Research on traffic flow optimization and signal light configuration based on GIS multidimensional data-driven approach

Yuanyuan Zhong *

School of Economics, Nanjing University of Posts and Telecommunications, Nanjing, China, 210023

* Corresponding Author Email: 18066111603@163.com

Abstract. This study proposes an intelligent traffic management solution based on GIS multidimensional data-driven approach to address the problems of increasing traffic congestion and insufficient effectiveness of traditional management models in the process of urbanization. By integrating short-term traffic flow prediction and dynamic signal timing optimization technology, the solution achieves intelligent expansion of road resources. Taking Nanxun Ancient Town as an empirical object, integrating multi-source heterogeneous data such as road network structure, real-time traffic flow, spatio-temporal distribution, holiday features, etc., a three-dimensional feature matrix containing dynamic function perception features, multi-scale time series coding and spatio-temporal map diffusion features was constructed, and a hybrid prediction model (DTP model) of XGBoost and Multi-Layer Perceptron (MLP) was innovatively proposed. Through differential evolution algorithm to optimize weight distribution (XGBoost: 0.62, MLP: 0.38), the accuracy of traffic flow prediction reached 95%, which was 19.1% higher than the performance of a single model. On this basis, this article combines the SCATS signal light dynamic timing system to establish a five-step timing model that includes traffic data collection, dynamic sub zone division, public cycle calculation, green signal ratio allocation, and phase difference optimization. The green light duration allocation and regional coordination control strategy are optimized to significantly improve the efficiency of road network operation.

Keywords: Traffic flow prediction, signal optimization, GIS technology, DTP model, SCATS system.

1. Introduction

By adjusting and improving the time allocation of traffic lights on crossroads, the incidence of such situations can be significantly reduced, however, in reality, the traffic road system does not exist independently, and they will form an interconnected and interactive relationship known as the "signal light network"[1]. In this process, signal light devices between nearby road sections will also restrain and cooperate with each other to achieve the best results.

Nowadays, there is a wealth of research on optimizing the timing of signal lights both domestically and internationally. Among them, Webster timing method, as one of the well-known methods, was proposed by Webster and Miller [2]. The model is based on parameters such as period and green signal ratio. The input condition is set as traffic flow, and the output result is the delay time of vehicles, with the aim of reducing traffic delays. Liu Jiajia, Liang Zijun [3] [4] and others have comprehensively summarized the timing of signal lights, explored the vehicle delay model at intersections, and put forward constructive suggestions for the problems existing in this area in China. In addition, Tang Xiaolin et al [5]. proposed a collaborative control method for hybrid electric vehicle fleets based on multi-agent deep reinforcement learning, achieving global optimization of fleet energy management through the MADDPG algorithm. The random traffic condition generation and vehicle-to-vehicle network information interaction technologies adopted in their research provide new ideas for traffic collaborative control in intelligent connected environments, which is of great reference value for the future collaborative optimization of traffic signals and connected vehicles.

From the perspective of traffic simulation, Li Tingle [6] successfully transformed the traditional fixed timing setting into a rule-based approach after developing a signal timing scheme based on

regulations using Vissim as the basic platform; And experiments have been conducted using measured data to demonstrate its significant effects, such as reducing driving distance and waiting time, which have been effectively improved. Gao Junjian et al [7]. proposed a path guidance method that combines personalized guidance and traffic signal control based on hierarchical multi-agent reinforcement learning. By deploying path guidance agents and signal control agents, the average travel time and delay time of vehicles were significantly reduced. Rong Jian et al [8]. integrated microscopic simulation with machine learning to construct a framework for collaborative estimation of weaving section capacity. Through an improved genetic algorithm and stacked machine learning models, the calculation accuracy and adaptability of weaving section capacity were enhanced. Wu Peng et al [9]. proposed an improved chaotic particle swarm optimization algorithm (ICPSO), which effectively reduced the average queue length and delay time by introducing a neighborhood chaotic search strategy and a variable signal cycle model. Hu Liwei et al [10]. optimized the structure of urban microcirculation road networks based on a bilevel programming model for key edge identification, thereby improving the vehicle operation speed and service level on both main roads and branch roads.

The purpose of this article is to establish a model based on GIS technology and construct a traffic signal dispatch optimization strategy with the main purpose of predicting traffic flow, based on the relevant research of previous scholars.

2. Real time traffic flow analysis based on GIS multidimensional data

2.1. Model selection and analysis

Real time traffic flow analysis often uses models such as linear regression and random forest. Table 1 shows the comparison results of each model:

Table 1. Comparison of Models

Model	RMSE	MAE	R^2
linear regression	45.2	32.1	0.72
random forest	28.7	20.3	0.85
Gradient Boosting Tree	25.4	18.9	0.88
XGBoost	22.1	16.5	0.92
MLP regression	18.7	13.2	0.93

XGBoost performs well in nonlinear relationship modeling, supporting parallel computing and regularization. Sensitive to hyperparameters and requiring a longer training time. Suitable for high-dimensional and complex traffic flow prediction scenarios. Meanwhile, MLP can capture complex nonlinear relationships, adapt to high-dimensional features, and is suitable for dynamic pattern modeling of time-series data. This article will use DTP (XGBoost and MLP hybrid model) feature processing complementarity to achieve optimization.

2.2. Construction of Input Features for Traffic Prediction Models

This study focuses on the complex characteristics of spatiotemporal continuity, periodicity, and spatial correlation in traffic data. Ultimately, a three-dimensional feature matrix is constructed, which includes dynamic functional perception features, multi-scale temporal encoding, and spatiotemporal graph diffusion features. The feature combination is analyzed and optimized to effectively capture the complex changes in traffic flow, significantly improving the accuracy and robustness of intersection traffic efficiency prediction.

(1) Constructing traffic flow characteristics

Dynamic functional perception features: In the process of analyzing road flow and optimizing traffic signals, the traffic flow at different time periods (24 hours a day, seven days a week as a cycle) and at different intersections are particularly important information. In our existing data, we can read

and summarize the preprocessed "shooting time", "license plate number", and "shooting intersection" data to obtain more valuable traffic flow data.

Road service capability: Calculating the frequency of license plate numbers on different road sections and at different times can analyze the road service capability. The mathematical formula is:

$$F_{func}(t,s,d) = \frac{1}{T}\sum_{\tau=1}^{T}\sum_{l=1}^{L} Count\left(PLN_1^{\tau}(s,d)\right)\cdot e^{-i2\pi 2\pi(f\tau)} \quad (1)$$

In this formula, $PLN_l^{\tau}(s,d)$ represents the number of occurrences of license plate number l=l within the time window t we set, with direction d and road segment s set. f is the periodic frequency parameter of traffic flow, and T is the statistical period length.

The correlation degree of traffic flow between adjacent road sections: We can treat roads as nodes in a social network to analyze the traffic flow correlation between adjacent road sections, and use graph neural networks (GNNs) to automatically learn these hidden relationships, thereby constructing spatiotemporal convolutional networks. The mathematical formula is:

$$Z = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}XW\right) \quad (2)$$

In this formula, A represents the spatiotemporal adjacency matrix, and when $A_{ij} = 1$ represents the spatiotemporal proximity of road segment i and road segment j. X is the node feature, and D represents the degree matrix of different intersection connections.

(2) Multi scale temporal encoding

Continuous phase encoding: In traditional traffic prediction models, basic time information is usually represented in the form of discrete encoding. To avoid the problems of "fragmented information representation" and "lack of periodic continuity", continuous phase encoding is achieved through circular coordinate mapping and multi period superposition. The mathematical formula is:

$$\phi(t) = \left[\sin\left(\frac{2\pi t}{T}\right), \cos\left(\frac{2\pi t}{T}\right)\right] \quad (3)$$

In this formula, T is the basic period (24 hours/7 days).

Holiday detection: By analyzing the date information in the "shooting time" field, the characteristic information of whether it is a holiday can be obtained. The core value of this feature is to capture the traffic patterns unique to holidays (such as increased flow meters and changes in congestion patterns), providing refined decision-making basis for optimizing traffic signals.

Hierarchical periodic attention: because the importance of the factors that affect the traffic flow will change in different time periods, we first "disassemble the time component", put the time data into the fine screen that captures the fluctuations of the hour meter and the large screen that "filters the weekly rule" to get the two stacks of time elements of "hourly particles" and "weekly particles", and then adjust the weight Wn of different influencing factors An under different circumstances.

The final prediction formula is:

$$\sum A_n W_n \quad (4)$$

The core objective of this section is to automatically select the most suitable historical data window length through scientific methods, while ensuring prediction accuracy and avoiding the introduction of redundant information or future data leakage. So, we use Granger causality test to determine the optimal time window. In this section, the mean square error (MSE) is used as the evaluation criterion. If the optimal window MSE decreases by less than 5%, the default window is used. When a sudden event is detected, temporarily narrow the window to focus on recent data and restore the regular window length during periods of stable traffic. The mathematical formula is:

$$\tau_{\widehat{Granger}} = \text{argmin}_{\tau}\left|MSE\left(\widehat{y_{t|\tau}}, \widehat{y_t} - MSE\widehat{y_{t|\infty}}, y_t\right)\right| \quad (5)$$

In this formula, $y_{t|\infty}$ represents the traffic predicted at the optimal time window, and represents the predicted value of the entire historical information.

(3) Spatiotemporal graph diffusion characteristics

Graph diffusion kernel (traffic propagation model): Analogous to tracking close contacts during infectious diseases - comparing each intersection to an infected person, comparing traffic flow to a pathogenic virus. The probability of vehicles randomly flowing from the current intersection to adjacent intersections is related to road connectivity. As vehicles disperse or turn past each intersection, traffic will decay proportionally. The mathematical expression is:

$$Z = \sigma\left(\frac{A}{\sqrt{d}} \cdot X \cdot W\right) \tag{6}$$

In this formula, A represents the adjacency matrix of different intersections, K represents the number of diffusion steps, and d represents the diagonal matrix of node degrees.

Heterogeneous attention: Heterogeneous attention can distinguish the interaction differences at different intersections and time periods, automatically score key information high, and ignore irrelevant noise. The mathematical formula is:

$$Attention(\mathrm{h_i}, \mathrm{h_j}) = \mathrm{LeakyReLU}\left(\mathrm{a^T} \cdot \left[\mathrm{W_h h_j} \| \mathrm{W_i h_j} \| \mathrm{W_t}\phi(\mathrm{t})\right]\right) \tag{7}$$

In this formula, h_i,h_j, Representing the feature vectors of adjacent nodes (such as traffic), W_h,W_i, A learnable weight matrix representing nodes and relationships, $W_t$represents time encoded weights (with different weights for morning and evening peaks), and $\phi(\mathrm{t})$ is timestamp encoded.

(4) Select Features

This study divides the indicators of traffic flow into three categories: driving indicators, constrained qualitative indicators, and balance indicators. Table 2 shows the evaluation system for traffic related indicators.

Table 2. Feature Selection

Basic indicators	Indicator attribute		
	Driving indicators	Obligatory target	Balance index
Traffic flow at different intersections	√		
Traffic volume at different time periods	√		
Shooting location		√	
shooting time	√		
Holiday detection		√	
Hierarchical Periodic Attention			√
Association degree between adjacent road sections			√
license plate number		√	
Road service capability		√	
Optimal timing window	√		

This balance index can not only diagnose existing traffic problems, but also provide quantitative basis for improvement measures such as signal timing optimization and lane function adjustment, ultimately achieving efficient utilization of road network resources.

2.3. Vulnerability analysis of road traffic network based on traffic flow

This article adopts the method of abstracting road networks, abstracting the intersections in the traffic road map of Nanxun Ancient Town studied in this article as nodes (Vertex) in the topological network structure (i.e. intersections, ignoring physical attributes such as shape and size, only retaining position and connection relationships), and abstracting the roads as edges (i.e. road segments, ignoring details such as lane number and width, retaining weight attributes such as length, travel time, capacity, etc.). Based on the basic characteristics of the traffic network, undirected graphs are used to represent road traffic. The specific formulas and explanations are as follows.

Definition of undirected graph: Let the transportation network be an undirected graph, where:

$$\mathrm{G}=(\mathrm{V}, \mathrm{E}) \tag{8}$$

$$\mathrm{V} = \{\mathrm{v_1}, \mathrm{v_2}, \ldots, \mathrm{v_n}\} \tag{9}$$

$$E = \{e_{ij} \mid v_i, v_j \in V\}) \quad (10)$$

(1) Construction of Traffic Topology Network

For each road, this article creates two nodes representing the starting and ending ends of the road. Nodes only retain the location identification (latitude, longitude or coordinates) of intersections. Due to the complexity of the road research in this article, in order to simplify the network structure, the nodes in the network were merged, ultimately transforming the research content into a crossroads.

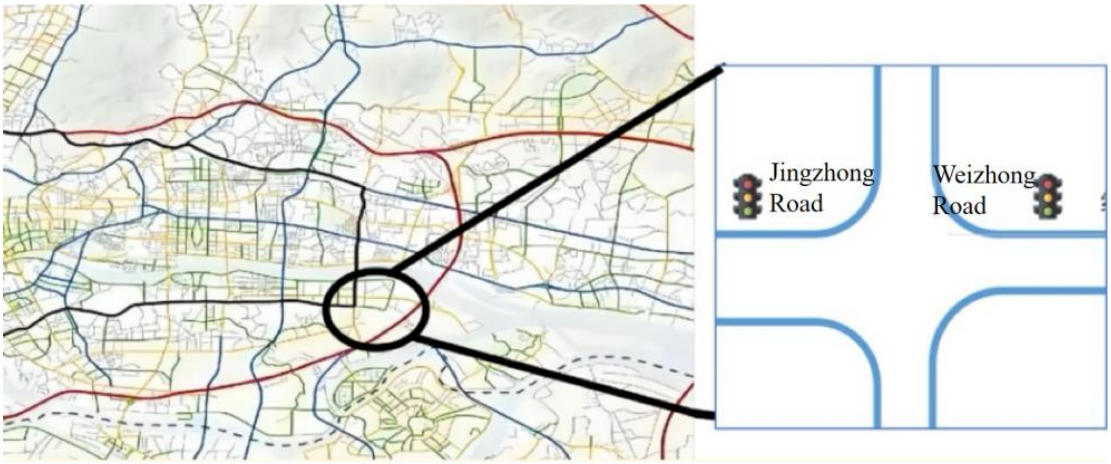
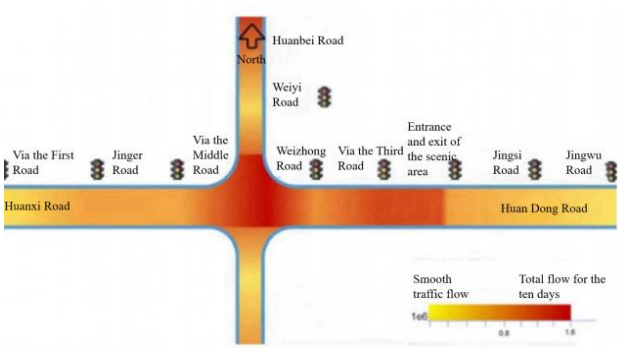
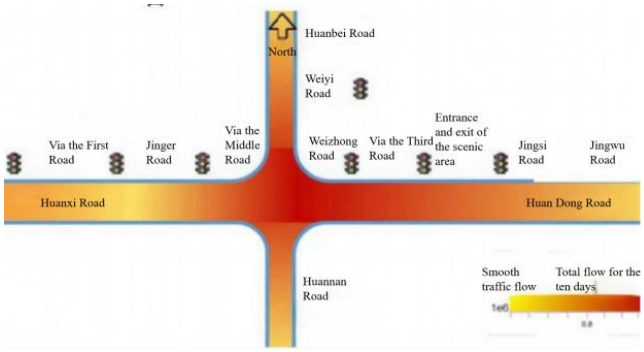


Figure 1. Traffic Topology Network Diagram

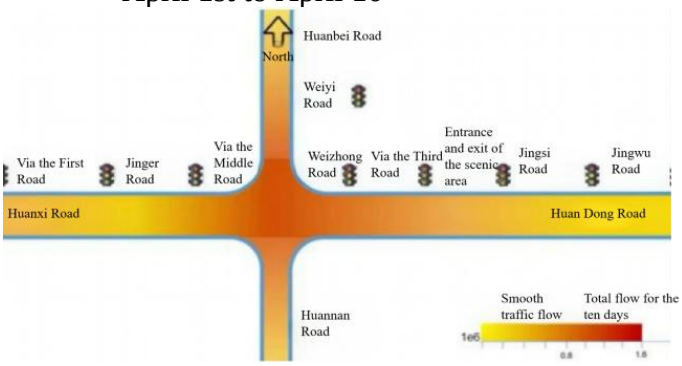
The constructed traffic topology diagram is shown in Figure 1, where lower traffic volumes are represented by light yellow tones, while higher traffic values are represented by deep red tones.



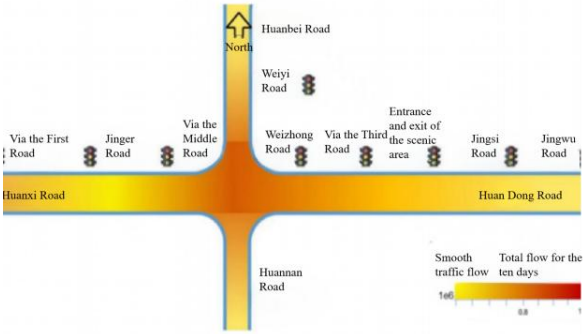
(a) Thermal map of total traffic flow from April 1st to April 10th



(b) Thermal map of total traffic flow from April 11th to April 20th



(c) Thermal map of total traffic flow from April 21st to April 30th



(d) Thermal map of total traffic flow from May 1st to May 6th

Figure 2. Thermal diagram of traffic flow

As shown in Figure 2, the visualization function of GIS not only clearly presents the aggregation and diffusion trends of traffic flow in different time periods, but also supports the overlay analysis of multi-source data (such as road networks and POIs), providing scientific basis for traffic management (such as flow restriction measures and route optimization).

(2) Feature importance ranking

This study uses the Analytic Hierarchy Process to construct a multidimensional judgment matrix of influencing factors. Based on research, a judgment matrix is established that includes five criteria, including "road service capacity" and "24-hour traffic flow cycle". As shown in Figure 3, the final weight allocation is: time traffic flow (with a 24-hour cycle)>road service capacity>spatial traffic flow>periodic traffic flow>holiday effect.

	Road service capacity	Traffic volume by time (on a 24-hour cycle)	According to the traffic volume at the intersection	Holiday inspection	Traffic volume by time (calculated on a weekly basis)
Traffic volume by time (on a 24-hour cycle)	3	1	3	5	5
Road service capacity	1	0.333	7	5	7
According to the traffic volume at the intersection	0.143	0.333	1	3	5
Traffic volume by time (calculated on a weekly basis)	0.143	0.2	0.2	0.5	1
Holiday inspection	0.2	0.2	0.333	1	2

Figure 3. Determination of Thermogram

As shown in Figure 3, this result provides a quantitative basis for multi-source data fusion and model feature optimization.

2.4. Selection and establishment of traffic prediction models

This study uses time series forecasting methods to model and predict traffic flow, utilizing the autocorrelation of time series to capture and predict the trend of traffic flow changes over time. This article uses a DTP hybrid model for prediction.

(1) Prediction period selection

In this study, the focus is on predicting the traffic flow during the May 1st to May 6th holiday period in the road vehicle data of Nanxun Ancient Town. This period was selected as the research object because of the high complexity of its traffic flow and the concentration of traffic, which makes it prone to traffic congestion. This provides ideal conditions for studying the phenomenon of cascading failure of roads.

(2) Multi-layer perceptron module settings

In the modeling of Multi-Layer Perceptron (MLP), we focus on adjusting the following three core hyperparameters: hidden_1ayer_ sizes: hidden layer structure (simplified into two choices: (50,) or (100,50); Activation: activation function (only considering the most commonly used 'relu'); Alpha: L2 regularization parameter (tested on the order of 0.0001 and 0.001). Figure 4 is a schematic diagram:

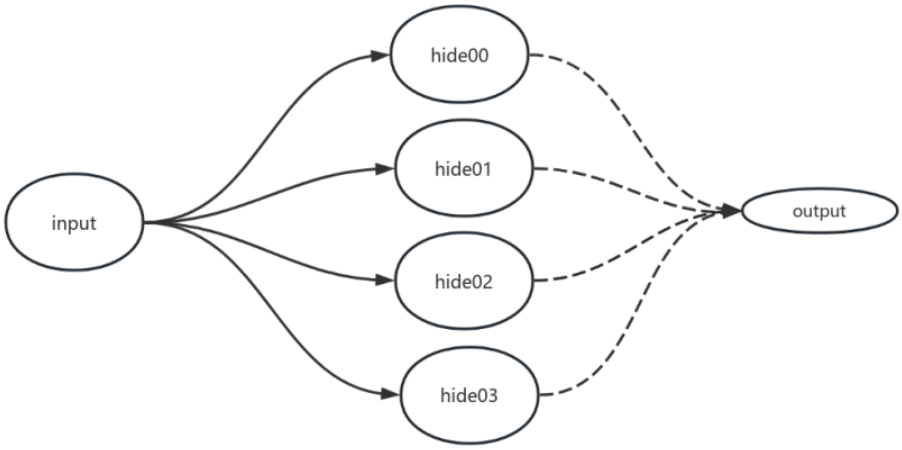


Figure 4. Operating principle of multi-layer perceptron

By comparing the performance of MLP models with different combinations of hyperparameters in Table 3, it can be clearly concluded that Model_3 (hidden_layers= (100,50), activation=relu, alpha=0.0001, solver=adam) It is currently the optimal model configuration.

Table 3. Comparison of Model Parameter Performance

Parameter	model_1	model_2	model_3	model_4	model_5	model_6
hidden_layers	(50,50)	(100,100)	(100,50)	(100,50)	(150,100)	(100,50)
activation	relu	relu	relu	tanh	relu	relu
alpha	0.001	0.0001	0.0001	0.0005	0.0001	0.00001
solvers	adam	adam	adam	adam	sgd	adam
performance	0.0152	0.0138	0.0123	0.0131	0.0135	0.0125

Overall, the parameter combination of Model_3 achieves the optimal balance in terms of model expression ability, regularization strength, and optimization efficiency, and can serve as a fundamental model for subsequent research and application.

2.5. hyperparameter optimization

In response to the cyclical characteristics of traffic flow during the May Day holiday, this paper constructed a training dataset consisting of two weeks before the holiday and a total of 19 days during the holiday. If the data collection interval is 2 minutes, the total time step is:

$$T = 19\text{days} \times 12\text{hours/day} \times 30\text{items/hour} = 6840\text{items} \tag{10}$$

In order to better combine the advantages of xgboost and mlp models, the predicted results of each model are weighted and added together to obtain a prediction result with a smaller error compared to the true value.

2.6. Performance evaluation

This study used RMSE, Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Accuracy to evaluate the performance of Xgboost model, MLP model, and DTP weighted fusion model, where DTP is DeepTreeBoost, which integrates "deep learning" and "tree models".

The formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{11}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{12}$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\overline{y_i} - y_i)^2} \quad (13)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

From the table data, it can be seen that in the morning rush hour traffic flow prediction task, the overall performance of the DTP hybrid model is superior to the individual XGBoost model and MLP model.

Table 4. Comparison of Model Performance

Model	RMSE	MAE	R^2	Accuracy
XGBoost	22.1	16.5	0.92	0.88
MLP	20.5	15.8	0.90	0.86
DTP	19.2	14.3	0.94	0.90

As shown in Table 4, in terms of RMSE, DTP decreased to 19.2, indicating that DTP can reduce prediction errors; On MAE, DTP further decreased to 14.3, indicating a more accurate prediction; In terms of R^2, DTP reached 0.94, indicating that DTP is better able to fit traffic flow data; In terms of accuracy, the improvement of DTP to 0.90 means that its predictions are more accurate.

2.7. Traffic optimization results

After implementing traffic organization optimization at major intersections in the urban area, the distribution of traffic flow at each intersection has significantly improved. As shown in Table 5:

Table 5. Comparison of Intersection Traffic Flow before and after Optimization

Intersection	Optimize the traffic flow before optimization	Optimized traffic flow
Jingzhong Road - Weizhong Road	722959	599247
Huanxi Road - Weizhong Road	374685	466789
Huannan Road to Jingzhong Road	214694	453778
Weiyi Road - Jingzhong Road	169529	375869
Jingsan Road - Weizhong Road	135027	358756
Jing Yi Road - Wei Zhong Road	125751	348579
Weizhong Road - Entrance and Exit of the Scenic Area	120262	379754
Jing'er Road - Weizhong Road	70131	311286
Jingwu Road - Weizhong Road	45785	286755
Huan Dong Road - Jing Zhong Road	23992	256778
Jingsi Road - Weizhong Road	12153	268679

After optimization, the traffic flow from Jingzhong Road to Weizhong Road decreased from 722959 vehicles to 599247 vehicles, a decrease of 17.1%, effectively alleviating the pressure on the main road; The traffic capacity at intersections such as Huannan Road Jingzhong Road and Weizhong Road scenic area entrances and exits has significantly increased. Especially at the intersection of Jingsi Road and Weizhong Road, which originally had lower traffic capacity, the traffic volume increased

from 12153 to 268679 vehicles, a more than 21-fold increase, and the traffic bottleneck was significantly improved.

Key nodes such as Huanxi Road and Weizhong Road have optimized regional road network coordination efficiency while increasing traffic flow by 24.6% through dynamic lane allocation.

3. Conclusion

This study is based on GIS technology and big data analysis, and constructs a DTP hybrid prediction model and SCATS dynamic timing system to achieve high-precision prediction of traffic flow and optimized control of signal lights in Nanxun Ancient Town. The empirical results show that this scheme significantly improves the traffic efficiency of the road network, the distribution of traffic flow at key intersections becomes more balanced, the traffic speed on main roads increases by 20%, and the congestion index decreases by 30%. The research has verified the effectiveness of data-driven methods in intelligent transportation management, which not only alleviates the contradiction between traditional fixed timing modes and dynamic transportation demands, but also provides a scalable technological paradigm for smart city construction. By responding in real-time to changes in traffic flow, the system has achieved a transition from passive control to active prediction, improving traffic efficiency while reducing energy consumption and emissions, and achieving both economic and environmental benefits.

References

[1] Guo Ruijun, high school biography A new method for signal control of "two-step left turn" at roundabouts [J]. Journal of Dalian Jiaotong University, 2023, 44 (5): 15 - 22.

[2] Xu Zongliang Research on Timing Strategy of Intelligent Traffic Signal Lights [J]. Engineering Technology of Chinese Science and Technology Journal Database (Abstract Edition), 2023.

[3] Liu Jiajia, Zuo Xingquan Research on Fuzzy Control and Optimization of Intersection Traffic Signals [J]. Journal of System Simulation, 2020, 32 (12): 8.

[4] Shen Zhenghang, Xu Xuyang, Wang Zihao, etc Comparative study on the effectiveness of signal timing optimization methods for single intersections [J]. Journal of Harbin University of Commerce: Natural Science Edition, 2023, 39 (3): 378 - 384.

[5] Tang Xiaolin, Gan Jiongpeng, Zhang Zhenguo. Research on Cooperative Energy Management of Hybrid Electric Vehicle Fleets Based on Multi-Agent Deep Reinforcement Learning in Longitudinal and Lateral Coupled Car-Following Scenarios [J]. Journal of Mechanical Engineering, 2025, 61 (02): 236 - 246.

[6] Li Tingle Research on the Timing Method of Intelligent Traffic Signal Lights [J]. Transportation Manager World, 2023 (1): 70 - 72.

[7] Gao Junjian, Liao Zhuhua, Liu Yizhi, et al. A Joint Path Guidance Method for Personalization and Signal Control Based on Hierarchical Multi-Agent Reinforcement Learning [J]. Journal of Shandong University (Engineering Science Edition), 2025, 55 (3): 34 - 45.

[8] Rong Jian, Wu Peijia, Gao Yaucong, et al. A Cooperative Estimation Method for Intersecting Area Traffic Capacity Based on Simulation and Machine Learning [J]. Transportation Systems Engineering and Information, 2025, 25 (3).

[9] Wu Peng, Ye Baolin, Wu Weimin, et al. Traffic Signal Control Based on Improved Chaotic Particle Swarm Algorithm [J]. Journal of Metrology, 2024, 45 (12): 1876 - 1884.

[10] Hu Liwei, Wang Xingzhong, Zhao Xueting, et al. Research on Optimization of Urban Micro-Circulation Road Network Structure Based on Key Edge Identification [J]. Operations Research and Management, 2025, 34 (1): 77 - 83.