Universiteit Leiden

Master Computer Science

Network-aware vs. network-agnostic flight delay propagation in airport networks

Name:	Bogdan Aioanei
Student ID:	s3268322
Date:	26/07/2023
Specialisation:	Artificial Intelligence
Supervisors:	dr. Frank Takes, dr. Akrati Saxena and Maxwell McNeil

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Air congestions and delays naturally occur. Considering the fact that aircrafts perform multiple flights per day, a delay occurring at a given time will propagate, if uncompensated, to the following flights and will interfere with the normal operations of other aircrafts. Understanding how delays propagate and having the ability to predict when a delay will appear is therefore of increasing importance. This thesis provides a comparison study between two methods used to model and predict flight delays within a real world airport network using two different epidemic transmission models. The key difference between the two is that one method uses information regarding individual airport-to-airport connections, i.e. it is network-aware, while the other is network-agnostic. The two methods are tested on the task of predicting the number of airports experiencing above normal delays. The experiments use real-world airport domestic flight data in the USA between May and September 2018. The experiments provide insight into the prediction accuracies of the two methods as well as how the models react to the abnormal delays occurring within the data set. This thesis, as far as we know, is the only such comparison study and provides concrete arguments choosing one delay propagation method over the other. Among other things, we find that fine-grained methods like the network-aware approach struggle when the the delays observed in the historical flight data have high variance. Overall, this work can be regarded as a stepping stone towards similar comparison studies covering flight delay propagation. One such study could be aggregating the delays of multiple smaller airports and analyzing the performance of the delay prediction models when replacing the smaller airports with that single aggregated entity.

Contents

1 Introduction

Air travel has become an increasingly important method of transport in society. The increase in popularity of air travel, whether it is for business or leisure purposes, translates into the increase in overall demand. This rapid growth of the air travel market has made the ability to travel quickly and reliably a difficult task to achieve [1].

Air traffic delays are a significant contributor to the current inefficiencies within the flight traffic network [1]. Delays occur when a flight is unable to depart or arrive at its destination at its scheduled time. The reasons for which flight delays occur include weather conditions, mechanical problems, air traffic congestion, crew availability and security concerns. Although air traffic demand has been seriously impacted in the last years due to the Covid-19 Pandemic, the effects of the pandemic are gradually decreasing. Considering that airspace is a limited resource and the air traffic demand is continuously increasing, understanding flight delays, their effect and finding ways to deal with them are important and non-trivial tasks that need to be addressed to reach a more efficient state of what can be called the Air Traffic Network. The Air Traffic Network, or the Airport Network, is defined as the directed graph consisting of a set of nodes and a set of edges, where the nodes are represented by actual airports and an edge exists between two nodes if there is direct flight connecting the respective airports. Figure 1 presents a simplified example of the airport network, using some real airports located in the USA.

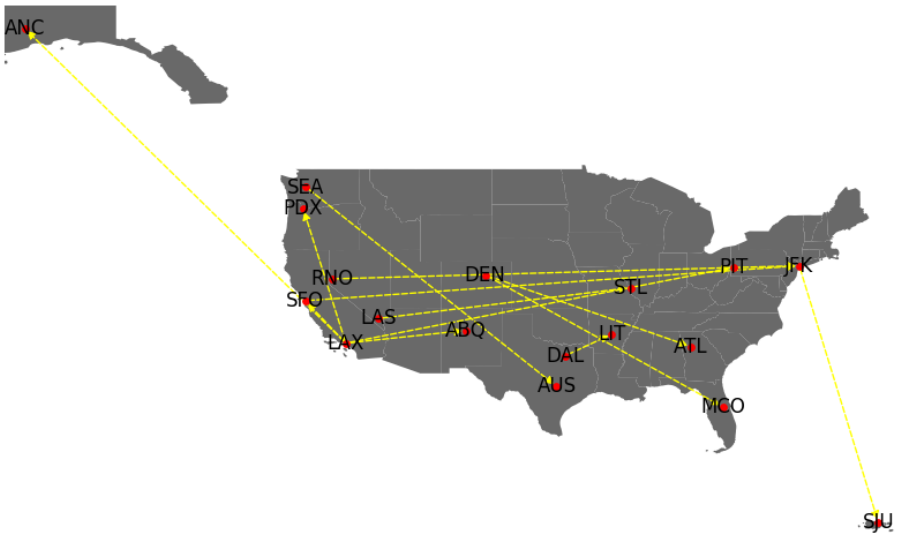


Figure 1: Simplified example of the Airport Network in the United States of America

The importance of reducing flight delays is best highlighted when looking at the general unfolding of the flight process, from the scheduling phase to the flight operations phase. Flights are planned in such a way as to maintain the same aircraft in the air for as long as possible. Keeping an aircraft grounded at an airport is extremely expensive and therefore an airline will chain together multiple flights for the same aircraft. As a result, an aircraft will accumulate, in one or multiple trips, on average, 8 hours, or more, of flight time per day. Chapter 4 presents this exact scenario in more detail.

Flights are performed as fast as safety restrictions allow and therefore, flight itinerary schedules have relatively small error margins with respect to how much normal operations can be impeded before causing delays. As such, whenever a delay occurs, it means that the error

margin for the specific flight phase (i.e. taxi-out, departure, en-route flight, approach, landing, taxi-in) has been exceeded. If the delay can not be absorbed by the error margins of the next flight phase, it will propagate to all remaining phases or to the next flights. This is how delays propagate through a single flight itinerary. However, other aircrafts can be, and usually are, affected by these delays. This is because airports generally perform operations at maximum capacity, or close to maximum capacity. This means that a delayed inbound airplane i.e., an aircraft which intends to land, but missed its initial allocated landing time-slot, at one of these airports will interfere with other inbound airplanes causing further delays, like a domino effect. The same effect can be observed for outbound airplanes. This is broadly defined as the ripple effect produced by flight delays [2]. Delays are an economic inconvenience for all stakeholders involved and in order to minimize the effects of delays, addressing the problem of delay propagation is necessary.

The long-lasting solutions to the problem of delays and the effects of their propagation through the network refer to improving the overall system before the delays can even occur. The measures that can be taken before departure include improving airport infrastructure and technology, optimizing flight schedules and routes, and improving communication and coordination between airlines and air traffic control. Because the safety aspect is prevalent in all of the normal operations in the air travel industry, the implementation of these measures can prove to be a slow and difficult process.

The solutions with a more immediate effect are the often short-term and taken mid-operations. While the pre-flight measures are similar to searching for a cure to a disease, the mid-flight decision making is more akin to alleviating the symptoms. The short-term measures are limited in scope: implementing big measures can lead to negative, yet unseen, effects while small measures may not have enough impact. In order to find the right balance, having accurate knowledge of the situation within the network at all times is a prerequisite. Therefore information is indispensable for reducing the effect of delays mid-operations. More concretely, accurately forecasting when and where flight delays occur can prove to be a game-changer. Forecasting generally refers to the process of making decisions based on past and present data by analyzing trends and identifying patterns [3]. Similarly, flight delay forecasting can be defined as the process of identifying the likelihood and the duration of delays before they occur. However, this is a non-trivial task due to the variety of reasons for which delays happen and their often hard-to-predict behavior.

The process of delay forecasting can be split into two main parts:

- Predicting **initial delays**, i.e. predicting when and where a delay will first appear within the airport network;
- Predicting **reactionary delays**, i.e. propagating the initial delay through the network and predicting its effects on the other nodes;

Initial delays are difficult to identify ahead of time. They are the first delays that occur in the flight itinerary and are unprovoked by other delays. As previously mentioned, delays happen due to a variety of reasons, from weather conditions and human or mechanical errors to bottlenecks created by other delayed flights. In the process of modeling the appearance of delays, working with unreliable or confidential data is required. For example, while weather forecasts are publicly available, they are always accompanied by a degree of uncertainty. Another example would be regarding the delays occurring due to mechanical issues observed during the pre-flight aircraft check-ups. In order to model these types of delays, one would require aircraft

status data, which, unfortunately is confidential. Propagating delays through the network however is a much more approachable task from a complexity standpoint. The reactionary delays are caused by previous delayed flights, whether they are flights within the same itinerary or exterior interference. While not representing a complete solution to the problem, obtaining reliable information regarding reactionary delays ahead of time is key to all stakeholders and therefore reliability of forecasting is the main aspect in the study of delay propagation.

This research deals with the study of flight delay forecasting within a real airport network, using statistical epidemic-based delay propagation models. The choice for epidemic-based models was a natural one considering the long history of research performed in this field and similarities between flight delays and epidemic diseases. This thesis explores two different solutions to issue of forecasting, in the form of delay propagation models. The two models skip entirely the source identification step (i.e. initial delay prediction) and focus only on the propagation part (i.e. modelling reactionary delays). Both approaches associate the process of delay propagation with the spreading of epidemic diseases using two different epidemiological models: a network-agnostic model [4] and a network-aware model [5]. In the case of these epidemic processes, the individuals in the network are represented by the airports and the disease is represented by the flight delay.

The network-agnostic model [4] implements a Susceptible-Infected-Recovered-Susceptible (SIRS) epidemic model which, as the name suggests, does not take into account the different connections between airports. Instead, it combines the propagation dynamics of the SIRS model with the time-dependent, numerical representation of the network in order to forecast delay occurrences. A particularity of this approach lies in the fact that the infection, recovery and immunity-loss rates, which are characteristic to epidemic processes, are captured at a network level, i.e. population level.

On the other hand, the network-aware model [5] implements a Susceptible-Infected-Susceptible (SIS) epidemic model, which takes into account the connections between airports. Contrary to the network-agnostic model, for each pair of connected airports, the infection and recovery rates are computed using historical data. Besides integrating network-level information, such as the airport-to-airport connectivity, another difference between the two models resides in the epidemic models themselves. The network-agnostic model proposes a SIRS model, which confers a momentary immunity to the recovered individuals. On the contrary, the network-aware model proposes the SIS model, which does not grant immunity at all. Every individual no longer considered to be infected immediately becomes susceptible.

The research thesis focuses on the comparison of network-agnostic and network-aware models. Thus, the research question we are concentrating on is:

- **RQ1:** *How does a network-aware epidemic model compare to a network-agnostic one in modelling flight delay propagation using real-world airport flight data?*

In the remainder of this thesis, in order to thoroughly address the proposed problem, we detail the related work that has been done in this field in Chapter 2. Following that, in Chapter 3, we will provide some definitions, metrics used and explain the basic principles and dynamics of the epidemic processes. Chapter 4 presents and overview of the real-world data sets used for the experiments. Chapter 5 details our approach, the implementation of the two models and the assumptions used. In Chapter 6, the experimental setup is presented and the results obtained. Lastly, in Chapter 7, we draw conclusions from the experiments and propose suggestions for future research.

2 Background and Related Work

In this chapter we discuss how other works have approached the task of flight delay propagation.

The research in the topic of flight delay propagation has seen a multitude of approaches and directions over the years. Among them, we distinguish two main categories: mathematical models and complex networks models. Each category is represented by works with both advantages and disadvantages.

The most common mathematical models are referred to as queuing models. The work presented in [6] introduces the idea of modelling the airport network as using an analytical queuing model called Queuing Engine. Each airport in the network is treated as a queuing system. The model is stochastic and updates the flight schedules corresponding to individual airports at every iteration. The approach captures the ripple effect created by flight delays by taking into account various factors such as flight schedules, weather conditions and gate allocations. The work presented in [7] builds upon the idea of using the Queuing Engine and adds a Link Transmission Model to the pipeline for computing delay propagation. The Link Transmission Model is used to compute delays at individual air sectors and propagates them to the delays felt at airport landings or departures. The model uses real-time flight data and air-sector characteristics to identify when a bottleneck or a sector overload occurs and propagates the resulting delay towards the airports receiving the respective incoming flights. Similarly to [7], [8] approaches the task of delay propagation from the perspective of en-route congestion. Simply put, en-route congestion is a phenomenon that appears when too many airplanes fly along the same designated air routes, thus exceeding the airspace sector's capacity. The approach uses historical data to learn a network composed of airports, congestion points and air corridors. The learnt network is used to approximate the real-world airport network. Following this [8] employs a similar stochastic and dynamic queuing network model to compute flight delays and track their propagation through the network previously learned. Although all the models which use the Queuing Engine perform well and manage to accurately identify and propagate delays, they have issues with scaling and employ the usage of confidential sector and airline related data. Due to the iterative process on which they are based, the number of airports and the number of flights queuing for landing or departure play a big role when assessing the time complexity of these models. Moreover, the usage of confidential data limits their applicability, i.e. individuals studies cannot hope to reproduce their implementation.

On the other hand, complex networks models trade off some of their accuracy for applicability. This category is most commonly represented by spread models. The work presented in [9] proposes an approach more akin to graph theory and network science. The delay propagation dynamics are captured using a spatial-temporal network. The nodes are represented by the airports and an edge between two nodes is active at a given time if there exists an ongoing flight between the two respective airports. In order to gauge the flight delay propagation within the network, the authors propose three metrics: magnitude, severity and speed. The metrics are computed using time dependent modified adjacency matrices, which, instead of capturing connectivity between nodes, are used to monitor the delays between nodes at a given time.

The more common sub-category of the complex networks approaches is represented by the epidemic spread models. Epidemiology studies how infectious diseases spread in a population with the help of spread models generally denoted by epidemic models [10]. These models generally follow the same outline:

1. **Compartmentalization of the population:** the individuals of the population are split

into pre-defined categories. The most common partitions are the susceptible, infected and recovered categories.

2. **Implementation of transmission dynamics:** a system of equations is established in order to model the changes in the network. Ordinary differential equations are the most common choice.

3. **Simulation and interpretation:** once the model is formulated, it can be simulated over time to predict the evolution of the epidemic disease.

The ingenuity behind the approaches within this sub-category resides in how similar the propagation of flight delays is to the propagation of a disease. Thus, the superimposition of the two perspectives becomes visible. By treating delays as a contagious disease that can spread from one flight to another, using epidemic models to monitor this dynamic offers a new and exciting perspective. The work presented in [11] introduces the three most general epidemic models, two of which serve as basis for the models introduced in flight delay studies: SIS model and SIR model.

The "Susceptible-Infected-Recovered (SIR) model", is used to model diseases which confer temporary or permanent immunity to the recovered individuals [11]. The work done in [12] applies this type of model on the various types of artificially generated airport networks. The research is directed towards understanding the role that airport traffic, airport connection and the level of airport turnaround services play in propagating delays. To this objective, the SIR model is simulated on various network configurations. Similarly to [12], [4] employs the usage of a modified SIR model. However, the research more closely integrates the epidemic model with the airport network taking into account various network-level characteristics and using real-world flight data. The paper implements a SIRS model, Susceptible-Infected-Recovered-Susceptible, model, which allows re-infection of already recovered nodes. This resembles the behavior observed in the real world, where airports suffering from delays are not exempt to further delays in the future.

The "Susceptible-Infected-Susceptible (SIS) model", is another basic epidemic model which is characterised by the fact that diseases don't confer immunity to recovered individuals. The work presented in [5] implements this type of epidemic model using two separate airport network perspectives: flight centered and airport centered. The flight centered approach treats individual flights as nodes while the airport centered treats individual airports as nodes.

This research compares the network-agnostic approach [4] with the network-aware approach [5]. While the various iterations of the Queuing Engine model presented in the previously described works show tremendous potential, that potential is only achievable in the right context. Without the necessary hardware capabilities and the wide plethora of data sources, most of which being confidential, this particular approach looses its appeal. On the other hand, the epidemic spread models have minimal data and medium hardware requirements. The choice of which epidemic models to employ was also a natural one. Considering the context of flight delays and how at any given time an airport can experience delays, regardless of previous ones, models which confer long-term or permanent immunity to recovered individuals are ruled out. As such, the SIS and the temporary immunity SIR models represent the best choices. Furthermore, considering the stochastic nature of flight delays, observing whether information regarding network connectivity plays a role in the overall predictions is an interesting and novel direction to pursue. Both of the originating researches for the network-aware [5] and network-agnostic [4] presented promising results on real-world data, as such, making them ideal candidates for the intended comparison.

3 Preliminaries

This research thesis uses terminology from the field of network science and aviation. This chapters gives the definitions and notations of certain terms and concepts and details the assumptions used.

A *network*, also called a *graph*, consists of a set of nodes and a set of edges connecting those nodes. It is denoted by $G = \{V, E\}$, where V is the set of nodes and E is the set of edges. In the case of an *airline network*, the set of nodes is represented by the set of unique airports and an edge exists between two nodes if there is a direct flight connecting the respective airports. From hereon, we formally define N as the total number of airports within the network:

$$N = |V| \tag{1}$$

Naturally, the airport network changes its structure over time, i.e. during different time periods, different airports may be active. In order to capture this time-evolving characteristic, the definition of the network is adapted to:

$$G = \{G_{t_0}, G_{t_1}, ..., G_{t_k}\} \tag{2}$$

Here $[t_0, t_k]$ represents the time interval in which the network was observed and G_{t_i} represents the snapshot of the network at time t_i . G_{t_i} is a directed subgraph and captures the active airports and the flights that either depart or arrive within the time interval $[t_i, t_{i+1}]$. In this research, each snapshot is of equal time length.

In general epidemiological studies, the observed population at time t is grouped into three distinct classes [11]:

- *susceptible class*, which contains all the individuals which can be infected by the disease, and is denoted with $S(t)$;
- *infective class*, which contains all the individuals which are infected by the disease, and is denoted with $I(t)$;
- *recovered class*, which contains all the individuals which were previously infected but have since recovered, and is denoted with $R(t)$;

While the susceptible and infective classes are always studied, the recovered class can be omitted depending on the objective of the study and the epidemiological model used [11].

Similarly to the aviation network, the epidemiological systems are time-evolving. This means that in an epidemiological study, the individuals within a population can switch between the three classes, depending on the spread of the disease. To this extent, we define the following coefficients: α, the infection rate, β, the recovery rate, γ, the immunity-loss rate [4].

This research focuses on applying epidemiological models to study the propagation of flight delays within the airport network and therefore, we associate the flight delays with the infectious disease and the set of airports with the individuals in the population. An airport having one delayed flight is not necessarily considered to be infected. In this research we use two metrics to determine whether an airport is considered delayed or infected: Normal Release Rate, denoted by NRR, and Average Flight Delay, denoted by AFD [13]. The normal release rate is a time-dependent metric, computed for each individual airport, and represents the fraction of flights operating normally:

$$NRR_i(t) = \frac{N - n_{dd}}{N} \tag{3}$$

Here N and n_{dd} represent the total number of departing flights and the total number of delayed flights at airport i, at time t [4].

The average flight delay is also a time-dependent metric, computed for each individual airport, and represents the average delay of both arriving and departing flights:

$$AFD_i(t) = \frac{\sum_k^{N+M} d_k}{N+M} \tag{4}$$

Here N and M represent the total number of departing and arriving flights, while d_k represents the delay (in minutes) of flight k, at airport i, at time t [4]. Using the NRR and AFD metrics we will later define the threshold for determining whether an airport is considered delayed, or infected, in chapter 5.1.

4 Data

In this chapter we introduce the data set used for the experiments presented in this thesis. In order to thoroughly investigate the accuracy and reliability of the proposed delay propagation epidemic models, we perform case studies on real world flight operation data within US airspace. The US domestic flight schedules have been obtained from the Bureau of Transportation Statistics website [1]. The data set contains all of the flight itineraries between May 2018 and September 2018. The year 2018 was chosen for this study due to the fact that the aviation industry has been severely affected by the Covid-19 Pandemic, between 2020 and 2022. The Bureau of Transportation statistics reports that between 2020 and 2022, the number of flights has decreased between by approximately 15% to 70% compared to the levels observed in 2019 [14]. This means that airports have been operating at a fraction of their potential capacity. Therefore, studying the airport network within this time period would not paint an accurate representation of the delay propagation during the airports' normal operations.

The data set contains a total of 2.17 million flights, spanning 153 days. Out of the total number of flights, 2.14 million flights had a departure delay higher than 15 minutes. Only a small percentage of these flights managed to fully recover the lost time mid-flight as 2.13 million flights still presented an arrival delay higher than the 15 minute threshold. Table 1 provides a brief description of each of the most important fields in the data set.

Field Name	**Field Description**
FlightDate	Flight Date (yyyy-mm-dd)
DayofMonth	Day of Month
Tail_Number	Unique identifier code of aircraft
Flight_Number_Operating_Airline	Unique identifier code of flight
OriginAirportID	Unique identifier code of origin airport
DestAirportID	Unique identifier code of destination airport
CRSDepTime	Scheduled Departure Time (local time: hhmm)
DepTime	Actual Departure Time (local time: hhmm)
DepDelay	Difference in minutes between scheduled and actual departure time
CRSArrTime	Scheduled Arrival Time (local time: hhmm)
ArrTime	Actual Arrival Time (local time: hhmm)
ArrDelay	Difference in minutes between scheduled and actual departure time
CRSElapsedTime	Scheduled elapsed time of flight, in minutes
ActualElapsedTime	Actual elapsed time of flight, in minutes

Table 1: Description of main fields within the data set

Table 2 presents the mean and standard deviation of several important fields within the data set. While the mean values of departure and arrival delays are within the 15 minute threshold, the standard deviation suggests a wide range of values.

	Flight Duration (min)	Departure Delay	Arrival Delay	Distance (miles)	Indegree	Outdegree
mean	110.7	12	7.8	799.4	16	16
std	70.6	47.5	49.6	602.2	27	27

Table 2: Mean and standard deviation of some important fields within the data set

Figure 2 displays the distributions of the departure and arrival delays, measured in minutes, for the US domestic flights within the entire study period (May - September, 2018). The hori-

[1]BTS data source `https://www.transtats.bts.gov/TableInfo.asp?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr&V0s1_b0yB=D`.

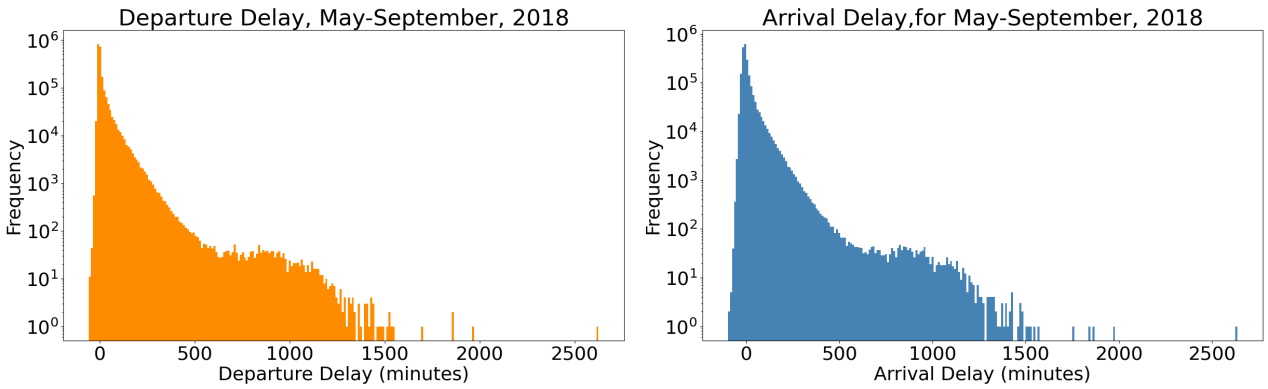


Figure 2: The distributions of delays, measured in minutes, on departure and arrival of US domestic flights, for the period of May to September, 2018

zontal axis represents the delay, in minutes, while the vertical axis is log-scaled and displays the frequency. Both distributions are right-skewed and we can immediately observe how common the delays are. The widespread interval of delay values confirms the initial suspicions: while the mean delay value is relatively small, a significant number of big delays still occur. While the majority of the flights have delays between 0 and 500 minutes, i.e. approximately 8.3 hours, there is also an important number of flights with even higher delays.

While it may be true that the delays happening in the departure phase generally carry over in the arrival phase as well, this does not mean that there exists a uniformity in when and where delays occur. Figure 3 presents the daily total departure delays throughout the data set. The monthly averages are displayed as dashed threshold horizontal lines. The delay values are given in minutes. The figure presents a high variance in total daily delays. This confirms the stochastic nature of delays. Although the months present similar monthly averages, with the exception of August, none of them present any evident periodicity with respect to their delays, even though airlines generally schedule their flights periodically. This in and of itself represents a major factor for which flight delay prediction is a considerably difficult task.

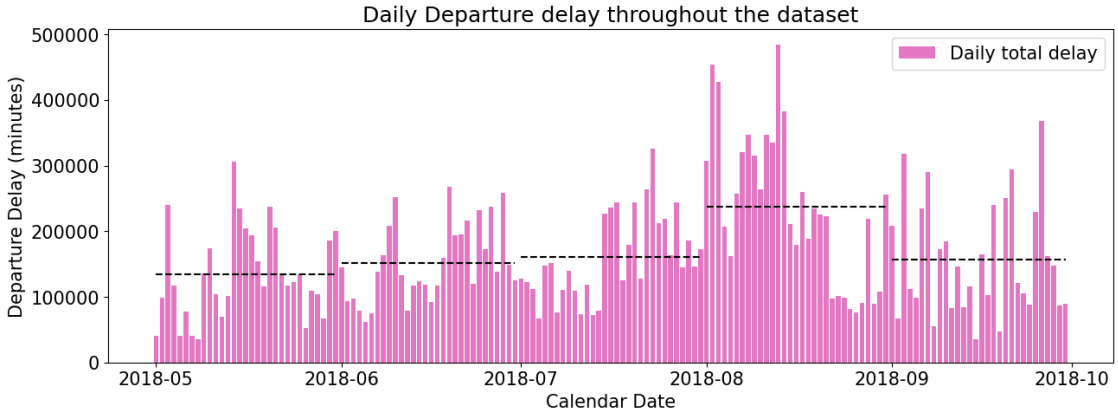


Figure 3: Daily total departure delay throughout the entire data set with monthly averages

Figure 4 displays the flight time distribution for all of the flights within the data set. The vertical axis is log-scaled and shows the frequency, while the horizontal axis displays the flight-time, in minutes. The majority of the flights have a total duration lower than 200 minutes, thus, when considering the fact that the majority of flights have delays higher than 15 minutes, the need for a more efficient network is further accentuated.

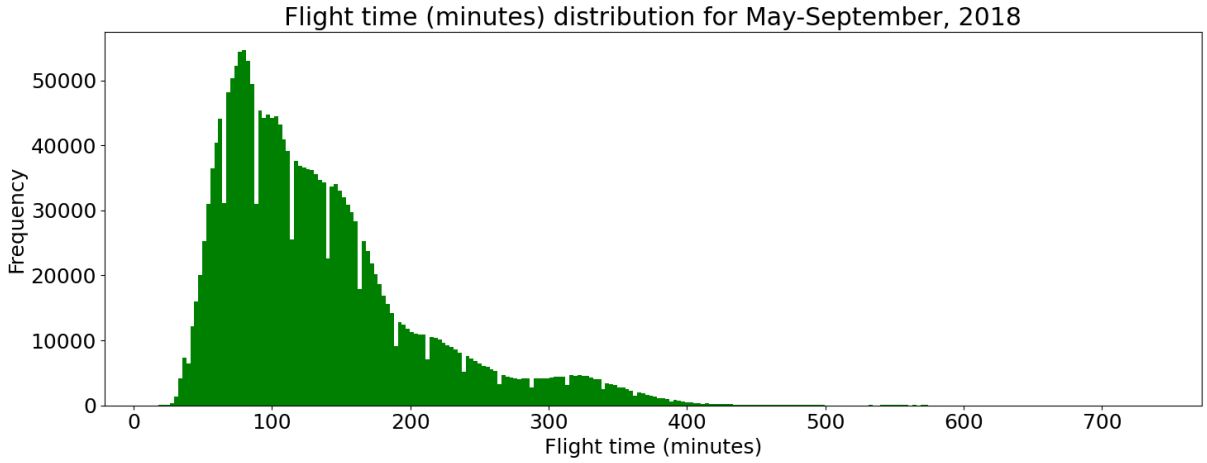


Figure 4: The flight time distribution for all flights within the period of May to September, 2018

Figure 5 presents the daily average flight time of individual aircrafts, measured in hours, throughout the entire data set. The mean over the 5 month period is at 8.2 hours and there is little deviation from this mean for the majority of the data set.

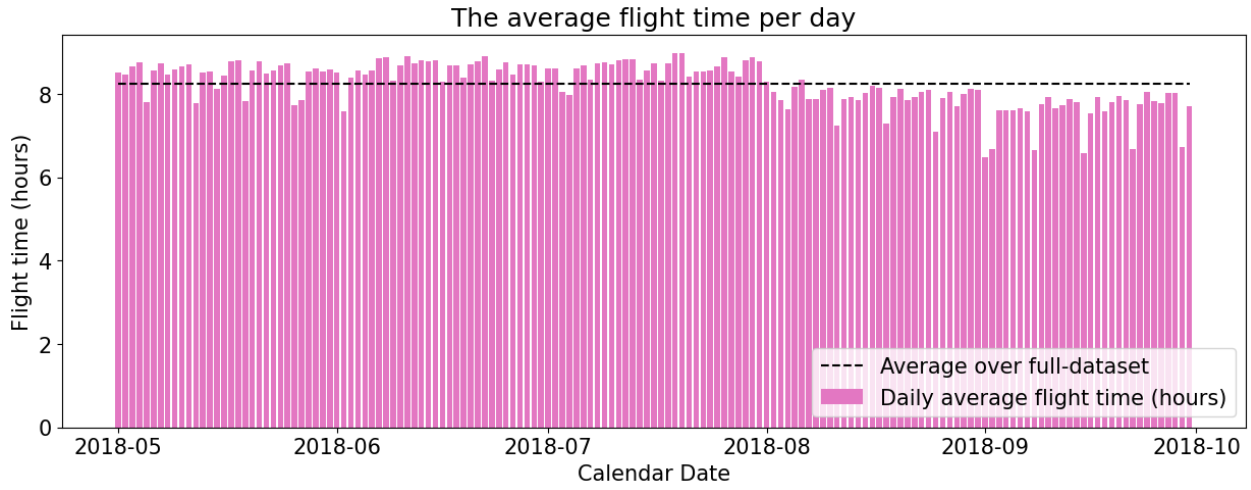


Figure 5: Average daily flight time of aircrafts throughout the data set, measured in hours

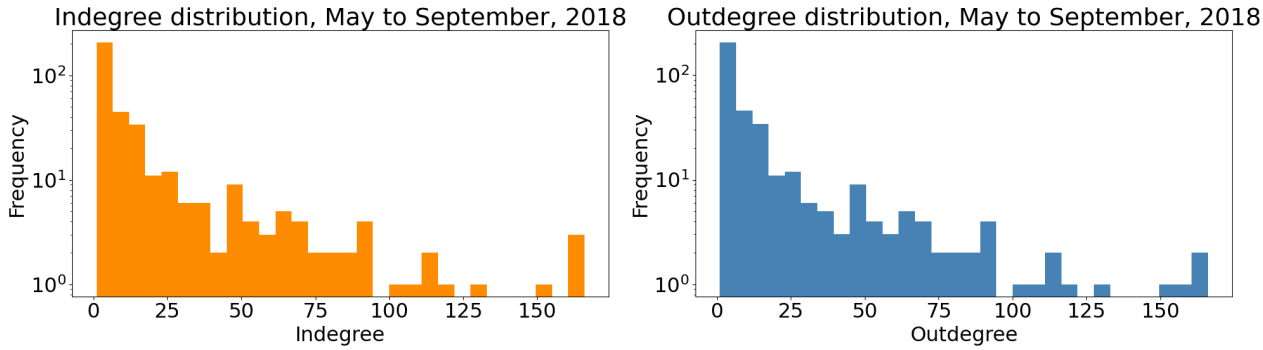


Figure 6: The indegree and outdegree distributions of the airport network, for the full period of May to September, 2018

Figure 6 displays the indegree and outdegree distributions of the airport network, for the full period of five months. It is worth noting that there are no multiple edges, i.e. they are only counted once. The network displays obvious hubs, with some airports having more than 150 incoming and outgoing edges. Moreover, it can be said that the overall network is generally well connected. Table 2 reveals that the network has mean indegrees and outdegrees of 16. Therefore the network is connected enough to facilitate the rapid transmission of delays from one airport to another.

5 Methodology

In this chapter, we discuss in detail the methodology used. First we describe the main objective of the research and the subsequent assumptions made. Following that, we will go into the metrics designed to support the results. Finally we detail the implementation of the network-agnostic and network-aware epidemiological models.

5.1 Objective and assumptions

The objective of the research paper is to compare two approaches, however, not all the output information is relevant. In order to perform a fair comparison, the capabilities of each models should be taken into account. While both models are able to predict both the number of infected and susceptible nodes, this research focuses only on the infected airports. Knowing that an airport is susceptible to being infected is not a relevant piece of information, since, as previously mentioned in Chapter 1, flight delays are stochastic in nature: they appear due to a variety of mostly unpredictable reasons. As such, at any given time, all non-infected airports can be regarded as susceptible or recovered, depending on their previous infection status. Thus, the output of the models and the subsequent experiments performed only take into account the number of infected airports at any given time, in order to decrease computational time.

Another assumption relates to how we define an airport being infected. Naturally, an airport experiencing only one delayed flight during a busy period of the day does not constitute grounds to it being considered infected. Similarly, as seen in Chapter 4, Figure 2, delays greatly vary in magnitude and thus, a delay of a greater magnitude cannot possibly have the same effect as a considerably smaller one. Thus, we remain consistent with the assumptions presented in [4]. In order to consider an airport being infected, or delayed, two separate conditions must apply simultaneously:

1. the normal release rate, NRR 3, must be lower than the threshold value of **0.7**;
2. the average flight delay, AFD 4, must be higher than **15 minutes**;

Moreover, in order for an airport to be considered recovered, at time t, the following conditions must apply:

1. the airport was infected at $t-1$;
2. $NRR \geq 0.7$ or $AFD \leq 15$ minutes at time t;

In all the other cases, the airport can be considered to be susceptible to infection.

It is worth noting that these parameters are worth experimenting with in order to assess the robustness of the models.

5.2 Metrics

The network-aware [5] and network-agnostic [4] models have been firstly implemented in research papers with different objectives and thus, in order to perform a proper comparison of the two approaches, not all of the capabilities of the network-aware model have been used. While network-aware is able to flag individual airports suffering from delays, only the aggregated number of infected airports is used.

The chosen metric is the Mean Absolute Error between the predicted number of infected airports and the actual one, at any given time, t.

5.3 Network-agnostic epidemiological model

Network-agnostic uses a very general and all-encompassing epidemiological model: the SIRS model. The infectious disease, in this case the flight delay, is modeled to confer temporary immunity to the recovered individual. The dynamics of the SIR model are described by system of equations in 5:

$$\begin{cases} \frac{dS(t)}{dt} = -\alpha(t)S(t)I(t) + \gamma(t)R(t) \\ \frac{dI(t)}{dt} = \alpha(t)S(t)I(t) - \beta(t)I(t) \\ \frac{dR(t)}{dt} = \beta(t)I(t) - \gamma(t)R(t) \\ S(t) + I(t) + R(t) = 1 \end{cases} \tag{5}$$

Here $S(t)$, $I(t)$ and $R(t)$ represent the fractions of susceptible, infected and recovered nodes. Attached to these variables, the coefficients $\alpha(t)$, $\beta(t)$ and $\gamma(t)$ represent the infection, recovery and immunity-loss rates [4]. All of the described variables and coefficients are time-dependent and airport-dependent. The system of equations is used to track the evolution of susceptible, infected and recovered individuals over time.

Before applying the system of equations in 5, the flight data needs to be preprocessed. Due to the nature of flight operations, time can not be perceived as a continuous variable and, instead, it needs to be discretized. The flights in the data set are grouped into individual time intervals, setting the length of consecutive intervals as a constant. The specific length of the time interval is treated as a hyperparameter to be later explored in Section 6.

After preprocessing, the data set is split into training and testing partitions and we begin the process of computing the necessary parameters and variables for the SIR model. Using the historical flight data from the training set, we compute the fractions of susceptible, infected and recovered nodes, $S(\Delta t)$, $I(\Delta t)$ and $R(\Delta t)$, for every time interval Δt. The following step is to compute the values for $\frac{dS(t)}{dt}$, $\frac{dI(t)}{dt}$ and $\frac{dR(t)}{dt}$ using Euler's Method [15]. Equations 6, 7 and 8 describe this process:

$$\frac{dS(t)}{dt} = \frac{S(t_{i+1}) - S(t_i)}{t_{i+1} - t_i} \tag{6}$$

$$\frac{dI(t)}{dt} = \frac{I(t_{i+1}) - I(t_i)}{t_{i+1} - t_i} \tag{7}$$

$$\frac{dR(t)}{dt} = \frac{R(t_{i+1}) - R(t_i)}{t_{i+1} - t_i} \tag{8}$$

Moving forward, we replace the computed values for $S(t)$, $I(t)$, $R(t)$, $\frac{dS(t)}{dt}$, $\frac{dI(t)}{dt}$ and $\frac{dR(t)}{dt}$ in the system of equations 5. The goal of this process is to ultimately compute the coefficients $\alpha(t)$, $\beta(t)$ and $\gamma(t)$. At the end of the learning process, we have access to the infection, recovery and immunity-loss coefficients computed from historical training data, for each time interval and for the entire network. Algorithm 1 gives an overview of the learning process.

In order to make predictions regarding the number of airports infected within the network at a given time, a reverse process is used on the test data set. Similarly to the training process, the $S(t)$, $I(t)$, $R(t)$ fractions are computed for time t, from the test data. Moving forward, the respective derivatives, $\frac{dS(t)}{dt}$, $\frac{dI(t)}{dt}$ and $\frac{dR(t)}{dt}$, are computed using the system of equations in 5 coupled with the $\alpha(t)$, $\beta(t)$ and $\gamma(t)$ coefficients previously identified in the learning procedure.

Algorithm 1 Learning Procedure for the network-agnostic model

Input: training data set, D, length of time interval, Δt
group the flights in D into time intervals of length Δt
for every Δt **do**
    compute $S(t)$, $I(t)$, $R(t)$
end for
for every Δt and $\Delta(t+1)$ **do**
    compute $\frac{dS(t)}{dt}$, $\frac{dI(t)}{dt}$, $\frac{dR(t)}{dt}$ using equations 6, 7,8
end for
for every pair $(S(t), \frac{dS(t)}{dt})$, $(I(t), \frac{dI(t)}{dt})$, $(R(t), \frac{dR(t)}{dt})$ **do**
    extract $\alpha(t)$, $\beta(t)$ and $\gamma(t)$ from the system of equations 5
end for
Return: $\alpha(t)$, $\beta(t)$ and $\gamma(t)$

Finally, the prediction is computed using Euler's Method [15]:

$$S(t_{i+1}) = S(t_i) + (t_{i+1} - t_i)\frac{dS(t)}{dt} \tag{9}$$

$$I(t_{i+1}) = I(t_i) + (t_{i+1} - t_i)\frac{dI(t)}{dt} \tag{10}$$

$$R(t_{i+1}) = R(t_i) + (t_{i+1} - t_i)\frac{dR(t)}{dt} \tag{11}$$

5.4 Network-aware epidemiological model

Contrary to the network-agnostic version, network-aware uses a SIS model. This time, the individual nodes in the network are no longer considered to be immune following the recovery from the delay. All recovered airports revert back to being susceptible to a new infection. The dynamics of the SIS model can be described as:

$$\begin{cases} \frac{dS(t)}{dt} = -\alpha(t)S(t)I(t) + \beta(t)I(t) \\ \frac{dI(t)}{dt} = \alpha(t)S(t)I(t) - \beta(t)I(t) \end{cases} \tag{12}$$

, where, similarly to the network-agnostic case, $S(t)$ and $I(t)$ represent the fractions of susceptible nodes. Attached to these variables, the coefficients $\alpha(t)$ and $\beta(t)$ represent the infection and recovery coefficients [5].

The awareness of the model comes from the fact that it takes into account the individual airport-to-airport connections from the network. The fractions of susceptible and infected nodes are modeled as vectors and the infection and recovery coefficients are modeled as matrices. To that extent, given time t, the infection dynamic of the SIS model becomes:

$$\dot{p} = (\mathbf{B} - \mathbf{D})p - [(\mathbf{B}p) \bullet p] \tag{13}$$

In equation 13, $p = (p_1, p_2, ...p_N)^T$ is the state vector of the system, where N is the total number of nodes. The state vector of the system conveys a similar information as the fraction of infected nodes, $I(t)$, described in the network-agnostic model, however, a different notation is used for outlining a clear separation between the two: $I(t)$ is a network-level parameter, while p is a vector containing all the individual parameters within the network. The "$\bullet$" notation

represents the Hadamard product (i.e. element-wise product). The $N \times N$, diagonal matrix $\mathbf{D}$ captures the recovery rates of each airport. The individual recovery rates are learned from historical flight data. $\mathbf{B}$ is a $N \times N$ matrix, containing the infection rates for all pairs of airports within the network. Each infection rate can be regarded as a normalized probability of a node being infected by another node, or the normalized inflow of infections. Equation 14 describes the computation process of each infection rate:

$$B_{ij} = \frac{N_{ij}}{N_i} \tag{14}$$

Here N_{ij} denotes the number of flights coming from airport j to airport i, while N_i denotes the total number of flights coming to airport i.

The state vector, $p = (p_1, p_2, ...p_N)^T$, is airport dependent and time dependent. It captures how infected a particular airport is at a given time:

$$p_i(t) = \frac{I_{arr} + I_{dep}}{N_i} \tag{15}$$

Here I_{arr} and I_{dep} represent the number of infected flights at arrival and departure stages at time t, while N_i represents the total number of flights at airport i, either arriving or preparing to depart.

The learning procedure of the model takes a similar route to the network-agnostic variant. The preprocessed data is split into training and testing partitions. The end-goal of the training is to infer the $\mathbf{B}$ and $\mathbf{D}$ matrices. Firstly, the state vector p is computed for every time interval, Δt, after which, using Euler's Method [15], the respective derivatives are computed, $\dot{p} = (\dot{p_1}, \dot{p_2}, ..., \dot{p_N})^T$. The components $[B_{ij}]$ of the $\mathbf{B}$ matrix do not need to be learned, they can directly be computed for each pair of airports using equation 14. The final step is to replace all the known parameters in equation 13 and extract the recovery rate matrix, $\mathbf{D}$.

After computing the infection rate and recovery rate matrices, predicting future infections is done by reversing the training process. Similarly to the network-agnostic approach, the state vector, p, is computed from the historical flight data in the test set. Together with the matrices $\mathbf{B}$ and $\mathbf{D}$, using equation 13 we compute the derivative, $\dot{p}$. Finally, using Euler's Method [15], we obtain the prediction value.

6 Experiments and Results

In this section we will discuss the experimental setup and the results obtained. Firstly, we will go into the hyperparameters that we will test and what are the types of experiments we designed in order to answer the research question. Moving forward we will present the results obtained and go to further explore the robustness of the models.

6.1 Experimental setup

The practical objective of delay propagation models, in general, is to give information to all stakeholders involved in the decision-making process. The more accurate this information is, the more informed decisions can be made. Accurate information is mandatory when preparing for avoidable delays. Thus, we devised two types of straightforward experiments to perform, in order to draw a complete comparison between the models, according to the overall objective:

- **Short term prediction:** given historical flight data prior to time t, predict the number of infected airports in the entire network at time $t + 1$;
- **Long term prediction:** given historical flight data prior to time t, predict how far into the future is the horizon for which the forecasting is still accurate, i.e. $t + 1$, $t + 2$, $t + 3$ etc.;

The hyperparameters of the models are the training and prediction window length, Δt, and the amount of training data used to infer the epidemic characteristic coefficients, L. The size of the training data, L, is expressed in number of days. It was chosen so due to the cyclical nature of flight operations: the same flights are continuously repeated with few additions based on the current season (e.g. tourist destinations are more popular during summer periods). Thus, in order to capture this repeated nature, full days worth of flights have been used as training data.

6.2 Short term prediction - Results

The short term prediction experiments serve the purpose of determining whether the models can be used for immediate responses to delays. Giving the models the most recent situation in the network, as well as historical flight data from previous days or weeks, from which to learn, should induce a low prediction error.

The first set of experiments in this category explores the prediction error of the two models, when using:

- $\Delta t \in [15, 30, 45, 60, 90, 120, 150, 180, 210]$ (minutes);
- $L \in [1, 7, 14, 21]$ (days);

In order to have an accurate representation of the capabilities of the two models, the testing was performed over a full day of flights in the following manner:

1. The day used for testing was split chronologically into time intervals of length, Δt;
2. For every interval in the day, the data up to that time interval as well as the data from the previous $L_i \in [1, 7, 14, 21]$ days was used to compute the model-specific parameters. For example, in order to predict the number of infected airports between the hours 12 PM and 13 PM, the models would use all of the flight data from the test day between 12 AM and 12 PM, as well as the previous L_i days;

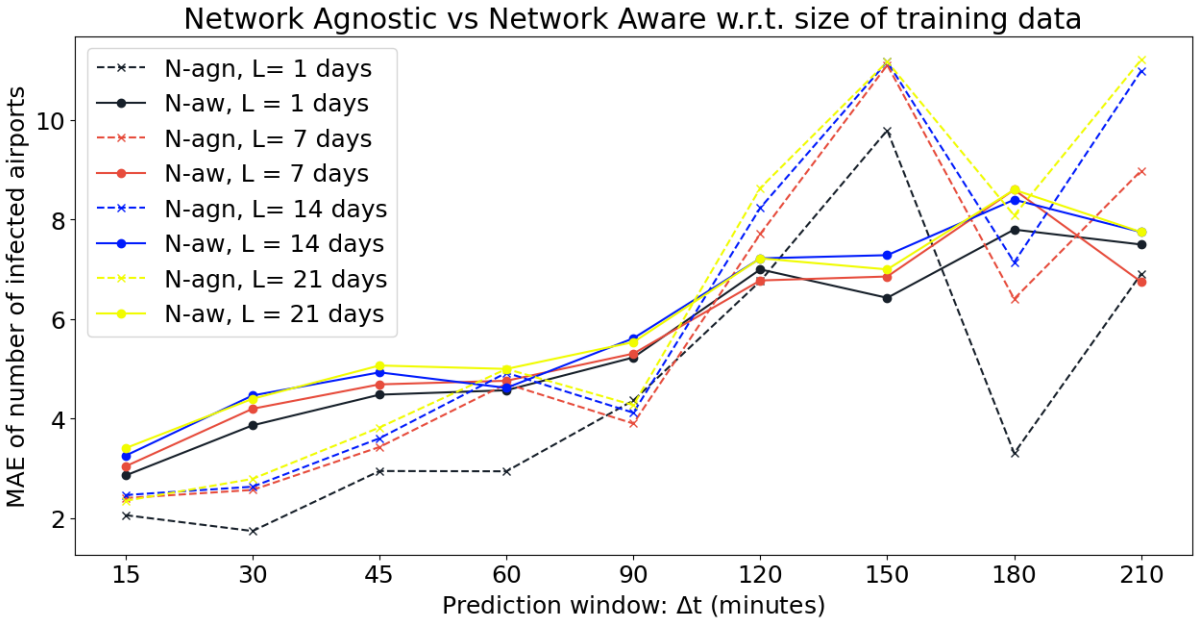


Figure 7: Short term prediction error w.r.t size of training data and prediction window length

3. The number of infected airports was predicted at every time interval, Δt, and the mean absolute error for the whole day was recorded;

The results of the first experiment were obtained using a single test day. The test day was randomly selected from a pool of close to average days, in terms of total daily delay.

Figure 7 presents the results obtained for this first run. The vertical axis represents the mean absolute error of the predicted number of infected airports versus the actual one, while the horizontal axis represents the prediction window length, Δt. The results of the network-aware model are displayed in continuous lines with circular markers while the results of the network-agnostic model are displayed in dashed lines, with cross markers. The results of the models when using the same value for L are displayed in the same color.

We first notice that the mean absolute error in number of airports increases with respect to the prediction window, for both models. This is natural since, the longer the prediction window, the more active airports exist concurrently in the network. Moreover, while this increase is steady for the network-aware approach, the network-agnostic model begins to heavily fluctuate after the 90-minute prediction window threshold. This is a somewhat expected behavior since flights can depart and land in the same prediction window and thus, the state of the network is more difficult to model when only looking at susceptible and infected nodes. On the other hand, the network-aware approach does not suffer from this drawback, since the approach models individual airport-to-airport connections by keeping track of all the flights ongoing in the network. On the other side of the coin however, when flights cannot begin and finish in the same window due to how short the window is, the network-agnostic approach outperforms the network-aware one. When predicting shorter windows, there is less movement in the network to be predicted and thus it becomes easier to model the overall network behavior.

The general idea that figure 7 paints is that the network-agnostic model performs relatively better when using shorter prediction intervals, while the network-agnostic model performs better, on average, when using longer ones. The effects of the number of training days used is not particularly visible. The mean errors obtained are within a small deviation interval. Although

the amount of training data used did not present a prominent effect in the results, the theoretical perspective for the network-aware model points out otherwise. Having more historical data regarding airport-to-airport connections should help the network-aware approach better approximate the airport-specific infection and recovery rates. In order to test this out, a new set of short term prediction experiments have been performed. The goal of this second set of experiments is to observe the prediction error when inputting longer prediction windows and more training data:

- $\Delta t \in [120, 150, 180, 210, 240, 270, 300, 330, 360, 390, 420]$ (minutes);
- $L \in [16, 19, 24, 26, 28]$ (days);

Figure 8 presents the results for this second run of tests. The results have been averaged in a similar manner as in Figure 7 over the course of 7 randomly selected days from the 5-month period available.

The small difference in performance with regards to the amount of training data, L, observed in Figure 7 is more obviously displayed here. While some minimal variation in results exists for the network-agnostic model, The network-aware model does not exhibit any of it. This suggests that either the performance variation has been lost due to averaging over multiple days, or the network-aware model actually is not sensitive to the amount of historical flight data used to infer the model-specific coefficients.

The results displayed in Figure 8 are contrasting with respect to the theoretical expectations. As previously stated, when given more data, the network-aware method should be able to more accurately extract the average infection and recovery rates of individual airport-to-airport connections and thus clearly outperform the network-agnostic approach. The results, however, do not completely paint this picture. While it is true that the results are more consistent when varying the prediction window, the network-aware model does not always outperform the agnostic approach.

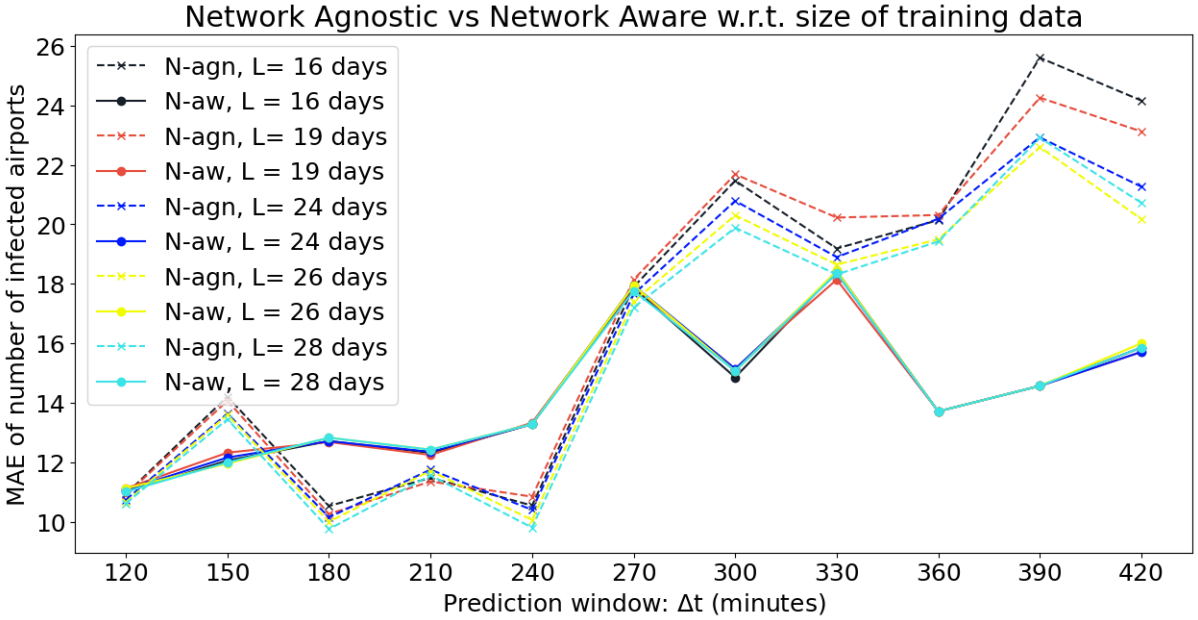


Figure 8: Short term prediction error w.r.t larger training data corpus and longer prediction window lengths

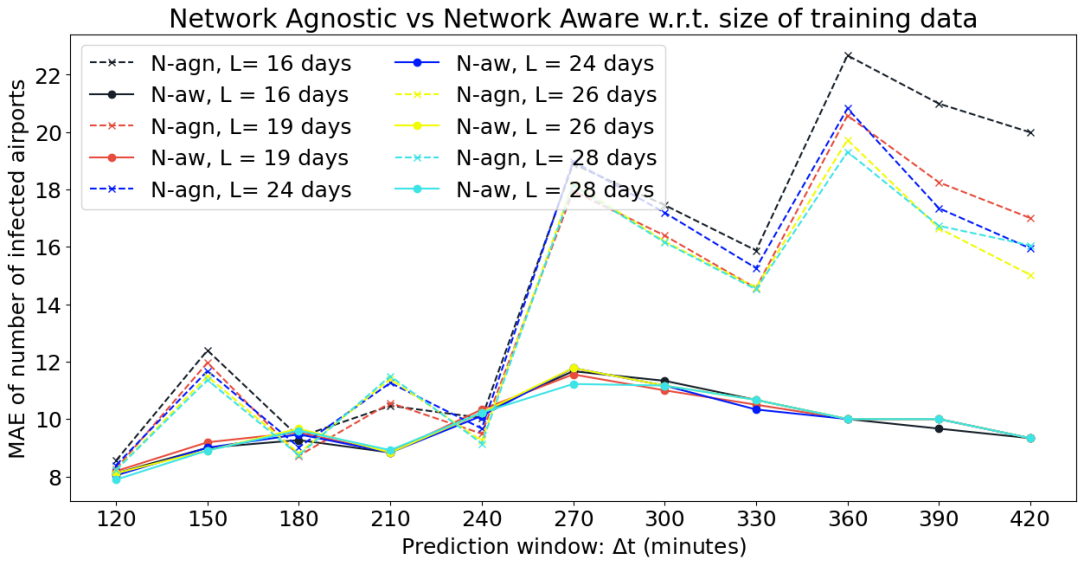


Figure 9: Short term prediction error w.r.t larger training data corpus and longer prediction window lengths. Averaged results over test days where network-aware performed better on average.

In order to more clearly understand the cause of this unexpected behavior, the results of each test day have been individually observed. Upon inspection, two scenarios have been identified: either the network-aware model clearly outperforms its counterpart or it performs marginally worse.

Figure 9 presents the averaged results over the days in which the network-aware model performed, on average, better than the network-agnostic model. The results confirm some of the previous theoretical assumptions:

- The network-aware model is capable of outperforming the agnostic counterpart when using longer prediction windows, Δt;
- The variation in the amount of training data has an insignificant effect in performance

Over the span of all Δt values, the network-aware approach behaved in a consistent manner, the results having little to no variation when considering both the L and Δt hyperparameters. On the other hand the network-agnostic approach began to struggle heavily after the $\Delta t = 240$ threshold.

Figure 10 presents the averaged results over the days in which the network-aware model performed either similar or worse than the network-agnostic model. This time, both models showed significant fluctuation in performance with the increase of Δt. This suggests that something happens during the four days covered by figure 10 that leads to the inconclusive results observed in Figure 8.

In order to fully determine the reason for which the actual results do not correlate with the theoretical intuition, the context of the test days should be individually investigated. Figure 11 presents the total departure delay throughout each test day. The delay was aggregated over time windows of length equal to 180 minutes. The horizontal axis presents the time window, while the vertical axis displays the delay measured in minutes. Each test day is numbered with respect to the first day in the data set, i.e. May 1st 2018. The test days in which network-aware was the better performing model are displayed with solid lines, while the days favored

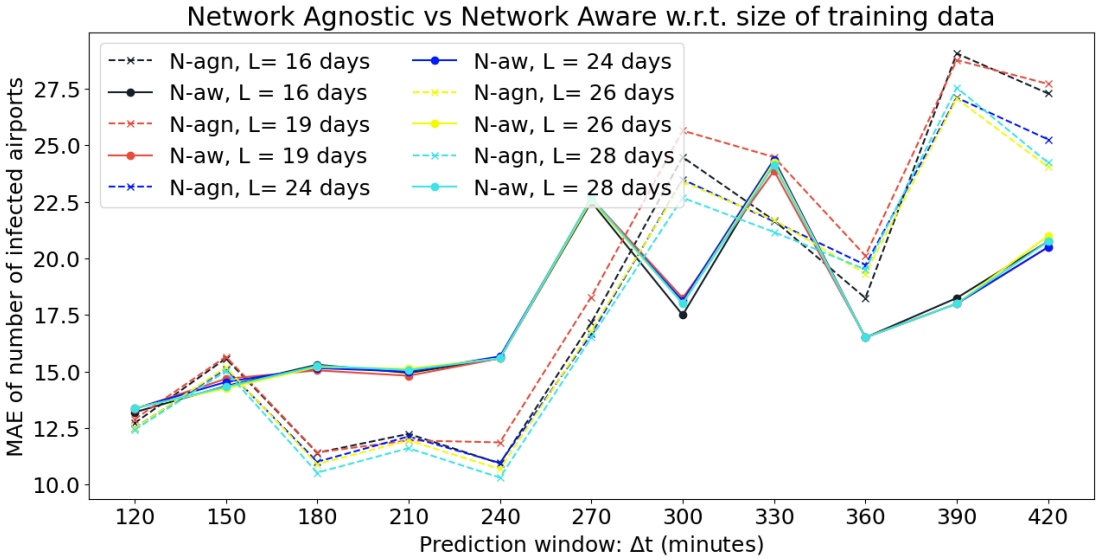


Figure 10: Short term prediction error w.r.t larger training data corpus and longer prediction window lengths, Averaged results over the poorly performing test days

by network-agnostic are displayed in dashed lines. The trend is immediately visible. In the

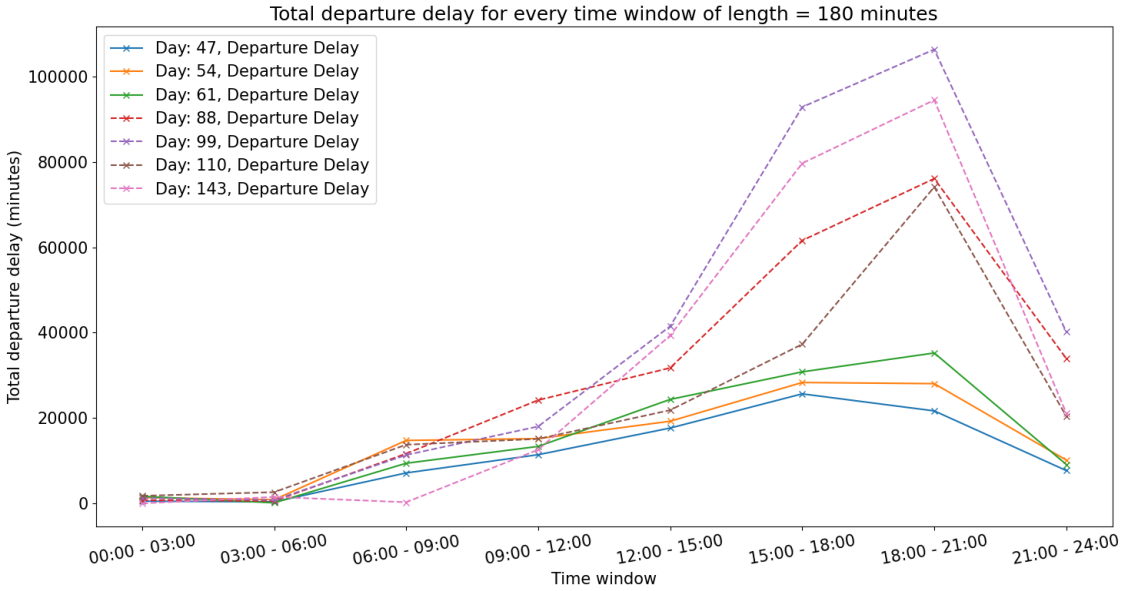


Figure 11: Departure delay throughout the test days aggregating over windows of length equal to 180 minutes

test days when network-aware performed better, the departure delays experienced in the network have a somewhat typical profile: we start with few delays that build up throughout the day, dying down in the later hours of the day, with no significant fluctuations from one time window to another. On the other hand, in all of the test days when network-agnostic performed similarly or better, we observe major increases in total delay between 15:00 and 21:00. Although the same conditions apply to both models, network-aware seems more affected by them. This suggests that network-agnostic is less sensitive to unexpected major disturbances in the network. This behavior might be explained by the fact that network-agnostic computes it's model-specific coefficients based on the entire network, not taking into account individual connections and thus, the contribution of such outliers is diminished when averaging over all airports.

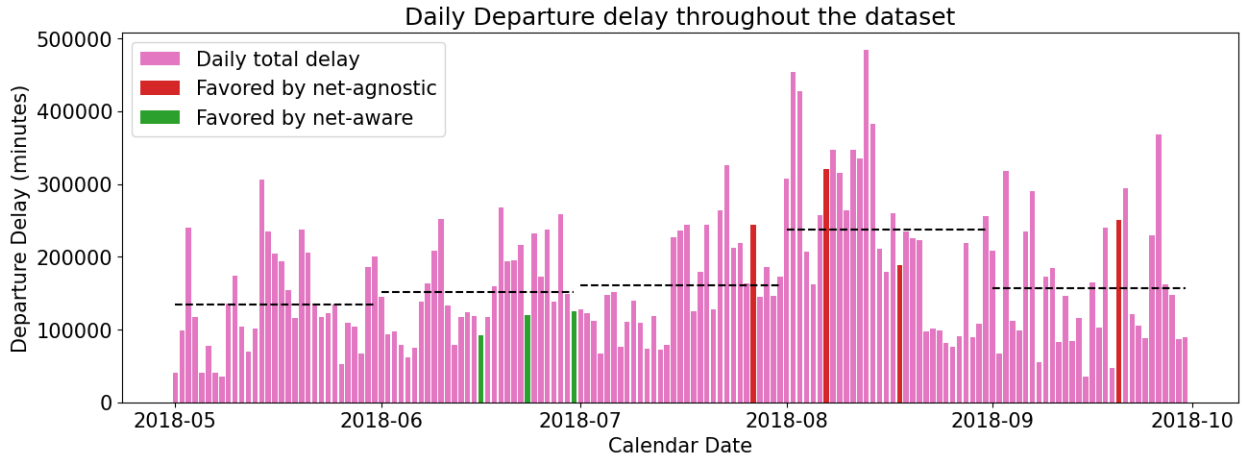


Figure 12: Daily total departure delay throughout the data set, highlighting the monthly averages and the test days.

Comparing the test days with the other days in the data set might provide another insightful perspective. Figure 12 presents the total daily departure delay throughout the data set. The horizontal axis displays the calendar days, while the vertical axis displays the total departure delay in minutes. The monthly averages are displayed using dashed horizontal threshold lines. The days favored by network-agnostic are displayed in red, while the days favored by network-aware are displayed in green.

Immediately visible is the distribution of the test days: all of the days in which network-agnostic outperformed network-aware are during or after the busiest periods of the data set. Moreover, when comparing the monthly average with the total delay experienced in the test days, apart from one test day, we can observe another trend: the days favored by network-aware have experienced delays below average, while the days favored by network-agnostic have experienced considerably above average delays. Another important observation is regarding the training data used. Each model was trained on the previous L days. If the training days used are not similar enough, or heavily fluctuate with respect to the test day, the models won't be able to accurately predict delays. Taking into consideration the fact that this context is applicable to both models suggests that there are other factors contributing to the lack of sensitivity to outliers of the network-agnostic approach.

In order to clearly determine the reason behind the lack of sensitivity to major disruptions observed for network-agnostic, we will look at how both models make predictions for each individual test day. Figure 13 presents the predictions' over and under-shoot for each test day, color-coded to both models. The horizontal axis presents the Δt value used, while the vertical axis presents the prediction error, i.e. the number of misidentified infected airports. The positive y-values mean that the model predicted more infected airports than the actual number, i.e. overshoot, while the negative y-values mean that the model predicted less infected airports, i.e. undershoot. The results for network-agnostic are in blue, while the results for network-aware are in orange. The undershoot and overshoot regions are displayed with a red and, respectively, green background. The results have been averaged over the course of the testing day, for the specific Δt used.

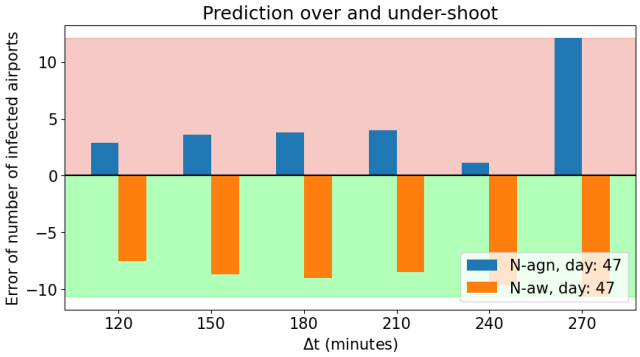
The results in Figure 13 show that network-agnostic overshot its predictions for all Δt values in five of the seven test days. In the remaining two test days, it overshot its predictions for

two, out of the six, Δt values tested.

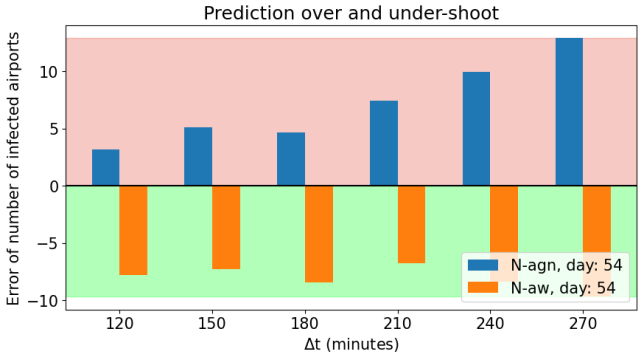
Table 3 presents the prediction overshoot values for network-agnostic, averaged over the seven test days. The results show that network-agnostic, on average, overshoots its predictions for every Δt value. This clearly indicates a tendency to predict more infections than there actually are. This predisposition explains why network-agnostic performs better in above average delay scenarios and performs worse in normal or below average delay scenarios. Because the model tends to predict more infections, in the test days with major unexpected delays, the prediction error is naturally smaller. Figure 13 also displays the consistency in prediction error of network-aware, across all Δt values. Nevertheless, with the exception of $\Delta t = 270$, network-agnostic generally performed better, achieving similar or lower prediction errors.

Δt	Mean overshoot
120	2.63
150	3.89
180	2.22
210	3.05
240	2.26
270	9.07

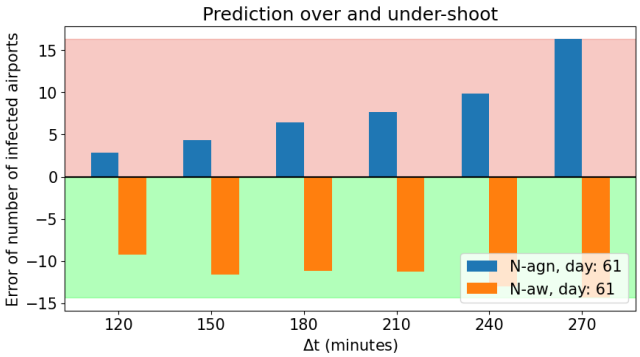
Table 3: Prediction overshoot values for network-agnostic, averaged over the 7 test days



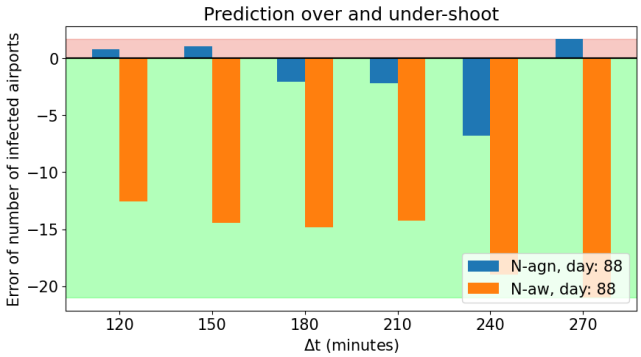
(a) Prediction over and under-shoot for test day: 47



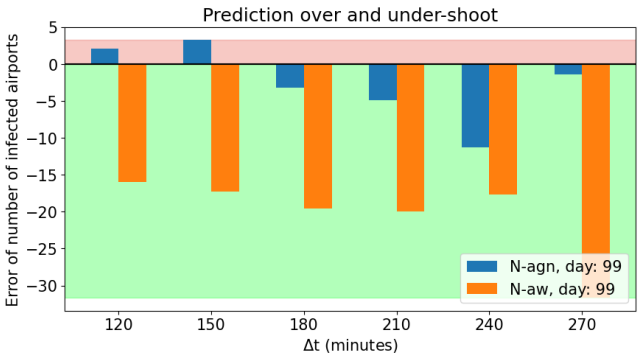
(b) Prediction over and under-shoot for test day: 54



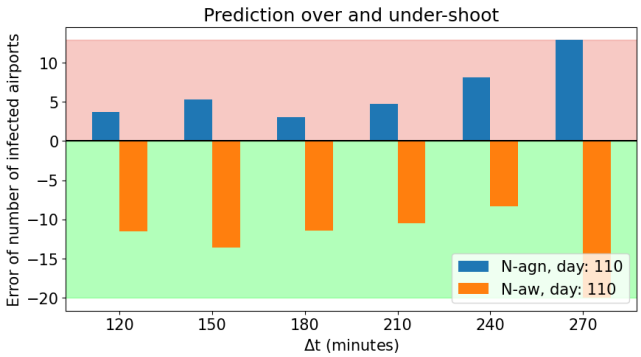
(c) Prediction over and under-shoot for test day: 61



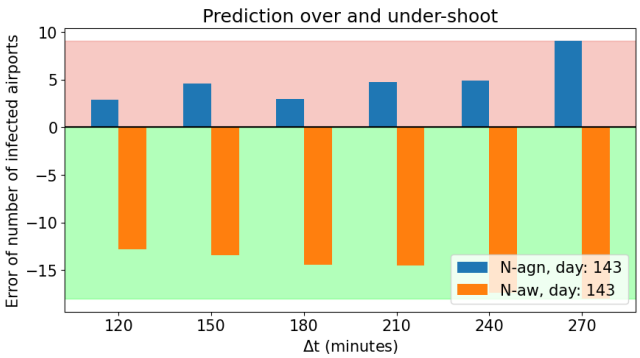
(d) Prediction over and under-shoot for test day: 88



(e) Prediction over and under-shoot for test day: 99



(f) Prediction over and under-shoot for test day: 110



(g) Prediction over and under-shoot for test day: 143

Figure 13: Prediction over and under-shoot for the network-agnostic and network-aware models, for each test day.

6.3 Long term prediction - Results

The long term prediction experiments serve the purpose of determining whether the models can be used for long-term planning. The idea is to cycle the output of the models back as input, in order to predict further into the future. Needless to say, the errors measured in each time step will propagate into future values, until the prediction is no longer valuable. The goal of these experiments is to determine whether the models can be used with reasonable accuracy up to a certain point in time.

From a practical standpoint, recycling the outputs of the model into inputs for new predictions indefinitely is not viable. Predicting using the more recent data will always lead to a lower prediction error and thus, a limit for the time steps ahead was chosen. Taking into account the duration of the prediction windows, the selected horizon limit has been set to $\Delta t+4$.

As in the short term prediction experiments, we will explore the capabilities of each model when varying the prediction window length, Δt, and the amount of training data, L:

- $\Delta t \in [60, 120]$ (minutes);
- $L \in [16, 19, 24, 26, 28]$ (days);

The size of the prediction window, Δt, was chosen to be on the medium side. Experimenting with values of $\Delta t = 120$ minutes, when using a horizon of $\Delta t + 4$ already means predicting 8 hours into the future. Network-agnostic and network-aware are meant to be used for obtaining information for short-term planning. As such, predicting further into the future, with longer prediction windows only invites more uncertainty. From a practical standpoint, Δt values higher than 120 minutes are not particularly relevant to long-term predictions. For a similar reason, smaller values are also less relevant because their $\Delta t+4$ predictions are generally covered by single predictions performed with medium Δt values.

Figure 14 displays the results for all L_i values and $\Delta t = 60$.

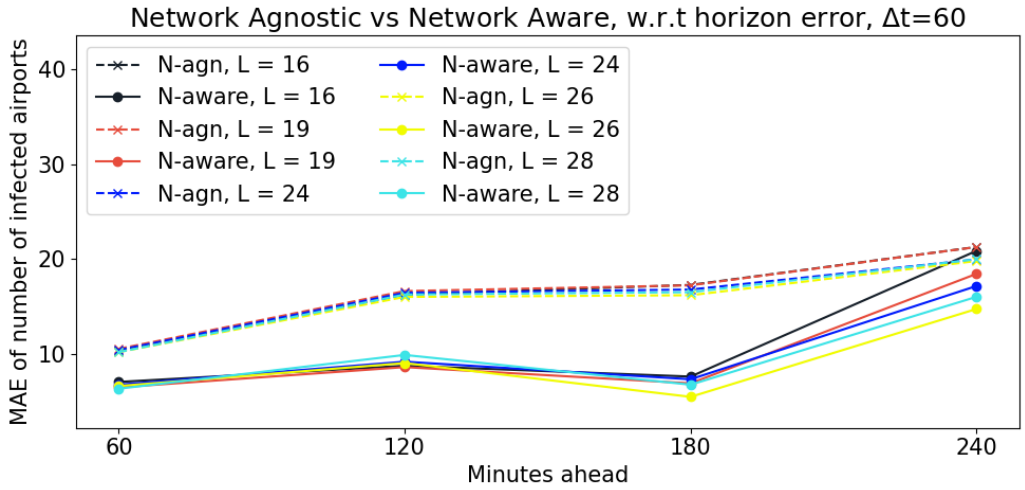


Figure 14: Long term prediction error w.r.t L, when $\Delta t = 60$ and the horizon is $\Delta t + 4$

While both approaches manage to maintain a relatively small increase in overall error, the network-aware model (continuous lines) outperforms the network-agnostic model (dashed lines) on all accounts by a reasonable margin. To put into perspective, at the time steps $\Delta t+2$, $\Delta t+3$ and $\Delta t + 4$, the errors observed when using the network-agnostic approach are comparable to

the short-term prediction errors when using $\Delta t \geq 270$ in figure 8. On the other hand, the network-aware approach maintains errors similar to the ones observed in the short-term prediction experiments up until the $\Delta t + 3$ mark.

Similarly to the short-term prediction experiments, we observe similar results for all L values, up until the $\Delta t + 4$ point. This further confirms the previous assumption that the amount of training data used is not a good indicator of performance or accuracy.

Figure 15 displays the results for all L_i values and $\Delta t = 120$. This time however, the situation is somewhat reversed. Even though the network-agnostic approach maintains a slightly smaller error until the $\Delta t + 3$ point, both models can be said to have a poor performance. For both approaches, the errors after the $\Delta t + 2$ point are not similar to the short-term prediction errors.

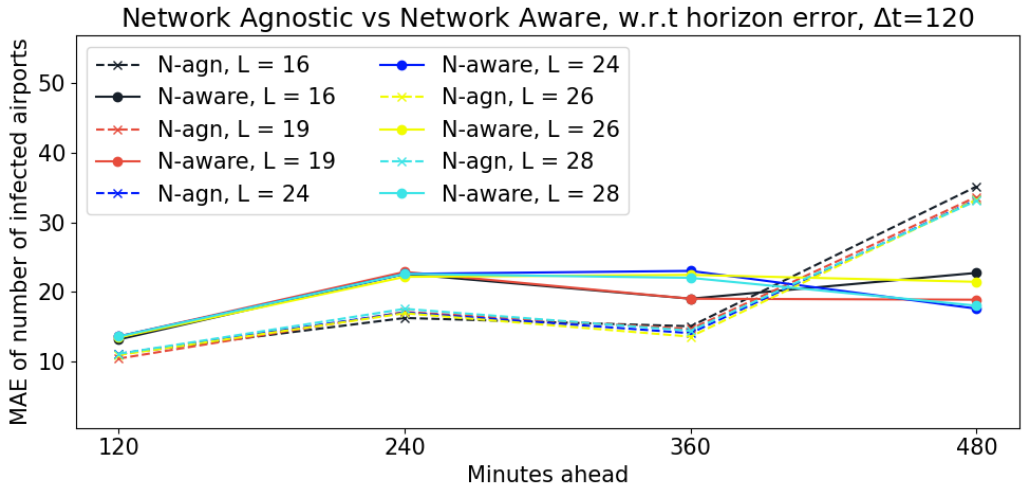


Figure 15: Long term prediction error w.r.t L, when $\Delta t = 120$ and the horizon is $\Delta t + 4$

7 Conclusion

The main objective of this research is to assess the efficiency and compare two approaches that attempt to solve the task of flight delay propagation within a real world airport network. Both models, network-agnostic and network-aware, make use of epidemiological models as their backbone. The two models represent fundamentally different approaches: the agnostic approach is not concerned with individual airport-to-airport connections while the aware variant does incorporate this information. Coupling the variety of reasons for which delays form with the high complexity and dynamic nature of the network itself builds a solid case for considering flight delays as stochastic events. Therefore, a comparison between the two approaches with regards to how well they manage to solve the problem of predicting extensive airport delays is well founded. As such, short-term and long-term prediction experiments have been performed.

The short-term experiments consisted in applying the delay propagation models to the task of predicting the number of infected airports in the network at the next time step. The hyperparameters being tested were the size of the prediction window, Δt, and the size of the training data used to infer the model-specific coefficients, L. The variation in the amount of training data produced no significant effects on the overall performance of the two models. For time windows shorter than 60 minutes, network-agnostic produced more accurate predictions. The theoretical intuition pointed to network-aware as being the favorite when using larger prediction windows. This hypothesis was tested and although network-aware performed marginally better and was more stable, the evidence was not strong enough as to correlate the results with the theoretical hypothesis. A robustness investigation was performed to identify the causes of the fluctuation in performance. The results show that network-aware always predicts less infected airports than the actual number while network-agnostic predicts, in the majority of investigated cases, more infections. This explains the surprising results obtained for the longer prediction windows.

The long-term experiments consisted in recycling the output predictions as inputs for the two models in order to predict the situation in the network further into the future. The same hyperparameters were tested and the results were similar in nature to the short-term experiments. The variation in the amount of training data used did not affect the prediction quality by significant margins. Network-aware produced a more consistent prediction error for all prediction window values, however, it did not always return the lower errors. Nevertheless, the best results were obtained when using a prediction window of size 60 minutes and the network-aware approach. The errors were comparable to the ones obtained in the short-term experiments, where the outputs were not cycled back as inputs and the most recent data was used.

The two models have been analyzed taking into account a variety of settings: prediction window size, amount of training data, tendency of overshoot or undershoot, reusability of predictions and behavior with respect to unexpected disturbances in the network. Contrary to the theoretical intuition, the network-agnostic approach performs better and is more adaptable. The lack of periodicity in delays causes the network-aware model to have difficulty in generalizing the propagation of delays. On the other hand, the network-agnostic approach is less affected by this, since it aggregates the entire network to an average behavior.

The models are not without limitations. While simplicity and intuitiveness are generally positive aspects, they can also be regarded as factors contributing to the errors observed. Raising the complexity of the models by factoring in multiple avenues of information might lead to lower errors while still maintaining a reasonable computing time and applicability. Another

visible limitation is regarding the number of test days experimented on. A higher number of test days could have painted a more complete picture with respect to the capabilities of each model. Experimenting with the two constraints that define whether an airport is infected or not could have also shown hidden capabilities or preferences of each model. Addressing these limitations can be the subject of future research done in this field.

References

[1] N. R. C. U. T. R. B. C. for a Study of Public-Sector Requirements for a Small Aircraft Transportation System, *Future Flight: A Review of the Small Aircraft Transportation System Concept*. Special report (National Research Council (U.S.). Transportation Research Board)), 2002.

[2] R. Beatty, R. Hsu, and J. Rome, "Preliminary evaluation of flight delay propagation through an airline schedule," *Air Traffic Control Quarterly*, vol. 7, 1998.

[3] F. Petropoulos, D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. Ben Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, J. Browell, C. Carnevale, J. L. Castle, P. Cirillo, M. P. Clements, C. Cordeiro, F. L. Cyrino Oliveira, S. De Baets, A. Dokumentov, J. Ellison, P. Fiszeder, P. H. Franses, D. T. Frazier, M. Gilliland, M. S. Gönül, P. Goodwin, L. Grossi, Y. Grushka-Cockayne, M. Guidolin, M. Guidolin, U. Gunter, X. Guo, R. Guseo, N. Harvey, D. F. Hendry, R. Hollyman, T. Januschowski, J. Jeon, V. R. R. Jose, Y. Kang, A. B. Koehler, S. Kolassa, N. Kourentzes, S. Leva, F. Li, K. Litsiou, S. Makridakis, G. M. Martin, A. B. Martinez, S. Meeran, T. Modis, K. Nikolopoulos, D. Önkal, A. Paccagnini, A. Panagiotelis, I. Panapakidis, J. M. Pavía, M. Pedio, D. J. Pedregal, P. Pinson, P. Ramos, D. E. Rapach, J. J. Reade, B. Rostami-Tabar, M. Rubaszek, G. Sermpinis, H. L. Shang, E. Spiliotis, A. A. Syntetos, P. D. Talagala, T. S. Talagala, L. Tashman, D. Thomakos, T. Thorarinsdottir, E. Todini, J. R. Trapero Arenas, X. Wang, R. L. Winkler, A. Yusupova, and F. Ziel, "Forecasting: theory and practice," *International Journal of Forecasting*, vol. 38, no. 3, 2022.

[4] S. Li, D. Xie, X. Zhang, Z. Zhang, and W. Bai, "Data-driven modeling of systemic air traffic delay propagation: An epidemic model approach," *Journal of Advanced Transportation*, vol. 2020, pp. 1–12, 2020.

[5] B. Baspinar and E. Koyuncu, "A data-driven air transportation delay propagation model using epidemic process models," *International Journal of Aerospace Engineering*, vol. 2016, pp. 1–11, 2016.

[6] N. Pyrgiotis, K. M. Malone, and A. Odoni, "Modelling delay propagation within an airport network," *Transportation Research Part C: Emerging Technologies*, vol. 27, pp. 60–75, 2013.

[7] Q. Wu, M. Hu, X. Ma, Y. Wang, W. Cong, and D. Delahaye, "Modeling flight delay propagation in airport and airspace network," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3556–3561, 2018.

[8] Y. Lin, L. Li, P. Ren, Y. Wang, and W. Szeto, "From aircraft tracking data to network delay model: A data-driven approach considering en-route congestion," *Transportation Research Part C: Emerging Technologies*, vol. 131, p. 103329, 2021.

[9] Q. Cai, S. Alam, and V. N. Duong, "A spatial–temporal network perspective for the propagation dynamics of air traffic delays," *Engineering*, vol. 7, no. 4, pp. 452–464, 2021.

[10] D. K. Arnett and S. A. Claas, "Chapter 35 - introduction to epidemiology," in *Clinical and Translational Science*, pp. 527–541, 2009.

[11] H. W. Hethcote, *Three Basic Epidemiological Models*, pp. 119–144. Springer Berlin Heidelberg, 1989.

[12] H. Zhang, W. Wu, S. Zhang, and F. Witlox, "Simulation analysis on flight delay propagation under different network configurations," *IEEE Access*, vol. 8, pp. 103236–103244, 2020.

[13] "Aviation System Performance Metrics (ASPM)." `https://aspmhelp.faa.gov/index/Aviation_System_Performance_Metrics_(ASPM).html`. [Accessed 12-06-2023].

[14] "The Week in Transportation — Bureau of Transportation Statistics." `https://www.bts.gov/covid-19/week-in-transportation`. [Accessed 12-06-2023].

[15] D. F. Griffiths and D. J. Higham, *Euler's Method*, pp. 19–31. London: Springer London, 2010.