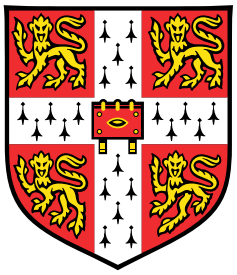
Probabilistic model for Highway Incident Detection with Real-time Traffic Data



Zhening Huang

Department of Engineering
University of Cambridge

This report is submitted for *FIBE2 CDT Mini Project*

Girton College April 2021

Probabilistic Model for Highway Incident Detection with Real-time Traffic Data

Zhening Huang

Abstract

Efficient incident detection on the Strategic Road Network (SRN) is a demanding task that necessitates the development of incident detection tools. This report proposes a framework to detect incidents on Highways solely by analysing the real-time traffic data. The suggested incident detection algorithm has the potential to reduce the detection timeframe and provide early information for the emergency response. This is particularly needed for areas like major A roads, where Highways England's intelligent detectors are less deployed. Two algorithms are proposed: one uses simple Gassuain Distribution (SGD) and detects incidents based on the continuous outlier data; the other uses Gaussian Process Regression (GPR) to predict the distribution of normal time traffic flow and implements the Generalized Likelihood Ratio Test (GLRT) to extract anomalous data based on the test statistic. Algorithm one (SGD) achieves 72% and 50% precision on the training and testing dataset respectively and algorithm two (GPR-GLRT) sets thresholds to achieve 80% precision and achieved 61% precision in the testing dataset.

Keywords— incident detection; Gaussian Process; Generalized Likelihood Ratio Test; traffic data analysis

Contents

List of Figures

1 Introduction

Incidents on highways can significantly affect road capacity and decrease the reliability of highways. The impact of disruptions can be minimized through clearing incidents rapidly and retrieving the smooth flow of traffic. Highways England oversees the Strategic Road Network (SRN), which consists of 4400 miles of motorway and major A roads.[1] Within Highways England's Incident management (IM) system, detecting incidents promptly plays a key role in the success of the emergency response step that follows[2].

Currently, incidents on highways are detected by various means including roadside technologies and reports from participants or other parties[3]. This system functions effectively in regions where Highways England intelligence is well deployed but there is significant room for improvement in regions where these technologies are less equipped. This project proposes a tool for incident detection, which solely utilizes data from traffic counters and could therefore provide an alternative solution in regions where sensors and cameras are limited. The occurrence of incidents on the network typically causes delay and reduction in traffic volume and therefore, it can be reflected in real-time traffic data. Based on this assumption, two algorithms that use statistical approaches to extract the average traffic flow parameters and identify anomalous data are developed in this project.

1.1 Aims and objectives

The aim of this project includes:

1. Reviewing current incident detection methods used on the SRN and identifying the benchmark parameters for the development of new incident detection algorithms.
2. Developing a probabilistic model that can effectively utilize traffic count data to extract incident information.
3. Assessing the performance of the proposed algorithm by using historic traffic data and incident records in a geographic area of the SRN, provided by Highways England.
4. Suggesting how the proposed algorithm can be integrated into the existing traffic management system and discussing the opportunities for further improvement.

1.2 Review of current detection techniques on the SRN

Currently, incidents are detected by various means including roadside technology, reports from the police in the form of 999 calls, calls directly to Highways England and reports by service providers as well as vehicle recovery organizations[3]. Depending on the type of incident, other stakeholders such as emergency services and Local Highway Authorities may be involved in incident detection and response. Motorways and major A roads have different monitoring and detecting approaches. Overall, traffic officer (TO) virtual patrolling supports early detection of disruptions in both types of networks and is the key in finding and responding to routine incidents and preventing escalation. Smart motorways and Stop Vehicle Detection (SVD), which are currently deployed in part of the network, can further improve the efficiency of monitoring and detection on motorways[2]. However, compared to motorways, major A roads have fewer patrolling systems and rely more on traffic flow technology and incidents reported externally, which may cause a delay in the responses of emergency circumstances [3]. The delayed response of incidents that impact running lanes causes significant reputational issues for Highways England.

The current incident management Key Performance Indicator (KPI) requires Highways England to clear at

least 85% of incidents on motorway networks within one hour.[2] The lane impacted duration is calculated from the time that the Highways England Control Centre operator operates the 'lane impacted' button in Control Works [2]. **However, there is no measuring matrix for the incident detection time, the period from the occurrence of incidents to when the 'lane impacted' button was pressed.** This is because the real occurrence time is usually vague and hard to track. As per the discussion with Highways England traffic management experts, there is a consensus among highway operators that **detecting incidents within a 15 to 30 minutes time frame after the actual occurrence of the incident could be an improvement for the current system.** Therefore "15 to 30 minutes" will be a benchmark target for the proposed algorithm.

2 Algorithm Framework

Two algorithms are developed in this project. Both algorithms contains two steps: the first step is detecting the anomalous data points from the data pool; the second step is setting thresholds to filter data and generate a potential incident list.

2.1 Simple Gaussian distribution based anomalies detection

Anomaly Detection

This is a simple traffic pattern extraction algorithm. The traffic volume distribution $y_i(x)$ at each time point is considered as an independent Gaussian.

$$y_i(x) \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

By analysing the traffic count data for for the period of a year, the mean and variance for the Gaussian distribution can be obtained and the 95% confidence interval for the Gaussian distribution is:

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} < y(x) < \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

A simple anomaly detection can be done by examining if $y_i(x)$ is located within the confidence interval. Since a traffic incident typically only reduces the traffic volume, only the lower boundary is considered, i.e.

$$y(x)_{anomaly} < \mu - 1.96\frac{\sigma}{\sqrt{n}}$$

Anomaly Filtering

Once all the anomalies are extracted, filtering can be carried out by setting a threshed n and **anomaly points can only be identified as potential traffic incidents, if the the n subsequent points are all detected as anomalies.** The value of n can be determined by trading off between associated precision and recall.

2.2 Gaussian Process based Generalized Likelihood Ratio Test

Gaussian Process

Gaussian Process is a generalization of a multivariate Gaussian distribution to infinitely many variables[4]. Here the traffic volume data can be formulated into a typical regression problem and Gaussian Process Regression can be used to extract traffic patterns.

A Gaussian Process is fully specified by a mean function $m(x)$ and covariance function $k(x, x')$):

$$m(x) = E[f(x)] \qquad k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))$$

$$f(x_i) \sim \mathcal{GP}(m(x), k(x, x'))$$

To apply Gaussian Process models to traffic data, we need to take into account the noise on the recorded traffic volume, which is given by

$$y(x_i) = f(x_i) + \epsilon_i$$

where ϵ_i is a random noise variable which is considered to have a Gaussian distribution, such that

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \qquad p(y_i|x_i, f_i) \sim \mathcal{N}(f_i, \sigma_\epsilon^2)$$

Likelihood Given N observations $X_1, X_2, ..., X_N$, the joint distribution of the traffic volume is also a Gaussian. The probability density of the Gaussian samples follows a Gaussian Distribution. Thus the likelihood function can be written as:

$$\begin{aligned} p(Y|X, F) &= \prod_{i=1}^{N} p(y_i|x_i, f_i) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp(-\frac{(y_i - f(i))^2}{2\sigma_\epsilon^2}) \\ &= (\frac{1}{\sqrt{2\pi}\sigma_\epsilon})^N \exp(-\frac{\|Y - F\|^2}{2\sigma_\epsilon^2}) \\ &= \mathcal{N}(F, \sigma_\epsilon^2 I_N) \end{aligned}$$

where I_N is the identity matrix.

Prior In the Bayesian formalism, we need to specify a prior over the function F, expressing our belief about the function before we look at the observations[5]. We put a zero mean Gaussian Prior with covariance function on the function F, such that

$$F \sim \mathcal{N}(0, K(x, x'))$$

Bayes' Rule The posterior distribution of traffic volume is determined by using Bayes' rules as follows[6]:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} \qquad p(F|X, Y) = \frac{p(Y|X, F)p(F)}{p(Y|X)}$$

Therefore the posterior distribution is:

$$p(F|X, Y) \propto p(Y|X, F)p(F) = \mathcal{GP}(m_{post}, k_{post})$$

where

$$m_{post}(x) = k(x, x)[K[x, x] + \sigma_{noise}^2 I]^{-1} Y$$

$$k_{post}(x, x') = k(x, x') - k(x, x') - k(x, x)[K(xx) + \sigma_{noise}^2 I]^{-1} k(x, x')$$

For a new observation X^* with corresponding output Y^*, the joint distribution according to the prior is:

$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X)+\sigma_n^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}\right)$$

If there are n training points and n^* test points, then $K(X, X^*)$ denotes the n^* matrix of the covariance evaluated at all pairs of training and test points, and similarly for the other entries $K(X,X), K(X^*,X^*)$and $K(X^*,X)$. The conditional distribution of Y* given Y is:

$$Y_*|X_*, X, Y, F \sim \mathcal{N}\left(\begin{array}{c} (K(X^*,X)[K(X,X)+\sigma_{noise}^2 I]^{-1}Y, \\ K(X^*,X^*)-K(X^*,X)[K(X,X)+\sigma_{noise}^2 I]^{-1}K(X^*,X) \end{array}\right)$$

Covariance Function In this problem, The Squared Exponential(SE) is used as the covariance function, and the formula is listed as follows:

$$k(x,x') = \exp\left(-\frac{1}{2}(\frac{x-x'}{l})^2\right)$$

Parameter l defines the characteristic length scale. This covariance function is infinitely differentiable, which means that the GP with this covariance function has mean square derivatives of all orders, and is thus very smooth[4]. The Square exponential function is the most widely used covariance function in the field[5].

Marginal Likelihood and Hyperparameters Selection The marginal likelihood refers to the likelihood of the data, where the functions values have been marginalised out, shown here for the training points X, data vector Y and function F:

$$p(Y|X) = \int p(Y|X,F)p(F)dF$$

so log marginal likelihood can be obtained analytically:

$$\log p(Y|X) = -\frac{1}{2}Y^T[K+\sigma_n^2 I_N]^{-1}Y - \frac{1}{2}\log|K+\sigma_n^2 I_N| - \frac{N}{2}\log(2\pi)$$

Typically, hyperparameters, (here are $\sigma_n$and l) are set by maximising the log marginal likelihood[6]. The motivation behind using the marginal likelihood is the automatic trade off between the data fit and the complexity of the model that is achieved when maximising the marginal likelihood. Once the hyperparameters are set and the predictive distribution is obtained, the residual noise vector can be calculated as

$$E = Y - \hat{Y}$$

where $\hat{Y}$ is the predictive traffic volume from the GPR. The anomaly detection can be done by using hypothesis testing, particularly generalized likelihood ratio testing in this case.

Hypothesis Testing

Data Cleaning and Pre-processing The available dataset is split into training data and testing data. We initially will pre-process the training data to remove all the anomalous traffic records. This is done by deleting traffic data recorded around logged incident events. The cleaned training data is then considered with no anomalies present and can represent the normal behaviour. This data is used to form a "normal model" where the input X_{train} is the time steps in the day and the input Y_{train} are the traffic volume records. The GPR is then implemented to obtain a residual noise sample predictive Y and E_{normal}.Test data can then be compared

to the training model to test for anomalies. The input X_{test} and Y_{test} are trained with GPR and a residual noise vector for testing data is obtained $E_{testing}$

At each time step, a test model is constructed from a sliding window of w previous data points and the number of testing point depends on the nature of the anomalies being detected. The generalized likelihood ratio test is used here to form the hypothesis testing, as described below.

Generalized Likelihood Ratio Test The GLRT is known as a powerful test among all invariant tests.It is a hypothesis testing technique which has been utilized successfully in model-based fault detection[7]. Focusing on the traffic volume problem, $E \in \mathcal{R}_N$ is the residual noise so should be formed by a Gaussian distribution with the mean as 0 and the variance as $\sigma_2 I_N$. This is then set a null hypothesis. The alternative hypothesis is that E is taken from a Gaussian distribution with nonzero mean θ, i.e.$N \sim (\theta, \sigma_2 I_N)$, and the unknown parameter θ is obtained for the maximum likelihood estimation. The hypothesis test can be expressed as

$$\mathcal{H}_0 : \{E \sim \mathcal{N}(0, \sigma^2 I_N)\}$$

$$\mathcal{H}_1 : \{E \sim \mathcal{N}(\theta, \sigma^2 I_N)\}$$

while the likelihood for the sliding window of both hypothesis can be expressed as:

$$lik(\mathcal{H}_0) = \frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} \exp\{-\frac{1}{2} E^T \Sigma^{-1} E\}$$

$$lik(\mathcal{H}_1) = \frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} \exp\{-\frac{1}{2} (E-\theta)^T \Sigma^{-1} (E-\theta)\}$$

The GLRT statistic compares the likelihood of both hypothesis, and the log likelihood ratio is calculated as [8]:

$$\begin{aligned}
T(E) &= 2 \log \frac{\max_{\theta \in \mathcal{R}} lik(\mathcal{H}_1)}{lik(\mathcal{H}_0)} \\
&= 2 \log \Big\{ \max_{\theta} \exp\{-\frac{1}{2}(E-\theta)^T \Sigma^{-1} (E-\theta)\} / \exp\{-\frac{1}{2}(E)^T \Sigma^{-1} (E)\} \Big\} \\
&= \frac{1}{\Sigma} \Big\{ \min_{\theta} (E-\theta)^T \Sigma^{-1} (E-\theta) + E^T \Sigma^{-1} E \Big\} \\
&= \frac{1}{\Sigma} \Big\{ (E-\hat{\theta})^T \Sigma^{-1} (E-\hat{\theta}) + E^T \Sigma^{-1} E \Big\} \\
&= \frac{1}{\Sigma} \Big\{ E^T \Sigma^{-1} E \Big\}
\end{aligned}$$

where $\hat{\theta} = \arg\min_{\theta} (E-\theta)^T \Sigma^{-1} (E-\theta) = E$ is the maximum likelihood estimate of θ. In the deviation above, maximising the likelihood function is equivalent to maximizing its natural logarithm since the logarithmic function is a monotonic function. The GLRT then decides between hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ as follows,

$$\delta(E) = \begin{cases} \mathcal{H}_0 & \text{if } T(E) < t_\alpha \\ \mathcal{H}_1 & else \end{cases}$$

Threshold setting Now the statistical testing framework is set up, this decision threshold must be set in a way to produce desirable performance. To select an appropriate threshold for the test statistics shown above, it is crucial to find their distributions. In both hypothesises, E follows N-dimensional Gaussian Distribution $X \sim \mathcal{N}(0, \Sigma)$, and therefore the test statistic follows the generalized chi-squared distribution χ^2, i.e.

$$T(Y) = \frac{1}{\Sigma} \Big\{ E^T \Sigma^{-1} E \Big\} \sim \chi^2$$

Typically setting the false alarm rate of a classifier is desirable. The Cumulative distribution function (CDF) of this test statistic could be used to give a desired significance level (i.e. false positive rate under the null hypothesis). Inspired by Glyn-Davies work[9], the threshold t_α can be set according to CDF F to achieve a desired false positive rate α:

$$\alpha = 1 - F(t_\alpha)$$

Here the CDF can be approximated with the empirical statistic of the training data. The overall GPR-based GLRT fault detection algorithm is shown in the flowchart (Figure 1).

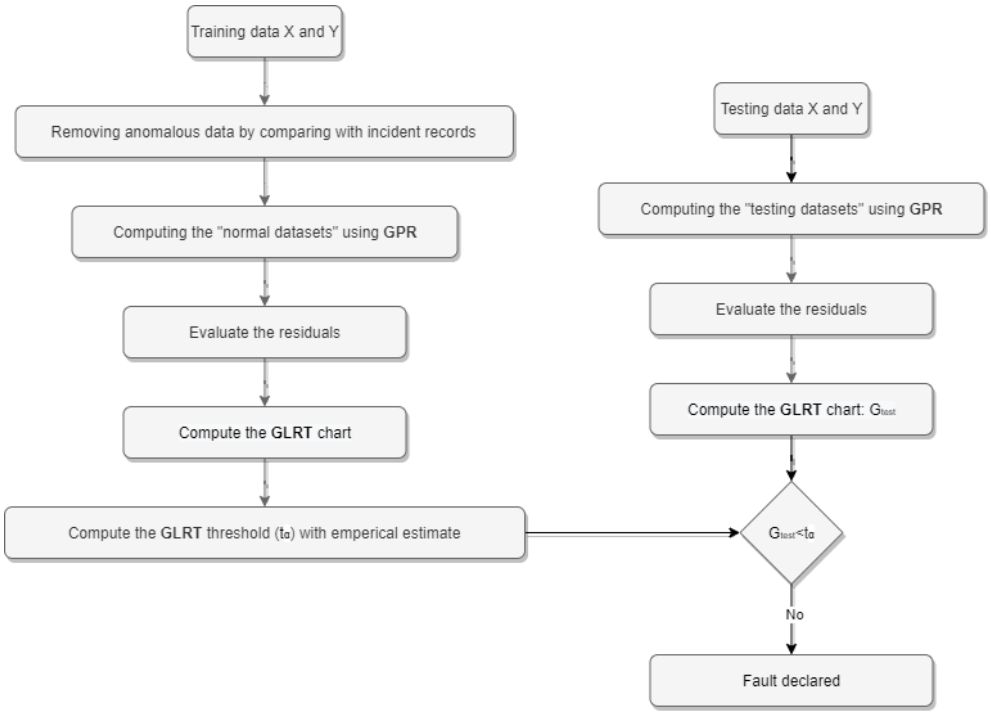


Figure 1: GPR-based GLRT fault detection scheme

3 Experimental Results

3.1 Data Processing

The road of interest (Figure 2) is Northbound section of M1 motorway, from Junction 6a to 10, with a total length of around 9 miles. The main reason for this particular choice is that this is a straight motorway without many sharp curves, junctions and roundabouts. Hence, it is assumed that disruptions that occur on any part of the road affect the whole section and be captured by different traffic counters deployed ((Figure 2).

In terms of data inputs, traffic count data is available on the MIDAS database, which can be accessed from the Halogen system, and incident records can be extracted from Highways England internal incident logs. Traffic count data contains the type (categorised by vehicle length) and the number of vehicles passing through the traffic counter on each lane for every minute and the dataset is available for the last three years. The extracted incident records document all incident events reported on the study road, including times, dates, locations and rough descriptions of incident types. The timeframe for the available incident records is from 1st January 2019 to 1st January 2020.

Raw traffic count datasets contain some null data points and these values are replaced by the average traffic

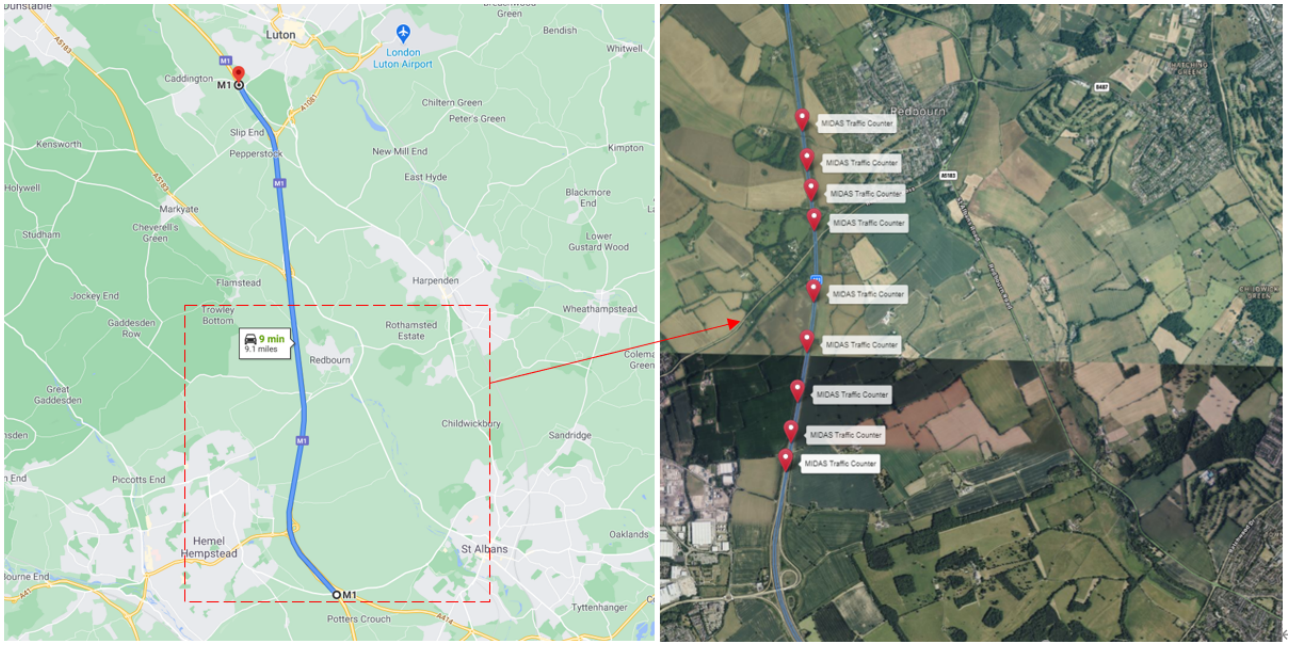


Figure 2: Road of interest (Left) and Locations of traffic counter along the road (Right)

volume at the same time on the other days. To simplify the input data, traffic volumes on different lanes are summed together to form a total road traffic volume, which is used in the later analysis. Moreover, the one-minute traffic count data exhibits large variances on different days, increasing the difficulty in pattern extraction and further analysis. Therefore, the data was condensed to multiple-minute interval volume, which shows less variance and is more tractable in further steps. Figure 3 shows the 15-minute interval traffic volume for the A2 motorway during the first 10 weeks of 2020 (prior to the national lockdown). As shown in Figure 3, there is a clear traffic pattern for each day of the week. On weekdays, traffic volumes are significantly higher than those of the weekend while Saturday and Sunday also have different shapes of traffic flows. As a result, the traffic flow on different days of the week needs to be studied separately. Moreover, traffic flow on special days in the year, e.g. Bank holidays, is expected to be distinct from that of normal days, and will be separated out at this stage.

The Further demonstration of the proposed algorithm focuses on traffic data on Tuesday in 2019 (Figure 4a). The available dataset is split into a training set and a testing set. The testing set is extracted evenly throughout the year to avoid bias. The following shows how the dataset is split:

- Training data: datasets in January, February, April, May, July, August, October, November
- Testing data: datasets in March, June, September, December

3.2 Simple Gaussian Distribution

Figure 4b shows the 95% confidence interval of training data by considering traffic volume at each timestep as an independent Gaussian distribution. A large number of data points are located outside the confidence interval zone (Figure 5a). When the threshold is set to 4, i.e. only when 4 continuous points are below the confidence zone, the last point is detected as a potential incident event, and the model has the highest precision on training data. There are 46 remaining points after filtering (Figure 5b), of which 33 have matches in real incident records, giving a precision rate of 72%. The recall is not considered here as the incident records contain more than 1000 incidents on Tuesday without proper indication of their severity and categories. The testing dataset is examined

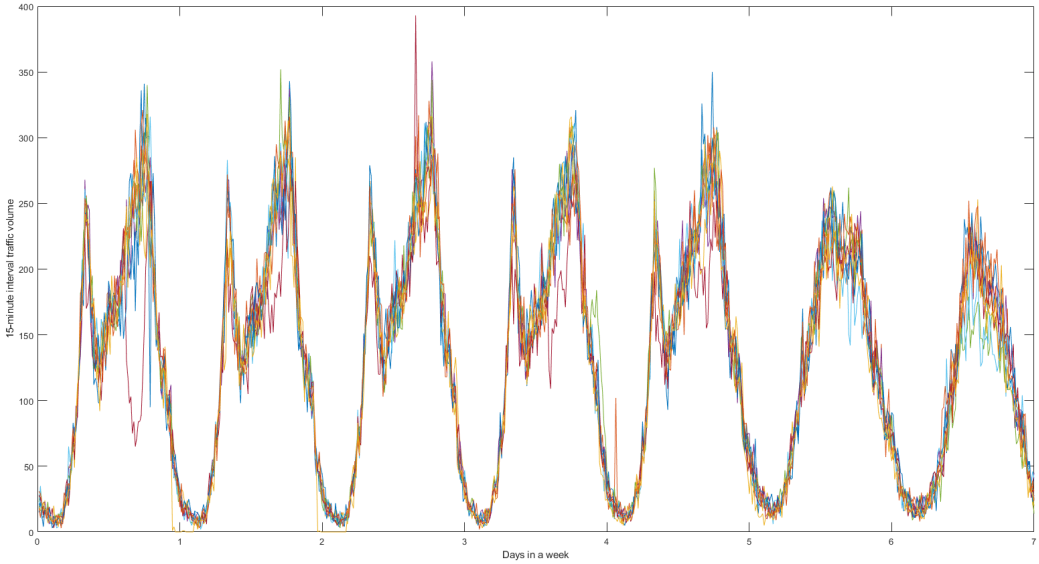
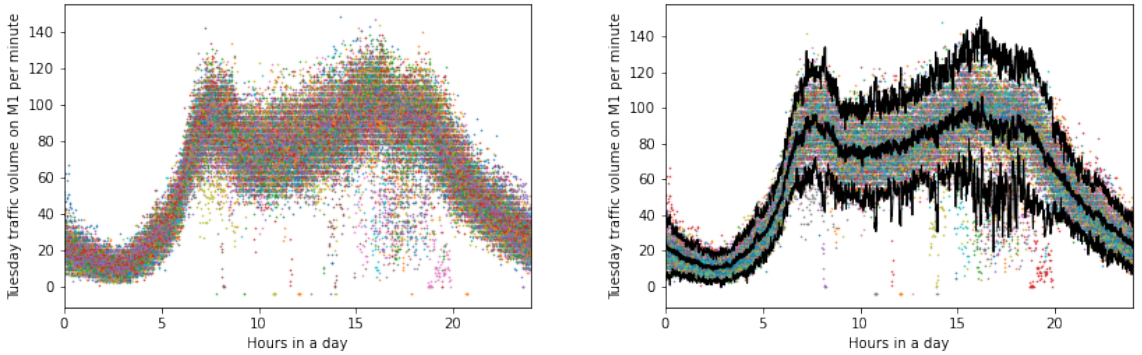
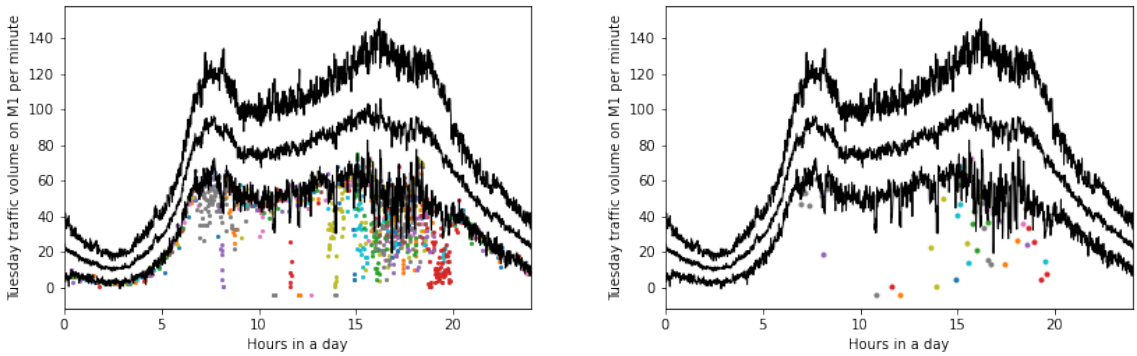


Figure 3: 15-minute interval traffic volume on A2 (before the National Lockdown



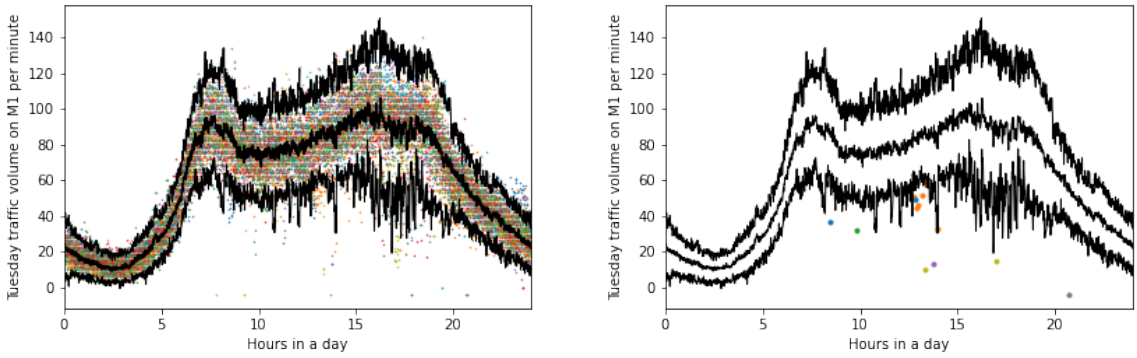
(a) Data for entire year (b) 95% confidence interval for training dataset

Figure 4: SGD-Training: Tuesday traffic volume on M1 per minute



(a) Detected anomalous from 95% confidence interval (b) Filtered anomalies (threshold: 4)

Figure 5: SGD-Training: Anomalous Traffic data



(a) Testing dataset raw data (b) Filtered anomalies (threshold: 4)

Figure 6: SGD-Testing: Testing dataset results

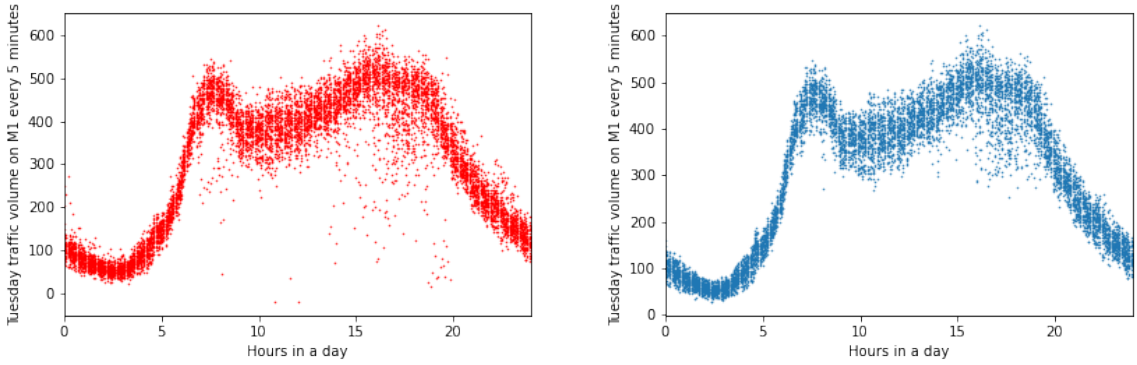
in this model(Figure 6a), which yields a precision rate of 50% (Figure 6b). The average incident detection time, which is calculated by averaging the time difference between recorded incidents and detected incidents, is 19.4 minutes.

3.3 GPR-based GLRT

To reduce the computational cost, the traffic data is condensed to 5-minute traffic volumes. Ideally, before implementing the training dataset on the GRP, all data points captured during the incident period need to be removed to create a "fault-free" dataset. Due to imperfect incident records and time constraints of the project, a simplified approach is used here: the raw traffic data is cleaned by filtering out data points that are outside of the 95% confidence interval, similar to the simple GD method above. The training dataset before and after the cleaning process is shown in Figure 10 below.

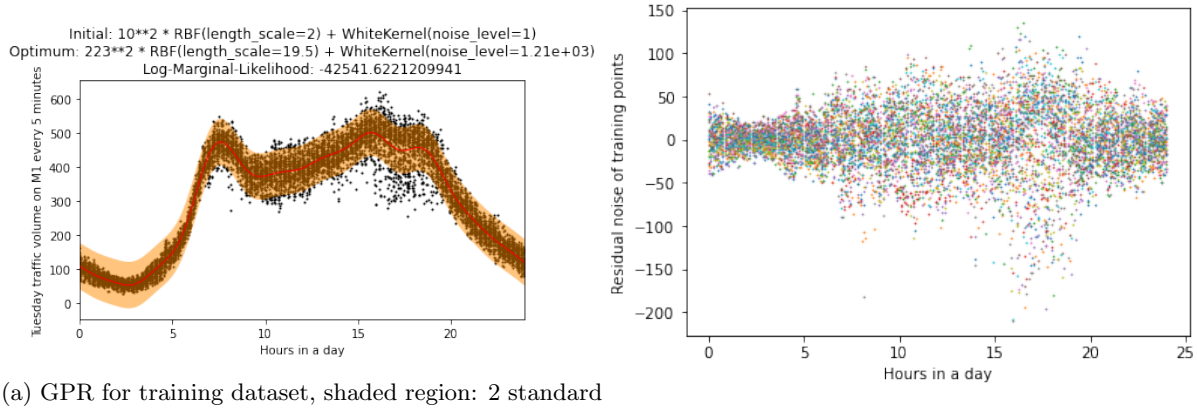
After implementing the GPR on the cleaned training dataset, the prediction of the normal traffic pattern on Tuesday is yielded (Figure 8a), and the residual noise of the training dataset is also calculated (Figure 8b). The GLRT statistic with a sliding window of 1 is computed for all the training points (Figure 9b), and the empirical PDF is shown in Figure 9a. The CDF of the GLRT statistic indicates that $t_\alpha = 1.42$ when the false alarm rate is set as 20%. The threshold of the GLRT statistic is plotted on Figure 9b.

The testing dataset is also trained with GPR. Figure 10a and Figure 10b show the prediction and residual noise.



(a) raw training dataset (b) "fault-free" assumed dataset

Figure 7: GPR-GLRT-Training: Five-minute interval traffic volume on M1



(a) GPR for training dataset, shaded region: 2 standard deviations

(b) Residual noises distribution for all training data

Figure 8: GPR-GLRT-Training: GPR of training data and the residual noise distribution

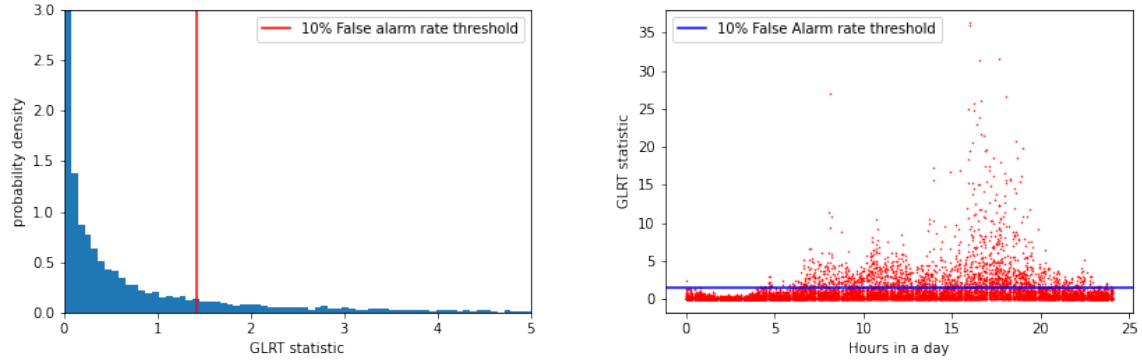
The GLRT for the testing dataset is carried out also with a sliding window of 1 (Figure 11). By extracting the information of detected anomalies and comparing it with incident records, a precision of 61% is yielded.

4 Conclusion and Recommendation

Two proposed algorithms have demonstrated their function in incident detection. The GRP-GLRT model achieves a better result as it considers the level of an anomaly for each data point while the SGD model only conducts a simple binary detection. The SGD model however has a much lower computational cost, making it an ideal algorithm for preliminary anomaly classification. Due to the time limit of the project, the sliding window is only set as 1 in the GRP-GLRT model. With a larger sliding window, the result is expected to be improved as it considers the covariance of different data points in the hypothesis test. Overall, the GRP-GLRT is a promising algorithm for incident detection on highways, although a larger dataset at other locations on the SRNs needs to be tested to evaluate the performance.

The nature of the project is only to prove the concept of the algorithm and verify its preliminary effectiveness, so many steps and techniques in the model are simplified. To further enhance the algorithm and prove its function on the entire network, the followings steps can be carried out in the future:

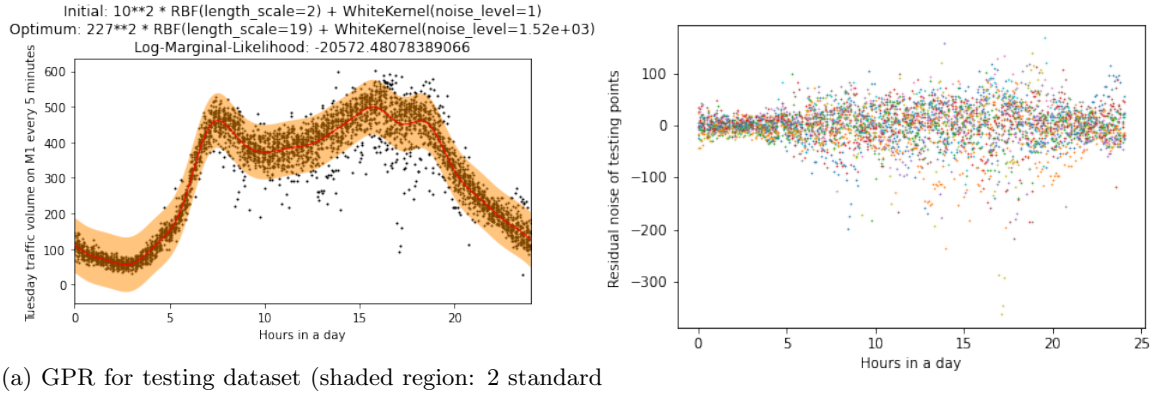
- Using a larger dataset over a wider study area (i.e. More than one roads) to verify the algorithm perfor-



(a) Empirical PDF based on training samples

(b) GLRT Statistic of residual noise

Figure 9: GPR-GLRT-Training: Hypothesis testing



(a) GPR for testing dataset (shaded region: 2 standard deviations)

(b) Residual noises of Testing points

Figure 10: GPR-GLRT-Testing: GPR and residual noise distribution

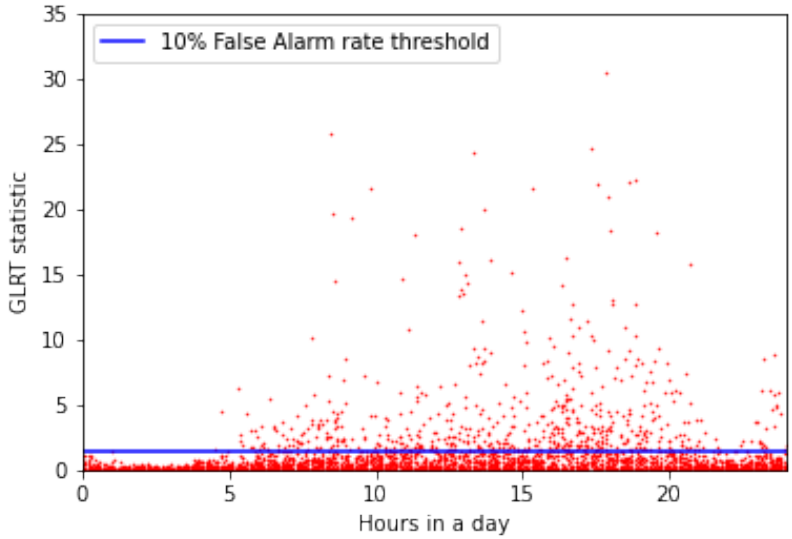


Figure 11: GPR-GLRT-Testing: GLRT Statistic of residual noise

mance

- Integrating the detection results from different traffic counters on the same road to improve detection performance
- Using the traffic counter data on different lanes to capture more details about the traffic changes
- Using a more detailed incident dataset with a different severity level of incident records. This will be useful to identify with which level of severity, incidents can effectively be detected from the algorithm.
- Sub-sampling the dataset during different seasons to extract more subtle differences between the traffic patterns at a different time of the year.
- Utilizing the average speed parameter on the traffic counter datasets to enrich features of training datasets
- Increasing the size of the sliding window to consider covariance between each data point.
- The empirical PDF estimation can be improved with available bootstrapping and data augmentation techniques.
- Fezai et al. [7] introduced an online-reduced GPR which has the advantages of improving the computational efficiency by decreasing the dimension of the kernel matrix. This could be implemented in the model to reduce the computational cost.

Regarding how to integrate the developed model into Highways England's current traffic management system, the following suggestions are proposed:

- The traffic count data is currently stored in the MIDAS system with a standardized format, making it easy to be processed and analysed with an external algorithm. As a result, it is feasible to develop an API that can integrate the traffic count data into the developed algorithm at the entire network scale. Operating an API like this may require significant computational power, so optimizing the algorithm at the local level is the essential starting point. If the false alarm rates can be reduced to 20% and the detection time can be shortened to around 15 minutes, investment in this incident detection system could prove worthwhile.
- The traffic pattern on the Strategic Road Network changes over time and therefore it is crucial to update the traffic pattern accordingly to maintain its effectiveness. This could be achieved by retraining the model periodically with the previous N month's traffic data. Last year, under the impact of the pandemic, the traffic pattern underwent a dramatic change, providing a good database for developing and testing such an adapting technique.
- Current incident detection records do not have a clear severity indicator and the excessive uncategorized incident records make the recall calculation (True positive/True negative+ False negative) hard and meaningless. Only on Tuesday, more than 15,000 incident events were logged in the records and many of them are too slight to be reflected on the traffic data. In future, developing a comprehensive incident records system that categorises the incidents by severity can significantly benefit the development of incident detection algorithms.
- The deployment of connected vehicles and the intelligent network in the future will generate more accurate and detailed data for traffic volume and speed. The proposed algorithm applies to different types of data and has the potential to achieve higher performance with a better dataset. Therefore, future data generated from new highway intelligence can further enhance the algorithm.

References

[1] Highways England. Highways England Strategic Road Network Initial Report. 2017.

[2] Transport Research Laboratory. Framework for transport-related technical and engineering advice and Lot 2 : Road Transport Package Order 687 (4 / 45 / 12) - Algorithm Improvement. 687, 2016.

[3] Elliott Asset Management. Highways England and Incident Management Study. 2018.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation And Machine Learning. MIT Press, 2005.

[6] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 04-2011 edition, 2011. In press.

[7] R. Fezai, M. Mansouri, K. Abodayeh, H. Nounou, and M. Nounou. Online reduced gaussian process regression based generalized likelihood ratio test for fault detection. *Journal of Process Control*, 85:30–40, 2020.

[8] Chiranjivi Botre, Majdi Mansouri, Mohamed Nounou, Hazem Nounou, and M. Nazmul Karim. Kernel pls-based glrt method for fault detection of chemical processes. *Journal of Loss Prevention in the Process Industries*, 43:212–224, 2016.

[9] Alex Glyn-Davies. Anomaly detection in streaming data with gaussian process based stochastic differential equations. 2021.

Appendix: GPR-GLRT codebase

Listing 1: GPR-GLRT codebase

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
from matplotlib.colors import LogNorm
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import RBF, WhiteKernel

'''import traffic count data'''

# splite data to training and testing test

'''data condensing '''
def datacondensing(orginal,compactingfactor):
    mergeddata=np.sum(orginal[:,compactingfactor*(i):compactingfactor*(i+1)], axis=1)
    if a%compactingfactor==0:
        for i in range (mergedsize):
            mergeddata=np.sum(orginal[:,compactingfactor*(i):compactingfactor*(i+1)], axis=1)
            if newdataset is None:
                newdataset=mergeddata
```

```
            else:
                newdataset=np.vstack((newdataset,mergeddata))
        return (newdataset.transpose())
'''clean training data by reomving anomalous points'''

# GPR
plt.figure()
kernel = 100 * RBF(length_scale=2, length_scale_bounds=(1, 1e3)) \
    + WhiteKernel(noise_level=1, noise_level_bounds=(1e-10, 1e+5))
gp = GaussianProcessRegressor(kernel=kernel,
                              alpha=0.0).fit(X_training, y_training)

X_pre = np.linspace(0, int(1440/compactingfactor), int(1440/compactingfactor))

y_gpr, y_std = gp.predict(X_pre[:, np.newaxis], return_std=True)

#find the residual
for i in range (len(sum_flow_yearly)):
    residual[i,:]=[i,:]-y_gpr
#GLRT
for i in range (len(residual)):
    GLRT[i,:]=y_std**-2*residual[i,:]**2
#PDF
fig = plt.figure()
ax = fig.add_subplot(111)
n, bins, rectangles = ax.hist(GLRT, 500,density=True)
fig.canvas.draw()
#CDF
n_bins = 50
fig, ax = plt.subplots(figsize=(8, 4))
n, bins, patches = ax.hist(GLRT, n_bins, density=True, histtype='step',
                           cumulative=True, label='Empirical')
```

Full Codebase of this project is available at:
https://github.com/ZheningHuang/Incident-detection-with-traffic-data