A novel untapped flight segment flow prediction framework based on graph deep learning and heuristic algorithm for sustainable transport development

Ms. Linlin Chen

Department of Management Science, School of Management,

Xiamen University, Xiamen, China

Email: 415949507@qq.com

Dr. Shuihua Han

Department of Management Science, School of Management,

Xiamen University, Xiamen, China

Email: hansh@xmu.edu.cn

Dr. Shivam Gupta

Department of Information Systems, Supply Chain Management & Decision Support,

NEOMA Business School,

59 Rue Pierre Taittinger, 51100 Reims, France

Email: shivam.gupta@neoma-bs.fr

Dr. Uthayasankar Sivarajah

School of Management, University of Bradford,

Richmond Road, Bradford, BD7 1DP, United Kingdom

Email: u.sivarajah@bradford.ac.uk

Dr. Fred A. Yamoah *

School of Business and Law, Buckinghamshire New University, United Kingdom

Email:

(* : Corresponding Author)

"This is an original manuscript of an article published by Taylor & Francis in [Journal of the Operational Research Society] on 9th December 2025, available at: https://doi.org/10.1080/01605682.2024.2433191."

A novel untapped flight segment flow prediction framework based on graph deep learning and heuristic algorithm for sustainable transport development

Abstract

This study explores a novel machine learning framework for predicting the flow of untapped flight segments, focusing on the unique challenges posed by the absence of historical flow data in airline networks. Utilizing a real-world datasets from a major airline, we evaluate the performance of a graph deep learning-based approach that combines Multi-Graph Attention Networks (MGAT) and Long Short-Term Memory (LSTM) networks, as well as Nondominated Sorting Genetic Algorithm II. The results demonstrate that the proposed framework significantly outperforms traditional models in accurately predicting passenger flow for new flight segments, particularly when compared to statistical benchmarks like time-series models that rely on historical flow data. Moreover, we find that optimizing the affinity coefficients within MGAT using the NSGA- II not only enhances predictive accuracy but also improves the interpretability of the model. Finally, we provide an in-depth analysis of the key factor that influence the predicted outcomes, highlighting the critical role of market competition in untapped segment operations.

Keywords: Flow Prediction; Heuristic Algorithm; Multi-Graph Attention Networks; Spatial-Temporal Dimension; Untapped Flight Segment

1 Introduction

In the aviation industry, predicting transportation flow is crucial for optimizing resource allocation and supporting sustainable development. Transportation flow in this paper refers to the passenger demand. Accurate flow forecasting, particularly for new flight segments, enables airlines to strategically open new routes and capture new markets (Suh & Ryerson, 2019). With the aviation sector rapidly recovering post-pandemic, passenger volumes have reached pre-pandemic levels, and the flow for air travel continues to rise. For example, according to America Airlines, passenger volumes in late 2023 exceeded those of 2019, highlighting the potential for new market opportunities (Suh & Ryerson, 2019).

Airlines frequently explore untapped routes, which are routes not previously operated by them but possibly served by competitors[1]. These routes present significant business opportunities, but many airlines still rely on flight tests and operator experience to make these decisions. Although it is easy to implement, these traditional approaches lack the precision required to accurately assess the profitability of new routes and inform effective scheduling decisions (J. Yu, 2021). This underscores the need for more accurate and data-driven flow forecasting tools.

Currently, there has been increasing interest in air traffic flow prediction (Cai, Shen, Luo, & Li, 2023; Yan, Yang, Wu, & Lin, 2023). Existing air traffic flow prediction methods are generally categorized into model-driven and data-driven approaches (Cai et al., 2023; Yan et al., 2023). Model-driven methods, such as gravity model (Birolini, Jacquillat, Cattaneo, & Antunes, 2021) and global vector autoregressive (GVAR) model (Gunter & Zekan, 2021), require defining numerous influencing

[1] https://airwaysmag.com/how-airlines-launch-new-routes/

factors, which complicates model construction and may limit accuracy. Data-driven approaches generally based on historical time series data (Gui, Zhou, Wang, Liu, & Sun, 2020; Z. Zhang et al., 2020). It have harnessed the power of machine learning to predict traffic flow (Lv, Lv, Cheng, Zhu, & Rashidi, 2023), including air traffic flow prediction (Xu, Pang, & Liu, 2023), which has the capacity to decipher intricate patterns and relationships, enabling more flexible modeling of complex air traffic system (Yan et al., 2023). However, for untapped flight segments, the lack of historical data presents a significant challenge, as it prevents the direct construction of time-series models traditionally used in these methods (Caserta & D'Angelo, 2024). To our knowledge, most recent studies only focus on the existing flight segment flow prediction (Gui et al., 2020; Z. Zhang et al., 2020), and there is little research that concentrates on predicting the flow of untapped flight segment.

To address this challenge, this study proposes a novel framework specifically designed to predict flow for untapped flight segments in the absence of historical data. Drawing inspiration from the work of Quan, Liu, Dou, Xiong, and Ge (2012), which utilized graph attention networks (GAT) for location optimization in the hospitality industry, we adapt this concept to the aviation context. We analogize the prediction of untapped flight segment flow to a recommendation problem, where GAT is employed as the foundational module of our framework. GAT excels in learning the importance of connections between nodes and integrating neighborhood information, making it suitable for capturing complex relationships within air traffic networks (S. Liu & Jiang, 2022; C. Zhang, Yu, & Liu, 2019).

Given that flight segments with similar characteristics but no direct connections may still share important passenger flow patterns, we introduce a multi-graph structure within the GAT framework. This structure allows the model to draw relevant information from these related segments, effectively enabling it to infer the likely flow for untapped segments by learning from the available data of similar, existing segments (L. Liu et al., 2022). Indeed, there have been many studies showing that the multi-graph can describe graph information better than single graph (Chai, Wang, & Yang, 2018; Lee & Rhee, 2022; M. Liang et al., 2022), but the research on multi-graph on GAT (MGAT) may be insufficient.

Furthermore, to effectively mine both spatial and temporal information, we integrate Long Short-Term Memory (LSTM) networks into the framework. LSTM is particularly well-suited for capturing temporal dependencies, which is essential given the dynamic nature of passenger demand over time. Recently, spatial-temporal dependencies of transportation networks have received a lot of attention (C. Zhang et al., 2019), and many expanded machine learning methods have been presented, including TS-STNN (Lv et al., 2023), CNN-LSTM (Crivellari, Beinat, Caetano, Seydoux, & Cardoso, 2022), STGCN (Huo et al., 2023; B. Yu, Yin, & Zhu, 2017) and so on. These methods advance the development of machine learning algorithm on the data-driven approach. However, they generally partition the dataset from the temporal dimension, which requires the historical flow data of all nodes on graph structure. Recognizing the limitations of existing data partitioning methods, we propose a new dataset partitioning strategy in our framework that does not rely on historical flow data for untapped segments.

Finally, to optimize the parameters within the MGAT, we employ the Nondominated Sorting Genetic Algorithm II (NSGA-II). Multi-graph structure usually involves the problem of relevant parameters assign (Lee & Rhee, 2022; M. Liang et al., 2022), but it is challenging to assign the optimal parameters combination manually. There have been some studies which demonstrated that introducing genetic algorithm (GA) into graph convolution network (GCN) can enhance the algorithm performance (Y. Liu, Liu, & Li, 2022). However, they only applied GA to optimize the general parameters of GCN, such as network architecture and other hyperparameters, which did not make the results interpretable. This study introduce NSGA-II to automatically and effectively determine the optimal parameters, particularly for the integrated affinity degree in the association graph, making the model's outputs interpretable and robust.

The contribution of this work lies in the following aspects.

1) Addressing the historical flow data challenge: The proposed MGAT-LSTM-NSGA-II framework directly tackles the challenge of predicting flow for untapped flight segments. specifically, multi-graph structure is introduced in MGAT component to mine the spatial dependence by constructing topology graph and association graph simultaneously, LSTM component aims to capture the temporal change information from historical data of existing routes.
2) Innovative data partitioning approach: We introduce a novel data partitioning strategy that allows the model to operate effectively without relying on historical data, addressing a key limitation of existing data-driven approaches.
3) Improved interpretability and optimization: By defining and optimizing the integrated affinity degree within the association graph using NSGA- II, our framework not only improves prediction accuracy but also enhances interpretability, offering insights into the key factors influencing flow on untapped routes.

The innovative approach allows airlines to make informed decisions on route expansion, reducing reliance on costly and environmentally unfriendly flight tests, and promoting the sustainable development of air transportation (Y. Liang et al., 2023).

The rest of the paper is organized as follows. Section 2 reviews existing research related to our work. Section 3 defines the problem of untapped flight segment flow prediction. Section 4 presents the proposed methodology. Section 5 evaluates the proposed approach through experiments and analyze the influence of node features on the results. Section 6 discusses the results and makes conclusions.

2 Literature review

In this section, we briefly review the recent literature relevant to our work, including air traffic flow prediction, which is the main problem this paper focusing on, and graph deep learning in traffic prediction which is the key methodology this paper adopting. Finally, we conducted a summary and analysis of the key relevant literature.

2.1 Air traffic flow prediction

The prediction of air traffic flow is meaningful for air traffic planning (Leandro, Andrade, & Kalakou, 2021), so that many studies on aviation networks optimization take air traffic forecasting as an important part (Birolini et al., 2021). From perspective of the prediction area, air traffic flow prediction can be divided into four levels, including sector, route, route points and airport (Cai et al., 2023). This paper mainly focuses on the air traffic flow prediction of route to provide advice for new airline routes operation decision.

From a supply and demand perspective, the variables considered in the prediction model are roughly divided into two categories: supply-related data and demand-related data (Hsiao & Hansen, 2011). Some studies consider supply-related data as the forecasting model input, where the supply-related data refers to the segments historical data generated by flights or other flight features (Birolini et al., 2021). For example, Cai et al. (2023) presented a multi-step spatial-temporal prediction model based on historical fight flow data. However, they ignore the influence of market feature.

Recently, numerous researchers have pay attention to the influence of socio-economic and geographical data, which refers to the statistics of demand-related data (Birolini et al., 2021). For example, J. Yu (2021) proposed the flight segments forecasting model, combining the historical price and demand of flight segment. Birolini et al. (2021) presented an empirical model to estimate passenger demand, considering segments historical data as well as other socio-economic and geographical data, like population, GDP and distance. Leandro et al. (2021) concluded the important variables in prediction model, including GDP, population, importance of tourism, and cost of ticket. Gui et al. (2020) built two traffic flow prediction models, considering impact factors of holiday, season, and average flow. Marazzo, Scherre, and Fernandes (2010) investigated the relationship between air transport demand and GDP.

These studies capture the socio-economic and geographical data, which can take market features into account (Birolini et al., 2021). However, the competition factor is not considered. Besides, they only focus on the problem of existing flight segment flow prediction, and most studies need historical data for model calibration (T. Li & Wan, 2019; Yan et al., 2023), ignoring the untapped flight segment prediction.

2.2 Graph deep learning in traffic forecasting

Deep learning has been widely used in traffic forecasting (Chang, Wu, Correia, Sun, & Feng, 2022; Makridakis et al., 2023). Especially, due to the complexity of spatial relationships and temporal dependencies, graph deep learning has attracted a lot of attention recently (Ta et al., 2022). In essence, the transportation network is a graphic organization, it is appropriate to express the traffic network as a graph mathematical form (Ye, Zhao, Ye, & Xu, 2022). Literature has pointed out that it is difficult to capture complex nonlinear correlations in traffic flow forecast based on traditional methods (Gui et al., 2020), while neural networks are capable to model high non-linearity and complex patterns inside time series data (F. Zhang, Fleyeh, & Bales, 2020). Therefore, more and more studies research on traffic forecasting based on graph deep learning, for instance, Graph convolutional networks (GCN) and its expanded algorithm have been applied widely in air traffic flow prediction to capture spatial dependencies in graphs (Huo et al., 2023; Lv et al., 2023).

As one of the most popular graph deep learning methods, GCN was first proposed by Kipf and Welling (2016). Many studies have applied GCN to traffic forecasting (J. Zhang, Chen, Cui, Guo, & Zhu, 2021). Currently, researchers have considered on other factors additional to the distance information in traffic networks, so that multi graph convolutional networks (MGCN) came into been (Lee & Rhee, 2022). Geng et al. (2019) presented MGCN for generic spatiotemporal forecasting by constructing three separate graphs, geo-distance graph, POI similarity graph and road network graph. Lee and Rhee (2022) incorporated distance graph, direction graph and positional relationship graph into GCN, and defined a set of partition filters to learn information. M. Liang et al. (2022) presented an advanced MGCN to take full advantage of the correlations of traffic flow in maritime graph, which includes distance graph, interaction graph, and correlation graph. However, most of them focus on the problem of node prediction, which is inconsistent with the route flow forecasting in this paper. According to a novel idea of graph construction proposed by J. Yu (2021), we regard flight segments as nodes and the association between flight segments as arcs. Different from existing literature, distance graph is not considered in this paper, but association graph and topology graph are considered.

In addition, GAT is also frequently used in the traffic forecasting, due to its effectiveness in computing the pair-wise attentional correlations (Velikovi et al., 2017). Compared with GCN, the attention mechanism in GAT is more appropriate for dealing with the relationship between geographically unrelated segments (Tang, Sun, Sun, Peng, & Gan, 2020). Many researchers have applied GAT to mine the complex spatial correlations within the traffic network (Tang et al., 2020; C. Zhang et al., 2019; K. Zhang, He, Zhang, Lin, & Li, 2020). GAT focuses on the features of neighbor nodes instead of geographical relationship (Tang et al., 2020). Therefore, to capture the information of both node features relationship and geographical relationship, multi-graph attention networks need to be considered. However, to our knowledge, there is few research on introducing multi-graph definition into GAT. Huang, Luo, Cao, Wen, and Zhong (2022) followed with interest with the multi-aspect traffic data and proposed three graph embeddings by specific graph attention designs, where the adaptively coupled is adopted according to graph embeddings' respective importance to prediction. Therefore, we compared our proposed method with Huang et al. (2022), and the results of numerical studies based on the real-world dataset from airline in China show that our MGAT outperforms that developed in Huang et al. (2022).

To integrate the temporal information as well as spatial information, the spatial-temporal integrated deep learning frameworks have been presented (Chang et al., 2022; B. Yu et al., 2017). For example, Lv et al. (2023) presented a spatial–temporal neural network based on tree structure for air traffic flow prediction. J. Zhang et al. (2021) proposed a deep learning architecture based on GCN and LSTM to forecast short-term passenger flow. On top of that, with the development of deep learning, some studies have combined heuristic algorithm and deep learning algorithm to optimize the results (Fei, Wu, Zhang, Zhang, & Chen, 2022; Lu et al., 2020). For instance, Y. Liu et al. (2022) applied Genetic algorithm to GCN to achieve automatic search of architecture and hyperparameters. These developments in graph deep learning are enlightening to this paper. However, these existing spatial-temporal integrated graph deep learning approaches belongs to the historical data-based method (H. Li, Jiao, & Yang, 2023), which can not directly solve the flow prediction of untapped flight segment due to the lack of historical data of untapped flight segments. Besides, the methods based on graph deep learning and heuristic algorithms do not attempt to make the results interpretable.

2.3 Overview and analysis of relevant research

To highlight the contributions of this study, we have compared our work with the most relevant literature, as shown in Table 1. Due to the limited research on air traffic flow prediction, we have also included some recent studies on traffic flow prediction for comparison.

Table 1. Overview of recent key literature

Author	Prediction Target	Input data	Methods and models	Spatial-temporal dependency	Multi-graph structure	Model interpretability
Gui et al. (2020)	Existing flight segment	Holiday, season, historical average flow	LSTM	×	×	×
J. Yu (2021)	Existing flight segment	Historical price and demand data	GCN &LSTM	√	×	×
Birolini et al. (2021)	All flight segment	Historical flow data, demographic, geographic and economic variables, flight frequency	Gravity model (GM)	×	×	√
Leandro et al. (2021)	All O/D pairs	GDP, population, importance of tourism, ticket price.	Multivariate liner regression (MLR) & generalized linear models (GLM)	×	×	√
J. Zhang, Chen, Cui, Guo, & Zhu, 2021	Passenger flow of existing traffic route	Historical inflow, weather condition, air quality	Residual network, GCN and LSTM	√	×	×
Huang et al. (2022)	Existing traffic prediction	Historical graph signals data, like traffic volume, speed and density.	Multi-relational synchronous graph attention network (MSGAT)	√	Channel, temporal and spatial relations	×
Cai et al. (2023)	Existing sector flow	Historical fight flow data and airspace data	Temporal attention & spatial dual- graph convolution	√	Adjacency graph & origin-destination graph	×
Lv (Lv et al., 2023)	Traffic flow of existing target	Historical flow data	Spatial–temporal neural network based on tree structure	√	×	×

	location					
This paper	All flight segment	Historical flow of existing flight segment, geographic and economic variables, flight frequency, and competition metric	NSGA II-MGAT-LSTM	√	Topology graph & association graph	√

From this comparison, it becomes evident that existing model-driven approaches, such as GM and MLR, although capable of predicting new flight segment flows with good interpretability, failing to capture information from both spatial and temporal dimension. On the other hand, data-driven models primarily based on machine learning methods, like GCN. While many of these studies have focused on the spatial-temporal structure of traffic, enabling better modeling, most of them regarded problem as an time series model, which can only be used if enough historical data exists (Caserta & D'Angelo, 2024), making them unsuitable for predicting future flows of new flight segment lacking historical data. Additionally, only a few studies have employed multi-graph structures to better represent the data, yet they lack further exploration into the interpretability of the model. Given the absence of historical flow data for new flight segments, there is a need for a novel prediction method that does not rely on historical data of the predicting target itself. Therefore, this study proposes the MGAT-LSTM-NSGA-II framework, which can capture information from spatial-temporal dimension and further discuss the interpretablity of model. Most importantly, it does not rely on historical traffic data specific to the new flight segments.

3 Problem statement and model construction

In air traffic networks, airlines face numerous route choices. Some routes are already part of their flight plans, which we refer to as existing routes, while others remain unoperated, referred to as untapped routes. The task of predicting the flow of untapped flight segments is crucial for determining whether it is valuable to initiate flights on these routes. However, the absence of historical flow data for untapped segments presents a significant challenge, despite the availability of historical data for existing segments.

This lack of data for untapped segments necessitates a method that can leverage the existing data to make predictions. To address this challenge, we model the air traffic network as a graph, where each flight segment is regarded as a node and each airport as a connection between segments. Let $G=(V,E)$ represent the topology graph of the air network, where V is the set of nodes representing flight segments, $V=p_1, p_2, ..., p_i, p_N$, N is the total number of nodes in graph structure, and E is the set of arcs representing connections between these segments. As shown in Figure 1, node P_5 denotes segment BD, which is an untapped flight segment, with links to segment BC, AD, and AB via airports B and D.

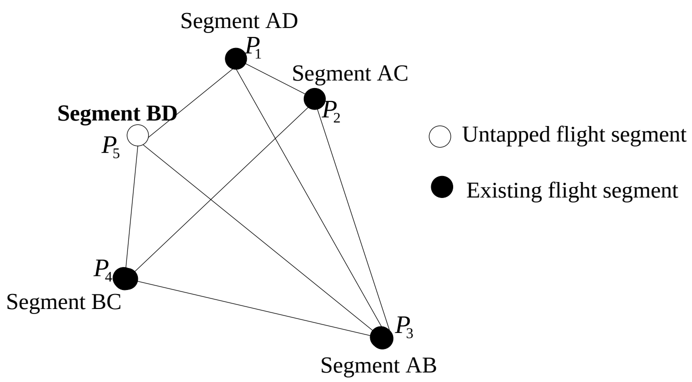

Figure 1. Illustration of topology graph structure

In real life, airlines need to evaluate untapped segments like BD to decide whether to launch flights. Although there is no historical flow data of segment BD, data from existing segments such as AD, AC, AB, and BC, along with route properties obtained from official statistics, can be leveraged. The objective is to predict the future flow of segment BD, aiding in the evaluation of its potential value.

This prediction task is complex due to the dynamic nature of the airline network, which has a high degree of uncertainty. As new segments open or close, the node statuses and adjacency relationships in graph *G* evolve over time. Thus, our approach needs to integrate information from both temporal and spatial dimensions to predict the flow of untapped segments.

We proposed a multi-graph attention network (MGAT) to aggregate spatial information from adjacent nodes. This component captures the spatial relationships and influences between different flight segments at each moment *t*, addressing the spatial aspect of the problem. Besides, to handle the temporal dependencies, a Long Short-Term Memory (LSTM) network is used to models the historical flow data of existing segments over time. This addresses the time-series aspect of the problem. Formally, the prediction problem can be expressed as equation (1):

$$\hat{Y}_{untapped}^{T+1,\ T+L} = LSTM\left[(MGAT(G_t), t = 0,1,\ldots,T), Y_{existing}^{tstep,T}\right] \tag{1}$$

where $G_t = (X_1^t, X_2^t, \ldots X_N^t)$,which represents the graph structure information at moment *t*, $MGAT(\cdot)$ is the MGAT operator to aggregator spatial information across at each moment *t*, $Y_{existing}^{tstep,T}$ denotes the historical flow information of existing segments over period *tstep* to *T*, which is used to train and test model. *tstep* is the input time step of LSTM, and $LSTM(\cdot)$ aggregates temporal information over every *tstep* historical moments to predict the flow in next *L* moments. Thus, the flow $\hat{Y}_{untapped}^{T+1,\ T+L}$ of the untapped segment over the future period *T*+1 to *T*+*L* is obtained by learning information from period *T*-*tstep* to *T* .

Additionally, there are key parameters in the multi-graph structure definition of MGAT that need to be determined, which play an important role in prediction efficiency. Since the values of these parameters will affect each other, it is difficult to confirm them manually. We regard the parameter searching process as an optimization problem and NSGA-II is used to solve the problem. The evaluation metrics for prediction model are used as the optimization objectives, and assuming the key parameter combination is $\boldsymbol{\delta}$, then the optimization model can be expressed as follows equation (2) to equation (6):

$$\min Z_1 = MAPE(\boldsymbol{\delta}) \tag{2}$$

$$\min Z_1 = MAE(\boldsymbol{\delta}) \tag{3}$$

$$\min Z_1 = RMSE(\boldsymbol{\delta}) \tag{4}$$

$$\text{S.t.} \quad \sum \boldsymbol{\delta} = 1 \tag{5}$$

$$\boldsymbol{\delta} \in [0,1] \tag{6}$$

where $MAPE(\cdot)$, $MAE(\cdot)$, and $RMSE(\cdot)$ are mainstream evaluation metrics equation in prediction problem, referring to mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) respectively (Huang et al., 2022). Equation (5) limits the sum of all parameters to be 1, and equation (6) defines the value range of parameters. These parameters are used to control the generation of multi-graph structure, we mainly focus on their relative sizes rather than their absolute sizes. Detailed descriptions will be provided in the Methodologies section.

The core motivation of this study is to leverage the historical flow data of existing segments to train the LSTM network, using the aggregated graph information produced by MGAT over time as input, where the key parameters that control multi-graph structure are determined by NSGA-II. This approach enables us to accurately predict the future flows for untapped segments, thus providing valuable insights for route planning.

4 Methodology

To predict the untapped flight segment future flow, the thinking of this paper is to use own and adjacent segment feature information as well as the historical flow data of existing segment.

Figure 2 shows the framework of the overall MGAT-LSTM-NSGA-II model, which includes MGAT, LSTM, and NSGA-II components. In this framework, NSGA-II serves as the main body to identify the optimal key parameters, with each parameters set leading to different prediction outcomes. Core MGAT-LSTM module is embedded into the fitness function calculation to forecast the future new flight segments flow, and the evaluation metrics of the prediction results are fed back into NSGA-II as fitness values.

The details of MGAT-LSTM module are shown in the right of the figure 2. The module's input consists of two parts: the spatio-temporal data of the airline network over past time steps ($0,1,\ldots,T$), and the key parameter combinations randomly generated by NSGA-II. Based on these key parameters, the input spatio-temporal data is represented as a multi-graph structure and fed into MGAT to aggregate spatial information. The aggregated data at each time step is then stacked into a time series and input into LSTM to extract temporal information, ultimately predicting the traffic flow for each flight segment in next L time steps ($tstep, tstep+1,\ldots,T+L$). The predicted flow of existing flight segments in historical time ($tstep, tstep+1,\ldots,T$) is compared with the actual flow to generate the model's evaluation metrics, which are then fed back into NSGA-II. A fixed sliding window prediction is used in LSTM, that is, input graph structure data at every *tstep* historical moments and output predicted traffic at the next L moments. Finally, this model can be used to predict the future ($T+1, T+2,\ldots,T+L$) flow for untapped flight segments. The detail description of algorithm is introduced in the following subsection.

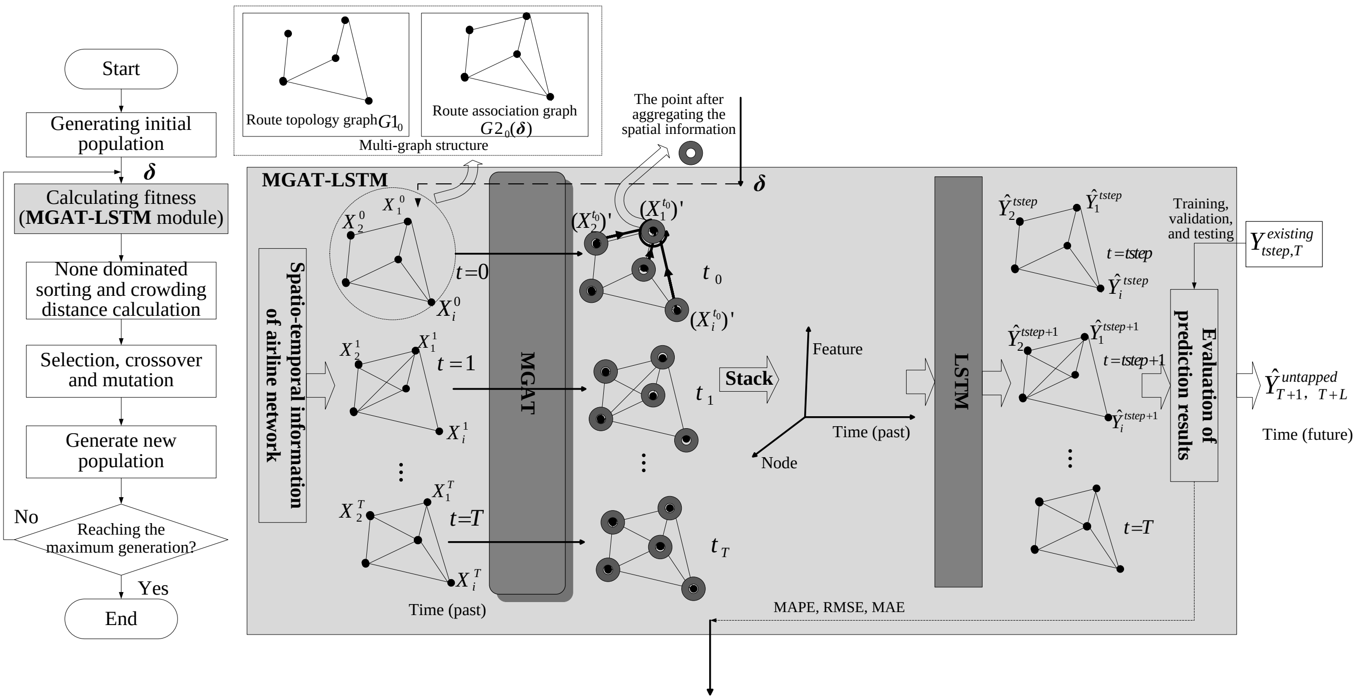

Figure 2. The framework of MGAT-LSTM-NSGA-II

4.1 Flight segment network graph

4.1.1 key factors and feature extraction

Refer to the gravity model, one of the most successful empirical models to forecast passenger demand, there are five key predictor factors needed to be considered, including distance, GDP, throughput, population, and flight frequency (Birolini et al., 2021; Boonekamp, Zuidberg, & Burghouwt, 2018). Since the new segment flow forecast considered in this paper is to support the development of new markets, we take into account important market competition factors. Though airfares can reflect the preference of consumers (Birolini et al., 2021), it is irrelevant to different routes of the same airline, therefore, the influence of airfares is not considered in this paper.

Thus, the feature of each point is considered from six aspects: distance, GDP per capita, airport throughput, population size, flight frequency and competition. Distance between origin and destination defines the route nature (h_i^{distance}), GDP per capita represents the economic distance between airports ($h_i^{\text{GDP}}=h_i^{\text{GDP\_origin}}-h_i^{\text{GDP\_destination}}$), airport throughput represents local passengers ($h_i^{\text{throughput}}=h_i^{\text{throughput\_origin}}\times h_i^{\text{throughput\_destination}}$), which reflects the market potential together with population size ($h_i^{\text{population}}=h_i^{\text{population\_origin}}\times h_i^{\text{population\_destination}}$), flight frequency refers to the number of flights by current airline on segment in week ($h_i^{\text{frequency}}$), which indicates the airline strategy direction, and competition reflects the degree of market saturation ($h_i^{\text{competition}}$). As shown in equation (7), h_i is defined as the feature of node i, and it determined by six different feature dimensions, $h_i \in \boldsymbol{R}^6$:

$$h_i=\{h_i^{\text{distance}},h_i^{\text{GDP}},h_i^{\text{throughput}},h_i^{\text{population}},h_i^{\text{frequency}},h_i^{\text{competition}}\} \tag{7}$$

Because nodes in the graph denote flight segments which have origin airport and determination airport, the feature of both GDP per capital, airport throughput, and population size are determined by both origin and destination (Birolini et al., 2021). Finally, the feature matrix of graph is a matrix with N rows and six columns, and is recorded as $\boldsymbol{H}^{N\times 6}$.

4.1. 2 multi-graph structure

Since the learning mechanism in GAT is to learn the information from its first-hop neighbors (C. Zhang et al., 2019), only define the topology graph is not sufficient, we redefine the adjacent relationship between flight segments. According to the literature(Quan et al., 2012), we define that the higher the similarity, the stronger the segment adjacency. Therefore, the association graph is introduced into this paper, and together with the topology graph, a multi-graph structure is constructed.

a) Topology graph

According to the flight networks in real life, the topology graph is constructed. Let $w_{ii'}$ denote the topology relationship between node p_i^{od} and node $p_{i'}^{o'd'}$ in topology graph (equation (8)), where symbols *o* and *o'* indicate origin airports, and symbols *d* and *d'* indicate destination airports:

$$w_{ii'}=\begin{cases}1, & if\ \ o=d'\ \ \text{or}\ \ o'=d\\ 0.5, & if\ \ o=o'\ \ or\ \ d=d'\\ 0, & otherwise\end{cases} \tag{8}$$

Topology graph implies the flow of passenger demand. If the origin airport of flight segment *i* is the same as the destination airport of segment *i'*, or the destination airport of flight segment *i* is the same as the origin airport of segment *i'*, it means that there may be a connection relationship between the two segments, and the weight is set as 1. While if the origin or destination airport of two flight segments is the same, there is a diversion or concurrent effect on flow, so the weight is set as 0.5. Otherwise, the weight is set as 0.

b) Association graph

Based on the idea of recommendation, the information of a new object can be learned from the similar one (Quan et al., 2012). Therefore, an association graph is constructed to describe the feature relationship between flight segments.

To reduce the impact causing by sample data differences at different feature dimension and make model training more efficient and robust, the vector with different features is normalized. As shown in equation (9), $h_i^{m}{}'$ denotes the normalized vector, $h_{\min}^{m}$ is the minimum logarithm value of feature dimension *m*, and $h_{\max}^{m}$ is the maximum logarithm value of feature dimension *m*:

$$h_i^{m}{}' = (\log(h_i^{m}) - h_{\min}^{m}) / (h_{\max}^{m} - h_{\min}^{m}), \quad m \in M \tag{9}$$

where *M* is the set of feature dimension, and $M = \{\text{Distance, GDP, Throughphput, Population, Frequency,Competition}\}$.

Let $s_{ii'}^{m}$ denotes the affinity relationship at feature dimension *m* between node p_i^{od} and node $p_i^{o'd'}$. As shown in equation (10), it can be calculated based on the Gaussian kernel (Quan, et al., 2012):

$$s_{ii'}^{m} = \exp(-\left\| h_i^{m}{}' - h_{i'}^{m}{}' \right\|^2 / 2\mu^2) \tag{10}$$

In equation (10), $\left\| h_i^{m}{}' - h_{i'}^{m}{}' \right\|^2$ is used to measure the distance of differences between two points, $\left\| h_i^{m}{}' - h_{i'}^{m}{}' \right\|^2 \in [0,1]$, and symbol μ represent affinity coefficient, which influences the affinity value between routes. According to Gaussian kernel function, too large coefficient will result in almost no differentiation in the results, while too small coefficient will result in the result approaching 0. Therefore, appropriate coefficient setting is crucial for representing true affinity relationship between points. To determine the appropriate coefficient, we give the distance ($\left\| h_i^{m}{}' - h_{i'}^{m}{}' \right\|^2$) different value to observe the change of affinity relationship with different coefficients, which is recorded as Figure 3.

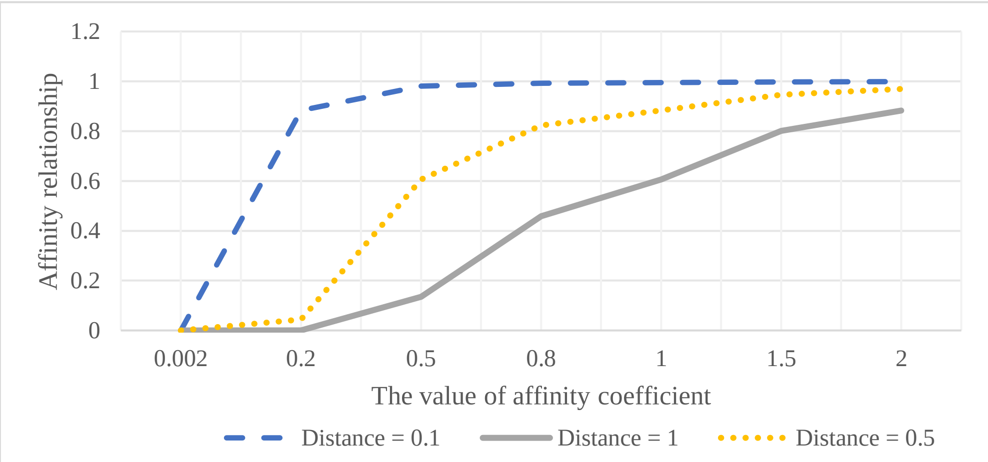

Figure 3. The sensitivity analysis of affinity coefficients μ

As shown in Figure 3, when affinity coefficient μ is 0.5, the affinity relationship among different distances are quite different. Therefore, we set the value of affinity coefficient as 0.5.

Based on the affinity relationship at each feature dimension, we determine the integrated affinity degree between node p_i^{od} and node $p_i^{o'd'}$ as the equation (11) shown:

$$s\_o_{ii'} = \sum_{m \in M} \delta_m s_{ii'}^{m} \tag{11}$$

As shown in equation (11), where δ_m is the key parameters that controls the influence degree from feature *m*, which is named as feature weight, and $\delta_m \in \boldsymbol{\delta}$, $\sum_{m \in M} \delta_m = 1$. The larger δ_m is, the greater the influence of affinity relationship at feature dimension *m* on integrated affinity degree.

Finally, to make the integrated affinity degree of each node sufficiently distinct to other nodes, we normalize the current integrated affinity degree $s\_o_{ii'}$ to obtain the final integrated affinity degree $s_{ii'}'$, which is similar to the process of assigning learning weights to each point, as shown in equation (12):

$$s_{ii'}' = (s\_o_{ii'}' - \min(s\_o_i)) / (\max(s\_o_i) - \min(s\_o_i)) \tag{12}$$

where $\min(s\_o_i)$ denotes the minimum of initial integrated affinity degree of node i, $\max(s\_o_i)$ denotes the maximum of initial integrated affinity degree of node i.

c) Extraction of adjacent matrix

Based on the topology graph and association graph, the adjacent matrix is defined as $\boldsymbol{A}^{N\times N}$. we represent the adjacent relationship $a_{ii'}$ as follows equation (13):

$$a_{ii'} = \begin{cases} w_{ii'} \times s_{ii'}, & w_{ii'} \times s_{ii'} > 0.4 \\ s_{ii'}, & s_{ii'} > 0.8 \quad and \quad w_{ii'} = 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

As shown in equation (13), when there is a topology connection between nodes, the adjacent relationship is determined by topology connection as well as association connection. When the affinity value between nodes is relatively large, we consider that there is a link between nodes although there is not a topology connection.

4.2 Spatial module: MGAT

After obtaining the relationship between nodes, the connection graph is constructed, and we further explore the spatial graph information by proposing the multi-graph attention networks (MGAT). GAT extends GCN by replacing its convolution operation with an attention mechanism, which can deceptively learn different correlations between different points (C. Zhang et al., 2019). The same as GAT, the core component of MGAT is graph attentional layer. Defined the node feature of layer l as a set $(H)^l = \{(h_1)^l, (h_2)^l, ..., (h_N)^l\}$, $(h_i)^l \in \boldsymbol{R}^6$, which is the input of the MGAT. Equation (14) shows the attention coefficient e_{ij} of nodes i and its neighbor node j. ϕ is a shared weight matrix, which is used to cast the input to an embedding (C. Zhang et al., 2019). Symbol β is the attention mechanism:

$$e_{ij} = \beta(\phi h_i^l, \phi h_j^l) \tag{14}$$

Equation (15) defined the normalized attention coefficient α_{ij} (C. Zhang et al., 2019), LeakyReLU is an activation function, which is to make the negative value as close to 0 as possible. Then, a softmax function is used to normalize the attention coefficients:

$$\alpha_{ij} = \text{softmax}(\text{LeakyReLU}(e_{ij})) \tag{15}$$

Consequently, the propagate process of MGAT can be represented by equation (16) (C. Zhang et al., 2019):

$$h_i^{l+1} = \sigma(\sum_{j\in N(i)} \alpha_{ij} \phi h_j^l) \tag{16}$$

where $N(\mathrm{i})$ is the set of neighbor nodes of node i. σ represents the activation function, where

ReLU function is used to confirm the value of output is bigger than 0.

Finally, we obtain the embedding as the output of spatial block MGAT, which is recorded as symbol Emb.

4.3 Temporal module: LSTM

Based on the spatial block MGAT, the points after aggregating spatial information at each moment can be obtained. They are stacked at time dimension to form a time-series data, and as the input of temporal block LSTM. LSTM model is used to capture the temporal information (Crivellari et al., 2022).

Assuming there is a historical period which has T moments. The MGAT model is used L times to obtain the embedding information at every moment. Then we get the input information for LSTM, that is embedding information, and $Emb \in \boldsymbol{R}^{N \times T \times 6}$. LSTM module designed integrates several layers: a convolutional neural network is used at first to extracts local patterns from the input sequence, followed by a pooling layer that reduces dimensionality and enhances robustness. The sequence is then fed into a LSTM layer, which captures temporal dependencies. To focus on important features across time steps, a multi-head attention mechanism is applied, enhancing the model's ability to learn long-range dependencies. Finally, a fully connected layer maps the processed feature to the output space, yielding the final prediction.

To enrich our sample, fixed sliding window prediction is adopted and time step is defined in this component, where time step is recorded as symbol *tstep*. Figure 4 shows the flow prediction process of node *i*. The process of LSTM in the context of our framework is a rolling prediction process, and aggregate data from the past *tstep* moments is used to predict flow of flight segment for the next L moments.

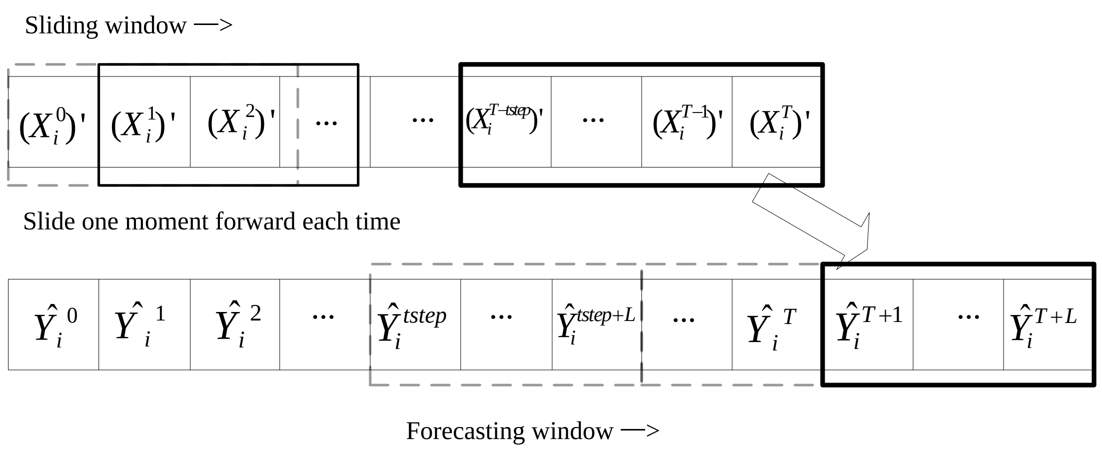

Figure 4. Illustration of the time step

Different from traditional flow prediction model, we divide the dataset not only from the time dimension, but also from the point set. Figure 5 shows the segmentation of data set.

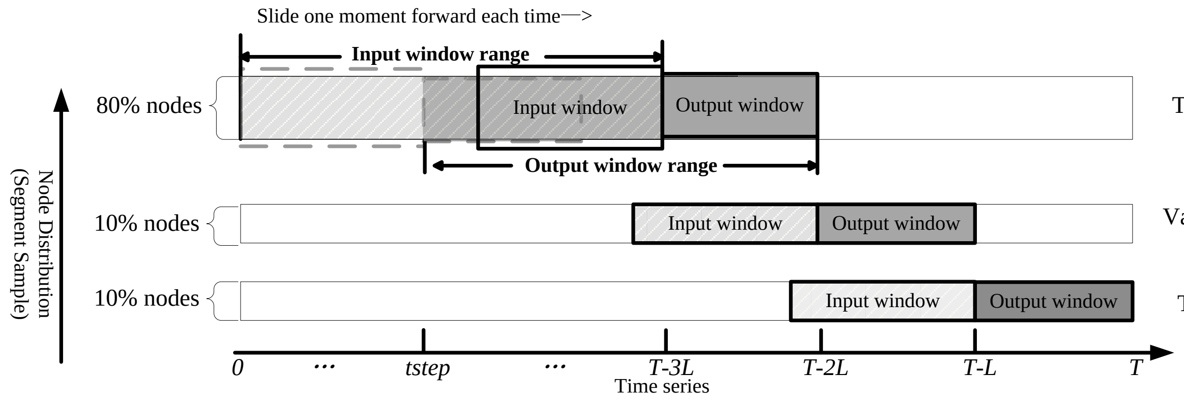

Figure 5. The segmentation of data set

As shown in figure 5, we use historical flow of 80% nodes from period *tstep* to *T-2L* to train the model, historical flow of 10% nodes from period *T-2L* to *T-L* to validate the model, and the historical flow of rest 10% nodes from period *T-L* to *T* to test the model. Each prediction flow in output window is obtained based on the aggregate information in the input window.

Because there were some flight segments opened or closed every month, some historical flow data are null. To ensure the efficiency of model training, mask is used to filter out nodes with empty flow in the output window. That is to say, in the process of each sliding training, it is necessary to judge if the flow of node in output window is null, if so, this node is not used to training model at this time.

To preserve the integrity of the graph, this dataset division only conducted on the training process of the LSTM. All points are included in the graph structure for spatial feature information aggregation, but when input into the LSTM layer, they are input according to the data set partitioning criteria. Points that are not in the corresponding dataset are also masked and do not participate in the gradient calculation of back propagation.

4.4 Optimization module: NSGA-II

Because there is an interaction between feature weight, it is hard to choose a proper parameter combination manually. NSGA-II is an algorithm which can generates population randomly and search for an optimal solution (Guo, Pu, Du, & Li, 2022). To make the results explicable and further discuss the influence from every affinity coefficient, this paper uses NSGA-II to search for the best parameter combination for MGAT.

a) Initial population: the population size of NSGA-II is determined as $6\times P$ at first, 0-1 coding is used to generate each individual randomly, and each individual indicates a coefficient combination. As shown in Figure 6, assign a gene fragment to the corresponding coefficient in sequence, and a coefficient corresponds to P gene fragment. By decoding, the coefficients are obtained.

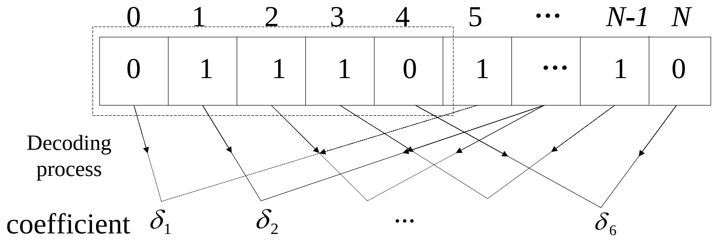

Figure 6. The example of individual coding

b) Calculate fitness: taking coefficient combination as an input to MGAT, the evaluation metric is returned. Because the purpose of this paper is to obtain the optimal prediction results, we set the fitness function based on the evaluation metrics. According to the equation (2) to equation (4), the fitness functions are determined as follow, and the goal of NSGA-II module is to find the optimal parameter combination $\boldsymbol{\delta}$, so as to obtain the optimal prediction results:

$$fitness_1 = \frac{1}{MAPE(\boldsymbol{\delta})} \tag{17}$$

$$fitness_2 = \frac{1}{MAE(\boldsymbol{\delta})} \tag{18}$$

$$fitness_3 = \frac{1}{RMSE(\boldsymbol{\delta})} \tag{19}$$

c) Selection, crossover and mutation: according to fitness, we rank individuals at first. To generate better offspring, we set the selection mechanism as that the higher the fitness, the higher the being selected probability. The individuals selected are regarded as parents to generate new individuals. Based on the random principle, we make crossover and mutation on the parental individuals.

d) Generate new population: use the above operation of selection, crossover and mutation to generate offspring which is the same size as the existing population. Then use the none dominated sort strategy to retain the population from origin population and new population. The size of new population keeps the same as the origin.

e) The stopping condition criteria: looping the above steps multiple times until the number of iterations reaches the maximum number of iterations.

5 Numerical studies and results analysis

5.1 Experiments setting

We conduct experiments on real-world datasets provided by an airline in China, which contains the information of airline networks, flight schedule and its corresponding historical flow. The time interval in this paper is month, to avoid the disruptions caused by external factors like the COVID-19 pandemic, we choose the data from 2015 to 2019 to conduct the experiments. In total, there are 60 months of data for 596 flight segments, equivalent to 35760 real-world data.

The data for the input feature of the model is taken from the airline mentioned above, Chinese Statistics Bureau, and Chinese Aviation Administration. The airline provides the distance and flight frequency feature data, where distance refers to the geographical distance between origin and destination airports. The flight frequency is the number of flights on each segment every week, which is obtained by the flight schedules. While the data of socio-economic features are provided by Chinese Statistics Bureau. Specifically, the population data is equal to the permanent population of the city where the origin and destination (OD) airports are located, and the GDP data refers to the GDP per capita of the city where the OD airports are located. Finally, Chinese Aviation Administration provides the throughput data of flight segment's OD airport, and the airline competition is measured by the number of airlines in the same flight segment.

Additionally, the output feature of the model is flight segment flow. We compared it with the actual flight segment's historical flow which is provided by the airline to evaluate the model.

Table 2 shows the data example of airline networks information and its corresponding origin and destination information.

Table 2. The data example of airline networks

O	D	Flow (people)	Distance (km)	Frequency (/week)	Population of origin (0,000)	Population of destination (0,000)	GDP of origin (00,000,000)	GDP of destination (00,000,000)	Throughput of origin (0,000)	Throughput of destination (0,000)	Competition

FOC	XMN	2057	320	2	780	40	122	139744.52	14760226	27413363	3
XMN	FOC	2522	320	14	429	780	139744	120414.10	27413363	14760226	3
JJN	SHA	3766	871	63	874	3428.14	373987	157138.06	8435805	45637882	11

We divide the dataset from both spatial and temporal dimension. From spatial dimension, the data is further divided into three parts: 80% nodes for training, 10% nodes for validating and 10% nodes for testing. From temporal dimension, the data is further divided into three parts: 36 month for train, 12 month for validating, and 12 month for testing. The dataset is randomly partitioned following the criteria of uniform distribution. Parameters of MGAT-LSTM-NSGA-II contains three parts, parameters in MGAT block, LSTM block and NSGA-II, which are set as Table 3 shown.

Table 3. Algorithm parameter settings for MGAT-LSTM-NSGA-II

Modules		Parameters	Value
Prediction model: MGAT-LSTM	Basic parameters of the model	Learning rate	0.0005
		Number of learning epoch	70
		Proportion of training node set	0.8
	MGAT block related	Number of convolution layers	1
		Number of heads	8
		Hidden size of MGAT	32
		Dropout rate	0.2
	LSTM block related	Hidden size of LSTM	16
		Time step: *tstep*	15
Optimization model: NSGA-II block		The range of affinity coefficients	(0, 1)
		Individual length: *N_len*	24
		Population size: *N_pop*	15
		Crossover probability	0.6
		Mutation probability	0.01
		Max generations	21

5.2 Baselines

In this section, the airline traffic flow data and statistical data are used to evaluate the performance of the proposed MGAT-LSTM-NSGA-II by comparing with several baselines, especially in the domain of flight demand prediction. The baselines are as follows

- Temporal attention & spatial dual- graph convolution (TAaDGCN) (Cai et al. (2023)): One of the state-of-the-art deep learning model in air traffic flow prediction, which considers dual-graph convolution network.
- Graphic Convolutional Neural Network and d the long short term memory (GCN-LSTM) (J. Yu, 2021) : One of the most popular framework of spatial-temporal graph deep learning for air traffic flow prediction.
- Spatio-Temporal Graph Convolutional Networks (STGCN) (B. Yu et al., 2017): A popular graph deep learning framework for traffic prediction, which takes spatio-temporal dependencies into account and performs well on multi-length predictions.
- Long short-term memory (LSTM) (Gui et al. (2020): A widely-used deep learning method for

predicting flight segment flow in time-series analysis, which can naturally capture the long-term and short-term time correlation.
- Gravity model (Birolini et al., 2021): One of the most successful empirical models to forecast passenger demand, which can be taken as one of the representatives of model-driven approach.

To ensure the fairness of the comparative experiments, we set the relative parameters of baselines the same as MGAT-LSTM-NSGA-II. Then we run algorithms through python 3.9. To reduce the influence of randomness, 10 independent runs are conducted based on MGAT-LSTM-NSGA-II and its comparative algorithms. Table 4 presents the average results of comparison between MGAT-LSTM-NSGA-II and other algorithms.

Table 4. The average results of comparison between MGAT-LSTM-NSGA-II and baselines

	MAE	RMSE	MAPE (%)
MGAT-LSTM- NSGA-II (this paper)	**1155.03**	**1524.26**	**30.80**
TAaDGCN (Cai et al., 2023)	2674.69	3444.52	49.35
GCN-LSTM (J. Yu, 2021)	2413.88	3269.81	48.83
STGCN (B. Yu et al., 2017)	3201.17	4182.32	59.81
LSTM (Gui et al., 2020)	2446.71	3241.19	49.00
Gravity model (Birolini, et al., 2021)	1398.10	2014.91	43.03

Table 3 shows that proposed MGAT-LSTM- NSGA-II model totally outperforms baselines, and its evaluation metric MAPE reach 30.80%, which indicates that the prediction model is relative well prepared.

TAaDGCN, GCN-LSTM and STGCN are spatio-temporal machine learning methods designed to capture both spatial and temporal dependencies in data, however, they tend to underperform in scenarios where historical flow data is unavailable. From the table 3, the average MAPE of TAaDGCN, GCN-LSTM and STGCN are higher than 45% and their MAE and RMSE are also relative high, indicating that they are not suitable for predicting flow on untapped flight segments. These methods typically rely on historical flow data either as a feature input or to construct adjacency matrices (Cai et al., 2023), and when the input feature data is replaced with socio-economic data, these models struggle to capture accurate traffic information. In contrast, the MGAT-LSTM-NSGA-II approach stands out by utilizing macro-level statistical data to characterize flight segments. Moreover, it compares the characteristics between flight segments to learn from the historical flow data of existing segments, thereby enabling more accurate predictions for untapped segments. As shown in Table 4, MGAT-LSTM-NSGA-II can reduce MAE by nearly 50% compared to other spatio-temporal machine learning methods.

LSTM, as a time-series model, primarily captures temporal patterns and relies heavily on historical flow data for accurate predictions. However, this reliance makes it less effective in scenarios where historical data is sparse or unavailable, such as in untapped flight segments. Additionally, LSTM lacks mechanisms to consider spatial relationships between data points, which are often crucial in networked systems like flight routes. In contrast, the proposed MGAT-LSTM-NSGA-II model integrates both spatial and temporal information through Multi-Head Graph Attention Networks and LSTM, enabling it to leverage spatial dependencies and learn from the characteristics of other flight segments. This dual capability allows MGAT-LSTM-NSGA-II to overcome the limitations of LSTM, making it better suited for predicting flow in untapped segments where historical data is not directly available.

As for gravity model, its performance depends on the model construction. The R-square (R^2) of gravity model in our comparison experiment is 91.5%, indicating that the model has a strong ability to explain the data, however, it still performs poor than MGAT-LSTM-NSGA-II proposed in this paper.

Compared with the existing machine learning method that require historical flow data as an input, the MAE and RMSE of gravity model are relatively good, which means that gravity model is more suitable for untapped fight segment flow prediction. Since gravity model is a regression model, it does not require historical flow data as an input. However, there are a lot of factors needed to be considered in the gravity model in general, and when there are limited data can be used, it may perform poorly. For instance, in our comparison, its MAPE is 39.71% higher than that our method. Thus, compared with the model-driven approach like gravity, the proposed method in this paper can capture complex spatial information based on graph structure, which is more flexible to model and does not require determining a large number of influencing factors as inputs (Blumenstock, Lessmann, & Seow, 2022). It indicates that the proposed framework can enhance prediction accuracy of untapped flight segment flow.

Since the goal of this study is to provide suggestions for the opening of new flight segment, we are more concerned with the overall traffic of new segment in the future. The figure 7 compares the total traffic of each flight segment in the next year predicted by different models and the actual annual traffic of the segments.

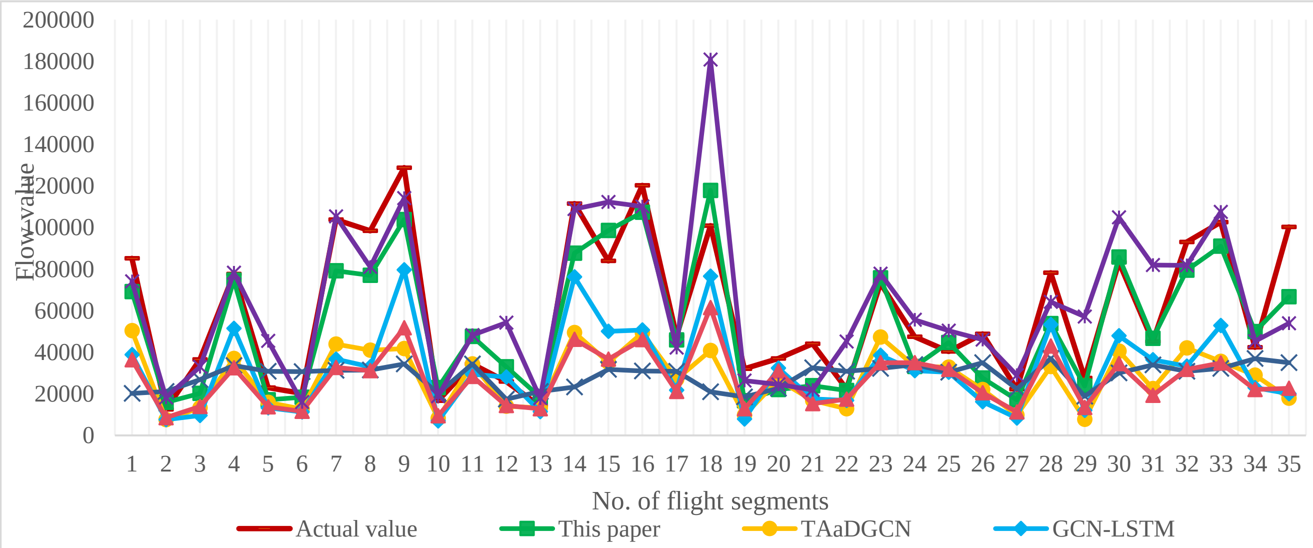

Figure 7. Comparison of the future annual total projections for each flight segment based on different methods

It can be seen from the figure 7, based on the method proposed in this paper, the prediction results for the future traffic of each segment are more accurate, and our forecast results can generally accurately capture the traffic levels of each flight segment. Although there are some errors, the predictions regarding the magnitude relationships between segments are relatively accurate, allowing us to correctly identify which segments have greater market potential. For example, to select the top three flight segments with the highest potential demand from the current segments for operation, the method proposed in this paper suggests choosing segments 18, 9, and 16. This recommendation is largely consistent with the actual segments 9, 16, and 14, and the actual flow for segments 18 and 14 is also relatively close.

5.3 The ablation study

To evaluate the effectiveness of three critical ingredients in the architecture of MGAT-LSTM-NSGA-II, we conduct ablation studies. Table 5 shows the experimental results in terms of the MAE, RMSE, and MAPE. To ensure the fairness of the comparative experiments, we set the same parameters as MGAT-LSTM- NSGA-II for all following methods. The affinity coefficients of MGAT-LSTM are randomly generated from 0 to 10.

Table 5. The ablation study of GA-MGAT-LSTM

	MAE	RMSE	MAPE
MGAT-LSTM- NSGA-II	**1155.03**	**1524.26**	**30.80**
MGAT-LSTM (Without GA)	1958.57	2608.76	42.81
GAT-LSTM (Without multi-graphs)	3120.10	4023.98	56.12
GA-MGAT (Without LSTM)	2696.11	3888.24	55.58

As shown in Table 5, MGAT-LSTM- NSGA-II proposed in this paper has the optimal results, where the MAE, RMSE, and MAPE are the lowest.

First, the GA component is proved effective for optimizing the affinity coefficient combination. Because the parameters combination is generated randomly, unsuitable parameters combination may result in false predictions. It indicates that it is hard to choose a proper parameters combination for MGAT-LSTM manually.

Second, the design of multi-graphs is proved very necessary to ensure the model accuracy. The idea of GAT is to aggregate information based on adjacent node features, however, the adjacent relationship only based on flight network is far from enough. It can be found from Table 5, the multi-graphs definition can reduce loss almost 50%.

Third, the LSTM component is confirmed essential for mining the information from temporal dimension. As shown in Table 5, the MAPE is 55.58%, upwards 50%, which indicates that the prediction results are poor. This is owing to the temporal information is ignored, and we only use each month spatial data to train the model. Thus, in untapped flight segment flow prediction, the information from temporal dimension is needed to be considered, which can capture the market changing information.

5.4 Explanation of flow prediction results

To analyze the different influence of different features on flight flow prediction, we plot the feature weight values corresponding to different solutions on the optimal frontier, as shown in the figure 7.

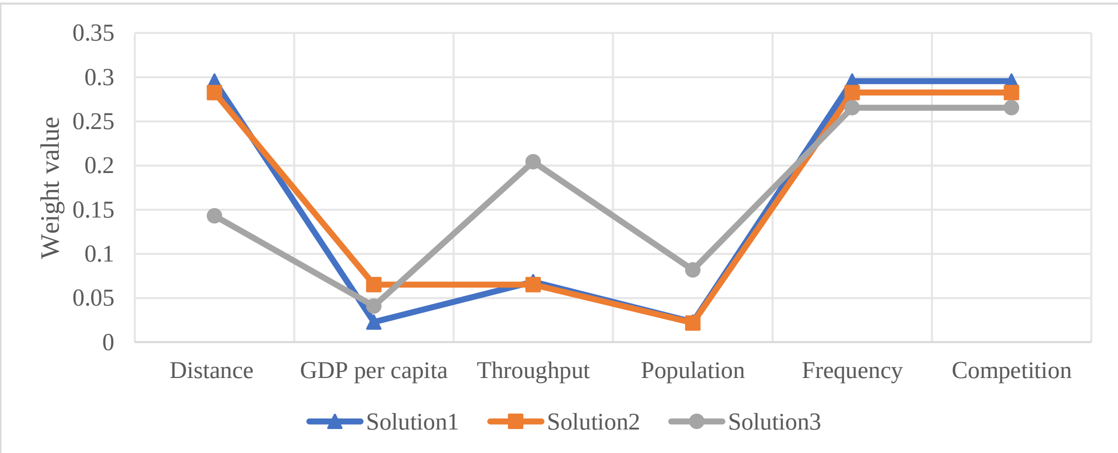

Figure 7. The affinity coefficients corresponding to the optimal population

From Figure 7, we can find that the weight value corresponding to the distance, throughput, frequency and competition almost higher than the value of population and GDP per capita. According to the feature weight coefficients and affinity degree defined in section 4.1.2, it indicates that the influence of distance, throughput, frequency and competition on prediction model are larger than population and GDP. It suggests that when it comes to an untapped flight segment flow prediction, the flow of the existing flight segment which has the similar distance, throughput, frequency and competition can provide more important reference for the untapped flight segment operation decision, especially the flight segments with similar distance, frequency and competition. It should be emphasized that the frequency here refers to the airline planning focus instead of the actual flight frequency.

On top of that, according to the gravity model (Birolini, et al., 2021) in section 5.2, where the coefficient for competitive factor is highly significant with a p-value of 0.000 and a substantial magnitude of -1085.0136. This suggests that competitive factor have a strong and statistically significant negative impact on the dependent variable, underscoring their critical importance in the model. The large t-value of -22.470 further reinforces the robustness and significance of this effect. This conclusion is consistent with the analysis results from feature weight value based on MGAT-LSTM- NSGA-II.

To sum up, the NSGA-II component not only can help seek solution automatically, but also can make the MGAT-LSTM- NSGA-II more explicable.

6 Conclusion

With the rapidly recovery of tourism industry, airlines around the world have started launching new airline routes. However, it is a complex work to evaluate whether the untapped airline routs are profitable. Due to the lack of historical data of untapped flight segments, traditional time series models which are usually utilized in traffic flow prediction are not applicable to untapped flight segment flow prediction. Referring to the idea of recommendation, we present MGAT-LSTM- NSGA-II, a novel framework combining graph deep learning and heuristic algorithm for untapped flight segments flow prediction. Accordingly, the idea of learning untapped flight segments flow information from similar existing flight segments is proposed. To be precise, based on the GAT, the MGAT is designed considering both topology graph and association graph, which is used to mine the complete graph information from spatial dimension, thereby the relationships between flight segments are captured. Then, LSTM is used to learn the information from temporal dimension to realize the prediction, where a new dataset partitioning method is presented to ensure the applicability of proposed method to predict untapped flight segments flow. Meanwhile, the NSGA-II is designed to optimize the affinity coefficients combination in MGAT automatically, which also makes the proposed approach explicable.

A real-life data study results based on our method and baselines have shown the outperformance of our method. Different from the model-driven approaches which have to consider a lot of factors to improve the prediction accuracy (Birolini et al., 2021; Hsiao & Hansen, 2011), our method constructs graph structures and dig information from spatial-temporal dimensions to reach scientific prediction. Besides, compared with other data-driven approaches that rely on own historical data (Huang et al., 2022; Lv et al., 2023), our method only requires some simple statistical data and other existing historical data, meanwhile, a new dataset partition method is proposed, which ensures its applicability of untapped flight segments flow prediction. The experimental results show that MGAT-LSTM- NSGA-II is suitable for untapped flight segments flow prediction.

Several managerial implications derived from this study are discussed below.

(1) When airlines want to launch new routes, they should pay more attention to the untapped routes which are similar to existing outstanding segments in terms of distance, OD throughput, frequency, and competition situation. The feature weight coefficients analysis results indicate that the distance, throughput, frequency and competition play important roles in flight segment flow prediction. Thus, airlines can gain a preliminary understanding of the untapped route value from these factors by existing routes.

(2) Airlines should consider the market change as well as the complex relationship between routes

when they evaluate the passenger demand of new routes. The market change can be reflected by temporal information. The baselines and ablation study demonstrate that the proposed approach is superior in predicting the untapped flight segment flow and all of its components are essential, including spatial information mining component and temporal information mining component. Hence, it is scientifical for airlines to adopt the framework proposed in this study to support the untapped routes operation decision.

(3) Finally, it suggests airlines to develop an untapped flight segment flow prediction system to make the new airline routes opening more effectively. In addition, the framework proposed by this paper can be expanded to new product demand forecasting in other industries. Managers can utilize this framework by adjusting the feature definition in association graph.

In future, we plan to explore more features of flight segments, and compare their different influences by affinity coefficients analysis, and try to build a systematic airline route prediction theory. Meanwhile, we consider expanding the scope of study to the international aviation network.

Declarations

Ethics approval and consent to participate: This research did not involve human participants and/or animals.

Consent for publication: The author(s) give their explicit consent that they obtained consent from the responsible authorities at the institute/organization where the work has been carried out, before the work is submitted.

Availability of data and material: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing Interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: This study was partially supported by the National Scientific Foundation of China (Project No. 71671152).

Authors' contributions: All the authors have made substantial contributions to the conception; design of the work; the acquisition, analysis, and interpretation of data. The authors drafted the work and revised it critically for important intellectual content. The authors approved the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgement: We are thankful to the reviewers and the editors for their valuable comments.

References

Birolini, S., Jacquillat, A., Cattaneo, M., & Antunes, A. P. (2021). Airline Network Planning: Mixed-integer non-convex optimization with demand-supply interactions. *TRANSPORTATION RESEARCH PART B-METHODOLOGICAL, 154*, 100-124. doi:10.1016/j.trb.2021.09.003

Blumenstock, G., Lessmann, S., & Seow, H.-V. (2022). Deep learning for survival and competing risk modelling. *Journal of the Operational Research Society, 73*(1), 26-38. doi:10.1080/01605682.2020.1838960

Boonekamp, T., Zuidberg, J., & Burghouwt, G. (2018). Determinants of air travel demand: The role of low-cost carriers, ethnic links and aviation-dependent employment. *TRANSPORTATION RESEARCH PART A-POLICY AND*

PRACTICE, 112, 18-28. doi:10.1016/j.tra.2018.01.004

Cai, K., Shen, Z., Luo, X., & Li, Y. (2023). Temporal attention aware dual-graph convolution network for air traffic flow prediction. *Journal of Air Transport Management, 106*. doi:10.1016/j.jairtraman.2022.102301

Caserta, M., & D'Angelo, L. (2024). Intermittent demand forecasting for spare parts with little historical information. *Journal of the Operational Research Society*, 1-16. doi:10.1080/01605682.2024.2349734

Chai, D., Wang, L., & Yang, Q. (2018). *Bike Flow Prediction with Multi-Graph Convolutional Networks*. Paper presented at the 26TH ACM SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS (ACM SIGSPATIAL GIS 2018). Proceedings Paper retrieved from

Chang, X., Wu, J., Correia, G. H. d. A., Sun, H., & Feng, Z. (2022). A cooperative strategy for optimizing vehicle relocations and staff movements in cities where several carsharing companies operate simultaneously. *Transportation Research Part E: Logistics and Transportation Review, 161*. doi:10.1016/j.tre.2022.102711

Crivellari, A., Beinat, E., Caetano, S., Seydoux, A., & Cardoso, T. (2022). Multi-target CNN-LSTM regressor for predicting urban distribution of short-term food delivery demand. *Journal of Business Research, 144*, 844-853. doi:10.1016/j.jbusres.2022.02.039

Fei, T., Wu, X., Zhang, L., Zhang, Y., & Chen, L. (2022). Research on improved ant colony optimization for traveling salesman problem. *MATHEMATICAL BIOSCIENCES AND ENGINEERING, 19*(8), 8152-8186. doi:10.3934/mbe.2022381

Geng, X., Li, Y., Wang, L., Zhang, L., Yang, Q., Ye, J., . . . Aaai. (2019). *Spatiotemporal Multi-Graph Convolution Network for Ride-Hailing Demand Forecasting*. Paper presented at the THIRTY-THIRD AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE / THIRTY-FIRST INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE CONFERENCE / NINTH AAAI SYMPOSIUM ON EDUCATIONAL ADVANCES IN ARTIFICIAL INTELLIGENCE. Proceedings Paper retrieved from

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. doi:10.1109/tvt.2020.2981959

Gunter, U., & Zekan, B. (2021). Forecasting air passenger numbers with a GVAR model. *Annals of Tourism Research, 89*. doi:10.1016/j.annals.2021.103252

Guo, J., Pu, Z., Du, B., & Li, Y. (2022). Multi-objective optimisation of stochastic hybrid production line balancing including assembly and disassembly tasks. *INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH, 60*(9), 2884-2900. doi:10.1080/00207543.2021.1905902

Hsiao, C.-Y., & Hansen, M. (2011). A passenger demand model for air transportation in a hub-and-spoke network. *Transportation Research Part E: Logistics and Transportation Review, 47*(6), 1112-1125. doi:10.1016/j.tre.2011.05.012

Huang, J., Luo, K., Cao, L., Wen, Y., & Zhong, S. (2022). Learning Multiaspect Traffic Couplings by Multirelational Graph Attention Networks for Traffic Prediction. *IEEE Transactions on Intelligent Transportation Systems, 23*(11), 20681-20695. doi:10.1109/TITS.2022.3173689

Huo, G., Zhang, Y., Wang, B., Gao, J., Hu, Y., & Yin, B. (2023). Hierarchical Spatio–Temporal Graph Convolutional Networks and Transformer Network for Traffic Flow Forecasting. *IEEE Transactions on Intelligent Transportation Systems, 24*(4), 3855-3867. doi:10.1109/tits.2023.3234512

Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. *ICLR*. doi:10.48550/arXiv.1609.02907

Leandro, F., Andrade, A. R., & Kalakou, S. (2021). Designing aviation networks under Public Service Obligations (PSO): A case study in Greece. *Journal of Air Transport Management, 93*. doi:10.1016/j.jairtraman.2021.102042

Lee, K., & Rhee, W. (2022). DDP-GCN: Multi-graph convolutional network for spatiotemporal traffic forecasting. *TRANSPORTATION RESEARCH PART C-EMERGING TECHNOLOGIES, 134*. doi:10.1016/j.trc.2021.103466

Li, H., Jiao, H., & Yang, Z. (2023). AIS data-driven ship trajectory prediction modelling and analysis based on machine learning and deep learning methods. *Transportation Research Part E: Logistics and Transportation Review, 175*. doi:10.1016/j.tre.2023.103152

Li, T., & Wan, Y. (2019). Estimating the geographic distribution of originating air travel demand using a bi-level optimization model. *Transportation Research Part E: Logistics and Transportation Review, 131*, 267-291. doi:10.1016/j.tre.2019.09.018

Liang, M., Liu, R. W., Zhan, Y., Li, H., Zhu, F., & Wang, F.-Y. (2022). Fine-Grained Vessel Traffic Flow Prediction With a Spatio-Temporal Multigraph Convolutional Network. *IEEE Transactions on Intelligent Transportation Systems, 23*(12), 23694-23707. doi:10.1109/tits.2022.3199160

Liang, Y., Lan, C., Dan, T., Qiaoqiong, Z., Yue, Y., & Lin, C. (2023). Research on supply and demand matching model of transportation modes in MaaS system of integrated passenger transport hub based on deep learning. *Soft Computing, 27*(9), 5973-5983. doi:10.21203/rs.3.rs-2550080/v1

Liu, L., Chen, J., Wu, H., Zhen, J., Li, G., & Lin, L. (2022). Physical-Virtual Collaboration Modeling for Intra- and Inter-Station Metro Ridership Prediction. *IEEE Transactions on Intelligent Transportation Systems, 23*(4), 3377-3391. doi:10.1109/tits.2020.3036057

Liu, S., & Jiang, H. (2022). Personalized route recommendation for ride-hailing with deep inverse reinforcement learning and real-time traffic conditions. *Transportation Research Part E: Logistics and Transportation Review, 164*. doi:10.1016/j.tre.2022.102780

Liu, Y., Liu, J., & Li, Y. (2022). Automatic search of architecture and hyperparameters of graph convolutional networks for node classification. *Applied Intelligence, 53*(9), 11104-11119. doi:10.1007/s10489-022-04096-w

Lu, W., Rui, H., Liang, C., Jiang, L., Zhao, S., & Li, K. (2020). A Method Based on GA-CNN-LSTM for Daily Tourist Flow Prediction at Scenic Spots. *ENTROPY, 22*(3). doi:10.3390/e22030261

Lv, Y., Lv, Z., Cheng, Z., Zhu, Z., & Rashidi, T. H. (2023). TS-STNN: Spatial-temporal neural network based on tree structure for traffic flow prediction. *Transportation Research Part E: Logistics and Transportation Review, 177*. doi:10.1016/j.tre.2023.103251

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Semenoglou, A.-A., Mulder, G., & Nikolopoulos, K. (2023). Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward. *Journal of the Operational Research Society, 74*(3), 840-859. doi:10.1080/01605682.2022.2118629

Marazzo, M., Scherre, R., & Fernandes, E. (2010). Air transport demand and economic growth in Brazil: A time series analysis. *Transportation Research Part E: Logistics and Transportation Review, 46*(2), 261-269. doi:10.1016/j.tre.2009.08.008

Quan, X., Liu, W., Dou, W., Xiong, H., & Ge, Y. (2012). Link Graph Analysis for Business Site Selection. *COMPUTER, 45*(3), 64-69. doi:10.1109/MC.2011.260

Suh, D. Y., & Ryerson, M. S. (2019). Forecast to grow: Aviation demand forecasting in an era of demand uncertainty and optimism bias. *Transportation Research Part E: Logistics and Transportation Review, 128*, 400-416. doi:10.1016/j.tre.2019.06.016

Ta, X., Liu, Z., Hu, X., Yu, L., Sun, L., & Du, B. (2022). Adaptive Spatio-temporal Graph Neural Network for traffic forecasting. *Knowledge-Based Systems, 242*. doi:10.1016/j.knosys.2022.108199

Tang, C., Sun, J., Sun, Y., Peng, M., & Gan, N. (2020). A General Traffic Flow Prediction Approach Based on Spatial-Temporal Graph Attention. *IEEE Access, 8*, 153731-153741. doi:10.1109/access.2020.3018452

Velikovi, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017). Graph Attention Networks. doi:10.48550/arXiv.1710.10903

Xu, Q., Pang, Y., & Liu, Y. (2023). Air traffic density prediction using Bayesian ensemble graph attention network (BEGAN). *Transportation Research Part C: Emerging Technologies, 153*. doi:10.1016/j.trc.2023.104225

Yan, Z., Yang, H., Wu, Y., & Lin, Y. (2023). A multi-view attention-based spatial–temporal network for airport arrival flow prediction. *Transportation Research Part E: Logistics and Transportation Review, 170*. doi:10.1016/j.tre.2022.102997

Ye, J., Zhao, J., Ye, K., & Xu, C. (2022). How to Build a Graph-Based Deep Learning Architecture in Traffic Domain: A Survey. *IEEE Transactions on Intelligent Transportation Systems, 23*(5), 3904-3924. doi:10.1109/tits.2020.3043250

Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. *27th International Joint Conference on Artificial Intelligence (IJCAI)*, 3634-3640.

Yu, J. (2021). A New Way of Airline Traffic Prediction Based on GCN-LSTM. *Frontiers in Neurorobotics, 15*. doi:10.3389/fnbot.2021.661037

Zhang, C., Yu, J. J. Q., & Liu, Y. (2019). Spatial-Temporal Graph Attention Networks: A Deep Learning Approach for Traffic Forecasting. *IEEE Access, 7*, 166246-166256. doi:10.1109/access.2019.2953888

Zhang, F., Fleyeh, H., & Bales, C. (2020). A hybrid model based on bidirectional long short-term memory neural network and Catboost for short-term electricity spot price forecasting. *Journal of the Operational Research Society, 73*(2), 301-325. doi:10.1080/01605682.2020.1843976

Zhang, J., Chen, F., Cui, Z., Guo, Y., & Zhu, Y. (2021). Deep Learning Architecture for Short-Term Passenger Flow Forecasting in Urban Rail Transit. *IEEE Transactions on Intelligent Transportation Systems, 22*(11), 7004-7014. doi:10.1109/tits.2020.3000761

Zhang, K., He, F., Zhang, Z., Lin, X., & Li, M. (2020). Graph attention temporal convolutional network for traffic speed forecasting on road networks. *Transportmetrica B: Transport Dynamics, 9*(1), 153-171. doi:10.1080/21680566.2020.1822765

Zhang, Z., Zhang, A., Sun, C., Xiang, S., Guan, J., & Huang, X. (2020). Research on Air Traffic Flow Forecast Based on ELM Non-Iterative Algorithm. *Mobile Networks and Applications, 26*(1), 425-439. doi:10.1007/s11036-020-01679-0