EXAMINING FIXED-LENGTH SEGMENTATION FOR TRAFFIC SAFETY ANALYSIS: OPTIMAL SEGMENT LENGTH BY SPECTRAL-BASED SEGMENTATION

Xi Zhao
Wuhan University of Technology, China
and
Clemson University, USA
Phone: +86 18971270282
E-mail: xiz@clemson.edu

Dominique Lord
Professor
Zachry Department of Civil Engineering
Texas A&M University
Phone (979) 458-3949
E-mail: d-lord@tamu.edu

Yichuan Peng, Corresponding Author
Assistant Professor
Tongji University
Department of Transportation Engineering
Phone: +86 13795316798
E-mail: yichuanpeng1982@hotmail.com
ABSTRACT
The Highway Safety Manual (HSM) presents a variety of methods for quantitative network segmentation. Existing approaches to conduct segmentation for safety analysis require engineering judgement and are subject to a lack of standard metrics for assessing segmentation performance. This paper presents a novel methodology for fixed-length segmentation. It provides general solutions to determining optimal segment lengths and innovating network segmentation methods for rigorous safety analysis. The methodology is based on spectral analysis of crash density in the spatial frequency domain (SFD) in which low frequency components represent trends while high frequency components represent details and randomness. 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 determine optimal segment lengths and the power spectral percentage (PSP) for assessing the segmentation performance. Based on those new analysis tools, 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. The methodology extended the knowledge of network segmentation and aggregation of crash data from a non-traditional perspective. Those new methods relying on this methodology are easy to implement and ready for practical application.

Keywords: crash distribution, segmentation, segment length, spectral analysis, sliding window, wavelet analysis
1 INTRODUCTION

Motor vehicle crashes are among the leading causes of death in the United States (WISQARS, 2018). 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 (Qin and Wellner, 2012). 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 prediction, and the development and application of crash modification factors (CMFs) (AASHTO, 2010). 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 segment length is one of the long-standing problems researchers and practitioners frequently confront in highway safety analysis. Thus, a critical question arises: “What is the optimal segment length for the aggregation of crash data for safety analysis?” 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 determine segment lengths, and many of them arbitrarily choose empirical lengths or simply refer to case study lengths in the Highway Safety Manual (HSM) (2010).

This paper describes how to conduct fixed-length segmentation in safety analysis, in particular those related to network screening. 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 a novel methodology for finding optimal segment lengths and improving network segmentation methods to support rigorous safety analysis.

2 BACKGROUND

2.1 Segmentation Approaches

The American Association of State Highway Transportation Officials (AASHTO) published the HSM, which provided analytical tools and techniques to improve safety decision making (AASHTO, 2010). Most 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 (Koorey, 2009). 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 overly sensitive 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). 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 (2010) and many studies (Bahar and Hauer, 2014; Cafiso et al., 2018; Koorey, 2009), 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 limited by the availability of roadway datasets that have a sufficient inventory. Koorey (2009) 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 et al. (2018) 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 (Dueker and Vrina, 1992). 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. (2014) 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.

### 2.2 Segment Length

Segment length has been a long-standing concern of researchers and varying lengths have been used in this field. 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 (Geyer et al., 2008). A segment length of 0.125 miles was used in New Zealand
(Koorey, 2009). 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 (AASHTO, 2010) for the sliding window 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 (1993) 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 (1994) 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 (1997) 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. (2011) empirically indicated that short segments length (less than 0.1 mile) lead to unreliable results in safety analysis. Cook et al. (2011) 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 (2012) 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 (2014) recommended to regroup sub-segments shorter than 0.1 miles to a length of 0.1 miles, as a minimum, and calculate a combined average CMF.

3 SPATIAL FREQUENCY DOMAIN ANALYSIS

3.1 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. (2014) 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 components 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.
3.2 One Dimensional Space of Linear Feature

Crashes are originally distributed in 3-dimensional space, and 2-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 \( X_i \) in original space \( V \), if the distance (analogous to the offset distance) from its location \( x_i \) to the 1-D space \( L \) (finite space determined by the linear feature) is less than a certain value \( \varepsilon \), then \( X_i \) is projected to \( Y_i \) in \( L \). The threshold \( \varepsilon \) can be determined by the width of the roadway and the accuracy of the crash location in the police reports (Sarasua et al., 2015), and it is also dependent on safety analysis tasks.

3.3 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 \( B_j \) on the linear feature can be determined by

\[
\hat{d}(y) = \frac{1}{h} \sum_{i=1}^{n} \sum_{j=1}^{m} I(y_i \in B_j) I(y \in B_j)
\]

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 an artifact of interval locations which has little effect on SFD analysis 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 (Hastie et al., 2009). 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 \( \hat{d}(y) \) by:

\[
\hat{d}(y) = \sum_{i=1}^{n} K_h(y - y_i) = \frac{1}{h} \sum_{i=1}^{n} K \left( \frac{y - y_i}{h} \right)
\]

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. Boroujerdan et al. (2014) 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 discretization processing can be mathematically equal to a histogram.

The output from the 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). Figure 1 (a, c, and e) shows the example of an output 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 signal or 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 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 practical applications, the 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).

![Figure 1](image_url) **Figure 1** The CDS and corresponding PSD of the linear feature of SC 146 in Greenville with the 2013 crash data at different sampling rates.

### 3.4 Spatial Frequency Domain

Spatial frequency is a measurement for events that are periodic across location in space (Boreman, 2001). 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

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

where \( N \) is the total number of samples of \( x_n \), and it is also the total number of samples of \( X_k \). The CDS \( x_n \) is a spatial domain representation which contains spatial information while \( X_k \) is an SFD representation which contains spatial frequency information. 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 indicate 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 on the linear feature. The spatial correlation of crash distribution is essentially governed by both deterministic factors and stochastic factors of transportation systems.

### 3.5 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

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

where \( \Delta t \) is sampling period. The range of \( f \) is determined by the Nyquist frequency \((\text{Shannon}, 1949)\). 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 band of its PSD according to the Nyquist–Shannon sampling theorem \((\text{Shannon}, 1949)\). As shown in Figure 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 power distribution within the high frequency band.

### 3.6 A Hypothesis for Low Frequency Components

As a type of presentation of the safety performance of transportation systems, a CDS should contain both systematic components and random components. Systematic components
indicate the major pattern in which crashes are distributed along the linear feature, which is presented as a 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 and randomness of occurrence of crashes, which are 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. Figure 1 (b, d, and f) shows the PSD of the corridor of SC 146 in Greenville with the 2013 crash data at sampling rates of 5 S-Hz, 10 S-Hz, and 20 S-Hz, respectively. The most significant part of this CDS’ power 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 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 section.

4 EXPERIMENTS DESIGN

4.1 Data Collection

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 location data. South Carolina began their deployment of one such system: the South Carolina Collision and Ticket Tracking System (SCCATTS) in 2010. SCCATTS provide 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. Figure 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).
4.2 Data Processing

Experiments were designed to explore spatial frequency characteristics of crash data using the SFDA. The experiments were implemented using ArcMap 10.2 and MATLAB R2016a. The procedure of the SFDA implemented in the experiments included the following steps:

1. Projecting crash instances recorded in crash data into the 1-D space determined by a linear shapefile feature of the roadway data.
2. Quantifying the projected crash data to generate the CDS of that linear feature. ($\varepsilon$ was set to 30 feet (Sarasua et al., 2015)).
3. Performing the DFT and SDE (periodogram) for the Fourier transform and the PSD of the CDS, respectively.

Besides PSD plots, three percentage metrics were used to quantitatively identify the proportion of power concentrated within low frequency bands (see Table 1): 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 data and roadway data were used in the experiments. The roadway data included linear features from nine corridors (see Table 1) which were selected from among the top eleven risky corridors that were identified as dangerous and representative in a project sponsored by SCDOT (Sarasua et al., 2015).
<table>
<thead>
<tr>
<th>Corridor</th>
<th>Length (mi)</th>
<th>Crash Count (crash)</th>
<th>Average Density (crash/mi)</th>
<th>Description</th>
<th>Power% (0–0.5 S-Hz /0–2 S-Hz) (1 mi /0.25 mi)</th>
<th>Power% (0–2 S-Hz /0–5 S-Hz) (0.25 mi/0.1 mi)</th>
<th>Power% (0–3 S-Hz /0–5 S-Hz) (0.1667 mi/0.1 mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC 146 Greenville</td>
<td>11.5</td>
<td>917</td>
<td>79.74</td>
<td>suburban</td>
<td>86.60%</td>
<td>84.47%</td>
<td>89.24%</td>
</tr>
<tr>
<td>SC 9 Spartanburg</td>
<td>17.5</td>
<td>421</td>
<td>24.06</td>
<td>suburban</td>
<td>77.85%</td>
<td>77.34%</td>
<td>87.79%</td>
</tr>
<tr>
<td>US 1 Lexington</td>
<td>29.5</td>
<td>637</td>
<td>21.60</td>
<td>rural / suburban</td>
<td>73.65%</td>
<td>83.26%</td>
<td>92.22%</td>
</tr>
<tr>
<td>US 1 Richland</td>
<td>13.5</td>
<td>797</td>
<td>59.04</td>
<td>suburban</td>
<td>84.09%</td>
<td>78.29%</td>
<td>86.47%</td>
</tr>
<tr>
<td>US 17 Berkeley</td>
<td>33.0</td>
<td>195</td>
<td>5.91</td>
<td>rural / suburban</td>
<td>57.53%</td>
<td>65.94%</td>
<td>79.15%</td>
</tr>
<tr>
<td>US 17 Horry (1)</td>
<td>18.5</td>
<td>515</td>
<td>27.84</td>
<td>suburban / urban</td>
<td>84.56%</td>
<td>82.95%</td>
<td>89.70%</td>
</tr>
<tr>
<td>US 17 Horry (2)</td>
<td>21.5</td>
<td>486</td>
<td>22.60</td>
<td>rural / suburban</td>
<td>54.04%</td>
<td>74.31%</td>
<td>85.01%</td>
</tr>
<tr>
<td>US 25 Greenville</td>
<td>53.5</td>
<td>1017</td>
<td>19.01</td>
<td>rural / suburban</td>
<td>77.71%</td>
<td>73.23%</td>
<td>86.65%</td>
</tr>
<tr>
<td>US 52 Florence</td>
<td>30.0</td>
<td>295</td>
<td>9.83</td>
<td>rural / suburban</td>
<td>78.08%</td>
<td>73.81%</td>
<td>83.60%</td>
</tr>
</tbody>
</table>
5 RESULTS AND DISCUSSION

5.1 Computation Results

The CDSs quantified based on the roadway data and the crash data are shown in Figure 3; and the corresponding PSD for each CDS is shown in Figure 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 1. According to the first metric, 7 corridors had 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 had over half (54.04% and 57.53%) power concentrated in the lowest 1/4 bandwidth. According to the second metric, 8 corridors had 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 had over half (65.94%) power concentrated in the lowest 2/5 bandwidth. According to the third metric, all nine corridors had 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 US 17 in Berkeley County had considerable proportion of power within relatively high frequency bands. The reason was that it had significantly fewer crashes than other corridors.
Figure 3 CDSs of nine analyzed corridors at the sampling rate of 10 S-Hz

Figure 4 The PSD corresponding to each CDS in Figure 3.
5.2 Hypothesis and Findings
PSD plots and all metrics justified the hypothesis that the CDS of a corridor had major power concentrated in low frequency bands. Although the extents of power spectral concentration varied for different corridors, all CDSs presented 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 low power distributed in the corresponding bands. The meaning of high frequency components makes them even less impacting in analyzing crash data because higher frequency components more likely represent randomness of occurrence of crashes as noise existing regardless of locations in the spatial domain. Thus, adequately filtering out high frequency components can remove fine details and randomness while preserving the primary information and a certain level of detailed information of crash distribution. Similarly, using a low sampling rate to quantify CDSs may cause loss of fine details and randomness that contain very limited power.

5.3 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, the 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 SC 146 Greenville (see Figure 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 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 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 relatively high frequency bands (like the corridor US 17 in Berkeley County), a carefully chosen threshold frequency can still lead to a practical 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, improvements in inference power could be very limited if the optimal PSSL is determined by a carefully chosen threshold frequency.

5.4 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. In the example of the corridor SC 146 in Greenville County, the segment length of 1 mile can preserve 86.60% (see Table 1) power compared to a baseline of 0.25 mile, which can be represented by the power spectral percentage (PSP) as $PSP_{0.25mi}^{1mi} = 86.60\%$. The corresponding PSSL can be represented as $PSSL_{0.25mi}^{86.60\%} = 1mi$. Similarly, for the corridor US 1 in Lexington County, a $PSSL_{0.1mi}^{83.26\%}$ of 0.25 miles would imply a $PSP_{0.1mi}^{0.25mi}$ 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. (2011) conducted a sensitivity analysis which employed ranking list shifts of segments ranked with the Iowa DOT scoring method. The output of ranking list shifts is not intuitive in its interpretation. Qin and Wellner (2012) conducted a sensitivity analysis, which calculated the miles of high-risk highway that can be identified with different segment lengths. This method can barely provide sufficient information for making practical decision. Boroujerdian et al. (2014) 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 macroscopically describe the extent to which the segment length can capture detailed 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 Equation (5), 2) the power would not be evenly distributed in the SFD for any real data as shown in Figure 4, and 3) components at different frequencies represent different spectral characteristics.
6 SPECTRAL-BASED SEGMENTATION

6.1 Low-Pass Filtering vs Sliding Window

The sliding window (SW) method uses a window to conceptually move along the road segments from beginning to end (AASHTO, 2010). 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 proposed 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 of crash distribution in order to avoid impacts of uncertainty and randomness on safety analysis. The implementation is straightforward: (1) use SFDA to determine the threshold frequency (or 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 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 the following advantages:

1. The LPF method allows window size (or PSSL) to be a rational number multiple of the sampling/interval. On the other hand, the SW method can only use a window size of integer multiple of sample/interval. As a result, the LPF method is capable of identifying and applying segment lengths that is beyond the limitation of integer range. The authors believe that the theoretically optimal segment length is more likely to be a rational number multiple of sample/interval rather than an integer multiple of sample/interval for real data.

2. The LPF method performance can be reliably evaluated. The proposed PSP provides a numerical metric to evaluate segmentation methods. The LPF method can use PSP to estimate ideal performance before processing and actual performance afterward. On the other hand, the SW method can only be evaluated afterward, and its performance evaluation is more biased than the LPF method that uses a properly designed filter (see the following experiments).

3. The LPF method outperforms the SW method in identifying high-risk segments. The capability of identifying and applying optimal segment length in an efficient manner within an extended range enable the LPF method superior performance in all aspects. Additionally, the LPF method outperforms the SW method in identifying top-risk segments even if they use exactly same length (see the following experiments).

An example based on the corridor SC 9 in Spartanburg County is shown in Figure 5. Figure 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. Figure 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 Equation
(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 same 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. The Butterworth LPF provides clean results because it adequately filters out unwanted components and preserves desired components. An ideal low-pass filter with a threshold frequency of 1.67 S-Hz can capture 75.81% power of the CDS (the $PSP_{0.1\,mi}^{0.3\,mi}$ of 75.81%). Compared with above results, a Butterworth LPF with a threshold frequency of 3 S-Hz (equivalent to a PSSL of 0.1667 mi) preserves finer crash distribution information (see Figure 5 (g) and (h)), and it can capture 87.79% power of the CDS (the $PSP_{0.1\,mi}^{0.1667\,mi}$ of 87.79%).
Figure 5 The corridor SC 9 in Spartanburg County using sliding window method and low-pass filtering method (Butterworth).

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. (2014). Similar SW and LFP segmentations were evaluated using the data of four corridors that have different average crash densities (see Figure 6 and Table 2).
Unsurprisingly, the LPF method with a threshold frequency of 3 S-Hz (equivalent to a PSSL of 0.1667 mi) can aggregate crashes in a more efficient manner than other two segmentation methods. Figure 6 clearly shows the comparison. Compared with the graph, the proposed PSP metric provide more intuitive evaluation about their performance (see Table 2).

Interestingly, the LPF method with a threshold frequency of 1.67 S-Hz can aggregate more crashes within the top-risk samples/intervals than the SW method with a window of 0.3 miles. Specifically, using those two methods to rank and identify high-risk regions, the top-risk regions identified by the LPF method have more crashes than those identified by the SW method, even though they use equivalent PSSL. However, for regions that have relatively fewer crashes, the SW method shows higher aggregate rates of crashes than the LPF method. Those two methods have similar performance for regions that have very few crashes. For example, regarding the corridor SC 9 in Spartanburg county, the LPF method can better identify top 5.14% risky regions than the SW method, and those regions have 29.69% crashes of all crashes on the corridor (see Table 2). All tested data present the similar patterns. 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 same PSSL.
Figure 6 Percentages of aggregated crashes by percentages of length of top-risk regions
Table 2 Performance evaluation based on percentages of aggregated crashes and power spectral percentage

<table>
<thead>
<tr>
<th></th>
<th>SC 146 Greenville</th>
<th>SC 9 Spartanburg</th>
<th>US 1 Richland</th>
<th>US 52 Florence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregated Length (LPF won SW)</td>
<td>26.09%</td>
<td>5.14%</td>
<td>8.88%</td>
<td>12.67%</td>
</tr>
<tr>
<td>Aggregated Crashes (LPF won SW)</td>
<td>78.08%</td>
<td>29.69%</td>
<td>25.09%</td>
<td>50.85%</td>
</tr>
<tr>
<td>PSP (0.3 mi window)</td>
<td>80.26%</td>
<td>74.71%</td>
<td>73.69%</td>
<td>69.33%</td>
</tr>
<tr>
<td>PSP (1.67 S-Hz low-pass)</td>
<td>84.84%</td>
<td>77.93%</td>
<td>73.30%</td>
<td>70.45%</td>
</tr>
<tr>
<td>PSP (3 S-Hz low-pass)</td>
<td>90.13%</td>
<td>89.45%</td>
<td>87.32%</td>
<td>84.61%</td>
</tr>
</tbody>
</table>

6.2 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 (Boroujerdian et al., 2014). 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 (low frequency) components indicate the trend and basic structure of crash distribution while small scale (high frequency) components indicate fine details and randomness. A rigorous wavelet-based segmentation requires proper understanding and interpretation of spatial wavelet analysis of crash data and the results, which is lacking in existing research (Boroujerdian et al., 2014).

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 (Gao and Yan, 2011). The pseudo-frequency $f$ corresponding to a scale $a$ is determined by the center frequency $f_c$ of the chosen wavelet function and the sampling period $\Delta t$ of the CDS, as

$$f = \frac{f_c}{a \Delta t} \quad (6)$$

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 detail. 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 analysis using scalogram, overly-large 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 effective interpretation of spatial wavelet analysis.

An example based on the corridor SC 9 in Spartanburg County is shown in Figure 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 Figure 7 (e)) or the corresponding frequency band of 0.08–2.5 S-Hz (see Figure 7 (e)). The wavelets’ energy distribution in terms of frequency has the same pattern as the PSD of the CDS: the most significant part of the energy is concentrated within the low frequency band (large scale range) (see Figure 7 (b, d, and f)). In this spatial wavelet analysis, 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 identify high-risk segments are critically dependent on proper interpretation of the wavelet transform of crash data and the results. Boroujerdian et al. (2014) 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 the methodology without theoretical 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. As shown in Figure 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 globally riskier because they do neither include more fundamental trends contained by larger scale components nor finer details contained by smaller scale components.

Based on the SFDA and the spatial wavelet analysis discussed above, the authors propose an improved method to divide and identify globally high-risk segments by integrating the components within a proper scale range (or frequency band) for more comprehensive crash distribution information. Firstly, a proper lower limit of the scale range is required depending on the extent to which the details are needed (see Figure 8 (b and e) for the results using different lower limit of scale range for integrating components). Secondly, a proper upper limit of the scale range is required to avoid overweighting the fundamental trend (see Figure 8 (h and k) for the results that overweight components at large scales). Thirdly, a normalization process is required to allow components at different scales to be weighted equally. The normalization method is dividing all coefficients by the corresponding scales. The improved method can provide intuitive information for dividing and identifying globally high-risk segments (see Figure 8 (c, f, i, and l) for the results using different scale ranges). It is worth mentioning that using very small-scale range can lead to results similar to using a single scale, however the normalization process can avoid biases towards overweighted components. The improved method can be implemented using various scale range for effective decision making depending on the demands of safety analysis tasks.
Figure 7 The corridor SC 9 in Spartanburg County using spatial wavelet analysis
7 CONCLUSIONS

Proper network segmentation is necessary for conducting rigorous safety analysis. Traditional approaches to determine segment lengths in safety analysis require engineering judgment and are subject to a lack of standard metrics to assess performance. Alternatively, this paper proposes a novel methodology for fixed-length segmentation in traffic safety analysis. This methodology provides general solutions for determining segment lengths and improving network segmentation, based on the natural characteristics of crash distributions and the extent to which detailed information about crash distributions is needed.

This research demonstrates the feasibility of exploring natural characteristics of crash data using spectral analysis. The analysis implies 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. Based on those findings, the paper proposes the PSSL for determining optimal segment lengths and the PSP for assessing the performance of segmentation. The methodology also leads to the low-pass filtering method that outperforms the sliding window method in network screening and an improved wavelet-based method that identifies high-risk segments properly.

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 $PSP_{0.25 \text{mi}}$ 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 the energy of the crash distribution and achieve a $PSP_{0.1 \text{mi}}^{0.1667}$ of 83.50%–92.22%.

Applying the proposed methodology in practical safety analysis is not as complex as the theoretical elaboration in this paper. Computing optimal PSSL and corresponding PSP, the low-pass filter 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 crash location information, and the extent to which professionals understand crash distribution in spatial frequency domain. 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 (Lord and Mannering, 2010). 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 studied in order to reveal the nature of crash incidence in transportation systems.

ACKNOWLEDGMENTS

The author would like to acknowledge Wayne Sarasua and Kweku Brown for literary input and providing data. The author would like to acknowledge Stephen Fry for literary input. The author would like to acknowledge Douglas Dawson for constructive comments.

REFERENCES


