

# Enhancing traffic flow monitoring with machine learning integration on cloud data warehousing

Peiyuan Yang<sup>1a,5,\*</sup>, Zhou Chen<sup>1b</sup>, Guangze Su<sup>2</sup>, Han Lei<sup>3</sup>, Baoming Wang<sup>4</sup>

<sup>1a</sup>Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

<sup>1b</sup>Software Engineering, Zhejiang University, Hangzhou, China

<sup>2</sup>Information Studies, Trine University, Phoenix, USA

<sup>3</sup>Computer Science Engineering, Santa Clara University, Santa Clara, USA

<sup>4</sup>Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA

<sup>5</sup>williamypy@gmail.com

\*corresponding author

**Abstract.** With urbanization and rising vehicle ownership rates, global traffic congestion and accidents have become pressing issues. This paper delves into leveraging cloud data warehousing and machine learning to tackle traffic flow monitoring and prediction, along with their potential in intelligent transportation systems. Through comprehensive analysis and case studies, it highlights how modern technology can enhance urban traffic management and services. Initially, the paper underscores the importance of traffic monitoring and prediction, identifying shortcomings in traditional approaches and advocating for machine learning solutions. It then reviews traditional traffic monitoring methods and explores machine learning's role in traffic flow prediction, illustrating its widespread application and evolving trends. Subsequently, the paper delves into the pivotal role of data warehousing in machine learning, encompassing data integration, management, cleaning, and multidimensional analysis. Additionally, it discusses the significance of relation matrices in graph convolution and presents experimental designs and model results for traffic flow prediction. Finally, the paper summarizes research findings, emphasizing the significance and future prospects of machine learning model design and cloud data warehousing in intelligent transportation systems.

**Keywords:** Traffic flow monitoring, Machine learning, Cloud data warehousing, Intelligent transportation systems, Traffic prediction.

## 1. Introduction

With the acceleration of urbanization and the continuous improvement of per capita vehicle ownership rate, traffic flow monitoring and forecasting has become a key link in today's intelligent transportation system. In the existing transportation system, traffic congestion, frequent accidents and environmental pollution are increasingly prominent problems, both developed and developing countries are facing these challenges. Therefore, to address these challenges, many cities are beginning to explore more efficient traffic management strategies and smarter transportation services. Therefore, accurate

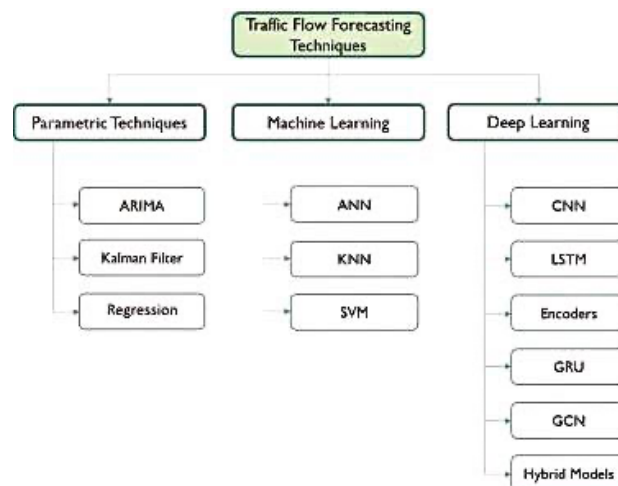
monitoring and forecasting of traffic flow is particularly important. [1] Real-time understanding of traffic conditions can not only help traffic management departments to formulate more flexible traffic control strategies, deploy traffic resources and guide traffic flow in advance, and reduce traffic congestion, but also provide scientific basis for traffic planning departments to carry out road reconstruction and traffic facility deployment.

Therefore, a machine learning-based approach has become an effective means to solve this challenge. [2-3] Machine learning technology can better deal with complex nonlinear problems, and can comprehensively consider the historical regularity of traffic flow data and the spatial correlation of road network, so as to improve the accuracy and performance of traffic flow prediction. This article will explore the potential of cloud data warehousing and machine learning in solving traffic flow monitoring and forecasting, as well as their application prospects in intelligent transportation systems. Through in-depth analysis and case studies of these technologies, we can gain a better understanding of how modern technology can be used to improve urban traffic management and services to address growing traffic challenges.

## 2. Related work

### 2.1. Traditional traffic flow monitoring

Traditional traffic flow monitoring usually adopts manual direct observation or video analysis. Due to Urban traffic flow can be analyzed from two dimensions of time and space through big data [3]. In terms of time, we can better grasp the periodic rules of traffic flow and the evolution characteristics of traffic flow under sudden situations by observing the changes of traffic flow over time from more than one year to as short as a second. In terms of space, this paper analyzes the route selection and starting and ending distribution of traffic in the flow process, and at the same time understands the quantity fluctuation of traffic flow in different sections, which enables researchers to examine urban traffic operation from a global perspective. [4] The system captures the traffic situation on the city's main roads through surveillance cameras, converts the video information into digital data, and records key information such as the driving path, speed and time of each vehicle. This data is stored in cloud data warehouses, forming a vast dataset of traffic flows. By analyzing these data, it is possible to understand the traffic conditions in different time periods and regions, identify congestion points and accident locations, and provide decision support for traffic management departments.



**Figure 1.** Development of traffic flow monitoring methods

Although the traditional traffic flow monitoring method adopts direct observation or video analysis, it has certain advantages. First of all, accurate traffic flow information [5] can be obtained through manual direct observation or video analysis. Secondly, due to the targeted and flexible way of manual

monitoring, a specific area can be monitored according to needs and the traffic situation in the area can be effectively mastered. In addition, the efficiency of manual data processing is low, which can not meet the rapidly changing traffic flow information processing needs. Therefore, it is necessary to introduce advanced technologies such as machine learning to make up for the shortcomings of traditional traffic flow monitoring methods and achieve a more comprehensive, accurate and efficient monitoring and analysis of urban traffic flow.

### *2.2. Machine learning and traffic flow monitoring*

Traffic flow prediction is an important part of intelligent transportation system and has been widely used in intelligent transportation subsystem. For example, Advanced Traveler Information Systems (ATIS), Advanced Traffic Management Systems (ATMS), etc. [6]. On the one hand, accurate traffic flow prediction can provide travelers with accurate road information, so as to effectively avoid congested sections and save travel time. On the other hand, the traffic management department can use the results of traffic flow prediction to conduct traffic guidance in advance to avoid too much congestion in a certain section. Therefore, in recent years, traffic flow forecasting has become one of the research hotspots in the field of traffic.

In recent years, with the improvement of computer computing power, the field of artificial intelligence and deep learning has been rapidly developed, and many applications have been produced in the field of traffic flow prediction. Common deep learning algorithms include deep residual network, recurrent neural network, convolutional neural network, etc. Example is China's traffic management system. In many Chinese cities, traffic surveillance cameras and on-board sensors are widely used on urban roads to monitor the movement of vehicles in real time. This data is aggregated to a central server, analyzed and processed to generate traffic flow heat maps and congestion warnings. [7-8] Through this information, traffic management departments can formulate traffic control measures, guide traffic flow, ease congestion and improve traffic efficiency. Therefore, the accumulation and analysis of real-time traffic monitoring data provides important support for urban traffic management, while the integration of machine learning and cloud data warehouses further improves the efficiency and accuracy of data processing and analysis.

### *2.3. Data warehousing and machine learning*

Data warehouses play an integral role in machine learning to predict traffic flow, and their importance is reflected in several ways. First of all, the data warehouse integrates various traffic data sources, such as traffic flow detection equipment, vehicle sensors, GPS [9] systems, etc., and gathers scattered data in one place to build a complete data set. This integration greatly improves the accessibility and availability of data, making it easier for analysts and decision makers to access the data they need for analysis and decision-making. Machine learning models based on data warehouses can be trained and optimized using historical data to make accurate predictions about future traffic flows. [10-11] This forecasting model based on historical data can not only help traffic management departments better understand the changing trend of traffic flow, but also provide scientific basis for traffic planning and decision-making, and improve the operating efficiency and service quality of urban traffic system.

The role of data warehouse in machine learning to predict traffic flow is multifaceted, including data integration and storage, data management and cleaning, data analysis and modeling. By making full use of the advantages of data warehouse, we can realize the comprehensive perception and accurate prediction of traffic flow data, and provide important support for the intelligent governance of urban traffic.

### *2.4. Relevant matrix*

The relation matrix is crucial to the convolution of graphs because it determines how node features are clustered and updated. In order to reduce the computational cost and improve the efficiency, we chose to include only the first 20 rows and 20 columns of data, which represent the first 20 most important features in the station. Through the comprehensive use of similarity matrix and adaptive matrix, we can

grasp the spatial dependence and dynamic change of traffic flow data more comprehensively, and provide more effective support for traffic management and optimization. The definition is as follows:

$$A_{ij}^0 = \exp(-\frac{\|x_i^5 - x_j^5\|^2}{\varepsilon^2}) \quad (1)$$

Adaptive matrix: Two learnable node embedding matrices E1 and E2 are randomly initialized, and the spatial dependence between the two embedding layers is calculated by multiplying E1 and E2, and the adaptive matrix is obtained:

$$\overline{A_{ada}} = ReLU(E_1 E_2^T) \quad (2)$$


$$\overline{A_{ada}} = norm(\overline{A_{ada}}) \quad (3)$$

Normalization of the adaptive matrix:

$$D = \sum_j A_{ada} \quad (4)$$

$$D_1 = \frac{1}{D} \quad (5)$$

$$\tilde{A} = D_1 \overline{A_{ada}} \quad (6)$$

 The definition of traffic flow prediction involves learning from historical traffic graph signal data with P time steps to develop a mapping function F. This function enables the prediction of future traffic graph signal data with Q time steps. In essence, the process entails analyzing the historical traffic data to understand patterns and trends, thereby enabling the accurate prediction of future traffic flow dynamics. This predictive capability is crucial for informing traffic management strategies and optimizing resource allocation in transportation systems:

$$X_{t-p+1:t} \in R^{p \times N \times d}, X_{t+1:t+Q} \in R^{Q \times N \times d} \quad (7)$$

### 3. Methodology

#### 3.1. Experimental design

The experiment utilized two datasets: NYCTaxi and NYCBike, which are stored in a data warehouse and contain valuable information about transportation in New York City during a specific time period.

**NYCTaxi:** This dataset comprises approximately 35 million records of taxi transactions in New York City spanning from April 1, 2016, to June 30, 2016. It includes detailed information such as pick-up and drop-off times, locations, travel distances, and other attributes. For the experiment, the data from April 1, 2016, to June 2, 2016, was utilized as the training set to train the models. The data from June 3, 2016, to June 16, 2016, served as the validation set for evaluating model performance during training. The remaining data was designated as the test set to assess the generalization ability of the trained models.

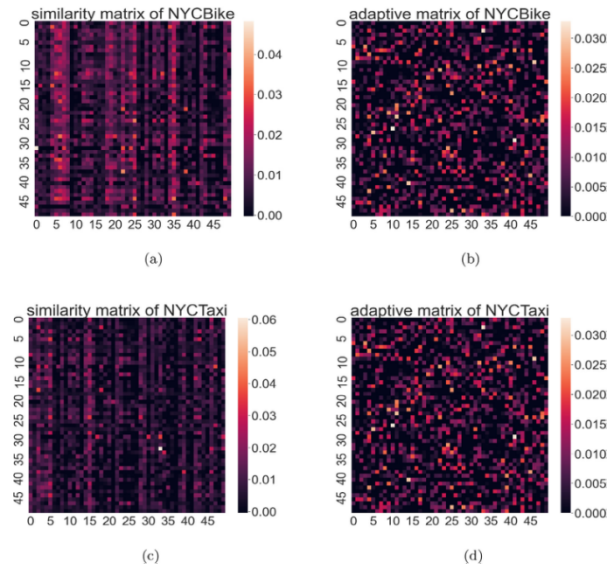
**NYCBike:** This dataset captures bike share orders in New York City over the same time period, containing information about pick-up and drop-off points, timestamps, and trip durations. By partitioning the datasets into training, validation, and test sets, researchers can effectively train machine learning models on historical data, validate their performance, and assess their generalization ability on unseen data. This structured approach to data management and experimentation is crucial for conducting rigorous and reproducible research in machine learning and data analysis.

As can be seen from Table 1, the proposed model has achieved the best performance and can accurately predict the future traffic flow of different traffic modes.

**Table 1.** The comparative results of different models on two datasets.

Method	NYCBike				NYCTaxi	
HA	MAE		PCC	RMSE	MAE	PCC
XGBoost	5.2003	3.4617	0.1669	29.7806	16.1509	0.6339
FC-LSTM	4.0494	2.469	0.4861	21.1994	11.6806	0.8077
DCRNN	3.8139	2.3026	0.5675	18.0708	10.22	0.8645
STGCN	3.2094	1.8954	0.7227	14.7626	8.4274	0.9122
STG2Seq	3.6042	2.7605	0.7316	22.6489	18.4551	0.9156
Graph WaveNet	3.9843	2.4976	0.5152	18.045	9.9415	0.865
CCRNN	3.2943	1.9911	0.7003	13.0729	8.1037	0.9322
MDRGCN(ours)	2.8383	1.7404	0.7934	9.5631	5.4979	0.9648
imporved	2.7393	1.6857	0.8077	8.7813	5.0758	0.9705
RMSE	3.48%	3.14%	1.80%	8.17%	7.67%	0.59%

In Figure a, notable patterns emerge, particularly in columns 6-9, 25, and 35, indicating a strong correlation between the traffic flow at these stations and that of others. Conversely, Figure b illustrates a dispersed color distribution in the heat map of the adaptive matrix.



**Figure 3.** Model training results

This variance signifies the adaptability of the matrix in capturing diverse traffic patterns and fluctuations, emphasizing its utility in accommodating the inherent randomness of traffic flow. By amalgamating insights from both relationship matrices, the study captures the interplay between regularity and randomness in inter-station traffic flow. Furthermore, this combined analysis enables the extraction of spatial dependencies inherent in traffic flow, offering a comprehensive understanding of traffic dynamics across different modes.

### 3.2. Research result

First, the machine learning perspective can be viewed in terms of model design, feature engineering, training process, and evaluation metrics. [12-13] The innovation of MDRGCN model lies in the introduction of multi-modal dynamic graph convolution module, which can effectively capture the flow characteristics of different traffic modes, and combine them through dynamic fusion module, while using gated cycle unit to realize the fusion of time and space dependence. The advantage of this model

design is that it can better handle the complexity and multimodal nature of traffic data. In addition, the ablation experiment verifies the validity of each module of the model, which further enhances the reliability of the model.

From a cloud data warehouse perspective, the study utilized publicly available datasets such as NYCTaxi and NYCBike to conduct experiments. These data sets are often stored in the cloud, where researchers can easily access and manage them through cloud data warehouses. At the same time, conducting experiments on the cloud platform can make full use of cloud computing resources, accelerate the speed of experiments and reduce costs.

#### 4. Conclusion

In conclusion, this paper has explored the integration of cloud data warehousing and machine learning techniques to address the challenges of traffic flow monitoring and prediction in urban areas. Traditional methods face limitations in handling the complexities of traffic data, prompting the adoption of machine learning solutions. Through extensive analysis and case studies, it has been demonstrated that machine learning, coupled with cloud data warehousing, offers a promising avenue for improving urban traffic management and services. Overall, the integration of cloud data warehousing and machine learning presents a viable solution for addressing the challenges of modern urban transportation systems.

Looking ahead, the future of intelligent transportation systems lies in further advancements in cloud data warehousing and machine learning technologies. Cloud data warehouses will continue to play a pivotal role in providing scalable and accessible data storage solutions for handling the vast amounts of transportation data generated daily. With the advent of edge computing and IoT devices, real-time data collection and analysis will become more efficient, enabling proactive traffic management and incident response. Additionally, the evolution of machine learning algorithms, particularly in the realm of deep learning and reinforcement learning, holds promise for even more accurate and adaptive traffic prediction models. These advancements will not only improve the efficiency of urban transportation systems but also pave the way for autonomous vehicle integration and smart city initiatives. Thus, by harnessing the synergies between cloud data warehousing and machine learning, we can expect significant advancements in intelligent transportation systems, leading to safer, more efficient, and environmentally sustainable urban mobility solutions.

#### References

- [1] May, Adolf Darlington. Traffic flow fundamentals. 1990.
- [2] Treiber, Martin, and Arne Kesting. "Traffic flow dynamics." *Traffic Flow Dynamics: Data, Models and Simulation*, Springer-Verlag Berlin Heidelberg (2013): 983-1000.
- [3] Tian B, Yao Q, Gu Y, et al. Video processing techniques for traffic flow monitoring: A survey[C]//2011 14th international IEEE conference on intelligent transportation systems (ITSC). IEEE, 2011: 1103-1108.
- [4] Williams, Nigel, Sebastian Zander, and Grenville Armitage. "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification." *ACM SIGCOMM Computer Communication Review* 36.5 (2006): 5-16.
- [5] Gui, G., Zhou, Z., Wang, J., Liu, F., & Sun, J. (2020). Machine learning aided air traffic flow analysis based on aviation big data. *IEEE Transactions on Vehicular Technology*, 69(5), 4817-4826.
- [6] Phan, Trung V., et al. "DeepGuard: Efficient anomaly detection in SDN with fine-grained traffic flow monitoring." *IEEE Transactions on Network and Service Management* 17.3 (2020): 1349-1362.
- [7] Lv, Yisheng, et al. "Traffic flow prediction with big data: A deep learning approach." *Ieee transactions on intelligent transportation systems* 16.2 (2014): 865-873.
- [8] Kumar, K., Parida, M., & Katiyar, V. K. (2015). Short term traffic flow prediction in heterogeneous condition using artificial neural network. *Transport*, 30(4), 397-405.

- [9] Wu, Y., Tan, H., Qin, L., Ran, B., & Jiang, Z. (2018). A hybrid deep learning based traffic flow prediction method and its understanding. *Transportation Research Part C: Emerging Technologies*, 90, 166-180.
- [10] Mitchell, Tom M. "Machine learning and data mining." *Communications of the ACM* 42.11 (1999): 30-36.
- [11] Kumar, Arun, Matthias Boehm, and Jun Yang. "Data management in machine learning: Challenges, techniques, and systems." *Proceedings of the 2017 ACM International Conference on Management of Data*. 2017.
- [12] Vartak, M., Subramanyam, H., Lee, W. E., Viswanathan, S., Husnoo, S., Madden, S., & Zaharia, M. (2016, June). ModelDB: a system for machine learning model management. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (pp. 1-3).
- [13] Polson, Nicholas G., and Vadim O. Sokolov. "Deep learning for short-term traffic flow prediction." *Transportation Research Part C: Emerging Technologies* 79 (2017): 1-17.