Development of a Functional Conflict-based Safety Performance Function for Signalized Intersections Using the pNEUMA Data

Tianyu Shen, M.S.
Ph.D. Student
Department of Transportation and Urban Infrastructure Studies
Morgan State University, Baltimore, MD, U.S. 21251
Email: tishe7@morgan.edu

Di Yang, Ph.D., Corresponding author
Assistant Professor
Department of Transportation and Urban Infrastructure Studies
SMARTER University Transportation Center
Morgan State University, Baltimore, MD, U.S. 21251
Email: di.yang@morgan.edu

Kun Xie, Ph.D.
Associate Professor
Department of Civil and Environmental Engineering
Old Dominion University, Norfolk, VA, U.S. 23529
Email: kxie@odu.edu

Hong Yang, Ph.D.
Associate Professor
Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, VA, U.S. 23529
Email: hyang@odu.edu

Xianfeng (Terry) Yang, Ph.D.
Associate Professor
Department of Civil and Environmental Engineering
University of Maryland, College Park, MD, U.S., 20742
Email: xtyang@umd.edu

Mansoureh Jeihani, Ph.D.
Professor
Department of Transportation and Urban Infrastructure Studies
SMARTER University Transportation Center
Morgan State University, Baltimore, MD, U.S., 21251
Email: mansoureh.jeihani@morgan.edu

Word Count: 6681 words + 3 table (250 words per table) = 7431 words

Submitted August 1st, 2024

ABSTRACT

Conflict-based Safety Performance Functions (SPFs) are commonly used to model the relationship between traffic conflicts and various traffic parameters, typically based on data aggregated at specific temporal levels, such as hourly, 15-minute intervals, or per signal cycle. However, such temporal data aggregation is insufficient for investigating safety risk changes within signal cycles. This research proposes a new functional conflict-based Safety Performance Function using the Functional Data Analysis (FDA) approach. In this approach, the number of conflicts and their corresponding exposure and safety risk factors are modeled as functions with respect to time within signal cycles, rather than being aggregated. Functional data smoothing is applied to smooth the data, and functional linear regression is employed to develop the functional conflict-based SPF. The results indicate significant temporal variation in the effects of safety risk factors, specifically the number of moving vehicles and backward-forming shock wave speed, on traffic conflicts. A comparative study is conducted to evaluate the proposed functional conflict-based SPF against traditional aggregated SPFs, demonstrating the superiority of the functional approach. The proposed functional conflict-based SPF shows potential for designing more effective proactive safety management strategies.

Keywords: Functional data analysis, Functional linear regression**,** Functional conflict-based SPFs

INTRODUCTION

Traffic safety at urban signalized intersections is a critical concern due to the significant loss of lives in traffic crashes around the world (*1*). The primary factors contributing to the high safety risk at signalized intersections are the complex traffic signal changes and conflicting traffic movements (*2*). These factors often lead to repeated stop-and-go situations, which in turn forms shock waves (*3*). Thus, it is vital to properly model and investigate the relationship between traffic safety risk and its risk factors at signalized intersections. Traditionally, crash-based safety performance functions (SPFs) are often developed to examine the relationship between crash frequency and its corresponding exposure and safety risk factors (*4*), where traffic crashes, exposure (often traffic volume), and safety risk factors are often aggregated annually (*5*). The decision of using annual aggregation level is mostly because of the rarity of traffic crashes, which requires relatively long period of time to accumulate enough crash samples for the development of SPFs.

Despite the valuable contribution of crash-based SPFs developed for signalized intersections, the rarity of crashes poses difficulties in terms of capturing more detailed safety patterns at higher granularity at signalized intersections. To overcome this challenge, traffic conflicts have gained popularity in assessing traffic safety risk at signalized intersections in recent years (*4*). Compared to traffic crashes, traffic conflicts occur more frequently and can be automatically extracted from detailed vehicle trajectory data (*6*). As a result, conflict-based SPFs have been developed to model the relationship between the number of traffic conflicts and their exposure and safety risk factors at more detailed temporal aggregation levels, such as hourly (*7; 8*), 15-minute (*9*), and the signal cycle level (*3; 10*). Studies that used traffic signal cycle as the temporal aggregation level (also the smallest temporal aggregation level) have shown that shock wave parameters, such as shock wave speeds, areas, etc., have statistically significant associations with the number of traffic conflicts (*3; 10*).

However, there is evidence in past literature that suggests that safety risk varies at different time points within signal cycles (*3; 11*), and safety risk factors, such as shock wave parameters, also changes within signal cycles (*10*). Furthermore, traffic conflicts and safety risk factors extracted from high-granular vehicle trajectory data possess very detailed spatio-temporal information. Thus, aggregating traffic conflicts and safety risk factors at the signal cycle level still may not be optimal and could result in potential information loss, thereby hindering a comprehensive understanding of safety risk at signalized intersections.

Therefore, this research proposes a new functional conflict-based SPF for signalized intersections, aiming to address the limitation of temporal data aggregation in current safety literature. Specifically, we propose to employ the Functional Data Analysis (FDA) approach in statistics and model the number of conflicts and its corresponding exposure and safety risk factors as functions with respect to time within signal cycles. Among various techniques in the FDA framework, we propose to use the functional data smoothing method for modeling the number of conflicts and its corresponding exposure and safety risk factors, and the functional linear regression method for establishing the functional conflict-based SPF between the number of conflicts and its exposure and safety risk factors, including shock wave parameters. As a case study, the pNEUMA dataset, an open dataset developed by Barmpounakis and Geroliminis (*12)*, is used and one pre-timed signalized intersection is chosen for the analysis. By modeling conflicts and safety risk factors as time series rather than aggregating them, more detailed understanding of their relationship can be uncovered, which can benefit both researchers and practitioners on the development of proactive safety evaluation and management strategies.

LITERATURE REVIEW

Conflict-based SPFs at signalized intersections

Conflict-based SPFs at signalized intersections are statistical models that model the relationship between the number of traffic conflicts (dependent variable) and exposure and various safety risk factors (independent variables) (*3*). The developed conflict-based SPFs can better quantify the potential safety risk and facilitate proactive safety assessments at signalized intersections. Different types of conflicts have been investigated for the development of conflict-based SPFs, such as left-turn conflict (*8; 9*), rear-

end conflict (*3; 8; 10*), or a combination of different types of conflicts (*7*). Different Surrogate Safety Measures (SSMs) have also been explored when developing conflict-based SPFs, such as time to collision (TTC) (*3; 9; 10*), modified time to collision (MTTC) (*3*), deceleration rate to avoid crash (DRAC) (*3*).

Additionally, many temporal data aggregation levels have been used during the development of conflict-based SPFs. For example, Sacchi and Sayed (*8)* used the hourly aggregation level and developed conflict-based SPFs in predicting the number of specific types of conflicts (i.e., rear-end and left-turn). El-Basyouny and Sayed (*7)* proposed a two-phase model that models first the average hourly conflicts with exposure, area type (e.g., urban or suburban), the number of through lanes and the presence of right and left turn lanes; and second the collisions based on conflicts, which indicated a significant proportional relationship exists between conflicts and collisions.

In addition to aggregating the data hourly, Zhang et al. (*9)* used 15-minute data aggregation level developed conflict-based SPFs to model left-turn conflicts. Specific safety risk factors related to left-turn characteristics such as the presence of white line extension for the left-turn lane, the average turning radius for the left-turn traffic movement, and the green time allocated to the left-turn movement were explored.

Besides, two studies (*3; 10*) aggregated the data at the signal cycle level and developed conflict-based SPFs that can account for more detailed shock wave parameters that are unique to signalized intersections. Specifically, the following shock wave parameters are explored, including shock wave area and backward-forming shock wave speed, as well as maximum queue length, platoon ratio, and traffic volume. A positive effect for shock wave area and a negative effect for backward-forming shockwave speed on traffic conflicts were identified.

As can be seen from the above discussion, the temporal aggregation levels have become more detailed over the years, with the traffic signal cycle level being the smallest aggregation level. By continuously improving the granularity of the temporal data aggregation levels, more detailed safety risk patterns at signalized intersections have been unveiled. However, even with the signal cycle level, the relationship between the traffic conflicts and safety risk factors, such as shock wave parameters, at each time point inside signal cycles remains unexplored. This research addresses the identified gap by proposing to use the FDA approach to develop a functional conflict-based SPF, aiming to capture more detailed relationships between shock wave parameters and the number of traffic conflicts within signal cycles.

Applications of FDA in transportation

FDA is a statistical framework designed for analyzing curves or functions over a continuum, which provides a comprehensive characterization of time series data. Representative studies that applied FDA approach in the transportation domain are summarized in **TABLE 1**. Among all the fourteen identified transportation studies, observed time series data were converted into functional curves using functional data smoothing methods as a preliminary step, while different FDA techniques were adopted for subsequent analysis based on different research purposes.

Many of these studies focused on predicting or analyzing traffic flow (*13-17*), while only two studies investigated traffic safety at signalized intersections from a functional perspective. Yang et al. (*11)* introduced the FDA approach to the transportation safety field and explored the safety risk levels for different traffic movements at different time points within signal cycles. The positivity constraint has been added during the functional data smoothing process in modeling traffic safety risk due to the nonnegativity property of the safety risk values. Based on the findings, Yang et al. (*18)* further explored the use of the FDA approach in detecting safety-related anomalies for proactive safety monitoring.

Besides, limited studies adopted functional regression in transportation domain. Briefly, Yang et al. (*19)* adopted nonparametric functional linear regression to identify the differences in driver response behavior to the speed compliance warning between the treatment and control groups; Shah et al. (*20)* focused on forecasting day-ahead traffic flow by using functional autoregression. Crawford, Watling and Connors (*17)* introduced functional linear regression to analyze systematic variations in daily traffic flow profiles based on known explanatory factors such as the day of the week and the season.

As can be seen from the above discussion, no studies have explored the relationship between functional traffic safety risk and safety risk factors, such as shock wave parameters, at signalized intersections, which is addressed in this research.

TABLE 1 Previous studies using FDA method

ID	Study	Research topic	Variable modeled as functions	FDA Approaches
1	Chiou (*13)*	Predicting traffic flow.	Traffic flow	Functional data smoothing, FPCA (Functional Principal Component Analysis), probabilistic functional classification, functional linear regression.
2	Guardiola, Leon and Mallor (*15)*	Analyzing different patterns of traffic flow.	Traffic flow	Functional data smoothing, FPCA.
3	Chiou et al. (*14)*	Impute missing values in traffic flow and detect outliers.	Traffic flow	Functional data smoothing, FPCA, functional bagplot and functional highest density region boxplot.
4	Sudweeks (*21)*	Detecting dangerous driving behaviors from videos recorded by cameras installed in vehicles.	Yaw rate	Functional data smoothing, Curve registration, Functional classification.
5	Seya, Yoshida and Tsutsumi (*22)*	Ex-post identification of the geographical extent of an area benefiting from a transportation project	Land price	Functional data smoothing, functional ordinary Kriging, functional clustering.
7	Wagner-Muns et al. (*16)*	Predicting traffic flow	Traffic flow	Functional data smoothing, FPCA.
8	Zhong et al. (*23)*	Predicting link travel time.	Travel time	Functional data smoothing, FPCA.
9	Hu et al. (*24)*	Analyzing drivers' behavior response.	Vehicle trajectories (longitude and latitude)	Functional data smoothing.
10	Yang et al. (*18)*	Detecting safety-related anomalies for proactive safety monitoring	Traffic safety risk measured by traffic conflicts	Functional data smoothing, Functional depth measures, Bivariate score depth, Bivariate score density.
11	Yang et al. (*11)*	Characterizing differences of safety risk levels for different traffic movements	Traffic safety risk measured by traffic conflicts	Functional data smoothing, FANOVA
12	Yang et al. (*19)*	Analyzing time-dependent driver response behavior to Connected Vehicle (CV) warnings	Vehicle speed time series profile following speed compliance warning	Functional data smoothing, FPCA, Nonparametric functional linear regression
13	Shah et al. (*20)*	Forecasting day-ahead traffic flow	Traffic flow data collected at 15-minute intervals over the course of a day	Functional data smoothing, Functional autoregression,
14	Crawford, Watling and Connors (*17)*	Predicting traffic flow	Traffic flow	Functional data smoothing, Functional linear regression.

DATA

The pNEUMA data, an open dataset developed by Barmpounakis and Geroliminis (*12)* and available at https://open-traffic.epfl.ch, is used in this research for developing the functional conflict-based SPFs. Briefly, this dataset contains vehicle trajectories extracted from a swarm of 10 drones covering a 1.3 km^2 area with over 100 km-lanes of road network and around 100 busy intersections in central Athens, Greece. Many recent studies have explored this dataset from different perspectives, and it has been proven to be highly useful for investigating urban transportation problems.

Data pre-processing

In this research, vehicle trajectories at the signalized intersection of Alexandras Avenue and Mpoumpoulinas Street, is selected as a case study, as illustrated in **Figure 1**. Specifically, data collected between 8:30 AM and 9:30 AM on October 24, 2018 (one-hour) is processed. Geometrically, the Alexandras Avenue is separated by a median, with three lanes in either direction including one bus lane. The eastbound direction has been selected for subsequent analysis because our exploratory analysis of the westbound direction shows fewer stopping vehicles during the red interval, which poses challenges to extract traffic signal cycles (please see below for more detailed discussion). A 250 ft buffer is used to select vehicle trajectories belong to this signalized intersection, which has been commonly used in a past literature (*25*). As a result, a total number of 1114 vehicle trajectories are selected, and the trajectories are then smoothed by the moving average approach (with the window size equal to 12) to remove noise and outliers.

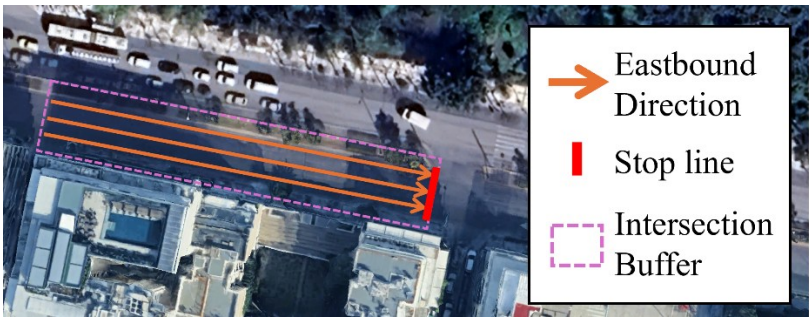

Figure 1 Study location.

Lane and signal cycle identification

To develop the proposed functional conflict-based SPFs, lane and signal timing information is needed. However, as discussed in (*26; 27*), pNEUMA does not provide signal and lane information but it is given that all the signalized intersections employ the pre-timed signal control strategy. Thus, with detailed vehicle trajectories, it is feasible to infer both the locations of lane markings and the signal cycles and phases as demonstrated in previous studies (*26; 27*).

For lane identification, this research adopts the assumption and algorithm discussed in Barmpounakis, Sauvin and Geroliminis (*26)* that vehicles tend to drive in the center of a lane and motorcycles should be excluded before lane identification due to their frequent travel near or on the lane markings. The lateral distances of vehicles from the median are visualized through the density of vehicle trajectory points. The midpoints between consecutive peaks are then identified as lane markings, as illustrated in **Figure 2**. Vehicle trajectories are then divided into the three lanes accordingly.

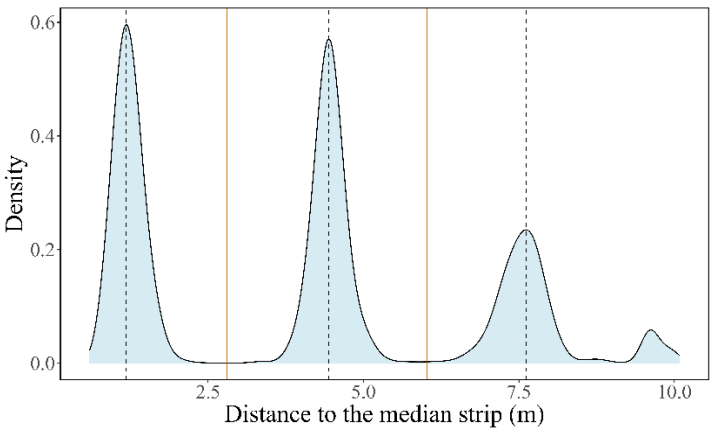

Figure 2 Density of the lateral distances of vehicles from the median with peak locations (black) and the identified lane markings (orange) highlighted.

For traffic signal extraction, this research follows the method discussed in a previous study (*27*) and applies the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering method to extract signal cycles directly from trajectory data. Similar to lane identification, motorcycles are also excluded before signal extraction due to their relatively erratic behavior that may jeopardize the signal extraction process, such as weaving through stopped traffic to jump to the front of a queue and only stopping for a red signal after already passing the stop line. Briefly, we first cluster the vehicle trajectories at the stop line to identify vehicles that have successfully passed the stop line during the green signal. Then, the time difference between the two first vehicles of each pair of consecutive clusters is identified as the signal cycle length. To avoid misidentification of the end of the green interval due to queue spill-overs that frequently occur in the chosen intersection, the maximum time difference between the first and last vehicles of each cluster is identified as the length of the green interval. As a result, a total of 20 complete signal cycles are identified. The average signal cycle length is 90.02 seconds while the maximum green interval is 51.92. Thus, 90 seconds and 52 seconds are used as the final signal cycle length and length of the green interval, respectively.

METHODS

In this section, we formally introduce the SSM and the shock wave parameters used in this research for developing the functional conflict-based SPF, followed by the FDA approach that includes the functional data smoothing and the functional linear regression techniques. In this research, data are not aggregated per signal cycle but are instead collected at a temporal resolution of one second. From a spatial perspective, lane-level data are obtained, which is the most detailed level in literature (*3*).

SSM and safety risk factors

Time to Collision

TTC, a very commonly used SSM proposed by Hayward (*28)*, is used in this research to identify traffic conflicts. It is defined as the time required for two vehicles to collide if they continue at their present speeds and on the same path. Mathematically:

$$\mathrm{TTC}=\frac{D_{1\text{-}2}}{v_2 - v_1}, \quad \forall v_2 > v_1 \tag{1}$$

, where D_{1-2} is the relative distance between the leading vehicle and the following vehicle; v_1 is the speed of the leading vehicle; and v_2 is the speed of the following vehicle. Surrogate Safety Assessment Model (SSAM), a commonly used tool to extract traffic conflicts from vehicle trajectories, is used to extract TTC and identify traffic conflicts accordingly (*29*). TTC threshold is set as 1.5 seconds, which has been widely

used in previous studies (*3; 7; 10*). The number of traffic conflicts per second is then obtained for subsequent modeling.

The number of vehicles and the number of moving vehicles

The number of vehicles per second is obtained by counting the number of vehicles for each lane within the intersection buffer area, while the number of moving vehicles per second is also gathered by counting the number of vehicles with speed larger than zero for each lane within the intersection buffer area. The reason for including the number of moving vehicles as one of the safety risk factors is that (*30*) has shown that the traffic safety risk may increase due to the turbulence of the moving vehicles.

Shock wave parameters

Shock wave parameters are also extracted in this research due to significant relationships identified between shock wave parameters and traffic conflicts in past literature (*3; 10*). Thus, this research extracted the following shock wave parameters, namely queue length and shock wave speeds, as safety risk factors in developing the proposed functional conflict-based SPF. Specifically, queue length is defined as the number of vehicles in the queue (*31*) and two shock wave speeds–the backward-forming shock wave speed S_1 and the backward-recovery shock wave speed S_2–are extracted based on the definitions in Daganzo (*32)* and are illustrated in **Figure 3**. By definitions, the extracted S_1 and S_2 are negative. For easier interpretation, this research uses the absolute values of S_1 and S_2 for subsequent modeling.

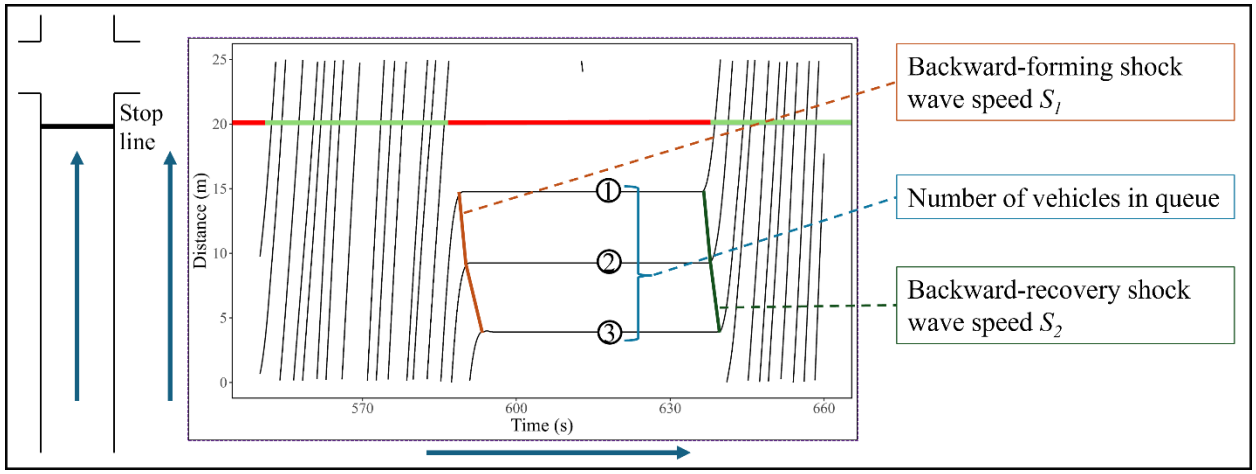

Figure 3 Demonstration of shock wave parameters.

Functional data analysis

The FDA approach constitutes a natural extension of the commonly used univariate and multivariate approaches in transportation safety research. Compared to these traditional approaches, FDA offers a distinct advantage by modeling the whole time series as a functional observation (*33*). Two major techniques in FDA are employed in our research: 1) the functional data smoothing, where each time series measure observation is modeled as a mathematical function with respect to time; and 2) the function linear regression that models the relationships between the number of traffic conflicts and its exposure and safety risk factors (all in functional forms).

Functional data smoothing

A function $w(t)$ with respect to time t can be constructed through a linear combination of a set of basis functions $\phi_k(t)$, $k=1,...,K$:

$$w(t)=\sum_{k=1}^{K}c_k\phi_k(t)=\mathbf{c}'\phi(t) \tag{2}$$

, where t is the time argument defined over $T=[T_{\min},T_{\max}]$ where $T_{\min}$ and $T_{\max}$ define the boundaries of the domain. The parameters $c_1,c_2,...,c_K$ are the coefficients of the expansion. In the matrix expression, the vector $\mathbf{c}$ denotes a K dimensional column vector containing the coefficients corresponding to each basis function and $\phi(t)$ denotes a K dimensional column vector containing all the basis functions at time t.

B-spline basis system

The B-spline basis system, a commonly used system for modeling nonperiodic data (*34*), is adopted in this research because no periodicity is observed in variables. The standard process for constructing B-spline basis functions is used in this study. Briefly, each basis function $\phi_k(t)$ in the B-spline basis system is a spline function that is generally constructed by firstly dividing the function domain into a number of subintervals separated by values called breakpoints and then specifying a polynomial over each subinterval. The breakpoints are specified to coincide with observed data points, i.e., at each second during the cycle, which is consistent with previous studies (*11; 35*). The order is set at 4, with subinterval having the same order, which adheres to the conventional practices (*11; 15*). One interior knot is put at each breakpoint to ensure that the function values and the first and second derivative values match between adjacent polynomial, while 4 knots (the same as the polynomial order) are assigned for the boundary points of the function domain (*34*). After determining the order and knots, K can be calculated as the sum of the order and the number of interior knots (*34*).

The roughness-penalized fitting criterion

The coefficients $c_1,c_2,...,c_K$ are estimated by minimizing the fitting criterion that is defined as:

$$\text{PENSSE}_{\lambda}(\mathbf{c})=\sum_{m=1}^{n}\left[y_m-\mathbf{c}'\phi(t_m)\right]^2+\lambda\int_{T_{\min}}^{T_{\max}}\left[D^2w(t)\right]^2dt \tag{3}$$

, where $\sum_{m=1}^{n}\left[y_m-\mathbf{c}'\phi(t_m)\right]^2$ is the traditional sum of squared errors (SSE). t_m is the time that the m^{th} value of the function $w(t)$ is observed. y_m is the observed function value at time t_m. n is the total number of observed function values, i.e., the total number of seconds in a signal cycle in this study. In the second term, λ is the hyperparameter, often called the smoothing parameter in FDA. $\int_{T_{\min}}^{T_{\max}}\left[D^2w(t)\right]^2dt$ is the penalty term that represents the roughness of the whole functional curves and is added to avoid overfitting to the data.

The optimal smoothing parameter λ is often obtained as the one that minimizes the generalized cross-validation (GCV) criterion using grid-search (*36*), which is denoted as:

$$\text{GCV}(\lambda)=\left(\frac{n}{n-df(\lambda)}\right)\left(\frac{\text{SSE}}{n-df(\lambda)}\right) \tag{4}$$

, where $df(\lambda)$ is the effective degree of freedom of the fit defined by λ and mathematically, $df(\lambda)=\text{trace}\left[\mathbf{\Phi}\left(\mathbf{\Phi}'\mathbf{\Phi}+\lambda\mathbf{R}\right)^{-1}\mathbf{\Phi}'\right]$ where $\mathbf{\Phi}$ is an n by K matrix contains the basis function values $\phi_k(t_j)$ at each observed time point. $\mathbf{R}=\int_{T_{\min}}^{T_{\max}}\phi(t)'\phi(t)dt$ is the order K symmetric roughness penalty matrix. Even though the optimal smoothing parameter λ is often chosen as the one that minimizes the

GCV, it is usually the case that GCV values change slowly near the minimum value, which indicates that the data are not particularly informative about the underlying true value of λ. Ramsay, Hooker and Graves (*34)* recommended using the researchers' own judgments and testing a range of values near the minimum value to enhance the utility of the modeling results, which is adopted in this research.

The positivity constraint

Since all the variables modeled in this research are non-negative, a positive constraint is imposed during the smoothing process using the exponential transportation (*34*). A positive smoothing function $x(t)$ can then be defined as:

$$x(t) = e^{w(t)} \tag{5}$$

By convention, the roughness of the positive smoothing function $x(t)$ is still defined as the roughness of its logarithm $w(t)$ (*37*). The roughness-penalized fitting criterion with the positivity constraint using the size of the second derivative is thus:

$$\text{PENSSE}_{\text{pos},\lambda}(\mathbf{c}) = \sum_{p=1}^{n}\left[y_p - \exp\left(\mathbf{c}'\phi(t_p)\right)\right]^2 + \lambda\int\left[D^2 w(t)\right]^2 dt \tag{6}$$

The optimal value of the smoothing parameter λ is obtained using the GCV method described above.

Functional linear regression

To model the relationship between the functional number of conflicts and the functional exposure and safety risk factors, we proposed to use the concurrent functional linear model, in which both the dependent variable $y(t)$ and the independent variables $x(t)$ are defined on the same function domain t and the value of the response variable $y(t)$ is predicted solely by the values of functional variables at the same time t. Then, the regression model is denoted as:

$$y_i(t) = \beta_0(t) + \sum_{j=1}^{q-1} x_{ij}(t)\beta_j(t) + \varepsilon_i(t) \tag{7}$$

, where $x_{ij}(t)$ denotes the j^{th} independent variable of the i^{th} observation. j ranges from 1 to $q-1$, where q is the number of variables in the model, including the intercept term, and i ranges from 1 to N, where N denotes the number of observations for each variable. $y_i(t)$ denotes the i^{th} observation of the dependent variable. $\beta_0(t)$ is the intercept function. $\beta_j(t)$ is the coefficient function for the corresponding $x_{ij}(t)$ functions. $\varepsilon_i(t)$ is the error term. Let the N by q functional matrix $\boldsymbol{Z}$ contain these x_{ij} functions, and let the vector coefficient function β of length q contain each of the regression functions. The concurrent functional linear model in matrix notation can be denoted as $y(t) = \boldsymbol{Z}(t)\beta(t) + \varepsilon(t)$, where y is a functional vector of length N containing the response functions, and ε is the corresponding residual functions.

To estimate the regression coefficients, the weighted regularized fitting criterion is used:

$$\text{LMSSE}(\beta) = \int \boldsymbol{r}(t)'\boldsymbol{r}(t)dt + \sum_{j}^{q}\lambda_j\int\left[L_j\beta_j(t)\right]^2 dt \tag{8}$$

where $\int \boldsymbol{r}(t)'\boldsymbol{r}(t)dt$ is the integral of the squared residuals, where $\boldsymbol{r}(t)$ is denoted as $\boldsymbol{r}(t) = \boldsymbol{y}(t) - \boldsymbol{Z}(t)\boldsymbol{\beta}(t)$. The second term $\sum_{j}^{q}\lambda_j\int\left[L_j\beta_j(t)\right]^2 dt$ is the regularization term that prevents overfitting. λ_j is a regularization parameter. L_j is a linear operator applied to $\beta_j(t)$. $\beta_j(t)$ are estimated

by minimizing equation (8). For more detailed explanation about the functional linear regression, please refer to (*34*). Bootstrapping method given in (*38*) is employed to obtain the confidence intervals for the estimated functional coefficients. Specifically, 1000 bootstrap samples are generated by sampling with replacement from all the observed functional subjects and 0.05 significance level is adopted to construct the confidence interval. The summary statistics of the variables used for developing the functional conflict-based SPFs are shown in **TABLE 2**.

TABLE 2 Summaries of Data Statistics

Variable	Description	Unit	Mean	SD	Min	Max
NV	The number of vehicles	-	10.30	4.02	0	26
NMV	The number of moving vehicles	-	4.51	2.70	0	15
Q	The number of vehicles in queue	-	5.80	5.07	0	21
S_1	The backward-forming shock wave speed	m/s	0.60	1.43	0	15.29
S_2	The backward-recovery shock wave speed	m/s	0.70	1.70	0	15.33
Conflict count	The number of conflicts	-	0.88	1.24	0	11

RESULTS and DISCUSSION

Determination of the smoothing parameter λ

As discussed in method section, the first step in functional data smoothing is to determine the smoothing parameter λ. To test a wide range of λ values, a grid search is conducted for each functional variable by specifying the logarithms of from -2 to 2 with 0.1 increments. The optimal λ for the backward-recovery shock wave speed, the number of vehicles, the number of moving vehicles, the backward-forming shock wave speed, the number of vehicles in queue, and the number of conflicts is -0.3, 0.5, 0.5, 0.5, 0.5, 1, respectively.

Smoothed functional curves

Using the optimal λ values, the smoothing functional curve for each variable is shown in **Figure 4** with the end of the green interval (at 52 seconds) labeled as a vertical dotted line. As can be seen from the figure, for the number of conflicts, there are noticeable conflicts around the beginning of green interval (0 s), which is consistent with a previous study (*3*). The functions approach zero during the middle of the green interval (i.e., around 25 seconds), which is probably because a stationary traffic state is observed at that time (i.e., vehicles traveling without interruptions). The number of conflicts reaches its peak near the beginning of the red interval (52 seconds), likely due to the deceleration of vehicles. Subsequently, the number of conflicts declines, which is probably because of the formation of the queue, i.e., an increasing number of vehicles become stationary. Overall, the smoothed curve of the number of conflicts exhibits patterns consistent with those observed in a previous study (*11*), where higher values are associated with signal changes.

Signal changes also affect the patterns of the number of vehicles, moving vehicles and vehicles in queue. As can be seen from the figure, the trends for the number of vehicles and vehicles in queue are similar, both decreasing after the start of the green interval. This decrease occurs as vehicles at the front of the queue clear the intersection, while those at the back remain stationary. After approximately 25 seconds, these values increase again, which is probably due to queue spill-overs. During the red interval, the values stabilize, which indicates a stationary queue within the intersection. Conversely, the number of moving vehicles shows an expected opposite trend, with some vehicles moving during the red interval, likely due to motorcycles bypassing the queue. Additionally, signal changes influence the backward-

recovery shock wave speed and the backward-forming shock wave speed. As shown in the figure, the backward-recovery shock wave speed increases immediately after the start of the green interval, indicating the vehicles at the front of the queue have started moving. The value decreases to zero approximately 25 seconds later, which implies that the queues are cleared. On the other hand, the backward-forming shock wave speed exhibits positive values after 25 seconds from the start of the green interval, which is likely due to the queue spill-overs. These findings, specifically the temporal variations of safety variables within signal cycles, have rarely been explored in past literature due to the aggregation of temporal data.

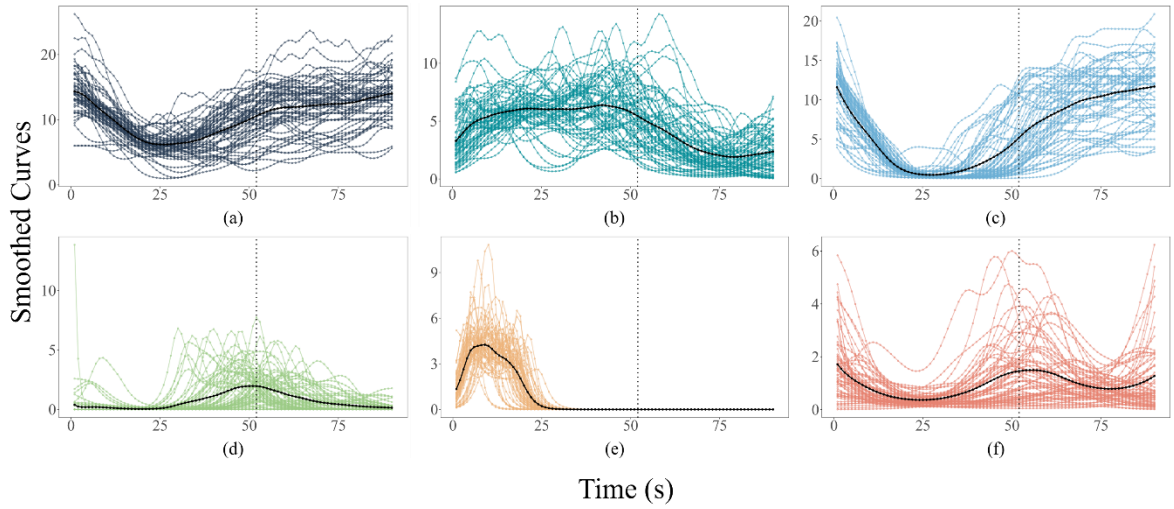

Figure 4 Smoothed functions for (a) the number of vehicles, (b) the number of moving vehicles, (c) the number of vehicles in queue, (d) backward-forming shock wave speed, (e) backward-recovery shock wave speed, (f) the number of conflicts.

Functional conflict-based SPFs

The functional conflict-based SPFs are developed by performing the functional linear regression approach discussed in the method section. All the safety risk factors discussed above are tested and only the ones that are statistically significant according to 0.05 significance level are retained. As a result, the final functional conflict-based SPF includes the number of moving vehicles and the backward-forming shock wave speed, as shown below.

$$y_i(t) = \alpha(t) + \beta_1(t)NMV_i(t) + \beta_2(t)S_{1i}(t) + \varepsilon_i(t) \quad (9)$$

The estimated functional regression coefficients corresponding to the number of moving vehicles and the backward-forming shock wave speed as well as their 95% confidence interval are illustrated in **Figure 5**.

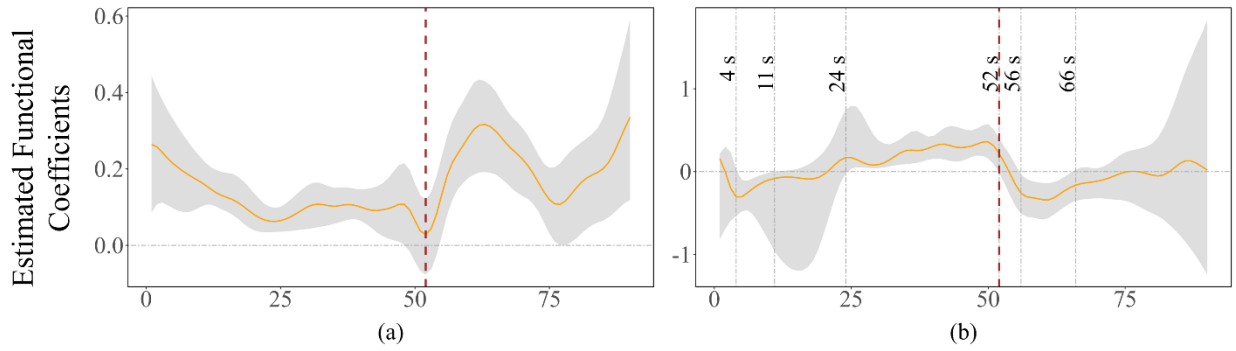

Figure 5 The estimated regression coefficient function and the corresponding 95% confidence interval for (a) the number of moving vehicles ($\hat{\beta}_1(t)$), and (b) the backward-forming shock wave speed ($\hat{\beta}_2(t)$), respectively (the start of the red interval is labeled with red dotted lines).

As can be seen from the figure, the $\hat{\beta}_1(t)$ is statistically significant according to the 0.05 significance level and positive for most time points. This indicates that a higher number of moving vehicles is associated with an increased number of traffic conflicts throughout the entire signal cycle. Notably, the estimated positive effects vary at different time points, which has not been observed in previous literature due to temporal data aggregation. Specifically, the largest positive effects between the number of moving vehicles and traffic conflicts occur at two periods: immediately after the red interval begins (i.e., approximately at 62 seconds) and immediately before the green interval starts (i.e., approximately at 90 seconds). This finding can be attributed to the fact that during the beginning of the red interval, vehicles decelerate and come to a full stop. As stopped vehicles cannot generate turbulence to the traffic flow, the number of moving vehicles can have large impacts on the traffic safety risk during this time. As for the time duration immediately before the green interval begins, as suggested in a previous study (*11*), some vehicles may begin accelerating prematurely, which results in elevated safety risk during this time.

Compared to $\hat{\beta}_1(t)$, $\hat{\beta}_2(t)$ exhibits both positive and negative associations with the number of conflicts within the entire signal cycles. Specifically, positive associations are found between the backward-forming shock wave speed and the number of traffic conflicts from 24 to 52 seconds, which corresponds to the second half of the green interval. This positive effect may be attributed to the spill-over effects in the queue downstream of the intersection. Vehicles approaching the intersection, despite seeing a green signal, may be unexpectedly required to stop due to the congestion spill-over, thereby resulting in an increase of safety risk. Conversely, from 56 to 60 seconds, i.e., at the beginning of the red interval, higher backward-forming shock wave speeds indicate that vehicles are coming to a full stop more rapidly. Given that drivers are aware of the signal being red during this time and are expected to brake accordingly, the faster the vehicles coming to a full stop, the less turbulence the vehicles will create, which results in lower safety risk.

In summary, the proposed functional conflict-based SPF for signalized intersections reveals new insights from a temporal perspective that are not apparent if data aggregation, a commonly used method in past literature, is employed. These findings can facilitate the development of more targeted proactive safety management strategies, such as controlling for safety risk factors that may increase safety risks at specific times during signal cycles by issuing warnings to drivers.

Comparison with aggregated data regression

As discussed in the Introduction section, previous studies often developed conflict-based SPFs with data aggregated at signal cycle level. Thus, to further demonstrate the contributions of the proposed functional conflict-based SPF, conflict-based SPF based on data aggregated at the signal cycle level is also developed as a comparison. Specifically, to be consistent with the developed functional conflict-based SPF above, the same safety risk factors, namely the number of moving vehicles and backward-forming shock wave speed, are included in the aggregated model as shown below.

$$\ln(Y) = \beta_0 + \beta_1 NMV + \beta_2 S_1 + \varepsilon \quad (10)$$

The estimated regression coefficients and the corresponding p-values are summarized in **TABLE 3**.

TABLE 3 Summaries of aggregated data regression

Variable	Estimate	Standard Error	P-value
Intercept	2.22	0.26	5.03×10^{-12} *

Number of moving vehicles	0.04	0.01	6.29×10^{-9} *
Backward-forming shock wave speed	0.14	0.04	5.69×10^{-4} *

*: Statistically significant at 0.05 significance level

As can be seen from the table, the estimated regression coefficients for both the number of moving vehicles and the backward-forming shock wave speed are statistically significant at the 0.05 level and positive. However, due to the data aggregation, conflict-based SPF using data aggregated at the signal cycle level fails to capture the temporal variations in the relationships between traffic conflicts and safety risk factors. This is especially true for the backward-forming shock wave speed, as the estimated functional coefficient is positive during some periods and negative during others as discussed above. Thus, aggregating data at the signal cycle level can significantly impair a comprehensive understanding of safety risks at signalized intersections.

CONCLUSIONS

This research proposes a new method to model traffic safety risk at signalized intersections by developing a functional conflict-based safety performance function (SPF). Compared to current literature where traffic conflicts and safety risk factors are often aggregated using some temporal aggregation levels, such as at the signal cycle level, this research proposes to model the traffic conflict and safety risk as functions with respect to time within signal cycles using the functional data analysis (FDA) approach in statistics. The use of the FDA approach allows for a more detailed examination of the temporal variations in safety risk within signal cycles. Specifically, the functional data smoothing technique is employed to convert observed time series data into continuous functional curves using the B-spline basis function system. A roughness-penalized fitting criterion is used to estimate the smoothed functional curves. The functional linear regression technique is then applied to model the relationships between the functional forms of the number of conflicts and its corresponding exposure and safety risk factors.

The pNEUMA dataset is used in this study and vehicle trajectories from the signalized intersection of Alexandras Avenue and Mpoumpoulinas Street between 8:30 AM and 9:30 AM on October 24, 2018, is analyzed. In addition to the traffic conflicts identified by the time-to-collision, safety risk factors extracted from this dataset include the number of vehicles, number of moving vehicles, queue length, backward-forming shock wave speed, and backward-recovery shock wave speed. The findings indicate that the number of moving vehicles and backward-forming shock wave speed are significant safety risk factors in modeling the number of traffic conflicts, with their effects varying across different time points within the signal cycle. Based on the proposed functional conflict-based SPF, detailed temporal effects are uncovered within signal cycles and time periods where safety risks are increased can be identified accordingly, which provides valuable insights for developing targeted proactive safety management strategies. Future research directions include exploring other surrogate safety measures, such as modified Time to Collision (MTTC), testing additional safety risk factors that may be relevant to modeling safety risk at signalized intersections, and investigating the impact of different levels of traffic conflict severity.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. 2401655. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: D. Yang; data collection: T. Shen; analysis and interpretation of results: T. Shen, D. Yang; draft manuscript preparation: T. Shen, D. Yang, K. Xie, H. Yang, X. Yang, M. Jeihani. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] Sharafeldin, M., A. Farid, and K. Ksaibati. Injury severity analysis of rear-end crashes at signalized intersections. *Sustainability,* Vol. 14, No. 21, 2022, p. 13858.

[2] Guo, F., X. Wang, and M. A. Abdel-Aty. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention,* Vol. 42, No. 1, 2010, pp. 84-92.

[3] Essa, M., and T. Sayed. Full Bayesian conflict-based models for real time safety evaluation of signalized intersections. *Accident Analysis & Prevention,* Vol. 129, 2019, pp. 367-381.

[4] Barhoumi, O., M. H. Zaki, and S. Tahar. A Formal Approach to Road Safety Assessment Using Traffic Conflict Techniques. *IEEE Open Journal of Vehicular Technology,* Vol. 5, 2024, pp. 606-619.

[5] Wang, K., S. Zhao, and E. Jackson. Investigating exposure measures and functional forms in urban and suburban intersection safety performance functions using generalized negative binomial - P model. *Accident Analysis & Prevention,* Vol. 148, 2020, p. 105838.

[6] Shekhar Babu, S., and P. Vedagiri. Proactive safety evaluation of a multilane unsignalized intersection using surrogate measures. *Transportation Letters,* Vol. 10, No. 2, 2018, pp. 104-112.

[7] El-Basyouny, K., and T. Sayed. Safety performance functions using traffic conflicts. *Safety Science,* Vol. 51, No. 1, 2013, pp. 160-164.

[8] Sacchi, E., and T. Sayed. Conflict-based safety performance functions for predicting traffic collisions by type. *Transportation research record,* Vol. 2583, No. 1, 2016, pp. 50-55.

[9] Zhang, X., P. Liu, Y. Chen, L. Bai, and W. Wang. Modeling the Frequency of Opposing Left-Turn Conflicts at Signalized Intersections Using Generalized Linear Regression Models. *Traffic Injury Prevention,* Vol. 15, No. 6, 2014, pp. 645-651.

[10] Essa, M., and T. Sayed. Traffic conflict models to evaluate the safety of signalized intersections at the cycle level. *Transportation Research Part C: Emerging Technologies,* Vol. 89, 2018, pp. 289-302.

[11] Yang, D., K. Ozbay, K. Xie, H. Yang, and F. Zuo. A functional approach for characterizing safety risk of signalized intersections at the movement level: An exploratory analysis. *Accident Analysis & Prevention,* Vol. 163, 2021, p. 106446.

[12] Barmpounakis, E., and N. Geroliminis. On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment. *Transportation Research Part C: Emerging Technologies,* Vol. 111, 2020, pp. 50-71.

[13] Chiou, J.-M. Dynamical functional prediction and classification, with application to traffic flow prediction. 2012.

[14] Chiou, J.-M., Y.-C. Zhang, W.-H. Chen, and C.-W. Chang. A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics,* Vol. 2, No. 2, 2014, pp. 106-129.

[15] Guardiola, I. G., T. Leon, and F. Mallor. A functional approach to monitor and recognize patterns of daily traffic profiles. *Transportation Research Part B: Methodological,* Vol. 65, 2014, pp. 119-136.

[16] Wagner-Muns, I. M., I. G. Guardiola, V. Samaranayke, and W. I. Kayani. A functional data analysis approach to traffic volume forecasting. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 19, No. 3, 2017, pp. 878-888.

[17] Crawford, F., D. Watling, and R. Connors. A statistical method for estimating predictable differences between daily traffic flow profiles. *Transportation Research Part B: Methodological,* Vol. 95, 2017, pp. 196-213.

[18] Yang, D., K. Ozbay, K. Xie, H. Yang, F. Zuo, and D. Sha. Proactive safety monitoring: A functional approach to detect safety-related anomalies using unmanned aerial vehicle video data. *Transportation Research Part C: Emerging Technologies,* Vol. 127, 2021, p. 103130.

[19] Yang, D., K. Ozbay, J. Gao, and F. Zuo. A Functional Approach for Analyzing Time-Dependent Driver Response Behavior to Real-World Connected Vehicle Warnings. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 24, No. 3, 2023, pp. 3438-3447.

[20] Shah, I., I. Muhammad, S. Ali, S. Ahmed, M. M. Almazah, and A. Al-Rezami. Forecasting day-ahead traffic flow using functional time series approach. *Mathematics,* Vol. 10, No. 22, 2022, p. 4279.

[21] Sudweeks, J. D. Using Functional Classification to Enhance Naturalistic Driving Data Crash/Near Crash Algorithms. 2015.
[22] Seya, H., T. Yoshida, and M. Tsutsumi. Ex-post identification of geographical extent of benefited area by a transportation project: Functional data analysis method. *Journal of Transport Geography,* Vol. 55, 2016, pp. 1-10.
[23] Zhong, R., J. Luo, H. Cai, A. Sumalee, F. Yuan, and A. H. Chow. Forecasting journey time distribution with consideration to abnormal traffic conditions. *Transportation Research Part C: Emerging Technologies,* Vol. 85, 2017, pp. 292-311.
[24] Hu, X., Y. Yuan, X. Zhu, H. Yang, and K. Xie. Behavioral responses to pre-planned road capacity reduction based on smartphone GPS trajectory data: A functional data analysis approach. *Journal of Intelligent Transportation Systems,* Vol. 23, No. 2, 2019, pp. 133-143.
[25] Kabir, R., S. M. Remias, S. M. Lavrenz, and J. Waddell. Assessing the impact of traffic signal performance on crash frequency for signalized intersections along urban arterials: A random parameter modeling approach. *Accident Analysis & Prevention,* Vol. 149, 2021, p. 105868.
[26] Barmpounakis, E., G. M. Sauvin, and N. Geroliminis. Lane detection and lane-changing identification with high-resolution data from a swarm of drones. *Transportation research record,* Vol. 2674, No. 7, 2020, pp. 1-15.
[27] Zhou, Q., R. Mohammadi, W. Zhao, K. Zhang, L. Zhang, Y. Wang, C. Roncoli, and S. Hu. Queue profile identification at signalized intersections with high-resolution data from drones.In *2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, IEEE, 2021. pp. 1-6.
[28] Hayward, J. C. Near miss determination through use of a scale of danger. 1972.
[29] Gettman, D., L. Pu, T. Sayed, S. G. Shelby, and S. Energy. Surrogate safety assessment model and validation.In, Turner-Fairbank Highway Research Center, 2008.
[30] Ma, W., Z. He, L. Wang, M. Abdel-Aty, and C. Yu. Active traffic management strategies for expressways based on crash risk prediction of moving vehicle groups. *Accident Analysis & Prevention,* Vol. 163, 2021, p. 106421.
[31] Reilly, W. Highway capacity manual 2000. *Tr News*, No. 193, 1997.
[32] Daganzo, C. F. *Fundamentals of transportation and traffic operations*. Emerald Group Publishing Limited, 1997.
[33] Chebana, F., S. Dabo-Niang, and T. B. Ouarda. Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research,* Vol. 48, No. 4, 2012.
[34] Ramsay, J., G. Hooker, and S. Graves. Functional data analysis with R and MATLAB. Science+ Business Media. *Inc., New York*, 2009.
[35] Crawford, F., D. P. Watling, and R. D. Connors. A statistical method for estimating predictable differences between daily traffic flow profiles. *Transportation Research Part B: Methodological,* Vol. 95, 2017, pp. 196-213.
[36] Craven, P., and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische mathematik,* Vol. 31, No. 4, 1978, pp. 377-403.
[37] Ramsay, J., and B. W. Silverman. *Functional Data Analysis.* Springer, New York, 2005.
[38] Şentürk, D., and D. V. Nguyen. Varying coefficient models for sparse noise-contaminated longitudinal data. *Statistica Sinica,* Vol. 21, No. 4, 2011, p. 1831.