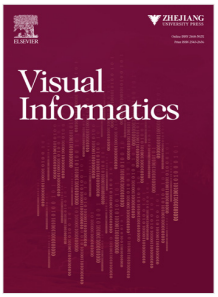
Journal Pre-proof



TCEVis: Visual analytics of traffic congestion influencing factors based on explainable machine learning

Jialu Dong, Huijie Zhang, Meiqi Cui, Yiming Lin, Hsiang-Yun Wu, Chongke Bi

PII: S2468-502X(23)00053-0
DOI: https://doi.org/10.1016/j.visinf.2023.11.003
Reference: VISINF 193

To appear in: *Visual Informatics*

Received date : 26 July 2023
Revised date : 4 November 2023
Accepted date : 6 November 2023

Please cite this article as: J. Dong, H. Zhang, M. Cui et al., TCEVis: Visual analytics of traffic congestion influencing factors based on explainable machine learning. *Visual Informatics* (2023), doi: https://doi.org/10.1016/j.visinf.2023.11.003.

Title Page (with Author Details)

TCEVis: Visual Analytics of Traffic Congestion Influencing Factors based on Explainable Machine Learning

Jialu Dong, School of Information Science and Technology, Northeast Normal University, Changchun 130117, China.

Huijie Zhang, School of Information Science and Technology, Northeast Normal University, Changchun 130117, China. (Corresponding author: zhanghj167@nenu.edu.cn)

Meiqi Cui, School of Information Science and Technology, Northeast Normal University, Changchun 130117, China, and School of Computing Science, Baicheng Normal University, Baicheng 137000, China.

Yiming Lin, School of Information Science and Technology, Northeast Normal University, Changchun 130117, China.

Hsiang-Yun Wu, St. Pölten University of Applied Sciences, Austria.

Chongke Bi, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.

TCEVis: Visual Analytics of Traffic Congestion Influencing Factors based on Explainable Machine Learning

Jialu Dong[a], Huijie Zhang[a,*], Meiqi Cui[a,b], Yiming Lin[a], Hsiang-Yun Wu[c] and Chongke Bi[d]

[a]School of Information Science and Technology, Northeast Normal University, Changchun 130117, China.
[b]School of Computing Science, Baicheng Normal University, Baicheng 137000, China.
[c]St. Pölten University of Applied Sciences, Austria.
[d]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.

ARTICLE INFO

Keywords:
Visual analytics
Speed prediction
Explainable
Traffic congestion

ABSTRACT

Traffic congestion is becoming increasingly severe as a result of urbanization, which not only impedes people's ability to travel but also hinders the economic development of cities. Modelling the correlation between congestion and its influencing factors using machine learning methods make it possible to quickly identify congested road segments. Due to the intrinsic black-box character of machine learning models, it is difficult for experts to trust the decision results of road congestion prediction models and understand the significance of congestion-causing factors. In this paper, we present a model interpretability method to investigate the potential causes of traffic congestion and quantify the importance of various influencing factors using the SHAP method. Due to the multidimensionality of these factors, it can be challenging to visually represent the impact of all factors. In response, we propose TCEVis, an interactive visual analytics system that enables multi-level exploration of road conditions. Through three case studies utilizing actual data, we demonstrate that the TCEVis system offers advantages for assisting traffic managers in analyzing the causes of traffic congestion and elucidating the significance of various influencing factors.

1. Inroduction

Due to accelerated economic expansion and urbanization, the number of motor vehicles is increasing rapidly. This presents substantial challenges and stresses for urban road traffic. The inadequacies of conventional road planning and design have begun to manifest themselves, resulting in increasingly severe traffic congestion issues. It also contributes to environmental pollution (Lu et al., 2021). In addition, it inhibits the intelligent development and modernization of urban transportation systems (Ajay et al., 2022). In the sphere of transportation, addressing the problem of traffic congestion has become an urgent and crucial concern. The objective is to accommodate diverse travel needs while effectively reducing energy consumption. Traditional methods for identifying the causes of congestion rely largely on time-consuming and costly human labor. Numerous studies use traffic prediction models to predict future traffic conditions (Pan et al., 2019; Kosugi et al., 2022). On the basis of historical traffic data, machine learning (ML) techniques are widely employed to analyze traffic patterns and identify road segments likely to experience congestion. However, machine learning (ML) models are frequently perceived as "black boxes" (Linardatos et al., 2021), making it challenging for experts to trust the model's decision-making capability and comprehend why congestion occurs. If drivers have access to information about the causes of congestion, they can dynamically modify their routes to mitigate its effects.

There are a number of factors that contribute to traffic congestion, including inadequate road infrastructure and unpredictability. To enhance the accuracy of predicting congestion, it is essential to take into account the interaction between these various factors. However, manually identifying all of the causes of traffic congestion events in map applications can be laborious and time-consuming. Visualization techniques offer significant advantages in the analysis of diverse and heterogeneous data sources, enabling users to explore the data through intuitive visual representations. Several research investigations on visual analytics of traffic data have been conducted (Ferreira et al., 2013; Sobral et al., 2019; Clarinval and Dumas, 2022), including the use of heat maps to display traffic flow (Song and Miller, 2012) and glyphs to depict the causes of traffic congestion on roads. These visualizations aid traffic administrators in analyzing traffic problems effectively. By employing interactive operations, visual analytics systems considerably aid users in discovering the underlying relationships between congestion and its influencing factors. This lays the groundwork for the development of effective traffic management solutions.

We collaborated with experts in the fields of transportation and machine learning to gain a deeper understanding of the causes of traffic congestion. To determine whether traffic congestion exists on the target roadways, we applied a model based on machine learning to estimate the short-term speed of roads in the future. However, because of the model's opacity, it is challenging for traffic managers to comprehend why congestion arises. Therefore, we used the SHAP method (Lundberg and Lee, 2017) to explain the speed prediction model and quantify the degree of influence of multi-source factors on congested roads. The method aids traffic managers to analyze the relationship between multi-source influencing factors and traffic congestion. We produced the interactive

* Corresponding author: Huijie Zhang
E-mail addresses: zhanghj167@nenu.edu.cn (H. Zhang)

visual analytics system **TCEVis**. In TCEVis, there are six different views available, including the global view, relationship view, monitor view, map view, matrix view, and temporal view. These views encourage two levels of investigation: comparative analysis of several causes of traffic congestion and temporal analysis of a specific cause of traffic congestion. Through case studies that make use of actual data, we have successfully demonstrated the effectiveness of our methodology in analyzing the factors that contribute to traffic congestion and improving traffic management. Our contributions can be summarized as follows.

- We presented a technique that examined the root causes of traffic congestion using the interpretability of machine learning. Our method successfully simulated the relationship between congestion and many contributing factors. When congestion occured, we specifically utilized the SHapley Additive exPlanation (SHAP) method to quantify the extent of their influence in an understandable manner.
- We developed TCEVis, a visual analytics system with a number of interactive views that facilitated the understanding of the causes of traffic congestion. The system enabled users to compare and examine the factors contributing to traffic congestion on different roadways. Additionally, it offerd assistance for performing granular analyses of traffic congestion on certain roadways.
- The usefulness of TCEVis was demonstrated in three case studies utilizing actual data for analyzing the similarity of influencing factors on road congestion, identifying the causes of congestion on a single road, and evaluating whether congestion was cyclical. These case studies demonstrated how TCEVis effectively and accurately completed certain analytical tasks.

2. Related work

2.1. Urban traffic prediction

Experts in the field of traffic flow analysis and planning can benefit from traffic prediction to better manage traffic congestion by analyzing the spatial and temporal variations in traffic flow.

Abadi et al. (Abadi et al., 2015) presented a technique for rapidly forecasting traffic flow on all links in a traffic network. They developed an autoregressive model that considered both past flow data and flow uncertainty. However, this work had limitations due to insufficient data to conduct additional testing in both incident and regular traffic scenarios.

The intricate spatial and temporal relationships inherent in traffic data were considered when forecasting traffic. Zhang et al. (Zhang et al., 2021) introduced an innovative model known as the Evolving Temporal Graph Convolutional Network (ETGCN), recognizing the limitations of GCN in capturing the evolving spatial correlations within the road network. This model aimed to learn the spatial-temporal correlations and their varying states for predicting traffic speeds within a road network. Li et al. (Li and Lasenby, 2021) proposed a spatial-temporal graph attention network adapted from the attention mechanism for road speed prediction. The model included a multi-headed map attention module to capture the spatial correlation between road segments. Wang et al. (Wang et al., 2022) introduced the transport flow model, which effectively analyzed the spatial-temporal correlation through the fusion of graph attention and time attention layers. The model is an attention-based spatial-temporal graph neural network model (AST-GAT). However, this research did not address the influence of multi-scale information on prediction.

The road network structure was also considered a significant factor in traffic prediction. To capture the temporal dynamics within the speed dataset, Lee et al. (Lee et al., 2019) utilized a straightforward LSTM model. To enhance prediction accuracy, they incorporated DeepWalk-generated road network structure data, encompassing network features, peak-hour information, and spatial-temporal road speed. To completely exhibit the geographical information, the structure of the road network was illustrated as a graph. From this graph representation, a graph convolutional neural network optimized for traffic forecasting was introduced (Guo et al., 2020).

In order to address the challenge of traffic congestion, this research focused on traffic speed prediction. We incorporated a variety of data points that affect traffic speed, including the spatial organization of the road network, temporal characteristics of traffic flow data, weather conditions, air quality, and more, with the aim of enhancing prediction accuracy.

2.2. Model interpretability

Machine learning's prediction abilities are getting better as computing power grows rapidly. People are now seeking a deeper knowledge of the underlying causes behind model decisions since they are no longer content with simply relying on the performance of the present models. As a result, the interpretation of machine learning models has emerged as a significant area of research. Only decisions resulting from interpretable models are likely to obtain public approval, especially in high-risk fields like healthcare, finance, and justice.

Two popular methods for interpreting models are post-hoc explanation and self-explanation. The interpretability technique employed in this article utilized the SHAP model, which has been widely recognized and validated across numerous research domains (Lundberg and Lee, 2017). To investigate the impact of road and environmental factors on accident severity and to clarify the XGBoost model, Li et al. (Li et al., 2022) introduced SHAP values. Ji et al. (Ji et al., 2022) utilized the SHAP model to determine the SHAP values of factors affecting road travel, enabling the assessment of the degree to which each element influences the dependent variable, ultimately improving prediction accuracy. Bialek

et al. (Białek et al., 2022) applied SHAP to energy consumption projection models, providing valuable and easy-to-understand insights into the internal workings of these models, which can be used for evaluating their reliability and planning for future development. Li et al. (Li, 2022) demonstrated that locally interpreted machine learning models can replace spatial statistical models, especially when dealing with complex geographic and non-spatial factors, surpassing the capabilities of traditional spatial statistical models. Lin et al. (Lin and Gao, 2022) investigated the interpretability of risk detection models and introduced the clustered SHAP approach for evaluating various firm capacities, including profitability, liquidity, and more. Using the SHAP-XGBoost algorithm, Zhang et al. (Zhang et al., 2023) developed a framework for explaining the landslide sensitivity assessment models. They addressed model generalizability variation across different landscapes and analyzed the regional features and geographical heterogeneity of landslide influences.

The use of visual analytics tools for examining the interpretability of machine learning models has been a subject of recent research. Kahng et al. (Kahng et al., 2017) developed an interactive visual analytics system for real-time interpretation of massive deep learning models. Users of this technology were capable of examining the outcomes of complex deep neural network models at both the instance and subset levels. Zhao et al.'s (Zhao et al., 2018) proposed a system for graphical explanations of random forest models and predictions, contributing to a clearer understanding of the model's internal mechanisms. Cheng et al. (Cheng et al., 2021) designed the visual analytics tool VBridge, which facilitated the interpretation of machine learning methods in the decision-making process, particularly in real clinical scenarios.

2.3. Traffic conditions visualization

Visualization technology is used to construct traffic management systems that enable the coordination of multiple traffic elements, increasing the level of information and efficiency in urban traffic management. Liu et al. (Liu et al., 2019) designed a visual analytics system to support users in examining movement patterns in trajectory data. There are several difficulties in managing urban transportation, with one of the main problems being traffic congestion. Zhao et al. (Zhao et al., 2022) designed a system that displays the uncertainty of bus travel time. To assist users with perception and decision making by calculating bus uncertainty information.

To enable effective congestion studies, Lee et al. (Lee et al., 2019) developed the VSRivers visualization to concurrently present volume and congestion conditions. The primary goal was to assist users in comprehending and predicting congestion issues, including their underlying causes and propagation patterns. However, it is essential to note that this method did not account for environmental variables such as weather and air quality in the anticipation of traffic congestion.

Addressing the impact of traffic on air quality has emerged as a growing concern in the field of traffic visualization, given that vehicle transportation significantly contributes to air pollution. Chiara et al. (Bachechi et al., 2021) developed a visual analytics dashboard that provided an effective approach for assessing urban traffic data in both spatial and temporal dimensions. This dashboard facilitated the analysis of traffic congestion in specific locations at specific times and allowed for the evaluation of the impact of currently operating vehicles on urban air quality. However, it is worth noting that some additional visualizations contrasting the spatial distribution of traffic over time were not created.

Approaches employing deep learning have consistently demonstrated strong performance in predicting traffic flow. However, the lack of transparent nature of these models has constrained experts' ability to comprehend the influence of input data on outcomes. Jiang et al. (Jiang et al., 2022) proposed a solution named TrafPS to address this challenge. This system was grounded in the Shapley value and was designed to facilitate the interpretation of predicted traffic flow. However, only considering the number of roads when clustering the urban grid omitted other characteristics that might have provided more information. Jin et al. (Jin et al., 2022) collaborated with experts to develop AttAnalyzer, a system that allowed users to investigate how deep learning models generate predictions.

We explored the temporal fluctuations in traffic patterns as well as the effects of weather and holidays on traffic conditions, drawing inspiration from the work mentioned above. Additionally, we utilized model interpretability methodologies to assist specialists in comprehending the extent to which certain elements influenced traffic situations. These findings contributed to the development of improved management approaches.

3. Overview

3.1. Tasks and requirements

To determine specific analysis tasks, we worked with three domain experts. A traffic manager with ten years of expertise in traffic management is the first expert (E1). E1 offered insightful requirements and useful insights in the field of traffic congestion studies. E2 has eight years of relevant experience in developing transportation-related software and has a high level of application of machine learning techniques. The third expert (E3) , who represents a transport firm, has six years of experience in the construction of smart transport systems. According to the experts' conversations, the main challenge with traffic management right now is congestion. Previous researches have used machine learning techniques to forecast short-term future road speeds, identify congested regions, and put remedial measures in place. However, it can be difficult to intuitively investigate the many elements due to the varied sources of influencing factors. Additionally, the model's intrinsic black box character makes it difficult to explain how different elements specifically affect

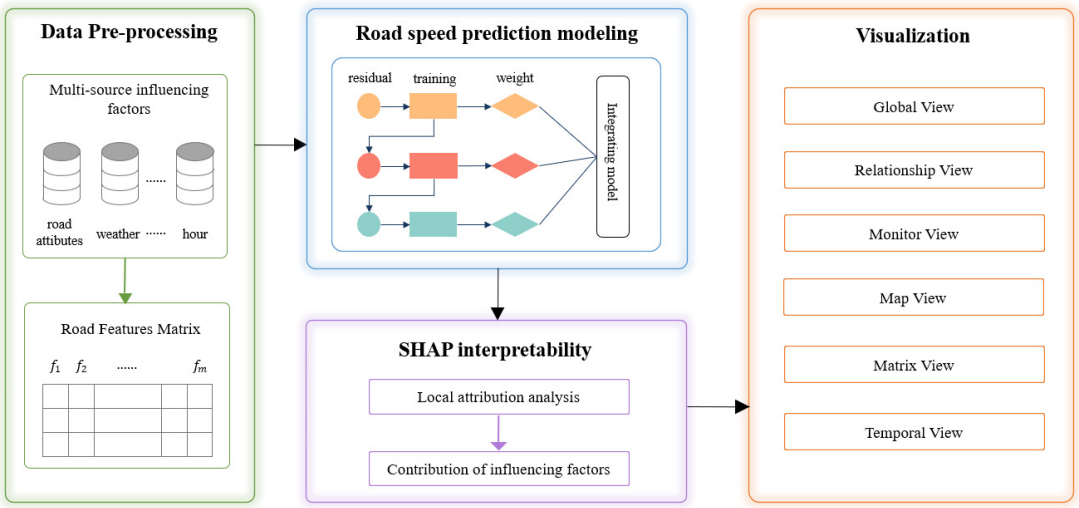


Fig. 1: The analysis pipeline of the system includes data pre-processing, road speed prediction modeling, SHAP interpretability and visualization. The data pre-processing component integrates the multiple influencing factors from various sources and constructs a road feature matrix as the input for the road speed prediction model. The SHAP model interpretability method quantifies the importance of the influencing factors on road speed. Finally, a visual analytics system is designed to assist users in analyzing the causes of road congestion.

road speed. It is crucial to take into account the impact of particular places at particular times on road speeds while analysing the causes of congestion.

The following are the issues that require attention:

Q1: Where are the congested roads mainly concentrated?

Q2: When does traffic congestion mainly occur?

Q3: How to describe the level of impact of several factors on the anticipated congested roads?

Q4: How can the probable relationship between the range of influence factors' values and the degree of their influence be analysed?

The following requirements have been established for a visual analytics system to investigate the variables affecting traffic congestion. The system's analytical pipeline is displayed in Fig. 1.

R1: The system should enable mapping congested roads' actual locations and investigating the spatial correlation of congestion occurrence (Q1).

R2: The system ought to track changes in traffic conditions over time and investigate the temporal relationships between congestion occurrence (Q2).

R3: The system should enable investigation of the degree of which multi-source factors affect road speed and allow comparison of the possible potential causes of congestion on different roads (Q3).

R4: The relationship between the domain of influence factor values and their level of influence on road speed should be highlighted by the system (Q4).

3.2. Data description

We use four datasets in our research to account for the effects of multiple factors on road speed: taxi trajectory data, air quality data, meteorological and weather data, and road network data.

Taxi trajectory data is a collection of information on the recorded movements of taxis in a city during the month of October 2021. This data includes details like the trajectory ID, vehicle ID, time, latitude and longitude, vehicle speed, etc.

Air quality data refers to information gathered from China's General Environmental Monitoring Station. This data contains measurements of different air pollutants taken in real-time, including "AQI", "PM2.5", "PM10", "SO_2", "NO_2" and "CO". Their 24-hour sliding averages are also provided. Additionally, the data includes 8-hour sliding averages and 24-hour maximum values for "O_3", as well as real-time concentrations and 24-hour maximum values for "O_3".

Meteorological and weather data are details about past weather conditions that can be found on historical websites. The maximum temperature, minimum temperature, meteorological conditions, and wind direction are all included in this data.

Road network data consists of a dataset extracted from OpenStreetMap that provides information about a city's road network. This dataset contains various details such as the section ID, the start and end latitude and longitude coordinates of each road section, the length of each section, the level of the road section and the name of the section if available.

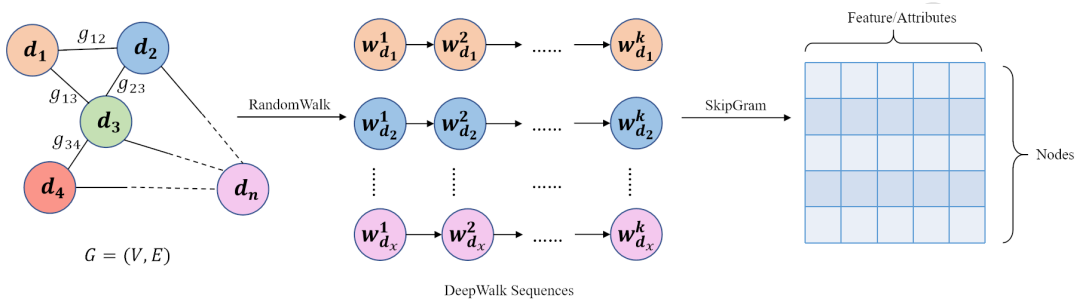


Fig. 2: The road embedding representation is generated using the DeepWalk algorithm.

3.3. Data Preprocessing

3.3.1. Road Network Matching

To solve the issue of poor matching between the trajectory data and the real road, we implemented a segmentation process for the trajectory points. Two main factors, namely the time interval and the distance interval between adjacent trajectory points, are considered in this study. The majority of the time interval is within 5 seconds. To accommodate any potential missing time intervals, the detector is permitted to sample up to 5 missing times. As a result, the time threshold is established at 25 seconds. The distance threshold has been established at 400 metres, considering the fault tolerance of the detector. Overall, these segmentation techniques enhance the precision and dependability of the trajectory data by imposing suitable constraints on both time and distance intervals among adjoining trajectory points.

Taking into consideration sampling errors, there is a possibility that the trajectory points could experience minor deviations. The matching of the road network is performed using the ST-Matching algorithm, which considers topological information such as the distance between GPS sampling points and roads.

3.3.2. Road Location Relationship

By examining the potential relationship between spatial road locations and congestion, we employ the DeepWalk algorithm to produce an efficient spatial representation of the original road network. Construct a road network graph $G = (V, E)$, where road segments are defined as nodes $V = (d_1, d_2, ..., d_x)$ and x represents the number of nodes. The edges of G are represented by $E = \{g_{ij}\}^x_{i,j}$, where i, j denote the road segment indices, and $g_{i,j} = 1$ if two segments are connected. Otherwise, $g_{i,j} = 0$.

The structure of the road network under DeepWalk is presented in (Fig. 2). There $w_{d_i} = \left\{ w^1_{d_i}, w^2_{d_i}, ..., w^k_{d_i} \right\}$, w_{d_i} represents the random walk with the ith road as the root, k denotes the length of the random walk. By considering the impact of neighboring roads on the target road, this study represents each road with a five-dimensional embedding feature vector.

Table 1
Specific composition of the road feature vector.

Influencing feature	Dimension
Velocity of the target road and the four shortest adjacent roads at time t	5
Target road embedding representation	5
Target road speed variation entropy	1
Target road level	1
Air quality of a city at time t	15
The weather conditions in a city at time t	14
The wind force at time t in a city	9
Temperature of a city at time t	2
The specific hour in the 24 hours to which the moment t belongs	24
Is it a holiday at time t	1

3.3.3. Road Feature Vector

Our study examines the effects of different factors on road speed, such as speed entropy, weather, air quality, holidays, and further variables. We construct a road feature vector as an input to the model for predicting short-term future road speed, and the feature vector is shown in Table 1. To distinguish the internal relationship among features, we propose a self-defined mapping method for the multidimensional spatial-temporal data in this study. This method maps similar influencing features into a shared space, thereby presenting them as one influencing factor.

3.3.4. Speed Performance Index

Road speed is commonly utilized as the principal indicator of traffic conditions. To evaluate traffic conditions, the road Speed Performance Index is utilized. This is calculated by dividing the actual vehicle speed by the maximum allowable travel speed. This can be represented by the following equation.

$$R_v = \frac{v}{v_{max}} \times 100 \tag{1}$$

where R_v represents the speed performance index and $R_v \in [0, 100]$, v denotes the mean speed of the current road on one time slice, and v_{max} denotes the maximum driving speed of the current road across all time slices. We have established three thresholds, namely 25, 50 and 75, based on the speed performance index as the classification standards for urban road traffic conditions, which is illustrated in Table 2.

Table 2
Classification of speed performance index

Speed performance index	Traffic status level	Traffic status description
(0,25]	Serious congestion	Low average speed and poor road traffic condition.
(25,50]	Light congestion	Slightly lower average speed and slightly poor road traffic condition.
(50,75]	Smooth	Slightly higher average speed and slightly better road traffic condition.
(75,100]	Very Smooth	High average speed and good road traffic condition.

4. Explain the degree of influence factors

4.1. SHAP Explainable Method

In the SHAP explainable method, the impact of each input feature on the prediction results is assessed by analyzing its marginal contribution to the interpretation of individual instances. This method is built on the concepts of cooperative game theory and local interpretation. According to Lundberg et al. (Lundberg et al., 2018), the method of additive feature attribution includes an explanatory model g, which is a linear function of binary variables.

$$g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z'_i \tag{2}$$

where M represents the input features number, φ_i is the SHAP value associated with feature i, and z' is a boolean value in the range of 0 to 1, indicating whether the corresponding feature is observed or not.

The mapping function h is used to evaluate how missing features affect model f. z' denotes a binary pattern reflecting these missing features. $f_x(S) = f(h_x(z'))$ where the set of non-zero indices in z' is called S. Each feature i attribute value is determined as follows:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \tag{3}$$

where N represents the set of all input features and $f_x(S)$ denotes the prediction result over a subset S of the features. $\frac{|S|!(M-|S|-1)!}{M!}$ denotes the weight under the corresponding feature subset S.

4.2. Result

The SHAP method examined the impact of road features on road speed and determined a Shapley value for each feature.

Road attribute influencing factors. The results of speed prediction were affected by road variables, as indicated by the results (Fig. 3). The vertical axis represented the multiple influencing factors, while the horizontal axis displayed the degree of influence. And the color coding indicated the initial value of the factor. As the initial values of the influencing factors increased, their influence values also increased. For example, the estimated road speed was more positively impacted when the target "road level" was higher.

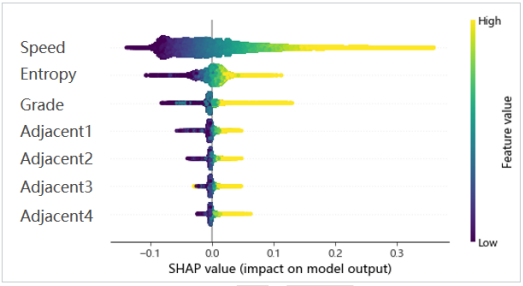


Fig. 3: SHAP abstract result map of road attributes factors.

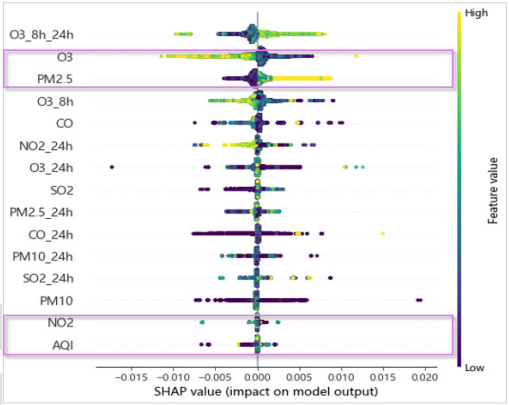


Fig. 4: SHAP abstract result map of air quality influencing factors.

Air quality influencing factors. The impact of "O_3" and "PM2.5" on speed prediction was demonstrated more clearly in Fig. 4. A negative effect became apparent when the feature value of "O_3" increased, while a positive effect was observed when the feature value decreased. The impact of "PM2.5" was completely opposite to that of "O_3". The features "NO2" and "AQI" did not exhibit any noticeable patterns in their influence on the speed prediction.

Hour influencing factors. When comparing the impact of road speed throughout the 24-hour, it became evident that the morning and evening peak hours, specifically 7:00 and 17:00, had a significant negative impact on road speed (Fig. 5). The early morning and late-night hours had a positive impact on road speed, whereas the rest of the day had minimal impact.

5. Visual design

We designed TCEVis, an interactive visual analytics system that assisted traffic managers in analyzing the causes of traffic congestion and evaluating its impact. The road speed prediction model was constructed using Python, and

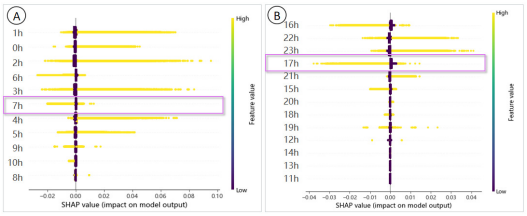


Fig. 5: SHAP abstract result map of hour influencing factors.

the system was developed with the Vue framework and D3.js. The system had six visual views that illustrated the degree to which multiple factors that could contribute to congestion affected the speed of the road.

5.1. Global view

Global view (Fig. 6 A) showed basic data information including number of roads, number of samples, and time range. It also encoded the average degree of influence for multiple influencing factors through bar charts and used pie charts to represent the degree of influence of specific features within the influencing factors. To support the exploration of the response relationship between the degree of influence of influencing factors or specific features and the range of their original value domains, interactive operations were designed where clicking on a text label relationship view updated the corresponding result.

5.2. Relationship view

Relationship view (Fig. 6 B) demonstrated the response relationship between the influencing factors or specific features in terms of their original value domains and the degree of influence in a two-dimensional space. The horizontal axis represented the original value domain and the vertical axis corresponded to the degree of influence. Each square block represented the distribution of the degree of influencing factors or specific features affecting road speed within the corresponding value range. In addition, the color of each square block indicated the number of instances that appeared, with darker colors indicating higher numbers.

5.3. Monitor view

Monitor view (Fig. 6 C) supported exploring the occurrence of traffic congestion on different moments, with two selection drop down lists offering date and time choices that could be personalized by users. With the date and time selected, the number of roads that were currently experiencing congestion was displayed, along with the level of influence of each factor on each road. If there were no congested roads, an alert box would appear.

In addition, it was available to select different roads for detailed analysis at the current moment, by clicking on the button in front of the congested roads. At the same time, the real geographic location of the corresponding roads was displayed in map view and matrix view was updated with detailed information about the selected roads.

5.4. Map view

Map view (Fig. 6 D) depicted the traffic conditions on the road, which were categorized by four colors codes to represent varying levels of congestion: "very congested", "slightly congested", "smooth", and "very smooth". Additionally, an input box in the upper right corner allowed for the entry of a specific road ID for exploration, and the location of the target road was then displayed on the map. Clicking on a road would activate a card displaying fundamental information about that road, encompassing its road ID, Speed Performance Index(SPI), speed, road grade, and speed entropy.

5.5. Matrix view

Matrix view (Fig. 6 E) adopted a matrix design, where each row represented a road, with the road ID marked on the left side. Each column represented an influencing factor or feature, with the degree of influencing factor or feature encoded in color. Initially, it displayed the top 21 roads in terms of combined influence, and by default, it explored the degree of influencing factors on the road's speed. Switching buttons allowed the exploration of the degree of influence for features under an influencing factor. The color of the rectangle encoded the magnitude of the influence, with values displayed upon hovering over the rectangle. Matrix view enabled comparing the differences in the influencing factors of multiple roads at both macro and micro levels. Furthermore, it supported temporal analysis of a single road. By clicking on the road ID, the temporal view would display the condition of the target road throughout the entire time range, along with changes in the degree of impact of multiple factors.

The initial design was presented using a scatter plot (Fig. 7), with the scatter size encoding the corresponding degree of influence. However, visually comparing the magnitude of each degree of influence proved to be challenging, and the view suffered from obfuscation, which significantly impacted the visual effect.

5.6. Temporal view

The traffic conditions of a road change over time, and a temporal view (Fig. 6 F) was designed to analyze how the degree of influence of different influencing factors on the speed of the target road changed over time. It comprised three components: (1) an overview of the traffic conditions on the target road throughout the entire time, (2) the degree of influence of different influencing factors on four adjacent time slices, and (3) the degree of influence and the raw values of features under an influencing factor at the current moment.

For a target road selected from matrix view, temporal view initially presented its traffic conditions over the entire time range. The horizontal axis represented time, while the vertical axis indicated the Speed Performance Index, encoded in bar charts for each time slice. The four traffic conditions were color-coded to match those encoded in map view. Hovering over a bar displayed the current date and

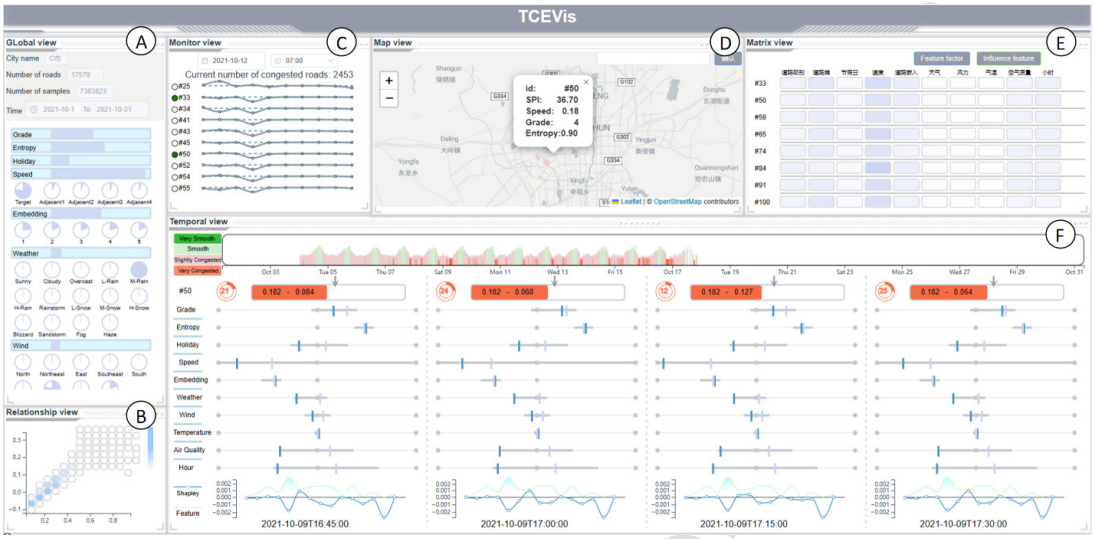


Fig. 6: Visual analytics system TCEVis for influencing factors of traffic congestion. (A) Global view provides a description of data and the importance of global influencing factors. (B) Relationship view shows the importance distribution of different values of feature. (C) Monitor view presents the number of congested roads appearing at different moments. (D) Map view represents the real geographical locations of roads and related attribute information. (E) Matrix view compares the effect of different influencing factors and specific features on road speeds. (F) Temporal view is used to analyze the importance of different influencing factors on a single road speed in time sequence.

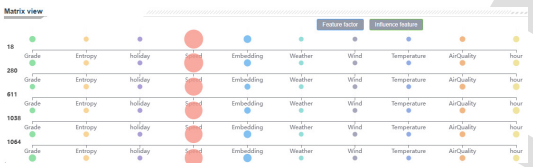


Fig. 7: Initialization design for matrix view.

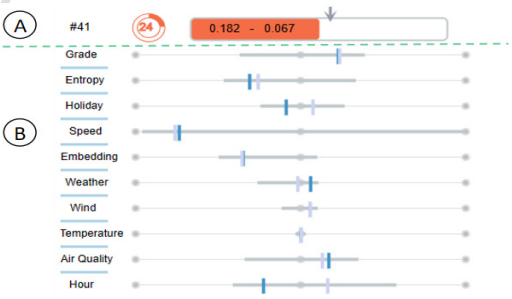


Fig. 8: The design of a time slice in temporal view.

time, and clicking on a bar allowed the selection of the time slice of interest for detailed analysis.

After selecting a time slice, the target road ID and information from four adjacent time slices were displayed, including the previous moment, the current moment, and the next two moments, enabling comparative analysis between adjacent moments. In the detailed view at each moment in time, the circular ring (Fig. 8 A) encoded the current Speed Performance Index as well as the specific value. The bar (Fig. 8 A) indicated the predicted current road speed, "0.182" representing the predicted average speed and marked with an arrow. Additionally, a "+" or "- " followed by a value indicated the overall degree of influence of all influencing factors on the current speed. At the same time, the degree of influence of different influencing factors at the current moment was presented in detail (Fig. 8 B), with each row representing one influencing factor. The horizontal axis (Fig. 9) represented the range of values for the degree of influence, with dots indicating the triple equidistant points. The dot in the center represented 0, with a negative number on the left and a positive number on the right. The gray bar (Fig. 9 A) indicated that the distribution of the degree of influence of the influencing factor on the target road speed extended over the entire time horizon. The purple vertical line (Fig. 9 B) indicated the average degree of influence of the influencing factor on the target road speed, and the blue vertical line (Fig. 9 C) indicated the degree of influence at the current moment.

To support more granular analysis for an influencing factor, clicking on the influencing factor label displayed the raw values and degree of influence of the specific features under the current influencing factor at the corresponding moment in time. The horizontal axis represented the specific features under the current influencing factor, and the vertical

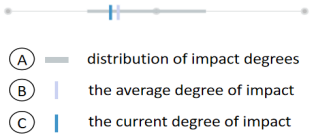


Fig. 9: Design details for single influencing factor analysis.

axis contained two values: the original value of the feature and its degree of influence, respectively. The green area plot indicated the original value of the feature, the blue curve represented the degree of influence, and the specific feature name was displayed when the mouse hovered over the dot.

6. Case Study

We conducted experiments based on real trajectory data to explain the degree to which multiple sources of influencing factors contributed to traffic congestion. We utilized TCEVis to explore the potential causes and differences in the occurrence of traffic congestion.

6.1. Explain the degree of influencing factors

From global view, we observed the average degree to which different influencing factors, such as speed, weather, and air quality had an impact on road speed. The degree of influence of these factors was calculated by the SHAP interpretable method and could help to explore the fine-grained causes under the influencing factors. In the "Speed" factor (Fig. 10 A), the speed of the "Target" road at the previous moment had a greater influence on the current speed. It meant that the speed on the target road was affected by the historical speed situation, so the temporal nature of congestion propagation was considered. Within the "Weather" factor (Fig. 10 A), moderate rain (M-Rain) had the largest impact on road speed. Compared to other weather conditions, moderate rain had the largest SHAP value, indicating a focus on monitoring road congestion during moderate rain. Similarly, the "Air Quality" factor (Fig. 10 B) had larger SHAP values for "O_3" and "PM2.5". The "Hour" factor (Fig. 10 B) also had larger SHAP values at 6:00 - 7:00 and 16:00 - 17:00. When managing congestion, it is important to pay attention to the concentrations of air pollutants "O_3" and "PM2.5", as well as to the two main time periods of the day. As mentioned above, road traffic congestion was affected by the common influence of multi-source factors. Recognizing the different degree of influence could help traffic managers to develop targeted measures to alleviate traffic congestion.

6.2. Compare road congestion at different times

As traffic congestion is a sequential process, traffic management policies should consider congestion at different times. We investigated traffic congestion from monitor view by randomly selecting six times on several days, at 6:00, 7:00, 13:00, 16:00, 17:00 and 20:00. Taking the two days

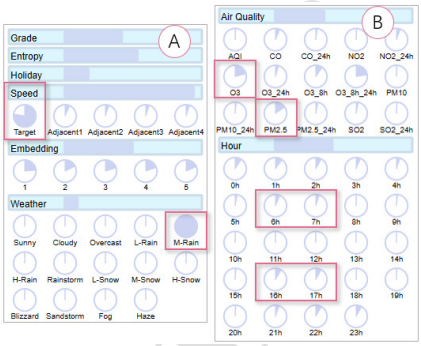


Fig. 10: The average impact of influencing factors and influencing features.

10th and 12th October 2021 for example, the analysis revealed that the road congestion was highest at 17:00. On the 10th day, the number of congested roads reached a peak of 2332 (Fig. 11), while on the 12th day, the value of the maximum congested road was 3072 during the same hour (Fig. 12). Moreover, the number of congested roads was relatively high at 7:00 on 12th day, with a value of 2453 (Fig. 12).

In addition, it is possible to analyze the degree of influence of influencing factors on congested roads at different times by the curves corresponding to each road in monitor view. We observed that the "Speed" factor was the main cause of congestion on congested roads. As indicated in Fig. 11 and Fig. 12, the influencing factor of "Speed", enclosed by the blue ellipse, had a negative maximum SHAP value. This suggested that the historical speed and the speed of neighbouring roads had the largest adverse impact on the target road speed. The parts enclosed by the orange ellipse in Fig. 11 and Fig. 12 were the "Entropy" influencing factor, which had a large negative SHAP value, indicating that the entropy of speed change played a role in reducing road speed.

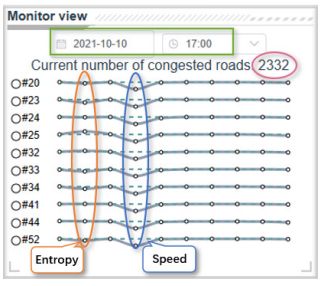


Fig. 11: Congestion analysis for 2021-10-10 in moniter view.

6.3. Analyze road congestion pattern

Urban entertainment centers are significant congestion areas and complex traffic conditions require special attention, so we selected a SHENZHEN street close to a SAIDE

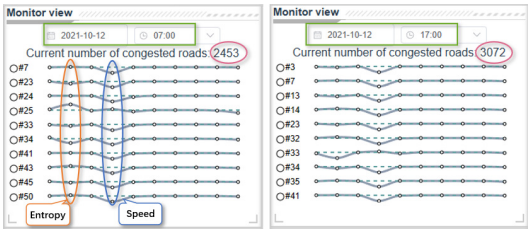


Fig. 12: Congestion analysis for 2021-10-12 in moniter view.

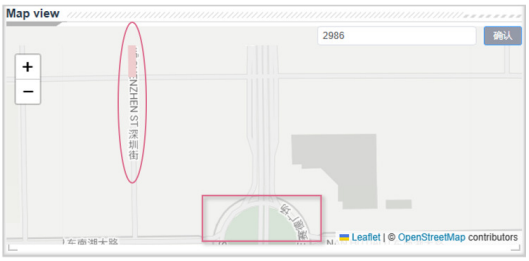


Fig. 13: The location of SHENZHEN Street in map view.

square in map view (Fig. 13) for detailed analysis. From matrix view (Fig. 14), we found that the road ID of SHENZHEN street was 2986. The SHAP values of "Speed" and "Air Quality" were larger and the color of rectangles was deeper, indicating that these two influencing factors had a larger influence on the speed of the road. Further exploration of the specific features under the two influencing factors revealed that the historical speed (Target) of road had the greatest impact on future speed in the "Speed" factor (Fig. 14). The "O_3" and "PM2.5" features had a greater impact in the "Air Quality" factor (Fig. 14).

We explored whether congestion had a pattern, using a week as a cycle. As displayed in Fig. 15, a time of congestion on SHENZHEN street was chosen for analysis using temporal view, namely "2021-10-08 19:15:00". The road was found to be "Very Congested" at this time, as indicated by the green arrow in Fig. 15, and there was also "Very Congested" at the next moment. The "Speed" factor had the greatest impact on the speed at two moments, as shown by the black dotted circle in Fig. 15. There was probably an impact of adjacent road speeds in addition to the impact of historical speed. Then, we observed the same moment next week for comparative analysis "2021-10-15 19:15:00". In Fig. 16, we discovered that only "Slightly Congested" occurred and the next moment was also "Slightly Congested". The impact of the "Speed" factor reduced, as evidenced by the blue vertical line in the black dotted circle in Fig. 16. Therefore, there was no regular pattern of congestion on this road. To verify this assumption, news reports were consulted which confirm a traffic accident transpired on the night of 8th October at the intersection of SHENZHEN Street and KUNSHAN Road.

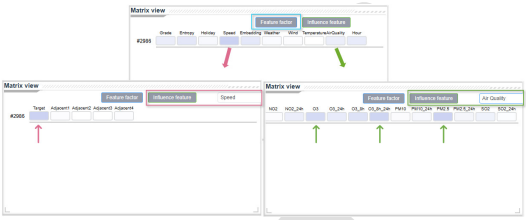


Fig. 14: Analyzing the degree to which influencing factors and influencing features affect SHENZHEN Street in matrix view.

Fig. 15: Analyzing the congestion of SHENZHEN Street at "2021-10-08 19:15:00" in temporal view.

Fig. 16: Analyzing the congestion of SHENZHEN Street at "2021-10-15 19:15:00" in temporal view.

7. Evaluation

In order to evaluate the usefulness of our TCEVis system, we invited twenty users to test our system. We allowed users to actually use our system and then conducted semi-structured interviews with the users to get an evaluation of the system.

Participants. Two traffic managers with some experience in congestion management. Eighteen participants with a background in data visualization and visual analysis (nine females and nine males).

Procedures. We provided users with system-related videos and documentation to help them learn how to use the system. Then, when users were familiar with the system, they could freely explore it according to their needs. To gain deeper insights, we conducted semi-structured interviews with each user after user exploration, which lasted approximately thirty minutes.

Feedback. We asked traffic managers for their opinions on the usefulness of the system for congestion decision making. Both agreed that visualizations were useful for congestion analysis. Traffic manager A commented, "There are many factors that affect congestion and cause us a cognitive load, and visualizing the degree of influence of different factors can help to manage the problem of traffic congestion. The analysis of congestion pattern of SHENZHEN Street

is correct, and this analysis can help me to better explore the congestion pattern of other roads." Traffic manager B said, "Previously, when we predicted road congestion using modelling, we struggled to understand what was causing the congestion, but with this system it is clear to see the degree of influence of different factors. For example, "O_3" and "PM2.5" in air quality have a significant impact on congestion and air pollution management needs to be taken into account in the subsequent improvement of traffic conditions." We also interviewed participants to obtain their views on the design and use of the system. All participants agreed that the system was easy to use. Three participants commented, "The design of temporal view is clear and concise, which made it possible to explore changes in the timing of congestion causes." However, some participants had suggestions for some views of the system. P7 suggested, "The system should add the ability to rank congested roads according to the degree of influence of specific influencing factors, so that the corresponding roads for different needs can be easily found." P15 also suggested, "When displaying basic road information on map view, glyphs can be designed."

8. Discussion and limitations

The majority of users indicated that our system was well designed and easy to understand, and that it effectively facilitated the analysis of traffic congestion causes. In addition, users have also made suggestions to our system, they would like to add the operation of filtering the influencing factors to facilitate targeted exploration.

There are also some limitations, such as the use of the widely adopted SHAP explanatory model for explaining the causes of traffic congestion without a comparison with other explanatory methods. In the future, we plan to explore a wider range of explanatory methods. We are also aware of scalability issues with the visualization and intend to adapt it for more datasets and incorporate additional interactive features.

9. Conclusion

We proposed a visual analytics system, TCEVis, to assist experts in gaining insight into the spatial and temporal evolution of congestion, identifying the specific causes of congestion occurrences, and explaining the significance of multiple influencing factors. When constructing the traffic congestion prediction model, we considered multiple influencing factors from various sources to enhance the accuracy of predictions. Meanwhile, we introduced the SHAP interpretable method to quantify and analyze the influence of different influencing factors on traffic congestion. Finally, we conducted case studies using real cab trajectory data to demonstrate that the approach presented in this paper offered a more comprehensible explanation of the potential causes of traffic congestion.

CRediT authorship contribution statement

Jialu Dong: Writing - original draft, Writing - review & editing, Data analyses, Visualization. **Huijie Zhang:** Writing - review & editing, Visualization. **Meiqi Cui:** Writing - original draft, Writing - review & editing, Data analyses,Visualization. **Yiming Lin:** Writing - review & editing. **Hsiang-Yun Wu:** Writing - review & editing. **Chongke Bi:** Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval

This study does not contain any studies with human or animal subjects performed by any of the authors.

Acknowledgments

This research was supported by National Natural Science Foundation of China under grant number 42171450, Key R&D Project of Science and Technology Development Plan of Jilin Province under Grant 20210201074GX and National Natural Science Foundation of China under grant number 62377008.

References

Abadi, A., Rajabioun, T., Ioannou, P.A., 2015. Traffic flow prediction for road transportation networks with limited traffic data. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS 16, 653–662.

Ajay, P., Nagaraj, B., Pillai, B.M., Suthakorn, J., Bradha, M., 2022. Intelligent ecofriendly transport management system based on iot in urban areas. ENVIRONMENT DEVELOPMENT AND SUSTAINABILITY .

Bachechi, C., Po, L., Rollo, F., 2021. Big data analytics and visualization in traffic monitoring. Big Data Res. 27, 100292.

Białek, J., Bujalski, W., Wojdan, K., Guzek, M., Kurek, T., 2022. Dataset level explanation of heat demand forecasting ann with shap. Energy 261, 125075.

Cheng, F., Liu, D., Du, F., Lin, Y., Zytek, A., Li, H., Qu, H., Veeramachaneni, K., 2021. Vbridge: Connecting the dots between features and data to explain healthcare models. IEEE Transactions on Visualization and Computer Graphics 28, 378–388.

Clarinval, A., Dumas, B., 2022. Intra-city traffic data visualization: A systematic literature review. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS 23, 6298–6315.

Ferreira, N., Poco, J., Vo, H.T., Freire, J., Silva, C.T., 2013. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS 19, 2149–2158. IEEE VIS Arts Program (VISAP), Atlanta, GA, OCT 13-18, 2013.

Guo, K., Hu, Y., Qian, Z., Liu, H., Zhang, K., Sun, Y., Gao, J., Yin, B., 2020. Optimized graph convolution recurrent neural network for traffic prediction. IEEE Transactions on Intelligent Transportation Systems 22, 1138–1149.

Ji, D., Dong, Q., Zhang, Y., 2022. Urban road passenger interpretation based on mlp and shap, in: 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), IEEE. pp. 512–519.

Jiang, Y., Feng, Z., Wang, H., Fan, Z., Song, X., 2022. Trafps: A visual analysis system interpreting traffic prediction in shapley. arXiv preprint arXiv:2203.06213 .

Jin, S., Lee, H., Park, C., Chu, H., Tae, Y., Choo, J., Ko, S., 2022. A visual analytics system for improving attention-based traffic forecasting models. IEEE transactions on visualization and computer graphics 29, 1102–1112.

Kahng, M., Andrews, P.Y., Kalro, A., Chau, D.H., 2017. A cti v is: Visual exploration of industry-scale deep neural network models. IEEE transactions on visualization and computer graphics 24, 88–97.

Kosugi, Y., Matsunaga, I., Ge, H., Michikata, T., Koshizuka, N., 2022. Traffic congestion prediction using toll and route search log data. 2022 IEEE International Conference on Big Data (Big Data) , 5971–5978.

Lee, C., Kim, Y., Jin, S., Kim, D., Maciejewski, R., Ebert, D., Ko, S., 2019. A visual analytics system for exploring, monitoring, and forecasting road traffic congestion. IEEE transactions on visualization and computer graphics 26, 3133–3146.

Li, C., Wu, X., Zhang, Z., Ma, Z., Zhu, Y., Chen, Y., 2022. Freeway traffic accident severity prediction based on multi-dimensional and multi-layer bayesian network, in: 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), IEEE. pp. 1032–1035.

Li, D., Lasenby, J., 2021. Spatiotemporal attention-based graph convolution network for segment-level traffic prediction. IEEE Transactions on Intelligent Transportation Systems 23, 8337–8345.

Li, Z., 2022. Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. Computers, Environment and Urban Systems 96, 101845.

Lin, K., Gao, Y., 2022. Model interpretability of financial fraud detection by group shap. Expert Systems with Applications 210, 118354.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2021. Explainable ai: A review of machine learning interpretability methods. ENTROPY 23.

Liu, H., Jin, S., Yan, Y., Tao, Y., Lin, H., 2019. Visual analytics of taxi trajectory data via topical sub-trajectories. Visual Informatics 3, 140–149.

Lu, J., Li, B., Li, H., Al-Barakani, A., 2021. Expansion of city scale, traffic modes, traffic congestion, and air pollution. CITIES 108.

Lundberg, S., Erion, G., Lee, S.I., 2018. Consistent individualized feature attribution for tree ensembles .

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

Pan, Z., Liang, Y., Wang, W., Yu, Y., Zheng, Y., Zhang, J., 2019. Urban traffic prediction from spatio-temporal data using deep meta learning. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining .

Sobral, T., Galvao, T., Borges, J., 2019. Visualization of urban mobility data from intelligent transportation systems. SENSORS 19.

Song, Y., Miller, H.J., 2012. Exploring traffic flow databases using space-time plots and data cubes. TRANSPORTATION 39, 215–234.

Wang, Y., Jing, C., Xu, S., Guo, T., 2022. Attention based spatiotemporal graph attention networks for traffic flow forecasting. Information Sciences 607, 869–883.

Zhang, J., Ma, X., Zhang, J., Sun, D., Zhou, X., Mi, C., Wen, H., 2023. Insights into geospatial heterogeneity of landslide susceptibility based on the shap-xgboost model. Journal of Environmental Management 332, 117357.

Zhang, Z., Li, Y., Song, H., Dong, H., 2021. Multiple dynamic graph based traffic speed prediction method. Neurocomputing 461, 109–117.

Zhao, W., Wang, G., Wang, Z., Liu, L., Wei, X., Wu, Y., 2022. A uncertainty visual analytics approach for bus travel time. Visual Informatics 6, 1–11.

Zhao, X., Wu, Y., Lee, D.L., Cui, W., 2018. iforest: Interpreting random forests via visual analytics. IEEE transactions on visualization and computer graphics 25, 407–416.

Ethical Approval

This study does not contain any studies with human or animal subjects performed by any of the authors.

Author Contributions Section

Jialu Dong: Writing - original draft, Writing - review \& editing, Data analyses, Visualization.

Huijie Zhang: Writing - review \& editing, Visualization.

Meiqi Cui: Writing - original draft, Writing - review \& editing, Data analyses,Visualization.

Yiming Lin: Writing - review \& editing.

Hsiang-Yun Wu: Writing - review \& editing.

Chongke Bi: Writing - review \& editing.

Declaration of interests

☑ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: