Frontiers in Traffic and Transportation Engineering (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ftte.2024.040109
ISSN 2616-2334 Vol. 4 Num. 1

Research on traffic congestion prediction based on lasso and ridge regression

Chengze Shang

La Salle College, Hong Kong, China

Keywords: Traffic Congestion, Lasso Regression, Ridge Regression

Abstract: In order to optimize the effective utilization of road resources, reduce economic losses, and enhance the efficiency of traffic management, accurate prediction of road congestion is of paramount importance. This study focuses on the problem of traffic congestion prediction and thoroughly analyzes traffic congestion data from both temporal and spatial dimensions. In the spatial dimension, violin plots are employed to analyze the spatial variations of traffic congestion, revealing significant differences in congestion levels among different road coordinates. Subsequently, this paper selects data from the 8 AM time frame and utilizes a scatter plot matrix to explore the correlations between traffic congestion values at different coordinate points. In the temporal dimension, noticeable differences in traffic patterns between weekdays and non-working days are observed. Weekdays exhibit two distinct traffic peaks, whereas non-working days have a single, longer-lasting peak. For the traffic congestion prediction, two algorithms, Lasso and Ridge regression, are employed, and the existing data is subjected to predictive analysis. To further enhance the performance of the models, this paper employs grid search to identify the optimal hyperparameters for the models. The research findings demonstrate that both models yield highly accurate predictions, with minimal differences between them. Specifically, the Mean Squared Error (MSE) is 131.563, the Root Mean Squared Error (RMSE) is 11.470, and the Mean Absolute Error (MAE) is 8.044. These evaluation metrics validate the effectiveness and reliability of this research in the field of traffic congestion prediction, providing robust data support for future traffic management endeavours.

1. Introduction

Nowadays, the increasingly severe traffic congestion has a significant impact on people's work and daily lives. Improving transportation issues and accelerating urban modernization have become urgent tasks. Intelligent transportation systems have garnered significant attention as a crucial means of addressing transportation problems. Among them, the construction of traffic indicators and the description and prediction of traffic flow are of paramount importance. Accurate measurement and description of traffic indicators contribute to a deeper understanding of traffic conditions and provide a foundation for traffic management and planning. Meanwhile, traffic flow prediction helps anticipate future conditions and implement strategies to alleviate congestion and optimize traffic flow [1].

Niukai et al. [2] proposed a traffic congestion level prediction model based on a backpropagation neural network. The study employed partial correlation analysis to identify the weight relationships

between different variables and subsequently established a neural network model to predict congestion levels. Tianyu et al. [3], on the other hand, used KNN-VA and KNN-RBF models as velocity prediction benchmarks and combined the two models through boosting fusion to obtain more accurate predictions.

In this study, a comprehensive analysis of traffic congestion is conducted from both temporal and spatial perspectives. In the spatial dimension, the study examines the variations and correlations of traffic congestion values. In the temporal dimension, the periodicity of time and the differences between weekdays and non-working days are analyzed. Finally, the lasso and ridge regression algorithms are employed to predict traffic congestion.

2. Algorithmic principle

2.1 Ridge Regression

Ridge Regression is a regression analysis method used to handle collinear data and analyze high-dimensional data. Its main advantage is the ability to address multicollinearity issues, providing stable estimation results that are less susceptible to the influence of outliers. Additionally, Ridge Regression can handle high-dimensional datasets by introducing a regularization term to constrain the model's complexity, thus avoiding overfitting [4].

For Ordinary Least Square (OLS):

$$\beta^{OLS} = argmin\{\frac{1}{2n}(y - X\beta)^2\} \tag{1}$$

In Equation 1, y represents the target variable, X denotes the data matrix, β represents the model parameters, and n represents the sample size.

OLS estimates the coefficients of the model by minimizing the squared error between actual values and predicted values. However, OLS does not incorporate regularization and is susceptible to overfitting. Additionally, when the number of variables exceeds the number of data points, it becomes impossible to solve the above equation.

Ridge Regression is an improvement upon Ordinary Least Squares (OLS) by introducing an L2 norm penalty term in the loss function.

$$\beta^{ridge} = argmin\{\frac{1}{2n}(y - X\beta)^2 + \lambda \parallel \beta \parallel 2\} \tag{2}$$

In Equation 2, y represents the target variable, X denotes the data matrix, β represents the model parameters, and λ represents the L2 regularization parameter.

2.2 Lasso Regression

Lasso Regression utilizes L1 regularization to constrain the complexity of the model and has the ability to automatically perform feature selection during the fitting process. By shrinking the weights of features that have a lesser impact on predicting the target variable, Lasso automatically selects the most relevant features. Lasso can provide a sparse model, making it particularly suitable for datasets with a large number of features [5].

Lasso Regression incorporates an L1 norm penalty term into Ordinary Least Squares OLS

$$\beta^{lasso} = argmin\{\frac{1}{2n}(y - X\beta)^2 + \lambda \parallel \beta \parallel 1\} \tag{3}$$

In Equation 3, y represents the target variable, X denotes the data matrix, β represents the model parameters, and λ represents the L1 regularization parameter.

Lasso achieves variable selection by compressing the coefficients of variables and setting some

regression coefficients to zero. The application scenarios of Lasso regression include datasets with a large number of features but only a few features that have a significant impact on the target variable, as well as cases where a more concise model with only relevant features is desired. Its advantages include automatic feature selection and generating sparse models. However, it has drawbacks such as instability in handling highly correlated features and potential difficulties in handling datasets with a large number of correlated features. When feature selection and obtaining a concise model are required, Lasso regression can be chosen. When dealing with multicollinearity and retaining all features, Ridge regression can be chosen. In practical applications, appropriate regularization parameters can be selected to optimize model performance using methods such as cross-validation. Additionally, for solving Lasso regression, the LARS (Least Angle Regression) algorithm can be used. The LARS algorithm determines the Lasso path, which is a series of Lasso solutions at different regularization parameters, by gradually adding variables. Its advantage lies in efficiently computing the entire path without requiring multiple iterations like traditional coordinate descent algorithms.

3. Model Construction

3.1 Source of Data

The data for this study was obtained from 12-hour traffic flow measurements in Chicago for the months of April to September in a certain year. The traffic congestion at each monitoring point was measured every 20 minutes. The dataset includes the coordinates (x and y) of the monitoring points. Specifically, there are 3 possible values for x [0, 1, 2] and 4 possible values for y [0, 1, 2, 3, 4]. There are 8 road directions: EB (eastbound), NB (northbound), WB (westbound), SB (southbound), NW (northwest), SE (southeast), SW (southwest), and NE (northeast). Additionally, the dataset includes the congestion level. A portion of the data is presented in Table 1:

Table 1: Traffic Congestion Dataset

time	x-coordinates	y-coordinates	direction	congestion
04-01 08:00:00	0	0	EB	70
04-01 08:20:00	0	1	EB	18
04-01 08:40:00	0	1	NB	60
……	……	……	……	……
05-29 08:00:00	0	0	NB	43
05-30 08:00:00	0	0	EB	42
……	……	……	……	……

Based on the combination of x and y coordinates along with the direction, there are a total of 65 drivable roads.

To gain insights into the distribution pattern of the data, a statistical analysis was conducted on the congestion levels (ranging from 1 to 100) for the entire dataset, as illustrated in Figure 1.

The histogram presented in Figure 1 showcases the distribution characteristics of the traffic congestion data and provides insights into its normality. The histogram exhibits a typical bell-shaped curve, indicating that the dataset adheres to the assumption of normal distribution. Moreover, the congestion levels of the traffic data consistently fall within the range of 0 to 100, and no outliers were observed.

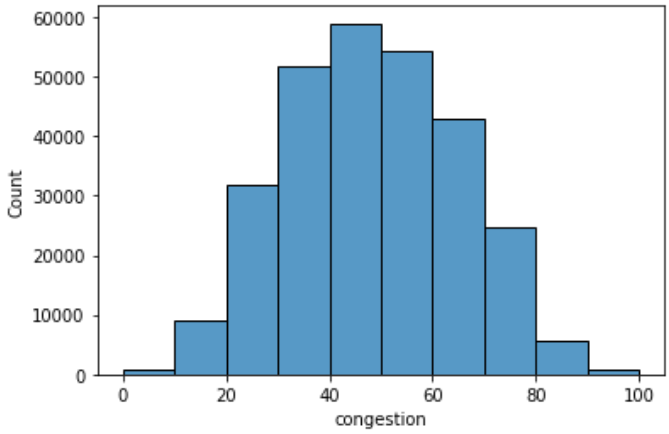


Figure 1: Histogram of normality test for traffic congestion data

3.2 Analysis of Spatial Variability in Traffic Congestion Levels

Violin plots are a type of chart used to visualize data distributions. They provide a comprehensive view of the distribution characteristics and statistical summaries of multiple datasets. This type of plot combines features of box plots and density plots, offering enhanced insights into the data. In this study, based on the data collected on April 10th, we generated violin plots for two roads with coordinates (x=0, y=0) and (x=2, y=2). These plots, depicted in Figure 2 and Figure 3 respectively, aimed to explore the impact of different road directions on congestion levels. In Figure 2, for the coordinates (x=0, y=0), congestion levels for different directions are primarily concentrated between 40 and 50. The distribution appears relatively concentrated and falls within the medium to lower range. On the other hand, Figure 3 represents the congestion levels for the coordinates (x=2, y=2), where the values are predominantly concentrated around 60. This distribution exhibits a more concentrated pattern and encompasses both high and low congestion values. From both plots, it is evident that congestion levels vary across different directions.

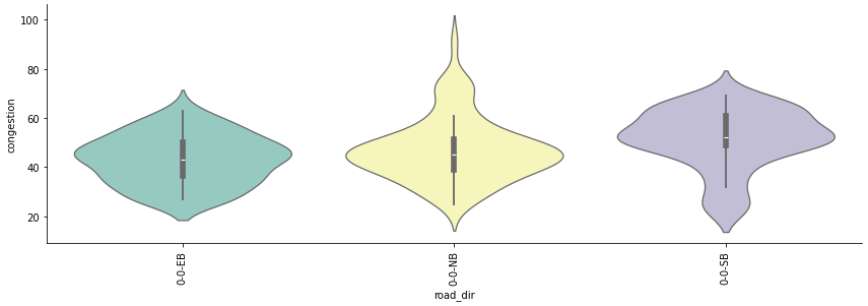


Figure 2: Distribution of road congestion values in different directions at x=0y=0 coordinates

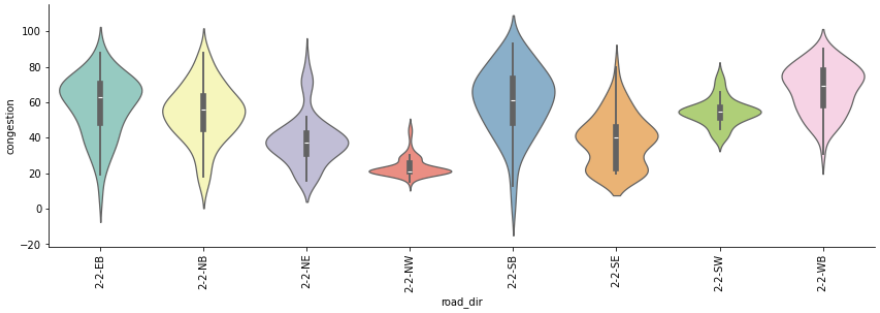


Figure 3: Distribution of road congestion values in different directions under x=2y=2 coordinates

3.3 Spatial correlation analysis of traffic congestion values

To investigate spatial correlation, this study employed a scatterplot matrix to analyze the frequency of different congestion levels at different detection points at 8:00 AM on a specific day. In this matrix, the X-axis represents different traffic congestion values, while the Y-axis represents the frequency of corresponding traffic flow values. The matrix includes various coordinate positions (e.g., x=0y=0, x=0y=1, etc.). Based on the results depicted in Figure 4, most combinations of variables exhibit irregular patterns, with only a few combinations (e.g., x2y0 and x1y0) displaying a certain linear increasing relationship. In Figure 5, most combinations of variables demonstrate a significant linear increasing relationship. Additionally, most scatterplots exhibit similar increasing trends, meaning that when one coordinate combination shows a high frequency of congestion values, a similar increasing relationship is observed in another combination (e.g., x=1y=2, x=1y=3, and x=1y=3, x=1y=2). This indicates the presence of positive spatial correlation among congestion values in different spatial locations. Therefore, considering spatial factors is essential when constructing models.

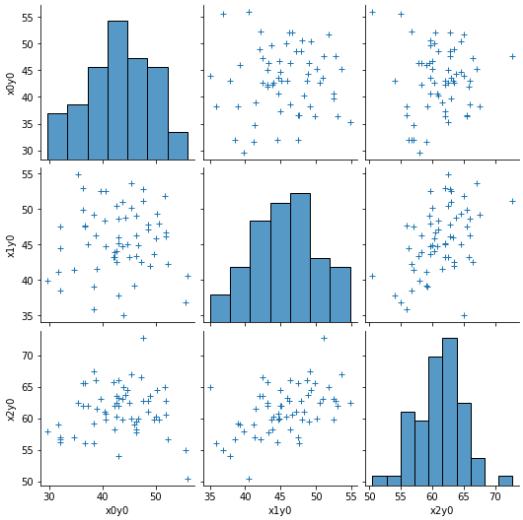


Figure 4: Scatterplot matrix with different coordinates (x=[1, 2], y=0)

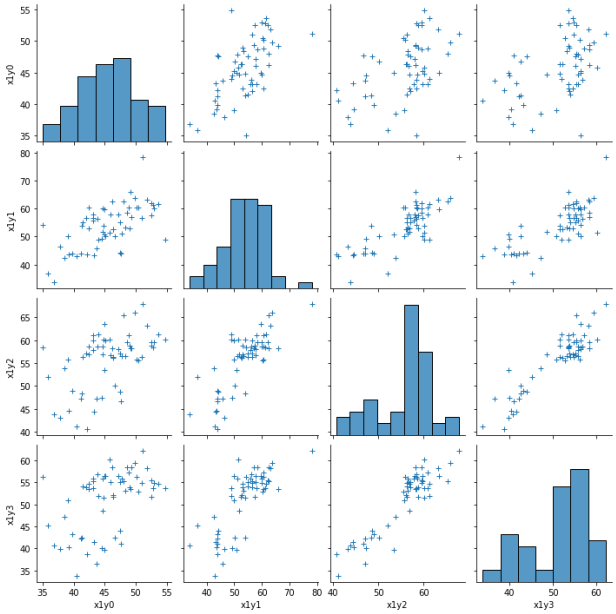


Figure 5: Scatterplot matrix with different coordinates (x=[1, 2, 3], y=[0, 3])

3.4 Temporal Regularity Analysis of Traffic Congestion Values

There is a noticeable difference in traffic congestion between weekdays and non-working days. On weekdays, traffic congestion primarily occurs during peak commuting hours, with two peaks, especially in the morning and evening. This is because people tend to leave early to arrive at work on time, resulting in concentrated traffic pressure within a short period. Additionally, weekday traffic trips are more purpose-driven, and drivers often aim to reach their destinations within specific time frames. This can lead to behaviors such as speeding, lane cutting, and lane changing, further exacerbating congestion. Therefore, congestion during weekdays tends to last longer, with a lower probability of dispersing. In contrast, non-working days exhibit a more evenly distributed pattern of traffic travel time, without distinct peak periods. Without the time constraints of work, people have more flexibility in their travel, resulting in relatively lower traffic pressure. Therefore, congestion levels on non-working days are usually lower than on weekdays, and congestion duration is relatively shorter. Furthermore, during weekdays, congestion tends to concentrate on main arterial roads within the urban road network, while on weekends and holidays, it is more likely to occur in commercial areas, parks, and public leisure and entertainment zones. Therefore, there are significant differences in traffic congestion between weekdays and non-working days, including congestion levels, duration, and occurrence areas. These differences are primarily influenced by factors such as travel times, travel purposes, and traffic pressure. The traffic congestion data from Monday to Sunday is plotted in Figure 6, as shown below.

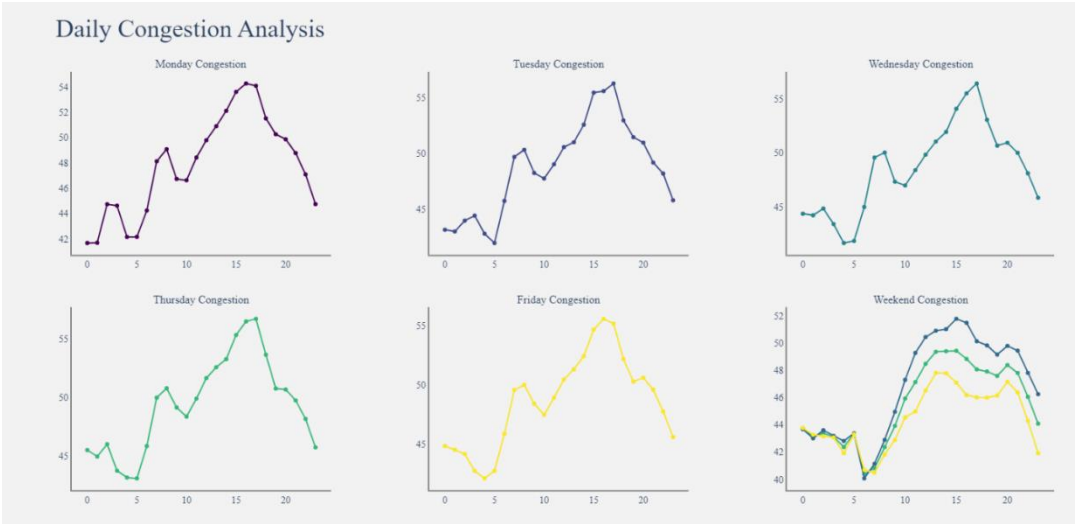


Figure 6: Daily Trend of Traffic Congestion Levels on working and Non-working Days

By observing the data presented in Figure 6, it becomes evident that there are two significant traffic peaks during weekdays (Monday to Friday). In the morning, from 8:00 to 9:00 AM, there is a sudden surge in road congestion as people leave for work or school. In the afternoon, from 4:00 to 6:00 PM, traffic volume increases again as it is the time for people to head home after work, forming the second peak. It is worth noting that the afternoon peak is generally more severe than the morning peak, which could be attributed to higher traffic volume and more complex road conditions during the evening rush hour. In contrast, non-working days (Saturday and Sunday) exhibit a different pattern of traffic congestion. There are no distinct morning and evening peaks during these days. Instead, there is a relatively prolonged peak period starting from around 1:00 to 2:00 PM and lasting until the evening, around 7:00 to 8:00 PM, followed by a rapid decline in traffic congestion. This indicates that on non-working days, people's travel times and purposes are relatively dispersed, and there is no concentrated peak travel period as observed on weekdays. In summary, weekdays (Monday to Friday) show a

certain regularity in traffic congestion patterns, while Saturdays and Sundays display similar patterns. This regularity suggests a cyclical nature of traffic congestion over time, which is crucial for understanding and predicting traffic congestion patterns and formulating effective traffic management strategies.

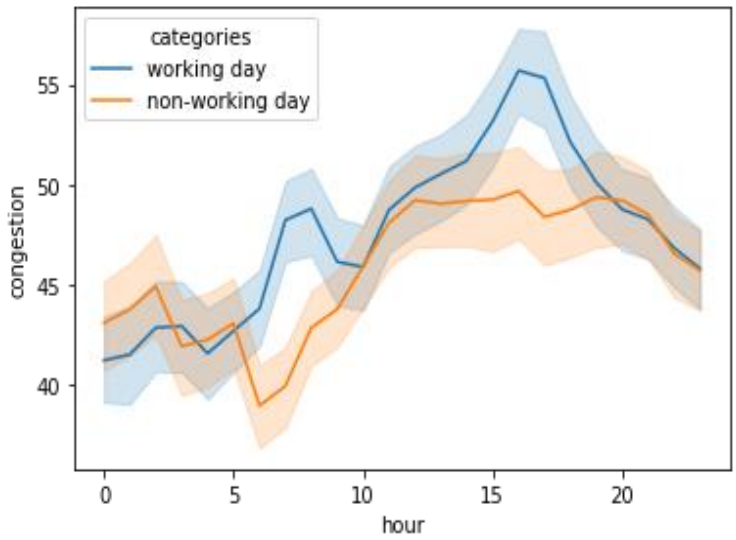


Figure 7: Comparison of Congestion Levels between Weekdays and Non-working Days

According to Figure 7, there are noticeable differences in traffic congestion between non-working days and weekdays, not only in terms of regularity but also in congestion values. It is worth noting that during the morning peak period on non-working days, there is a slight decrease in congestion levels. This may be attributed to the fact that most people do not have the habit of going out early in the morning on non-working days. Furthermore, compared to the congestion values during the evening peak on weekdays, the congestion levels during the evening peak on non-working days are lower but generally more stable, encompassing multiple time periods. This could be due to the increased flexibility in people's travel times on non-working days, as they are not constrained by work schedules. Consequently, the distribution of traffic flow becomes more even, reducing the severity of congestion. These findings highlight the distinct characteristics of traffic congestion between non-working days and weekdays. Understanding these differences is crucial for devising effective traffic management strategies tailored to each context. By leveraging the insights provided by Figure 7, policymakers and transportation authorities can implement measures to mitigate congestion during peak periods on weekdays and ensure smoother traffic conditions on non-working days.

3.5 Modeling of Traffic Congestion Prediction

3.5.1 Data segmentation

Randomly selecting 80% of all samples as the training set and allocating the remaining 20% as the test set, the trained model is utilized to obtain predicted values for the test set data.

3.5.2 Model hyperparameterization

In the paper, experimental investigations were conducted using two prediction models, Lasso and Ridge, with the optimal hyperparameters of these models determined through grid search. The parameter configurations for the Lasso algorithm are presented in Table 2, while the parameter configurations for the Ridge algorithm are presented in Table 3.

Table 2: Lasso Hyperparameter Settings

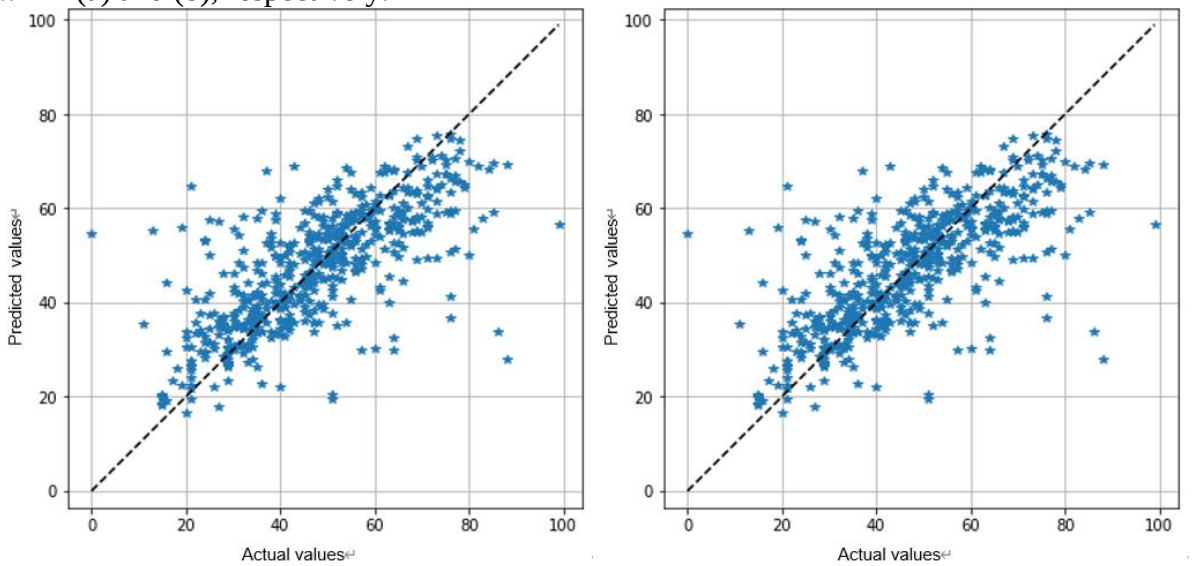
Parameter name	parameter value
alpha	0
fit_intercept	True
positive	False
precompute	False
copy_X	True
max_iter	1000

Table 3: Ridge Hyperparameter Settings

Parameter name	parameter value
alpha	1
fit_intercept	True
normalize	False
copy_X	True
max-iter	None

3.5.3 Model prediction results

To simulate the models with the aforementioned hyperparameters and generate the predicted results using Python, the predictions for the Lasso and Ridge algorithms are presented in Figure 8, as shown in (a) and (b), respectively.



(a) Lasso Prediction result (b) Ridge Prediction Result

Figure 8: Model prediction results

The scatter plot in Figure 8, which depicts the relationship between the actual values and the predicted values, is utilized to evaluate the performance of the regression models. By observing this plot, one can intuitively assess the accuracy and bias of the model predictions. Upon examining the distribution of the scatter plot in Figure 8, it is apparent that the majority of the data points roughly follow a linear pattern. However, a few scattered points deviate significantly from the main cluster, representing outliers. This suggests that while the accuracy of both models' predictions is acceptable, there is still room for improvement.

3.5.4 Evaluation

During the model evaluation phase, different evaluation metrics focus on different aspects of model performance. For example, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) measure the squared differences between the predicted values and the actual values, while Mean Absolute Error (MAE) focuses on the absolute differences between the predicted values and the actual values. By considering multiple metrics, we can avoid the limitations of relying on a single metric and gain a more comprehensive understanding of the model's performance across different aspects. In this study, we calculated the MSE, MAE, and RMSE between the predicted values and the actual values, and used these three metrics to evaluate the models. The specific numerical values of the prediction errors for different models are presented in Table 4.

Table 4: Comparison of Prediction Error Metrics

Model	RMSE	MSE	MAE
Lasso	11.470	131.563	8.044
Ridge	11.470	131.563	8.044

3.5.5 Analysis of results

The analysis of Table 1 and Fig. 8 shows that the two algorithms predict similar results.

4. Conclusions

In this paper, an analysis of traffic congestion data was conducted, focusing on both the spatial and temporal dimensions. Regarding the spatial dimension, violin plots were initially employed to explore the differences in traffic congestion across different spatial locations. The analysis revealed notable variations in congestion levels among these locations. Subsequently, a scatter plot matrix was constructed, depicting the positive correlation between congestion values in different spatial coordinates. These findings underscore the necessity of considering spatial factors when constructing models. Regarding the temporal dimension, distinct disparities in traffic congestion were observed between weekdays and non-working days. To predict and simulate the traffic congestion, two algorithms were employed in this study. The results indicate that the predictions yielded by both models were similar, with MSE, RMSE, and MAE values of 131.563, 11.470, and 8.044, respectively.

Although the LASSO and Ridge algorithms, being a linear regression method, yielded acceptable prediction results, there is room for improvement. In the next step, nonlinear regression techniques, decision tree regression, and ensemble learning algorithms such as XGBoost regression will be employed to enhance the accuracy of traffic congestion prediction. By incorporating these advanced methodologies, it is anticipated that the predictive accuracy of the models will be further enhanced.

References

[1] Cheng Shuxiang, Xiong Shiqi, Liu Jun, et al. Highway Bridge Construction Cost Prediction Model Based on Ridge Regression Optimization Algorithm [J]. Construction Economics, 2023, 44 (S2):225-229.

[2] Niu Kai, et al. Prediction of Urban Traffic Congestion Level Based on BP Neural Network [J]. Tianjin Construction Science and Technology, 2021, 31(05):7-9.

[3] Tian Yu, et al. Traffic Congestion Prediction Model Based on Integrated Learning [J]. Computer & Telecommunication, 2020(04):60-63+70.

[4] Xiaoyu W, Xingyuan W, Bin M, et al. High-performance reversible data hiding based on ridge regression prediction algorithm [J]. Signal Processing, 2023, 204.

[5] Meihong S, Wenjian W. A network Lasso model for regression [J]. Communications in Statistics - Theory and Methods, 2023, 52(6).