Dean&Francis
ISSN 2959-6157

Estimation of Urban Road Network OD Matrix Based on Graph Convolutional Network

Tianbo Wu*

Chang'an Dublin International College of Transportation at Chang'an University, Chang'an University, Xi'an, 710018, China
*Corresponding author: tianbo.wu@outlook.com

Abstract:

The Origin-Destination (OD) matrix is important in the intelligent traffic system, but most of the estimators do not predict the spatiotemporal dependence of traffic. In the current state of the art of deep learning, this study utilizes new neural network-based traffic forecasting methods, Graph Convolutional Networks (GCN) and temporal graph convolutional network (T-GCN). This paper uses taxi data from the New York City Yellow Taxi Trip Record in February 2025. Based on the prediction performance of both models, the trends obtained by these models fit the original data, which supports the application of the proposed models to passenger flow prediction. The GCN model is the best of the two available models as well as the best in terms of the spatial and temporal dependence, which shows that both modeling spatial and temporal dependence is necessary for accurate OD flow prediction. This provides a powerful data-driven tool in urban traffic analysis and provides information about travel demand patterns.

Keywords: GCN; T-GCN; OD Matrix Estimation

1. Introduction

Traffic flow behaviors in rapid urbanization and increasing urban transportation systems are crucial for better travel efficiency, congestion reduction, and intelligent transportation management [1]. A common tool is the origin–destination (OD) matrix, which describes the traffic demand distribution by estimating the number of trips from each origin zone to each destination area within a period [2]. Proper estimation of OD matrices can allow transportation designers and operators to analyze travel patterns, predict traffic conditions and devise effective control and optimization strategies [3-5]. However, accurate and timely OD matrices are hard to obtain in large-scale and dynamic urban networks where data are often incomplete.

Classical OD estimators (e.g., gravity models, entropy-maximization, statistical regression models, etc.) are often linear and independent, and do not capture the spatio-temporal relationships inherent to real-world traffic [6]. Such models tend to treat the urban areas as isolated zones, even if the traffic demand in one area is influenced by surrounding areas and network links. Moreover, they frequently rely on aggregated or survey-based data and lack the time-grained details required for real-time applications [7].

Hence, they fail in today‘s data-rich urban environments. With big data and deep learning technology, data-driven approaches have proven to be very useful in OD estimation. For example, the graph convolutional networks (GCN) have shown the capacity to capture spatial correlations of non-infinite structures such as road networks [8]. Whereas traditional convolution neural networks (CNN) are limited to grid data, GCN works on graph-structured data and can model relationships among connected nodes (e.g., road intersections, traffic zones) by their connectivity and feature similarity [9]. In the recent work, Temporal graph convolutions (T-GCN), which couple GCN with accelerated recurrent units (GRU), showed the state-of-the-art performance in capturing spatio-temporal correlations [10]. By combining GCN and temporal modeling such as recurrent neural networks or GRU, researchers can jointly model spatial and temporal relationships and achieve more accurate and robust OD predictions [11].

This research establishes an OD matrix estimation framework based on Graph Convolutional Neural Networks, using the New York City (NYC) taxi trip dataset as the research object. The NYC taxi dataset provides a rich and fine detail of millions of taxi trips, including pickup and drop-off time, locations, and distances. Such are the properties of the NYC taxi to analyze urban mobility dynamics and validate OD estimation models. By modeling the urban road network as a spatial graph where nodes correspond to traffic analysis zones (TAZs) and edges correspond to road links or proximity, the GCN model could learn how trips change over time and how they depend on network topology.

2. Data Description

2.1 Data Source

This study utilizes the New York City (NYC) Yellow Taxi Trip Record dataset published by the Taxi and Limousine Commission (TLC) [12]. The dataset provides detailed trip-level information, including pickup and drop-off timestamps, geographic zone identifiers (PULocationID, DOLocationID), trip distances, and passenger counts. This dataset selected for analysis corresponds to the entire month of February 2025, containing over ten million taxi trips across approximately 263 Taxi Zones. As shown in Table 1, the data are stored in CSV files, including the datetime and location of origin and destination.

Table 1. Attribute information for raw data

Field	Instruction	Data Type
Vendor ID	Taxi ID	Date
Pickup_datetime	Pick-up time	Date
Dropoff_datetime	Drop-off time	Date
PULocationID	Pick-up location ID	Integer
DOLocationID	Drop-off location ID	Integer
Fare_amount	Basic fare	Floating
Mta_tax	Traffic tax	Floating
Tip_amount	Tip fee	Floating
Total_amount	Total fee	Floating
Congestion_surcharge	Congestion surcharge	Floating
Airport_fee	Airport fee	Floating

2.2 Indicators Selection

The primary indicator for OD estimation is the trip flow volume between each pair of zones within each hour, which reflects the intensity of human mobility across the urban network. This study considers Vendor ID as passenger travel, PULocationID as origin and DOLocationID as destination. The OD heat map of raw data is shown in Figure 1.

ISSN 2959-6157

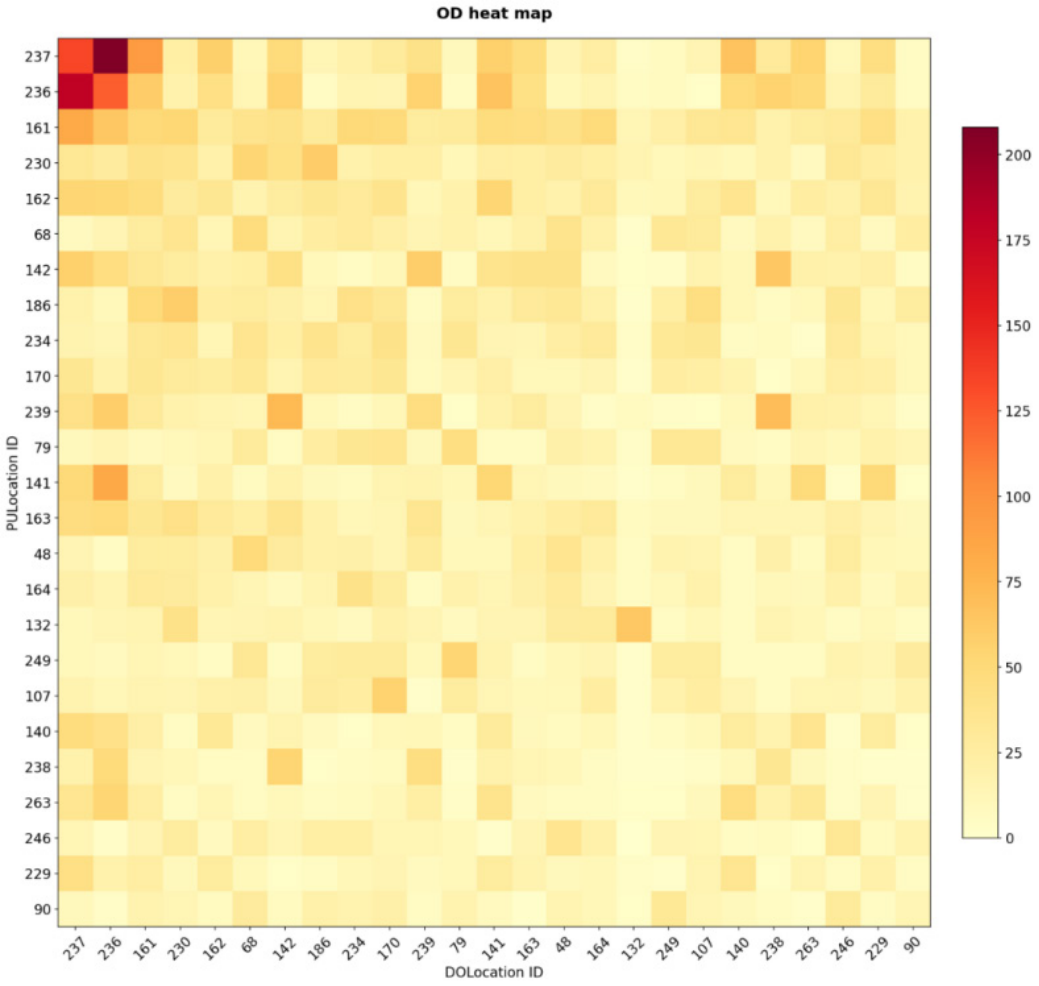

Fig. 1 Original OD heat map (Picture credit: Original)

This heat map shows the OD flows between different locations represented by PULocation ID and DOLocation ID. The color intensity, ranging from light yellow to dark red, corresponds to the truth flow volume in the raw dataset from 0 to more than 200.

Prior to modeling, the data are preprocessed by removing abnormal records (e.g., zero or negative distances), and all trip counts are normalized using min-max scaling to ensure stable neural network training. Missing or extremely sparse OD pairs are imputed with zero flow values to maintain a consistent matrix dimension.

3. Methods

3.1 GCN Model

To estimate and predict the dynamic OD matrix, this study applies GCN as a basic model and progresses to the T-GCN model based deep learning framework that combines both spatial dependencies and temporal evolution. The city road network is represented as a graph G = (V, E), where each node $v_{ij} \in$ V corresponds to a Taxi Zone and each edge $e_{ij} \in$ E represents the spatial adjacency between zones derived from the dataset. The adjacency matrix A encodes the connectivity of the graph, while the OD flow matrix X_t at time t provides node-level attributes describing trip volumes.

A two-layer GCN is used to capture spatial correlations among zones through spectral graph convolution:

$$H^{(l+1)} = \sigma(\sum_{k=1}^{K} D^{-\frac{1}{2}} A^{K} D^{K-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (1)$$

Where $A = A + I$ is the normalized adjacency matrix, $H^{(l)}$ is the input feature at layer l, and $W^{(l)}$ denotes learnable weights. The structure of GCN is shown as Figure 2. Assuming that node 1 is a central road. The blue nodes indicate the roads connected to the central road. This model obtains the spatial features by obtaining the topological relationship between road 1 and its surrounding roads. Af-

ter multiple convolutions, the model result is outputted.

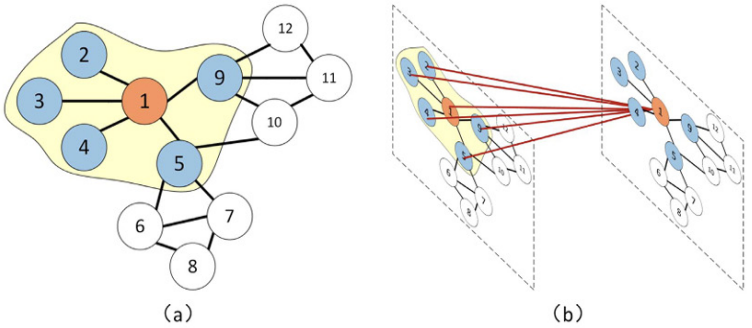

Fig. 2 Structure of GCN model [4]

3.2 T-GCN Model

To incorporate temporal dependencies, the output of the GCN is fed into a recurrent structure such as GRU. It can form a temporal GCN (T-GCN) model, which is a combined GCN and spatial and time data. This hybrid architecture enables the network to learn both the spatial correlations between zones and the temporal evolution of travel demand. The model is trained to minimize the mean squared error between predicted and actual OD flows. The overview of the T-GCN model is presented in Figure 3.

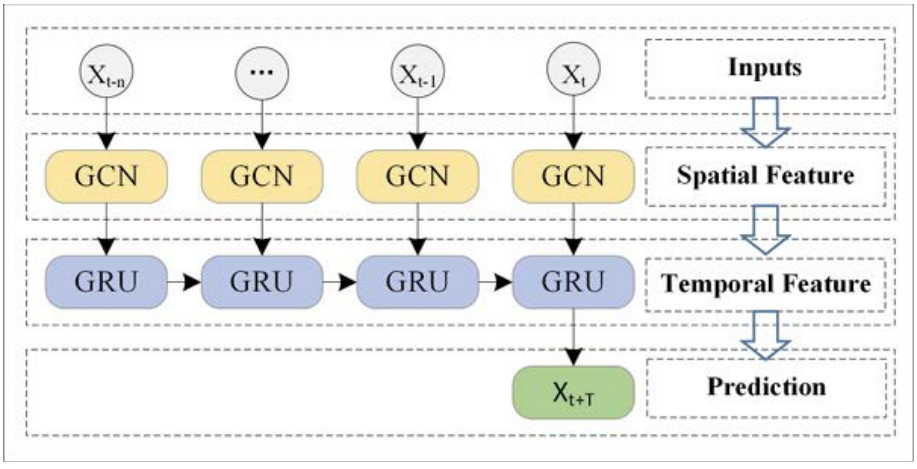

Fig.3 Architecture of T-GCN model [4]

3.3 Model Process

The complete modeling process is shown in Figure 4. After data cleaning and aggregating taxi trips into an hourly OD flow matrix, the urban network is represented as a graph where nodes correspond to taxi zones and edges denote spatial adjacency. For the GCN model, the current OD matrix serves as input to predict next-hour flows through graph convolution. The T-GCN model extends this by processing historical OD sequences, where the GCN captures spatial dependencies and the GRU models temporal evolution. Both models are trained to minimize prediction error, with outputs inverse-normalized to obtain final OD estimates. In addition, the predicted result will be visualized and compared with the original OD matrix. Finally, model performance will be evaluated through MAE, RMSE, and R^2 metrics.

ISSN 2959-6157

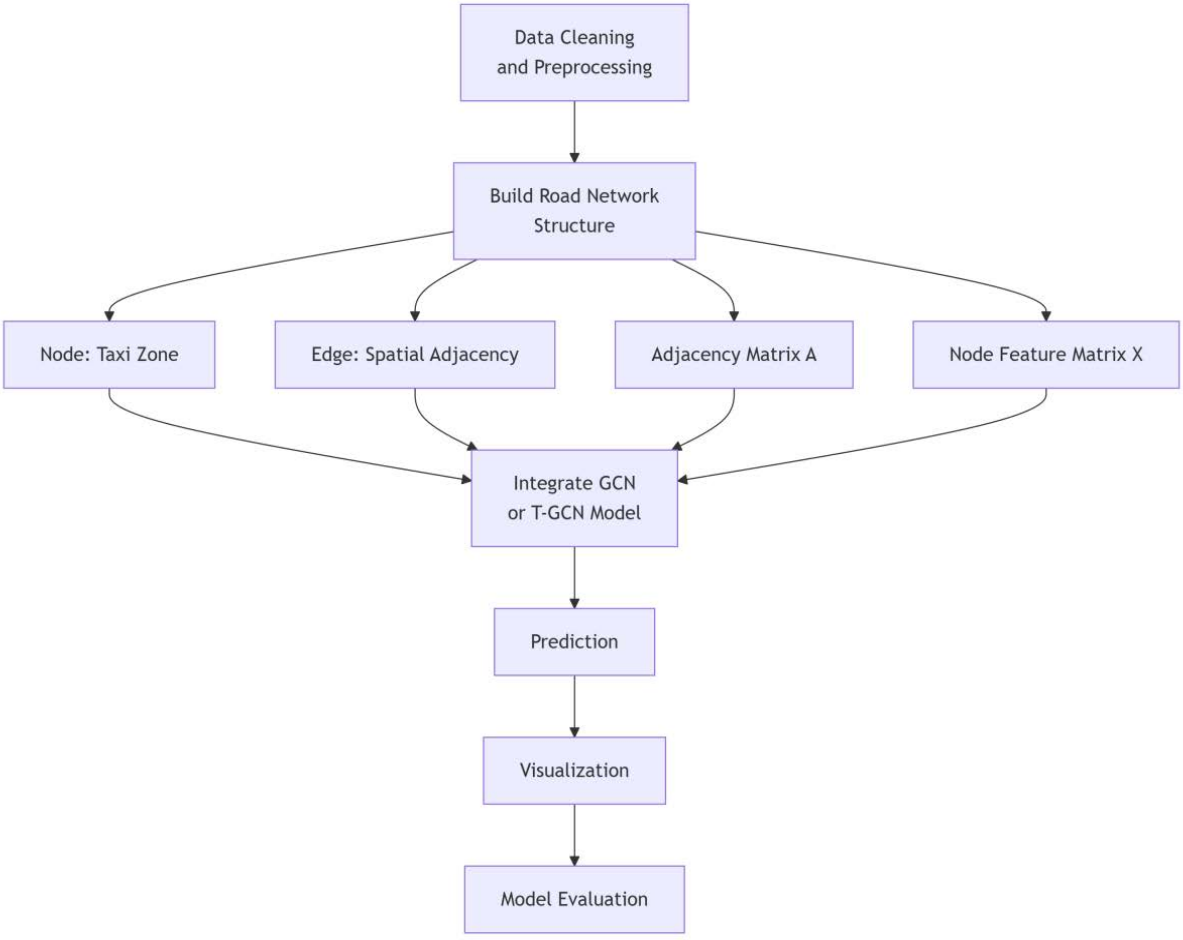

Fig. 4 Modeling process (Picture credit: Original)

4. Results and Discussion

4.1 GCN Model Result

The proposed GCN framework is implemented using PyTorch in Python and transformed into a heat map, as shown in Figure 5. The thermodynamic diagram indicates the estimation of OD flow through using the top 30 location IDs.

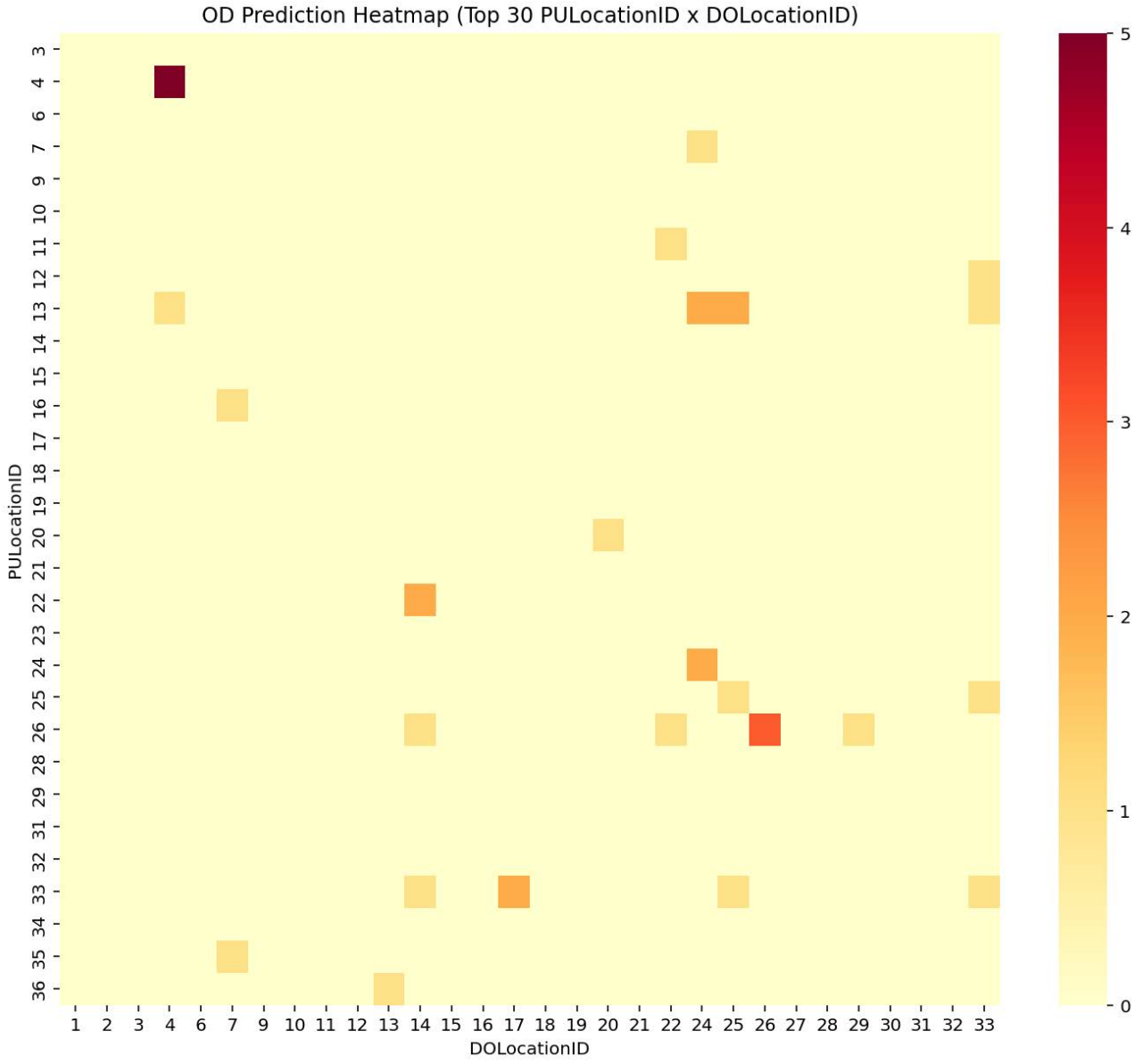

Fig. 5 GCN OD matrix (Picture credit: Original)

This heat map represents the predicted traffic flow intensity, which is divided into 5 levels between different zone pairs by using the spatial topology of the road network effectively, where the color intensity of each cell is proportional to the predicted trip volume from origin to destination. Through a qualitative comparison with the original truth OD heat map (Fig. 1), it shows a high consistency in the overall spatial pattern.

4.2 T-GCN Model Result

For the T-GCN model, after data cleaning, there are five districts within the OD network regardless of unknown districts. The distribution of these five districts is presented in Figure 6.

ISSN 2959-6157

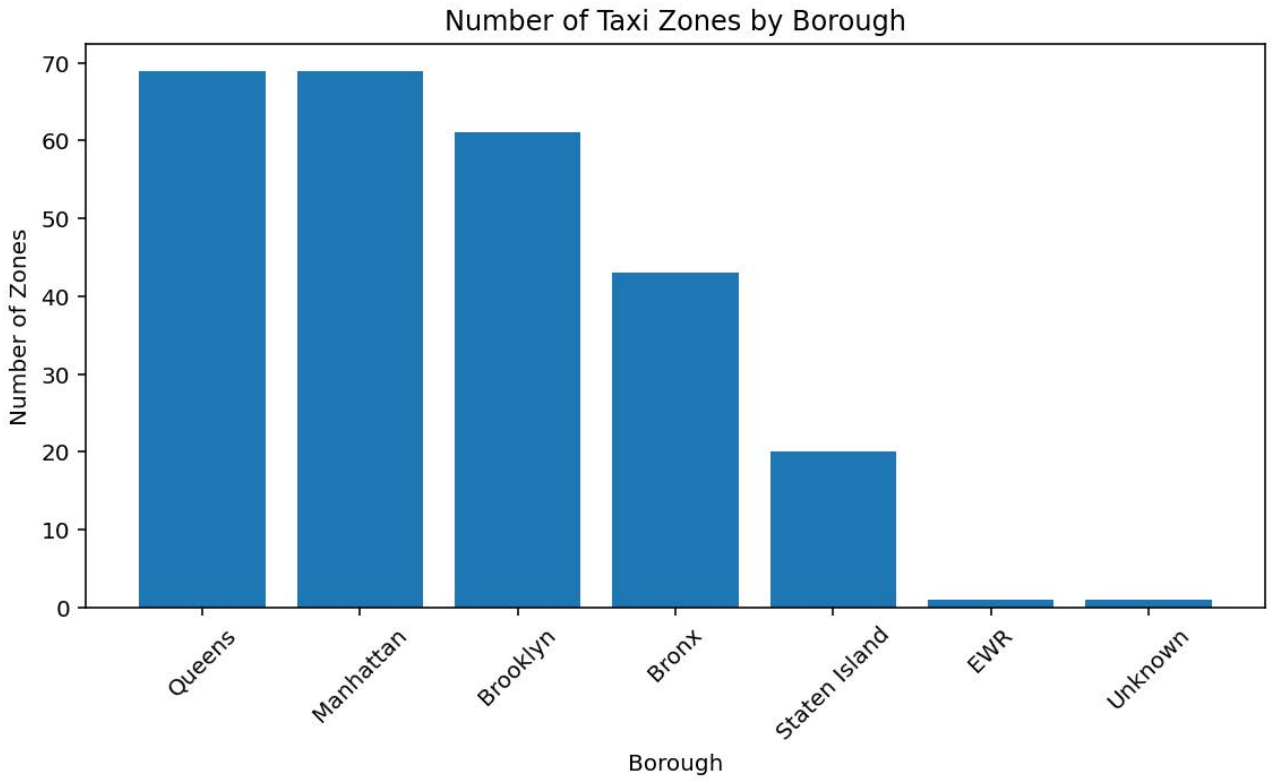

Fig. 6 Distribution of districts (Picture credit: Original)

Figure 7 below shows that morning periods (7-9 a.m.) exhibit strong flows from residential zones (particularly location ID 100-120 and 220-240) toward central business districts, while evening periods (5-7 p.m.) show the reverse pattern.

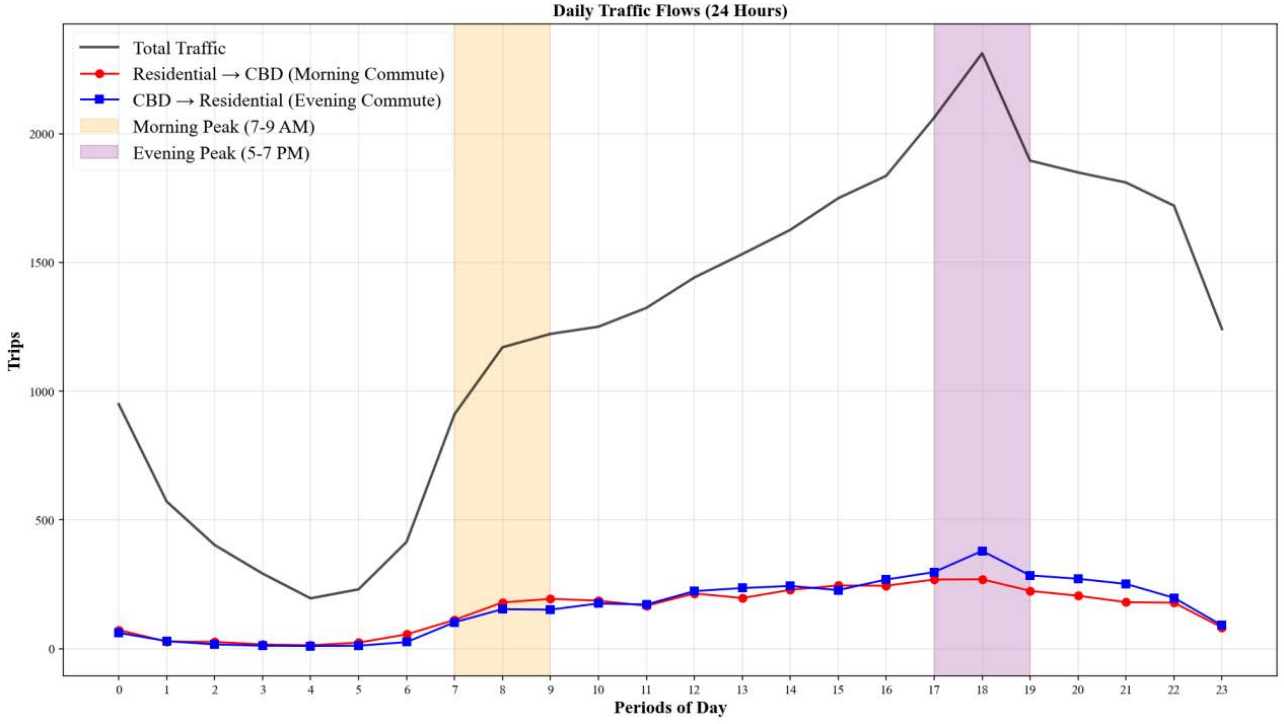

Fig. 7 Traffic flows in periods of the day (Picture credit: Original)

After multiple attempts, it is found that the best sequence length for temporal modeling to 10 hours (8 a.m. to 6 p.m.). Combining the spatial data and temporal data, the T-GCN prediction of OD flows is developed. To facilitate the presentation of the diagram, the results of T-GCN prediction are taken as logarithms as shown in Figure 8.

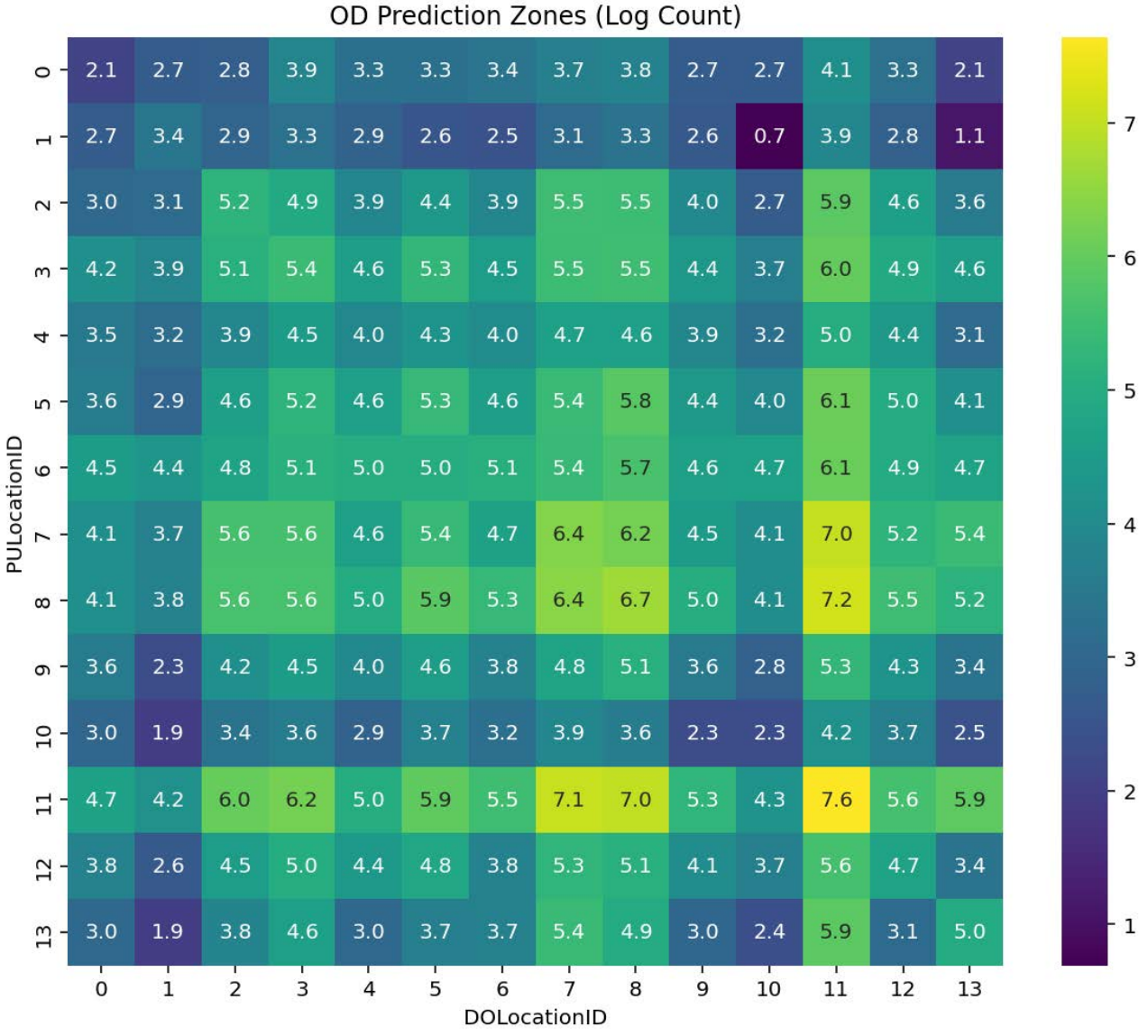

Fig. 8 T-GCN OD matrix (Picture credit: Original)

Each cell contains the logarithm of the predicted trip count for a specific OD pair. This result reveals a distinct spatial structure in the predicted demand. In certain zones, particularly those with higher indices such as row 11, column 11, with a log value of 7.6, serve as significant traffic attractors. It indicates hotspots of travel activity. Conversely, some zones exhibit markedly lower values. It signifies these zones are peripheral areas with sparse travel demand.

Model evaluation involves reviewing one or more existing models and evaluating their performance based on their category using various methods. To assess the accuracy and feasibility of the GCN and T-GCN model, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) are compared between these two models by using the following algorithms. The results are shown in Table 2 below:

ISSN 2959-6157

Table 2. Indicators for the evaluation of models

Model	MAE	RMSE	R^2
GCN	8.95	14.23	0.812
T-GCN	6.12	10.34	0.912

The R^2 values are close to each other, but T-GCN is still slightly higher than the GCN model, which indicates that the T-GCN model still has an excellent ability to elucidate variability compared to the GCN model. It reveals that both spatial and temporal components are crucial for accurate OD estimation.

4.3 Discussion

Experimental results reveal that the proposed T-GCN is significantly better than the single GCN model in predicting the urban road network OD matrix (e.g., smaller MAE and RMSE values and R^2) due to the fact that the T-GNN can capture the spatial as well as temporal relationship of traffic flow. While the GCN models the spatial relationship between the traffic analysis regions using graph convolution, the GCNN cannot model the dynamics of travel demand with time, and T-GCN learns a recurrent structure (GRU) modelling temporal sequences, which can reflect regular patterns (morning and evening peak flows).

5. Conclusion

This study selects Yellow Taxi Trip data from New York City, as people travel from source to destination. Through evaluating the OD matrix using the GCN model and T-GCN model, it shows that the performance of the proposed models is superior. T- GCN with MAE 6.12, RMSE 10.34, and R^2 0.912 shows good performance compared to pure GCN. This clearly shows that temporal dynamics are crucial for an accurate OD flow prediction. Visualization of prediction results, such as OD heatmaps and logarithm flow matrix, confirmed that the models were able to predict important travel patterns, such as morning or evening rush hour flows between residential and city districts.

Whereas this research has certain limitations. The models' performances may be influenced by external factors not included in the current framework, such as weather conditions, public events, and public traffic data. Based on these results, future work will focus on incorporating these multi-source data to enhance the model's stability and practicality. Other graph learning architectures could be explored in order to improve the estimation accuracy and interpretation.

References

[1] Wang Y, Xu Z, Zhao S, Zhao J, Fan Y. Performance degradation prediction of rolling bearing based on temporal graph convolutional neural network. Journal of mechanical science and technology, 2024, 38(8): 4019-4036.

[2] Xing X, Wang B, Ning X, Wang G, Tiwari P. Short-term OD flow prediction for urban rail transit control: A multi-graph spatiotemporal fusion approach. Information fusion, 2025, 118: 102950.

[3] Duan S, Huang P, Chen M, Wang T, Sun X, Chen M, Dong X, Jiang Z, Li D. Semi-supervised classification of fundus images combined with CNN and GCN. Journal of Applied Clinical Medical Physics, 2022, 23(12): e13746-n/a.

[4] Zhao L, Song Y, Zhang C, Liu Y, Wang P, Lin T, Deng M, Li H. T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction. IEEE transactions on intelligent transportation systems, 2020, 21(9): 3848-3858.

[5] Lu Z, Rao W, Wu Y, Guo L, Xia J. A Kalman filter approach to dynamic OD flow estimation for urban road networks using multi-sensor data. Journal of advanced transportation, 2015, 49(2): 210-227.

[6] Pamula T, Zochowska R. Estimation and prediction of the OD matrix in uncongested urban road network based on traffic flows using deep learning. Engineering applications of artificial intelligence, 2023, 117: 105550.

[7] Rong J, Xu W, Wen Y. A spatiotemporal model for urban taxi Origin–Destination prediction based on Multi-hop GCN and Hierarchical LSTM. Alexandria engineering journal, 2025, 128: 905-917.

[8] Sun M, Tian Y, Wang X, Huang X, Li Q, Li Z, Li J. Transport causality knowledge-guided GCN for propagated delay prediction in airport delay propagation networks. Expert systems with applications, 2024, 240: 122426.

[9] Sharma S, Mawane N, Kuraganti C K, M D G, Taware M, Dixit Y C, Mishra S, Krishnapuram R, Ramesh R. Enhanced ETA Predictions with T-GCN on Optimized Road Segments. IEEE, 2024: 1.

[10] Guo Y, Huang J, Jiang X. Time series prediction based on the variable weight combination of the T-GCN-Luong attention and GRU models. Scientific reports, 2025, 15(1): 21945-12.

[11] Wang X, Zhang Y, Zhang J. Large-Scale Origin–Destination Prediction for Urban Rail Transit Network Based on Graph Convolutional Neural Network. Sustainability, 2024, 16(23): 10190.

[12] New York City Taxi and Limousine Commission (TLC). Yellow Taxi Trip Records, 2025. Retrieved from https://www.nyc.gov