Traffic Incident Detection Using Probabilistic Topic Model

Akira Kinoshita
The University of Tokyo
2-1-2 Hitotsubashi, Chiyoda,
Tokyo, Japan
kinoshita@nii.ac.jp

Atsuhiro Takasu, Jun Adachi
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda,
Tokyo, Japan
{takasu,adachi}@nii.ac.jp

ABSTRACT

Traffic congestion is quite common in urban settings, and is not always caused by traffic incidents. In this paper, we propose a simple method for detecting traffic incidents by using probe-car data to compare usual and current traffic states, thereby distinguishing incidents from spontaneous congestion. First, we introduce a traffic state model based on a probabilistic topic model to describe traffic states for a variety of roads, deriving formulas for estimating the model parameters from observed data using an expectation–maximization algorithm. Next, we propose an incident detection method based on our model, which issues an alert when a car's behavior is sufficiently different from usual. We conducted an experiment with data collected on the Shuto Expressway in Tokyo over the 2011 calendar year. The results showed that our method discriminates successfully between anomalous car trajectories and the more usual, slowly moving traffic. However, our method does sometimes classify abnormally fast-moving cars as traffic incidents.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining, Spatial databases and GIS*

General Terms

Algorithms

Keywords

Anomaly detection, automatic incident detection, probabilistic topic model, probe-car data, traffic state estimation

1. INTRODUCTION

Automatic incident detection (AID) is a crucial technology in intelligent transport systems, particularly in terms of reducing congestion on freeways [10]. Traffic incidents often cause traffic congestion, causing great inconvenience and economic loss to society. A technology that can detect traffic incidents in real time and alert people accordingly would therefore be a desirable way of reducing these ill effects.

Against this background, there have been many studies on AID, e.g., [2, 13]. Most of the approaches exploit data sent from stationary sensors and cameras installed on roads. However, the installation and maintenance of such sensors is expensive, with only the main routes likely to have them [17]. On the other hand, probe-car data (PCD), on which we focus in this paper, are becoming increasingly important, as the number of probe cars and the size of the associated data archives increase. PCD includes timestamps and the locations of vehicles, and may contain additional values such as the probe cars' speed and direction. Although a PCD system cannot monitor all cars, it enables traffic administrators to watch a vast area at a lower cost than by using stationary sensors. In addition, a PCD system can follow a probe car's sequence of movements in detail, which is hard to achieve via stationary sensors, and trajectory mining can be applied to the collected data.

Using PCD for freeways, it is easy to detect any reduction in speed, which sometimes implies congestion, by analyzing the speeds of the probe cars. However, this method is less applicable to local streets where there are many crossings and traffic lights that cause cars to stop frequently but normally. Moreover, speed reduction is not always an abnormal circumstance, even on freeways, and is not always caused by incidents such as accidents, which we would regard as sudden and unusual traffic events in this paper.

There are two types of congestion: spontaneous and abnormal [2]. Detecting spontaneous congestion is less important, as it originates in road design and urban planning. Any road may have potential bottlenecks, such as upslopes, curves, junctions, tollgates, and narrow sections. Vehicles are likely to slow down at the bottlenecks, with vehicular gaps shortening and drivers in the following cars having to brake. Congestion will then occur even without a traffic incident [7]. Spontaneous congestion also occurs when the traffic demands exceed the traffic capacity of such bottlenecks, and it is not resolved until the demand drops below the capacity [12]. The drivers may be familiar with the locations of such potential bottlenecks, and they can avoid them. On the other hand, abnormal congestion originates in traffic incidents, which need to be detected in real time to prevent or resolve any sudden heavy congestion.

In this paper, we propose an AID method for detecting traffic incidents by discovering abnormal car movements, distinguishing such movements from those occurring in spontaneous congestion. Our method measures differences between the current and usual traffic states, and has two aspects; namely, traffic state estimation and anomaly detection. First, we employ a probabilistic topic model [4] to model generation of PCD, which is influenced by hidden traffic situations, such as "smooth" and "congested." The model introduces a single set of several hidden component states, that are associated with probabilistic distributions over the PCD values, and all the road segments have their respective mixing coefficients. Using archived PCD, maximum-likelihood parameters of the model are estimated by an expectation–maximization (EM) algorithm. The estimated model reflects the usual state over the whole observation period. Our incident detection method simply follows the intuitive meaning of "anomaly." To detect incidents, the proposed method estimates the hidden state behind an observed PCD value and compares this current state with the usual state. If the current state is significantly different from the usual state, it is recognized as an anomaly.

We conducted an experiment using PCD observed for three of the routes of the Shuto Expressway system in Tokyo over the 2011 calendar year. The experiment showed that the proposed method can be effective for AID.

The main contributions of this paper are as follows.

- We propose a method for estimating traffic states by applying a probabilistic topic model to PCD, whereby road segments are characterized in terms of their expected performance.
- We propose a new method for detecting anomalous car trajectories according to the differences between the estimated states behind the trajectory and the usual states indicated by the learned model, whereby the detection is conducted adaptively in terms of the segments.
- Our experiment showed that the usual traffic state could be estimated using the observed PCD, and that our AID method had good selectivity for anomalous behavior by cars encountering incidents.

2. RELATED WORK

Although many studies have considered the traffic state estimation problem, there is no general agreement about a formal definition of a "traffic state." Some research estimates the traffic state in terms of vehicular speed [11, 19], and this kind of estimation characterizes states, i.e., quantized speeds, as "free" or "congested" [6]. Yoon et al. [17] proposed two feature values based on vehicular speed to detect a "bad" traffic state, i.e., slow traffic. In contrast, Kerner et al. [8] used travel time. Xia et al. [15] used a clustering method to identify congested traffic in a feature space involving traffic flow, speed, and occupancy, which has been well studied in traffic engineering [12].

AID can be considered to be an application of anomaly or outlier detection. Zhu et al. [20] applied the outlier detection methods to feature vectors carefully extracted from PCD using heuristics. If an incident occurs, cars upstream of the incident will travel slower and downstream cars will travel faster. In addition, a car passing before the incident will travel faster at that position than one passing just after the incident. If $v(d, t, l)$ is the vehicular speed in link l at time t on date d, Zhu et al. proposed the following four features: $v(d, t, l)$, $v(d, t, l) - v(d, t-1, l)$, $v(d, t, l-1)$ and $v(d, t, l+1) - v(d, t, l)$, where link $l-1$ is the next link upstream of l, and $l+1$ is the next link downstream. These feature vectors are filtered using the heuristics above and analyzed by distance-based outlier detection. In another AID study, Akatsuka et al. [2] proposed an alternative feature vector. From the viewpoint of machine learning, AID can be regarded as a classification problem. Abdulhai et al. [1] used neural networks, and Yuan et al. [18] used support vector machines, to classify the observed vectors from stationary sensors as being incident based or otherwise. AID can also be regarded as an application of the change-point detection problem in time-series analysis, with Wang et al. [13] developing a hybrid method using time-series analysis and machine learning.

Table 1: Notation

Notation	Definition
K	Number of traffic states.
k	Index of a traffic state.
S	Number of segments.
s	Index of a segment.
x_{sn}	n-th data in the s-th segment.
N_s	Number of observations in the s-th segment.
$\boldsymbol{\theta}_k$	Parameter of the k-th distribution.
$\boldsymbol{\pi}_s$	Mixing coefficient vector in segment s.
Λ	$(\{\boldsymbol{\pi}_s\}_{s=1,\cdots,S}, \{\boldsymbol{\theta}_k\}_{k=1,\cdots,K})$.
$\sigma(s, x)$	Traffic state in s when x was observed.
$\sigma(s)$	Usual traffic state in s.
$d(s, x)$	Divergence of $\sigma(s, x)$ from $\sigma(s)$.
X_s	Set of data observed in the s-th segment, i.e., $X_s = \{x_{s1}, x_{s2}, \cdots, x_{sN_s}\}$.
X	Whole set of data, i.e., $X = \{X_1, \cdots, X_S\}$.
X_c	Data sequence from car c, i.e., $X_c = \langle (s_1, x_1), (s_2, x_2), \cdots, (s_{N_c}, x_{N_c}) \rangle$.
$D(X_c)$	Divergence of X_c.

In this paper, we regard the AID problem as an anomaly detection problem. Previous work exploits characteristics of congested traffic, such as slowdown, in which vehicular speed decreases even in the absence of a traffic incident. We take another approach to follow the intuitive meaning of "anomaly"; namely, an event different than usual. For this purpose, the traffic should be described by a probabilistic model. We therefore exploit the idea of probabilistic topic models, which was originally studied in the field of natural language processing [5, 4]. The proposed method estimates both a set of traffic states over an entire route and the mixing coefficients for each road segment, with a traffic state corresponding to a topic.

3. METHODOLOGY

Table 1 summarizes the notations used in this paper.

This section describes our traffic state model and incident detection method. We first introduce a method for applying a probabilistic topic model to PCD. Our task is to estimate the model parameters using a PCD archive and to identify incidents by comparing the usual and current traffic states, which are obtained from the learned model.

3.1 Traffic State Model

Intuitively, we can identify some traffic states as "smooth" or "congested" regardless of location. Vehicles travel fast in smooth states and behave in a stop-and-go fashion in heavily congested states. When observing the speed of a probe car, the value is likely to be small if the car is in "congested traffic," or large if the traffic is "smooth." The value will also be affected by geographical conditions, such as curves and slopes. In short, the behavior of a car is affected by the surrounding traffic state, and the observed values for the probe car will change, whereas the traffic state is latent and varies according to the time and place. This relation between traffic states and PCD can be modeled using the latent Dirichlet allocation [5], the simplest topic model [4].

Traffic states are strongly related to roads, so we introduce the *segment* as the unit for watching traffic. The segment is defined independently of the PCD by the spatiotemporal space of observation. For example, one such segment could be defined as the section between Interchanges A and B on the inbound direction of Route 3 between 6 a.m. and 9 a.m. PCD includes timestamps and location data, that are obtained via GPS and are represented by longitude and latitude, and each probe-car observation can be assigned to a predefined segment.

PCD also has information on values such as speed and direction that can be recorded directly in the PCD or calculated using sequential observation. Here, all the observations are aggregated for each segment, and a set X_s of the observed data for the s-th segment is obtained. The symbol x_{sn}, the n-th value of X_s, might have either a scalar or a vector value. For simplicity in this paper, we assumed that x_{sn} was a scalar value, but our method could be extended to observe vector values.

Our model associates a traffic state with a probability distribution. Let K be the number of states, with the k-th traffic state corresponding to the parameter $\boldsymbol{\theta}_k$. The probability distribution for the s-th segment, given by $p(x|s)$, is described in terms of a mixture of these K distributions and can be described as follows:

$$p(x|s) = \sum_{k=1}^{K} \pi_{sk} p(x|\boldsymbol{\theta}_k), \tag{1}$$

where π_{sk} is the mixing coefficient for the k-th state and satisfies the conditions:

$$0 \leq \pi_{sk} \leq 1, \ \sum_{k=1}^{K} \pi_{sk} = 1 \tag{2}$$

for each s. The state parameters $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K\}$ are identical for all segments, but the mixing coefficient vector $\boldsymbol{\pi}_s = (\pi_{s1} \cdots \pi_{sK})^\mathsf{T}$ is different for each segment. By using a global $\boldsymbol{\theta}_k$, we can compare and characterize segments in terms of local $\boldsymbol{\pi}_s$. For example, straight sections are dominated by "smooth" states, with sections that include tollgates that are dominated by "congested" states.

Finally, for each segment, the generative process for this model was as follows.

1. Choose a hidden state $k \sim$ multinomial probability distribution Multi($\boldsymbol{\pi}_s$).
2. Generate the value $x_{sn} \sim p(x_{sn}|\boldsymbol{\theta}_k)$.

3.2 Parameter Estimation

Our model is described by a mixture distribution, with its maximum-likelihood parameters estimated by an EM algorithm, using X as training data [3]. For simplicity, we introduce the symbol Λ as a set of all parameters in the model. For the entire set X of observed data, the likelihood under the model introduced above is given by the following equation:

$$L(X) = \prod_{s=1}^{S} \prod_{n=1}^{N_s} \sum_{k=1}^{K} \pi_{sk} p(x_{sn}|\boldsymbol{\theta}_k). \tag{3}$$

The update equations are derived by considering the maximization of the following Q function under constraint (2):

$$Q(X, \Lambda, \hat{\Lambda}) = \sum_{s=1}^{S} \sum_{n=1}^{N_s} \sum_{k=1}^{K} p(k|x_{sn}, \hat{\Lambda}) \log p(k, x_{sn}|\Lambda), \tag{4}$$

where

$$p(k|x_{sn}, \hat{\Lambda}) = \frac{\hat{\pi}_{sk} p(x_{sn}|\hat{\boldsymbol{\theta}}_k)}{\sum_{k=1}^{K} \hat{\pi}_{sk} p(x_{sn}|\hat{\boldsymbol{\theta}}_k)} \equiv \gamma_{snk} \tag{5}$$

$$p(k, x_{sn}|\Lambda) = \pi_{sk} p(x_{sn}|\boldsymbol{\theta}_k), \tag{6}$$

and $\hat{\Lambda}$ refers to the parameters estimated in the previous EM iteration.

This Q is maximized by introducing Lagrange multipliers and setting its partial derivative to zero. The update equation for $\boldsymbol{\theta}_k$ is then derived by solving the equation:

$$\sum_{s=1}^{S} \sum_{n=1}^{N_s} \frac{\gamma_{snk}}{p(x_{sn}|\boldsymbol{\theta}_k)} \frac{\partial}{\partial \boldsymbol{\theta}_k} p(x_{sn}|\boldsymbol{\theta}_k) = 0. \tag{7}$$

For example, assume a Poisson distribution for p when any x_{sn} values, e.g., speed, are nonnegative integers, then:

$$p(x_{sn}|\boldsymbol{\theta}_k) \equiv p(x_{sn}|\lambda_k) = \frac{{\lambda_k}^{x_{sn}} e^{-\lambda_k}}{x_{sn}!}, \tag{8}$$

where λ_k is both the mean and variance, and is the only parameter of p. In this case, by solving equation (7), the update equation for λ_k is derived as:

$$\lambda_k = \frac{\sum_{s=1}^{S} \sum_{n=1}^{N_s} \gamma_{snk} x_{sn}}{\sum_{s=1}^{S} \sum_{n=1}^{N_s} \gamma_{snk}}. \tag{9}$$

For the mixing coefficient $\boldsymbol{\pi}_s$ for the s-th segment, we obtain, regardless of p, the equation:

$$\pi_{sk} = \frac{\sum_{n=1}^{N_s} \gamma_{snk}}{N_s}. \tag{10}$$

We now have the EM algorithm for estimating the parameters of our traffic state model: After generating Λ at random, the EM iteration alternates between the E step, which calculates all γ_{snk} using equation (5), and the M step, which updates Λ according to equations (7) and (10), until the log likelihood $\log L(X)$ converges.

3.3 Incident Detection

We have now described our traffic state model and its parameter estimation method. Given the estimated parameter

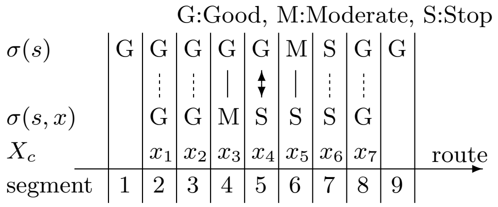


Figure 1: Concept of divergence comparison

Λ and the value x observed in segment s, the posterior distribution is given by $p(k|x,s)$. We now define the *current traffic state* when x was observed, denoted by $\sigma(s,x)$, as the maximum probable state given x. Using the posterior distribution with Bayes' theorem, $\sigma(s,x)$ is estimated as:

$$\sigma(s,x) = \arg\max_k \{\pi_{sk} p(x|\boldsymbol{\theta}_k)\}. \tag{11}$$

Meanwhile, the learned model itself reflects the usual state over the whole observation period because the parameters are estimated to fit the distribution in the dataset. We can therefore define the *usual traffic state* for the s-th segment, denoted by $\sigma(s)$, as the maximum probable state:

$$\sigma(s) = \arg\max_k \pi_{sk}. \tag{12}$$

We now have the usual and the current traffic states for each segment. Figure 1 describes our idea of incident detection via *divergence comparison*. For example, the usual state $\sigma(s)$ may indicate smooth traffic in a straight midnight segment, congested traffic in a rush-hour segment, or stop-and-go traffic in segments that contain tollgates for any time of day. If $\sigma(s)$ indicates congested traffic and $\sigma(s,x)$ is also congested, the current traffic remains usual and would not be considered an anomaly. If the usual state $\sigma(s)$ indicates free-flowing traffic and the current state $\sigma(s,x)$ indicates stop-and-go traffic, then it would be suspected that an anomaly caused by an incident has occurred.

Our AID method measures the degree of anomaly for each probe car's trajectory. Assume that a probe car c traverses a road, observing a set of N_c values. Let X_c be the sequence of data such that each is a tuple of segment and value observed by c as described in Table 1. Let x_n be an observed value in a segment s_n. Of course, we can count how many times $\sigma(s_n,x_n)$ differs from $\sigma(s_n)$, but this approach regards major differences in the same light as minor differences, which perhaps stem from individual variation rather than from a traffic incident. We therefore introduce the *divergence* of the current state from the usual state, denoted by $d(s_n,x_n)$, to quantify the difference between the two states. Because in our model each state is associated with a probability distribution, we measure this difference in terms of the Kullback–Leibler divergence of the current state's distribution from the usual state's distribution. The k-th state corresponds to the probability distribution $p(x|\boldsymbol{\theta}_k)$, and therefore:

$$d(s_n,x_n) = \sum_x p(x|\boldsymbol{\theta}_{\sigma(s_n,x_n)}) \log \frac{p(x|\boldsymbol{\theta}_{\sigma(s_n,x_n)})}{p(x|\boldsymbol{\theta}_{\sigma(s_n)})}, \tag{13}$$

where p is discrete. For example, assume Poisson distribution for p as equation (8). The divergence is derived as:

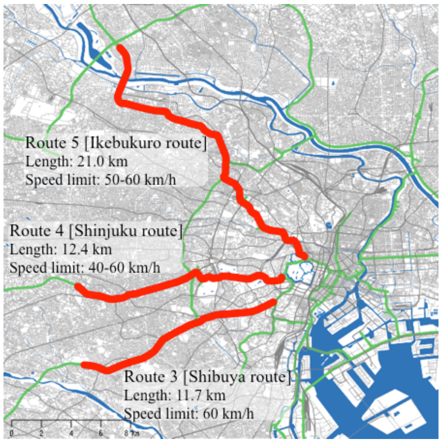


Figure 2: Three routes of Shuto Expressway within the Tokyo area

$$d(s_n,x_n) = \lambda_{\sigma(s_n)} - \lambda_{\sigma(s_n,x_n)} + \lambda_{\sigma(s_n,x_n)} \log \frac{\lambda_{\sigma(s_n,x_n)}}{\lambda_{\sigma(s_n)}}. \tag{14}$$

The behavior of a car c is determined as anomalous if the estimated state behind the observed data sequence X_c is quite different from the usual state. We define the divergence of X_c from the usual state, denoted by $D_{\mathrm{all}}(X_c)$, as:

$$D_{\mathrm{all}}(X_c) = \sum_{n=1}^{N_c} d(s_n,x_n). \tag{15}$$

The more a car behaves differently from its usual behavior, the larger $D_{\mathrm{all}}(X_c)$ will be. $D_{\mathrm{all}}(X_c)$ is considered as a score of the degree of anomaly, with c being determined as anomalous when $D_{\mathrm{all}}(X_c)$ is sufficiently large, i.e., larger than a predefined threshold.

There are two points to consider about $D_{\mathrm{all}}(X_c)$. First, $D_{\mathrm{all}}(X_c)$ will also increase the longer the car c runs, and any car would eventually be determined as being anomalous. We therefore define the normalized divergence $D(X_c)$ as the sum of the largest N divergences $d(s_n,x_n)$ if N_c is not less than N. Otherwise, $D(X_c)$ is equivalent to $D_{\mathrm{all}}(X_c)$. We have used D instead of D_{all} in the rest of the paper. Second, when a car generates values periodically, no observation or multiple observations in a trajectory can be assigned to a single segment. Our idea of divergence comparison in Figure 1 assumed that one segment corresponded to one current state, which might require interpolation or aggregation of data for each segment.

4. EXPERIMENT

4.1 Dataset and Preprocessing

Our probe-car dataset was obtained from probe cars traveling on three routes on the Shuto Expressway system in Tokyo during 2011. The route information is displayed in Figure 2, with the three routes being shown as thick red lines on a map of Tokyo.

Data preprocessing comprised four phases: 1) segment definition, 2) map matching, 3) trajectory identification, and 4) interpolation. These procedures are described below.

4.1.1 Segment Definition

Traffic state information is strongly related to geographical conditions. We defined road segments by partitioning each route on the expressway every 50 m for estimation at a finer level of granularity. The direction was noted. This experiment did not consider temporal partitioning, even though the traffic in some places changed considerably over time. Therefore, each segment represented a certain 50-m length of roadway for a certain direction on a certain expressway route, and all data for a segment were treated without any consideration of time.

4.1.2 Map Matching

Despite the above definition of a segment being based on an expressway route, location data in PCD were described in terms of the two-dimensional (2-D) coordinates of longitude and latitude, with the original observation not being related to any particular segment. It was therefore necessary to identify the segment that the probe car was in from the time and position for every observation, even though in this experiment we did not consider the timestamps. Map matching is a technology for identifying the road segment on which the vehicle is traveling and for locating the vehicle within that segment [9], and several methods have been proposed [14, 16]. In this experiment, map matching was conducted in the simplest way: a probe car's observation was matched with the nearest segment to the car's location. The direction was estimated from the angular difference between the probe car's heading azimuth in the PCD and the segment's azimuth for each direction, and then choosing the direction that gave the smaller angle.

4.1.3 Trajectory Identification

After map matching, each probe-car observation whose location was represented by coordinates in the 2-D space was matched with the nearest-neighbor segment, as defined in the first phase of preprocessing. However, the observations form a collection of punctuated data, with each observation being separate from the others. Therefore, the continuous movement of the car, i.e., its trajectory, is not directly available. To identify trajectories, we grouped all observations in the probe-car dataset by the car's ID and sorted them by timestamp for each group, before concatenating them in chronological order whenever the time gap between two consecutive observations was 10 min or less. A probe car does not always travel the entire length of a route, because it can enter or exit the route at intermediate junctions. For a car traveling on a single route, its trajectory can be visualized in terms of a time–space diagram [12]. Figure 5 is an example of such a diagram and will be described in detail later.

After the trajectory identification, we labeled each trajectory, using the traffic log made available by the administrator of the Shuto Expressway. This traffic log is recorded via stationary sensors on or alongside the roads every 5 min, together with notations about incidents such as accidents and construction. A trajectory was labeled as anomalous whenever a car passed a stationary sensor that had recorded an incident at that time. Table 2 summarizes the statistical information for our probe-car dataset after trajectory identification. The number of anomalies means the number of trajectories for a probe car passing the scene of an incident when the incident occurred but does not indicate the number of unique incidents.

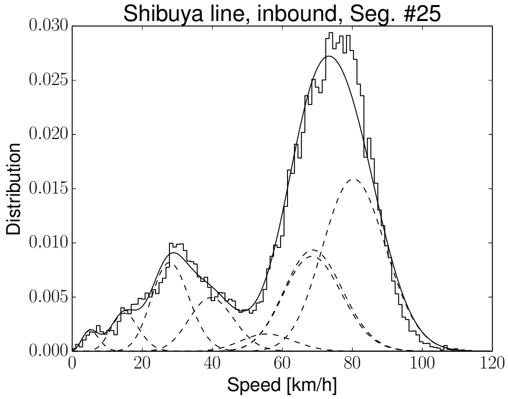


Figure 3: Histogram of speed of probe cars in a segment and estimated Poisson mixture

4.1.4 Interpolation

We used a probe car's speed as the observed value in this experiment. However, as mentioned in Section 3.3, our detection method estimated the current state for each trajectory for each segment that the car had passed. Our 50-m segment was too short for fast-moving probe cars to conduct observations in every segment, whereas a slow-moving car generated multiple data in a single segment. We therefore formed an observation sequence for a trajectory by linear interpolation, giving a sequence of consecutive observations at 50-m intervals.

4.2 Parameter Estimation

In this experiment, the observed values represented the speed of probe cars as nonnegative integers. We therefore assumed a Poisson distribution for the probability distribution corresponding to each traffic state.

We also assumed K, the number of traffic states, to be 8. In a preliminary experiment, we estimated the parameters of our traffic model while varying the value of K up to 100, and we used the Akaike information criterion (AIC) to evaluate the model. However, the effect of K was substantially less than that of the likelihood for improving the AIC, with AIC being almost the same regardless of K. If K is assumed to be large, there is a tendency for multiple states to have almost the same distribution.

We implemented the EM algorithm described in Section 3.2 using OpenMP for multiprocessing. The estimation was executed on our 32-core Xeon computer for each route of the Shuto Expressway. It took about 1 min for each direction of the Shibuya and Shinjuku routes, and about 2 min for each direction of the Ikebukuro route. Figure 3 shows the actual histogram for a segment of the inbound Shibuya route as a step line chart and the estimated Poisson mixture as a solid curved line. Each of the eight Poisson distributions was multiplied by the mixing coefficients π_{sk}, and each is also shown in Figure 3 as dashed curves. The estimated curve almost fits the actual histogram for the training dataset.

4.3 Incident Detection

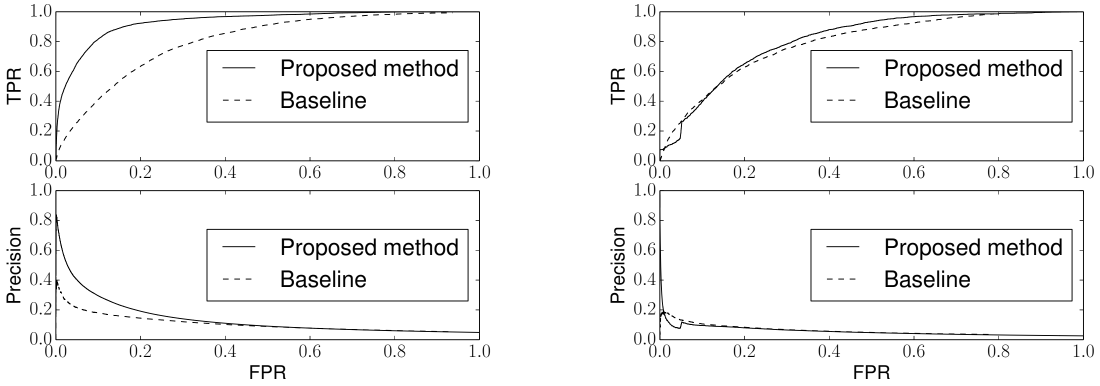
Using the estimated traffic model, we examined whether the proposed method could identify anomalous trajectories. We calculated the divergence for each trajectory and sorted the trajectories in order of their divergence. The divergence

Table 2: Statistics on trajectories in our probe-car dataset

Routes	Shibuya route		Shinjuku route		Ikebukuro route	
	Inbound	Outbound	Inbound	Outbound	Inbound	Outbound
Period	January 1, 2011 – December 31, 2011 (365 days)					
# of trajectories	100,581	95,386	95,293	88,345	128,789	114,942
# of anomalies	4,259	2,475	4,365	3,891	6,089	5,603
Average travel distance [km]	5.7	5.8	6.4	6.7	7.5	8.1

Table 3: AID results

Routes		Shibuya route		Shinjuku route		Ikebukuro route	
		Inbound	Outbound	Inbound	Outbound	Inbound	Outbound
AUC	Our method	0.912	0.812	0.927	0.919	0.902	0.933
	Baseline [20]	0.802	0.794	0.846	0.780	0.823	0.805



(a) Ikebukuro route, outbound (the best case)

(b) Shibuya route, outbound (the worst case)

Figure 4: ROC curves (upper frames) and precision vs. false positive rate (lower frames)

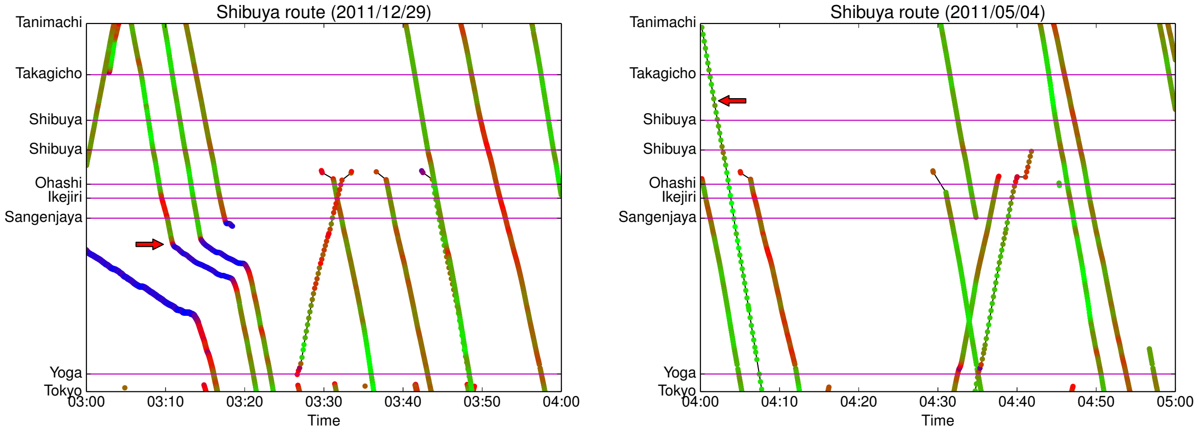
of a trajectory was calculated by summing the top N divergences among the observations. In a preliminary experiment, we conducted the detection for several values of N, obtaining the best result when N was 20. Because we were using 50-m segments, the divergences of trajectories were normalized to 1-km equivalents.

For comparison, we implemented a second method based on Zhu et al. [20], which was described in Section 2. The method was modified to enable its application to our dataset, and although it detected outlier segments represented by the pair of time and position, our system was evaluated in terms of anomalous cars. Therefore, we judged that a detection event was successful if the detected car was labeled as an anomaly in our dataset, even if the detected segment for the detected car was not a segment involving an incident.

Our detection method gives an alert when the divergence of a trajectory exceeds a given threshold, and the compared method gives an alert when the average distance of a feature vector from other vectors exceeds a given threshold. The lower the threshold, the more alerts will be issued. We evaluated the selectivity performance of the two methods in terms of a receiver-operating characteristic (ROC) curve. An ROC curve is drawn by plotting the true positive rate (TPR), which is equivalent to recall, against the false positive rate (FPR). The area under the curve (AUC) indicates the discrimination performance, with larger AUC values indicating better discrimination.

The results are displayed in Table 3 and Figure 4. Table 3 reports the AUC of the proposed and baseline methods on our probe-car datasets. The results showed that our method had better selectivity for cars that have passed incident locations, despite using fewer heuristics about anomalies than the baseline method. Figure 4 shows the ROC curves in the upper frames and the precision against FPR in the lower frames. Although the ROC curve should connect points (0,0) and (1,1), that of the baseline method broke off before (1,1) was reached, because the method filtered out some feature vectors, with the number of subject trajectories being less than the total number of trajectories. The AUC of the baseline method was calculated by interpolating linearly between the right-hand end of the ROC curve and (1,1). Figure 4(a) shows the curves for the outbound Ikebukuro route, which was the best case in our experiment. The precision exceeded 80% for the worst 1,000 trajectories. Figure 4(b) shows the curves for the outbound Shibuya route, which was the worst case.

Figure 5 shows examples of trajectories for the outbound Shibuya route that had much divergence in terms of their time–space diagram. Each plot shows the position of a probe car against time. The position is represented as the distance from the origin of the line: the bottom corresponds to the Tokyo interchange, the westernmost along the Shibuya route, and the top corresponds to the Tanimachi (easternmost) junction. Horizontal pink lines indicate the positions of interchanges and junctions. The inbound direction is the direction from the bottom to the top in this diagram. Therefore, trajectories downward and to the right involve traveling along the outbound Shibuya route. The color of the plot



(a) True positive example: a car involved in an accident

(b) False positive example: an abnormally fast car

Figure 5: Time–space diagrams for probe cars

indicates the speed of the probe car at that point. Green represents high speed (100 km/h), red is moderate speed (50 km/h), and blue is "almost stopped" (0 km/h). The color changes gradually according to the speed. The trajectory marked with an arrow in Figure 5(a) was the most anomalous trajectory, with this car being directly affected by an incident. The diagram shows that this car was "stop-and-go" between Sangenjaya and Yoga. On the other hand, the marked trajectory in Figure 5(b) was ranked as no. 301 among the anomalous trajectories. This car did not encounter an incident. The diagram shows the car traveling rapidly along the route.

5. DISCUSSION

In Figure 4(b), the TPR of the proposed method was sluggish when the TPR was around 0.1, indicating that the proposed method rarely detected anomalous cars correctly even when the threshold was lowered to some degree. Cars whose trajectories became anomalous at this point traveled rapidly, with the car of the marked trajectory in Figure 5(b) being an example of such cars. This car traversed the route before dawn, when the traffic is usually smooth. One of the possible reasons for such false positives is that our experiment did not consider temporal partitioning in the segment definition, even though the traffic changed considerably over time. The spatial length of a segment, as well as parameters K and N, should also be determined in future studies.

The following discussion demonstrates another analysis on the abovementioned false positives based on the estimated traffic states. In this experiment, we used the speed of the probe car, and the traffic state was represented by the Poisson distribution, which was characterized by the mean and variance parameter λ. The stacked area chart in Figure 6 shows the estimated mixing coefficients for the eight Poisson distributions for each segment of the outbound Shibuya route. The horizontal axis shows the position along the route, and cars travel from left to right. The colored areas show the mixing coefficient for each state varying with position. They are in order of λ, with the bottommost being the slowest, and the topmost being the fastest. From Tanimachi to Ikejiri, the top three fastest states were dominant, which means that cars usually travel quickly in this section. However, from Ikejiri to Yoga, the coefficients for the faster states decrease as the slower states begin to dominate, because the cars usually travel more slowly in this section. Therefore, although the marked trajectory in Figure 5(b) does not seem to include any incidents, this behavior was quite different from the usual running pattern, and our method identified this as an anomaly. It is noteworthy that our traffic state model has enabled this sort of analysis, with every segment being characterized using a single set of traffic states. Although we used the data sequence to give observations at 50-m intervals for each probe car, stationary sensors can also generate similar data except for tracking information for each car. Because parameter estimation does not require such information, this road characteristics analysis can be conducted using stationary sensors, and its output might be applied to other problems; e.g., route guidance.

The Shuto Expressway system has many bottlenecks, such as curves and narrow sections that involve frequent changes in vehicular speed, unlike freeways. We speculate that this is the reason that our intuitive method found that "unusual" car behavior worked better than a heuristic method that pays attention to changes in speed. On the other hand, a significantly fast car can be surely determined as an anomaly if its behavior is statistically unusual relative to the past observations, although this kind of "unusualness" is not a problem for drivers. Anomalies accompanying a slowdown in vehicular speed can be regarded as a subset of the anomalies discussed in this paper. The administrator and drivers have the option to filter the outcome of our detection algorithm using additional heuristics. However, a particular incident is hard to detect by the proposed method if the traffic behavior in the incident is just like regular spontaneous congestion. We are currently conducting investigations into detailed issues as a further study, expanding our dataset sphere from only three routes to all the routes on the Shuto Expressway system.

6. CONCLUSION

We have studied the problem of detecting traffic incidents using probe-car data. Although congestion can be detected by monitoring vehicular speeds, it is chronic in some spots and does not necessarily indicate the occurrence of an in-

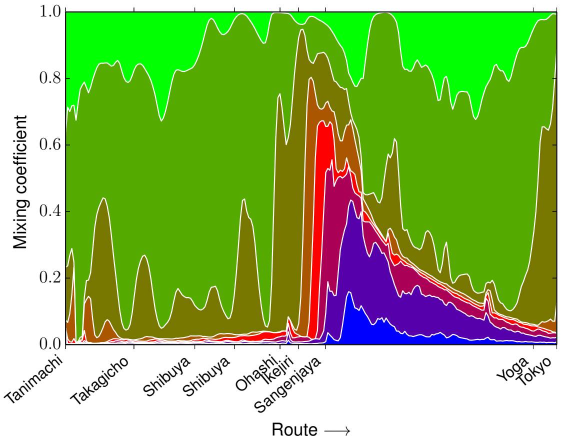


Figure 6: Mixing coefficients for eight Poisson distributions for each segment of the outbound Shibuya route

cident. To detect traffic incidents, we propose an approach that compares the current traffic state with the usual one for that location in terms of anomalous car movements, using a probabilistic topic model to describe the state of monitored traffic. We proposed an incident detection method that measured the difference between the usual and current states. Our method was applied to real probe-car data that were collected on the Shuto Expressway system in Tokyo, and the discrimination performance was evaluated. The results showed that our method could discriminate trajectories affected by incidents from other trajectories, although abnormally fast cars were also reported as anomalies, giving a low precision for certain routes.

7. ACKNOWLEDGMENTS

The traffic log used in our experiment as the ground truth for incident occurrence was made available by Metropolitan Expressway Co., Ltd.

8. REFERENCES

[1] B. Abdulhai and S. G. Ritchie. Enhancing the universality and transferability of freeway incident detection using a bayesian-based neural network. *Transportation Research Part C: Emerging Technologies*, 7(5):261–280, 1999.

[2] H. Akatsuka, A. Takasu, K. Aihara, and J. Adachi. Highway incident detection based on probe car data. In *International Conference on Information Systems (Information Systems 2013)*, pages 103–110, 2013.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 9. Springer, 2006.

[4] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, Apr. 2012.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] C. de Fabritiis, R. Ragona, and G. Valenti. Traffic estimation and prediction based on real time floating car data. In *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems*, pages 197–203, 2008.

[7] East Nippon Expressway Co., Ltd. Generation mechanism of traffic congestion caused by traffic concentration. `http://www.e-nexco.co.jp/activity/safety/mechanism.html`. Accessed: 2013-12-06.

[8] B. Kerner, C. Demir, R. Herrtwich, S. L. Klenov, H. Rehborn, M. Aleksic, and A. Haug. Traffic state detection with floating car data in road networks. In *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*, pages 44–49, 2005.

[9] M. A. Quddus, W. Y. Ochieng, and R. B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, 2007.

[10] J. M. Sussman. ITS: A short history and a perspective on the future. In *Perspectives on Intelligent Transportation Systems (ITS)*, pages 3–17. Springer US, 2005.

[11] S. Tao, V. Manolopoulos, S. Rodriguez Duenas, and A. Rusu. Real-time urban traffic state estimation with a-gps mobile phones as probes. *Journal of Transportation Technologies*, 2(1):22–31, 2012.

[12] M. Treiber and A. Kesting. *Traffic Flow Dynamics*. Springer Berlin Heidelberg, 2013.

[13] J. Wang, X. Li, S. Liao, and Z. Hua. A hybrid approach for automatic incident detection. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1176–1185, 2013.

[14] C. E. White, D. Bernstein, and A. L. Kornhauser. Some map matching algorithms for personal navigation assistants. *Transportation Research Part C: Emerging Technologies*, 8(1âĂŞ6):91–108, 2000.

[15] J. Xia, W. Huang, and J. Guo. A clustering approach to online freeway traffic state identification using its data. *KSCE Journal of Civil Engineering*, 16(3):426–432, 2012.

[16] J.-s. Yang, S. Kang, and K.-s. Chon. The map matching algorithm of gps data with relatively long polling time intervals. *Journal of the Eastern Asia Society for Transportation Studies*, 6:2561–2573, 2005.

[17] J. Yoon, B. Noble, and M. Liu. Surface street traffic estimation. In *Proceedings of the 5th international conference on Mobile systems, applications and services*, MobiSys '07, pages 220–232, New York, NY, USA, 2007. ACM.

[18] F. Yuan and R. L. Cheu. Incident detection using support vector machines. *Transportation Research Part C: Emerging Technologies*, 11(3–4):309–328, 2003. Traffic Detection and Estimation.

[19] Y. Yuan, J. W. C. Van Lint, R. Wilson, F. van Wageningen-Kessels, and S. Hoogendoorn. Real-time lagrangian traffic state estimator for freeways. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):59–70, 2012.

[20] T. Zhu, J. Wang, and W. Lv. Outlier mining based automatic incident detection on urban arterial road. In *Proceedings of the 6th International Conference on Mobile Technology, Application & Systems*, Mobility '09, pages 29:1–29:6, New York, NY, USA, 2009. ACM.