





Paper

# Predicting traffic breakdown on expressways using linear combination of vehicle detector data

Rikuto Shigemi <sup>1</sup>, Hiroyasu Ando <sup>2</sup>, Kentaro Wada <sup>1</sup>,  
and Risa Mukai <sup>3</sup>

<sup>1</sup> Faculty of Engineering, Information and Systems, University of Tsukuba  
1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305-8577, Japan

<sup>2</sup> Advanced Institute for Materials Research, Tohoku University  
2-1-1 Katahira, Aoba-ku, Sendai-shi 980-8577, Japan

<sup>3</sup> Hanshin Expressway Co., Ltd.  
3-2-4 Nakanoshima, Kita-ku, Osaka 530-0005, Japan

Received October 14, 2022; Revised December 28, 2022; Published April 1, 2023

**Abstract:** Traffic congestion is closely associated with various social issues that must be solved urgently. With the recent advancement of machine learning technologies, diverse methods for predicting traffic congestion have been developed. Specifically, traffic prediction using deep learning can provide highly accurate performance. Nevertheless, several difficulties remain because of the complexity of deep learning models: particularly, they require large amounts of data and computational power. For this study, we strive to achieve traffic prediction precision using a simple linear model. Instead of improving complex models, we select training data appropriately with a linear model and then verify the feasibility of prediction by exploring “data complexity”. The prediction results imply that the linear model is as precise as deep learning even with fewer number of data and parameters. We use actual data from expressways collected using detectors.

**Key Words:** detector data, explainability, linear model, predicting congestion, regression

## 1. Introduction

### 1.1 Background

Automobiles remain a major mode of transportation, playing an important role in the social system. Nevertheless, it is also true that they can produce difficulties such as traffic congestion and collision accidents. In Japan, traffic congestion in urban areas causes an annual time loss of about 3.81 billion hours [1]. As a solution to these shortcomings of automobile use, Intelligent Transport Systems (ITS) are being promoted all over the world. For such systems, development of information infrastructure for road traffic using AI and IoT technologies is progressing rapidly [2, 3]. Given that background, the



prediction of traffic congestion using machine learning techniques has attracted much attention [3–6]. One of those technologies that predict future traffic congestion is based on various current and past information related to road networks. If the prediction technology is developed to a practical level, then it can be expected to contribute to the development of ITS in a mode of facilitating traffic flow and selecting optimal routes that avoid traffic congestion. Currently, the major technology for traffic congestion prediction is based on deep learning models with big data. Unfortunately, these models have some shortcomings. Most importantly, they require huge amounts of data and tuning algorithms according to the model complexity. As presented herein, we discuss the usefulness and significance of a prediction model with a linear combination of detected data.

## 1.2 Earlier work

Traffic congestion is, generally speaking, classified into two types based on its characteristics: recurring and non-recurring congestion. Many studies have been undertaken to predict each type of congestion using machine learning with traffic data. This study specifically examines recurring congestion, which results from the bottleneck’s inability to cope with excessive traffic during peak hours such as commuting [7]. When predicting recurring traffic congestion, accurate prediction of the time and scale of periodic traffic congestion is a challenge. For this purpose, it is important to capture the characteristics of spatiotemporal variations in traffic data. Recently, as shown by the DEEPLSTM method reported by [8] and the DCRNN method reported by [9], the mainstream approach is to model nonlinear dynamics and improve prediction accuracy by creating more complex deep learning models based on CNN and RNN techniques. In contrast to this avenue of research, this paper presents consideration of predicting traffic congestion using a linear model. To this aim, we demonstrate the possibility of making predictions by exploiting physical phenomena in the real world as a part of computational processes.

## 1.3 Objective

For this study, we perform prediction of traffic congestion on expressways using a linear model based on earlier studies [10]. We examine whether the accuracy can be improved by modifying the training data compared with earlier studies. The purpose of this study is twofold. The first is to show that the essential information for traffic congestion can be focused on successfully from the data complexity by selecting the data, thereby increasing the prediction accuracy. This improved prediction enables us to propose a methodology for prediction that does not rely on the complexity of the model, but rather finds the computational process from the complexity inherent in the data. This results in significantly reducing the length of time for training. In fact, the learning time was about 1,900 times shorter than that of deep learning as indicated in [10]. Secondly, we demonstrate the usefulness and importance of linear models for predicting traffic congestion. Specifically, we demonstrate the high explanatory ability of the linear model in the prediction by investigating the causes of changes in accuracy through parameter analysis. The subject of this study is the No. 11 inbound line of the Hanshin Expressway.

## 1.4 No. 11 inbound line of the Hanshin Expressway

The No. 11 inbound line, an expressway connecting Osaka Airport and central Osaka City, is characterized by heavy traffic and frequent congestion. The bottlenecks caused by traffic congestion are particularly at 0.0 kp, 4.0 kp, and 10.0 kp (in the case of the No. 11 inbound line, each point is distinguished as  $x$  [kp], where  $x$  [km] represents the distance from the junction point with No.1 line). Among these bottlenecks, generated congestion patterns at 0.0 kp and 4.0 kp are assessed for this study. The 0.0 kp point is the merging of the No. 1 line and the No. 11 inbound line, as shown in Fig. 1, where traffic flows from the No. 11 inbound line to the main flow of the No. 1 line. This road arrangement leads to congestion starting from the right lane. The 4.0 kp point is the sag in which the road vertical gradient changes from downhill to uphill. The slight deceleration of the lead vehicle in the group is amplified and transmitted to the rear, thereby causing congestion. An entrance from the surface street is located near the 4.0 kp point, but the inflow from this point has little effect on the occurrence of traffic congestion.

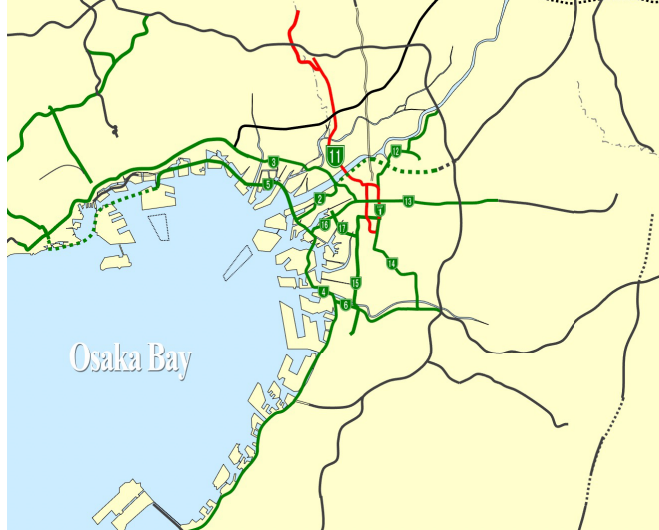


Fig. 1. Hanshin Expressway map. The No. 11 line is shown as red.

## 2. Model

As described in the preceding section, this study uses a linear combination of the data to predict traffic congestion occurrence. Additionally, we divide the occurrence of traffic congestion at a particular location into two categories based on spatial characteristics: traffic breakdown (i.e., bottleneck activation at this location) and spillback from the downstream active bottleneck. For those categories, we devise different training data. This section presents a description of the following four points.

- Data Summary
- Outline of the linear model
- Prediction targets and their correctness criteria
- Definitions of the two types of congestion and the training data used to forecast them

### 2.1 Data summary

For use in this study, a prediction model was constructed using data of vehicle detectors installed on the Hanshin Expressway. First, as explanatory variables, per-detector traffic data recorded every 5 min at the installation points are used. Traffic data of four types are recorded for each detector: average velocity, traffic flow, high-vehicle traffic flow, and detector occupancy. Additionally, as the explained variable, daily traffic congestion report data are used. The data include records of the presence or absence of traffic congestion at that location every 30 s.

Among the many detectors located on the road, the appropriate detectors are selected for the task. Its traffic data are used as the explanatory variable. To match the time series with the explanatory variables, the data are processed so that if traffic congestion occurs once every five minutes, that five-minute period is regarded as congested. Alternatively, if there is no traffic congestion, that five-minute period is not regarded as congested.

### 2.2 Linear model

The model we use in this study is a linear model. Therefore, we use ridge regression in the form of Eq. (1) to learn the parameter  $W_{out}$ . As for the teacher signal, we use  $X(t) = [x_1^1(t), \dots, x_K^1(t), x_1^2(t), \dots, x_K^2(t), \dots, x_1^J(t), \dots, x_K^J(t), 1]$ , and  $U(t) = [u_1(t), \dots, u_K(t)]$ . Here,  $x_k^j(t)$  is the set of traffic data type  $j$  collected by detectors at the location  $k$  and time  $t$ .  $u_k(t)$  represents a binary state of traffic at the location  $k$  and time  $t$ , namely 0 is non-congested and 1 is congested. The predicted state of  $u_k(t)$  is denoted by  $\hat{u}_k(t)$ . By using the learned parameter  $W_{out}$ , we can calculate the prediction vector  $Y(t + \tau) = [y_1(t + \tau), y_2(t + \tau), \dots, y_K(t + \tau)]$  by Eq. (2), where  $y_k(t + \tau)$  is the continuous variable for prediction. Then, we compute binary  $u_k(t + \tau)$  by the threshold in Eq. (3).

$$W_{out} = (X^T X + \lambda I)^{-1} X^T U, \quad (1)$$

$$Y(t + \tau) = X(t)W_{out}, \quad (2)$$

$$\hat{u}_k(t) = \begin{cases} 1, & y_k(t) \geq 0.5, \\ 0, & y_k(t) < 0.5, \end{cases} \quad (3)$$

Here,  $\tau$  is the prediction time. As can be seen by the equations and the summary of dataset, it is used only a single time point in the space data for prediction. However, the single time point has time series information due to the ten-minute time lag  $\tau$ . As might be apparent from these equations and from the data summary, only spatial data at one point in time are used for prediction. In addition, when estimating parameters, we do not train backward in the time-series direction as RNNs do. However, because a lag  $\tau$  exists between the explanatory variable and the target variable at a single point in time, only the time-series information in the data at this single point is used for learning.

### 2.3 Prediction targets and their correctness criteria

The objective of this traffic congestion forecasting is to ascertain whether traffic congestion will occur at a target location 10 min later. Therefore, label identification is performed for congested and non-congested traffic. The accuracy is measured by the percentage of correct answers to the occurrence of traffic congestion. It should be noticed that it is important to evaluate the prediction precision by the accuracy of prediction regarding congestion occurrence label. Here, the congestion occurrence label corresponds to the congestion labels which satisfy a certain condition. Specifically, the conditions are (1) the label is located at the beginning of the congestion, (2) there are no congestion labels in the last 30 minutes before the occurrence. By introducing the congestion occurrence label, we can focus on the occurrence of traffic congestion out of the whole traffic congestion.

Moreover, when we consider the application of the proposed method to the real world, it is sufficient that the time at which we conduct prediction is before the congestion occurrence and the predicted occurrence time is earlier than the actual occurrence. According to the two points, we allow the time interval for the evaluation criteria. Specifically, the definition of the correct prediction is that the actual traffic congestion occurs between 5 minutes earlier and 30 minutes later than the prediction time point. Moreover, under this criterion, we use the evaluation score, Recall, as the ratio of the number of correctly predicted congestion occurrences out of the actual occurrence. On the other hand, we use Precision as the ratio of the number of actual congestion occurrences out of the total predicted occurrences. Accordingly, we use F1 that is the harmonic mean of Recall and Precision.

### 2.4 Prediction of traffic breakdown at bottlenecks

Congestion at a bottleneck is defined as the occurrence of traffic congestion, which is the starting point of a series of congestion events. By the definition of congestion, the traffic data at the very beginning of the predicted congestion are always non-congested. Compared to other congestion labels, the explanatory variables at that point in time are similar to those at the non-congested label. This similarity suggests that the key to this forecasting task is to learn the difference between ‘‘congestion near non-congestion’’ and ‘‘non-congestion near congestion’’. Therefore, this task uses training data consisting only of data before and after the occurrence of traffic congestion. Specifically, the training data are defined as a set of three steps: before, after and the onset of the time point at which the target variable corresponds to a traffic congestion occurrence. This definition is intended to eliminate data on traffic congestion [non-congestion] with sufficiently low [high] speed, which is far different from the traffic congestion occurrence, from the training data.

### 2.5 Prediction of congestion spillback

Congestion spillback is caused by downstream congestion reaching the point of interest. Therefore, an important part of this task is to learn the difference between the traffic data in the vicinity when downstream congestion arrives and when it does not. Therefore, the training data used here are not selected as those used for predicting traffic breakdown at bottlenecks. All traffic data during active traffic periods are used as training data.

### 3. Data Sampling

Three methods of sampling exist with respect to traffic congestion occurrence of one type: traffic breakdown was tested for accuracy improvement and was compared with results found from earlier studies. As described in this paper, the definitions of the training data differ among verifications because we fix the model and aim to improve the accuracy only by appropriately selecting the training data. Therefore, this section presents details of how to select the training data used for each verification. Unless otherwise noted, the traffic data used are divided into two parts: data of the period from April 2015 through November 2017, which are used as training data; and data of the period from April 2018 through March 2019, which are used as test data. For the training data, we devised each verification within that period. This facilitates comparisons by aligning the period with that of an earlier study [10–12].

#### 3.1 Sampling immediately before and after traffic congestion

For this verification, the training data are devised in the time-series direction to predict traffic congestion occurrence. Specifically, the training data are defined as a set of three steps in the time-series direction around the timing when traffic congestion occurs. Therefore, the training data consist of data immediately before and after the occurrence of traffic congestion. The purposes of using the sampling data are to eliminate data on traffic conditions far different from the occurrence of congestion from the training data, and to increase the ratio of data around the occurrence of traffic congestion in the training data. To evaluate the change in accuracy resulting from this sampling, parameter estimation was applied to training data of three types. Then the prediction accuracy was compared with that obtained using the test data. Each training dataset used is shown in Table I.

**Table I.** Definition of training data used in 3.1.

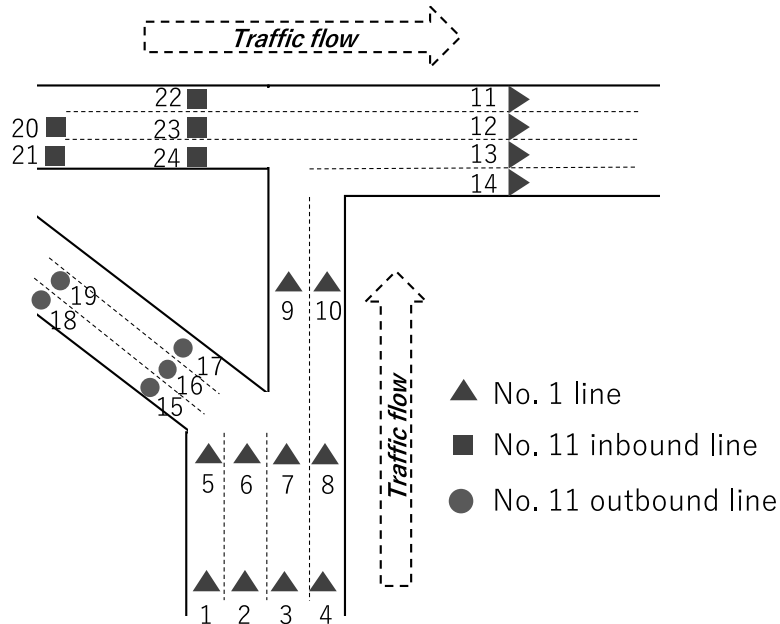
Training data	A	B	C
Definition	Vehicle detector data for No. 11 line		
Explained variable	Daily traffic congestion report data for forecasted locations		
Period	April 1, 2015 ~ November 30, 2017		
Constraints	None	6am-21pm only	3 time points before and after traffic congestion

Training data C is the newly proposed training dataset, which consists of a set of seven points in time: 10 min before the occurrence of the traffic congestion and three points before and after the occurrence of traffic congestion. There are two reasons for using these as the training data. The first is to balance traffic congestion labels and non-traffic congestion labels. The training data used for an earlier study [10, 11] (similar to A) were imbalanced data with a very large proportion of non-traffic congestion labels. Subsequently, we surmised that the accuracy could be improved by balancing the data. The second reason is improvement of the quality of each label. For this prediction task, predicting the occurrence of traffic congestion correctly is important. Therefore, we estimated that increasing the proportion of data near the occurrence of traffic congestion in the training data and making each label closer to the occurrence of traffic congestion would improve the prediction performance of the boundary. As comparison targets for data C, we prepared data A, which are the same as in the earlier study [10, 11]. We also prepared data B, which are intermediate between A and C and which consist of training data only during periods of sufficient traffic flow. The obtained results were compared with those reported from an earlier study [11].

#### 3.2 Refine the detector data

For this verification, we investigate the training data in both time and spatial directions to predict the occurrence of traffic congestion at 0.0 kp. First, for the time-series direction, the same division

of training data as in Section 3.1 is used. In the spatial direction, two additional modifications are made: The explanatory variables comprise only vehicle detectors in the vicinity of the prediction target point. The detector data are treated by lane rather than by aggregating data of all lanes. The purpose of these modifications is accurate observation of the intersections of traffic flow caused by the merging of routes and traffic conditions that vary from lane to lane. The spatial arrangement of the vehicle detectors used for this verification is shown in Fig. 2. Symbols in Fig. 2 indicate the location of the vehicle detectors used in this verification. Their shape indicates the route to which they belong. The number next to the detector is the identification number of that detector.



**Fig. 2.** Detectors used in Section 3.2.

To evaluate the change in accuracy attributable to this refinement, parameter estimation was performed for training data of two types. Then the respective prediction accuracies were compared. The training data definitions are presented in Table II. Of the contents of Table II, pattern  $\Gamma_2$  is the method proposed here. Pattern  $\Gamma_1$ , by contrast, uses detector data from the same locations, but aggregates the detectors (lanes) at each location into one detector dataset per location used as an explanatory variable. The obtained results were also compared with those of an earlier study [12].

**Table II.** Definitions of training data used for Section 3.2.

	Explanatory variable		Devices for time sequence direction	number of points in time
	Definition	Total		
$\Gamma_1$	Consolidate data by location	32	3 steps before and after a traffic congestion	6519
$\Gamma_2$	Per detector	96		

### 3.3 Prediction of congestion spillback

For this validation, prediction is performed at 9.0 kp, where traffic congestion occurs because of the extension of traffic congestion originating at 4.0 kp. The intention here is to learn as much traffic congestion and non-traffic congestion data as possible using detector data at all spatial points and times on the No. 11 inbound and outbound lines. We use the dataset for learning the same as data A in Table I in Section 3.1. The results obtained were then compared to those reported from an earlier study [11].

## 4. Results

This section presents results obtained from each of the three verifications described in the Data sampling section herein.

### 4.1 Results for temporal data sampling

The prediction accuracy obtained for each pattern is presented in Table III. However, for this study, only the bottleneck point of 4.0 kp, where the results are characteristic, is shown. From this table, it is apparent that, at 4.0 kp, data C produces the most accurate prediction. Furthermore, the prediction results obtained using training data C are compared with those of previous studies [11], as shown in Table IV. The congestion prediction using training data C is apparently more accurate than the prediction obtained from an earlier study at 4.0 kp [11], even though the data in the time-series direction are far fewer.

**Table III.** Prediction accuracy of training data.

Training Data	Precision	Recall	F1 value
A	0.685	0.662	0.673
B	0.72	0.718	0.719
C	0.75	0.878	0.809

**Table IV.** Results of Section 3.1 compared with the accuracy of earlier studies (Convolutional neural networks (CNN) models).

Model	Precision	Recall	F1 value	Length of time series
Linear model(C)	0.75	0.878	0.809	13886
CNN	0.74	0.856	0.794	280497

### 4.2 Results for spatial data sampling

The results obtained for accuracy are presented in Table V. The F1 value in the training data for  $\Gamma_2$  is more than 0.1 points higher than that for  $\Gamma_1$ . The F1 value in the training data for  $\Gamma_2$  is almost equal that obtained using Graph convolutional neural networks (GCN) in the earlier study, even though they have fewer parameters.

**Table V.** Results of Section 3.2 with the accuracy of earlier studies (GCN).

Model	Precision	Recall	F1 value	Number of parameters
Linear model( $\Gamma_1$ )	0.374	0.154	0.219	32
Linear model( $\Gamma_2$ )	0.51	0.268	0.351	96
GCN	-	-	0.385	2163

### 4.3 Results for predicting traffic spillback congestion

The accuracy verification results are presented in Table VI. The results indicate that the linear model can predict the 9.0 kp stretch, where traffic congestion occurs frequently because of the extension, with accuracy that is superior to that achieved using the CNN model.

**Table VI.** Results of 3.3 compared with the accuracy reported from an earlier study (CNN).

Model	Precision	Recall	F1 value
CNN	0.718	0.818	0.763
Linear model	0.758	0.837	0.796

## 5. Discussion

### 5.1 Discussion for temporal data sampling

Based on the verification results, we discuss the training data tendency of C by analyzing the estimated parameters (Elements of  $W_{out}$ ). The estimated  $W_{out}$  from each training data are sorted in ascending order of absolute value. Then the trends are analyzed. The top 20 parameters are presented in Table VII.

**Table VII.** Parameters  $W_{out}$  estimated for training data at 4.0 kp.

A		B		C	
Variable (DataType,Point[kp])	Wout	Variable (DataType,Point[kp])	Wout	Variable (DataType,Point[kp])	Wout
('d_occupancy', 4.5)	0.85	('avg_velocity', 4.1)	-0.959	('avg_velocity', 4.1)	-0.682
('d_occupancy', 4.1)	0.746	('flow', 3.5)	-0.388	('avg_velocity', 3.5)	-0.337
('avg_velocity', 4.1)	-0.714	('flow', 5.1)	0.366	('d_occupancy', 4.1)	0.321
('flow', 3.5)	-0.428	('avg_velocity', 4.5)	-0.36	('flow', 5.1)	0.306
('flow', 5.1)	0.36	('d_occupancy', 4.1)	0.36	('avg_velocity', 4.5)	-0.295
('flow', 4.1)	0.224	('d_occupancy', 4.5)	0.359	('flow', 6.5)	0.261
('d_occupancy', 3.0)	0.22	('avg_velocity', 5.1)	-0.258	('flow', 3.5)	-0.235
('avg_velocity', 5.1)	-0.206	bias	0.195	('flow', 3.0)	-0.208
('flow', 3.0)	-0.18	('flow', 3.0)	-0.194	('avg_velocity', 3.0)	-0.192
('d_occupancy', 3.5)	0.164	('flow', 4.1)	0.169	('flow', 6.2)	0.169
('avg_velocity', 1.0)	0.161	('d_occupancy', 3.0)	0.136	('flow', 7.0)	0.149
('avg_velocity', 1.4)	0.134	('avg_velocity', 1.0)	0.135	('flow', 8.6)	0.144
('flow', 6.5)	0.132	('avg_velocity', 1.8)	0.125	bias	0.142
('flow', 4.5)	-0.113	('flow', 6.5)	0.123	('flow', 8.0)	0.142
('flow', 16.7)	-0.112	('d_occupancy', 11.5)	0.103	('flow', 4.1)	0.133
('avg_velocity', 0.5)	0.111	('avg_velocity', 0.5)	0.092	('flow', 1.8)	-0.133
('flow', 17.8)	0.109	('avg_velocity', 1.4)	0.088	('flow', 4.5)	0.132
('d_occupancy', 5.5)	0.096	('d_occupancy', 1.8)	0.087	('flow', 5.5)	0.131
bias	0.09	('avg_velocity', 12.4)	0.086	('avg_velocity', 7.0)	-0.125
('high_flow', 10.5)	0.089	('d_occupancy', 1.4)	0.081	('avg_velocity', 8.0)	-0.116

Comparison of datasets A, B, and C indicates that the ratio of detector occupancy (d-occupancy) becomes smaller in the order of A, B, and C. The ratio of average velocity (avg-velocity) and flow (flow) becomes larger instead. For the following reason, C has a smaller ratio of detector occupancy than the other two.

Congestion and non-congestion labels applied in this study indicate whether the location will be congested 10 min later. Therefore, when one location has a congestion label, the traffic data in the vicinity of the location may have two conditions: non-congested and congested. They are designated respectively as congestion occurrence and as congestion continuation. Predicting the occurrence of traffic congestion necessitates prediction of traffic congestion from the traffic data when the location is not congested. For that reason, data with congestion occurrence labels are important. Particularly at bottlenecks, data with congestion occurrence labels are more important because locations around the bottlenecks are in a non-congested state. However, the percentage of congestion occurrence labels is very small. Therefore, if all the congestion labels are trained uniformly, then the parameter is emphasized by the congestion-continuation labels. It should be noticed that the congestion occurrence labels are assigned at a single time point, although the evaluation for correct prediction has a certain time interval. This is because it is sufficient that the time performing prediction is before the



**Table VIII.** Average detector occupancy for each label in 4.1 kp.

4.0 kp label	Mean normalized detector occupancy (4.1 kp)
Congestion continuation	0.375
Congestion occurrence	0.152
Non-congestion	0.101

occurrence and the predicted time is earlier than the actual occurrence.

As might be apparent from Table VIII, for data with congestion continuation labels, the average detector occupancy is extremely high relative to the non-congestion labels. Seeing the occupancy of neighboring detectors is crucially important to ascertain the congestion label. Therefore, the detector occupancy weights are greater for the training data of A and B.

The training data for C are a set of data for seven time points, which include three time points both before and after the time at which traffic congestion occurs. Therefore, out of the seven time points, three time points before the occurrence of a traffic congestion are labeled as non-congested and three time points after are labeled as congested or non-congested, depending on the continuation status of the traffic congestion. Consequently, the number of time points with a congestion label is one to four out of seven, including the time at which the traffic congestion occurs. Therefore, it can be inferred that the parameters for predicting the occurrence of traffic congestion are estimated more accurately than in the case of A and B. The mean of the values of detector occupancy under the congestion occurrence label is not different from that under the non-congestion label. Here, it is considered that the weight of detector occupancy became smaller. The weights of average velocity and traffic flow became larger instead.

Based on the discussion presented above, one can infer that the proposed method is suitable for predicting congestion at bottlenecks because the training data are balanced in the ratio of congested and non-congested labels. The proportion of data with congestion labels with points that are non-congested is greater than in cases A and B.

**Table IX.** Top 15 absolute values of all parameters estimated using training data  $\Gamma_2$  in 4.2, in ascending order.

Variable (DataType, Detector No.)	Wout
bias	0.315
(d_occupancy, 23)	0.237
(flow, 12)	-0.232
(d_occupancy, 24)	0.204
(avg_velocity, 12)	-0.185
(avg_velocity, 11)	-0.184
(avg_velocity, 13)	-0.181
(avg_velocity, 23)	-0.176
(flow, 11)	-0.149
(avg_velocity, 14)	-0.145
(flow, 21)	0.144
(flow, 24)	-0.132
(flow, 14)	-0.117
(high_flow, 22)	0.116
(flow, 22)	0.107

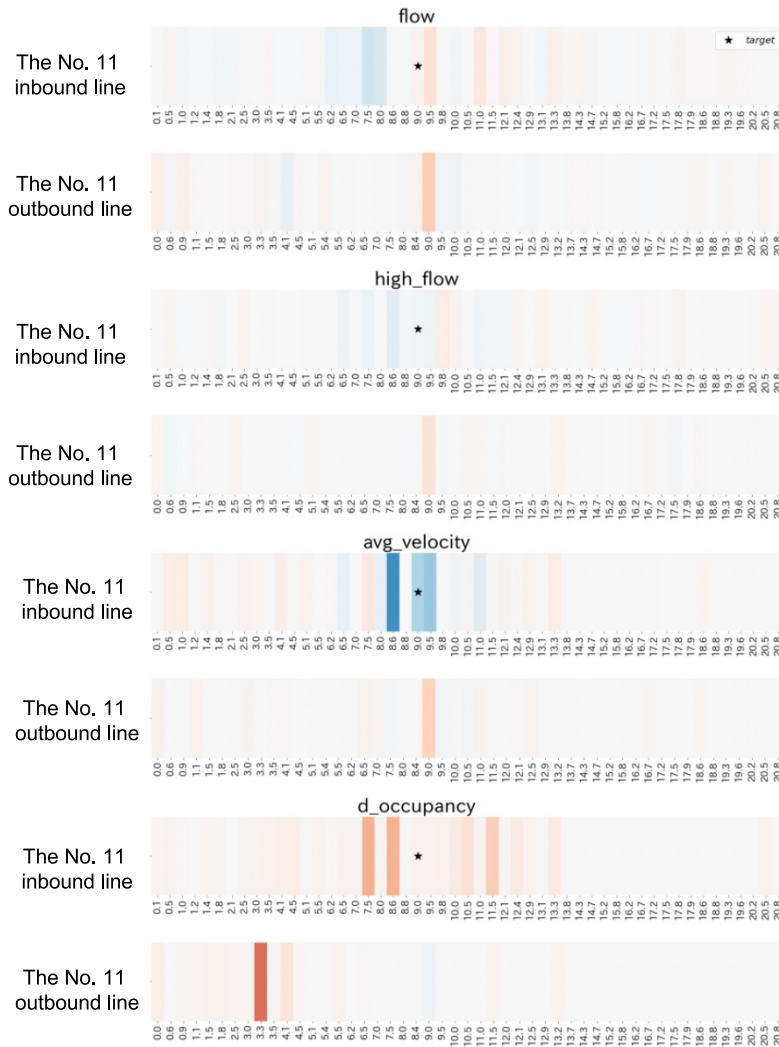
## 5.2 Discussion for spatial data sampling

To discuss reasons for the improved accuracy in prediction using training data  $\Gamma_2$ , we can analyze the parameters (Elements of  $W_{out}$ ) in the case of  $\Gamma_2$ . The parameters estimated from the training data are sorted in ascending order of absolute value. Then their tendencies are analyzed. The top 15 parameters are shown in Table IX.

From Table IX and the positions of detectors represented in Fig. 2, we can see the importance of the detector data for the two lanes from the right (23,24) on the 0.1 kp of No. 11 inbound line and the data for all lanes (11,12,13,14) on the 3.2 kp of No. 1 line. Only the right lane of the 0.1 kp of No. 11 inbound line is specifically examined. Less attention is devoted to the left lane (22), probably because only the right lane shows signs of congestion 10 min before the onset of traffic congestion. Accordingly, one can infer that training data  $\Gamma_2$ , by treating the vehicle detector data lane by lane, enables more accurate observation of traffic flow from the right lane of the 0.1 kp of No. 11 inbound line to the 3.2 kp of No. 1 line. It improves the predictive accuracy by extracting precursor of the occurrence of traffic congestion sensitively at the 0.0 kp of the No. 11 inbound line.

## 5.3 Discussion for predicting traffic spillback congestion

Next we discuss the factors which made the linear model prediction as accurate as deep learning (CNN) at 9.0 kp, where the congestion spillback is the cause of the congestion. First, the magnitudes of the estimated parameters are depicted in Fig. 3, where red is positive and blue is negative. Darker colors



**Fig. 3.** Parameters assigned to each explanatory variable in the linear model in 4.3. Red signifies positive and blue stands for negative. Darker colors represent greater values.

colors signify the magnitudes of the absolute values of  $W_{out}$ . Star symbols shown in each graph denote the predicted location. Figure 3 shows that  $W_{out}$  is larger near the star symbols in the average velocity and the detector occupancy. The results indicate that prediction in this model emphasizes data near the location of interest.

In light of these results obtained using  $W_{out}$  analysis, probably the linear model predicts whether traffic congestion will extend 10 min later at the target location in consideration of traffic data such as upstream traffic flow and average speed, while recognizing traffic congestion downstream of the target location. Consequently, because this linear model is good at predicting the extension of traffic congestion, it might have been able to outperform the deep learning model in terms of accuracy at 9.0 kp, which is not the occurrence location of traffic congestion.

## 6. Conclusion

---

We have demonstrated that the accuracy of predicting traffic congestion on expressways can be improved by selecting the training data appropriately. Moreover, by analyzing the parameters, we were able to explain the factors and processes responsible for improvement of the prediction accuracy. Subsequently, we were able to establish that the linear model is highly explanatory in its predictions.

As described in this paper, we have strived to reduce the volume of the training data in the time-series and spatial directions. Because these modifications drastically reduce the amount of data to be trained while maintaining the precision of prediction performance, they are not expected to function in deep learning models that require big data for training. Consequently, the learning time of the current method was much shorter than that of deep learning. It can be anticipated that the performance of deep learning with the reduced dataset would be much worse than our method due to the lack of the amount of data.

Future works must underscore the effectiveness of the linear model for predicting the occurrence of traffic congestion using the following two steps: extraction of the set of data characterized by the important parameters for prediction, and identification of what characteristic data are captured from those parameters and used to predict traffic congestion. Furthermore, by analyzing the parameters, we can explain which quantity improves accuracy, i.e. explainability, by virtue of the linear properties of the prediction method.

## Acknowledgments

---

This research was partially supported by JST MIRAI No. JPMJMI19B1, SECOM Science and Technology Foundation, AMED under Grant Number JP21zf0127005 and Hanshin Expressway Company, Ltd.

## References

---

- [1] Ministry of Land, Infrastructure, Transport and Tourism, "Traffic congestion countermeasures in urban areas - Road Improvement for Urban Regeneration -," 2013, <https://www.mlit.go.jp/common/000043136.pdf>. (Viewed October 4, 2022).
- [2] Ministry of Land, Infrastructure, Transport and Tourism "ITS overall concept," <https://www.mlit.go.jp/road/ITS/j-html/5Ministries/>. (Viewed October 4, 2022).
- [3] X. Ma, H. Yu, Y. Wang, and Y. Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory," *PLoS ONE* 10(3): e0119044, 2015. DOI: 10.1371/journal.pone.0119044
- [4] M. Fouladgar, M. Parchami, R. Elmasri, and A. Ghaderi, "Scalable deep traffic flow neural networks for urban traffic congestion prediction," *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017. DOI: 10.1109/IJCNN.2017.7966128
- [5] Y. Liu and H. Wu, "Prediction of road traffic congestion based on random forest," *2017 Tenth International Symposium on Computational Intelligence and Design (ISCID)*, pp. 361–364, 2017. DOI: 10.1109/ISCID.2017.216

- [6] M. Chen, X. Yu, and Y. Liu, "PCNN: Deep convolutional networks for short-term traffic congestion prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 11, pp. 3550–3559, November 2018. DOI: 10.1109/TITS.2018.2835523
- [7] N. Kumar and M. Raubal, "Applications of deep learning in congestion detection, prediction and alleviation: A survey," *Transportation Research Part C: Emerging Technologies*, vol. 133, 2021. DOI: 10.1016/j.trc.2021.103432
- [8] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep learning: A generic approach for extreme condition traffic forecasting," *2017 SIAM International Conference on Data Mining*, pp. 777–785, 2017. DOI: 10.1137/1.9781611974973.87
- [9] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," 2017. arXiv preprint arXiv:1707.01926
- [10] T. Okamoto, H. Ando, K. Wada, R. Mukai, Y. Nishiumi, and D. Tamagawa, "Predicting traffic breakdown in urban expressways based on simplified reservoir computing," *AAAI 21 Workshop on AI for Urban Mobility*, 2021.
- [11] R. Mukai, S. Sakuragi, D. Tamagawa, M. Yamamoto, G. Hatayama, T. Hirano, S. Kase, K. Suzuki, Y. Kojima, and T. Teramae, "A study on traffic congestion prediction method on hanshin expressway using machine learning (CNN)," *17th ITS Symposium in Japan*, 2019.
- [12] T. Teramae, R. Mukai, K. Suzuki, Y. Kojima, and A. Abe, "Jam congestion prediction on hanshin expressway with graph convolutional networks," *18th ITS Symposium in Japan*, 2020.