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

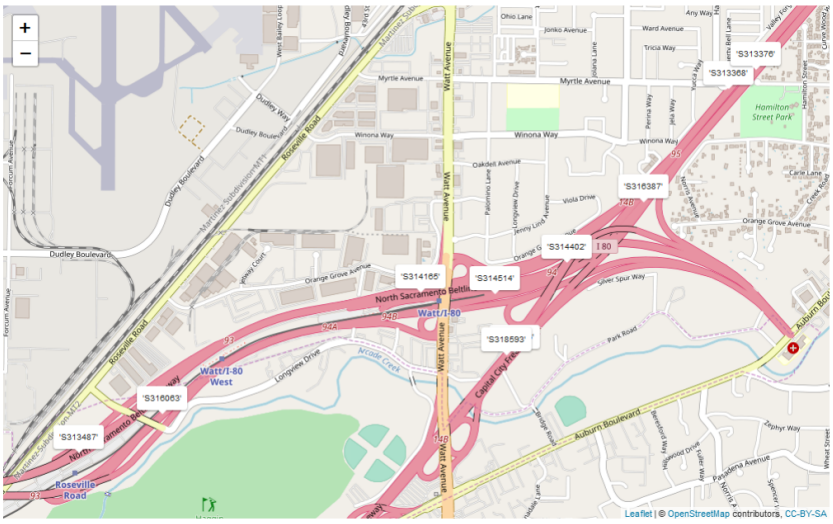

Figure S.1. Sensors with Known Locations in District 3 of Sacramento, Mainly Located on Highways

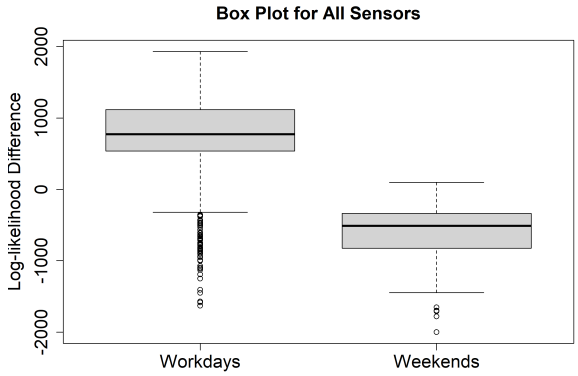

Figure S.2. Boxplots of Log-likelihood Differences of Workdays and Weekends from All Sensors

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

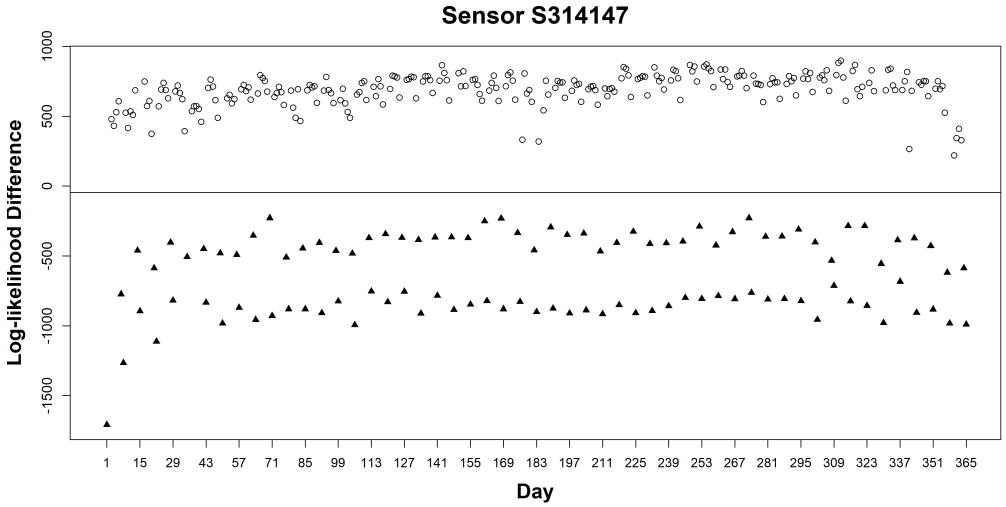

Figure S.3. SVC Classification for Sensor S314147: Straight line shows the threshold −45.53; hollow circles represent workdays; solid triangles indicate weekends

	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.

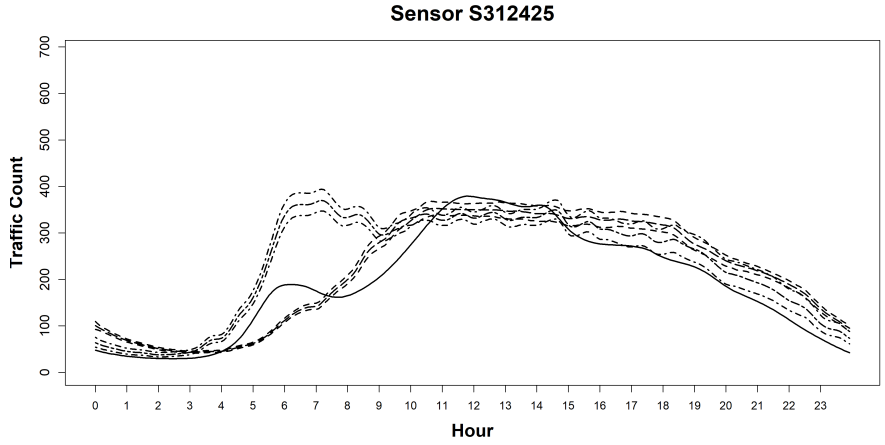

Figure S.4. Mean Response Curves and Confidence Bands with the MLK Day for Sensor S312425: Dotted lines represent the mean response curves and the confidence bands of workdays and weekends, respectively; solid line is the mean response curve of the Martin Luther King Jr. Day's

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.

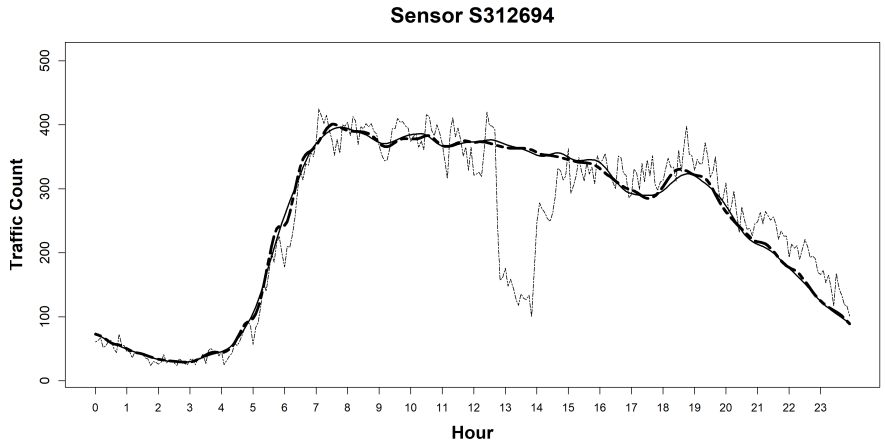

Figure S.5. Comparison of Mean Response Curves from Real Traffic Data and Simulated Data for Sensor S312694: Bold dotted line represents the estimated mean response curve from real workday traffic data; bold solid line is the estimated mean response curve from simulated data with inserted accidents;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

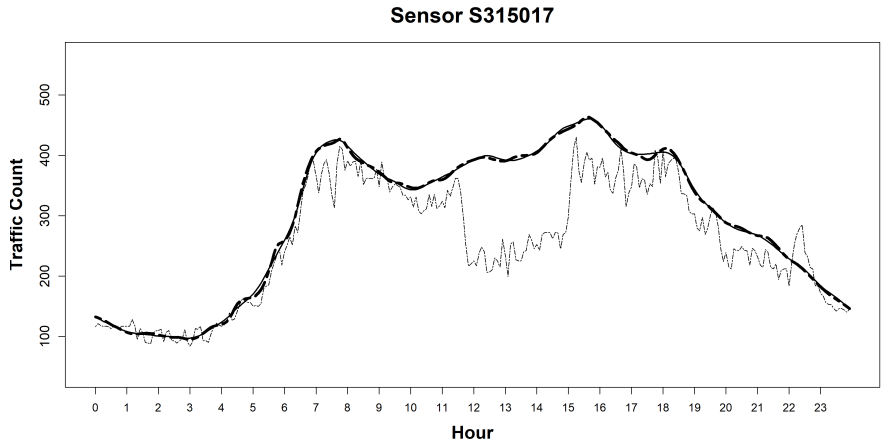

Figure S.6. Comparison of Mean Response Curves from Real Traffic Data and Simulated Data for Sensor S315017: Bold dotted line represents the estimated mean response curve from real workday traffic data; bold solid line is the estimated mean response curve from simulated data with inserted accidents; 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

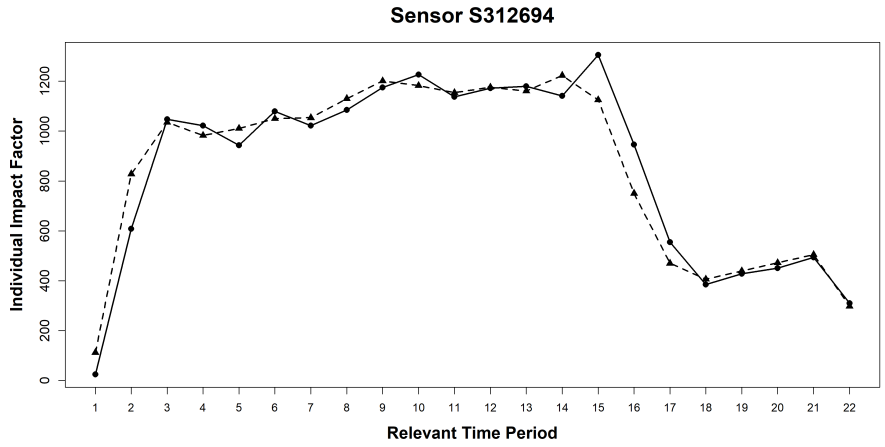

Figure S.7. Impact Factors Restricted to 5-min Intervals against Interval Index: Real broken lines with solid circles represent the 5-min impacts factors calculated with the step function. The dotted broken lines with solid triangles represent the 5-min impact factors calculated by the piecewise linear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

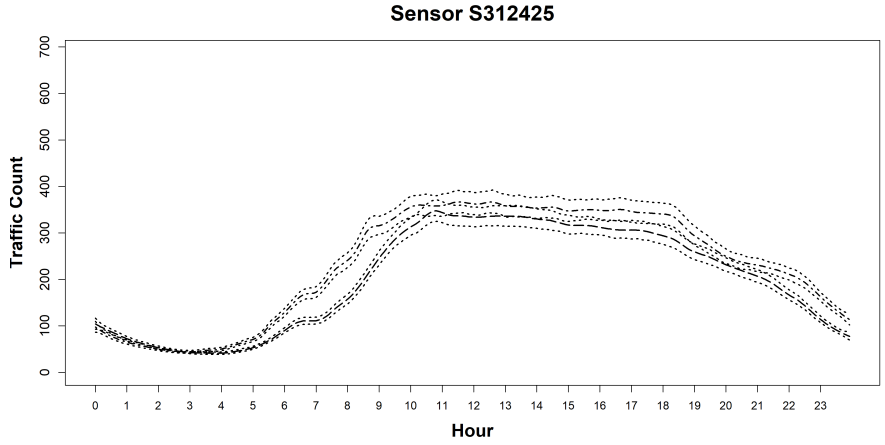

Figure S.8. Mean Response Curves and Confidence Bands on Saturdays and Sundays for Sensor S312425: Higher curves are the Saturday's. The lower curves are the Sunday's.

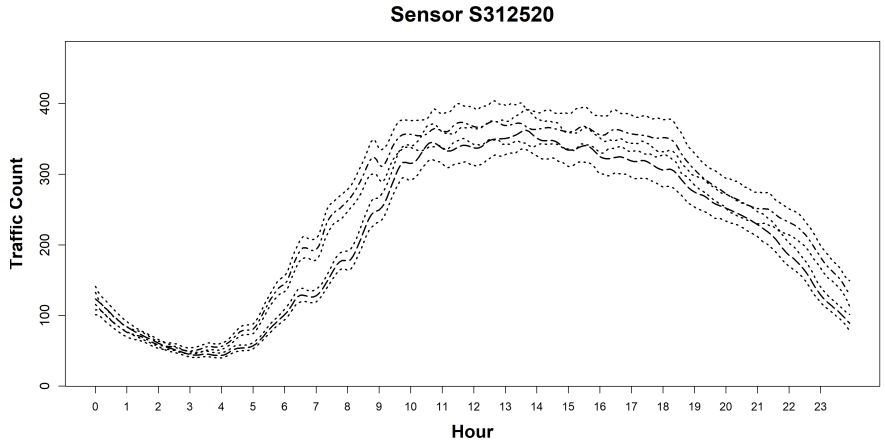

Figure S.9. Mean Response Curves and Confidence Bands on Saturdays and Sundays for Sensor S312520: Higher curves are Saturday's. The lower curves are Sunday's

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

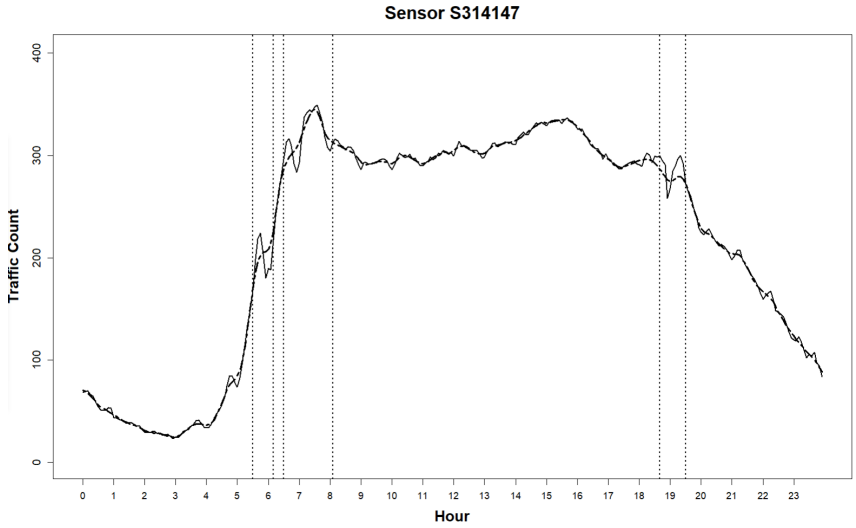

Figure S.10. SSANOVA-Based Mean Response Curve vs FDA-Based Mean Response Curve for Sensor S314147 Workday Traffic Flow Data: Dotted line represents the SSANOVA-based mean response curve and solid line is the FDA-based mean response curve

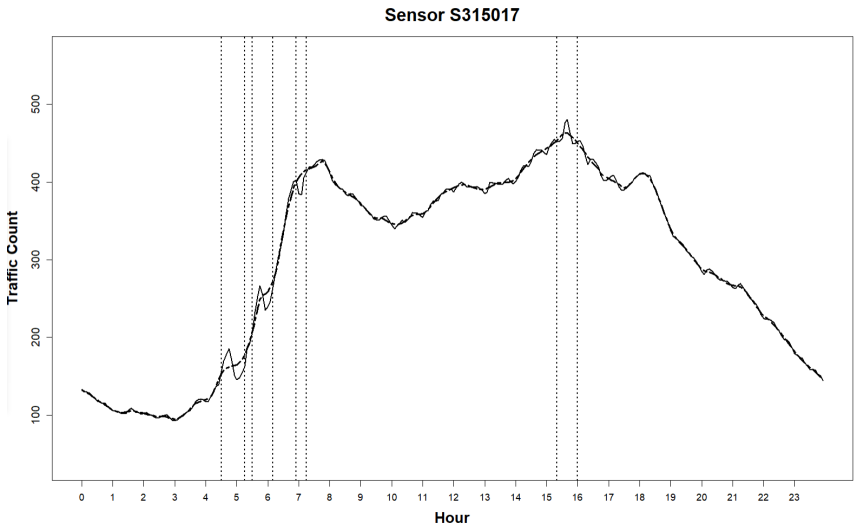

Figure S.11. SSANOVA-Based Mean Response Curve vs FDA-Based Mean Response Curve for Sensor S315017 Workday Traffic Flow Data: Dotted line represents the SSANOVA-based mean response curve and solid line is the FDA-based mean response curve

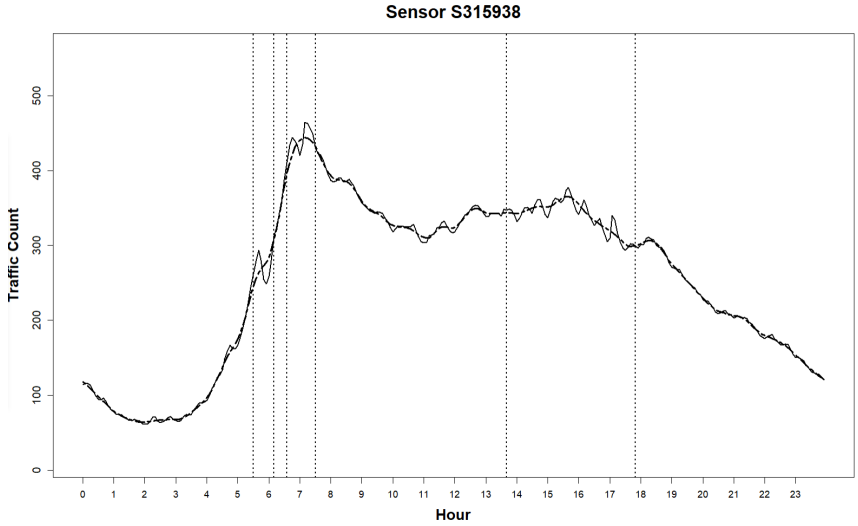

Figure S.12. SSANOVA-Based Mean Response Curve vs FDA-Based Mean Response Curve for Sensor S315938 Workday Traffic Flow Data: Dotted line represents the SSANOVA-based mean response curve and solid line is the FDA-based mean response curve

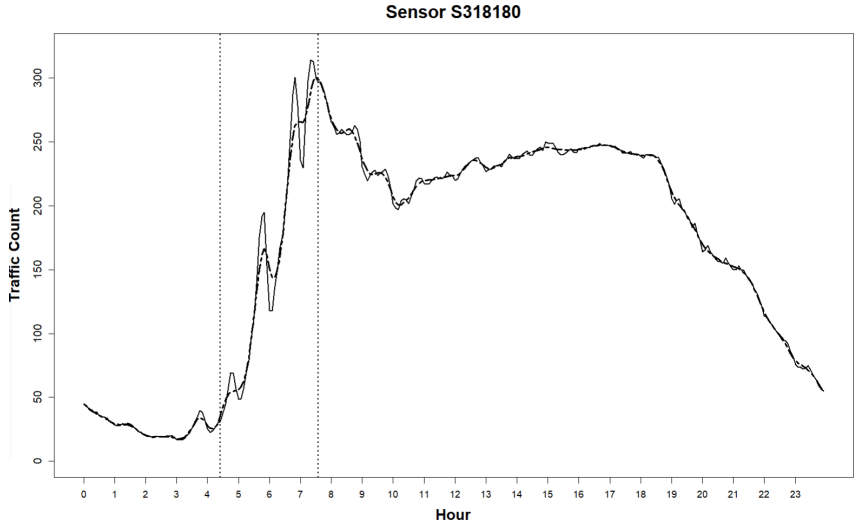

Figure S.13. SSANOVA-Based Mean Response Curve vs FDA-Based Mean Response Curve for Sensor S318180 Workday Traffic Flow Data: Dotted line represents the SSANOVA-based mean response curve and solid line is the FDA-based mean response curve

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

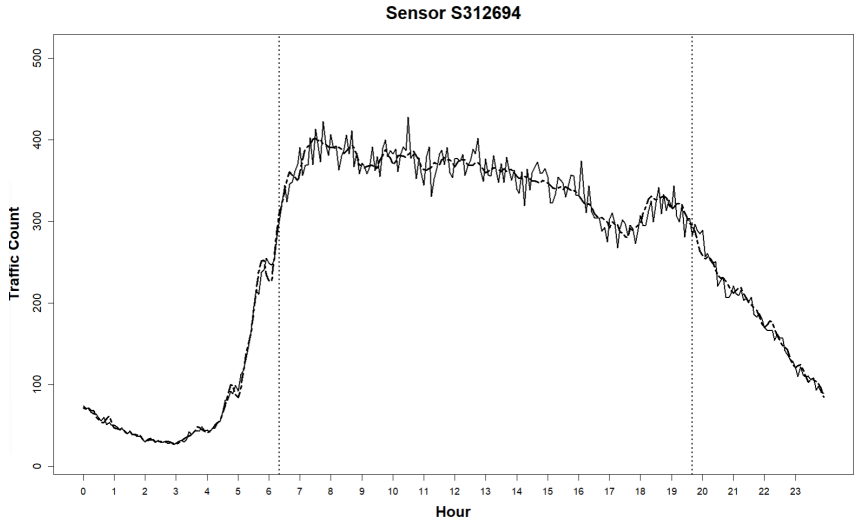

Figure S.14. Comparison of FDA-Based Mean Response Curves from Real Traffic Data and Simulated Data for Sensor S312694: Dotted line represents the mean response curve from real workday traffic data. The solid line is the mean response curve from simulated data with inserted accidents

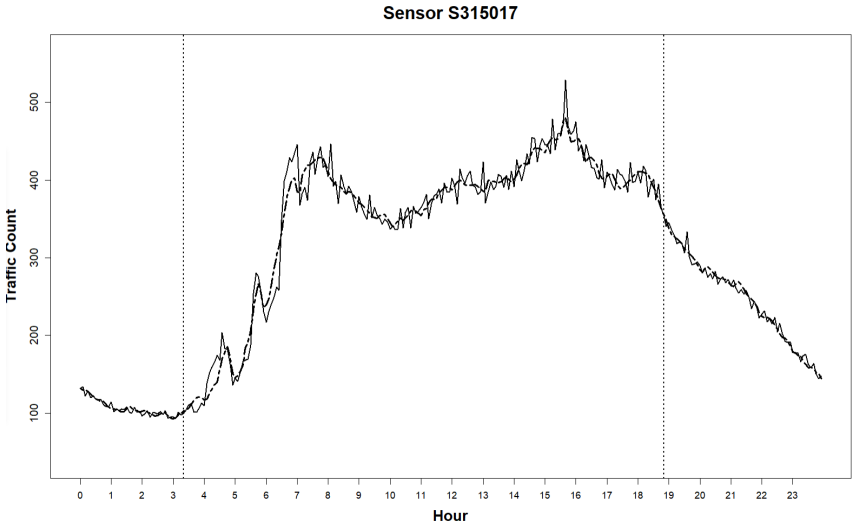

Figure S.15. Comparison of FDA-Based Mean Response Curves from Real Traffic Data and Simulated Data for Sensor S315017. The dotted line represents the mean response curve from real workday traffic data. The solid line is the mean response curve from simulated data with inserted accidents