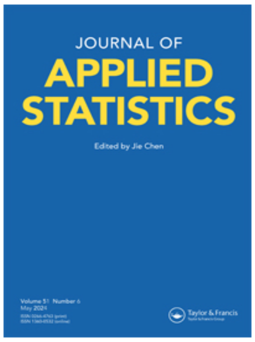

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cjas20

Smoothing regression and impact measures for accidents of traffic flows

Zhou Yu, Jie Yang & Hsin-Hsiung Huang

To cite this article: Zhou Yu, Jie Yang & Hsin-Hsiung Huang (2024) Smoothing regression and impact measures for accidents of traffic flows, Journal of Applied Statistics, 51:6, 1041-1056, DOI: 10.1080/02664763.2023.2175799

To link to this article: https://doi.org/10.1080/02664763.2023.2175799

JOURNAL OF APPLIED STATISTICS
2024, VOL. 51, NO. 6, 1041–1056
https://doi.org/10.1080/02664763.2023.2175799

Taylor & Francis
Taylor & Francis Group

Smoothing regression and impact measures for accidents of traffic flows

Zhou Yu[a], Jie Yang[a] and Hsin-Hsiung Huang[b]

[a]Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, IL, USA; [b]Department of Statistics and Data Science, University of Central Florida, Orlando, FL, USA

ABSTRACT
Traffic pattern identification and accident evaluation are essential for improving traffic planning, road safety, and traffic management. In this paper, we establish classification and regression models to characterize the relationship between traffic flows and different time points and identify different patterns of traffic flows by a negative binomial model with smoothing splines. It provides mean response curves and Bayesian credible bands for traffic flows, a single index, and the log-likelihood difference, for traffic flow pattern recognition. We further propose an impact measure for evaluating the influence of accidents on traffic flows based on the fitted negative binomial model. The proposed method has been successfully applied to real-world traffic flows, and it can be used for improving traffic management.

ARTICLE HISTORY
Received 28 July 2022
Accepted 27 January 2023

KEYWORDS
Negative binomial regression model; smoothing spline ANOVA model; traffic flow pattern; holiday traffic flow; impact factor of accidents

1. Introduction

In this paper, the traffic flow represents a sequence of numbers of vehicles passing by a given location during sequential time intervals. Analyzing traffic flows at different locations and time points can reveal the patterns and behaviors of the entire transportation system. Identifying regular traffic flows and anomalies not only provides the information about past and current traffic conditions but also helps with improving traffic controls and future route designs. It is of critical needs to develop statistical models with high accuracy and low computational cost to be applied practically and efficiently, as well as being incorporated with spatial and temporal characteristics of the traffic flows [9].

In early studies, traffic flow data have been mainly used for detecting sensor malfunction, solve issues of data collection in process of vehicle congestion, estimate velocity and forecast travel times on freeway networks [3,4]. However, one major challenge for modeling traffic flow data is that overdispersion frequently occurs, since vehicles in urban and signalized area have high fluctuating arriving and leaving rates during each traffic season [10,11]. It reduces accuracy and causes false conclusion on the significance of correlated variables if it is not appropriately handled. In this paper, we propose a negative binomial model (1)

CONTACT H.H. Huang ✉ hsin.huang@ucf.edu

Supplemental data for this article can be accessed online at https://doi.org/10.1080/02664763.2023.2175799.

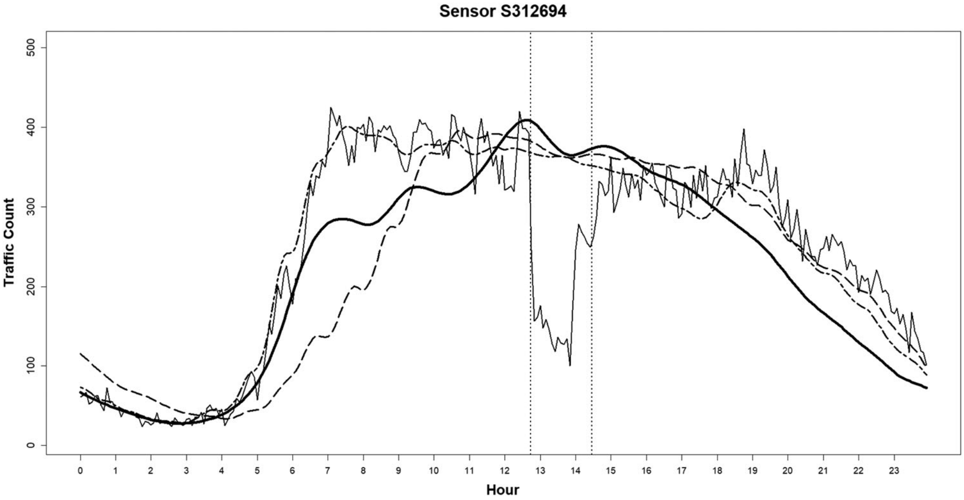

Figure 1. Visualization of Traffic Patterns and Impact of an Accident Recorded by Sensor S312694: Dotted lines represent the mean traffic flows of workday and weekend groups; bold smooth solid line represents the mean traffic flow on Martin Luther King Jr. Day; thin fluctuating solid line denotes the recorded traffic flow on 02/10/2017; two vertical dashed lines represent the start and end times of a reported accident.

with temporal patterns for analyzing traffic flow data, which can capture both the mean traffic flow and the influence of overdispersion well. More specifically, for $j = 1, \ldots, n$, at the time point x_j, the number of cars Y_j passing by a given location follows a negative binomial model with constant parameter $\nu > 0$, temporal parameter $p(x_i) \in (0, 1)$, and predictor function $\eta(x_i) = \log[p(x_i)/(1 - p(x_i))]$ with logit link

$$f(Y_j \mid x_j) = \exp\left\{-(\nu + Y_j)\log\left(1 + e^{\eta(x_i)}\right) + \nu\eta(x_i) + \log\left[\frac{\Gamma(\nu + Y_j)}{Y_j!\,\Gamma(\nu)}\right]\right\} \qquad (1)$$

For exponential family (including negative binomial distributions) smoothing splines, [16] obtains a lower-dimensional approximation of the estimates. For complex and massive data, [31] proposed a procedure that randomly selects a subset of basis functions to reduce the computational cost. Such an approximation approach was also adopted by [15] for generalized linear models. In this paper, we develop an algorithm by utilizing the lower-dimensional approximation to obtain a smoothing spline estimate for $\eta(x_i)$.

The second challenge that we need to address is different traffic patterns. Although some studies on cluster analysis have developed a test-based procedure that performs unsupervised clustering [30], our purpose is to model different traffic patterns due to workday, weekend, and holiday rather than clustering. Hence this unsupervised clustering is not the focus of this challenge. According to our analysis in Section 4.1, the traffic flow patterns of workday and weekend are significantly different and need to be modeled separately (see the two dash lines in Figure 1), while the traffic flows on most holidays follow a similar pattern as of weekends except the Martin Luther King Jr. Day (see the bold smooth solid line in Figure 1).

The third challenge that we will address is the impact of accidents to traffic flows. Accident leads to a critical problem for traffic control and transportation management since

Table 1. Sample records by sensor S312425: Column S312425 lists traffic flow counts; NA in column Holidayname means Not-a-holiday.

Date	S312425	Timestamp	Hour	Holidayname	Holiday	Weekend
2017-01-01	61	2017-01-01 23:50:00	23:50:00	NA	FALSE	TRUE
2017-01-01	67	2017-01-01 23:55:00	23:55:00	NA	FALSE	TRUE
2017-01-02	64	2017-01-02 00:00:00	00:00:00	New Year	TRUE	FALSE
2017-01-02	48	2017-01-02 00:05:00	00:05:00	New Year	TRUE	FALSE
2017-01-02	51	2017-01-02 00:10:00	00:10:00	New Year	TRUE	FALSE

it causes traffic congestion and affects road safety [24]. For example, a reported accident occurred on 10 February 2017 caused a significant reduction of observed traffic flow (see the thin fluctuating solid line in Figure 1) between 12:44 pm and 14: 28 pm. In the literature, the existing studies in this domain [29] tend to use an analytical approach and a weak proxy for traffic congestion [22] or measure purely the congestion time [25]. For example, linear relationships [24] and U-shaped relationships [28] have been used to characterize the associations between the flow of traffic and the levels of accidents, which may not be realistic or suitable in many applications since they do not fit nonlinear traffic flow data well without a large number of features [2].

The statistical models and inference used in this paper are different. For each group of daily traffic flows (such as workday or weekend), we fit a separate negative binomial regression model (1). The temporal effects on traffic flows are captured by smoothing splines built on a lower-dimensional space, called the effective model space. We construct Bayesian credible bands of the mean traffic flow curves based on the fitted smoothing spline negative binomial model for identifying anomalies and evaluating the impact of accidents.

2. Data

The data here contains 365 daily traffic flows in 2017 collected by 10 inductive loop road sensors of the California Department of Transportation [7]. The traffic performance measurement system (PeMS) currently functions as a statewide repository for traffic data gathered by thousands of automatic sensors [3,4]. In this paper, the traffic data can be downloaded through the Caltrans PeMS website (https://pems.dot.ca.gov/).

The traffic flow data were recorded over 5-min periods from time 00:00:00 to 23:55:00 by 10 sensors with ID labels S312425, S312520, S312694, S312942, S314147, S315017, S315938, S317814, S318180, S318566, respectively. It contains 105,120 records with 16 entries, including Date, Sensor ID, Time, Hour, Holiday Label and Weekend Label (see the sample records in Table 1 of sensor S312425).

Each sensor records the number of vehicles passing by or remaining over the sensor within each 5-min period [8]. Figure 2 displays the measurements of the traffic flow counts. The sensors record traffic flow data on workdays and weekends (see Table 1). Overall the traffic flows are significantly different for workdays and weekends (see Figure S.2 in Supplementary Material). More details can be found in Section 3.2.

The 10 sensors here are actually a subset of a larger collection of sensors that are distributed in and around District 3 of Sacramento, CA. These sensors are located generally on Interstate 80 (I-80) Highway without exact coordinates informed. For readers' reference, a

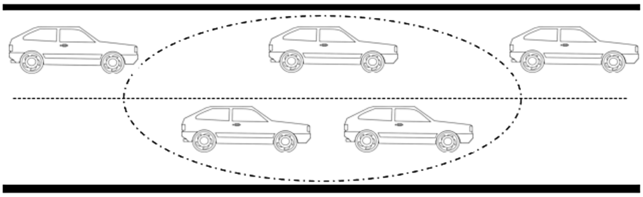

Figure 2. Traffic flow count: dotted circle denotes the 5-min neighborhood that vehicles remain or pass over the sensor.

map with some sensors along with their locations can be found in Figure S.1 in Section S.4 of Supplementary Material.

Due to various reasons, the data recorded and transferred by the sensors could be missing from time to time. For our data, there are 140 missing records in total on 3 days, 03/12/2017, 08/15/2017 and 10/15/2017.

For each of 08/15/2017 and 10/15/2017, only one record is missing at 22:45:00 or 11:55:00. We apply the *last-value-carried-forward* strategy [26] to impute the single missing data by duplicating the counts from their prior 5-min periods. For other missing data imputation strategies, please see, for example, Chapter 25 of [14]. For 03/12/2017 with more missing data involved, we apply the *linear interpolation plus round-off* strategy [12] to impute the missing data from 02:00:00 to 02:55:00. More details on our missing data imputation are relegated to Tables S.1, S.2, and S.3 in Section S.3 of Supplementary Material.

3. Methodology

3.1. Negative binomial smoothing spline ANOVA model

The smoothing spline analysis of variance (SS-ANOVA or SSANOVA [15]) has been used in applications that require a statistical technique to determine whether the shapes of multiple curves are significantly different from one another. It does not return a p-value to determine significance. Instead, it provides a Bayesian credible band for the mean response curve.

Typically, the generalized functional linear models [18,21] and functional ANOVA models [18,20] can be used to perform this. After applying the functional data analysis (FDA) to the traffic flow data, we find that the overall patterns of these two estimated mean response curves are almost the same. However, especially during certain time periods, the SSANOVA-based mean response curves are smoother than the FDA-based curves. More importantly, SSANOVA-based mean response curves are more robust than FDA-based curves for the sensors S314147, S315017 and S318180, which recorded numerous accidents (see Table 3 and Table S.9). Hence we prefer SSANOVA in this paper. More details on the FDA applications and comparisons can be found in Section S.9 of Supplementary Material.

The SSANOVA model that we use in this study is to model the 5-min-period traffic flow counts as negative binomial random variables with constant number ν_k of successes and time-dependent success rate $p(t,k)$ as a penalized smoothing spline for each specified group k.

More specifically, we denote Y_{tki} as the number of vehicles passing over a sensor at the tth 5-min period, where $t = 1, \dots, 288$ is the time index indicating $5t$ min after midnight, $k = 0$ for workday or 1 for weekend, $i = 1, \dots, n_k$ is the day index in group k. We assume that $Y_{tki} \sim \text{NB}(\nu_k, p(t,k))$, that is, a negative binomial distribution with parameters $\nu_k > 0$ and $p(t,k) \in (0,1)$. The probability that $Y_{tki} = y$ is

$$f(Y_{tki} = y; \nu_k, p(t,k)) = \frac{\Gamma(\nu_k + y)}{y!\Gamma(\nu_k)} p(t,k)^{\nu_k}[1 - p(t,k)]^y \tag{2}$$

where $y = 0, 1, 2, \dots$. Following [15], we take the logit link for $p(t,k)$. That is, the predictor $\eta(t,k) = \log \frac{p(t,k)}{1-p(t,k)}$, and the mean response $\mu(t,k) = E(Y_{tki}) = \frac{\nu_k(1-p(t,k))}{p(t,k)} = \nu_k e^{-\eta(t,k)}$. Model (1) can be obtained by writing $p(t,k)$ in terms of $\eta(t,k)$. Given each group index $k \in \{0,1\}$, the procedure for estimating the mean response curve $\mu(t,k)$, $t = 1, \dots, 288$ is described as Algorithm 1.

```
**Algorithm 1** Estimating the mean response curve $\mu(t,k)$
**Input:** Data $\{Y_{tki} \mid t = 1, \dots, 288;\ k = 0, 1;\ i = 1, \dots, n_k\}$.
**Output:** Estimated mean response curves $\hat{\mu}(t,k)$, $t = 1, \dots, 288$; $k = 0, 1$.
**Steps:** For each $k = 0, 1$, do
**1:** Rewrite the data $(Y_{tki})_{ti}$ from its matrix form to a long vector $(y_1, y_2, \dots, y_n)$, where $n = 288n_k$; denote the time point $x_j$ as $5(t-1)$ correspondingly, $j = 1, \dots, n$;
**2:** Calculate the maximum likelihood estimate $\nu_{mle}$ of $\nu$ when $p(t,k) \equiv p$, a constant;
**3:** Obtain $\lambda$ that minimizes the generalized approximate cross-validation score function (see equation (5.42) in [15] or Section S.1 in Supplementary Material);
**4:** Obtain $\eta_\lambda$ which minimizes the penalized likelihood functional (see (5.1) in [15])
        $$\frac{1}{n}\sum_{j=1}^{n}\left\{(y_j + \nu_{mle})\log\left(1 + e^{\eta(x_i)}\right) - \nu_{mle}\eta(x_i)\right\} + \frac{\lambda}{2}J(\eta) \tag{3}$$
**5:** Calculate $\hat{p}(t,k) = \frac{\exp\{\eta_\lambda(5(t-1))\}}{1+\exp\{\eta_\lambda(5(t-1))\}}$, $t = 1, \dots, 288$;
**6:** Calculate the mean response curve $\hat{\mu}(t,k) = \nu_{mle}\frac{1-\hat{p}(t,k)}{\hat{p}(t,k)}$, $t = 1, \dots, 288$.
```

3.2. Log-likelihood difference

By fitting the negative binomial SSANOVA model in Section 3.1 on workday traffic flows ($k = 0$) and weekend traffic flows ($k = 1$) separately, we obtain $\hat{\nu}_k$ and $\hat{p}(t,k)$, $k = 0, 1$ and $t = 1, \dots, 288$. Given any daily traffic flow data $\mathbf{Y} = \{y_t, t = 1, \dots, 288\}$ we propose a single index for pattern recognition of traffic flows, called the *log-likelihood difference* (LLD) with

$$\text{LLD}(\mathbf{Y}) = l(\hat{\nu}_0, \hat{p}(\cdot, 0) \mid \mathbf{Y}) - l(\hat{\nu}_1, \hat{p}(\cdot, 1) \mid \mathbf{Y})$$

where $l(\hat{\nu}_k, \hat{p}(\cdot,k) \mid \mathbf{Y}) = \sum_{t=1}^{288} \log f(Y_{tki} = y_t; \hat{\nu}_k, \hat{p}(t,k))$. For instance, if a workday data $\mathbf{Y}$ (e.g. Tuesday) is considered, it tends to produce a larger $l(\hat{\nu}_0, \hat{p}(\cdot,0) \mid \mathbf{Y})$ than $l(\hat{\nu}_1, \hat{p}(\cdot,1) \mid \mathbf{Y})$, and $\mathrm{LLD}(\mathbf{Y})$ tends to be positive. On the contrary, a weekend traffic flow (e.g. Sunday) is expected to produce a negative LLD. More technical details is relegated to Section S.2 in Supplementary Material.

3.3. Pattern recognition and validation

For any daily traffic flow data, the log-likelihood difference (LLD) proposed in Section 3.2 provides a single-number index for pattern recognition such as workday versus weekend. As an illustration, in this section we utilize Support Vector classifier (see, e.g. [17]) to show how our LLD can be used for identifying traffic flow patterns.

In this study, the traffic flows can be naturally grouped into two classes, the workday class and the weekend class. To separate these two classes based on the one-dimensional LLD, a boundary point, called a *threshold*, can be determined by the Support Vector Classifier (SVC) algorithm [23]. If the LLD is greater than the threshold, the traffic flow has a pattern closer to workdays. Conversely, an LLD smaller than the threshold indicates that the traffic flow more likely follows the weekend pattern.

If the threshold makes workdays and weekends well separable (as we will see in Section 4.1), it is a clear indication that the underlying statistical model (negative binomial SSANOVA) captures the traffic flow pattern accurately and efficiently. To validate the well separation, we apply the fivefold cross-validation to evaluate the prediction error on new data and prevent overfitting or selection bias [5,6]. Note that in the five-fold cross-validation, the whole dataset is randomly partitioned into five equal-sized subsets [13], and each subset will be used as the testing data, while the rest four subsets are used as the training data. The threshold obtained by the SVC algorithm for predicting the testing data is determined by the training data only.

3.4. Bayesian credible bands

By applying Algorithm 1, we obtain the negative binomial parameter estimates $\hat{\nu}_k$ and $\hat{p}(t,k)$ for each k and t. It is known that a point estimate alone is insufficient in practice because of lacking an assessment of the estimation precision and an adequately justified interval estimate is a rarity in nonparametric functional estimation [15]. We adopt the Bayesian credible intervals of [27] based on a Bayesian model (see also Section 2.5 in [15]). More specifically, we use the `R` package `gssanova` to obtain 95% Bayesian credible bands for our mean traffic flow $\nu(t,k)$. In Section 4.3, we will show how to use the Bayesian credible bands for evaluating accident impacts.

4. Applications

4.1. Workday and weekend traffic flow patterns

Given any date, it is clear whether it is a workday or a weekend, while it is not clear whether its traffic flow pattern is workday or weekend. For example, given that the workday is also a holiday, does its traffic flow follow workday's or weekend's pattern?

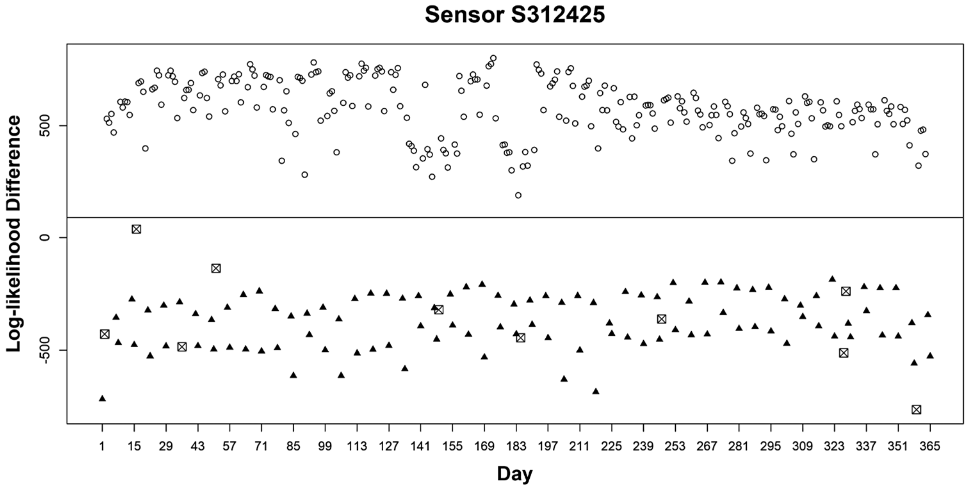

Figure 3. Log-likelihood difference (LLD) of daily traffic flow by sensor S312425 vs. day index in year 2017: Each point represents a traffic flow on a day in 2017; each hollow circle indicates a workday; each solid triangle indicates a weekend; each little square with cross indicates a holiday; the horizontal line represents the threshold 90.42 obtained by the SVC algorithm

Table 2. Thresholds obtained for individual sensors by SVC after removing holiday data.

Sensor ID	S312425	S312520	S312694	S312942	S314147
Threshold	90.42	164.63	−33.39	209.67	−45.53
Sensor ID	S315017	S315938	S317814	S318180	S318566
Threshold	61.28	67.01	62.17	87.62	103.10

In this section, we use sensor S312425 as an illustration. Figure 3 shows the visualization of the log-likelihood differences (LLD, see Section 3.2) against the day index of the year 2017. Each daily traffic flow data corresponds to a single point in Figure 3. Almost all weekend points (solid triangles) are at the bottom and almost all workday points (hollow circles) are at the top. As mentioned in Section 3.3, the good separation implies that LLD identifies the patterns of workday and weekend very well.

An interesting question is whether the traffic flows on holidays follow the pattern of workdays or weekends. Although 9 out of 10 holidays are workdays (see Table S.4 in Supplementary Material for the list of holidays under consideration), their traffic flows (see little squares in Figure 3) show quite similar patterns to weekends except for Martin Luther King Jr. Day (Monday, 16 January 2017), whose little square stays in the middle. We will revisit this special holiday in Section 4.2.

As mentioned in Section 3.3, we utilize the SVC algorithm, which is commonly used for linearly separable data [19], to obtain a threshold for separating workday and weekend patterns after removing all the 10 holidays' data. Note that the thresholds for different sensors could be different (see Table 2). For instance, the threshold for sensor S312425 is 90.42 (see also the horizontal line in Figure 3), which means if the LLD of a traffic flow is larger than 90.42, then its pattern is closer to workdays; otherwise, it is more like weekends.

To validate the threshold as a single-number indicator for classifying the patterns of daily traffic flows, we define an *error rate* (ER), the relative frequency of classification errors

Table 3. Error counts and rates of fivefold cross-validation with holidays treated as weekends based on negative binomial SSANOVA.

	Error						Total Error	
	Workday (no holidays)		Weekend (no holidays)		Holidays Only			
Sensor ID	Error Count	Error Rate (%)	Error Count	Error Rate (%)	Error Count	Error Rate (%)	Error Count	Error Rate (%)
S312425	0	0	0	0	0	0	0	0
S312520	5	1.99	0	0	1	10	6	1.64
S312694	0	0	0	0	1	10	1	0.27
S312942	0	0	0	0	0	0	0	0
S314147	0	0	0	0	1	10	1	0.27
S315017	1	0.40	0	0	1	10	2	0.55
S315938	0	0	1	0.96	0	0	1	0.27
S317814	1	0.40	0	0	1	10	2	0.55
S318180	0	0	0	0	1	10	1	0.27
S318566	0	0	0	0	2	20	2	0.55

occurred, as the ratio of the total number of misclassified daily traffic flows to the total number of daily traffic flows under consideration.

Since most holidays have a similar pattern as the weekends, we label $k = 1$ for all the 10 holidays for validation purpose. That is, class label $k = 0$ for 251 workdays excluding holidays, and $k = 1$ for 104 non-holiday weekends plus 10 holidays. In total, we have 365 days for year 2017.

The summarized error counts based on fivefold cross-validations (see Section 3.3 and Section S.5 of Supplementary Material) for all 10 sensors based on the thresholds determined by the SVC algorithm (see Table 2) are provided in Table 3. The error rates are quite small for workdays (seven zeros, two 0.4% and one 1.99%) and extremely small for weekends (1 error in total). Almost all errors associated with holidays are caused by Martin Luther King Jr. Day that have been mentioned previously. The overall ER is below 1% for 9 sensors and 1.64% for sensor S312520. Generally speaking, the SVC algorithm performs very well with LLD on these ten sensors, which further validates the effectiveness of the negative binomial SSANOVA model (1).

4.2. A special holiday: martin luther king jr. Day

In this section, we look into the a special holiday, Martin Luther King Jr. Day (Monday, 16 January 2017), which appears to be an outlier among the 10 holidays. According to our log-likelihood difference, this holiday neither likes a workday nor a weekend (see Figure 3).

Based on the negative binomial SSANOVA model (see Section 3.1), we obtain the mean response curve and the corresponding 95% Bayesian credible interval (see Section 3.4) for each class with $k = 0$ or $k = 1$ (see Figure 4). We relegate a justification on the robustness of the estimated mean response curve to Section S.6 of Supplement Material. The fitted mean response curve based solely on the traffic flow data on the Martin Luther King Jr. (MLK) Day is obtained and plotted in Figure 4 as well (more illustrations can be found in Figure S.4 in Section S.4 of Supplement Material).

Based on Figure 4 (see also Figure 1), the Martin Luther King Jr. Day's mean response curve (the solid line) is quite different from workdays and weekends. Especially between 3 am and 9 am, its mean response curve stays in the middle of workdays (higher curves) and

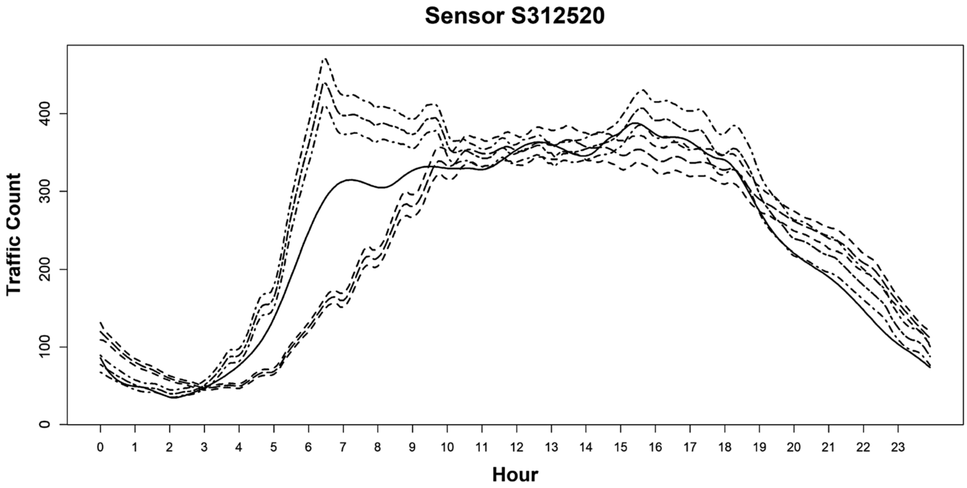

Figure 4. Mean response curves and credible bands with MLK day vs. hours of a day at sensor S312520: Dotted lines represent mean response curves and credible bands of the workday and weekend groups, respectively; solid line represents the mean response curve of the Martin Luther King Jr. Day's traffic flow

weekends (lower curves), which is significant since it is totally outside the credible bands of workday and weekend. It implies that the traffic flow on the Martin Luther King Jr. Day holiday stands by itself as a distinct cluster other than workdays and weekends.

One possible explanation is that public schools and government agencies consider the Martin Luther King Jr. Day a recognized holiday. However, most private schools and industries may not give their employees the day off. Therefore, some people may choose to rest, while some others still have to drive to schools or offices, which makes the Martin Luther King Jr. Day stand alone.

4.3. Measuring accident impact

Based on the negative binomial SSANOVA model described in Section 3.1, in this section we propose an impact factor for measuring or evaluating the impact of an accident on traffic flows, which can be potentially used for analyzing accident data [2] and identifying important factors for reducing accident impacts on traffic and transportation [29].

By applying Algorithm 1 in Section 3.1, we obtain from historical data the estimated mean curve of traffic flow $\hat{\mu}(x)$, representing the mean number of vehicles passing by a specified sensor at a time point x, along with a 95% credible band $(\hat{\mu}_L(x), \hat{\mu}_R(x))$, where $x \in [0, 60 \times 24)$ is the number of minutes after midnight.

Given an observed traffic flow $y(x)$ with a reported accident starting at time x_L and ending at time x_R, we define the *impact factor* of the accident as

$$I = \int_{x_L}^{x_R} \left| y(x) - \hat{\mu}(x) \right| \mathrm{d}x \tag{4}$$

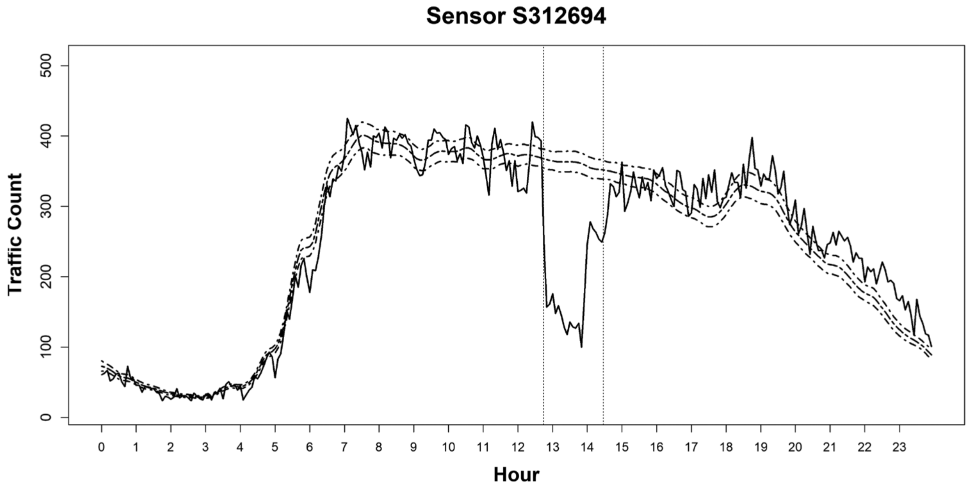

Figure 5. Visualization of an accident occurred on 02/10/2017 (Friday) from 12:44:00 to 14:28:00 by sensor S312694: dotted lines represent the mean response curve and its credible bands of the workday group; solid line stands for the real traffic flow on 02/10/2017; two vertical dashed lines represent the beginning and ending time of this reported accident.

and the *impact intensity* of the accident as

$$R = \frac{I}{x_R - x_L} = \frac{1}{x_R - x_L} \int_{x_L}^{x_R} \left| y(x) - \hat{\mu}(x) \right| \mathrm{d}x. \tag{5}$$

Intuitively speaking, the impact factor I represents the absolute cumulative changes of traffic flows compared with the mean traffic flows estimated from historical data, which is essentially the change of number of cars passing by the specific sensor over the accident period. The impact intensity R is the average change of number of cars per minute [1]. Since accidents typically reduce traffic flows significantly (see, e.g., Figure 5), our impact factor I could be roughly interpreted as *how many less cars passing by the sensor due to the accident* (up to a constant depending on the sampling frequency of the data).

Since $\hat{\mu}(x)$ is estimated from historical data, then both I and R are essentially estimated values. We can calculate 95% credible intervals based on $\hat{\mu}_L(x)$ and $\hat{\mu}_R(x)$ for the true I and R, respectively. That is, the left ends of credible intervals can be obtained by replacing $\hat{\mu}(x)$ in (4) and (5) with $\hat{\mu}_L(x)$, and the right ends of credible intervals uses $\hat{\mu}_R(x)$ instead of $\hat{\mu}(x)$.

In practice, we typically only observe traffic flow $y(x)$ at discrete time points $x = x_1, x_2, \ldots$, such as, $x = 0, 5, 10, \ldots$ minutes as in our dataset. When data points are sampled frequently such as in every 1, 5 or 10 min, we recommend a *step function* for $y(x)$ used in (4) and (5), that is, $y(x) = y(x_j),\ x \in [x_j, x_{j+1})$ for $j = 1, 2, \ldots$. In this case, suppose $x_l \leq x_L < x_{l+1}$ and $x_r \leq x_R < x_{r+1}$ for some indices $l < r$. Then $I = (x_{l+1} - x_L)|y(x_l) - \hat{\mu}(x_l)| + \mathbf{1}_{\{r-l>1\}} \cdot \sum_{j=l+1}^{r-1} (x_{j+1} - x_j)|y(x_j) - \hat{\mu}(x_j)| + (x_R - x_r)|y(x_r) - \hat{\mu}(x_r)|$.

For traffic data collected less frequently, such as every 20, 30, even 60 min, we recommend linear interpolation for generating a *piecewise linear function* $f(x)$, that is, $f(x) = \frac{x_{j+1}-x}{x_{j+1}-x_j} f(x_j) + \frac{x-x_j}{x_{j+1}-x_j} f(x_{j+1}),\ x \in [x_j, x_{j+1})$ for $j = 1, 2, \ldots$, where $f(x_j)$ here is defined as

Table 4. Impact factors, rates, and accident categories of reported accidents based on step function.

Sensor ID	Accident date	Accident duration (min)	Impact factor	Impact rate (%)	Accident category
S312425	09/05/2017	173	2010	0.60	Minor
S312520	09/26/2017	105	1008	0.26	Minor
S312694	02/10/2017	104	18736	5.22	Severe
S312942	03/22/2017	124	7579	2.02	Moderate
S314147	01/12/2017	212	16605	5.27	Severe
S314147	05/05/2017	165	6699	2.13	Moderate
S314147	07/18/2017	108	2636	0.84	Minor
S315017	02/06/2017	211	33089	7.77	Severe
S315017	04/05/2017	102	3234	0.76	Minor
S315938	05/30/2017	169	10957	2.98	Moderate
S315938	12/01/2017	122	6366	1.73	Moderate
S317814	07/26/2017	178	15861	6.86	Severe
S318180	09/05/2017	104	3051	1.28	Moderate
S318180	10/28/2017	122	4236	2.18	Moderate
S318566	06/12/2017	128	4487	1.03	Moderate

$f(x_j) = |y(x_j) - \hat{\mu}_{(}x_j)|$. In this case, we obtain that, $I = (x_{l+1} - x_L)[f(x_L) + f(x_{l+1})]/2 + \mathbf{1}_{\{r-l>1\}} \cdot \sum_{j=l+1}^{r-1}(x_{j+1} - x_j)[f(x_j) + f(x_{j+1})]/2 + (x_R - x_r)[f(x_R) + f(x_r)]/2$.

For the traffic data considered in this paper, the results based the step function $y(x)$ and the piecewise linear function $f(x)$ are almost identical (see Figure S.7 in Section S.7 of Supplementary Material). Therefore, in this study we use the step function $y(x)$ to calculate the impact factor I and the *impact rate* $\frac{I}{\int_0^{1440} \hat{\mu}(x)\mathrm{d}x}$, which is the relative impact of the accident with respect to the total number of cars passing by the sensor during the full day on average (see Table 4).

The accident categories listed in Table 4 are defined according to the impact rate or the relative impact of the accident. In this study, we call an accident *minor* if its impact rate $\leq 1\%$, *moderate* if $1\% <$ impact rate $\leq 5\%$, and *severe* if the impact rate $> 5\%$. For example, Sensor S315017 involves two reported accidents, occurred on 6 February and 5 April 2017, respectively. According to our analysis (see Table 4), the accident on April 5 is rated minor, while the accident on February 6 is rated severe.

To validate our impact measure, we plot the two accidents with Sensor S315017 in Figures 6 and 7, respectively. The accident on February 6 (see Figure 6) begins at 11:42:00 and ends at 15:13:00, lasting 211 minutes. Throughout this duration of the accident, the real traffic flow is far below the mean traffic flow and its credible bands (dotted line). The calculated impact rate is 7.77, which is a severe impact on the traffic flow. The corresponding impact factor 33,089 implies that so many vehicles were forced to choose other routes, which could significantly increase the traffic pressure and cause chaos in the neighborhood. The accident on April 5, however, has a different story. It lasted 102 min from 10:45:00 to 12:27:00 and did not cause serious traffic problems. According to Figure 7, the accident duration can be roughly divided into two parts. During the first 51 min (from 10:45:00 to 11:36:00), the real traffic flow (solid line) deviates significantly from the mean curve and its credible bands (dotted lines). Nevertheless, after that, the traffic flow soon recovers to a good state in the next 51 min (from 11:36:00 to 12:27:00), which falls within the credible bands. It is a clear indication that overall the accident on April 5 did not cause a big impact on the traffic flow. The corresponding impact rate is only 0.76 with an impact factor 3,234, about one-tenth of the accidents on February 6.

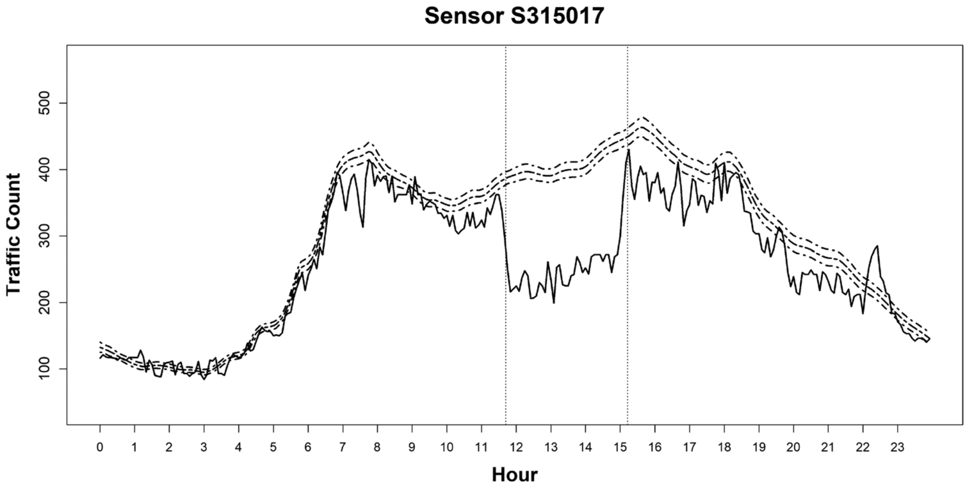

Figure 6. Traffic flow on 02/06/2017 recorded by sensor S315017: Dotted lines represent the mean response curve and 95% credible bands of workdays; solid line denotes the real traffic flow on 02/06/2017, Monday; two vertical dashed lines denote an accident occurred between 11:42:00 and 15:13:00.

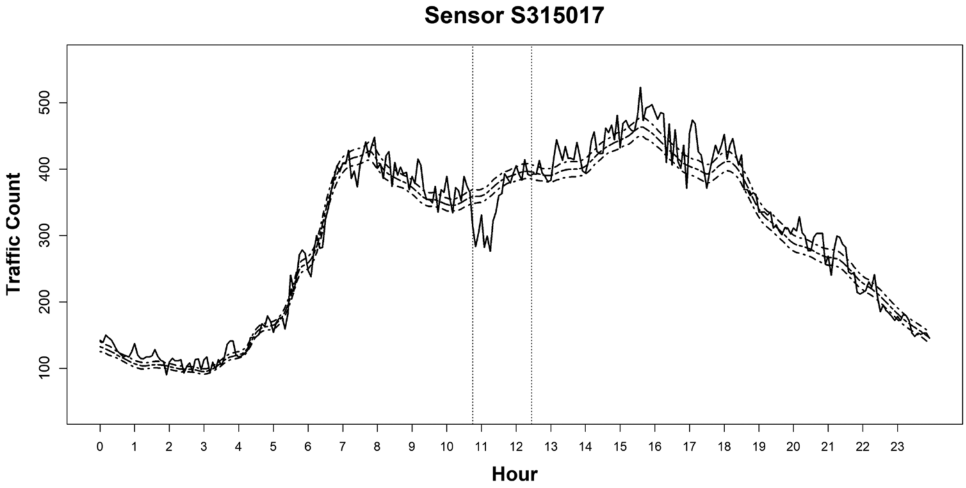

Figure 7. Traffic flow on 04/05/2017 recorded by sensor S315017: Dotted lines represent the mean response curve and 95% credible bands of workdays; solid line denotes the real traffic flow on 04/05/2017, Wednesday; two vertical dashed lines denote an accident occurred between 10:45:00 and 12:27:00.

Comparing Figure 6 and Figure 7, we conclude that impact factors and rates can measure the impact of an accident fairly well. The measures quantify the impact of an accident and provide an intuitive but precise index, the number or rate of affected cars. It can draw attentions from the traffic and transportation authorities and help them to discover good solutions such as in dealing with the accident on 04/05/2017 and reduce the impact of severe accident such as the one on 02/06/2017. With the precise measurement of the

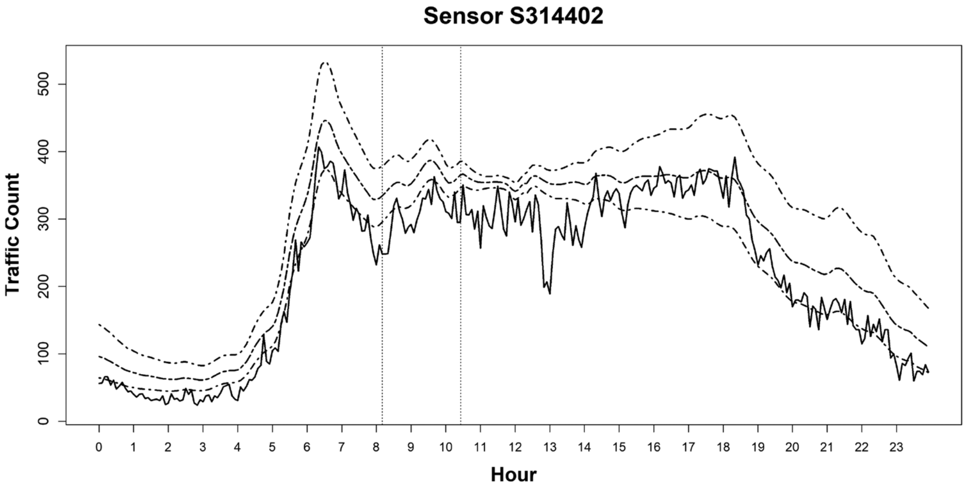

Figure 8. Traffic flow on 02/06/2017 recorded by sensor S314402: Dotted lines represent the mean response curve and 95% credible bands of workdays; solid line denotes the real traffic flow on 02/06/2017; two vertical dashed lines indicate an accident occurred between 08:10:00 and 10:26:00.

accident impact, more statistical models and further data analysis could be applied to predict possible impact at the beginning of an accident and recommend the best strategy for recovering the traffic flow and reducing the accident impact.

4.4. Investigation of accidents on 6 February 2017

From Table 4, we note that the accident with the highest impact rate (7.77) occurred on February 6, which deserves further investigations to explore possible causes of this severe accident. Through the inquiry, we find out that there was a basketball game of the National Basketball Association (NBA), Chicago Bulls versus Sacramento Kings in Sacramento on that day. To analyze this accident with the help of actual geographical information, we calculate the impact rate by sensor S314402 whose latitude and longitude information is available (while sensor S315017 is not). Figure 8 shows the visual effect of the accident occurred on February 6 around this sensor.

The impact rate of the accident occurred between 08:10:00 and 10:26:00 at sensor S314402 is 1.93, which indicates that it is a moderate accident. As we can see in Figure 8, it seems that the impact of the accident on traffic flow lasted much longer than the reported period from 08:10:00 to 10:26:00. The location of sensor S314402 is near Exit 14B of I-80 highway and close to the Sacramento McClellan Airport. Therefore, a reasonable explanation is that lots of the NBA fans from different states and even different countries, especially the Chicago Bulls, came to Sacramento to watch the game. The arrivals of a large number of fans had increased the pressure on highway traffic near the airport, which had also led to this accident, making the traffic flows on this day unusual. We investigate the impact of the accident on this date, by calculating the impact rates of multiple sensors that recorded this accident. Table 5 shows the results for multiple accidents reported on 6 February 2017.

Table 5. Impact measures for accidents occurred on 6 February 2017, recorded by six sensors.

Impact measures for accidents on 06-Feb-2017				
Sensor ID	Accident duration (min)	Impact factor	Impact rate (%)	Accident category
S312566	115	10165	3.47	Moderate
S313393	48	3484	0.71	Minor
S313405	43	3702	0.80	Minor
S314402	136	7341	1.93	Moderate
S315017	211	33089	7.77	Severe
S318566	203	12142	2.79	Moderate

All the six sensors in Table 5 are distributed near the Sacramento McClellan Airport. According to our measurements, the six reported accidents are rated as minor (2), moderate (3), or severe (1) (see Table 5). These are good examples of the important role that the proposed impact factors and rates for assessing the impact of accidents on traffic flow play in analyzing accidental data and identifying important factors for reducing the impact of accidents on traffic and transportation.

5. Conclusion

For the daily traffic flow data, we develop an effective negative binomial smoothing spline ANOVA model whose success rate is a function of time. One major task in this study is the pattern recognition of daily traffic flows. The most critical challenge in this pattern recognition task is how to properly address the traffic patterns of holidays. Although 9 out of the 10 federal holidays in 2017 fall in the range of workdays, including Monday (6), Tuesday (1), Thursday (1) and Friday (1), the traffic flows during holidays are significantly different from the usual workdays, which is reasonable since many people do not have to work or go to school during holidays.

It is natural to group the traffic flow patterns into workdays and weekends. Based on the proposed negative binomial smoothing spline ANOVA model, we propose the log-likelihood difference (LLD) as a single index for identifying the two patterns. The fivefold cross-validation described in Section 4.1 shows that the LLD index with the SVC works very well in identifying the traffic patterns. According to the threshold determined by the SVC algorithm, the traffic flow patterns of most holidays are classified as weekends except for the Martin Luther King Jr. Day.

As a conclusion in Section 4.2, the traffic flow pattern of the Martin Luther King Jr. Day is a mixture of workdays and weekends due to a significant portion of the population still need to go to work or school. We recommend that the Martin Luther King Jr. Day stands along as a distinct cluster in terms of traffic flow patterns.

Based on the proposed negative binomial smoothing spline ANOVA model, we could go further on identifying distinct patterns of traffic flows. For example, according to our further analysis (see Section S.8), the traffic flow patterns on Saturday and Sunday are actually significantly different (see Figures S.8 and S.9 in Supplementary Material). Similarly, the workday group can be further divided into subgroups as well, which will provide more precise descriptions for traffic flow patterns. With $K \geq 3$ possible classes of traffic flows, instead of LLD, we would recommend a K-tuple $(l(\hat{\nu}_1, \hat{p}(\cdot, 1) \mid \mathbf{Y}), \ldots, l(\hat{\nu}_K, \hat{p}(\cdot, K) \mid \mathbf{Y}))$

for traffic flow classification or pattern recognition. A different classifier other than SVC may be applied with the K-tuple.

Another major application based on the proposed statistical model is to properly evaluate the impact of an accident on traffic flows. The thresholds that we recommended in Section 4.3 are 1% and 5% for using impact rates to separate minor, moderate and severe accidents. Depending on different circumstances, the users may define their own thresholds based on the impact rates and name their own class labels.

Disclosure statement

The authors declare no conflict of interest. All authors reviewed the results and approved the final version of the manuscript.

Funding

This research was supported in part by the National Science Foundation grants, DMS-1924792 and DMS-1924859.

ORCID

Hsin-Hsiung Huang http://orcid.org/0000-0001-7150-7229

References

[1] F. Amato, M. Pandolfi, A. Alastuey, A. Lozano, J. Contreras González, and X. Querol, *Impact of traffic intensity and pavement aggregate size on road dust particles loading*, Atmos. Environ. 77 (2013), pp. 711–717.

[2] S. Ardekani, E. Hauer, and B. Jamei, Traffic impact models. *Chapter 7 in Traffic Flow Theory, Oak Bridge National Laboratory Report*, 1992.

[3] P. Bickel, C. Chen, J. Kwon, J. Rice, P. Varaiya, and E. van Zwet, Traffic flow on a freeway network. In *Nonlinear Estimation and Classification*, Springer, 2003, pp. 63–81.

[4] P.J. Bickel, C. Chen, J. Kwon, J. Rice, E. Van Zwet, and P. Varaiya, Measuring traffic. *Statistical Science*, 2007, pp. 581–597

[5] M.W. Browne, *Cross-validation methods*, J. Math. Psychol. 44 (2000), pp. 108–132.

[6] G.C. Cawley and N.L.C. Talbot, *On over-fitting in model selection and subsequent selection bias in performance evaluation*, J. Mach. Learn. Res. 11 (2010), pp. 2079–2107.

[7] C. Chen, K. Petty, A. Skabardonis, P. Varaiya, and Z. Jia, *Freeway performance measurement system: mining loop detector data*, Transp. Res. Rec. 1748 (2001), pp. 96–102.

[8] S.Y. Cheung, S. Coleri, B. Dundar, S. Ganesh, C.-W. Tan, and P. Varaiya, *Traffic measurement and vehicle classification with single magnetic sensor*, Transp. Res. Rec. 1917 (2005), pp. 173–181.

[9] C. Daganzo and C.F. Daganzo, *Fundamentals of transportation and traffic operations*, Vol. 30, Pergamon Oxford, 1997.

[10] Y.-A. Daraghmi, C.-W. Yi, and T.-C. Chiang, Space-time multivariate negative binomial regression for urban short-term traffic volume prediction. In *12th International Conference on ITS Telecommunications*, IEEE, 2012, pp. 35–40.

[11] Y.-A. Daraghmi, C.-W. Yi, and T.-C. Chiang, Mining overdispersed and autocorrelated vehicular traffic volume. In *5th International Conference on Computer Science and Information Technology*, IEEE, 2013, pp. 194–200.

[12] I. Fried, *Discretization and round-off errors in the finite element analysis of elliptic boundary value problems and eigenvalue problems.* Ph.D. diss., Massachusetts Institute of Technology, 1971

[13] T. Fushiki, *Estimation of prediction error by using k-fold cross-validation*, Stat. Comput. 21 (2011), pp. 137–146.

[14] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Analytical Methods for Social Research. Cambridge University Press, 2006.

[15] C. Gu, *Smoothing Spline ANOVA Models*, Springer, 2013.

[16] C. Gu and P. Ma, *Generalized nonparametric mixed-effect models: computation and smoothing parameter selection*, J. Comput. Graph. Stat. 14 (2005), pp. 485–504.

[17] T. Hastie, R. Tibshirani, J.H. Friedman, and J.H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer, 2009.

[18] J.Z. Huang, *Functional anova models for generalized regression*, J. Multivar. Anal. 67 (1998), pp. 49–71.

[19] T. Joachims, Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 217–226.

[20] P.A. Morettin, A. Pinheiro, and B. Vidakovic, Functional anova. In *Wavelets in Functional Data Analysis*, Springer, 2017, pp. 71–88.

[21] H.-G. Müller and U. Stadtmüller, *Generalized functional linear models*, Ann. Stat. 33 (2005), pp. 774–805.

[22] R.B. Noland and M.A. Quddus, *Congestion and safety: A spatial analysis of London*, Trans. Res. Part A Policy Pract. 39 (2005), pp. 737–754.

[23] A. Pradhan, *Support vector machine-a survey*, Int. J. Emerg. Technol. Adv. Eng. 2 (2012), pp. 82–85.

[24] A.E. Retallack and B. Ostendorf, *Current understanding of the effects of congestion on traffic accidents*, Int. J. Environ. Res. Public. Health. 16 (2019), pp. 3400.

[25] M.A.P. Taylor, J.E. Woolley, and R. Zito, *Integration of the global positioning system and geographical information systems for traffic congestion studies*, Trans. Res. Part C Emerg. Technol. 8 (2000), pp. 257–285.

[26] J. Twisk and W. de Vente, *Attrition in longitudinal studies: how to deal with missing data*, J. Clin. Epidemiol. 55 (2002), pp. 329–337.

[27] G. Wahba, *Bayesian "confidence intervals" for the cross-validated smoothing spline*, J. R. Stat. Soc. Ser B (Methodological) 45 (1983), pp. 133–150.

[28] C. Wang, M. Quddus, and S. Ison, *A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the uk*, Transportmetrica A Trans. Sci. 9 (2013), pp. 124–148.

[29] C. Wang, M.A. Quddus, and S.G. Ison, *Impact of traffic congestion on road accidents: A spatial analysis of the m25 motorway in England*, Accid. Anal. Prev. 41 (2009), pp. 798–808.

[30] A.Z. Zambom, J.A. Collazos, and R. Dias, *Functional data clustering via hypothesis testing k-means*, Comput. Stat. 34 (2019), pp. 527–549.

[31] J. Zhang, H. Jin, Y. Wang, X. Sun, P. Ma, and W. Zhong, *Smoothing spline Anova models and their applications in complex and massive datasets*, in *Topics in Splines and Applications*, Y.K.N. Truong and M. Sarfraz, eds., IntechOpen, London, UK, 2018, pp. 63–82.

Smoothing Regression and Impact Measures for Accidents of Traffic Flows

Zhou Yu[a], Jie Yang[a], and Hsin-Hsiung Huang[b]

[a]University of Illinois at Chicago, Chicago, Illinois, USA and
[b]University of Central Florida, Orlando, Florida, USA

Supplementary Materials

S.1. Technical details for Algorithm 1

In this section, we provide more technical details on obtaining λ in Step 3 of Algorithm 1:

3: Obtain λ that minimizes the generalized approximate cross-validation score function (see equation (5.42) in [15]):

$$\begin{aligned} V(\lambda) &= -\frac{1}{n}\sum_{j=1}^{n}\{(\nu_{mle}+y_j)\log(1-p_\lambda(x_j))+\nu_{mle}\eta_\lambda(x_j)\} \\ &\quad + \alpha\frac{\mathrm{tr}(A_w W^{-1})}{n-\mathrm{tr}A_w}\frac{1}{n}\sum_{j=1}^{n} y_j p_\lambda(x_j)\{(\nu_{mle}+y_j)p_\lambda(x_j)-\nu_{mle}\} \end{aligned} \tag{S.1}$$

where η_λ minimizes the penalized likelihood functional (3)

$$\frac{1}{n}\sum_{j=1}^{n}\left\{(y_j+\nu_{mle})\log\left(1+e^{\eta(x_i)}\right)-\nu_{mle}\eta(x_i)\right\}+\frac{\lambda}{2}J(\eta)$$

(see expression (5.1) in [15]), and $p_\lambda(x_j)$ are produced via the equation:

$$p_\lambda(x_j)=\frac{\exp\{\eta_\lambda(x_j)\}}{1+\exp\{\eta_\lambda(x_j)\}}$$

Here more notations in (S.1) need to be clarified. The α is known as the fudge factor. The value of α can be 1 or 1.4, and $\alpha = 1.4$ is generally preferred to $\alpha = 1$ [15]. In this study, $\alpha = 1.4$ is used to overcome the undersmoothing issue of GCV (generalized cross-validation) while maintaining its good performance [15]. The W is a diagonal matrix with elements $\tilde{w}_j$, $j = 1, \ldots, n$, that is $W = \mathrm{diag}\{\tilde{w}_1, \ldots, \tilde{w}_n\}$, where

$$\tilde{w}_j=\frac{\nu_{mle}\exp\{\eta_\lambda(x_j)\}}{[1-\exp\{\eta_\lambda(x_j)\}]^2}$$

The A_w is an $n \times n$ matrix defined as

$$A_w = I - n\lambda F_2(F_2^T Q_w F_2 + n\lambda I)^{-1}F_2^T$$

where F_2 is an orthogonal matrix with $F_2^T F_2 = I$ (see expression (3.5) in [15]), $Q_w = W^{1/2}QW^{1/2}$, and Q is a square matrix (see expression (2.16) in [15]). After all, λ is obtained as Step 3.

S.2. Technical details for Section 3.2

In this section, we provide more technical details on log-likelihood difference (LLD, see Section 3.2).

Given the class label $k \in \{0,1\}$ and a daily traffic flow data $\mathbf{Y} = \{y_t, t = 1, \ldots, 288\}$, the likelihood function of the negative binomial model is

$$L(\nu_k, p(\cdot,k) \mid \mathbf{Y}) = \prod_{t=1}^{288} \frac{\Gamma(\nu_k + y_t)}{y_t!\Gamma(\nu_k)} p(t,k)^{\nu_k} [1 - p(t,k)]^{y_t}$$

In terms of $\eta(t,k) = \log \frac{p(t,k)}{1-p(t,k)}$, the log-likelihood function

$$\begin{aligned}
l(\nu_k, p(\cdot,k) \mid \mathbf{Y}) &= l(\nu_k, \eta(\cdot,k) \mid \mathbf{Y}) \\
&= \sum_{t=1}^{288} \left\{ \log \frac{\Gamma(\nu_k + y_t)}{y_t!\Gamma(\nu_k)} + \nu_k \eta(t,k) + \nu_k \log[1 - p(t,k)] + y_t \log[1 - p(t,k)] \right\} \\
&= \sum_{t=1}^{288} \left\{ \log \frac{\Gamma(\nu_k + y_t)}{y_t!\Gamma(\nu_k)} + \nu_k \eta(t,k) + (\nu_k + y_t) \log[1 - p(t,k)] \right\} \\
&= \sum_{t=1}^{288} \left\{ \log \frac{\Gamma(\nu_k + y_t)}{y_t!\Gamma(\nu_k)} + \nu_k \eta(t,k) - (\nu_k + y_t) \log[1 + e^{\eta(t,k)}] \right\} \\
&= \sum_{t=1}^{288} \log[\Gamma(\nu_k + y_t)] - 288 \log[\Gamma(\nu_k)] - \sum_{t=1}^{288} \log(y_t!) + \nu_k \sum_{t=1}^{288} \eta(t,k) \\
&\quad - \sum_{t=1}^{288} (\nu_k + y_t) \log[1 + e^{\eta(t,k)}]
\end{aligned}$$

Then the difference of log-likelihood between the workday class ($k = 0$) and the weekend class ($k = 1$) is

$$\begin{aligned}
\mathrm{LLD}(\mathbf{Y}) &= l(\hat{\nu}_0, \hat{p}(\cdot,0) \mid \mathbf{Y}) - l(\hat{\nu}_1, \hat{p}(\cdot,1) \mid \mathbf{Y}) \\
&= \sum_{t=1}^{288} \log \frac{\Gamma(\hat{\nu}_0 + y_t)}{\Gamma(\hat{\nu}_1 + y_t)} - 288 \log \frac{\Gamma(\hat{\nu}_0)}{\Gamma(\hat{\nu}_1)} + \sum_{t=1}^{288} [\hat{\nu}_0 \hat{\eta}(t,0) - \hat{\nu}_1 \hat{\eta}(t,1)] \\
&\quad - \sum_{t=1}^{288} \left\{ \log \frac{[1 + e^{\hat{\eta}(t,0)}]^{\hat{\nu}_0}}{[1 + e^{\hat{\eta}(t,1)}]^{\hat{\nu}_1}} + y_t \log \frac{1 + e^{\hat{\eta}(t,0)}}{1 + e^{\hat{\eta}(t,1)}} \right\}
\end{aligned}$$

S.3. More tables for Sections 2 and 4.1

The results of the missing data imputation mentioned in Section 2 are provided in Table S.1. We apply the *last-value-carried-forward* strategy to handle these missing data and use bold font for "22:45:00" column and "11:55:00" column indicating imputed data.

In Tables S.2 and S.3 mentioned in Section 2, the imputed data, marked in bold font, are obtained by the *linear interpolation plus round-off* strategy.

In Table S.4, we list the holidays in Year 2017 used in our analysis mentioned in Section 4.1.

S.4. More figures for Sections 2 and 3.2

By courtesy of Figure 1 of [30] with the copyright permission, Figure S.1 displays the locations of ten sensors in District 3 of Sacramento, mentioned in Section 2.

Mentioned in Section 2, Figure S.2 shows the boxplots of the log-likelihood differences of workdays and weekends, respectively, which visually implies the possibility of using the difference for separating traffic flows of workdays and weekends (see also Section 3.2).

Using Sensor S314147 as an illustration, Figure S.3 shows how the threshold obtained by the SVC algorithm separates the two classes. For better visualization, we extend this one-dimensional space to

Sensor ID	Missing Time Periods					
	08/15/2017			10/15/2017		
	22:40:00	**22:45:00**	22:50:00	11:50:00	**11:55:00**	12:00:00
S312425	95	**95**	91	325	**325**	335
S312520	152	**152**	149	386	**386**	377
S312694	132	**132**	135	395	**395**	400
S312942	138	**138**	131	358	**358**	362
S314147	131	**131**	123	311	**311**	258
S315017	190	**190**	188	308	**308**	379
S315938	236	**236**	229	276	**276**	277
S317814	110	**110**	107	214	**214**	206
S318180	78	**78**	74	217	**217**	210
S318566	136	**136**	141	406	**406**	394

Table S.1. Imputed Missing Data Using the Last-value-carried-forward Strategy with Bold Font Indicating Imputed Ones

Sensor ID	Missing Time Periods						
	03/12/2017						
	01:55:00	**02:00:00**	**02:05:00**	**02:10:00**	**02:15:00**	**02:20:00**	**02:25:00**
S312425	39	**38**	**38**	**37**	**37**	**36**	**36**
S312520	53	**53**	**54**	**54**	**55**	**55**	**55**
S312694	61	**61**	**61**	**61**	**62**	**62**	**62**
S312942	70	**70**	**69**	**69**	**68**	**68**	**67**
S314147	66	**67**	**68**	**68**	**69**	**70**	**71**
S315017	145	**145**	**145**	**145**	**145**	**145**	**145**
S315938	57	**56**	**55**	**55**	**54**	**53**	**52**
S317814	66	**66**	**66**	**66**	**66**	**66**	**66**
S318180	39	**39**	**38**	**38**	**38**	**37**	**37**
S318566	83	**84**	**84**	**85**	**85**	**86**	**87**

Table S.2. Imputed Missing Data Using the Linear Interpolation plus Round-off Strategy with Bold Font Indicating Imputed Ones (Part I)

a two-dimensional space by introducing the day index as the x-coordinate. The day index here is from 1 to 365 corresponding to the date from January 1 to December 31, 2017. Therefore, the threshold we have obtained will become a straight line which is not influenced by the day index.

Using the negative binomial smoothing regression ANOVA model, we obtain the mean response curve for the Martin Luther King Jr. Day (see Figures S.4 mentioned in Section 3.2). The mean response curves associated with workdays and weekends are also obtained, respectively, as well as the 95% Bayesian credible bands.

S.5. 5-fold cross-validation in Sections 3.3 and 4.1

As mentioned in Sections 3.3 and 4.1, the 5-fold cross validation is applied for estimating the error rate and avoiding the potential risk of overfitting. The 5 equally sized subsets here are from workdays and weekends after removing holidays. To maintain the same data structure as the preceding datasets, the workday data and weekend data are treated separately and divided into 5 equal parts, respectively. By this way, the numbers of workdays and weekends in each subset (see Table S.5) can be kept with a ratio of 5:2.

Table S.6 provides the error rates for each of the ten sensors based on the 5-fold cross-validation. Note that all holidays are removed. It is consistent with the Workday and Weekend columns of Table 3.

Sensor ID	Missing Time Periods						
	03/12/2017						
	02:30:00	**02:35:00**	**02:40:00**	**02:45:00**	**02:50:00**	**02:55:00**	03:00:00
S312425	**35**	**35**	**34**	**34**	**33**	**33**	32
S312520	**56**	**56**	**56**	**57**	**57**	**58**	58
S312694	**62**	**62**	**62**	**63**	**63**	**63**	63
S312942	**67**	**66**	**66**	**65**	**65**	**64**	64
S314147	**71**	**72**	**73**	**74**	**74**	**75**	76
S315017	**146**	**146**	**146**	**146**	**146**	**146**	146
S315938	**52**	**51**	**50**	**49**	**49**	**48**	47
S317814	**65**	**65**	**65**	**65**	**65**	**65**	65
S318180	**37**	**37**	**36**	**36**	**36**	**35**	35
S318566	**87**	**88**	**89**	**89**	**90**	**90**	91

Table S.3. Imputed Missing Data Using the Linear Interpolation plus Round-off Strategy with Bold Font Indicating Imputed Ones (Part II)

Holiday	Date
New Year's Day Observed	Monday, January 2, 2017
Martin Luther King Jr. Day	Monday, January 16, 2017
Superbowl Sunday	Sunday, February 5, 2017
Presidents' Day	Monday, February 20, 2017
Memorial Day	Monday, May 29, 2017
Independent Day	Tuesday, July 4, 2017
Labor Day	Monday, September 4, 2017
Thanksgiving Day	Thursday November 23, 2017
Day After Thanksgiving	Friday, November 24, 2017
Christmas Day	Monday, December 25, 2017

Table S.4. List of the holidays in Year 2017 used in our analysis.

Table S.6 shows that there are 6 sensors have no prediction error at all; 3 sensors have only 1 prediction error; and 1 sensor (S312520) has 5 prediction errors. Overall the estimated prediction error rates based on 5-fold cross-validation are fairly low. It indicates that the SVC algorithm with log-likelihood difference works indeed very well.

S.6. Simulation study on robustness of estimated mean response curve

By applying Algorithm 1, we obtain the mean response curve $\hat{\mu}(t,k)$ for the kth group of daily traffic flow data. As mentioned in Section 4.2, in this section we use simulation study to check the robustness of the estimated mean response curve.

We use the workday traffic flow data recorded by Sensors S312694 and S315017 for illustrations. For each sensor, by applying Algorithm 1 to the group of workday traffic flow data $\{Y_{t0i}, t = 1,\ldots,288, i = 1,\ldots,n_0\}$, we obtain the parameter estimates $\hat{\nu}_0$ and $\hat{p}(t,0)$. Assuming that the negative distribution $f(y;\hat{\nu}_0,\hat{p}(t,0))$ as defined in (2) is the true distribution, we simulate a new dataset $Y'_{t0i} \sim f(y;\hat{\nu}_0,\hat{p}(t,0))$, $t = 1,\ldots,288$ and $i = 1,\ldots,n_0$. To show that the estimated mean response curve is not much affected by potential outliers, we artificially add some outliers from reported real traffic accidents. More specifically, we insert traffic flow data during the reported accident period into the simulated dataset with randomly chosen dates (see Table S.7).

In Figure S.5 (for Sensor S312649) and Figure S.6 (for Sensor 315017) we show both the mean response curves estimated from the real traffic data and the simulated data with inserted accidental traffic flows. For readers' reference, in each figure we add one simulated traffic flow with an inserted

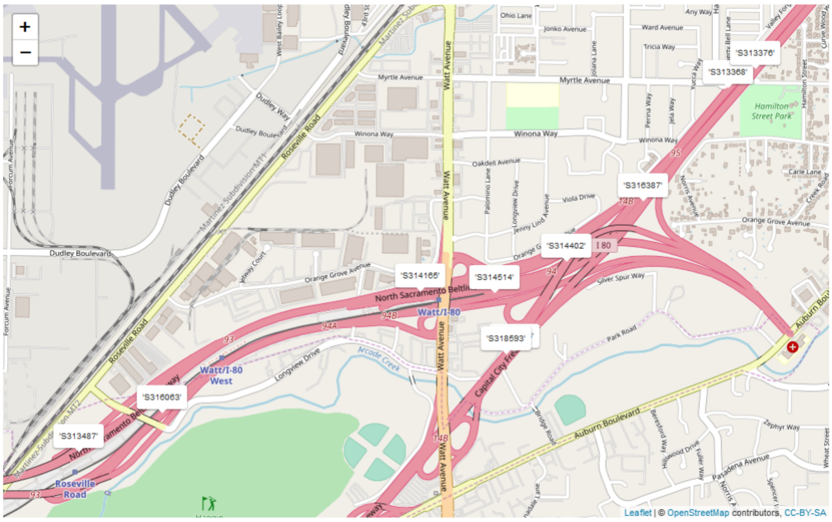

Figure S.1. Sensors with Known Locations in District 3 of Sacramento, Mainly Located on Highways

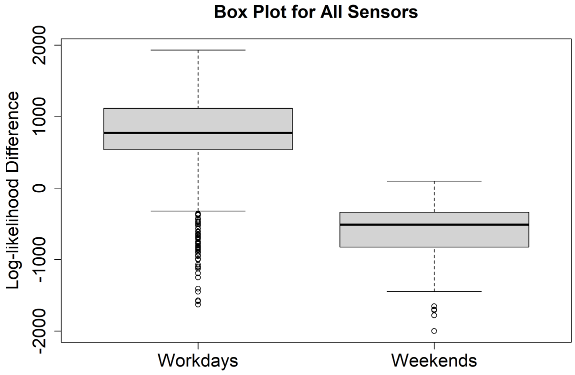

Figure S.2. Boxplots of Log-likelihood Differences of Workdays and Weekends from All Sensors

accident. The two inserted accidents displayed are real accidents originally occurred on 02/10/2017 (see Figure S.5) and 02/06/2017 (see Figure S.6). Both accidents have a long duration and are evaluated as *severe* according to our impact rate (see Section 4.3 and Table 4).

According to Figures S.5 and S.6, our estimated mean response curve is not affected by artificially added accidents, even if some of them are severe. It shows that our proposed model is reliable and the estimated mean response curve is fairly robust.

S.7. Step function versus linear function in Section 4.3

As mentioned in Section 4.3, the traffic flow $y(x)$ is actually observed only at discrete time points. To calculate the impact factor (4) and impact intensity (5), there are two methods: one is to use a step function (still denoted by $y(x)$; the other is to use a piecewise linear function $f(x)$. In this section, we compare these two methods.

Using Sensor S312694 as an illustration, the impact factors (I or a restricted version I_5 with $x_R - x_L = 5$ or less) and impact rates (or a restricted version $I_5/I \times 100\%$) are listed in Table S.8. It should be noted that the accident reported by S312694 was from 12:44:00 to 14:28:00 on February 10, 2017, with a total duration of 104 minutes. The total impact factor is 18735.97 with impact rate

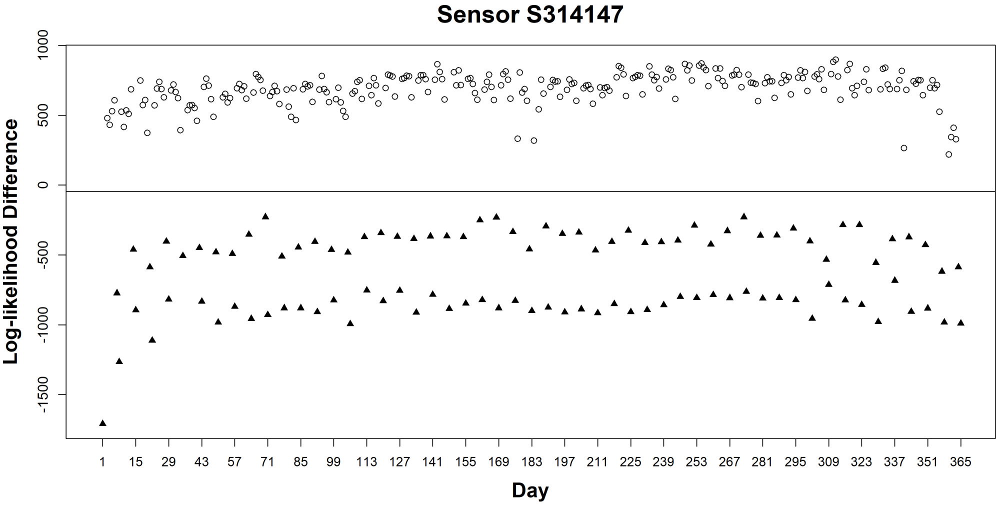

Figure S.3. SVC Classification for Sensor S314147: Straight line shows the threshold −45.53; hollow circles represent workdays; solid triangles indicate weekends

	5 Equal Sized Subsets				
Workday (except holidays)	51	50	50	50	50
Weekend (except holidays)	21	21	21	21	20
Subset	72	71	71	71	70

Table S.5. Sizes of 5-fold Cross-validation Subsets: 251 workdays (without holidays) are divided into 5 equal-sized subgroups; 104 weekdays (without holidays) are divided into 5 equal-sized subgroups; each subset of the 5-fold cross-validation consists of about 71 days

5.22% based on the step function or 18765.24 with impact rate 5.23% based on the piecewise linear function, which are pretty close.

To visually see the difference, in Figure S.7 we show the graph of individual impact factors (impact factors restricted to 5-min intervals) against the index of 5-min intervals. It can be seen that the two curves are fairly close to each other, which implies the step function and piecewise linear function provide essentially the same results in this study.

S.8. Further separating Saturdays and Sundays

As mentioned in Section 5, it is worthy of checking differences of traffic flow patterns between Saturdays and Sundays. Have obtained the log-likelihood differences and their visualization, a special feature emerges in the weekend cluster. There seems to be two paths and exists clear gaps between them. The mean response curves and confidence bands are implemented again in this case. The weekend data is divided into Saturday's and Sunday's. Figures S.8 and S.9 show that the pattern of Saturdays is significantly different from Sundays'.

It can be found that the mean response curves and corresponding confidence bands of Saturday and Sunday do not overlap during a considerate amount time, especially between 4am and 9am. There is indeed a significant gap, although the overall trends are similar. Figures S.8 and S.9) show that although Saturdays and Sundays both belong to weekends, their traffic flow patterns are still significantly different. After checking the 24-hour data of weekends, it can be seen that the traffic counts on Saturday tend to be higher than Sunday's, which indicates that more travels involve on Saturdays than Sundays.

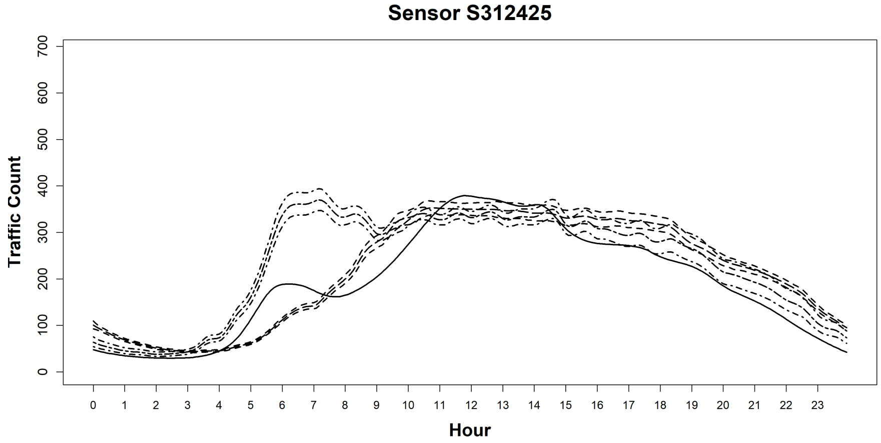

Figure S.4. Mean Response Curves and Confidence Bands with the MLK Day for Sensor S312425: Dotted lines represent the mean response curves and the confidence bands of workdays and weekends, respectively; solid line is the mean response curve of the Martin Luther King Jr. Day's

Sensor ID	5-Fold Cross-Validation Error	
	Error Count	Error Rate (%)
S312425	0	0
S312520	5	1.41
S312694	0	0
S312942	0	0
S314147	0	0
S315017	1	0.28
S315938	1	0.28
S317814	1	0.28
S318180	0	0
S318566	0	0

Table S.6. Estimated Error Rates in Separating Patterns of Weekdays and Weekends Based on 5-fold Cross-validation without Holidays

S.9. Functional data analysis

As mentioned in Section3.1, the traffic flows can be considered as a random function which the functional data analysis (FDA) [31] could be applied to. Moreover, generalized functional linear models [18, 21] and functional ANOVA [20] could be alternative approaches.

After implementing FDA to all the traffic flow data recorded by 10 sensors, the comparison results of these two methods are obtained. In general, although the mean response curves obtained by both have almost the same pattern, the SSANOVA-based mean response curves are smoother than the FDA-based curves for all 10 sensors. It is important to note that the FDA-based mean response curves are very rough during some time periods. Since Sensors S314147, S315017, S315938 and S318180 recorded many accidents and the duration of these accidents are long (see Table 4), we use these sensors as illustrations to show the comparison of the mean response curves. From Figures S.10, S.11, S.12 and S.13, it can be seen that the FDA-based mean response curves are overall rougher than the SSANOVA-based curves, especially in the marked time periods.

In Section4.1, we perform Support Vector Classifier (SVC) and 5-fold cross-validations to validate the threshold for classification and measure the error count based on the SSANOVA (see Table 3). Similarly, we need to apply these two methods to the FDA. The summarized error counts based on

Original Accident Date	Accident Start Timestamp	Accident Duration (mins)	Inserted Date in Simulated Dataset
09/05/2017	01:45:00	173	01/19/2017
09/26/2017	00:17:00	105	02/06/2017
02/10/2017	12:44:00	104	03/22/2017
03/22/2017	06:31:00	124	04/14/2017
01/12/2017	11:05:00	212	05/17/2017
05/05/2017	09:21:00	165	06/06/2017
07/18/2017	12:08:00	108	07/18/2017
02/06/2017	11:42:00	211	07/31/2017
04/05/2017	10:45:00	102	08/24/2017
05/30/2017	05:51:00	169	09/05/2017
12/01/2017	08:30:00	122	09/20/2017
07/26/2017	05:23:00	178	10/09/2017
09/05/2017	16:56:00	104	11/17/2017
10/28/2017	16:14:00	122	11/28/2017
06/12/2017	16:47:00	128	12/19/2017

Table S.7. Reported Traffic Accident Information and Randomly Inserted Date in Simulated Dataset

5-fold cross-validations for all 10 sensors based on the thresholds determined by the SVC algorithm are provided in Table S.9.

Comparing the error rates in Table 3 and Table S.9, the classification results in Sensors S314147, S315017 and S318180 are different. For Sensor S314147, the FDA-based SVC misclassifies one more workday data point than the SSANOVA-based SVC. For Sensor S315017, the FDA-based SVC misclassifies by two more workday data points. And for Sensor S318180, the FDA-based SVC misclassifies by two more workday data points than the SSANOVA-based SVC. Overall, the SSANOVA performs more reliably than the FDA, especially for workday's traffic flow data, and for weekend's data, both perform equally well.

In Section 4.3, we calculate the impact factors and impact rates of reported accidents for 10 sensors (see Table 4). For comparison, after obtaining the FDA-based mean response curves, we also calculate the FDA-based impact factors and rates for these reported accidents. The results are presented in Table S.10. In general, the impact factors and rates based on these two methods are roughly the same. Since only 15 accidents are reported in these 10 sensors, these impact factors and rates will be approximately the same when the overall patterns of the estimated mean response curves are the same, even the SSANOVA-based curves are smoother than the FDA-based curves.

Furthermore, we analyzed the robustness of FDA-based mean response curves needs to be analyzed in comparison to the SSANOVA-based curves in Section S.6. In Section S.6, we use simulation study to check the robustness of the estimated mean response curve and use the workday traffic data recorded by Sensor S312694 and S315017 for illustrations. Therefore, for a fair comparison, we use the same simulated dataset and the same sensors to implement the analysis on the robustness of FDA-based mean response curves. The visualization results can be found in Figures S.14 and S.15.

For Sensor S312694 (see Figure S.14), after inserting multiple accidents, its mean response curve (based on simulated data) appears to fluctuate significantly, and this curve becomes rougher compared to the real data mean response curve. In particular, this roughness is very obvious within marked long time period from 06:20:00 to 19:40:00. This situation indicates poor robustness of the FDA-based mean response curve. It also happens with Sensor S315017 (see Figure S.15), the estimated mean response curve obtained by simulated data becomes very rough and lasts for a long time from 03:20:00 to 18:50:00. From the performance of these two sensors, we find that the FDA-based curves lack robustness, so we do not recommend this method in our analysis.

In summary, the SSANOVA method provides smaller error counts and rates, and the estimated mean response curves obtained by this method are smoother and more robust than those obtained by the FDA method. Therefore, we prefer SSANOVA in this paper.

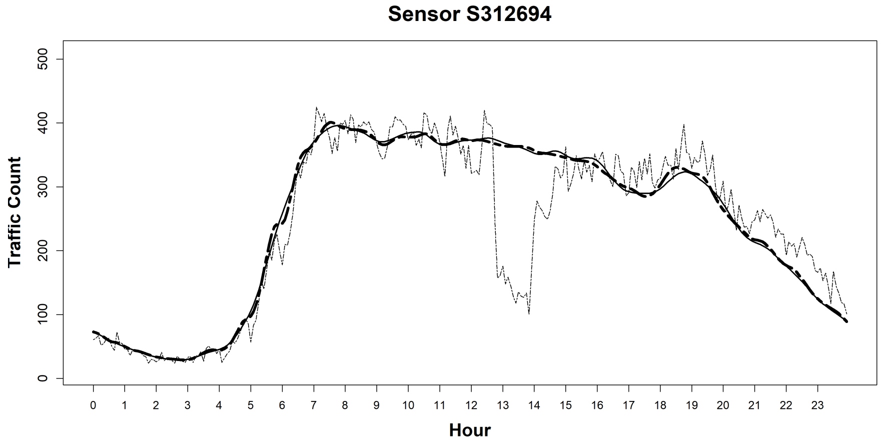

Figure S.5. Comparison of Mean Response Curves from Real Traffic Data and Simulated Data for Sensor S312694: Bold dotted line represents the estimated mean response curve from real workday traffic data; bold solid line is the estimated mean response curve from simulated data with inserted accidents; thin dashed line is one simulated traffic flow curve with inserted accident originally occurred on 02/10/2017

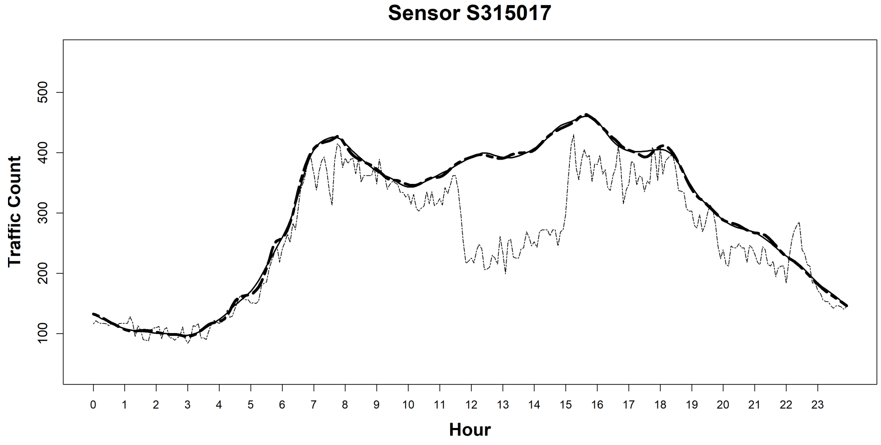

Figure S.6. Comparison of Mean Response Curves from Real Traffic Data and Simulated Data for Sensor S315017: Bold dotted line represents the estimated mean response curve from real workday traffic data; bold solid line is the estimated mean response curve from simulated data with inserted accidents; thin dashed line is one simulated traffic flow curve with inserted accident originally occurred on 02/06/2017

Timestamp	Impact Factor on Step Function	Impact Rate (%)	Impact Factor on Piecewise Linear Function	Impact Rate (%)
12:44:00 - 12:45:00	24.09	0.13	111.93	0.60
12:45:00 - 12:50:00	608.46	3.25	828.05	4.41
12:50:00 - 12:55:00	1047.63	5.59	1035.00	5.52
12:55:00 - 13:00:00	1022.37	5.46	982.61	5.24
13:00:00 - 13:05:00	942.85	5.03	1011.04	5.39
13:05:00 - 13:10:00	1079.22	5.76	1050.44	5.60
13:10:00 - 13:15:00	1021.66	5.45	1053.47	5.61
13:15:00 - 13:20:00	1085.29	5.79	1130.35	6.02
13:20:00 - 13:25:00	1175.40	6.27	1201.15	6.40
13:25:00 - 13:30:00	1226.90	6.55	1182.22	6.30
13:30:00 - 13:35:00	1137.54	6.07	1154.68	6.15
13:35:00 - 13:40:00	1171.81	6.25	1175.80	6.27
13:40:00 - 13:45:00	1179.80	6.30	1160.58	6.18
13:45:00 - 13:50:00	1141.35	6.09	1223.31	6.52
13:50:00 - 13:55:00	1305.26	6.97	1125.51	6.00
13:55:00 - 14:00:00	945.75	5.05	750.23	4.00
14:00:00 - 14:05:00	554.71	2.96	469.65	2.50
14:05:00 - 14:10:00	384.60	2.05	406.23	2.16
14:10:00 - 14:15:00	427.87	2.28	439.16	2.34
14:15:00 - 14:20:00	450.45	2.40	471.95	2.52
14:20:00 - 14:25:00	493.46	2.63	504.66	2.69
14:25:00 - 14:28:00	309.51	1.65	297.24	1.58
Total	**18735.97**	**5.22**	**18765.24**	**5.23**

Table S.8. Comparisons of Impact Factors and Rates Based on Step Function or Piecewise Linear Function for Sensor S312694: Impact factor of each 5-min is restricted to 5-min intervals; impact rate of each 5-min is the percentage of the restricted impact factor relative to the (total) impact factor; (total) impact rate is the percentage of impact factor relative to the total area under the estimated mean response curve

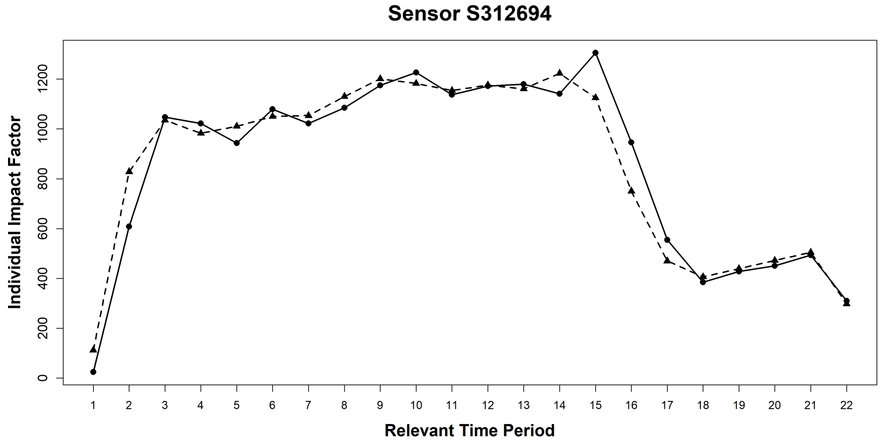

Figure S.7. Impact Factors Restricted to 5-min Intervals against Interval Index: Real broken lines with solid circles represent the 5-min impacts factors calculated with the step function. The dotted broken lines with solid triangles represent the 5-min impact factors calculated by the piecewise linear function

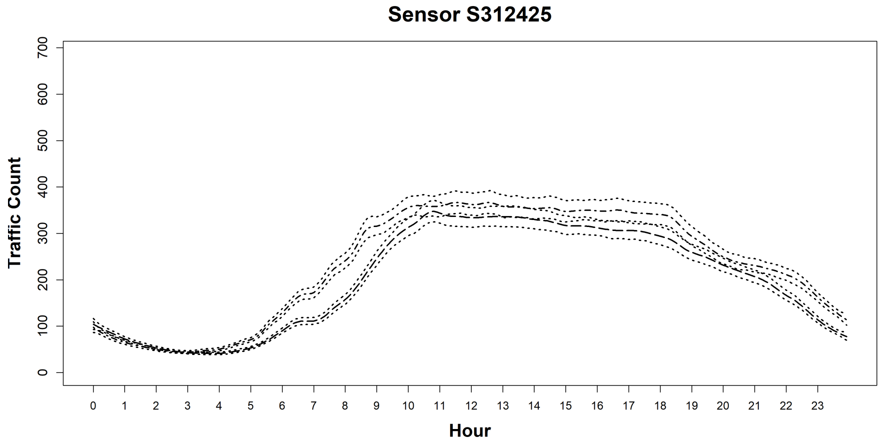

Figure S.8. Mean Response Curves and Confidence Bands on Saturdays and Sundays for Sensor S312425: Higher curves are the Saturday's. The lower curves are the Sunday's.

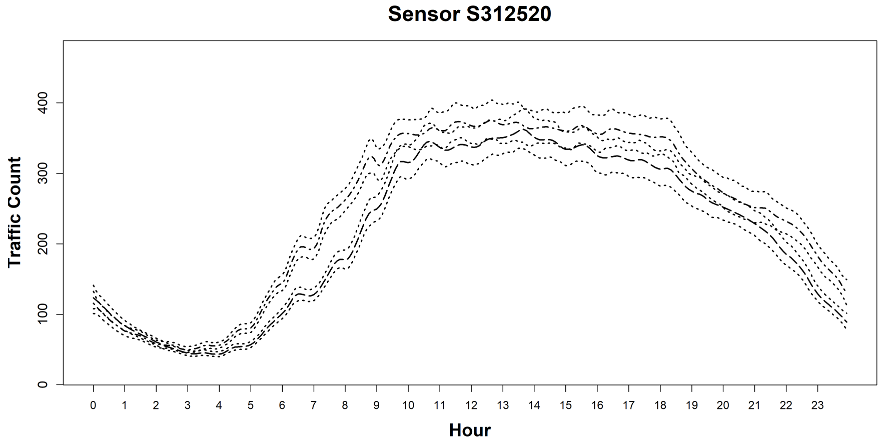

Figure S.9. Mean Response Curves and Confidence Bands on Saturdays and Sundays for Sensor S312520: Higher curves are Saturday's. The lower curves are Sunday's

Sensor ID	Error						Total Error	
	Workday (no holidays)		Weekend (no holidays)		Holidays Only			
	Error Count	Error Rate (%)	Error Count	Error Rate (%)	Error Count	Error Rate (%)	Error Count	Error Rate (%)
S312425	0	0	0	0	0	0	0	0
S312520	5	1.99	0	0	1	10	6	1.64
S312694	0	0	0	0	1	10	1	0.27
S312942	0	0	0	0	0	0	0	0
S314147	1	0.40	0	0	1	10	2	0.55
S315017	3	1.20	0	0	1	10	4	1.10
S315938	0	0	1	0.96	0	0	1	0.27
S317814	1	0.40	0	0	1	10	2	0.55
S318180	2	0.80	0	0	1	10	3	0.82
S318566	0	0	0	0	2	20	2	0.55

Table S.9. Error Counts and Rates of 5-fold Cross-validation with Holidays Treated as Weekends Based on Functional Data Analysis

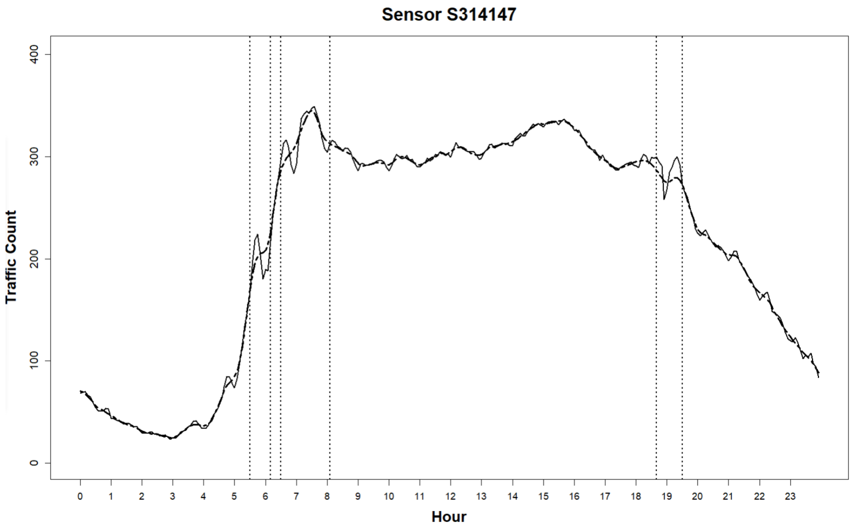

Figure S.10. SSANOVA-Based Mean Response Curve vs FDA-Based Mean Response Curve for Sensor S314147 Workday Traffic Flow Data: Dotted line represents the SSANOVA-based mean response curve and solid line is the FDA-based mean response curve

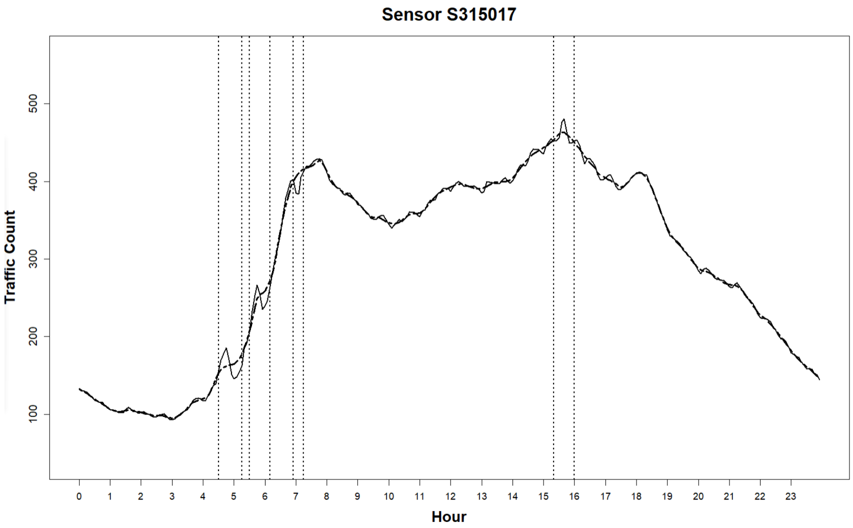

Figure S.11. SSANOVA-Based Mean Response Curve vs FDA-Based Mean Response Curve for Sensor S315017 Workday Traffic Flow Data: Dotted line represents the SSANOVA-based mean response curve and solid line is the FDA-based mean response curve

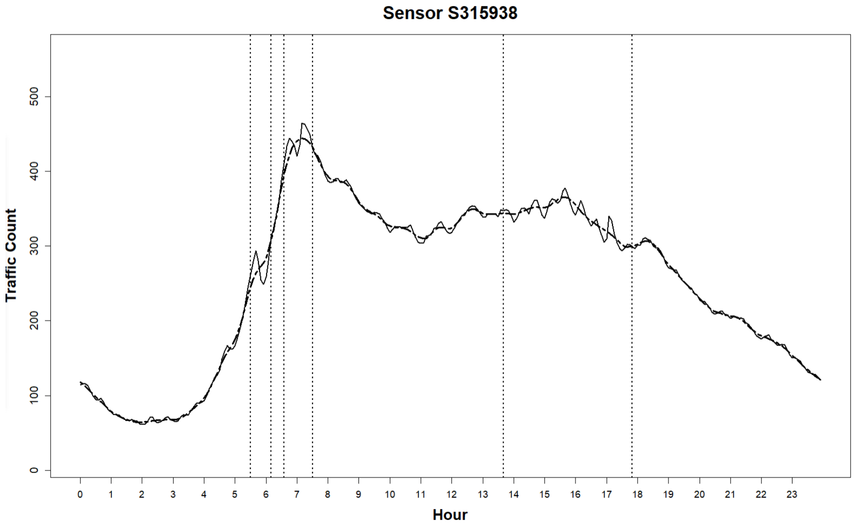

Figure S.12. SSANOVA-Based Mean Response Curve vs FDA-Based Mean Response Curve for Sensor S315938 Workday Traffic Flow Data: Dotted line represents the SSANOVA-based mean response curve and solid line is the FDA-based mean response curve

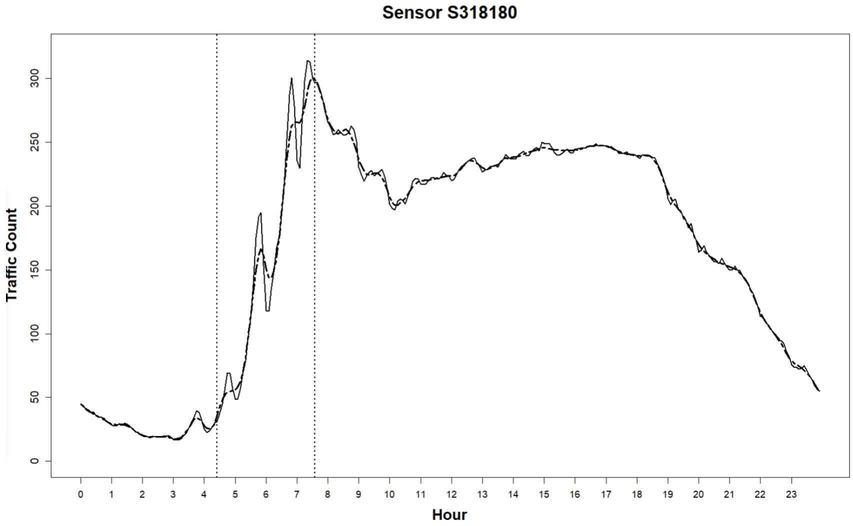

Figure S.13. SSANOVA-Based Mean Response Curve vs FDA-Based Mean Response Curve for Sensor S318180 Workday Traffic Flow Data: Dotted line represents the SSANOVA-based mean response curve and solid line is the FDA-based mean response curve

Sensor	Accident		Smoothing Regression		Functional Data Analysis	
Sensor ID	Accident Date	Accident Duration (mins)	Impact Factor	Impact Rate (%)	Impact Factor	Impact Rate (%)
S312425	09/05/2017	173	2010	0.60	1957	0.58
S312520	09/26/2017	105	1008	0.26	940	0.24
S312694	02/10/2017	104	18736	5.22	18744	5.22
S312942	03/22/2017	124	7579	2.02	7630	2.03
S314147	01/12/2017	212	16605	5.27	16937	5.67
S314147	05/05/2017	165	6699	2.13	6939	2.36
S314147	07/18/2017	108	2636	0.84	2671	0.85
S315017	02/06/2017	211	33089	7.77	33085	7.76
S315017	04/05/2017	102	3234	0.76	3251	0.76
S315938	05/30/2017	169	10957	2.98	10883	2.95
S315938	12/01/2017	122	6366	1.73	6372	1.73
S317814	07/26/2017	178	15861	6.86	15979	6.92
S318180	09/05/2017	104	3051	1.28	3024	1.26
S318180	10/28/2017	122	4236	2.18	4063	1.70
S318566	06/12/2017	128	4487	1.03	4402	1.01

Table S.10. Impact Factors and Rates of Reported Accidents Based on Smoothing Regression and Functional Data Analysis

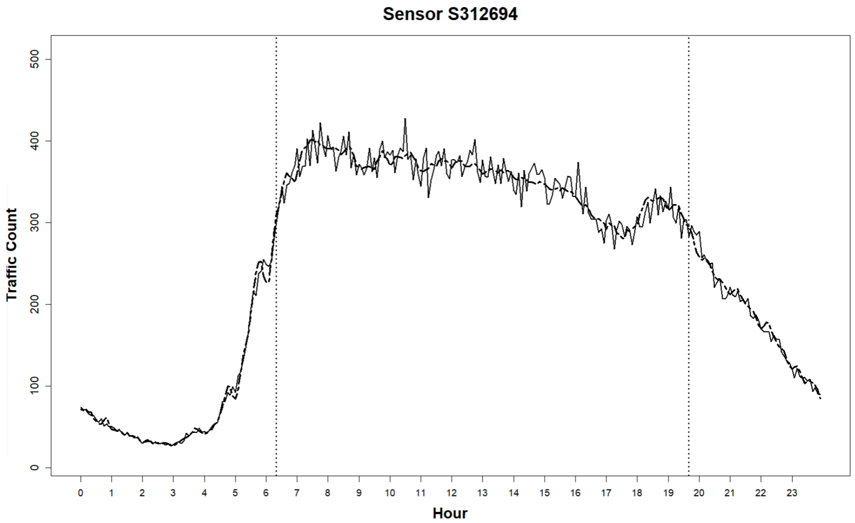

Figure S.14. Comparison of FDA-Based Mean Response Curves from Real Traffic Data and Simulated Data for Sensor S312694: Dotted line represents the mean response curve from real workday traffic data. The solid line is the mean response curve from simulated data with inserted accidents

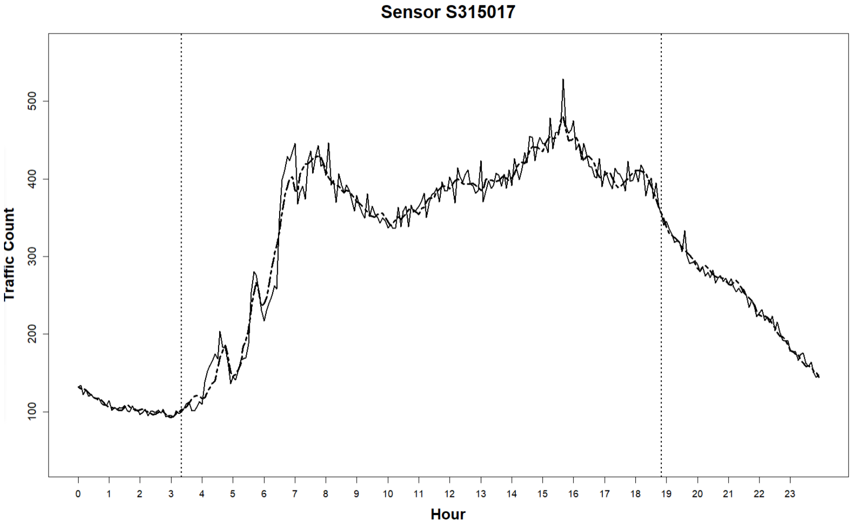

Figure S.15. Comparison of FDA-Based Mean Response Curves from Real Traffic Data and Simulated Data for Sensor S315017. The dotted line represents the mean response curve from real workday traffic data. The solid line is the mean response curve from simulated data with inserted accidents