TITAN: A Spatiotemporal Feature Learning Framework for Traffic Incident Duration Prediction

Kaiqun Fu[1], Taoran Ji[1], Liang Zhao[2], Chang-Tien Lu[1]
{fukaiqun,jtr}@vt.edu,lzhao9@gmu.edu,ctlu@vt.edu
[1]Virginia Tech
[2]George Mason University

ABSTRACT

Critical incident stages identification and reasonable prediction of traffic incident duration are essential in traffic incident management. In this paper, we propose a traffic incident duration prediction model that simultaneously predicts the impact of the traffic incidents and identifies the critical groups of temporal features via a multi-task learning framework. First, we formulate a sparsity optimization problem that extracts low-level temporal features based on traffic speed readings and then generalizes higher level features as phases of traffic incidents. Second, we propose novel constraints on feature similarity exploiting prior knowledge about the spatial connectivity of the road network to predict the incident duration. The proposed problem is challenging to solve due to the orthogonality constraints, non-convexity objective, and non-smoothness penalties. We develop an algorithm based on the alternating direction method of multipliers (ADMM) framework to solve the proposed formulation. Extensive experiments and comparisons to other models on real-world traffic data and traffic incident records justify the efficacy of our model.

CCS CONCEPTS

• **Information systems → Data mining**; • **Computing methodologies → Feature selection**; • **Applied computing → Transportation**.

KEYWORDS

intelligent transportation systems, feature learning, incident impact analysis

ACM Reference Format:
Kaiqun Fu[1], Taoran Ji[1], Liang Zhao[2], Chang-Tien Lu[1]. 2019. *TITAN*: A Spatiotemporal Feature Learning Framework for Traffic Incident Duration Prediction. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3347146.3359381

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11…$15.00
https://doi.org/10.1145/3347146.3359381

1 INTRODUCTION

The studies of early detecting the traffic incidents and estimating the impact of the non-recurrent congestions caused by traffic incidents have become increasingly important research topics due to the significant social and economic losses generated. A one-minute reduction on congestion duration produces a 65 US dollars gain per incident [1]. Although non-recurrent congestion is hard to predict due to its nature of randomness, the studies on impact and duration of the traffic incidents are still ones of the major focuses for the traffic operators. The vast deployment of transportation traffic speed sensors and Traffic Incident Management Systems (TIMS) make the traffic speed data and traffic incident records ubiquitously accessible for the transportation operators. With the abundance of the traffic data sources, an efficient multi-task learning model can be implemented to provide an accurate prediction on incident duration.

Incident duration is the time elapsed from the incident occurrence until all evidence of the incident has been removed from the incident scene. From the perspective of traffic management and operation, the life cycle of a traffic incident is split into five stages: Detection, Verification, Response, Clearance, and Recovery [20]. Figure 1 shows the life cycle of a traffic incident. However, the five-stage life cycle separation cannot be used directly as the temporal features for the traffic incident duration prediction. To accurately estimate the duration of a traffic incident in its early stages, the transportation operators and first responders encounter three major challenges: **1) No explicit high-level temporal features**: Although the conventional five-stage life cycle separation is effective for the purposes of traffic management, such five-stages cannot be considered as temporal features in traffic incident duration prediction task. It is important to group the critical time point features in the early stages of the incident forming higher level time periods that can perform as a better indicator for predicting the incident duration. **2) Hard to predict the influence of incident**: In the research field of Traffic Incident Management, one of the most essential tasks is to estimate the impact of the traffic incident in terms of its temporal duration at early stages. However, the performances of the conventional time series based methods are limited by their incapability of identifying higher level temporal features. **3) Spatial connectivity of the road networks is rarely considered**: The traffic congestion cascades within the road network. As a consequence, the traffic patterns of incidents in their early stages are similar when the traffic incidents are topologically closer from the perspective of the road networks. Traffic incidents that are spatiotemporally closer should share more similar traffic speed patterns. However, this spatial correlation between traffic incidents is rarely considered in the previous studies [15].

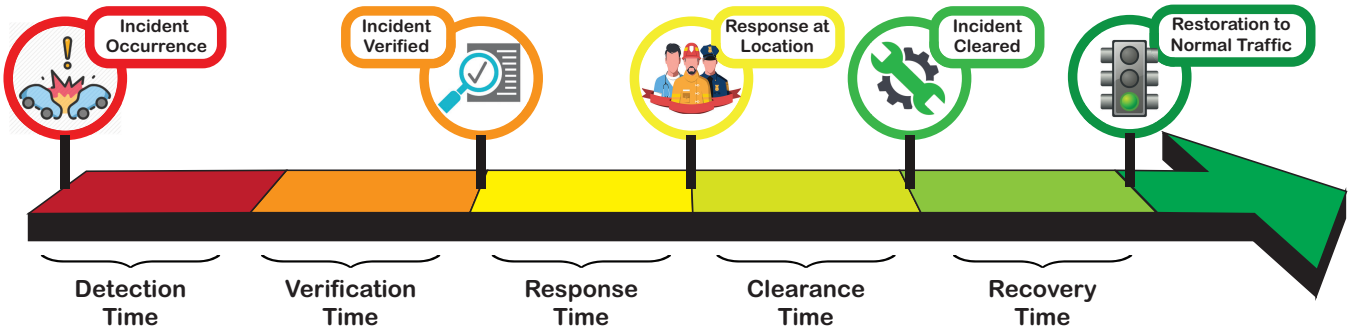


Figure 1: From the perspectives of traffic management and transportation operations, the life cycle of a traffic incident is separated into five stages: Detection, Verification, Response, Clearance, and Recovery

The existing methods are mostly infeasible to solve these challenges. Current feature learning methods such as ℓ_1-norm regularized methods such as Lasso [28] have properties in terms of feature selection. However, strong assumptions on the design matrix are required [39]. Zhan et al. [32] propose an M5P tree algorithm to predict the clearance time of traffic incident based on the geometric, and traffic features. Feature learning algorithms for biomarker identification [40] and social event indicators [36] are proved to be effective while finding higher level features. However, most of them focus on learning important feature sets from attributes and does not apply to our encountered problem due to expensive computation. In these studies, they considered the duration of an incident to quantify the impact. However, their quantification strategies are designed to capture the one-time impact of the incident, instead of the time-varying nature of impact at different locations. Multi-task learning based spatiotemporal model plays an important role while considering the connectivity of the road networks. Multi-task based spatiotemporal models focus on regression and classification problems such as county income prediction [33], social unrest event forecasting [37], and even service disruption detection for transit networks [10]. However, none of the previously proposed methods is capable of modeling the spatial connectivity between features at a higher level. Therefore, most of the existing models are not suitable for our traffic incident duration prediction problem.

To address these challenges, we propose a **T**raffic **I**nciden**T** Dur**A**tion Predictio**N** (*TITAN*) model based on both sparse feature learning and multi-task learning framework. Our main contributions are:

• **Formulating a novel machine learning framework for traffic incident duration prediction using temporal features**. In contrast to existing works, we formulate the problem of traffic incident duration prediction for transportation systems as a multi-task supervised learning problem. In the proposed methods, models for different road segments are learned simultaneously by restricting all road segments to exploit a common set of features.

• **Modeling traffic speed similarity among road segments via spatial connectivity in feature space**. Based on the cascading nature of the traffic congestion in road networks, specifically designed constraints are proposed to model traffic speed similarities among data for spatiotemporally correlated road segments. These similarities in feature space are driven by spatial connectivity.

• **Proposing a sparse feature learning process to identify groups of temporal features at a higher level**. According to the nature of the traffic incidents, the traffic speed fluctuation in the early stages of the incidents is always important while estimating the impact and duration of the traffic incident. In the proposed model, constraints with sparsity and orthogonality are introduced to extract grouped important temporal features at a higher level.

• **Developing an efficient algorithm to train the proposed model**. The underlying optimization problem of the proposed multi-task model is a non-smooth, multi-convex, and inequality-constrained problem, which is challenging to solve. By introducing auxiliary variables, we develop an effective ADMM- based algorithm to decouple the main problem into several sub-problems which can be solved by block coordinate descent and proximal operators.

The rest of our paper is structured as follows. Related works are reviewed in Section 2. In Section 3, we describe the problem setup of our work. In Sections 4 and 5, we present a detailed discussion of our proposed *TITAN* model for predicting durations of traffic incidents, and its solution for parameter learning. In Section 6, extensive experiment evaluations and comparisons are presented. In the last section, we discuss our conclusion and directions for future work.

2 RELATED WORKS

In this section, we provide a detailed review of the current state of research for traffic incident analysis problem. There are several threads of related work of this paper: traffic incident impacts analysis, urban event forecasting, and spatiotemporal multi-task learning.

Traffic Impacts Analysis. The applications of conventional statistical methods have addressed its effectiveness in the traffic incident duration time prediction problems. The statistical methods fall into several branches: Bayesian classifier [7], discrete choice model [16], linear/non-parametric regression [23], hazard-based duration model [17]. In the recent decade, the Traffic Incident Management Systems (TIMS) have been deployed by traffic control centers in various cities and highways to alleviate the influence of traffic incidents on traffic conditions [19]. The historical traffic data obtained corresponds to traffic incidents play an important role in predicting the traffic incident durations. A new research

field based on data-driven algorithms and supported by real-world traffic data availability has recently emerged for traffic incident duration prediction with increasing research popularity. Various data mining and machine learning approaches have been employed to estimate and predict traffic incident duration time. Some of these approaches are the following: Lee et al. [14] proposed a genetic algorithm on traffic incident duration time prediction problems; Kim et al. and Zhan et al. [32] applied decision trees and classification tree models on the same problem and achieved improvements; Valenti et al. [29] proposed a support vector machine related method that utilizes the temporal features of the traffic data; artificial neural networks [30] is another highlighted direction for traffic incident duration prediction. In recent years, the research field of Intelligent Transportation Systems (ITS) have addressed its attention towards the hybrid methods [12] to predict traffic incident durations.

Urban Event Forecasting. To predict and detect the occurrence and impact the traffic incidents as urban events have received increasing attention in recent years. A large body of traditional work for event forecasting has focused on the early detection of events such as earthquakes [25], disease outbreaks [34], and transit service disruption [10], while event forecasting methods predict the incidence of such events in the future. Temporal events are the major focuses of the most existing event forecasting methods, with no interest in the geographical dimension, such as stock market movements [5] and elections [18]. A handful of works started to address the urban event prediction problem on a spatiotemporal resolution. For example, Zhao et al. [35] proposed a multi-task learning framework that models forecasting tasks in related geo-locations concurrently and; Gerber et al. [9] utilized a logistic regression model for spatiotemporal event forecasting, the urban event predictions with true spatiotemporal resolution. One limitation of these existing studies is that the temporal dimension is considered to be independent of the spatial dimension, and any interactions between the two are ignored. Our proposed *TITAN* model addresses the importance of the topology dimension, which is derived from the spatial dimension. We propose a multi-task learning framework with orthogonal constraints to model the interactions between the temporal and topological dimensions.

Spatiotemporal Multi-task Learning. Multi-task learning (MTL) refers to models that learn multiple related tasks simultaneously to improve overall performance. Recent decades have witnessed proposals for many MTL approaches [38]. Evgeniou et al. [8] proposed a regularized MTL formulation that constrains the models of each task to be close to each other. Task relatedness can also be modeled by constraining multiple tasks to share a common underlying structure (e.g., a common set of features) [3], or a common subspace [2]. Zhao et al. [35] designed a multi-task learning framework that models forecasting tasks in related geolocations. MTL approaches have been applied in many domains including computer vision and biomedical informatics. Our work, to the best of our knowledge, is the first paper to address the feasibility of combining multi-task learning and orthogonal regularization techniques to resolve traffic incident duration prediction and critical phases learning problems.

3 PROBLEM STATEMENT

Assume that we are given a collection of traffic incidents $\mathcal{I}$ from the traffic incident management system (TIM). For each traffic incident i in $\mathcal{I}$, we find the spatially correlated traffic sensor s, and its traffic speed reading at time interval τ: $\mathbf{v}_s(\tau)$, the granularity of the time interval is 1 minute. Given an incident record, and the traffic speed readings of its corresponding traffic speed sensor, the main objective of this paper is to predict the future impact of this given incident in terms of the temporal duration of this traffic incident.

Definition I: *Traffic speed in detection time and early verification time.* Suppose the verification time of the traffic incident is in time interval τ_v, we define and extract two important time periods **respond time** (*time between incident occurrence τ_o and incident verification time τ_v*) and **early verification time** (*a short period after the traffic incident verification time τ_v*) for feature construction. The traffic speeds for both time periods are extracted as: **(1) Traffic speed in detection time**: the previous h readings: $\mathbf{v}_s(\tau_v - 1), \mathbf{v}_s(\tau_v - 2), ..., \mathbf{v}_s(\tau_v - h)$ and **(2) Traffic speed in early verification time**: the succeeding t readings $\mathbf{v}_s(\tau_v), \mathbf{v}_s(\tau_v + 1), ..., \mathbf{v}_s(\tau_v + t)$.

Given the collection of traffic incidents, we first filter the collection with a selection Φ of arterial roads. This produces the targeted traffic incidents collection $\mathcal{I}^+$. Then based on which traffic incident takes place at the arterial road, $\mathcal{I}^+$ is grouped into $\{\mathcal{I}_r^+\}^{r\in\Phi}$, for example, $\Phi = \{\text{I-270, I-295, I-395, I-495, I-66, I-95}\}$.

We adopt a combination of traffic speed readings in *detection time* and *early verification time* $\mathcal{F} = \{\mathbf{v}_s(\tau_v - h), ..., \mathbf{v}_s(\tau_v + t)\}$ as the training features. For each traffic incident subcollection $\mathcal{I}_r^+$, we construct the training input $\mathbf{X}_r$ and the label $\mathbf{Y}_r$. The problem is then formulated as solving the mapping:

$$F_r(\mathbf{X}_r) \rightarrow \mathbf{Y}_r \tag{1}$$

where $\mathbf{X}_r \in \mathbb{R}^{n_r \times p}, p = h + t; \mathbf{Y}_r \in \mathbb{R}^{n_r}$. n_r is the number of traffic incident records for one arterial road; p represents the feature dimension of the training data, which is a combination of the detection time and the verification time; F_r is the learning model for inferring the traffic incident duration in the subcollection $\mathcal{I}_r^+$.

Consider that our problem is to predict the duration of the traffic incidents if there is a historical traffic speed reading for the corresponding collection of target traffic incidents $\mathcal{I}^+$, then it fits into the scope of the regression problem. For instance, learning the function F_r can be modeled as a regression problem with a least square loss function, and the model parameters $\mathbf{w}_r$ can be learned by solving the following optimization problem:

$$\underset{\mathbf{W}_r}{\operatorname{argmin}}\ \mathcal{L}_r = \|\mathbf{X}_r\mathbf{W}_r - \mathbf{Y}_r\|_2^2/n_r + \lambda_{\mathbf{W}}\|\mathbf{W}_r\|_1 \tag{2}$$

where $\lambda_{\mathbf{W}}$ controls the sparsity of the grouped features, n_r is the total number of data points in $\mathcal{I}^+$. Moreover, as inspired by the spatial correlations of traffic incidents introduced by the connectivity between road segments, the subproblem F_r defined in Section 3 to a regression problem under a multi- task learning framework. The proposed model should be encouraged to capture hidden patterns among road segments and to maintain sparsity in feature space. Mathematically, this consideration inspires us to use the $\ell_{2,1}$

norm [4] to perform joint feature selection:

$$\underset{\mathbf{W}}{\operatorname{argmin}} \sum_{r=1}^{|\Phi|} \|\mathbf{X}_r\mathbf{W}_r - \mathbf{Y}_r\|_2^2/n_r + \lambda_{\mathbf{W}}\|\mathbf{W}_r\|_{2,1} \tag{3}$$

where each column of $\mathbf{W}$, which represented by $\mathbf{W}_r$, denotes the model parameters for F_r. In this way, we can further model the relatedness among the road segments with parameter matrix $\mathbf{W}$. The overview of the *TITAN* model is represented in Figure 3. The following subsections address the details of the constraints on orthogonality and spatial connectivity.

4 MODEL

To identify the critical temporal features for traffic incident duration prediction, orthogonal constraints are applied to the *TITAN* model; to properly model the correlations between the traffic incidents based on the connectivity between the arterial roads, we apply a multi-task learning framework while designing the model.

4.1 Group Feature Learning

In the studies of Traffic Incident Management (TIM), one important task is to identify the key response time points and periods of traffic incidents. Assume that a two-vehicle collision occurs at 5:15 pm on the road segment of *Interstate 66*, based on the traffic speed readings from the traffic sensor, the transportation agencies want to learn how much impact the traffic incident will introduce to the local transportation system in terms of duration in time. The traffic speed readings of 5 minutes and 15 minutes after the traffic incident's occurrence play an important role in predicting the duration of the traffic incident.

Definition II: *Groups of key time points for a traffic incident.* The group assignment information is represented in a vector, and the ith group of time points is denoted by $\mathbf{q}_i \in \mathbb{R}^p$. If the jth time point feature belongs to this group, then the jth component of $\mathbf{q}_i$ is non-zero and the relative magnitude represents the 'importance' of the feature in this group. For training data $\mathbf{X}_r$ for one specific road segment, the new features generated by the group assignment is given by $\mathbf{X}_r\mathbf{q}_i$. Assume that there are k groups of features and the group structure is denoted by $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_k]$, and the generalized new features are given by $\mathbf{X}_r\mathbf{Q}$. To assign physical meaning to each generated group, the elements of $\mathbf{Q}$ have to be non-negative.

The new model vector for the grouped features is denoted by $\mathbf{w}_r \in \mathbb{R}^k$. The resulting formulation of the key feature group identification problem is then defined by:

$$\begin{aligned}\underset{\mathbf{Q},\mathbf{W}}{\operatorname{argmin}}\ \mathcal{L} = &\sum_{r=1}^{|\Phi|} \|\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r\|_2^2/n_r + \lambda_{\mathbf{W}}\|\mathbf{W}\|_{2,1} \\ &\text{s.t. } \mathbf{Q} \geq 0, \|\mathbf{q}_i\|_1 \leq \theta, i = 1, ..., k,\end{aligned} \tag{4}$$

where θ the parameter that controls the sparsity of each assigned group in $\mathbf{Q}$. The ℓ_1-norm in the constraint determines the length of the column in $\mathbf{Q}$ to be θ, which makes the group matrix $\mathbf{Q}$ easy to be interpreted.

By solving Equation 4, the model learns the group structure of the data features. However, the features may be largely overlapped because the proposed constraint does not consider any restrictions on feature overlapping. Such group overlapping is not ideal in our problem setting of traffic incident duration prediction problem. Because our selection of features is based on a time sequence of traffic speed readings, the consecutiveness of the features always provides a physical meaning.

In the research of traffic incident management, the lifetime of an incident generally consists of five different stages: incident detection, verification, response, clearance, and recovery. Because all stages do not overlap with each other, we impose the orthogonal constraints $\mathbf{q}_i^T\mathbf{q}_j = 0$ to control the overlapping conditions among the groups. The original nonnegative constraint $\mathbf{Q} \geq 0$ between all i,j is also applied. In terms of simplicity and interpretation, we normalize the group assignments and assume that the columns of $\mathbf{Q}$ are of length 1 for ℓ_2 norm. The constraint can further be expressed by $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. We use the ℓ_1 norm regularization to control the sparsity on $\mathbf{Q}$. The improved formulation of group feature learning can be given by:

$$\begin{aligned}\underset{\mathbf{Q},\mathbf{W}}{\operatorname{argmin}}\ \mathcal{L} = &\sum_{r=1}^{|\Phi|} \|\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r\|_2^2/n_r \\ &+ \lambda_{\mathbf{W}}\|\mathbf{W}\|_{2,1} + \lambda_{\mathbf{Q}}\|\mathbf{Q}\|_1 \\ &\text{s.t. } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}, \mathbf{Q} \geq 0\end{aligned} \tag{5}$$

4.2 Spatial Connectivity in Feature Space

In real-world transportation systems, different road segments are spatially related by intersections or interchanges. That is, two or more road segment may share similar traffic speed pattern during the traffic incidents. For instance, traffic congestion on *Interstate 495* could not only cause traffic pattern change at local road segments but also lead to traffic pattern change on other arterial roads that have close spatial correlations (e. g. *Interstate 66* and *US Route 7*). This spatial relatedness caused by network failure cascade [13, 26] results in similar traffic speed fluctuations; therefore, a similar pattern of traffic incident durations.

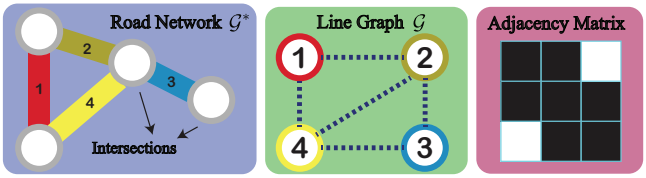


Figure 2: Road Segments Connectivity Shown by Adjacency Matrix. The left figure shows an example of the road network, the edges represent the road and the vertices represent the intersections; the middle figure shows the converted line graph of the road network, the vertices represent the roads; the right figure shows the adjacency matrix generated from the line graph.

Definition III: *Traffic incident spatial correlations.* With prior knowledge such as the road network connectivity, we assume that the traffic incidents are spatially correlated with each other. Given a road network $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$, where the vertices set $\mathcal{V}^*$ represents the union collection of the intersections and interchanges, and

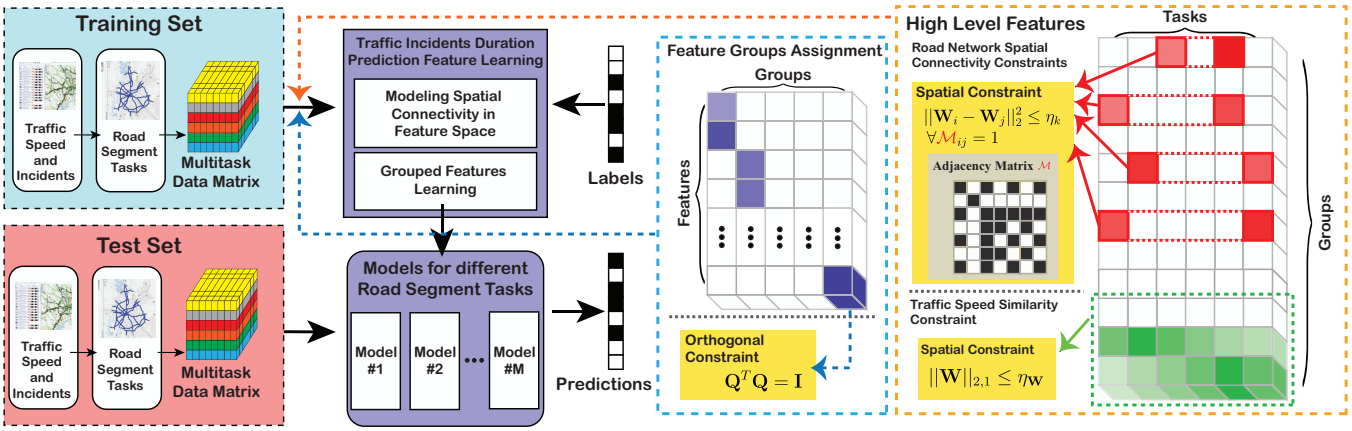


Figure 3: A Schematic View of the Traffic Incident Duration Prediction Model (TITAN). Similarities among temporal features are modeled by two major factors: spatial connectivity between arterial roads and the orthogonal constraint on Q. In particular, arterial roads connectivity constraints encourage the model to decrease differences between spatially related arterial roads in feature space. The orthogonal constraint encourages the model to identify groups of critical temporal features that are most influential to the prediction results.

the edges set $\mathcal{E}^*$ represents the collection of roadblocks. In order to model the connectivity of the road network, we transform the original road network graph $\mathcal{G}^*$ to its line graph $\mathcal{G} = \mathsf{L}(\mathcal{G}^*) = (\mathcal{V}, \mathcal{E})$, where the vertices set $\mathcal{V}$ represents the roads, and the edges set $\mathcal{E}$ represents the connectivity between roads. The adjacency matrix $\mathcal{M}$ of the line graph $\mathcal{G}$ reflects the overall connectivity of the roads. The roads connectivity and the line graph transformation is shown in Figure 2. Mathematically, we improve the model with constraints on parameters among different tasks:

$$\begin{aligned} \underset{\mathbf{Q},\mathbf{W}}{\operatorname{argmin}}\ \mathcal{L} &= \sum_{r=1}^{|\Phi|} \|\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r\|_2^2/n_r + \lambda_{\mathbf{W}}\|\mathbf{W}\|_{2,1} \\ \text{s.t. } &\|\mathbf{W}_i - \mathbf{W}_j\|_2^2 \leq \eta_k, \eta_k \geq 0, \forall \mathcal{M}_{ij} = 1 \end{aligned} \tag{6}$$

where each constraint with η_k forces the Euclidean distance between model parameters for a specific pair of road segments to be within a range. As defined in Section 3, $\mathcal{M}$ is the adjacency matrix that models the connectivity between road segments.

Combining the models represented by Equations 5 and 6, we obtain our proposed *TITAN* model. By moving the non-trivial constraints that are correlated to spatial connectivity into the objective function, we can obtain an equivalent regularized problem, which is easier to solve:

$$\begin{aligned} \underset{\mathbf{Q},\mathbf{W}}{\operatorname{argmin}}\ &\sum_{r=1}^{|\Phi|} \|\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r\|_2^2/n_r + \lambda_{\mathbf{W}}\|\mathbf{W}\|_{2,1} \\ &+ \lambda_{\mathbf{Q}}\|\mathbf{Q}\|_1 + \frac{1}{2}\sum_{ij} \mathcal{M}_{ij} \cdot \lambda_k \|\mathbf{W}_i - \mathbf{W}_j\|_2^2 \\ \text{s.t. } &\mathbf{Q}^T\mathbf{Q} = \mathbf{I}, \mathbf{Q} \geq 0 \end{aligned} \tag{7}$$

where λ_k is trade-off penalty balancing the value of the loss function and the regularizers. $\mathcal{M}$ is the adjacency matrix representing the road connectivity; $\mathcal{M}_{ij} \in \{0, 1\}$ denotes the connectivity information between the i-th road and the j-th road. Because the line graph $\mathcal{G}$ for road segments is undirected, the corresponding adjacency matrix $\mathcal{M}$ is a symmetric matrix. The coefficient $\frac{1}{2}$ is introduced to eliminate the repeatedly added lower triangular matrix.

5 PARAMETER LEARNING FOR *TITAN*

The objective function in Equation 7 is multi-convex and the regularizer $\ell_{2,1}$ is non-smooth. This increases the difficulty of solving this problem. A traditional way to solve this kind of problem is to use proximal gradient descent. But this approach is slow to converge. Recently, the alternating direction method of multipliers (ADMM) [6] has become popular as an efficient algorithm framework which decouples the original problem into smaller and easier to handle subproblems. Here we propose an ADMM-based Algorithm 1 which can optimize the proposed models efficiently. In particular, primal variables are updated on Line 4, dual variables on Line 5, and Lagrange multipliers on Line 6. Line 7 calculates both primal and dual residuals.

5.1 Augmented Lagrangian Scheme

First, we introduce an auxiliary variable $\mathbf{U}_{\mathbf{Q}} = \mathbf{Q}$ and $\mathbf{U}_{\mathbf{W}} = \mathbf{W}$ into the original problem 7 and obtain the following equivalent problem:

$$\begin{aligned} \underset{\Theta}{\operatorname{argmin}}\ &\sum_{r=1}^{|\Phi|} \|\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r\|_2^2/n_r + \lambda_{\mathbf{W}}\|\mathbf{U}_{\mathbf{W}}\|_{2,1} \\ &+ \lambda_{\mathbf{Q}}\|\mathbf{U}_{\mathbf{Q}}\|_1 + \frac{1}{2}\sum_{ij} \mathcal{M}_{ij} \cdot \lambda_k \|\mathbf{W}_i - \mathbf{W}_j\|_2^2 \\ \text{s.t. } &\mathbf{U}_{\mathbf{Q}} = \mathbf{Q}, \mathbf{U}_{\mathbf{W}} = \mathbf{W}, \mathbf{Q}^T\mathbf{Q} = \mathbf{I}, \mathbf{Q} \geq 0 \end{aligned} \tag{8}$$

where $\Theta = \{\mathbf{W}, \mathbf{Q}, \mathbf{U}_{\mathbf{W}}, \mathbf{U}_{\mathbf{Q}}\}$ is the set of variables to be optimized. Then we transform the above problem into its augmented

Algorithm 1: An ADMM-based solver for TITAN.

Input: $\mathbf{X}, \mathbf{Y}$
Output: $\mathbf{W}, \mathbf{Q}$
Initialize $\mathbf{W}^{(0)}$, $\mathbf{Q}^{(0)}$, $\mathbf{U}_\mathbf{W}^{(0)}$, $\mathbf{U}_\mathbf{Q}^{(0)}$, $\mathbf{\Lambda}_1^{(0)}$, $\mathbf{\Lambda}_2^{(0)}$, $\mathbf{\Lambda}_3^{(0)}$;
Initialize $\rho = 1$, $\epsilon^p > 0, \epsilon^d > 0$, MAX_ITER;
for $k = 1$: MAX_ITER **do**
  Update $\mathbf{W}^{(k)}$, $\mathbf{Q}^{(k)}$ with BCD using Equations 12 and 13;
  Update $\mathbf{U}_\mathbf{W}^{(k)}$ and $\mathbf{U}_\mathbf{Q}^{(k)}$ with Equations 16;
  Update $\mathbf{\Lambda}_1^{(k)}$, $\mathbf{\Lambda}_2^{(k)}$, and $\mathbf{\Lambda}_3^{(k)}$ with Equations 17;
  Compute p and d by Equations 18;
  if $p < \epsilon^p$ *and* $d < \epsilon^d$ **then**
    break;
  end
end

Lagrangian form as follows:

$$\begin{aligned}\underset{\Theta}{\operatorname{argmin}} & \sum_{r=1}^{|\Phi|} \|\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r\|_2^2/n_r + \lambda_\mathbf{W}\|\mathbf{U}_\mathbf{W}\|_{2,1} \\ & + \lambda_\mathbf{Q}\|\mathbf{U}_\mathbf{Q}\|_1 + \sum_{ij} \mathcal{M}_{ij}\cdot\lambda_k\|\mathbf{W}_i - \mathbf{W}_j\|_2^2 \\ & + \langle\mathbf{\Lambda}_1, \mathbf{W} - \mathbf{U}_\mathbf{W}\rangle + \langle\mathbf{\Lambda}_2, \mathbf{Q} - \mathbf{U}_\mathbf{Q}\rangle + \langle\mathbf{\Lambda}_3, \mathbf{I} - \mathbf{Q}^T\mathbf{Q}\rangle \\ & + \frac{\rho}{2}\|\mathbf{W} - \mathbf{U}_\mathbf{W}\|_F^2 + \frac{\rho}{2}\|\mathbf{Q} - \mathbf{U}_\mathbf{Q}\|_F^2 + \frac{\rho}{2}\|\mathbf{I} - \mathbf{Q}^T\mathbf{Q}\|_F^2\end{aligned} \tag{9}$$

where $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, and $\mathbf{\Lambda}_3$ are the Lagrangian multipliers. With this step, we decouple the original problem into two easier to handle problems in which seven variables $\mathbf{W}$, $\mathbf{Q}$, $\mathbf{U}_\mathbf{W}$, $\mathbf{U}_\mathbf{Q}$, $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, and $\mathbf{\Lambda}_3$ will be optimized individually. Note that the coefficient $\frac{1}{2}$ is omitted according to the optimization problem, and $\|\cdot\|_F$ is the Frobenius norm.

5.2 Parameter Optimization

The Lagrangian form in Equation 9 is separated based on the primal variables and the dual variables, where the problem of solving the primal variables $\mathbf{W}$ and $\mathbf{Q}$ is smooth and convex:

$$\begin{aligned}\underset{\mathbf{W},\mathbf{Q}}{\operatorname{argmin}} & \sum_{r=1}^{|\Phi|} \|\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r\|_2^2 + \sum_{ij} \mathcal{M}_{ij}\cdot\lambda_k\|\mathbf{W}_i - \mathbf{W}_j\|_2^2 \\ & + \langle\mathbf{\Lambda}_1, \mathbf{W} - \mathbf{U}_\mathbf{W}\rangle + \langle\mathbf{\Lambda}_2, \mathbf{Q} - \mathbf{U}_\mathbf{Q}\rangle + \langle\mathbf{\Lambda}_3, \mathbf{I} - \mathbf{Q}^T\mathbf{Q}\rangle \\ & + \frac{\rho}{2}\|\mathbf{W} - \mathbf{U}_\mathbf{W}\|_F^2 + \frac{\rho}{2}\|\mathbf{Q} - \mathbf{U}_\mathbf{Q}\|_F^2 + \frac{\rho}{2}\|\mathbf{I} - \mathbf{Q}^T\mathbf{Q}\|_F^2\end{aligned} \tag{10}$$

5.2.1 Update $\mathbf{W}$. We define Equation 10 as objective function $\mathcal{Q}$ which is multi-convex. In particular, $\mathcal{Q}$ of $\mathbf{W}_r$ is convex where all other $\mathbf{W}_{r'\neq r}$ are fixed. This kind of problem can be decoupled into subproblems using block coordinate descent (BCD) [31], in which each $\mathbf{W}_r$ is updated by solving the following sub-optimization problems:

$$\mathbf{W}_r \leftarrow \underset{\mathbf{W}_r}{\operatorname{argmin}}\ \mathcal{Q}. \tag{11}$$

$\mathcal{Q}$ is smooth and convex for each $\mathbf{W}_r$ and can be solved by gradient descent as follows:

$$\frac{\partial\mathcal{Q}}{\partial\mathbf{W}_i} = \mathcal{P}(i) + 2\sum_{ij}\mathcal{M}_{ij}\cdot\lambda_k(\mathbf{W}_i - \mathbf{W}_j) \tag{12}$$

where according to the BCD algorithm, the $\partial\mathcal{Q}_\mathbf{W}/\partial\mathbf{W}_i$ is calculated in sequence, from $i = 1$ to k. And the $\mathcal{P}(r)$ is defined as follows:

$$\mathcal{P}(r) = 2\mathbf{Q}^T\mathbf{X}_r^T(\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r) + \mathbf{\Lambda}_1^r + \rho(\mathbf{W}_r - \mathbf{U}_\mathbf{W}^r)$$

where $\mathbf{\Lambda}_1^r$ and $\mathbf{U}_\mathbf{W}^r$ are the r-th columns of the corresponding Lagrangian multiplier and dual variable.

5.2.2 Update $\mathbf{Q}$. Similarly, the objective function $\mathcal{Q}$ of $\mathbf{Q}$ is also smooth and convex. Because there are no constraints defined between the columns of $\mathbf{Q}$, the problem can be solved by gradient decent directly based on the objective function 10, and the gradient of $\mathbf{Q}$ is calculated by:

$$\begin{aligned}\frac{\partial\mathcal{Q}}{\partial\mathbf{Q}} =& 2\sum_{r=1}^{|\Phi|}\mathbf{X}_r^T(\mathbf{X}_r\mathbf{Q}\mathbf{W}_r - \mathbf{Y}_r)\mathbf{W}_r^T + \mathbf{\Lambda}_2 \\ & + 2\mathbf{Q}\mathbf{\Lambda}_3 + \rho(\mathbf{Q} - \mathbf{U}_\mathbf{Q}) + 2\rho\mathbf{Q}(\mathbf{Q}^T\mathbf{Q} - \mathbf{I})\end{aligned} \tag{13}$$

and the primal variable $\mathbf{Q}$ is then updated with a step size of α:

$$\mathbf{Q}^+ \leftarrow \mathbf{Q} - \alpha\cdot\frac{\partial\mathcal{Q}}{\partial\mathbf{Q}} \tag{14}$$

Now that the primal variable $\mathbf{W}$ is taken care of, the dual variable $\mathbf{U}_w$ is updated as follows:

$$\mathbf{U}_w^+ \leftarrow \underset{\mathbf{U}_w}{\operatorname{argmin}}\ \lambda_5\|\mathbf{U}_w\|_{2,1} + \frac{\rho}{2}\|\mathbf{U}_1 + \mathbf{W} - \mathbf{U}_w\|_2^2. \tag{15}$$

Note that this problem is the definition of proximal $\operatorname{prox}_{f_1,1/\rho}(\mathbf{U}_1 + \mathbf{W})$, where f_1 is the non-smooth function $\lambda_5\|\mathbf{U}_w\|_{2,1}$. The proximal operator can be solved efficiently using [21].

5.2.3 Update Dual Variables. Now that primal variables $\mathbf{Q}$ and $\mathbf{W}$ is taken care of, the dual variables $\mathbf{U}_\mathbf{Q}$ and $\mathbf{U}_\mathbf{W}$ are updated as follows:

$$\begin{aligned}\mathbf{U}_\mathbf{W}^+ &\leftarrow \operatorname{prox}_{f_1,1/\rho}(\mathbf{\Lambda}_1 + \mathbf{W}), \\ \mathbf{U}_\mathbf{Q}^+ &\leftarrow \operatorname{prox}_{f_2,1/\rho}(\mathbf{\Lambda}_2 + \mathbf{Q})\end{aligned} \tag{16}$$

where f_1 is the non-smooth function $\lambda_\mathbf{W}\|\mathbf{U}_\mathbf{W}\|_{2,1}$ and f_2 is the non-smooth function $\lambda_\mathbf{Q}\|\mathbf{U}_\mathbf{Q}\|_1$. The proximal operator can be solved efficiently using proximal operators [21].

Next, the Lagrangian multipliers $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$, and $\mathbf{\Lambda}_3$ are updated as follows:

$$\begin{aligned}\mathbf{\Lambda}_1^+ &\leftarrow \mathbf{\Lambda}_1 + \rho(\mathbf{W}^+ - \mathbf{U}_\mathbf{W}^+) \\ \mathbf{\Lambda}_2^+ &\leftarrow \mathbf{\Lambda}_2 + \rho(\mathbf{Q}^+ - \mathbf{U}_\mathbf{Q}^+) \\ \mathbf{\Lambda}_3^+ &\leftarrow \mathbf{\Lambda}_3 + \rho(\mathbf{Q}^{+T}\mathbf{Q}^+ - \mathbf{I})\end{aligned} \tag{17}$$

Finally, primal and dual residuals are calculated with:

$$\begin{aligned}p &= \|\mathbf{W}^+ - \mathbf{U}_\mathbf{W}^+\|_2 + \|\mathbf{Q}^+ - \mathbf{U}_\mathbf{Q}^+\|_2 + \|\mathbf{Q}^{+T}\mathbf{Q}^+ - \mathbf{I}\|_2 \\ d &= \rho\left(\|\mathbf{U}_\mathbf{W}^+ - \mathbf{U}_\mathbf{W}\|_2 + \|\mathbf{U}_\mathbf{Q}^+ - \mathbf{U}_\mathbf{Q}\|_2\right).\end{aligned} \tag{18}$$

where p is primal residual, and d is dual residual.

6 EXPERIMENT

In this section, we present the experiment environment, dataset introduction, evaluation metrics and comparison methods, extensive experimental analysis on predictive results, and discussions on the learner features.

6.1 Experiment Setup

6.1.1 Experiment Environment. We conducted our experiments on a machine with Intel Core i7-4790 3.6 GHz, the computational power of this CPU is 4.13 Gflops per core. For real-world traffic incident analysis problems, time requirements should be an important factor. The most time-consuming process of our proposed *TITAN* model is at the training stage. The training stage learns the parameters for temporal features $\mathbf{W}$ and the orthogonal groups of the temporal features $\mathbf{Q}$. A matrix multiplication $\mathbf{XQW}$ will generate the prediction rapidly. In the validation and testing stages, our prediction for a single data point is generated in less than 0.003 seconds.

6.1.2 Dataset and Feature Settings. We evaluate our proposed Traffic Incident Duration Prediction model using two real-world traffic data sources. **1) Traffic incident records with reported duration**. We collect 43,923 records of traffic incidents in the year 2018 from three major transportation agencies in the Washington DC Metropolitan area: Washington DC, Virginia State and Maryland State departments of transportation. From the collected traffic incident records, we select 29,075 traffic incidents that take places on the six major arterial roads in the region: I-270, I-295, I-395, I-495, I-66, and I-95. In the selected data frame, the time duration of the traffic incidents are recorded in minutes, and we utilize the duration as the ground truth. From the selected incidents 80% of the records serve as the training set, while the rest serve as the testing set. **2) INRIX traffic speed data**. We leverage the traffic speed readings from the traffic sensors as the training features. Given the location and verification time of the traffic incidents, we collect traffic speed readings of nearby traffic sensors.

The connectivity of the road network determines the number of tunable parameters in our *TITAN* model. According to the selected arterial roads in our experiment, seven hyperparameters can be tuned. During the experiment, we observe that the value of the loss function is significantly larger than regularizers, which means a large penalty should be used to balance the loss function and the regularizers.

6.2 Comparison Methods

To evaluate the performance of the traffic incident duration prediction, 5 comparison methods are considered in our experiment: ℓ_2 regulized linear regression (ridge regression), ℓ_1 regulized linear regression (LASSO), support vector regression (SVR), Naïve multi-task learning model (nMTL), and feature refiner method (FeaFiner).

• ℓ_2 **Regulized Linear Regression (Ridge)** [23]. Ridge regression is an extension for linear regression. It's a linear regression model regulized on ℓ_2 norm. The λ parameter is a scalar that controls the model complexity; the smaller λ is, the more complex the model will be. In our implementation, λ is searched from $\{10, 100\}$. This model only considers the temporal features on duration prediction. No multi-task for arterial road connectivity and grouped temporal features are considered.

• ℓ_1 **Regulized Linear Regression (LASSO)** [24, 27]. This is a classic way to conduct cost-efficient regressions by enforcing the sparsity of the selected features. It has been proved to be effective in the field of event detection [24]. It includes a parameter λ that trades off the regularization term; typically, the larger this parameter is, the fewer the selected features will be. In our experiment, λ is searched from $\{1, 10, 100\}$. The feature configurations applied by this model is the same as the ridge regression model.

• **Support Vector Regression (SVR)** [27]. Support vector regression provides solutions for both linear and non-linear problems. In our experiment implementation, we utilize non-linear kernel functions (RBF kernel) to find the optimal solution for incident duration prediction problem. The model parameters are selected with $c = 1$ and $\epsilon = 0.1$. This model considers similar temporal features with ridge regression and LASSO methods, no multi-task features for connectivity is considered.

• **Naïve Multi-task Learning Model (nMTL)** [37]. We implement the fundamental settings of the naive multi-task learning model for event detection. This comparison method is regularized with $\ell_2 1$ constraint between tasks. The training tasks of this model are split by the arterial roads. The correlations between tasks are intuitively constrained by ℓ_2 norm, and within each task, the importance of the features are constrained by ℓ_1 norm. The penalty parameter λ is searched from $\{1, 10, 100\}$.

• **FeaFiner** [40]. FeaFiner regression model with a capability of learning feature clusters. This method learns an optimal sparse feature grouping for general regression problems. However, there are no multi-task properties supported. In our implementation of this method, we apply this method on the complete set of traffic incidents, and the target feature is selected to be the temporal features. In the parameter initialization, we select the parameter $k = 30$ for the k-Mean clustering.

6.3 Evaluation Metrics

To quantify and validate model performance on traffic incident duration prediction, we adopt root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). These metrics are widely utilized in the field of traffic duration prediction studies [11, 15, 22, 41], it reflects the predictive performance of the proposed model. Equations 19, 20, and 21 represent the calculations of the selected evaluation metrics:

$$RMSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2} \tag{19}$$

$$MAE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i}^{N}|y_i - \widehat{y}_i| \tag{20}$$

$$MAPE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i}^{N}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{21}$$

where N is the total number of records; $\boldsymbol{y}$ is the predicted traffic incident durations represented in vector; $\hat{\boldsymbol{y}}$ is the ground truth value of the corresponding record, which is also represented in vector. y_i

Table 1: Traffic Incident Duration Prediction Comparisons (RMSE (Min), MAE (Min), MAPE (%))

Method	I-270			I-295			I-395		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Ridge	92.4709	76.4666	96.3826	89.1404	69.1273	87.3530	84.6881	65.5869	83.3106
LASSO	90.8535	73.8732	90.3336	76.4372	58.8515	70.1599	72.4028	55.8695	68.8993
SVR	87.8016	72.9036	88.7639	72.4579	53.9583	68.6843	68.4456	50.0854	62.6849
nMTL	**70.7942**	59.9754	82.8141	55.4657	42.6052	55.3893	57.2953	43.3107	**41.2034**
FeaFiner	77.0080	**57.5550**	81.4397	63.3036	50.1060	62.6381	51.6727	40.8695	47.4805
TITAN	73.1291	59.5265	**81.3789**	**46.0873**	**34.3043**	**52.9296**	**46.2329**	**38.9277**	42.3794
Method	I-495			I-66			I-95		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Ridge	69.9718	52.2384	81.2393	80.4118	62.5392	85.3443	76.0088	64.6172	80.1281
LASSO	60.0119	48.5583	75.6027	68.0900	60.7429	77.9394	84.5617	58.7706	69.6493
SVR	58.9676	46.7641	71.5021	72.7470	59.0808	71.1609	62.8689	54.7717	68.8999
nMTL	52.5722	40.5422	63.6820	60.6244	48.4900	58.4887	57.1166	45.1327	**49.4991**
FeaFiner	56.3049	44.0023	44.9048	62.5098	50.4090	56.4438	55.6806	46.0073	56.0013
TITAN	**47.7131**	**31.7725**	**37.1649**	**53.7001**	**44.3786**	**40.9370**	**52.6403**	**40.5345**	49.9848

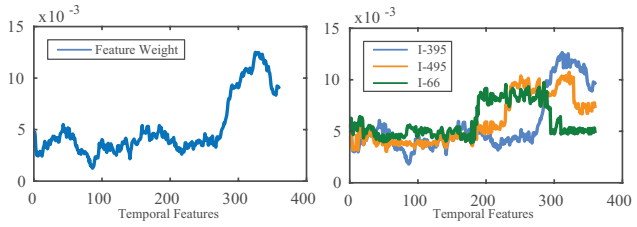
and $\widehat{y}_i$ are the i^{th} predicted result and the i^{th} ground truth value respectively.

6.4 Incident Duration Prediction Analysis

6.4.1 TITAN Performance Analysis on Spatial Connectivity. Table 1 summarizes the comparisons of our proposed method to the competing methods for the task of traffic incident duration prediction. From the experimental results, we can justify our application of a multi-task learning framework for predicting the incident duration. In general, *TITAN* outperforms the single task models (LR, SVR, and FeaFiner) on RMSR, MAE, and MAPE. This result shows that the spatial correlations between the road segments can improve the performance of the traffic incident duration prediction. The *TITAN* model outperforms the nMTL on RMSE and MAE. These results demonstrated that for the traffic incident duration prediction problem, only ℓ_1 regularizers is insufficient, detailed spatial connectivity between the road segments should also be considered.

6.4.2 TITAN Performance Analysis on Feature Groups Learning. TITAN Performance Analysis on Feature Groups Learning. Among the comparison methods, the FeaFiner [40] method considers the orthogonal constraint that is capable of grouping low-level features into a high-level feature representation. The original FeaFiner applies the Ridge and LASSO as the original problem settings. Thus, the results presented in Table 1 can be categorized by whether the orthogonal constraints are considered or not. The methods consider orthogonal constraints are FeaFiner and *TITAN*; the methods do not consider the orthogonal constraint are Ridge, LASSO, and SVR. By comparing these two categories, we learn that the overall performance of the methods consider the orthogonal constraint is better than the methods do not consider the orthogonal constraint. However, the overall performance increase is not as significant as the performance increase from the spatial connectivity constraint introduced by the framework of multi-task learning.

6.4.3 Performance Analysis between Training Tasks. The results in Table 1 show that the model performance for traffic incident duration prediction is not the same across different road segments. For instance, the prediction performances of all the comparison methods on highway I-270 only have slight differences between each other. This is because the highway I-270 only has one spatial connectivity to the rest of the road segments, and the constraint of Euclidean distance for I-270 only shares a limited connection between the other columns of the feature matrix $\mathbf{W}$. In contrast, our model for the highway I-495 outperforms the comparison methods, because the subtask for I-495 shares feature similarity with all other subtasks.



(a) TITAN w/o Orth. Const. (b) TITAN with Orth. Const.

Figure 4: Feature Learning Results on Q

6.5 Feature Groups Assignment Analysis

The orthogonal constraint ensures the proposed model to learn a group of highlighted features that play an important role in predicting the traffic incident durations. In our experiment, we also

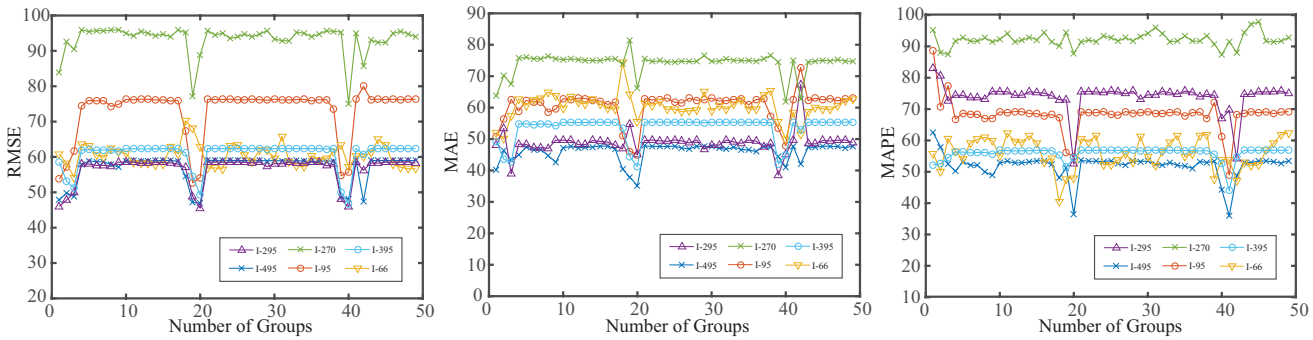


Figure 5: Illustration of how the number of grouped temporal features will affect the performance of the TITAN model. The performance is evaluated in terms of RMSE, MAE, and MAPE respectively.

study the results of the learned group features empirically. In the experiment, we set the number of groups to be 10, and we also apply two conditions: 1) *TITAN* with orthogonal constraint and 2) *TITAN* without orthogonal constraint. Figure 4 shows the learned feature groups assignments for both experimental conditions. We can find the learned features with orthogonal constraint overlap less than the learned feature assignment without orthogonal constraint.

While experimenting without the orthogonal constraint, we found that the model has a preference for grouping the low-level features into one feature assignment for every group $\mathbf{q}_i$. Figure 4(a) shows the single feature group assignment for the model without orthogonal constraint. From Figure 4(a), we can find that for the model without orthogonal constraint, temporal features with higher indexes are assigned with higher weights (>300). This result is reasonable because this can be interpreted as the duration of the traffic incident can be better inferred with the most recent traffic speed readings.

To compare with the model with orthogonal constraint, Figure 4(b) shows the learned feature group assignment for several subtasks. We can find the most weighted feature group by checking the weights in the learned variable $\mathbf{W}$. For example, in Figure 4(b), we demonstrate top weighted group for three subtasks (I-495, I-66, and I-395). From Figure 4(b), we find that the top assigned feature group for different arterial roads differ from each other slightly. This result shows that the most critical temporal features for predicting the traffic incident duration for different roads differ. This observation is valuable for the transportation operators and first responders. In Figure 4(b), we can observe that the high-level features of the subtask I-495 have a shift comparing to the subtask of I-395. The 10 minutes' shift indicates that to predict the duration of an incident on I-495, the traffic speed readings of 10 minutes in advance have higher importance.

6.6 Case Studies

During the experiments, several interesting facts revealed by using the proposed approach were discovered. Here we discuss the details towards the identification of the critical phases for traffic incidents and the influences of the connectivity between the arterial roads.

6.6.1 Critical Phases Identification for Traffic Incidents. According to the experiment results on the correlations between the number of groups and the performance of the *TITAN* mode, we discover the optimal number of groups for the temporal features. The physical meaning of the number of groups in this experiment, corresponding to the number of phases will be identified for the traffic incidents. As mentioned in Section I, the life cycle of the traffic incident is conventionally grouped into five phases: detection, verification, response, clearance, and recovery. Although such grouping strategy is efficient in the perspective of transportation management and operations, it cannot provide useful temporal feature grouping to predict the traffic incident durations. From this experiment, we can study how the performance of the *TITAN* model will be affected with respect to the number of feature groups. As shown in Figure 5, we illustrate the RMSE, MAE, and MAPE obtained by varying the number of the groups from 1 to 50; and the color-coded lines representing different arterial roads in the experiment. From Figure 5, we learn that for most of the arterial roads, the *TITAN* model reaches the best performance when the number of groups in the ranges of 18-20 and 40-43. This experiment result indicates that the conventional five-phase definition of traffic incident life cycle may not provide informative input to the traffic incident duration prediction problems.

6.6.2 Influences of Arterial Road Connectivity. The performance differences between the arterial roads can be observed in Figure 5. In Figure 5, the general prediction performance of the arterial road Interstate 495 outperforms the rest of the arterial roads, and the arterial road Interstate 270 has the worst duration prediction results overall. This comparison result reveals that the connectivity between different arterial roads plays an important role while predicting the traffic incident duration. Because the more connection with other arterial roads means the more information shared with other train tasks in the training stage. The Interstate 495 intersections with all other arterial roads, while the Interstate 270 only intersects with the Interstate 495.

7 CONCLUSION

This paper presents a novel traffic incident duration prediction and feature learning model *TITAN*. The proposed model is designed

based on the multi-task learning framework for prediction, and a sparse feature learning framework for higher feature groups identification. The proposed *TITAN* model outperforms the existing traffic incident duration prediction models because of two major advantages in model design: 1) consideration of the connectivity between road segments within the urban road networks; 2) the learned higher level features provide a better predictive pattern for the problem of traffic incident duration prediction. Extensive experiments on real-world datasets with comparisons of the baseline methods justify the performance of *TITAN* model. By applying the orthogonal constraint, the proposed model is capable of identifying groups of higher level features which can be further considered as the critical evolution stages of the traffic incident. Such functionality provided by our proposed model is helpful for the transportation operators and first responders while judging the influences of the traffic incidents.

REFERENCES

[1] Martin W Adler, Jos van Ommeren, and Piet Rietveld. 2013. Road congestion and incident duration. *Economics of transportation* 2, 4 (2013), 109–118.

[2] Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* 6, Nov (2005), 1817–1853.

[3] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2007. Multi-task feature learning. In *Advances in neural information processing systems*. 41–48.

[4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2008. Convex multi-task feature learning. *Machine Learning* 73, 3 (2008), 243–272.

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3, 1 (2011), 1–122.

[7] Stephen Boyles, David Fajardo, and S Travis Waller. 2007. A naive Bayesian classifier for incident duration prediction. In *86th Annual Meeting of the Transportation Research Board, Washington, DC*. Citeseer.

[8] Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 109–117.

[9] Matthew S Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.

[10] Taoran Ji, Kaiqun Fu, Nathan Self, Chang-Tien Lu, and Naren Ramakrishnan. 2018. Multi-task Learning for Transit Service Disruption Detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 634–641.

[11] Asad J Khattak, Jun Liu, Behram Wali, Xiaobing Li, and ManWo Ng. 2016. Modeling traffic incident duration using quantile regression. *Transportation Research Record* 2554, 1 (2016), 139–148.

[12] Woon Kim and Gang-Len Chang. 2012. Development of a hybrid prediction model for freeway incident duration: a case study in Maryland. *International journal of intelligent transportation systems research* 10, 1 (2012), 22–33.

[13] Agus Trisnajaya Kwee, Meng-Fen Chiang, Philips Kokoh Prasetyo, and Ee-Peng Lim. 2018. Traffic-Cascade: Mining and Visualizing Lifecycles of Traffic Congestion Events Using Public Bus Trajectories. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 1955–1958.

[14] Ying Lee and Chien-Hung Wei. 2010. A computerized feature selection method using genetic algorithms to forecast freeway accident duration times. *Computer-Aided Civil and Infrastructure Engineering* 25, 2 (2010), 132–148.

[15] Ruimin Li, Francisco C Pereira, and Moshe E Ben-Akiva. 2018. Overview of traffic incident duration analysis and prediction. *European Transport Research Review* 10, 2 (2018), 22.

[16] Pei-Wei Lin, Nan Zou, and Gang-Len Chang. 2004. Integration of a discrete choice model and a rule-based system for estimation of incident duration: a case study in Maryland. In *CD-ROM of Proceedings of the 83rd TRB Annual Meeting, Washington, DC*.

[17] Doohee Nam and Fred Mannering. 2000. An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice* 34, 2 (2000), 85–102.

[18] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*.

[19] Nicholas Owens, April Armstrong, Paul Sullivan, Carol Mitchell, Diane Newton, Rebecca Brewster, and Todd Trego. 2010. *Traffic incident management handbook*. Technical Report.

[20] Kaan Ozbay and Pushkin Kachroo. 1999. Incident management in intelligent transportation systems. (1999).

[21] Neal Parikh, Stephen Boyd, et al. 2014. Proximal algorithms. *Foundations and Trends® in Optimization* 1, 3 (2014), 127–239.

[22] Hyoshin Park, Ali Haghani, and Xin Zhang. 2016. Interpretation of Bayesian neural networks for predicting the duration of detected incidents. *Journal of Intelligent Transportation Systems* 20, 4 (2016), 385–400.

[23] Srinivas Peeta, Jorge L Ramos, and Shyam Gedela. 2000. Providing real-time traffic advisory and route guidance to manage Borman incidents on-line using the hoosier helper program. (2000).

[24] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, et al. 2014. 'Beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1799–1808.

[25] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.

[26] Zhen Su, Lixiang Li, Haipeng Peng, Jürgen Kurths, Jinghua Xiao, and Yixian Yang. 2014. Robustness of interrelated traffic networks to cascading failures. *Scientific reports* 4 (2014), 5413.

[27] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[28] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 1 (2005), 91–108.

[29] Gaetano Valenti, Maria Lelli, and Domenico Cucina. 2010. A comparative study of models for the incident duration prediction. *European Transport Research Review* 2, 2 (2010), 103–111.

[30] Eleni I Vlahogianni and Matthew G Karlaftis. 2013. Fuzzy-entropy neural network freeway incident duration modeling with single and competing uncertainties. *Computer-Aided Civil and Infrastructure Engineering* 28, 6 (2013), 420–433.

[31] Yangyang Xu and Wotao Yin. 2013. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences* 6, 3 (2013), 1758–1789.

[32] Chengjun Zhan, Albert Gan, and Mohammed Hadi. 2011. Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. *IEEE Transactions on Intelligent Transportation Systems* 12, 4 (2011), 1549–1557.

[33] Xuchao Zhang, Liang Zhao, Arnold P Boedihardjo, Chang-Tien Lu, and Naren Ramakrishnan. 2017. Spatiotemporal event forecasting from incomplete hyper-local price data. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 507–516.

[34] Liang Zhao, Jiangzhuo Chen, Feng Chen, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2015. Simnest: Social media nested epidemic simulation via online semi-supervised deep learning. In *2015 IEEE International Conference on Data Mining*. IEEE, 639–648.

[35] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1503–1512.

[36] Liang Zhao, Junxiang Wang, and Xiaojie Guo. 2018. Distant-supervision of heterogeneous multitask learning for social event forecasting with multilingual indicators. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[37] Liang Zhao, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2016. Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2085–2094.

[38] Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Malsar: Multi-task learning via structural regularization. *Arizona State University* 21 (2011).

[39] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. 2013. Modeling disease progression via multi-task learning. *NeuroImage* 78 (2013), 233–248.

[40] Jiayu Zhou, Zhaosong Lu, Jimeng Sun, Lei Yuan, Fei Wang, and Jieping Ye. 2013. Feafiner: biomarker identification from medical data through feature generalization and selection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1034–1042.

[41] Yajie Zou, Kristian Henrickson, Dominique Lord, Yinhai Wang, and Kun Xu. 2016. Application of finite mixture models for analysing freeway incident clearance time. *Transportmetrica A: Transport Science* 12, 2 (2016), 99–115.