National College of Ireland

Comparison of machine learning in Intelligence Traffic System

MSc Research Project
Short term traffic flow prediction

Hongyi Yan
Student ID: 19207433

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …Hongyi Yan……

Student ID: …19207433……

Programme: …short traffic flow prediction…… **Year:** …2021……..

Module: …machine learning……

Supervisor: …Jorge Basilio……

Submission Due Date: …08.16……

Project Title: …Comparison of machine learning in Intelligence Traffic System…

Word Count: …6866…… **Page Count**…19……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …Hongyi Yan……

Date: …08.16……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	√
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	√
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	√

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparison of machine learning in Intelligence Traffic System

Hongyi Yan
19207433

Abstract

After the previous literature survey, it is found that short-term traffic flow prediction is very important in Intelligent Traffic System(ITS). In the introduction and literature review of this paper, the research direction and value will be determined. High precision prediction results play a positive role in traffic data transmission and congestion intelligent regulation. In the introduction part, the structure and development of ITS will be described in detail. In the data exploration stage, the basic analysis of the data set will be carried out. At this part, the actual analysis will be carried out in combination with the reality and research value, visual analyses is also included. In addition, ARIMA and LSTM are used to predict time series. The seasonal concept ARIMA and ordinary ARIMA are introduced to compare the results to prove whether there is seasonality in the traffic data. In order to compare the prediction accuracy with the traditional machine learning algorithm, the more promising deep learning algorithm is also implemented. Finally, the results of this study are summarized, and the future research direction is discussed to improve the traffic flow prediction accuracy.

1 Introduction

Because of the improvement of human life quality and the development of economy, purchasing the private cars has become quite common. This is because the development of technology reduce the cost of making cars. It is quite convenient to have a car for daily life. Therefore, if the economic conditions permit, more families will choose to buy a car as a daily travel tool (Yang et al.; 2019).

But when the number of cars on the road is increasing, the traffic situation will become more and more complex. Low speed driving and traffic congestion are the sequelae of it, and even crash events that threaten human health will occur. Traffic congestion means that in a period of time, the load capacity of traffic remains unchanged, but due to the increase of traffic demand, the total traffic flow of a certain section is greater than the traffic capacity of the road. And eventually lead to the car can not pass smoothly on the lane, appear slow driving phenomenon (Yanguo; 2015).

In fact, in the early stages, the government hoped to reduce traffic congestion by constantly widening roads and other basic ways. However, with the development of society, the complex urban transportation system and high labour costs have made the traditional way insufficient to meet the needs of people's high quality of life. Especially in the early urban construction, the traffic planning of some important areas (CBD) has been completed, even if it is unreasonable, it is difficult to replan and construct.

The control of Government Transportation Department is regarded as a solution to traffic congestion, which was tried in the middle of last century. There are a lot of human subjective factors in the early control, and the most classic one is the signal control. This method is low-cost and easy to implement, and can have good performance without a lot of data analysis and experimental research (Zhao et al.; 2012). With the development of technology, the excellent computing power of computer is more and more expected to participate in the process of traffic control. With the development of intelligent theory, transportation intelligent system appears.

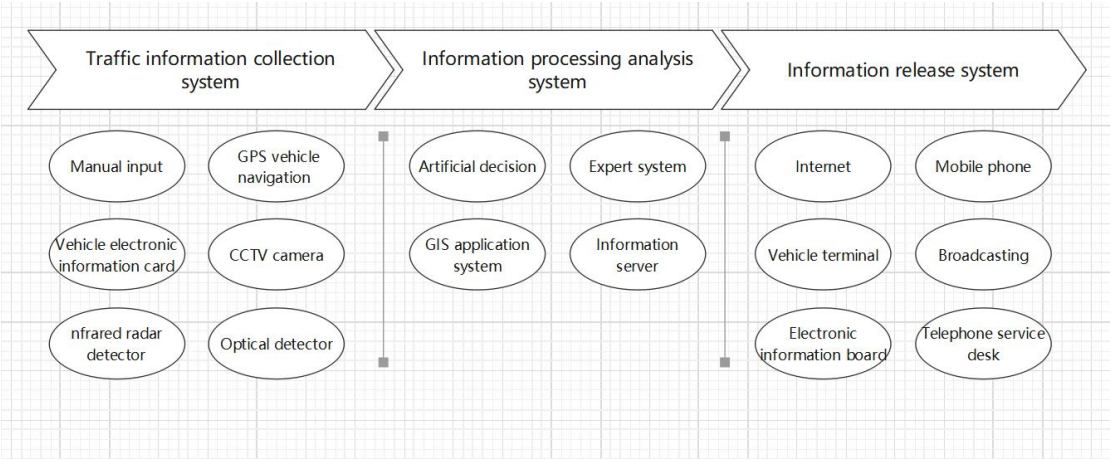

Figure 1: The whole ITS system

The figure 1 is a summary of the core technologies in ITS. This is a complex program that runs in coordination with many aspects. It needs expensive hardware and excellent software to work. Even in recent years, the theory makes the system more complex.

To sum up, it can be seen that the deployment cost of the whole system is very high, and the application value will not be worth in some economically underdeveloped areas. With the increase of population, but constrained by the lack of budget, semi intelligent system is proposed and applied. Of course, in this system, the proportion of artificial regulation will be relatively large.

In addition, by the previous survey, the signal time control solution has the longest history, and still exists in today's era. Its deployment cost is low. It only needs to introduce a traffic prediction system on the basis of having traffic lights control system. No matter the complex system or the single function which is mentioned above can not work without the underlying core technology, traffic flow prediction. Only when the short-term traffic flow forecast is accurate, it will be of guiding significance to various complex traffic control systems.

Therefore, the research on traffic flow prediction has been very popular. Especially under the background of the continuous improvement of machine learning and deep learning theory in recent years, more accurate traffic flow prediction becomes possible.

To sum up, traffic flow prediction is the basic technology of advanced intelligence traffic system(ITS). This paper hopes to improve the operation efficiency of ITS by enhancing the prediction accuracy of short-term flow. The model is established with time as a variable. Therefore, the most basic time series will be applied. In addition, neural networks and deep learning algorithms will also be applied to compare the prediction results of different algorithms through parameter optimization. Finally, a relatively good model which performs better in traffic flow prediction is obtained. As for the existence of seasonality, it is also the research objective of this paper.

Research Question:

1. Which time series model is better for short-term traffic flow prediction? Is the neural network model LSTM more suitable than the ARIMA model?

2. Whether the original model can be improved by introducing the definition of seasonality?

Next, some papers will be quoted which is include the application of machine learning in the field of transportation, and compare their research results. Through these ideas to determine the methodology of this study.

2 Literature review

Internet-of-Vehicles (IoV) is regarded as a direct technology to solve road congestion, including vehicle to vehicle (V2V) or networking a single vehicle with infrastructure. A common IoV system is composed of vehicles and road infrastructure in a certain section(Qu et al.; 2015). In fact, the structure of this vehicle network is more conducive to the management and scheduling together of traffic police, or it will be more clear about the current traffic flow and future trend (Sun and Samaan; 2020).

In this mode, when there is a sudden increase in vehicle flow (traffic congestion), the interaction and transmission of vehicle information will become very slow. Because IoV has extremely high topological activity, this system needs to be continuously selected in the face of congestion. It is a repetitive behaviour, that is, doing a lot of useless work. Therefore, if the traffic flow can be predicted in advance, this continuous selection behaviour can be

reduced and the transmission efficiency of data in this network can be accelerated (Sun and Boukerche; 2020).

As for the macro-control level, traffic intelligent system is essential. After obtaining the current road information, predict the future traffic conditions. It directly regulates the parts that will be congested, such as the change of tidal Lane (Boukerche et al.; 2020). The time control of the signal lamp mentioned above is also one of them, which can enhance the dredging capacity of the road section which will be congested by squeezing the traffic time in other driving directions. It is worth noting that the regulation of traffic intelligence should also be completed in cooperation with the IOV mentioned above (Sun and Samaan; 2020).

Auto Regressive Integrated Moving Average (Johnston and DiNardo; 1963) is the most common algorithm for processing time series. It can describe the correlation between different data points, and take into account the differences between the values. ARIMA combines three basic methods (AR, I, MA), so it is a relatively stable algorithm (Li and De Moor; 2002).

By learning the regular of historical data and modelling with ARIMA, Dehuai Zeng found that the accuracy of short-term traffic prediction is relatively high, and proved that it has research value (Zeng et al.; 2008). In this experiment, a hybrid model is established to predict, and it is found that the short-term prediction efficiency is higher. The advantages of ARIMA and artificial neural network in linear and nonlinear are used to predict the distribution, and then the results are summarized. The final results show that in the short-term forecasting problem, the prediction stationarity of the hybrid model may be better than that of the single model.

As for the influence of weather factors on traffic, people's research direction has also changed. In fact, we have to admit that extreme weather has something to do with traffic flow and traffic congestion. However, there has been still no clear definition of how it affects. In recent years, scholars hope to establish a simulated weather model for traffic flow prediction. In this process, there are a lot of complex mathematical formulas, which is the description of the traffic structure (Du et al.; 2018). Shengdong Du proposed a logical framework which is called HMDLF and integrated it into deep learning model. This will be used to simulate various traffic conditions and events, and finally transfer the predicted traffic state to convolution neural network for traffic flow prediction. Of course, this model not only simulates the weather variables, but also involves the high contingency events such like the vehicle collision. But these factors are difficult to be simulated, and with the development of technology, the short-term weather prediction accuracy has been relatively high. Therefore, this step may be redundant, because the short-term weather prediction accuracy is already high, we can choose this kind of weather data into the model for traffic flow prediction.

In addition, in the actual data set, data pre-processing is also very important, such as the filling of null values. S. Narmadha proposes to use stacked denoise autoencoder technology to fill in the null values of time series (Narmadha and Vijayakumar; 2021). In addition, in his research, it is proved that the prediction accuracy of the model based on the data of using SDAE null filling technology is higher than that without null processing. It uses the most measurable indicator of RMSE. The filling method with ELU activation function used by him has higher accuracy than ordinary SDAE. In addition, S. narmadha believes that convolutional neural network is more accurate in predicting the spatial variables of traffic

prediction, and LSTM has a strong ability to capture the time characteristics of time series (Narmadha and Vijayakumar; 2021). Therefore, he combined the two ideas and found that the prediction accuracy was improved. This shows that the two algorithms can be combined when the data dimension is high and involves time and space.

3 Methodology

3.1 CRISP-DM

Cross-Industry Standard Process for Data Mining(CRISP-DM) is a standard process which is developed by early SPSS and other companies in this industry. A total of six stages are iterated. Each module can be interspersed, but generally proceed in order(figure 2). The research method of this paper will also follow this step.

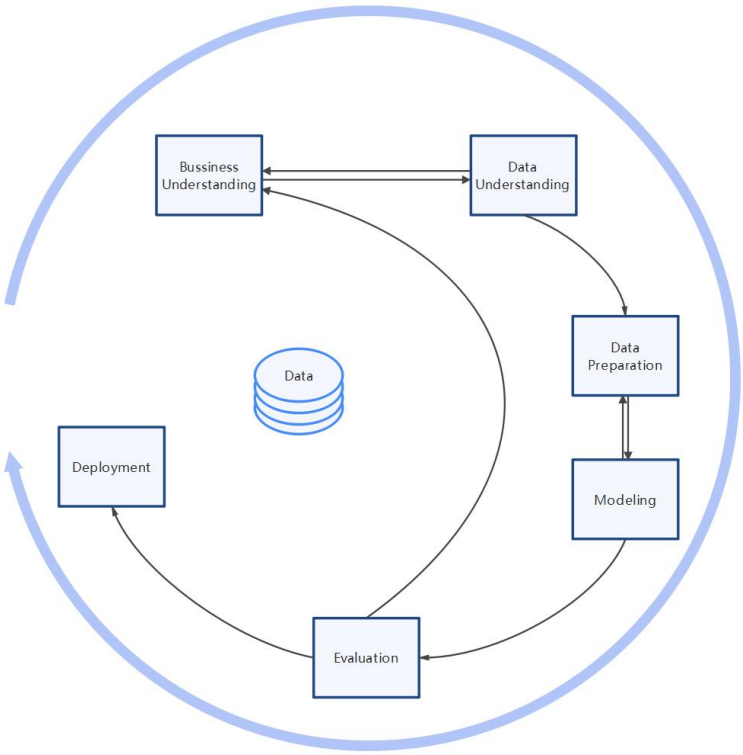

Figure 2: the CRISP-DM process

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

In fact, the first and second item (business understanding and data understanding) have been introduced before this article. These two are mainly combined with the actual situation to analyse the application and value of the project. This helps to guide data mining to a more valuable direction. And in this process of understanding, there is an iterative situation. Analyse the high-value direction, and then find the appropriate data set for the next step of

understanding. The understanding of data also needs to be combined with business knowledge. Finally, summarize the value of the research.

3.2 Data Preparation

This research chooses Python language and Anaconda as the working environment. The first is to read the original data into the working environment. Then, the basic data exploration process is carried out to clarify the practical significance represented by each column of variables. Because this paper selects time series, it is necessary to consider which column is regarded as time variables for research. After it is determined, we can draw a line chart and study the changes of variables, such as whether there is a trend, etc. This step is the visualization of data, which is carried out by using the plot() function. The exploration of data analyse needs to be combined with business understanding. Finally, the data pre-process needs to be implemented, including null value processing and so on. This study deleted the inappropriate weekend data in this step, and regarded the C1 variable as traffic flow (short-term).

3.3 Time series

Time series is a classic concept in statistics, which is the data points measured in a fixed time interval. In this sequence, it can be understood that each data point is associated with the previous data. And there may be a trend.

3.4 ARIMA

Autoregressive Integrated Moving Average(ARIMA) model is a very popular model, which is applied in time series. ARIMA can accurately describe the differences between the data points and the relationship between the data series.

The establishment of this model requires the following assumptions:

1. The input data should be univariate data.

2. First of all, the data should be stable, which means that the mean and variance of the data should not change over time. Generally, in order to realize the stationarity of data, the difference method will be selected.

Of course, ARIMA has three components: autoregressive term(p), difference term(d) and moving average term(q).

3.4.1 AR

It can describe the relationship between the current value and the historical value. The historical data is used to predict the future series which is about the variables themselves, on the premise of meeting the stationarity.

The autoregressive process of p order formular is defined as follows:

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$$

- y_t is the current value

- μ is a constant term
- P is the lag order
- γ_i is the autocorrelation coefficient
- $\in_t$ is error

3.4.2 MA

MA is similar to autoregressive model, its prediction is based on the change of residual term for training and modelling. The accumulation of error terms has been paid more attention in the Moving Average. In general, using MA will be helpful to alleviate the random fluctuation of the model prediction results.

The formula of autoregressive process of order q is as follows:

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

3.5 Seasonal-ARIMA

In fact, the above part can be regarded as a non seasonal ARIMA model. After introducing the concept of seasonality, it will need to make a simple deformation of the original model.

$$SARIMA(p, d, q) \times (P, D, Q, S)$$

The reason for introducing seasonality is that there are peak and low periods in the trend of traffic flow. And this regular pattern goes on with time, with a period of 24 hours (Seasonal). Therefore, it is necessary to establish seasonal ARIMA models to study whether they have better performance.

3.5.1 SAR

Seasonal autoregression (P), which is similar to ordinary AR, but has seasonal AR values. The SAR value is determined by observing the partial correlation map. For example, observing how many cycles the truncation of PACF occurs. Simply put, it needs to observe the time lag on the multiple of the length of the season. For example, if the period is 12, it is necessary to observe the appearance of lags with values of 12, 24 and 36, determine whether there is hysteresis, and determine the order of hysteresis (Permanasari et al.; 2013).

3.5.2 SMA

The seasonal moving average index is generated on the basis of the original moving average combined with the seasonal theory. By establishing the value of SMA, we can not only eliminate long-term trend changes, but also reduce accidental changes and periodic cyclic changes. Similarly, it is also based on the changes of the original error term for memory and continuous training to reduce the error that resemble to SMA (Olsson and Soder; 2008). In addition, it can accurately reflect some seasonal changes. Therefore, it is necessary to establish an accurate value of SMA, which can improve the prediction accuracy of seasonal time series.

3.6 ACF and PACF

ACF is a complete autocorrelation function, while partial autocorrelation function (PACF) is not comprehensive. ACF can provide the autocorrelation value of any sequence which has hysteresis. Generally speaking, It can describe the correlation between the current value and historical value (different timing) of the data. As mentioned earlier, time series can include various components. Such as periodicity, positive or negative trends, seasonality and residuals, etc. ACF will constantly look for these components in the process of establishment, and describe the direct and indirect correlation.

The autocorrelation function formula is as follows:

$$ACF = \rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)}$$

The value range of ρ_k is [-1, 1].

The establishment of PACF is not the same as ACF to find the current value and its lag relationship, but to study the correlation with the next lag value by calculating the residual. PACF eliminates the influence of intermediate value $a_{(n-1)}$, $a_{(n-2)}$,……$a_{(n-a+1)}$, and in fact, there are some random variables, a-1 in total, are related to $x_{(n-a)}$.

In conclusion, PACF only strictly explores the correlation between these two variables.

The following is the value solution for ACF and PACF:

Table 1: The parameter selection method

Model	ACF	PACF
AR(p)	Attenuation tends to zero (geometric or oscillatory)	P-order posterior truncate
MA(q)	q-order posterior truncate	Attenuation tends to zero (geometric or oscillatory)
ARMA(p,q)	q-order posterior attenuation tends to zero (geometric or oscillatory)	q-order posterior attenuation tends to zero (geometric or oscillatory)

3.7 LSTM

The RNN is a kind of neural network structure in which the form of a ring keeps circulating in order to remember something. It is good at modelling sequence data. Of course, time series is also included. The essence of RNN is that it can have the ability of memory like human beings, relying on the previous learning (memory) combined with the current input and output data. This is in line with the concept of time series. The later data is related to the previous data. It can be said that continuous circulation enables RNN to have the ability of memory and use this past memory to predict future data.

But this memory ability also has disadvantages. That is, the internal forgetting logic is too simple. There is only a single tanh layer, resulting in a large number of forgetting from the first part of the memory to the later part(Luo et al.; 2019). This situation is not expected in this study. Because the traffic flow change is strongly related to the previous data, it can not be forgotten although it is a long time ago. This will affect the prediction accuracy. Therefore, a more complex RNN is proposed, which is the LSTM model.

The interior has a more complex structure. This includes input and output gates. Ordinary RNN can only simply stack the contents that need to be remembered, which is very different from LSTM. A sigmoid function and operation structure are composed of a gate, which is used to control the transmission of information. In addition, it can control the forgetting of unimportant things and the long-term memory of information that needs to be remembered. LSTM is especially useful when you need to have a good memory of the previous content with a long time series. However, due to the introduction of a lot of content, there are more parameters, which also makes the training process more difficult.

4 Implementation

4.1 Exploratory Data Analysis

The data set which is selected is from the open information platform of Queensland, and it records the change of traffic flow of a certain road section from 2006 to now. Considering the conditions are not same of different roads, the same data of SITE is selected for visual analysis. Firstly, the large amount of data and relatively recent data is selected. The dataset is as follows:

	SITE	Direction	DAY	TIME	C1	C2	C3	C4-13	Ped	Bike
297229	2015WH09	Entry	12/03/2015	0:00:00	0	0.0	0.0	0.0	NaN	NaN
297230	2015WH09	Entry	12/03/2015	0:15:00	0	0.0	0.0	0.0	NaN	NaN
297231	2015WH09	Entry	12/03/2015	0:30:00	0	0.0	0.0	0.0	NaN	NaN
297232	2015WH09	Entry	12/03/2015	0:45:00	0	0.0	0.0	0.0	NaN	NaN
297233	2015WH09	Entry	12/03/2015	1:00:00	0	0.0	0.0	0.0	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
297896	2015WH09	Entry	18/03/2015	22:45:00	0	0.0	0.0	0.0	NaN	NaN
297897	2015WH09	Entry	18/03/2015	23:00:00	1	0.0	0.0	0.0	NaN	NaN
297898	2015WH09	Entry	18/03/2015	23:15:00	0	0.0	0.0	0.0	NaN	NaN
297899	2015WH09	Entry	18/03/2015	23:30:00	0	0.0	0.0	0.0	NaN	NaN
297900	2015WH09	Entry	18/03/2015	23:45:00	0	0.0	0.0	0.0	NaN	NaN

Figure 3: The head display of dataset

But it is interval statistics, statistics every 15 minutes. Different kinds of cars are counted separately, C1 is the most common type of vehicle, and it has the most counts, the trend is larger than other. Therefore, the research value of C1 will be very high.

Through the exploration, it is found that the Direction variable records the condition of the car, the Entry represents the number of cars entering the road in this period, and the Exit represents the number of cars leaving the road in this period. Delete data which is leaving the lane (Direction = Exit), and only count the number of cars entering the section. In this way, we can get the number of cars entering this section every 15 minutes, and take this value as the traffic flow for subsequent research.

SITE stands for the names of different road sections. The concept of time is defined by time and day. Time is recorded every 15 minutes. This shows that there is a fixed interval for data recording, which is in line with the definition of time series. C1 and other variables represent different models. Through the null value query and the establishment of line chart,

it is found that except C1, there are a large number of null values in other variables, and the trend of line chart is not obvious. Therefore, C1 was selected for subsequent analysis.

Through the visualization of line graph, it is found that there are two days of traffic data very irregular. The rush hour comes later than usual, but it lasts for a short time. After looking up the calendar, we found that it was March 14 and 15 of 2015, which were weekends. Weekend traffic flow data has great uncertainty, so in this step, the weekend data will be deleted and only the data of five working days will be retained. Then draw a line chart again to see their relationship(figure 4).

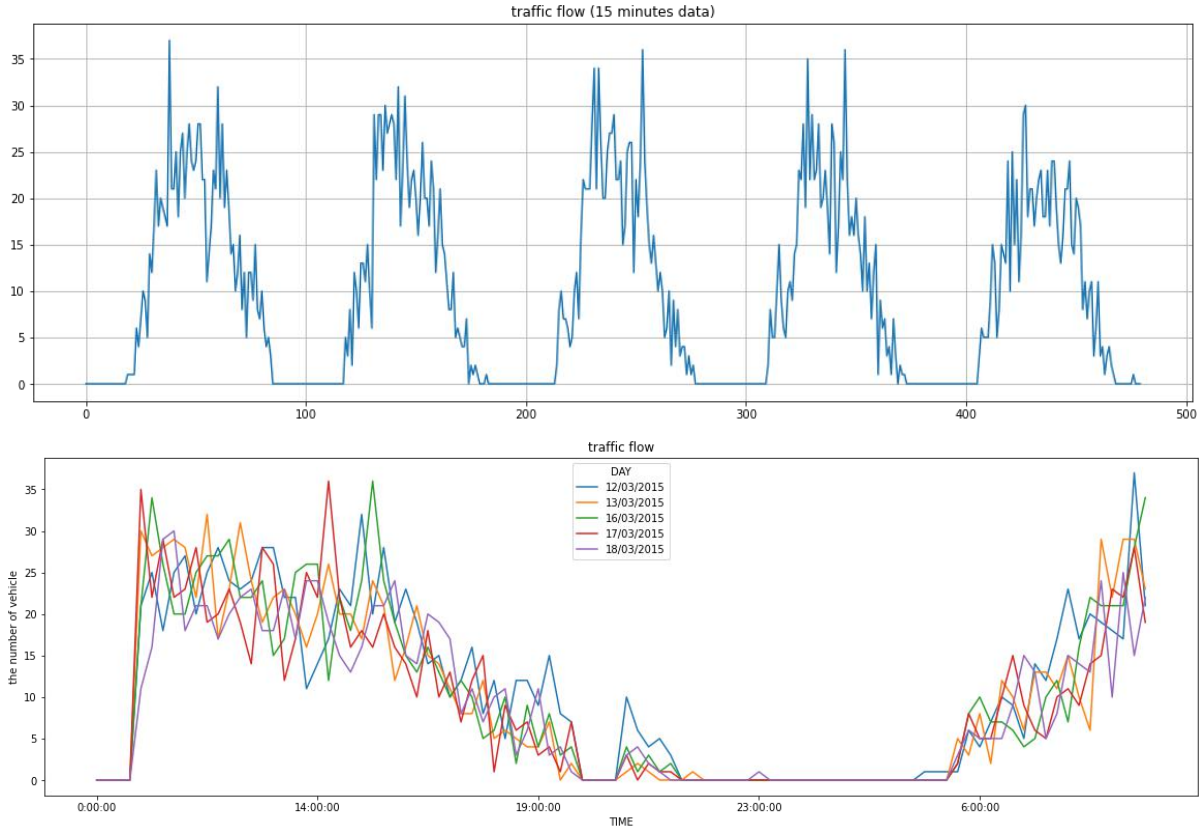

Figure 4: The visualization of data

Then, through visualization of the data, it is found that there is a certain rule in the traffic flow change, which can be considered to use time series for modelling. The line graph shows the change of traffic flow within 24 hours, and different lines represents different dates. Therefore, it can be seen that the traffic volume in the daytime will be much larger than that in the evening, and there is a rule. For example, in the morning and evening peak, traffic flow will be relatively large, it lead to the line presents a concave curve in the middle. This is because the traffic flow in this area will be lower than that on both sides in the middle of the two peak periods.

4.2 ARIMA

4.2.1 Difference

Through the difference() function to get a stationary time series, the following figure 5 is about the original data and the first-order difference, second-order difference image. Through

observation, the data is stable in the first-order(second) difference. In addition, choosing a higher difference sequence will lead to more data loss. Therefore, the first-order difference is chosen.

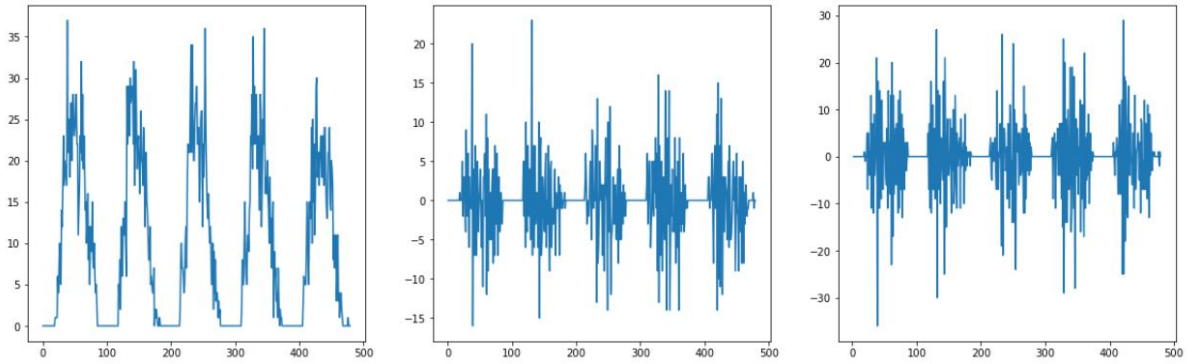

Figure 5: The results after difference

4.2.2 AR and MA

Autoregressive model is to describe the relationship between current value and historical value on the premise of ensuring data stability. The fundamental principle is to use historical time data to predict their own future. The P and Q values of the model were determined by observing the autocorrelation function and partial autocorrelation function.

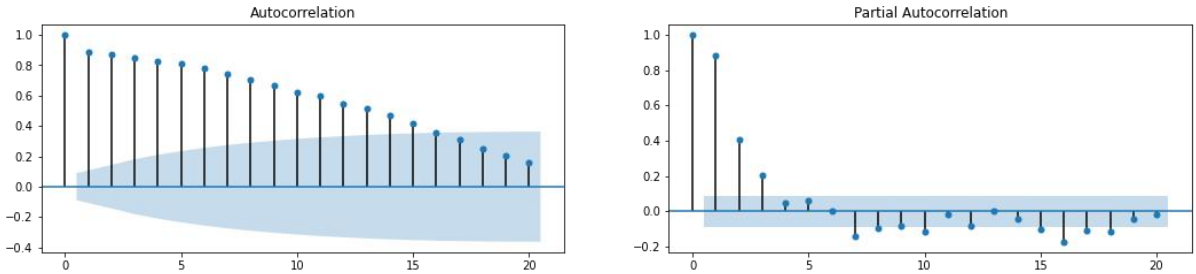

Figure 6: ACF and PACF diagrams

In the figure 6 of ACF, it is a first-order tailing and slowly decays to zero. It is obvious that there is a fast decaying trend between 0 and 1, and then the trend slows down. Observing the PACF graph, we find that it is truncated and decrease rapidly when the value is two, and oscillates around 0. The parameters which is confirmed of the model are as follows:

$$ARIMA(2,1,1)$$

After modelling, use the plot_diagnostics() function to observe the performance of the model. The results are shown in the figure below. When selecting appropriate parameters, the values of AIC and BIC are selected as measurement standards. In this link, the logical structure of the model will not be considered again (in order to control variables), and only the parameters will be tested continuously. For example, in the graphs of ACF and PACF, there are fuzzy values of P and Q. Therefore, in this optimization process, I have selected ARIMA (1,1,1) and other parameters for modeling. However, in the final comparison, it is found that the ARIMA(2,1,1) model is the most appropriate. In a word, the prediction accuracy of the model obtained by modeling with AIC and BIC criteria is not necessarily

very high (this can not be used as the final result to consider the quality of the model), but the model with low AIC and BIC values must be better than the model with high values. This only applies when selecting parameters.

ARIMA Model Results

Dep. Variable:	D.y	**No. Observations:**	479
Model:	ARIMA(2, 1, 1)	**Log Likelihood**	368.061
Method:	css-mle	**S.D. of innovations**	0.112
Date:	Tue, 27 Jul 2021	**AIC**	-726.122
Time:	21:23:10	**BIC**	-705.263
Sample:	1	**HQIC**	-717.922

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.601e-06	0.003	0.001	0.999	-0.005	0.005
ar.L1.D.y	-0.2271	0.145	-1.564	0.118	-0.512	0.057
ar.L2.D.y	-0.0857	0.083	-1.029	0.304	-0.249	0.078
ma.L1.D.y	-0.3329	0.141	-2.362	0.018	-0.609	-0.057

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.3254	-3.1486j	3.4162	-0.3134
AR.2	-1.3254	+3.1486j	3.4162	0.3134
MA.1	3.0043	+0.0000j	3.0043	0.0000

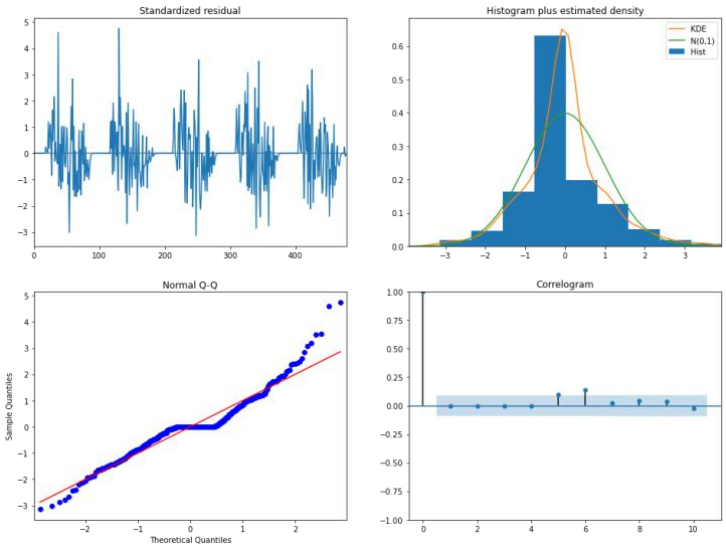

Figure 7: Evaluation of ARIMA(2,1,1) model

4.3 SARIMA

In fact, the traffic flow on the same road is regular. In the morning and evening peak of every day, as well as the low traffic in the early morning. These trends exist every day. Therefore, this trend can be regarded as a season, and this cycle should be once a day.

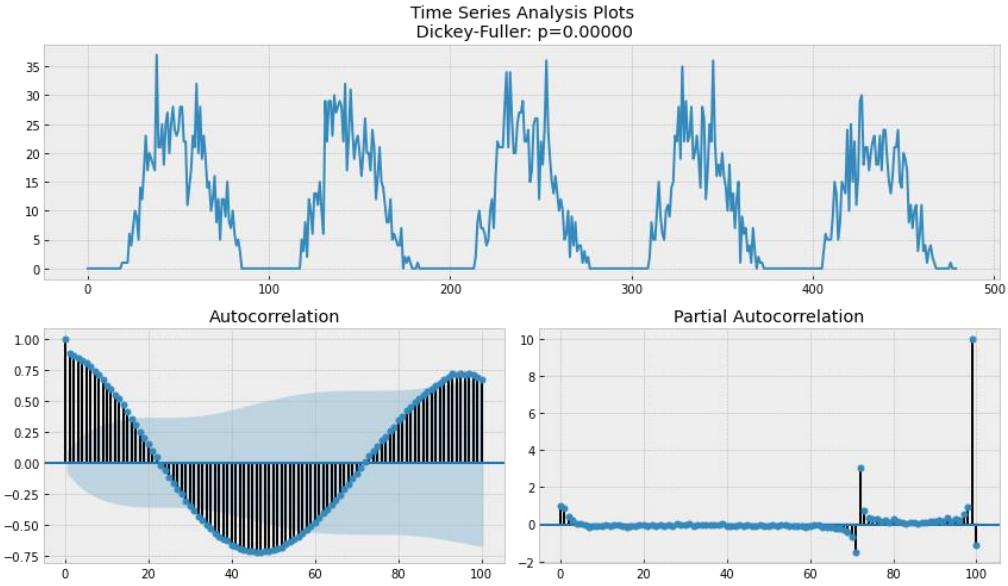

Figure 8: Observation of time lag and seasonality

The figure 8 is a time sequence diagram obtained by visualizing the time series with long lag. Through observation, we find that there is a periodicity. The periodicity is calculated as follows:

$$Seasonal = n \times t$$

n is the time interval of the time series statistics, that is, 4 Statistics per hour. t refers to the time, which is 24 hours when the season is one day.

$$Seasonal = 24 \times 4 = 96$$

Therefore, the value of seasonal is 96.

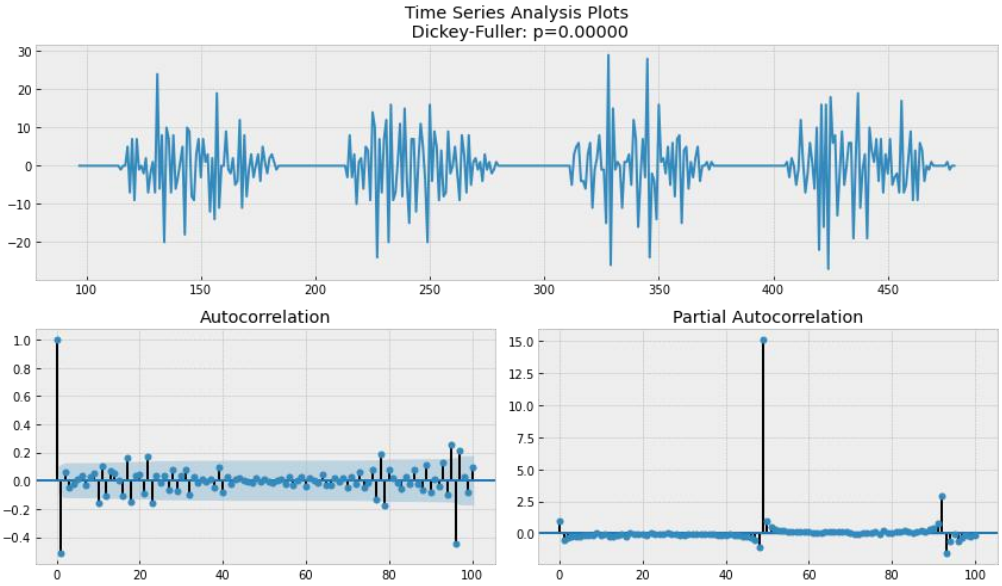

Figure 9: ACF and PACF charts with seasonality

The figure 9 is a visualization inspection after introducing the idea of seasonality. In general, the sequence presents a stable state. Of course, it may need first-order difference. After the comparison, the first-order difference is carried out on the basis of seasonality. This needs to be modelled later to compare their results. In addition, some information can also be obtained from the diagrams of ACF and PACF. First, the seasonality disappears, and even when the lag value is very high, there is still no law. Therefore, it is more certain that the sequence is stable.

Therefore, there is seasonality, and the value of S is determined to be 96.

By observing the above figure, it is found that ordinary p and q values become oscillatory attenuation after the first order. This is the same as the identification of non seasonal ARIMA. Just observe the ACF and PACF diagrams to get the appropriate p and q values. This has been explained in principle and applied in practice, so there is no more explanation ($p = 1, q = 1$). Because the first-order difference is made, it is determined that the value of d is 1.

Next, the SAR (P) and SMA (q) values are determined. According to the multiple principle of finding seasons, it is found that there are obvious changes in both graphs at order 96, that is, the original stable state has changed. In other words, there are changes in the first season, and the values of P and Q are determined as 1(Olsson and Soder; 2008).

In addition, the values of D are only 0 and 1. When using seasonality, it need to be set the value to 1.

Finally, the model parameters are established as follows for modelling.

$$SARIMA(1,1,1) \times (1,1,1,96)$$

The following are the final results of SARIMA, which are still measured and optimized by AIC and BIC criteria.

SARIMAX Results

Dep. Variable:	y	No. Observations:	480
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 96)	Log Likelihood	270.819
Date:	Wed, 28 Jul 2021	AIC	-531.637
Time:	00:09:04	BIC	-511.897
Sample:	0	HQIC	-523.807
	- 480		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0743	0.044	1.698	0.089	-0.011	0.160
ma.L1	-0.9980	0.088	-11.326	0.000	-1.171	-0.825
ar.S.L96	-0.0706	0.085	-0.827	0.408	-0.238	0.097
ma.S.L96	-0.9955	10.713	-0.093	0.926	-21.993	20.002
sigma2	0.0091	0.097	0.094	0.925	-0.181	0.200

Ljung-Box (Q):	43.89	Jarque-Bera (JB):	75.96
Prob(Q):	0.31	Prob(JB):	0.00
Heteroskedasticity (H):	1.18	Skew:	0.07
Prob(H) (two-sided):	0.35	Kurtosis:	5.18

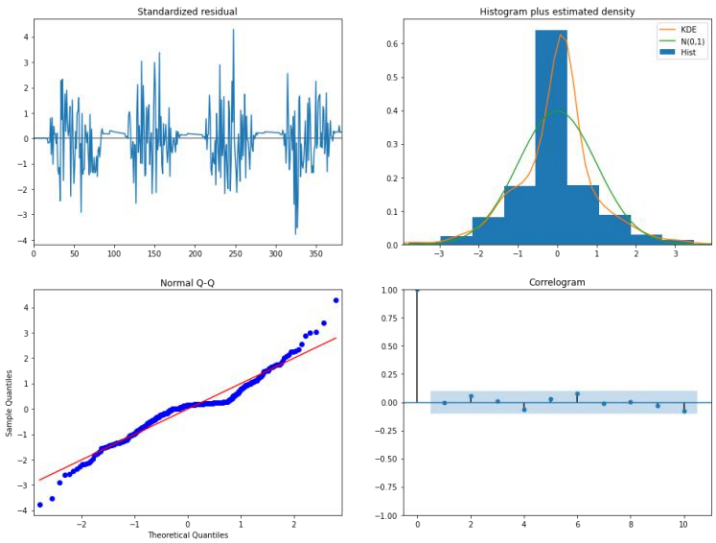

Figure 10: The evaluation of SARIMA

4.4 LSTM

4.4.1 Data Normalization

The standardization of data is conducive to improve the accuracy of the model and the convergence speed of the training model. The reason for adopting this process is that the complexity of LSTM model is too high, the amount of data is relatively large, and the computational power of computer is insufficient. Therefore, it may be important to implement normalization. Min-Max standardization is adopted in this paper. This is to change the original data so that the range of the data falls on [0,1].

$$x = \frac{x - min}{max - min}$$

4.4.2 Data Preparation

The standardized data is used to assign the data set and target value. Finally, it is divided when the feature dimension is 2 (look_back = 2). Make dataX as the data set and dataY as the target value.

The training set and test set are divided according to the ratio of 7:3. The len () function is called here to calculate the position of the data with a proportion of 0.7. Then, the data after this position is regarded as the test set, and the data before this position is regarded as the training set.

After the partition, the data needs to be converted into the input dimension and output dimension of LSTM. The LSTM algorithm encapsulated in the Pytoch package will be called soon, so the training set and test set need to be transformed into the dimension of the function. The reshape () function is called here. Adjust the training data to (-1, 1, 2) and the dimension of the target data to (-1, 1, 1).

4.4.3 Structure Design

The encapsulated LSTM code in torch.nn package is called, and initialization inherits this part. First, setting the time slice (number of features), which is determined by the previous look_back, so it is set to 2. The hidden layer defines a total of 6 layers, which is a parameter with high accuracy after testing. Of course, the value of this parameter is affected by the computing power. The higher value indicates a higher complexity of the model. At the same time, it will increase the difficulty of training and the training time. Two LSTM models are defined in series connection. The second model accepts the results of the first LSTM.

Next, the prediction results are transmitted to a linear layer. The dimension of the data received by this layer is 6, while the dimension of the output data is 1. Of course, this also involves the mismatch between input and output dimensions of different layers. Therefore, it also need to define a forward() function to convert dimensions. For example, convert the data from LSTM layer from 3D to 2D.

4.4.4 Parameter Optimization

The optim.Adam() function in the torch package is selected for model optimization. At the same time, the loss function, mean absolute error (MAE) and root mean square error (RMSE) are selected as the visualization of the training process. According to the figure 11, it can be found that with the increase of training times, the index values are getting lower and lower. It proves that the predict accuracy of the long short term memory model is getting better and better, with the increase of the epoch value.

Table 2: Training result on LSTM

Epoch	Loss	MAE	RMSE
100	0.01123	0.072	0.011
200	0.01087	0.073	0.011
300	0.00909	0.065	0.009
400	0.00826	0.060	0.008
500	0.00789	0.059	0.008
600	0.00773	0.058	0.008
700	0.00790	0.059	0.008
800	0.00708	0.055	0.007
900	0.00691	0.055	0.007
1000	0.00698	0.054	0.007

5 Evaluation

RMSE: It is the square root of the ratio of the square of the difference between the predicted value and the real value to the number of predictions. The full name is Root Mean Square Error. In this research, it is used to measure the difference between the predicted value and the real value of the final model, which can be regarded as the prediction accuracy. The formula is as follows:

$$\mathrm{RMSE}(X,h)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)})-y^{(i)})^2}$$

In this step, RMSE is used as the evaluation method for the three models mentioned above, and the performance capabilities of different algorithms in the time series are compared. The results are as follows:

Table 3: Evaluation and comparison of all algorithms

Algorithm name	ARIMA	SARIMA	LSTM
RMSE value	0.385	0.121	0.007

In fact, in this comparison process LSTM is normalized which lead to the result of RMSE is much smaller than that of the other two. After analysis, it is found that the ARIMA and SARIMA models are normalized again. This step also improves the prediction accuracy.

Therefore, the result which is predicted from the LSTM model is closest to the real value.

After analysis, it is found that the RMSE value of LSTM is the lowest. The result which is predicted from the LSTM model is closest to the real value. Therefore, LSTM is selected as the final prediction model. The model is applied to the test set and the visualization results are as follows:

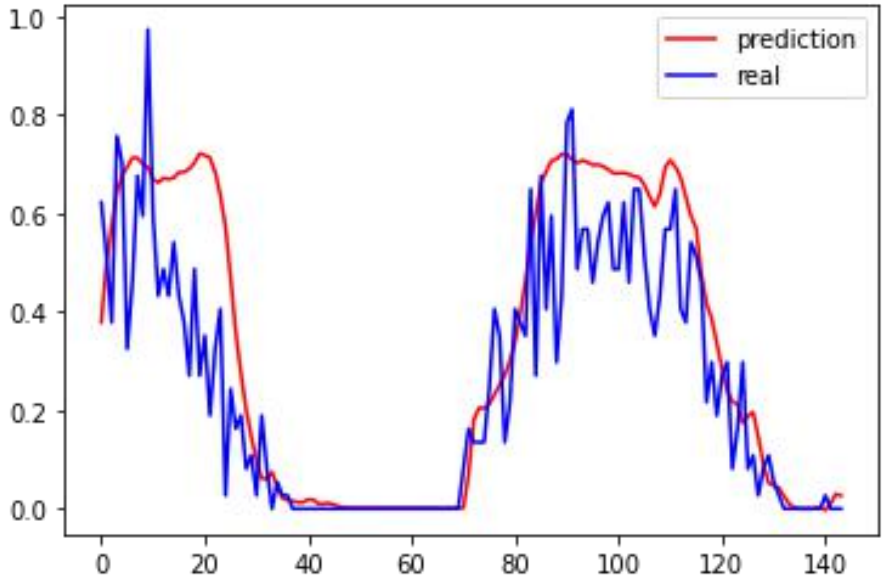

Figure 12: Evaluation and comparison of algorithms

Generally speaking, the prediction results are good, and the two obvious big fluctuations (trends) have been learned. However, there are many small fluctuations in the real data, or fluctuations rising and fluctuations falling, which do not perform well at the time of reaching these small peaks.

In addition, SARIMA performs better than non-seasonal ARIMA model. This is equivalent to answering the previous question. Introducing the concept of seasonality is feasible to improve the prediction accuracy, and the effect of this improvement is obvious.

6 Conclusion and Future Work

From the evaluation in the previous part, we can have a very clear answer to the research question. First of all, according to our conjecture, there is a certain law in the traffic flow in the city. It is not an upward trend of fluctuation, at least it does not exist in short-term data. In

short-term data, seasonal changes will be very obvious. And this is completely affected by the peak travel time of going to and from work every day. This is the same as we imagined at the beginning.

However, during the study, it was found that the situation on weekends became very different. Compared with the peak twice a day on weekdays, the peak on weekends became short and extreme. This is because the travel of citizens on weekends has great subjectivity, which is applicable both in time and space. This also caused huge trouble to the research. After consulting the relevant articles, it was found that many scholars had relevant puzzles, and pointed out the particularity of the weekend. Therefore, in order to improve the accuracy of the model, this part of irregular data is eliminated finally. In fact, from the value of research, it can also be found that the prediction value of weekend traffic flow is not high. At present, countries still focus on vigorously controlling and regulating the traffic on weekdays. Therefore, this operation is feasible in order to improve the prediction accuracy of traffic flow on weekdays.

In addition, when it comes to the comparison between SARIMA and ARIMA, the final evaluation result is that the seasonal ARIMA prediction result will be better. Of course, this result is based on the optimization of both models. The introduction of the seasonality concept makes the RMSE of the original model decrease significantly. Especially after normalization, the value of the original data becomes smaller, so the value of the evaluation result of RMSE becomes smaller. It can be considered that the gap between ARIMA and SARIMA is larger than it seems. This will more forcefully demonstrate that seasonality may be an important part of this subject. In other words, the daily punctual peak which is very similar to the seasonality has a great impact on the traffic flow.

In addition, in the literature review, it is found that LSTM performs well in short-term time series and can be used to predict it. This paper designs a model with two LSTM series connections and a linear fitting layer. Therefore, after implementation, LSTM can learn about various special situations (peaks) because of its high complexity of internal structure and strong memory ability. Therefore, in the final evaluation, it can be concluded that LSTM is the best among these models. Back to the initial question, the newer neural network model LSTM will perform better than the traditional classical ARIMA. However, looking at the broken line diagram between the final predicted value and the real value, it can be seen that LSTM did not learn this fluctuating trend (up or down) in the process of training. LSTM only gives a general trend, which makes the predicted value very close to the real value, while the broken line graph of the real value is not so smooth and has many peaks. Therefore, from this point of view, the LSTM model may also be optimized and improved.

In addition, the concept of seasonality may be used as a separate factor for LSTM model training. From the conclusion obtained above, seasonality is very strong. Therefore, if we can combine this idea in LSTM and enhance the seasonal weight, the trained model may perform better.

References

Boukerche, A., Tao, Y. and Sun, P. (2020). Artificial intelligence-based vehicular traffic

flow prediction methods for supporting intelligent transportation systems, Computer Networks 182: 107484.

Du, S., Li, T., Gong, X. and Horng, S. J. (2018). A hybrid method for traffic flow forecasting using multimodal deep learning, International Journal of Computational Intelligence Systems 13(1).

Johnston, J. and DiNardo, J. (1963). Econometric methods.

Li, B. and De Moor, B. (2002). Identification of influential observations on total least squares estimates, Linear algebra and its applications 348(1-3): 23–39.

Narmadha, S. and Vijayakumar, V. (2021). Spatio-temporal vehicle traffic flow prediction using multivariate cnn and lstm model, Materials Today: Proceedings .
URL: https://www.sciencedirect.com/science/article/pii/S2214785321031692

Olsson, M. and Soder, L. (2008). Modeling real-time balancing power market prices using combined sarima and markov processes, IEEE Transactions on Power Systems 23(2): 443–450.

Permanasari, A. E., Hidayah, I. and Bustoni, I. A. (2013). Sarima (seasonal arima) implementation on time series to forecast the number of malaria incidence, 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 203–207.

Qu, F., Wu, Z., Wang, F.-Y. and Cho, W. (2015). A security and privacy review of vanets, IEEE Transactions on Intelligent Transportation Systems 16(6): 2985–2996.

Sun, P. and Boukerche, A. (2020). Ai assisted data dissemination methods for supporting intelligent transportation systems, Internet Technology Letters 4(1).

Sun, P. and Samaan, N. (2020). A novel vanet-assisted traffic control for supporting vehicular cloud computing, IEEE Transactions on Intelligent Transportation Systems pp. 1–11.

X. Luo, D. Li, Y. Yang, S. Zhang, Spatiotemporal traffic flow prediction with KNN and LSTM, J. Adv. Transp. 2019 2019 1-10, Hindawi, 4145353.

Yang, H., Wang, J., Lin, Y., Zhai, G., Liu, X. and Tao, S. (2019). Effect of average car price on city-level private car ownership: A study based on panel data analysis, 2019 5th International Conference on Transportation Information and Safety (ICTIS), pp. 1196–1201.

Yanguo, H. (2015). Research on Traffic congestion mechanism and Traffic Control Method for Urban Road, PhD thesis, South China University of Technology.

Zeng, D., Xu, J., Gu, J., Liu, L. and Xu, G. (2008). Short term traffic flow prediction using hybrid arima and ann models, 2008 Workshop on Power Electronics and Intelligent Transportation System, pp. 621–625.

Zhao, D., Dai, Y. and Zhang, Z. (2012). Computational intelligence in urban traffic

signal control: A survey, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(4): 485–494.