GATA - Graph Attention Networks in Aviation

Using a graph-based model to analyze the distribution of reactionary delays across a fleet network

Master Thesis
Aerospace Engineering - Control & Operations
February-December 2024

Constança Miranda de Andrade Veiga

Graph Attention Networks in Aviation

Using a graph-based model to analyze the distribution of reactionary delays across a fleet network

Master Thesis
Aerospace Engineering - Control & Operations

by

Constança Miranda de Andrade Veiga

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday December 18, 2024.

Student number:	4786327	
Project duration:	February 4, 2024 – December 4, 2024	
Thesis committee:	Dr. M. Ribeiro,	TU Delft, supervisor
	Dr. ir. E. van Kampen,	TU Delft
	Dr. J. Sun,	TU Delft
	Dr. M. Carré,	Swiss International Air Lines

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Thesis Contents

List of Abbreviations

Abbreviation	Definition
ATA	Actual Time of Arrival
ATC	Air Traffic Control
ATD	Actual Time of Departure
ABM	Agent Based Modelling
ANN	Artifical Neural Network
ARR	Arrival
ATFM	Air Traffic Flow Management
ATM	Air Traffic Management
BN	Bayesian Network
BiGRU	Bidirectional Gated Recurrent Unit
BusinessPax	Business-class Passengers
CCM	Convergent Cross Mapping
CDM	Collaborative Decision Making
CMI	Conditional Mutual Information
CN	Connection Network
CO2	Carbon dioxide
CPN	Cancellation Propagation Network
DCN	Delay Causality Network
DEP	Departure
DN	Delay Network
DC-SIC	Delay Causality Strong & Independent Causality
DPT-BN	Delay Propagation Tree - Bayesian network
DT	Time
4DTA	4D Trajectory Adjustments
EcoPax	Economy Passengers
EcoMaxPax	Premium Economy Passengers
ETD	Estimated Time of Departure
ELU	Exponential Linear Unit
EM	Expectation-Maximization algorithm
ESPL	Exponentially Truncated Shifted Power Law
ETFMS	Enhanced tactical flow management system
FirstPax	First-class Passengers
GANN	Graph Attention Neural Network
GAT	Grapth Attention Network
GBDT	Gradient Boosting Decision Tree
GCN	Graph Convolutional Neural network
GCKI-SPI	Granger Causality Kernel - SPI
GCT	Granger Causality Test
IID	Independent and Identically Distributed
LightGBM	Light Gradient Boosting Method

Abbreviation	Definition
LM	Levenberg Marquart (optimization algorithm)
LSTM	Long short-Term Memory
MAE	Mean Average Error
MAPE	Mean Absolute Percentage Error
MCDM	Multiple-Criteria Decision Making
MGT	Minimum Ground Time
ML	Machine Learning
MLCN	Multi-Layer flight Connection Network
MLIL-NN	Multi-Level Input Layer Neural Network
MLP	Multi-Layer Perceptron
MSTAGCN	Multiscale Spatial-Temporal Adaptive Graph Convolutional Neural Network
NewCat	Network-wide Congestion Assessment Tool
NN	Neural network
NOC	Network Operations Center
PAX	Passengers
REACT	Reactionary
RMSE	Root Mean Square Error
RNGC	Refined Nonlinear Granger Causality
SCHED	Scheduled
SHAP	SHapley Additive exPlanations
SPL	Shifted Power Law
SPI	Systematic Path Isolation
STA	Scheduled Time of Arrival
STD	Scheduled Time of Departure
SVR	Support Vector Machine
TOPSIS	Technique for Order Preferred by Similarity to Ideal Solution
UTC	Coordinated Universal Time
WLR	Weighted Linear Regression
XGBoost	Extreme Gradient Boosting

I. Introduction

In the interconnected world of air transportation, delays are rarely isolated incidents. Instead, they propagate through airline networks, affecting a wide range of stakeholders, including passengers, airlines, and airport operations. This thesis delves into the complexities of delay propagation within passenger airline networks, examining where and how delays spread and their severity on airline operations. By focusing on hub-and-spoke networks, where flights are routed through central hub airports, the research explores the vulnerabilities of these systems to cascading delays. Such delays not only impact the hub but ripple through connecting flights, amplifying disruptions.

Swiss International Air Lines (SWISS), Switzerland's flag carrier and a prominent hub-spoke operator, provides the foundation for this thesis. Collaboration with SWISS enables a detailed exploration of operational challenges posed by delays. The efficiency of spoke airports within the network plays a crucial role in mitigating delay propagation, underscoring their importance alongside the central hub. Delays originating at spoke airports or upstream flights often trigger a chain reaction, leading to widespread network disruptions. Although buffer times are integrated into airline schedules to address such issues, insufficient buffers can aggravate the problem, causing missed connections and affecting the daily operational strain of the network. These dynamics highlight the need for advanced models to better understand and predict delay propagation.

The thesis focuses on reactionary delays, which stem from earlier disruptions and present a unique challenge within hub-and-spoke networks. These delays are particularly critical in European airline operations, where thousands of flights daily experience significant disruptions. Understanding the distribution and impact of reactionary delays is essential, as even minor delays can escalate, straining operational efficiency and passenger satisfaction. The study aims to analyze how delays propagate through interconnected flights and assess the potential of advanced predictive models to address this issue. This includes examining the role of confidential airline data and ensuring model outputs are interpretable for operational decision-making.

The aim of the Master thesis, meaning the research question is the following:

Research Question

To what extent can we accurately determine the reactionary delay distribution over a fleet network, taking into account the effects of spoke airports?

The structure of this report is divided into two main sections. The first section, **II. Scientific Article**, presents the scientific article detailing the research conducted throughout the thesis, the results and discussion. The second section, **III. Literature Study**, covers the literature review conducted beforehand, which aimed to analyze previous studies, identify the research gap, and ultimately define the research question.

II. Scientific Paper

GATA: Graph Attention Networks in Aviation

Using a graph-based model to analyze the distribution of reactionary delays across a fleet network

Constança Miranda de Andrade Veiga[1,2]

Supervisors: Marta Ribeiro[1], Marie Carré[2]

[1] Department of Control & Operations, Faculty of Aerospace Engineering, TU Delft

[2] Operations Research and Air Traffic Management, Swiss International Air Lines

Abstract

Reactionary delays are a critical challenge in airline operations, especially within hub-and-spoke networks, where disruptions at spoke airports can propagate and amplify throughout the fleet. This study evaluates the capability of a Graph Attention Network (GAT) model to predict reactionary delay distributions within a fleet network Using operational data from Swiss International Air Lines' short-haul fleet, the GAT model integrates node-level features, such as flight-specific parameters, and edge-level features, including rotational dependencies and passenger connections, to capture the spatial-temporal dynamics of delay propagation. The learnt attention weights provide further insights into which flights and connections are most critical. The GAT model achieved a reliable predictive accuracy, particularly on medium-delay days, of a root mean squared error of 15.59 minutes and a mean absolute error of 10.50 minutes. However, the model's performance declines in scenarios involving irregular disruptions or extreme delays, highlighting its reliance on routine patterns. Nonetheless, the results show promise in integrating GATs for delay predictions and identification of critical flights and connections into airline operation tools.

1 Introduction

Airline punctuality is critical not only for operational efficiency but also for passenger satisfaction and cost control. Delays in air traffic operations can be categorized into primary delays and reactionary delays. Primary delays are initial disruptions caused by factors such as technical issues, weather conditions, or air traffic control restrictions. Reactionary delays, on the other hand, occur when earlier disruptions affect later flights, and they can significantly amplify the impact of the initial delay [12].

Reactionary delays present a unique challenge, especially within hub-and-spoke airline networks where spoke airports can introduce unpredictable complexities due to arrival delays and extended ground times [15]. Spoke airports often have limited resources and operational constraints, making them more susceptible to delays that can propagate through the network. In the European air traffic network, reactionary delays account for approximately 45% of total delays, sometimes exceeding primary delays in their cumulative effect [13]. This is particularly concerning because even minor disruptions can escalate, affecting numerous downstream flights and leading to significant operational strain [18].

Understanding the distribution and impact of these reactionary delays is very important. Hub-spoke networks suffer more from reactionary delays than point-to-point networks because delays originating at spoke airports can quickly spread to hub airports, which are central nodes connecting multiple routes [11]. This interconnectedness means that a single delay at a spoke airport can have a ripple effect, causing widespread disruptions across the network. By contrast, point-to-point networks are less susceptible to such cascading delays due to their less interconnected nature.

This research explores the potential of a Graph Attention Network (GAT) model to accurately determine reactionary delay distributions within a fleet network, specifically examining the role that spoke airports play in delay propagation. To achieve this goal, several sub-questions are formulated. First, the extent to which reactionary delays contribute to operational strain across airline networks is assessed, with particular attention to how delays propagate through interconnected flights. Additionally, we explore whether incorporating key categories of confidential airline data can address the challenges of quantifying uncertainty in reactionary delay predictions is also explored, thereby improving model accuracy across diverse airport operations. Another focus is on the interpretability of model outputs, especially regarding feature importance, to ensure usability for operational decision-making. Finally, the possibility of creating a delay prediction model that supports real-time updates without sacrificing predictive accuracy is examined, enabling airlines to respond more effectively to rapidly evolving operational scenarios.

This research aims to address limitations in current delay prediction methodologies by adopting a Graph Attention Network (GAT) model that builds upon more traditional data-driven approaches. Since delays propagate through interconnected flights within an airline network, features from neighboring flights are crucial for accurate prediction. The GAT architecture is particularly suited for this purpose, as it dynamically adjusts the importance of connected nodes (flights) based on evolving conditions, enabling the model to capture both temporal and spatial dependencies. Unlike traditional Graph Convolutional Networks (GCNs), which assign static importance to connections, GATs continuously update the relevance of connected flights, allowing the model to more effectively capture shifting patterns and relationships within the network. Additionally, GATs can better incorporate edge features, adding a layer of detail to the model's understanding of flight-to-flight interactions.

By leveraging this dynamic, graph-based approach and utilizing real operational data from Swiss International Air Lines, the model can address the complex, real-time nature of delay propagation, especially in hub-and-spoke networks where spoke airports play a critical role. The study analyzes the entire short-haul fleet of Swiss International Air Lines (SWISS), comprising of approximately 350 flights per day connecting Zurich Airport, the central hub, to numerous spoke airports across Europe. The dataset includes detailed flight information such as scheduled and actual arrival and departure times, minimum ground times, aircraft types, and confidential connecting passenger data. Furthermore, the attention weights within the GAT model highlight the most critical flights, offering insights into which connections have the greatest impact on delay propagation. This interpretability is essential for operational decision-making, as it allows Swiss Airlines to identify and prioritize interventions on key flights that could mitigate widespread disruptions and enhance overall network performance.

The remainder of this paper is organized as follows: Section 2 discusses related work and outlines the unique contributions of this study. Section 3 presents the problem and research question. Section 4 details the methodology used, including data sources, model architecture, and training process. Section 6 presents the results, highlighting key performance metrics, feature importance, and examining the model's effectiveness across various delay scenarios. In Section 7, the model's performance with real-world airline data is validated, while in Section 8 the results and practical implementations are discussed, followed by recommendations for future research.

2 Related Work

Predicting delay propagation in air traffic networks is a complex challenge that has been approached through various modeling techniques over the years. These models can be categorized into mathematical, statistical, and machine learning approaches, each offering unique strengths in capturing the intricacies of delay dynamics. Subsection 2.1 explores the mathematical and statistical methods, while Subsection 2.2 focuses on the machine learning techniques. The section concludes by highlighting current research gaps and outlining the direction of this study.

2.1 Mathematical and Statistical Methods

Early efforts to understand delay propagation relied on mathematical models aimed at quantifying the cascading effects of delays within air traffic networks. A foundational concept in this domain is the "delay multiplier" introduced by Beatty et al., which estimates the systemic impact of an initial delay as it propagates through connected flights [2]. This concept helps in understanding how delays can amplify across a network. Another significant mathematical approach involves Monte Carlo simulations that apply statistical distributions to model the variability in ground processes and their start times [14]. These simulations are known for their statistical approach and flexibility across various delay scenarios, enabling the estimation of the probability of different delay outcomes under uncertain conditions.

Building on these mathematical foundations, more sophisticated models like Delay Propagation Trees and Bayesian Networks have been developed to account for the non-independent and identically distributed (non-IID) nature of flight delays. These models capture complex causal relationships among various factors contributing to delays, enabling a more detailed analysis of how delays spread across the network. For example, a study using a Delay Propagation Tree model with a Bayesian Network (DPT-BN) examined multiple connecting sources, aircraft, cabin crew, pilots, and passenger connections to model delay propagation in an airline network [22]. Using a two-year dataset from an Asia-Pacific airline, the model reconstructed network connections and evaluated delay dynamics. The incorporation of conditional probability distributions for each delay source allowed the model to account for uncertainties without relying on deterministic assumptions.

While these mathematical models provide valuable insights, they often lack detailed validation on diverse datasets. Studies like [22] focus primarily on a single network subset, limiting the ability to generalize findings to other networks. Testing the models on additional networks or under varying operational scenarios could better validate their robustness.

Statistical methods have also played a crucial role in predicting delay propagation. These models often involve regression analysis and time-series forecasting, leveraging historical delay data to identify patterns and probabilistic factors that contribute to delays. Linear regression models, for instance, have been used to capture the relationship between early and later delays in the day, providing insights into critical periods for delay evolution at different airports [5]. This approach is favored for its simplicity and ease of interpretation, providing insights into the root causes.

Advanced statistical techniques include Granger causality and its extensions like Refined Non-linear Granger Causality (RNGC). These methods have been used to construct Delay Causality Networks (DCNs) that map the flow of delays across airports, offering a deeper understanding of the dynamics of delay spread [10, 16]. Propagation Trees have also been used to track the spread of delays from individual flights through a network, identifying critical junctures where delays are most likely to propagate.

Despite their effectiveness, these statistical models often lack consideration of real-time data and dynamic operational factors. For example, [5] could be improved by incorporating real-time weather conditions and live traffic patterns to enhance adaptability and accuracy. Similarly, studies like [10] and [16] do not account for seasonal and peak-period variations or important factors like ground-handling features, which can significantly impact delay propagation.

2.2 Machine Learning

With the rise of big data and increased computational power, machine learning (ML) models have become prominent in predicting delay propagation due to their ability to uncover complex patterns within large datasets. Traditional methods struggle to capture the non-linear and interdependent factors influencing delays, whereas ML models can learn from extensive data, identifying intricate relationships among various features. Supervised learning techniques such as Gradient Boosting Decision Trees (GBDT) and Random Forest have been widely adopted for their capacity to handle complex, nonlinear relationships [1]. For instance, studies have demonstrated that LightGBM, a variant of GBDT, outperforms other models in predicting take-off times, especially when processing vast amounts of flight data [7].

A notable study conducted by EUROCONTROL proposed an ensemble machine learning approach (PETA) to improve the estimated time of arrival predictions for flights. The model performed well closer to departure time, achieving a Mean Absolute Error (MAE) of 12.1 minutes. However, its performance declined for long-term predictions, with an MAE of 19.9 ± 25.1 minutes at 4–6 hours before departure and 19.6 ± 26.9 minutes beyond 6 hours. The study utilized an extensive dataset covering a three-month period and included all intra-European Civil Aviation Conference flights, leveraging historical traffic and meteorological data to enhance prediction accuracy [9].

Another example is the application of Random Forest Regressor (RFR) to predict departure delays across Colombia's airport network, analyzing data from over 350,000 flights [1]. The model used an extensive dataset including scheduled and actual departure times, delay durations, airline information, and IATA delay codes. It achieved a Root Mean Square Error (RMSE) of 33.8 minutes, which the authors suggest indicates effectiveness in handling the variability inherent in delay prediction. However, in a real-world airline application, particularly for short-haul operations, this level of accuracy would not meet the precision required to support operational decision-making effectively. While these ML models demonstrate strong predictive performance, they often lack interpretability, which can be a drawback for operational decision-making. Models like Random Forest are inherently "black-box," making it challenging to understand which factors most influence delay predictions.

Deep learning models, particularly neural networks, have further advanced the field by capturing both spatial and temporal dependencies in-flight data. Convolutional Neural Networks (CNNs) [4] and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks [19], have been employed to model the complex relationships between inputs (e.g., weather conditions, airport congestion) and outputs (e.g., delay times). For example, [19] uses an LSTM to predict delays at the airport-level by modeling the temporal dynamics within each airport. By using historical delay data, the model forecasted future delays at multiple look-ahead intervals, achieving RMSEs ranging from 6.31 minutes (30-minute look-ahead) to 7.73 minutes (180-minute look-ahead).

Another study introduced the Multiscale Spatial-Temporal Adaptive Graph Convolutional Neural Network (MSTAGCN) to predict flight delays across a network of 74 major Chinese airports, covering over 90% of the nation's air traffic [4]. The model incorporated spatial and temporal dependencies to capture delay dynamics, achieving an RMSE of 10.37 minutes for one-hour-ahead predictions.

Despite their successes, these deep learning models often lack real-world application testing. Many results are based on historical data and simulated conditions. Live implementation is important to assess the models' real-time performance, especially in unpredictable scenarios like sudden weather changes or peak traffic. Additionally, some models make simplifying assumptions, such as treating delays as independent across aircraft sequences [8], which may not reflect the interconnected nature of delay propagation due to shared resources.

A notable advancement in research is the use of Graph Attention Networks (GATs), which excel at modeling complex relationships within networks by applying attention mechanisms to focus on the most relevant nodes and edges, such as airports and flights in an airline network. GATs dynamically adjust to changing network conditions, making them ideal for real-time delay prediction by capturing spatial-temporal interactions that reveal how delays propagate through interconnected flights and airports. For instance, the Spatial-Temporal Gated Multi-Attention Graph Network (STGMAGNet) was developed to predict network-wide airport delays up to 24 hours in advance [26]. The model, designed to incorporate spatial-temporal correlations and weather impacts, used data from 75 major U.S. airports. STGMAGNet addressed uncertainties like weather, airport congestion, and fluctuating schedules through an external module, achieving an RMSE of approximately 30.7 minutes for arrival delays and 34.1 minutes for departure delays. In this case, airports are treated as nodes and connections as flights, which is suitable for airport-level predictions but not directly applicable to airline-level applications. In the context of an airline, flights are more appropriately modeled as nodes, as this allows for the prediction of both their arrival times and the extraction of attention weights to understand their impact on delay propagation.

2.3 Research Gap

While various approaches have been developed to predict delay propagation, there are key research gaps that this thesis seeks to address. First, existing models often lack access to connecting passenger data due to confidentiality constraints, limiting a comprehensive understanding of delay propagation. Another gap is the focus on U.S.-based air traffic data, with limited studies on European networks, suggesting the need for more diverse geographic analysis. Furthermore, current models rarely incorporate multiple interconnected factors, such as aircraft, crew, and passenger connections, in a unified approach. This thesis also addresses the need for a dynamic prediction model with a real-time, adaptable prediction horizon that supports operational decision-making in rapidly changing conditions. Lastly, the spatial-temporal dynamics of delay propagation remain underexplored, with limited emphasis on how delays evolve across different network regions. This paper aims to contribute to these areas by developing a comprehensive, dynamic model that accounts for both multi-factor dependencies and the spatial-temporal aspects of delay propagation.

3 Problem Definition

Delay propagation in air traffic networks poses significant challenges to airlines, airports, and passengers. Reactionary delays, where an initial delay causes subsequent delays in connected flights, are particularly disruptive. These delays can cascade through the network, amplifying the impact of the original disruption and affecting the overall efficiency of airline operations.

Research Question

To what extent can we accurately determine the reactionary delay distribution over a fleet network, taking into account the effects of spoke airports?

This thesis aims to address the gaps described in Section 2.3 by developing a comprehensive, dynamic model that accounts for both multi-factor dependencies and the spatial-temporal aspects of delay propagation. By leveraging the capabilities of Graph Attention Networks, the aim is to capture the complex, interconnected nature of delays in air traffic networks. A significant feature of the GAT model is its computation of attention weights, which represent the most critical flights and connections within the network. These weights enable the model to focus on flights that have a greater influence on delay propagation, effectively identifying key nodes that contribute most to reactionary delays.

By focusing on the spatial-temporal dynamics and incorporating multiple interconnected factors, this research aims to develop a model that not only predicts delay propagation more accurately but also identifies the most critical flights through the analysis of attention weights. This approach will deepen the understanding of delay mechanisms and ultimately contribute to more resilient and efficient airline operations.

4 Methodology

GATs are inherently complex, making it crucial to thoroughly understand their architecture in order to effectively comprehend their functionality and identify appropriate applications. This section is structured as follows. Subsection 4.1 describes the setup of the graph model, followed by a summary table of features in Subsection 4.2. The architecture of the GAT model is outlined in Subsection 4.3. Subsection 4.4 details the training procedure. Lastly, Subsection 4.5 discusses the hyperparameter optimization conducted for the training.

4.1 Problem Definition and Graph Representation

Predicting flight delay propagation requires capturing the complex network of dependencies between flights within an airline's network. A graph-based representation is well-suited for this purpose, as it models the interconnections and interactions among flights, enabling a structured analysis of how delays spread across the network. By representing the fleet network as a graph, the dynamics of delay propagation can be effectively simulated and studied.

Let $G = (V, E)$ be a directed graph where:

- $V = \{v_1, v_2, ..., v_n\}$ represents the set of flights, with each node v_i corresponding to a unique flight. Nodes serve as individual flight representations within the network, encapsulating specific flight attributes and statuses.
- $E = \{e_{ij}\}$ represents the set of edges, where an edge e_{ij} exists if flight v_i is connected to flight v_j through operational relationships. Edges capture the dependencies between flights that may influence delay propagation.

In this graph representation, nodes and edges represent the fundamental elements of the flight network. Each flight (node) can have multiple connections (edges) to other flights, which reflect either shared aircraft (rotation) or passenger transfers (connecting passengers). A small representation of the graph structure is depicted in Figure 1. Connections based on passenger transfers can involve multiple passengers and flights, creating a web-like network structure. This structure captures the complex ways in which delays in one flight can propagate through numerous other flights, resembling a complex, interconnected web.

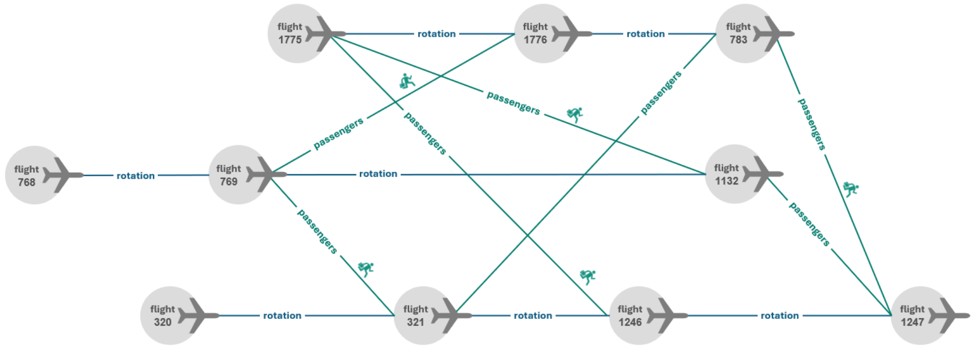


Figure 1: Illustration of the flight network structure, showing nodes (flights), edges (connections), and their interdependencies.

More specifically, a rotation connection indicates that flights share the same aircraft, meaning the aircraft operates flight v_i and then proceeds to operate flight v_j. This connection captures the dependency where a delay in flight v_i can directly impact flight v_j. Additionally, Connecting Passengers signifies that passengers from flight v_i are transferring to flight v_j. Delays in the initial flight can cause missed connections or require holding the subsequent flight.

Each node v_i is associated with a feature vector $\mathbf{h}_i$ that includes both categorical and numerical attributes pertinent to the flight. Similarly, each edge e_{ij} is associated with a feature vector $\mathbf{e}_{ij}$, capturing the specifics of the relationship between connected flights, such as the type of connection and the number of connecting passengers.

By modeling the flight network as a graph with these nodes and edges, it is possible to effectively simulate and analyze how delays propagate through the network due to the interconnected nature of flights.

4.2 Node, Edge and Graph Features

Each component of the GAT model utilizes specific node, edge, and graph features to capture complex relationships within the network. Node features capture individual flight characteristics as detailed in Table 1, while edge features describe interactions between connected flights, such as passenger connections or rotational dependencies (Table 2). Additionally, graph features, shown in Table 3, capture temporal aspects of the dataset, helping the model in accounting for both daily and monthly operational cycles, such as capturing high-peak seasons in the Summer months.

The feature vector $\mathbf{h}_i$ for each node v_i is composed of the following components:

Table 1: Features for Flight Nodes

Features	**Type**
Aircraft type	Categorical
Departure airport	Categorical
Arrival airport	Categorical
Delay codes	Categorical
Estimated time of departure (ETD)	Numerical (time)
Scheduled departure time	Numerical (time)
Scheduled arrival time	Numerical (time)
Actual departure time	Numerical (time)
Actual arrival time	Numerical (time)

The feature vector $\mathbf{e}_{ij}$ for each edge e_{ij} (representing the connection between node i and node j) includes:

Table 2: Features for Flight Edges (connections)

Features	Type
Connection type	Categorical
Number of connecting passengers	Numerical
Destination	Categorical
Alternative	Categorical
Rotation	Categorical
Crew	Categorical
Number of connections per class	Numerical
Connection time	Numerical (minutes)
Minimum ground time	Numerical (minutes)
Fleet buffer	Numerical (minutes)
Flight number	Numerical

Graph features are attributes that characterize the entire graph rather than individual nodes or edges. In this model, the graph features include the day and month associated with each graph, where each graph represents a single day. Defining day and month as graph features allows the model to capture seasonal patterns and daily trends that may influence flight delays. This setup is advantageous because it enables the model to recognize time-based effects that impact all flights on a given day, such as increased demand during specific months or differences between weekday and weekend schedules, rather than attributing these effects to individual flights or connections.

The graph feature vector **g** includes:

Table 3: Graph-Level Temporal Features

Features	Type
Day	Numerical (time)
Month	Numerical (time)

In Subsection 5.1, it is shown how the temporal features are encoded such that the model understands their cyclical nature.

4.3 Model Architecture

The GAT model leverages multi-layer attention-based graph convolutions to predict arrival times by learning from the relationships and dependencies among flights. The model is implemented using the PyTorch Geometric library, with three attention layers that progressively refine the node-level embeddings through neighborhood aggregation. The architecture of the model has two primary components that are further explained in the following sub-sections and can be seen in Figure 2.

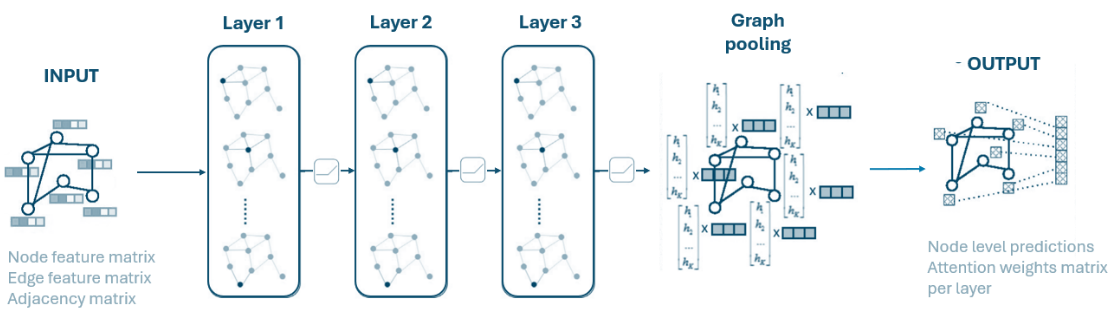


Figure 2: Model architecture used during training and predictions

4.3.1 Attention Mechanism

The first layer in the model applies the graph attention mechanism, which assigns attention coefficients α_{ij} to each edge e_{ij}. These coefficients quantify the importance of neighboring nodes in determining the representation of a target node. By learning which neighboring nodes are most impactful, the model can focus on the relevant relationships and filter out the least important connections in the network. The attention coefficients are computed as follows;

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k\right]\right)\right)} \tag{1}$$

where a is a weight vector, W is a learnable weight matrix, $\mathcal{N}(i)$ denotes the neighbors of node i, and $\|$ represents concatenation [20]. These coefficients determine the influence of neighboring nodes on the target node during the aggregation process. The LeakyReLU activation function introduces a small, non-zero gradient for negative input values, allowing the model to retain information from these inputs rather than setting them to zero. This helps prevent issues with inactive nodes that can occur with standard ReLU, where nodes stop updating during training [24].

To explain Equation 1 in simpler terms, the attention weights of the network (Figure 3) are computed as follows. Each node feature vector is transformed by the matrix W, enabling the model to learn unique embeddings for nodes in relation to their neighbors. The concatenated features of a target node and its neighbor are scored by the weight vector a, yielding a scalar value that indicates the importance of their relationship. This score is then normalized across all neighbors using a *softmax* function, ensuring that the attention coefficients sum to 1 within each neighborhood. This normalization balances the influence of each neighbor, allowing the model to focus dynamically on the most relevant relationships.

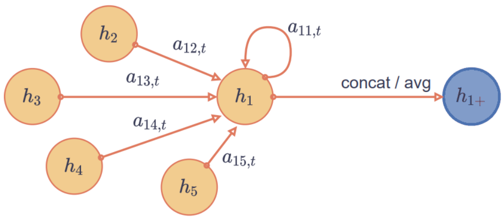


Figure 3: GAT neighborhood attention for Node 1: attention coefficients from neighbors are used to compute its aggregated feature (h1+).

In Figure 3, the GAT attention mechanism, h_i represents the hidden feature vector of node i at a specific layer, capturing the node's current state. The attention coefficient $a_{ij,t}$ indicates the importance of neighboring node j to node i, allowing the model to weigh each neighbor's influence based on learned relevance. Self-loops (e.g., $a_{11,t}$) enable nodes to retain their own information in addition to neighboring influences. After calculating attention-weighted features from neighbors, these are combined via concatenation or averaging to form the updated node representation h_{1+} for the layer. Here, t denotes the layer or step in the model.

Through this process, the attention mechanism allows the GAT model to focus more on flights or nodes that have a greater impact on the target flight's arrival time. For instance, if a flight is influenced by delays in certain airports or connections, the attention mechanism will assign higher weights to those connections, thereby amplifying their contribution to the prediction.

4.3.2 Layers

Following the attention mechanism, three Graph Attention Convolutional (GATv2Conv) layers are used to further process the aggregated node features. Each layer utilizes attention heads to focus on the most relevant neighbors of each node, thereby capturing important spatial and temporal dependencies.

The GATv2Conv layers improve the power of the model by allowing dynamic weights for the neighbors [3]:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}^{(l)} \mathbf{h}_j^{(l)}\right) \tag{2}$$

where l indicates the layer number, σ is a non-linear activation function, and α_{ij} are the attention weights from the previous layer. Furthermore, h represents the hidden feature vector (or embedding) of a node in the graph at a specific layer l. For example, $h_i^{(l)}$ is the feature vector of node i at layer l, while $h_i^{(l+1)}$ is the updated feature vector at the next layer. And, $\mathcal{N}(i)$ denotes the neighborhood of node i, which includes all nodes j that are directly connected to i in the graph. These neighboring nodes contribute to the update of i's feature vector through the attention mechanism.

A layer applies a weighted aggregation of features from neighboring nodes, with α_{ij} representing the attention weights (or importance) of each neighbor j relative to node i [3]. In this implementation, the Exponential Linear Unit (ELU) is used as the non-linear activation function. ELU ensures smooth gradients for negative inputs and reduces the likelihood of inactive node updates, which enhances the model's convergence and overall learning capability [6]. The non-linear activation function σ is applied after aggregation to introduce non-linearity into the model. Introducing non-linearity is an important feature of neural network architectures as it allows the model to approximate complex, non-linear relationships within the data. This capability enables the network to effectively capture intricate patterns and dependencies, such as spatial-temporal interactions in delay prediction, that cannot be modeled with purely linear operations.

The first layer takes a concatenated node and graph-level features as input, extending each node's feature dimension. With 16 attention heads, it aggregates neighborhood information in detail, producing an output transformed to 16 times the hidden dimension. This output feeds into the second layer, which uses 8 attention heads to refine node representations by focusing on the most relevant features. This multi-head attention principle can be visualized in Figure 4.

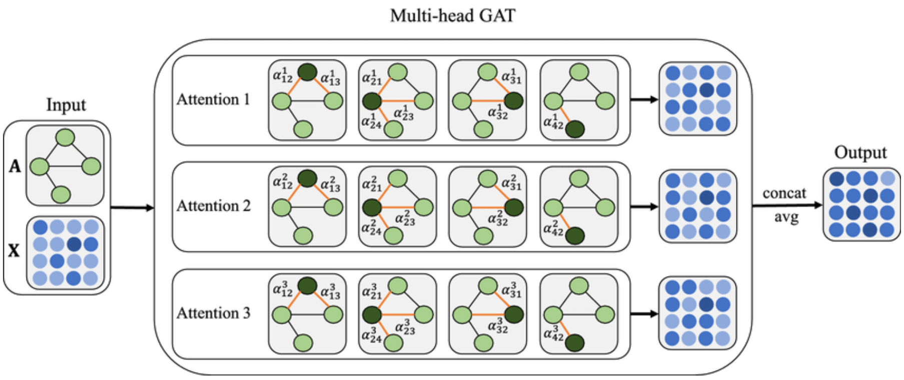


Figure 4: Illustration of multi-head attention in a GAT model [23]

The final layer of the network, with a single attention head, is a linear layer that outputs the predicted arrival time $\hat{t}_{\text{arr},i}$ for each flight v_i [3]:

$$\hat{t}_{\text{arr},i} = \mathbf{w}^\top \mathbf{h}_i^{(L)} \tag{3}$$

where w is a weight vector, and $h_i^{(L)}$ is the final representation of the node after L layers.

The GAT model's structure, with its layered attention mechanisms and integration of edge features, allows it to capture complex dependencies across flights more effectively than traditional models, which primarily rely on isolated flight data. By utilizing these advanced features, the GAT model gains a detailed understanding of the interconnected flight network, improving its accuracy in predicting arrival times.

4.4 Training

The GAT model is trained using historical flight data from SWISS, covering the period from January 1st to September 30th 2024. This dataset includes detailed records of flight schedules, actual departure and arrival times, connecting passenger data, and various other features for the airline's short-haul fleet, with approximately 350 flights per day. Using this extensive data, the model learns to predict delay propagation across interconnected flights, enhancing understanding of how delays impact subsequent flights.

To support accurate arrival time predictions, the GAT model was trained over 1000 epochs with batch processing and iterative optimization. Training and validation data loaders were established with a batch size of 10 (with shuffling) for training and 1 for validation. Moreover, the model architecture integrates both node and edge features, as well as graph-level features, capturing the complex interactions within the flight network that influence delay patterns.

For optimization, Mean Squared Error (MSE) was selected as the loss function, minimizing the difference between predicted and actual values. The Adam optimizer, with a learning rate of 0.002, was used to ensure stable convergence through adaptive weight updates. Additionally, dropout, applied at 0.1 after each of the first two layers, enhanced generalization, while Exponential Linear Units (ELU) as activation functions enabled the model to capture non-linear relationships inherent to delay dynamics.

During training, each epoch involved forward passes through mini-batches, calculating the loss on labeled nodes only, and backpropagating gradients for weight updates. Validation was conducted every five epochs with dropout disabled to ensure reliable assessment, logging RMSE, MAE, and MAPE metrics to evaluate prediction accuracy on unseen data. Moreover, attention weights, particularly in the second and third layers, were extracted as they highlight which connections between flights most influence arrival times. This training approach provided valuable insights into model dynamics, and the trained parameters were saved and preserved for future predictions.

4.5 Hyperparameter Optimization

As previously mentioned, the model used employs a GAT model with three layers to predict flight arrival times by capturing spatial dependencies through attention mechanisms on graph-structured data. The model architecture was initially designed with two layers, a learning rate of 0.005, a weight decay of 0.0005, and a dropout rate of 0.6, as indicated in the relevant literature. Table 4 details the values used for the hyperparameter tuning of the model.

Table 4: Summary of GAT Model Hyper-parameter Fine-tuning.

Hyper-parameter	**Value(s)**	**Optimum**	**Description**
Batch Size	1–100	10	The number of examples used in each iteration before updating the model parameters. A smaller batch size improved accuracy but increased training time slightly.
Epoch	100–1000	500	The number of complete passes through the training dataset. A larger number of epochs increases the risk of overfitting, while a smaller number risks underfitting.
Drop-out	0.1–0.6	0.2	The dropout technique is used to prevent overfitting by randomly setting a fraction of input units to zero at each update during training.
Learning Rate	0.001–0.1	0.002	Controls the step size at each iteration while moving toward a minimum of the loss function. A lower learning rate can provide more precise convergence but slows down training.
Number of Layers	1, 2, 3, 4	3	The number of hidden layers in the model. A smaller number limited the model's capacity, while too many led to overfitting.

Through hyperparameter tuning, the final architecture was optimized to three layers, with each layer incorporating progressively fewer attention heads. Multiple attention heads in the initial layers capture diverse relational patterns, which are refined as the layers progress. A dropout rate of 0.1 was applied to prevent overfitting.

The transition from a two-layer to a three-layer architecture enhanced the model's ability to capture complex dependencies and intricate relationships, which are critical in networks where indirect connections are important. Attempts to increase the model depth to four layers resulted in a worse performance of the model due to the bias-variance trade-off. This means that increasing the model's capacity to four layers introduced higher variance, reducing generalization capability and leading to reduced performance.

Additionally, the final model configuration omits weight decay, which improves performance by allowing the model parameters to adjust more freely during training. While weight decay is often useful for preventing overfitting, in this case, it appeared to limit the model's ability to adapt to the underlying patterns in the data. The finalized three-layer configuration effectively balances model complexity and generalization, capturing critical predictive patterns in the data. This setup yielded the best performance among the tested configurations.

5 Data

The dataset used for training captures detailed flight and operational information for arrival time prediction, incorporating both node and edge attributes relevant to each flight. Each graph instance represents a network of flights in a day, with nodes corresponding to individual flights and edges capturing the connections (such as sequential flight rotations or passenger transfers) between them. The dataset includes features including scheduled and actual departure/arrival times (encoded trigonometrically), airport and flight characteristics, and delay-related factors (e.g., baggage, weather, ATC delays). These attributes are processed and standardized to enhance model learning and generalization.

The dataset is preprocessed to encode temporal features using trigonometric functions to capture their cyclical nature, as discussed in Subsection 5.1. Both node and edge features are normalized where necessary to ensure that the model can learn effectively from the data.

The graph structural metrics of the training and test datasets are summarized in Table 5. The training dataset spans 292 days and covers a broader range of structural characteristics, with an average of 592 nodes and 1552 edges per graph. The test dataset, by comparison, represents a more focused set of 15 days, with each graph containing around 342 nodes and 1443 edges on average. Both datasets exhibit scattered connectivity, with low graph density and clustering coefficients, reflecting the typical structure of flight networks.

Table 5: Graph Structural Metrics for Training and Test Datasets

Metric	**Train Dataset**	**Test Dataset**
Mean Degree	8.74	8.74
Max Degree	22.0	22.0
Min Degree	0.0	0.0
Graph Density	0.0091	0.0278
Average Clustering Coefficient	0.0020	0.0029
Average Nodes per Graph	592	342
Average Edges per Graph	1552	1443
Largest Graph (Nodes)	773	356
Smallest Graph (Nodes)	332	332
Largest Graph (Edges)	2076	1620
Smallest Graph (Edges)	850	1192

This table highlights the structural complexity of each dataset, showcasing the diverse node degrees and sparse nature of the flight network, which can influence model training and performance across various delay scenarios.

5.1 Trigonometric Encoding of Time Features

To effectively capture the cyclic nature of time-based features (such as scheduled and actual departure and arrival times), trigonometric encoding is applied. Times within a 24-hour period are cyclic, meaning that the difference between times, such as 23:00 and 01:00, should be small rather than large to reflect their proximity within the daily cycle. Trigonometric encoding addresses this by converting time features into sine and cosine values, enabling the model to recognize the cyclic continuity of time.

The trigonometric encoding for each time-based feature t (in minutes since midnight) is given by:

$$t_{\text{sin}} = \sin\left(\frac{2\pi t}{1440}\right), \quad t_{\text{cos}} = \cos\left(\frac{2\pi t}{1440}\right)$$

where 1440 represents the total number of minutes in a day. This encoding maps the time values to a unit circle, preserving the cyclic relationship between different times of day. For instance, values at the start and end of the day are close in angular terms, allowing the model to interpret temporal proximity accurately.

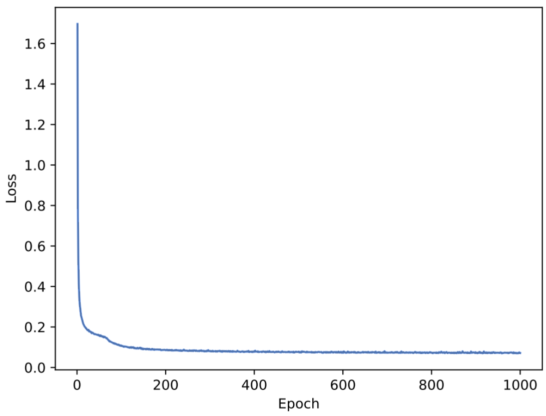
This approach improves the model's ability to learn from temporal patterns within the flight schedule, enhancing predictions across different times of day.

6 Results

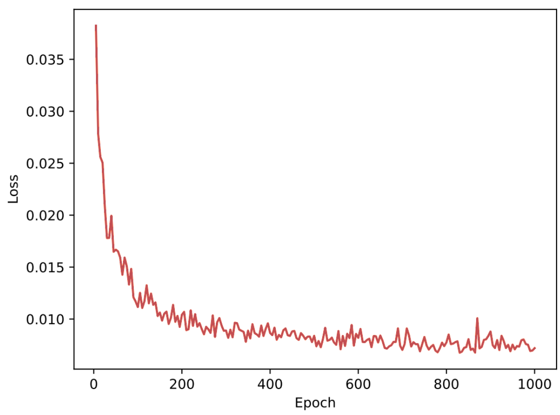
This section presents the model's performance, highlighting its training and testing accuracy, as well as evaluation metrics that measure prediction quality. Through an analysis of loss curves (Subsection 6.1), RMSE, MAE, and MAPE (Subsection 6.2), the model's generalization capability and its effectiveness across different delay scenarios is assessed. Subsection 6.3 depicts the evolution of the model's uncertainty as it is run throughout the day at different time points. Finally, in Subsection 6.4, feature importance and the correlation analysis amongst features are analyzed.

6.1 Training and Testing Performance

The training loss decreases rapidly during the initial epochs, stabilizing around a low value of 0.0702 by the 1000th epoch, as seen in Figure 5a. This behavior indicates that the model effectively captures the underlying patterns within the training data, achieving a robust fit without overfitting. The test loss (Figure 5b) follows a similar trend, reaching a stable value of 0.0071, which is lower than the training loss. This outcome suggests that the model generalizes effectively to unseen data, indicating that the model architecture is well-suited to this task. The lower test loss may imply that the test data is slightly less complex than the training data, or it may reflect the model's effectiveness in learning essential patterns while avoiding overfitting. Overall, these results highlight the model's strong generalization capabilities, showing that the chosen architecture and hyperparameters are effective in capturing data patterns, making the model reliable for predictions on new data.



(a) Train loss curve



(b) Test loss curve

Figure 5: Loss curves for training and testing datasets

6.2 Performance Measures

The Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) are used to evaluate the performance of the models.

The RMSE, MAE, and MAPE metrics calculated for the validation dataset indicate the prediction errors across all flights. These values represent the overall prediction accuracy for the entire day, made just before each flight day begins. The RMSE and MAE values are close at around 16.35 minutes, suggesting a consistent error margin across both metrics. The relatively low MAPE (2.61%) indicates that the model's percentage error remains minimal, pointing to a strong alignment between predicted and actual arrival times for the majority of flights.

For a closer examination, Figure 6 illustrates the RMSE per day, reflecting the model's performance across various days within the test period. Here, we observe that the daily RMSE fluctuates slightly but generally hovers around 15-20 minutes. This steady performance suggests that the model handles variations in daily flight data effectively. However, the error bars indicate some level of variability, which could be associated with specific high-delay days.

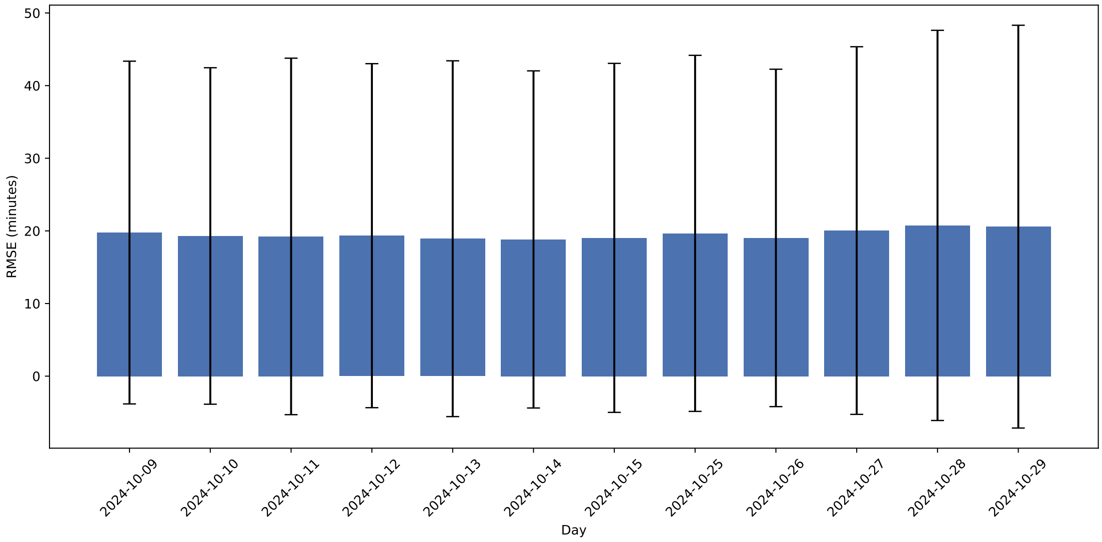


Figure 6: Mean and standard deviation of RMSE per day

When comparing these results with other studies, it is found that the model performs similarly well on low-delay days, where RMSE remains within the 15-20 minute range. Notably, higher error bars on certain days may imply increased difficulties in prediction accuracy due to factors like severe delays or operational disruptions. In line with prior studies, this model demonstrates robust prediction capabilities under typical conditions, with some performance drop-offs on days of extreme delays. Nonetheless, the overall RMSE remains within acceptable limits, making this model a useful tool for daily operational use.

Table 6 highlights the GAT model's performance across various delay day categories. The model performs best on medium delay days, with an accuracy of 85.7% and relatively low error metrics. This higher accuracy is likely due to the large number of medium delay examples in the dataset, which enables the model to learn patterns more effectively for moderate delays. In contrast, the accuracy for low delay days is lower because these instances are less frequent in the dataset, providing fewer examples for the model to learn from. High delay days remain challenging, as seen with the highest errors and lower accuracy (31.7%), reflecting the model's sensitivity to variability in delay intensity. Furthermore, the RMSE is consistently higher than the MAE across all delay day categories, highlighting the impact of outliers on the overall performance metrics.

Table 6: Performance Metrics by Delay Day Category

Test Days	**MAE [min]**	**RMSE [min]**	**Accuracy**
High delay days	27.39	37.56	31.7%
Medium delay days	10.50	15.59	85.7%
Low delay days	13.45	23.90	72.3%
All test days	16.30	26.52	65.3%

6.3 Uncertainty Throughout the Day

In order to evaluate the performance of the Graph Attention Network model in predicting flight arrival times along the day, the model was run every two hours between 02:00 and 16:00 UTC across multiple dates to assess how prediction errors evolve throughout the day. By analyzing the Root Mean Square Error (RMSE), average delay, and average number of nodes, the study aims to understand the model's strengths and areas for potential improvement in handling temporal and operational variability.

The RMSE values, which generally range around 25 minutes, remain stable across the time blocks, with only an occasional peak reaching 60 minutes as seen from Figure 7. Individual curves for different dates show some variability, reflecting the influence of day-specific factors on prediction performance. The average RMSE, however, follows a smooth and consistent trend, highlighting the model's robustness in handling diverse temporal conditions.

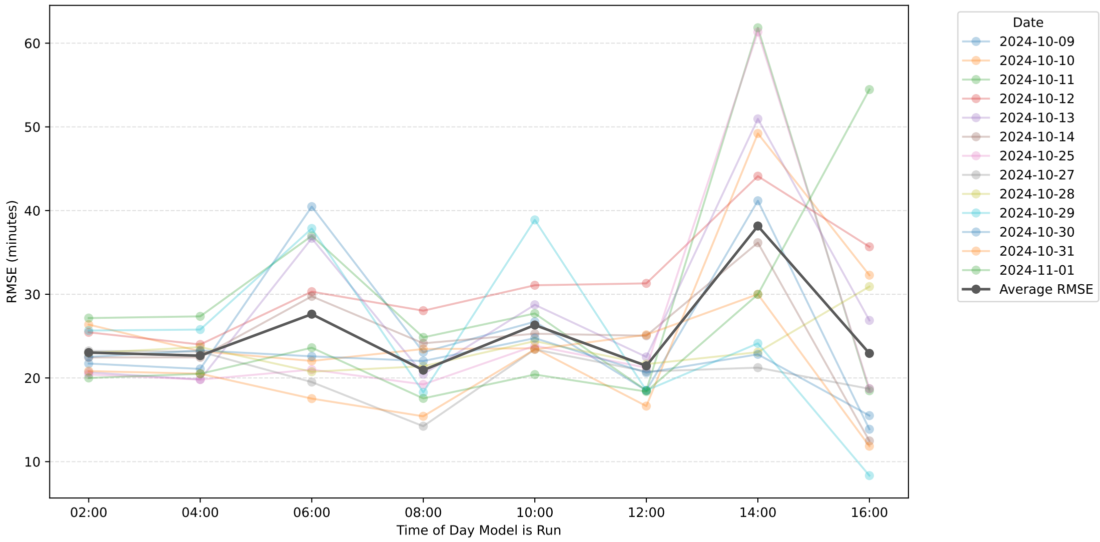


Figure 7: Prediction error as a function of inference time in UTC.

An interesting observation emerges when comparing the model's performance at 02:00 and 14:00 UTC. Despite having less updated information at 02:00, the model achieves a lower RMSE compared to when it's run at 14:00. This can be attributed to the fact that the model is trained on data available at 02:00, making it more effective at predicting the entire day's flights based on early information. At 14:00, although the model has access to more recent data, it only predicts flights from 14:00 onwards and faces increased complexity due to accumulated delays and operational disruptions that have occurred throughout the day. Furthermore, the average number of nodes, representing unique flight numbers, fluctuates across time blocks, peaking around midday and decreasing in the early morning and late afternoon (Figure 33b). Additionally, when observing the average delay in Figure 33a (Appendix A), there is a gradual increase from approximately 11 to 13 minutes as the day progresses. This trend likely reflects operational factors, such as accumulating disruptions during peak flight periods. Despite this increase, the GAT model maintains consistent performance, suggesting it effectively captures patterns in the data even under varying conditions.

Moreover, the higher RMSE at 14:00 UTC reflects the model's decreased performance in handling the greater variability and complexity of the network later in the day. This suggests that while the model is robust in predicting flights based on early-day information, its performance diminishes during peak operational periods. To further examine this, Figure 8 shows the variance of the prediction error. The shape of the boxplots in Figure 8 reveals the variation in model performance across the day. At 02:00, the compact box and shorter whiskers demonstrate minimal variability, indicating consistent prediction accuracy during early hours. Conversely, at 14:00, the broader box and extended whiskers reflect increased variability and a higher range of errors, showcasing the model's challenges in dealing with greater network complexities and accumulated delays later in the day.

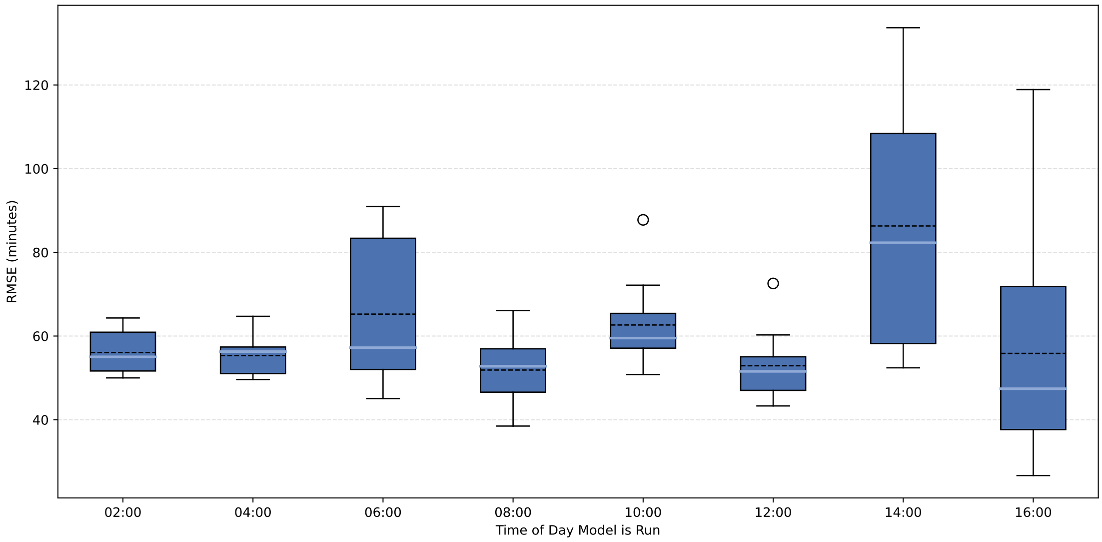


Figure 8: Variance in the prediction error as a function of inference time in UTC.

At 02:00 UTC, the model is tasked with predicting all the flights for the day, covering both peak and off-peak periods. This diverse mix of flight timings may contribute to error distributions that are less extreme, as the model's predictions include flights with varying levels of complexity and potential delays. In contrast, at 14:00 UTC, the predictions are limited to the remaining flights of the day. These flights are typically during peak operational periods and often reflect significant accumulated delays from earlier disruptions. This combination of peak-time pressure and delay propagation increases the complexity of predictions and likely explains the observed higher RMSE and variability in performance.

The results highlight both the key strengths and limitations of the GAT model. The stability in RMSE across time blocks demonstrates its ability to generalize well and handle temporal variation effectively, reflecting robustness in capturing relational and contextual patterns within the dataset. This stability is particularly noteworthy given the rising average delay and fluctuating node counts throughout the day, indicating that the model can adapt to varying levels of flight activity without significantly impacting prediction accuracy. However, the limited reduction in uncertainty throughout the day points to potential limitations in the model's sensitivity to temporal dynamics. While the consistent RMSE suggests that time of day may not be a major determinant of flight delays, it may also indicate that the model is not fully capturing important temporal relationships. Incorporating other features, such as weather conditions and air traffic trends, may enhance the model's sensitivity to time-based patterns and further reduce predictive uncertainty.

6.4 Features

An analysis of the model's features reveals important information about its strengths and limitations in prediction. This section provides an in-depth look at the key features used in the model, including an analysis of their importance (subsubsection 6.4.1) and their correlations with one another (subsubsection 6.4.2).

6.4.1 Feature Importance

GNN Explainer, a method often implemented within graph network frameworks such as PyTorch Geometric, provides interpretability for Graph Neural Networks (GNN) by identifying important substructures and features that contribute most to a model's predictions [25]. It does this by optimizing a mask over the graph's edges and features to maximize the mutual information between the prediction and the identified subgraph, effectively learning the smallest subset of edges and node features that strongly influence the model's output.

An initial plot of the node feature importance for the GAT model's prediction of arrival times was generated (Figure 34 in Appendix A). This plot highlights that the latest estimated arrival time (*ARR ACTUAL DT*) has the highest importance among all features. This outcome is expected, as predicting arrival time is the model's primary objective, making *ARR ACTUAL DT* a crucial baseline feature. It can also be observed that the delay codes have an average importance of zero because they are non-existent in the testing dataset used for this analysis. To better understand the impact of the other features, a filtered plot, Figure 9, was generated, excluding *ARR ACTUAL DT* and the delay codes to focus on the impact of the remaining features.

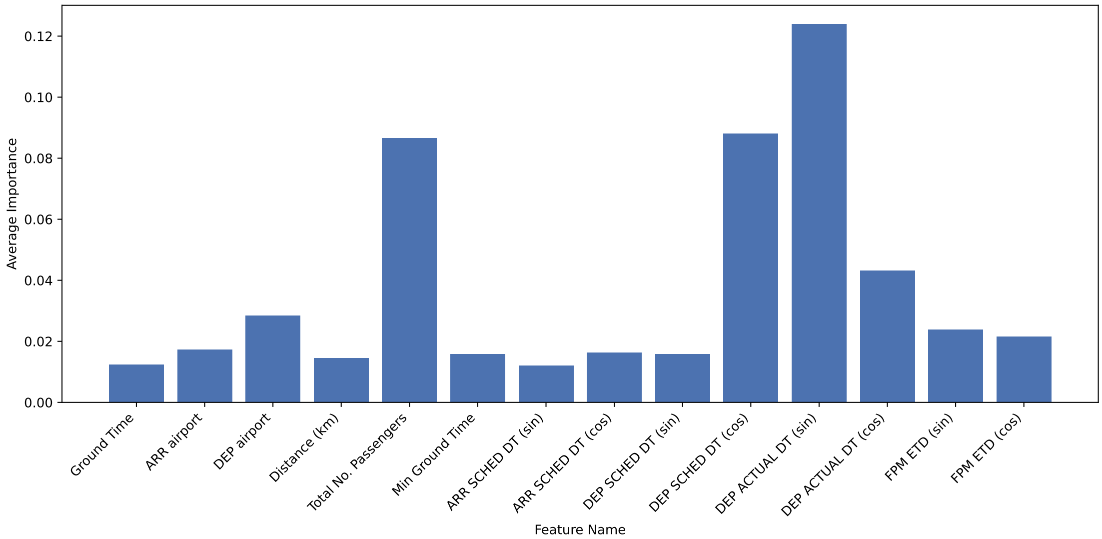


Figure 9: Average node feature importance across samples

From Figure 9, several observations arise regarding feature importance. Ground time and total number of passengers show relatively high importance, suggesting that both the aircraft's turnaround time and passenger load significantly influence arrival predictions. This aligns with operational patterns where increased passenger counts and ground handling requirements are known contributors to delays. Additionally, the departure airport also scores highly, indicating that certain airports may be more susceptible to delays, such as high-traffic hubs.

Scheduled time encodings represented by the encoded values for both scheduled arrival and departure times (e.g., *ARR SCHED DT (sin)* and *DEP SCHED DT (cos)*) hold moderate importance. These features capture daily and weekly cycles, such as peak travel hours, that can influence delay likelihood.

This analysis reinforces the structural and temporal dependencies within the flight network and highlights the operational factors that most significantly impact arrival time predictions. For instance, total number of passengers, a confidential feature to the airline, has a significant impact on the prediction.

In the analysis of edge feature importance, as presented in Figure 10, several observations become apparent regarding the influence of these features on predicting arrival delays. Connection type stands out, with Rotation connections showing notably higher importance than Passenger connections. This reflects the impact of consecutive flights sharing the same aircraft: if an earlier flight is delayed, the subsequent flight that uses the same aircraft will likely be delayed as well. This strong dependency highlights the importance of aircraft rotation, as delays in one segment can cascade into subsequent flights within the network.

Passenger-related features, such as the number of economy passengers (*EcoMaxPax*, *EcoPax*) and group sizes, demonstrate moderate importance. This suggests that flights with higher numbers of passengers, especially in economy class or traveling in groups, may require more processing time, impacting turnaround efficiency. Notably, the number of first-class passengers is relatively high in importance compared to other passenger types. This may be due to the fact that connecting flights are more likely to wait for first-class passengers, thereby increasing the likelihood of delays. On the other hand, features such as HON Member and Senator statuses show minimal importance, suggesting that frequent flier status has little effect on delay outcomes across the network.

Special service features, such as the number of wheelchairs (*No. Wheelchairs*) and the number of unaccompanied minors (*No. Unaccompanied Minors*), show low importance in the model. However, this may not necessarily indicate that these services are irrelevant to delay predictions. Instead, it could be attributed to their relatively low frequency in the dataset, with wheelchairs appearing in only about 10% of cases and unaccompanied minors in approximately 5%. The limited number of instances may prevent the model from accurately learning their impact on network-wide delays. Lastly, minimum connection time demonstrates moderate importance, highlighting the risk associated with short passenger connection times. Short connection times increase the likelihood of delays if there is a disruption at any point in the network.

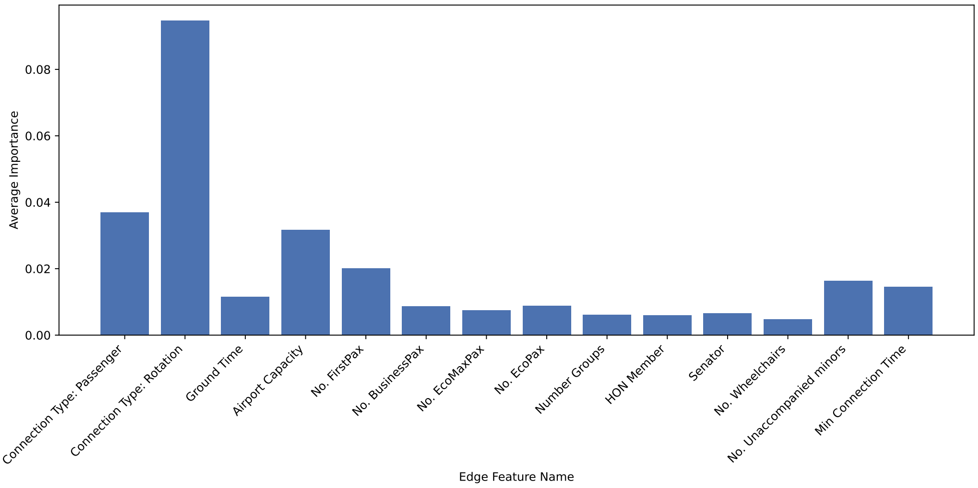


Figure 10: Average edge features importance across samples

In summary, this edge feature analysis highlights the importance of operational factors, such as aircraft rotation, ground time, and connection times, in influencing delays across the network. These results highlight the interconnectedness of flights sharing an aircraft and the impact of operational efficiency on delay predictions.

6.4.2 Correlation Analysis

The correlation analysis of node, edge, and graph features offers insights into how various relationships may impact delay propagation predictions, highlighting key dependencies as well as features that contribute to the model's predictive framework. First, the correlation analysis of the node features used in the GAT model can be seen in Figure 11. In this figure, acronyms include ARR (Arrival), DEP (Departure), DT (Time), ETD (Estimated Time of Departure), PAX (passengers), REACT (reactionary) and ATC (Air Traffic Control). SCHED stands for scheduled times, while "sin" and "cos" represent the encoded sine and cosine terms of timestamps, providing cyclic representations of time features.

From Figure 11, several significant insights were identified. The flight distance (*Distance (km)*) feature shows a strong positive correlation with both the total number of passengers (*Total No. Passengers*) (0.98) and minimum ground time (*Min Ground Time*) (0.98). This suggests that longer-haul flights, often transporting more passengers, are typically operated by larger aircraft, which require extended turnaround times. This relationship implies that flight distance is a substantial factor in turnaround time due to additional requirements for refueling, maintenance, and boarding. Furthermore, a high correlation between total number of passengers and minimum ground time (0.87) reinforces this, as larger aircraft with higher passenger loads typically require longer for boarding and de-boarding.

Scheduled and actual times correlations also show meaningful patterns. High correlations are observed within both scheduled and actual time features; for example, *ARR SCHED DT (sin)* and *DEP SCHED DT (sin)* have a strong positive correlation (0.93), as do *ARR ACTUAL DT (sin)* and *DEP ACTUAL DT (sin)* (0.93). This indicates that flights departing and arriving in similar time frames follow periodic patterns, aligning with the expected schedule-based flow of flights between interconnected hubs. This is expected given that actual times are essentially derived from scheduled times, adjusted by any delays.

It is important to note that in the training dataset, the actual departure time is the same as the *FPM ETD*, as the latter reflects the final planned departure time after the flight has departed. However, until the actual time of departure is available, the actual departure time is essentially a prediction, and *FPM ETD* is the most updated estimate on the day of operations. Including *FPM ETD* is crucial, particularly in the validation and testing dataset, where it provides the latest and most reliable information for departure predictions.

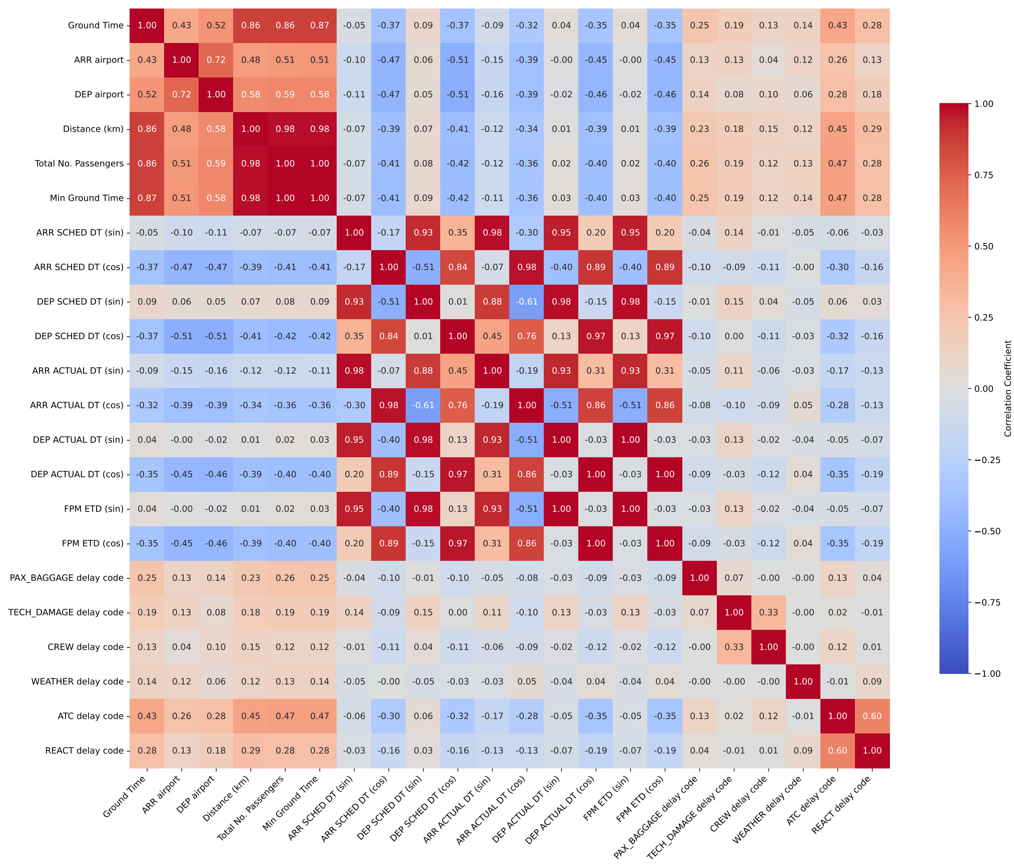


Figure 11: Node features correlation matrix

Temporal features encoded as sine and cosine transformations of scheduled and actual timestamps for arrivals and departures exhibit consistent negative correlations with each other (e.g. *ARR SCHED DT (sin)* versus *ARR SCHED DT (cos)* at -0.37), reflecting expected periodicity in scheduling and adherence to cyclical patterns. Furthermore, certain delay codes, such as the ATC delay code, display moderate positive correlations with ground time (0.43), suggesting that air traffic control delays may contribute to extended ground durations.

Furthermore, edge features reveal further structural insights into delay propagation in Figure 12. Acronyms include *FirstPax* (first-class passengers), *BusinessPax* (business class passengers), *EcoMaxPax* (economy premium passengers), and *EcoPax* (economy passengers. HON Member and Senator refer to high-status frequent flier members.

Notably, connection type exhibits a perfect negative correlation between passenger and rotation (-1.00), indicating that these categories are mutually exclusive by definition. Additionally, ground time shows a moderate correlation with both *Connection Type Passenger* (-0.58) and *Connection Type Rotation* (0.58). This relationship suggests that different connection types may influence ground time requirements, with rotation connections typically associated with longer ground times due to additional operational needs for aircraft rotation as compared to passenger-only connections. Furthermore, passenger composition and group features reveal several relationships. For example, the number of groups displays a positive correlation (0.32) with the number of economy passengers, indicating that economy passengers are more likely to travel in groups compared to other classes.

Minimum connection time shows a notable correlation with ground time (-0.42), reflecting that flights with tighter connection schedules may experience shorter ground times, likely to meet transfer requirements. Furthermore, frequent fliers and special service indicators like HON member and Senator show low correlations with operational features, implying that while they are tracked, they do not heavily influence ground time or capacity in the network. These correlations highlight the structural differences across connection types, passenger demographics, and operational needs in the network.

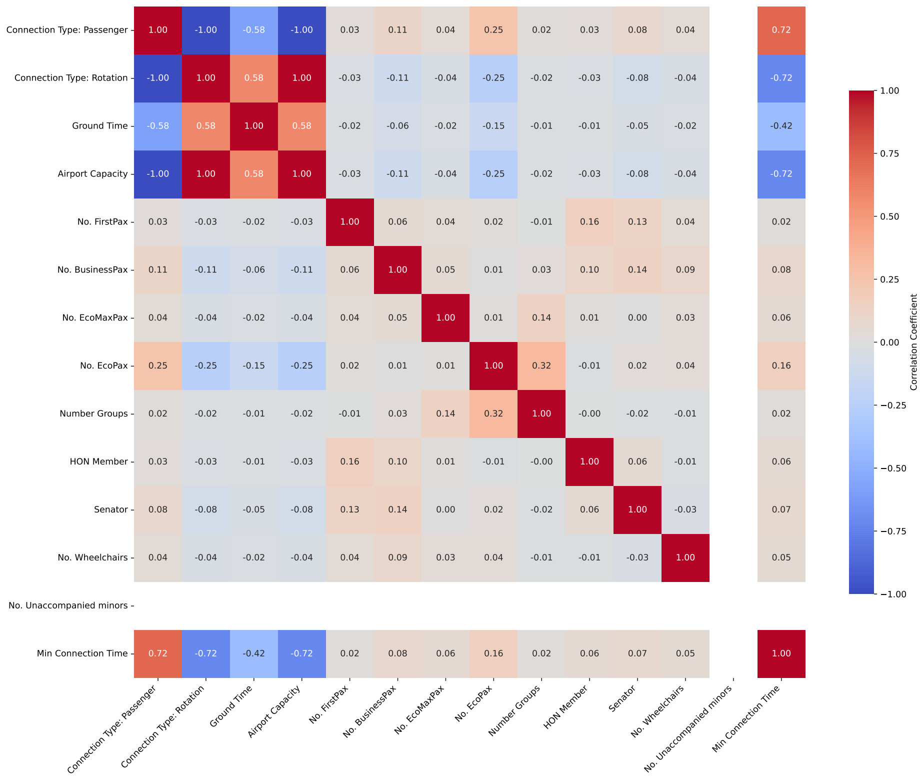


Figure 12: Edge features correlation matrix

For graph features, Figure 13, the independence of *Day (sin)* and *Day (cos)* from *Month (sin)* and *Month (cos)* supports the model's ability to track daily and monthly cycles separately, preserving valuable temporal distinctions. This separation is critical, as it allows the model to adapt to both short-term and seasonal patterns without merging them, ultimately enabling a more nuanced representation of temporal effects on delay propagation.

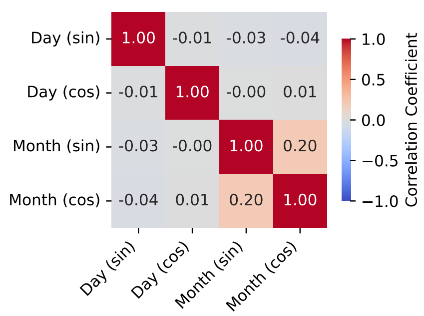
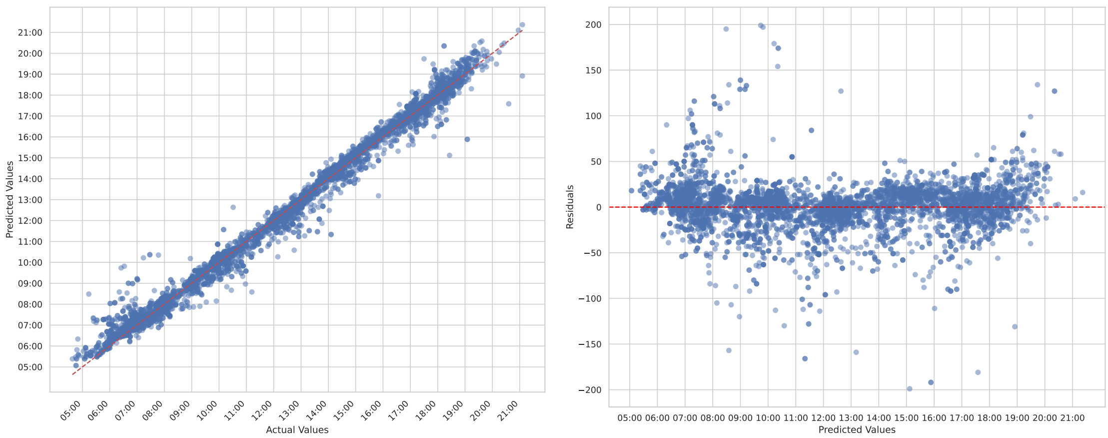


Figure 13: Graph features correlation matrix

These findings clarify how each feature type supports the model's understanding of delay propagation, showing which variables are more closely connected and which act independently, potentially influencing the prediction.

7 Validation

To assess the model's generalization ability, a validation dataset from October 9th to October 30th 2024 was used, providing a period distinct from the training data. This approach allows for an unbiased evaluation of the model's performance with new data, providing insight into its effectiveness in predicting delays under varying real-world conditions. The model was run at 02:00 UTC, just before the day of operations began, aligning predictions closely with the start of each day's schedule. The scatter plot of predicted versus actual values (Figure 14a) reveals an encouraging alignment along the diagonal, indicating that the model captures the overall trend in arrival times. However, several notable outliers deviate from this line, suggesting cases where prediction accuracy is reduced. This observation is further supported by Figure 14b, where residuals generally cluster around zero but show pronounced deviations for specific data points. These outliers suggest that certain conditions may impact the model's accuracy more significantly, and they will be examined in detail in a subsequent section to identify potential factors influencing these discrepancies.



(a) Prediction vs Actual Arrival time

(b) Residuals

Figure 14: Result analysis for validation dataset

In analyzing the scatter plot (Figure 14a), a slight increase in prediction uncertainty for the first flights of the day is observed. The GAT model predicts arrival times by using interactions between nodes and edges. However, these early flights, represented as the initial nodes in the graph, have no incoming edges and, therefore, depend only on their node-specific features without prior information from preceding flights. This limited data may contribute to the model's occasional difficulty in accurately predicting arrival times for these early flights, as it lacks the contextual data that later flights have.

This section is structured as follows; Subsection 7.1 presents various validation plots based on features used, Subsection 7.2 analyzes the model's accuracy in predicting delay propagation, followed by Subsection 7.3 which compares the GAT model to an already existing model used at SWISS. Finally, Subsection 7.4 presents unforeseen cases that the model cannot predict and Subsection 7.5 evaluates the attention weights outputted by the model. It is important to note that, to protect confidentiality, all flight numbers and airport codes (except ZRH, the SWISS hub) have been altered and assigned random values, in this section.

7.1 Analyzing Prediction Errors Across Key Parameters

In order to validate the model the following section includes a series of plots that analyze prediction errors across various factors relevant to delay propagation. Each plot focuses on examining how certain parameters (departure time, departure airport, aircraft type, travel distance, and node degree) affect prediction accuracy.

Figure 15 shows a trend in prediction error across different departure times along the day when the model is run before the day of operations begins at 02:00 in the morning (UTC). Errors tend to fluctuate more during peak hours, such as early morning and late afternoon. This temporal variation suggests the model may struggle to capture complexities associated with peak operational times, likely due to scheduling bottlenecks and intensified traffic network interactions. This finding highlights the need for time-dependent adjustments or temporal embeddings in future model iterations to capture the cyclical patterns of airport congestion and other time-sensitive factors. The higher error observed for the final flights of the day could stem from additional variables involved for the last wave of arrivals back to Zurich. Given that these flights often carry passengers with critical long-haul connections, certain operational adjustments, such as prioritizing arrival times or even flying at higher speeds, may be introduced by the Operations Center to ensure the flights arrive on time. These complexities add variability that the model cannot predict, resulting in larger prediction errors.

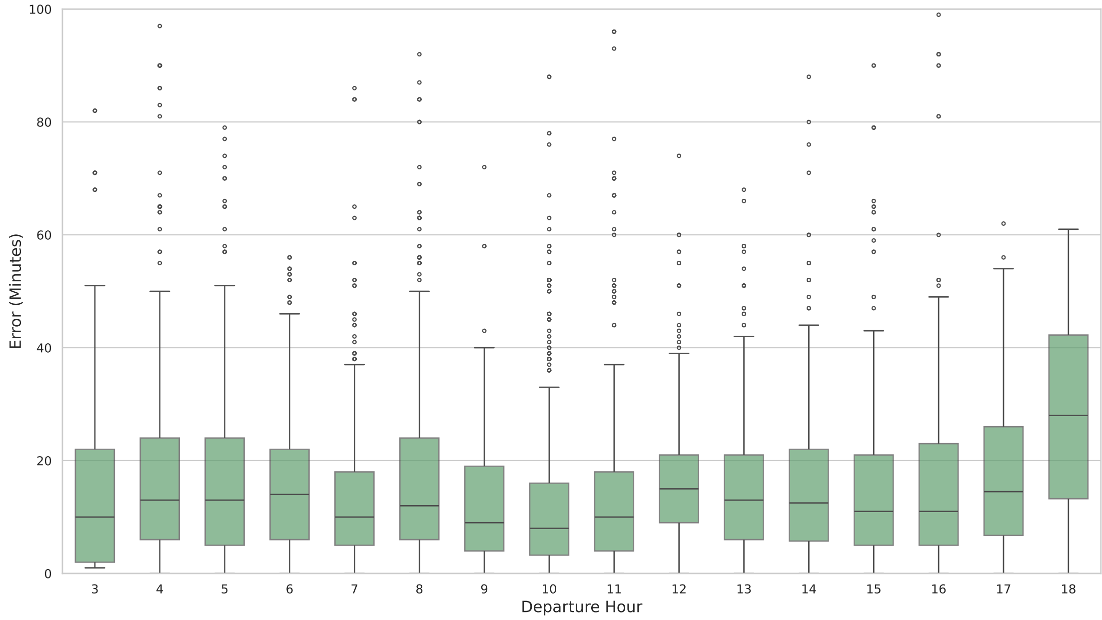


Figure 15: Error vs Departure time (hourly in UTC)

An analysis of error versus departure airport, as seen in Figure 16, further illustrates the geographical influence on model predictions. Each airport reflects a unique operational environment, with variability in delays likely influenced by factors such as airport size, congestion levels, and connectivity. For instance, larger international hubs, such as CDG, exhibit higher error margins, potentially due to the greater complexity associated with larger volumes of connecting passengers and intricate scheduling interdependencies. This further strengthens the idea that airport-specific features are crucial to understanding the model.

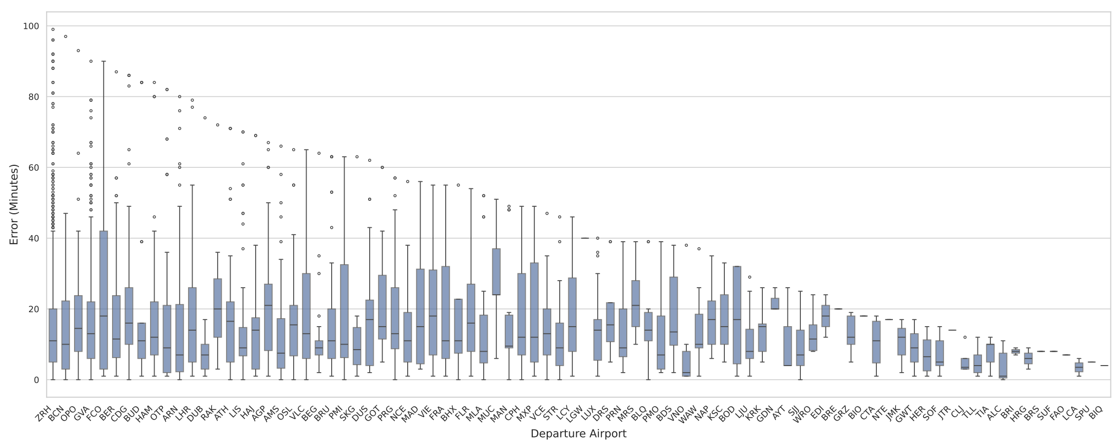


Figure 16: Error vs Departure airport

From Figure 16, a few irregular cases are identified. For this reason, the historical delay average per airport is computed and shown in Figure 17. By comparing these two, it can be observed that the model has a tendency to overpredict delays. The model performs with varying accuracy across the airports, reflecting their unique delay patterns. For instance, FCO shows a median error close to 20 minutes, with whiskers reaching up to 90 minutes, indicating high variability in prediction accuracy. This corresponds with FCO's historical delays, which, while generally low, show a wide range. The model's occasional overpredictions appear to affect FCO's unpredictability, where delays can unexpectedly increase.

In contrast, BHX presents a different scenario: the model maintains low prediction errors (median around 15 minutes) despite the airport's historically high and variable delays. This suggests that BHX's delay patterns, although elevated, are consistent enough for the model to predict reliably, achieving stable, low error rates. This capability highlights the model's adaptability in accurately forecasting delays, even in high-delay contexts when patterns are predictable.

AMS offers a clearer view of this pattern of adaptability. With the lowest median error (approximately 10 minutes), AMS's prediction error reflects its stable historical delay profile, marked by low median delays and narrow variability. The model's accuracy here underscores its efficiency in environments where delays are predictable, maintaining low error rates without the need for frequent overpredictions. On the other hand, CDG demonstrates how the model manages high-delay environments with a tendency toward overprediction. Despite CDG's high and variable historical delays, the model achieves low errors (median around 15 minutes) with limited variability, suggesting that the model has successfully adapted to CDG's recurring delay trends. This conservative overprediction approach likely minimizes underestimations in variable environments, helping stabilize accuracy even in high-delay scenarios like those seen at CDG and FCO.

Overall, airports with consistent delay patterns, such as AMS and BHX, demonstrate lower prediction errors, whereas airports with more unpredictable delays, like FCO and CDG, experience a higher frequency of overpredictions.

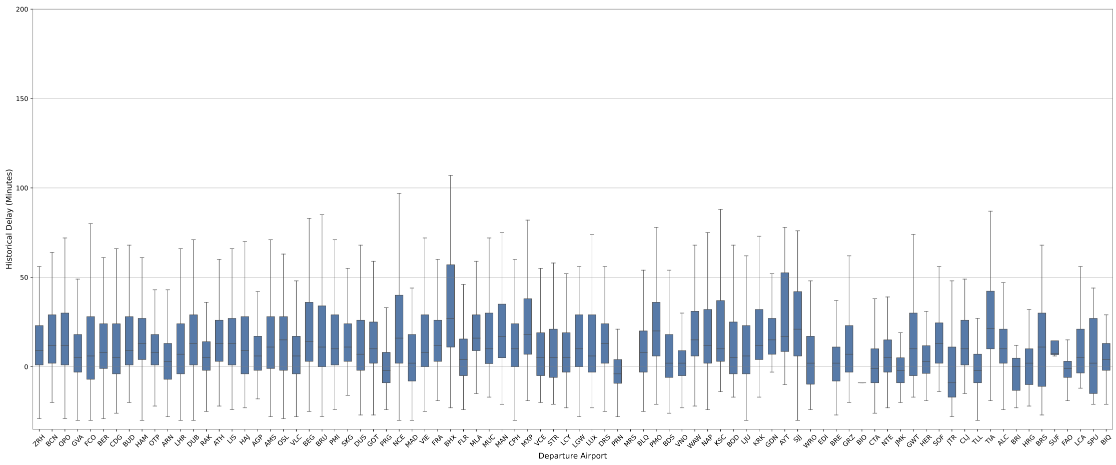


Figure 17: Average delay distribution for historical dataset

The Error vs. Aircraft Type plot (Figure 18) reveals clear differences in prediction errors across various aircraft models, suggesting that certain operational characteristics unique to each type may impact delay patterns more significantly than currently captured. Larger (32N, 320) or more frequently utilized aircraft types yield slightly higher errors due to their increased likelihood of encountering rotational delays or maintenance requirements. More specifically, larger aircraft like the 32N (Airbus 320 Neo) often require greater turnaround times due to higher passenger numbers and cargo capacity, increasing their susceptibility to propagation delays. Additionally, incorporating maintenance-related features, such as fleet age or maintenance frequency, could enhance prediction accuracy by accounting for the likelihood of technical checks or repairs, which tend to be higher for older aircraft and may impact reliability. Furthermore, adding a feature that captures the typical recovery speed of each aircraft type would enable the model to distinguish between types better equipped to make up for flight delays.

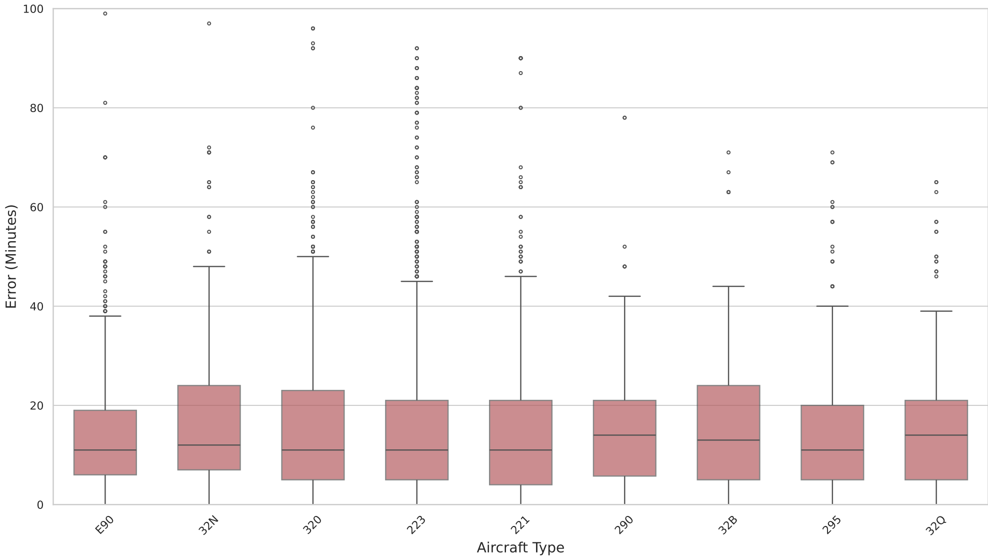


Figure 18: Error vs Aircraft type

Moreover, Figure 19 suggests an upward trend in prediction error with increasing travel distance, hinting that longer flights may introduce additional uncertainties impacting the model's accuracy. Longer routes often encounter more variable conditions, such as weather shifts and air traffic control delays, which can accumulate over extended distances and lead to higher prediction errors. However, it is also important to note that data points become sparse at these greater distances, limiting the ability to derive a fully reliable conclusion from this trend. While the pattern indicates that extended flights might inherently carry more prediction challenges, the limited data in this range means these findings should be interpreted with caution, as additional long-distance data could further clarify or adjust this observed relationship.

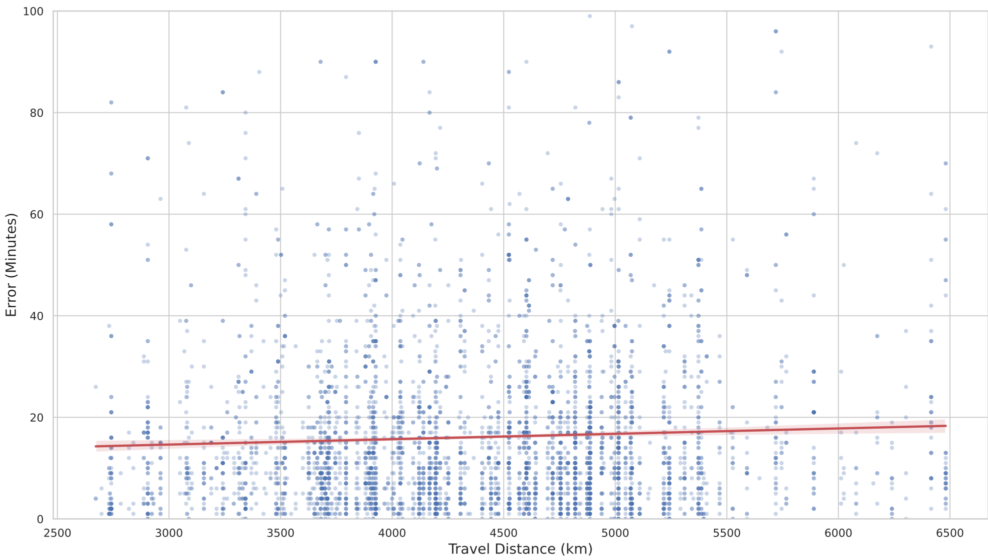


Figure 19: Error vs Travel distance with trend line

Furthermore, Figure 20 indicates that prediction error tends to increase with node degree, especially for flights with over 20 connections. However, a closer look at the data points reveals that some of the highest RMSE values occur at lower node degrees. This suggests that while flights with high connectivity bring prediction challenges due to their role in spreading delays, flights with fewer connections also prove difficult for the model. It's likely that flights with lower node degrees, having fewer connections, provide the model with less information about the flight network, leading to more uncertain predictions. Thus, both highly connected flights, with their intricate delay patterns, and low-degree flights, with limited data on surrounding connections, introduce distinct prediction challenges, pointing to areas for further refinement in the model.

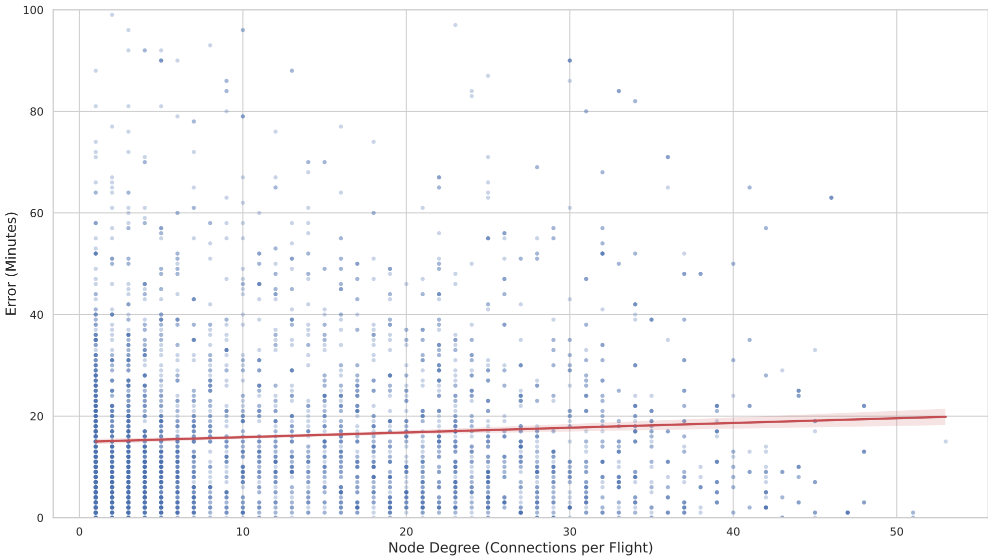


Figure 20: Error vs Node Degree with trend line

7.2 Propagation Delay

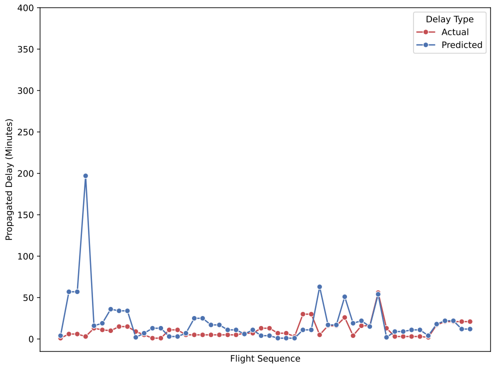
The use of a Graph Attention Network model aims to leverage its structural properties to capture the intricate dynamics of delay propagation across a short-haul fleet network throughout the day. The GAT model's graph-based approach theoretically allows it to detect dependencies between flights, where delays in one node may ripple through subsequent flights. By predicting arrival times and comparing the model's output with actual delay propagation, this section examines whether the GAT model can adequately predict these cascading delay effects, which are critical for understanding operational efficiency in a networked airline context.

The delay propagation curves (Figure 21) provide a side-by-side comparison of real and predicted delays for individual aircraft, with red indicating actual delay propagation and blue representing the model's predictions. These predictions are generated by running the model at 02:00 UTC. Certain aircraft, such as Figure 21b, demonstrate a strong alignment between predicted and actual delay patterns, suggesting that the GAT model effectively captures the propagation phenomenon for these cases. For HBAZC, the model's predicted curve mirrors the real propagation trend, accurately following the rises and falls in delay as they unfold across the day. This successful alignment indicates that the model has likely captured the sequence of dependencies between flights, suggesting an effective application of the GAT's attention mechanism in tracking delays as they propagate through consecutive flights.

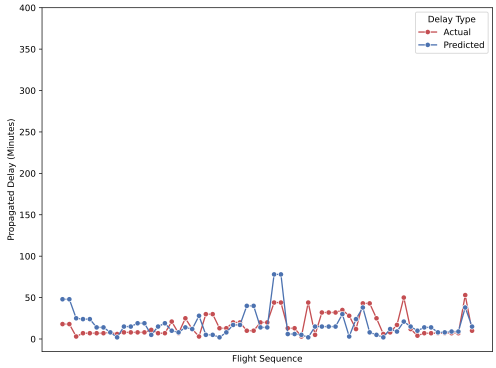
Another example of effective delay prediction is observed in Figure 21f, where the predicted delay curve closely resembles the actual propagation trajectory. The model's ability to track the fluctuations in delay for HBJCF demonstrates its potential to identify delay dependency patterns within a structured network, as this aircraft's operational pattern aligns well with the model's graph-based relational approach. The consistency of the model's predictions with actual delays for these aircraft reinforces the hypothesis that the GAT model can, under certain network configurations, capture the critical relationships necessary for modeling delay propagation. The same is seen for aircraft YLAAT in Figure 21h, where the model understands the delay propagation trend but cannot fully capture the extent of the delay.

However, not all cases display such alignment. In Figure 21a, the GAT model's predictions diverge significantly from reality. The predicted delay curve lacks the sensitivity to capture actual delay spikes, remaining relatively flat in comparison to the pronounced peaks and troughs in the real data. This discrepancy may indicate that HBIJM's flight connections or operational rotation introduce complexities that the GAT model, in its current configuration, is unable to fully comprehend. The model's inability to account for these fluctuations suggests a limitation in capturing delay propagation for aircraft with irregular or complex rotations, where interdependencies between flights may not be straightforward or consistently represented within the model's attention mechanisms.

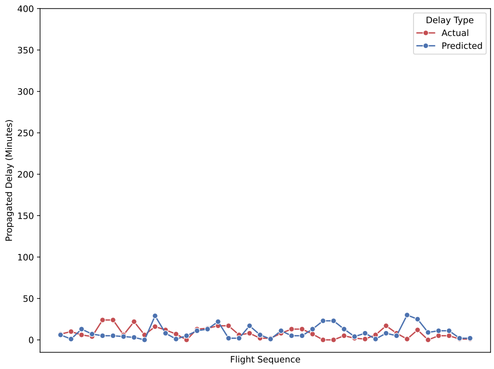
Similarly, for HBIOO in Figure 21d, the predicted delays do not closely track the real propagation pattern, particularly at the beginning. The model's output remains restrained, failing to capture the actual variations in delay. This gap could imply that the model's graph structure struggles to generalize to aircraft with less predictable or atypical patterns of delay propagation. These inconsistencies reveal that while the GAT model may perform well in certain structured network scenarios, it may lack the versatility required to accurately model propagation for more irregular cases, potentially due to limited information or insufficient representation of key operational features within the dataset.



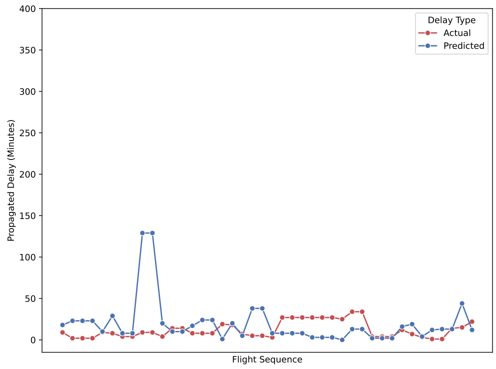
(a) Aircraft HBIJM



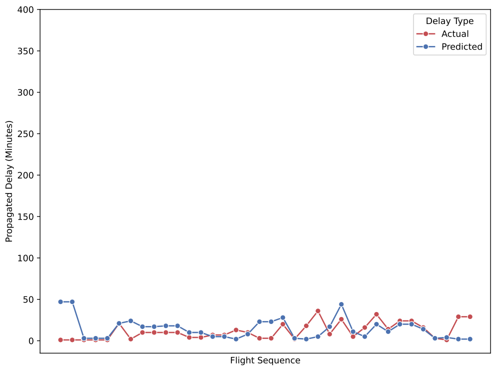
(b) Aircraft HBAZC



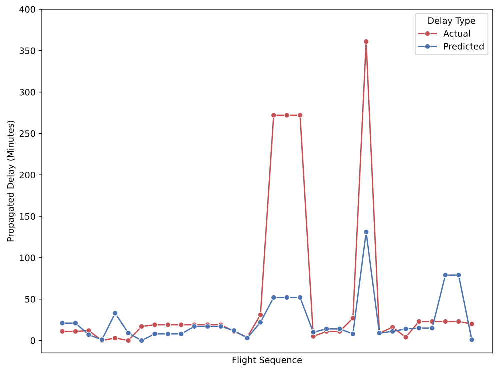
(c) Aircraft HBJBF



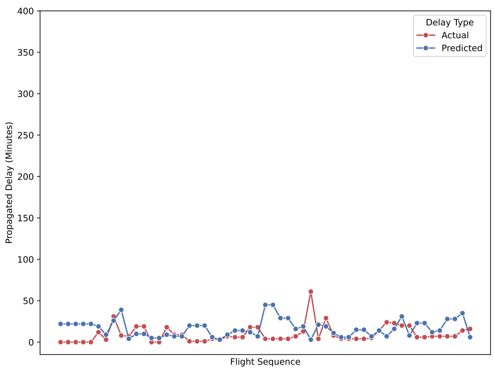
(d) Aircraft HBIOO



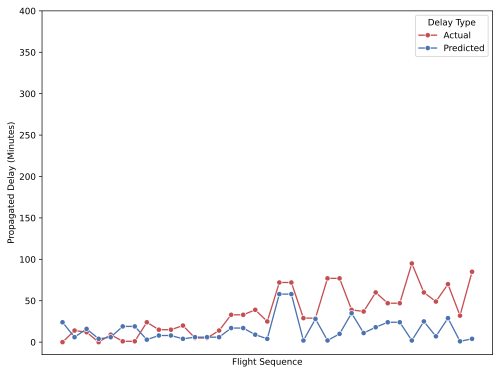
(e) Aircraft HBJCB



(f) Aircraft HBJCF



(g) Aircraft HBJDG



(h) Aircraft YLAAT

Figure 21: Comparison of actual and predicted propagated delays per flight sequence for selected aircraft

7.3 Comparison with SWISS's XGBoost Model

Given the model was trained and tested using SWISS's data, the GAT model is compared with the existing machine learning model currently used by SWISS for arrival time prediction. The existing model at SWISS employs XGBoost, a gradient-boosting framework optimized for efficiency and performance, to predict arrival times on a per-flight basis before the start of the day. While XGBoost leverages a set of features primarily related to the structural characteristics of each flight, as outlined in Table 7, it relies on a more limited set of features than the GAT model. This comparative analysis examines the impact of these additional features in the GAT model and assesses the potential performance improvements derived from leveraging a graph-based approach to incorporate relational and temporal dependencies between flights.

The GAT model and the SWISS machine learning model differ in the features they utilize. Notably, the GAT model does not include average ground time and maximum ground time. Furthermore, features such as ground time before flight and number of flights without a break are not explicitly represented as node or edge features in the GAT framework. Instead, the graph structure implicitly captures these aspects through the connections and interactions between flights, taking advantage of relational and temporal dependencies in the model.

Table 7: Features for XGBoost SWISS Model

Features	**Type**
Departure time of day (min)	Numerical (time)
Minimum ground time	Numerical
Ground time before flight	Numerical
Maximum ground time	Numerical
Average ground time	Numerical
Departure month	Categorical
Number of flights without break	Numerical
Off-block to on-block time (scheduled)	Numerical (time)
Departure weekday	Categorical
City pair	Categorical
Departure airport IATA code (scheduled)	Categorical

The SWISS model's feature importance analysis, shown in Figure 22, highlights a strong reliance on the temporal and structural characteristics of each flight. Departure time of day ranks as the most critical feature, likely reflecting congestion patterns that impact delay. Departure airport and previous ground time are also influential, suggesting that origin-specific factors and preceding schedules play a key role in predictions. Additional features like departure month and city pair help capture seasonal and route-specific patterns, while offblock-onblock time and ground time metrics underscore the importance of efficient turnaround operations.

In comparing the GAT (Figure 11) and SWISS model (Figure 22), several key similarities and differences arise in feature importance. Departure times (both scheduled and actual) are the most influential features in both models, underscoring the importance of timing in delay prediction. Departure airport ranks as the second most significant feature for each model, demonstrating the model's emphasis on location-related characteristics. The total number of passengers is also highly ranked in the GAT model, highlighting its added focus on passenger load, a feature not included in the SWISS model. Minimum ground time scores are low in both models, suggesting limited predictive value. A key difference is that departure month and day are handled differently: they are categorical in the SWISS model, while in the GAT model, these are integrated within the graph structure, capturing relational and time-based dependencies across flights.

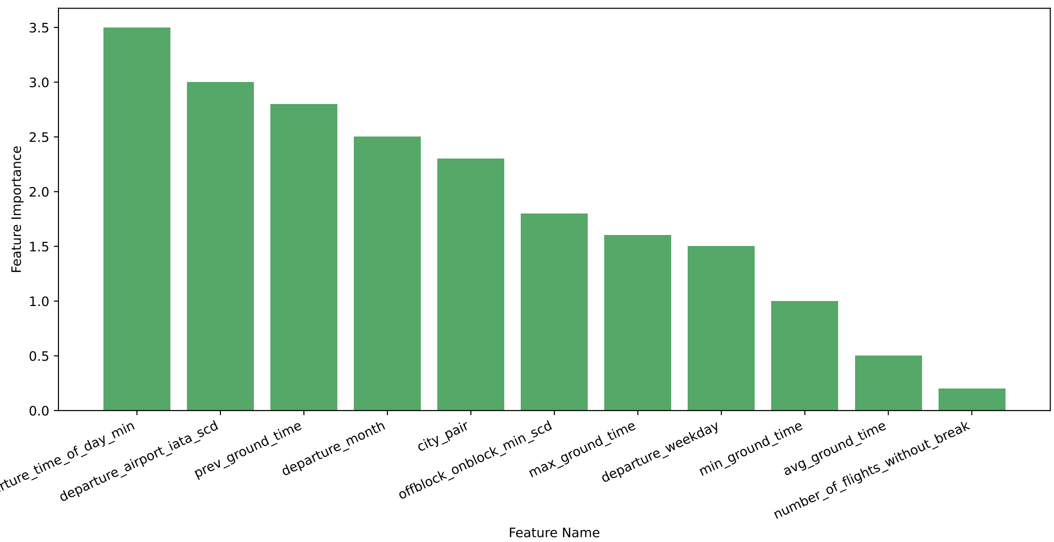


Figure 22: Feature importance of SWISS Model

Moreover, in Table 8, the RMSE comparison shows that both the SWISS model and the GAT model achieve nearly identical performance in predicting arrival times, with minimal day-to-day differences. RMSE values across the days indicate close accuracy levels, with slight variations where each model alternates in having a marginally better result. Additionally, the standard deviation of RMSE remains consistent between models, suggesting similar variability in prediction errors. These findings imply that, despite the GAT model's added complexity with graph-based features, it does not significantly outperform the simpler XGBoost model, indicating that the existing feature set may already capture the essential dynamics for accurate prediction.

Table 8: RMSE Comparison between SWISS XGBoost Model and GAT Model by Day

Date	**SWISS Model** (RMSE $\pm$ SD)	**GAT Model** (RMSE $\pm$ SD)
22-10-2024	17.52 $\pm$ 12.06	17.18 $\pm$ 10.50
23-10-2024	18.28 $\pm$ 13.00	16.98 $\pm$ 11.63
24-10-2024	18.03 $\pm$ 13.18	17.77 $\pm$ 11.94
25-10-2024	18.72 $\pm$ 12.76	17.76 $\pm$ 11.95
26-10-2024	27.88 $\pm$ 22.39	18.95 $\pm$ 12.65
27-10-2024	22.77 $\pm$ 16.43	22.93 $\pm$ 15.01
28-10-2024	20.55 $\pm$ 14.69	21.15 $\pm$ 13.34
29-10-2024	18.03 $\pm$ 13.12	22.78 $\pm$ 14.08
30-10-2024	19.35 $\pm$ 13.76	21.98 $\pm$ 13.72
31-10-2024	16.30 $\pm$ 10.05	19.98 $\pm$ 13.17

The performance comparison between the GAT and SWISS models shows differences across time blocks and days. Table 8 further breaks down the error per time-block (these plots are shown in Figure 35 Appendix B). The SWISS model often performs better in the earlier blocks, suggesting it handles morning operations or less complex scenarios more effectively. In contrast, the GAT model tends to perform better in later blocks, where accumulated delays or network complexities may have a larger impact. However, this pattern is not consistent; on some days, such as October 31, SWISS outperforms GAT across all blocks, while on others, their performance is mixed or similar in specific blocks. These results suggest that each model's performance depends on external factors and operational conditions on any given day.

To further understand Table 8, specifically why the SWISS model sometimes outperforms the GAT model and vice versa, a simple plot of average delay per day and its variability is presented in Figure 23. The differences in performance between the SWISS and GAT models appear to correlate with the average delay and its variability across days. On days with higher average delays and greater variability, such as October 26, the GAT model outperforms the SWISS model significantly, likely due to its ability to leverage graph-based features to capture complex interdependencies within the data. Conversely, on days with lower average delays, such as October 31, the SWISS model performs better, suggesting that its simpler structure is more effective under stable, low-delay conditions. On days with moderate average delays and lower variability, such as October 22 and October 23, both models achieve similar performance, indicating that they handle stable delay scenarios comparably well. These findings highlight how the characteristics of the delay distribution influence each model's effectiveness.

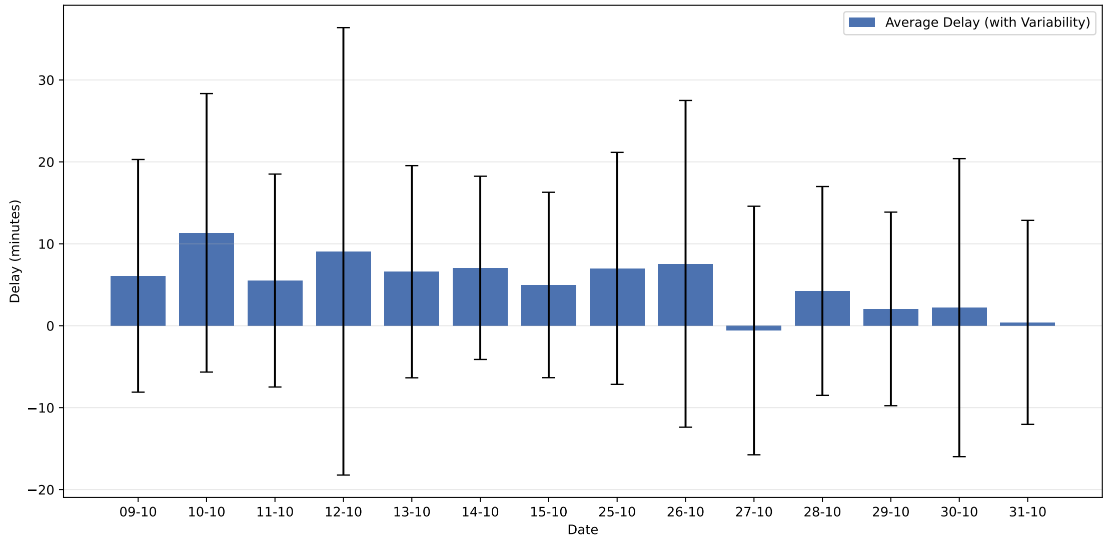
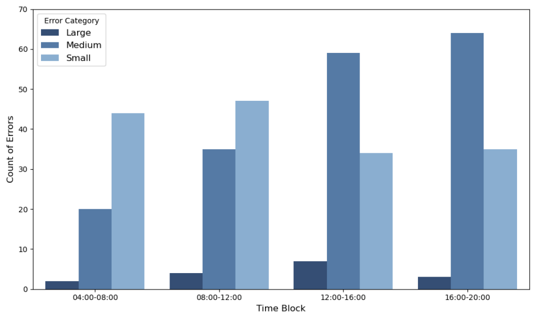
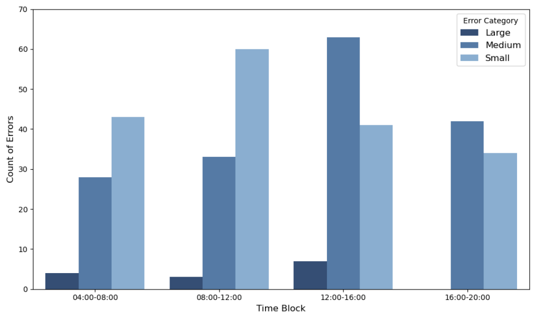


Figure 23: Average delay per day of the comparison dataset

The performance of the GAT model in predicting delays was further evaluated by examining the distribution of error categories (large, medium, and small) across different time blocks and days. First, the SWISS model shows relatively low counts of large errors, with notable increases during midday and late afternoon, particularly on October 25 and 26 (Figures 24a and 25a, respectively). In contrast, the GAT model exhibits a more even distribution of large errors but experiences spikes on specific days, such as October 27 (Figure 41b in Appendix C) during time block 08:00-12:00 and time block 16:00-20:00, and October 26 (Figure 25b) during time block 12:00-16:00.



(a) SWISS Model

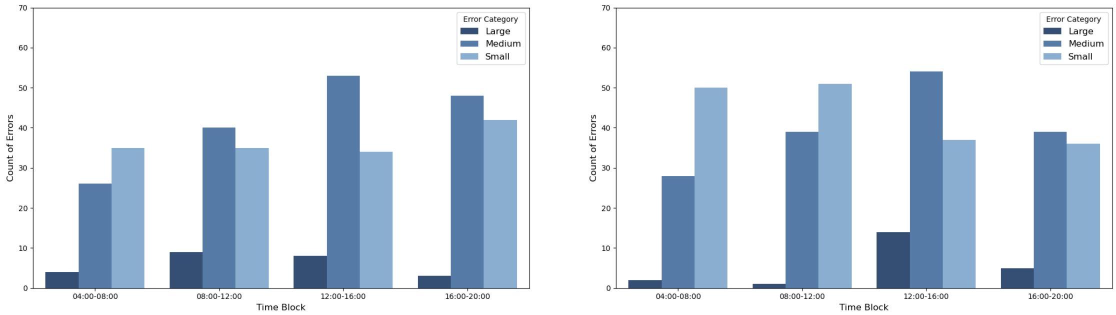


(b) GAT Model

Figure 24: Error category distribution for October 25

Secondly, medium errors dominate both models' distributions. The SWISS model records the highest counts during the middle of the day, with October 25 (Figure 24a) and 27 (Figure 41a in Appendix C) showing steady increases from morning to evening. The GAT model, while more uniform across time blocks, sees higher counts between 08:00-16:00 but performs relatively better during the evening compared to the SWISS model.

Third and last, small errors are generally the most frequent type of error, highlighting the models' accuracy. The SWISS model consistently achieves higher small error counts, particularly during early hours. On days like October 22 (Figure 36a) and 23 (Figure 37a), it maintains strong performance in minimizing larger deviations. The GAT model exhibits lower small error counts between 04:00-18:00 but performs comparably between 08:00-16:00, with more evenly distributed small errors on days such as October 25 and 26, as shown in Figures 24b and 25b, respectively.



(a) SWISS Model (b) GAT Model

Figure 25: Error category distribution for October 26

Overall, the SWISS model excels at minimizing large errors in the morning and consistently achieves higher small error counts, indicating greater overall accuracy. However, it struggles with medium errors in later time blocks. The GAT model shows more consistent performance across time blocks but occasionally experiences spikes in large errors during certain operational periods.

Examining the flight-by-flight error comparisons between the SWISS model and the GAT model provides further insight into each model's strengths and limitations. Figure 26, Figure 27 and Figure 28 show this comparison for a few time-blocks per day. The error plots across multiple days demonstrate that both models achieve a similar level of accuracy, as reflected in the overall RMSE values. For most flights, the errors remain within a comparable range, reinforcing the observation that the SWISS model's feature set is generally sufficient for accurate arrival time predictions.

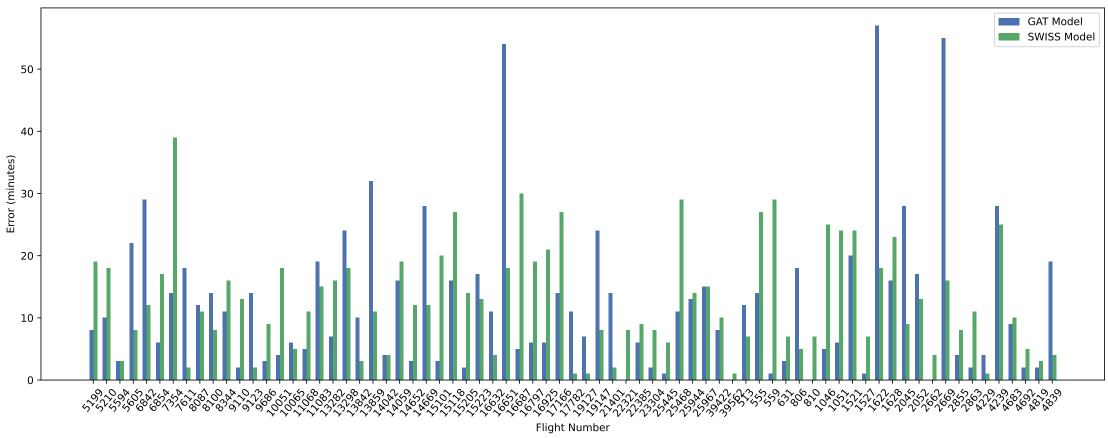


Figure 26: Comparison of GAT model with SWISS model for 31/10 between 12h-16h

The sharp error peaks observed in the GAT model, specifically in Figure 26, can be attributed to its sensitivity to specific relational and operational features within the dataset. For instance, highly connected flights, such as longer short-haul routes (Flight 16651 and 288), appear to drive the model toward overestimating delays. In these cases, the GAT model may overemphasize the influence of tight ground time constraints or the propagation of potential delays through numerous connections, even when these delays do not occur.

Additionally, the first flights of the day, such as Flight 1622, pose a challenge for the model due to the absence of prior delay information, requiring it to depend on generalized delay trends rather than flight-specific characteristics, often leading to inaccuracies. Another significant factor is the model's difficulty in accounting for corrections, such as in the case of Flight 2669, a short, high-frequency connection, where a departure delay of 25 minutes was offset by flying high-speed, resulting in an on-time arrival. These examples indicate that the GAT model's error peaks stem from an over-dependence on relational patterns and historical data without adequately incorporating dynamic factors such as operational adjustments or the unique characteristics of first flights in a rotation.

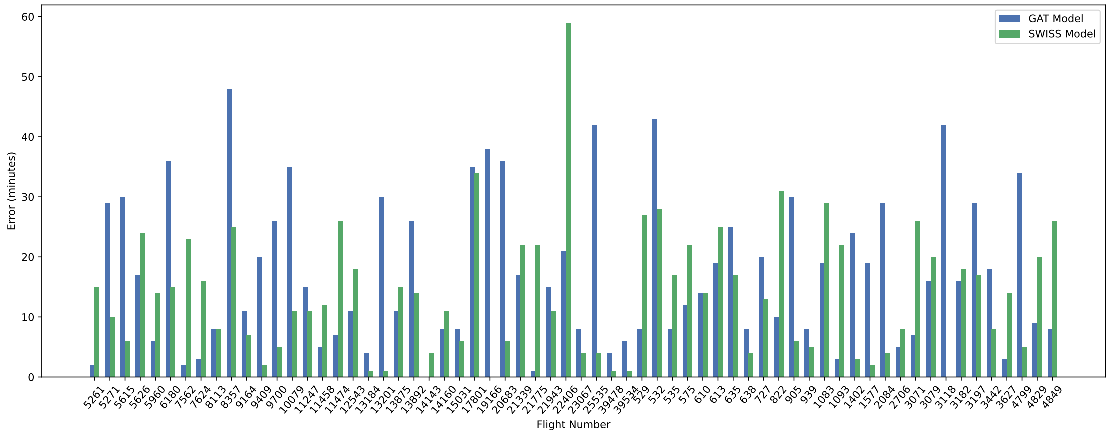


Figure 27: Comparison of GAT model with SWISS model for 31/10 in 16h-20h

On the other hand, while the GAT model shows the most peaks for specific scenarios, there are some cases where the SWISS model experiences significant sharp errors. Looking at Figure 27, Flight 22406 landing in Zurich demonstrates a case where the SWISS model predicted the delay inaccurately by 58 minutes, compared to the GAT model's smaller error of 20 minutes. In reality, the flight experienced a total delay of 70 minutes, caused by multiple factors: delay code 93 (rotation delay due to the late arrival of the incoming aircraft, contributing 11 minutes), delay code 33 (delays caused by a lack or breakdown of loading equipment, accounting for 44 minutes), and delay code 80 (missed ATC slot due to preceding process delays, adding 15 minutes). The SWISS model's large error indicates difficulty in accurately handling the combined impact of these distinct disruptions on the overall delay. In contrast, the GAT model, with its ability to incorporate relational features and connections, likely processed the interaction between these delay causes more effectively, resulting in a more accurate prediction. This case highlights the SWISS model's challenges in scenarios involving multiple overlapping delay factors, where the GAT model's graph-based approach appears to offer an advantage.

Another example illustrating the GAT model's limitations is Flight 39703 in Figure 28, which had a 14-minute delay despite departing on time due to ATC restrictions. With a ground time of 70 minutes from the previous rotation, well above the minimum required, there was no clear reason for a significant delay. Yet, the GAT model overpredicted the delay by 47 minutes, while the SWISS model only had a 5-minute error. This suggests that the GAT model may over-rely on certain features, such as extended ground times or network connections, leading to exaggerated predictions in cases without evident operational disruptions.

However, the GAT model shows occasional strengths, particularly on days when the SWISS model exhibits higher variability in its errors. On these days, the GAT model's predictions appear more stable, with fewer large error spikes. This suggests that the GAT model's incorporation of relational features may enable it to better capture inter-flight dependencies, such as delays that propagate across consecutive flights. For instance, when a delay on one flight impacts the following flights on the same route or involving the same aircraft, the GAT model's graph structure appears to capture these cascading effects more effectively than the SWISS model.

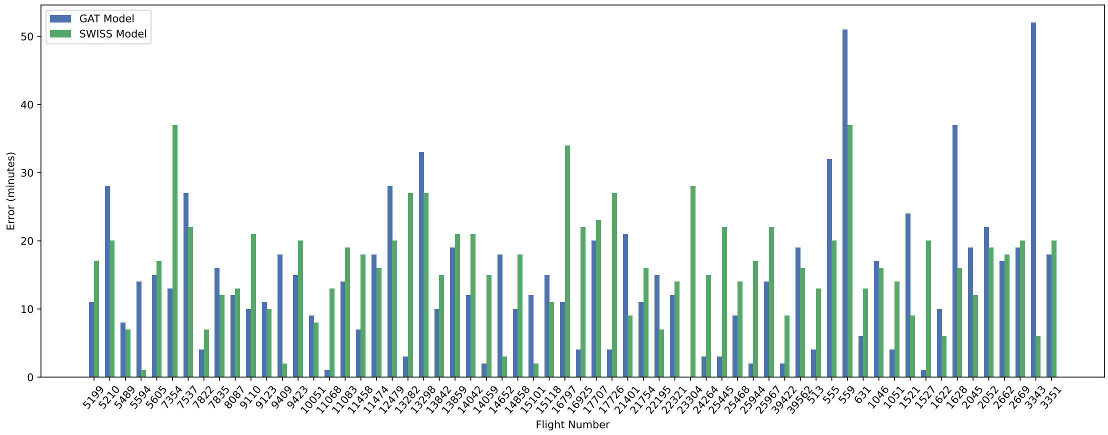


Figure 28: Comparison of GAT model with SWISS model for 27/10 in 12h-16h

The comparison between the GAT and SWISS models highlights distinct strengths and weaknesses in their predictions. For Flight 22195 (64-minute delay due to adverse weather), the GAT model performed better with a 50-minute error compared to the SWISS model's 74-minute error (Figure 47 from Appendix D). Similarly, for Flight 1402 (79-minute delay due to procedural issues), the GAT model had a small error of 5 minutes, while the SWISS model's error was 72 minutes (Figure 46 from Appendix D). However, for Flight 11474, which had no delay and few network connections, the GAT model overpredicted with a 55-minute error, while the SWISS model had a much smaller error of 8 minutes (Figure 46). These examples suggest the GAT model handles complex disruptions better, leveraging its ability to interpret network connections and dependencies, but tends to overpredict for isolated flights with minimal connectivity. In contrast, the SWISS model performs more consistently in simpler scenarios but struggles with events involving multiple factors, often overestimating delays.

While the overall performance of the GAT model does not significantly surpass the SWISS model, there are subtle advantages that become evident in complex, interconnected scenarios. For flights with high delay propagation, situations where delays are likely to impact subsequent flights due to shared resources or scheduling conflicts, the GAT model sometimes exhibits a more consistent error margin. This could point to the strength of the GAT model's graph-based approach, which leverages both temporal and relational dependencies to predict delays more robustly under certain conditions.

Another notable aspect is the presence of specific flights where the GAT model significantly outperforms the SWISS model. Although these outliers do not alter the overall performance metrics drastically, they highlight unique conditions, such as flights from particular airports or during peak times, where the GAT model's relational features align more closely with actual delay dynamics. This suggests that the GAT model has the potential to better handle intricate cases in airline operations, particularly when delays are interdependent across the network.

The differences in performance between the GAT and SWISS models most likely arise from the distinct features each model uses to predict delays. The SWISS model relies heavily on flight-specific features, such as departure time, ground time metrics, and categorical details like city pair and departure airport. These features focus on static operational characteristics, enabling the SWISS model to perform well in simpler scenarios where delays are primarily driven by individual flight attributes. In contrast, the GAT model incorporates relational and temporal dependencies, leveraging graph structures to capture interactions between flights, such as shared resources, connections, or cascading delays. This allows the GAT model to excel in complex scenarios where delays result from interdependencies across the flight network. However, this reliance on relational features can also lead to overpredictions in isolated cases or flights with minimal connectivity, highlighting the trade-offs in each model's feature set.

In summary, while the GAT model's added complexity does not yield a major accuracy advantage over the simpler, feature-constrained SWISS model, it demonstrates potential strengths in capturing relational and temporal dependencies that could be valuable in specific scenarios. The SWISS model's focus on structural and operational features is well-suited for day-to-day predictions, but the GAT model's graph-based structure may offer enhanced resilience in managing cascading delays and other interconnected challenges within the airline's network. This analysis presents the GAT model as a complementary approach that could offer incremental benefits in complex delay scenarios, though without a significant performance improvement overall.

7.4 Model Performance in Handling Unforeseen Events

Predictive models, such as GATs, offer potential solutions by forecasting delays based on historical data and flight interdependencies. However, certain unforeseen and irregular operations, particularly those related to Aircraft on Ground (AOG) incidents, pose challenges to the predictive capabilities of these models. This analysis goes into the statistics of heavily delayed flights over a period of 20 days, highlighting the presence of AOG incidents and other critical delay causes that remain unpredictable by the GAT model.

Figure 29 provides examples of real-world disruptions in flight schedules, showcasing how unpredictable events such as AOG incidents and rotational delays impact subsequent flights. Each example begins with a scheduled itinerary (in green) and compares it against actual timings (in gray) and the GAT model's predictions (in blue). Disruptive flights, marked in red, show how a single unexpected delay can cascade through the rotation, causing successive flights to deviate significantly from their planned schedule. For instance, in Example 1, an AOG event at ZHK leads to a substantial delay in returning to Zurich, affecting the subsequent scheduled flight.

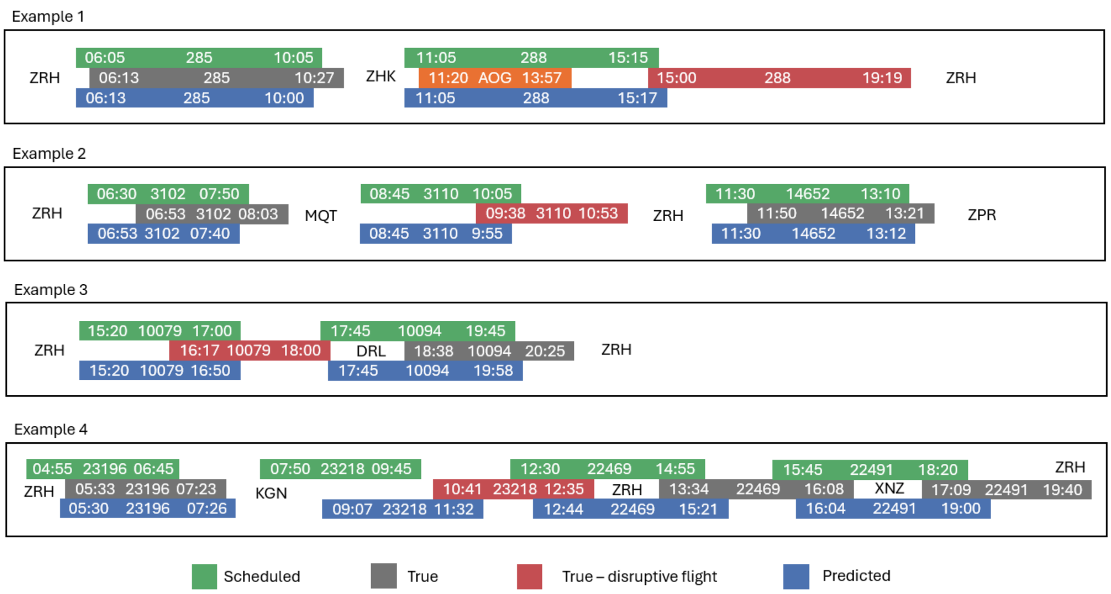


Figure 29: Examples where GAT model is unable to predict unforeseen delays

Additionally, in Example 2, a baggage sorting system failure and adverse weather at the destination delayed Flight 3102 departing Zurich, creating rotational delays for Flight 3110. In Example 3, procedural and maintenance issues delay Flight 10079 in Zurich, which then propagates delays to Flight 10094 departing DRL. Moreover, Example 4 further demonstrates how ATFM constraints and severe weather at departure stations result in compounding delays across flights 23196, 23218, and 22469, disrupting the entire rotation. These examples underscore the GAT model's limitation in predicting such disruptions, as it relies on historical patterns and cannot account for irregular, real-time operational challenges. This emphasizes the importance of integrating live data or other predictive tools into the GAT model to enhance their ability to capture the cascading nature of unforeseen delays in complex fleet networks.

Although only a small selection of flights is shown above, the analysis covers a span of 21 days, during which a total of 392 heavily delayed flights were recorded. Among these, 40 flights were classified as AOG incidents. On average, approximately 18.7 heavily delayed flights occurred per day, with 1.9 AOG flights daily.

Table 9: Overall Statistics of Heavily Delayed Flights

Metric	**Value**
Total Days Analyzed	21
Total Heavily Delayed Flights	392
Total AOG Flights	40
Average Heavily Delayed Flights per Day	18.7
Average AOG Flights per Day	1.9

Delays can be attributed to various reasons beyond AOG incidents. The distribution of these delay causes provides insights into the complexity and multifaceted nature of flight operations. Understanding these reasons is essential for identifying gaps in predictive modeling. In particular, delays were distributed across categories such as technical issues, rotation delays, flight operations and crew scheduling, and external factors like weather and industrial action. Table 10 shows the number of flights per delay reason.

Table 10: Distribution of Delay Reasons

Delay Reason	**Number of Flights**
AOG (Aircraft on Ground)	40
Technical Issues (Code 89)	50
Flight Operations & Crew (Code 81)	25
Immigration, Customs, Health (Code 84)	15
Weather (Codes 63f, 13)	10
Industrial Action (Code 16c)	5
Ground Handling (Codes 65, 83)	5
Air Traffic Control (Code 39b)	3
Others	14
Total	**132**

Technical issues, which involve unexpected equipment malfunctions not covered by routine maintenance, pose significant challenges to the model's predictive accuracy. These challenges are further intensified by delays related to flight operations and crew scheduling, including unscheduled crew shortages or operational mishaps, which introduce an additional layer of unpredictability. External factors, such as weather disruptions, industrial action, and ground handling problems, add to this unpredictability by contributing to delays that are sporadic and difficult to foresee. These unforeseen events require immediate operational responses, which are challenging for predictive models like GAT to forecast accurately. Consequently, while GAT models excel in leveraging historical patterns, their predictive power is limited by the inherent randomness and complexity of real-time disruptions.

Additionally, AOG incidents, characterized by sudden technical failures causing aircraft to become unserviceable, are inherently unpredictable and can lead to substantial delays and cancellations. AOG flights often lead to cascading delays affecting multiple subsequent flights within a rotation. On average, each AOG incident impacts approximately 3.5 other flights, amplifying the disruption within the flight network. This ripple effect highlights the inherent difficulty in predicting delays that originate from such unforeseen operational setbacks.

The analysis underscores the limitations of the GAT model in forecasting delays caused by unforeseen and irregular operations, particularly AOG incidents and technical issues. These factors contribute significantly to flight delays but remain difficult to predict due to their random and sudden nature. Understanding the frequency and impact of these delays is crucial for enhancing operational resilience and improving the predictive capabilities of flight delay models.

Overall, the analysis yields several key insights. An average of approximately 1.9 AOG incidents occurs daily, indicating a consistent presence of aircraft technical issues within the operational schedule. Additionally, 18.7 flights experience significant delays each day, reflecting ongoing challenges in maintaining punctuality. The impact of AOG incidents is substantial, with each occurrence typically affecting 3 to 4 subsequent flights, thereby disrupting the overall flight schedule and amplifying delay propagation across the network.

7.5 Attention Weights

The attention weights generated by the GAT model offer an insight into the model's internal prioritization, highlighting the most significant connections within the flight network. The GAT model's attention mechanism identifies certain flights as critical, assigning significant weights to specific connections within the network. While some patterns in these attention assignments are clearly identifiable, others are less straightforward. Thus, analyzing the operational context of these flights can help uncover potential reasons for their prioritization.

Across the datasets, a consistent pattern emerges with high attention weights assigned to the "first flights of the day." These flights are fundamental in initiating the day's rotation cycle. Since they are the first in their respective rotations, any delays can ripple through the rest of the day, affecting downstream connections. This phenomenon, known as the "snowball effect," means that the punctuality of first flights is critical to maintaining operational stability. The model seems to inherently recognize this operational dependency, assigning high attention weights to these flights to account for their potential cascading impact on subsequent rotations.

To further analyze the model's evaluation of flight importance, the "first flights of the day" are excluded, and the subsequent 20 flights with the highest attention weights are examined. Tables summarizing these results are provided below, while the full tables can be found in Appendix E. These flights are analyzed in relation to designated "priority flights," identified through a multi-criteria decision-making tool that accounts for factors such as connectivity, passenger load, and operational importance. Comparing the model's prioritization with this established list of critical flights allows for an assessment of how well the attention weights correspond to operationally significant flights.

Table 11: Highest Attention Weights for 12/10 (Excluding Initial Flights of the Day)

Flight Number	Attention Weight	Criticality	Origin	Destination	Edges
39506	0.999982	Not Priority	GVA	ZRH	16
24264	0.999936	Low Criticality	KXP	GVA	6
2103	0.999910	Initial Flight of Day	LHT	ZRH	7
4654	0.999887	Initial Flight of Day	NFX	ZRH	22
7354	0.999438	-	TKH	ZRH	7
2728	0.999048	Initial Flight of Day	VWB	ZRH	15
16614	0.999009	Dense Rotation	JMC	GVA	2
27681	0.998917	Low Criticality	WRN	GVA	1

For example, flights such as 6081 (Table 13) and 4770 (Table 15) show high attention weights, likely due to their numerous connections or proximity to high-passenger loads at major transit points. Delays on these flights could disrupt not just direct connections but also lead to misconnections for many passengers, underscoring their operational criticality. The attention mechanism also highlights flights that operate within tightly packed rotations. For instance, Flight 9123 (as shown in Figure 30) and FLight 4770 (Figure 32) have dense scheduling with minimal ground time, leaving little room for recovery in case of delays. Such flights, particularly those with short turnaround times at busy airports, are highly sensitive to even minor delays. The model likely assigns higher attention to these flights to reflect the operational risk associated with their tight schedules, where any disruption could necessitate rapid adjustments to maintain the network's overall timing integrity.

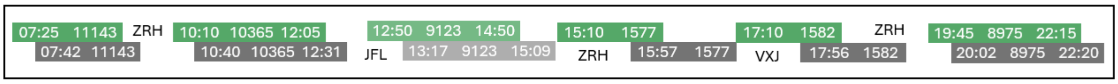


Figure 30: Rotation of critical Flight 9123 (green represents scheduled and gray actual times)

Some flights, such as 24264 and 3142 (Table 13), receive high attention weights even though they do not immediately appear to have high criticality or connectivity on the surface. This raises interesting questions about the model's internal criteria. It is possible that these flights have historical patterns of delays or incidents that lead to elevated attention weights. Alternatively, their roles within specific rotation cycles might make them more sensitive to disruptions than initially apparent. For instance, a flight may be connecting a secondary hub or providing a special service, but if its delay history shows it has a tendency to propagate disruptions, the model might prioritize it accordingly. In addition, flights such as 24264 might have fewer connections straightforwardly but could hold significance due to their geographical locations or the strategic importance of their schedules. The model may have recognized dependencies on these flights, especially if delays on these routes impact subsequent rotations involving more critical hubs.

Table 12: Lowest Attention Weights for 12/10 per node

Flight Number	Attention Weight	Criticality	Origin	Destination	Edges
39422	0.118740	Non-critical Flight	ZRH	GVA	4
5873	0.132116	High Connecting Times	GVA	PLW	18
39929	0.140600	Non-critical Flight	ZRH	GVA	2
1691	0.141458	High Turnaround times	ZRH	YNZ	4
532	0.152608	High Connecting Times	VSL	ZRH	5
14042	0.155336	High Connecting Times	ZRH	CMT	5
10051	0.170371	Semi-critical Flight	ZRH	DRL	10

As shown in Table 12, the flights with the lowest attention weights typically exhibit characteristics such as high ground-time buffers, minimal connections, or long connecting times, which reduce their overall operational criticality. For instance, flights such as 39422 and 39929 had non-critical roles in the network with limited connections, allowing sufficient time to recover from minor delays without impacting subsequent rotations. An exception to this pattern is Flight 10051, categorized as semi-critical. Despite its relatively low attention weight, 10051 experienced a delay in reality, which had a knock-on effect on subsequent flights due to its low scheduled ground-time buffers. However, although these downstream flights had numerous connections, no misconnections occurred in reality, indicating that the network's operational structure mitigated the potential impact of the delay.

Table 13: Highest Attention Weights for 25/10 (Excluding Initial Flights of the Day)

Flight Number	Attention Weight	Criticality	Origin	Destination	Edges
24264	0.999896	Moderate Criticality	KXP	GVA	2
6081	0.999404	High Connectivity	HFR	ZRH	17
16072	0.999269	High Connectivity	DXB	ZRH	18
3142	0.999115	Medium Connectivity	MQT	GVA	6
810	0.999090	Rotational Dependence	TRN	ZRH	22
5179	0.998873	Initial Flight of Day	ZRH	FZY	2
16632	0.998140	Low Impact	GVA	JMC	1

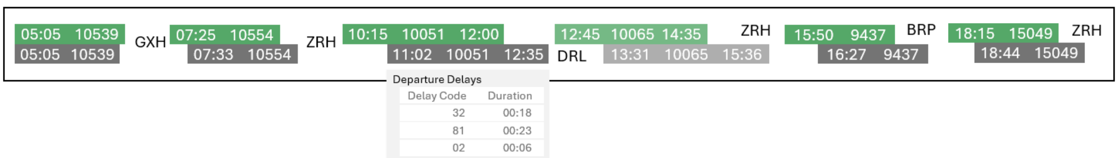


Figure 31: Rotation of critical flight 10065 (green represents scheduled and gray actual times)

In the context of rotations, the delay codes provide an additional layer of insight. The presence of delay codes such as "32" and "02", as seen in Figure 31, in close proximity to high-attention flights indicates that these flights are not only operationally dense but also prone to frequent delays due to specific recurring issues (e.g., technical problems, ATC restrictions). By factoring in these delay codes, the model might be associating certain flights with elevated operational risk, thus assigning them greater attention to account for these repeated disruptions. This pattern suggests the model's ability to adaptively weigh historical data, underscoring the importance of these flights even if their immediate connectivity appears low.

Table 14: Lowest Attention Weights for 25/10 per node

Flight Number	Attention Weight	Criticality	Origin	Destination	Edges
39354	0.112145	Non-critical Flight	ZRH	GVA	2
19127	0.116978	Semi-critical Flight	ZRH	XTR	12
39929	0.135073	Non-critical Flight	ZRH	GVA	3
2885	0.145401	Priority flight	ZRH	MQF	10
11458	0.146821	Priority flight	ZRH	GRZ	4
6795	0.152544	Non-critical Flight	ZRH	WBT	17
23325	0.154728	De-Priority flight	ZRH	KGN	1

The flights with the lowest attention weights for the 25th of October, as shown in Table 14, reflect varying levels of operational significance. Non-critical flights like 39534 and 39929 rank low in attention due to their limited connections and long ground-time buffers, reducing their impact on the broader network. Similarly, Flight 6795, despite having a significant number of connections, is non-critical by the GAT model because its long ground-time buffers and extended passenger connecting times minimize the risk of disruptions.

Flights identified as priority or semi-critical by the TOPSIS tool (described in Appendix F) illustrate differences in how the GAT model evaluates importance. For instance, 11458 is marked as a priority by TOPSIS because its paired flight has two large group connections to international destinations. While TOPSIS puts a strong emphasis on passenger group numbers and short connecting times, the GAT model focuses more on the total number of connections (edges). Semi-critical flights, such as 19127, further highlight these differences. Although its paired flight, 19147, has a considerable number of passenger connections, the longer connecting times ensure delays have a minimal impact on subsequent operations. In contrast, flights such as 11474 and 2885, with 52 and 72 short connections respectively, are more heavily weighted by the GAT model due to their dense network connectivity. Finally, Flight 23325 was identified as a de-priority flight by the TOPSIS tool, aligning with its low attention weight. With minimal connections and low operational importance, this flight ranks low in both frameworks. These patterns demonstrate that the GAT model prioritizes flights based on network connectivity, offering a perspective that differs from passenger-focused evaluations like those provided by TOPSIS.

Table 15: Highest Attention Weights for 04/11 (Excluding Initial Flights of the Day)

Flight Number	Attention Weight	Criticality	Origin	Destination	Edges
7488	0.999999	Priority Flight	ZRH	YRP	3
4770	0.999930	High Passenger Connectivity	NFX	ZRH	15
19108	0.999851	Priority Flight	XTR	ZRH	16
16797	0.999843	Priority Flight	JMC	ZRH	23
7354	0.999757	High Criticality	TKH	ZRH	21
1588	0.999634	Initial Flight of Day	VXJ	ZRH	2
19127	0.999160	High Passenger Connectivity	XTR	ZRH	1

It should be emphasized that flights identified as "Priority" through a multi-criteria decision-making tool also appear among the top attention weights, indicating alignment between the model's attention assignments and established criticality ratings. In Figure 32, Priority flights are indicated with a red triangle. These priority flights often serve routes with critical connections, potentially carrying numerous high-value passengers and connecting flights. The model's attention weights reflect the increased operational importance of these flights, validating the decision-making tool's emphasis on their criticality. For example, Flights 19108 and 19127 demonstrate both high criticality and operational risk due to their extensive passenger connections and priority status.

By analyzing the attention weights in combination with Figure 32, it can clearly be seen that the model's attention weights predominantly highlight flights with high connectivity and flights integral to rotations. By cross-referencing these flights with the Priority flights from the multi-criteria decision-making tool, significant overlap can be observed, suggesting that the model has successfully identified flights that could propagate delays across the network. However, due to the inherent lack of transparency of the attention mechanism, additional validation is essential to ensure accurate interpretation and to confirm that the model's prioritization aligns with operational objectives.

Table 16: Lowest Attention Weights for 04/11 per node

Flight Number	Attention Weight	Criticality	Origin	Destination	Edges
6912	0.055620	Non-critical Flight	ZRH	WBT	8
822	0.068699	Semi-critical	ZRH	TRN	2
19166	0.104608	Semi-critical	ZRH	XTR	2
3118	0.106501	High Connecting Times	ZRH	MQT	7
2885	0.112978	Non-critical Flight	ZRH	MQF	3
2084	0.115062	Priority Flight	ZRH	LHT	4
2893	0.116590	High Connecting times	MQF	ZRH	15

The flights with the lowest attention weights for in Table 16 reflect the GAT model's emphasis on connectivity over other operational features. For instance, the 2084-2091 rotation was flagged as a priority by the TOPSIS tool due to flight 645's large group connections to international destinations and many short connections. However, the GAT model assigns low importance, as it does not recognize the criticality of large passenger numbers connecting to international flights at the end of the day. Flight 19166 is classified as semi-critical because its paired flight has five critical international connections with low connecting times, requiring a high-speed rotation. Similarly, 822, also semi-critical, operates within a tightly packed schedule with minimal ground-time buffers. Its paired flight, 826, carries 17 connections, including two tight ones, adding to its operational importance. These cases once more highlight the GAT model's focus on the number of connections rather than passenger-related metrics or tight schedules.

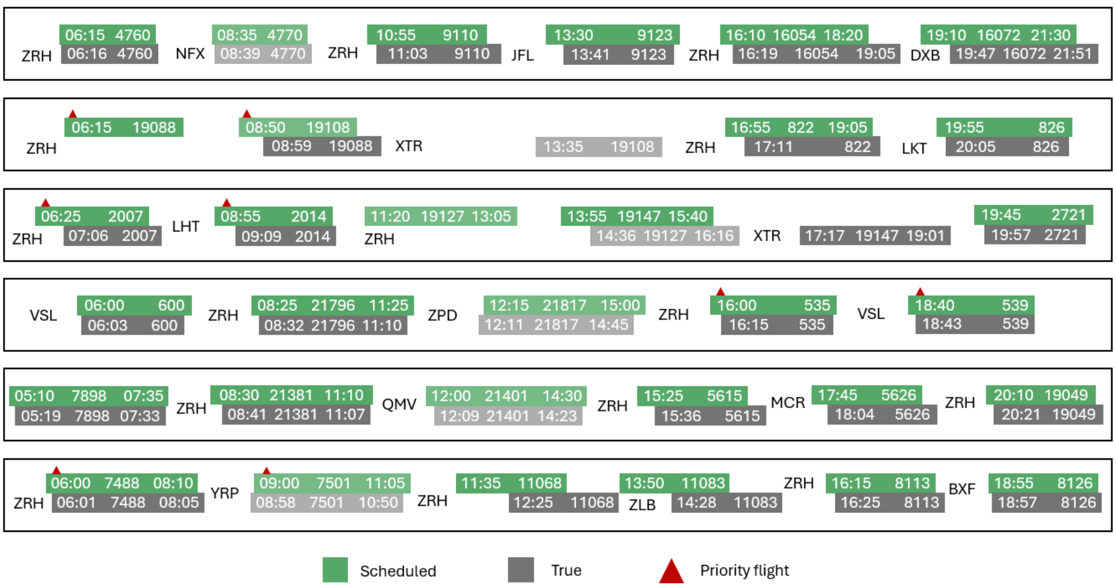


Figure 32: Critical flights and their rotations for 25/10

In conclusion, the model's high-attention nodes reveal a notable trend: flights with high connectivity (i.e., multiple connections or groups of connecting passengers) frequently receive elevated attention scores. For instance, flights with numerous connections to major hubs (e.g., ZRH) or closely packed rotations often appear in the top ranks. This focus aligns with operational intuition, as these flights represent points in the network where disruptions could cascade, impacting large numbers of passengers and subsequent flights. Additionally, flights identified as "Priority flights" by the decision-making tool generally appear within the high-attention subset, indicating a strong alignment between the model's attention outputs and established criticality metrics.

7.5.1 Operational Use of Attention Weights

The following section investigates the attention weights assigned by the GAT model for flights across all waves. The analysis explores whether high-attention flights align with critical operations, either due to priority classification (determined by the TOPSIS multi-decision criteria method, which selects 10 priority flights per wave) or their connectivity. Trends in departure and arrival airports are also analyzed to understand the influence of hub dynamics on attention allocation.

TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) is used to rank flights from most critical to least critical based on a comprehensive set of parameters such as rotation buffers, passenger connections, curfew performance, and slot allocation. By evaluating each flight against these key performance factors, TOPSIS calculates the relative closeness of each flight to an ideal solution that maximizes benefits (e.g., high passenger connections) and minimizes costs (e.g., risk of curfew violation). These top-ranked flights are considered priority flights by SWISS, representing the flights that are allocated more resources to minimize delay. Further details about TOPSIS can be found in Appendix F.

Waves represent structured periods of flight arrivals and departures designed to optimize hub connectivity and resource allocation. At Zurich Airport (ZRH), flights are grouped into five waves, defined by their scheduled times in Local Time:

- Wave 1: 06:00 to 08:30
- Wave 2: 08:31 to 10:30
- Wave 3: 10:31 to 14:00
- Wave 4: 15:31 to 19:00
- Wave 5: 19:01 to 21:45

The GAT model showed varying consistency when compared to the priority flights across the operational waves. Waves 3 and 5 demonstrated the strongest alignment, with 60% of high-attention flights classified as priority. This highlights the operational importance of midday and evening flights in connecting passengers and cargo across global destinations. Waves 2 and 4 also showed substantial alignment, with 50% of high-attention flights classified as priority, reflecting the critical roles of morning and early afternoon operations. By contrast, Wave 1, covering early morning flights, displayed the lowest alignment, with only 40% of high-attention flights classified as priority. This may indicate the GAT model's reduced focus on early morning operations or fewer critical flights during this period.

The following table provides an overview of the priority flight ratios and average attention weights for priority and non-priority flights in each wave:

Table 17: Priority Flight Ratio and Attention Weights by Wave

Wave	Priority Flight Ratio	Average Weight (Priority)	Average Weight (Non-Priority)
Wave 1	40%	0.7102	0.8203
Wave 2	50%	0.7190	0.8396
Wave 3	60%	0.7415	0.8127
Wave 4	50%	0.7254	0.7938
Wave 5	60%	0.6842	0.8012

This variation in alignment suggests that the GAT model effectively identifies critical flights in midday and early afternoon waves but may require refinement for morning operations.

The relationship between attention weights and flight connectivity, measured by the number of edges (i.e., connections), was evaluated for each wave. Across most waves, the correlation between these variables was not very strong, indicating that the GAT model does not rely heavily on connectivity when assigning attention weights. Notably, Wave 3 demonstrated the strongest positive correlation (0.589), suggesting greater emphasis on connectivity during the midday period when the network's efficiency relies on maintaining high-connectivity links. This makes sense, as Wave 3 flights carry the most connecting passengers. Wave 2 also showed a moderate positive correlation (0.358), reflecting some consideration of connectivity during morning operations. For other waves, including Wave 1 (0.204), Wave 4 (0.167), and Wave 5 (0.278), the correlation remained relatively low, indicating that the model appears to prioritize factors such as priority classification or operational importance over connectivity.

The table below summarizes the correlation between attention weights and edge count across waves:

Table 18: Correlation Between Attention Weights and Edge Count by Wave

Wave	Correlation (Weight vs. Edge Count)
Wave 1	0.204
Wave 2	0.358
Wave 3	0.589
Wave 4	0.167
Wave 5	0.278

These results indicate that while connectivity plays a role in specific waves, such as Wave 3, it is not the only contributor to the attention weights.

An analysis of departure and arrival airports highlights the dominance of ZRH as the central hub for high-attention flights. Across all waves, Zurich consistently serves as either a primary arrival or departure point, emphasizing its strategic importance in maintaining network efficiency. Geneva Airport (GVA) also plays a key supporting role, particularly during Wave 1 and Wave 5, where it connects regional hubs to Zurich.

In Wave 1, high-attention flights primarily arrive at Zurich from regional hubs, reflecting the wave's role in consolidating regional traffic early in the day. Wave 2 continues this pattern, with arrivals at Zurich from regional destinations. By Wave 3, the focus shifts to longer-range connectivity, with high-attention flights departing Zurich to destinations such as KGN and TRN, while Wave 4 prioritizes connections to Eastern Europe. Finally, in Wave 5, the model's attention weights decrease, primarily reflecting reduced operational intensity and focusing on regional connections.

The following table summarizes the trends in attention weights by departure and arrival airports for key waves:

Table 19: Trends in Attention Weights by Departure and Arrival Airports

Wave	Departure Airport	Arrival Airport	Average Weight
Wave 1	GVA	ZRH	0.8203
Wave 2	BDS	ZRH	0.7404
Wave 3	ZRH	KGN	0.7415
Wave 4	ZRH	DRL	0.7254
Wave 5	ZRH	GVA	0.6842

These patterns reinforce the importance of Zurich and Geneva as critical nodes in the network, with Zurich playing a dominant role across all operational periods.

The GAT model demonstrates a strong ability to align attention weights with critical flights during midday and early afternoon waves, particularly in Wave 3. The weaker alignment in Wave 5 highlights potential areas for improvement in recognizing critical evening operations. While the model incorporates connectivity to some extent, particularly in Wave 3, its reliance on other features, such as priority classification, underlines its nuanced decision-making process.

To determine if the GAT model assigns stronger attention weights to flights in certain waves, the average weights across all flights in each wave are calculated. The table below shows the average attention weights for flights in each wave, including both priority and non-priority flights:

Table 20: Average Attention Weights Across Waves

Wave	Average Weight
Wave 1	0.7984
Wave 2	0.8113
Wave 3	0.8262
Wave 4	0.7896
Wave 5	0.7602

From Table 20, it is evident that Wave 3 has the highest average attention weight (0.8262), indicating that midday flights receive slightly stronger attention compared to earlier or later waves. This aligns with the operational importance of Wave 3, which connects many international destinations and likely requires more sophisticated attention allocation.

In addition to analyzing high-attention flights, the flights with the lowest attention weights were examined to understand how the GAT model de-prioritizes operations. These flights often correspond to passenger connections with reduced urgency, characterized by sparse edge connections, long layovers, and minimal downstream impact. This analysis highlights the model's ability to focus on critical network operations while de-prioritizing less impactful ones.

Across all waves, flights with the lowest attention weights consistently exhibit limited influence on passenger connectivity. For example, in Wave 1, flights arriving at ZRH from spoke airports such as GHF and LHT involve layovers exceeding 200 minutes and play a minor role in immediate passenger transfers. In Wave 2, similar patterns emerge, with low-weight flights originating from secondary airports like NKJ or VSL, demonstrating reduced urgency compared to morning or midday connections. By Wave 3, when midday operations prioritize high-connectivity links, low-weight flights, such as MCR to ZRH and LHT to ZRH, remain those with fewer passenger dependencies. Later waves, including Wave 5, further illustrate the model's selective attention allocation, as evening flights, such as ZRH to MQF or ZRH to MQT, typically involve short-haul connections with minimal time-sensitive transfers.

A temporal pattern emerges, with attention weights declining steadily throughout the day. This reflects the reduced urgency of passenger flows during later waves, validating the model's ability to adjust to operational priorities dynamically. Earlier waves focus on strengthening regional connectivity, while midday operations achieve greater weight distribution for high-priority transfers. The analysis confirms the GAT model's effective prioritization of critical operations while systematically de-prioritizing flights with fewer passenger connections and limited downstream impact.

8 Discussion

The GAT model has shown notable strengths in predicting delay propagation within typical operational conditions, particularly for flights with high connectivity and consistent scheduling patterns. By effectively capturing networked dependencies, the model leverages relational data to anticipate how delays may affect connected flights. The model's use of attention weights is effective in identifying flights essential for maintaining network stability, such as those that initiate daily rotations or have numerous connections. This capability to discover key flights highlights the potential of the GAT model in detecting flights that could significantly impact the fleet network if delayed.

However, the GAT model encounters limitations when tasked with predicting irregular delays caused by unexpected factors, such as technical issues, crew shortages, or sudden weather changes. Such cases often lack the predictability that the model relies on, indicating a gap in its capability to generalize on less predictable events. Although the model captures dependencies within the flight network, it relies heavily on historical data patterns, which limits its adaptability in real-time, unexpected scenarios. This reliance reduces its effectiveness in handling disruptions that lack a consistent historical precedent.

When compared to SWISS's XGBoost model, the GAT model exhibited comparable RMSE values across test days, with both models performing well under typical operational conditions. However, the GAT model showed a distinct advantage in scenarios involving cascading delays, achieving an RMSE of 18.95 minutes compared to the SWISS model's 27.88 minutes on a particularly challenging operational day. This highlights the advantage of using a graph-based approach to capture interdependencies between connected flights more effectively than purely gradient-boosting models. Conversely, the SWISS model outperformed the GAT model on days with lower average delays, achieving an RMSE of 16.30 minutes compared to the GAT model's 19.98 minutes on an operationally stable day. This suggests that while the SWISS model remains efficient for typical day-to-day operations, the GAT model performs better in predicting networked delays that propagate across multiple connections.

Additionally, the GAT model's performance aligns with and surpasses other models from existing research used in delay prediction for long-term prediction (before the day of operations begins). For instance, Random Forest models applied to Colombia's airport network achieved an RMSE of 33.8 minutes, and MSTAGCN models on the U.S. airport network yielded 30.7 minutes. By comparison, the GAT model's RMSE of 15.59 minutes on medium-delay days underscores its ability to adapt to the complex dynamics of European airline networks. Despite this, the model does not outperform specialized sequential prediction models, such as LSTMs, which achieve RMSE values between 6.31 and 7.73 minutes for short look-ahead predictions. However, these models lack the GAT model's comprehensive understanding of the network structure of flights and hence are not appropriate for modeling delay propagation across interconnected flights.

The attention mechanism embedded in the GAT model has proven insightful for identifying flights with high operational impact. The model often assigns higher attention weights to flights that influence network stability, such as those that initiate daily operations or those with significant passenger transfer connections. This alignment with operational priorities underscores the relevance of the model's attention outputs in highlighting flights where delays may propagate widely through the network. The variation in attention weights across waves highlights the operational significance of midday flights (Wave 3), which exhibit the strongest alignment with priority flights and the highest average attention weight. Conversely, early morning flights display lower alignment, as the model equally weighs the first flights of the day, given their importance and significant impact on the propagation of delays. Moreover, Zurich and Geneva emerge as critical nodes within the network, reinforcing the central role of these hubs in maintaining operational efficiency. Outliers, such as non-priority flights receiving unexpectedly high attention weights, suggest that the model might identify unquantified operational risks or factors not currently captured by the multi-criteria decision-making tool that selects priority flights.

Lastly, the model's performance is influenced by the quality and scope of available data, including the absence of real-time operational data, such as maintenance records and updated crew rotations. The lack of this data may restrict the model's ability to adjust dynamically to evolving operational conditions. Consequently, this limitation may partly explain discrepancies between predicted and actual delays in certain flights or aircraft types with atypical schedules. Expanding the model's data inputs to incorporate weather forecasts or maintenance data could enhance its predictive accuracy and allow for more adaptive responses in future applications.

8.1 Model Utilization in an Airline Setting

The GAT model provides SWISS with an advanced tool for operational decision-making by leveraging historical data to predict arrival times and understand delay propagation within the network. Unlike the static TOPSIS method, which selects priority flights based solely on 11 predefined features without accounting for historical delay patterns, the GAT model captures the complex interactions between flights and their connections. By training on historical data, the GAT model accurately predicts arrival times while generating attention weights that highlight the most critical nodes (flights) and edges (connections between flights) in the network. These attention weights identify high-impact flights that, if delayed, could significantly disrupt subsequent operations due to their numerous connections or pivotal roles in daily rotations.

With this information, SWISS can proactively pinpoint critical flights and allocate additional resources to them to prevent delays. For instance, if a flight with a high attention weight is predicted to experience a delay, it can be flagged as a priority. Addressing potential issues on these flights can minimize the ripple effects of delays across the network, enhancing overall operational efficiency.

In contrast to the static TOPSIS method, the GAT model offers a dynamic, data-driven approach that adapts to real-time conditions and historical delay propagation. This enables SWISS to make more informed decisions by focusing on flights that are most likely to impact network stability. By combining predicted arrival times with attention weights, the airline can better allocate resources, monitor critical flights more closely, and maintain on-time performance across the network.

Overall, the GAT model improves upon basic models by not only predicting delays but also identifying which flights are most critical within the network. This dual capability supports more effective resource allocation and prioritization, helping to reduce operational strain and enhance the robustness of airline operations.

8.2 Future Work

Building on the GAT model's current performance, several enhancements can be implemented to further improve its predictive accuracy and operational utility for SWISS.

A key area for future improvement is the integration of real-time operational data into the GAT model. Incorporating data such as maintenance checks, crew rotations, and real-time weather conditions and forecasts, especially impactful events like snowstorms or changing wind directions could strengthen the model's ability to anticipate delays from sudden disruptions. Real-time data would allow the GAT model to respond dynamically rather than relying solely on historical patterns, reducing prediction gaps and improving adaptability in rapidly changing operational environments.

Additionally, the model's performance could benefit from integrating airport-specific features, such as airport capacity, runway availability, and ground congestion, as these factors significantly influence delay patterns. Including more detailed weather data would further enhance the model's accuracy, as weather conditions often play a critical role in determining delay likelihood. By accounting for these variables, the model would provide more precise predictions specific to each spoke airport's operational context. Given Zurich's role as a central hub and Geneva's critical support, incorporating explicit hub-related metrics could provide additional insights. Features capturing hub-specific dynamics like transfer passenger volumes, hub connectivity indices, or wave-specific priorities such as curfew risks and final flight connectivity could enhance the model's ability to predict delays and allocate resources effectively.

Another promising direction is the inclusion of maintenance-related features, such as aircraft maintenance frequency, fleet age, and operational constraints of certain aircraft. These details could enhance prediction accuracy by accounting for the likelihood of technical checks or repairs, which tend to be higher for older aircraft and may impact reliability. Furthermore, adding a feature that captures the typical recovery speed of each aircraft type would enable the model to distinguish between types better equipped to make up for delays in flight.

Adjusting the model's sensitivity to rare but high-impact scenarios would be beneficial. This can be achieved by incorporating a broader range of irregular operational cases into the training data to improve the model's attention allocation. Future refinements could involve incorporating explicit hub-related metrics or exploring additional features that capture evening wave-specific priorities, such as curfew risks or final flight connectivity. These insights not only validate the model's utility but also highlight areas where its outputs could be leveraged to enhance resource allocation and operational planning.

Furthermore, refining the analysis of propagation delays would strengthen the understanding of delay dynamics. By examining specific types of connections, such as international versus domestic flights, that are most prone to cascading delays, the model could yield actionable insights for targeted operational interventions, helping to reduce the impact of delays on high-risk routes and enhancing the resilience and efficiency of the network.

For SWISS, refining the TOPSIS multi-decision criteria method to better adapt to evolving operational priorities would also improve alignment with real-world requirements. Future work could involve integrating the GAT model's insights into the TOPSIS method through two key approaches. First, the attention weights generated by the GAT model could be used to dynamically adjust the weights of the criteria in TOPSIS, ensuring that the decision-making process reflects real-time network dynamics and evolving operational priorities. Second, the TOPSIS criteria could be refined by incorporating delay predictions and critical node/edge insights provided by the GAT model. This could include introducing new performance factors, such as predicted delay impact or propagated delay significance, as well as enhancing existing criteria like connection importance based on the GAT's analysis of key routes and nodes. These enhancements would improve the method's alignment with operational realities and decision-making accuracy.

In summary, integrating real-time operational data, incorporating additional airport and flight-specific features, expanding the model's applicability, refining delay propagation analysis, adjusting sensitivity to rare events, and enhancing existing decision-making methods represent significant opportunities for future work. These enhancements have the potential to improve the model's predictive accuracy, provide deeper insights into delay dynamics, and support more effective resource allocation and operational planning. Pursuing these directions will make the GAT model an even more valuable tool for enhancing the resilience and efficiency of SWISS's operations.

9 Conclusion

The primary objective of this study was to evaluate the capability of a Graph Attention Network model to predict reactionary delay distributions within a fleet network, with specific attention to the impact of spoke airports in a hub-and-spoke structure. The GAT model was chosen for its ability to incorporate node-level and edge-level dependencies, allowing it to capture complex, interconnected delay patterns across the network. By analyzing these dependencies, this work aimed to improve the accuracy of delay predictions and provide a more actionable model for operational decision-making.

Through this model, the results reveal that node-level features are the primary drivers of accurate predictions, with edge-level attributes contributing contextual information, where nodes represent the flights, and node-level features are operational parameters linked to the flight. This approach enabled the GAT model to identify key connections such as rotational dependencies, high-volume passenger transfers, and critical spoke-hub flights that contribute to delay propagation. These connections are particularly significant for airlines where spoke airports play a major role in network disruptions. Performance metrics, including RMSE and MAE, confirm the model's accuracy, especially under moderate delay conditions. In fact, the model achieved the highest predictive reliability on medium-delay days with an RMSE of 15.59 minutes and an MAE of 10.50 minutes, with accuracy declining slightly on days of extreme or unpredictable delays, indicating the model's sensitivity to routine operational patterns. On a high-delay day, RMSE values averaged around 37.56 minutes. The RMSE of 15.59 minutes being higher than the MAE of 10.50 minutes suggests that while the model generally performs well, some larger errors (outliers) are disproportionately inflating the RMSE compared to the average error represented by the MAE.

The GAT model's attention weights play a crucial role in enhancing operational decision-making by highlighting critical nodes (flights) and edges (connections) that significantly influence delay propagation within the fleet network. These weights provide a quantitative measure of the relative importance of each flight and connection, enabling a focused analysis of high-impact areas such as rotational dependencies and pivotal spoke-hub routes. By integrating these insights with the existing TOPSIS model used by SWISS Airlines, the combined approach can refine the prioritization of daily critical flights. For instance, the delay predictions and attention-driven identification of impactful flights can be used to augment TOPSIS criteria, including refining the assessment of "connection importance" or introducing new metrics such as the predicted severity of propagated delays. This integration ensures that the selection of priority flights is not only data-driven but also enriched by the nuanced understanding of network dynamics provided by the GAT model.

Further analysis showed that, while effective in most cases, the model struggles with irregular disruptions caused by unpredictable operational events, such as technical issues or severe weather, suggesting a need for weather forecasts and historical data and prediction on unpredictable events. Information on airport capacity, maintenance checks, crew availability, and adverse weather conditions, such as snowstorms, could enhance the model's responsiveness to sudden disruptions. Additionally, airport-specific features, such as runway availability and ground congestion levels, may improve predictions for flights originating from or arriving at highly congested hubs.

Moreover, the effectiveness of the GAT model is closely tied to the volume of training data available. Although a year's worth of data was utilized, research suggests that GAT models perform best when trained on significantly larger datasets, enabling them to capture hidden patterns in highly interconnected networks [21]. Expanding the dataset to include more than two years would not only capture seasonal trends and provide additional data on low-frequency flights but also account for variability in external factors, such as weather patterns or disruptions caused by strikes, which can differ significantly from year to year. By incorporating a wider range of scenarios, such as seasonal variations, low-frequency flight schedules, and external factors like strikes or unusual weather patterns, the model becomes less sensitive to outliers or atypical events from any single year.

While the GAT model demonstrated strength in capturing localized delay propagation, it sometimes showed limitations in predicting the cascading effects across distant nodes in highly interconnected networks. This is partly due to the model's reliance on local attention mechanisms, which prioritize immediate neighbors over distant connections. Studies indicate that this localized focus can reduce accuracy when predicting delays that propagate through several layers of the network [17]. Addressing this limitation may involve integrating global attention mechanisms or hybrid approaches (e.g. combining GAT with LSTM) to extend the model's ability to account for long-range dependencies within the network.

In conclusion, this research underscores the GAT model's potential to capture delay propagation dynamics within a fleet network. The model not only accurately predicts arrival times by understanding how delays spread through the network but also provides valuable insights through its attention weights. These attention weights highlight the most critical nodes (flights) and edges (connections between flights), allowing airlines to identify high-impact flights that, if delayed, could significantly strain operational stability. By combining predicted arrival times with attention weight analysis, airlines can pinpoint these critical flights and allocate additional resources to them to prevent delays from occurring. This proactive approach equips airlines with better insights for managing disruptions, enabling more effective resource allocation and enhancing overall network resilience. With further enhancements, such as expanded datasets and architectural refinements, the GAT model could serve as an invaluable tool for delay forecasting, helping airlines navigate increasingly complex operational environments.

Acknowledgments

I would like to thank David Graber for his invaluable insights and expertise on Graph Attention Networks. I also extend my gratitude to Leonardo Caranti for providing the SWISS model's data and offering a comprehensive overview of the model's structure and feature importance.

References

[1] J. M. Anguita and O. D. Olariaga. Prediction of departure flight delays through the use of predictive tools based on machine learning/deep learning algorithms. *The Aeronautical Journal*, (128):111–133, 2024.

[2] R. Beatty, R. Hsu, L. Berry, and J. Rome. Preliminary evaluation of flight delay propagation through an airline schedule. *Air Traffic Control Quarterly*, 7(4), 1999.

[3] S. Brody, U. Alon, and Y. Eran. How attentive are graph attention networks? *International Conference on Learning Representations*, 2022.

[4] K. Cai, Y. Li, Y. Fang, and Y. Zhu. A deep learning approach for flight delay prediction through time-evolving graphs. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[5] A. Churchill, D. Lovell, and M. Ball. Flight delay propagation impact on strategic air traffic flow management. *Transportation Research Record*, (2177):105–113, 2010.

[6] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *International Conference on Learning Representations*, 2016.

[7] R. Dalmau, F. Ballerini, H. Naessens, S. Belkhoura, and S. Wangnick. An explainable machine learning approach to improve take-off time predictions. *Journal of Air Transport Management*, 95, 2021.

[8] R. Dalmau, G. Murgese, Y. Wandeler, R. Correira, and A. Marsden. Early detection of night curfew infringements by delay propagation with neural networks. 2021.

[9] R. Dalmau, A. Trzmiel, and S. Kirby. Peta: Combining machine learning models to improve estimated time of arrival predictions. *SESAR Innovation Days*, 2023.

[10] W. Du, M. Zhang, Y. Zhang, X. Cao, and J. Zhang. Delay causality network in air transport systems. *Transportation Research Part E*, 118(466-476), 2018.

[11] V. M. Eguíluz, E. Hernández-García, F. Plo, and J. J. Ramasco. Data-driven modelling of the tree of reactionary delays. *6th International Conference on Research in Air Transportation ICRAT*, 2014.

[12] EUROCONTROL. Coda digest: Delays to air transport in europe—annual 2013, 2014.

[13] EUROCONTROL. All-causes delays to air transport in europe - quarter 2 2023, 2023.

[14] H. Fricke and M. Schultz. Delay impacts onto turnaround performance. 2009.

[15] H. Fricke and M. Schultz. Delay impacts onto turnaround performance optimal time buffering for minimizing delay propagation. *8th USA/Europe Air Traffic Management Research and Development Seminar*, 2009.

[16] Z. Jia and X. Cai. Delay propagation network in air transport systems based on refined nonlinear granger causality. *Transportmetrica B*, 10(4):1–13, 2022.

[17] K. Lin, X. Xie, W. Weng, and X. Du. Global-local graph attention: unifying global and local attention for node classification. *The Computer Journal*, 67, 2024.

[18] N. Pyrgiotis. A stochastic and dynamic model of delay propagation within an airport network for policy analysis. *Massachusetts Institute of Technology*, 2012.

[19] J. Sun, T. Dijkstra, K. Aristodemou, V. Buzetelu, T. Falat, T. Hogenelst, N. Prins, and B. Slijper. Designing recurrent and graph neural networks to predict airport and air traffic network delays. *10th International Conference for Research in Air Transportation FAA & Eurocontrol*, 2022.

[20] P. Velickovi, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *International Conference on Learning Representations*, 2018.

[21] A. G. Vrahatis, K. Lazaros, and S. Kotsiantis. Graph attention networks: A comprehensive review of methods and applications. *A Comprehensive Review of Methods and Applications. Future Internet*, 16, 2024.

[22] C. Wu and K. Law. Modelling the delay propagation effects of multiple resource connections in an airline network using a bayesian network model. *Transportation Research*, (122):62–77, 2019.

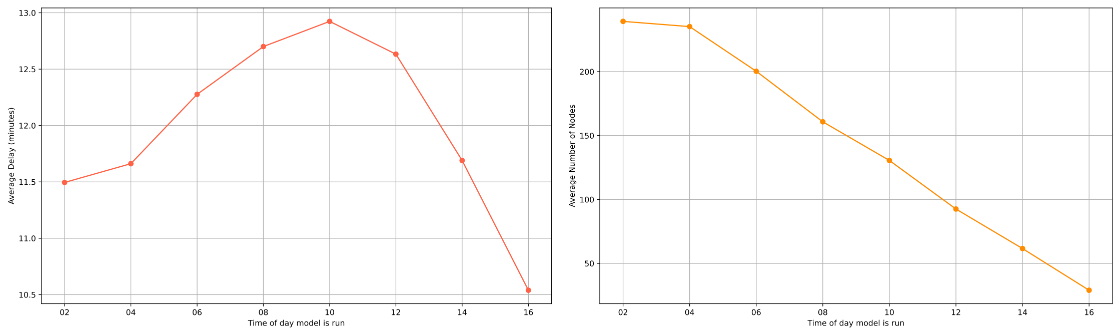
[23] S. Xiao, W. Shiping, D. M. Yuanfei, and G. Wenzhong. Graph neural networks in node classification: survey and evaluation. *Machine Vision and Applications*, 33(1), 2022.

[24] B. Xu, W. Naiyan, T. M. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. 2015.

[25] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *33rd Conference on Neural Information Processing Systems*, 2019.

[26] H. Zheng, Z. Wang, C. Zheng, Y. Wang, X. Fan, W. Cong, and M. Hu. A graph multi-attention network for predicting airport delays. *Transportation Research Part E*, 2024.

Appendix A



(a) Average delay by time-block (b) Number of nodes per time-block

Figure 33: Information about the dataset at each time-block the model is run

From Figure 34 it is important to note that all delay code features (e.g., *PAX BAGGAGE delay code, TECH DAMAGE delay code, CREW delay code, WEJMCER delay code, ATC delay code,* and *REACT delay code*) show zero importance. This result can be explained by the fact that these codes are typically unknown before a flight departs and, therefore, do not influence predictions made before departure. Consequently, including these delay code features is irrelevant to the model's performance. Despite this, these features were included during training to leverage all available data, maintain dataset consistency, explore potential correlations or indirect impacts with other features, and ensure the model could fully learn patterns without excluding potentially useful information.

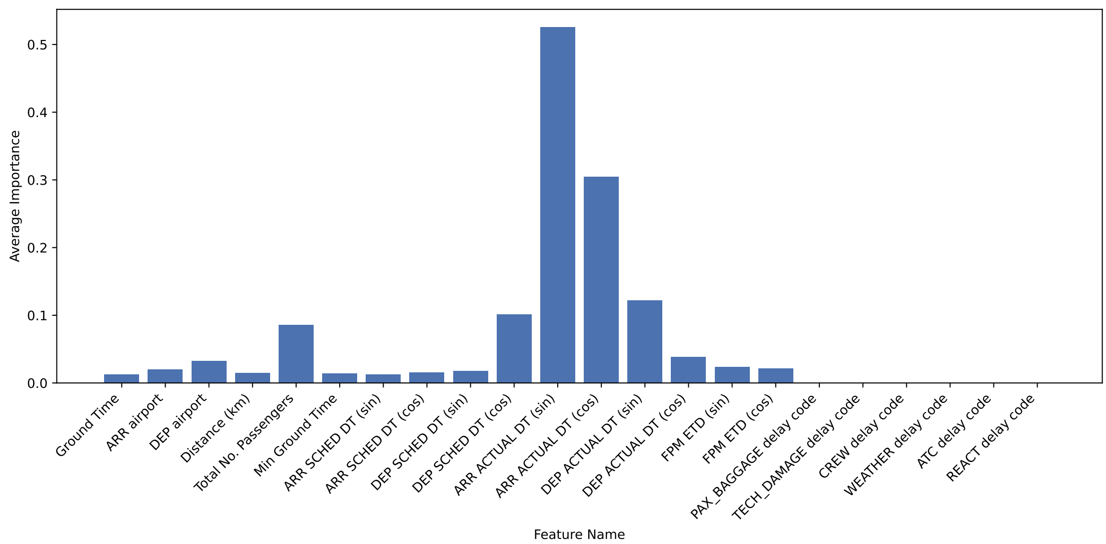
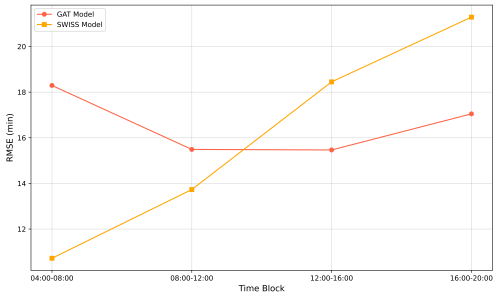
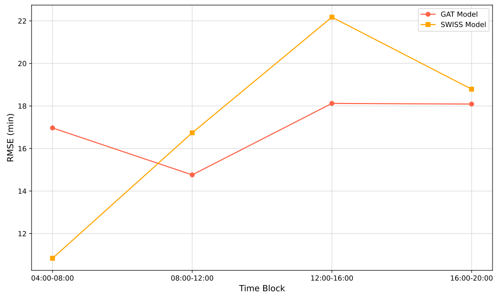


Figure 34: Average node feature importance across samples

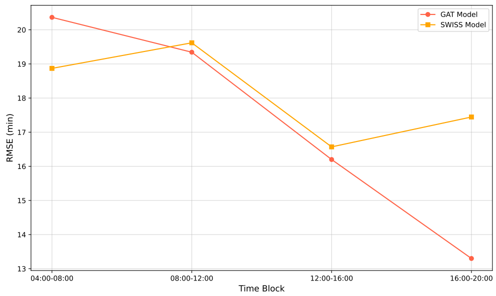
Appendix B



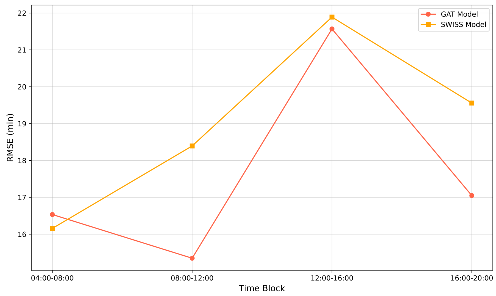
(a) 2024-10-22



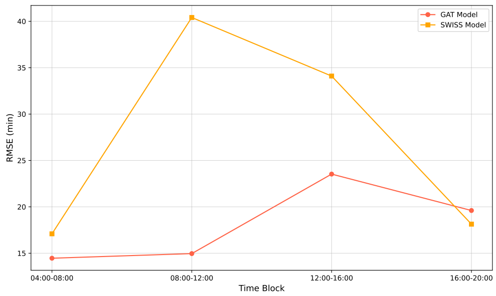
(b) 2024-10-23



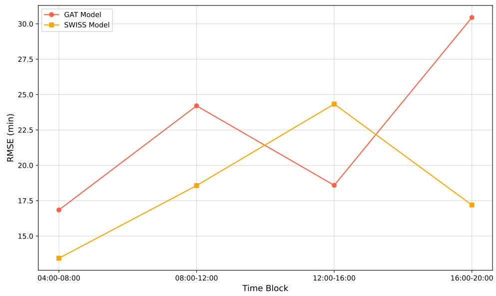
(c) 2024-10-24



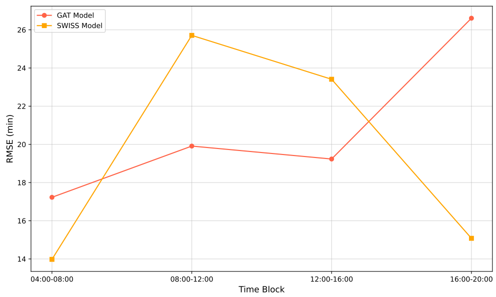
(d) 2024-10-25



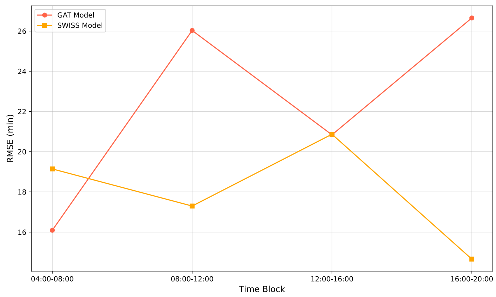
(e) 2024-10-26



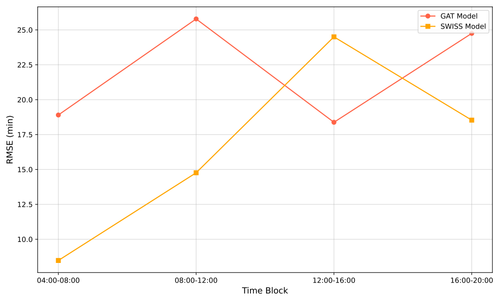
(f) 2024-10-27



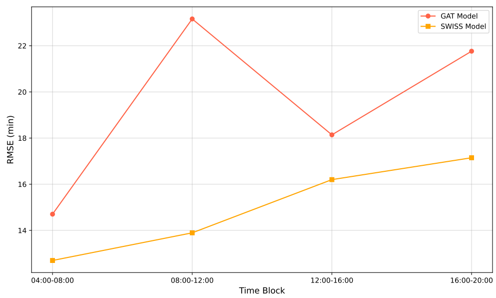
(g) 2024-10-28



(h) 2024-10-29



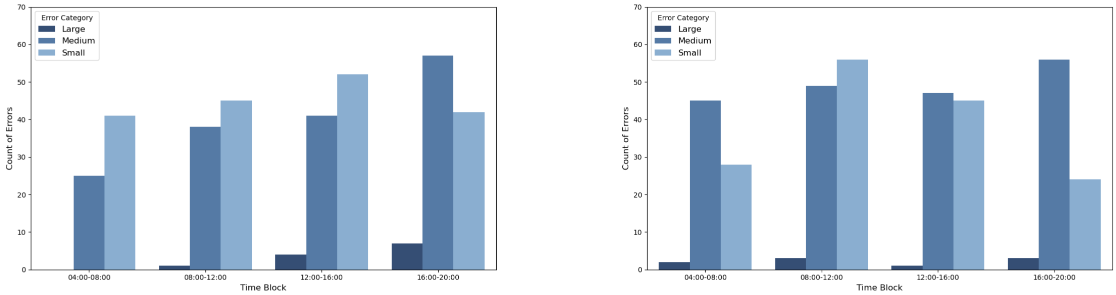
(i) 2024-10-30



(j) 2024-10-31

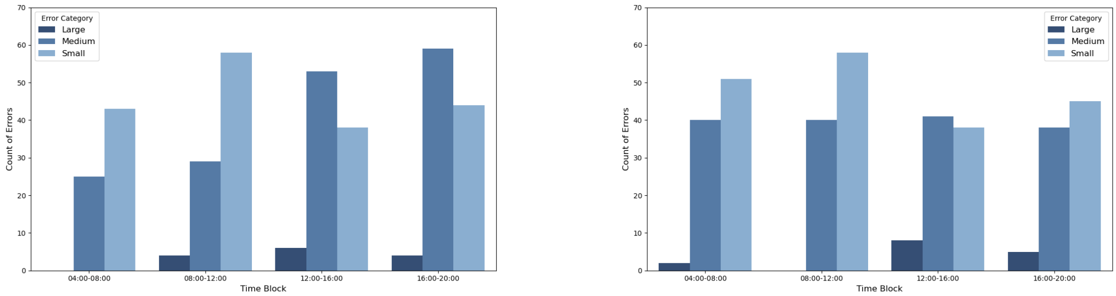
Figure 35: RMSE Comparison Between SWISS and GAT Models for Specific Timeframes

Appendix C



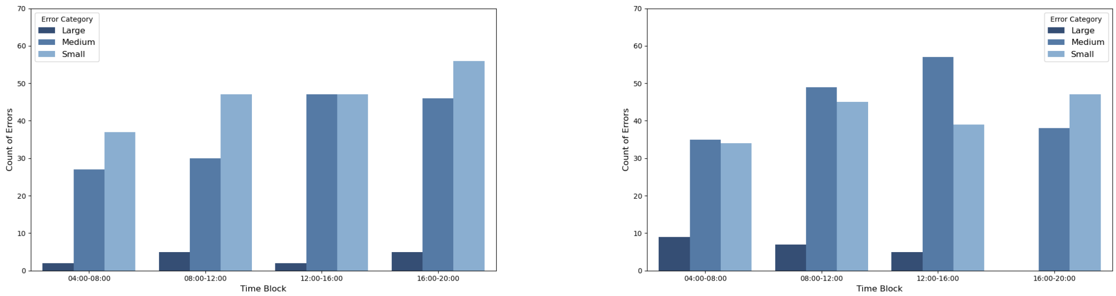
(a) SWISS Model (b) GAT Model

Figure 36: Error category distribution for October 22



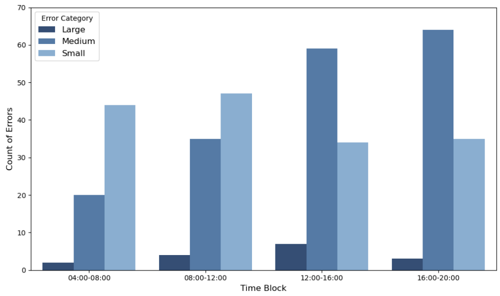
(a) SWISS Model (b) GAT Model

Figure 37: Error category distribution for October 23

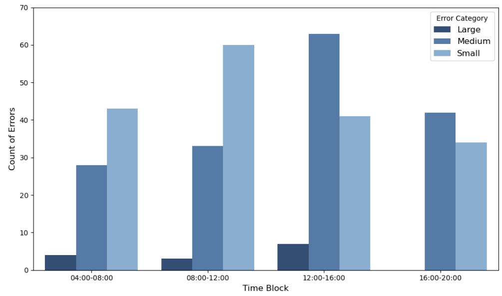


(a) SWISS Model (b) GAT Model

Figure 38: Error category distribution for October 24

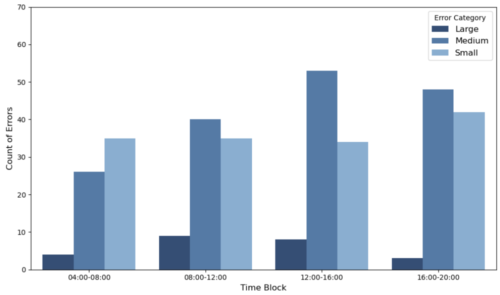


(a) SWISS Model

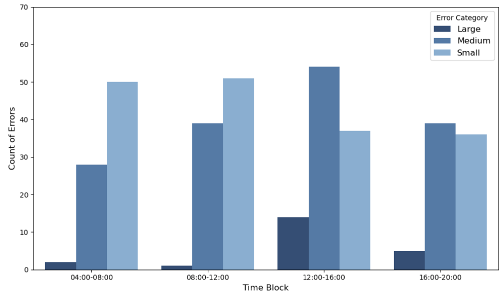


(b) GAT Model

Figure 39: Error category distribution for October 25

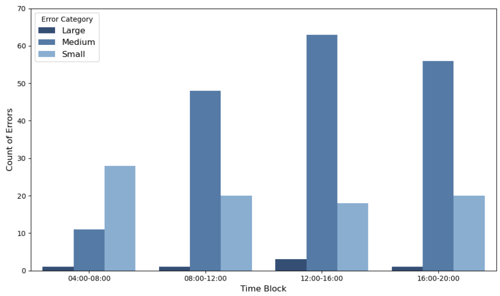


(a) SWISS Model

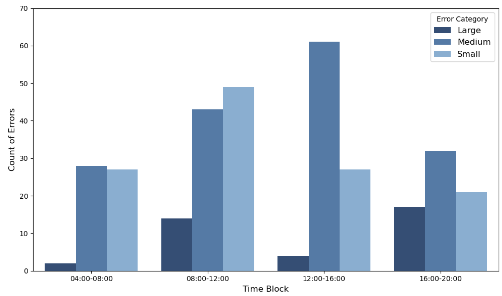


(b) GAT Model

Figure 40: Error category distribution for October 26

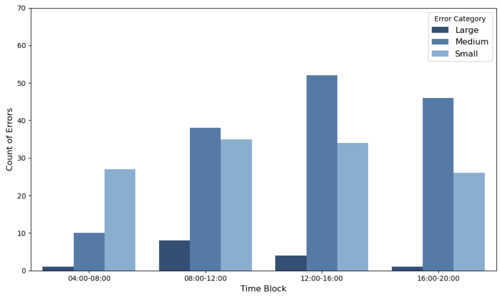


(a) SWISS Model

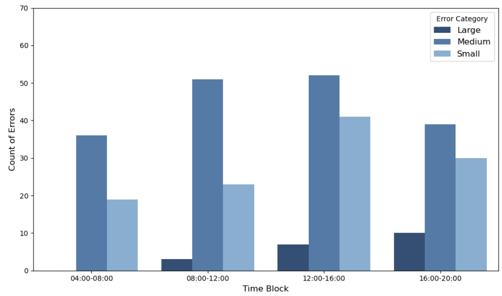


(b) GAT Model

Figure 41: Error category distribution for October 27

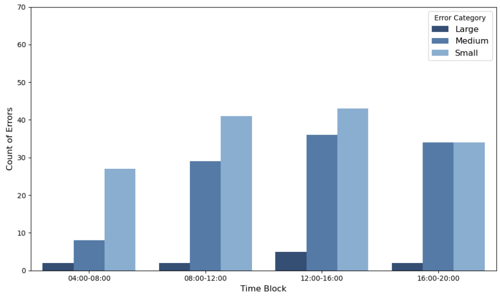


(a) SWISS Model

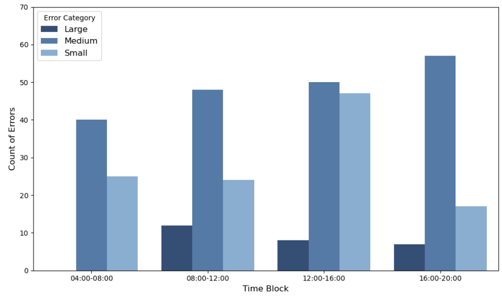


(b) GAT Model

Figure 42: Error category distribution for October 28

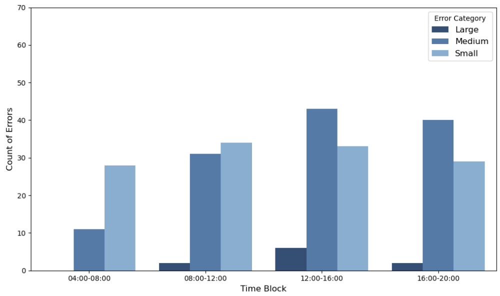


(a) SWISS Model

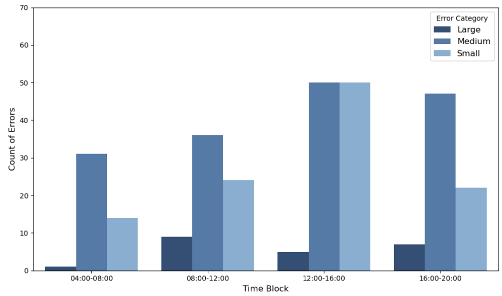


(b) GAT Model

Figure 43: Error category distribution for October 29

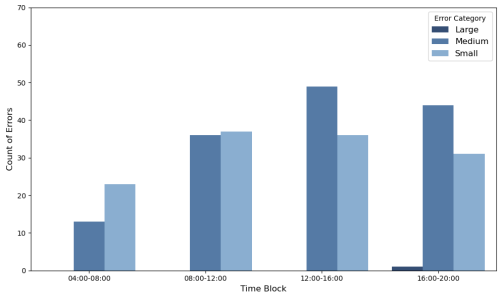


(a) SWISS Model

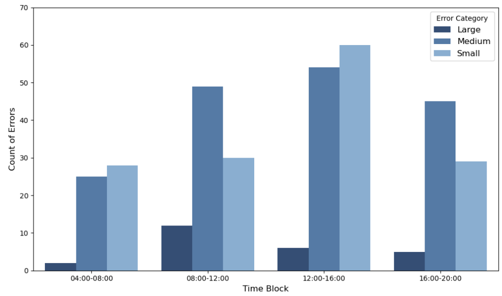


(b) GAT Model

Figure 44: Error category distribution for October 30



(a) SWISS Model



(b) GAT Model

Figure 45: Error category distribution for October 31

For Figures 36, 37, 38, 39, 40, 41, 42, 43, 44, 45 small errors account for errors less than 10 minutes, medium errors are errors between 10 to 40 minutes, and, lastly, large errors are those larger than 40 minutes.

Appendix D

It is important to mention that the flight numbers in the following figures have been anonymized, and the flight numbers do not correspond to real flights.

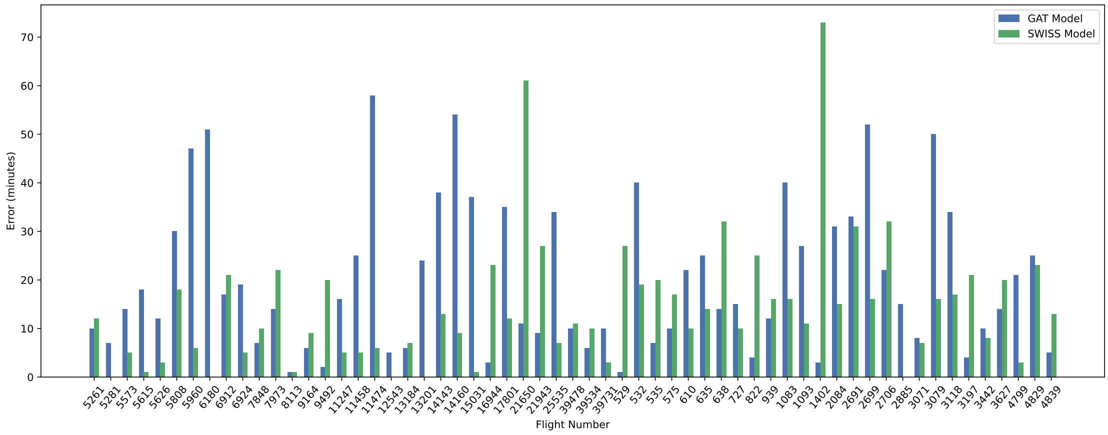


Figure 46: Comparison of GAT model with SWISS model for 30/10 in 16h-20h

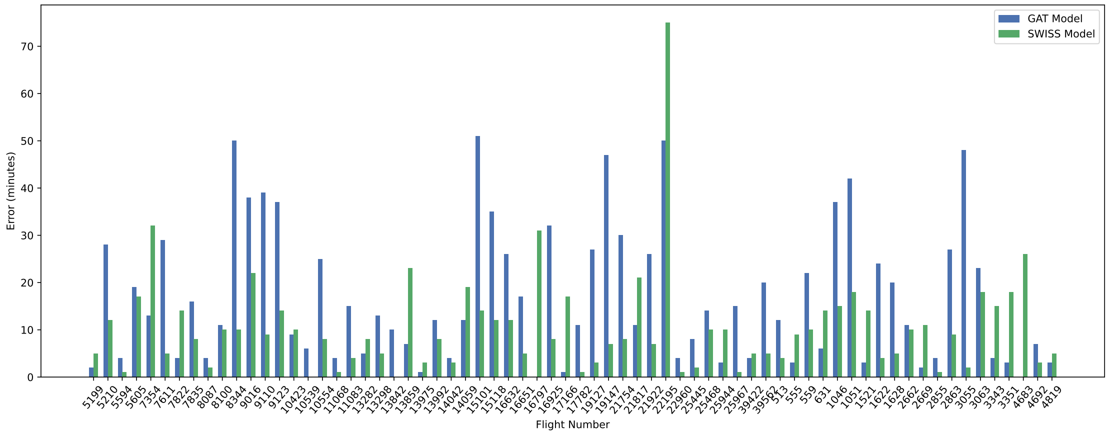


Figure 47: Comparison of GAT model with SWISS model for 29/10 in 12h-16h

Appendix E

The flight numbers and airport codes (except ZRH and GVA) in the following figures have been changed to protect confidentiality and do not reflect real values.

In the following tables, the arrow symbol (→) represents a passenger connection. The number preceding the arrow indicates the connecting time in minutes, and the code following the arrow represents the destination of the connecting flight.

Table 21: 12/10/2024 - Wave 1

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
5722	5722	1	05:00	1	—	PLW	ZRH
5851	5851	1	04:55	1	—	PLW	GVA
6070	6070	1	05:15	1	—	HFR	ZRH
8037	8037	1	05:10	1	Priority Flight	BXF	ZRH
9409	9409	1	04:55	1	—	FKJ	ZRH
9686	9686	1	04:50	1	—	ZMK	ZRH
9994	9994	1	05:10	1	—	DRL	ZRH
10251	10251	1	05:10	1	—	LKT	GVA
10539	10539	1	05:05	1	—	GXH	ZRH
11038	11038	1	05:25	1	—	ZLB	ZRH
12416	12416	1	05:10	1	—	YKZ	ZRH
13925	13925	1	05:00	1	—	ZRH	GYP
14194	14194	1	04:05	1	—	ZRH	CMT
14686	14686	1	05:00	1	—	ZPR	ZRH
14789	14789	1	04:50	1	Priority Flight	NKJ	ZRH
15083	15083	1	04:15	1	—	ZRH	BRP
15170	15170	1	04:55	1	—	VXH	ZRH
16054	16054	1	04:40	1	—	DXB	ZRH
19088	19088	1	05:10	1	—	XTR	ZRH
20440	20440	1	05:00	1	—	TRF	ZRH
7898	25898	1.03e-06	04:10	1	244min → WRH	GHF	ZRH
7898	14858	8.46e-07	04:10	1	203min → NKJ	GHF	ZRH
2103	17782	3.98e-07	04:45	1	292min → PYD	LHT	ZRH
7898	14009	2.19e-07	04:10	1	219min → TKV	GHF	ZRH
23088	10079	1.83e-07	04:00	1	582min → DRL	CPR	ZRH
7898	17782	1.72e-07	04:10	1	238min → PYD	GHF	ZRN
2728	16925	1.69e-07	05:00	1	254min → JMC	VWB	ZRH
2728	1010	1.47e-09	05:00	1	193min → BKC	VWB	ZRH
7898	23153	7.74e-10	04:10	1	418min → KGN	GHF	ZRH
2103	25944	1.90e-10	04:45	1	592min → WRH	LHT	ZRH

Table 22: 12/10/2024 - Wave 2 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
3142	3142	0.967776	06:50	2	—	GVA	MQT
628	628	0.921088	07:55	2	—	GVA	VSL
801	801	0.803752	06:50	2	—	ZRH	FRP
23651	23651	0.758332	07:40	2	—	ZRH	XLB
25378	25378	0.746926	07:40	2	—	ZRH	WQK
14807	14807	0.740449	07:35	2	Prio Flight	ZRH	NKJ
15188	15188	0.729746	07:40	2	Prio Flight	ZRH	VXH
8049	8049	0.725521	07:40	2	Prio Flight	ZRH	BXF
9423	9423	0.699962	07:25	2	Prio Flight	ZRH	FKJ
17147	17147	0.699274	07:40	2	Prio Flight	HLB	ZRH
23218	23218	0.691551	07:50	2	—	ZRH	KGN
13022	13022	0.677294	07:00	2	Prio Flight	NJK	ZRH
10554	10554	0.673501	07:25	2	—	ZRH	GXH
16596	16596	0.664423	08:20	2	—	JMC	GVA
20460	20460	0.656946	08:10	2	—	ZRH	TRF
3335	3335	0.653221	07:45	2	Prio Flight	ZRH	FRB
786	786	0.653049	07:25	2	—	ZRH	WJN
770	770	0.649965	07:45	2	—	JMC	ZRH
786	786	0.628172	07:50	2	—	ZRH	ZTN
818	818	0.621064	07:35	2	Prio Flight	ZRH	LKT
14807	22385	9.60e-05	07:35	2	102min → XNZ	NKJ	ZRH
506	16962	8.74e-05	07:40	2	321min → JMC	VSL	ZRH
8723	22385	8.28e-05	07:45	2	294min → XNZ	FRP	ZRH
14704	6912	7.46e-05	07:30	2	381min → WBT	ZPR	ZRH
8049	22385	7.07e-05	07:40	2	114min → XNZ	BXF	ZRH
19108	2001	7.02e-05	07:35	2	296min → TSK	XTR	ZRH
20319	25944	5.27e-05	07:45	2	389min → WRH	LHT	ZRH
4770	15240	5.20e-05	07:40	2	369min → VXH	NFX	ZRH
628	638	3.48e-05	07:55	2	402min → VSL	VSL	GVA
16072	25944	1.24e-05	07:35	2	324min → WRH	DXB	ZRH

Table 23: 12/10/2024 - Wave 3 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
39506	39506	0.999982	08:50	3	—	ZRH	GVA
24264	24264	0.999936	09:45	3	—	GVA	KXP
7354	7354	0.999438	12:00	3	—	ZRH	TKH
16614	16614	0.999009	11:55	3	—	GVA	JMC
27681	27681	0.998917	10:20	3	—	GVA	WRN
17166	17166	0.998641	10:40	3	—	ZRH	HLB
810	810	0.998453	12:00	3	Prio Flight	ZRH	TRN
21401	21401	0.998188	11:15	3	Prio Flight	ZRH	QMV
22385	22385	0.992121	11:20	3	—	XNZ	ZRH
28106	28106	0.991041	09:10	3	—	GVA	MKZ
1051	1051	0.974658	11:15	3	—	ZRH	FRL
21754	21754	0.971279	10:45	3	—	ZPD	ZRH
631	631	0.966201	10:55	3	—	TSK	ZRH
631	631	0.966201	10:55	3	—	VSL	GVA
23347	23347	0.921423	10:45	3	—	ZRH	KGN
27916	27916	0.888651	08:50	3	—	GVA	TSK
19127	19127	0.878497	10:20	3	—	XTR	ZRH
16797	16797	0.877966	11:20	3	—	ZRH	JMC
23131	23131	0.751164	09:50	3	Prio Flight	ZRH	KGN
5873	822	0.000197	08:45	3	346min → KGN	GVA	PLW
2662	2728	0.000188	10:30	3	980min → ZRH	ZRH	VWB
5563	610	0.000172	08:45	3	247min → WRN	MCR	ZRH
5563	22385	0.000172	08:45	3	76min → XNZ	MCR	ZRH
5873	6959	0.000159	08:45	3	104min → QMV	GVA	PLW
1594	22385	0.000121	08:35	3	78min → XNZ	VXJ	ZRH
12479	12432	0.000113	10:25	3	1174min → ZRH	ZRH	YKZ
5884	4683	9.51e-05	08:45	3	989min → HFR	PLW	GVA
2045	2103	1.61e-06	10:55	3	963min → ZRH	ZRH	LHT
7835	7898	1.59e-06	10:35	3	917min → ZRH	ZRH	GHF

Table 24: 12/10/2024 - Wave 4 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
15223	15223	0.997948	12:35	4	—	ZRH	VXH
10065	10065	0.997674	12:45	4	—	ZRH	DRL
9123	9123	0.997103	12:50	4	—	ZRH	JFL
8100	8100	0.996624	13:10	4	Prio Flight	ZRH	BXF
559	559	0.996530	12:45	4	—	ZRH	VSL
25468	25468	0.996042	12:45	4	—	ZRH	WQK
14026	14026	0.995725	12:35	4	—	ZRH	TKV
7973	7973	0.995374	13:05	4	—	GVA	BXF
5605	5605	0.993798	12:55	4	Prio Flight	ZRH	MCR
21922	21922	0.991910	13:10	4	—	ZPD	GVA
14669	14669	0.991838	12:40	4	Prio Flight	ZRH	ZPR
1628	1628	0.989617	12:50	4	—	ZRH	VXJ
14059	14059	0.988390	12:50	4	—	ZRH	CMT
21650	21650	0.987923	13:35	4	—	GVA	QMV
15118	15118	0.987740	13:05	4	—	ZRH	BRP
17801	17801	0.986257	13:10	4	—	ZRH	PYD
8344	8344	0.984130	14:00	4	—	TKH	GVA
1697	1697	0.983685	12:35	4	—	ZRH	YNZ
19147	19147	0.979281	12:55	4	—	ZRH	XTR
8164	8164	0.978598	13:45	4	—	BXF	ZRH
14026	7885	0.006027	12:35	4	332min $\rightarrow$ GHF	TKV	ZRH
14059	6180	0.006009	12:50	4	182min $\rightarrow$ HFR	CMT	ZRH
14026	6912	0.005315	12:35	4	122min $\rightarrow$ WBT	TKV	ZRH
559	25944	0.005004	12:45	4	49min $\rightarrow$ WRH	VSL	ZRH
13298	25944	0.003037	12:50	4	83min $\rightarrow$ WRH	NJK	ZRH
39562	25944	0.002185	13:00	4	97min $\rightarrow$ WRH	GVA	ZRH
1697	25944	0.001843	12:35	4	111min $\rightarrow$ WRH	YNZ	ZRH
1697	3102	2.10e-08	12:35	4	960min $\rightarrow$ MQT	YNZ	ZRH
39534	24242	1.67e-08	15:55	4	660min $\rightarrow$ KXP	ZRH	GVA
1527	3102	1.24e-09	13:00	4	875min $\rightarrow$ MQT	TNY	ZRH

Table 25: 12/10/2024 - Wave 5

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
5626	5626	0.675679	17:10	5	—	ZRH	MCR
5808	5808	0.549024	17:10	5	Prio Flight	PLW	ZRH
25967	25967	0.513789	17:20	5	Prio Flight	ZRH	WRH
6180	6180	0.475332	17:25	5	Prio Flight	HFR	ZRH
14160	14160	0.458135	17:05	5	—	ZRH	CMT
6924	6924	0.440593	17:30	5	Prio Flight	ZRH	WBT
4849	4849	0.403910	17:45	5	—	ZRH	NFX
1582	1582	0.375740	17:10	5	—	ZRH	VXJ
11262	11262	0.350832	17:10	5	Prio Flight	ZRH	ZLB
13892	13892	0.317144	17:25	5	—	ZRH	GYP
2893	2893	0.287483	17:10	5	—	ZRH	MQF
5808	532	0.275760	17:10	5	Prio Flight	PLW	ZRH
6180	16779	0.082693	17:25	5	Prio Flight	HFR	ZRH
6180	21629	0.056502	17:25	5	Prio Flight	HFR	ZRH
5626	5594	0.040937	17:10	5	970min $\rightarrow$ MCR	ZRH	MCR
14160	14143	0.020154	17:45	5	Rotation	ZRH	CMT
6180	24242	4.47E-10	17:25	5	Prio Flight	HFR	ZRH
6924	3102	1.59E-10	17:30	5	Prio Flight	ZRH	WBT

Table 26: 25/10/2024 - Wave 1

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
5531	5531	1	05:15	1	—	MCR	ZRH
5851	5851	1	04:55	1	—	PLW	GVA
6070	6070	1	05:15	1	—	HFR	ZRH
7898	7898	1	04:10	1	—	ZRH	GHF
8037	8037	1	05:10	1	—	BXF	ZRH
9029	9029	1	04:55	1	—	JFL	ZRH
9409	9409	1	05:00	1	—	FKJ	ZRH
9994	9994	1	05:10	1	—	DRL	ZRH
10423	10423	1	04:50	1	—	FKN	ZRH
10539	10539	1	05:05	1	—	GXH	ZRH
11024	11024	1	04:20	1	—	ZRH	ZLB
11128	11128	1	04:50	1	—	BWR	ZRH
12416	12416	1	05:10	1	—	YKZ	ZRH
13925	13925	1	05:00	1	—	ZRH	GYP
14109	14109	1	06:05	1	—	CMT	ZRH
14194	14194	1	04:05	1	—	ZRH	CMT
14482	14482	1	04:40	1	—	NKJ	GVA
14789	14789	1	04:45	1	—	NKJ	ZRH
15083	15083	1	04:15	1	—	ZRH	BRP
15170	15170	1	04:55	1	Priority Flight	VXH	ZRH
2728	17147	2.07e-06	05:00	1	81min → HLB	VWB	ZRH
7452	17782	1.34e-06	04:25	1	209min → PYD	TKH	ZRH
2728	15205	9.54e-07	05:00	1	221min → VXH	VWB	ZRH
5179	5210	8.93e-07	05:30	1	372min → ZRH	ZRH	FZY
7452	16090	4.98e-07	04:25	1	224min → DXB	TKH	ZRH
2728	14858	8.12e-08	05:00	1	221min → NKJ	VWB	ZRH
600	638	5.00e-08	05:00	1	158min → KGS	VSL	ZRH
6070	13875	2.81e-09	05:15	1	168min → SBZ	ZRH	HFR
2728	178	4.68e-10	05:00	1	292min → PVG	VWB	ZRH
2728	638	1.72e-10	05:00	1	185min → KGS	VWB	ZRH

Table 27: 25/10/2024 - Wave 2 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
9423	9423	0.999455	07:25	2	—	ZRH	FKJ
6081	6081	0.999404	07:00	2	—	ZRH	HFR
3142	3142	0.999115	06:50	2	—	GVA	MQT
509	509	0.996081	08:30	2	—	VSL	ZRH
22746	22746	0.994750	07:25	2	—	GVA	CPR
25378	25378	0.985673	07:40	2	—	ZRH	WQK
427	427	0.856538	07:30	2	—	ZRH	ZQF
5542	5542	0.830489	07:35	2	—	ZRH	MCR
39450	39450	0.816612	08:00	2	—	ZRH	GVA
5189	5189	0.777946	07:40	2	—	ZRH	FZY
11143	11143	0.718704	07:25	2	—	ZRH	BWR
10009	10009	0.714352	07:35	2	—	ZRH	DRL
21277	21277	0.706048	07:40	2	Prio Flight	ZRH	QMV
19108	19108	0.687459	07:35	2	Prio Flight	ZRH	XTR
8049	8049	0.670164	07:40	2	Prio Flight	ZRH	BXF
12432	12432	0.641435	07:35	2	—	ZRH	YKZ
10554	10554	0.638026	07:25	2	—	ZRH	GXH
5862	5862	0.636328	06:55	2	—	GVA	PLW
10438	10438	0.629992	07:35	2	—	ZRH	FKN
14499	14499	0.628823	07:25	2	—	GVA	NKJ
17203	22385	1.42e-05	07:35	2	57min → XNZ	HLB	ZRH
12432	22385	1.36e-05	07:35	2	135min → XNZ	VWB	ZRH
15188	22385	1.18e-05	07:40	2	297min → XNZ	VXH	ZRH
14945	22385	1.05e-05	07:40	2	578min → XNZ	VXJ	ZRH
10009	22385	1.03e-05	07:35	2	380min → XNZ	DRL	ZRH
19108	22385	1.01e-05	07:35	2	324min → XNZ	FKN	ZRH
506	22385	1.01e-05	07:40	2	803min → XNZ	VSL	ZRH
11143	22385	9.88e-06	07:25	2	882min → XNZ	BWR	ZRH
770	22385	8.29e-06	07:50	2	568min → XNZ	MQF	ZRH
5542	22385	5.18e-06	07:35	2	957min → XNZ	MCR	ZRH

Table 28: 25/10/2024 - Wave 3 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
24264	24264	0.999896	09:45	3	—	GVA	KXP
16072	16072	0.999269	10:55	3	—	ZRH	DXB
810	810	0.999090	12:00	3	Prio Flight	ZRH	TRN
7354	7354	0.997889	12:00	3	—	ZRH	TKH
7822	7822	0.997683	11:15	3	Prio Flight	ZRH	GHF
21401	21401	0.996209	11:15	3	Prio Flight	ZRH	QMV
22385	22385	0.993760	11:10	3	Prio Flight	XNZ	ZRH
17166	17166	0.985150	10:40	3	—	ZRH	HLB
23347	23347	0.985110	10:45	3	—	ZRH	KGN
555	555	0.975028	10:05	3	Prio Flight	VSL	ZRH
7611	7611	0.973612	08:35	3	—	GHF	GVA
2662	2662	0.966612	10:30	3	—	VWB	ZRH
14858	14858	0.945376	10:00	3	—	NKJ	ZRH
513	513	0.937180	11:05	3	—	ZRH	VSL
1046	1046	0.888486	09:50	3	—	FRL	ZRH
5884	5884	0.864937	10:55	3	—	GVA	PLW
5489	5489	0.828623	11:45	3	—	MCR	GVA
23282	23282	0.811139	10:10	3	—	KGN	ZRH
16090	16090	0.810247	10:30	3	—	DXB	ZRH
7961	7961	0.767690	10:15	3	—	BXF	GVA
39929	638	9.38e-05	09:55	3	295min → VSL	ZRH	GVA
2833	610	8.73e-05	08:45	3	274min → WRN	MQF	ZRH
1594	610	5.18e-05	08:35	3	228min → WRN	VXJ	ZRH
39506	610	4.06e-05	08:50	3	268min → WRN	GVA	ZRH
4790	610	3.85e-05	08:55	3	233min → WRN	NFX	ZRH
4790	22385	2.37e-05	08:55	3	57min → XNZ	NFX	ZRH
25876	22385	2.22e-05	08:40	3	74min → XNZ	WRH	ZRH
2833	22385	2.09e-05	08:45	3	98min → XNZ	MQF	ZRH
1594	418	1.96e-05	08:35	3	249min → TJP	VXJ	ZRH
2662	2728	6.58e-08	10:30	3	923min → ZRH	ZRH	VWB

Table 29: 25/10/2024 - Wave 4 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
9561	9561	0.996934	12:40	4	—	ZRH	YLB
9123	9123	0.995363	12:50	4	—	ZRH	JFL
10065	10065	0.994614	12:45	4	—	ZRH	DRL
23564	23564	0.994544	13:45	4	—	GVA	KGN
14059	14059	0.991098	12:50	4	—	ZRH	CMT
21943	21943	0.989868	13:30	4	—	GVA	ZPD
9451	9451	0.987710	12:55	4	Prio Flight	ZRH	FKJ
21650	21650	0.986838	13:35	4	—	GVA	QMV
14669	14669	0.985951	12:40	4	Prio Flight	ZRH	ZPR
9506	9506	0.979518	13:05	4	—	ZRH	JNY
5605	5605	0.978375	12:55	4	—	ZRH	MCR
4692	4692	0.976536	12:55	4	—	ZRH	NFX
23003	23003	0.972769	13:00	4	—	ZRH	CPR
13859	13859	0.972016	12:55	4	—	ZRH	GYP
7501	7501	0.964469	12:55	4	Prio Flight	ZRH	YRP
17801	17801	0.963321	13:10	4	—	ZRH	PYD
7848	7848	0.959135	13:50	4	—	ZRH	GHF
11083	11083	0.939541	13:50	4	—	ZRH	ZLB
39562	1159	0.011594	13:00	4	111min $\rightarrow$ ZGR	GVA	ZRH
11083	2885	0.010724	12:55	4	181min $\rightarrow$ MQF	ZLB	ZRH
11083	21339	0.009698	12:55	4	178min $\rightarrow$ QMV	ZLB	ZRH
39562	6912	0.005337	13:00	4	132min $\rightarrow$ WBT	GVA	ZRH
14876	6912	0.003144	12:45	4	85min $\rightarrow$ WBT	NKJ	ZRH
529	600	1.86e-07	13:30	4	810min $\rightarrow$ ZRH	ZRH	VSL
5916	6070	5.99e-09	16:45	4	133min $\rightarrow$ TLS	GVA	PLW
22406	8975	1.22e-09	15:00	4	119min $\rightarrow$ GYZ	XNZ	ZRH
5916	9409	9.06e-10	16:45	4	163min $\rightarrow$ FKJ	GVA	PLW
23304	2728	1.76e-10	13:00	4	889min $\rightarrow$ VWB	KGN	ZRH

Table 30: 25/10/2024 - Wave 5

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
944	944	0.692511	17:45	5	—	ZRH	DRK
13892	13892	0.687162	17:25	5	—	ZRH	GYP
11262	11262	0.642274	17:10	5	—	ZRH	ZLB
14160	14160	0.604937	17:05	5	—	ZRH	CMT
3079	3079	0.604636	17:10	5	Prio Flight	ZRH	MQT
6924	6924	0.598723	17:30	5	—	ZRH	WBT
1402	1402	0.578614	17:10	5	—	VXJ	GVA
6180	6180	0.486306	17:25	5	Prio Flight	HFR	ZRH
39788	39788	0.438663	18:10	5	—	ZRH	GVA
2893	2893	0.416810	17:10	5	Prio Flight	ZRH	MQF
3450	3450	0.401451	17:45	5	—	ZRH	FRB
1582	1582	0.390887	17:10	5	—	ZRH	VXJ
5808	5808	0.318944	17:10	5	Prio Flight	PLW	ZRH
5808	4839	0.199243	17:10	5	Prio Flight	PLW	ZRH
6180	21629	0.127438	17:25	5	Prio Flight	HFR	ZRH
2893	2885	0.083601	17:10	5	Prio Flight	ZRH	MQF
5808	5873	0.067159	17:10	5	Prio Flight	PLW	ZRH
6180	22981	0.035598	17:25	5	Prio Flight	HFR	ZRH
6180	24242	1.59E-08	17:25	5	Prio Flight	HFR	ZRH
6180	22725	9.37E-09	17:25	5	Prio Flight	HFR	ZRH
6180	3134	7.31E-09	17:25	5	Prio Flight	HFR	ZRH
5808	6070	6.72E-09	17:10	5	Prio Flight	PLW	ZRH
5808	424	3.71E-09	17:10	5	Prio Flight	PLW	ZRH
5808	9409	1.81E-09	17:10	5	Prio Flight	PLW	ZRH
2893	25355	1.08E-10	17:10	5	Prio Flight	ZRH	MQF

Table 31: 04/11/2024 - Wave 1

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
5179	5179	1	06:20	1	Prio Flight	FZY	ZRH
5489	5489	1	05:50	1	—	MCR	GVA
5531	5531	1	06:10	1	—	MCR	ZRH
6070	6070	1	06:10	1	—	HFR	ZRH
7898	7898	1	05:10	1	—	ZRH	GHF
8037	8037	1	06:05	1	Prio Flight	BXF	ZRH
8202	8202	1	05:50	1	—	ZRH	BXF
8696	8696	1	04:05	1	—	ZRH	XKJ
9029	9029	1	06:00	1	Prio Flight	JFL	ZRH
9409	9409	1	06:20	1	—	FKJ	ZRH
9994	9994	1	05:55	1	Prio Flight	DRL	ZRH
11024	11024	1	05:25	1	—	ZRH	ZLB
13462	13462	1	05:45	1	—	ZRH	NJK
16669	16669	1	06:05	1	—	JMC	GVA
17839	17839	1	04:30	1	—	ZRH	PYD
19088	19088	1	06:15	1	Prio Flight	XTR	ZRH
21629	21629	1	06:00	1	—	ZRH	PYD
22725	22725	1	05:00	1	—	CPR	GVA
23672	23672	1	05:55	1	Prio Flight	XLB	ZRH
25355	25355	1	06:10	1	—	WQK	ZRH
17018	5199	2.06e-06	05:05	1	243min → FZY	JMC	ZRH
17018	17707	8.90e-07	05:05	1	185min → TPN	JMC	ZRH
17018	1046	7.98e-07	05:05	1	211min → VSL	JMC	ZRH
5531	5626	3.95e-07	06:10	1	627min → ZRH	GVA	MCR
17018	806	3.56e-07	05:05	1	155min → TRN	JMC	ZRH
5179	5271	2.46e-07	06:20	1	604min → ZRH	ZRH	FZY
17018	9492	6.64e-08	05:05	1	290min → JNY	JMC	ZRH
17018	2662	3.86e-08	05:05	1	290min → VWB	JMC	ZRH
17018	2084	1.08e-09	05:05	1	527min → LHT	JMC	ZRH
4760	4731	1.40e-10	06:15	1	591min → ZRH	ZRH	NFX

Table 32: 04/11/2024 - Wave 2 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
11458	11458	1	07:45	2	Prio Flight	TLY	ZRH
39394	39394	1	06:40	2	—	ZRH	GVA
649	649	1	07:00	2	—	GVA	VSL
1588	1588	0.999634	06:35	2	Prio Flight	VXJ	ZRH
6081	6081	0.989730	08:10	2	Prio Flight	ZRH	HFR
4239	4239	0.985836	08:30	2	Prio Flight	ZRH	XMC
1024	1024	0.747564	08:10	2	Prio Flight	ZRH	FRL
10009	10009	0.739169	08:25	2	Prio Flight	ZRH	DRL
2825	2833	0.630983	08:05	2	—	MQF	VSL
22746	22746	0.600988	07:30	2	—	GVA	CPR
2825	2825	0.591950	08:05	2	—	MQF	VSL
649	2728	0.559708	07:00	2	—	GVA	VSL
5499	5499	0.558563	08:25	2	—	GVA	MCR
14109	14126	0.540307	06:35	2	—	CMT	ZRH
39366	39506	0.537248	07:55	2	—	GVA	ZRH
3142	3142	0.498309	08:00	2	—	GVA	MQT
39394	600	0.415783	06:40	2	—	ZRH	GVA
5552	5563	0.413938	07:45	2	Prio Flight	MCR	ZRH
1065	1069	0.411830	06:55	2	Prio Flight	FRL	ZRH
14109	14109	0.388796	06:35	2	—	CMT	ZRH
10009	2885	0.000120	08:25	2	344min → JFL	DRL	ZRH
10009	9164	0.000116	08:25	2	359min → MQF	DRL	ZRH
649	638	5.73e-05	07:00	2	500min → VSL	VSL	GVA
3142	3182	5.63e-05	08:00	2	450min → MQT	MQT	GVA
39394	3071	5.60e-05	06:40	2	488min → MQT	GVA	ZRH
5499	24242	4.16e-05	08:25	2	196min → KXP	MCR	GVA
22746	610	3.72e-05	08:25	2	357min → VSL	CPR	GVA
39366	24242	3.35e-05	08:25	2	297min → KXP	ZRH	GVA
649	610	2.55e-05	08:25	2	381min → VSL	VSL	GVA
2825	2893	2.27e-05	08:25	2	532min → ZRH	ZRH	MQF

Table 33: 04/11/2024 - Wave 3 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
4770	4770	0.999930	08:35	3	—	ZRH	NFX
19108	19108	0.999851	08:50	3	Prio Flight	ZRH	XTR
16797	16797	0.999843	11:45	3	Prio Flight	ZRH	JMC
7354	7354	0.999757	12:00	3	—	ZRH	TKH
19127	19127	0.999160	11:20	3	—	XTR	ZRH
21401	21401	0.998872	12:00	3	Prio Flight	ZRH	QMV
21922	21922	0.998634	10:50	3	—	ZPD	GVA
7501	7501	0.998312	09:00	3	Prio Flight	ZRH	YRP
21650	21650	0.996599	09:10	3	—	GVA	QMV
16687	16687	0.995488	10:20	3	—	GVA	JMC
806	806	0.993252	10:00	3	—	TRN	ZRH
21754	21754	0.991576	11:35	3	—	ZPD	ZRH
8087	8087	0.981964	11:15	3	—	BXF	ZRH
285	285	0.981681	11:50	3	—	CAI	ZRH
10051	10051	0.943320	11:05	3	—	DRL	ZRH
23694	23694	0.925909	09:00	3	Prio Flight	ZRH	XLB
16632	16632	0.916080	09:30	3	—	JMC	GVA
5594	5594	0.886285	11:30	3	—	MCR	ZRH
22174	22174	0.867596	10:15	3	—	XNZ	GVA
5189	5189	0.837367	08:40	3	Prio Flight	ZRH	FZY
5189	285	2.36e-05	08:40	3	137min → CAI	FZY	ZRH
39506	285	2.35e-05	09:35	3	98min → CAI	GVA	ZRH
2833	6912	2.15e-05	09:55	3	410min → WBT	MQF	ZRH
506	285	1.88e-05	08:55	3	82min → CAI	VSL	ZRH
892	13842	1.49e-05	08:55	3	71min → GYP	GRN	ZRH
13959	822	1.08e-05	09:40	3	369min → LKT	TKV	ZRH
39422	638	1.00e-05	11:55	3	218min → VSL	ZRH	GVA
7501	7488	9.96e-07	09:00	3	1156min → YRP	YRP	ZRH
16797	2825	1.32e-07	11:45	3	969min → MQF	JMC	ZRH

Table 34: 04/11/2024 - Wave 4 Flights

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
22195	22195	0.994728	13:40	4	—	GVA	XNZ
559	559	0.991580	13:45	4	Prio Flight	ZRH	VSL
16651	16651	0.991341	12:55	4	—	GVA	JMC
1628	1628	0.981689	13:40	4	Prio Flight	ZRH	VXJ
9123	9123	0.975523	13:30	4	Prio Flight	ZRH	JFL
786	786	0.975327	14:10	4	Prio Flight	ZRH	ZTN
3182	3182	0.973644	16:40	4	—	MQT	GVA
5776	5776	0.972856	13:45	4	—	ZRH	PLW
7439	7439	0.966244	14:45	4	—	TKH	ZRH
21943	21943	0.957915	14:20	4	—	GVA	ZPD
10065	10065	0.949253	13:35	4	—	ZRH	DRL
11083	11083	0.939541	13:50	4	—	ZRH	ZLB
15118	15118	0.936737	13:50	4	Prio Flight	ZRH	BRP
14669	14669	0.931760	14:00	4	Prio Flight	ZRH	ZPR
13859	13859	0.924298	14:15	4	—	ZRH	GYP
4829	4829	0.922140	14:45	4	—	ZRH	NFX
10554	10554	0.913616	13:45	4	—	ZRH	GXH
7848	7848	0.865282	14:30	4	—	ZRH	GHF
1051	1051	0.856766	13:30	4	—	ZRH	FRL
5605	5605	0.856292	13:50	4	Prio Flight	ZRH	MCR
5605	9164	0.009013	13:50	4	55min → JFL	MCR	ZRH
3351	13184	0.007921	13:15	4	116min → NJK	FRB	ZRH
6854	822	0.007016	13:00	4	207min → LKT	WBT	ZRH
6854	7562	0.006815	13:00	4	140min → YRP	WBT	ZRH
6854	3118	0.006657	13:00	4	204min → MQT	WBT	ZRH
6854	13184	0.002877	13:00	4	160min → NJK	WBT	ZRH
4692	1588	1.08e-08	13:55	4	919min → VXJ	NFX	ZRH
16944	2007	2.24e-09	15:40	4	741min → LHT	JMC	ZRH
16944	6070	5.32e-10	15:40	4	709min → HFR	JMC	ZRH
14669	7488	1.60e-10	14:00	4	843min → YRP	ZPR	ZRH

Table 35: 04/11/2024 - Wave 5

Flight number *from*	**Flight number *to***	**Weight**	**Scheduled departure**	**Wave number**	**Check**	**Arrival Airport**	**Departure Airport**
9178	9178	0.936101	18:45	5	—	ZRH	JFL
14026	14026	0.675474	17:55	5	—	ZRH	TKV
5808	5808	0.537133	17:50	5	Prio Flight	PLW	ZRH
6180	6180	0.480122	18:05	5	Prio Flight	HFR	ZRH
14160	14160	0.429022	17:55	5	Prio Flight	ZRH	CMT
6924	6924	0.422781	18:25	5	Prio Flight	ZRH	WBT
613	613	0.418286	17:25	5	—	GVA	VSL
532	575	0.400573	17:10	5	Prio Flight	ZRH	VSL
3079	3079	0.363138	17:45	5	Prio Flight	ZRH	MQT
4731	4731	0.341944	17:40	5	—	ZRH	NFX
6180	24242	0.339139	18:05	5	Prio Flight	HFR	ZRH
613	638	0.331923	17:25	5	—	GVA	VSL
575	575	0.272122	17:25	5	Prio Flight	VSL	ZRH
39731	39731	0.261272	18:05	5	—	ZRH	GVA
5626	5626	0.254825	17:45	5	—	ZRH	MCR
3197	3197	0.237252	18:10	5	—	MQT	GVA
1402	1402	0.236267	18:35	5	—	VXJ	GVA
5271	5271	0.231615	17:30	5	—	ZRH	FZY
4731	4799	0.205462	17:40	5	—	ZRH	NFX
13201	13201	0.204020	18:00	5	—	ZRH	NJK
2893	2885	0.067248	18:05	5	Rotation	MQF	ZRH
3079	3118	0.066413	17:45	5	1314min → MQT	MQT	ZRH
6924	6912	0.063698	18:25	5	Rotation	WBT	ZRH
14160	285	0.007785	17:55	5	828min → CAI	CMT	ZRH
14026	2825	1.07e-06	17:55	5	706min → MQF	TKV	ZRH
13201	2825	1.95e-08	18:00	5	681min → MQF	NJK	ZRH
6180	21629	4.53e-09	18:05	5	1339min → MCR	ZRH	HFR
5808	5531	8.14e-10	17:50	5	76min → LHT	ZRH	PLW
5271	5179	6.85e-10	17:30	5	708min → FZY	FZY	ZRH
13201	4760	1.98e-13	18:00	5	648min → NFX	NJK	ZRH

Appendix F

The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a robust Multiple-Criteria Decision Making (MCDM) method that identifies the best option from a set of alternatives based on their distance from an ideal solution. This method is particularly effective in situations where decision-making involves multiple, often conflicting, criteria. In the context of the internship project, TOPSIS was employed to dynamically determine priority flights by evaluating each flight against four key performance factors: passenger connections, rotation buffers, curfew performance, and slot presence. The method has four steps which are detailed as follows.

The process begins with the construction of a decision matrix comprising the alternatives (flights) and the criteria (performance factors), followed by the normalization of this matrix to eliminate the effects of disparate units of measurement among criteria.

	Fleet number	Rotation buffer	Flight buffer	Pre Flight buffer	High connex	High VIP connex	Short connex	Rotation connex pax	# groups	Close to curfew	Slot
Weights	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11
Impact	-	-	-	-	+	+	+	+	+	+	+

Table 36: Initial weights per parameter and their respective impact

where *Short connex* are connecting passengers with a connecting time of less than 60 minutes, *Rotation buffer* is the ground time buffer for an aircraft in a day, *Pre Flight buffer* is the ground time buffer before the flight takes-off, *Flight buffer* is the ground time buffer after the flight lands, *High VIP connex* are connecting passengers which are HON Members, First-class passengers, Business class passengers, wheelchairs and unaccompanied minors. *Rotation connex pax* is the number of connecting passengers in the upcoming flights of the rotation.

Each criterion is then weighted according to its importance in the decision-making process, reflecting the priorities of airline operations management. Table 36 shows the initial weighting distribution used and the impact of each variable. Meaning if the parameter has a negative sign it means that the lower the value the more critical it is in the model (higher score). On the other hand if the impact is positive it means the higher the value is the more critical it is.

TOPSIS calculates the geometric distance of each alternative from the ideal (best possible) and negative-ideal (worst possible) solutions. The ideal solution maximizes the benefit criteria (e.g., high number of passenger connections, optimal rotation buffers) and minimizes the cost criteria (e.g., risk of curfew violation, lack of slot presence), while the negative-ideal solution does the opposite.

$$d_{iw} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{wj})^2} \tag{4}$$

where d_{iw} is the worst distance calculated of an ith row. t_{ij} is element value. t_{wj} is the ideal worst for that column

$$d_{ib} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{bj})^2} \tag{5}$$

where d_{ib} is the best distance calculated of an ith row. t_{bj} is the ideal best for that column

The priority of each flight is determined based on its relative closeness to the ideal solution, with those closest to the ideal solution ranked as higher priority. Hence Equation 6 is performed per row to determine the ranking score of each flight.

$$TOPSIS_{score} = \frac{d_{iw}}{(d_{ib} + d_{iw})} \tag{6}$$

In conclusion, the higher the score the more prioritized the flight.

Using the TOPSIS method lets us fairly rank flights by looking at different important factors all at once, making sure that the choice of which flights to prioritize is data-driven and matches what SWISS is aiming to achieve operationally. This strategy enables efficient modifications to the prioritization, allowing for real-time adjustments to the selection of priority flights as operational conditions evolve throughout the day.

9.1 Results

The sensitivity analysis conducted in this study aimed to determine the relative importance of various parameters influencing flight prioritization. Table 37 presents the weights assigned to each parameter following the analysis.

	Fleet number	Rotation buffer	Flight buffer	Pre Flight buffer	High connex	High VIP connex	Short connex	Rotation connex pax	# groups	Close to curfew	Slot
Weights	0.04	0.05	0.1	0.15	0.07	0.08	0.2	0.11	0.12	0.04	0.04

Table 37: Final weight per parameter

Verified with the Network Operations Center to ensure the weights are logical. For instance, the significance of the number of short connecting passengers is emphasized, particularly for SWISS operations. Conversely, the lower weight assigned to the fleet ground time buffer reflects its role primarily as an indicator of fleet planning density.

III. Literature Study

Literature Study

Master Thesis
February 2024 - December 2024

by

Constança Miranda de Andrade Veiga

Student number:	4786327
Thesis Duration:	February, 2024 - December, 2024
Faculty:	Faculty of Aerospace Engineering, Delft
Department:	Control & Operations

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
ATA	Actual Time of Arrival
ATD	Actual Time of Departure
ABM	Agent Based Modelling
ANN	Artifical Neural Network
ATFM	Air Traffic Flow Management
ATM	Air Traffic Management
BN	Bayesian Network
BiGRU	Bidirectional Gated Recurrent Unit
CCM	Convergent Cross Mapping
CDM	Collaborative Decision Making
CMI	Conditional Mutual Information
CN	Connection Network
CO2	Carbon dioxide
CPN	Cancellation Propagation Network
DCN	Delay Causality Network
DN	Delay Network
DC-SIC	Delay Causality Strong & Independent Causality
DPT-BN	Delay Propagation Tree - Bayesian network
4DTA	4D Trajectory Adjustments
EM	Expectation-Maximization algorithm
ESPL	Exponentially Truncated Shifted Power Law
ETFMS	Enhanced tactical flow management system
GANN	Graph Attention Neural Network
GBDT	Gradient Boosting Decision Tree
GCN	Graph Convolutional Neural network
GCKI-SPI	Granger Causality Kernel - SPI
GCT	Granger Causality Test
IID	Independent and Identically Distributed
LightGBM	Light Gradient Boosting Method
LM	Levenberg Marquart (optimization algorithm)
LSTM	Long short-Term Memory
MAE	Mean Average Error
MAPE	Mean Absolute Percentage Error
MCDM	Multiple-Criteria Decision Making
MGT	Minimum Ground Time
ML	Machine Learning
MLCN	Multi-Layer flight Connection Network
MLIL-NN	Multi-Level Input Layer Neural Network
MLP	Multi-Layer Perceptron

Abbreviation	Definition
MSTAGCN	Multiscale Spatial-Temporal Adaptive Graph Convolutional Neural Network
NewCat	Network-wide Congestion Assessment Tool
NN	Neural network
NOC	Network Operations Center
RMSE	Root Mean Square Error
RNGC	Refined Nonlinear Granger Causality
SHAP	SHapley Additive exPlanations
SPL	Shifted Power Law
SPI	Systematic Path Isolation
STA	Scheduled Time of Arrival
STD	Scheduled Time of Departure
SVR	Support Vector Machine
TOPSIS	Technique for Order Prefered by Similarity to Ideal Solution
WLR	Weighted Linear Regression

Symbols

Symbol	Definition	Unit
d_{iw}	Worst distance calculated for i^{th} row	-
t_{ij}	Element value for row i column j	-
t_{wj}	Ideal worst value for column j	-
d_{ib}	Best distance calculated for i^{th} row	-

1. Introduction

In the interconnected world of air transportation, delays are not merely isolated incidents; they propagate through the network, affecting a wide number of stakeholders, from passengers to airlines. In this thesis, the complex issue of delay propagation within passenger airline networks is explored, focusing on how and to where delays propagate and how severe they are to the airline's operational functioning.

Swiss International Air Lines, commonly referred to as SWISS, is the flag carrier of Switzerland, operating a wide network of domestic and international flights. Through collaboration with SWISS Airlines—a prominent hub-spoke carrier—it is possible to shed light on the intricacies of airline operations in the face of delays. Hub-spoke networks, characterized by a central 'hub' airport through which flights are routed, are especially vulnerable to the snowball effects of delays. In such systems, delays at the hub can lead to widespread impacts, affecting both direct and connecting flights. Additionally, the efficiency of operations at spoke airports can significantly influence the overall network's ability to mitigate delays, highlighting their crucial impact on delay propagation. This model contrasts with point-to-point systems, where flights operate directly between destinations, reducing but not eliminating the potential for delay propagation.

The interaction between various airline resources including passengers, aircraft, crew, and airport facilities is identified as a contributing factor to the complexity and unpredictability of delay propagation. Delays originating at upstream airports can initiate a chain reaction, leading to widespread disruptions across the network. The introduction of buffer times within airline schedules is a strategic approach to mitigate these issues. However, when these buffers are insufficient, a cascading effect ensues, affecting subsequent operations, such as missed connections, and adding to the challenge of delay management.

The thesis highlights the challenge of delay propagation within European airline operations, with an average of 1328 flights per day experiencing significant delays [1]. Despite considerable efforts to mitigate delays, research indicates that many flights are susceptible to delay propagation, with its impact varying widely across the network. With this, the aim of this literature study is to present the current problem, analyze the current state of the art and devise a research proposal. This sets the foundation for the Master thesis.

This literature study is structured in the following way. Chapter 2 details the current problem in the aviation industry concerning flight network operations. Chapter 3 presents the available data for this research, along with the tools that have been developed and are available for validation or as an additional resource. The current state-of-the-art with regards to the models and features used, and the uncertainties considered, is discussed in Chapter 4. Having defined the problem and what has currently been done, both the research gaps and the research question are proposed in Chapter 5. To finalize, Chapter 6 presents the organizational plan for the duration of the Master thesis.

2. Current Problem

The airline industry, characterized by its complex operational frameworks and extensive global networks, faces significant challenges in maintaining punctuality and efficiency. Among these challenges, flight delays emerge as a notable concern, impacting not only airlines and airports but also stakeholders and passengers. This challenge becomes more pronounced during peak operational periods, as evidenced by the substantial increase in delays during the summer months.

Late aircraft arrivals precipitate delayed turnaround processes, triggering a cascade of financial repercussions. These include passengers missing their connections, incurring curfew fees, and affecting customer satisfaction negatively. Additionally, the increased workload for airport staff and operational stakeholders further compounds the issue, emphasizing the need for effective management strategies to mitigate these delays.

Building on the preceding discussion, this chapter explores the complex, multifaceted nature of flight delays, focusing on the operational impacts, the prioritization of flights, and the cascading effects of these delays within the network.

2.1. Identifying High-Impact Flights

Certain flights are more critical than others due to reasons including the volume of connecting passengers, proximity to airport curfews, or their potential to significantly impact other critical flights within the network. Recognizing these flights and understanding their role within the network's operational integrity is crucial. By pinpointing these key flights, airspace users and stakeholders can collaborate more effectively, prioritizing resources to minimize disruptions and maintain operational continuity.

Flights with a high number of connecting passengers are particularly critical due to the domino effect that delays can have on passengers' subsequent travel plans. A delay in a single flight can lead to missed connections, affecting potentially hundreds of passengers. This not only disrupts the travel plans of individuals but also impacts the airline's reputation and operational costs, as accommodations and alternative arrangements must be made.

Airports often impose strict operating curfews to minimize noise pollution in surrounding areas during late-night and early-morning hours. Flights scheduled close to these curfew times carry a higher risk, as delays could result in significant fines for the airline, operational restrictions, or even the cancellation or rescheduling of flights to the following day. This inconveniences passengers and incurs additional costs for the airline. Prioritizing these flights for on-time departure requires careful planning and real-time adjustments to avoid the repercussions of curfew violations.

Moreover, some flights serve as critical nodes within the airline's network, linking major hubs or feeding passengers to long-haul flights. Delays in these flights can have cascading effects, disrupting the airline's operation. For example, a delay in an early-morning flight could affect the aircraft's utilization for the rest of the day, leading to subsequent delays across the network. Recognizing these critical flights and prioritizing resources, such as allocating backup aircraft or crew, is essential for maintaining punctual operations.

2.2. Challenges in Flight Prioritization

At SWISS Airlines, a critical aspect of managing flight delays involves the selection of priority (prio) flights by the Network Operations Center (NOC) at the Operations Center. Prio flights are identified daily to ensure the allocation of additional resources, aiming to minimize potential delays and operational disruptions. This allocation involves not only internal efforts from SWISS, such as those by the NOC and Flight Dispatch to secure improved slots but also extends to encompass comprehensive support from airport services. These include ground operations, turnaround teams, luggage handling, and catering, ensuring a cohesive approach to mitigating delays.

However, the current process for selecting prio flights is manual and occurs only once at the start of the operational day. This approach is rigid, and the selections often become outdated as the day progresses, failing to adapt to real-time operational changes. The lack of flexibility in identifying prio flights underscores the urgent need for a dynamic, responsive system capable of adjusting to the evolving operational landscape throughout the day.

Section 3.2 presents a novel, static prio flight tool that is able to compute the prio flights per wave based on a selected number of parameters extracted in real-time using a multi-criteria decision-making model.

2.3. The Operational Dilemma of Complex Airline

With a daily operational load of approximately 350 flights, catering to 45,000 passengers and spanning 150 destinations, SWISS Airlines faces a significant challenge manually monitoring and prioritizing flights due to its extensive network. Identifying flights that are critical requires navigating a complex array of factors, including aircraft rotation, crew availability, passenger transfers, the cascade effects of delays, and airport curfews. This complexity underscores the importance of identifying key routes and flights within the network to adeptly address operational obstacles.

Adding to this complexity, previous research has found that a delay in a single flight can, on average, affect four subsequent flights[1] [2]. This insight illustrates the risk of significant ripple effects throughout the network. When you think about the computational effort needed to handle

[1] "The diameter of an airport-flight network is four, which means that it will take a minimum of four flights to propagate a delay to the entire network of the airline."

connections across hundreds of flights, the task becomes overwhelming. Such a task suffers from the curse of dimensionality, where the complexity grows exponentially with each additional flight and variable introduced into the system, rendering the task overwhelmingly intricate for human operators to manage effectively manually.

Based on a simplified model of flight delay propagation, where one delayed flight on average impacts four other flights, the total number of flights potentially impacted over just three "layers" of propagation reaches 84. When considering the complexity of managing these interconnected delays, with at least seven variable factors (e.g. rotation parameters, crew schedules, passenger connections, delay propagation, curfews) for each flight, the computational scale of the challenge becomes apparent. In this scenario, operators would need to consider approximately 588 distinct variables across all affected flights.

This exemplifies the immense difficulty for human operators to quantify and mitigate the operational impacts manually, underscoring the necessity for sophisticated, automated systems to manage such complexity efficiently.

2.4. The Systemic Impact of Delay Propagation

"Delay propagation dynamics has also been studied as a process that heavily relies on the interconnectivity pattern of the air transportation network." [3]

The impact of delay propagation extends beyond individual airlines, affecting the entire aviation ecosystem. Delays not only diminish passenger satisfaction but also trigger a snowball effect on subsequent flights, airport operations, and the broader network. A single delayed flight can initiate a chain of reactionary delays across the network, exacerbating operational inefficiencies and financial losses. At airports, particularly those with strict curfews like Zurich, airlines face significant penalties when curfew breaches occur. Moreover, the ripple effects of delays extend to crew schedules and can lead to missed passenger connections, further deteriorating customer trust and satisfaction. To further emphasize the effects of delay dispersion, Figure 2.1 illustrates the interconnectedness of different delay propagation factors on a network of flights.

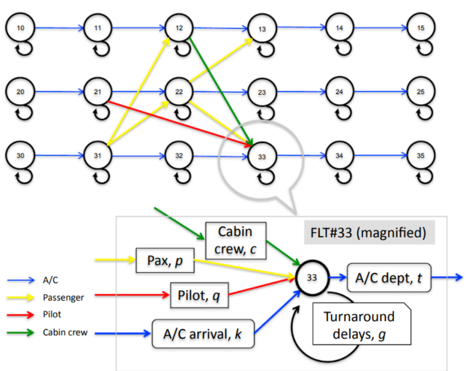


Figure 2.1: Delay propagation factors [4]

2.5. Conclusion

The challenges highlighted in this chapter underscore the complex and interconnected nature of the operational framework within the airline industry. Addressing these challenges necessitates an innovative, multi-faceted approach that leverages data-driven solutions, real-time data, and collaborative strategies to improve decision-making and operational efficiency. The development of solutions to mitigate flight delays and their cascading effects is crucial for the sustainability and success of the aviation sector.

3. Current Models and Data

Given the opportunity to do the Master thesis with SWISS Airlines. This brings value to the research given the detailed and abundant database the airline has, with data not publicly available, expertise from industry, and possible application to a real-case scenario. Meaning the research is not only relevant for academics but also for the airline industry.

This chapter is then structured as follows. The primary data available from SWISS airlines is discussed in Section 3.1. In addition, it is important to highlight two models that have been developed in the previous six months related to this topic. Section 3.2 showcases a static model used to determine the priority flights of the day. Section 3.3 presents a method of estimating turnaround time at different airports based on historical data.

3.1. Data Available

In this section, the comprehensive datasets obtained from SWISS are presented, pivotal for the model. These datasets are divided into four distinct categories, each serving a unique purpose in operational planning and analysis. An in-depth understanding of these datasets is crucial for developing strategies that address the number of challenges faced by the airline industry today.

3.1.1. Flight Operational Data

This dataset encompasses all pertinent information regarding flight operations, featuring real-time data on flights including boarding times, gate positions, and various time metrics crucial for day-to-day airline operations. Aircraft-specific information is visible to only airlines, be this reserve aircraft, maintenance events, and/or specific restrictions. In addition, crew information (rotation, duty time limits, possible extensions, qualifications) is confidential to the airline.

The dataset spans data from yesterday, today, and tomorrow, providing a near-term operational perspective that is essential for dynamic decision-making.

3.1.2. Connecting Passenger Database

The second dataset focuses on connecting passengers, a critical component of network airlines like SWISS. At SWISS on average only 41% of the passengers on long-haul flights are local passengers and 50% of all passengers on long-haul flights are connecting passengers originating from short-haul destinations and 9% are connecting passengers originally coming from intercontinental destinations.

The database details the number of connecting passengers per class, including specific counts for infants and children, and provides information on their onward connections. This dataset spans approximately 23 days, with data for 10 past days and 13 future days, offering insights into short-term passenger flow and connectivity within the network. This information is vital for optimizing passenger models and managing resources effectively to accommodate the intricate needs of connecting travelers which are key factors in the aviation industry. Most passenger information such as passenger real numbers, passenger connections, and passenger status (booking class, wheelchair, groups, minor not accompanied, etc.), are confidential to airlines. Hence, it is important to leverage this data to create a more accurate model.

3.1.3. Flight Information with Historical Data

Comprising both real-time and historical flight information, the third dataset offers a comprehensive view of flight operations over the past four years, albeit with slightly fewer features than the first dataset. This combination of current and historical data is invaluable for trend analysis, forecasting, and long-term strategic planning. By examining patterns and outcomes over a significant period, SWISS can identify potential areas for operational improvement and develop strategies to enhance efficiency and reliability.

3.1.4. Delay Codes Database

The fourth dataset is dedicated to understanding the specifics of flight delays, containing detailed delay codes for each flight. These codes elucidate the reasons behind delays, quantifying the impact in minutes and allowing for a granular analysis of delay causes. This dataset is instrumental in identifying operational bottlenecks, improving turnaround times, and implementing measures to prevent future delays. By analyzing the reasons for delays, the model can target specific operational challenges, enhancing punctuality and reducing the cascading effects of delays on the network.

The datasets from SWISS Airlines serve as a foundational element for the research, allowing for the creation of a model that offers predictions as accurate and realistic as possible.

3.2. Prio dePrio Flight Selection Model

The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a robust Multiple-Criteria Decision Making (MCDM) method that identifies the best option from a set of alternatives based on their distance from an ideal solution. This method is particularly effective in situations where decision-making involves multiple, often conflicting, criteria. In the context of the internship project, TOPSIS was employed to dynamically determine priority flights by evaluating each flight against four key performance factors: passenger connections, rotation buffers, curfew performance, and slot presence. The method has four steps which are detailed as follows.

The process begins with the construction of a decision matrix comprising the alternatives (flights) and the criteria (performance factors), followed by the normalization of this matrix to eliminate the effects of disparate units of measurement among criteria.

	Fleet number	Rotation buffer	Flight buffer	Pre Flight buffer	High connex	High VIP connex	Short connex	Rotation connex pax	# groups	Close to curfew	Slot
Weights	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11
Impact	-	-	-	-	+	+	+	+	+	+	+

Table 3.1: Initial weights per parameter and their respective impact

where *Short connex* are connecting passengers with a connecting time of less than 60 minutes, *Rotation buffer* is the ground time buffer for an aircraft in a day, *Pre Flight buffer* is the ground time buffer before the flight takes-off, *Flight buffer* is the ground time buffer after the flight lands, *High VIP connex* are connecting passengers which are HON Members, First-class passengers, Business class passengers, wheelchairs and unaccompanied minors. *Rotation connex pax* is the number of connecting passengers in the upcoming flights of the rotation.

Each criterion is then weighted according to its importance in the decision-making process, reflecting the priorities of airline operations management. Table 3.1 shows the initial weighting distribution used and the impact of each variable. Meaning if the parameter has a negative sign it means that the lower the value the more critical it is in the model (higher score). On the other hand, if the impact is positive it means the higher the value is the more critical it is.

TOPSIS calculates the geometric distance of each alternative from the ideal (best possible) and negative-ideal (worst possible) solutions. The ideal solution maximizes the benefit criteria (e.g., high number of passenger connections, optimal rotation buffers) and minimizes the cost criteria (e.g., risk of curfew violation, lack of slot presence), while the negative-ideal solution does the opposite.

$$d_{iw} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{wj})^2} \tag{3.1}$$

where d_{iw} is the worst distance calculated of an ith row. t_{ij} is element value. t_{wj} is the ideal worst for that column

$$d_{ib} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{bj})^2} \tag{3.2}$$

where d_{ib} is the best distance calculated of an ith row. t_{bj} is the ideal best for that column

The priority of each flight is determined based on its relative closeness to the ideal solution, with those closest to the ideal solution ranked as a higher priority. Hence Equation 3.3 is performed per row to determine the ranking score of each flight.

$$TOPSIS_{score} = \frac{d_{iw}}{(d_{ib} + d_{iw})} \tag{3.3}$$

In conclusion, the higher the score the more prioritized is the flight.

Using the TOPSIS method lets us fairly rank flights by looking at different important factors all at once, making sure that the choice of which flights to prioritize is data-driven and matches what SWISS is aiming to achieve operationally. This strategy enables efficient modifications to the prioritization, allowing for real-time adjustments to the selection of priority flights as operational conditions evolve throughout the day.

3.2.1. Results

The sensitivity analysis conducted in this study aimed to determine the relative importance of various parameters influencing flight prioritization. Table 3.2 presents the weights assigned to each parameter following the analysis.

	Fleet number	Rotation buffer	Flight buffer	Pre Flight buffer	High connex	High VIP connex	Short connex	Rotation connex pax	# groups	Close to curfew	Slot
Weights	0.04	0.05	0.1	0.15	0.07	0.08	0.2	0.11	0.12	0.04	0.04

Table 3.2: Final weight per parameter

Verified with the Network Operations Center to ensure the weights are logical. For instance, the significance of the number of short connecting passengers is emphasized, particularly for SWISS operations. Conversely, the lower weight assigned to the fleet ground time buffer reflects its role primarily as an indicator of fleet planning density.

3.3. Groundtime Estimation Model

In this section, I will delve into the methodology employed to estimate the average ground time for aircraft turnarounds at outstations. This analysis is based on an extensive dataset encompassing a full year's worth of historical data, with the primary goal of estimating ground time based on two significant factors: the departure airport and the aircraft type.

The distinction of aircraft type in this analysis is particularly important as it accounts for the varying sizes of airplanes, which inherently affects the range of passenger numbers. This approach marks a pivot from previous methodologies that relied on passenger numbers as a predictor for ground time. My analysis, supported by the calculation of the R^2 value (a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model), indicated that the correlation between passenger numbers and ground time was relatively weak. This insight led to the adoption of aircraft type as a more reliable indicator, providing a nuanced understanding of how different airplane capacities influence turnaround times.

Recognizing the potential distortion caused by extreme outliers on the overall analysis, a strategic approach to outlier removal was implemented. Specifically, flights exhibiting extreme ground times—defined as delays exceeding 90 minutes or early arrivals/departures by more than 30 minutes—were excluded from the dataset. This filtering process aims to mitigate the influence of anomalous data points, thereby enhancing the accuracy and reliability of the ground time estimates.

To address the challenge of non-constant variance (heteroscedasticity) within the data—a common occurrence in real-world datasets where the variability of a variable is unequal across the range of values—a weighted linear regression (WLR) model was utilized. WLR adjusts the influence of individual data points based on their variance, assigning more weight to points with lower variance. This method is particularly advantageous in this context as it compensates for the heterogeneity in ground times, ensuring that the model remains robust and sensitive to the nuances in the data. By optimizing the weights applied to each observation, the WLR model facilitates a more precise estimation of ground times, reflective of the underlying data distribution.

To validate the effectiveness of this model, I closely examined the p-value and R^2 value of the regression results. The p-value offered insight into the statistical significance of the model's findings, ensuring that the observed relationships were not due to chance. Meanwhile, the R^2 value, a measure of the model's explanatory power, indicated how well the variations in ground time could be accounted for by the departure airport and aircraft type. Together, these statistical measures served as vital benchmarks for assessing the quality and reliability of the ground time estimation, guiding the refinement of the model to better serve operational planning and efficiency improvement efforts at SWISS Airlines.

The estimated average ground time is used as a baseline for determining the minimum ground time required in the turnaround buffer calculations. By adopting this model-driven approach for the minimum ground time, I ensure that the scheduling practices are grounded in realistic and achievable benchmarks, thereby enhancing the overall reliability and efficiency of airline operations. This strategy allows for the inclusion of a calculated buffer, optimizing the use of time and resources, and reducing the likelihood of delays.

4. Current State-of-the-Art

In this chapter, the current state-of-the-art is examined. In particular how models predict initial flight delays that affect not only the next flight but also have a lesser but significant impact on later flights. Various methodologies, including statistical models, operations research, and machine learning algorithms, are reviewed for their contributions to improving air traffic management and decision-making processes regarding flight schedules and airport operations. The literature points out how crucial it is to predict problems early to avoid making decisions that are not the most appropriate or efficient. In particular, the value of making predictions before the day of operation, which helps airlines plan better and make adjustments to flights ahead of time. Additionally, this chapter delves into several analytical approaches, such as mathematical and data-driven models—including Bayesian networks and machine learning—to enhance delay prediction capabilities. Recent advancements have leveraged historical data and innovative machine learning techniques, improving the precision of delay forecasts and enriching strategies for smart flight scheduling.

This chapter is outlined in the following way. First, the models that have been used in the topics of delay propagation, estimating delays, and modeling flight networks are explained. Section 4.1 presents the mathematical models, followed by statistical models in Section 4.2. Right after, data-driven models are in Section 4.3, and, lastly, machine learning models are in Section 4.4, which includes supervised learning and deep learning. Furthermore, at the end of each section, the models are summarized and a score is given, either red, yellow, or green.

- Green: model should *strongly* be considered for thesis
- Yellow: model should *perhaps* be considered for thesis
- Red: model is *not appropriate* and hence should not be considered for thesis

To conclude the Chapter, Section 4.5 summarizes the features used in past research, Section 4.6 illustrates the type of data used in studies, Section 4.7 describes the uncertainties frequently used, and Section 4.8 details the usual computational times needed to compute. Section 4.9 presents several Mean Absolute Errors for different models.

4.1. Mathematical models

Initial research in the 1990s introduced concepts such as delay multipliers and delay trees to understand the scale and impact of delay propagation in air transport networks. These studies aimed at measuring the overall effect of initial delays on the network, suggesting that delays in one part of the system could propagate and affect other parts in a cascading manner.

Beatty et al. developed the concept of a delay multiplier to estimate the true system impact of a delayed flight. This concept considers the cumulative effect of delays across all connected flights, showing that large initial delays early in the day are particularly disruptive and that the delay multiplier grows non-linearly with the size of the initial delay [5].

Ten years later, in 2009, research utilized a Monte Carlo Simulation approach, employing statistical distributions for each ground process duration and starting times [6]. This method is used for its ability to provide statistically significant results across various delay categories. The advantage of this approach is its statistical rigor and adaptability to different scenarios. However, the reliance on historical data and predefined constraints may limit the model's ability to predict unforeseen delays.

Following studies by Lovell et al. [7] and Churchill et al. [8] refined methodologies for differentiating flight data to separate propagated and queuing effects, focusing on subtracting upstream delays from empirical records and adjusting downstream schedules accordingly. These models contributed to understanding the separation of new delay from propagated delay, offering a more detailed view of delay dynamics.

By using two models, a microscopic delay propagation model and a macroscopic delay propagation model, a comprehensive analysis of both spatial and temporal mechanisms for transmitting delays across the air traffic system can be 'made' [7]. The microscopic model tracks individual aircraft through their daily operations, decomposing flight delays into propagated and new categories. This model provides detailed insights but requires extensive data and computational power. The macroscopic model, on the other hand, analyzes aggregate measures of airport performance to understand the temporal evolution of delays. This approach allows for a broader understanding of delay propagation patterns over time with less detailed data requirements.

The findings reveal that propagated delays account for a significant portion of total flight delays, with implications varying across different airports. The microscopic analysis indicates that propagated delays constitute between 20% and 30% of total delays [7]. The macroscopic analysis demonstrates that the impact of earlier delays on later delays varies throughout the day and differs across airports, with some periods showing a higher marginal cost of delay. These results suggest that strategic planning efforts must consider the specific characteristics of delay propagation at individual airports and the overall network.

More recently, the use of Bayesian networks, a type of probabilistic graphical model, has resurfaced. Recent studies have shown the effectiveness of using Bayesian Networks (BN) to model delay propagation, taking into account non-IID (independent and identically distributed) flight delay distributions and the stochastic occurrence of flight delays. These models have been applied at both individual route levels and airport levels to understand the causal relationships between different factors contributing to flight delays [4]. This non-IID-based model aims to better represent the complexity of airline operations by accounting for various resource connections.

4.1.1. Agent-Based Models

Agent-based modeling (ABM) is a simulation modeling technique that uses autonomous agents to explore the actions and interactions of individuals within a system [9]. These agents can represent individuals, groups, or entities with the ability to make decisions and interact with each other in a virtual environment. It is widely used to model complex systems dynamically and to analyze the potential outcomes of different scenarios.

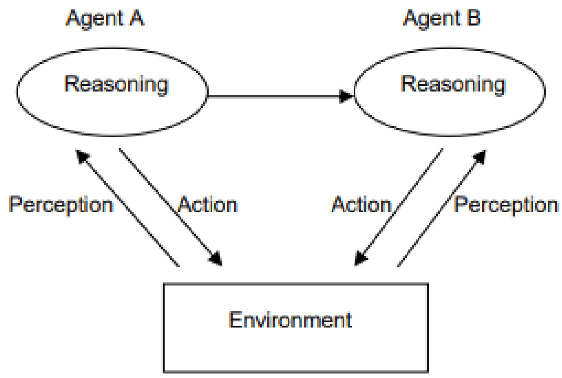


Figure 4.1: Basic principle of Agent-based modeling [10]

Figure 4.1 presents a basic model illustrating the interaction between two agents and their environment, offering the most fundamental depiction of how agent-based modeling (ABM) is applied. The agents gather data from the environment, which shapes their understanding and perception of the environmental condition, and then, based on this, the agents decide on what actions to take.

At least six studies have applied agent-based modeling to delay propagation in a flight network. The following details these papers, highlighting how they use the model, its strengths and weaknesses, and improvement possibilities for the future.

- **Campanelli et al. (2015) [11]**:
  - *Use*: Centers on aircraft as fundamental units incorporating mechanisms to simulate aircraft rotations, passenger connections, slot reallocation, and swapping.
  - *Strength*: Utilizes empirical data to recreate flight schedules, airport capacities, & passenger connectivity patterns with primary delays. Beneficial for its realism & potential to closely replicate observed delay patterns.
  - *Weakness*: Over-reliance on accurate historical data and predefined parameters may limit flexibility to account for unforeseen operational challenges.
  - *Improvement*: Incorporation of empirical data for airport capacities and passenger connectivity.
- **Campanelli et al. (2014) [12]**:
  - *Use*: Simulates delay propagation where airports are nodes and direct flights create links between them.
  - *Strength*: Based on Complex Systems theory, suitable for analyzing systems with a large number of interacting components.

  - *Weakness & Improvement*: Simplifications and assumptions such as the inability to recover delay en-route and the fixed re-scheduling threshold could be addressed in future work.

- **Fleurquin et al. (2015) [13]**:

  - *Use*: NewCat (Network-wide Congestion Assessment Tool) simulates the propagation of delays across the air transportation system.
  - *Strength*: Ability to incorporate real flight schedule data, making predictions grounded in empirical observations. Predictive power and adaptability to various scenarios.
  - *Improvement*: Generalize the problem of system resilience to perturbations, introducing metrics to quantify the impact of such perturbations and the robustness of the air-traffic system.

- **Guleria et al. (2019) [14]**:

  - *Use*: Multi-agent based model where each flight is considered an agent that interacts within an airport environment.
  - *Strength*: Allows for an abstract yet detailed analysis of how individual components (flights) interact within the dynamic environment of airports. Flexibility & ability to model emergent behaviors from simple agent rules.
  - *Weakness*: Potential challenges in scaling the model for larger networks or incorporating other significant factors like crew scheduling and passenger flow.

- **Wang et al. (2021) [15]**:

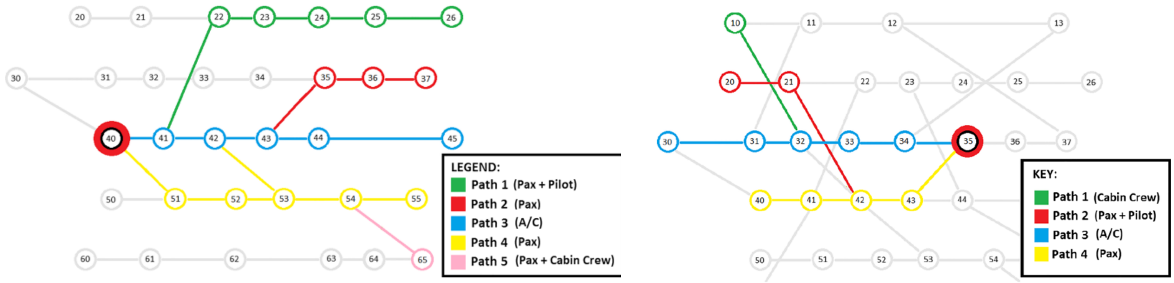
  - *Use*: Predicts individual flight delays across the entire air traffic network.
  - *Strength*: Offers high precision in delay prediction by integrating real-time information and detailed parameter models. Capable of simulating the operation of numerous flights and airports efficiently.
  - *Weakness*: Performance of the model is closely tied to the accuracy of the underlying parameter estimation models.

- **Gurtner et al. (2021) [16]**:

  - *Use*: Models individual flights & passengers within a realistic cost framework for airlines, capturing network-wide effects across the EU air transport system.
  - *Strength*: Ability to simulate the complex interactions between different stakeholders (airlines, passengers, etc.) and the detailed output it provides.

Additionally, it is implied that further research could enhance the understanding of network-wide impacts and support the development of more effective, equitable delay management solutions across the air transport system. [16]

4.1.2. Delay Propagation Trees



(a) Propagation paths

(b) Contributory paths

Figure 4.2: Front and Backwards Propagation [4]

The surge of these delay propagation trees stems from the need for models that incorporate multiple connecting sources and passenger connections without relying on IID-based assumptions, pointing out that such simplifications could lead to overestimation of buffer times and increased scheduling costs for airlines [4]. Nonetheless, the application of this model with a larger network is still to be done. Figure 4.2 shows clearly how a propagation path in a tree looks like, and the degree to which it can spread out based on the variables involved (passenger, pilots, aircraft, and cabin crew).

Furthermore, the Delay Propagation Tree - Bayesian Network (DPT-BN) model is introduced to address the shortcomings of existing approaches by incorporating stochastic variables into the delay propagation analysis. This model enables the detailed examination of how delays propagate through an airline's network by considering the non-independent and identically distributed (non-IID) nature of delay profiles across flights. The DPT-BN model's application revealed that flight delays have heterogeneous propagation effects, and delay profiles are non-IID in nature [17].

The DPT-BN model showcased the potential of using a probability-based model to identify 'weak links' within a network. These 'weak links' are defined as turnaround operations that fail to improve with additional buffer times, and airports that see considerable variation in ground operational efficiency over the course of a day due to unforeseen disruptions.

Model	Advantages	Disadvantages	Overall conclusion	Score
Delay Multipliers	Early understanding of the scale and impact of delay propagation.	Limited by the lack of detailed operational data and computational tools available at the time.	Moderate & Laid the groundwork for more complex delay propagation modeling efforts.	
Monte Carlo Simulation	Provides statistically significant results across various delay categories with statistical rigor.	Reliance on historical data and predefined constraints may limit predictive capabilities for unforeseen delays.	Offers a robust statistical approach to modeling delay propagation with adaptability to different scenarios.	
Microscopic and Macroscopic Models	Detailed insights into individual and aggregate delay propagation patterns.	Microscopic model requires extensive data & computational resources; macroscopic model may lack detailed operational insights.	Provides a comprehensive analysis of delay propagation, emphasizing the importance of both detailed and broad perspectives.	
Bayesian Networks	Models complex causal relationships and the non-IID nature of flight delays effectively.	May be complex to implement and require extensive data for accurate modeling.	Demonstrates the effectiveness of probabilistic models in understanding delay propagation at individual and airport levels.	
Agent Based Models	Dynamically models complex systems and explores different scenarios through simulation.	Can be computationally intensive and require detailed modeling of agent behaviors.	Offers valuable insights into the actions and interactions within the air transport system, with flexibility in scenario analysis.	
Delay Propagation Trees - Bayesian	Identifies 'weak links' in the network and models stochastic variables in delay propagation analysis.	Application to larger networks and integration with real-time operational data could be challenging.	Highlights the potential of probabilistic models in analyzing delay propagation and optimizing network performance.	

Table 4.1: Summary of Mathematical Methods to Model Delay Propagation

4.2. Statistical

Previous studies in flight delay prediction often relied on statistical analysis and probabilistic models to assess and forecast delays. These approaches typically involved analyzing historical delay data to identify patterns and probabilistic factors contributing to delays. For example, regression analysis and time-series forecasting have been common techniques for predicting flight delays based on historical trends.

4.2.1. Pure statistical

To estimate flight departure delays, a non-parametric method combined with a mixture distribution model can be used. Non-parametric methods have the ability to capture daily and seasonal trends, and a mixture model can estimate residual errors [18]. This combination addresses the complex nature of flight delays, which are influenced by numerous factors and exhibit non-linear trends and patterns that vary over time. This model aims to provide not just point estimates but estimates of the entire distribution. The model is structured as follows [18];

- Non-parametric method: smoothing spline model, allowing to treat time as a continuous factor.
- Mixture model to estimate residuals.

- Expectation-Maximization (EM) algorithm to estimate mixture components.
- Genetic algorithm used to overcome the challenges of local optima associated with the mixture distribution.

It is important to note that Markov chain Monte Carlo can be used alternatively to perform the optimization task, as stated by Tu et al, as it resembles the model just discussed while offering a broader parameter space than methods that are purely deterministic [18]. Another way of enhancing the model would be to implement dynamic updating techniques, which allow for the integration of new data as it becomes available, thereby aiming to continuously maintain or improve its predictive accuracy over time.

Churchill et al. points out the lack of in-depth research on the temporal evolution of delay propagation at individual airports and aims to fill this gap by comparing airports regarding their delay propagation characteristics [8]. A linear regression model is used to capture the relationship between early and later delays in the day, providing insights into the critical periods for delay evolution at different airports. This approach is defended for its simplicity and ability to produce results that can be continually related to causal factors. While the simplicity of the model facilitates easy application and interpretation, it may not capture all complexities of delay propagation, such as specific operational details or unpredictable external factors [8]. For this reason, the exploration of more complex models and/or incorporating additional data sources is important to better understand delay propagation mechanisms.

Later, researchers shifted towards more analytical models. For instance, an analytical model alongside a joint discrete-continuous econometric model can be used to quantify propagated and newly formed delays across flight nodes (departure or arrival points). Allowing for a more detailed analysis of delay propagation patterns and mitigation strategies. This approach is notable for incorporating both flight and ground buffers as variables, aiming to differentiate between propagated and newly formed delays and to understand how these delays are absorbed. These models provide robust results across different scenarios of buffer absorption and buffer values [19]. An econometric analysis can provide the following; an understanding of the conditions under which propagated delays are likely to occur, how the size of total propagated delay evolves downstream, and the influencing factors.

To uncover universal statistical patterns of delay propagation, two dynamic models of delay propagation were developed to classify airlines based on their delay propagation patterns: one exhibiting a shifted power law (SPL) distribution and another showing an exponentially truncated shifted power law (ETSPL) distribution. The advantage of these models lies in their ability to derive universal metrics from big data, offering insights into airlines' operational efficiency in delay mitigation [20]. Nevertheless, investigating the applicability of the models to other airlines and over different time frames is needed for full validation. Developing more refined models or incorporating other predictive techniques might enhance the accuracy and applicability of the findings [20].

4.2.2. Granger causality

Granger causality is a statistical concept used to determine whether one time series can predict the future values of another time series [21]. Its application in the past couple of years in the aviation world has increased substantially. Granger causality provides an advanced method for examining propagation patterns and causality that goes beyond basic correlation to investigate the dynamics of delay spread. Two key concepts form the basis of the causality relationship [22]:

- The cause happens prior to its effect.
- The cause has unique information about the future values of its effect.

In 2018, a study used the Granger causality test to create a Delay Causality Network (DCN) [23]. This method helped examine how delays spread among airports by identifying cause-and-effect relationships in time series data. Granger causality was chosen because it can show how one airport's delays can predict another's, mapping out the flow of delays. The study pointed out that further research could look into more detailed DCNs that consider the strength of these relationships, giving a clearer picture of how delays move through airport networks.

In 2022, Refined Nonlinear Granger Causality (RNGC) method was introduced, which utilizes a low-dimensional approximation of conditional mutual information (CMI) to improve the estimation of causal relationships in high-dimensional, non-linear time series data [24]. This methodology is particularly suited to analyze the complex interactions that lead to delay propagation in air transport systems. Using RNGC, the authors construct delay propagation networks that map out the causal relationships between airports, as visualized in Figure 4.3. These networks serve as a basis for analyzing how delays spread across the air transport system. The RNGC method overcomes the dimensionality challenge inherent in traditional methods, offering a more accurate detection of causality within multivariate time series. In other words, the built-in robustness in handling the non-linearity and high-dimensionality of delay data is a major advantage of RNGC.

For future reference, it is important to note that the RNGC method is combined with complex network theory to analyze the global structure of delay propagation networks [24].

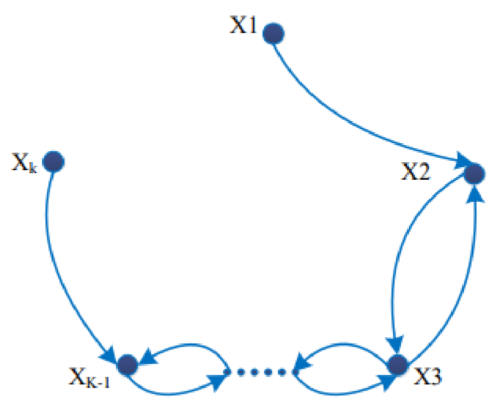


Figure 4.3: Causality Network [24]

A year later, Delay Propagation Networks (DPNs) and Cancellation Propagation Networks (CPNs) for each airline during specific time events are constructed, employing Granger Causality to detect patterns of delay causality between airports [25]. This method is advantageous because it allows for the identification of causal relationships and patterns that are not immediately apparent, enabling a deeper understanding of delay and cancellation dynamics. However, one can infer that the complexity of establishing causality and the requirement for large datasets to produce reliable results may be potential limitations.

Recently, Systematic Path Isolation (SPI) has been highlighted as an effective tool for selecting causal pathways in time-series data [26]. More specifically, a Systematic Path Isolation (SPI)-based causal inference method that incorporates the Granger test and the Kernel-based test, catering to both linear and non-linear relationships within the aviation system. This method is favored for its ability to handle complex data without explicit feature extraction and its scalability and effectiveness in selecting causal pathways in time-series data, making it well-suited for the air transportation network.

The GCKCI-SPI selection algorithm is utilized to determine the relationships among delay causalities [26]. The complex network theory (from [27]) is employed to construct delay propagation networks based on these causal relationships. This approach ensures that the analysis captures a comprehensive picture of delay causality, accounting for the multitude of factors and interactions that contribute to delay propagation across the network.

4.2.3. Propagation Trees

Propagation trees are useful for tracking individual flight delays through the network and studying the impact of airline schedules on delay propagation. Utilized for tracking the spread of delays from a single flight through the network. Key findings include identifying primary delays' early reduction as crucial for controlling delay propagation and determining key buffers limiting delay spread. Studies utilizing propagation trees to track individual flight delays through the network offer insights into how airline schedules impact delay propagation.

AhmadBeygi et al. investigated delay propagation using a tree structure approach for airlines' networks, analyzing how delays spread through the network over a day [28]. Their focus was on understanding the root causes of delays and the mechanisms through which delays were transmitted across flights.

The Delay propagation tree concept is a tool effective for estimating how delays spread through a network due to interconnected resources like aircraft and crew. This model is static, however, and does not account for the stochastic nature of delay propagation, such as delay amplification or absorption, which are critical for understanding and improving airline schedule robustness against delays. Moreover, although simple and easy to construct, it could lead to an overestimation of propagated delays [17].

Model	Advantages	Disadvantages	Overall conclusion	Score
Non-parametric and Mixture Model	Captures daily and seasonal trends; estimates entire distribution.	May not fully capture complex operational details or unpredictable external factors.	Effective for estimating flight departure delays, offering flexibility and capturing complex patterns.	
Linear Regression Model	Simple, facilitating easy application and interpretation.	May not capture all complexities of delay propagation.	Provides insights into critical periods for delay evolution but may require more complex models for comprehensive analysis.	
Analytical and Econometric Model	Differentiates between propagated and newly formed delays; provides robust results across scenarios.	Requires detailed analysis and data for buffer absorption and values.	Offers detailed insights into delay propagation patterns and mitigation strategies.	
Dynamic Models (SPL and ETSPL)	Derives universal metrics from big data; offers insights into operational efficiency in delay mitigation.	Full validation requires applicability testing across various airlines and time frames.	Useful for uncovering statistical patterns of delay propagation and comparing airline efficiency.	
Granger Causality	Examines propagation patterns and causality beyond basic correlation.	Complexity in establishing causality and large data requirement.	Effective for analyzing how delays spread among airports and producing Delay Causality Networks.	
Refined Non-Linear Granger Causality	Handles non-linearity and high-dimensionality; improves causal relationship estimation.	High computational resources & sophisticated data processing.	Enhances accuracy in detecting causality within multivariate time series and analyzing complex interactions.	
Propagation Trees	Tracks the spread of delays through networks; identifies key buffers limiting delay spread.	Static model; might overestimate propagated delays.	Useful for tracking individual flight delays but requires dynamic updating to account for stochastic nature of delay propagation.	

Table 4.2: Summary of Statistical Methods

4.3. Data-driven models

Past research extensively addressed flight delays, employing data-driven models to estimate flight delay distributions at non-European airports, analyzing historical flight data to determine delay statistics for major US airports, and employing statistical models to estimate flight departure delay distributions and seasonal trends at specific airports. This section focuses mainly on network models. Neural networks are also data-driven models but given that most used in research are multi-neural networks hence they are described in the Deep learning section.

4.3.1. Network models

In 2021, Q. Cai et al. pinpointed one significant gap in existing studies is the limited use of spatial-temporal properties in analyzing delay propagation [29]. Most prior research does not leverage dynamic network modeling to explore how delays propagate over time and across different locations within the air traffic network. This gap means that traditional studies might not offer a comprehensive view of delay propagation dynamics, overlooking the details of how delays evolve and affect various parts of the network differently [29].

A spatial-temporal network model can be used to analyze delay propagation dynamics, with airports as nodes and dynamic edges representing flight connections and delays. This model aims to capture the complex dynamics of how delays spread through the network over time and space and provide insights into delay propagation magnitude, severity, and speed [29]. This approach stands out for its ability to offer a fine-grained, dynamic perspective on delay propagation.

Expanding the use of the network-based approach for analyzing delay propagation in real time could assist in strategic air traffic management (ATM) and collaborative decision-making (CDM). Moreover, by calculating estimated temporal delays for specific airports, stakeholders in aviation can implement CDM strategies more effectively to lessen the effects of delay propagation, particularly in major hub airports.

Numerous previous studies have primarily focused on delay causality within strongly connected systems using methods like the Granger causality test (GCT), which, however, show limitations in complex, weakly connected systems like air traffic networks, as discussed in subsection 4.2.2. There is also an interest in addressing challenges posed by the growing volume of data in aeronautics by integrating DC-non-SIC (Delay Causality non-Strong Independent Causality) with new data science methodologies [27]. Hence, there is a need to explore delay propagation without these strong assumptions, using methods that can account for weak causal relationships and external interferences.

One way of solving the above issue is to employ a Convergent Cross Mapping (CCM) and complex network theory to analyze delay propagation [27]. CCM is chosen for its ability to identify and measure causality in complex systems from time series data, regardless of the system's connectivity strength. The framework involves using CCM to identify delay causality without relying on strong causality assumptions and employing complex network theory to analyze these causal relationships. Through this approach, able to address gaps in the current understanding of delay propagation, particularly in regional systems characterized by dense interconnections and shared resources [27].

Moreover, to enable a more detailed comprehension of the dynamics of congestion as they evolve, a novel multistage and multi-event model for analyzing and predicting congestion propagation by dividing the process into different stages and considering the congestion connection/degree among flights can be used [30]. The model differentiates between heterogeneous and homogeneous network models based on the distribution of flight connections.

- **Strength:** its ability to describe congestion propagation and its causes at different operational stages and the simplified version of the model aims for quick short-term predictions, beneficial for flight dispatchers
- **Weakness:** the initial model's complexity and time-consuming nature for prediction, which the simplified models seek to address by reducing accuracy slightly for significant gains in prediction speed

Based on the strengths and weaknesses, future research directions include considering the complexity and probability of event coupling in congestion propagation models and exploring further simplifications or enhancements to improve prediction accuracy and speed [30]. The ultimate goal is to better understand congestion propagation mechanisms and develop models that can be quickly and accurately used for strategic and tactical decision-making.

Network science techniques, including centrality measures and percolation analysis, are applied to models to pinpoint central or influential nodes and connections. Chauhan et al. used this idea to employ a network science approach to address flight delay propagation within an airline network, focusing on identifying the most disruptive elements like airports, flights, and connections that potentially or historically cause significant disruptions [2]. The approach models the airline's data through connection networks (CNs) that represent flight schedules, delay networks (DNs) that focus on historical operational delays, and multilayer flight connection networks (MLCNs) that analyze different types of connections (crew, tail, and passenger) in detail. It is worth mentioning that the methodology differentiates between potential disruptive elements derived from schedules (CNs) and actual disruptive elements based on historical data (DNs), providing a comprehensive view of both theoretical and practical delay propagation aspects.

A study on delay propagation in air traffic networks employed a sophisticated methodology combining delay time series analysis, transfer entropy, and complex network theory to uncover spatial and temporal patterns of delay spread among airports. Table 4.3 presents the methodology used in chronological order from left to right.

Methodology			
Delay time series analysis	Transfer entropy	Complex network theory	Logistic regression model
created for each airport to capture operational conditions	used to measure the strength of delay propagation between airports, establishing causality relationships	construction of delay propagation networks, with airports as nodes and causality relationships as edges	applied to predict future delay propagation based on identified network features

Table 4.3: Combination of models used to forecast delay propagation [31]

The delay propagation network facilitated the examination of network properties through techniques such as k-core decomposition, analysis of network topology features, and community detection. This analysis uncovered that air traffic networks manifest a multi-layered structure where a core layer of highly interconnected airports plays a pivotal role in the propagation of delays across the system [31].

Model	Advantages	Disadvantages	Overall conclusion	Score
Network models	Enable dynamic modeling of delay propagation over time and space, offering insights into propagation magnitude, severity, and speed.	Limited use in existing studies, especially in leveraging spatial-temporal properties.	Effective for providing a comprehensive view of delay propagation dynamics, essential for strategic ATM and collaborative decision-making.	
Spatial Temporal Network Models	Fine-grained, dynamic perspective on delay propagation. Can analyze delay propagation dynamics, with airports as nodes and dynamic edges representing connections and delays.	Requires complex data analysis and modeling to accurately capture delay dynamics.	Offers significant insights into how delays spread across the network, which is vital for implementing effective collaborative decision-making strategies in aviation.	
CCM & Complex Network Theory	Ability to identify and measure causality in complex systems without strong causality assumptions. Applicable in regional systems characterized by dense interconnections.	May involve sophisticated data processing and analysis. Complexity in interpreting network causal relationships.	Addresses gaps in understanding delay propagation, especially useful in analyzing weak causal relationships and external interferences in complex, weakly connected systems like air traffic networks.	
Multistage & Multi-event Models	Describes congestion propagation and its causes at different operational stages. Simplified version of the model aims for quick short-term predictions, beneficial for flight dispatchers.	Initial model complexity and time-consuming nature for prediction. Simplified models reduce accuracy for gains in prediction speed.	Highlights the importance of considering event coupling in congestion propagation models and suggests future research towards simplifying or enhancing models for better accuracy and speed.	

Table 4.4: Summary of Network Models for Analyzing Delay Propagation in Air Traffic Networks

4.4. Machine learning

The application of machine learning (ML) techniques to flight delay prediction represents a significant advancement in the field. ML models, both in regression and classification contexts, have been developed to predict flight delays based on a wide array of input features, including operational, environmental, and temporal variables. These models learn from historical data to identify complex patterns and relationships that traditional statistical methods might not capture.

This section is split in two; supervised learning and deep learning in subsection 4.4.1 and subsection 4.4.2, respectively. Supervised learning techniques, also known as ensemble models, comprise of Random Forest, gradient boosting decision trees, and a few others. Alternatively, deep learning comprises of convolution neural networks, recurrent neural networks, and graph neural networks.

4.4.1. Supervised Learning

Gradient Boosting Decision Trees (GBDT) are a powerful and widely used machine learning method that combines the concepts of gradient boosting and decision trees. Gradient Boosting is an ensemble technique that builds models in a sequential manner [32]. Each new model focuses on correcting the errors made by the previous models in the sequence. Decision Trees are simple models that make decisions based on asking a series of questions based on the features of the data. This method stands out for its prediction speed and accuracy, especially with large and complex datasets.

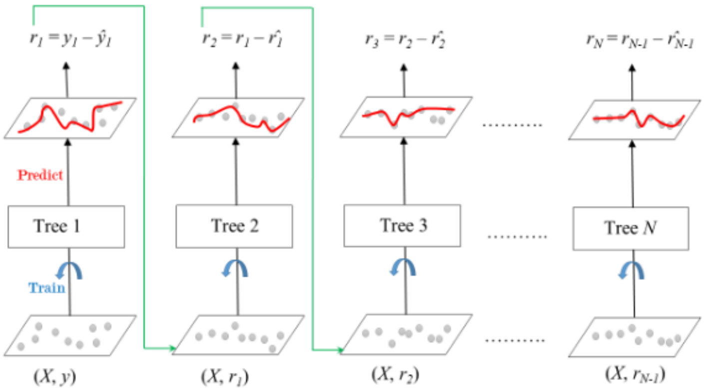


Figure 4.4: Gradient Boosted Trees for Regression [33]

However, traditional versions of GBDT require scanning all data points for each feature to calculate the information gained from all potential split points. As a result, their computational demands grow with both the number of features and data instances, making these methods time-intensive for large datasets [34]. For this reason, LightBGM, a novel GBDT algorithm has been developed featuring two innovative approaches: Gradient-based One-Side Sampling and Exclusive Feature Bundling. These techniques are designed to efficiently handle large datasets and a high number of features, respectively [34].

In 2021, to predict take-off times a Gradient Boosted Decision Trees (GBDT) model, specifically leveraging LightGBM due to its efficiency and scalability, is implemented. The choice is motivated by LightGBM's ability to process extensive data volumes, manage noise, and automatically identify significant features without extensive manual intervention [35]. The GBDT model demonstrated a significant improvement in predicting take-off times, this improvement was particularly notable one hour before the scheduled departure time, indicating the model's effectiveness in capturing and analyzing complex patterns affecting take-off times. The study also identified the most influential factors affecting predictions, including ATFM regulations, weather conditions, and the operational status of previous flight legs.

Several studies have been focused on comparing ensemble methods, which are a set of machine learning techniques that develop a robust learner from a set of weak learners, such as decision trees [35] [36]. Table 4.5 details the findings of three different studies that compared the performance of machine learning algorithms. These models were selected due to their varying underlying mechanisms (gradient boosting, neural networks, and decision forests, respectively), allowing a comprehensive evaluation of different machine learning approaches [36].

<table>
<tr><td colspan="2">Study</td><td colspan="4">Yet at al. (2022)</td></tr>
<tr><td colspan="2">Models compared</td><td>Light GBM</td><td>Support Vector Machine</td><td>Extremely Randomized Trees</td><td>Multiple Linear regression</td></tr>
<tr><td rowspan="2">Best model</td><td>Name</td><td colspan="4">LightGBM</td></tr>
<tr><td>Reason</td><td colspan="4">- Achieved an accuracy rate of 0.8655 with a MAE of 6.65min for a 1-hour forecast horizon
- Outperformed other models & improved from past research results by reducing the MAE by 1.83min
- Aggregate characteristics, such as the number of planned departures and the expected delay time of departures before the prediction period, heavily influence the prediction accuracy</td></tr>
</table>

<table>
<tr><td colspan="2">Study</td><td colspan="3">Lambelho et al. (2020)</td></tr>
<tr><td colspan="2">Models compared</td><td>LightGBM</td><td>Multi-Layer perceptron</td><td>Random forests</td></tr>
<tr><td rowspan="2">Best model</td><td>Name</td><td colspan="3">LightGBM</td></tr>
<tr><td>Reason</td><td colspan="3">- Achieved accuracy rates of:
up to 79.4% for departure delays, 79.1% for arrival delays, and 98.7% for cancellations
- Faster computing speed, higher prediction accuracy & capability of handing large scale data</td></tr>
</table>

<table>
<tr><td colspan="2">Study</td><td colspan="4">Anguita et al. (2024)</td></tr>
<tr><td colspan="2">Models compared</td><td>Random forest</td><td>Support vector machine</td><td>Multi-Layer perceptron</td><td>+ 7 others</td></tr>
<tr><td rowspan="2">Best model</td><td>Name</td><td colspan="4">Random forest</td></tr>
<tr><td>Reason</td><td colspan="4">demonstrated the best performance, with prediction accuracy measured using the RMSE</td></tr>
</table>

Table 4.5: Summary of comparative studies and their findings [37][36][38]

In summary, LightGBM was found to provide the best results, chosen for its efficiency in handling large datasets and speed in training and prediction, making it suitable for real-time prediction tasks. LightGBM's high efficiency and accuracy make it particularly effective for the predictive modeling of flight delays, where real-time data processing is crucial. And, while not explicitly discussed, a potential limitation of using LightGBM could be its complexity in tuning for optimal performance, requiring extensive knowledge and experience.

Another key finding of Anguita et al. is the analysis of feature importance identified key flight attributes influencing delay predictions, allowing for dataset simplification without sacrificing prediction accuracy [38]. Moreover, permutation importance is used to rank the importance of flight attributes when computing the flight delay. Future research could focus on incorporating additional flight attributes into the models to further improve prediction accuracy.

4.4.2. Deep learning

Deep learning, a subset of machine learning involving neural networks with multiple layers, has been explored for its potential to model the high-dimensional, nonlinear relationships inherent in flight delay data. Deep learning models have been applied to predict flight delays by leveraging large datasets and capturing the intricate dependencies among delay factors.

Artificial Neural Networks

Models based on probability, statistics, and operations research do not fully address the challenges posed by the complexity and dynamic nature of air traffic. ANN techniques offer promising solutions due to their capability to handle nonlinear problems and adapt to changes in air traffic demand and capacity. However, one of the main challenges in using Artificial Neural Networks (ANNs) for flight delay prediction is handling nominal variables, which are prevalent in air traffic data (e.g., airport codes, and day of the week). The commonly used 1-of-N encoding for nominal variables introduces multi-collinearity[1] and increases the complexity of input data, thereby affecting the performance and interpretability of ANNs [39].

Khanmohammadi et al. introduced a novel ANN model, termed Multi-Level Input Layer Neural Network (MLIL-NN), designed to overcome the limitations of handling nominal variables in traditional ANNs [39]. This model features a multi-level input layer that facilitates the direct processing of nominal variables without converting them to numeric formats, thus avoiding the introduction of artificial orderings. This approach not only improves the model's performance by preventing multi-collinearity but also enhances interpretability, allowing for a clearer understanding of the relationship between input and output variables. The MLIL-NN model demonstrated superior performance compared to the traditional gradient descent backpropagation approach, with lower root mean squared error (RMSE) and faster training time. Future work could explore integrating this model with fuzzy logic to handle the complexity and enhance real-world applicability further [39].

By comparing the following three regression models, Partial least square regression, Random forests, and Neural networks, it is found that neural networks produce the lowest RMSE [40]. In addition, Neural Networks (NN) regression analysis for delay prediction is preferred for its superior performance in handling non-normal and collinear data, which are common in air traffic delay datasets. This approach allows for complex computations and provides better predictive power compared to other statistical models, addressing the limitations of Ordinary Least Squares Regression due to data non-normality and variable collinearity[2][40].

Research on traffic prediction using deep learning methods, particularly the graph convolutional neural network (GCN), has shown success in ground transportation [41][42][43]. Examples include the use of GCN for complex non-linear relationship extraction in general graphs and intelligent methods developed for ground transportation traffic forecasting.

Previous studies on flight delay prediction have largely focused on single-airport scenarios, overlooking the complex spatial interactions among airports in a network. The literature reveals that understanding these interactions is crucial, as delays can propagate through the network due to interconnected resources like aircraft, crew, and passengers [41]. Although some works have begun to explore spatial dependencies, they often fall short in quantitatively predicting delays within these dynamic spatial interactions.

[1] when two or more independent variables have a high correlation with one another in a regression model
[2] correlation between predictor variables

Long Short-Term Memory

A study examined three different machine learning models aimed at predicting delays at various levels: individual flights, airports, and the network of airports. These include statistical regression models, recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, and spatial-temporal graph attention neural networks (GATs). The models were trained and tested using data from the EUROCONTROL research data archive, covering the top 50 European airports over two years. The results showed that the models could predict delays with an error of approximately 5 minutes or less for look-ahead times of up to 3 hours, which represents a significant advancement compared to existing prediction models [44]. Important takeaways are:

- The airport delay model employs an LSTM network to capture the dynamics of delay evolution, using past data to predict future states.
- The most innovative contribution is the development of dynamic spatial-temporal graph attention (DST-GAT) neural networks for network-level delay predictions.
  - This model considers not only the temporal features of individual airports but also the relationships between them, represented as a graph.
  - The edges of this graph are dynamic, reflecting the changing nature of flight connections and their impact on delays.
  - The DST-GAT model uses two types of adjacency matrices to represent relationships between airports: one based on geographical proximity and the other on flight connections. These matrices are updated dynamically to more accurately model the delay propagation in the network.
- The model demonstrates superior prediction accuracy and is capable of studying delay propagation across the network, providing insights into how delays at one airport can affect others.

The importance of selecting appropriate model parameters and the challenges in balancing model complexity with the available data is highlighted as important factors to consider when conducting research.

Graph Convolutional Neural Network

A model based on a Graph Convolutional Neural Network (GCN) enhanced with a temporal convolutional block and an adaptive graph convolutional block has been introduced to address the time-evolving and periodic nature of airport networks [41]. The model, named Multi-scale Spatial-Temporal Adaptive Graph Convolutional Neural Network (MSTAGCN), captures both spatial and temporal dependencies by processing sequences of graph snapshots. The model's effectiveness is attributed to its ability to model the time-evolving structure of airport networks and to adaptively capture spatial interactions among airports, even in the absence of direct flight connections.

- **Advantage**: its ability to outperform benchmark methods by capturing the dynamic interactions within the network
- **Disadvantage**: the complexity of implementing and training such a model could be considered a disadvantage, requiring extensive computational resources and data pre-processing

Graph Convolutional Neural Network's have scarcely been used in aviation, nevertheless, a wide range have been used to model railway networks. For instance, Graph Convolutional-Long short-term memory (GC-LSTM) was designed to accurately forecast the inflow and outflow of passengers within a high-speed rail network by effectively capturing the complex spatial dependencies dictated by the network's topological structure and temporal dependencies influenced by traffic dynamics and exogenous factors like holidays [42]. GCNs are utilized to manage the graph-structured data reflecting the network topology, while LSTMs address temporal patterns and dynamics. Furthermore, this model effectively captured the graph-based spatial and temporal dependencies in the network, demonstrating superior performance over traditional models like ARIMA and even over other neural network-based models that do not incorporate both spatial and temporal data as comprehensively.

Another paper utilized the Spatial-Temporal Graph Convolutional Network (STGCN) model to predict railway delays [43]. Four key takeaways from this research are:

- The STGCN model consistently demonstrated lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across various testing scenarios compared to linear regression (LR) and multi-layer perceptron (MLP). This is further elaborated in Section 4.9. This indicates a higher predictive accuracy, particularly in capturing the complexities of delay propagation in railway networks.
- One of the significant strengths of the STGCN model is its ability to effectively integrate both spatial and temporal dimensions of the data. This integration allows the model to understand and predict the cascading effects of delays across the network, which are influenced by both spatial connections (e.g., geographical layout and connectivity of stations) and temporal factors (e.g., the timing of delays and their progression over time).
- The model's graph-based approach allows it to capture the nonlinear relationships and dependencies within the network, which are often missed by traditional models that treat data points independently or do not fully capture the network's topology.

The Graph Convolutional Neural Network model, though primarily developed for railway networks, holds significant potential for application in flight networks due to the shared complexities of spatial and temporal dynamics in transportation logistics.

Deep Learning with Levenberg-Marquart optimization algorithm

A model was proposed that combines deep learning (specifically, stacked denoising autoencoders) with the Levenberg-Marquart (LM) optimization algorithm [45]. First and foremost, deep learning is chosen for its ability to handle complex datasets and automatically extract relevant features, while the LM algorithm optimizes the model by finding proper values for weights and biases, enhancing prediction accuracy. The model can effectively process large volumes of data, manage noise, and automatically identify significant features without extensive manual feature selection, making it highly suitable for complex prediction tasks like flight delays. Specifically, the proposed model achieved higher accuracy and precision on both imbalanced and balanced datasets when compared to models that either did not use denoising autoencoders or did not optimize with the LM algorithm [45]. The results underline the effectiveness of combining deep learning with optimization algorithms for predicting flight delays. Yazdi et al. suggests applying the proposed model to other datasets and examining the impact of additional variables on prediction accuracy [45].

Recurrent Neural Networks

To predict ATFM (Air Traffic Flow Management) delay evolution using historical data, a machine learning model (model utilizes a hierarchical structure and recurrent neural networks (RNN), specifically bi-directional GRUs) architecture is designed [46]. This research offers two significant insights;

- The choice of model was driven by its potential for short-term deployment and benefits, leveraging existing ATFM strategies and the extensive data collected by the network manager.
- Recommended to explore strategies to improve the performance of the trend classification model, particularly in detecting cases where delays increase.

Alternatively, night curfew infringements are often caused by a delay propagation along the sequence of flight legs. Hence, a two-stage model utilizing state-of-the-art machine learning techniques to predict the risk of night curfew infringement by forecasting the propagation of arrival delays across sequential flights of an aircraft is introduced. Although a different research focus, Dalmau et al. also introduced a model utilizing a bidirectional Gated Recurrent Unit (BiGRU) neural network to predict the probability of night curfew infringements [47]. The model performs predictions per aircraft and can be queried at any time of the day. A few reasons why BiGRU's are used are:

- Due to BiGRU's efficiency in handling sequential data and its ability to capture the temporal dynamics of flight schedules and delays.
- This model is distinct in its ability to compute the distribution of predicted in-block times for the last flight of the sequence, thereby estimating the risk of curfew infringement (and learn from historical delay patterns).

Yu et al. addresses the challenge of forecasting traffic under extreme conditions, such as peak hours and post-accident scenarios, which traditional linear models and early neural network models struggle with due to their unpredictability and non-linearity [48]. For this reason, a deep learning framework based on Long Short Term Memory (LSTM) units, which is capable of capturing both short-term and long-term traffic patterns is used.

Advantages

- The Deep LSTM model demonstrates superior performance in peak-hour forecasting, achieving as low as 5% Mean Absolute Percentage Error (MAPE), significantly outperforming traditional models and early neural network architectures.
- The LSTM-based models are proficient at learning both short-term and long-term dependencies in traffic patterns. This capability is crucial for accurately predicting conditions that are influenced by various factors over time.
- The inclusion of time-related features (e.g., time of day, day of the week) in the Deep LSTM model aids in capturing the recurring nature of peak-hour traffic, enhancing the model's predictive power.

- Uses an innovative technique for examining the model which reveals how the neural network processes information. Thus, showing the model's capability to memorize historical patterns and adapt to disruptions.

Disadvantages

- Deep learning models, especially those with multiple layers like Deep LSTM, require substantial computational resources for training. Hence, this might limit its applicability in environments with restricted computational capabilities.
- While the paper mentions the use of dropout and L2 regularization to mitigate over-fitting, deep neural networks' complexity inherently increases the risk of over-fitting. This requires careful tuning of regularization parameters and model architecture.
- Despite the proposed model inspection method, deep learning models remain relatively black-box compared to simpler, more interpretable models. Understanding why the model makes specific predictions can be challenging, limiting trust and applicability in critical decision-making processes.

Interestingly, RNN has a sort of memory over previous computations and uses this information when processing the current input of the sequence. In a delay propagation model, improving the model's architecture could involve incorporating a two-dimensional recurrent neural network. This enhancement would allow the hidden state to propagate across both flights and time [47].

Model	Advantages	Disadvantages	Overall conclusion	Score
Artificial Neural Networks	Can handle nonlinear problems and adapt to changes in air traffic demand and capacity. Improves performance and interpretability, especially with MLIL-NN for nominal variables.	Handling of nominal variables increases complexity and can affect performance without specific adaptations.	ANNs, especially with innovations like MLIL-NN, offer effective solutions for modeling complex and dynamic air traffic systems.	
Long Short-Term Memory	Captures dynamics of delay evolution and interactions among airports with high accuracy. DST-GAT for network-level predictions considers temporal and spatial relationships.	Complexity and computational demands for training, especially with advanced models like DST-GAT.	Provide accurate delay predictions at individual, airport, and network levels, capturing complex temporal and spatial dependencies.	
Graph Convolutional Neural Network	Captures dynamic interactions within airport networks effectively. Adapts to temporal and spatial dependencies without direct flight connections.	Implementing & training the model requires extensive computational resources and data preprocessing.	Excel in modeling the time-evolving structure of airport networks and adapting to spatial interactions (especially MSTAGCN).	
Deep Learning w/ Levenberg Marquart optimization	Handles complex datasets and automates feature extraction. Enhances prediction accuracy with optimized weight and bias values.	Complexity in tuning for optimal performance and computational resources for training.	The combination of deep learning with optimization algorithms like LM offers precise and effective flight delay predictions.	
Recurrent Neural Network	Handles sequential data well, capturing temporal dynamics of flight schedules and delays. Efficient in short-term deployment and predicting delay evolution.	Requires strategies to improve performance, especially in trend classification for delay increases.	RNNs, particularly with hierarchical structures and bi-directional GRUs, are powerful in predicting delays & analyzing sequential data.	

Table 4.6: Summary of Deep Learning Models for Flight Delay Prediction

4.5. Features

A shift towards metrics focused on passengers rather than aircraft has been motivated, aiming to assess the impact of disruptions from a passenger perspective.

Research	Departure & arrival time	Historical data	Weather	Flight plans	Airport capacity	Cockpit crew	Cabin crew	Connecting passenger data	
AhmadBeygi et al. (2008)	✓					✓			[28]
Campanelli et al. (2014)	✓				✓			✓	[11]
Hao et al. (2014)	✓	✓	✓						[49]
Fleurquin (2015)	✓	✓			✓				[13]
Khanmohammadi et al. (2016)	✓	✓		✓					[39]
Wu et al. (2019)	✓	✓							[4]
Guo et al. (2020)	✓	✓	✓	✓	✓				[50]
Cai et al. (2021)	✓	✓							[29]
Falco et al. (2021)	✓			✓	✓				[51]
Dalmau et al. (2021)	✓			✓					[35]
Jia et al. (2022)	✓	✓							[24]
Sismanidou et al. (2022)	✓	✓		✓					[40]
Chauhan et al. (2023)	✓	✓		✓		✓		✓	[2]
This paper	✓	✓		✓		✓		✓	

Table 4.7: Common features used across multiple studies

Throughout the years researchers have pinpointed the importance of several features in enhancing, improving, and making it more relevant, and predominant, to involving crew and connecting passenger data in the models. In Table 4.7 it can clearly be seen what are the features most researchers use, which is strongly correlated to what data they have access to, and what the Master thesis research aims to include. As discussed in chapter 3, connecting passenger data and cockpit crew information is a strong set of features part of the model and part of the research gap.

For example, AhmadBeygi et al., looked at the relationship between aircraft and crew schedules and the potential for delays to spread across the network [28]. In addition, it further recommends the consideration of cabin crew and passenger connections in the analysis to provide a more comprehensive view of delay propagation. To strengthen this point, 8 years later the thesis findings by Pablo Fleurquin underscore the critical role of internal mechanisms, such as aircraft rotation and passenger connectivity, in the propagation of delays within the network [13]. The reason is that besides the aircraft, flight crew, and cabin crew delays also propagate forward. This includes delays caused by waiting for connecting passengers, bags, or cargo from delayed arriving flights [49].

Previous studies have overlooked modeling passenger connections, despite their significant contribution to flight delays and schedule disruptions [4] [19]. Future research avenues should delve into a thorough analysis of delay propagation attributed to both passenger and crew connections, alongside exploring the effects of maintenance and mechanical checks on delays.

R. Dalmau et al. take this a step further by suggesting the exploration of additional input features that could further refine prediction accuracy, such as passenger boarding status, crew schedules, and runway configurations [35]. Advocating for increased data sharing among airlines and airports to enhance model inputs and, consequently, prediction accuracy.

However, it must be noted that in 2022, Sismanidou et al. examined the role of connecting passengers in delay propagation within the US air transport system, focusing on hub-and-spoke networks where delays are more likely to propagate due to interconnected flights [40]. Focusing, exclusively, on the impact of connecting passengers on departure delays and delay propagation, is an area that has received limited attention in previous research. Future studies could enlarge the scope of the analysis to include all different delay parameters in one single model as well as determining the proportion of delays caused by connecting passengers.

The aim of extending the feature set for the prediction algorithms is to enhance the accuracy of predictions further.

4.6. Data

The literature surveyed underscores the predominance of studies focused on the US system, even by European researchers. A gap in modeling and simulating attempts specifically tailored to European flight operations can be clearly observed across the numerous papers I read.

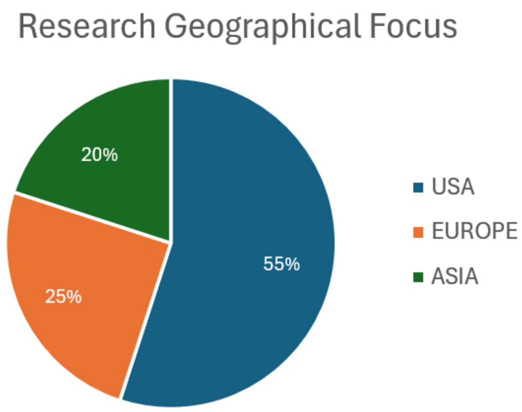


Figure 4.5: Geographic Scope of Earlier Research Papers

It can be seen from Figure 4.5 that a limited number of studies have analyzed European airline planning and traffic data for delay propagation patterns. Figure 4.5 shows the distribution of the geographical scope of past research.

Further division about what kind of approach was used, airport focus, airline focus, or accessed general public flight data. As can be seen from Figure 4.6, airline focus hasn't been a strong focus. Probably due to the lack of data with this respect, data privacy and protection. Public flight data consists of data extracted from either EUROCONTROL, CODA, ETFMS data, Flightradar24.com, Federal Aviation Administration, Bureau of Transportation Statistics, and a few others.

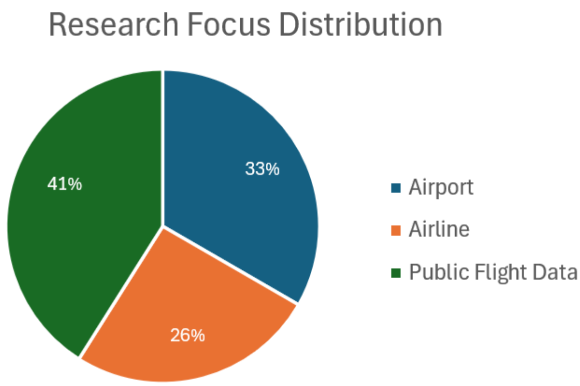


Figure 4.6: Distribution of Papers by Focus Area: Airports, Airlines, and Public Flight Data

Figure 4.6 displays the research direction chosen by researchers in their studies. It reveals that a significant majority concentrated on modeling airport networks, while a smaller number dedicated their efforts to airline fleets. Others faced limitations due to data access, making the most of what was available through public flight data.

The research highlights the significant impact small airports can have on delay propagation, suggesting that their limited resources and mitigation measures can contribute to the spread of delays. The findings show that delays propagated by large airlines tend to spread more broadly, while those propagated by small airlines spread faster [24]. This distinction is crucial for tailoring delay mitigation strategies to specific airline sizes and operational scopes. Key characteristics identified that should be taken into account include the disproportional importance of airports in delay propagation relative to their hub status or size, the presence of small-world network properties, and the ability to divide systems into separable sub-systems based on delay propagation perspectives [27].

Campanelli et al. suggests further exploration of individual flight impacts and the characteristics that contribute to significant delay propagation. It also highlights the importance of considering the interplay between an airline's connectivity pattern and the time of day in understanding delay dynamics [3].

4.7. Uncertainties

Model uncertainties have been addressed in different ways depending on the model used. These uncertainties are inherent in the aviation system and can vary in magnitude and impact over time, making them challenging to model accurately. A significant number of studies have identified similar uncertainties [3] [18] [27] [46] [52] [47]. Here are some examples:

1. **Weather conditions**: Uncertainties in weather forecasts, such as unexpected storms or severe weather patterns, can disrupt flight schedules and lead to delays.
2. **Operational issues**: Uncertainties related to airline operations, such as equipment malfunction, crew scheduling problems, or maintenance issues, can also cause delays.
3. **Air Traffic Control**: Delays can arise due to inefficiencies or disruptions in air traffic control systems.

4. **Passenger Behavior**: Passenger-related uncertainties, such as late arrivals, missed connections, or security incidents, can impact flight schedules and contribute to delays.
5. **Airline Network Interactions**: The complex interactions within the airline network, including code-sharing agreements, flight connections, and scheduling decisions, introduce uncertainties that can impact delay propagation.
6. **Airport Infrastructure**: Uncertainties related to airport capacity, runway availability, and ground handling services can influence the efficiency of flight operations and contribute to delays.
7. **Regulatory Factors**: Changes in regulations, security protocols, or airspace restrictions may introduce uncertainties that affect the flow of air traffic and lead to delays.

These uncertainties are inherent in the air transportation system and can interact in complex ways, making it challenging to predict and manage delay propagation. Different studies have considered these uncertainties in different ways, and below highlights some of their approaches.

Campanelli et al. analyze and quantify uncertainties through agent-based simulations, which simulate the dynamic behavior of the air transportation network under various initial conditions and disruptions [3]. By studying how delays propagate through the network, the paper seeks to improve our understanding of its robustness and identify strategies for mitigating the impact of uncertainties on flight schedules. Furthermore, by considering a mixture distribution to capture residual errors, a model can account for the inherent uncertainty in departure delay data, allowing for a more comprehensive analysis of delay dynamics [18].

Uncertainties in a regional air transport system can be modeled by proposing a two-stage analytical framework without relying on the strong and independent causality (SIC) assumption. By employing a convergent cross mapping (CCM) method, it is possible to capture the delay propagation effect, which is called delay causality based on the non-SIC assumption (DC-non-SIC) [27]. This method allows for the detection of weak delay causality and the influence of external interferences, such as weather conditions and air traffic control.

Moreover, the application of complex network theory to a delay causality graph highlights the system's vulnerability to delay propagation, underscoring the importance of modeling uncertainties in understanding and mitigating the impact of delays in air transport operations [27].

On the other hand, Dalmau et al. use two strategies for addressing uncertainties in predictions from a machine learning model [46].

- Firstly, Poisson regression variants are utilized to predict the expected ATFM (Air Traffic Flow Management) delay and its probability distribution, accommodating the zero-inflated nature of delays, thus offering a more comprehensive representation of uncertainty.
- Secondly, ordinal regression is employed to forecast the trend of ATFM delays, categorizing them as increase, decrease, or remain stable, with associated probabilities indicating the likelihood of each trend.

These approaches collectively enhance the model's ability to provide probabilistic estimates of delay evolution, facilitating more informed decision-making for airspace stakeholders.

Likewise, uncertainties can be modeled through empirical investigations using regression models and an algorithm that considers the variability and standard deviation of historical parameters [52]. Findings suggest that both the mean and the variance of historical flight times significantly influence strategic decisions like setting the scheduled travel time for flights and operational decisions such as the amount of fuel loaded.

An alternative approach to modeling the above uncertainties is to incorporate a two-stage approach. The first stage predicts the propagation of arrival delay along a sequence of flights using a neural network trained on historical data [47]. The second stage uses the distribution of predicted in-block times to compute the probability of night curfew infringement analytically. Uncertainties in the predictions are modeled through the distribution of predicted in-block times, derived from the neural network's output, allowing for an analytical computation of the risk of night curfew infringement. This approach acknowledges the inherent uncertainties in predicting flight delays and their propagation by focusing on the distribution of possible in-block times rather than a single value, thus offering a probabilistic assessment of curfew infringement risks.

Another way to analyze uncertainties is related to the stochastic (random) nature of two key factors in airport operations:

1. **Arrival Punctuality of Inbound Aircraft:** This uncertainty pertains to the variability in the actual arrival times of incoming flights compared to their scheduled arrival times. Factors such as weather conditions, air traffic congestion, and aircraft performance can all contribute to variations in arrival times.
2. **Performance of Ground Services:** This encompasses the uncertainties associated with the efficiency and effectiveness of ground handling operations, including aircraft turnaround times. Ground services cover a range of activities such as baggage handling, refueling, cleaning, and maintenance, each of which can be subject to various operational challenges and delays.

Through modeling these uncertainties probabilistically, [53] captures the inherent variability in these factors, in addition to assessing their impact on aircraft turnaround and schedule punctuality. This approach offers a more realistic representation of the operational environment, enabling better decision-making and optimization of airport operations.

4.8. Prediction Horizon

In the domain of flight delay prediction, a variety of computational models have been developed, each designed to operate over specific temporal horizons. These models leverage advanced machine learning techniques, including recurrent and graph neural networks, to predict delays with varying degrees of lookahead time. Here, we categorize these methodologies into four main temporal themes: "Hours to Minutes Before Departure," "Minutes to Seconds Before Departure," "Continuous Updates," and "Retrospective Analyses."

Hours to Minutes Before Departure

Predictive Modeling Approaches

These methods involve predictions made at four distinct time horizons: 15 hours, 6 hours, 3 hours before departure, and at departure time, showcasing the model's ability to adapt its predictions as the departure time approaches [51]. The selection of model parameters indicates a focus on data from up to three hours past to forecast future delays up to three hours ahead [44].

- Utilization of LSTM networks considers various time steps in the past, emphasizing the capability to anticipate future states within a three-hour window [17].
- The Granger Causality approach and the Spatio-Temporal Graph Dual-Attention neural network illustrate how hourly delays are influenced by the delays in the three preceding hours, underscoring a three-hour prediction window for practical planning and scheduling [25] [50].
- A recommendation emerges for a dynamic analysis of airline networks, suggesting an hourly examination of operations to enhance schedule resilience against disruptions [2].

Minutes to Seconds Before Departure

Complex vs. Simplified Models

The distinction between complex and simplified models highlights the trade-off between detail and computational efficiency [30].

- The complex model, while offering a nuanced view of congestion propagation, requires approximately 300 seconds for predictions, a timeframe considered too lengthy for immediate decision-making.
- Conversely, the simplified model, through assumptions of homogeneity among flights or operational simplifications, reduces prediction time to approximately 10 seconds, aligning better with the needs of real-time applications and rapid response scenarios.

Continuous Updates

Real-Time Predictive Adjustments

Emphasizing advancements in prediction methodologies, certain models, specifically leveraging LightGBM, are updated continuously from the moment the initial flight plan is received [35]. This approach significantly improves prediction accuracy, demonstrating a 30% reduction in take-off time prediction errors one hour before scheduled departure.

Retrospective Analyses

Historical Data Evaluation

Unlike predictive models, retrospective analyses focus on understanding past patterns of delay propagation. By analyzing historical operational data, these studies uncover insights into an airport's recovery capabilities from disruptions, offering a valuable perspective on delay management strategies without the pressure of real-time prediction [8] [2].

- The detailed examination of data such as flight numbers, airports, schedules, and crew details from a six-month period provides a foundation for these retrospective analyses, contributing to a comprehensive understanding of operational dynamics.

In summary, the exploration of flight delay prediction models reveals a landscape where temporal precision and prediction horizon vary significantly across approaches. From hours before departure to real-time and retrospective analyses, each methodology offers unique insights and challenges, pointing towards potential gaps such as the need for faster, yet detailed, real-time predictive models or more nuanced retrospective analyses to inform future delay mitigation strategies.

4.9. Mean Absolute Error

The evaluation of different predictive models in the context of airport and air traffic delay predictions reveals some clear trends regarding their performance, as measured by the Mean Absolute Error (MAE). The MAE is a metric used to quantify the accuracy of a model's predictions by calculating the average magnitude of the errors in these predictions. Table 4.8 details different MAE values found across literature with respect to delay prediction using different models.

Publication	Model used	Context	MAE	Source
Sun et al. (2022)	Dynamic Spatial Temporal Graph Attention Networks	Arrival delay prediction	<5 min	[44]
		Departure delay prediction	<4 min	
	Random forest	Single-event flight arrival delay prediction	3.8 - 7.7 min	
Dalmau et al.	Deep belief network	Predict delays of particular flights	8.5 min	[35]
Murgese et al. (2021)	Gated Recurrent Units (recurrent neural network)	Arrival delay prediction w/ 0 flight legs ahead	9 min	[47]
		Arrival delay prediction w/ multiple flight legs ahead	18 min	
Dalmau et al. (2021)	Basic and Poisson regression model	Prediction w/ 4h time horizon	4-26 min	[46]
		Prediction w/ <1h time horizon	10 min	
Cai et al.(2021)	MSTAGCN	Predict delays 1h ahead	5.8 min	[41]
		Predict delays 2h ahead	6.1 min	
		Predict delays 3h ahead	6.3 min	

Table 4.8: Mean Absolute Errors Across Research

Graph Convolutional Networks (GCNs), particularly Dynamic Spatial-Temporal Graph Attention Networks (DST-GAT) and MSTAGCN have shown exceptional performance, achieving the lowest MAE values among the studied models. These results underscore the effectiveness of GCNs in capturing the complex spatial-temporal relationships inherent in air traffic networks, making them highly suitable for accurate short-term and network-level delay predictions [41].

Recurrent Neural Networks (RNNs), like Gated Recurrent Units (GRUs), demonstrate moderate performance with MAE values increasing as the complexity of the prediction scenario increases, particularly when predicting delays over multiple flight legs. This variability highlights the challenges RNNs face in scenarios with increasing temporal dependencies.

Regression Models and Deep Belief Networks show varied performance with generally higher MAE values compared to GCNs. However, they still represent viable options depending on the specific characteristics and requirements of the prediction task.

In conclusion, while different models have their strengths and weaknesses, Graph Convolutional Networks stand out for their robustness and lower error rates, suggesting that they might be the best approach for enhancing the accuracy of flight delay predictions across various scenarios within the aviation industry.

The Spatial-Temporal Graph Convolutional Network (STGCN) model outperformed traditional statistical models in predicting delays, showcasing lower MAE and Root Mean Square error (RMSE) values across various testing conditions [43]. Table 4.9 clearly shows that the STGCN outperformed other statistical models, such as linear regression and multi-layer perceptron (MLP). This was attributed to its capability to capture the dependencies between nodes and the temporal dynamics within the railway network effectively.

Model	**MAE (10min)**	**MAE (30min)**	**MAE (60min)**
Linear Regression	0.279	0.337	0.338
Multi-Layer Perceptron	0.331	0.340	0.327
STGCN	0.250	0.282	0.270

Table 4.9: Mean Absolute Error Comparison between models [43]

5. Research Proposal

Establishing a robust foundation for the research proposal and precisely defining the research question starts with identifying the research gap, as detailed in Section 5.1. Subsequently, Section 5.2 specifies the central research question and related sub-questions, establishing a structured approach towards addressing the research aims. This chapter lays the foundation for the Master's thesis.

5.1. Research Gap

Drawing from the analysis in Chapter 4, it is possible to identify the following gaps in the existing literature.

- **GAP-1: Impact of connecting passengers on delay propagation**
  - Due to confidentiality reasons, researchers typically do not have access to connecting passenger data, which limits the comprehensiveness of delay propagation studies. This restriction is problematic as it prevents researchers from conducting a comprehensive analysis. Connecting passenger information is crucial for a holistic understanding of delay impacts and improving airline operations.
- **GAP-2: Lack of cabin crew data available for research in the context of flight delays and their propagation**
  - Due to privacy and protection reasons, researchers often lack access to crew data, restricting the depth of studies on delay propagation.
- **GAP-3: Exploration of Delay Propagation in Non-U.S. Networks**
  - The literature predominantly focuses on U.S. data and networks for studying delay propagation. There's a significant gap in research that analyzes European airline planning and traffic data for delay propagation patterns. This suggests a need for more geographically diverse studies that can provide insights applicable across different air traffic management systems and operational contexts.
- **GAP-4: Comprehensive Modeling of Delay Propagation**
  - Current models do not consider multiple connecting factors, for example, the aircraft, cabin crew, pilots, and passenger connections. Including these factors would offer a more comprehensive and realistic modeling of delay propagation, accounting for the complex interdependencies.
- **GAP-5: Lack of a Dynamic Prediction Horizon**
  - A notable gap exists in developing models that can predict the real-time evolution of delay for individual flights. There's a need for a solution that can be deployed in the short term

and utilize the vast amount of data collected to support airspace users in decision-making processes.
  - Furthermore, it is crucial to develop a computationally inexpensive model that can be executed multiple times a day. This requirement ensures the model's adaptability to the constantly changing operational landscape, enabling more accurate and timely decision-making processes.

- **GAP-6: Spatial-Temporal Dynamics of Delay Propagation**
  - Traditional studies might not offer a comprehensive view of delay propagation dynamics, as well as modeling reactionary delays. Thus, overlooking the implications of how delays evolve and affect various parts of the network differently. A novel approach would introduce a dynamic network perspective to study air traffic delay propagation using spatial-temporal networks.

5.2. Research Question

Building upon the research gap outlined in Section 5.1 and the problem statement described in Chapter 2, while also considering the insights provided by SWISS International Air Lines in Chapter 3, a hub-spoke airline, the following research question has been defined.

To what extent can we correctly determine the reactionary delay distribution over a (fleet) network, taking into account the effects of spoke airports?

To approach this comprehensive research question in a structured manner, it can be divided into several sub-questions:

1. To what extent do reactionary delays contribute to operational strain within airline networks?

- This question will be addressed by analyzing historical delay data from SWISS Airlines to quantify the operational impact. A multiple linear regression model can be used to correlate the extent of reactionary delays with indicators of operational strain, such as increased turnaround times and missed passenger connections. The analysis will help understand the magnitude of the impact of reactionary delays on operational efficiency.

2. What are the primary challenges in quantifying uncertainty in reactionary delay predictions across diverse airport operations?

- Expert interviews with airline operation managers and air traffic controllers will be conducted to identify the variety of factors contributing to delays and their perceived unpredictability. This qualitative insight will be supplemented with a quantitative analysis of variance in delay data across different airports to identify patterns of uncertainty.

3. To what extent does incorporating key categories of confidential airline data address the challenges of quantifying uncertainty in reactionary delay predictions across diverse airport operations, thereby enhancing the models' predictive accuracy?

- With access to confidential airline data, a more detailed predictive model will be built. Techniques like feature engineering will be used to integrate this data into existing models, and the incremental improvement in predictive accuracy will be measured.
- An analysis will be conducted to identify which types of confidential data (e.g., connecting passenger data and crew data) have the highest predictive value for modeling delays. This will involve data mining techniques to uncover hidden patterns and relationships.

4. How can uncertainty modeling techniques be optimized within a given model to accurately predict reactionary delays in a multi-airport environment?

- This sub-question will involve integrating and fine-tuning specific uncertainty modeling techniques directly into the chosen model for predicting reactionary delays, aiming to maximize forecasting accuracy in the context of a multi-airport environment.
- To focus on the development of models that can predict delays even with limited data, focusing on identifying patterns and factors specific to spoke airports.

5. To what extent can the outcomes of reactionary delay models be interpretable and usable by end-users, emphasizing feature importance?

- A user study will be designed to test different model visualization techniques, such as interactive dashboards, with operational staff from SWISS Airlines.

6. To what extent is it possible to design an accurate delay prediction model that ensures real-time update capability without compromising accuracy?

- Following the literature review, a comparative analysis will be conducted to evaluate the strengths and weaknesses of these frameworks in the context of delay prediction for airline operations. Key evaluation criteria will include latency (the time it takes to update predictions based on new data), scalability (the system's ability to handle increasing volumes of data without degradation in performance), and accuracy (the precision of the delay predictions).

6. Organization

This chapter outlines a strategic plan for completing my Master's thesis over the next nine months, from April 2024 to December 2024. Recognizing the need to balance academic, professional, and personal commitments, this plan is designed to ensure timely progress and high-quality research outcomes.

Work Schedule and Academic Commitments

I am currently required to complete 7 credits from two courses: Systems Identification and Bio-Inspired Decision Making. Both courses are assignment-based, for which I attended lectures last year and attempted the projects. To fulfill these academic requirements and maintain progress in my thesis, I will work at an 80% capacity. This arrangement translates to dedicating four days a week to my thesis, with the remaining day allocated to completing the necessary coursework and fulfilling my responsibilities as an intern at SWISS Airlines.

Thesis Structure and Timeline

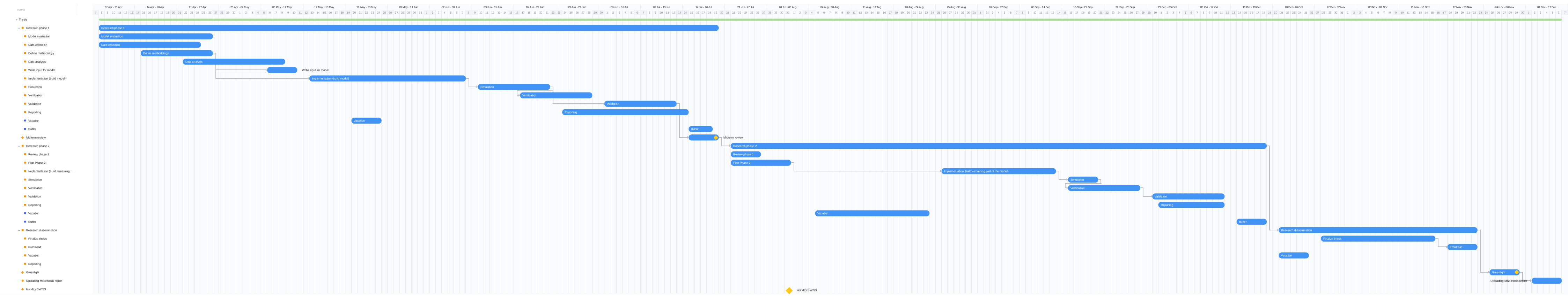
The thesis is structured into four main phases: Literature Review, Research Phase 1, Research Phase 2, and Dissemination. Each phase is critical for the development and completion of the thesis.

1. **Literature review**: This initial phase involves a comprehensive review of existing literature to establish a solid foundation for the research and to define the research question.
2. **Research Phase 1**: The focus is on creating a basic model. Activities include model evaluation, data collection, data analysis, writing input for the model, implementation (building the model), simulation, verification, and validation, leading to the drafting of the report and minimum product file.
3. **Research Phase 2**: This phase also serves as a critical checkpoint to evaluate the progress made towards answering the initial research questions. It's an opportunity to assess whether the research is on the right track. In this phase, the scope of the initial model is broadened by incorporating additional variables, expanding the number of airports under study, and enhancing the model with more comprehensive features.
4. **Dissemination**: The final phase involves preparing the thesis for submission and planning for its presentation.

6.1. Gantt Chart

A detailed Gantt chart is presented on the next page, delineating the timeline for each activity in particular within the Research Phases 1 and 2, including designated periods for model evaluation, data collection and analysis, implementation, simulation, verification, and validation.

Master Thesis Gantt Chart



7. Conclusion

In the interconnected world of air transportation, delays are not isolated incidents but rather pervasive issues that cascade through networks, significantly impacting various stakeholders including passengers and airlines. This thesis delves into the complexities of delay propagation within passenger airline networks, analyzing the pathways and severity of these delays on operational functionality. The extent to which delays not only inconvenience passengers but also lead to a range of challenges for airlines, including complexities in network management, safety concerns, financial losses, and environmental impacts is examined. The literature study conducted revealed a single delay can cascade through the entire network, with early delays having a more pronounced effect. This highlights the importance of proactive measures to manage and mitigate delays. The primary goal of this report is to present the outcomes of the literature review and lay the foundation for formulating an innovative research question for the thesis, aiming to address these critical challenges in delay management within the airline industry.

The literature search revealed that statistical and mathematical models provide foundational insights, but machine learning approaches, particularly using historical data, show superior performance in predicting delays more accurately. Bayesian networks and agent-based models offer nuanced understandings of delay propagation, demonstrating how interconnected variables influence delay dynamics across the airline network. The chapter emphasizes that early predictions using these advanced models enable more effective scheduling and operational decisions, significantly enhancing air traffic management. These findings underscore the shift towards more data-driven and predictive approaches in managing flight delays, which has become a pivotal strategy in modern air traffic systems.

Based on the current state-of-the-art, several research gaps have been identified in the field of flight delay propagation. First, there is a significant gap in the availability and use of connecting passenger data, which hinders comprehensive studies on how these connections influence delay propagation across networks. Additionally, there is a noted lack of accessible data on cabin crew movements, which limits the depth of analysis possible in understanding how crew scheduling impacts delays. Another gap identified is the limited research on delay propagation outside of U.S. networks, suggesting a need for studies that incorporate data from other regions, such as Europe, to understand regional differences in delay dynamics. Furthermore, the models currently used do not account sufficiently for multiple interconnected factors like aircraft, crew, and passenger schedules in their delay predictions. This oversight can affect the accuracy of predicting how delays propagate throughout an airline network. Lastly, there is a need for the development of real-time, dynamic predictive models that can adapt quickly to operational changes and provide accurate delay forecasts, which would be invaluable for decision-making processes in air traffic management.

Hence, based on the research gaps and current problem, the main research question of this thesis investigates how reactionary delays are distributed across hub-spoke airline networks, specifically examining the interconnectedness of flights as nodes and their connecting legs as edges within this complex network model. The research question is;

To what extent can we correctly determine the reactionary delay distribution over a (fleet) network, taking into account the effects of spoke airports?

This question is pivotal as it delves into modeling the flight network to visually and analytically observe the propagation of delays and how various factors—such as aircraft turnaround times, crew scheduling, and passenger transfers—are interconnected and contribute to systemic inefficiencies. By exploring these dynamics, the research aims to uncover how delays at specific nodes (spoke airports) impact the broader network, potentially causing cascading delays that ripple through interconnected flights and airports. The objective is to refine predictive models and optimize operational strategies to mitigate the impact of these delays. This enhanced understanding of the structural and dynamic aspects of delay propagation is critical for airlines striving to improve operational efficiency and customer satisfaction. The insights derived from this study not only advance the theoretical framework for managing airline delays but also provide airlines with robust tools to help real-time decision-making and strategic planning in airline operations.

One class of models that stands out from the literature are graph convolutional networks (GCNs) as they can efficiently handle data characterized by complex network structures, which are intrinsic to systems like airlines. These networks model the essential components—flights as nodes and the segments between them (legs) as edges, incorporating the detailed network topology into the analysis. By employing graph convolutions, models such as the Spatial-Temporal Graph Convolutional Network (STGCN) are equipped to directly utilize this structural data. This integration is crucial as it allows the model to capture both the spatial relationships among the nodes (flights) and the temporal dynamics along the edges (legs). Consequently, GCNs are particularly effective in predicting how an initial delay at one node (flight) ripples through to other parts of the network over time. Their scalability ensures robust performance even as they are applied to increasingly large and complex networks, offering a potent tool for enhancing operational planning and efficiency in sectors where precise coordination and timing are critical.

References

1. Rosenow, J., Michling, P., Schultz, M. & Schonberger, J. Evaluation of Strategies to Reduce the Cost Impacts of Flight Delays on Total Network Costs. *Aerospace* **7** (2020).
2. Chauhan, V. *et al.* Network science approach for identifying disruptive elements of an airline. *Data Science and Management* **6,** 110–121 (2023).
3. Campanelli, B., Ramasco, J., Fleurquin, P. & Eguiluz, V. Trees of Reactionary Delay: Addressing the Dynamical Robustness of the US Air Transportation Network. *SESAR Innovation Days* (2014).
4. Wu, C. & Law, K. Modelling the delay propagation effects of multiple resource connections in an airline network using a Bayesian network model. *Transportation Research,* 62–77 (2019).
5. Beatty, R., Hsu, R., Berry, L. & Rome, J. Preliminary Evaluation of Flight Delay Propagation through an Airline Schedule. *Air Traffic Control Quarterly* **7** (1999).
6. Fricke, H. & Schultz, M. Delay Impacts onto Turnaround Performance (2009).
7. Churchill, A., Lovell, D. & Ball, M. Flight Delay Propagation Impact on Strategic Air Traffic Flow Management. *Transportation Research Record,* 105–113 (2010).
8. Churchill, A., Lovell, D. & Ball, M. Examining the temporal evolution of propagated delays at individual airports: case studies (2007).
9. Castiglione, F. Agent Based Modelling. *Scholarpedia* **1,** 1562 (2006).
10. Janssen, M. Agent-Based Modelling. *Internet Encyclopaedia of Ecological Economics* (2005).
11. Campanelli, B. *et al.* Modeling Reactionary Delays in the European Air Transport Network. *SESAR Innovation Days* (2014).
12. Campanelli, B. *et al.* Modelling Delay Propagation Trees for Scheduled Flights. *ATM Research and Development Seminar* (2015).
13. Fleurquin, P. Systemic propagation of delays in the air-transportation network (2015).
14. Guleria, Y., Cai, Q., Alam, S. & Li, A. A Multi-Agent Approach for Reactionary Delay Prediction of Flights. *IEEE Access* **7** (2019).
15. Wang, C., Hu, M., Yang, L. & Zhao, Z. Prediction of air traffic delays: An agent-based model introducing refined parameter estimation methods. *PLoS ONE* **16** (2021).
16. Gurtner, G., Delgado, L. & Valput, D. An agent-based model for air transportation to capture network effects in assessing delay management mechanisms. *Transport Research Part C* **133** (2021).
17. Wu, W. & Wu, C. Enhanced delay propagation tree model with Bayesian Network for modelling flight delay propagation. *Transportation Planning and Technology* **41,** 319–335 (2018).
18. Tu, Y., Ball, M. & Jank, W. Estimating Flight Departure Delay Distributions-A Statistical Approach with Long-Term Trend and Short-Term Pattern. *Journal of the American Statistical Association* **103,** 112–125 (2008).
19. Kafle, N. & Zou, B. Modeling flight delay propagation: A new analytical-econometric approach. *Transportation Research Part B* **93,** 520–542 (2016).
20. Wang, Y. *et al.* Universal patterns in passenger flight departure delays. *Scientific Reports* **10** (2020).
21. Seth, A. Granger causality. *Scholarpedia* **2,** 1667 (2007).
22. Granger, C. Testing for causality: A personal viewpoint. *Journal of Economics Dynamics and Control* **2,** 329–352 (1980).
23. Du, W., Zhang, M., Zhang, Y., Cao, X. & Zhang, J. Delay causality network in air transport systems. *Transportation Research Part E* **118** (2018).
24. Jia, Z. & Cai, X. Delay propagation network in air transport systems based on refined nonlinear Granger causality. *Transportmetrica B* **10,** 1–13 (2022).
25. Bombelli, A. & Sallan, J. Analysis of the effect of extreme weather on the US domestic air network. A delay and cancellation propagation network approach. *Journal of Transport Geography* **107** (2023).

26. Feng, D., Hao, B. & Lai, J. Tracing delay network in air transportation combining causal propagation and complex network. *International Journal of Intelligent Networks* **5,** 63–76 (2024).

27. Guo, Z., Hao, M., Yu, B. & Yao, B. Detecting delay propagation in regional air transport systems using convergent cross mapping and complex network theory. *Transportation Research Part E* **157** (2022).

28. AhmadBeygi, S., Cohn, A., Guan, Y. & Belobaba, P. Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management* **14,** 221–236 (2008).

29. Cai, Q., Alam, S. & Duong, V. A spatial-temporal network perspective for the propagation dynamics of air traffic delays, Engineering (2021).

30. Dai, X., Hu, M., Tian, W. & Liu, H. Modeling Congestion Propagation in Multistage Schedule within an Airport Network. *Journal of Advanced Transporation* **2018** (2018).

31. Chen, S., Du, W., Liu, R. & Cao, X. Finding spatial and temporal features of delay propagation via multi-layer networks. *Physica A* **614** (2023).

32. Hastie, T., Tibshirani, R. & Friedman, J. *Boosting and Additive Trees: The Elements of Statistical Learning* 2nd ed. (Springer, New York, 2009).

33. Aksoy, N. & Genc, I. Predictive models development using gradient boosting based methods for solar power plants. *Journal of Computational Science* **67** (2023).

34. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree (2017).

35. Dalmau, R., Ballerini, F., Naessens, H., Belkhoura, S. & Wangnick, S. An explainable machine learning approach to improve take-off time predictions. *Journal of Air Transport Management* **95** (2021).

36. Wang, C., Hu, M., Yang, L. & Zhao, Z. Prediction of air traffic delays: An agent-based model introducing refined parameter estimation methods. *Journal of Air Transport Management* **82** (2020).

37. Ye, B., Liu, B., Tian, Y. & Wan, L. A Methodology for Predicting Aggregate Flight Departure Delays in Airports Based on Supervised Learning. *Sustainability* **12,** 1–13 (2022).

38. Anguita, J. M. & Olariaga, O. D. Prediction of departure flight delays through the use of predictive tools based on machine learning/deep learning algorithms. *The Aeronautical Journal,* 111–133 (2024).

39. Khanmohammadi, S., Tutun, S. & Kucuk, Y. A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport. *Procedia Computer Science,* 237–244 (2016).

40. Sismanidou, A., Tarradellas, J. & Suau-Sanchez, P. The uneven geography of US air traffic delays: Quantifying the impact of connecting passengers on delay propagation. *Journal of Transport Geography* (2022).

41. Cai, K., Li, Y., Fang, Y. & Zhu, Y. A Deep Learning Approach for Flight Delay Prediction through Time-Evolving Graphs. *IEEE Transactions on Intelligent Transportation Systems* (2021).

42. He, Y., Zhao, Y., Wang, H. & Tsui, K. L. GC-LSTM: A Deep Spatiotemporal Model for Passenger Flow Forecasting of High-Speed Rail Network. *IEEE 23rd International Conference on Intelligent Transportation Systems* (2020).

43. Heglund, J., Taleongpong, P., Hu, S. & Tran, H. T. Railway Delay Prediction with Spatial-Temporal Graph Convolutional Networks. *IEEE 23rd International Conference on Intelligent Transportation Systems* (2020).

44. Sun, J. *et al.* Designing Recurrent and Graph Neural Networks to Predict Airport and Air Traffic Network Delays. *10th International Conference for Research in Air Transportation FAA & Eurocontrol* (2022).

45. Yazdi, M., Kamel, S., Chabok, S. & Kheirabadi, M. Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data* (2020).

46. Dalmau, R., Genestier, B., Anoraud, C., Choroba, P. & Smith, D. A Machine Learning Approach to Predict the Evolution of Air Traffic Flow Management Delay. *ATM Research and Development Seminar* (2021).

47. Dalmau, R., Murgese, G., Wandeler, Y., Correira, R. & Marsden, A. Early Detection of Night Curfew Infringements by Delay Propagation with Neural Networks (2021).

48. Yu, R., Li, Y., Shahabi, C., Demiryurek, U. & Liu, Y. Deep learning: A generic approach for extreme condition traffic forecasting. *Proceedings of the 2017 SIAM International Conference on Data Mining,* 777–785 (2017).

49. Hao, L., Hansen, M., Zhang, Y. & Post, J. New York, New York: Two ways of estimating the delay impact of New York airports. *Transportation Research,* 245–260 (2014).

50. Guo, Z. *et al.* SGDAN—A Spatio-Temporal Graph Dual-Attention Neural Network for Quantified Flight Delay Prediction. *Sensors* **20** (2020).

51. Falco, P. & Delgado, L. Prediction of reactionary delay and cost using machine learning. *Airline group of the International Federation of Operational Research Society* (2021).

52. Hao, L. Quantifying the Impact of Flight Predictability on Strategic and Operational Airline Decisions. *University of California, Berkeley* (2015).

53. Wu, C. & Caves, R. Aircraft operational costs and turnaround efficiency at airports. *Journal of Air Transport Management* **6,** 201–208 (2000).