Counterfactual explanations for deep learning-based traffic forecasting

Wang, Rushan; Xin, Yanan; Zhang, Yatao; Perez-Cruz, Fernando; Raubal, Martin

DOI
10.1016/j.commtr.2025.100176

Publication date
2025

Document Version
Final published version

Published in
Communications in Transportation Research

Contents lists available at ScienceDirect

Communications in Transportation Research

journal homepage: www.journals.elsevier.com/communications-in-transportation-research

Full Length Article

Counterfactual explanations for deep learning-based traffic forecasting

Rushan Wang [a,b,*], Yanan Xin [a,c], Yatao Zhang [a], Fernando Perez-Cruz [d], Martin Raubal [a]

[a] *Institute of Cartography and Geoinformation, ETH Zurich, Zurich, 8092, Switzerland*
[b] *WSL Institute for Snow and Avalanche Research SLF, Davos Dorf, 7260, Switzerland*
[c] *Department of Transport and Planning, TU Delft, Delft, 2600, the Netherlands*
[d] *Institute for Machine Learning, ETH Zurich, Zurich, 8092, Switzerland*

ARTICLE INFO

Keywords:
Traffic forecast
Deep learning
Counterfactual explanations
Explainable artificial intelligence

ABSTRACT

Deep learning models are widely used in traffic forecasting and have achieved state-of-the-art prediction accuracy. However, their black-box nature presents challenges for interpretability and usability, particularly when predictions are significantly influenced by complex urban contextual features. This study aims to leverage an explainable artificial intelligence (AI) approach, counterfactual explanations, to enhance the explainability of deep learning-based traffic forecasting models and elucidate their relationships with various contextual features. We present a comprehensive framework that generates counterfactual explanations for traffic forecasting. The study first implements a graph convolutional network (GCN) to predict traffic speed based on historical traffic data and contextual variables. Counterfactual explanations are generated through a multi-objective optimization process, with four objectives, validity, proximity, sparsity, and plausibility, each emphasizing different aspects of optimization. We investigated the impact of contextual features on traffic speed prediction under varying spatial and temporal conditions. The scenario-driven counterfactual explanations integrate two types of user-defined constraints, directional and weighting constraints, to tailor the search for counterfactual explanations to specific use cases. These tailored explanations benefit machine learning practitioners who aim to understand the model's learning mechanisms and traffic domain experts who seek insights for necessity factors to alter traffic condition. The results showcase the effectiveness of counterfactual explanations in revealing traffic patterns learned by deep learning models and explaining the relationship between traffic prediction and contextual features, demonstrating its potential for interpreting black-box deep learning models.

1. Introduction

Accurate traffic forecasting is integral to build intelligent transportation systems, which can help alleviate traffic congestion, improve traffic operation efficiency, and reduce carbon emissions (Meena et al., 2020). Research on traffic forecasting has focused on capturing the temporal and spatial dependencies in traffic data and predicting dynamic traffic states such as traffic flow, traffic speed, and traffic demand. Over the last few years, the focus of traffic forecasting methods has shifted from using classical statistical techniques (Lee and Fambro, 1999; Wu et al., 2004; Zarei et al., 2013) to data-driven machine/deep learning methods such as recurrent neural network, long short-term memory, or graph neural network (Polson and Sokolov, 2017). The performance of traffic forecasting benefited significantly from the advancement of deep learning techniques and artificial intelligence (AI) (Yin et al., 2022). A considerable number of studies have demonstrated the exceptional performance of deep learning algorithms in reducing predictive errors in traffic forecasting. However, challenges arise with the black-box nature of these deep learning models. The lack of interpretability and explainability makes it difficult for machine learning developers to understand the learning mechanisms of these models (Xin et al., 2023). Furthermore, it is also challenging for domain experts to utilize these models and derive insightful understandings of traffic dynamics due to the opacity of the models (Jonietz et al., 2022). These challenges hinder the adoption of deep learning models in practice (Fernandez et al., 2020).

Recently, the issues of interpretability and explainability in AI gained increasing attention from researchers (Marcinkevics and Vogt, 2023). To address this challenge, explainable artificial intelligence (XAI) techniques are proposed to enhance machine learning (ML) models' interpretability and explainability, making the output of these models more comprehensible to humans (Edwards and Veale, 2017). XAI methods are generally categorized into two types: global explanations and local

[*] Corresponding author. Institute of Cartography and Geoinformation, ETH Zurich, Zurich, 8092, Switzerland.
E-mail address: ruswang@ethz.ch (R. Wang).

https://doi.org/10.1016/j.commtr.2025.100176
Received 26 August 2024; Received in revised form 29 October 2024; Accepted 17 November 2024

explanations. Global explanations provide insights into the overall model behavior, identifying which input variables have a significant influence on the model's predictions. However, applying global explanations to neural networks is challenging due to the vast number of parameters and the complexity of their nonlinear relationships. This limitation prompts the use of local explanations instead. Local explanations, on the other hand, focus on interpreting the model's behavior within a specific region of interest. These methods typically involve using simpler surrogate models to approximate the complex model's decisions at a local level, providing more interpretable information such as feature importance scores (Lundberg and Lee, 2017). However, these techniques suffer from an inherent fidelity-interpretability trade-off due to the use of a simpler model for generating explanations.

In traffic domain, road-traffic management, including tasks like traffic flow forecasting and congestion control, involves high-stake decisions that impact public safety and efficiency. Traditional deep learning models, while accurate, often function as “black boxes”, making it difficult for traffic managers and policymakers to understand and trust their decisions.

Counterfactual explanations (CFEs) as a local explanation method can maintain consistency with the original machine learning model, offering insights into the inner workings of machine learning models (Wachter et al., 2018). CFEs reveal the minimal changes required in the original input features to alter the model's prediction, thus providing understanding without sacrificing fidelity or complexity.

In our study, CFEs are particularly advantageous since we are interested in determining the minimal change in the input to obtain a desired alternative prediction. CFEs are straightforward to understand and can be used to provide users with a course of action to alter the prediction if they receive unfavourable decisions. These explanations establish a relationship between the input features and the decision, making them highly valuable for users to comprehend, interact with, and utilize these models.

Currently, there is a significant lack of study in applying XAI techniques in the domain of traffic forecasting (Li et al., 2023; Xin et al., 2022; Yang et al., 2023). It is not straightforward to apply counterfactual methods developed in non-spatial domains to spatiotemporal data analysis due to the high complexity and dimensionality of spatiotemporal data (Xin et al., 2022). Moreover, due to the complex spatial and temporal dependency in traffic domain, single feature change can induce effect on different spatial and temporal dimension, which also requires more comprehensive study. Thus, one of the core objectives of this study is to explore the potential and limitations of counterfactual explanations in deep learning-based traffic forecasting applications.

The study is guided by the following research questions.

- What is the impact of input variables on deep learning-based traffic forecasting?
- How can we modify the input variables to achieve the desired prediction for various scenarios?

This study involves training and explaining a deep-learning model for traffic forecasting. Particularly, by applying the XAI technique, the study contributes to our understanding of how the model produces predictions, and how variations in input contextual features can affect predicted results. The second key contribution of our study is the application of CFEs on spatiotemporal prediction tasks, where the spatiotemporal dependencies are critical. In this context, we conduct a thorough evaluation of the impact of the counterfactual features on the spatiotemporal traffic dynamics. Another contribution of this study is the proposal of scenario-driven counterfactual explanations, where we demonstrate and validate different methods to integrate user prior knowledge or constraints in generating counterfactuals.

In summary, this study proposes a framework to tackle the lack of explainability of black-box traffic forecasting models. By streamlining the procedures of generating and examining counterfactual explanations in deep learning-based traffic forecasting, this study offers valuable insights for future studies in this direction.

2. Related work

2.1. *Deep learning in traffic forecasting*

It is an important research topic to analyze the non-linear and complex spatiotemporal patterns of traffic dynamics in order to make accurate traffic predictions (Yin et al., 2022). Statistical and traditional machine learning models are two major representative data-driven methods for traffic prediction. This includes methods such as historical average (HA), auto-regressive integrated moving average (ARIMA) (Williams and Hoel, 2003), support vector regression (Chen et al., 2015), and random forest regression (Johansson et al., 2014). However, one of the disadvantages of traditional approaches is that most of the applied features need to be carefully selected and processed by a domain expert to reduce the complexity of the feature space and make the underlying patterns easier to extract.

Over the last few years, deep learning-based methods have unlocked the potential of artificial intelligence in traffic prediction (Lv et al., 2015). Deep learning models exploit much more features and complex architectures than classical methods and can achieve better performance. Recurrent neural networks (RNNs) stand out as particularly effective in time series forecasting (Prasad and Prasad, 2014; Ramakrishnan and Soni, 2018). Additionally, a series of studies have applied CNN to capture spatial correlations in traffic networks from two-dimensional spatiotemporal traffic data (Li and Shahabi, 2018). However, the CNN-based approach is not optimal for traffic foresting problems that have a graph-based data type.

Over the past few years, graph neural networks (GNNs) have emerged as a cutting-edge deep learning technique, demonstrating state-of-the-art performance in numerous applications (Wu et al., 2019a,b). Due to their capability of modeling non-euclidean graph structures, GNNs are particularly well-suited for traffic forecasting tasks where complex spatial dependencies need to be captured (Jiang and Luo, 2021). These include, for instance, the diffusion convolutional recurrent neural network (DCRNN) (Li et al., 2017), temporal graph convolutional network (T-GCN) (Zhao et al., 2018), and graph WaveNet (Wu et al., 2019a,b) models.

In traffic prediction studies, contextual data has been widely recognized as an important input to improve traffic prediction performance (Zhang et al., 2023, 2024). Some commonly used external variables include weather conditions, events, and time information (Yin et al., 2022). One previous study (Liao et al., 2018) incorporated auxiliary data, such as crowd map queries and road intersections, along with geographical and social variables, into an encoder-decoder sequence learning framework for traffic forecasting. In another study (Zhu et al., 2020), researchers categorized these influencing factors as either dynamic or static attributes and designed an attribute-augmented unit that seamlessly integrates these variables into a spatiotemporal graph convolution model, which enhanced the model's forecasting capabilities. Classifying contextual data into spatial and temporal contextual features, Zhang et al. (2023) proposed a multimodal context-based graph convolutional neural network (MCGCN) to embed spatial and temporal contexts and incorporate them into traffic speed prediction for better performance.

2.2. *Explainable AI in traffic domain*

Previous studies have demonstrated the value of XAI in traffic applications (Gaur and Sahoo, 2022). For example, Xu et al. (2014) developed an interpretable model to predict short-term traffic flow, helping identify key road segments contributing to congestion, but their model lacks the flexibility to analyze hypothetical scenarios. Similarly, Kruber et al. (2018) used a modified random forest to categorize traffic situations, creating static visual representations that do not suggest specific actions to improve conditions. SHAP has been applied to deep learning models like LSTMs to explain feature contributions in traffic prediction (Barredo-Arrieta et al., 2019), but it is limited to post-hoc interpretations and cannot capture complex feature interactions or provide guidance on altering inputs. Ma et al. (2017) used gradient boosting trees to prioritize factors influencing incident clearance times, but this only highlights feature importance without exploring the effects of feature changes.

In contrast, counterfactual explanations (CFE) overcome these limitations by not only identifying key features but also suggest what should be different in the input instance to change the outcome of an AI system (Diakopoulos, 2020; Wachter et al., 2018). However, there is lack of study on counterfactual explanations in traffic domain.

In recent years, CFEs have been applied in various tasks to enhance the interpretability of machine/deep learning models (Fernandez et al., 2020). It has already been widely used in image classification, where generative models such as GANs and variational autoencoders (VAE) are used to implement interventions and generate realistic CFEs (Covert et al., 2021; Liu et al., 2019; Parafita and Vitrià, 2019; Singla, 2022). Other than image data, CFEs have also been utilized for text data (Jung et al., 2022), speech data (Zhang and Lim, 2022), time-series data (Ates et al., 2021), graph data (Prado-Romero et al., 2022), etc.

Numerous methods are developed for generating CFEs, each with its specific focus and application. For instance, the FACE method (Poyiadzi et al., 2020) aims to produce plausible CFEs by building feasible paths between data points associated with opposing predictions. On the other hand, DiCE (Mothilal et al., 2019) is designed primarily for differentiable models and is especially useful for handling continuous features. Another innovative approach is the Bayesian-optimization-based counterfactual explanations (Spooner et al., 2021), which employ probabilistic methods to generate counterfactuals. Additionally, multi-objective counterfactual (MOC) (Dandl et al., 2020) was proposed recently that conceptualizes the counterfactual search as a multi-objective optimization problem, which broadens the scope and applicability of CFEs in complex scenarios. In this study, we used MOC due to its ability to produce a varied set of counterfactuals, offering multiple options for actionable feature adjustments based on different objective trade-offs.

3. Methods

3.1. Traffic forecasting model

Graph convolutional networks (GCNs) have demonstrated significant efficacy in traffic forecasting tasks as discussed in Section 2.1. In this study, the traffic forecasting model is built to predict the future traffic speed for each road segment of the traffic graph. Specifically, the definitions of traffic graph and graph-based traffic forecasting are as follows.

- Traffic graph: A graph $G = (V, E, A)$ can be utilized to describe the topological structure of the road network, and each road segment is treated as a node, where V is a set of road nodes, $V = \{v_1, v_2, \ldots, v_N\}$, where N is the number of the nodes, and E is a set of edges. The adjacency matrix A is used to represent connections between road segments, $A \in R_{N\times N}$.
- Graph-based traffic forecasting: The spatiotemporal traffic forecasting task can be defined as to find a function f which generates $y = f(\chi, \varepsilon; G)$, where y is the traffic state to be predicted, $\chi = \{\chi_1, \chi_2, \ldots, \chi_T\}$ is the historical traffic state defined on graph G, T is the number of time steps in the historical window size, and ε represents the external factors.

Inspired by the temporal graph convolutional network model (Zhao et al., 2018) and AST-GCN model (Zhu et al., 2020), this study adopted a similar model. Fig. 1 shows the architecture of the deep learning model we used.

For each input unit at time step t, traffic speed data χ_t and contextual data ε_t are concatenated as enhanced feature matrix X_t. Together with adjacency matrix A, they are fed into the graph convolutional network (GCN), which can capture the spatial dependence of the data. The modeling process of GCN can be expressed as Zhu et al. (2020):

$$gc_{l+1} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}gc_l W_l\right) \tag{1}$$

where σ is the activation function; $\widetilde{A} = A + I$ represents a matrix with self-connection structure; $\widetilde{D}$ is a degree matrix; W_l denotes the weight matrix of the l-th convolutional layer; c_l is the output representation; and $gc_0 = X$, X is the feature matrix.

To capture the temporal features, the architecture combines GCN and GRU models. Specifically, the feature matrics are fed into a series of GCNs to generate time-varying features. Then the feature series are used as input of GRUs to model the temporal dependence and derive hidden traffic states.

As shown in Fig. 2, h_{t-1}denotes the output at time $t-1$, gc is graph convolution process, u_t and r_t are update gate and reset gate at time t, and h_t denotes the output at time t. The specific calculation process is shown below, where W and b are the weights and deviations in the training process:

$$u_t = \sigma(W_u \cdot [gc(X_t, A), h_{t-1}] + b_u) \tag{2}$$

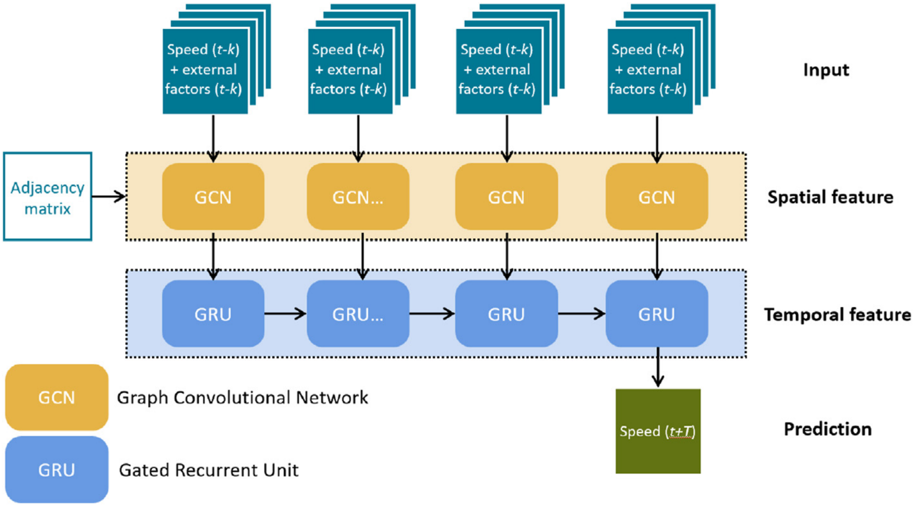

Speed (t-k) + external factors (t-k)
Speed (t-k) + external factors (t-k)
Speed (t-k) + external factors (t-k)
Speed (t-k) + external factors (t-k)
Input
Adjacency matrix
GCN
GCN…
GCN
GCN
Spatial feature
GRU
GRU…
GRU
GRU
Temporal feature
GCN
Graph Convolutional Network
GRU
Gated Recurrent Unit
Speed (t+T)
Prediction

Fig. 1. Architecture of the deep learning model used in this study for traffic forecasting.

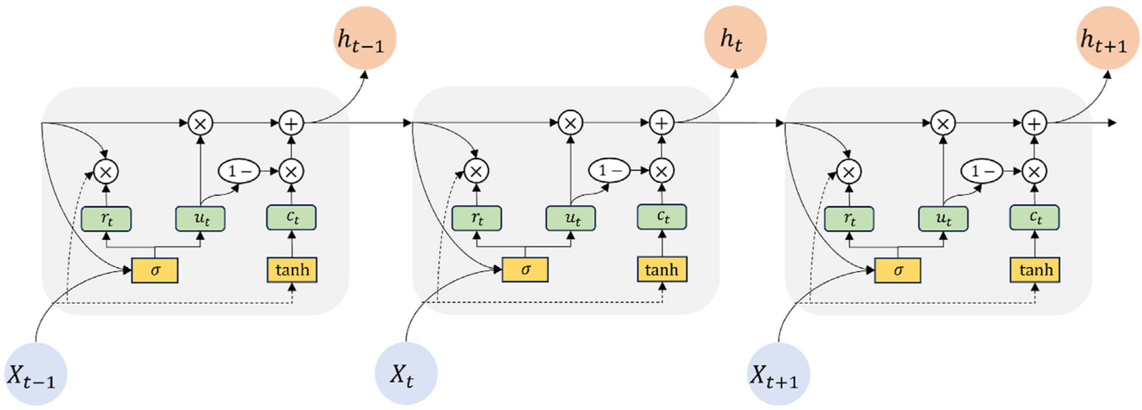


Fig. 2. Architecture of the gated recurrent unit (GRU) model.

$$r_t = \sigma(W_r \cdot [gc(X_t, A), h_{t-1}] + b_r) \tag{3}$$

$$c_t = \tanh(W_c \cdot [gc(X_t, A), (r_t, h_{t-1})] + b_c) \tag{4}$$

$$h_t = u_t \times h_{t-1} + (1 - u_t) \times c_t \tag{5}$$

During the training process, the loss function is set to minimize the variation between the real traffic speed and the predicted speed.

$$\text{Loss} = ||y_t - \hat{y}_t|| + \lambda L_{\text{reg}} \tag{6}$$

where y_t and $\hat{y}_t$ are the ground truth and prediction, respectively, L_{reg} is the L1 regularisation term to avoid overfitting, and λ is a hyperparameter.

The following metrics were used to evaluate the prediction accuracy of the model.

- Root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{7}$$

- Mean absolute error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{8}$$

- Accuracy:

$$\text{Accuracy} = 1 - \frac{||y - \hat{y}||_{\text{F}}}{||y||_{\text{F}}} \tag{9}$$

where $||\cdot||_{\text{F}}$ is the Frobenius norm.

- Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{t=1}(y_t - \hat{y}_t)}{\sum_{t=1}(y_t - \bar{y}_t)} \tag{10}$$

- Explained variation (VAR):

$$\text{Var} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)} \tag{11}$$

This measures the proportion to which the proposed model accounts for the variation in real traffic states, which is mainly used to measure the predictive ability of the model.

3.2. Multi-objective optimization to select CFEs

Counterfactual explanations (CFEs) provide insights into what minimal changes in input features can lead to a desired alternative prediction. Given a classifier b that outputs the decision $y = b(x)$ for an instance x, a counterfactual explanation seeks to find an instance x' such that the decision for b on x' is different from y, i.e., $b(x') \neq y$, while the difference between x and x' is minimal.

When generating CFEs, there can be multiple possibilities to conduct changes in input features to achieve the desired alternative prediction. Therefore, different criteria or objectives are proposed to help select optimal CFEs. Existing approaches to generate counterfactual explanations often rely on optimizing a single weighted sum of multiple objectives, making it difficult to balance different objectives. Following the approach proposed by (Dandl et al., 2020), this study considers the task of generating counterfactual explanations as a multi-objective optimization problem, which allows for the generation of a diverse set of CFEs.

Multi-objective optimization is a mathematical technique used for solving problems involving competing objectives. In the context of counterfactual explanations, the goal is to optimize for multiple criteria simultaneously, rather than aggregating them into a single metric.

To guide the search for counterfactuals, we employed four key criteria, which are

- **Validity**: A counterfactual is valid if it produces a predicted outcome closely approximating the target speed, which is an artificially defined or desired speed that serves as a reference for model predictions.
- **Proximity**: The ideal counterfactual should differ minimally from the original feature set, thereby ensuring that the changes suggested are modest and realistic.
- **Sparsity**: A counterfactual gains in feasibility when the number of altered features is minimized.
- **Plausibility**: For a counterfactual explanation to be considered plausible, it should be close to the nearest observed data points.

It is important to recognize that a counterfactual example, while perhaps optimal in feature space, may not be practically feasible due to real-world constraints. Therefore, users should also have the flexibility to specify constraints on specific features, including:

- **Range constraints**: These define feasible ranges for each feature. For instance, a constraint might specify that “Speed limit on the road should be larger than 30 km/h.”
- **Mutable variables**: Alternatively, users may specify which variable can be altered in the search for a counterfactual explanation.

The presence of multiple objectives in a problem gives rise to a set of optimal solutions, known as Pareto-optimal solutions. Without additional information, it is hard to say which Pareto-optimal solution is better than the others. To efficiently address this problem, we used the non-dominated sorting genetic algorithm II (NSGA-II), a fast multi-objective evolutionary algorithm used in Deb et al. (2002).

In this study, the performance of a counterfactual is represented by a vector of quantitative measures, corresponding to the criteria outlined above. Lower values of the metrics signify better counterfactuals.

For the generation of counterfactuals, the search process plays a critical role. In this study, Gaussian mutation is utilized, with predefined standard deviations assigned to each feature. This process ensures that only a small change will be added to the features each time. The process of generating counterfactual explanations can be summarized into the following steps.

1) **Identify target outcome**

Given that the entire road graph contains 3169 road segments, we narrowed its focus to optimizing speed on a single, selected road segment in each experiment. This targeted approach allows for a more manageable and detailed examination of the generated counterfactuals.

2) **Determine search space**

The search space under consideration is constrained by two key dimensions. The first involves identifying which nodes within the network have features amenable to modification for generating counterfactual explanations. The second aspect focuses on delineating the permissible range within which these counterfactual features can be altered. By establishing these constraints, we create a well-defined scope for generating meaningful and feasible counterfactual explanations.

3) **Define objective function**

In line with previously outlined criteria, the objective function is constructed as follows:

Let $f: X \rightarrow \mathbb{R}$ denote the prediction function, X^{obs} represents the observed feature space, and y_{target} is the predetermined target speed. A counterfactual explanation x' for a given observation x aims to meet four key criteria: validity, proximity, sparsity, and plausibility. The overarching goal is to minimize a four-component loss function as defined by Molnar (2022):

$$L(x, x', y_{\text{target}}, X^{\text{obs}}) = \left(o_1(f(x'), y_{\text{target}}),\ o_2(x, x'),\ o_3(x, x'),\ o_4(x', X^{\text{obs}})\right) \quad (12)$$

where each component captures one of the aforementioned criteria.

- **Validity**: The objective function o_1 evaluates the distance between the predicted speed $f(x')$ and the target speed y_{target}:

$$o_1(f(x'), y_{\text{target}}) = \left|f(x') - y_{\text{target}}\right| \quad (13)$$

- **Proximity**: The objective function o_2 measures the L1-norm between the original and counterfactual features, x and x':

$$o_2(x, x') = \|x - x'\|_1 \quad (14)$$

- **Sparsity**: The objective function o_3 captures the sparsity of the changes needed to convert x into x' by computing the L0-norm:

$$o_3(x, x') = \|x - x'\|_0 \quad (15)$$

- **Plausibility**: The final objective o_4 evaluates the plausibility of the counterfactual explanation x' within the observed feature space X^{obs}. This is calculated by averaging the Euclidean distances between x' and its k nearest neighbors in X^{obs} in an n-dimensional feature space:

$$o_4(x', X^{\text{obs}}) = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\sum_{j=1}^{n}\left(x'_j - x^{\text{obs}}_{\text{nearest},i,j}\right)^2} \quad (16)$$

where $k = 3$ in this study.

4) **Searching the counterfactual explanations**

The NSGA II is employed to generate a set of counterfactual explanations that satisfy all four objectives. The selection of the most suitable CFE from this set is also a crucial aspect of our approach. To facilitate this, an evaluation score y_{e} is defined, as shown in Eq. (17). This evaluation score serves as a multi-objective trade-off criterion. Users can adjust the weights λ_1, λ_2, λ_3, λ_4 to prioritize specific objectives. For instance, if users value the effectiveness of a CFE in altering the predicted speed over the cost incurred in modifying the features, they might assign a higher weight to the validity objective (o_1).

$$y_e = \lambda_1\frac{o_1}{\max(o_1)} + \lambda_2\frac{o_2}{\max(o_2)} + \lambda_3\frac{o_3}{\max(o_3)} + \lambda_4\frac{o_4}{\max(o_4)} \quad (17)$$

5) **Evaluating the counterfactual explanations**

After the generation and selection of counterfactual explanations, a comprehensive evaluation is essential to understand the generated counterfactuals and assess their performance. It is crucial to verify that the counterfactual explanations actually achieve the desired speed improvement for the targeted road segment. Beyond the targeted road segment, it is also necessary to ensure that localized changes do not negatively impact the speed prediction in other road segments of the network.

3.3. *Scenario-driven counterfactual explanations*

To incorporate user prior constraints effectively, this study proposes an adjustment to the cost function. Specifically, we modified the proximity objective, as represented in Eqs. (18) and (19), to enable the exploration of different scenario settings.

$$o_2'(x, x') = \sum_{i\neq E}\left|x_i - x_i'\right| + \lambda\sum_{i=E}\left(d_i(x_i - x_i')\right) \quad (18)$$

$$o_2'(x, x') = \sum_{i\neq E}\left|x_i - x_i'\right| + \lambda\sum_{i=E}\left|x_i - x_i'\right| \quad (19)$$

where E represents the feature space that the user wishes to remain unchanged. By incorporating a large weight λ, we introduce a significant penalty, steering the generated counterfactual explanations towards user-defined preferences. For Eq. (18), d_i is a direction indicator, $d_i = 1$ if an increase in x_i is preferred, $d_i = -1$ if a decrease in x_i is preferred. This study proposes two distinct mechanisms for integrating user-specific preferences into the counterfactual explanations.

- **Directional constraints**: Users have the option to specify the direction—either increase or decrease—in which they would like specific features to change. For instance, if the user wants to increase the number of nearby POIs, by setting a large penalty for any generated CFEs where the number of POIs is decreased, the algorithm can tend to generate CFE with a larger number of nearby POIs.

- **Weighting constraints**: Users can assign weights to individual features to prioritize their importance during the counterfactual generation process. For instance, if a user prefers not to alter the number of lanes on road segments, applying a larger penalty for CFEs where the number of lanes is modified will encourage the algorithm to generate CFEs that maintain the current number of lanes, focusing changes on other features instead.

4. Experiments and results

The traffic speed data was provided by HERE technologies,[1] which offers a record of traffic speed observations on different road segments.

In this study, the road graph is located in Thousand Oaks, California, USA, as shown in Fig. 3, which consists of 3169 road segments. The data were collected from January 1st to 30th, 2019, at 5-min intervals. Fig. 4 shows the average speed of all the road segments within the study period. A noticeable temporal pattern emerges, where lower speeds appear during the daytime and a distinct weekly pattern exists with different speed variations between weekdays and weekends.

Contextual data is of great importance to traffic prediction. In this study, several contextual features were collected, which can be classified into static features and dynamic features. Static features are location-based, which vary with regard to different road segments. Based on findings from previous studies (Section 2.1), this study included nearby POI data, speed limit data, and lane configuration of each road segment as static features. Particularly, the POIs include the nearby gas station, charging station, parking lot, and restaurant. Dynamic features are time-based features that change over time. In our study, dynamic features such as the day of the week, hour of the day, and weather condition data (e.g., temperature, wind speed, precipitation, humidity) are included. Table 1 summarizes all the contextual features involved in the study.

The overall performance metrics for the traffic forecasting model are detailed in Table 2. The accuracy reached 91.24%, indicating a decent prediction performance.

4.1. Generating counterfactual explanations

Fig. 5 displays the locations of Node A on a suburban road, Road I, which are the focusing road segments in this experiment. Fig. 6 illustrates the speed of each road segment on Road I from 6:00 to 8:00, January 10th, 2019.

Specifically, the target of this experiment is to increase the average predicted speed for the road segment Node A from 28 to 56 km/h. The prediction uses input data from 8:00 to 8:55 on January 10th, 2019 to predict the traffic speed from 9:00 to 9:55. Modifications are restricted to road segments situated within Road I. In this experiment, only the static features of each road segment are considered for modification. Based on feature values present in the dataset, the specific ranges for the changeable features are set as follows.

- Number of POIs: Range from 0 to 36.
- Number of Lanes: Range from 1 to 6.
- Speed Limit: Range from 40 to 120 km/h.

It is important to note that the speed limit is constrained to remain the same across all segments within Road I to be more realistic.

4.1.1. Objective distributions and correlations

Fig. 7 shows the distribution of the objectives for the set of counterfactual explanations generated in this experiment. The distribution patterns reveal insights into the relationships among different objectives.

Validity–Proximity: As illustrated in Fig. 7a, there appears to be a negative correlation between the validity loss and the proximity loss, which suggests that as counterfactual predictions become closer to the target speed, the divergence of the generated counterfactual features from the original features increases. This relationship reflects a key characteristic of the model: generally, larger changes in input features tend to produce more significant differences in output results, thereby increasing the likelihood of reaching the target outcome, and vice versa. However, Fig. 7a also highlights a small subset of generated counterfactual features that achieve both low validity and low proximity losses, which is precisely the desired outcome. These instances demonstrate the potential of our method to generate effective counterfactuals that are both accurate and minimally divergent from the original input, thereby preserving interpretability and usability.

Validity–Plausibility: Similar observations can be made from Fig. 7b, where validity loss and plausibility loss are negatively correlated. This implies that when the counterfactual predictions become closer to the target speed, they tend to deviate more from observed points in the feature space.

Proximity–Plausibility: Fig. 7c depicts an overall positive correlation between proximity loss and plausibility loss. Generally, a greater proximity loss is accompanied by a larger plausibility loss. However, an interesting cluster of points exists in the bottom-right corner of this figure. These points show that there are counterfactual explanations that differ substantially from the original features but still maintain an overall close distance to observed data points.

4.1.2. Evaluation of the most optimal counterfactual explanations

Different weight parameters can be assigned to each objective function in Eq. (17) to find the optimal counterfactual explanation for a particular interest or purpose. As a case study, we investigated the results where $\lambda_1 = 1$, $\lambda_2 = 0.2$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.6$. This choice of weights reflects the relative importance of different criteria in the evaluation score. Particularly, validity is prioritized as the most critical factor and is assigned the highest weight. Plausibility also holds significance, but to a lesser extent, so it was assigned a weight of 0.6. Given that sparsity was considered less crucial for this particular study, it was given a lower weight of 0.2. Additionally, since proximity and plausibility are interrelated, we assigned proximity a smaller weight of 0.2 to ensure a balanced evaluation.

The optimal counterfactual explanation with the given weights produces the objective scores outlined in Table 3. The validity score shows a minimal deviation of 1.3496 km/h from the target speed. Regarding sparsity, a total of 32 features were altered across all road segments on Road I.

Table 4 shows the speed prediction change with this optimal counterfactual explanation. With the counterfactual features, the speed prediction increases significantly and is very close to the target speed of 56 km/h.

Fig. 8 shows the comparison between original features and the selected counterfactual features (the number of POIs and the number of lanes) for each road segment on Road I. The counterfactual features in Fig. 8 suggest that a general increase in POIs at certain locations of the road network is associated with higher speed prediction. Given other counterfactual features, the number of lanes only needs minor modification at a few locations to achieve the target speed, as in Fig. 8b. The original speed limit is 72 km/h, while the counterfactual speed limit is 105.62 km/h.

4.2. Spatial comparison

The type of road facility (e.g., highway, urban road, or suburban road) is widely acknowledged as an important factor influencing traffic patterns (Yazici et al., 2014). In light of this, to gain deeper insights into how the deep learning model predicts speed differently across different types of roads, this section compares the counterfactual explanations generated for three distinct types of road segments, i.e., a suburban road, an urban road, and a highway, represented by Node A, Node B, and Node C

[1] HERE Technologies: https://developer.here.com/products/platform/data.

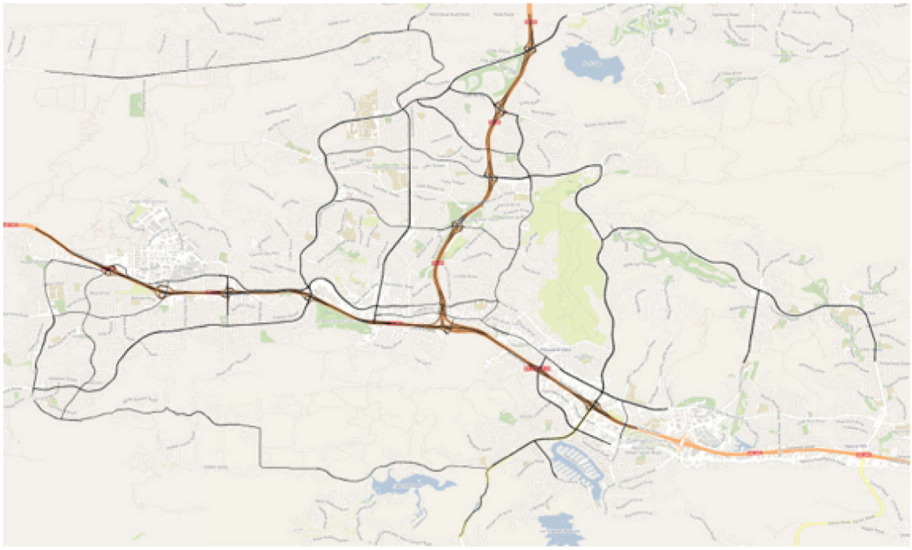


Fig. 3. Location of road network (dark line).

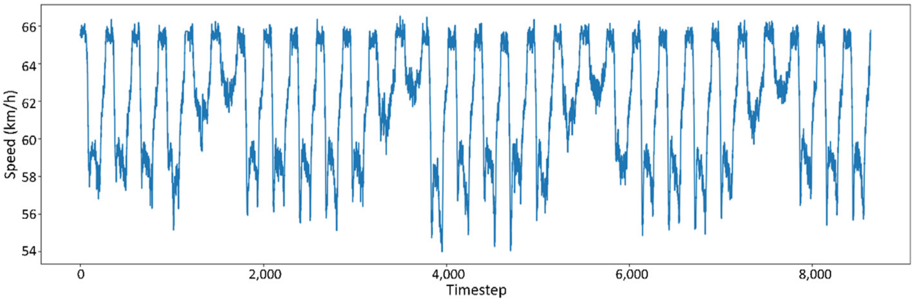


Fig. 4. Average speed for all the 3169 road segments from January 1st to 30th, 2019.

Table 1
Summary of the contextual data in this study and their encoding method.

Class	Contextual data	Encoding method
Static feature	Number of POIs	Integer
	Speed limit	Integer
	Number of lanes	Integer
Dynamic feature	Day of the week	One-hot encoding
	Hour of the day	Sin-cos encoding
	Temperature	Float
	Wind speed	Float
	Precipitation	Float
	Humidity	Float

Table 2
Traffic forecasting model performance.

Metrics	RMSE	MAE	Accuracy	R^2	VAR
Performance	5.7473	2.9876	91.24%	0.9282	0.9291

respectively. Fig. 9 displays the locations of the two additional nodes, Node B and Node C. Fig. 10 shows the speeds of the three nodes on January 10th, 2019.

For all three road segments, the target was set identically: to increase the predicted average speed on each node between 9:00 and 10:00 to reach 56 km/h. The initial average speeds recorded were 28.2 km/h for Node A, 49 km/h for Node B, and 20.18 km/h for Node C. To achieve the target speed, counterfactual explanations were generated and selected for each node following the procedures outlined in Section 3.2. To compare the impact of the generated counterfactual features on the daily pattern, we generate counterfactual predictions for each node for the entire day and display the results respectively in Fig. 11.

Figs. 11a and 11b reveal that the generated counterfactual explanations for node A (suburban road) and node B (urban road) managed to increase the predicted speed, particularly for the targeted duration (9:00–10:00). However, Fig. 11c shows that the counterfactual explanation for node C (highway) did not result in a substantial speed increase. This demonstrates that static features, including the number of POIs, the number of lanes, and speed limits, do not exert a significant influence on predicting highway speeds. Therefore, in the following experiments, we will only focus on Node A and Node B for subsequent analyses.

Since we only generated counterfactual features at local road segments to increase the predicted speed on a particular node, it is uncertain whether the generated counterfactuals will negatively impact predicted traffic in other parts of the road network. In this section, we evaluate the global impact of counterfactual explanations on the speed prediction for the entire traffic network.

Fig. 12 shows the difference between the counterfactual speed prediction and the original speed prediction. In Fig. 12a the speed increase is mainly distributed on the urban road I. The counterfactual features only have a minimal negative impact on the speed prediction of other locations, with a maximum decrease of 6.9 km/h in predicted speed. In contrast, Fig. 12b shows that the counterfactuals generated for Node B on the urban road also broadly change the predicted traffic speed in other road segments. In addition, the negative impact caused by counterfactuals at Node B (urban road) is larger than those at Node A (suburban road). The largest speed decrease reaches 21.3 km/h with the counterfactual features generated for urban roads.

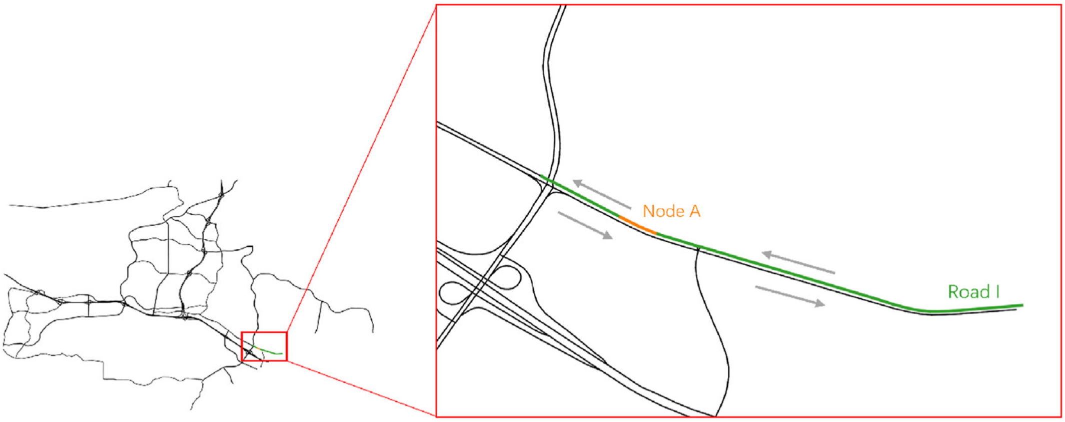


Fig. 5. Location of Node A and Road I (a suburban road).

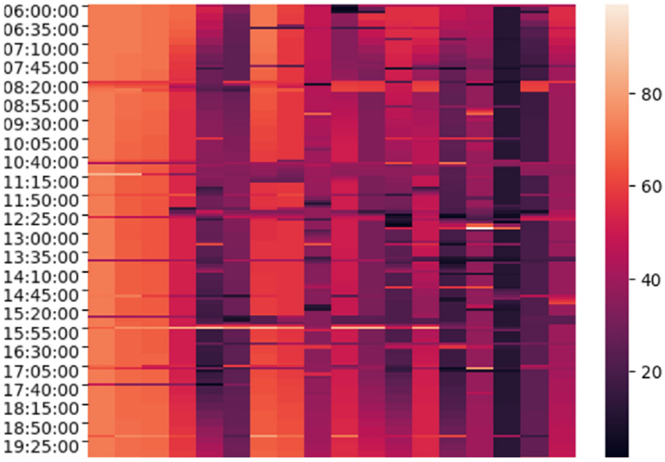


Fig. 6. Speed variation for each road segment on Road I on January 10, 2019. The color bar indicates the speed (km/h), each column shows the speed for one road segment, and the traffic flow is from the right side to the left side.

4.3. *Temporal comparison*

Temporal setting also significantly influences traffic patterns. To examine these effects, we compared counterfactuals generated for five time slots.

- **Morning** Jan 10th (Thursday): 8:00–10:00
- **Noon** Jan 10th (Thursday): 12:00–14:00
- **Afternoon** Jan 10th (Thursday): 15:00–17:00
- **Evening** Jan 10th (Thursday): 18:00–20:00
- **Weekend** Jan 13th (Sunday): 8:00–10:00

We generated counterfactual explanations for Node A on the suburban road and Node B on the urban road during each of these time slots. The most optimal counterfactual explanations for each temporal setting across all nodes in the road segment are selected. Summing over the difference across all nodes, Fig. 13 compares the total difference between the counterfactual features and original features for each setting.

4.3.1. *Comparison of number of POIs*

Fig. 13a illustrates the variations in the counterfactual number of POIs for both Node A and Node B across the selected time slots.

Node A: The counterfactual features for Node A show a consistent increase in the number of POIs across all time slots. This trend suggests that the model associates a higher number of POIs with lower congestion levels on suburban roads. This increase is more pronounced during weekends, indicating that during weekends, the number of POIs has a

Table 3
Objective value for the selected counterfactual explanation.

Validity, o_1	Proximity, o_2	Sparsity, o_3	Plausibility, o_4
1.3496	172.5508	32	0.9862

Table 4
Average speed prediction from 9:00 to 10:00, January 10th, 2019.

Original prediction	Counterfactual prediction	Target
30.10 km/h	54.65 km/h	56 km/h

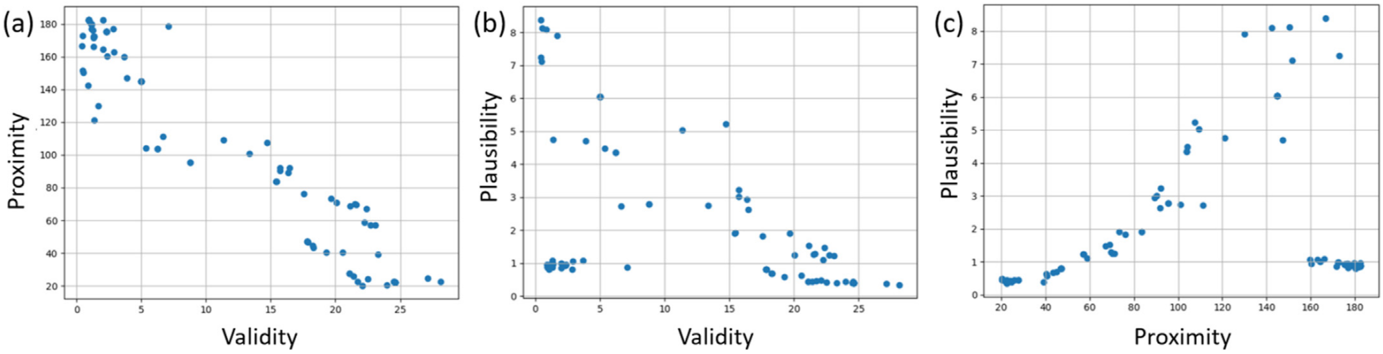


Fig. 7. Objective distribution for the group of counterfactual explanations. (a) Distribution between validity and proximity; (b) distribution between validity and plausibility; (c) distribution between proximity and plausibility.

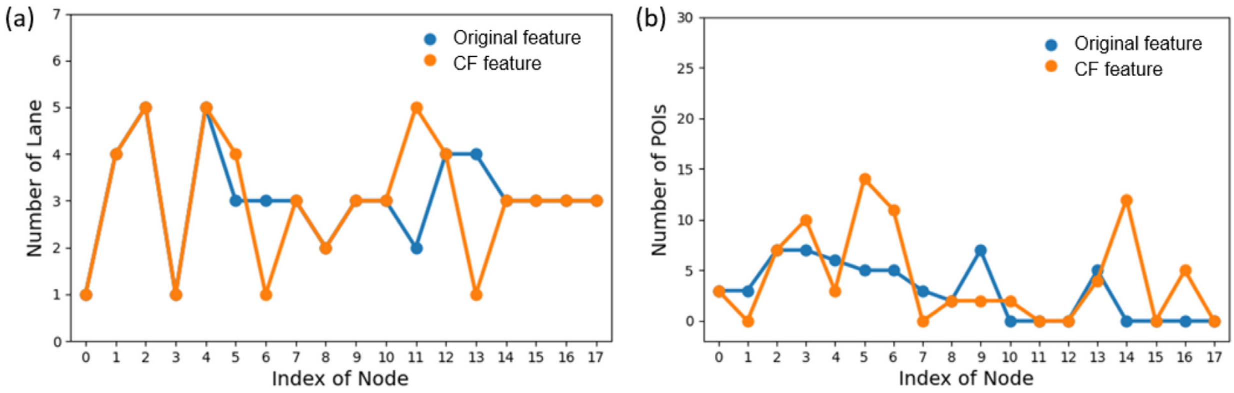


Fig. 8. Comparison between original features (in blue) and counterfactual features (in orange) for each road segment on Road I. The x-axis represents individual road segments and is arranged to follow the direction of traffic flow. (a) Comparison of the number of POIs; (b) comparison of the number of lanes.

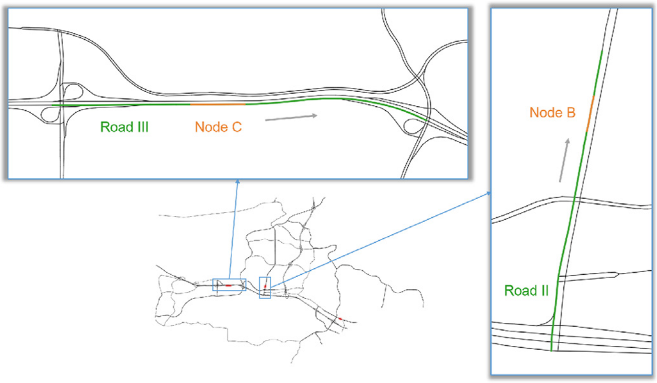


Fig. 9. Location of Node B and Node C. Node B is located on an urban road, Node C is located on a highway.

stronger influence on the speed of suburban roads.

Node B: On the other hand, for Node B, which is located on an urban road, Fig. 13 reveals that the counterfactual explanations generally advocate for a reduction in the number of POIs. This can be attributed to the high original count of nearby POIs, which likely contribute to traffic congestion. Thus, reducing the number of POIs is suggested to mitigate traffic demand. However, it is worth noting that, in the afternoon setting, the counterfactual number of POIs stays relatively consistent, which can be interpreted that in the weekday afternoon, the number of POIs has a small impact on the traffic of urban roads.

4.3.2. *Comparison of number of lanes*

If we compare the difference between the counterfactual number of lanes and the original number of lanes, as shown in Fig. 13b. There are no substantial changes for most time slots in both Node A and Node B. However, in the afternoon on the suburban road, the number of lanes drops by 19 compared to the original number of lanes, with most reduction occurring in the upstream part of the road segment. This can be interpreted that in the afternoon on the suburban road, counterfactual explanations suggest a decrease in the number of lanes (node index 0–10) before node B (index 11), thereby limiting the volume of cars and enabling smoother traffic flow.

4.3.3. *Comparison of speed limit*

Fig. 13c presents counterfactual speed limits for each setting. For Node A, the speed limit increases for all time slots except on the weekend, implying that changing the speed limit may not be effective on suburban roads during this period. For Node B, the counterfactual speed limits remain fairly consistent throughout weekdays but drop on weekends, possibly due to lower congestion levels.

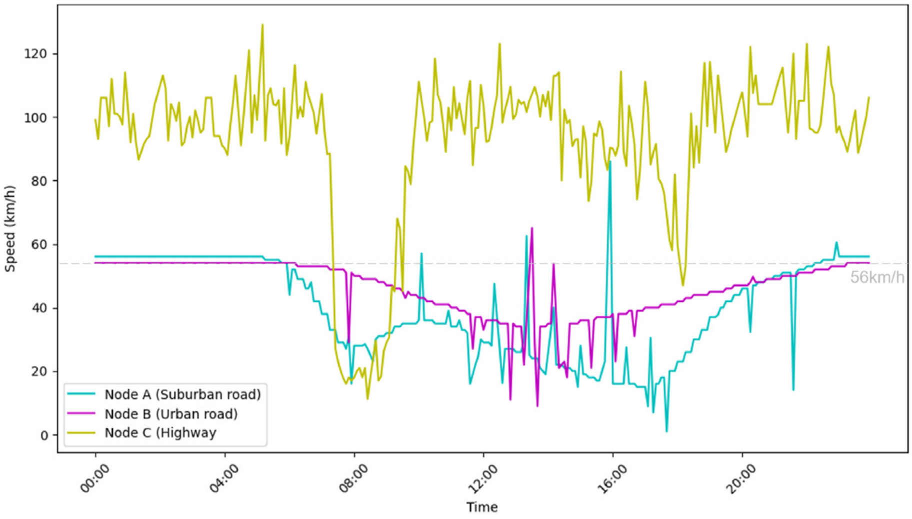


Fig. 10. Speed of Node A, Node B, and Node C on January 10th, 2019. The gray dashed line indicates the target speed of 56 km/h for generating counterfactuals.

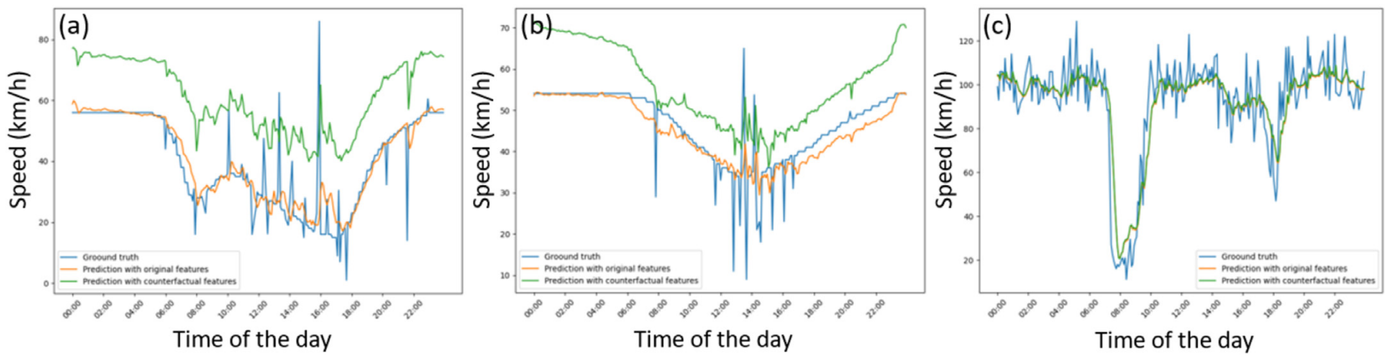


Fig. 11. Comparison of ground truth speed, original speed prediction, and counterfactual speed prediction for (a) Node A-suburban road, (b) Node B-urban road, and (c) Node C-highway on January 10, 2019.

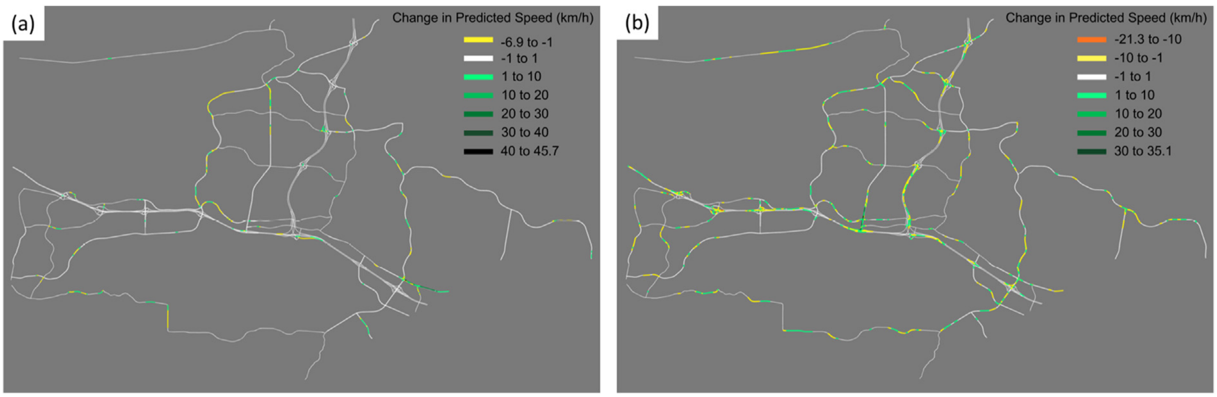


Fig. 12. Difference between original speed prediction and counterfactual speed prediction (km/h). Negative values indicate decrease in predicted speed, and positive values indicate increase in predicted speed: (a) impact of CFE for Node A; (b) impact of CFE for Node B.

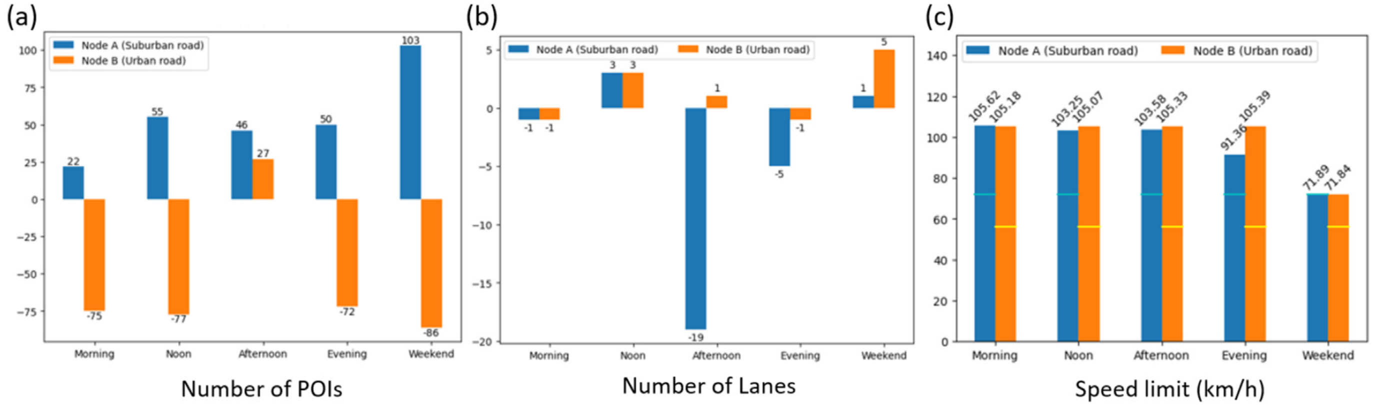


Fig. 13. Comparison of the difference between counterfactual and original features in different temporal settings for Node A and Node B. (a) Total difference in the number of POIs; (b) total difference in the number of lanes; (c) counterfactual speed limit (the original speed limit on Node A is 72 km/h, and on Node B is 56 km/h).

4.4. Experiments on scenario-driven counterfactual explanations

4.4.1. Directional constraints

Directional constraints allow users to specify the desired direction of feature change—either an increase or a decrease. In the scope of this experiment, several scenario-specific constraints are evaluated and compared. We focus on Node A on the suburban road for demonstration. The objective is to enhance the predicted speed between 9:00 and 10:00 to reach 56 km/h.

Despite the additional requirement on the direction of feature change, we want to ensure that the generated counterfactual explanations achieve the desired prediction (i.e., low validity loss) and are close to the feature space of the observational data (i.e., low plausibility loss). Therefore, we examined the validity and plausibility scores of scenario-based counterfactuals, as shown in Fig. 14. Fig. 14 shows even with the directional constraint, the distribution of the two objective scores falls within a similar range as the one without directional constraint.

In addition, we examine the distribution of the counterfactual explanations in terms of their total feature changes in Fig. 15 and the Appendix. The scatter plot visualizes the cumulative feature changes for each generated counterfactual explanation. In the 2D scatter plot, the axes represent the variations in the number of POIs and the number of

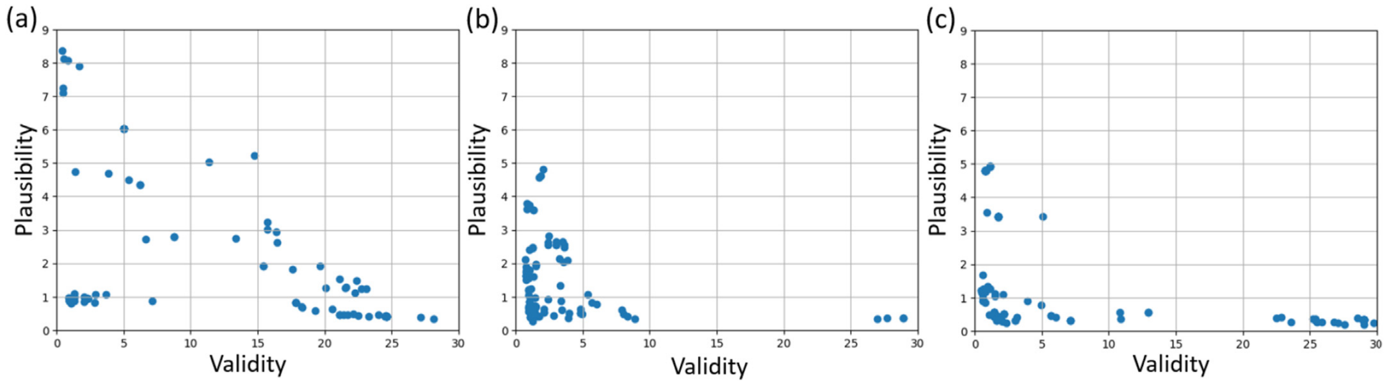


Fig. 14. Objective distribution (validity vs. plausibility) for different directional constraint settings: (a) no scenario constraint; (b) constraints on the number of POIs decreasing and the number of lanes increasing; (c) constraints on the number of POIs increasing.

lanes. Larger values on these axes signify greater differences between the counterfactual and original features. The 3D scatter plot adds a z-axis to display changes in speed limits. The color bar shows the validity score associated with each counterfactual, with a brighter color denoting a better performance of the counterfactual explanation. Detailed analyses of the three scenarios are presented below.

Scenario A: No directional constraints In this baseline scenario, counterfactual explanations were generated from Section 4.1. The scatter plot and its corresponding linear interpolation suggest that counterfactual explanations involving a greater increase in the number of POIs and a decrease in the number of lanes tend to yield superior performance, as evidenced by lower validity loss. This observation aligns well with previous findings specific to suburban roads. The 3D scatter plot illustrates that the larger the increase in the speed limit, the better the performance of the counterfactual.

Scenario C: Increase in POIs City planners may, at times, wish to enhance the infrastructure surrounding roads by introducing additional amenities like parking spaces, restaurants, or gas stations. However, they often aim to do this without adversely impacting road traffic. For this scenario, the aim is to increase the number of POIs and see how it affects the predicted traffic. Consequently, large penalties were applied to counterfactual features that proposed a decrease in POIs. The scatter plot indicates a shift in the distribution of the difference in the number of POIs for the generated counterfactual explanations. This shift leans towards a higher count, suggesting that the counterfactual explanations, under this constraint, tend to propose a greater number of POIs compared to the unconstrained baseline. Meanwhile, the distribution concerning the difference in the number of lanes remains unchanged.

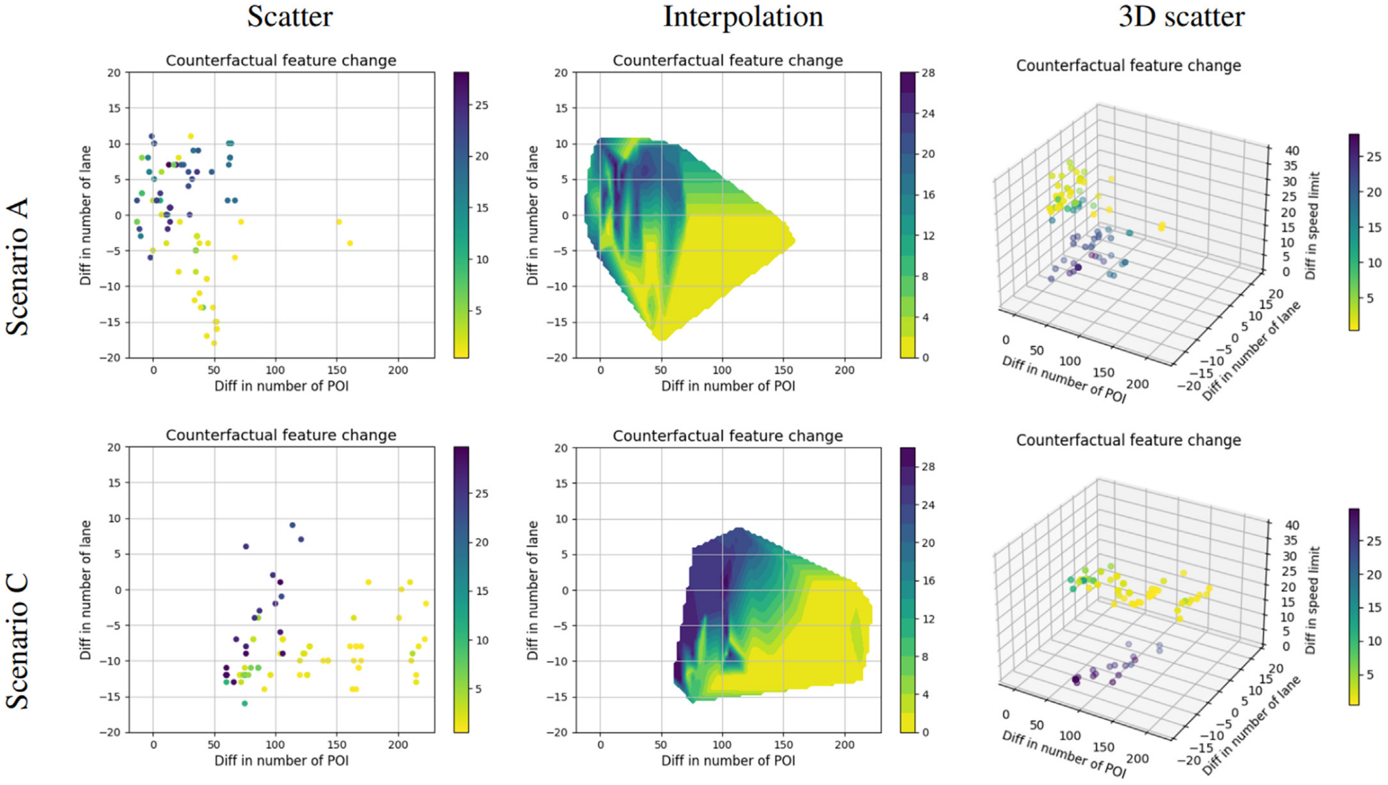


Fig. 15. Results for various directional constraints. Scenario A has no extra constraint; scenario B involves a decrease in the number of POIs and an increase in the number of lanes; scenario C involves an increase in the number of POIs. The “Scatter” column displays a scatter plot of the total feature change, where the color bar represents the validity score—the brighter the color, the better the counterfactual performance. The “Interpolation” column provides a linear interpolation based on the scatter plot data. The “3D Scatter” column presents a 3-dimensional scatter plot incorporating total feature changes, including variations in speed limit as z-axis.

4.5. Weighting constraints

User experience and expertise can guide the assignment of importance to different features, effectively serving as another layer of constraint. In this study, the target is consistently set for node B, an urban road. The aim is to improve the predicted speed between 9:00 and 10:00 to achieve a target speed of 56 km/h. Fig. 16 visualizes the results of the generated counterfactual explanations under different constraints. It is worth noting that the scatter plot in Fig. 16 displays the absolute differences between the original and counterfactual features.

Scenario D: No Weighting Constraints In this scenario, the counterfactual explanations generally perform better with a larger change in the number of lanes, while there is no discernible trend for the change in the number of POIs. As for the 3D scatter plot, it fails to indicate any significant correlation between variations in speed limit and the performance of the counterfactual explanations in terms of validity.

Scenario F: Preserve Number of Lanes In this setup, a higher weight is allocated to the number of lanes with the objective of minimizing alterations to this attribute. Both the scatter plot and the interpolation exhibit a constricted distribution range for the absolute difference between the original and counterfactual number of lanes. This outcome substantiates the effectiveness of this weighting strategy. It is noteworthy that when modifications to the number of lanes are restricted, the distribution of changes in the number of POIs also becomes more condensed. Compared to Scenario D, the scatter points are markedly clustered towards smaller differences in both lanes and POIs' counts. Moreover, adding this constraint appears to enhance the overall validity performance of the counterfactuals.

5. Discussion

5.1. Impact of contextual data on traffic forecasting

To affirm the assumption that incorporating contextual features enhances traffic forecasting, we undertook a systematic performance evaluation of models trained on various datasets. All models underwent training across 80 epochs for a fair comparison. The baseline model, trained exclusively on speed data, serves as the point of reference. Subsequently, we incorporated each contextual feature into the training data at a time and compared the resultant model's performance with that of the model trained using both speed and all contextual data.

As illustrated in Table 5, the comprehensive model that incorporates all contextual features demonstrates best performance across all the evaluation metrics. It exhibits the lowest values for RMSE (9.7578), MAE (6.4914), and Loss (95.2140) while achieving the highest scores in Accuracy (85.12%), R^2 (0.7931), and VAR (0.7940). At the same time, the model trained without any contextual data exhibited the least effective performance. It is worth noting that although these contextual features contribute to model accuracy, their overall enhancement of predictive performance is relatively limited, resulting in a modest reduction of merely 0.4 km/h in error, which suggests their role might be less critical in terms of model training. However, the utility of these features is notably underscored through the application of counterfactual explanations. With CFEs, it is possible to alter the prediction results with minor changes in the input contextual features, which can tell us the importance of input features in terms of sensitivity.

5.2. Comparison of CFEs in various spatial and temporal configurations

5.2.1. Impact of contextual features on highway traffic

Counterfactual explanations generated for highway road segments failed to yield improvements in speed. This suggests that the static features investigated in this study, namely the number of POI, the number of lanes, and speed limits, do not substantially influence traffic patterns on highways within the scope of this road network.

In the case of nearby POIs, their presence appears to have negligible impact on highway speeds, as highways generally lack direct access to these facilities. Regarding the number of lanes and speed limits, isolated adjustments to these parameters on specific highway segments seem ineffective at altering overall predicted speed. This is likely because highway traffic speed at a specific time is highly dependent on near historical traffic speeds and inflow conditions; altering the static attributes of only a section of the highway would not significantly impact the overall traffic demand or the carrying capacity of the entire highway network. Therefore, it will not increase the predicted speed in this situation.

5.2.2. Impact of contextual features on suburban road

When aiming to increase predicted speeds on suburban road segments, counterfactual explanations suggest an increase in the number of POIs nearby. This is because the model associates road segments with a higher density of nearby POIs with lower levels of traffic congestion.

The geographical location of a suburban road appears to significantly influence its traffic patterns. For instance, suburban roads adjacent to residential neighbourhoods may experience lighter traffic but with more nearby POIs. In contrast, other suburban roads might be part of arterial routes and, despite having fewer nearby POIs, experience higher traffic volumes, leading to increased congestion or reduced speeds. It is likely the deep learning model captured these associations, therefore the CFE recommends increasing the number of nearby POIs when trying to improve predicted speeds on specific suburban roads. This alteration makes these road segments contextually similar to quieter, residential suburban roads, where lower traffic volumes and less congestion are observed.

With regard to the number of lanes, the CFE does not suggest any significant modifications, except for the case of weekday afternoons, when the original traffic is the most congested and experiences the lowest speed. During these hours, the counterfactual explanations recommend reducing the number of lanes. Specifically, by reducing the number of lanes at the beginning of the road segment, less traffic would be able to enter the road segment, leading to more free traffic flow and overall higher speeds.

During weekends, the CFEs did not recommend alterations to the speed limit. This suggests that speed limits are not a significant factor affecting suburban road traffic forecasting during these times.

5.2.3. Impact of contextual features on urban road

In contrast to the suburban road, when targeting to increase speeds on urban road segments, counterfactual explanations suggest a decrease in the number of POIs nearby.

This discrepancy between urban and suburban roads could be interpreted in two ways. Firstly, it reflects the inherently different traffic patterns between suburban and urban settings. Secondly, it is important to note that the initial number of POIs near the studied urban road segments is already quite high. Unlike in suburban areas where an increase in POIs seems to alleviate congestion, urban roads appear to benefit from a reduction in POIs, presumably because fewer attractions would lead to less traffic. Interestingly, an exception arises during weekday afternoons, where the counterfactual explanations do not recommend a reduction in the number of POIs for urban roads. This could be because, during these peak hours, the number of POIs does not have a significant influence on the speed of traffic on urban roads.

5.3. Effectiveness of scenario-driven counterfactual explanations

The experimental results, obtained by incorporating various scenario constraints into the counterfactual explanation generation process, are highly promising for several reasons.

Firstly, all generated counterfactual explanations demonstrate reasonable validity and plausibility scores. This indicates that the method retains its efficacy to reach the set target even when additional constraints are applied, thereby affirming the feasibility and effectiveness of

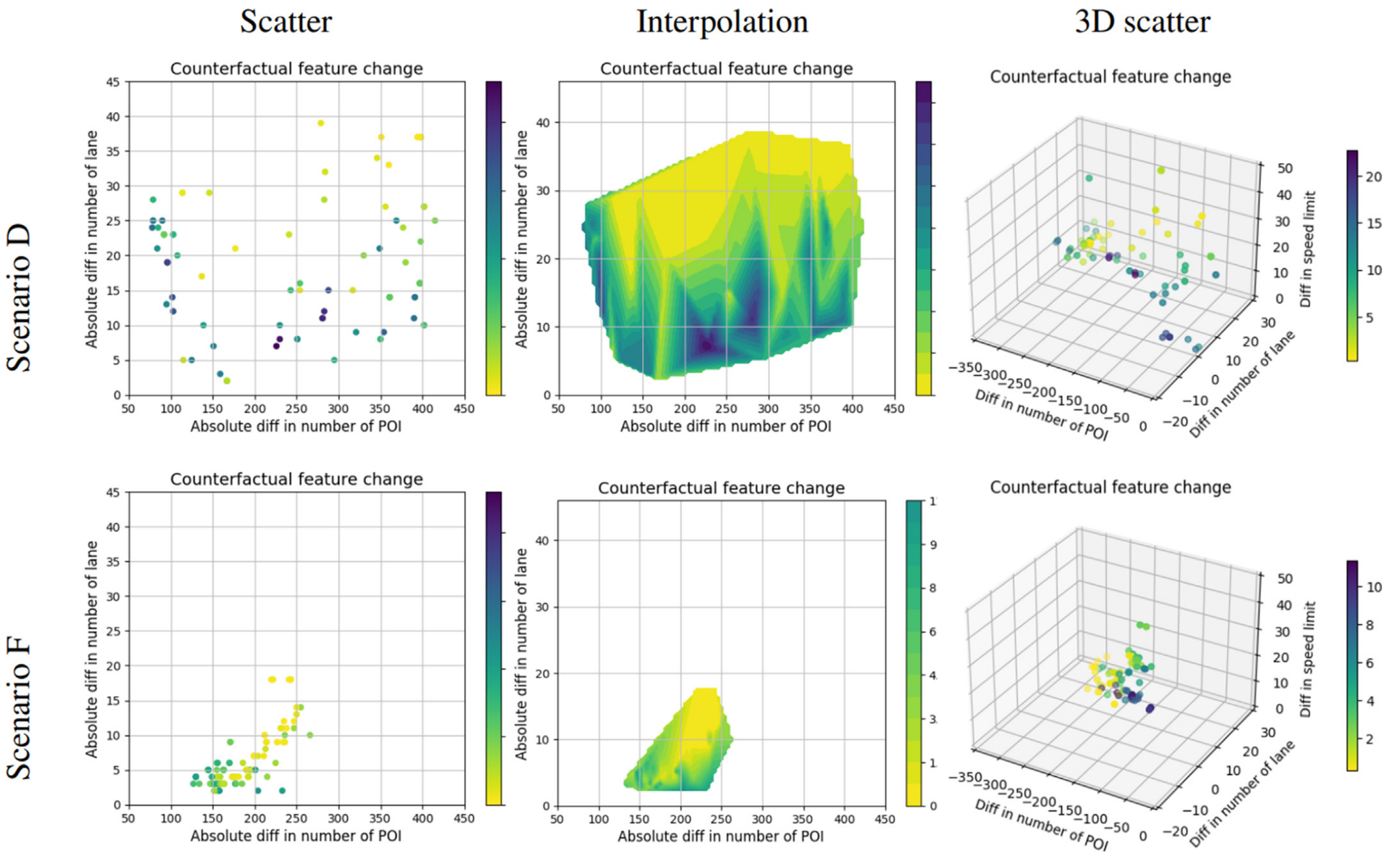


Fig. 16. Results for various weighting constraints. Scenario D has no extra constraint; scenario E preserves the number of POIs; scenario F preserves the number of Lanes; scenario G preserves both the number of POIs and the number of lanes; scenario H preserves the speed limit.

Table 5
Traffic forecasting model performance with different training datasets. Baseline indicates the model trained with only speed data. Full data shows the model trained with speed data and all contextual features. The "Loss" metric presents the loss value for the test data.

Metrics	Baseline	Number of lanes	Number of POI	Speed limit	Temperature	Precipitation	Wind	Humidity	Hour of day	Day of week	Full data
RMSE	10.2676	10.1596	10.2214	9.9265	10.2018	10.2208	10.2025	10.2295	10.2362	10.1841	**9.7578**
MAE	6.8945	6.6427	6.7095	7.0003	6.6097	6.6260	6.6018	6.7878	6.8190	6.8221	**6.4914**
Accuracy	84.35%	84.50%	84.42%	84.87%	84.45%	84.42%	84.44%	84.40%	84.39%	84.47%	**85.12%**
R^2	0.7709	0.7754	0.7727	0.7874	0.7738	0.7729	0.7737	0.7724	0.7722	0.7747	**0.7931**
VAR	0.7719	0.7754	0.7727	0.7939	0.7745	0.7732	0.7744	0.7727	0.7727	0.7757	**0.7940**
Loss	105.4242	103.2179	104.4773	98.5349	104.0777	104.4658	104.0910	104.6418	104.7800	103.7168	**95.2140**

the approaches proposed in this study.

Secondly, some constraints facilitate more efficient counterfactual generation. On the one hand, the collection of generated Counterfactual Explanations generally exhibits lower validity loss, implying proper performance in aligning the predicted speeds with target speeds. On the other hand, underweighting constraints in Fig. 16, not only do the colors in the set of CFEs become more vibrant, but the scatter points also converge within a smaller area. This indicates increased efficiency after adding the scenario constraint, as the algorithm is more adept at identifying optimal counterfactuals within a constrained search space.

In summary, the integration of user-defined prior knowledge into post-hoc explanations has proven to be valuable. This not only addresses the initial research questions posed but also has profound implications for future work in the field of Explainable AI.

5.4. Limitations and potential work

The use of deep learning models, coupled with Counterfactual Explanations, provides a powerful combination for uncovering complex relationships between variables. These relationships may be too subtle or intricate for humans to notice, thus highlighting the novel capabilities of explainable AI and deep learning in data analysis.

However, similar to all data-driven approaches, the effectiveness of this method depends on the quality and diversity of the training data. A limitation of our study is that the generalizability of the model may be constrained by the dataset, which is restricted to a limited number of road segments and contextual features. While our results are promising, broader applicability to different environments or traffic conditions may require more diverse data to ensure robustness and transferability of the generated counterfactuals.

One potential avenue for mitigating these limitations involves the incorporation of domain-specific knowledge into the data-driven models. This can enhance the generalizability and reliability of the model's recommendations. In light of this, scenario-driven counterfactual explanations are proposed. While our work demonstrates that scenario-driven counterfactual explanations offer considerable benefits in the context of integrating prior constraints, a key question that remains is how to ensure the practical utility and broader applicability of these methods in real-world settings.

While this study focuses primarily on the space of machine learning models, it is important to note that the insights gained can still be highly valuable in real-world applications. For instance, AI model developers in traffic forecasting can leverage our approach to detect potential vulnerabilities, prevent adversarial attacks, and enhance model robustness. Similarly, urban planners can use our model to gain a deeper understanding of which factors exert the greatest influence on traffic predictions, supporting more informed decision-making. In future work, we aim to refine our approach by collaborating more closely with practitioners to explore the practical challenges of deploying counterfactual explanations in real-world environments.

6. Conclusions

We introduce a comprehensive framework that advances the use of counterfactual explanations in spatiotemporal prediction tasks, effectively bridging the gap between theoretical understanding of models and their practical implications for generating insights.

In this study, a deep learning-based traffic forecasting model was trained at first, using the state-of-the-art architecture, attribute augmented spatiotemporal graph convolutional networks. Subsequently, we generated diverse sets of counterfactual explanations by targeting various spatial and temporal settings.

On the one hand, by suggesting minimal alterations to input features, counterfactual explanations enhance our understanding of the model's behavior and elucidate the role of various contextual variables in deep learning-based traffic forecasting. This provides invaluable insights for AI practitioners, aiding in a deeper comprehension of what the model has learned from the data. More specifically, by examining a variety of spatial settings—such as suburban roads, urban roads, and highways, as well as different time slots, this study reveals that the impact of static contextual features on traffic speed is influenced by distinct spatial and temporal conditions. On the other hand, this study advances the field by introducing scenario-driven counterfactual explanations, which offer domain experts like urban planners insightful recommendations tailored to specific scenarios. By integrating user-defined constraints into our framework, we can provide insights that are directly applicable to a range of real-world conditions. Specifically, we introduce two methods for incorporating these scenario constraints: directional and weighting constraints. Both approaches effectively align the generated counterfactual explanations with users' prior knowledge and expectations, thereby making the search for optimal solutions more efficient. Importantly, we observed that some scenarios, particularly those incorporating weighting constraints, expedited the generation process and yielded more precise and useful CFEs. This is manifested through a more focused distribution of CFEs, indicating a clearer pathway for the algorithm to identify optimal counterfactual conditions.

Although this study has successfully leveraged counterfactual explanations to interpret traffic forecasting models and provided valuable insights via scenario-driven counterfactuals, several promising avenues for future research exist. Upcoming investigations could focus on.

- **Cross-validation with Other Deep Learning Models** Although CFEs is a model agnostic method to generate explanations and this study primarily contributes to the general approach of applying CFEs to spatiotemporal prediction tasks, future research should expand this approach to other traffic forecasting models to evaluate and compare the performance and interpretability across different architectures.
- **Geographic Generalizability:** The current framework relies heavily on data from a specific geographical region. Future studies should aim to validate and adapt the model across diverse geographical settings, thereby assessing its ability to generalize the identified correlations between contextual features and traffic behaviors.
- **Fine-Grained Feature Analysis:** While the present study broadly examines the impact of POIs on traffic dynamics, subsequent research should delve into how different categories of POIs individually influence traffic patterns.
- **Inclusion of Dynamic Temporal Elements:** This study primarily focuses on altering static features for generating counterfactuals. Future research should expand the scope to include conducting counterfactuals on time-dependent features, potentially unveiling intricate, time-sensitive patterns that impact traffic conditions. This would entail the development of time-series counterfactual explanations, which is still an under-explored area in current literature.
- **Collaboration with Domain Experts:** Future work should actively involve domain experts, such as urban planners, to better incorporate real-world insights and practical constraints in the modeling process. This collaboration will improve the model's applicability and utility in decision-making processes.

CRediT authorship contribution statement

Rushan Wang: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yanan Xin:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Yatao Zhang:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Fernando Perez-Cruz:** Writing – review & editing, Funding acquisition, Conceptualization. **Martin Raubal:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Replication and data sharing

The main code for this project can be found in the GitHub repository: https://github.com/RushanWang1/CFforTrafficForecast.git.

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.

Acknowledgements

We would like to thank HERE Technologies for providing the traffic data used in this study. The study is supported by the Hasler Foundation under the project Interpretable and Robust Machine Learning for Mobility Analysis (No. 21041). The study is partially conducted at the Future Resilient Systems program at the Singapore-ETH Centre, supported by the National Research Foundation (NRF) Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

Appendix. Results for directional constraints

Scenario A: No directional constraints In this baseline scenario, counterfactual explanations were generated from Section 4.1. The scatter plot and its corresponding linear interpolation suggest that counterfactual explanations involving a greater increase in the number of POIs and a decrease in the number of lanes tend to yield superior performance, as evidenced by lower validity loss. This observation aligns well with previous findings specific to suburban roads. The 3D scatter plot illustrates that the larger the increase in the speed limit, the better the performance of the counterfactual.

Scenario B: Decrease in POIs, increase in Lanes In this scenario, the counterfactual explanations are generated with directional constraints to reduce the number of POIs and increase the number of lanes. Based on the results in Fig. A1, regarding the change in the number of POIs for each counterfactual, the distribution range remains relatively stable. In contrast, the distribution range for the change in the number of lanes broadens, with an increasing number of counterfactuals reflecting a lane increase. Another noteworthy observation is that when this constraint is applied, the resulting counterfactual explanations tend to be associated with brighter colors on the validity score scale, implying lower validity loss. This suggests that these constrained counterfactuals generally outperform those generated under the original, unconstrained setting.

Scenario C: Increase in POIs City planners may, at times, wish to enhance the infrastructure surrounding roads by introducing additional amenities like parking spaces, restaurants, or gas stations. However, they often aim to do this without adversely impacting road traffic. For this scenario, the aim is to increase the number of POIs and see how it affects the predicted traffic. Consequently, large penalties were applied to counterfactual features that proposed a decrease in POIs.

The scatter plot indicates a shift in the distribution of the difference in the number of POIs for the generated counterfactual explanations. This shift leans towards a higher count, suggesting that the counterfactual explanations, under this constraint, tend to propose a greater number of POIs compared to the unconstrained baseline. Meanwhile, the distribution concerning the difference in the number of lanes remains unchanged.

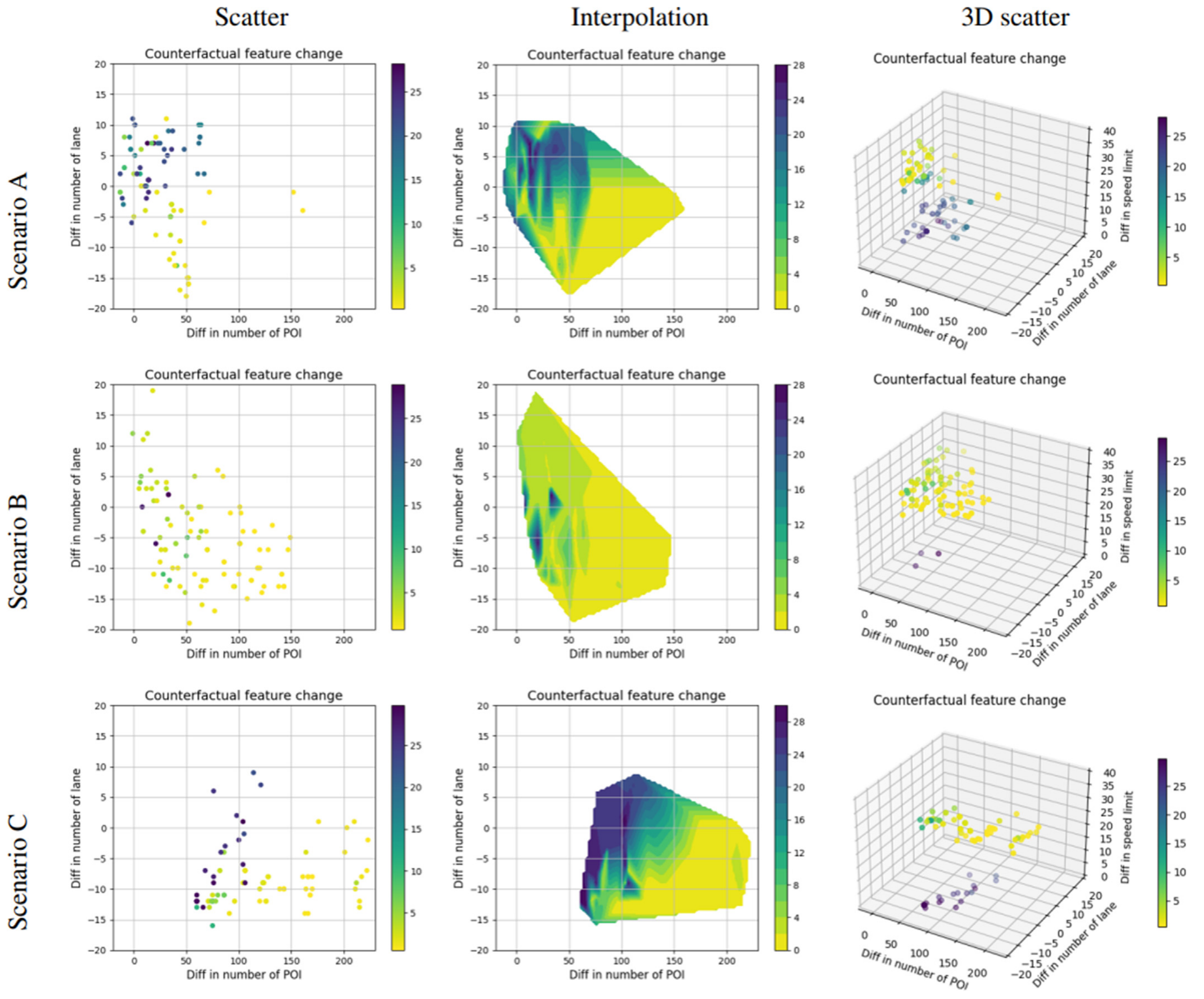


Fig. A1. Results for various directional constraints.

Appendix. Results for weighting constraints

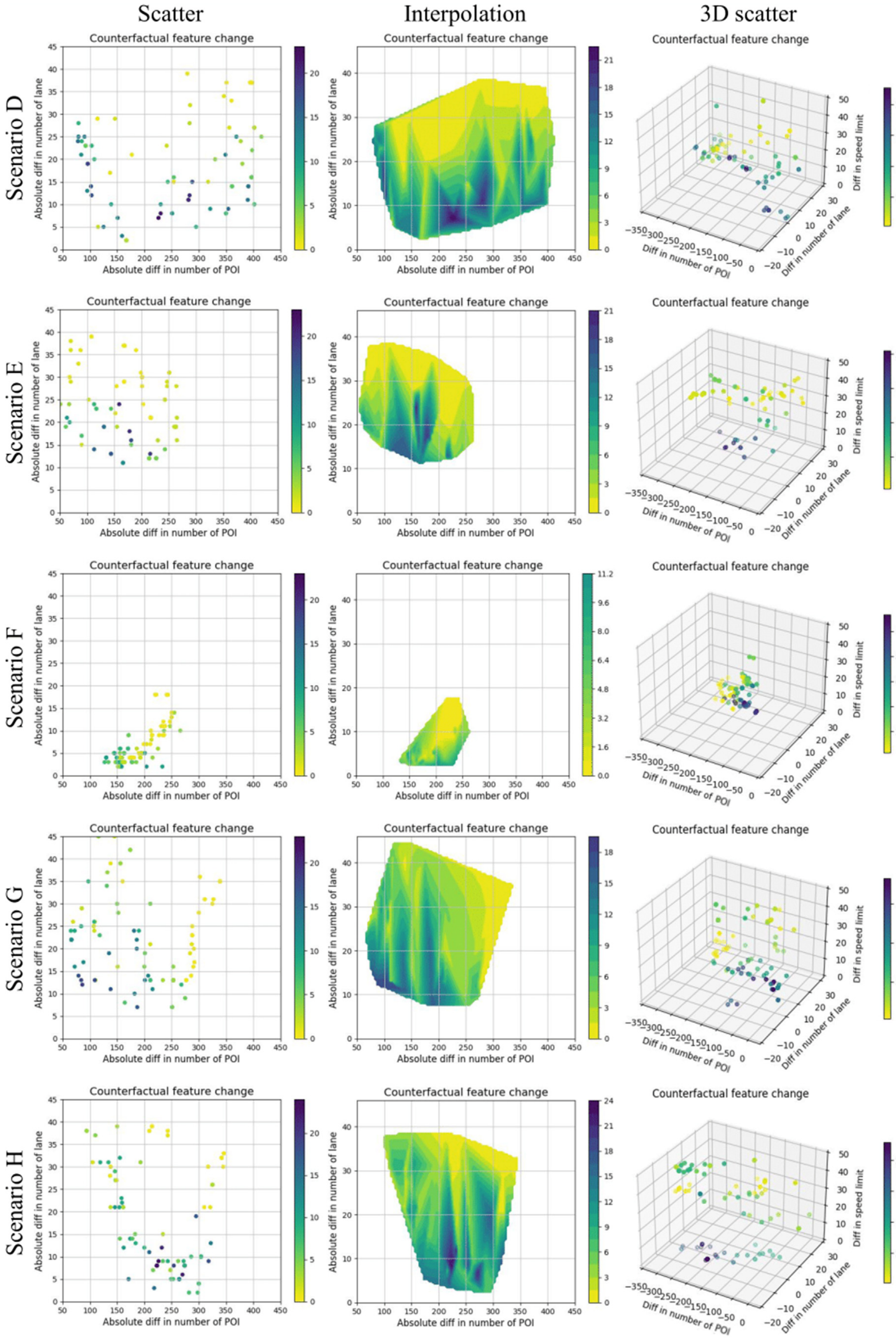


Fig. A2. Results for various weighting constraints.

Fig. A2 visualizes the results of the generated counterfactual explanations under different constraints. It is worth noting that the scatter plot in Fig. A2 displays the absolute differences between the original and counterfactual features.

Scenario D: No Weighting Constraints In this scenario, the counterfactual explanations generally perform better with a larger change in the number of lanes, while there is no discernible trend for the change in the number of POIs. As for the 3D scatter plot, it fails to indicate any significant correlation between variations in speed limit and the performance of the counterfactual explanations in terms of validity.

Scenario E: Preserve Number of POIs In this configuration, we assign a higher weight to the number of POIs to discourage substantial alterations to this feature. The scatter plot and its corresponding interpolation reveal a narrower distribution range for the absolute difference between the counterfactual and original number of POIs, validating the efficacy of this weighting approach. Noticeably, when the changes to the number of POIs are constrained, the distribution of alterations in the number of lanes tends to cluster towards the higher end of the range.

Scenario F: Preserve Number of Lanes In this setup, a higher weight is allocated to the number of lanes with the objective of minimizing alterations to this attribute. Both the scatter plot and the interpolation exhibit a constricted distribution range for the absolute difference between the original and counterfactual number of lanes. This outcome substantiates the effectiveness of this weighting strategy. It is noteworthy that when modifications to the number of lanes are restricted, the distribution of changes in the number of POIs also becomes more condensed. Compared to Scenario D, the scatter points are markedly clustered towards smaller differences in both lanes and POIs' counts. Moreover, adding this constraint appears to enhance the overall validity performance of the counterfactuals.

Scenario G: Preserve Both Number of POIs and Number of Lanes In this setup, significant weights are allocated to both the number of POIs and lanes, guiding the model to focus on modifying speed limits. Interestingly, the scatter plot shows that this constraint only moderately limits alterations in the number of POIs. Moreover, it does not restrain changes in lane count. With respect to counterfactual performance, more effective counterfactuals seem to be concentrated in areas showing larger differences in the number of POIs. As for the overall performance of the set of counterfactual explanations, a decrease in validity loss suggests enhanced efficacy.

Scenario H: Preserve Speed Limit This scenario attaches a high weight to the speed limit, directing the model to search for counterfactuals that predominantly alter other features while keeping the original speed limit intact. Based on the scatter plot, this constraint does not yield a noticeable impact on the magnitude of changes in any specific counterfactual features. In terms of performance, higher-quality counterfactuals are more likely to be located in regions showing substantial differences in the number of lanes.

References

Ates, E., Aksar, B., Leung, V.J., Coskun, A.K., 2021. Counterfactual explanations for multivariate time series. In: 2021 International Conference on Applied Artificial Intelligence (ICAPAI), Halden, Norway, vol. 2021, pp. 1–8.

Barredo-Arrieta, A., Laña, I., Del Ser, J., 2019. What lies beneath: a note on the explainability of black-box machine learning models for road traffic forecasting. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 2232–2237.

Chen, R., Liang, C.Y., Hong, W.C., Gu, D.X., 2015. Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. Appl. Soft Comput. 26, 435–443.

Covert, I.C., Lundberg, S., Lee, S.-I., 2021. Explaining by removing: a unified framework for model explanation. J. Mach. Learn. Res. 22, 1–90.

Dandl, S., Molnar, C., Binder, M., Bischl, B., 2020. Multi-Objective Counterfactual Explanations. PPSN XVI. Springer International Publishing, pp. 448–469.

Deb, K., Agrawal, S., Pratap, A., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6, 182–197.

Diakopoulos, N., 2020. Accountability, Transparency. In: The Oxford handbook of ethics of AI. Oxford University Press.

Edwards, L., Veale, M., 2017. Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. Duke L. & Tech. Rev. 16, 18.

Fernandez, C., Provost, F.J., Han, X., 2020. Explaining data-driven decisions made by AI systems: the counterfactual approach. arXiv preprint arXiv:2001.07417. https://arxiv.org/abs/2001.07417.

Gaur, L., Sahoo, B.M., 2022. Explainable Artificial Intelligence for Intelligent Transportation Systems: Ethics and Applications. Springer Nature.

Jiang, W., Luo, J., 2021. Graph neural network for traffic forecasting: a survey. Expert Syst. Appl. 207, 117921.

Johansson, U., Bostrom, H., Lofstrom, T., Linusson, H., 2014. Regression conformal prediction with random forests. Mach. Learn. 97, 155–176.

Jonietz, D., Sester, M., Stewart, K., Winter, S., Tomko, M., Xin, Y., 2022. Urban mobility analytics: report from dagstuhl seminar 22162. Dagstuhl Reports 12, 26–53.

Jung, H.-G., Kang, S.-H., Kim, H.-D., Won, D.-O., Lee, S.-W., 2022. Counterfactual explanation based on gradual construction for deep networks. Pattern Recogn. 132, 108958.

Kruber, F., Wurst, J., Botsch, M., 2018. An unsupervised random forest clustering technique for automatic traffic scenario categorization. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 2811–2818.

Lee, S., Fambro, D.B., 1999. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. Transp. Res. Rec. 1678, 179–188.

Li, Y., Shahabi, C., 2018. A brief overview of machine learning methods for short-term traffic forecasting and future directions. SIGSPATIAL Special 10, 3–9.

Li, Y., Yu, R., Shahabi, C., Liu, Y., 2017. Graph convolutional recurrent neural network: data-driven traffic forecasting. arXiv preprint arXiv:1707.01926 7 (8). http://arxiv.org/abs/1707.01926.

Li, J., Xin, Y., Hong, Y., Raubal, M., 2023. Interpreting Deep Learning Models for Traffic Forecast: A Case Study of UNet.

Liao, B., Zhang, J., Wu, C., McIlwraith, D., Chen, T., Yang, S., Guo, Y., Wu, F., 2018. Deep sequence learning with auxiliary information for traffic prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 537–546.

Liu, S., Kailkhura, B., Loveland, D., Han, Y., 2019. Generative counterfactual introspection for explainable deep learning. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1–5.

Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. http://arxiv.org/abs/1705.07874.

Lv, Y., Duan, Y., Kang, W., Li, Z.X., Wang, F., 2015. Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transport. Syst. 16, 865–873.

Ma, X., Ding, C., Luan, S., Wang, Y., Wang, Y., 2017. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. IEEE Trans. Intell. Transport. Syst. 18, 2303–2310.

Marcinkevics, R., Vogt, J., 2023. Interpretable and explainable machine learning: a methods-centric overview with concrete examples. Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. 13, e1493.

Meena, G., Sharma, D., Mahrishi, M., 2020. Traffic prediction for intelligent transportation system using machine learning. In: 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), pp. 145–148.

Molnar, C., 2022. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, second ed. Online https://christophm.github.io/interpretable-ml-book.

Mothilal, R.K., Sharma, A., Tan, C., 2019. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 607–617.

Parafita, Á., Vitrià, J., 2019. Explaining visual models by causal attribution. arXiv preprint arXiv:1909.08891. https://arxiv.org/abs/1909.08891.

Polson, N.G., Sokolov, V.O., 2017. Deep learning for short-term traffic flow prediction. Transport. Res. C Emerg. Technol. 79, 1–17.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P., 2020. FACE: feasible and actionable counterfactual explanations. In: Proceedings of the AAAI/ACM Conference on AI. Ethics, and Society, pp. 344–350.

Prado-Romero, M.A., Prenkaj, B., Stilo, G., Giannotti, F., 2022. A survey on graph counterfactual explanations: definitions, methods, evaluation, and research challenges. ACM Comput. Surv. 56, 1–37.

Prasad, S.C., Prasad, P., 2014. Deep recurrent neural networks for time series prediction. http://arxiv.org/abs/1407.5949.

Ramakrishnan, N., Soni, T., 2018. Network traffic prediction using recurrent neural networks. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 187–193.

Singla, S., 2022. Deep learning for medical imaging from diagnosis prediction to its counterfactual explanation. arXiv 2209, 02929. https://arxiv.org/abs/2209.02929.

Spooner, T., Dervovic, D., Long, J., Shepard, J., Chen, J., Magazzeni, D., 2021. Counterfactual explanations for arbitrary regression models. arXiv preprint arXiv: 2106.15212. https://arxiv.org/abs/2106.15212.

Wachter, S., Mittelstadt, B., Russell, C., 2018. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. JL & Tech. 31, 841.

Wu, C.-H., Ho, J.-M., Lee, D.-T., 2004. Travel-time prediction with support vector regression. IEEE Trans. Intell. Transport. Syst. 5, 276–281.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S., 2019a. A comprehensive survey on graph neural networks. IEEE Transact. Neural Networks Learn. Syst. 32, 4–24.

Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C., 2019b. Graph WaveNet for deep spatial-temporal graph modeling. arXiv preprint arXiv:1906.00121. http://arxiv.org/abs/1906.00121.

Xin, Y., Tagasovska, N., Perez-Cruz, F., Raubal, M., 2022. Vision paper: causal inference for interpretable and robust machine learning in mobility analysis. In: Proceedings of the 30th International Conference on Advances in Geographic Information Systems, pp. 1–4.

Xin, Y., Hong, Y., Dirmeier, S., Perez-Cruz, F., Raubal, M.. Evaluating the robustness of deep learning models for mobility prediction through causal interventions. https://ethz.ch/content/dam/ethz/special-interest/mavt/csfm-dam/events/2023/symposium/posters/mp/xin-ikg-causal-interventions-presentation.pdf.

Xu, Y., Kong, Q.-J., Klette, R., Liu, Y., 2014. Accurate and interpretable Bayesian MARS for traffic flow prediction. IEEE Trans. Intell. Transport. Syst. 15, 2457–2469.

Yang, Y., Du, K., Dai, X., Fang, J., 2023. Counterfactual graph transformer for traffic flow prediction. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), pp. 521–527.

Yazici, M.A., Kamga, C., Ozbay, K., 2014. Highway versus urban roads: analysis of travel time and variability patterns based on facility type. Transp. Res. Rec. 2442, 53–61.

Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., Yin, B., 2022. Deep learning on traffic prediction: methods, analysis, and future directions. IEEE Trans. Intell. Transport. Syst. 23, 4927–4943.

Zarei, N., Ghayour, M.A., Hashemi, S., 2013. Road traffic prediction using context-aware random forest based on volatility nature of traffic flows. In: Intelligent Information and Database Systems: 5th Asian Conference, ACIIDS 2013, Kuala Lumpur, Malaysia, March 18-20, 2013, Proceedings, Part I, vol. 5, pp. 196–205.

Zhang, W., Lim, B.Y., 2022. Towards relatable explainable AI with the perceptual process. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1–24.

Zhang, Y., Zhao, T., Gao, S., Raubal, M., 2023. Incorporating multimodal context information into traffic speed forecasting through graph deep learning. IJGIS 37, 1909–1935.

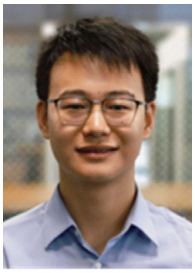
Zhang, Y., Wang, Y., Gao, S., Raubal, M., 2024. Context-aware knowledge graph framework for traffic speed forecasting using graph neural network. IEEE Trans. Intell. Transport. Syst. 26, 3885–3902.

Zhao, L., Song, Y., Deng, M., Li, H., 2018. Temporal graph convolutional network for urban traffic flow prediction method. arXiv preprint arXiv:1811.05320. http://arxiv.org/abs/1811.05320.

Zhu, J., Tao, C., Deng, H., Zhao, L., Wang, P., Lin, T., Li, H., 2020. AST-GCN: attribute-augmented spatiotemporal graph convolutional network for traffic forecasting. IEEE Access 9, 35973–35983.

Rushan Wang is a Ph.D. student at ETH Zurich. Her research focuses on the intersection of artificial intelligence and geoinformation. She holds the M.S. degree in Geomatics Engineering from ETH Zurich and the B.S. degree in Geographic Information Science from Wuhan University.

Yanan Xin is an Assistant Professor at the Department of Transport and Planning at TU Delft, co-Director of the DAI-MoND (Digitization and AI for Mobility Network Dynamics) lab. With a goal of promoting responsible AI transformation in transportation, her research interests and methodologies cover interpretable machine learning, spatial causal inference, mobility-based anomaly detection, and the intersection between mobility and energy. Before joining TU Delft, she was a Senior Assistant and Lecturer at ETH Zurich, leading the Mobility Information Engineering Lab. She received the M.S. degree of Urban Spatial Analytics from the University of Pennsylvania and the Ph.D. degree in Geography (Specialization: Geographic Information Science) from the Pennsylvania State University.



Yatao Zhang is a Ph.D. student at the Mobility Information Engineering lab at ETH Zurich and the Future Resilient Systems at the Singapore-ETH Center. His research interests lie in context-based spatiotemporal analysis, geospatial big data mining, and traffic forecasting. He received the B.S. and the M.S. degrees in Geographical Information Science from Sun Yat-sen University, Guangzhou, China, in 2017 and Wuhan University, Wuhan, China, in 2020, respectively.

Fernando Perez-Cruz received the Ph.D. degree in Electrical Engineering from the Technical University of Madrid. He is Titular Professor in the Computer Science Department at ETH Zurich and Head of Machine Learning Research and AI at Spiden. He has been a member of the technical staff at Bell Labs and a Machine Learning Research Scientist at Amazon. He has been a visiting Professor at Princeton University under a Marie Curie Fellowship and an associate professor at University Carlos III in Madrid.

Martin Raubal is a Professor of Geoinformation Engineering at ETH Zurich. His research interests focus on spatial decision-making for sustainability, more specifically he concentrates on analyzing spatio-temporal aspects of human mobility, spatial cognitive engineering and mobile eye-tracking to investigate visual attention while interacting with geoinformation and in spatial decision situations.