# Exploring the Possibility of Outlier Detection Using Functional Data Analysis for Proactive Safety Management

Di Yang*, Kaan Ozbay, Kun Xie, Hong Yang, Fan Zuo, Di Sha

*Department of Civil and Urban Engineering, New York University, 15 MetroTech Center 6th Floor, Brooklyn, NY, 11201, USA, dy855@nyu.edu*
*Department of Civil and Urban Engineering, New York University, 15 MetroTech Center 6th Floor, Brooklyn, NY, 11201, USA, kaan.ozbay@nyu.edu*
*Department of Civil & Environmental Engineering, Old Dominion University (ODU), 129C Kaufman Hall, Norfolk, VA 23529, USA, kxie@odu.edu*
*Department of Computational Modeling and Simulation Engineering, Old Dominion University, 4700 Elkhorn Ave, Norfolk, VA 23529, USA,* hyang@odu.edu
*Department of Civil and Urban Engineering, New York University, 15 MetroTech Center 6th Floor, Brooklyn, NY, 11201, USA, fz380@nyu.edu*
*Department of Civil and Urban Engineering, New York University, 15 MetroTech Center 6th Floor, Brooklyn, NY, 11201, USA, ds5317@nyu.edu*

**Keywords**: Functional Data Analysis; Surrogate Safety Measures; Outlier Detection;

## 1. Introduction

In the era of smart cities, transportation data that are previously deemed as too difficult or even impossible to acquire can now be obtained through the use of various emerging technologies, such as cameras mounted on drones, traffic cameras, smart phones, computer vision and so on. For traffic safety analytics, the result of this technological development is the possibility of quantifying a large number of surrogate safety measures (SSMs). This in turn fuels significant advancements of safety analytic methods as well as proactive safety management approaches. In this study, we will focus on exploring the use of real-world SSM data extracted from videos recorded from drones at an urban intersection for proactive safety management.

When using the SSM data for intersection-level safety assessment, microsimulation software and the SSAM software were generally used (Stevanovic *et al.* 2013, Vasconcelos *et al.* 2014, Astarita *et al.* 2019). Only several recent studies have explored the use of SSMs extracted from drone-recorded or street-level videos, for example, to investigate pedestrian-vehicle conflicts (Chen *et al.* 2017) or vehicle-vehicle conflicts (Shekhar Babu and Vedagiri 2018). Several limitations can be identified from these studies: a) normal and abnormal safety conditions were not distinguished, which might include confounding factors in safety analytics; b) most studies investigated aggregated conflicts for the whole intersection without exploring the difference of conflicts for different lane or signal timing types.

Thus, this study aims to explore the possibility of outlier detection using real-world SSM data. We propose to use functional data analysis (FDA) approach that can reduce noise in the data to identify outliers by obtaining smoothed curves and setting certain criteria. Conflict count series at each signal cycles, more specifically each green period, will be analyzed instead of aggregating at the intersection level. This study can serve as the first step in developing and refining the proactive safety management strategies.

## 2. Methodology

### 2.1. Time to Collision

Time to collision (TTC) defined as the time required for two vehicles to collide if they continue at their present speeds and on the same path is used in this study to identify conflicts (Hayward 1972). A conflict is identified if its TTC value is less than 1.5 second and the number of conflicts is then obtained for each second during the green period. The number of conflicts is divided by the number of vehicles per second to obtain the normalized conflict counts, which will be used in the following outlier detection analysis.

### 2.2. Functional Data Smoothing

A function $W(t)$ is generally built by constructing a linear combination of a set of basis functions $\Phi_k$, $k=1,...,K$:

$$W(t)=\mathbf{c}^T\mathbf{\Phi} \tag{1}$$

where, $t$ is the time argument ($t=1,2,...,T$). The vector $\mathbf{c}$ of length $K$ contains the coefficients $c_k$ corresponding to each basis and the bold $\mathbf{\Phi}$ denotes a vector of length $K$ containing all the basis functions. Because the shape of the normalized conflict counts does not display any apparent periodicity (see Figure 2), the B-spline basis system is used in this study. B-splines are piecewise polynomials and they are typically defined by the range of validity, the knots, and the order. Typically, knots are specified to coincide with the observed data points, which ensures the consistency across all functions estimated (Ramsay and Silverman 2005).

Positivity constraint is added to the function defined in Equation (1) as follows since the normalized number of conflicts can only be nonnegative values. Specifically, a positive smoothing function $x(t)$ can be defined as:

$$x(t)=e^{W(t)} \tag{2}$$

The coefficients are estimated by minimizing the least squares criterion and a roughness penalty term is usually added to avoid overfitting. Thus, the estimate of the function is obtained by finding the coefficients that minimizes the following penalized sum of squares:

$$PENSSE_{\lambda}(\mathbf{c})=(\mathbf{y}-\mathbf{\Phi c})^T(\mathbf{y}-\mathbf{\Phi c})+\lambda\int\left[D^2x(s)\right]^2 ds \tag{3}$$

where, $\mathbf{y}$ represents the vector of observed values at each time $t$. $\lambda$ is the smoothing parameter. $\left[D^2x(t)\right]^2$ is the square of the second derivative of a function with respect to time. The smoothing parameter $\lambda$ specifies the amount of smoothing and is chosen by minimizing the generalized cross validation (GCV) proposed by (Craven and Wahba 1979).

After smoothing functions have been estimated, the functional sample mean and variance are calculated by computing the sample mean and variance at each time $t$. The boundary for identifying outliers is

subsequently defined as the mean functions plus three times the functional standard deviation at each time $t$. If a smoothed curve is higher than this boundary for at least one time point, then this smoothed curve will be identified as an outlier.

## 3. Results

A 24-minute video recorded by the DJI Spark drone at an intersection in Brooklyn, NY between 6 AM to 7 AM on September 17th, 2019 was analyzed. Anonymous vehicle trajectories were extracted by Data From Sky (Data From Sky 2020). The layout of this intersection is shown in Figure 1. Conflicts were extracted for the two throughput lanes highlighted in Figure 1 with protected throughput signal phasing. 10 signal cycles were manually identified from the video.

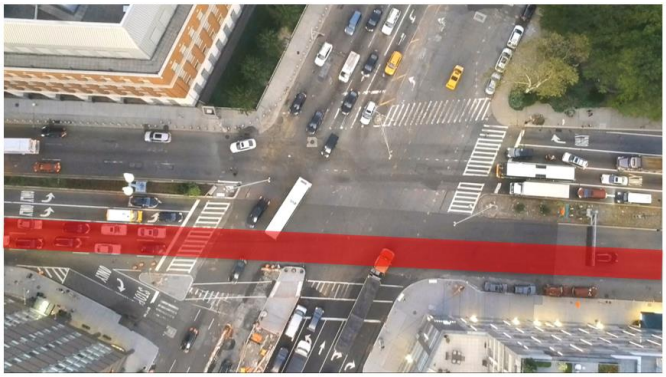

**Figure 1. The layout of the Flatbush & Tillary intersection (study approach highlighted)**

Conflict series of these 10 cycles are shown in Figure 2. The irregular shapes of these curves indicate the presence of noise in these series, which justifies the use of the FDA smoothing method. Two abnormal cycles (i.e., truck blocking the throughput approach and illegal vehicle right-turn resulting in blocking the throughput approach) were identified manually by carefully examining the video. These two cycles were excluded when estimating smoothing functions under the normal conditions.

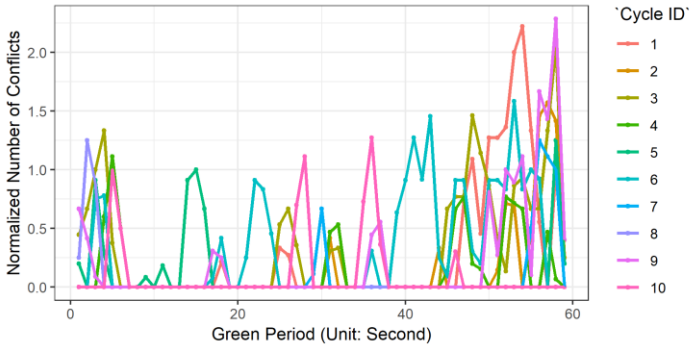


**Figure 2. Normalized number of conflicts trajectories the ten cycles**

By minimizing the GCV values, the optimal value of $\lambda$ is determined as 0.1. The estimated smoothing curve for each normal cycle (blue) along with the mean curve (green) are shown in Figure 3. Two peaks can be identified, which are at the start and the end of the green period. The peak at the start is generally caused by vehicles that accelerate before the leading vehicles in queues and have to decelerate to avoid

collision. The peak at the end is generally caused by the deceleration when the signal turns red. The constructed outlier detection boundary (purple) and the two outlier curves (orange) are also shown in Figure 3. No normal smoothed curves are falsely identified as outliers while only one of the two outliers is positively identified.

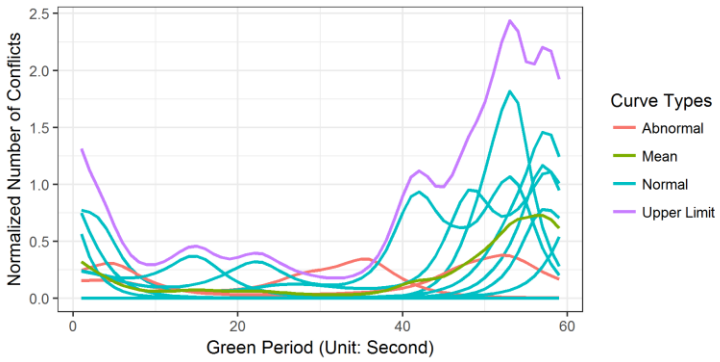


**Figure 3. Smoothing curves of normalized number of conflicts trajectories under normal condition and abnormal conditions along with the mean curve and the upper limit for outlier detection**

## 4. Conclusion

In this study, we take advantage of the SSM data extracted from drone-recorded videos and explore the possibility of using FDA for outlier detection. Two outlier cycles are identified out of the total ten cycles and our method can identify one of the outliers without falsely labeling normal cycles as outliers. Smoothed functions shown in Figure 3 can reduce noise, which can significantly reduce the false alarm rate, while noise in the raw data observations shown in Figure 2 will mask real outliers and hinder the identification process. This study can be extended by a) testing more cycles; b) constructing confusion matrix to formally evaluate the performance of the method; c) tuning the boundary threshold; d) discussing proactive safety management strategies at the cycle level, such as real time outliers detection and real time intervening.


## Acknowledgement

The authors would also like to thank Data From Sky (https://datafromsky.com/) for extracting vehicle trajectories from the drone-recorded videos for our analysis. The contents of this paper reflect views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents of the paper do not necessarily reflect the official views or policies of the funding agencies.


## References


Astarita, V., Festa, D.C., Giofrè, V.P., Guido, G., 2019. Surrogate safety measures from traffic simulation models a comparison of different models for intersection safety evaluation. Transportation research procedia 37, 219-226.

Chen, P., Zeng, W., Yu, G., Wang, Y., 2017. Surrogate safety analysis of pedestrian-vehicle conflict at intersections using unmanned aerial vehicle videos. Journal of advanced transportation 2017.

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Matematik 31, 377-403.
Data from Sky, 2020. Data from sky.
Hayward, J.C., 1972. Near miss determination through use of a scale of danger. Transportation Research Record: Journal of the Transportation Research Board 384, 24-34.
Ramsay, J., Silverman, B.W., 2005. Functional data analysis. Springer Series in Statistics, New York.
Shekhar Babu, S., Vedagiri, P., 2018. Proactive safety evaluation of a multilane unsignalized intersection using surrogate measures. Transportation letters 10 (2), 104-112.
Stevanovic, A., Stevanovic, J., Kergaye, C., 2013. Optimization of traffic signal timings based on surrogate measures of safety. Transportation research part C: emerging technologies 32, 159-178.
Vasconcelos, L., Neto, L., Seco, Á.M., Silva, A.B., 2014. Validation of the surrogate safety assessment model for assessment of intersection safety. Transportation Research Record 2432 (1), 1-9.