartanburg County using SW method and LPF method (Butterworth).

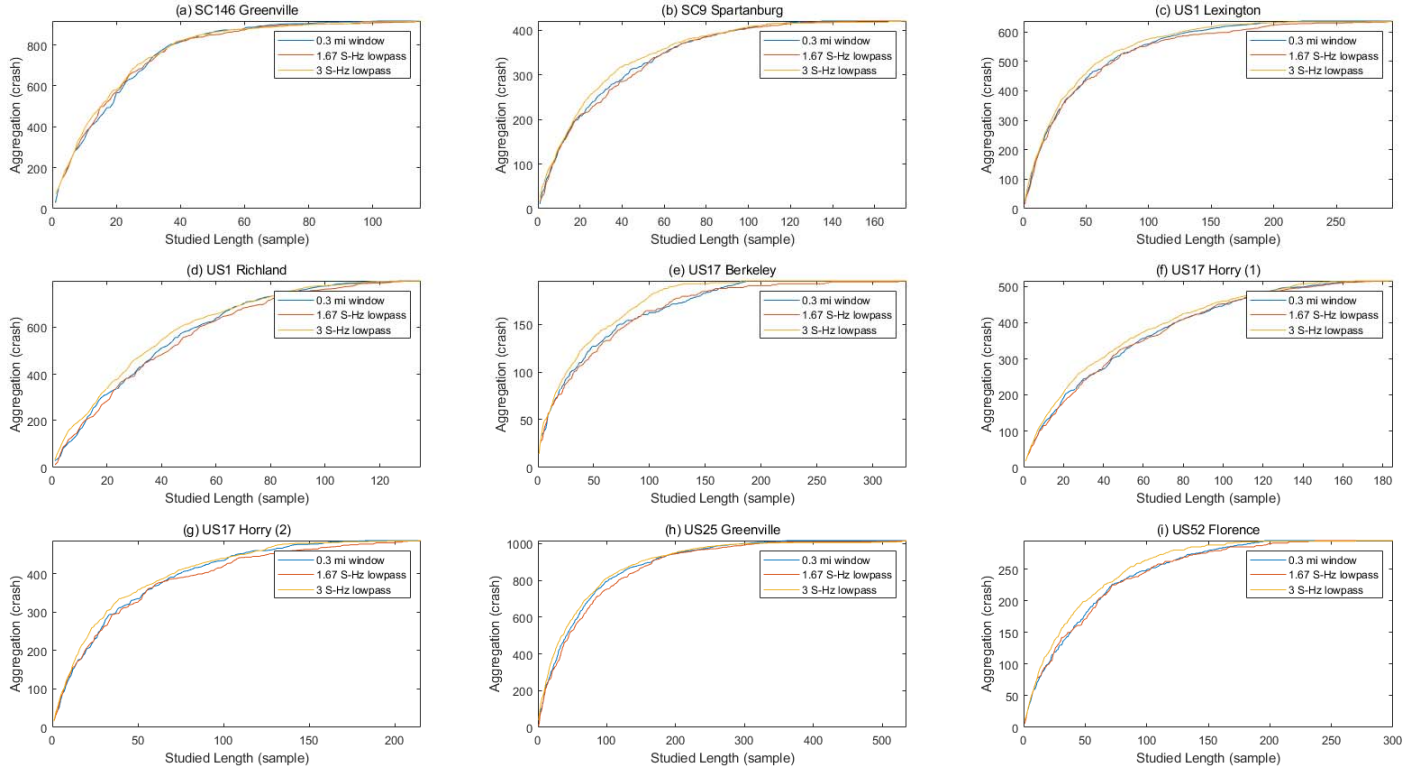


FIGURE 6 Percentages of aggregated crashes by percentages of length of top-risk region

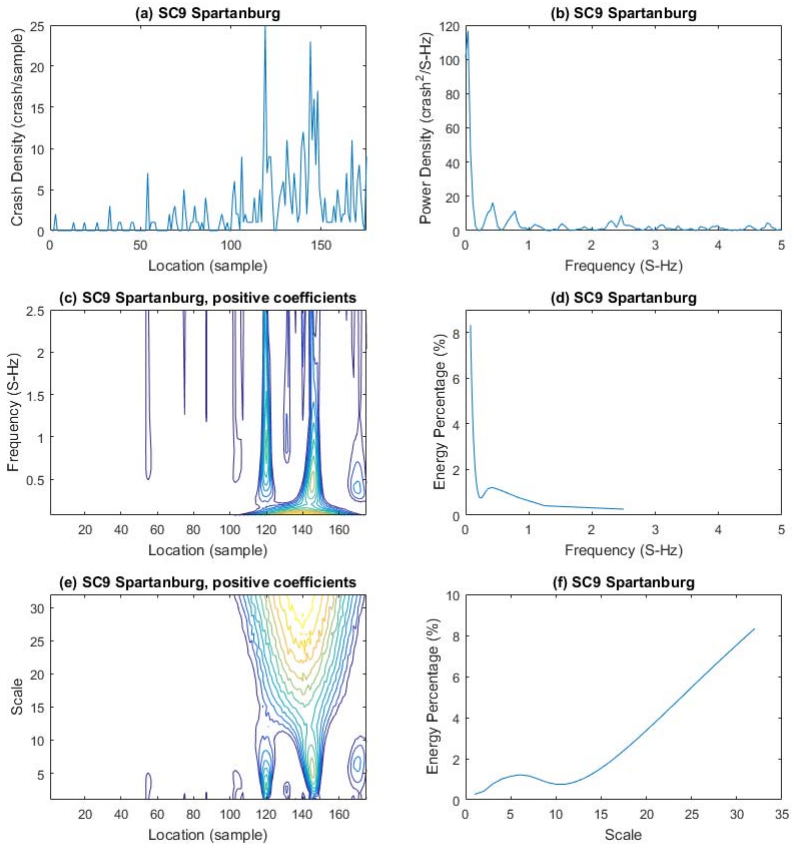


FIGURE 7 The corridor of SC 9 in Spartanburg County using spatial wavelet analysis

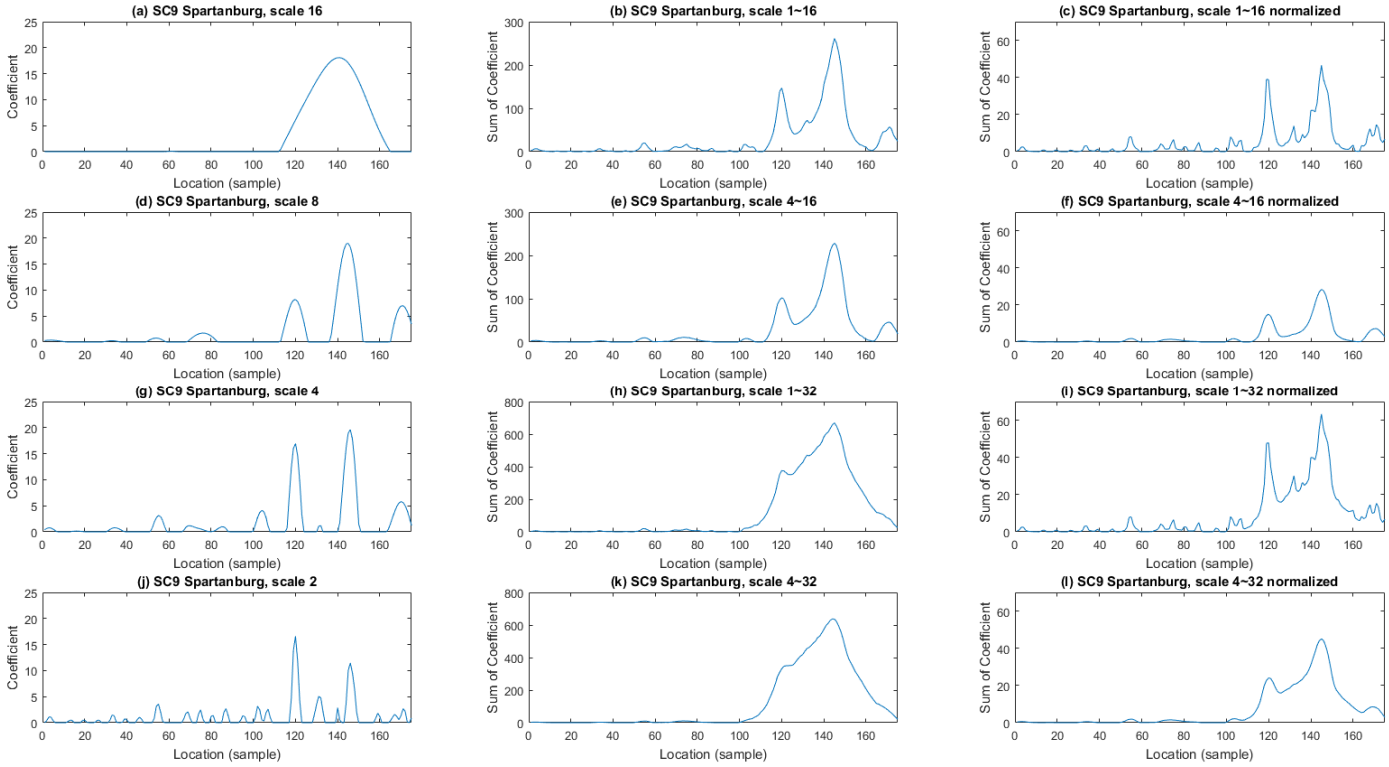


FIGURE 8 The corridor of SC 9 in Spartanburg County using different components for identifying high-risk segments