Short-term Spatio-temporal Prediction of Freeway Traffic Volume in California, U.S.A: By statistical modelling and Deep Learning

CEGE0042 Coursework Report

1. Introduction

Nowadays, studying the big data of traffic volume is essential in traffic engineering. Exploring the patterns behind the data is helpful for traffic resource allocation, traffic control, and transportation system enhancement (Mugdha, 2018). Moreover, considering the spatial and temporal aspects of the data allows a more comprehensive analysis of how traffic volume varies through different periods of time and locations, hence, spatio-temporal forecasting can be made.

This study aims to analyse and predict freeway traffic volume data in California, U.S.A., using both traditional statistical modelling and more advanced machine learning techniques. Our goal is to identify the most effective approach for capturing the spatio-temporal patterns present in the traffic data. Furthermore, by comparing the performance of models, we will discuss the strengths of each method based on their structural characteristics and their alignment with the data.

One example of common statistical regression models for spatial-temporal data is Geographically and Temporally Weighted Regression (GTWR). GTWR is an extended version of GWR, and the weight matrix based on distances is determined from coordinates (x, y, t), where (x, y) are the location and t is time point (Huang et al., 2010). It can capture local spatial-temporal variations by fitting separate regression models for each location and time point using dependent and independent variables. Hence, it is more flexible and adaptable to diverse datasets. Another common model is Space-Time Autoregressive Integrated Moving Average (STARIM). Unlike GTWR, STARIM is simpler and focuses more on analysing time series data with moderate spatial dependencies. Therefore, STARIM will be implemented in this study since the traffic volume data obtained is time series, and there are no other variables like weather conditions, congestion situation etc. So STARIM can capture a more specific analysis on spatial and temporal patterns.

A widely used machine learning model for analysing time-series data is the Long Short-Term Memory (LSTM) network, known for its ability to effectively capture long-term dependencies and model temporal patterns. To incorporate the spatial aspect in the model, networks such as Spatial LSTM (ST-LSTM) and Convolutional LSTM (CNN-LSTM) are employed (Tang et al., 2019; Zhao et al., 2021). In the research conducted by (Hu et al., 2021), a ST-LSTM architecture and a framework for analysing and predicting time series data with spatial relationships were introduced. Despite the paper's focus on environmental datasets, the innovative framework is still applicable to the objectives of this report, owing to the similar data structure, simple data pre-processing requirement, and excellent missing values handling.

This study uses the data of traffic volume from Caltrans PeMS. The data source collects real-time data from nearly 40,000 individual detectors spanning the freeway system across all major metropolitan areas of California. Our analysis will base on one month (02/01/2023-29/01/2023) of hourly traffic volume data, on freeway No.5 only. The data pre-processing results a data frame including 109 rows, excluding rows with missing values (each row indicate a time series for one detector), and 672 columns (each column indicate the volume for all 109 detector at the time point).

Consequently, we will first explore the traffic volume data in sections 2. Then follow the structure provided in the course material, lab3, to construct STARIMA for forecasting in section 3.1. And the re-implementation of the ST-LSTM framework presented in the paper (Hu et al., 2021) will be adapted to our analysis in section 3.2. Finally, the discussion between the two models, STARIMA and ST-LSTM will be in section 4.

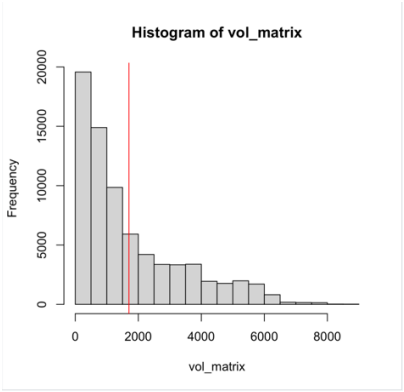

Fig.2.1 Histogram of volume data across all time points and detectors

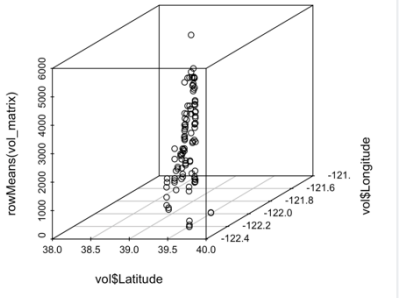

Fig.2.2 3D scatter plot of volume data across all time points and detectors

2. Exploratory Spatial Temporal Data Analysis

2.1 Non spatio-temporal characteristics

Fig.2.1 presents a highly right-skewed distribution in the traffic data, where most of the data is concentrated within 1000 and there are a very few special cases over 8000. From the 3D visualization shown in Fig.2.2, the mean of data for circled detectors in the middle tend to have higher traffic volume.

2.2 Temporal characteristics

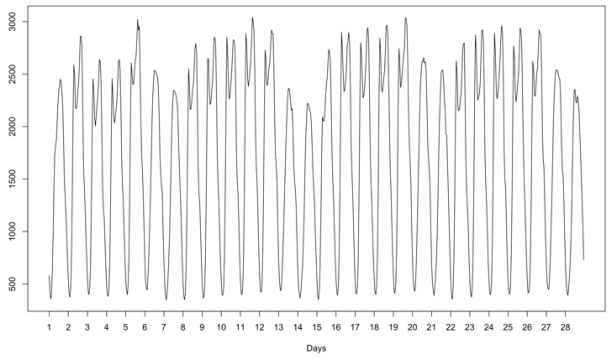

Fig.2.3 Time series of hourly average traffic volume in Freeway No.5

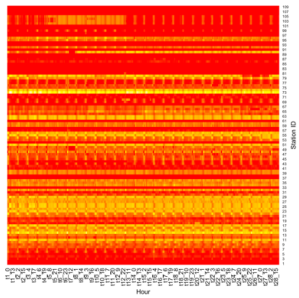

Fig.2.4 Heatmap of traffic volume

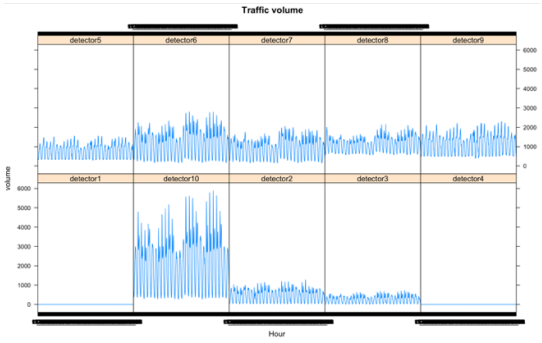

Fig.2.5 Time series of first 10 detectors

Fig.2.3 can be considered as a graph shows how the traffic volume for freeway No.5 changes along with time. A daily periodic pattern is presented, and we can notice that, within a week, the traffic volume on Tuesday to Friday has two noticeable peaks within one day, whereas Monday and Weekends do not have that significant characteristic. Furthermore, the traffic volume in Weekends tend to be lower than weekdays. This can also be observed from Fig.2.4, horizontally, we can see a very similar temporal pattern for every detector, where each yellow segments have identical interval with respect to the daily period pattern. However, significant spatial pattern cannot be observed from here, but we could say that detectors that are near to each other tends to have more similar pattern. But when we take a closer look to the first 10 detectors, as we can see from Fig.2.5, the difference between them can still be observed where detector 3 has lowest overall volume within 1000 but detector 10's volume reaches up to 6000.

2.3 Spatial characteristics

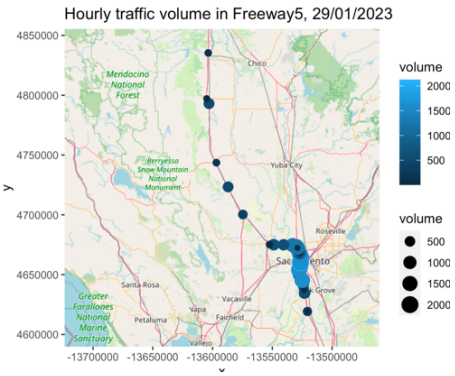

Fig.2.6 Map of hourly average traffic volume in Freeway No.5

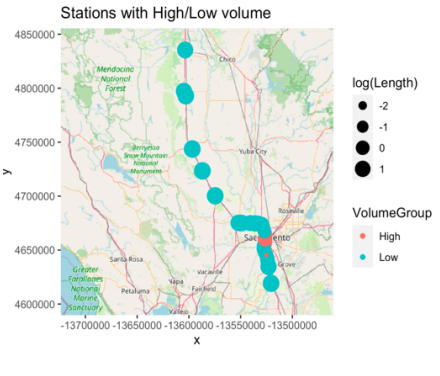

Fig.2.7 Map of hourly average traffic volume classified as volume groups

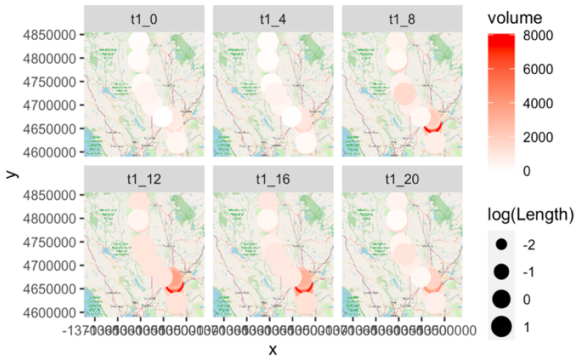

Fig.2.8 Map of hourly traffic volume at different time points

To have an overall realistic visualisation about how the volume data locate along Freeway No.5, Fig.2.6 is plotted. As result discussed from Fig.2.2, stations centered at city Sacramento, detect higher volume than those who are far away. Setting a threshold as 3000 to classify the volume data as 'high' and 'low' , Fig.2.7 detailly evident this observation. Fig.2.8 further shows how the volume change in different hours through out a day (02/01/2023). As we can see, the traffic volume starts to increase from around 8:00 until 16:00, and then decrease again around 20:00. This also coincides with the peaks pattern occurred Fig.2.3.

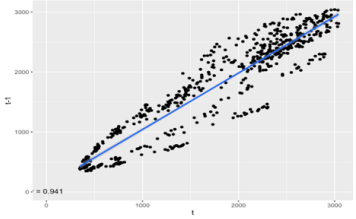

Fig.2.9 temporal autocorrelation of lagged variable

2.4 Autocorrelation Analysis

2.4.1 Temporal

In order to investigate the autocorrelation in temporal and spatial data, the 1 temporal lagged variable is used and Fig.2.9 indicates a strong, and positive autocorrelation, 0.941, in the temporal aspect of our traffic volume data.

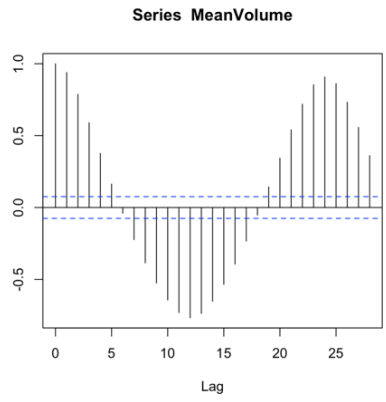

Fig.2.10 ACF of mean traffic volume

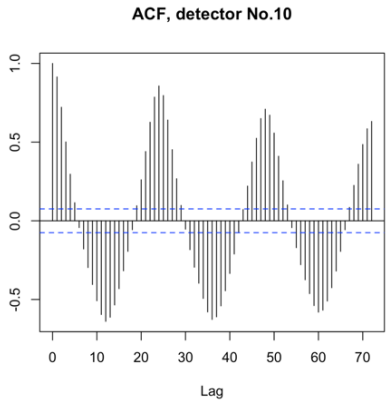

Fig.2.11 ACF of traffic volume at dector No.10

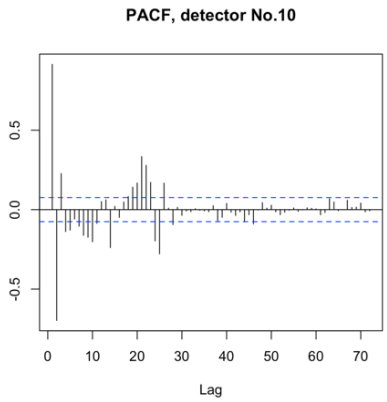

Fig.2.12 PACF of traffic volume at dector No.10

The temporal autocorrelation of the mean volume time series is shown in Fig.2.10 and we can notice that the period of autocorrelation is also 24 hours and within one period, smaller the lag, stronger the autocorrelation. Fig.2.11 and Fig.2.12 are the autocorrelation and the partial autocorrelation graphs for station 311844. And we can see for this specific detector, the autocorrelation shares the identical pattern as the mean volume, but in terms of PACF, lags that are within the daily period are more statistically significant, and the peaks are mostly centered around the 9th lag and the 23rd lag.

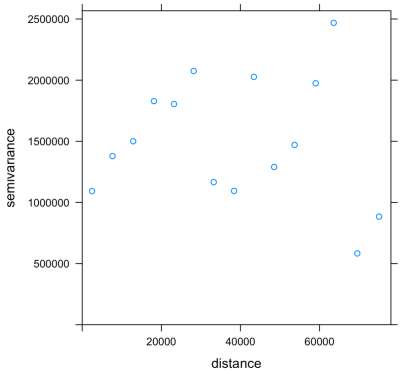

Fig.2.13 Semivariogram of traffic volume data

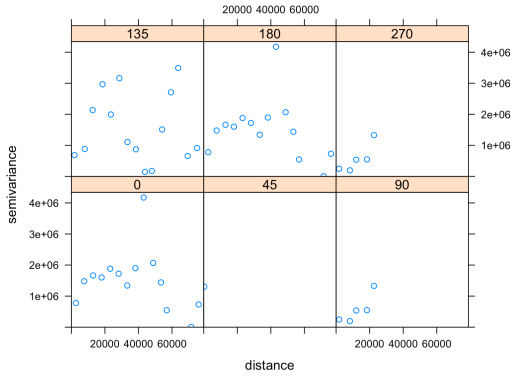

Fig.2.14 Directional semivariance of traffic volume data

2.4.2 Spatial

Since our traffic volume is point data, semi-variogram will be used to investigate to spatial autocorrelation. Unexpectedly, from Fig.2.13 the semi-variances for a relatively small distance range (within 20000) are higher than those for larger distance range (around 40000). And we cannot observe any trends from the overall semi-variogram. Meanwhile, we also investigate the semi-variance at six different directions as shown in Fig.2.14. And we could say that for the horizontal directions (0 and 180), the semi-variance decreases with the distance increase. This might due to the diverse traffic situation in the urban area.

2.4.3 Space-Time

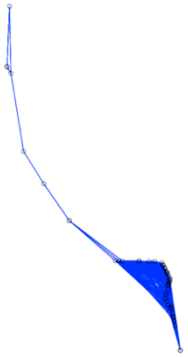

Fig.2.16 Connectivity graph

The distance-band is used to calculate the weight matrix for the traffic volume point data. The workflow by (Anselin & GrantMorrison, 2019) is re-implemented. Since we focus on the data along one freeway, the smallest k allowing all the data points in the connectivity graph Fig.2.16 is chosen. So k-nearest neighbors for k=3 are first find.A critical threshold is then computed by getting the maximum of distance between all the 3-NN neighbors. Then, a new neighbor list can be created, where for each point, its new neighbors are selected by the critical threshold. Hence, the weight matrix can be obtained.
Now, the space-time autocorrelation is ready to be analyzed. Fig.2.17 is the STACF of traffic volume data w.r.t the weight matrix obtained and 72 temporal lags. With the help of Augmented Dickey-Fuller test, a 5-order seasonal differencing is considered, and from Fig.2.18, we notice the autocorrelation drops noticeably.

For the space-time partial autocorrelation shown in Fig.2.19, we can see most of the partial autocorrelation are not that considerable except the four peaks around time lag 1 and lag 25. But add the seasonal difference 5 to it, most the autocorrelations are decreased under the confidence interval.

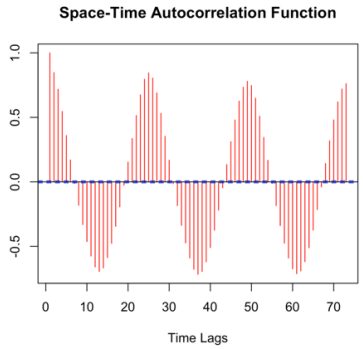

Fig.2.17 ST-ACF of traffic volume data

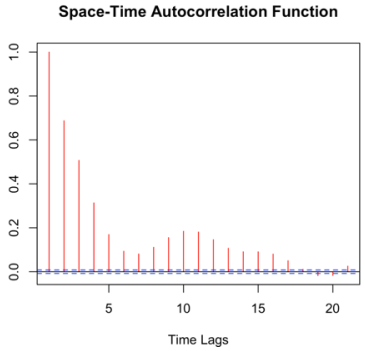

Fig.2.18 ST-ACF after seasonal differencing

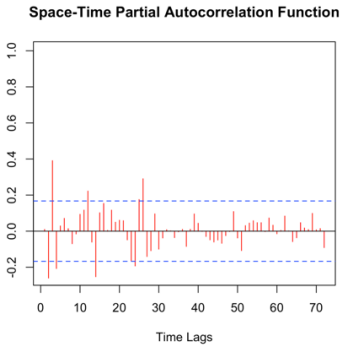

Fig.2.19 ST-PACF of traffic volume data

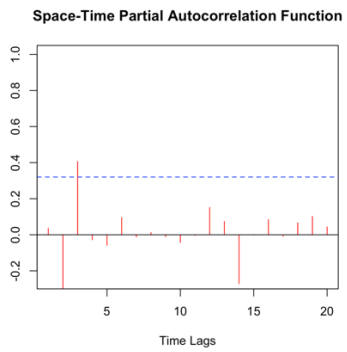

Fig.2.20 ST-PACF after seasonal differencing

3. Methodology

3.1 STARIMA

- **Parameters setting and Model Construction**

Given all the autocorrelation analysis and the weight matrix calculation, we are now ready to build the STARIMA model. According to Fig.2.18, the STACF after seasonal differencing keeps decreasing and reaches a relatively stationary state after 17 lags, so the space-time moving average order q will be set as 17. As shown in Fig.2.20, after 3 lags, the STPACF falls below the confidence interval, suggesting the order p to be 3. Therefore, the model is specified as $STARIMA(p = 3, d = 5, q = 17)$. And then, this model will be trained using data from the first 3 weeks (02/01/2023-22/01/2023)

- **Model Fitting Analysis**

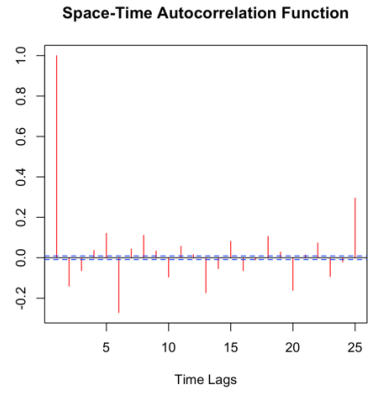

Fig.3.1 ST-ACF of fitting residulas

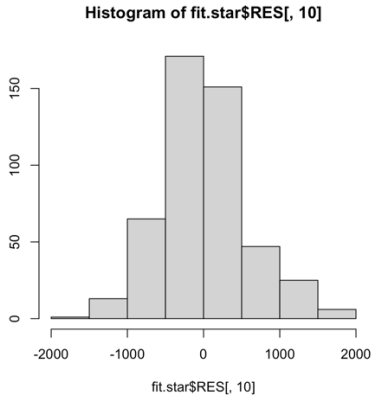

Fig.3.2 Histogram of fitting residulas at detecotR No.10

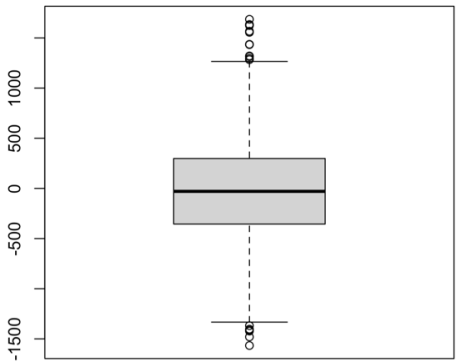
Fig.3.3 BoxPlot of fitting residulas at detecotR No.10

In Fig.3.1, although the space-time autocorrelations for model's fitting residuals are still considered as statistically significant, by the confidence interval, there is a considerable drop after time lag 1. To be more specifically, the histogram of residuals at detector No.10 is shown in Fig.3.2. Almost all the residuals are within 1000, and 50% of them are within 500.

- **Model Predicting Analysis**

To investigate the model predicting ability, Fig.3.4 plots the actual traffic volume as the black line and the predicted traffic volume as the red line together. We can see these two lines coincides quite well except the predicted peaks are always higher than the true value. Meanwhile, for Monday, the difference between the predicted gap and the actual lowest value between the two peaks is considerable.

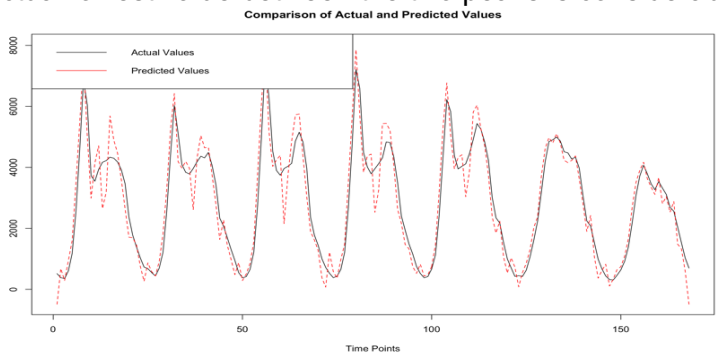

Fig.3.4 Actual v.s. predicted traffic volume at DETECTOR No.10

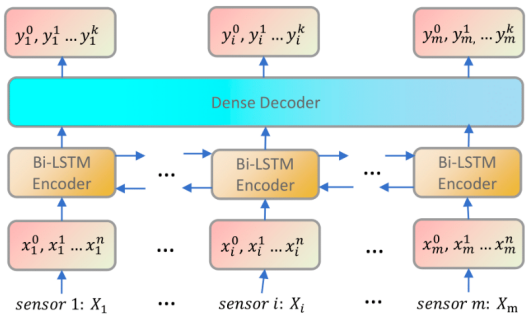

Fig.3.5 Structure of ST-LSTM

3.2 ST-LSTM

- **Parameters setting and Model Construction**

The structure of ST-LSTM constructed by (Hu et al., 2021) is presented in Fig.3.5, where a unidirectional LSTM is applied to each sensor's time series, for the temporal aspect. And between those individual sensors, a series of bidirectional layers are between distinct time series, connecting different sensors together. And this allowing the spatial aspect to be considered. Traffic volume data is normalized in this method by logarithm. A few modifications are added to the code for ST-LSTM, but there is no major change.

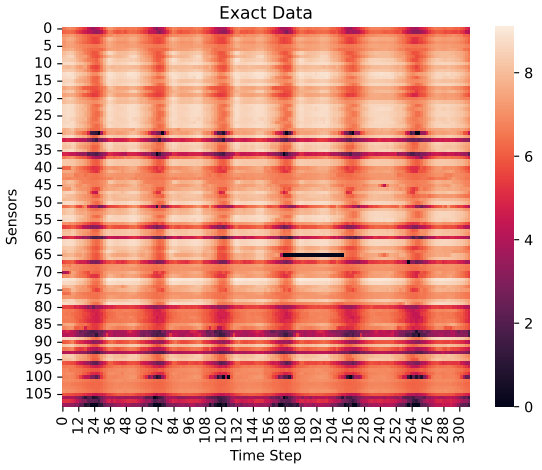

Fig.3.6 Heatmap of actual data point

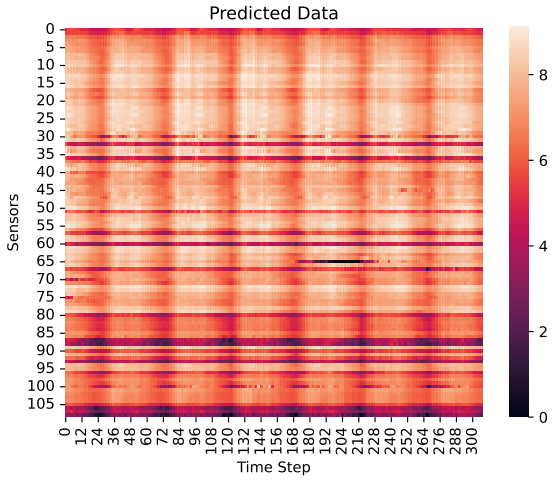

Fig.3.7 Heatmap of predicted data point

To ensure a fair comparison between ST-LSTM and STARIMA, the training data size is standardized as 0.75 of the total dataset, which equates to 3 weeks for training and 1 week for testing and predicting. The ST-LSTM architecture utilizes $train\_hour$ and $predict\_hour$ to create window periods for each time series, representing the duration of training and prediction in hours. To increase the granularity of the model's understanding of the data, we can set the $train\_hour$ and $predict\_hour$ to smaller values. If these values are shorter than the total time series, multiple windows are generated to cover the entire duration. Through several experiments, we determined that setting $train\_hour$ = 6 and $predict\_hour$ = 1 resulted in the best overall performance.

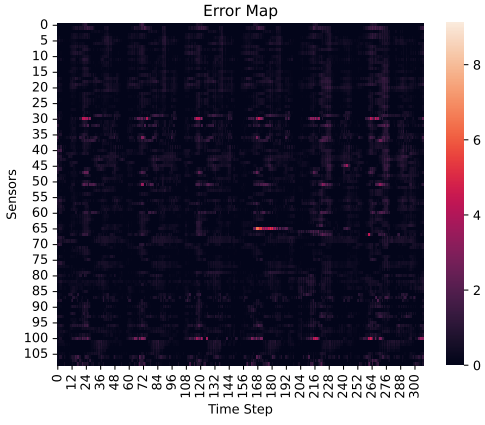

Fig.3.8 Heatmap of errors

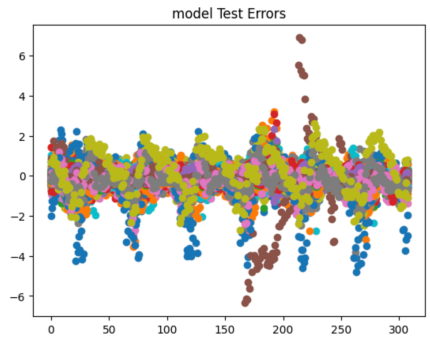

Fig.3.9 Scatter points of errors inn testing

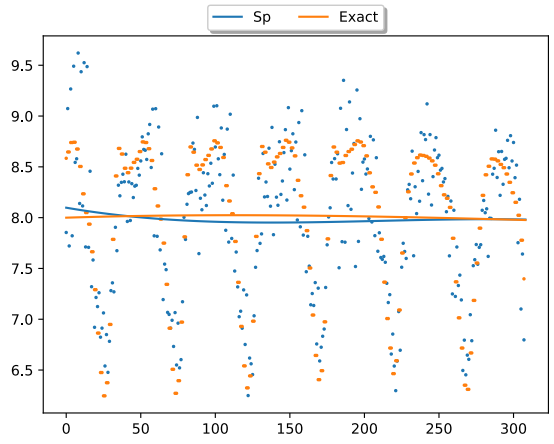

Fig.3.10 Actual v.s. predicted traffic volume at DETECTOR No.10

- **Model Fitting Analysis**

As shown in Fig.3.6 and Fig.3.7, we can see the heatmap of exact data and predicted data show an identical pattern and they are almost the same. With the help from the error map in Fig.3.8, we can see all errors are within 8 and most of the data have errors almost 0.

- **Model Predicting Analysis**

To compare the predictability of ST-LSTM with STARIM, the prediction results of the same detector 10 is shown in Fig.3.10. As we can see, the predicted point and the true data point coincide quite well also. Similar to STARIMA prediction, the ST-LSTM predicted peak values are also higher than those true peak values. And, it cannot accurately capture the gap between the two peaks within each weekday.

3.3 Comparison

Based on the fitting results, it can be inferred that ST-LSTM outperforms STARIMA, as the model produces small errors, demonstrating its capability in effectively capturing the underlying patterns and trends in the data. In terms of prediction results, both models demonstrate the ability to accurately predict traffic volume, and can effectively distinguish between the different rush hours and off-peak hours for weekdays and weekends.

- **Strengths**

STARIM has the advantage in its structure designed for time series, where the autoregressive order and moving average order can deal with the complex autocorrelation in both spatial and temporal perspectives. Also the seasonal differencing can deal with the periodic pattern in time series.

In the other hand, ST-LSTM capture the temporal dependencies within and between the time series data by using stacked bidirectional LSTM layers. This allows for a more accurate and robust modeling of complex spatio-temporal patterns and trends in the data.

One of the discussions point here is the different method to deal with spatial relationship between data points. STARIM relies on the weight matrix to model the spatial dependencies, and this allows a more specific, and targeted autocorrelation investigation between 'related' detectors. Whereas ST-LSTM uses the bi-directional method to capture the spatial relationships between detectors, allowing for a more comprehensive modelling that can also be applied for different kind of time series data.

- **Ease of implementation:**

For this study, STARIM requires more analysis before implementation, for examples, the weight matrix calculation and temporal, spatial and space-time autocorrelation analysis. Whereas the ST-LSTM only requires a simple data-preprocessing and parameters decisions. Hence, ST-LSTM is more user-friendly if more complex data is included.

- **Running time:**

Due to the different model architecture, ST-LSTM has longer tunning time around 8 minutes in this task. The training process for it involves multiple iteration through multiple windows for each time series, so it is very computationally intensive. However, STARIM is a lot faster, requiring inly x minutes.

To sum up, both models have their own strength in different perspectives, so they should be wisely chosen depending on different data type, learning, and forecasting tasks.

4. Discussion and Conclusion

There are still limitations in this study. From the data perspective, the Christmas holiday is included in the period we study. Hence, although the overall pattern is not supposed to be changed a lot, but the data might still be influenced by this factor. Additionally, missing values are excluded in this study. It is because the two models' treatments for the missing values are not the same, hence, to have a fair and consistent comparison between the two models, detectors containing missing values were removed. Doing this results in a smaller training dataset for the model and lose in the information.

From the methodology perspective, since the traffic volume data is transformed within ST-LSTM, so the fitting and predicting values are the logged values but not the actual volume values. This makes it difficult to compare the results to the untransformed values produced by STARIMA. While this issue can be resolved by inverse logarithm transformation, it is not included in this report due to length constraints. However, it is recommended for future work to make the comparison more reasonable.

In conclusion, this report provides a work frame of comparing the abilities of statistical model STARIMA and the deep learning model ST-LSTM in capture the spatial and temporal traffic volume data in Freeway No.5 in California. The strengths and limitations of the data itself and the two models are also discussed in detail.

References

Mugdha, P., 2018. *Engineering Notes.* [Online]
Available at: https://www.engineeringenotes.com/transportation-engineering/traffic-engineering/traffic-volume-studies-flow-characteristics-and-forecasting-engineering/48401

Anselin, L. & GrantMorrison, 2019. *Distance-based spatial weights.* [Online]
Available at: https://spatialanalysis.github.io/lab_tutorials/Distance_Based_Spatial_Weights.html#distance-band-weights

Caltrans, 2023. *Caltrans PeMS.* [Online]
Available at: https://pems.dot.ca.gov/?dnode=Clearinghouse&type=station_5min&district_id=3&submit=Submit

., O'Donncha, F., Palmes, P., Burke, M., Filgueira, R., & Grant, J. (2021). *A spatio-temporal LSTM model to forecast across multiple temporal and spatial scales*.

ng, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, *24*(3), 383–401. https://doi.org/10.1080/13658810802672469

, Q., Yang, M., & Yang, Y. (2019). ST-LSTM: A Deep Learning Approach Combined Spatio-Temporal Features for Short-Term Forecast in Rail Transit. *Journal of Advanced Transportation*, *2019*, 1–8. https://doi.org/10.1155/2019/8392592

, Z., Li, Z., Li, F., & Liu, Y. (2021). CNN-LSTM Based Traffic Prediction Using Spatial-temporal Features. *Journal of Physics: Conference Series*, *2037*(1), 012065. https://doi.org/10.1088/1742-6596/2037/1/012065