Online Surrogate Measures and its Application to Transportation Safety Analysis

Yilin Zhang

May 8, 2021

Background

Background

- Traffic crashes are rare events, with an average rate of one crash 6.8 per million vehicle miles traveled in the US.
- The **naturalistic driving study** (NDS) provides an unprecedented opportunity to evaluate crash risk.
- **Second Strategic Highway Research Program** (SHRP2) NDS, the largest NDS to-date, with more than 3,400 participants and 1 million hours of continuous driving data.
- The entire dataset comprised of about 2000 traffic crashes, about 8000 near crashes, and tons of thousands of normal driving segments.

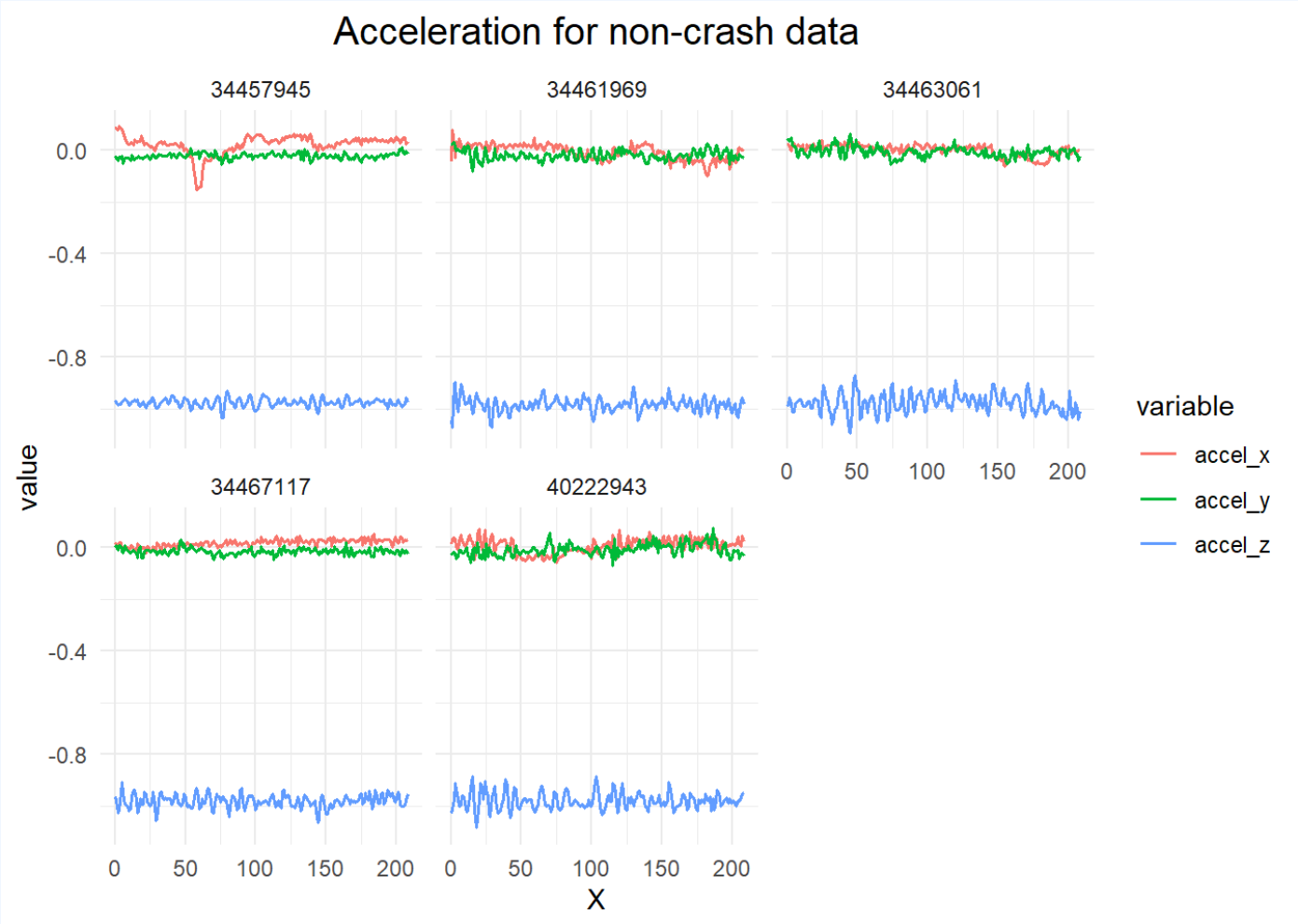
Surrogate measure

A broad spectrum of studies investigate identification and development of various surrogate measures, which helps to enhance the reliability of safety analysis.

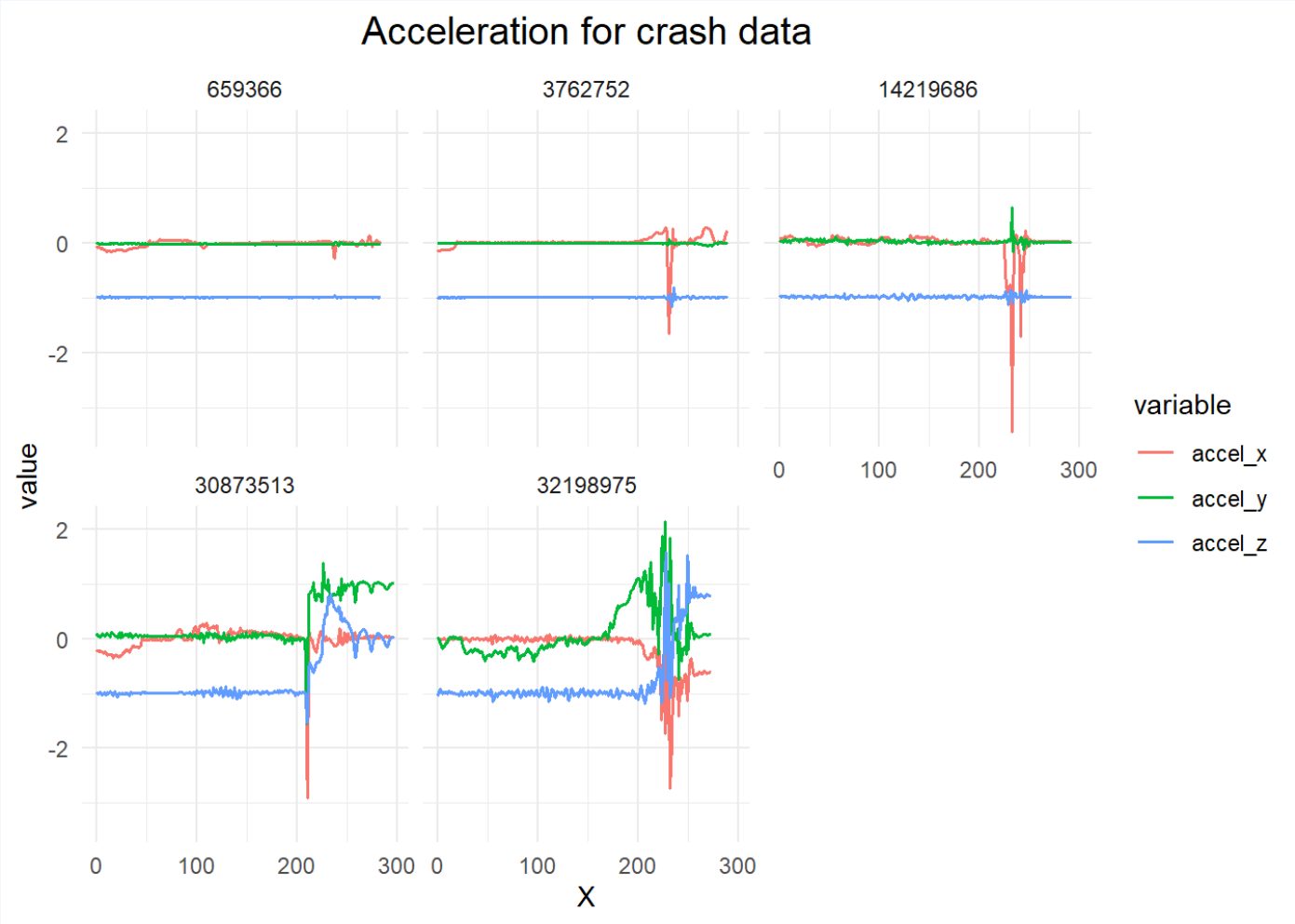
Our core concern is to propose surrogates based on three-dimension of acceleration.

- Propose new surrogate measures to predict the risk for crash
- Implement the surrogate measures with online version

Data


Acceleration for non-crash data
34457945
34461969
34463061
34467117
40222943
0.0
-0.4
-0.8
value
0
50
100
150
200
X
variable
accel_x
accel_y
accel_z

Data


Acceleration for crash data
659366
3762752
14219686
30873513
32198975
value
X
variable
accel_x
accel_y
accel_z

Data

Trainging Data: Each driving segment is about 200 time points, 60000 no risk segments and 1000 risky segments.

Testing Data: Each driving trip is about 2000 time points, 500 no risk trip and 100 risky trips.

Surrogates and online estimates

Surrogate Measure

We propose three surrogates that can be estimated via U statistics.

- Standard deviation measures variation of the distribution.
- Coefficient of Variation (CV) measures dispersion after standardization.
- Skewness measures asymmetry of the distribution.

Online estimation

Data $\{X_i\}_{i=1}^N$ flow in our machine in batch.

- Initially, we only have $\{X_i\}_{i=1}^n$ data in the memory
- For each time, r new samples are updated in the memory, replacing those oldest r ones

Online U statistics

Step 1: We calculate the U statistics based on all initial data, which is

$$\widehat{\theta}^{(0)} \stackrel{\text{def}}{=} \binom{n}{m}^{-1} \sum_{\{i_1,\ldots,i_m\}\in I_0} h(X_{i_1},\ldots,h_{i_m}).$$

Online estimation

Step 2: To refresh the U statistics in time t, we use the following iterative formula.

$$
\begin{aligned}
R^{(t)} &\stackrel{\text{def}}{=} \sum_{\{i_1,\ldots,i_m\}\in I_t} h(X_{i_1},\ldots,X_{i_m}),\\
C^{(t)} &\stackrel{\text{def}}{=} \sum_{k=1}^{m-1} \sum_{\{i_1,\ldots,i_k\}\in J_{t-1},\{i_{k+1},\ldots,i_m\}\in I_t} h(X_{i_1},\ldots,X_{i_m}) \text{ and}\\
\widehat{\theta}^{(t)} &\stackrel{\text{def}}{=} \left[\left\{t\binom{n}{m} - (t-1)\binom{n-r}{m}\right\}\widehat{\theta}^{(t-1)} + C^{(t)} + R^{(t)}\right]\\
&\qquad \left\{(t+1)\binom{n}{m} - t\binom{n-r}{m}\right\}^{-1}.
\end{aligned}
$$

Step 3: By transforming $\widehat{\theta}^{(t)}$ as $T(\widehat{\theta}^{(t)})$, we derive the new estimates for $T(\theta)$.

Online estimation

To derive the asymptotic properties for $\widehat{\theta}^{(t)}$, we denote $N_t \stackrel{\text{def}}{=} n + rt$, which is the sample size at time t. Then we can derive the asymptotic normality for our online estimates.

Theorem 1

Given $E\{h(X_1, \ldots, X_m)\}^2 < \infty$ and $\zeta_1 > 0$, we have

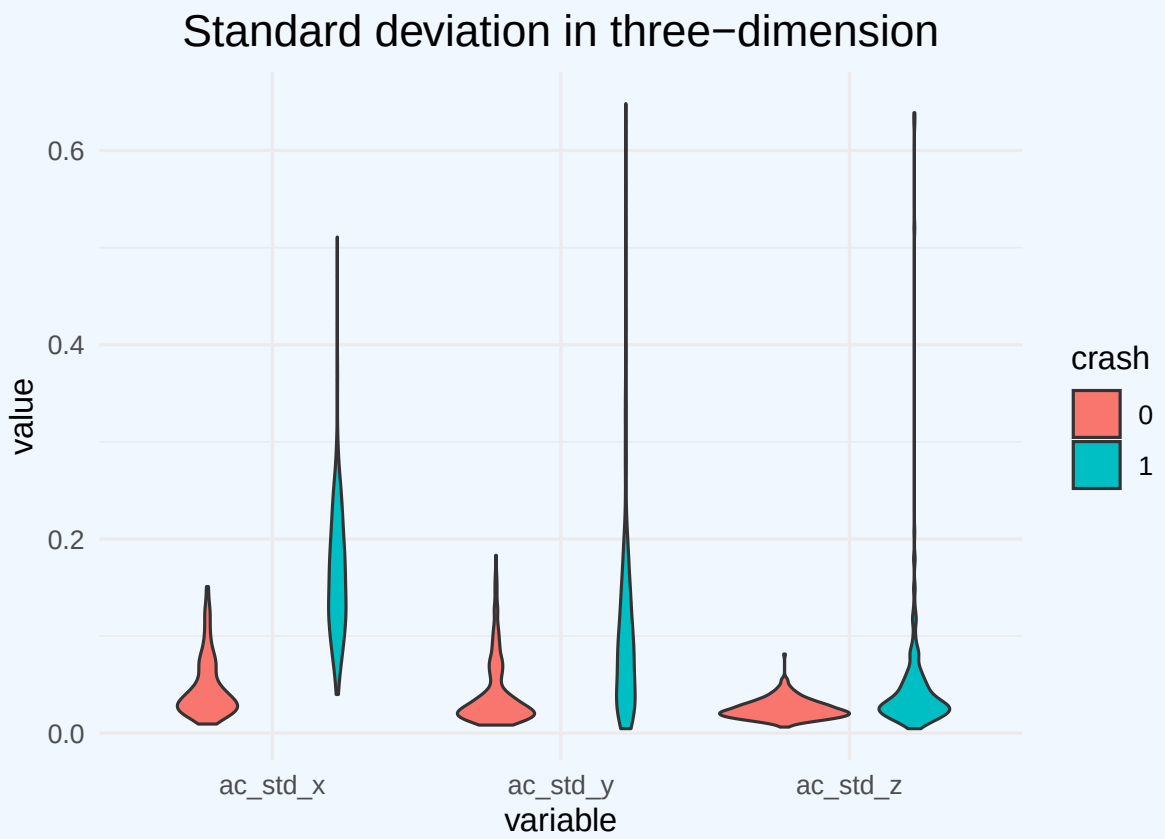
$$N_t^{1/2}\{T(\widehat{\theta}^{(t)}) - T(\theta)\} \xrightarrow{d} \text{Normal}\,(0, m\zeta_1),$$

as N_t goes to infinity.

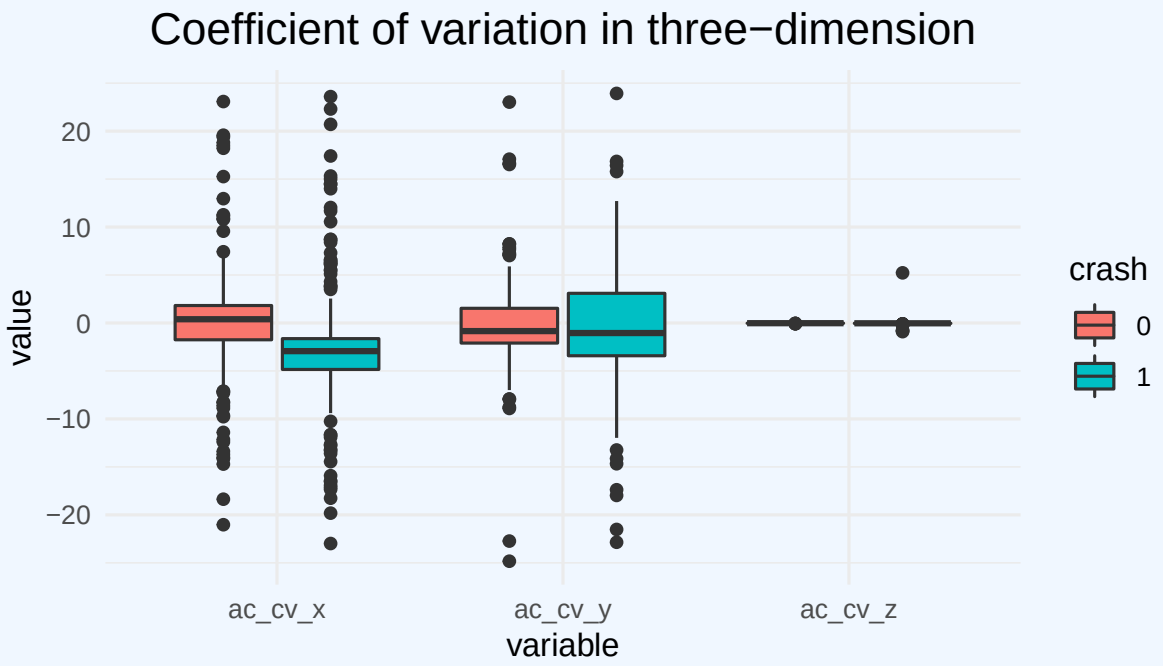
Risk prediction

Three surrogate measures

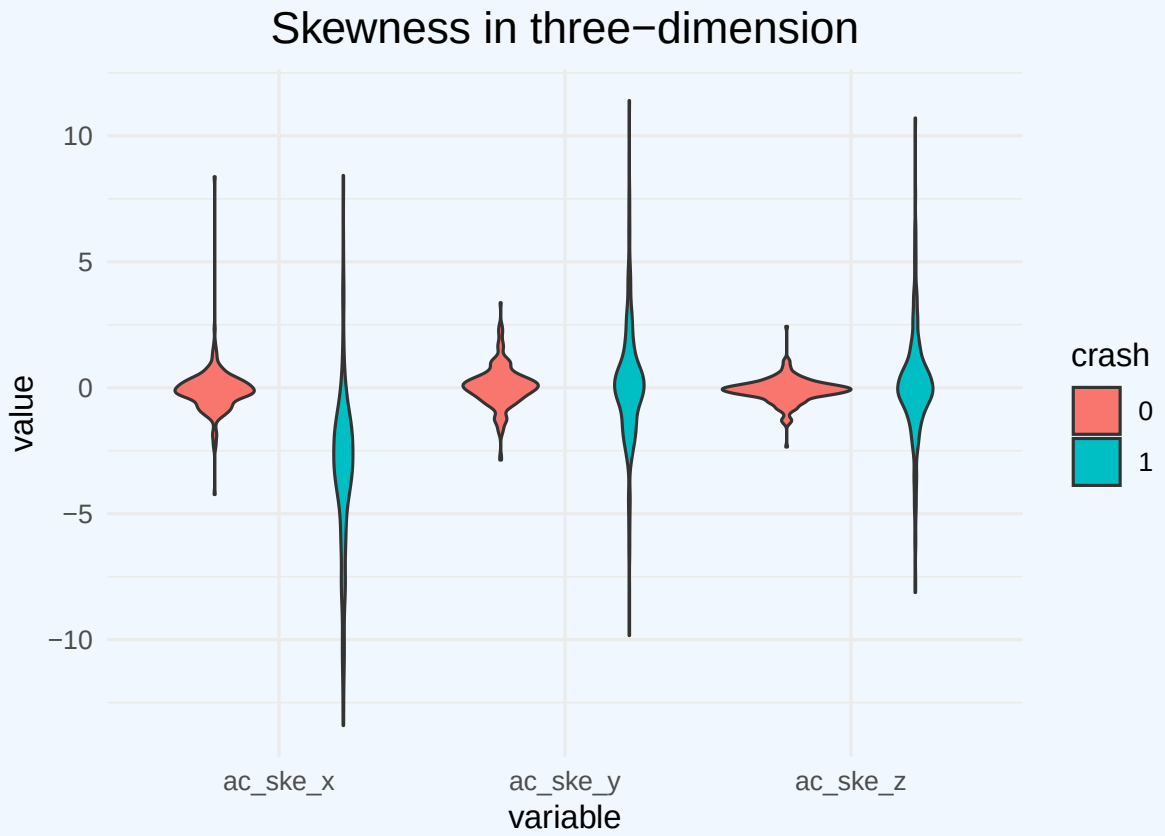
We here depict three violin of box plots for standard deviation, coefficient of variation and skewness.



Three surrogate measures



Three surrogate measures



Prediction model

We build the prediction model based on training and testing steps.

Training step: Build the benchmark model.

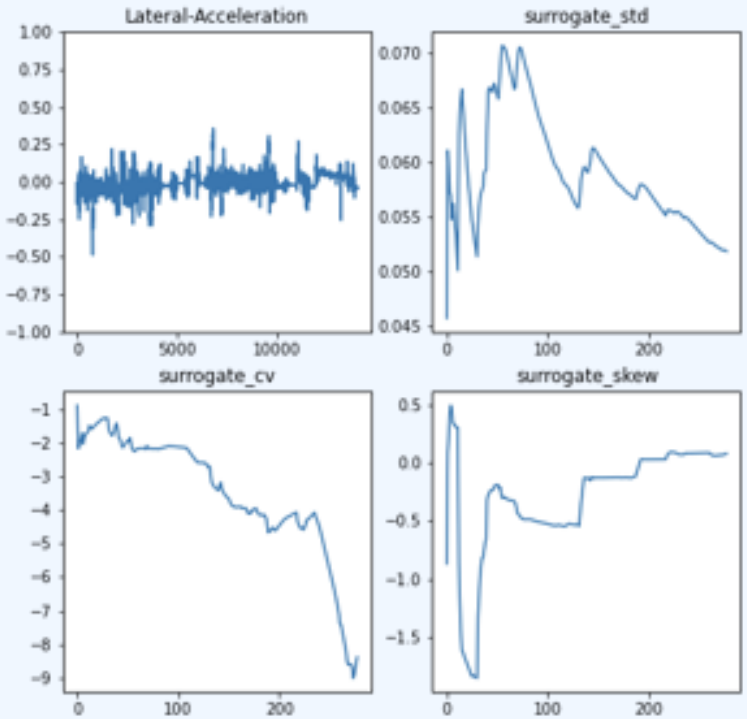
- Three surrogates with dimensions are input
- The safety benchmark model consists 61000 driving segments
- Each driving segment is about 200 time points
- 60000 baselines and 1000 crash
- Train a GBDT model as a safety benchmark model

Prediction model

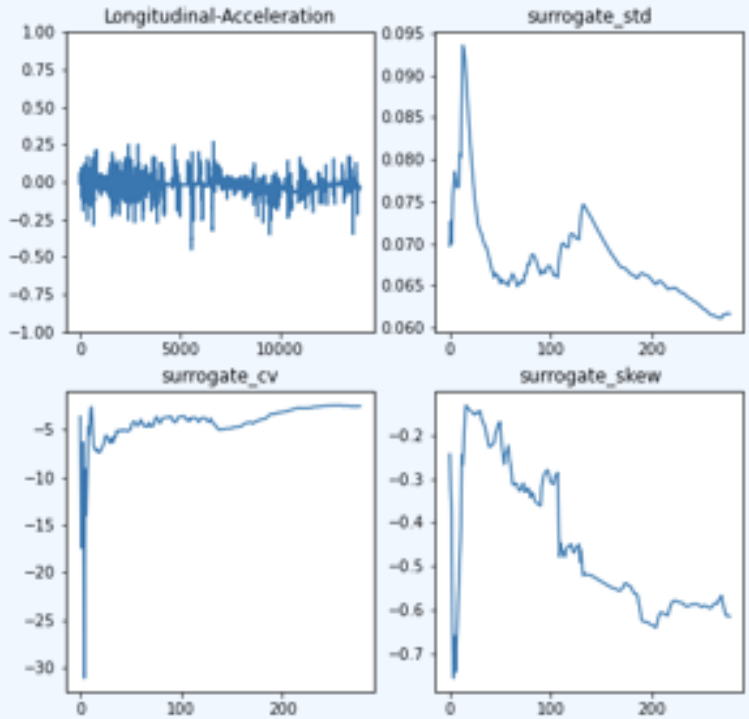
Prediction step: Predict use online estimation.

- Each driving trip is about 2000 time points, 500 no risk trip and 100 risky trips
- Initial with 200 time points and 50 time points renew each time
- Using the online estimation to extract features for every trip
- Then use the safety benchmark model to estimate timely risk

Prediction model



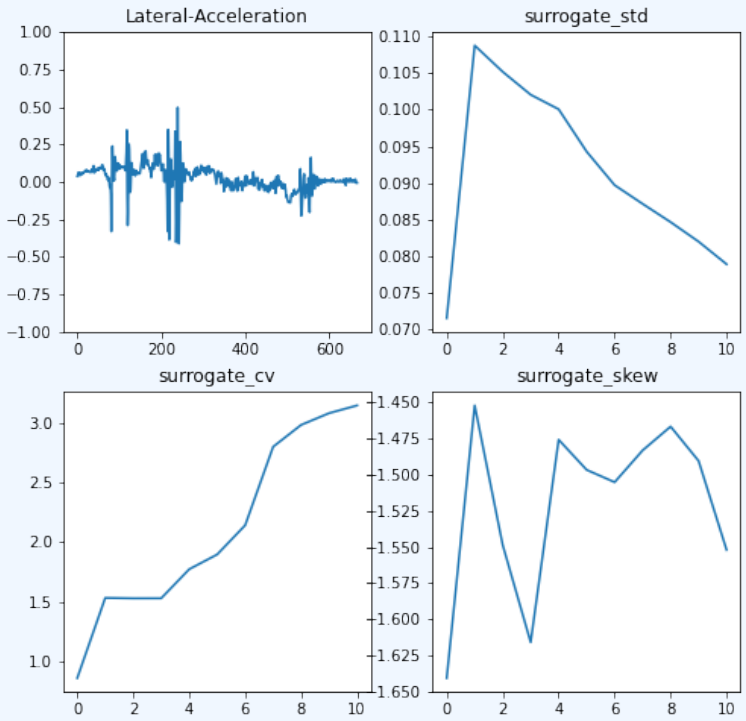
(a) Lateral



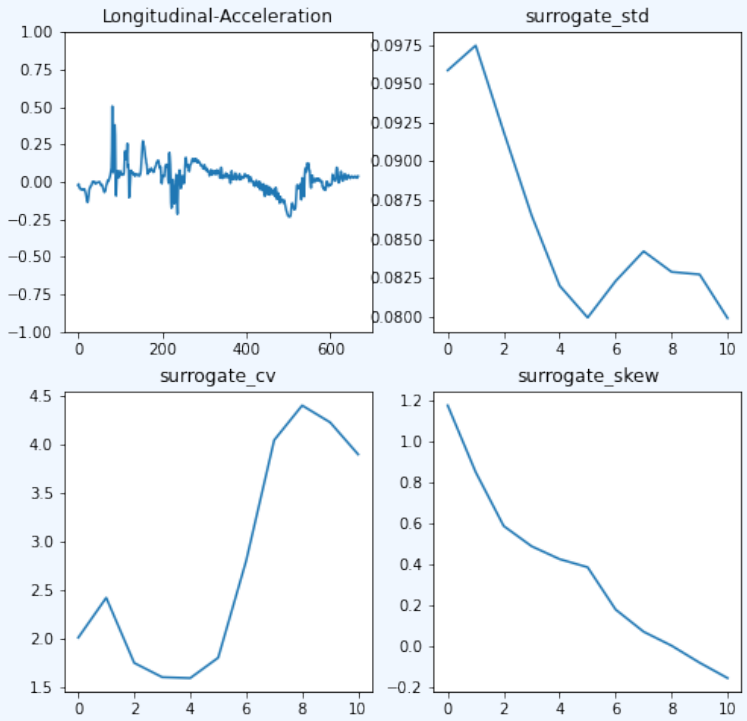
(b) Longitudinal

Figure: Online estimation for non-crash trip

Prediction model



(a) Lateral

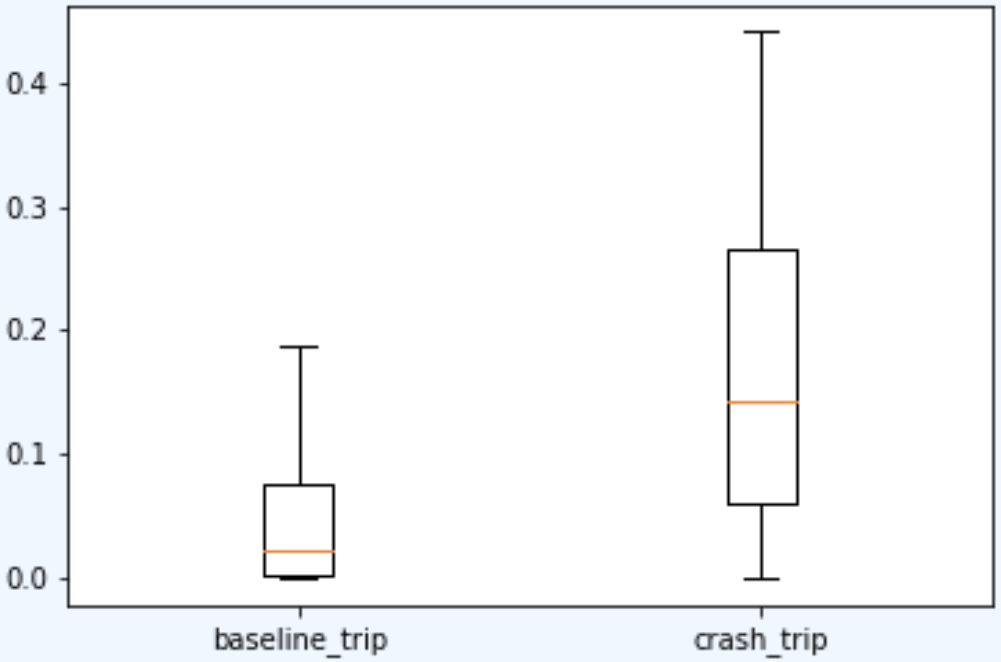


(b) Longitudinal

Figure: Online estimation for crash trip

Prediction model

Evaluation: We average online risk during a trip and depict box-plots for crash and non-crash trip.



Further work

Further work

1. More prediction models will be considered.
2. Old methods comparisons for example, maximum of acceleration
3. Robustness of model

Thanks for Listening !